Glossaire structuré – IA embarquée

Auteurs

Eliane Fourgeau	eliane.fourgeau@quantyss.com
Pierre Gaillard	pierre.gaillard@cea.fr
Bertrand Braunschweig	bertrand.braunschweig@irt-systemx.fr
Frédéric Cetlin	frederic.cetlin@leminnov.com

En collaboration avec :

CEA List - https://list.cea.fr/

Leminnov - www.leminnov.com

DeepGreen - www.deepgreen.ai

European Trustworthy AI Association - www.trustworthy-ai-association.eu

Version 1.0 du 29 Janvier 2026

Ce glossaire constitue un travail de documentation et de clarification de la terminologie associée à l'intelligence artificielle embarquée. Il a pour objectif de proposer un cadre de référence pédagogique destiné à faciliter la compréhension, le partage et l'appropriation des concepts, sans prétendre à l'exhaustivité.

Le contenu de ce document ne saurait engager la responsabilité des auteurs, ni refléter les positions officielles, stratégiques ou normatives des entreprises, institutions ou organisations ayant contribué à son élaboration. Il ne constitue ni un document de normalisation, ni un standard, ni une recommandation formelle, mais un outil d'aide à la lecture et à la structuration des connaissances dans un domaine en évolution rapide.

La rédaction de ce glossaire s'appuie sur une analyse croisée de sources variées, incluant notamment la taxinomie de Confiance.ai, des ressources encyclopédiques (Wikipédia en français et en anglais), des textes de réglementation européenne relatifs à l'IA, ainsi que des articles et publications techniques, pédagogiques et scientifiques.

Pour citer ce document: "Embedded France, Glossaire structuré IA embarquée version 1.0, Janvier 2026"

Table des Matières

1 INTRODUCTION GÉNÉRALE

Une initiative structurante pour la filière

Ce glossaire est une initiative du groupe de travail “IA Embarquée” d’Embedded France, portée par des experts issus d’organisations industrielles, académiques et institutionnelles positionnées au premier plan de la discipline. Il reflète une volonté collective de structurer un langage commun autour des technologies d’intelligence artificielle embarquée et s’inscrit dans une logique de transparence et de diffusion ouverte, afin de favoriser l’appropriation collective des concepts.

Une ambition claire, une posture humble

Ce document est avant tout un référentiel taxonomique, destiné à définir les termes essentiels qui structurent les pratiques, les architectures et les usages de l’IA embarquée. Il ne prétend pas à l’exhaustivité ni à la normativité, mais propose une base rigoureuse, contextualisée et opérationnelle, ouverte à l’enrichissement collectif.

Nous souhaitons qu’il reste un document vivant, évolutif, et accessible à tous les acteurs de la filière. C’est pourquoi nous le mettons en libre accès, dans une logique de transparence, de partage et de structuration ouverte.

Une exploration dans le temps : passé, présent, futur

Ce glossaire explore l’intelligence artificielle embarquée dans sa profondeur historique, ses usages actuels et ses trajectoires futures - depuis les systèmes experts et les architectures temps réel jusqu’aux modèles frugaux, distribués et génératifs, porteurs d’autonomie et de souveraineté technologique.

Conscient de la **rapidité et de la multiplicité des évolutions** dans ce domaine, le glossaire sera revu et mis à jour à intervalles réguliers, pour rester en phase avec les avancées technologiques et les besoins de la filière.

Une valeur intrinsèque fondée sur l’usage

Ce document ne se limite pas à une compilation de définitions:

- Il articule les concepts et architectures techniques avec leurs implications concrètes dans les systèmes embarqués, distribués et temps réel,
- Il décrit les composants matériels (processeurs, capteurs, interconnexions) et les environnements logiciels (optimisation, exécution),
- Il définit les paradigmes d’apprentissage (supervisé, fédéré, par transfert…) et les formes d’inférence embarquée,
- Il structure les cas d’usage en vision, audio, langage naturel, séries temporelles et interfaces intelligentes,et intègre les notions avancées (l’IA générative, l’IA agentique,...) comme les comportements émergents (IA physique, jumeaux numérique,...).

En dialogue avec des initiatives existantes

Ce glossaire s'inscrit en complémentarité d'autres travaux (liste non exhaustive) :

- Le glossaire du programme Confiance.ai (IRT SystemX), focalisé sur l'IA de confiance dans les systèmes critiques,
- Le référentiel terminologique de l'AI Act (Commission européenne), qui encadre les obligations réglementaires,
- Les fiches ISO/IEC JTC 1/SC 42, qui définissent les concepts standards de l'IA, y compris embarquée.

Il va plus loin :

- En ciblant **l'IA embarquée dans ses dimensions matérielles, logicielles, algorithmiques et applicatives**, avec une attention particulière portée à la frugalité, à l'explicabilité et à l'autonomie distribuée.
- En explorant l'intelligence artificielle embarquée **dans sa profondeur historique, ses usages actuels et ses trajectoires futures**.

En résumé, ce glossaire vise à proposer une **cartographie claire, opérationnelle et contextualisée** des notions fondamentales de l'IA embarquée, telles qu'elles sont mobilisées dans les projets industriels, académiques et institutionnels ; il poursuit plusieurs **objectifs complémentaires** :

- **Structurer un langage commun** : établir des définitions claires et partagées des concepts, architectures, modèles et outils propres à l'IA embarquée.
- **Faciliter la compréhension et l'appropriation** : offrir aux acteurs de la filière un référentiel accessible, qui relie les notions techniques à leurs usages concrets dans les systèmes embarqués.
- **Renforcer la crédibilité et la cohérence disciplinaire** : proposer une taxonomie rigoureuse qui consolide la place de l'IA embarquée dans le paysage technologique et académique.
- **Accompagner l'innovation et la souveraineté technologique** : fournir un cadre de référence utile aux projets industriels, aux travaux de recherche et aux initiatives institutionnelles.
- **Favoriser l'évolution collective** : permettre aux membres de la communauté d'enrichir et d'actualiser le glossaire, afin qu'il reste un outil pertinent et durable.

Note méthodologique sur les frontières taxonomiques

Il convient de préciser que les frontières entre intelligence artificielle (IA), IA embarquée et systèmes embarqués demeurent parfois diffuses ; les terminologies fondamentales de ces domaines se chevauchent en raison de leur histoire commune, de leurs interdépendances techniques et de l'évolution rapide des usages. Dans ce glossaire, certains termes repris ne sont pas strictement réservés à l'IA embarquée. Ils relèvent plus largement de l'IA ou de l'embarqué au sens général. Leur inclusion répond à une double nécessité :

- Assurer la cohérence et la continuité du vocabulaire: de nombreux concepts génériques constituent des briques indispensables pour comprendre et déployer l'IA dans des environnements embarqués.
- Faciliter l'appropriation par la communauté: en intégrant des définitions partagées, le glossaire permet aux acteurs de la filière de disposer d'un référentiel commun, évitant les ambiguïtés et renforçant la lisibilité des échanges.

Ce choix méthodologique reflète la volonté de proposer un document opérationnel et inclusif, qui articule les spécificités de l'IA embarquée avec les fondations plus larges de l'IA et des systèmes embarqués.

GRILLE DE LECTURE

La **grille de lecture** ci-après offre une vue d'ensemble du glossaire en présentant, pour chaque chapitre, les thèmes abordés, leur rôle dans le document et les liens qu'ils entretiennent avec les autres chapitres. Elle vise à guider le lecteur dans la compréhension progressive des notions clés de l'intelligence artificielle embarquée et à faciliter la navigation entre les différentes parties du glossaire.

Chapitre	Rôle dans le glossaire	Contenu principal couvert	Articulation avec les autres chapitres
1. Introduction générale	Définir le cadre, le périmètre et les objectifs du glossaire	Périmètre de l'IA embarquée, logique du document, méthode de travail	Fournit le cadre de lecture commun à l'ensemble des chapitres (2 à 13)
2. Cadre et concepts des données pour l'IA embarquée	Poser les fondations liées aux données dans l'IA embarquée	Données, jeux de données, qualité, flux, traçabilité	Constitue la base conceptuelle pour les chapitres traitant des modèles, de l'évaluation et du cycle de vie (4, 5, 9, 11)
3. Introduction à l'intelligence artificielle pour l'embarqué	Introduire les fondements de l'IA adaptés aux systèmes embarqués	Principes de l'IA, types d'apprentissage, inférence sous contraintes	Sert de socle conceptuel aux chapitres consacrés aux méthodes, architectures et paradigmes avancés (4, 6, 8)
4. Méthodes et modèles d'IA pour systèmes embarqués	Décrire les approches algorithmiques exploitables en embarqué	Modèles, méthodes d'apprentissage, inférence et adaptation	Prolonge les fondements de l'IA (3) et alimente les chapitres relatifs à l'exécution et à l'évaluation (6, 9, 11)
5. Données, acquisition et traitements pour l'IA embarquée	Décrire la chaîne de données depuis le capteur jusqu'au modèle	Capteurs, acquisition, prétraitement, structuration des flux	Complète le cadre des données (2) et s'articule avec les méthodes et architectures (4, 6)

Chapitre	Rôle dans le glossaire	Contenu principal couvert	Articulation avec les autres chapitres
6. Architectures matérielles et logicielles	Présenter les supports d'exécution de l'IA embarquée	Processeurs, accélérateurs, systèmes embarqués, OS	Supporte la mise en œuvre des modèles et outils (4, 7), et conditionne l'exploitation (11)
7. Outils, frameworks et environnements	Décrire les outils de développement et d'exécution	Frameworks, bibliothèques, toolchains embarquées	Met en œuvre concrètement les méthodes et architectures décrites précédemment (4, 6)
8. Paradigmes avancés et émergents de l'IA embarquée	Ouvrir sur les évolutions et approches non classiques	IA générative embarquée, agentique, IA physique, frugalité	Complète les fondements de l'IA (3) sans redéfinir les concepts, et éclaire les usages futurs (13)
9. Performances, métriques et évaluation	Qualifier et mesurer le comportement de l'IA embarquée	Latence, précision, consommation, robustesse	Apporte des critères transverses applicables aux modèles et architectures (4, 6), et alimente la gouvernance (11)
10. Sûreté, robustesse et confiance	Encadrer la fiabilité et la maîtrise des systèmes	Robustesse, sûreté de fonctionnement, confiance	Prolonge l'évaluation technique (9) par des notions de maîtrise et de fiabilité en exploitation (11)
11. Cycle de vie et gouvernance de l'IA embarquée	Structurer la conception, le déploiement et l'exploitation	Cycle de vie, validation, supervision, mises à jour	Fait le lien entre conception technique (4, 6), exploitation (9, 10) et conformité (12)

Chapitre	Rôle dans le glossaire	Contenu principal couvert	Articulation avec les autres chapitres
12. Réglementation, normes et cadres de référence	Présenter les cadres normatifs applicables	Réglementations européennes, normes et référentiels	Complète la gouvernance opérationnelle (11) en apportant le cadre réglementaire
13. Applications et cas d'usage de l'IA embarquée	Illustrer les usages concrets et sectoriels	Fonctions applicatives, secteurs, bénéfices et contraintes	Met en perspective l'ensemble des chapitres techniques à travers des usages concrets (2 à 12)

Les **règles de formulation** des définitions visent à garantir la cohérence, la lisibilité et l'usage opérationnel du glossaire. Chaque entrée présente un terme en français, accompagné de son équivalent en anglais lorsque cela est pertinent. Lorsque le terme est couramment désigné par un acronyme, celui-ci est précisé et intégré à la terminologie afin de faciliter la lecture et l'usage du glossaire. Dans certains cas, le terme anglais est privilégié lorsqu'il constitue la référence dominante dans la littérature scientifique, technique ou industrielle.

Les **définitions** sont formulées de manière concise et contextualisée, avec un lien explicite avec l'IA embarquée dès lors que le concept s'y applique. Lorsque cela apporte une meilleure compréhension, des précisions ou des exemples d'usage sont ajoutés. Les définitions issues de référentiels reconnus (tels que Confiance.ai, ISO/IEC ou cadres équivalents) sont reprises et adaptées afin d'assurer une continuité terminologique et une cohérence avec les pratiques existantes.

Les **notions** sont définies de manière non redondante au sein du glossaire : un terme est formalisé dans un chapitre de référence et peut être mobilisé dans d'autres chapitres sans être redéfini. L'ensemble des définitions adopte une approche descriptive et opérationnelle, sans caractère prescriptif ou commercial.

Certaines **traductions** ont été réalisées avec l'appui de l'outil DeepL. Par ailleurs, des outils d'intelligence artificielle générative (tels que Perplexity, ChatGPT, Copilot et Mistral) ont été utilisés comme aides à la recherche, à la structuration et à la reformulation des contenus, sous le contrôle et la validation des rédacteurs.

Cette **approche éditoriale** permet d'aborder le glossaire comme un ensemble structuré et cohérent, facilitant la compréhension des notions et leur mise en relation, depuis les fondations techniques jusqu'aux enjeux applicatifs et de gouvernance de l'IA embarquée.

2 Cadre, acquisition et traitement des données

Dans un système d'IA embarquée, les données, leur acquisition et leur traitement ne peuvent être considérés comme des éléments indépendants. Ils forment un ensemble cohérent de fonctions qui conditionnent la capacité du système à percevoir son environnement, à produire des décisions fiables et à fonctionner de manière autonome sous contraintes.

Les données constituent la matière première de l'IA embarquée. Leur nature, leur qualité et leur représentation déterminent ce que le système est en mesure de percevoir et d'interpréter. L'acquisition permet de relier ces données au monde réel, en assurant une production continue ou événementielle d'informations adaptées aux contextes d'usage. Le traitement des données, enfin, joue un rôle d'adaptation et de stabilisation, en transformant ces informations en entrées compatibles avec les modèles déployés sur des plateformes embarquées contraintes.

Cette articulation est spécifique à l'IA embarquée, car elle s'inscrit dans un environnement où les ressources sont limitées, les conditions d'exploitation variables et les exigences de temps réel et de sûreté élevées. Les choix effectués à chaque niveau influencent directement les autres, et l'équilibre entre ces trois dimensions conditionne la performance globale du système.

2.1 Cadre des Données

Ce chapitre regroupe la terminologie permettant de définir ce qu'est une donnée exploitable dans un système d'IA embarquée. Il précise les notions nécessaires pour caractériser la nature des données, leur structuration, leur représentation et leur usage dans le cadre de modèles déployés sur des plateformes contraintes, c'est-à-dire des dispositifs embarqués disposant de ressources limitées en calcul, mémoire, stockage et énergie, et soumis à des exigences de temps réel et de robustesse.

La compréhension de ces termes est essentielle pour qualifier la pertinence et la fiabilité des données utilisées par l'IA embarquée. Une définition claire des concepts liés aux données permet d'éviter des interprétations ambiguës et de garantir que les modèles reposent sur des informations adaptées aux contraintes d'exécution locale.

Ingénierie des données – Data Engineering

Discipline visant à organiser, structurer, sélectionner, transformer et tracer les données utilisées par des systèmes d'IA embarquée, de manière à garantir leur qualité, leur disponibilité, leur pertinence opérationnelle et leur traçabilité tout au long du cycle de vie du système, sous contraintes de temps réel, de ressources matérielles et de sûreté.

Donnée – Data

Information brute issue de l'observation du monde physique ou logique, principalement produite par des capteurs embarqués, et utilisée comme entrée des traitements et des modèles d'IA embarquée pour l'inférence locale et la prise de décision.

Donnée numérique – Digital Data

Représentation codée, généralement binaire, d'une information physique ou logique, destinée à être traitée par un système embarqué exécutant des algorithmes d'IA, dans des conditions de calcul, de mémoire et de latence contraintes.

Jeu de données – Dataset

Ensemble structuré de données utilisé pour l'entraînement, la validation ou le test de modèles d'IA, généralement constitué en amont du déploiement et servant de référence pour des modèles destinés à être exécutés sur des systèmes embarqués.

Échantillon – Sample

Instance individuelle d'un jeu de données représentant une observation unique (mesure capteur, image, signal, séquence), utilisée comme entrée d'un modèle d'IA embarquée lors de l'entraînement ou de l'inférence.

Caractéristique – Feature

Variable mesurable ou attribut dérivé d'un échantillon, décrivant un aspect pertinent de la donnée et servant de représentation d'entrée pour un modèle d'IA embarquée, souvent extraite ou calculée localement sous contraintes de ressources.

Espace de caractéristiques – Feature Space

Ensemble des caractéristiques utilisées pour représenter les données dans un modèle d'IA embarquée, définissant la dimension et la structure de l'entrée du modèle, avec un impact direct sur la complexité, la performance et le coût d'inférence.

Données de test – Test Data

Ensemble de données distinct utilisé pour mesurer de manière objective les performances finales d'un modèle d'IA, avant son intégration et son exploitation dans un système embarqué.

Étiquette – Label

Valeur, classe ou catégorie associée à un échantillon dans le cadre d'un apprentissage supervisé, servant de référence pour l'entraînement et l'évaluation de modèles d'IA embarquée.

Échantillonnage de données – Data Sampling

Processus de sélection d'un sous-ensemble représentatif de données à partir d'un ensemble plus large, utilisé pour limiter les volumes, équilibrer les classes ou adapter les données aux contraintes d'entraînement et de déploiement embarqué.

Données d'entraînement – Training Data

Sous-ensemble d'un jeu de données utilisé pour ajuster les paramètres d'un modèle d'IA, généralement exploité hors ligne avant le déploiement du modèle sur une plateforme embarquée.

Données de validation – Validation Data

Sous-ensemble de données utilisé pendant la phase d'entraînement pour évaluer les performances intermédiaires d'un modèle, ajuster les hyperparamètres et éviter le surapprentissage avant déploiement embarqué.

Fiche de données – Data Card or Datasheet for Datasets

Document décrivant de manière synthétique et normalisée les caractéristiques d'un jeu de données (origine, composition, modalités de collecte, usages prévus, limites, biais identifiés), afin de garantir la transparence, la traçabilité et la conformité des modèles d'IA embarquée qui en sont issus.

Flux de données – Data Stream

Suite continue ou événementielle de données produites par des capteurs embarqués et traitées localement par des algorithmes d'IA, souvent sous contraintes de latence, de bande passante et de consommation énergétique.

Données en temps réel – Real-Time Data

Données collectées et traitées dans des délais strictement bornés afin de permettre une réaction immédiate ou quasi immédiate d'un système d'IA embarquée, notamment dans des applications critiques ou autonomes.

Qualité des données – Data Quality

Ensemble des propriétés (précision, cohérence, complétude, fraîcheur, bruit) conditionnant la fiabilité des inférences produites par un système d'IA embarquée et sa capacité à fonctionner de manière sûre et robuste en conditions réelles.

Données manquantes ou partielles – Missing or Incomplete Data

Situation dans laquelle certaines informations attendues ne sont pas disponibles ou sont incomplètes, nécessitant des mécanismes de gestion adaptés afin de préserver la stabilité et la sûreté de l'IA embarquée.

Traçabilité des données embarquées - Embedded Data Traceability

Capacité à retracer l'origine, les transformations, les versions et les usages des données au sein d'un système d'IA embarquée, en support des exigences de sûreté, de certification, d'auditabilité et de conformité réglementaire.

Sécurité des données embarquées – Embedded Data Security

Ensemble des mécanismes techniques visant à protéger les données utilisées par l'IA embarquée contre les accès non autorisés, les altérations et les pertes, tout en respectant les contraintes de calcul et d'énergie du système.

Annotation de données – Data Annotation

Processus consistant à associer des informations descriptives (étiquettes, catégories, métadonnées) à des données, généralement réalisé en amont du déploiement embarqué, afin de permettre l'entraînement et l'évaluation de modèles d'IA supervisés destinés à l'inférence locale.

Augmentation de données – Data Augmentation

Technique consistant à générer artificiellement de nouveaux échantillons à partir de données existantes (transformations géométriques, bruit, variations de signal), afin d'améliorer la robustesse et la généralisation de modèles d'IA destinés à l'exécution embarquée.

2.2 Acquisition des données

Ce chapitre rassemble la terminologie relative aux dispositifs et mécanismes permettant de produire les données utilisées par l'IA embarquée, en décrivant les capteurs et les notions associées à leur fonctionnement dans des systèmes embarqués exposés à des environnements variables.

Les notions présentées permettent de comprendre comment les données sont adaptées aux exigences d'exécution de l'IA embarquée, tant en termes de format, de représentation que de robustesse face aux perturbations ou aux variations des signaux. Le traitement des données joue ainsi un rôle essentiel dans la stabilité et la prévisibilité du comportement du système en conditions opérationnelles.

Capteur – Sensor

Dispositif matériel embarqué mesurant une grandeur physique (température, pression, lumière, son, mouvement, etc.) et la convertissant en un signal électrique ou numérique exploitable par des traitements et des modèles d'IA embarquée, généralement en temps réel et sous contraintes de ressources.

Caméra imageur – Imager

Capteur optique convertissant la lumière en signaux numériques afin de produire des images exploitables par des algorithmes de vision artificielle embarquée pour la perception, la détection ou la classification.

Caméra événementielle – Event-Based Camera

Capteur optique asynchrone qui ne capture pas des images à fréquence fixe, mais génère des événements uniquement lors de variations locales de luminosité sur les pixels, permettant une perception embarquée à très faible latence et consommation énergétique.

Capteur LiDAR – Light Detection and Ranging

Capteur actif émettant des impulsions laser afin de mesurer la distance à des objets environnants et de produire une représentation tridimensionnelle de l'environnement, généralement sous forme de nuage de points, utilisée pour la perception et la localisation en IA embarquée.

Capteur RADAR – Radio Detection and Ranging

Capteur utilisant des ondes radio pour détecter la présence, la distance, la vitesse et parfois la direction d'objets, particulièrement robuste aux conditions environnementales (pluie, brouillard), et largement utilisé pour la perception embarquée.

Capteurs inertiels – Inertial Measurement Unit (IMU)

Ensemble de capteurs comprenant généralement un accéléromètre, un gyroscope et parfois un magnétomètre, permettant de mesurer les mouvements linéaires et rotationnels, essentiels pour la navigation, la stabilisation et la fusion multi-capteurs en IA embarquée.

Capteurs de localisation – Geolocation Sensors

Capteurs ou systèmes permettant d'estimer la position ou la trajectoire d'un système embarqué (GNSS, UWB, balises locales), souvent combinés à des capteurs inertiels pour améliorer la robustesse de la perception et de la navigation embarquée.

Capteurs acoustiques – Audio Sensors

Capteurs capturant des signaux sonores ou vibratoires (microphones, capteurs ultrasoniques), utilisés en IA embarquée pour la reconnaissance vocale, la détection d'événements acoustiques, le diagnostic ou la surveillance.

Capteurs de proximité et de distance courte – Proximity and Short-Range Sensors

Capteurs mesurant la distance à courte portée (ultrasons, infrarouge, ToF courte portée), utilisés en robotique, objets connectés et systèmes embarqués pour la détection d'obstacles et l'interaction avec l'environnement proche.

Capteurs environnementaux – Environmental Sensors

Capteurs mesurant des paramètres physiques ou chimiques de l'environnement (température, humidité, pression, gaz, particules, vibrations), utilisés en IA embarquée pour la surveillance, l'analyse contextuelle et la détection d'anomalies.

Capteurs biométriques et physiologiques – Biometric and Physiological sensors

Capteurs mesurant des signaux biologiques ou physiologiques (fréquence cardiaque, activité musculaire ou cérébrale, saturation en oxygène), utilisés dans des systèmes d'IA embarquée pour la santé, les wearables et le suivi humain.

Capteurs de force, pression et couple – Force, Pressure, and Torque Sensors

Capteurs mesurant des efforts mécaniques ou des pressions, utilisés en IA embarquée pour la robotique, l'interaction homme–machine et le contrôle de systèmes physiques.

Capteurs multispectraux et hyperspectraux – Multispectral and Hyperspectral Sensors

Capteurs optiques capturant des informations sur plusieurs bandes spectrales, permettant une analyse fine des matériaux ou de l'environnement, utilisés en IA embarquée dans des domaines spécialisés (agriculture, environnement, défense).

Capteur intelligent – Smart Sensor

Capteur intégrant des capacités locales de traitement, de filtrage ou d'extraction de caractéristiques, permettant de réduire les volumes de données transmises et d'optimiser les performances et la consommation énergétique de l'IA embarquée.

Synchronisation multi-capteurs – Multi-sensor Synchronization

Mécanisme assurant l'alignement temporel des données provenant de capteurs hétérogènes, indispensable à la fusion de capteurs et à la cohérence des inférences en IA embarquée.

Horodatage – Timestamping

Association d'une référence temporelle aux données issues des capteurs afin de garantir leur cohérence, leur traçabilité et leur exploitation correcte dans des systèmes embarqués temps réel.

Dérive capteur – Sensor Drift

Évolution progressive des caractéristiques d'un capteur liée à l'environnement, à l'usure ou aux conditions d'exploitation, susceptible de dégrader la qualité des données et les performances des systèmes d'IA embarquée.

2.3 Traitements des données

Ce chapitre rassemble la terminologie relative aux opérations appliquées aux données afin de les rendre utilisables par des modèles d'IA embarquée. Il décrit les transformations nécessaires pour adapter les données aux contraintes de calcul, de mémoire et de temps réel des systèmes embarqués.

Les notions abordées mettent en évidence l'importance du traitement des données dans la performance et la stabilité de l'IA embarquée. Une compréhension partagée de ces termes permet de concevoir des chaînes de traitement adaptées aux exigences opérationnelles.

Ensemble d'opérations appliquées aux données issues des capteurs embarqués, après acquisition et avant inférence, visant à transformer des données brutes en entrées exploitables par des modèles d'IA embarquée, tout en respectant les contraintes de latence, de ressources, de déterminisme et de sûreté du système.

Prétraitement embarqué – On-device preprocessing

Prétraitement réalisé localement sur la cible embarquée (microcontrôleur, SoC, DSP, NPU), constituant une partie intégrante de la chaîne d'inférence, et visant à améliorer la qualité des entrées du modèle tout en maîtrisant coût calcul, mémoire et énergie.

Nettoyage de données – Data Cleaning

Opérations visant à corriger, filtrer ou ignorer des données incohérentes, erronées ou non exploitables issues des capteurs, afin de limiter les erreurs d'inférence et les comportements indésirables de l'IA embarquée.

Gestion des données manquantes – Missing Data Handling

Mécanismes permettant de gérer l'absence ponctuelle de mesures (ignorance contrôlée, valeur de repli, estimation simple), afin de préserver un comportement stable et sûr du système embarqué.

Débruitage et filtrage – Denoising and Filtering

Traitements numériques visant à réduire le bruit et les perturbations des signaux capteurs (filtrage fréquentiel, lissage), afin d'améliorer la robustesse des entrées du modèle sans introduire de latence excessive.

Correction et compensation – Correction and Compensation

Ensemble de mécanismes logiciels visant à corriger des biais ou dérives résiduelles dans les données numériques (offset, dérive lente), afin de maintenir la cohérence des entrées du modèle dans le temps.

Mise à l'échelle des données – Data Scaling

Transformation des valeurs numériques afin de les adapter à la plage attendue par les modèles et accélérateurs embarqués, en lien direct avec les contraintes de quantification et de représentation numérique.

Quantification des entrées – Input Quantization

Conversion des données d'entrée vers des représentations à précision réduite (par exemple entiers), afin d'optimiser l'exécution sur cibles embarquées tout en maîtrisant l'erreur introduite sur l'inférence.

Mise en forme des données – Data Formatting

Transformation structurelle des données (dimensions, organisation des canaux, formats numériques) afin de correspondre exactement aux entrées attendues par le modèle d'IA embarquée.

Encodage des données – Data Encoding

Représentation d'informations discrètes ou catégorielles sous une forme exploitable par le modèle (codage binaire, embeddings compacts), avec un souci de frugalité et de compatibilité embarquée.

Extraction de caractéristiques – Feature Extraction

Processus consistant à transformer les données brutes en variables pertinentes réduisant la complexité des entrées, afin d'améliorer robustesse, explicabilité et efficacité de l'inférence embarquée.

Réduction de dimension – Dimensionality Reduction

Techniques visant à limiter le nombre de variables d'entrée tout en conservant l'information utile, afin de réduire charge calcul, mémoire et latence sur la cible embarquée.

Fenêtrage temporel – Temporal Windowing

Découpage d'un flux de données en segments temporels glissants ou fixes, permettant une analyse locale adaptée aux contraintes temps réel des systèmes embarqués.

Segmentation – Segmentation

Découpage logique d'une donnée (signal, image, séquence) afin de restreindre la zone ou la période d'intérêt analysée par le modèle, réduisant ainsi le coût de traitement embarqué.

Adaptation du taux d'échantillonnage – Sampling Rate Adaptation

Ajustement logiciel de la fréquence d'échantillonnage des données numériques afin d'adapter le volume d'information aux capacités de calcul et aux besoins de l'application embarquée.

Synchronisation et alignement temporel – Synchronization and Temporal Alignment

Alignement des flux de données issus de capteurs multiples afin de garantir la cohérence temporelle des entrées d'un modèle d'IA embarquée multimodal.

Filtrage de valeurs aberrantes – Outlier Filtering

Détection et traitement de mesures numériques anormales ou incohérentes afin d'éviter des inférences erronées ou instables dans des environnements embarqués perturbés.

Prétraitement déterministe – Deterministic Preprocessing

Chaîne de prétraitement conçue pour produire des résultats reproductibles avec des délais d'exécution bornés, indispensable à l'intégration de l'IA dans des systèmes embarqués temps réel et critiques.

Prétraitement frugal – Frugal Preprocessing

Prétraitement optimisé pour minimiser l'usage des ressources matérielles (calcul, mémoire, énergie), afin de préserver l'autonomie et la performance globale du système embarqué.

Chaîne de prétraitement – Preprocessing Pipeline

Enchaînement structuré et validé de traitements appliqués aux données entre l'acquisition et l'inférence, considéré comme un composant à part entière du système d'IA embarquée.

Validation du prétraitement – Preprocessing Validation

Processus visant à vérifier que la chaîne de prétraitement respecte les exigences fonctionnelles, temporelles et de sûreté, et qu'elle ne dégrade pas le comportement global de l'IA embarquée en conditions opérationnelles.

3 INTELLIGENCE ARTIFICIELLE DANS L'EMBARQUÉ

L' embarquée se distingue de l'IA « générale » par un ensemble de contraintes propres aux systèmes physiques et aux environnements opérationnels. Elle doit fonctionner localement, souvent en temps réel, sur des ressources limitées (calcul, mémoire, énergie), tout en garantissant robustesse, sécurité, fiabilité et conformité réglementaire. Ces contraintes influencent profondément le choix des modèles, des architectures matérielles, des techniques d'optimisation et des outils. Les chapitres qui suivent mettront en lumière ces spécificités, tout en rappelant les fondements de l'IA « en général », indispensables à l'exhaustivité didactique de ce glossaire

La figure ci-dessous illustre les principaux paradigmes de l'intelligence artificielle et les classant en deux catégories: à partir de données (apprentissage automatique), à partir de connaissances (IA symbolique). L'IA appelée « hybride » ou « neurosymbolique » est composée de ces deux paradigmes. Les méthodes statistiques conventionnelles, conceptuellement proches des réseaux de neurones, ne sont pas habituellement considérées comme faisant partie de l'IA: quelques considérations à ce sujet figurent dans les "Lignes directrices de la Commission sur la définition d'un système d'intelligence artificielle au sens du règlement (UE) 2024/1689 (règlement sur l'IA)": https://ec.europa.eu/newsroom/dae/redirection/document/118623

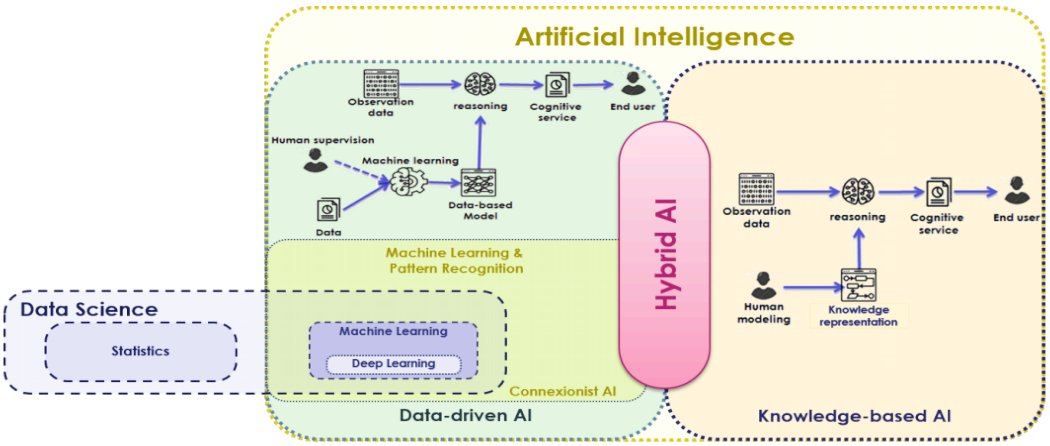

Figure reproduite avec l'aimable autorisation de Confiance.ai

Intelligence Artificielle (IA) – Artificial Intelligence (AI)

Ensemble de techniques mathématiques et informatiques permettant à une machine (un calculateur) de simuler des fonctions cognitives humaines (raisonnement, apprentissage, perception).

Système d'IA – AI System

Un système automatisé qui est conçu pour fonctionner à différents niveaux d'autonomie et peut faire preuve d'une capacité d'adaptation après son déploiement, et qui, pour des objectifs explicites ou implicites, déduit, à partir des entrées qu'il reçoit, la manière de générer des sorties telles que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels. (définition européenne officielle).

IA multimodale – Multimodal AI

IA capable de traiter simultanément plusieurs types de données (texte, image, audio, capteurs, …). Important pour les interfaces embarquées intelligentes (ex : assistants vocaux, lunettes connectées, ..).

Inférence – Inference

Phase d'exécution d'un modèle entraîné pour produire une prédiction ou une décision. L'inférence suit l'entraînement et précède l'évaluation. C'est le processus par lequel un modèle d'intelligence artificielle appliqué à des données nouvelles produit une prédiction, une décision ou une classification, sur la base de ce qu'il a appris lors de sa phase d'entraînement. C'est la mise en œuvre concrète du modèle dans un environnement réel ou embarqué, souvent en temps réel, sans réentraîner le modèle.

3.1 Grandes Familles de l'IA

Les approches fondamentales de l'intelligence artificielle décrivent les grands cadres conceptuels utilisés pour concevoir des systèmes intelligents. Dans le contexte de l'IA embarquée, ces approches conditionnent la manière dont les systèmes perçoivent, raisonnent et prennent des décisions sous contraintes de ressources, de temps réel et de sûreté.

IA connexionniste – Connectionist AI

Fondée sur l'apprentissage statistique et les réseaux de neurones ; c'est une approche de l'IA fondée sur des modèles inspirés du fonctionnement du cerveau humain, en particulier les réseaux de neurones artificiels. Elle excelle dans l'apprentissage à partir de données massives et dans le traitement de signaux complexes (image, son, texte).

IA symbolique – Symbolic AI

Approche reposant sur des représentations explicites des connaissances (règles, logiques, ontologies) et sur des mécanismes d'inférence. Elle privilégie la structure, l'explicabilité et la manipulation de concepts abstraits.

IA hybride – Hybrid AI

IA combinant les forces des deux précédentes pour répondre aux exigences de robustesse, d'explicabilité et d'adaptation: capacité d'apprentissage d'un côté, explicabilité et contrôle de l'autre. Elle est particulièrement pertinente pour les systèmes embarqués nécessitant robustesse, sûreté et adaptabilité.

3.1.1 Paradigmes d'apprentissage automatique

Les paradigmes d'apprentissage automatique définissent les différentes manières dont un système d'IA peut acquérir, adapter et exploiter des connaissances à partir des données. Dans le contexte de l'IA embarquée, le choix du paradigme d'apprentissage conditionne fortement l'architecture du système, son mode de déploiement et sa capacité à fonctionner sous contraintes de ressources, de temps réel et d'environnement.

Certains paradigmes reposent sur des phases d'entraînement hors ligne suivies d'une inférence locale, tandis que d'autres permettent une adaptation partielle ou continue au plus près du terrain, directement sur les dispositifs embarqués ou de manière distribuée.

Ces approches répondent à des enjeux variés tels que la limitation des données annotées, la préservation de la confidentialité, l'adaptation aux environnements évolutifs ou la coordination entre systèmes embarqués multiples.

Apprentissage Automatique – Machine Learning (ML)

Sous-discipline de l'intelligence artificielle qui consiste à concevoir des algorithmes capables d'apprendre à partir de données, sans être explicitement programmés pour chaque tâche. Les algorithmes identifient des motifs dans les données pour construire des modèles capables de faire des prédictions ou des décisions sur de nouvelles données.

Il existe différents types d'apprentissage:

Apprentissage supervisé – Supervised learning

Stratégie d'apprentissage où le modèle apprend à partir de données étiquetées (paires entrée-sortie), ajustant ses paramètres pour minimiser l'erreur entre ses prédictions et les étiquettes fournies. L'apprentissage supervisé est couramment utilisé pour les tâches de classification et de régression.

Apprentissage non supervisé – Unsupervised learning

Stratégie d'apprentissage où le modèle découvre les structures et les motifs inhérents aux données sans étiquette de référence. L'apprentissage non supervisé est utilisé pour des tâches comme le clustering (regroupement), la réduction de dimensionnalité et l'extraction de caractéristiques latentes.

Apprentissage semi-supervisé – Semi-supervised learning

Paradigme d'apprentissage combinant données étiquetées et non étiquetées, exploitant la structure des données non étiquetées pour améliorer les performances du modèle. L'apprentissage semi-supervisé est particulièrement utile lorsque l'étiquetage des données est coûteux ou limité.

Apprentissage par renforcement – Reinforcement learning

Stratégie d'apprentissage consistant à prendre des décisions optimales en interagissant avec un environnement, incluant des récompenses ou des pénalités selon ses actions. L'apprentissage par renforcement est appliqué aux tâches de contrôle, de robotique et de jeux, où le modèle doit apprendre une stratégie sans données étiquetées explicites.apprentissage par essais-erreurs avec récompenses selon les actions.

Apprentissage fédéré – Federated Learning

Stratégie d'apprentissage décentralisé permettant à plusieurs appareils ou serveurs de collaborer pour entraîner un modèle commun sans partager leurs données brutes. Chaque appareil (smartphone, capteur, robot, etc.) entraîne un modèle local sur ses propres données, puis envoie uniquement les paramètres du modèle (poids, gradients) à un serveur central. Le serveur agrège ces mises à jour pour créer un modèle global qui est ensuite redistribué à tous les appareils. Puisque les données restent localisées sur chaque appareil, l'apprentissage fédéré garantit une meilleure confidentialité et une conformité réglementaire renforcée (ex. RGPD). Aussi appelé apprentissage collaboratif décentralisé.

Apprentissage par transfert – Transfer Learning

Stratégie d'apprentissage qui consiste à réutiliser les connaissances acquises par un modèle sur une tâche source pour améliorer ses performances sur une tâche cible, généralement différente mais apparentée. Plutôt que d'entraîner un modèle à partir de zéro, on part d'un modèle pré-entraîné (souvent sur un grand jeu de données) et on l'adapte à une nouvelle tâche, souvent avec peu de données disponibles. Le fine-tuning qui consiste à réentraîner (partiellement ou totalement) un modèle pré-entraîné afin de l'adapter plus finement à la tâche cible est un exemple d'apprentissage par transfert. A titre d'exemple : un modèle entraîné à reconnaître des voitures dans des images peut être réutilisé (via fine-tuning) pour reconnaître des camions, avec peu de données supplémentaires. Cela reflète une forme d'apprentissage par analogie, proche du raisonnement humain.

Few-shot / Zero-shot Learning

Stratégie d'apprentissage visant à permettre à un modèle de généraliser à de nouvelles classes ou tâches à partir d'un nombre très limité d'exemples annotés (few-shot) ou sans aucun exemple direct (zero-shot) afin de réduire le coût de collecte et d'annotation des données.

Apprentissage en ligne – Online Learning

Technique d'apprentissage dans laquelle le modèle est mis à jour de manière incrémentale au fil de l'arrivée des données, sans repasser sur l'ensemble du jeu d'entraînement. Particulièrement utile en IA embarquée pour traiter des flux continus (capteurs, signaux, événements) et s'adapter en temps réel sous contraintes de ressources.

Apprentissage adaptatif – Adaptive Learning

Capacité d'un modèle ou d'un système à s'ajuster à des changements de contexte, de distribution ou d'environnement. Il peut s'appuyer sur l'apprentissage en ligne, mais inclut également des mécanismes plus larges (détection de dérive, adaptation de domaine, méta-apprentissage). Essentiel pour maintenir la performance et la robustesse de l'IA embarquée dans des conditions réelles évolutives.

3.1.2 Modèles connexionnistes

Les réseaux de neurones et l'apprentissage profond constituent aujourd'hui le socle algorithmique de nombreuses applications d'IA embarquée, notamment pour la perception, l'analyse de signaux et la prise de décision automatisée. Leur capacité à apprendre des représentations complexes à partir de données brutes en fait des outils puissants, mais leur mise en œuvre en embarqué nécessite des adaptations spécifiques pour respecter les contraintes de calcul, de mémoire, de latence et de consommation énergétique.

Réseau de neurones - Artificial Neural Network (ANN)

Modèle computationnel inspiré du fonctionnement du cerveau humain. Il est composé de neurones artificiels interconnectés, organisés en couches (entrée, cachées, sortie), qui transforment des données d'entrée en sortie via des calculs mathématiques. Chaque neurone reçoit des signaux d'entrée, les pondère, les transforme via une fonction d'activation, puis transmet le résultat. Le réseau apprend en ajustant les poids des connexions pour minimiser l'erreur entre la sortie prédite et la sortie réelle.

Apprentissage profond – Deep Learning

Sous-catégorie du machine learning qui utilise des réseaux de neurones artificiels profonds (avec plusieurs couches, ex. CNN, RNN) pour extraire automatiquement des représentations complexes à partir de données brutes. Il est particulièrement adapté aux données non structurées comme les images, les sons ou les textes.

3.1.3 Représentation des connaissances et IA symbolique

Les systèmes à base de connaissances constituent une approche de l'intelligence artificielle fondée sur la représentation explicite des savoirs et sur des mécanismes d'inférence logique. Dans le contexte de l'IA embarquée, ces systèmes apportent des capacités de raisonnement, de traçabilité et d'explicabilité particulièrement adaptées aux environnements contraints ou critiques.

Système à base de connaissances ou Système Expert – Knowledge-Based System or Expert System

Système de traitement de l'information qui aide à résoudre les problèmes dans un domaine d'application particulier en faisant des inférences à partir d'une base de connaissances. Le terme «système à base de connaissances» est parfois utilisé comme synonyme de «système expert», qui est normalement restreint aux connaissances d'experts. Certains systèmes à base de connaissances sont capables d'apprentissage. (ref. définition Confiance.ai).

Base de connaissances – Knowledge Base

Base de données contenant des règles d'inférence et des informations relatives à l'expérience et la compétence humaines dans un domaine particulier. Dans les systèmes autodidactes, la base de connaissances contient aussi des informations provenant de la résolution de problèmes antérieurs. (définition Confiance.ai).

Graphes de Connaissances – Knowledge Graphs

Dans le domaine de la représentation des connaissances, un graphe de connaissances (knowledge graph en anglais) est une base de connaissance modélisant les données sous forme de graphes. Les graphes de connaissances permettent de faciliter l'accès et l'intégration des données, d'ajouter du contexte et de la profondeur à d'autres techniques d'IA et servent de ponts entre les humains et les systèmes par exemple en des explications lisibles par des humains.

Resource Description Framework (RDF)

Modèle de graphe destiné à décrire formellement les ressources Web et leurs métadonnées, afin de permettre le traitement automatique de ces descriptions. Développé par le W3C, RDF est le langage de base du Web sémantique. Un triplet RDF est une association (sujet, prédicat, objet) : le sujet représente la ressource à décrire ; le prédicat représente un type de propriété applicable à cette ressource ; l'objet représente une donnée ou une autre ressource : c'est la valeur de la propriété.

Ontologie – Ontology

Modèle de données contenant des concepts et relations permettant de modéliser un ensemble de connaissances dans un domaine donné. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques ou des relations de subsomption.

Web Ontology Language (OWL)

Langage de représentation des connaissances construit sur le modèle de données de RDF. Il fournit les moyens pour définir des ontologies web structurées. Recommandation du W3C pour le web sémantique depuis 2012.

3.2 Modèles pour l'IA embarquée

Les modèles d'IA sont une représentation mathématique et algorithmique apprise à partir de données, utilisée par un système d'IA embarquée pour réaliser localement des tâches d'inférence telles que la perception, la classification, la prédiction ou la prise de décision. Dans un contexte embarqué, un modèle est conçu, adapté ou optimisé afin de pouvoir être exécuté sur des plateformes contraintes en calcul, mémoire, énergie et latence, tout en garantissant un comportement maîtrisé, reproductible et compatible avec les exigences opérationnelles du système.

3.2.1 Les principaux modèles d'IA

Les modèles d'IA constituent les briques algorithmiques permettant aux systèmes embarqués de percevoir, analyser et interpréter des données issues de capteurs ou de flux temporels. Dans le contexte de l'IA embarquée, le choix du modèle conditionne directement les performances, la latence, la consommation énergétique et la capacité du système à fonctionner en temps réel sur des plateformes contraintes.

Réseau de neurones convolutif – Convolutional Neural Network (CNN)

Réseau de neurones spécialisé dans le traitement de données visuelles (images, vidéos, signaux spatiaux). Il combine généralement trois types de couches : les couches de convolution qui extraient les caractéristiques locales via des filtres, les couches de pooling qui réduisent la dimensionnalité tout en préservant l'information pertinente, et les couches entièrement connectées qui produisent la sortie finale. Cette structure hiérarchique permet au réseau d'apprendre progressivement des motifs de complexité croissante, des éléments visuels simples aux concepts complexes. Ils sont particulièrement efficaces pour des applications comme la classification d'images, la détection d'objets ou la reconnaissance de visages, et peuvent être également adaptés pour le traitement des séries temporelles.

Réseau récurrent – Recurrent Neural Network (RNN)

Réseau de neurones conçu pour le traitement de données séquentielles ou temporelles. Les RNN se caractérisent par leur capacité à mémoriser les informations passées grâce à des connexions récurrentes, ce qui leur permet d'utiliser les sorties précédentes comme entrées pour les étapes suivantes. Les architectures les plus répandues, LSTM (Long Short-Term Memory) et GRU (Gated Recurrent Unit), ont été développées pour améliorer la gestion de la mémoire à long terme.

Modèle attentionnel - Attention-based model

Famille de modèles de réseaux de neurones intégrant un mécanisme d'attention, dont le rôle est d'attribuer dynamiquement des poids d'importance aux différentes parties d'une entrée (mots, instants temporels, régions d'image, etc.). Ce mécanisme permet au modèle de se concentrer sur les informations les plus pertinentes pour une tâche donnée et de mieux capturer les dépendances, y compris à longue distance. Les mécanismes d'attention peuvent être intégrés dans diverses architectures (réseaux récurrents, convolutifs ou autres) et sont largement utilisés en traitement du langage naturel, en vision par ordinateur et pour l'analyse de séries temporelles.

Transformer

Architecture de réseau de neurones appartenant à la famille des modèles attentionnels. Les Transformers traitent les données de manière parallèle, ce qui améliore l'efficacité de l'apprentissage et la modélisation des dépendances longues. Initialement développés pour le traitement du langage naturel, ils ont été étendus à d'autres domaines tels que la vision (Vision Transformers – ViT), l'audio et le multimodal. Des variantes optimisées et compactes (par exemple TinyBERT, MobileBERT, MobileViT) rendent leur déploiement possible sur des systèmes d'IA embarquée soumis à des contraintes de calcul, de mémoire et de consommation énergétique.

Réseau de neurones à impulsion – Spiking Neural Network 'SNN)

Réseau de neurones bio-inspirée où les neurones communiquent par impulsions discrètes (spikes) plutôt que par activations continues. Les neurones génèrent des impulsions lorsque leur potentiel de membrane dépasse un seuil, créant un système de communication événementiel et asynchrone comparable aux neurones biologiques. Ces réseaux traitent l'information de manière temporelle et événementielle, ce qui les rend particulièrement efficaces pour le traitement de données sensorielles et les tâches temps réel.

3.2.2 Les autres modèles

Au-delà des modèles neuronaux les plus couramment utilisés, l'IA embarquée mobilise également d'autres familles de modèles répondant à des besoins spécifiques de traitement, de raisonnement ou d'adaptation dynamique. Ces modèles se distinguent par leur capacité à représenter des structures complexes, des dépendances temporelles ou des relations probabilistes, tout en offrant des compromis variés entre expressivité, frugalité et robustesse.

Grand modèle de langage – Large Language Model (LLM)

Modèle d'IA, généralement basé sur des architectures de type Transformer, entraîné sur de vastes corpus textuels pour comprendre et générer du langage naturel. En IA embarquée, les LLM sont utilisés sous des formes adaptées (modèles réduits, quantifiés ou distillés) afin de respecter les contraintes de calcul, de mémoire et de latence des plateformes embarquées.

Petit modèle de langage – Small Language Model (SLM)

Modèle de langage de taille réduite, conçu pour des tâches spécifiques et nécessitant moins de ressources que les grands modèles de langage. En IA embarquée, les SLM sont privilégiés pour l'inférence locale en temps réel, grâce à leur empreinte mémoire limitée, leur latence maîtrisée et leur compatibilité avec des plateformes embarquées contraintes.

Réseau liquide – Liquid Neural Network (LNN)

Réseau de neurones récurrent dont la dynamique interne est décrite par des équations différentielles continues dans le temps. Contrairement aux réseaux récurrents classiques (RNN), ces réseaux adaptent leurs paramètres et leur comportement en fonction du signal d'entrée, ce qui leur confère une grande flexibilité et robustesse. Ils permettent de modéliser des systèmes dynamiques complexes avec peu de neurones et sont particulièrement adaptés aux applications embarquées en robotique, contrôle adaptatif et analyse de signaux non stationnaires.

Réseau de neurones de graphes – Graph Neural Network (GNN)

Réseau neuronal conçu pour apprendre à partir de données structurées sous forme de graphes. Chaque nœud agrège les informations de ses voisins pour construire une représentation globale. Les GNN sont utilisés pour la modélisation de réseaux de capteurs, la robotique, le traitement de données événementielles.

Modèle à espace d'états – State Space Model (SSM)

Réseau de neurones représentant les données temporelles à travers une séquence d'états cachés qui évoluent selon des équations différentielles. Un modèle à espace d'états maintient un état interne continu qui se met à jour en fonction des entrées reçues, permettant de capturer les dépendances temporelles complexes et les dynamiques long terme.

Réseau bayésien – Bayesian Network

Modèle probabiliste qui utilise l'inférence bayésienne pour les calculs de probabilité en utilisant un graphe acyclique orienté. (définition Confiance.ai).

3.3 Techniques d'optimisations pour l'IA embarquée

Les systèmes embarqués présentent des contraintes strictes en termes de mémoire, puissance de calcul, latence et consommation énergétique. Dans ce contexte, le déploiement de modèles d'intelligence artificielle classiques, tels que les réseaux de neurones profonds, nécessite des techniques d'optimisation adaptées afin de garantir une exécution efficace tout en maintenant des performances acceptables.

Ces techniques couvrent plusieurs axes complémentaires :

- **Optimisations numériques**, visant à réduire la précision des poids et des activations (**quantification**), afin de diminuer l'empreinte mémoire et d'accélérer l'inférence.
- **Optimisations de compression**, permettant de réduire le nombre de paramètres ou la complexité des modèles.
- **Optimisation d'architectures**, regroupant les approches permettant de définir ou d'adapter la structure du modèle aux contraintes matérielles et applicatives.

3.3.1 Optimisations numériques

Les optimisations numériques constituent un levier essentiel pour adapter les modèles d'intelligence artificielle aux contraintes des systèmes embarqués. En agissant directement sur la représentation et la précision des calculs, elles permettent de réduire l'empreinte mémoire, d'accélérer l'inférence et de diminuer la consommation énergétique, tout en préservant un niveau de performance compatible avec les exigences applicatives.

Quantification – Quantization

Méthode qui consiste à réduire la précision numérique des poids et des activations d'un modèle IA, en passant de formats flottants (ex : 32 bits) à des formats entiers plus compacts (ex : 8 bits, 4 bits, voire binaire). Elle permet de réduire la taille mémoire, accélérer l'inférence, et diminuer la consommation énergétique, tout en maintenant des performances acceptables. Essentielle dans l'embarqué pour le déploiement sur microcontrôleurs, NPU, DSP, où les ressources sont limitées.

Précision mixte – Mixed-Precision Quantization

Technique combinant plusieurs niveaux de précision numérique au sein d'un même modèle (ex : FP32, FP16, INT8) pour adapter la vitesse, la consommation mémoire et la fidélité des prédictions selon la sensibilité de chaque couche.

Binarisation – Binarization

Cas extrême de quantification où les poids et parfois les activations ne prennent que deux valeurs distinctes (ex. –1/+1 ou 0/1), remplaçant ainsi les multiplications par des opérations logiques élémentaires.

Ternarisation – Ternarization

Variante de la binarisation permettant aux poids de prendre trois valeurs possibles (ex. –1, 0, +1).

Quantification post-entraînement – Post-Training Quantization (PTQ)

Ensemble de techniques appliquant la quantification après l'entraînement complet du modèle.

Quantification pendant l'entraînement – Quantization Aware Training (QAT)

Ensemble de techniques intégrant la quantification lors de l'entraînement, permettant au modèle de s'adapter à la perte de précision et d'atteindre de meilleures performances.

3.3.2 Optimisations de compression

Les optimisations de compression visent à réduire la taille et la complexité des modèles d'intelligence artificielle afin de les rendre compatibles avec les contraintes des systèmes embarqués. En diminuant le nombre de paramètres à stocker et à traiter, ces techniques permettent de limiter l'empreinte mémoire, d'accélérer l'exécution et de réduire la consommation énergétique, tout en conservant des performances fonctionnelles acceptables.

Élagage – Pruning

Méthode qui consiste à supprimer les connexions ou neurones jugés non essentiels dans un réseau de neurones, en exploitant la sparsité naturelle des poids. Cela permet de réduire la complexité du modèle, accélérer les calculs, et économiser de la mémoire, sans perte significative de précision ; utile pour les rendre compatibles avec des architectures edge par exemple.

Elagage par magnitude – Magnitude Pruning

Suppression des poids de plus faible amplitude.

Élagage structuré – Structured Pruning

Suppression d'éléments structuraux du réseau, tels que des filtres convolutionnels, des canaux ou des couches entières.

Compression par factorisation – Low-Rank Factorization)

Technique approximant des matrices de poids par des facteurs de rang inférieur, réduisant le nombre de paramètres et les calculs.

Compression par encodage – Encoding-based Compression

Technique réduisant la taille du modèle stocké en appliquant des techniques d'encodage (ex. codage de Huffman, codage entropique) pour représenter les paramètres de manière plus compacte.

Compression par regroupement – Weight Clustering

Technique regroupant des poids en centres de clusters, puis remplaçant par des index. Il s'agit d'une forme de compression vectorielle permettant de réduire la mémoire...

Compression par partage de poids – Weight Sharing

Technique utilisant les mêmes poids dans plusieurs parties du réseau, réduisant ainsi la mémoire nécessaire et la complexité du modèle.

3.3.3 Optimisation d'architectures

Les optimisations d'architecture visent à adapter la structure même des modèles d'intelligence artificielle afin de répondre aux contraintes spécifiques des systèmes embarqués. En agissant sur l'organisation des réseaux, les chemins de calcul ou les mécanismes d'exécution, ces techniques permettent de trouver un compromis optimal entre précision, latence, consommation énergétique et robustesse.

Distillation – Distillation

Méthode où un modèle complexe ("teacher") transmet ses connaissances à un modèle plus simple ("student"), en lui apprenant à reproduire ses prédictions. Elle permet de conserver les performances d'un modèle lourd tout en réduisant sa taille et sa complexité.

Recherche d'architecture neuronale – Neural Architecture Search (NAS)

Méthode automatique pour chercher la meilleure architecture de réseau selon des contraintes (latence, précision, mémoire). Usage embarqué : permet de générer des modèles sur mesure pour une cible matérielle donnée.

Sortie anticipée – Early Exit

Technique qui ajoute des points de sortie intermédiaires dans le réseau, permettant une inférence anticipée si la confiance est suffisante. Usage embarqué : réduction de la latence et de la consommation énergétique, particulièrement utile dans les systèmes temps réel.

Optimisation topologique – Topological optimization

Ensemble de techniques modifiant la structure du modèle en fusionnant plusieurs opérations consécutives en une seule opération optimisée ou en modifiant leur ordonnançant leur exécution pour réduire les accès mémoire et la latence.

Découpage en blocs – Tiling

Ensemble de techniques d'optimisation consistant à partitionner les calculs et les données en blocs de taille adaptée à l'architecture du matériel cible, afin d'améliorer la localité mémoire, maximiser l'utilisation des mémoires rapides et exploiter efficacement le parallélisme matériel.

4 Architecture système pour l'IA embarquée

Les architectures et concepts de l'IA embarquée constituent le socle sur lequel reposent les systèmes intelligents distribués, frugaux et critiques. Ils décrivent à la fois les organisations matérielles (processeurs spécialisés, systèmes temps réel, calcul neuromorphique), les architectures logicielles et organisationnelles (edge computing, fog computing, AIoT, systèmes distribués, pipelines d'inférence), ainsi que des concepts et patterns architecturaux plus avancés (jumeaux numériques, architectures agentiques, architectures edge mesh ou fédérées).

Cette diversité reflète la nature hybride de l'IA embarquée :

- Elle combine des contraintes physiques fortes telles que la latence, la consommation énergétique ou la mémoire, avec des exigences logicielles élevées en matière de performance, de fiabilité et de maîtrise du comportement.
- Elle s'inscrit également dans des architectures hiérarchiques et distribuées (edge–fog–cloud), dans lesquelles les traitements peuvent être répartis entre la périphérie, les passerelles et les infrastructures centrales.
- Elle intègre des approches émergentes, telles que l'agentique, la fédération ou le calcul neuromorphique, ouvrant la voie à des systèmes plus autonomes, résilients et souverains.

L'objectif de ce chapitre est de proposer une cartographie claire et contextualisée des principales architectures et concepts mobilisés dans l'IA embarquée, afin de faciliter leur appropriation par la communauté industrielle, académique et institutionnelle.

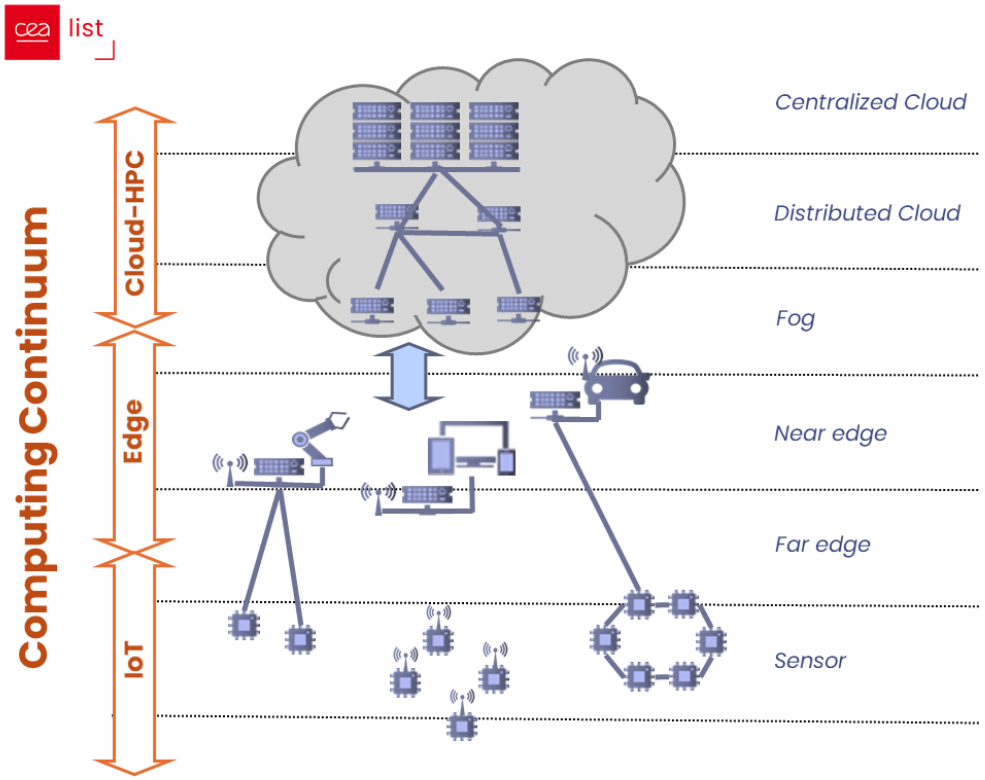

Figure : schéma représentant une architecture distribuée s'étendant du cloud HPC jusqu'aux capteurs IoT (© Denis Dutoit, List Tech Days 2025)

Continuum de calcul – Computing Continuum

Distribution continue du calcul et du traitement des données depuis les dispositifs embarqués jusqu'au cloud, permettant de choisir dynamiquement l'emplacement optimal selon latence, énergie et bande passante.

Cloud Computing

Architecture centralisée où les données et les calculs sont traités dans des centres de données distants. Dans le contexte de l'IA embarquée, le cloud est utilisé pour l'entraînement des modèles, le stockage et la gestion des services, en articulation avec les niveaux edge et fog.

Systèmes distribués – Distributed Systems

Répartition des traitements entre plusieurs nœuds du système, pour améliorer résilience et latence.

Fog Computing

Couche intermédiaire entre edge et cloud, permettant un traitement distribué et hiérarchisé des données, souvent au niveau des passerelles ou des réseaux locaux.

Architecture edge-to-cloud – Edge-to-cloud Architecture

Organisation hiérarchique où les traitements sont répartis entre périphérie (edge), intermédiaire (fog) et cloud, selon les contraintes de latence, de sécurité et de puissance de calcul.

Edge Computing

Paradigme de calcul décentralisé où le traitement des données est effectué au plus près des capteurs ou des appareils. Cela permet de réduire la latence et le bilan énergétique, d'améliorer la confidentialité des données et de diminuer la charge du réseau.

Edge AI

Déclinaison de l'IA spécifiquement exécutée sur des dispositifs en périphérie (edge devices), combinant modèles optimisés et contraintes embarquées.

Internet des objets – Internet of Things (IoT)

Réseau d'objets physiques connectés capables de collecter, traiter et échanger des données via Internet ou un réseau local. L'IA embarquée y permet une prise de décision locale, réduisant la dépendance au cloud et la latence. Les systèmes embarqués sont au cœur de l'IoT et des applications d'IA décentralisées.

Artificial Intelligence of Things (AIoT)

Convergence de l'IA et de l'IoT, où les objets connectés intègrent des capacités d'apprentissage et d'inférence locales.

TinyML

Déploiement de modèles de ML sur des microcontrôleurs et appareils embarqués à très faible puissance, faible coût et ressources limitées, pour de l'inférence locale en temps réel.

Systèmes cyber-physiques – Cyber-Physical Systems (CPS)

Intégration étroite entre composants physiques et logiciels, où l'IA assure perception et contrôle en temps réel.

Systèmes temps réel – Real-Time Systems

Architectures garantissant des réponses déterministes dans des délais stricts. Dans le contexte de l'IA embarquée, ils permettent d'exécuter des algorithmes de traitement et de prise de décision avec latence minimale, ce qui est crucial pour des applications critiques.

Edge Mesh

Architecture distribuée dans laquelle des nœuds edge (capteurs, microcontrôleurs, caméras, robots, passerelles) sont interconnectés en réseau maillé, leur permettant de communiquer et de coopérer directement sans dépendre systématiquement du cloud. L'edge mesh favorise une intelligence collective distribuée, améliore la résilience et la continuité de service, et renforce la souveraineté des données locales. Il est souvent associé à des approches d'apprentissage collaboratif ou fédéré.

5 LOGICIELS ET OUTILS POUR L'IA EMBARQUÉE

Après avoir présenté les architectures et concepts qui structurent l'IA embarquée, ce chapitre s'intéresse aux briques logicielles et aux outillages indispensables pour transformer ces concepts en solutions opérationnelles. Ces éléments constituent l'environnement de développement et d'exécution permettant d'assurer la portabilité, l'efficacité et la maîtrise des modèles d'IA sur des plateformes embarquées hétérogènes.

Les logiciels et outils décrits dans ce chapitre offrent des interfaces de développement adaptées aux contraintes spécifiques de l'embarqué, telles que le temps réel, la frugalité des ressources et les exigences de sécurité. Ils facilitent l'intégration, l'optimisation, l'exécution et la validation des modèles d'IA embarquée, y compris dans des systèmes à criticité élevée.

Ces briques logicielles ne peuvent toutefois être dissociées des composants matériels spécialisés qui les supportent — processeurs, accélérateurs, circuits dédiés et mémoires. Le chapitre suivant est consacré à ces fondations matérielles de l'IA embarquée et met en lumière leur rôle déterminant dans la performance, la consommation énergétique et la frugalité des systèmes.

Le présent chapitre propose une cartographie structurée des principaux éléments logiciels de l'IA embarquée, en distinguant notamment les langages, les formats de modèles, les frameworks, les kits de développement, les bibliothèques matérielles et les systèmes d'exploitation.

5.1 LANGAGES

Les langages de programmation utilisés en IA embarquée jouent un rôle déterminant dans la capacité à concevoir, intégrer et exécuter des modèles d'intelligence artificielle sur des plateformes contraintes. Ils conditionnent l'accès au matériel, la maîtrise des ressources (calcul, mémoire, énergie), ainsi que la performance et la fiabilité des systèmes.

Du développement bas niveau optimisé à l'intégration d'outils de prototypage et de déploiement rapide, ces langages permettent de couvrir l'ensemble du cycle de mise en œuvre de l'IA embarquée, en s'adaptant aux exigences de temps réel, de sûreté et de portabilité propres aux environnements embarqués.

C / C++

Langages compilés offrant un contrôle fin du matériel et des ressources, largement utilisés pour le développement d'IA embarquée, de drivers et d'optimisations bas niveau sur CPU, GPU et NPU.

Rust

Langage compilé moderne garantissant la sécurité mémoire et une gestion sûre de la concurrence, de plus en plus utilisé pour des applications embarquées critiques et des systèmes IA performants et fiables.

Python

Langage interprété très populaire pour le prototypage rapide de modèles IA ; en IA embarquée, il est utilisé via MicroPython sur des dispositifs contraints ou via CPython sur des plateformes embarquées pour l'exécution de modèles et de scripts de haut niveau.

Java

Langage orienté objet utilisé sur des plateformes mobiles et embarquées (Android, IoT), permettant de développer des applications intégrant des fonctionnalités IA avec une gestion automatique de la mémoire et un écosystème riche de bibliothèques.

JavaScript

Langage interprété principalement utilisé dans les environnements web et certains systèmes embarqués disposant d'un moteur JavaScript, permettant l'exécution de modèles IA, côté client, et le déploiement rapide d'applications IA interactives.

WebAssembly (Wasm)

Format binaire portable et sécurisé servant de cible de compilation pour des langages comme C, C++ ou Rust, permettant l'exécution de code compilé à haute performance dans des environnements web ou embarqués.

5.2 Gestion des modèles

Les formats de modèles constituent un élément clé de la chaîne de déploiement de l'IA embarquée. Ils déterminent la manière dont les modèles sont représentés, échangés et intégrés entre les phases d'entraînement, d'optimisation et d'exécution sur des plateformes embarquées hétérogènes.

En favorisant l'interopérabilité, la portabilité et l'optimisation des modèles, ces formats permettent d'adapter les architectures d'IA aux contraintes spécifiques de l'embarqué, telles que la limitation des ressources, les exigences de performance, de certification et de compatibilité matérielle.

5.2.1 Format des modèles

Dans le domaine de l'intelligence artificielle, un format de fichier pour un modèle définit la manière dont les poids, les structures du réseau et les métadonnées associées sont stockés et échangés. Ces formats jouent un rôle clé dans le déploiement, la portabilité et l'optimisation des modèles, en facilitant leur utilisation sur différents frameworks, plateformes matérielles et environnements.

Open Neural Network Exchange (ONNX)

Format ouvert et standard d'interopérabilité permettant de sérialiser et de transférer des modèles de machine learning entre différents frameworks d'apprentissage profond. Il offre une représentation abstraite et indépendante du framework, agissant comme un pont standardisé entre l'entraînement et le déploiement. Il facilite la portabilité sur diverses plateformes matérielles (CPU, GPU, appareils mobiles, edge) grâce à sa prise en charge par de nombreux runtimes d'inférence optimisés pour les systèmes embarqués. Un groupe de travail, nommé S-ONNX, a été mis en place afin de faire évoluer le format et mieux prendre en compte les contraintes liées aux processus de certification.

Neural Network Exchange Format (NNEF)

Standard ouvert permettant l'échange de modèles de réseaux de neurones entre outils d'apprentissage et d'inférence. Conçu pour l'interopérabilité sur systèmes embarqués, il facilite le déploiement optimisé sur diverses plateformes matérielles (CPU ARM, GPU mobiles, DSP, FPGA, ASIC).

GPT-Generated Unified Format (GGUF)

Format binaire conçu pour stocker efficacement les modèles de langage et leurs métadonnées. Il supporte nativement la quantification (4-bit, 8-bit) et le chargement partiel en mémoire, facilitant le déploiement de LLM sur des appareils embarqués à ressources limitées (edge devices, smartphones).

Neuromorphic Intermediate Representation (NIR)

Standard conçu pour faciliter l'interopérabilité entre différentes plateformes neuromorphiques. Il définit un ensemble de primitives de calcul événementielles, permettant de transférer des modèles sur des processeurs neuromorphiques (comme Intel Loihi, SpiNNaker) ou des accélérateurs spécialisés.

5.2.2 Bibliothèques de modèles

Les bibliothèques de modèles constituent un point d'entrée essentiel pour le développement de solutions d'IA embarquée, en mettant à disposition des modèles pré-entraînés pouvant être sélectionnés, évalués et adaptés à des contextes matériels et applicatifs spécifiques. Elles permettent d'accélérer les phases de prototypage et de déploiement tout en facilitant la réutilisation de modèles éprouvés.

Zoo de modèles – Model Zoo

Collection de modèles d'IA pré-entraînés. Le *model zoos* facilite la sélection et l'adaptation de modèles. Exemples : ONNX Model Zoo, OpenVINO Model Zoo, Qualcomm AI Hub ou Qualcomm sur Hugging Face, STMicroelectronics sur Hugging Face, Eclipse Aidge sur Hugging Face

Hugging Face

Plateforme collaborative d'IA open source proposant un vaste model zoo hébergeant des millions de modèles pré-entraînés.

Fiche descriptive d'un modèle – Model Card

Document non standardisé, décrivant généralement les caractéristiques techniques, les performances, les limites, les biais potentiels et les cas d'usage recommandés d'un modèle d'IA. Elle assure la transparence et guide les développeurs dans le choix et le déploiement approprié des modèles, notamment pour les applications embarquées.

5.3 Plateformes et kits de développements

Les plateformes dédiées à l'IA embarquée constituent l'environnement structurant permettant d'industrialiser le développement et le déploiement de modèles d'IA sur des systèmes contraints. Elles orchestrent l'ensemble de la chaîne logicielle, depuis la conception et l'entraînement des modèles jusqu'à leur optimisation, leur intégration et leur exécution sur des architectures matérielles hétérogènes.

En fournissant des outils, des bibliothèques et des méthodologies adaptées aux contraintes de l'embarqué, ces plateformes facilitent la portabilité, la performance et la fiabilité des solutions d'IA. Elles jouent également un rôle clé dans la traçabilité, la validation et, le cas échéant, la certification des systèmes, en assurant un lien cohérent entre les modèles, les outils logiciels et les composants matériels sur lesquels ils sont déployés.

5.3.1 Plateformes dédiées à l'IA embarquée

Plateforme – Platform

Ensemble structuré d'outils logiciels offrant une architecture cohérente pour concevoir, entraîner, optimiser et déployer des modèles d'intelligence artificielle. Elle fournit des fonctionnalités prêtes à l'emploi (gestion des données, des modèles et des métriques) et impose une méthodologie de développement. Dans le contexte de l'IA embarquée, il simplifie le déploiement sur du matériel aux ressources limitées.

Eclipse Aidge

Plateforme open source et collaborative hébergée par la Fondation Eclipse, développée principalement par des acteurs académiques et industriels européens. Cette plateforme fournit une chaîne outillée complète pour concevoir, entraîner, optimiser et déployer des modèles IA embarquée. Aidge permet de générer un code source optimisé et auditable pour diverses architectures matérielles, avec l'ambition de faciliter le déploiement sur des architectures hétérogènes. Elle offre également des fonctionnalités natives pour réaliser du benchmarking et pour répondre aux exigences de certifications (traçabilité, robustesse aux fautes matérielles, etc.).

TensorFlow, **LiteRT**

Développé par Google, TensorFlow est une plateforme complète qui permet de concevoir, entraîner, optimiser et déployer des modèles d'intelligence artificielle. Elle fournit des outils de gestion de données, des modules d'entraînement, et des interfaces pour l'inférence sur CPU, GPU ou dispositifs embarqués. Elle est extensible via des bibliothèques spécialisées (par exemple TensorFlow Federated, TensorFlow Decision Forests) qui ciblent des usages spécifiques, y compris sur matériel contraint (LiteRT).

PyTorch, ExecuTorch

Développé par Meta, PyTorch est une plateforme open source conçu pour la création, l'entraînement et le déploiement de modèles de deep learning. Elle permet de concevoir des réseaux neuronaux complexes, d'exécuter des calculs tensoriels, et d'optimiser des modèles sur CPU ou GPU. Executorch est une plateforme complémentaire à PyTorch, spécialisée dans l'inférence optimisée de modèles pré-entraînés, permettant leur déploiement efficace sur des systèmes embarqués ou contraints en ressources.

Edge Impulse

Acquis par Qualcomm, Edge Impulse est une plateforme dédiée à la création, l'entraînement et le déploiement de modèles d'intelligence artificielle optimisés pour les appareils embarqués (edge devices). Elle propose des outils, sous la forme d'une plateforme cloud, permettant de collecter des données, concevoir des modèles, et les intégrer facilement dans des dispositifs à ressources limitées comme les microcontrôleurs ou les capteurs IoT.

5.3.2 Kits de développement

Dans l'écosystème de l'IA embarquée, les fournisseurs de composants matériels mettent généralement à disposition leurs propres kits de développement afin de faciliter l'implémentation et l'optimisation des modèles sur leurs plateformes. Ces kits regroupent le plus souvent des bibliothèques spécifiques au matériel, des outils de configuration ainsi que des couches middleware.

Kit de développement (Software Development Kit - SDK)

Ensemble d'outils, de bibliothèques, mis à disposition pour permettre aux développeurs de créer des applications ou d'interagir avec une plateforme, un service ou un matériel spécifique.

Par exemple :

- **STM32Cube** : kit de STMicroelectronics pour ses microcontrôleurs STM32.
- **JetPack SDK** : kit de NVIDIA pour ses modules Jetson avec support CUDA & TensorRT
- **OpenVINO** : kit de Intel pour ses processeurs et accélérateurs.
- **Snapdragon Neural Processing SDK** : kit de Qualcomm pour ses SoC.

5.4 Bibliothèques matérielles

Les bibliothèques matérielles optimisées jouent un rôle clé pour accélérer les calculs critiques tels que les convolutions, les produits matriciels ou les activations. Elles fournissent des implémentations hautement performantes adaptées à chaque type de matériel — microcontrôleurs, CPU, GPU, NPU ou FPGA. Ces bibliothèques sont souvent intégrées dans les frameworks et runtimes cités précédemment et utilisées par les compilateurs pour générer du code natif optimisé, permettant de déployer des modèles IA efficacement.

Par exemple :

- **CMSIS-NN :** bibliothèque open source optimisée pour les Cortex-M de ARM
- **XNNPACK :** bibliothèque open source optimisée pour les CPU ARM et x86
- **cuDNN :** bibliothèque NVIDIA pour GPU

5.5 Compilateurs

Les compilateurs et chaînes d'optimisation jouent un rôle central dans la mise en œuvre de l'IA embarquée, en assurant la transformation des modèles et du code de haut niveau en exécutables optimisés pour des plateformes matérielles contraintes. Ils constituent le lien essentiel entre les abstractions algorithmiques de l'IA et l'exécution efficace sur des architectures spécifiques telles que les microcontrôleurs, processeurs embarqués, accélérateurs ou FPGA.

En combinant représentations intermédiaires, optimisations automatiques et techniques ciblées (quantification, fusion d'opérateurs, vectorisation, auto-tuning), ces outils permettent d'adapter les modèles d'IA aux contraintes de performance, de consommation énergétique et de latence propres à l'embarqué, tout en garantissant la portabilité et la maîtrise du comportement à l'exécution.

Compilateur – Compiler

Outil qui traduit le code source en code exécutable natif pour une plateforme matérielle spécifique. Il effectue généralement des optimisations (optimisation topologique comme la fusion de couche, vectorisation, pruning) pour améliorer l'exécution.

Low Level Virtual Machine (LLVM)

Infrastructure open source de compilation qui fournit un compilateur modulaire et des outils d'optimisation pour générer du code natif sur différentes architectures (x86, ARM, RISC-V, GPU). LLVM permet la représentation intermédiaire des programmes, facilitant l'analyse, transformation et optimisation du code pour des cibles variées, y compris embarquées & IA.

TVM

Framework open source de compilation automatique conçu pour transformer des modèles IA pré-entraînés en code bas niveau optimisé, déployable sur une grande variété de matériels — du cloud aux dispositifs edge (CPU, GPU, FPGA, microcontrôleurs). Il génère du code bas niveau en s'appuyant notamment sur LLVM ou CUDA, tout en appliquant des techniques avancées d'optimisation telles que la quantification, la fusion d'opérateurs et le pruning.

Multi-Level Intermediate Representation (MLIR)

Infrastructure de compilateur open-source fournissant une architecture générique et extensible pour la représentation intermédiaire de code à plusieurs niveaux d'abstraction. MLIR permet de construire des compilateurs et des outils d'optimisation pour le machine learning, facilitant la traduction de modèles de haut niveau vers du code optimisé pour différentes cibles matérielles (CPU, GPU, accélérateurs neuromorphiques). Son caractère multi-niveaux permet d'appliquer des optimisations progressives, du niveau sémantique du modèle jusqu'aux instructions machine.

Accelerated Linear Algebra (XLA)

Compilateur basé sur MLIR développé par Google pour TensorFlow et JAX, qui transforme les graphes computationnels en code optimisé pour CPU, GPU ou TPU, en appliquant fusion de graphes et élimination des opérations redondantes.

Il existe également des compilateurs dédiés pour certaines cibles matérielles, afin d'optimiser au mieux l'exécution des modèles sur le matériel spécifique.

Par exemple, on peut citer :

ARM-GCC

Compilateur largement utilisé pour les microcontrôleurs et processeurs ARM, offrant des optimisations pour la vitesse et la taille mémoire.

FINN

Compilateur développé par Xilinx, spécialement conçu pour générer du code et des pipelines optimisés pour les FPGA et accélérateurs AMD.

CUDA NVCC

Compilateur pour les plateformes NVIDIA telles que Jetson, Orin ou les GPU RTX, souvent en association avec TensorRT ou DeepStream pour l'inférence IA. Transforme les kernels CUDA en code exécutables hautement optimisés pour les GPU. En plus des techniques d'optimisation topologique, les compilateurs s'appuient généralement sur des techniques d'optimisation spécifiques

Auto-tuning

Technique où le compilateur ou runtime **teste différentes configurations de calcul et paramètres de kernel** pour choisir la combinaison qui maximise les performances sur le matériel cible (CPU, GPU, NPU).

Vectorisation – Vectorization

Optimisation qui transforme des opérations scalaires en opérations sur **vecteurs ou blocs de données**, permettant d'exploiter les instructions SIMD et d'accélérer les calculs sur CPU ou GPU.

Dépliage de boucles – Loop unrolling

Technique qui remplace une boucle par plusieurs itérations consécutives non itératives pour réduire le coût des instructions de contrôle et améliorer la parallélisation.

5.6 Runtime

Les moteurs d'inférence constituent la couche logicielle chargée d'exécuter les modèles d'intelligence artificielle sur les plateformes embarquées. Ils assurent le chargement des modèles, leur exécution efficace et l'exploitation optimale des ressources matérielles disponibles, en tenant compte des contraintes de latence, de consommation énergétique et de précision numérique propres à l'embarqué.

En s'appuyant sur des mécanismes d'optimisation et sur l'accès contrôlé au matériel sous-jacent, les moteurs d'inférence jouent un rôle central dans la performance et la fiabilité des systèmes d'IA embarquée. Ils permettent de rendre opérationnels des modèles issus de différents frameworks, tout en garantissant leur portabilité et leur intégration dans des environnements temps réel ou critiques.

Moteur d'inférence – Inference engine or Runtime

Couche logicielle responsable de charger et d'exécuter un modèle d'IA dans un environnement donné (cloud ou embarqué), en s'appuyant éventuellement sur l'OS pour accéder au matériel. Il optimise l'utilisation des ressources (CPU, GPU, mémoire), la latence et parfois la **précision** numérique (FP32, FP16, INT8), afin de permettre au modèle de fonctionner efficacement.

Les plateformes logicielles dédiées à l'IA embarquée proposent généralement un runtime. On peut également citer les runtime suivants :

ONNX Runtime

Moteur d'inférence open-source conçu pour l'inférence de modèles de machine learning au format ONNX. Il offre une exécution optimisée (quantification, optimisation topologique) et multiplateforme, compatible avec différentes architectures matérielles (CPU, GPU, accélérateurs).

TensorRT

Moteur d'inférence optimisé, intégré dans des frameworks embarqués comme JetPack SDK ou DeepStream permettant d'optimiser (quantification, optimisation topologique, pruning) et d'exécuter des modèles IA (issus de PyTorch, TensorFlow, ONNX…) sur les GPU NVIDIA, notamment ceux dédiées à l'embarqué (Jetson, Orin) Xavier, ou RTX) des environnements embarqués ou temps réel (Jetson, Orin, Xavier, ou RTX).

5.7 Systèmes d'exploitation

Les systèmes d'exploitation constituent la couche logicielle fondamentale permettant d'exécuter des applications d'IA embarquée en interaction directe avec le matériel.
Ils déterminent la manière dont les ressources de calcul, de mémoire et de périphériques sont allouées et maîtrisées, influençant directement la performance, la latence et la fiabilité des inférences réalisées localement.

Selon le niveau de contrainte et les exigences temporelles, l'IA embarquée peut s'appuyer sur des environnements sans système d'exploitation (bare metal), sur des distributions Linux embarquées ou sur des systèmes d'exploitation temps réel. Ces choix conditionnent la capacité du système à répondre aux contraintes de temps réel, de frugalité et de sûreté.

Système d'exploitation embarqué – Embedded Operating System

Logiciel de base gérant les ressources matérielles d'un système embarqué et fournissant les services nécessaires à l'exécution d'applications d'IA embarquée sous contraintes de temps réel, de mémoire et d'énergie.

Micro-ROS

Version allégée de ROS 2 adaptée aux microcontrôleurs, permettant l'intégration de fonctions de perception et de décision IA embarquée dans des systèmes très contraints en ressources.

5.7.1 Systèmes d'exploitation Linux

Yocto Project

Ensemble d'outils open source permettant de créer des distributions Linux personnalisées et optimisées pour systèmes embarqués, réduisant la taille du système d'exploitation et l'empreinte mémoire selon les besoins spécifiques de l'application.

Android / Android Things

Système d'exploitation basé sur Linux développé par Google, Android étant destiné aux smartphones et tablettes, tandis qu'Android Things est une version optimisée pour les appareils embarqués et l'IoT, permettant d'utiliser l'écosystème Android sur des microcontrôleurs et SoC pour objets connectés et dispositifs intelligents.

Ubuntu Core

Distribution Linux embarquée modulaire et sécurisée, conçue pour les dispositifs connectés et edge, facilitant le déploiement et la mise à jour de composants logiciels, y compris des applications d'IA embarquée.

Buildroot

Ensemble d'outils permettant de générer des systèmes Linux embarqués minimalistes et optimisés, adaptés au déploiement de pipelines d'IA embarquée sur des plateformes à ressources limitées.

5.7.2 Systèmes d'exploitation temps réel

Système d'exploitation temps réel – Real-Time Operating System (RTOS)

Système d'exploitation spécialisé conçu pour exécuter des tâches avec des contraintes temporelles strictes, garantissant que certaines opérations se produisent dans un délai déterminé et prévisible.

Free RTOS

Système d'exploitation temps réel (RTOS) open source léger conçu pour microcontrôleurs et processeurs embarqués, fournissant multitâche, gestion de queues, sémaphores et timers pour applications critiques en temps réel.

Zephyr RTOS

Système d'exploitation temps réel (RTOS) open source destiné aux appareils IoT et microcontrôleurs, offrant un support natif des périphériques, des communications réseau et des petites applications d'IA embarquée avec une empreinte mémoire optimisée.

Eclipse ThreadX

Système d'exploitation temps réel (RTOS) open source, complet et certifié pour les applications critiques, conçu spécifiquement pour les systèmes embarqués et les dispositifs de l'Internet des Objets (IoT) à ressources limitées (microcontrôleurs). Connu sous le nom d'Azure RTOS chez Microsoft, le projet a été transféré à la **Fondation Eclipse**.

ROS

Ensemble de bibliothèques, outils et services qui facilitent l'écriture de logiciels pour robots (capteurs, actuateurs, perception, planification, etc.)

Mbed OS

Système d'exploitation temps réel léger destiné aux microcontrôleurs, facilitant le développement d'applications IoT et d'IA embarquée ultra-contraintes, avec une gestion fine de l'énergie et des périphériques.

Middleware robotique – ROS 2

Framework logiciel distribué facilitant l'intégration de capteurs, d'algorithmes d'IA et d'actionneurs dans des systèmes robotiques embarqués, avec des garanties de communication temps réel et de fiabilité.

6 Matériels pour l'IA embarquée

Les matériels constituent la fondation physique de l'IA embarquée. Ils déterminent les capacités de calcul, la consommation énergétique, la latence et la fiabilité des systèmes. Sans une adéquation fine entre les composants matériels et les modèles logiciels, l'IA embarquée ne peut répondre aux contraintes de performance, de frugalité et de sécurité propres aux environnements critiques.

Ce chapitre explore les principales briques matérielles :

- Processeurs et accélérateurs spécialisés (CPU, GPU, FPGA, ASIC, NPU) qui assurent l'exécution des modèles d'IA dans des conditions embarquées.
- Circuits et architectures dédiés (neuromorphiques, hétérogènes, temps réel) qui optimisent la performance et l'efficacité énergétique.
- Mémoire et interconnexions qui garantissent la fluidité des flux de données et la rapidité des traitements.

L'objectif est de proposer une cartographie claire et contextualisée des matériels mobilisés dans l'IA embarquée, en montrant leur rôle dans la mise en œuvre des architectures présentées au chapitre 6 et leur articulation avec les outils logiciels du chapitre 7.

6.1 Architectures matérielles et paradigme de calcul

Les architectures matérielles et les paradigmes de calcul constituent les fondations physiques et computationnelles de l'IA embarquée. Ils déterminent la manière dont les modèles sont représentés, exécutés et optimisés sur des plateformes contraintes, en influençant directement les performances, la consommation énergétique et la capacité à respecter des exigences temps réel.

Dans le contexte de l'IA embarquée, ces notions couvrent à la fois les interfaces fondamentales entre logiciel et matériel, les philosophies de conception des processeurs, les mécanismes de calcul parallèle et vectoriel, ainsi que les structures de données manipulées par les modèles. Elles fournissent le cadre nécessaire pour comprendre comment les calculs d'IA sont effectivement réalisés sur des cibles embarquées hétérogènes, du microcontrôleur aux systèmes sur puce spécialisés.

Instruction Set Architecture (ISA)

Spécification définissant l'ensemble des instructions, registres, modes d'adressage et le modèle d'exécution visibles par le logiciel, constituant l'interface entre le matériel et les compilateurs, systèmes d'exploitation et runtimes d'IA.

Reduced Instruction Set Computer (RISC)

Philosophie de conception des processeurs reposant sur un jeu d'instructions simple et régulier, favorisant le pipeline, le parallélisme et l'efficacité énergétique, particulièrement adaptée aux systèmes embarqués.

Open Hardware

Approche de conception matérielle dans laquelle les spécifications, interfaces et parfois les descriptions matérielles sont ouvertes et librement accessibles, permettant la modification, la redistribution et la personnalisation des architectures matérielles, notamment pour les processeurs et accélérateurs IA.

RISC-V

ISA ouverte et libre de droits, fondée sur les principes RISC et conçue de manière modulaire, permettant l'ajout d'extensions spécialisées (calcul vectoriel, sécurité, accélération IA) pour couvrir un large spectre de processeurs, du microcontrôleur aux SoC pour l'IA embarquée.

ARM

Famille d'architectures de processeurs RISC propriétaires, largement utilisées dans les systèmes embarqués et mobiles, bénéficiant d'un écosystème logiciel mature et d'extensions dédiées au calcul vectoriel et à l'IA.

Tensor Operator Set Architecture (TOSA)

Spécification d'opérateurs tensoriels de haut niveau proposée par ARM et utilisée comme format intermédiaire dans les chaînes de compilation pour exprimer des opérations de réseaux de neurones indépendamment du matériel.

Tenseur – Tensor

Structure de données multidimensionnelle utilisée pour représenter les entrées, poids et sorties d'un modèle de deep learning. Un tenseur peut être vu comme une généralisation des matrices à un nombre arbitraire de dimensions (par exemple un tenseur 4D pour une image RGB en batch).

Disposition des données – Data layout

Organisation en mémoire des dimensions d'un tenseur. Les deux formats les plus courants sont **NCHW** (batch, canaux, hauteur, largeur) et **NHWC** (batch, hauteur, largeur, canaux). Le choix du layout influence fortement les performances selon l'architecture matérielle et la bibliothèque matérielle utilisée.

- **Lot (Batch)** : Ensemble d'échantillons traités simultanément par le modèle lors d'une itération d'entraînement ou d'inférence. Le traitement par lots améliore l'efficacité du calcul en exploitant le parallélisme matériel.
- **Canaux (Channels)** : Dimension représentant les différentes composantes d'une donnée. Pour une image, cela correspond généralement aux **canaux de couleur** (ex. : R, G, B). Pour un signal audio ou des données capteurs, cela peut représenter différentes sources ou dimensions de mesure.
- **Hauteur (Height)** : Taille verticale d'une donnée bidimensionnelle (par exemple, le nombre de pixels verticaux dans une image).
- **Largeur (Width)** : Taille horizontale d'une donnée bidimensionnelle (par exemple, le nombre de pixels horizontaux dans une image).

Multiply–Accumulate (MAC)

Unité de calcul effectuant une multiplication suivie d'une accumulation (a × b + c) en une seule opération, brique fondamentale des calculs matriciels et convolutifs en IA.

Calcul vectoriel – Vector Computation

Mode de calcul exploitant des registres et instructions vectorielles pour accélérer les opérations parallèles.

Single Instruction Multiple Data (SIMD)

Paradigme d'exécution dans lequel une instruction opère simultanément sur plusieurs données.

6.2 Processeurs et accélérateurs spécialisés

Les processeurs et accélérateurs spécialisés constituent le socle matériel sur lequel s'exécute l'IA embarquée. Ils déterminent la capacité des systèmes à traiter localement des modèles d'intelligence artificielle sous des contraintes strictes de performance, de latence, de consommation énergétique et de coût.

Dans le contexte de l'IA embarquée, le choix et la combinaison de ces unités de calcul conditionnent les architectures possibles, les techniques d'optimisation applicables et les cas d'usage adressables. Les notions présentées dans ce chapitre permettent de comprendre le rôle et la complémentarité des différents types de processeurs et d'accélérateurs mobilisés pour l'exécution efficace de l'IA au plus près du système physique.

Accélérateur IA embarqué

Unité de calcul matérielle dédiée à l'exécution efficace de modèles d'IA (réseaux neuronaux, calcul tensoriel), optimisée pour réduire la latence, la consommation énergétique et l'empreinte mémoire sur des plateformes embarquées.

Accélérateur neuronal ARM Ethos

Famille d'accélérateurs matériels conçus pour l'exécution de modèles d'IA embarquée sur architectures ARM, optimisés pour les réseaux neuronaux quantifiés et l'inférence à faible consommation énergétique.

Accélérateur TPU embarqué – Edge TPU

Accélérateur matériel spécialisé pour l'inférence de modèles d'IA embarquée, optimisé pour les opérations tensorielles et la quantification, permettant une exécution locale rapide et économe en énergie sur des dispositifs edge.

SoC IA embarqué Kendryte

Famille de systèmes sur puce intégrant des cœurs de calcul généralistes et des accélérateurs IA dédiés, illustrant l'intégration étroite du calcul IA dans des plateformes embarquées à faible consommation.

Central Processing Unit (CPU)

Processeur généraliste exécutant les instructions logiques et arithmétiques.

Microcontroller Unit (MCU)

Processeur intégré à faible consommation, adapté aux tâches simples ou modèles IA ultra-légers.

Graphics Processing Unit (GPU)

Processeur parallèle optimisé pour les calculs matriciels, utilisé pour l'entraînement et l'inférence IA.

Neural Processing Unit (NPU)

Accélérateur matériel généraliste conçu pour exécuter efficacement des modèles de deep learning, pas seulement des CNN. Il optimise les opérations typiques des réseaux neuronaux : multiplications matricielles, convolutions, activations, quantification, etc.

Convolutional Neural Network Accelerator (CNN Accelerator)

Cet accélérateur de réseau de neurones convolutifs est un bloc matériel dédié spécifiquement aux opérations de convolution et aux couches typiques des CNN. Il est plus spécialisé qu'une NPU, souvent plus simple, plus économe, mais moins polyvalent.

Digital Signal Processor (DSP)

Processeur spécialisé dans le traitement de signaux numériques (audio, capteurs).

Tensor Processing Unit (TPU)

Accélérateur IA conçu par Google pour les calculs tensoriels à grande échelle.

Vision Processing Unit (VPU)

Processeur dédié à la vision par ordinateur embarquée.

Processeur neuromorphique – Neuromorphic Processor

Circuit inspiré du fonctionnement du cerveau humain, il repose sur des architectures composées de neurones et synapses matériels, souvent implémentés sous forme de circuits événementiels (spiking) qui traitent l'information de façon asynchrone et massivement parallèle. Pour les systèmes embarqués, ces architectures offrent une très forte efficacité énergétique, une faible latence et un traitement local des données, ce qui les rend particulièrement adaptés aux applications contraintes en puissance et en temps réel, comme la perception et le contrôle autonome.

6.3 Circuits spécialisés et architectures embarquées

Les architectures matérielles intégrées définissent la manière dont les composants de calcul, de mémoire et d'interconnexion sont organisés physiquement pour exécuter des applications d'IA embarquée. Elles conditionnent la performance globale, l'efficacité énergétique, la compacité et la capacité d'intégration des systèmes intelligents dans des environnements contraints.

Dans le contexte de l'IA embarquée, ces architectures couvrent aussi bien les supports physiques et modulaires que les circuits intégrés spécialisés ou reconfigurables. Elles offrent différents compromis entre flexibilité, spécialisation et performances, et constituent un levier clé pour adapter les solutions d'IA aux exigences opérationnelles des systèmes embarqués.

Printed Circuit Board (PCB)

Carte à circuits imprimés sur laquelle sont assemblés les composants électroniques (CPU, GPU, mémoire, interfaces, etc.), servant de support physique et d'interconnexion pour les systèmes embarqués et modules IA.

Field Programmable Gate Array (FPGA)

Circuit reconfigurable matériellement, permettant une personnalisation fine des fonctions IA.

Application-Specific Integrated Circuit (ASIC)

Puce dédiée à une tâche spécifique, très efficace pour l'IA embarquée.

Massively Parallel Processor Array (MPPA)

Architecture parallèle composée de centaines de cœurs, adaptée aux applications critiques.

System on Chip (SoC)

Circuit intégré regroupant sur une seule puce CPU, GPU, NPU, mémoire et interfaces de communication, offrant une solution compacte et performante.

Network-on-Chip (NoC)

Infrastructure d'interconnexion interne à un SoC permettant la communication rapide entre processeurs, caches, accélérateurs et mémoires, essentielle pour exploiter efficacement la hiérarchie mémoire et maximiser le débit de données.

Architecture hétérogène – Heterogeneous Architecture

Plateformes matérielles combinant plusieurs types d'unités de calcul spécialisées, processeurs et accélérateurs au sein d'un même système. Cette organisation permet d'adapter chaque tâche au moteur de calcul le plus efficace, en optimisant le compromis entre performances, consommation énergétique et latence. Pour les systèmes embarqués contraints, les architectures hétérogènes offrent ainsi une bonne efficacité énergétique, une certaine flexibilité d'exécution et la capacité de répondre aux exigences du temps réel.

Chiplet

Circuit intégré modulaire qui contient un sous-ensemble de fonctionnalités bien défini (par exemple, des cœurs de processeur, de la mémoire, des entrées/sorties).

Computer on Module (CoM)

Carte électronique compacte intégrant processeur, mémoire et interfaces, conçue pour être intégrée dans des systèmes embarqués ; utilisé pour les architectures modulaires d'IA embarquée.

System on Module (SoM)

Module embarqué plus complet que le CoM, incluant souvent des composants spécifiques (GPU, FPGA, interfaces industrielles, …) ; utilisé pour les applications embarquées exigeantes (vision, robotique, …).

Intelligence Processing Unit (IPU)

Architecture parallèle conçue pour l'IA, optimisée pour les graphes de calcul.

Reconfigurable Processing Unit (RPU)

Processeur adaptable à la tâche souvent basé sur FPGA pour personnaliser l'accélération IA

6.4 Mémoire et interconnexion

Les mécanismes de communication, de gestion mémoire et d'interconnexion jouent un rôle central dans l'exécution efficace de l'IA embarquée. Ils assurent la circulation des données entre capteurs, processeurs, accélérateurs et modèles, tout en maîtrisant les contraintes de latence, de bande passante et de consommation énergétique.

Dans le contexte de l'IA embarquée, ces éléments constituent l'infrastructure invisible mais déterminante permettant d'orchestrer les calculs, de stocker les données et d'optimiser les transferts. Les notions présentées dans ce chapitre fournissent les clés pour comprendre comment middleware, mémoires et interfaces matérielles contribuent à la performance et à la fiabilité des systèmes intelligents embarqués.

Middleware
Logiciel intermédiaire qui assure la communication, la coordination et la gestion des données entre différentes applications, systèmes ou couches d'un environnement informatique. Il fournit des services communs (comme la gestion des messages, des transactions ou des ressources) permettant aux composants logiciels de fonctionner ensemble sans dépendances directes.

Static and Dynamic RAM (SRAM and DRAM)
Mémoire vive rapide pour le traitement temporaire des données IA.

Non-Volatile Memory (NVM)
Mémoire qui conserve les données même sans alimentation électrique. Elle s'oppose à la RAM (volatile).

Flash / eMMC / UFS / NAND
Types de mémoire non volatile (NVM) pour le stockage des modèles IA, firmware, logs.

Magnetoresistive and Resistive RAM (MRAM / ReRAM)
Mémoires émergentes, non volatiles, rapides et économes.

Caches L1 / L2 / L3 – L1 / L2 / L3 Caches
Mémoires rapides intégrées au processeur pour réduire la latence des accès aux données et instructions, organisées en niveaux hiérarchiques (L1 proche du cœur, L3 partagé).

Direct Memory Access (DMA)
Mécanisme matériel permettant le transfert de données entre mémoire et périphériques ou accélérateurs sans solliciter le CPU, réduisant la latence et la consommation.

In-Memory Computing (IMC)
Paradigme matériel dans lequel le calcul (souvent des opérations matricielles ou logiques) est effectué directement dans les cellules de mémoire, minimisant les transferts de données et la consommation d'énergie, et améliorant les performances pour l'inférence IA embarquée.

Interfaces de communication PCIe / MIPI / UART / SPI / I2C
Interfaces permettant le transfert de données et la communication entre processeurs, capteurs et modules IA dans les systèmes embarqués.

7 Communications & connectivité dans l'embarqué

La communication et la connectivité sont des éléments essentiels de l'IA embarquée. Elles permettent aux dispositifs de collecter, échanger et synchroniser des données, mais aussi de coopérer dans des environnements distribués, de relier capteurs, processeurs et infrastructures dans un continuum edge–fog–cloud. Cette connectivité, à la fois locale, distribuée et globale, prépare le terrain pour les évolutions du chapitre 8, où l'IA embarquée de demain se dessine à travers de nouvelles formes d'intelligence, de frugalité et de souveraineté technologique.

Dans le contexte de l'IA embarquée, la connectivité doit répondre à des contraintes spécifiques :

- **Latence et temps réel** : garantir des échanges rapides et fiables pour des applications critiques.
- **Frugalité énergétique** : optimiser les protocoles et les flux pour limiter la consommation des dispositifs.
- **Sécurité et fiabilité** : assurer la protection des données transmises et la robustesse des communications face aux perturbations.
- **Interopérabilité** : permettre l'intégration de multiples standards et protocoles (CAN, SPI, I²C, UART, MIPI, Ethernet, 5G, etc.).

Ce chapitre propose une cartographie des principaux protocoles, bus, réseaux et mécanismes de connectivité mobilisés dans l'IA embarquée, en montrant leur rôle dans la mise en relation des capteurs, des processeurs, des systèmes distribués et des infrastructures edge-fog-cloud.

7.1 Réseaux physiques

Les réseaux physiques constituent l'infrastructure matérielle (filaire ou radio) de la connectivité embarquée. Ils définissent comment les données circulent entre capteurs, processeurs et infrastructures.

Bus internes – Internal buses

Standards universels pour relier capteurs et processeurs à courte distance, omniprésents dans l'automobile, l'électronique embarquée et l'industrie pour leur fiabilité et leur faible coût. Ils assurent des échanges rapides, fiables et adaptés aux contraintes temps réel des systèmes embarqués. Par exemple : I²C, SPI, UART, CAN.

Réseaux industriels et Ethernet (Ethernet industriel, ProfiBus)

Technologies de communication filaires ou hybrides conçues pour les environnements industriels, où la fiabilité, la robustesse et le déterminisme sont essentiels. Ils assurent l'échange de données entre automates, capteurs, actionneurs et systèmes de contrôle, souvent dans des environnements soumis à des contraintes fortes (bruit électromagnétique, vibrations, températures élevées). Ils garantissent des communications stables, sûres et prévisibles, indispensables aux systèmes embarqués industriels et aux applications temps réel.

Réseaux sans fil courte portée – Short-range wireless networks

Technologies radio locales, adaptées aux environnements contraints et à faible consommation, très utilisées dans la domotique, les bâtiments intelligents et les objets portables. Ils assurent la communication locale entre capteurs et passerelles. Ils favorisent la mobilité et l'intégration dans des environnements contraints, avec une faible consommation énergétique. Par exemple : BLE, Zigbee, Thread.

Réseaux sans fil longue portée – Long-range wireless networks

Connectivité étendue entre dispositifs et infrastructures, comme par exemple : Wi-Fi, LTE/4G, 5G URLLC, NB-IoT, LTE-M, LoRaWAN. La 5G URLLC est essentielle pour les applications critiques nécessitant faible latence, tandis que NB-IoT, LTE-M et LoRaWAN ciblent les déploiements massifs d'objets connectés à faible consommation. Selon le protocole, ils privilégient le haut débit, la couverture ou l'efficacité énergétique, et permettent l'intégration edge–cloud.

Mesh networks

Architectures maillées permettant la résilience et l'extension de la couverture, utilisées dans les réseaux de capteurs distribués et les environnements urbains intelligents. 6LoWPAN, Thread mesh.

7.2 Protocoles applicatifs

Les protocoles applicatifs sont des couches logicielles qui définissent comment les données sont structurées, échangées et interprétées au-dessus des réseaux physiques. Ils assurent l'interopérabilité et la cohérence des systèmes distribués.

Protocoles IoT légers – Lightweight IoT protocols

Protocoles conçus pour la transmission frugale et scalable des données capteurs, largement adoptés dans les réseaux IoT.

Message Queuing Telemetry Transport (MQTT)

Protocole très léger basé sur le modèle publish/subscribe, conçu pour les environnements contraints en bande passante et énergie, massivement adopté dans l'IoT (domotique, capteurs distribués, télémétrie).

CoAP (Constrained Application Protocol)

Protocole inspiré de HTTP, mais optimisé pour les appareils à ressources limitées. Il fonctionne sur UDP, avec un modèle requête/réponse. Il est utilisé dans l'automatisation industrielle, la surveillance environnementale et les capteurs très contraints.

Advanced Message Queuing Protocol (AMQP)

Protocole riche, orienté messages, avec garanties de livraison et sécurité. Il est plus lourd que MQTT ou CoAP, donc moins adapté aux objets très contraints ; il est utilisé dans les environnements d'entreprise (finance, systèmes critiques, cloud).

Modbus

Protocole historique (années 1970), simple et robuste, basé sur un modèle maître-esclave, encore très répandu dans l'industrie, notamment pour connecter automates programmables (PLC), capteurs et actionneurs. Sa simplicité et son adoption massive en font une référence universelle, mais il reste limité en termes de sécurité et de richesse de données.

Open Platform Communications Unified Architecture (OPC UA)

Protocole moderne, orienté objet, conçu pour l'interopérabilité et la normalisation des échanges industriels. Il est largement adopté dans l'automatisation, l'industrie 4.0 et les architectures edge–cloud, car il permet de relier des systèmes hétérogènes. Il est extensible, sécurisé (intègre chiffrement et authentification), supporte des modèles de données riches et hiérarchiques.

Interopérabilité edge–fog–cloud – Edge–fog–cloud interoperability

Mécanismes normalisés permettant de fluidifier les échanges entre couches hétérogènes (edge, fog, cloud) et d'assurer la compatibilité entre systèmes, équipements et services distribués. Ils garantissent la continuité opérationnelle dans les architectures hybrides et facilitent l'intégration multi-technologies. Exemples : API REST, gRPC, OPC-UA.

REST

Interface de communication basée sur HTTP, utilisant des ressources identifiées par des URL ; il est universel et simple et s'impose comme le standard du web et du cloud pour exposer et consommer des services.

Google Remote Procedure Call (gRPC)

Protocole moderne basé sur HTTP/2 et Protobuf, orienté appels de fonctions distantes, qui répond aux besoins des architectures distribuées et des microservices, avec des échanges rapides et efficaces, y compris en streaming bidirectionnel.

8 Paradigmes avancés et émergents de l'IA embarquée

L'IA embarquée est une discipline en évolution rapide, portée par des avancées algorithmiques, matérielles et architecturales. Au-delà des approches classiques de l'apprentissage automatique et de l'inférence embarquée, de nouveaux paradigmes émergent et ouvrent des perspectives inédites pour les systèmes intelligents déployés sur des plateformes contraintes.

Ces paradigmes ne redéfinissent pas les fondements de l'intelligence artificielle introduits dans les chapitres précédents. Ils prolongent et enrichissent ces concepts en explorant des trajectoires avancées propres à l'IA embarquée, caractérisées par une plus grande autonomie, une intégration renforcée avec le monde physique et une prise en compte explicite des contraintes de ressources et d'impact environnemental.

8.1 IA générative embarquée

Les notions associées à l'IA générative embarquée décrivent les mécanismes permettant la production locale de contenus et de représentations à partir de modèles appris, dans un contexte de ressources limitées. Elles couvrent les types de modèles mobilisés, les moyens de guider la génération, les formes d'adaptation de modèles génériques à l'embarqué, ainsi que les risques spécifiques liés à la génération autonome de contenus. Ces éléments permettent de caractériser les usages génératifs compatibles avec l'exécution sur des plateformes embarquées et de les distinguer des approches centralisées.

IA générative – Generative AI

Approche de l'intelligence artificielle visant à produire du contenu nouveau (texte, image, signal, code) à partir de modèles appris. Dans le contexte de l'IA embarquée, l'IA générative est mise en œuvre sous des formes spécialisées et contraintes, permettant une génération locale adaptée aux limitations de calcul, de mémoire, de latence et de consommation énergétique des plateformes embarquées.

Système d'IA à usage général – General-Purpose AI System (GPAIS)

Système d'IA conçu pour être adaptable à une variété de tâches. En IA embarquée, ces systèmes sont généralement spécialisés, distillés ou restreints à des fonctions ciblées afin d'être exécutables localement et intégrables dans des applications embarquées spécifiques.

Modèle de fondation – Foundation Model

Modèle d'intelligence artificielle de grande taille, entraîné de manière auto-supervisée sur de vastes ensembles de données. Dans l'IA embarquée, les modèles de fondation servent principalement de point de départ à des déclinaisons compactes ou spécialisées, adaptées à l'inférence locale sur des systèmes contraints.

Ingénierie des prompts – Prompt Engineering

Démarche consistant à concevoir et structurer des requêtes permettant de guider le comportement d'un modèle génératif. En IA embarquée, le prompt engineering vise à maximiser la pertinence, la robustesse et la prévisibilité des sorties générées, tout en maîtrisant les coûts de calcul et de mémoire.

Retrieval-Augmented Generation (RAG)

Technique combinant un modèle génératif avec un mécanisme de recherche dans une base de connaissances. En IA embarquée, le RAG repose sur des bases locales, embarquées ou partiellement synchronisées, afin de limiter la dépendance à une connectivité distante tout en enrichissant les réponses générées.

Embedding

Représentation vectorielle compacte d'objets tels que des mots, phrases, images ou signaux, organisée de manière à refléter leurs similarités sémantiques ou fonctionnelles. Les embeddings sont utilisés en IA embarquée pour la recherche locale, la classification, la similarité ou la génération assistée.

Jeton – Token

Unité élémentaire de traitement utilisée par les modèles de langage pour représenter des fragments de texte ou de données symboliques. En IA embarquée, la gestion du nombre et de la taille des tokens est un levier essentiel pour maîtriser la latence et l'empreinte mémoire.

Hallucination

Production par un modèle génératif d'une réponse incorrecte, incohérente ou sans fondement factuel, présentée comme plausible. Dans les systèmes d'IA embarquée, les hallucinations constituent un risque critique, notamment dans les domaines de la santé, de la mobilité ou de la défense, et nécessitent des mécanismes de limitation, de détection ou de validation.

Joint-Embedding Predictive Architecture (JEPA)

Architecture d'apprentissage auto-supervisé visant à prédire l'évolution future d'un état ou d'une scène dans un espace latent, plutôt qu'à générer directement du contenu. JEPA est particulièrement adaptée à l'IA embarquée pour construire des modèles du monde permettant l'anticipation, la planification et l'adaptation à des environnements dynamiques.

Hierarchic Reasoning Model (HRM) et Tiny Recursive Model (TRM)

Deux architectures de modèles de petite taille, respectivement proposées par Sapient Intelligence et par Samsung, qui utilisent le raisonnement hiérarchique en boucle ou récursif pour résoudre des problèmes. Ces modèles sont particulièrement économes en espace et en temps de calcul et peuvent avoir un intérêt pour l'embarqué.

World models embarqués – Embedded World Models

Représentations internes permettant à un système d'IA embarquée de modéliser l'état courant et l'évolution probable de son environnement. Ces modèles soutiennent des capacités d'anticipation et de raisonnement essentielles pour les systèmes autonomes.

8.2 IA AGENTIQUE ET SYSTÈMES AUTONOMES

Les concepts relatifs à l'IA agentique décrivent les capacités d'autonomie décisionnelle des systèmes embarqués intelligents. Ils permettent de formaliser la notion d'agent, les mécanismes de prise de décision orientée objectif, les interactions entre agents et les architectures distribuées. Ces notions servent à décrire des systèmes embarqués capables d'agir de manière proactive, coordonnée et adaptative, en interaction directe avec leur environnement, sans dépendance permanente à une supervision externe.

Agentivité – Agency

Capacité d'un système d'IA à initier des actions, prendre des décisions et poursuivre des objectifs de manière autonome dans un environnement donné, sans supervision humaine continue.

Agent IA – AI Agent

Entité logicielle ou matérielle autonome capable de percevoir son environnement, de raisonner sur la base de ses observations et d'agir afin d'atteindre un objectif. En IA embarquée, un agent est généralement déployé sur une plateforme contrainte et agit en interaction directe avec le monde physique.

Systèmes multi-agents – Multi-Agent Systems (MAS)

Ensemble d'agents interagissant au sein d'un environnement partagé, par coopération, coordination ou négociation. En IA embarquée, ces systèmes sont utilisés pour la robotique collective, les réseaux de capteurs intelligents ou les dispositifs coopératifs distribués.

Architectures agentiques – Agentic Architectures

Architectures reposant sur des agents autonomes capables de perception, de planification et d'action, favorisant une prise de décision distribuée et contextuelle sur des systèmes embarqués.

Approche orientée objectif – Goal-Driven Approach

Paradigme dans lequel un agent agit en fonction d'objectifs explicites, en évaluant différentes options d'action et en adaptant sa stratégie en fonction de l'évolution de l'environnement.

Boucle perception–action – Perception–action loop

Cycle continu reliant la perception de l'environnement, la prise de décision et l'action, caractéristique des systèmes embarqués autonomes capables de s'adapter en temps réel à leur contexte.

Orchestration de l'IA – AI Orchestration

Coordination de plusieurs agents ou modèles d'IA afin de réaliser des tâches complexes de manière cohérente. Dans l'IA embarquée, l'orchestration concerne principalement des systèmes coopératifs distribués fonctionnant avec des ressources et une connectivité limitées.

8.3 IA PHYSIQUE ET INTERACTION AVEC LE MONDE RÉEL

Les notions associées à l'IA physique concernent l'incarnation matérielle de l'intelligence artificielle dans des systèmes embarqués capables de percevoir et d'agir sur le monde réel. Elles permettent de qualifier les formes d'intégration entre algorithmes, capteurs et actionneurs, ainsi que les outils de représentation et de simulation du comportement physique. Ces concepts sont essentiels pour décrire des systèmes où l'IA embarquée influence directement des processus physiques et mécaniques.

IA physique – Physical AI

Forme d'IA intégrée dans des dispositifs matériels capables de percevoir, décider et agir sur le monde réel. L'IA physique combine perception sensorielle, raisonnement et action motrice, et est optimisée pour fonctionner en temps réel sous contraintes embarquées.

Jumeau numérique – Digital Twin

Représentation virtuelle fidèle d'un système physique permettant la simulation, l'analyse et l'optimisation de son comportement. Dans l'IA embarquée, les jumeaux numériques sont utilisés pour anticiper les comportements, améliorer la résilience et soutenir la maintenance prédictive.

8.4 ÉCO-CONCEPTION ET SOBRIÉTÉ DE L'IA EMBARQUÉE

Les notions liées à l'éco-conception et à la sobriété formalisent des approches de conception de l'IA embarquée intégrant explicitement les impacts environnementaux et les contraintes de ressources. Elles décrivent les principes permettant d'évaluer, de réduire et de maîtriser la consommation énergétique, l'usage du calcul et des données, sur l'ensemble du cycle de vie des systèmes embarqués intelligents. Ces notions traduisent le passage des contraintes techniques de l'embarqué à des objectifs de conception assumés.

Éco-conception – Eco-design

Démarche consistant à intégrer la réduction des impacts environnementaux dès la conception des systèmes d'IA embarquée, sur l'ensemble de leur cycle de vie, en tenant compte des contraintes de calcul, d'énergie et de matériel.

IA verte – Green AI

Approche de l'intelligence artificielle visant à réduire l'impact environnemental des modèles et des systèmes d'IA, notamment en termes de consommation énergétique, d'émissions carbone et d'utilisation des ressources.

IA frugale – Frugal AI

Démarche de conception et d'usage de l'IA fondée sur la sobriété des moyens, la limitation des données et du calcul, et l'adéquation entre le besoin fonctionnel et la solution déployée, particulièrement adaptée aux systèmes embarqués contraints.

Sobriété computationnelle – Computational Sobriety

Principe visant à limiter le volume de calcul nécessaire à l'exécution d'un modèle, en privilégiant des architectures légères, des algorithmes efficaces et des stratégies d'inférence adaptées.

Co-optimisation matériel-logiciel – Hardware-software co-optimization

Approche consistant à co-concevoir le modèle, les algorithmes et l'architecture matérielle afin de réduire la consommation, le calcul et la mémoire, tout en garantissant les performances fonctionnelles.

Conception thermique – Thermal-Aware Design

Prise en compte des contraintes thermiques dans la conception des modèles et des pipelines d'inférence, afin d'éviter les surchauffes, throttling et dégradations de performance.

Conception durable – Sustainable Design

Conception visant à maximiser la durée de vie matérielle, la réparabilité, la modularité et la réutilisation, tout en minimisant l'impact environnemental global.

Économie circulaire – Circular economy

Stratégie visant à réduire les déchets et réutiliser les composants.

Impact environnemental – Environmental impact

Effet sur l'environnement lié à la production, utilisation et fin de vie d'un système.

Analyse du cycle de vie (ACV) – Life Cycle Assessment (LCA)

Évaluation des impacts environnementaux d'un produit ou système sur toutes les étapes, de la fabrication à la fin de vie.

Analyse multicritère – Multi-criteria Eco-assessment

Méthode d'évaluation intégrant simultanément performance, énergie, mémoire, carbone, coût matériel et durée de vie.

Conception orientée contrainte – Constraint-Driven Design

Approche de conception guidée par les contraintes de ressources (énergie, mémoire, calcul, thermique), plutôt que par la seule performance algorithmique.

9 Tests, validation & certification de l'IA embarquée

L'essor de l'IA embarquée, et plus encore ses évolutions futures (IA générative, agentique, edge-to-cloud, architectures émergentes), impose de garantir la **fiabilité, la sécurité et la conformité** des systèmes avant leur déploiement dans des environnements critiques. Les tests, la validation et la certification constituent ainsi des étapes incontournables pour instaurer la confiance et assurer l'acceptabilité industrielle, réglementaire et sociétale.

Ce chapitre propose une cartographie des méthodes de test, des métriques de validation et des dispositifs de certification, en montrant leur rôle central dans la crédibilité et la diffusion de l'IA embarquée.

Ces pratiques répondent à plusieurs enjeux majeurs :

- **Fiabilité technique** : vérifier la robustesse des modèles et des architectures dans des conditions réelles.
- **Sécurité et sûreté** : garantir que les systèmes embarqués ne présentent pas de risques pour les utilisateurs ou l'environnement.
- **Conformité réglementaire** : répondre aux normes européennes et internationales encadrant l'IA et les systèmes critiques.
- **Certification sectorielle** : obtenir des labels ou agréments spécifiques (automobile, aéronautique, médical, etc.) pour valider l'usage de l'IA embarquée.

9.1 Métriques

Les métriques constituent un outil fondamental pour caractériser, comparer et maîtriser les performances des systèmes d'IA embarquée. Elles permettent d'évaluer de manière objective la capacité d'un modèle à fonctionner efficacement sous contraintes, en prenant en compte non seulement la qualité des résultats produits, mais aussi les ressources matérielles mobilisées et les exigences temporelles et énergétiques.

Dans le contexte de l'IA embarquée, les métriques jouent un rôle structurant dans les choix d'architecture, d'optimisation et de déploiement. Elles permettent d'arbitrer entre précision, latence, consommation énergétique et usage mémoire, afin de garantir un comportement prévisible, robuste et compatible avec les environnements embarqués.

9.1.1 Métriques de mémoire

Empreinte mémoire – Memory Footprint
Quantité totale de mémoire RAM nécessaire pour charger et exécuter une inférence du modèle sur la cible matérielle.

Taille du modèle – Model Size
Taille du fichier contenant les poids du modèle IA, exprimée en kilo-octets ou méga-octets (voire en giga-octets). Elle détermine l'espace de stockage requis sur l'appareil.

Bande passante mémoire – Memory Bandwidth

Débit maximum de transfert de données entre la mémoire et le processeur, exprimé en gigaoctets par seconde (GB/s).

9.1.2 Métriques de performance temporelle

Latence – Latency

Délai entre la réception d'une donnée et la production du résultat par le système, généralement exprimée en millisecondes (ms) ou microsecondes (µs). Elle est critique dans les applications embarquées temps réel, où une décision tardive peut compromettre la sécurité ou la performance.

Latence extrême – Tail latency

Latence maximale observée pour un pourcentage donné des inférences (ex. 99e percentile), utilisée pour garantir la prévisibilité en temps réel.

Worst-Case Execution Time (WCET)

Durée maximale d'exécution garantie qu'un modèle IA peut prendre dans le pire des cas sur un système donné. Le WCET assure le respect des contraintes temps réel strictes.

Débit – Throughput

Quantité de données traitées par le système durant une période donnée, généralement exprimée en images par seconde (frame/s), échantillons par seconde, ou opérations par seconde (ops/s). Dans le contexte de l'IA embarquée, le débit mesure la vitesse d'inférence et permet de qualifier la réactivité globale du système.

9.1.3 Métriques énergétiques

Consommation énergétique – Power Consumption

Quantité d'énergie consommée par le système durant son fonctionnement, exprimée en milliwatts (mW) ou watts (W). C'est une contrainte majeure dans les dispositifs autonomes ou alimentés par batterie.

Énergie par inférence – Energy per inference

Quantité d'énergie nécessaire pour effectuer une seule inférence, exprimée en microjoules (µJ) ou millijoules (mJ). Cette métrique est essentielle pour évaluer l'efficacité énergétique globale sur la durée de vie de la batterie.

9.1.4 Métriques de précision

Précision – Accuracy

Proportion de prédictions correctes parmi l'ensemble des prédictions effectuées par le modèle.

Précision Top-1 – Top-1 Accuracy

Taux de prédictions où la classe prédite en première position correspond à la vérité terrain.

Précision Top-5 – Top-5 Accuracy

Taux de prédictions où la vérité terrain est incluse parmi les 5 premières classes prédites.

Intersection sur Union – Intersection over Union (IoU)

Métrique d'évaluation utilisée en détection et segmentation d'objets pour mesurer le recouvrement entre la zone prédite par le modèle et la zone réelle. Une valeur élevée indique une bonne précision de localisation.

Précision Moyenne – Mean Average Precision (mAP)

Moyenne de la précision sur différents seuils de chevauchement (Intersection over Union), utilisée pour évaluer les modèles de détection d'objets.

Intersection sur Union Moyenne – Mean Intersection over Union - (mIoU)

Moyenne du recouvrement entre les zones prédites et les zones réelles, utilisée pour évaluer les modèles de segmentation d'images.

Score F1 – F-mesure

Mesure de performance utilisée pour évaluer un classifieur, qui fournit un équilibre (la moyenne harmonique) entre le rappel et la précision. (définition Confiance.ai)

9.2 Benchmarks

Les benchmarks constituent des outils de référence pour évaluer et comparer de manière objective les performances des systèmes d'IA embarquée. Ils fournissent des cadres standardisés permettant de mesurer l'efficacité des modèles, des chaînes logicielles et des plateformes matérielles dans des conditions représentatives des contraintes de l'embarqué.

Dans le contexte de l'IA embarquée, les benchmarks jouent un rôle clé pour qualifier les compromis entre précision, latence, consommation énergétique et usage des ressources. Ils permettent d'orienter les choix technologiques, de comparer des solutions hétérogènes et de disposer de résultats reproductibles pour l'évaluation des matériels et des chaînes d'inférence déployées en périphérie.

MLPerf Edge et MLPerf Tiny

Benchmarks standardisés de MLCommons pour l'évaluation des performances de l'inférence IA sur plateformes embarquées. MLPerf Edge cible les systèmes edge (SoC, CPU, GPU, NPU) et mesure latence, débit, énergie et précision, tandis que MLPerf Tiny est dédié aux microcontrôleurs et environnements TinyML à fortes contraintes de ressources.

Ces benchmarks servent de référence industrielle pour comparer matériels et chaînes d'inférence embarquées.

AIoT Benchmarks

Suite de benchmarks développée par BenchCouncil destinée à l'évaluation des performances des systèmes d'intelligence artificielle pour l'Internet des objets (AIoT). AIoTBench propose un ensemble de charges de travail représentatives couvrant différents scénarios d'IA embarquée, tels que la vision par ordinateur, le traitement audio et l'analyse de données issues de capteurs et permet de comparer les modèles, les plateformes logiciels et les capacités du matériel.

10 IA EMBARQUÉE DE CONFIANCE

La confiance dans l'IA embarquée repose sur un ensemble d'attributs tels que fiabilité, robustesse, équité, transparence, etc. qui peuvent notamment contribuer à la certification et à la qualification (voir également au chapitre sur les normes et standards). Les termes ci-dessous forment le socle lexical de l'IA de confiance. Ils couvrent les exigences générales, les processus (certification, qualification, audit), les outils (SBOM, traçabilité etc.) et les critères (preuves, robustesse, explicabilité etc.).

10.1 GÉNÉRALITÉS SUR L'ÉTHIQUE

L'intégration de fonctions d'intelligence artificielle dans des systèmes embarqués soulève des enjeux éthiques spécifiques, liés à l'autonomie décisionnelle, à l'interaction directe avec le monde physique et à l'exécution locale sous contraintes. Contrairement aux systèmes centralisés, l'IA embarquée agit souvent en temps réel, au plus près des utilisateurs, des infrastructures ou des environnements critiques, ce qui renforce les exigences de maîtrise, de responsabilité et de confiance.

Ce chapitre aborde les principes et mécanismes permettant d'encadrer le comportement des systèmes d'IA embarquée afin d'en garantir un usage responsable, transparent et maîtrisé. Il met l'accent sur la prévention des dérives, la limitation des biais, la gestion des situations non prévues et l'alignement des décisions automatisées avec les cadres éthiques, réglementaires et sociétaux applicables aux systèmes embarqués intelligents.

10.2 FIABILITÉ

La fiabilité constitue un pilier fondamental de l'IA embarquée de confiance, en particulier lorsque les systèmes intelligents sont intégrés à des environnements critiques, autonomes ou soumis à des exigences de sûreté élevées. Elle recouvre l'ensemble des propriétés permettant de garantir qu'un système d'IA embarquée se comporte de manière prévisible, correcte et maîtrisée tout au long de son fonctionnement.

Dans le contexte de l'IA embarquée, la fiabilité ne se limite pas aux composants logiciels ou matériels : elle dépend également de la qualité et de la provenance des données, de la robustesse face aux perturbations et attaques, ainsi que de la capacité du système à être maintenu et sécurisé dans la durée. Les notions présentées dans ce chapitre fournissent un cadre structurant pour analyser, concevoir et évaluer des systèmes d'IA embarquée capables d'inspirer la confiance en conditions opérationnelles.

Fiabilité - Dependability

La capacité de fournir un service auquel on peut faire confiance.

Cela implique :

- **Disponibilité (Availability)** : Disponibilité pour un service correct
- **Fiabilité (Reliability)** : continuité d'un service correct,
- **Confidentialité (Confidentiality)** : absence de divulgation non autorisée d'informations,
- **Intégrité (Integrity)** : absence d'altérations inappropriées du système,
- **Maintenabilité (Maintainability)** : capacité à subir des modifications et des réparations.

Sécurité – Security

Existence simultanée de la disponibilité pour les utilisateurs autorisés uniquement, de la confidentialité et de l'intégrité.

Robustesse – Robutness

Aptitude d'un système IA embarqué à fonctionner correctement malgré des perturbations, des incertitudes ou des attaques.

Justesse – Correctness

Degré auquel un système, un composant ou un élément logiciel (code, documentation, spécifications) est exempt de fautes et répond correctement aux exigences définies..

Qualité des données – Data Quality

Ensemble des propriétés (précision, cohérence, complétude, fraîcheur, bruit) conditionnant la fiabilité des inférences produites par un système d'IA embarquée et sa capacité à fonctionner de manière sûre et robuste en conditions réelles.

Provenance des données – Data Provenance

Origine et historique des données utilisées pour entraîner ou alimenter un système d'IA, essentielle pour garantir leur fiabilité et conformité.

10.3 Biais et équité

Biais

Déséquilibre ou une distorsion systématique dans les résultats produits par un modèle, souvent causé par des préjugés présents dans les données ou les algorithmes.

Le biais n'est pas une erreur ponctuelle, mais un schéma répétitif qui reflète des défauts structurels dans le système. Il survient lorsque les décisions ou prédictions d'un système d'intelligence artificielle sont injustes, discriminatoires ou inexactes envers certains groupes ou individus.

Cela peut être dû à des données d'entraînement biaisées (ex. : surreprésentation d'un groupe, absence de diversité), des algorithmes mal conçus qui amplifient des corrélations injustes, des préjugés humains intégrés involontairement lors de la conception ou de l'annotation des données**.**

Equité algorithmique – Fairness

Principe selon lequel les systèmes d'IA et d'apprentissage automatique doivent produire des décisions qui ne discriminent pas injustement des individus ou des groupes, notamment sur la base de caractéristiques sensibles comme le sexe, l'origine ethnique ou l'âge.

En pratique, l'équité algorithmique renvoie à l'ensemble des méthodes et critères utilisés pour détecter, mesurer et corriger les biais dans les modèles, afin que les résultats soient aussi équitables que possible entre différents groupes.

10.4 Explicabilité et interprétabilité

L'explicabilité et l'interprétabilité sont des composantes essentielles de l'IA embarquée de confiance, en particulier lorsque les systèmes intelligents sont intégrés dans des environnements critiques, autonomes ou réglementés. Elles permettent de comprendre, d'analyser et de justifier les décisions produites localement par un modèle d'IA exécuté sous contraintes matérielles et temporelles.

XAI - IA explicable et interprétable (XAI Explainable AI)
Ensemble de techniques permettant de comprendre et justifier les décisions prises par une IA.

Interprétabilité – Interpretability
Capacité à comprendre et expliquer les décisions ou prédictions d'un modèle d'intelligence artificielle.

Explicabilité – Explainability
Capacité à fournir une justification compréhensible des décisions ou prédictions d'un modèle IA, souvent exigée dans les processus de qualification et de certification.

Selon le NIST, l'explicabilité doit obéir à quatre principes :

- **Explications**: le système fournit ou contient des preuves ou des raisons accompagnant les résultats et/ou les processus.
- **Pertinence** : le système fournit des explications compréhensibles pour les consommateurs visés.
- **Exactitude** : l'explication reflète correctement la raison pour laquelle le résultat a été généré et/ou reflète fidèlement le processus du système.
- **Limites** des connaissances : le système ne fonctionne que dans les conditions pour lesquelles il a été conçu et lorsqu'il atteint un niveau de confiance suffisant dans ses résultats.

10.5 Transparence et traçabilité

La traçabilité et l'auditabilité sont des exigences structurantes pour la mise en œuvre d'une IA embarquée de confiance, en particulier dans les secteurs soumis à des contraintes réglementaires, de sûreté ou de certification. Elles permettent de documenter, suivre et vérifier l'ensemble des éléments constitutifs d'un système d'IA embarquée, depuis les données et les modèles jusqu'aux logiciels et composants intégrés.

Traçabilité – Traceability
Capacité à suivre l'origine, l'évolution et les modifications d'un modèle, d'un logiciel ou d'un composant matériel, essentielle pour la qualification et la certification. Plus précisément pour l'IA embarquée, il s'agit de documenter les étapes de développement, d'entraînement et de déploiement d'un système d'IA, pour en garantir la transparence et l'auditabilité. Globalement, c'est un suivi des données, des décisions et des composants d'un système d'IA.

Audit

Examen systématique et indépendant des processus, des preuves et des résultats pour confirmer la conformité aux normes et réglementations.

Auditabilité – Auditability

Possibilité de vérifier, contrôler et évaluer le fonctionnement d'un système d'IA, notamment en cas d'incident ou de dérive ; Indispensable pour les dispositifs embarqués dans les secteurs règlementés.

Software Bill of Materials (SBOM)

Inventaire détaillé des composants logiciels (bibliothèques, dépendances, modules) utilisés dans un système embarqué. Cela sert à la traçabilité, à la cybersécurité et à la conformité réglementaire.

10.6 Cybersécurité

La cybersécurité constitue un enjeu majeur pour les systèmes d'IA embarquée, en particulier lorsqu'ils sont intégrés à des environnements critiques, autonomes ou cyber-physiques. Ces systèmes combinent des composants matériels, logiciels, algorithmiques et des données sensibles, ce qui élargit leur surface d'attaque et les expose à des menaces spécifiques liées à l'IA.

Dans le contexte de l'IA embarquée, la cybersécurité vise à protéger les modèles, les données et les mécanismes d'inférence contre les attaques, tout en garantissant la continuité de service, l'intégrité des décisions et la confidentialité des informations traitées. Les notions abordées dans ce chapitre fournissent un cadre pour analyser les menaces, concevoir des mécanismes de défense adaptés et renforcer la robustesse des systèmes d'IA embarquée face aux attaques intentionnelles ou accidentelles.

Cybersécurité des systèmes d'IA embarqués – Cybersecurity of Embedded Systems

Protection des modèles, les algorithmes et les données d'IA utilisées dans des secteurs cloisonnés où prédominent les systèmes critiques. Étant donné que les systèmes critiques dépendent fortement des systèmes d'IA embarqués pour leurs opérations autonomes, leur navigation et leur traitement des données, il est essentiel d'assurer leur sécurité. Par conséquent, la cybersécurité des systèmes d'IA embarqués consiste à protéger les systèmes critiques basés sur l'IA contre le piratage.

Système IA embarquée pour la cybersécurité – Embedded AI system for cybersecurity

Systèmes d'IA pour améliorer la cybersécurité des systèmes critiques en termes de sécurité. Les systèmes d'IA embarqués peuvent détecter, prévenir et répondre aux cyber menaces en temps réel, ce qui est crucial pour les composants des systèmes critiques.

Menaces – Threats

Violation potentielle de la sécurité. Les menaces qui pèsent sur les systèmes cyber-physiques comprennent la destruction d'informations et/ou d'autres ressources ; le vol, la suppression ou la perte d'informations et/ou d'autres ressources ; la corruption ou la modification d'informations ; la divulgation d'informations ; l'interruption de services.

Stratégies d'atténuation – Mitigation Strategies

Mise en œuvre de mesures de protection visant à réduire la vulnérabilité des actifs en limitant les conséquences des accidents et en prévenant les incidents de sécurité.

Confidentialité par construction – Privacy by Design

Approche qui intègre la protection des données personnelles dès les premières phases de conception d'un système d'IA. (ex : capteurs biométriques).

Sécurité par construction – Security by Design

Méthodologie qui intègre des mécanismes de sécurité dès la phase de développement d'un système, pour prévenir les vulnérabilités.

Couche de sécurité IA – AI safety layer

Mécanismes intégrés pour surveiller et corriger les comportements inattendus ou dangereux d'un modèle d'IA.

Environnement d'exécution sécurisé – Trusted Execution Environment (TEE)

Zone sécurisée pour exécuter du code sensible, isolée dans un processeur permettant d'exécuter un code sensible à l'abri des attaques.

Cyber vol de modèles – AI Model Theft

Extraction non autorisée, la rétro-ingénierie ou l'exploitation malveillante de modèles d'IA/ML déployés dans des systèmes critiques. Cela peut se produire en temps réel pendant le fonctionnement ou la maintenance des systèmes critiques.

Evaluation offensive – Red Teaming

Technique de test où une équipe simule des attaques ou des usages malveillants pour identifier les failles d'un système d'IA.

Perturbation Adverse – Adversarial Attack

Action malveillante consistant à modifier légèrement une entrée (par exemple en ajoutant un bruit imperceptible) afin de tromper un modèle d'apprentissage automatique. Ces perturbations restent généralement invisibles pour un humain, mais provoquent une erreur de prédiction du modèle.

Robustesse face aux attaques adverses – Adversarial Robustness

Capacité d'un modèle d'IA à résister aux attaques adverses, c'est-à-dire à maintenir des performances acceptables même lorsque ses entrées ont été intentionnellement perturbées. La robustesse est une propriété recherchée dans les systèmes critiques ou embarqués exposés à des manipulations potentielles.

Apprentissage adverse – Adversarial Training

Technique de défense consistant à intégrer des entrées trompeuses dans le processus d'entraînement afin d'augmenter la robustesse du modèle. Le modèle apprend ainsi à reconnaître et à neutraliser les perturbations malveillantes, ce qui en fait l'une des méthodes les plus efficaces pour améliorer la résistance aux attaques adverses.

Démarrage sécurisé – Secure Boot

Mécanisme de sécurité garantissant que seuls des logiciels authentifiés et intègres peuvent être exécutés lors du démarrage d'un système embarqué. En IA embarquée, le Secure Boot protège les modèles, les chaînes d'inférence et les composants critiques contre les modifications non autorisées dès la mise sous tension.

Environnement d'exécution de confiance – Trusted Execution Environment (TEE)

Environnement isolé matériellement au sein d'un processeur permettant l'exécution sécurisée de code et le stockage de données sensibles. En IA embarquée, le TEE est utilisé pour protéger les modèles, les clés cryptographiques et certaines décisions critiques contre les attaques logicielles ou matérielles.

10.7 Qualification, certification, preuves

La qualification constitue une étape clé pour démontrer qu'un système d'IA embarquée est apte à être déployé et exploité dans un contexte donné, en particulier lorsqu'il s'inscrit dans des environnements critiques, réglementés ou à forts enjeux de sûreté. Elle vise à établir, de manière structurée et documentée, que les modèles, les logiciels et les environnements d'exécution répondent aux exigences définies.

Qualification

Ensemble des activités visant à démontrer qu'un système, un composant ou un processus est apte à remplir sa fonction dans un contexte donné. En IA embarquée, cela inclut la validation des modèles et des environnements d'exécution.

Vérifiabilité – Verifiability

Capacité à évaluer une mise en œuvre des exigences afin de déterminer qu'elles ont été satisfaites (d'après ARP4754A).

Vérification et validation – Verification and Validation (V&V)

Vérification que le système est construit correctement, autrement dit qu'il respecte les exigences, spécifications et normes définies; La validation consiste à vérifier que l'on a construit le bon système, c'est-à-dire qu'il répond effectivement aux besoins et attentes des utilisateurs.

Conformité – Compliance

Démonstration qu'une caractéristique ou une propriété d'un produit satisfait aux exigences énoncées.

Certification

Attestation d'une tierce partie concernant des produits, des processus, des systèmes ou des personnes. Ce peut être une garantie écrite qu'un système ou un composant est conforme à ses exigences spécifiées et est acceptable pour une utilisation opérationnelle. Cela peut-être aussi une démonstration formelle qu'un système ou un composant est conforme aux exigences spécifiées et est acceptable pour une utilisation opérationnelle. Pour l'embarqué, c'est le processus de confirmation qu'un système ou un composant est conforme à ses exigences spécifiées et qu'il est acceptable pour une utilisation opérationnelle.

Preuves – Evidence

Documents, résultats de tests ou rapports qui démontrent la conformité d'un système aux exigences de sécurité, de robustesse ou de réglementation. Les "preuves" sont indispensables dans les processus de certification.

Dossier de garantie – Assurance Case

Argument pour justifier certaines affirmations concernant un système, sur la base de preuves concernant à la fois le système et l'environnement dans lequel il fonctionne. Les affirmations peuvent concerner n'importe quelle propriété du système, comme la fiabilité ou la sécurité..

Responsabilité – Accountability

Principe selon lequel les acteurs impliqués dans la conception et l'usage de l'IA doivent pouvoir être identifiés et répondre des décisions prises par les systèmes. Important pour les systèmes embarqués autonomes, notamment en cas d'incident ou de litige. L'imputabilité est une propriété qui garantit que les actions d'une entité ne peuvent être attribuées qu'à cette entité.

11 Cycle de vie et Gouvernance de l'IA embarquée

Ce chapitre regroupe la terminologie permettant de décrire les étapes, les processus et les mécanismes de gouvernance encadrant le cycle de vie des systèmes d'IA embarquée. Il se concentre sur les concepts nécessaires à la maîtrise opérationnelle de l'IA embarquée, depuis la conception jusqu'à la fin de vie, en intégrant les exigences de traçabilité, de validation et de responsabilité indispensables à la mise en œuvre effective des cadres réglementaires applicables.

Les notions présentées visent à structurer une approche cohérente et maîtrisée de l'IA embarquée dans des systèmes contraints, souvent critiques, où les choix techniques et organisationnels doivent être documentés, justifiables et vérifiables tout au long du cycle de vie.

11.1 Cycle de vie de l'IA embarquée

Cette partie décrit les différentes étapes permettant de concevoir, transformer, intégrer et exploiter un système d'IA embarquée sur une plateforme contrainte. Elle met en évidence les transitions successives entre un modèle conçu hors ligne et un système opérationnel embarqué, en tenant compte des adaptations nécessaires pour garantir une exécution fiable, performante et maîtrisée dans le temps.

Ces notions sont essentielles pour structurer une vision complète et cohérente du parcours d'un modèle d'IA embarquée, depuis sa conception jusqu'à sa fin de vie, en intégrant les réalités matérielles, opérationnelles et organisationnelles propres aux systèmes embarqués.

Cycle de vie de l'IA embarquée – Embedded AI Lifecycle

Ensemble des phases successives couvrant la conception, le développement, l'intégration, la validation, la conversion, l'optimisation, le déploiement, l'exploitation, l'évolution et la fin de vie d'un système d'IA embarquée, depuis les premières spécifications jusqu'au retrait du système.

Phase de conception – Design Phase

Étape durant laquelle sont définis les objectifs fonctionnels de l'IA embarquée, les contraintes matérielles et opérationnelles, ainsi que les exigences de performance, de robustesse, de traçabilité et de gouvernance nécessaires à son intégration dans un système embarqué.

Développement du modèle embarqué – Embedded Model Development

Processus consistant à concevoir, entraîner et adapter un modèle d'IA destiné à être exécuté sur une plateforme embarquée, en tenant compte des contraintes de calcul, de mémoire, de latence et de comportement déterministe.

Conversion du modèle – Model Conversion

Processus de transformation d'un modèle entraîné en un format et une représentation compatibles avec la cible embarquée (runtime, bibliothèque d'inférence, accélérateur matériel), en vue de son intégration et de son exécution locale.

Optimisation du modèle – Model Optimization

Ensemble de techniques visant à réduire le coût d'exécution d'un modèle destiné à l'embarqué (latence, consommation énergétique, empreinte mémoire), tout en maîtrisant l'impact sur les performances fonctionnelles.

Artefact de déploiement – Deployment Artifact

Ensemble versionné des éléments nécessaires à l'exécution d'un système d'IA embarquée sur la cible finale (modèle converti, paramètres, chaînes de traitement, configurations), constituant l'unité opérationnelle traçable et versionnée de déploiement.

Intégration système – System Integration

Phase durant laquelle le modèle d'IA, les chaînes de traitement associées et les mécanismes de supervision sont intégrés au système embarqué global, en interaction avec le matériel, les capteurs et les fonctions existantes.

Déploiement embarqué – On-device Deployment

Mise en service d'un système d'IA embarquée sur la cible finale, incluant la configuration, l'optimisation finale, la vérification du comportement attendu et la documentation de l'état initial du système.

Exploitation

Phase durant laquelle le système d'IA embarquée fonctionne en conditions réelles d'usage, produisant des inférences et des décisions locales, sous supervision et selon des règles de gouvernance définies.

Évolution et maintenance – Evolution and Maintenance

Ensemble des activités visant à corriger, adapter ou améliorer un système d'IA embarquée au cours de son exploitation, tout en garantissant la continuité de service, la traçabilité des changements et la conformité aux exigences initiales.

Fin de vie – End of Life

Phase durant laquelle un système d'IA embarquée est retiré du service, incluant la désactivation contrôlée des fonctions d'IA, l'archivage des éléments nécessaires à la traçabilité et la gestion des versions finales des modèles.

11.2 Validation et maîtrise du cycle de vie

Cette partie présente les concepts permettant de démontrer que l'IA embarquée se comporte conformément aux exigences définies. Elle couvre les activités nécessaires pour s'assurer que le système fonctionne correctement, de manière reproductible et dans des conditions représentatives de son usage réel.

Ces notions constituent un socle indispensable pour garantir que l'IA embarquée reste maîtrisée tout au long de son exploitation et pour disposer d'éléments objectifs permettant de justifier les performances et le comportement du système dans des contextes contraints ou critiques.

Validation de l'IA embarquée – Embedded AI validation

Processus visant à démontrer que le système d'IA embarquée répond aux exigences fonctionnelles et opérationnelles définies, dans des conditions représentatives de son usage réel sur la plateforme embarquée.

Vérification – Verification

Ensemble des activités permettant de s'assurer que le système d'IA embarquée est correctement implémenté par rapport aux spécifications définies, notamment en termes de comportement, de performances et de contraintes temporelles.

Tests embarqués – Embedded AI Testing

Tests réalisés sur la cible embarquée ou dans un environnement représentatif afin d'évaluer le comportement de l'IA embarquée dans des conditions proches de l'exploitation réelle.

Validation continue – Continuous Validation

Processus consistant à vérifier, durant l'exploitation ou à la suite d'une mise à jour ou recalibration, que le comportement de l'IA embarquée reste conforme aux exigences initiales, malgré l'évolution des conditions d'usage, des données d'entrée ou du contexte opérationnel.

11.3 Gouvernance de l'IA embarquée

Cette partie regroupe les concepts permettant d'encadrer les décisions techniques et organisationnelles liées à l'IA embarquée. Elle clarifie les responsabilités, les règles et les processus nécessaires à la gestion des modèles, des configurations et de leurs évolutions dans le temps.

Ces notions sont indispensables pour assurer une maîtrise durable de l'IA embarquée, faciliter la prise de décision et garantir une exploitation responsable, documentée et cohérente des systèmes intelligents embarqués.

Gouvernance de l'IA embarquée – Governance of Embedded AI

Ensemble des règles, processus et responsabilités permettant de maîtriser la conception, le déploiement, l'exploitation et l'évolution des systèmes d'IA embarquée, en assurant la cohérence des décisions techniques et organisationnelles tout au long du cycle de vie.

Responsabilité – Accountability

Principe selon lequel les acteurs impliqués dans la conception et l'usage de l'IA doivent pouvoir être identifiés et répondre des décisions prises par les systèmes. Important pour les systèmes embarqués autonomes, notamment en cas d'incident ou de litige. L'imputabilité est une propriété qui garantit que les actions d'une entité ne peuvent être attribuées qu'à cette entité.

Responsabilité opérationnelle – Operational Accountability

Définition des responsabilités liées à l'exploitation quotidienne, à la supervision et aux actions correctives sur un système d'IA embarquée, en particulier en cas de dégradation de performance ou d'incident.

Traçabilité du cycle de vie – Lifecycle Traceability

Capacité à retracer l'historique des décisions, des versions de modèles, des paramètres, des chaînes de traitement et des configurations tout au long du cycle de vie d'un système d'IA embarquée, afin de permettre l'audit et la justification des choix effectués.

Gestion des versions – Version Management

Processus permettant de maîtriser l'évolution des modèles, des paramètres, des chaînes de traitement et des configurations déployées sur des systèmes embarqués, en assurant l'identification claire des versions en exploitation.

EdgeOps

Ensemble des pratiques et processus d'exploitation dédiés aux modèles d'IA déployés sur des systèmes embarqués, couvrant le déploiement, la supervision, la maintenance, les mises à jour et la gestion des versions, sous contraintes de ressources et de connectivité.

11.4 Exploitation maîtrisée et supervision

Cette partie traite des concepts associés à l'observation et au suivi du comportement des systèmes d'IA embarquée en fonctionnement réel. Elle permet de comprendre comment les performances, la stabilité et le comportement de l'IA peuvent être surveillés et évalués dans la durée.

Ces notions sont essentielles pour maintenir un niveau de confiance opérationnelle dans l'IA embarquée, en identifiant les situations de dégradation et en assurant un comportement prévisible dans des environnements d'exploitation variables.

Supervision embarquée – Embedded Monitoring

Mécanismes permettant de surveiller le fonctionnement de l'IA embarquée à l'exécution, afin de détecter des comportements anormaux, des dérives de performance ou des situations non prévues.

Surveillance en exploitation – In-Operation Monitoring

Observation continue ou périodique du comportement et des performances de l'IA embarquée durant son fonctionnement réel, dans le respect des contraintes de calcul et d'énergie du système.

Télémétrie IA embarquée – On-Device AI Telemetry

Collecte et exposition, sur la cible embarquée, d'indicateurs relatifs au fonctionnement de l'IA (temps d'inférence, charge, qualité des entrées, scores de confiance, événements de dégradation), afin de permettre une analyse et une supervision compatibles avec les contraintes embarquées.

Dérive du modèle – Model Drift

Dégradation progressive des performances d'un modèle en exploitation, liée à l'évolution des données d'entrée, du contexte d'usage ou de l'environnement, pouvant nécessiter recalibration, mise à jour ou revalidation du système embarqué.

Gestion des dérives – Drift Management

Ensemble des mécanismes visant à détecter, qualifier et traiter les dérives de performance d'un système d'IA embarquée, incluant seuils, alertes, modes dégradés, recalibration et mise à jour versionnée.

Recalibration périodique – Periodic Model Recalibration

Ajustement planifié ou déclenché d'un système d'IA embarquée (paramètres, seuils, calibrations associées) afin de maintenir des performances conformes en exploitation, tout en assurant la traçabilité et la revalidation nécessaires.

Suivi post-déploiement – Post-Deployment Monitoring

Activité consistant à suivre dans la durée le comportement d'un système d'IA embarquée après sa mise en service, afin de confirmer le maintien des performances attendues et d'identifier les besoins d'ajustement ou de mise à jour.

Dégradation maîtrisée – Graceful Degradation

Capacité d'un système d'IA embarquée à adopter un comportement sûr, prévisible et documenté en cas de perte de performance, de données dégradées ou de conditions d'exploitation non nominales.

11.5 Mise à jour et continuité opérationnelle

Cette partie décrit les mécanismes permettant de faire évoluer un système d'IA embarquée après son déploiement, tout en préservant la continuité de service. Elle formalise les principes nécessaires pour introduire des modifications contrôlées et réversibles dans des environnements embarqués.

Ces notions jouent un rôle clé dans la pérennité des systèmes d'IA embarquée, en permettant leur adaptation progressive sans compromettre la stabilité ni la maîtrise opérationnelle.

Mise à jour – Over-the-Air update (OTA)

Mécanisme permettant de déployer à distance une nouvelle version d'un modèle, d'un paramétrage ou d'une configuration sur une cible embarquée, en assurant l'identification des versions et la continuité de service.

Retour arrière – Rollback

Capacité à restaurer une version antérieure d'un modèle ou d'une configuration sur une cible embarquée, notamment en cas de régression de performance, d'instabilité ou d'incident détecté en exploitation.

12 Réglementation & normes pour l'IA embarquée

12.1 Réglementations européennes

AI Act

Cadre réglementaire proposé par l'Union européenne pour encadrer l'usage de l'IA selon son niveau de risque, avec des obligations spécifiques pour les systèmes critiques.

Règlement - UE - 2024/1689 - EN - EUR-Lex

Textes de la loi | Loi sur l'intelligence artificielle de l'UE

Le Règlement européen sur l'intelligence artificielle : publics concernés, dates clés, conséquences pour les entreprises | Direction générale des Entreprises

Règlement Général sur la Protection des Données (RGPD)

Règlement européen qui harmonise les règles de traitement des données personnelles dans l'UE, entré en application le 25 mai 2018. Il renforce les droits des personnes et responsabilise les acteurs traitant des données.

Data Act

Entré en vigueur en janvier 2024, le Data Act de l'UE (Règlement (UE) 2023/2854) est pleinement applicable depuis septembre 2025 après l'expiration de sa période transitoire. Ce texte instaure un cadre harmonisé fixant de nouvelles règles sur l'accès, l'utilisation et le partage des données industrielles générées par les produits et services connectés dans l'UE, couvrant l'ensemble des secteurs économiques.

Chips Act

Le Chips Act européen (Règlement (UE) 2023/1781) vise à renforcer la capacité de l'UE en semi-conducteurs, soutenir la R&D et développer des capacités de production avancées. Son objectif est d'atteindre 20 % de la production mondiale de puces d'ici 2030. Il prévoit également un mécanisme de coordination pour sécuriser la chaîne d'approvisionnement en cas de crise. Il favorise l'émergence d'un écosystème européen capable de concevoir et produire les composants optimisés pour l'IA embarquée et les systèmes cyberphysiques.

Data Governance Act (DGA)

Le Data Governance Act (Règlement (UE) 2022/868) organise la gouvernance et les mécanismes de partage des données, en particulier la réutilisation de certaines données du secteur public, le cadre des intermédiaires de données et le partage « altruiste » de données.

Product Liability Directive

La nouvelle directive européenne sur la responsabilité du fait des produits (2024/2853) a été adoptée le 23 octobre 2024 et doit être transposée dans les droits nationaux des États membres d'ici le 9 décembre 2026. Elle modernise les règles de responsabilité pour inclure les produits numériques, l'IA, et les modèles économiques circulaires.

Cybersecurity Act (2019)

Un règlement européen adopté en 2019, qui crée le cadre européen de certification de cybersécurité pour les produits TIC. C'est lui qui a donné naissance aux futurs schémas de certification comme EUCC, EUCS, EU5G, etc. Il ne crée pas d'obligations directes pour les fabricants, mais il structure le cadre de confiance et les schémas de certification. Son objectif était d'harmoniser la certification de cybersécurité en Europe.

Cyber Resilience Act (CRA)

Règlement européen entré en vigueur le 10 décembre 2024, visant à améliorer la cybersécurité des produits comportant des éléments numériques (PDE) et à sécuriser l'internet des objets. Il impose des obligations légales de cybersécurité pour tous les produits comportant des éléments numériques (PDE), y compris les objets connectés, logiciels embarqués, systèmes industriels, etc. Il vise à garantir que tous les produits numériques vendus en Europe respectent un niveau minimal de cybersécurité, tout au long de leur cycle de vie.

12.2 Directives transverses impactantes

General Product Safety Regulation (GPSR)

Règlement (UE) 2023/988 sur la sécurité générale des produits, applicable depuis décembre 2024. Il impose aux fabricants de garantir la sécurité des produits de consommation, y compris ceux intégrant des logiciels ou de l'IA, et renforce les obligations de surveillance du marché.

Digital Product Passport (DPP)

Dispositif introduit par le paquet économie circulaire de l'UE, visant à fournir un passeport numérique contenant les informations essentielles sur un produit (composition, réparabilité, durabilité, traçabilité). Il s'applique progressivement à plusieurs catégories de produits, dont les équipements électroniques.

NIS2

Directive (UE) 2022/2555 renforçant les exigences de cybersécurité pour les secteurs essentiels et importants. Elle impose des obligations accrues en matière de gestion des risques, de sécurité des systèmes et de notification d'incidents, impactant directement les systèmes embarqués et les infrastructures critiques.

Radio Equipment Directive (RED)

La Directive 2014/53/UE sur les équipements radio intègre désormais des exigences renforcées en cybersécurité, protection des données et prévention des fraudes pour les objets connectés. Les actes délégués associés imposent des obligations spécifiques aux fabricants de dispositifs radio intégrant des fonctions logicielles ou d'IA.

European Cybersecurity Certification Scheme (EUCC)

Schéma européen de certification de cybersécurité pour les produits TIC, élaboré dans le cadre du Cybersecurity Act. Il fournit un cadre harmonisé pour certifier la sécurité des composants matériels et logiciels, pertinent pour les systèmes embarqués et les produits intégrant de l'IA.

12.3 Normes européennes en soutien de la réglementation sur l'IA

Les normes harmonisées européennes destinées à soutenir l'application de l'AI Act sont en préparation au sein du Joint Technical Committee 21 (AI) de l'organisation CEN/CENELEC. Le respect de l'ensemble de ces normes par les fournisseurs et déployeurs de systèmes d'IA vaudra présomption de conformité aux dispositions de la réglementation, principalement pour les systèmes qualifiés à haut risque.

Ces normes n'existent pas encore, cette section fournit l'ensemble des normes qui devraient être publiées courant 2026 et 2027. Elles s'appliqueront en totalité aux systèmes embarqués, même si leur contenu n'a rien de spécifique.

Des normes complémentaires pour les systèmes d'IA à usage général (comprenant l'IA générative, les LLM etc.) devraient être conçues dans un deuxième temps.

Quality Management Systems (prEN 18286)
Gestion de la qualité pour les systèmes d'IA

Risk Management (prEN 18228)
Maîtrise des risques pour les systèmes d'IA.

AI trustworthiness framework – Part 1
Logging, transparency and human oversight (prEN 18229-1): gestion de trois aspects de la confiance envers les systèmes d'IA: traçabilité, information aux utilisateurs et contrôle par les opérateurs humains.

AI trustworthiness framework – Part 2
Accuracy and robustness (prEN 18229-2): gestion de deux aspects de la confiance envers les systèmes d'IA: précision et robustesse.

Quality and governance of datasets (prEN 18284)
Norme sur la qualité et la gouvernance des bases de données utilisées dans les systèmes d'IA.

Managing Bias (prEN 18283)
Gestion (et réduction) des biais dans les données, afin de satisfaire aux exigences d'équité.

Cybersecurity specifications for AI Systems (prEN 18282)
Spécifications sur la cybersécurité des systèmes d'IA, analyse des menaces et des moyens d'y répondre.

Conformity Assessment (prEN 18285)
Évaluation de la conformité à l'ensemble des normes, faite par les organismes habilités.

Les deux normes suivantes, également en préparation, ne correspondent pas à une demande explicite de la Commission Européenne.

- Evaluation methods for accurate computer vision systems (prEN 18281)
- Natural Language Processing (prEN ISO/IEC 23282)

12.4 Normes internationales de gouvernance et gestion du risque

Les normes harmonisées européennes en préparation s'inspirent de ces normes internationales sans les reprendre intégralement car ces dernières concernent d'abord les organisations qui produisent des systèmes d'IA, alors que la réglementation européenne porte sur les produits. Le respect des normes ISO/IEC ne vaudra pas donc présomption de conformité à la réglementation européenne.

ISO/IEC 42001

Première norme internationale dédiée aux systèmes de management de l'intelligence artificielle (IA), publiée en décembre 2023, qui encadre le développement, l'utilisation et la gouvernance des systèmes d'IA pour une utilisation responsable, éthique et sécurisée.

ISO/IEC 5338

Publiée le 20 décembre 2023, elle fournit un cadre standardisé pour le développement et la gestion des systèmes d'IA, couvrant tous les aspects du cycle de vie des logiciels intégrant l'IA.

ISO/IEC 23894

Fournit un cadre complet pour identifier, évaluer et gérer les risques liés à l'intelligence artificielle, en s'appuyant sur les principes de l'ISO 31000 pour une intégration efficace dans les activités des organisations.

12.5 Normes sectorielles

L'essor de l'IA embarquée dans les systèmes critiques impose de s'appuyer sur des normes sectorielles établies. Ces normes sectorielles constituent le socle de confiance indispensable à l'intégration de l'IA embarquée dans les systèmes critiques. Qu'il s'agisse de l'automobile, de la santé, de l'aéronautique, de l'industrie, de l'IoT, de la défense de l'énergie ou du ferroviaire, elles encadrent la sécurité fonctionnelle, la cybersécurité et la conformité réglementaire. Ensemble, elles assurent que l'IA embarquée puisse être déployée de manière fiable, interopérable et souveraine dans des environnements exigeants.

12.5.1 Automobile

Cette synthèse met en évidence que dans l'automobile, les normes couvrent sécurité fonctionnelle, sûreté de fonctionnement des algorithmes, cybersécurité et gestion des mises à jour logicielles: quatre piliers essentiels pour l'IA embarquée.

ISO 26262 (Functional Safety)

Norme internationale de sécurité fonctionnelle pour les systèmes électriques et électroniques dans les véhicules. Elle définit les processus de développement, validation et vérification afin de réduire les risques liés aux défaillances matérielles ou logicielles. Indispensable pour tout système embarqué critique, y compris ceux intégrant de l'IA.

ISO/PAS 21448 (SOTIF – Safety Of The Intended Functionality)

Complément à l'ISO 26262, elle couvre les risques liés non pas aux défaillances techniques, mais aux limites de la conception et des algorithmes. Cruciale pour l'IA embarquée, car elle traite des comportements imprévus ou non anticipés des systèmes intelligents.

ISO/SAE 21434 (Cybersecurity for Road Vehicles)

Norme conjointe ISO/SAE qui définit les exigences de cybersécurité pour les véhicules connectés. Elle couvre tout le cycle de vie, de la conception à la maintenance, afin de protéger contre les attaques et intrusions. Essentielle pour les architectures embarquées intégrant IA et connectivité.

UNECE R155/R156 (Cybersecurity & Software Updates)

Règlements de la Commission Economique pour l'Europe (ONU) imposant aux constructeurs automobiles de mettre en place une gestion de la cybersécurité (R155) et des processus sécurisés de mise à jour logicielle (R156). Ils rendent obligatoires des pratiques de protection et de maintenance continue, directement liées aux systèmes embarqués intelligents.

12.5.2 Santé

Cette synthèse montre que dans le secteur santé, les normes couvrent sécurité des dispositifs, fiabilité des logiciels, cybersécurité et conformité réglementaire, toutes essentielles pour l'intégration de l'IA embarquée.

MDR (Medical Device Regulation)

Règlement européen qui encadre la mise sur le marché des dispositifs médicaux. Il impose des exigences strictes en matière de sécurité, performance et traçabilité, essentielles pour les systèmes embarqués intégrant de l'IA dans les dispositifs médicaux.

IVDR (In Vitro Diagnostic Regulation)

Complémentaire au MDR, il régule les dispositifs de diagnostic in vitro. Il garantit la fiabilité et la sécurité des systèmes embarqués utilisés dans les tests biologiques, notamment lorsque l'IA intervient dans l'analyse et l'interprétation des résultats.

IEC 62304

Norme internationale pour le cycle de vie des logiciels de dispositifs médicaux. Elle définit les processus de développement, maintenance et gestion des risques, cruciaux pour les logiciels embarqués intégrant des algorithmes d'IA.

IEC 60601

Norme de sécurité et performance pour les équipements électromédicaux. Elle couvre les aspects électriques, mécaniques et logiciels, garantissant que les dispositifs embarqués, y compris ceux enrichis par l'IA, ne présentent pas de danger pour les patients.

IEC 81001-5-1

Norme récente dédiée à la cybersécurité des logiciels de santé. Elle fixe les exigences pour protéger les systèmes embarqués contre les menaces numériques, un enjeu majeur pour l'IA médicale connectée et distribuée.

12.5.3 Aéronautique

En résumé, dans l'aéronautique, les normes couvrent trois piliers essentiels pour l'IA embarquée : sécurité logicielle, fiabilité matérielle et cybersécurité.

DO-178C (Software Considerations in Airborne Systems and Equipment Certification)
Norme de référence pour la certification des logiciels embarqués dans l'aéronautique. Elle définit les niveaux d'assurance requis selon la criticité des fonctions, avec des processus stricts de développement, vérification et validation. Indispensable pour tout logiciel embarqué critique, y compris ceux intégrant de l'IA.

DO-254 (Design Assurance Guidance for Airborne Electronic Hardware)
Cette norme est complémentaire de la DO-178C: elle s'applique au matériel électronique embarqué (circuits, FPGA, ASIC). Elle impose des méthodes rigoureuses de conception et de validation pour garantir la fiabilité des composants critiques. Elle est pertinente pour l'IA embarquée lorsqu'elle repose sur du hardware spécialisé.

DO-326A / DO-356A (Airworthiness Security Process & Information Security Guidance)
Normes dédiées à la cybersécurité aéronautique. Elles définissent les processus de gestion des menaces et des vulnérabilités, ainsi que les exigences de protection des systèmes embarqués contre les attaques. Elles sont cruciales pour l'IA embarquée connectée, qui doit rester sûre face aux risques cyber.

12.5.4 Industrie

Dans l'industrie, les normes se concentrent sur deux axes fondamentaux pour l'IA embarquée : sécurité fonctionnelle et cybersécurité des systèmes industriels.

IEC 61508 (Functional Safety of Electrical/Electronic/Programmable Systems)
Norme internationale de sécurité fonctionnelle qui couvre les systèmes électriques, électroniques et programmables. Elle définit les exigences de conception, validation et gestion des risques pour garantir que les systèmes critiques fonctionnent de manière sûre. Elle est pertinente pour l'IA embarquée lorsqu'elle est intégrée dans des chaînes industrielles automatisées.

IEC 62443 (Industrial Communication Networks – IT Security for Industrial Automation Systems)
Norme dédiée à la cybersécurité des systèmes industriels et des environnements OT (Operational Technology). Elle couvre l'ensemble du cycle de vie, de la conception à l'exploitation, et fixe des exigences pour protéger les architectures embarquées contre les menaces cyber. Elle est essentielle pour l'IA embarquée dans les usines connectées et l'industrie 4.0.

12.5.5 Internet of Things (IoT)

Dans l'IoT, cette norme constitue le référentiel central pour garantir la sécurité et la confiance des systèmes embarqués intelligents.

ETSI EN 303 645 (Cybersecurity for Consumer IoT)

Norme européenne qui définit les exigences de cybersécurité pour les objets connectés grand public. Elle fixe des règles minimales de protection (mots de passe robustes, mises à jour sécurisées, gestion des données personnelles) afin de réduire les vulnérabilités des dispositifs IoT. Elle est pertinente pour l'IA embarquée, car elle encadre la sécurité des capteurs, passerelles et appareils intelligents utilisés dans les environnements domestiques et urbains.

12.5.6 Défense

STANAG (Standardization Agreement – OTAN)

Les STANAG sont des accords de normalisation de l'OTAN qui définissent des standards techniques et opérationnels pour assurer l'interopérabilité entre les forces alliées. Dans le domaine de l'IA embarquée, ils concernent notamment la sécurité des systèmes, la fiabilité des communications et la compatibilité des équipements critiques. Ils garantissent que les solutions embarquées respectent des exigences communes de sûreté et de cybersécurité dans les environnements militaires.

12.5.7 Énergie / Smart Grid

Dans l'énergie, ces normes assurent l'interopérabilité, fiabilité et sécurité des communications dans les réseaux intelligents, permettant à l'IA embarquée de jouer un rôle clé dans l'optimisation et la résilience des infrastructures.

IEC 61850 (Communication Networks and Systems for Power Utility Automation)

Norme internationale qui définit l'architecture et les protocoles de communication pour les systèmes électriques intelligents. Elle permet l'interopérabilité entre équipements de différentes marques et assure la fiabilité des échanges temps réel dans les réseaux électriques. Elle est pertinente pour l'IA embarquée, car elle encadre la gestion locale et distribuée des données dans les smart grids.

IEEE 2030.5 (Smart Energy Profile 2.0)

Norme américaine qui spécifie les protocoles de communication pour l'interopérabilité des réseaux énergétiques distribués, notamment dans le cadre des énergies renouvelables et du pilotage de la demande. Elle favorise l'intégration de l'IA embarquée pour optimiser la gestion dynamique des flux énergétiques et la coordination des micro-grids.

12.5.8 Rail/Ferroviaire

Dans le ferroviaire, ces normes assurent que l'IA embarquée respecte les mêmes exigences de sécurité logicielle et électronique que les systèmes traditionnels, garantissant fiabilité et conformité dans un environnement critique.

EN 50128 (Software for Railway Control and Protection Systems)
Norme européenne qui définit les exigences pour le développement et la validation des logiciels utilisés dans les systèmes ferroviaires de contrôle et de signalisation. Elle s'appuie sur l'IEC 61508 mais l'adapte aux spécificités du ferroviaire. Elle est pertinente pour l'IA embarquée lorsque l'IA intervient dans la gestion du trafic, la signalisation ou les systèmes de sécurité.

EN 50129 (Safety Related Electronic Systems for Signalling)
Norme complémentaire qui encadre la sécurité des systèmes électroniques liés à la signalisation ferroviaire. Elle fixe les exigences de conception, d'évaluation et de certification pour garantir la sûreté de fonctionnement des architectures embarquées. L'IA embarquée y est concernée dès lors qu'elle participe à la prise de décision dans des systèmes critiques.

12.5.9 Compléments sectoriels pour le marquage CE

Automobile
Déjà couvert par ISO 26262, SOTIF, ISO/SAE 21434, UNECE R155/156. Il faut y ajouter le **Règlement général sur la sécurité des véhicules (EN 2019/2144)** qui impose certaines fonctions automatisées et sécurité électronique.

Santé
Il faut ajouter l'**ISO 14971 (Risk Management for Medical Devices)**, norme harmonisée indispensable pour le marquage CE des dispositifs médicaux.

Aéronautique
Pas de marquage CE direct (l'aéronautique relève de certifications EASA/FAA), mais EN 9100 (qualité aéronautique) peut être mentionnée comme norme harmonisée.

Industrie
IEC 61508 et IEC 62443 sont clés ; il faut leur ajouter la **Directive Machines (2006/42/CE / futur Règlement Machines)** et la **Directive ATEX (2014/34/UE)** dans le cas de systèmes embarqués en atmosphères explosives.

IoT
ETSI EN 303 645 est central. Il faut y ajouter la **Directive RED Radio Equipment Directive (2014/53/UE)**, déjà évoquée au chapitre 12.2; celle-ci est obligatoire pour tout dispositif radio connecté (Wi-Fi, Bluetooth, LTE, etc.). Elle conditionne le marquage CE des équipements IoT, en imposant des exigences de sécurité électrique, CEM et cybersécurité. Elle est pertinente et incontournable pour l'IA embarquée dès qu'il y a connectivité radio.

Défense

STANAG reste spécifique OTAN. Pas de marquage CE direct (produits militaires exclus) ; en revanche, pour les équipements dual-use (civil + militaire), les directives CE transversales (CEM, Basse Tension, RoHS, Machines, RED) s'appliquent pour un marquage CE.

Énergie / Smart Grid

IEC 61850 et IEEE 2030.5 sont sectoriels. Il faut ajouter la Directive Basse Tension et la Directive Compatibilité électromagnétique, qui sont toujours exigées pour le marquage CE des équipements électriques.

Rail/Ferroviaire

EN 50128 et EN 50129 sont sectoriels. Il faut y ajouter la **Directive Interopérabilité Ferroviaire (UE 2016/797)**, qui encadre la conformité CE des sous-systèmes ferroviaires.

13 Applications & cas d'usage

Ce chapitre regroupe la terminologie permettant de décrire, structurer et qualifier les applications et cas d'usage de l'IA embarquée. Il met l'accent sur les fonctions d'intelligence artificielle exécutées directement sur des systèmes embarqués intégrant capteurs, capacités de calcul locales et chaînes de traitement dédiées, afin de produire des décisions, des estimations ou des commandes en interaction avec le monde physique.

Les notions présentées visent à fournir un cadre de lecture commun pour relier les finalités opérationnelles de l'IA embarquée aux contraintes d'exécution sur cible, aux choix d'architecture et aux exigences d'exploitation dans des environnements réels.

13.1 Concepts généraux

Les notions regroupées permettent de définir ce qu'est une application d'IA embarquée et comment elle s'inscrit dans un système réel. Elles précisent la relation entre objectifs opérationnels, fonctions d'IA déployées sur la cible et scénarios d'exploitation, en tenant compte des contraintes propres aux plateformes embarquées. Ce cadre conceptuel sert à décrire de manière cohérente les usages de l'IA intégrée à des systèmes physiques.

Application
Ensemble cohérent de fonctions fournies par un système embarqué, dont certaines reposent sur des modèles d'IA exécutés sur la plateforme cible. Une application d'IA embarquée combine capteurs, traitements et logique décisionnelle afin de produire un comportement, une capacité ou un service opérationnel intégré au système.

Cas d'usage
Description structurée d'un scénario d'exploitation dans lequel une ou plusieurs fonctions d'IA embarquée contribuent à atteindre un objectif opérationnel identifié. Un cas d'usage précise le contexte, les entrées issues de capteurs, les sorties attendues et les contraintes d'exécution associées.

Fonction d'IA embarquée
Fonction logicielle déployée sur un système embarqué, mettant en œuvre une chaîne de traitement et un ou plusieurs modèles d'IA pour produire une sortie exploitable (information, alerte, estimation, décision ou commande) en interaction avec l'environnement.

Scénario opérationnel
Description d'un contexte réel d'utilisation caractérisé par un environnement, des conditions d'exploitation variables, des contraintes temporelles et des situations nominales ou dégradées auxquelles l'IA embarquée doit répondre de manière fiable.

Bénéfice de l'inférence locale
Avantage obtenu par l'exécution des modèles d'IA directement sur la cible embarquée, permettant de répondre à des exigences de réactivité, d'autonomie, de continuité de service, de résilience ou de maîtrise des flux de données.

13.2 Grandes familles d'applications sectorielles

La structuration par domaines d'activité met en évidence la diversité des contextes dans lesquels l'IA embarquée est déployée. Chaque famille sectorielle illustre des usages concrets reposant sur l'exécution locale de fonctions d'IA, permettant d'améliorer la réactivité, l'autonomie et la robustesse des systèmes dans des environnements industriels, critiques ou distribués.

Industrie, fabrication et procédés
Applications dans lesquelles l'IA embarquée analyse des données issues de capteurs industriels afin de réaliser des inspections, surveiller des procédés et détecter des anomalies directement sur les équipements. L'exécution embarquée permet une réactivité immédiate, une robustesse face aux environnements industriels et une réduction des flux de données vers des systèmes centraux.

Robotique autonome et collaborative
Applications où l'IA embarquée soutient la perception, la navigation, la manipulation et l'interaction entre robots et opérateurs. Le traitement embarqué garantit une latence compatible avec le mouvement et l'interaction physique, ainsi qu'une autonomie de fonctionnement dans des environnements dynamiques.

Mobilité, transport et véhicules intelligents
Applications intégrant l'IA embarquée pour interpréter l'environnement, assister la conduite, gérer des trajectoires et soutenir des fonctions d'autonomie. L'embarqué permet une prise de décision locale rapide et fiable, indépendante de la connectivité réseau.

Aéronautique et aérospatial
Applications embarquées à bord d'aéronefs, de drones ou de systèmes spatiaux utilisant l'IA pour la perception, la surveillance et l'assistance à la décision. L'IA embarquée apporte autonomie, résilience et continuité de service dans des environnements isolés ou fortement contraints.

Défense et systèmes critiques
Applications dans lesquelles l'IA embarquée contribue à la surveillance, à l'analyse situationnelle et à l'aide à la décision. Le traitement embarqué permet un fonctionnement robuste, déconnecté et maîtrisé, indispensable dans des contextes critiques.

Santé, dispositifs médicaux et pharma
Applications intégrant l'IA embarquée dans des dispositifs médicaux, des équipements de diagnostic ou des systèmes de monitoring. L'exécution locale favorise la réactivité, la continuité de fonctionnement et une meilleure maîtrise des données sensibles.

Énergie, réseaux et infrastructures critiques
Applications exploitant l'IA embarquée pour superviser, diagnostiquer et optimiser des équipements énergétiques ou des infrastructures distribuées. Le traitement embarqué permet une gestion locale efficace et une exploitation à grande échelle de dispositifs autonomes.

Télécommunications et équipements réseau
Applications dans lesquelles l'IA embarquée est utilisée pour optimiser l'utilisation des

ressources réseau, détecter des anomalies et améliorer la qualité de service sur des équipements en périphérie. L'embarqué permet des ajustements rapides au plus près du trafic et des équipements.

Construction, bâtiments et infrastructures intelligents
Applications utilisant l'IA embarquée pour la gestion énergétique, la surveillance d'équipements et la sécurité de bâtiments ou d'ouvrages. Le traitement embarqué facilite des décisions locales adaptées aux usages réels et à la variabilité des conditions.

Agriculture et environnement
Applications exploitant des capteurs et la vision embarquée pour surveiller l'environnement, analyser des cultures ou détecter des événements naturels. L'IA embarquée permet des dispositifs autonomes, sobres en énergie et capables de fonctionner sur de longues durées sans connectivité continue.

Objets connectés, wearables et électronique grand public
Applications intégrant des modèles d'IA compacts pour l'interaction, la personnalisation et la reconnaissance d'activités. L'exécution embarquée garantit une expérience réactive, une faible consommation énergétique et une indépendance vis-à-vis des infrastructures distantes.

Culture, médias et contenus interactifs
Applications dans lesquelles l'IA embarquée soutient l'analyse audiovisuelle, l'interaction temps réel, la réalité augmentée ou des expériences immersives. Le traitement embarqué permet des interactions fluides et immédiates sur des dispositifs autonomes.

13.3 Fonctions applicatives typiques

Les fonctions décrites correspondent aux capacités opérationnelles fournies par l'IA embarquée, telles que la perception de l'environnement, l'analyse de situations, l'estimation d'états ou l'interaction avec des utilisateurs. Elles constituent des briques fonctionnelles réutilisables, mises en œuvre directement sur la plateforme embarquée pour produire des décisions ou des commandes adaptées aux contraintes du système.

Perception embarquée
Capacité d'un système embarqué à interpréter son environnement physique à partir de données issues de capteurs (image, radar, LiDAR, audio, inertiel, biométrique, etc.), afin de produire une représentation exploitable pour la prise de décision ou la commande d'un système physique.

Détection
Fonction consistant à identifier la présence ou l'apparition d'objets, d'événements ou de situations d'intérêt dans une donnée d'entrée, par exemple la détection d'un défaut, d'un obstacle, d'un incident ou d'un changement d'état, avec une sortie généralement binaire ou localisée.

Classification
Fonction visant à associer une observation issue de capteurs à une catégorie prédéfinie, telle que le type d'objet, l'état d'un équipement, une condition environnementale ou une activité, afin d'orienter une décision ou une action locale.

Segmentation
Fonction permettant de découper une donnée, notamment une image ou un signal, en régions ou segments homogènes correspondant à des zones d'intérêt, afin de localiser précisément des objets, des défauts ou des phénomènes pertinents pour l'application embarquée.

Estimation
Fonction produisant une valeur continue représentant un état ou une grandeur physique, comme une position, une vitesse, une orientation, un niveau ou une intensité, à partir de mesures capteurs, pour alimenter une logique de contrôle ou de supervision.

Prédiction
Fonction consistant à anticiper l'évolution future d'un état, d'un comportement ou d'un phénomène observé, par exemple l'usure d'un équipement, une trajectoire probable ou une dérive de performance, afin de permettre une action proactive du système embarqué.

Détection d'anomalies
Fonction visant à repérer des écarts significatifs par rapport à un fonctionnement nominal appris ou défini, afin d'identifier précocement des défauts, des incidents ou des situations inhabituelles dans des environnements réels.

Reconnaissance d'activités
Fonction permettant d'identifier une activité, un geste ou un comportement à partir de séries temporelles ou de signaux multimodaux, tels que des données issues de capteurs inertiels, biométriques ou acoustiques, notamment dans les wearables et systèmes interactifs.

Interaction intelligente embarquée
Fonction permettant une interaction directe et réactive entre un utilisateur et un système embarqué, par la voix, les gestes, le contexte ou des interfaces sensorielles, avec un traitement local garantissant faible latence et continuité de service.

Boucle perception–décision–action
Organisation fonctionnelle dans laquelle les informations issues de la perception alimentent une décision locale, qui déclenche ensuite une action ou une commande sur le système ou son environnement, formant une boucle fermée exécutée sous contraintes de temps réel.

13.4 Contraintes transverses structurantes

Les contraintes présentées caractérisent les conditions dans lesquelles les applications d'IA embarquée doivent être conçues et exploitées. Elles portent notamment sur la nécessité d'un traitement local, la maîtrise des délais de réponse, l'autonomie énergétique, la robustesse face aux conditions réelles et la gestion de parcs de dispositifs. Ces contraintes influencent directement les choix techniques et l'architecture des solutions embarquées.

Traitement sur cible
Principe selon lequel les fonctions d'IA sont exécutées directement sur la plateforme embarquée, sans dépendance systématique à une infrastructure distante, afin de garantir réactivité, autonomie de fonctionnement et continuité de service dans des contextes contraints.

Temps réel applicatif
Exigence imposant que les résultats produits par l'IA soient disponibles dans un délai compatible avec l'usage, par exemple pour déclencher une action, assister un opérateur ou piloter un système physique, avec des délais de réponse bornés et maîtrisés.

Autonomie énergétique
Capacité d'un système embarqué intégrant de l'IA à fonctionner durablement avec un budget énergétique limité, qu'il soit alimenté par batterie, récupération d'énergie ou contraintes strictes de consommation.

Robustesse opérationnelle
Aptitude du système d'IA embarquée à maintenir un comportement utile et prévisible malgré la variabilité des conditions réelles, le bruit des capteurs, les perturbations de l'environnement ou des situations non nominales.

Connectivité limitée ou intermittente
Contexte dans lequel le système embarqué ne peut pas s'appuyer sur un accès réseau continu ou à haut débit, nécessitant des capacités de décision et de traitement autonomes directement sur la cible.

Déploiement à grande échelle
Déploiement d'une même fonction d'IA embarquée sur un grand nombre de dispositifs, impliquant une homogénéité de comportement, une maîtrise des versions et une exploitation industrialisée du parc embarqué.

13.5 Cas d'usage représentatifs

Les cas d'usage retenus illustrent des situations concrètes dans lesquelles l'IA embarquée apporte une valeur opérationnelle mesurable. Ils montrent comment des fonctions d'analyse, de détection ou de prédiction sont intégrées dans des systèmes embarqués pour répondre à des besoins tels que l'inspection, la maintenance, la surveillance ou l'assistance à la décision, sans dépendance permanente à des infrastructures distantes.

Inspection visuelle embarquée
Cas d'usage dans lequel un système embarqué analyse des images ou des flux vidéo issus de caméras afin de détecter des défauts, qualifier une conformité ou trier des objets, avec une exécution locale garantissant réactivité et robustesse en environnement industriel ou opérationnel.

Maintenance prédictive embarquée
Cas d'usage où des signaux issus de capteurs (vibration, acoustique, courant, température, pression) sont analysés par des modèles d'IA embarqués afin d'anticiper des défaillances, d'identifier des dérives de fonctionnement et de planifier des actions de maintenance adaptées.

Surveillance d'événements critiques
Cas d'usage consistant à détecter en temps utile des événements rares ou anormaux, tels qu'un incident, un comportement dangereux ou un dépassement de seuil, afin de déclencher une alerte ou une action immédiate du système embarqué.

Assistance à la conduite et à la mobilité
Cas d'usage dans lequel la perception multi-capteurs et la décision locale soutiennent l'assistance à la conduite, la navigation ou des fonctions d'autonomie partielle, avec des exigences élevées de temps réel et de fiabilité.

Optimisation locale d'énergie
Cas d'usage où un système embarqué ajuste dynamiquement des stratégies de gestion énergétique (charge, pilotage d'équipements, priorisation) à partir de mesures et d'estimations locales, afin d'améliorer l'efficacité énergétique sans dépendance continue à une infrastructure distante.

Aide à la décision en environnement isolé
Cas d'usage dans lequel l'IA embarquée fournit des recommandations, des alertes ou des diagnostics directement sur le terrain, dans des contextes à connectivité limitée ou inexistante, afin de soutenir l'action humaine ou automatisée.

Reconnaissance d'activités sur wearable
Cas d'usage où un dispositif portable analyse localement des données issues de capteurs inertiels, biométriques ou physiologiques pour identifier des activités, des gestes ou des états, tout en respectant des contraintes fortes d'autonomie énergétique.

Supervision locale d'infrastructures
Cas d'usage dans lequel des dispositifs embarqués déployés sur des infrastructures réalisent un diagnostic continu ou périodique de l'état d'équipements ou d'ouvrages, produisant des indicateurs exploitables même en conditions de connectivité intermittente.

13.6 Cas d'usage émergents

Les usages émergents reflètent l'évolution récente de l'IA embarquée vers des systèmes plus adaptatifs, coopératifs et responsables. Ils mettent en avant des approches visant l'adaptation locale des modèles, la coopération entre dispositifs, la sobriété des ressources et la prise en compte des enjeux environnementaux et sociaux, tout en conservant une exécution majoritairement embarquée.

Adaptation locale et apprentissage embarqué
Capacité limitée d'un système d'IA embarquée à ajuster certains paramètres ou seuils en exploitation afin de maintenir ses performances face à des environnements évolutifs, sous contraintes de ressources et de sûreté.

IA embarquée distribuée
Organisation dans laquelle plusieurs systèmes embarqués coopèrent pour produire une perception ou une décision collective, tout en conservant une exécution majoritairement locale.

IA embarquée frugale et responsable
Applications dans lesquelles l'IA embarquée est utilisée comme levier de sobriété numérique, en minimisant l'empreinte énergétique, les volumes de données traitées et la dépendance aux infrastructures distantes, tout en conservant une utilité opérationnelle.

IA embarquée pour la durabilité environnementale

Applications exploitant l'IA embarquée pour observer, analyser et agir localement sur des phénomènes environnementaux, à l'aide de dispositifs autonomes, distribués et faiblement consommateurs de ressources.

IA embarquée pour la confiance et l'acceptabilité sociale

Applications dans lesquelles l'IA embarquée contribue à renforcer la confiance des utilisateurs et des parties prenantes par des décisions locales maîtrisées, une transparence accrue du fonctionnement et une intégration respectueuse des contextes humains et sociaux.