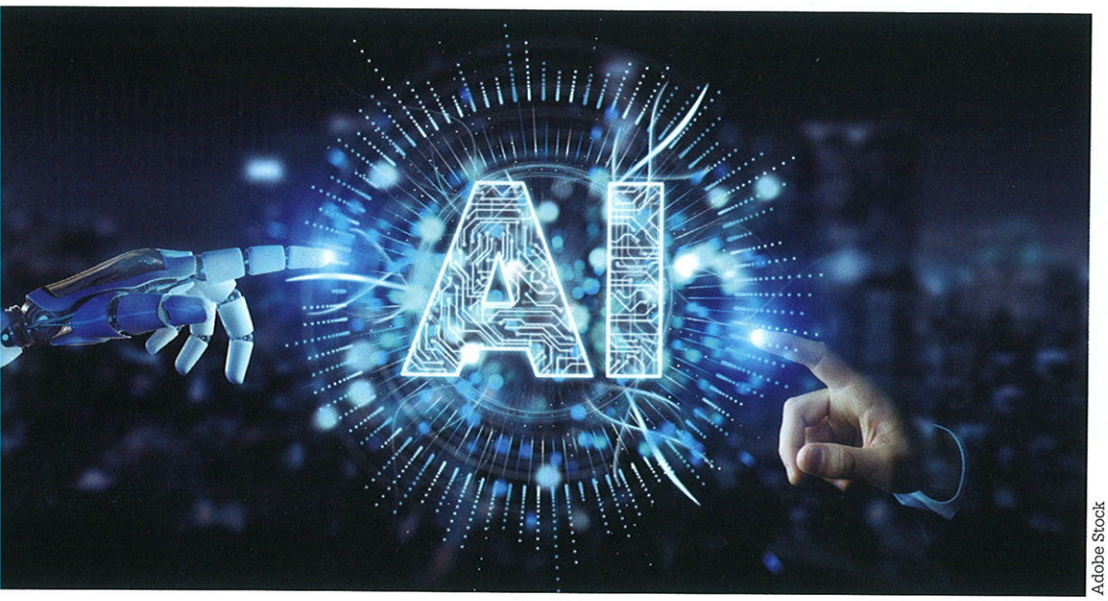
CONTRÔLE ET SURVEILLANCE

Placer l'IA au plus près de l'événement à surveiller

Pour le concepteur de matériels de surveillance, le choix de la disposition des unités de traitement des données issues des capteurs est un problème. Faut-il effectuer un traitement local, au plus proche des événements à surveiller, ou un traitement dans le cloud ou encore dans des unités relais ? Ce texte se veut une aide à la résolution de ce type de dilemme en apportant un éclairage sur les derniers développements en IA.

En un peu plus de dix ans, depuis que les chercheurs ont découvert de nouvelles techniques pour améliorer l'efficacité de l'apprentissage en profondeur (*deep learning*), ce dernier est devenu une technologie pratique qui constitue désormais la base d'un certain nombre d'applications faisant appel à l'intelligence artificielle (IA). Beaucoup de ces applications sont hébergées dans le cloud par l'intermédiaire de puissants serveurs. En effet, ces tâches impliquent parfois le traitement de sources comportant un très grand nombre de données, telles que des images, des vidéos et des fichiers audio. Ces serveurs font souvent appel aux performances supplémentaires fournies par les matériels d'accélération, qui s'étendent des unités de traitement graphique aux dispositifs personnalisés. Cela devient particulièrement important lors de processus qui exigent des calculs intensifs au cours desquels un réseau neuronal est formé à partir de nouvelles données.

En règle générale, les processus d'inférence qui utilisent un réseau formé pour évaluer de nouvelles données effectuent beaucoup moins de calculs que les processus formés. On trouve également des charges de travail impliquant des sources faisant appel à un volume plus faible de données, telles que les lectures de capteurs de dispositifs IoT, où la formation et l'inférence peuvent être effectuées sur du matériel moins performant. En conséquence, les concepteurs de systèmes constatent que les charges de travail ne doivent pas nécessairement être situées sur le cloud une fois le modèle d'IA formé, bien que de nombreux services actuels s'y trouvent encore pour des raisons commerciales. Au lieu de cela, le modèle formé peut être transféré vers une machine locale pour un traitement plus proche de la source des données.



Adobe Stock

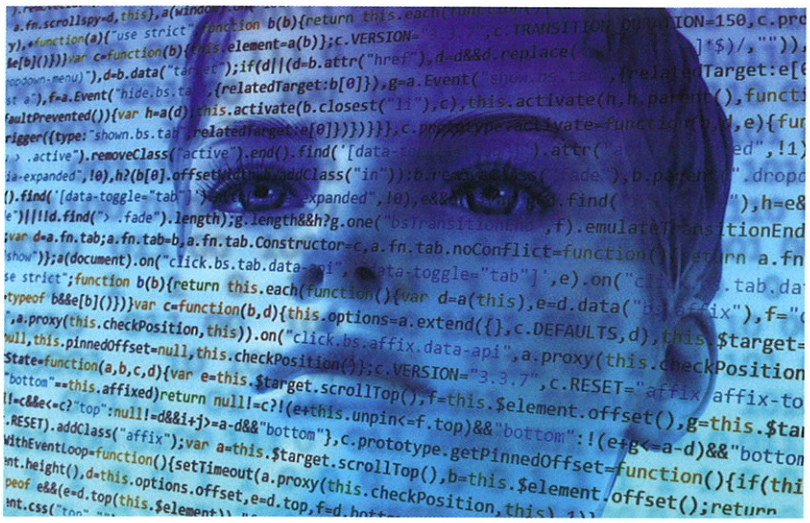
↗ Une fois le modèle d'IA formé, il peut être transféré dans une machine locale de façon à réaliser un traitement au plus proche de la source des données.

Les avantages de l'IA à la périphérie de l'événement

Il existe un certain nombre de raisons pour rapprocher les modèles d'IA de la périphérie du réseau. Un facteur important est la confidentialité et l'acceptation par les utilisateurs. Par exemple, les clients utilisant des dispositifs tels que les haut-parleurs intelligents sont de plus en plus préoccupés par le fait que leurs conversations privées sont régulièrement enregistrées et téléchargées vers des services dans le cloud qui pourraient être pris en charge localement.

Au sein des applications de contrôle industriel, qui commencent désormais à utiliser l'IA pour la surveillance de l'état des machines ou l'optimisation des processus, les préoccupations relatives à la confidentialité des données de production exigeront également que le volume de données le plus faible possible soit transféré vers le cloud. Pour de nombreuses applications industrielles, la fiabilité et la rapidité de la connexion au cloud posent aussi des problèmes. De nombreux systèmes, qu'ils soient situés en atelier ou sur des sites à distance, ne disposent pas des connexions à large bande passante nécessaires pour prendre en charge l'inférence basée dans le cloud.

Les systèmes de contrôle sont également touchés par une latence élevée au niveau des communications. Si les modèles d'intelligence artificielle sont utilisés dans des systèmes de contrôle en boucle fermée, tout retard dans le téléchargement d'une mise à jour à partir du cloud conduira à des inexactitudes et à de l'instabilité. Certains systèmes peuvent utiliser un mélange de traitement dans le cloud et de traitement local. Les caméras de surveillance, par exemple, préserveront la bande passante du réseau si elles identifient localement les menaces immédiates, puis font ensuite appel au cloud pour effectuer un traitement supplémentaire dans les situations non prises en charge par les



Geralt

← Pour de nombreuses applications industrielles, la fiabilité et la rapidité de la connexion au cloud posent des problèmes. En effet, de nombreux systèmes ne disposent pas des connexions à large bande passante nécessaires pour prendre en charge l'inférence basée dans le cloud.

modèles locaux. La durée de vie des batteries augmente également de manière significative lorsque l'IA se trouve en périphérie, car beaucoup moins de données sont envoyées sur le réseau, ce qui réduit les coûts liés au réseau et au cloud.

Application de l'IA dans la maintenance prévisionnelle

La maintenance prévisionnelle est actuellement le cas d'utilisation le plus populaire pour les applications industrielles connectées. Elle offre un retour sur investissement élevé en raison de sa capacité à réduire la fréquence des inspections sur site. Les temps d'arrêt de la machine peuvent également être réduits en identifiant avec précision le temps utile restant d'un composant en service, ce qui permet d'exploiter pleinement sa durée de vie opérationnelle disponible sans risquer une panne en cours de fonctionnement. La prévision de la durée de vie sur la base des données opérationnelles permet également d'optimiser la planification de la maintenance et d'identifier avec précision les besoins en pièces de rechange. Les clients utilisant la maintenance prévisionnelle constatent déjà des gains d'efficacité allant de 20 à 25 %. Avec la maintenance prévisionnelle, les capteurs d'une machine-outil peuvent détecter que des variations de température associées à une augmentation du bruit ou des vibrations, par le biais d'un ensemble de microphones et d'accéléromètres, indiquent un problème potentiel. Avec les techniques algorithmiques traditionnelles, il peut être extrêmement difficile de faire le lien entre un problème potentiel et plusieurs ensembles de lectures de capteurs en temps réel. Les modèles basés sur l'IA peuvent, grâce à des données de séries chronologiques associées à des informations obtenues en temps réel, effectuer une interprétation pour déterminer avec précision l'état du système. De tels modèles peuvent tirer parti de techniques telles que la fusion de capteurs pour déterminer si la situation nécessite une tâche de maintenance ou non. La complexité et les exigences de latence du modèle détermineront la possibilité d'exécuter le modèle localement.

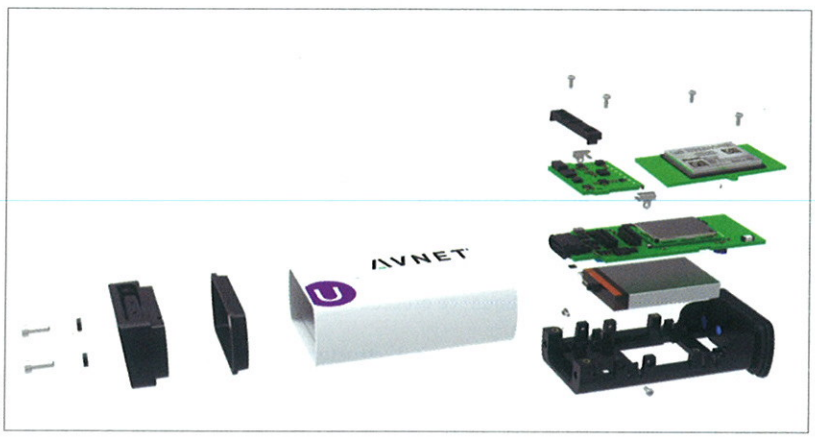
Du cloud à la passerelle ou à la périphérie

Il existe deux approches du traitement local. La première consiste à utiliser la puissance de traitement disponible sur le dispositif-même. La faisabilité dépend de la complexité du modèle d'IA et de la puissance de traitement disponible. Dans le cas des nœuds de capteurs de faible puissance, il ne sera pas possible d'exécuter un modèle à grande échelle d'apprentissage en profondeur, ne serait-ce que pour l'inférence. Toutefois, les algorithmes de l'apprentissage automatique pouvant prendre de nombreuses formes, il peut ne pas être nécessaire d'exécuter une tâche aussi lourde de manière locale uniquement.

La deuxième option consiste à transférer tout ou partie du traitement sur un autre dispositif. Par exemple, le dispositif peut exécuter un modèle d'IA simplifié qui effectue une analyse initiale des données. Dans le cas d'un système de surveillance à l'écoute des activités, ce modèle peut simplement déterminer si le signal est un bruit de fond, tel que du vent, si un individu passe à proximité ou si le son provient d'une fenêtre brisée. Des fonctions plus complexes peuvent être transférées vers une passerelle qui concentre les données de plusieurs caméras et d'autres dispositifs de sécurité. Les dispositifs peuvent exécuter un traitement préalable pour aider à rationaliser le travail du modèle sur passerelle.

Un matériel puissant, tel que les dispositifs à logique programmable Xilinx fournis par la plateforme Ultra96 d'Avnet, peut effectuer des inférences basées sur des modèles sophistiqués d'apprentissage en profondeur et d'autres algorithmes d'apprentissage automatique complexes. La passerelle locale peut même mettre à jour le modèle en collectant des données et en effectuant une formation par lots lorsque la charge de travail d'inférence est réduite. La passerelle peut également collecter des données importantes pour reformer les modèles et les transmettre vers un serveur cloud chaque jour, chaque semaine ou chaque mois.

Au niveau du dispositif, l'IA fournit la latence la plus faible au niveau



Farnell

← La maintenance prévisionnelle est actuellement le cas d'utilisation le plus populaire pour les applications industrielles connectées. Elle offre un retour sur investissement élevé en raison de sa capacité à réduire la fréquence des inspections sur site.

des communications. Toutefois, la vitesse de traitement plus rapide d'une passerelle locale peut s'avérer plus efficace que le processeur d'un dispositif moins performant et offrir les meilleurs paramètres de latence et de débit. Le transfert vers le cloud dépendra de facteurs tels que la bande passante disponible de la connexion internet et la distance avec les serveurs eux-mêmes : la vitesse finie de la lumière impose une limite à la faiblesse de la latence si le cloud est utilisé pour des charges de travail d'inférence.

Options technologiques d'apprentissage

Le développeur a le choix entre plusieurs types de modèles pour une application spécifique. Les premiers modèles d'IA utilisaient des structures telles que des arbres décisionnels ou des systèmes experts qui demandaient beaucoup de travail aux spécialistes de domaines pour relier les ensembles d'informations aux causes et aux résultats probables. L'apprentissage automatique a amélioré les arbres décisionnels en rendant possibles certaines techniques telles que les forêts d'arbres décisionnels. Une telle approche crée plusieurs arbres décisionnels à partir de données d'apprentissage et les analyse en parallèle pour calculer une moyenne représentant le résultat le plus probable sur la base des données d'apprentissage. Une forêt aléatoire est un exemple de système d'apprentissage supervisé : elle relie les données fournies par les développeurs du système aux informations à partir desquelles le modèle d'apprentissage automatique apprend à gérer les relations.

L'apprentissage en profondeur est un autre exemple de technologie d'apprentissage supervisé, car il repose sur l'étiquetage préalable des données de formation. Dans un système de classification d'images, par exemple, le modèle extrait des données provenant de l'image et les applique à une couche de neurones simulés. Les résultats de la première couche neuronale passent successivement à travers un grand nombre d'autres couches. Certaines combinent les résultats de plusieurs neurones de la couche précédente pour produire une seule valeur qui est transmise à la suivante. De cette manière, les réseaux neuronaux profonds effectuent une réduction dimensionnelle. Il s'agit d'une étape essentielle dans la conversion de données multidimensionnelles complexes telles que des images ou des fichiers audio de manière à pouvoir les utiliser pour fournir une classification finale. Le réseau neuronal est capable d'effectuer une classification, car il peut apprendre comment différents arrangements de coefficients dans les couches neuronales répondent à différentes images pour fournir des résultats correspondant aux étiquettes de formation.

La formation d'un modèle d'apprentissage automatique ne doit pas nécessairement reposer uniquement sur des données étiquetées. Des algorithmes d'apprentissage automatique non supervisés, tels que les regroupements, peuvent trouver des modèles dans les données sans aide supplémentaire. Un tel processus peut être très utile dans les systèmes de contrôle industriels où plusieurs capteurs sont utilisés ou lorsque le comportement des séries chronologiques des informations entrantes est important. Dans le cas de la surveillance des états d'une machine-outil, l'amplitude de la vibration n'indique pas forcément un problème, mais peut simplement être une conséquence du processus. Cependant, un schéma de mouvement dans les données de la série chronologique associé à des changements rapides de température peut indiquer un problème nécessitant une tâche de maintenance. Ces différences peuvent être révélées par la séparation des données en groupements faciles à différencier lorsque les données sources, si elles sont utilisées directement, ne montrent aucune tendance claire.

Un système d'apprentissage automatique non supervisé peut détecter des modèles dans les lectures de capteurs qui peuvent être utilisés pour réduire la quantité de données qui est ensuite transmise à un algorithme supervisé formé sur des machines fonctionnant sous différents types de contraintes et d'états. Étant donné que le regroupement permet de réduire la quantité de données devant être transférée, un modèle de conception utile pour la mise en œuvre dans les systèmes IoT consiste à effectuer cette activité sur le nœud du dispositif. Le modèle dérivé de l'apprentissage supervisé peut être exécuté sur la passerelle, voire dans le cloud, en fonction des compromis entre la latence, les performances et le coût.

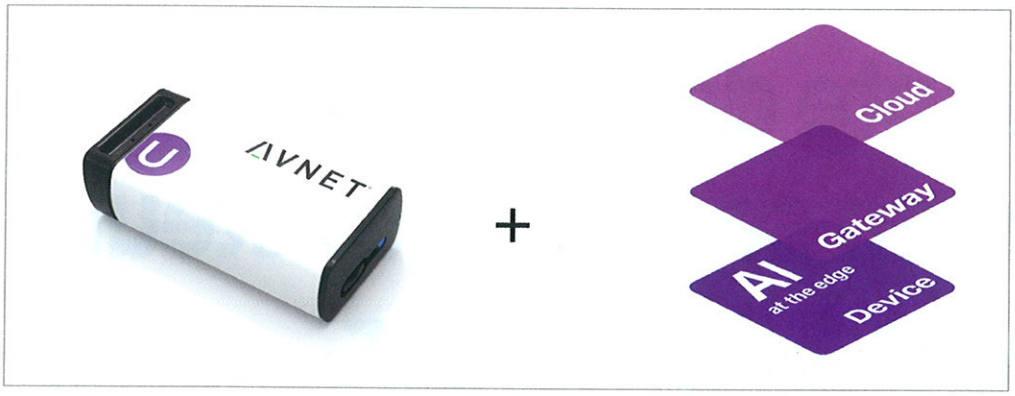
QUELQUES DÉFINITIONS

- → Processus d'inférence : méthode permettant de tirer des conclusions ou d'identifier des objectifs à partir de données collectées.
- → Processus formé : procédé qui a déjà été testé et utilisé.
- → Apprentissage automatique : procédé apprenant par lui-même à partir de données collectées, sans aide de procédés annexes.
- → Apprentissage en profondeur ou deep learning : une méthode d'apprentissage machine.
- → Apprentissage supervisé : le contraire de l'apprentissage automatique ; ce procédé nécessite l'aide de procédés annexes.
- → Arbre décisionnel : structure du processus de décision.
- → Forêt d'arbres décisionnels : ensemble d'arbres décisionnels.

Une seule solution : l'IA en périphérie

Un facteur important à prendre en compte pour l'apprentissage automatique en tant que technologie est que sa mise en œuvre nécessite souvent des connaissances et une expérience approfondies afin de tirer le meilleur parti des différentes formes d'IA disponibles. Octonion a développé une solution logicielle pour résoudre ce problème. Elle est intégrée au dispositif SmartEdge Agile d'Avnet, et permet aux ingénieurs de développer des solutions d'IA pour les systèmes IoT impliquant un traitement en périphérie important sans recourir à une aide coûteuse, ni acquérir une expertise approfondie. La solution fonctionne à trois niveaux : dispositif, passerelle et cloud. La partie matérielle du dispositif, Avnet SmartEdge Agile, est un nœud de capteur autonome à faible consommation. Ce dispositif permet d'accéder à différents types de capteurs, notamment des accéléromètres, des gyroscopes, des magnétomètres, ainsi qu'à des capteurs permettant de mesurer la pression, la température, l'humidité et la proximité. Un microphone est disponible pour enregistrer les informations audio entrantes.

La partie matérielle du capteur intelligent communique avec une passerelle locale à l'aide du BLE. Des versions prenant en charge d'autres technologies réseau orientées IoT telles que LoRaWAN, SigFox, les



Farnell

↗ Dans le dispositif SmartEdge Agile d'Avnet, les capteurs intelligents communiquent via Bluetooth LE avec un dispositif Android ou iOS sur passerelle ou avec un logiciel exécuté sur carte informatique. La couche dans le cloud peut être déployée sur AWS, sur Microsoft Azure ou avec des solutions de serveur personnalisées.

données cellulaires et le Wi-Fi sont en cours de développement. Le module sur passerelle peut prendre la forme d'un dispositif intelligent Android ou iOS ; ou bien, le logiciel sur passerelle peut être exécuté sur un dispositif Raspberry Pi ou une plateforme Linux similaire pouvant fournir un accès au cloud. La couche dans le cloud peut être déployée sur AWS, sur Microsoft Azure ou avec des solutions de serveur personnalisées. La coordination des différentes couches est assurée par l'environnement logiciel Brainium d'Octonion qui offre une prise en charge de bout en bout dans le cloud dans un environnement sans code. La sécurité est la clé de tous les déploiements IoT et un élément important du système afin de protéger la confidentialité des données utilisées pour la formation des modèles et lors des inférences. Le système développé pour Brainium utilise un chiffrement matériel AES, ainsi qu'un stockage de clé de cryptographie inviolable basé sur la partie matérielle pour protéger les données dès leur lecture par le MCU sur le dispositif SmartEdge Agile d'Avnet. Les images du micrologiciel utilisées par les dispositifs pour exécuter les fonctions d'IA sont également cryptées avec une validation de signature numérique. Lorsque les données sont envoyées, les canaux de communication utilisent le cryptage TLS pour empêcher les intrus d'observer les messages en périphérie dans le cloud.

L'environnement logiciel Brainium d'Octonion offre une combinaison d'apprentissage automatique non supervisé et supervisé. Par exemple, il peut être formé sur les anomalies dans la représentation des séries chronologiques des données de capteur brutes. De plus, il regroupera les données de différents scénarios pour identifier les modèles de données communs. Ces modèles de données peuvent être intégrés dans un constructeur de modèle flexible au sein du logiciel AI Studio afin que les développeurs puissent adapter le modèle à chaque cas d'utilisation cible. Pendant la procédure d'apprentissage, le dispositif collecte des échantillons de données, les sécurise et les crypte avant de les envoyer vers le logiciel AI Studio qui est exécuté dans le cloud. Ce logiciel apprend à partir des échantillons pour générer des modèles d'IA qu'il retransmet au dispositif périphérique pour le travail d'inférence. Tout dispositif appartenant au matériel de déploiement du client peut recevoir ce modèle d'IA et fonctionner en mode autonome pour effectuer la surveillance et l'analyse. L'environnement restant est cohérent, du prototypage à la production à grande échelle. Pour le déploiement, les clients peuvent utiliser le matériel SmartEdge Agile d'Avnet disponible dans le commerce ou adapter la conception à leurs besoins avec leurs propres implémentations. En exploitant le calcul intelligent des données en périphérie, la conception du SmartEdge Agile d'Avnet, alimenté par le logiciel Brainium, minimise le volume de données devant être transmis à partir de chaque dispositif périphérique. Un niveau de sécurité élevé est intégré au système là où les données sont transmises dans le cloud. Il en résulte un système qui fournit tous les outils nécessaires au développement et au déploiement rapides de systèmes compatibles avec l'IA.

CLIFF ORTMEYER, Global Head of Technical Marketing (Farnell)