Opinions Libres
le blog d'Olivier Ezratty

Et l'IA entra dans les smartphones !

Lors de l'IFA 2017 à Berlin, le chinois **Huaweî** annonçait le **Kirin 970**, le premier chipset mobile intégrant de l'IA, leur "Neural Processing Unit" (**vidéo**), issu de sa filiale de semiconducteurs HiSilicon. Il équipera les futurs smartphones de Huaweï, les Mate 10 qui seront lancés en octobre 2017. Ce Kirin 970 permet de traiter plus rapidement la reconnaissance d'images et probablement également, celle de la parole. C'est une véritable première même si elle ne comprend pas forcément de tour de force technologique particulier. C'est une forme d'innovation par l'intégration qui va dans le sens du vent et qui en précède d'autres.

Apple faisait de même la semaine suivante en lançant ses iPhone 8 et X qui intègrent un composant dédié à l'IA, le A11 Bionic Neural Engine. La tendance est donc bien lancée pour mettre de l'IA dans les smartphones sachant qu'elle avait déjà été entamée dans d'autres catégories d'objets connectés comme les caméras de surveillance.

Le leader mondial des chipsets mobiles, l'Américain **Qualcomm** proposait jusqu'à présent son architecture **Zeroth** pour faire tourner des fonctions de deep learning dans les smartphones. Dans la pratique, il ne s'agissait que d'un kit de développement logiciel (SDK) qui exploitait des fonctions standards d'un DSP Hexagon intégré dans leurs chipsets Snapdragon. Les processeurs mobiles de **Nvidia** comprenaient aussi un grand nombre de cœurs pour exécuter des applications de deep learning mais on les trouve plutôt dans les véhicules à conduite assistée et autonomes comme les Tesla que dans des smartphones.

Comme les annonces similaires se suivent en général, il ne serait pas étonnant que d'autres fournisseurs de chipsets mobiles s'y mettent, surtout Mediatek et Samsung, puis Qualcomm, qui pourrait bien faire une annonce de ce genre avant, pendant ou après le CES 2018.

Les réseaux de convolution du deep learning

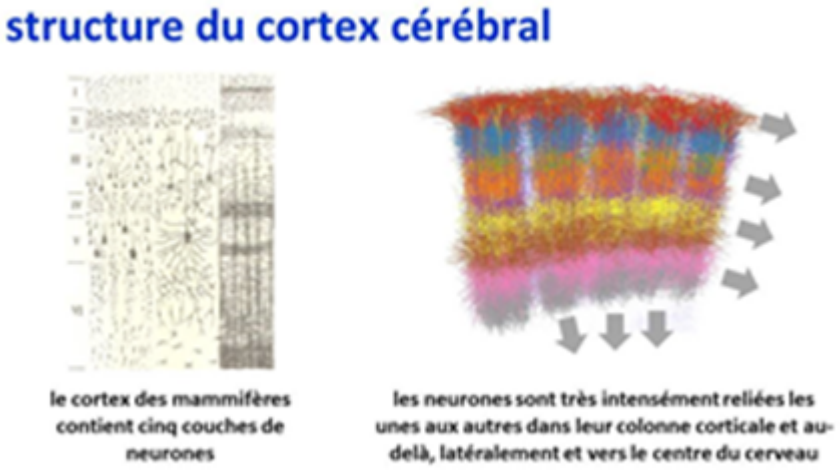
Pour comprendre à quoi peut servir ce Kirin 970 doué d'IA, il faut rentrer un peu dans le lard de l'une des techniques de l'IA, les réseaux de convolution. C'est une des techniques de deep learning qui est principalement utilisée pour reconnaitre le contenu d'images et détecter les phonèmes dans la reconnaissance de

la parole. Cela colle bien avec les applications de ce Kirin 970 présentées lors de son lancement.

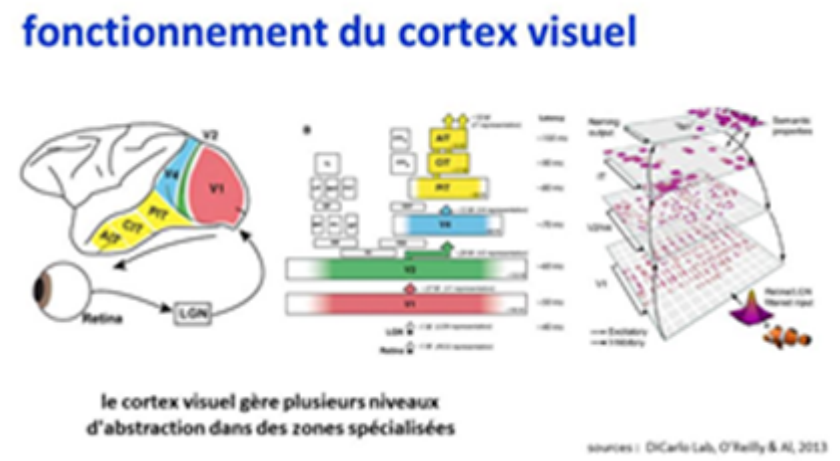
Les réseaux de neurones convolutionnels, CNN, ou ConvNets (convolutional neuron networks), ont été inventés par le français **Yann LeCun** en 1988 puis perfectionnés en 1989, 1998 et jusqu'à présent. Ils servaient au départ à la reconnaissance de chiffres, puis sont passés à la reconnaissance d'images. Il explique cela très bien dans la conférence inaugurale de sa chaire du Collège de France du 12 février 2016 (**vidéo**).

Le deep learning exploite des réseaux de neurones multicouches qui reconnaissent le contenu d'un objet complexe (son, image, vidéo, texte) en le décomposant progressivement d'abord en composantes de bas niveau, puis en montant progressivement le niveau d'abstraction jusqu'à aboutir au descriptif de l'objet.

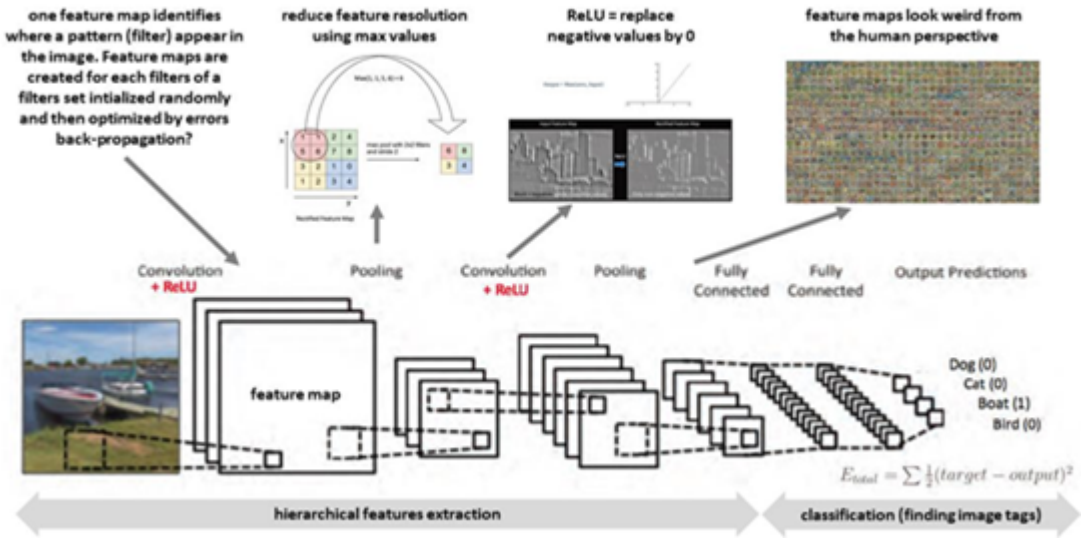
Les ConvNets perfectionnent ce modèle en s'inspirant fortement du fonctionnement du cortex visuel des mammifères qui est structuré, de près, dans des colonnes corticales faites de cinq couches de neurones et qui, de loin, comprend des aires spécialisées qui élèvent progressivement le niveau d'abstraction des objets reconnus. On a pu le vérifier sur des … chats !



Contrairement au cortex humain, les ConvNets qui font de la reconnaissance d'images utilisent des représentations à très basse résolution. Les algorithmes utilisés sont cependant si puissants qu'ils permettent de générer des taux de reconnaissance d'images meilleurs que ceux de l'Homme ! Qu'est-ce que cela serait si la résolution utilisée était la même que dans l'œil et le cortex humains ! On y arrivera certainement un jour et on pourra alors parler d'IA d'œil de lynx.



Les ConvNets utilisent plusieurs techniques enchainées les unes avec les autres avec des filtres et des “feature maps” qui consistent à identifier des formes dans les images. Chaque “feature map” est une cartographie de l’apparition d’un filtre dans l’image analysée. Un ConvNet utilise un jeu de plusieurs filtres initialisé aléatoirement. Les filtres sont ensuite affinés par différentes techniques dont la rétropropagation d’erreurs dans l’ensemble du réseau, un mécanisme qui est appliqué pour toutes les images d’un jeu d’entrainement qui peut comprendre des millions d’images. C’est très consommateur de ressources machine et l’annonce de Huaweï ne précise d’ailleurs pas si le Kirin 970 sert à l’exécution d’un réseau de neurones ou également à son entrainement qui est la partie la plus consommatrice de ressources. Un réseau de neurones convolutionnel ne reconnait pas magiquement un chat ou un visage ! On lui a indiqué auparavant de quoi il s’agissait !

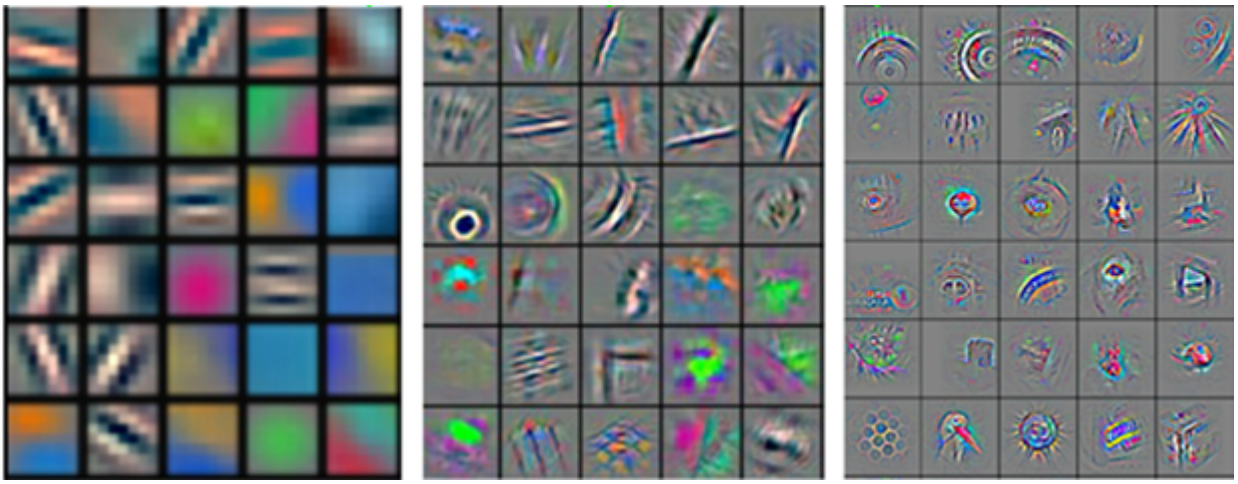


Chaque feature map se voit appliquée une réduction de résolution (pooling) puis une suppression des valeurs négatives (ReLU pour Rectified Linear Units) pour réduire la quantité de travail à réaliser sur les couches suivantes. Le processus est répété sur plusieurs niveaux ou couches, chaque “feature map” issue d’un niveau devenant une image qui subit un traitement équivalent dans le niveau suivant. C’est assez abstrait à comprendre, j’en conviens bien ! J’ai mis moi-même beaucoup de temps à comprendre !

A la fin de l’histoire, la dernière couche de “feature maps” est reliée à une liste de tags avec une probabilité de correspondance via quelques couches de neurones dites « fully connected », à savoir que tous les neurones d’une couche sont liées à celles de la couche suivante. C’est là qu’un chat ou un bateau sont reconnus dans l’image et que plusieurs objets peuvent être reconnus dans une même image. La dernière couche de cet empilement est un ensemble de neurones dont le nombre est égal au nombre d’objets différents à reconnaitre. Il peut être très grand mais doit rester raisonnable pour tenir compte des capacités du matériel. Ainsi, les meilleurs moteurs de reconnaissance d’images comme ceux de Google n’ont-ils pour l’instant que quelques dizaines de milliers de classes d’objets dans cette dernière couche de réseaux de neurones.

A chaque couche d’un réseau convolutionnel, le nombre de “feature maps” augmente et leur taille diminue. Les “feature maps” étant optimisées automatiquement, leur forme n’est pas interprétable par le cerveau humain. C’est la magie des ConvNets : ils créeent des niveaux de représentations hiérarchiques intermédiaires des images qui optimisent leur reconnaissance avec une vision plus statistique que sémantique, sans que l’on puisse comprendre comment ils fonctionnent pas à pas. C’est particulièrement éloquent avec la reconnaissance d’images, cf l’exemple *ci-dessous*. D’où le fameux soucis de “non explicabilité” des algorithmes du deep learning. Cela ne va pas nous empêcher de dormir pour la reconnaissance d’une image mais peut être délicat pour d’autres usages comme ceux qui sont liés au traitement du langage, même si ceux-ci exploitent généralement d’autres formes de réseaux de neurones qu’on appelle les réseaux de neurones récurrents (RNN)

et qui n'exploitent pas par ces mystérieuses "feature maps".



Le nombre et la taille des filtres de chaque couche du ConvNet sont décidés par le développeur en fonction de la nature des images à traiter et des objets à y reconnaitre. Plus ils seront grands, plus le calcul sera long pour l'entrainement du système. Les concepteurs de réseaux de neurones passent donc beaucoup de temps à peaufiner leurs modèles pour améliorer les performances de l'entrainement.

On peut distinguer les ConvNets selon le nombre de dimensions des données reconnues : **1D** (une dimension) pour le texte, la reconnaissance de genre de musique, des prédictions temporelles sur une seule variable, **2D** (deux dimensions) pour les images, pour la reconnaissance de la parole qui associe fréquence audio et temps, puis **3D** (trois dimensions) pour le traitement de vidéos et d'imagerie médicale 3D. Rien d'empêche d'ajouter d'autres dimensions si l'usage l'exige.

Pour en savoir plus, voici une bonne explication des ConvNets en trois parties : A Beginner's Guide To Understanding Convolutional Neural Networks de Adit Deshpande (un étudiant aux USA), **partie 1**, **partie 2** et **partie 3**, datant de 2016.

Dans la pratique, l'exécution rapide d'un réseau de neurones convolutionnel nécessite de pouvoir effectuer les opérations mathématiques évoquées : des multiplications de matrices générant des matrices, la réduction de résolution de matrices, et la multiplication de vecteurs par des matrices de poids synaptiques pour générer un nouveau vecteur, correspondant à la fin du processus de reconnaissance ("fully connected"). Ce sont des primitives mathématiques relativement simples au regard d'autres mécanismes mathématiques déjà mis en œuvre dans les processeurs mobiles comme les techniques d'analyse du signal à base de transformées de Fourrier ou les algorithmes de compression de vidéo type H264 ou H265.

Il est aussi important de pouvoir gérer de manière optimale l'entrainement du système si toutefois le système embarqué a cette capacité. Pour ce faire, il faut pouvoir gérer la propagation des erreurs en remontant le circuit à partir de l'arrivée. Ce n'est pas trop compliqué pour la multiplication de vecteurs par des matrices synaptiques qui peuvent être inversées. L'entrainement peut être nécessaire pour reconnaitre un nouveau type d'objet ou la voix de l'utilisateur même si certains modèles permettent de s'en passer.

Mais je n'ai évoqué ici qu'une seule des méthodes de ConvNet : le tagging d'images. Il existe bien d'autres méthodes adaptées à d'autres usages, comme le découpage d'une image en objets ou la modification d'images en imitant d'autres styles d'images. Le deep learning est une discipline bouillonnante et la diversité des réseaux de neurones associée est étonnante. Un processeur peut normalement en supporter la diversité s'il est bien conçu.

L'IA dans le Kirin 970

Lors de l'annonce à l'IFA, les caractéristiques exactes de ce NPU du Kirin 970 n'ont pas été vraiment précisées. On savait que c'est un schmillblick d'IA. C'est à peu près tout tout. Côté spécifications, il a une puissance de calcul de 1,92 TeraFlops et la capacité à reconnaitre 2000 images en une minute alors que la génération précédente traditionnelle de chipsets Kirin en reconnaissait seulement une soixantaine. Soit dit en passant, on peut se demander à quel usage cela correspond. Une petite division met la puce à l'oreille : 2000/60 donne 33,33, pas loin de 30 images par secondes, qui est le nombre d'image moyen d'une vidéo. Cela veut donc dire que le NPU peut suivre en temps réel un objet dans une vidéo. Aux usages et aux applications d'en faire quelque chose !

Grâce à **cet article** bien renseigné paru après le lancement, on a pu en savoir un peu plus sur ce NPU. Il utilise une conception de circuit provenant d'une startup chinoise de Beijing, **Cambricon Technology**. Cette société a été créée en 2016 et vient de lever au mois d'aout la bagatelle de $100M, auprès d'un investisseur public chinois qui ressemble à notre Bpifrance. HiSilicon n'a pas utilisé telle que un bloc de processeur neuromorphique de Cambricon Technology. Ils ont travaillé ensemble pour le personnaliser et l'intégrer dans le Kirin 970 et notamment pour l'adapter au processus de fabrication du chipset qui est en intégration à 10 nm, fabriqué par TSMC à Taïwan. Cambricon Technology est à l'origine de la conception du Cambrian 1A, un chipset neuromorphique pour l'exécution de réseaux de neurones dans des terminaux.


CHINA MONEY NETWORK

Investment, Venture Capital

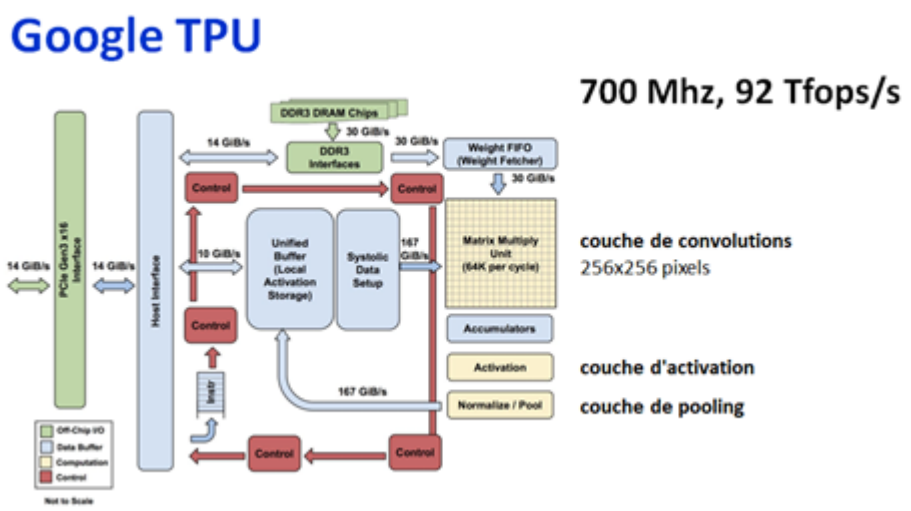
China's State Development & Investment Corp Leads $100M Round In AI Chip Maker Cambricon

PAN YUE

August 18, 2017 — 22:59 HKT

De l'architecture du NPU intégré dans le Kirin 970, on ne sait qu'une chose : il contient des multiplicateurs de matrices de 3×3 pixels, sans que leur nombre soit précisé. Or 3×3 ressemble à la taille d'un filtre typique de réseau convolutionnel et non pas celle d'une image ou de feature map.

On peut comparer cela avec d'autres processeurs neuromorphiques dédiés aux réseaux de neurones. La première génération de Tensor Processing Units (TPU) de **Google** annoncée en 2016 comportait un multiplicateur de matrices de 256×256 pixels. C'est d'ailleurs la résolution habituelle de traitement d'images pour leur reconnaissance. Dans la dernière génération 2017, ils ont deux multiplicateurs de 128×128 pixels. Chez **Nvidia**, la dernière génération de GPU adaptée à ces traitements, le V100, en plus de milliers de cœurs destinés à des opérations élémentaires sur des nombres entiers et flottants, ils comprennent 640 cœurs dotés de multiplicateurs de matrices de 4×4 pixels. C'est une des composantes des processeurs neuromorphiques mais pas la seule.



Techniquement parlant, le Kirin 970 un **ASIC**, à savoir un circuit dont toutes les portes logiques définissant son fonctionnement sont gravés dans le dur. L'alter égo des ASIC sont les variantes de **FPGA** dont les portes sont programmables par logiciel. C'est une technique retenue pour certains chipsets neuromorphiques, notamment ceux qui sont produits par des startups comme le français **Scortex**. Un FPGA coute moins cher à produire en petite série qu'un ASIC mais il est moins rapide et consomme plus d'énergie. Un ASIC n'est intéressant que si le composant est produit en volume. Cela rappelle, par analogie la différence entre l'impression 3D qui est intéressante en faible volume par rapport à la fabrication plus traditionnelle avec des moules (et pressage ou injection) qui ne sont rentabilisés qu'avec un gros volume de production.

Comme les smartphones de Huaweï sont produits en dizaines de millions d'exemplaires, la question du choix de la technologie ASIC ne se posait même pas. De leur côté, les TPU de Google sont des ASIC tandis que les processeurs BrainWave de **Microsoft**, qu'il utilise dans ses propres datacenters, sont en FPGA **Intel** en 10 nm, des Predix 10, issus du rachat d'**Altera** fin 2015. Ils en ont optimisé l'architecture, notamment au niveau de la mémoire, pour qu'ils ne soient pas pénalisés vis à vis des ASIC.

En termes de puissance, les chipsets de Nvidia et Google sont bien plus puissants que le NPU du Kirin 970 : celui de Google fait 92 TFlops/s et le Nvidia atteint 120 TFLOPS, c'est-à-dire, 60 fois la puissance du Kirin 970 ! Les Nvidia sont produits en technologie 12 nm et ont 21,5 milliards de transistors, 4 fois plus que le total du Kirin 970 qui en fait 5,5. Mais ne comparons pas des choux et des carottes : le Kirin 970 est fait pour des smartphones et de la basse consommation tandis que les TPU de Google et le V100 de Nvidia sont conçus pour des serveurs et consomment plus de 100W.

On n'a aucune information sur les autres parties du NPU du Kirin 970. Mais elles devraient logiquement également comprendre :

- Une **matrice de synapses** reliant deux vecteurs de neurones pour gérer la dernière étape de traitement des réseaux convolutionnels (dite "fully connected"). Ca parallélise bien les traitements de ces réseaux de neurones. La taille de cette matrice est importante car elle indiquera en gros le nombre d'objets différents qui peuvent être reconnus par le système. En gros, cela ressemble conceptuellement à cela…

- Des **zones de traitements** pour d'autres étapes des réseaux convolutionnels, comme pour le pooling qui gère la réduction de résolution d'images d'étapes intermédiaires. Mais on peut éventuellement utiliser les multiplicateurs de matrice existant.

- De la **mémoire rapide** pour que les paramètres du réseau de neurones soient rapidement accédés et éventuellement modifiés par le multiplicateur de matrices et la matrice de synapses.

Du côté du logiciel, le Kirin 970 supporte les applications de deep learning réalisées avec la bibliothèque open source **TensorFlow** originaire de Google et sa déclinaison pour Android mobile, TensorFlow Lite, ainsi qu'avec **Caffe**. Ce sont deux des outils les plus courants de développement d'applications d'IA couvrant le machine learning et toutes les formes de réseaux de neurones de deep learning.

Le NPU du Kirin 970 occuperait une surface équivalente à un seul cœur du Kirin 970 sachant que celui-ci en comprend 8 de traditionnels, des cœurs ARM, plus 12 cœurs de GPU ARM MALI G72. Ces cœurs de CPU sont précisément quatre cœurs 64 bits A73 tournant à 2,40 GHz et quatre autres, A53 tournant à 1,80 GHz dans ce qu'ARM appelle l'architecture BIGlittle associant des cœurs puissants et rapides et des cœurs moins rapides et moins puissants, et moins consommateurs d'énergie, permettant d'augmenter l'autonomie des smartphones.

L'un des intérêts de ce NPU n'est pas seulement la vitesse de reconnaissance d'images, mais aussi sa faible consommation. Elle sera utile pour les applications de reconnaissance de la parole. la startup française **Snips** propose déjà une solution de reconnaissance de la parole embarquée dans un smartphone, en exploitant des chipsets mobiles standards. Elle pourra éventuellement exploiter les capacités de ce processeur et de ses équivalents pour améliorer la performance de la reconnaissance de la parole en local, et surtout, en diminuant la consommation d'énergie associée. Le fait de ne pas avoir à faire un aller et retour avec un serveur est en tout cas un avantage indéniable pour la reconnaissance de la parole.

Ce Kirin 970 a d'autres fonctionnalités qui n'ont rien à voir avec l'IA, qui nous intéressent moins ici :

- Un modem "double data" qui permet d'utiliser la connexion cellulaire data sur deux cartes SIM en même temps
- Un modem LTE Cat18 supportant 1,2 GBits/s, un débit auquel vous n'aurez jamais accès en pratique, comme pour votre fibre. Le premier à supporter le Cat18 semble être Qualcomm avec son modem X20 annoncé début 2017.
- La compression de vidéo 4K, en 2160p60 pour le décodage et en 2160p30 pour l'encodage, semble-t-il en HDR (haute dynamique) histoire d'être en phase avec les TV qui supportent le HDR depuis quelques temps.
- Son processeur d'image (Imaging DSP) qui n'a rien à voir avec le NPU permettra de mieux gérer la prise de photos, notamment en mouvement et en basse lumière.

Patient numéro un

Nous avons donc là un produit hybride qui intègre un design de chipset neuromorphique couplé au design d'un chipset classique de smarphones.

Il est fascinant d'observer la sauce marketing de l'intelligence artificielle ! Vu de près, ce processeur magique faisant de l'IA est nouveau parce qu'il contient des multiplicateurs de matrices de 3×3 valeurs. Ce qui n'a rien d'intelligent et est très basique d'un point de vue conceptuel. L'intelligence est située dans l'organisation des réseaux de neurones convolutionnels qui exploitent ces multiplicateurs de matrice. L'IA est dans le logiciel ! Mais on pourra mieux juger sur pièces lorsque les spécifications détaillées de ce NPU seront publiques. Au pire, après la sortie des smartphones l'intégrant et après que les geeks de **Chipworks** aient analysé leur

processeur au microscope.

Heureusement, quelques exemples concrets d'usages étaient indiqués lors du lancement : de la réalité augmentée à faible consommation, de la compréhension du langage aussi bien dans les images que dans la voix, de la vision artificielle et une amélioration des automatismes de prise de photos en fonction des caractéristiques des objets ou personnes photographiés. Et la peinture en temps réel d'une photo qui consiste à appliquer un style de dessin au pinceau sur une image, un des usages des réseaux de neurones convolutionnels à double entrée, l'image de départ et le style à lui appliquer, ce qui est tout à fait secondaire. Le tout sans véritables démonstration. Il faudra au minimum attendre l'annonce en octobre des smartphone Mate 10 intégrant le Kirin 970 pour appréhender cette IA mobile.

Qui sera le suivant ? Ce processeur va être le patient numéro 1 d'une longue série auquel ses successeurs et copies seront comparés. On imagine la manière dont Apple pourrait intégrer cette innovation révolutionnaire dans ses iPhone 8, 9 ou 10 selon leur rapidité d'action. Il est probable qu'au moment de ce lancement, on aura oublié les balbutiements de ce pionnier qu'est Huaweï.

Patient numéro deux

Petite mise à jour du 12 septembre 2017 : finalement, Apple a intégré un "Neural Engine" dans ses nouveaux iPhone 8 et iPhone X dont l'existence avait été dévoilée en mai 2017. Il s'agit du "A11 Bionic Neural Engine", une fonctionnalité qui est intégrée dans le nouveau chipset des iPhone 8 et X, "A11 Bionic". Il sert notamment à SIRI ainsi qu'au login par détection 3D du visage avec la version miniaturisée du capteur 3D de PrimeSense qui est intégrée dans l'iPhone X. Il permet aussi de gérer des emoji qui captent les mouvements du visage de l'utilisateur. Il semble que le Neural Engine soit aussi utilisé dans les fonctions photo avancées de ces iPhones, comme le mode portrait qui modifie l'éclairage du visage pour simuler un éclairage de studio.

L'A11 Bionic Neural Engine complète le nouveau GPU de ces A11 qui remplace les Power VR de l'Anglais Imagination Technology. Apple a indiqué que les capacités de ce "Neural Engine" étaient de 900 millions d'opérations par seconde qui représentent visiblement la moitié de la capacité de 1,92 TFlops du Kirin 970 de Huawei. Apple précisait aussi que c'était un composant à deux cœurs, ce qui ne veut pas dire grand chose pour un composant neuromorphique. Est-ce à dire qu'il ne comporte que deux multiplicateurs de matrices (ou tenseurs), ce qui serait léger ou pas en fonction de la taille des matrices ?

A ce stade, aucune information n'a filtré sur les caractéristiques précises de ce Neural Engine. Il faudra attendre pour savoir ce qu'il a dans le ventre et les interfaces qu'il propose aux développeurs et le support éventuel des nombreux SDK type TensorFlow, Theano et Caffe. Il est probable qu'il supporte les **APIs BNNS** de développement de réseaux de neurones. Ces APIs supportent la création de réseaux convolutionnels avec des

couches de convolution, de pooling et des couches de neurones “fully connected” permettant notamment de reconnaître des images ou des phonèmes pour la reconnaissance vocale. Il y a aussi les APIs **Core ML**, annoncées lors de la dernière conférence développeurs Apple.

On peut par contre s’attendre à de belles innovations côté applications car avec tous ces capteurs et processeurs, ces nouvelles générations de smartphones ont un beau potentiel à révéler, notamment dans l’univers assez large du bien être et de la santé. Qui plus est, sans nécessiter d’envoyer ses données personnelles dans le cloud. Ce n’est que le début d’une nouvelle ère applicative pour la mobilité !

Cet article a été publié le 7 septembre 2017 et édité en PDF le 17 mars 2024.