Chapitre 1: L’intelligence artificielle embarquée

L'apprentissage automatique a ouvert des portes à de nombreuses applications. Il devient de plus en plus omniprésent, étant ainsi présenté dans presque tous les domaines tels que la santé, l'agriculture, la surveillance, entre autres. Le concept de l'apprentissage automatique est simple et peut être expliqué comme suit : faire apprendre à la machine à réaliser une tâche à partir de données, qui sont des observations de la réalisation de cette tâche. Une des techniques de l'apprentissage automatique est devenue la star : c'est l'apprentissage profond. L'apprentissage profond repose sur des réseaux neuronaux artificiels. Ces réseaux neuronaux artificiels sont inspirés du cerveau humain. C'est en fait un modèle formel de neurone biologique, qui peut être composé de plusieurs couches de neurones. Ainsi, un modèle d'apprentissage profond est un modèle qui comporte plusieurs couches de neurones. Chaque couche est responsable de l'apprentissage d'une représentation, en d'autres termes, de caractéristiques pouvant être utilisées pour accomplir une tâche donnée. Les modèles d'apprentissage profond sont utilisés dans presque toutes les tâches : vision par ordinateur, traitement du langage naturel, et bien d'autres. Ces modèles sont entraînés sur d'énormes volumes de données et nécessitent beaucoup de puissance de calcul, étant également très volumineux. Une fois entraînés, ces modèles sont déployés dans le cloud pour être utilisés dans des situations réelles.

Dans certaines applications critiques nécessitant un traitement en temps réel ou une faible latence, exécuter le modèle dans le cloud n'est pas une solution adaptée. Dans le contexte de l'Internet des objets (IoT), les données sont collectées par l'appareil en périphérie et ensuite envoyées au cloud, ce qui pose également un problème de latence. Dans certains cas, les données collectées par les appareils finaux sont sensibles, et la confidentialité est nécessaire. De plus, dans certains environnements, il n'y a pas d'accès au cloud en raison d'une absence de couverture réseau, ou les systèmes sont destinés à fonctionner dans des environnements difficiles. Pour toutes ces raisons, un nouveau paradigme a émergé : l'Edge AI. L'Edge AI est une intelligence artificielle exécutée sur l'appareil ou simplement à la source des données, afin de pallier toutes ces limitations du cloud, et aussi pour des raisons économiques et environnementales.

Chapitre 2: Applications

- Voice assistant
- Wild animal surveillance
- Autonomous vehicles
- Smart city
- Speech recognition
- Containers trackers

Chapitre 3: Contraintes

Systèmes embarqués sont des systèmes contraints en ressources mémoire, en puissance de calcul. D'ailleurs la programmation de ces systèmes exigent une connaissance de la plateforme matérielle pour une bonne utilisation des ces ressources. Les modèles d'IA sont volumineux et gourmand en puissance de calcul dont très energivores. Seuls les modèles optimisés correspondant aux contraintes des systèmes embarqués sont utilisés. Donc des techniques d'optimisation visant à reduire l'empreinte memoire des ces modèles et/ou assurant la bonne gestion de la communication entre RAM et la mémoire Flash, tout en gardant une précision satisfaisante, doivent être utilisée.

Techniques d'optimisation
Plusieurs techniques d'optimisation existent. Elles optimisent généralement l'empreinte memoire du modèle, la consommation énergétique, ou architecture des modèles d'IA.

Chapitre 4: Techniques de Compression du modèle existant

- Quantification
- Pruning
- Knowledge distillation
- Neural search
- Mise en place de Modèles/Blocks respectant les contraints du système embarqué
- Outils Soft-hard
- Micro-controllers, Specialized hardware, Powerful embedded systems
- TF lite, Edge impulse, KubeEdge
- End-to-end Edge AI projects
- Projets opensource