Distillation de connaissances de CNN dans une infrastructure de Edge Computing

Cédric MARON[1,2], Virginie FRESSE[1], Karynn MORAND[2]

[1] Laboratoire Hubert Curien, 18 rue Professeur Benoît Lauras Bâtiment F, 42000 Saint-Etienne, France
[2] SEGULA Technologie, 1 Rue des Combats du 24 Août 1944, 69200 Vénissieux, France

Résumé – Avec l'intérêt grandissant pour la compression de réseaux de neurones, plusieurs techniques permettant d'obtenir des réseaux compacts et efficients ont vu le jour. L'une d'entre elles, la distillation de connaissances, permet de transférer les connaissances d'un réseau (enseignant) vers un réseau (élève) lors de la phase d'apprentissage de ce dernier. La distillation de connaissances est généralement réalisée dans le Cloud car les architectures de réseaux enseignants sont souvent trop lourdes pour être hébergées sur des ressources de Edge. Ainsi, la distillation est souvent impossible dans un contexte purement Edge. Dans cet article, nous proposons une nouvelle méthode de distillation adaptée à une infrastructure de Edge Computing utilisant des architectures de réseaux élève et enseignants de tailles réduites et équivalentes. Les expérimentations réalisées montrent une augmentation de la précision du réseau élève de 0,8% sur la base de données CIFAR10 par rapport à un entrainement sans distillation.

Abstract – With the growing interest in neural network compression, several techniques generating efficient networks have emerged. One of them, knowledge distillation, aims to transfer knowledge from a network (teacher) to a network (student) during its training phase. Knowledge distillation is generally carried out in the Cloud because teacher network architectures are often too heavy to be hosted on Edge computing resources, making distillation impossible in a pure Edge context. In this paper, we propose a new distillation method adapted to an edge computing infrastructure using student and teacher network architectures of reduced and equivalent sizes. The experiments show an increase in the accuracy of the student network of 0.8% on the CIFAR10 database compared to training without distillation.

1 Introduction

Il est largement admis que la précision d'un réseau de neurones est étroitement liée à sa taille et à sa complexité architecturale. Cette observation s'explique par le fait que les réseaux de neurones de tailles importantes sont capables de modéliser des relations plus complexes entre les données d'entrée et les sorties attendues.

De manière générale, la précision, la consommation mémoire et le nombre d'opérations de calcul par seconde sont liés. Ainsi un réseau ayant une grande précision aura une consommation mémoire et/ou une vitesse d'inférence supérieure à un réseau ayant une précision moindre.

Dans un contexte de Edge Computing, les ressources de calcul sont souvent multiples mais généralement limitées en termes de mémoire et de puissance de calcul. Ainsi le choix de l'architecture de réseau de neurones doit être réalisé en fonction des ressources disponibles dans cette infrastructure. Des méthodes de compression de réseaux sont utilisées afin d'obtenir des réseaux toujours plus efficients. Parmi ces méthodes, la distillation de connaissances permet de transférer les connaissances d'un réseau enseignant vers un réseau élève ce qui permet d'accroitre la précision du réseau élève par rapport à un entrainement sans distillation.

Dans un contexte de distillation classique, l'architecture du réseau élève est choisie en fonction des contraintes matérielles comme la puissance de calcul et la mémoire disponible. Le réseau enseignant quant à lui est choisi uniquement dans le but d'obtenir la meilleure précision possible. De manière générale plus la précision du réseau enseignant est élevée, plus la distillation de connaissances sera efficace sur le réseau élève sous réserve que le réseau élève ait une capacité architecturale permettant d'imiter le réseau enseignant [1].

Le réseau enseignant nécessite souvent trop de puissance de calcul et de mémoire pour pouvoir être exécuté sur les ressources présentes dans le Edge. Ainsi, afin de réaliser la distillation de connaissances, il est en général nécessaire de faire appel à des services de Cloud Computing mettant en œuvre des GPUs haut de gamme.

Dans cet article, nous proposons une méthode permettant de réaliser l'entièreté de la distillation dans une infrastructure de Edge. La contribution présentée est une méthode de distillation de connaissances multi-enseignants à partir de réseaux enseignants mono-classe ayant des architectures restreintes et de tailles similaires à l'architecture du réseau élève. Nos expérimentations montrent un gain en précision pour le réseau élève de 0,8% sur la base de données CIFAR10 par rapport à un entrainement classique sans distillation.

2 Distillation

2.1 Méthodes classiques de distillation

Il existe plusieurs méthodes de distillation de connaissances [2]. La distillation de connaissances basée sur les logits (prédictions non normalisées) de sorties consiste à entraîner un réseau élève ayant une architecture restreinte à générer des logits similaires à

un modèle enseignant plus complexe et ayant une grande précision. Cette méthode est souvent utilisée dans le cadre de la classification d'images.

La distillation de connaissances basée sur les cartes de patterns générées par des couches intermédiaires d'un réseau consiste à entrainer un réseau élève à extraire des cartes de motifs similaires à un modèle enseignant. Cette méthode est souvent utilisée dans le cadre de la segmentation d'image.

La distillation de connaissances basée sur les gradients est une méthode permettant d'améliorer la robustesse des deux méthodes précédentes. En effet, les gradients générés lors de l'entrainement d'un réseau permettent de savoir quelles parties du réseau sont les plus actives. Par conséquent, il est possible d'utiliser cette information pour faire en sorte qu'un réseau élève puisse reproduire le fonctionnement des parties les plus actives du réseau enseignant.

Il est également possible d'appliquer ces différentes méthodes en utilisant plusieurs réseaux enseignants. Cela permet d'avoir un transfert de connaissances plus stable, étant donné que les connaissances de différents réseaux enseignants sont agrégées.

2.2 Schémas de distillation

Il existe trois schéma de distillation :

- La distillation offline consiste à distiller les connaissances d'un réseau enseignant pré-entrainé vers un réseau élève.
- La distillation online consiste à entrainer conjointement un réseau élève et un réseau enseignant tout en réalisant la distillation en parallèle.
- La self-distillation consiste à réaliser de la distillation entre les couches intermédiaires d'un même réseau.

Le schéma de distillation utilisé dans cet article est le schéma offline car c'est le schéma le plus répandu et le plus simple à implémenter.

2.3 Distillation dans une infrastructure de Edge

Les méthodes de distillation offlines présentées dans la littérature se focalisent majoritairement sur l'utilisation de réseaux enseignants de tailles supérieures au réseau élève comme le montre le Tab 1. Ainsi notre méthode se différencie par le fait qu'elle consiste en l'utilisation de réseaux enseignants et élève de tailles similaires.

Tab 1 : Présentation du ratio du nombre de paramètres entre réseaux élèves et réseaux enseignants pour différentes techniques de distillation offline.

Methode	Enseignant	Elève	Ratio Nb Params
CTKD [3]	WRN-40-1 WRN-40-2	WRN-16-1 WRN-16-2	3.29 3.19
TOFD [4]	ResNet152 ResNet152	ResNeXt50-4 MobileNetV2	2.40 17.17
AdaIN [5]	ResNet26 WRN-40-2	ResNet8 WRN-16-2	4.63 3.19
FN [6]	ResNet110 ResNet56	ResNet56 ResNet20	2.0 3.15
Notre	LeNet	LeNet	1.0

2.4 Méthode proposée

La méthode proposée est une méthode de distillation offline multi-enseignant basée uniquement sur les logits de sortie. Elle se place dans un contexte où la distillation est réalisée uniquement avec les ressources de calculs disponibles dans le Edge comme le montre la Figure 1 et où l'utilisation de réseaux enseignants de tailles supérieures au réseau élève est impossible.

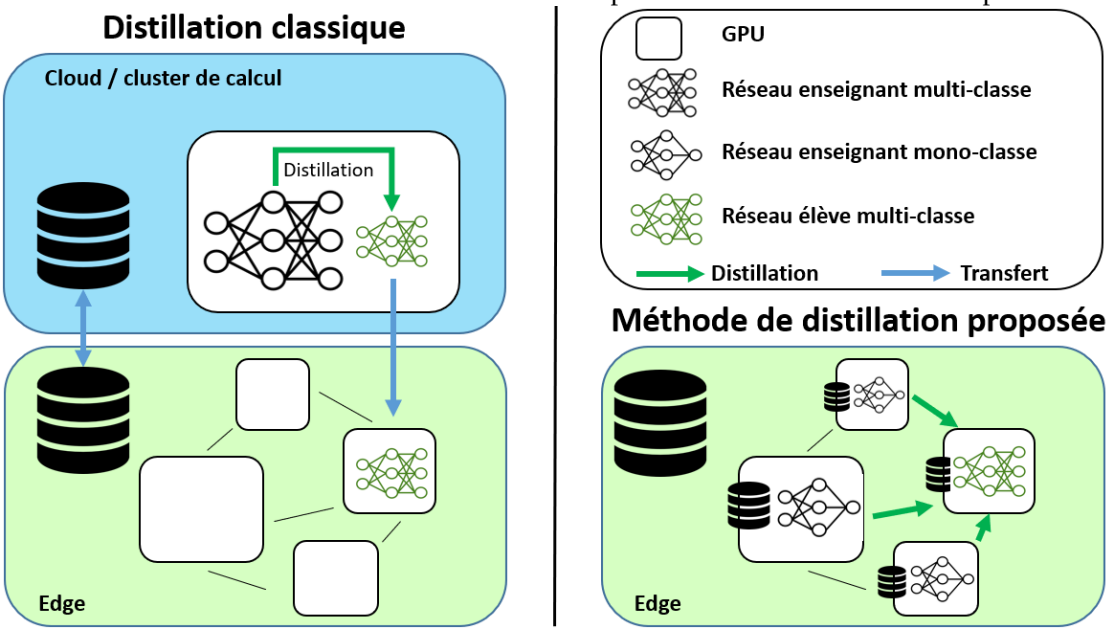

Figure 1 : Schéma de contextualisation de notre approche. A gauche, distillation classique utilisant un réseau enseignant multi-classe de taille importante et des ressources externes vs, à droite, notre approche utilisant les ressources présentes dans le Edge et des réseaux enseignants mono-classe pour réaliser de la distillation de connaissances

L'objectif de la méthode proposée est de déterminer si la combinaison de plusieurs réseaux enseignants mono-classes peut permettre d'améliorer l'apprentissage d'un réseau élève tout en ayant des architectures de réseaux élève et enseignants de tailles restreintes et similaires.

Dans un premier temps, les réseaux enseignants mono-classes sont entrainés à classifier sur une unique classe. Chaque réseau est réparti sur les différentes ressources de calcul disponibles afin de réaliser l'entrainement en parallèle. Une fois entrainés, les logits de sorties de chaque réseau enseignant sont stockés pour chaque image de la base de données d'entrainement.

Dans un second temps, le réseau élève est entrainé en utilisant à la fois la vérité terrain de la base de données et les sorties agrégées des différents réseaux enseignants mono-classe en suivant la méthode d'agrégation présentée Figure 2.

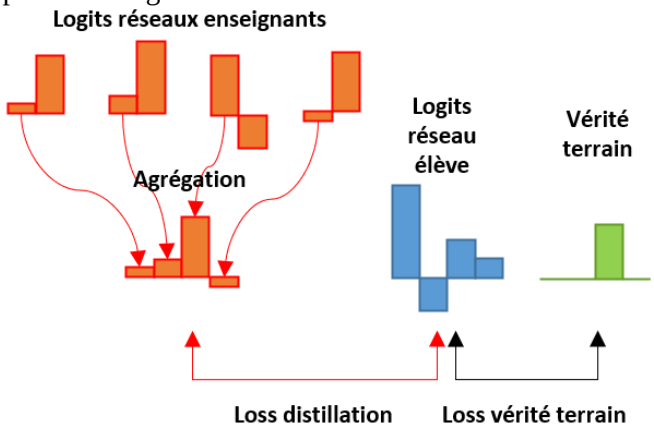

Figure 2 : Agrégation des logits de classe principale des différents réseaux enseignants mono-classe puis calcul de la loss de distillation en parallèle du calcul de la loss vérité terrain

La méthode d'agrégation proposée consiste en une agrégation des logits de classe principale des n réseaux enseignants afin de recréer un vecteur de n logits correspondant à chaque classe de la base de données. Cette combinaison de n logits est ensuite utilisée pour distiller les connaissances acquises par les réseaux enseignants en calculant la loss (fonction de coût) de distillation entre les logits des enseignants et les logits de l'élève. Dans notre cas, la fonction de coût de distillation est MSE (Mean Squared Error) et la fonction de coût vérité terrain est CrossEntropy. La fonction de coût de distillation est mise à zéro pour les cas où le logit maximal des n réseaux enseignants ne correspond pas à la vérité terrain.

3 Expérimentation et résultats

Les tests sont réalisés pour évaluer le gain en précision que permet la méthode proposée sur le réseau élève. Pour ce faire, deux comparaisons sont menées. La première est faite par rapport à une distillation classique avec un réseau enseignant multi-classe. La seconde comparaison est réalisée par rapport au réseau élève sans distillation car la méthode se place dans un contexte où l'utilisation de réseaux enseignants de tailles supérieures au réseau élève n'est pas possible.

3.1 Cadre de l'expérimentation

Les entrainements des réseaux enseignants multi-classe/mono-classe et du réseau élève sont réalisés avec une taille de batch de 96 images sur 100 epochs avec la base de données CIFAR10 qui comporte des images appartenant à 10 classes différentes réparties en 50 000 images d'entrainement et 10 000 images de test. La méthode d'optimisation Stochastic Gradient Descent est utilisée avec un learning rate de 0.001 et un momentum de 0.9. Les tests sont réalisés sur un ordinateur doté d'un processeur i7-9700 CPU, 32Go de RAM et une carte graphique Nvidia Quadro P5000 16Go.

Les architectures utilisées lors des tests sont des architectures personnalisées de type LeNet. L'architecture du réseau élève (51 880 paramètres) et des réseaux enseignants mono-classes (51 752 paramètres) sont identiques à l'exception de la sortie des réseaux enseignants mono-classe qui comporte 2 sorties au lieu de 10. Les 2 classes en sortie correspondent soit à la classe à prédire soit à « autre ». L'architecture du réseau multi-classe a une profondeur identique au réseau élève mais est cependant beaucoup plus large (4 187 018 paramètres).

Lors des expérimentations, les réseaux enseignants sont d'abord entrainés puis distillés sur le réseau élève. Les trois fonctions de coût MSE (Mean Square Error), MAE (Mean Averaged Error) et CE (Cross Entropy) sont également comparées afin d'évaluer laquelle permet d'obtenir le meilleur transfert de connaissances. Seule une pondération 1 :1 avec la fonction de coût vérité terrain (CE) a été évaluée. Les expérimentations présentées sont le résultat d'une unique série de tests.

3.2 Distillation classique

D'après les expérimentations présentées dans le Tab 2, les fonctions de coût MSE (distillation) et CE (vérité terrain) donnent les meilleurs résultats. L'évolution de la précision sur la base de données de test est présentée dans la Figure 3. La distillation avec le réseau enseignant multi-classes permet d'augmenter la précision du réseau élève de 2,3% par rapport à un entrainement sans distillation (68,0% à 70,3%).

Tab 2 : Distillation du réseau multi-classe sur le réseau élève avec différentes fonctions de coût

		Loss distillation			Elève sans distillation
		CE	MAE	MSE	
Loss VT	NONE	66,59	67,88	69,89	
	CE	65,81	70,04	70,34	67,98

3.3 Méthode de distillation proposée

Les expérimentations présentées dans le Tab 3 montrent que les fonctions de coût MSE (distillation) et CE (vérité terrain) donnent les meilleurs résultats. Les

tests réalisés permettent d'obtenir un gain de précision de 0,8% par rapport à l'entrainement du réseau élève sans distillation (68,0% à 68,8%). L'évolution de la précision sur la base de données de test est présentée Figure 3. Il est notable que la distillation des réseaux mono-classe sur le réseau élève ne génère aucun overfitting. En effet, l'augmentation de la précision est continue et le maximum de précision est atteint à la fin de l'entrainement.

Tab 3 : Distillation des réseaux mono-classe sur le réseau élève avec différentes fonctions de coût

		Loss distillation			Elève sans distillation
		CE	MAE	MSE	
Loss VT	NONE	58,1	58,25	57,94	
	CE	66,7	68,57	68,76	67,98

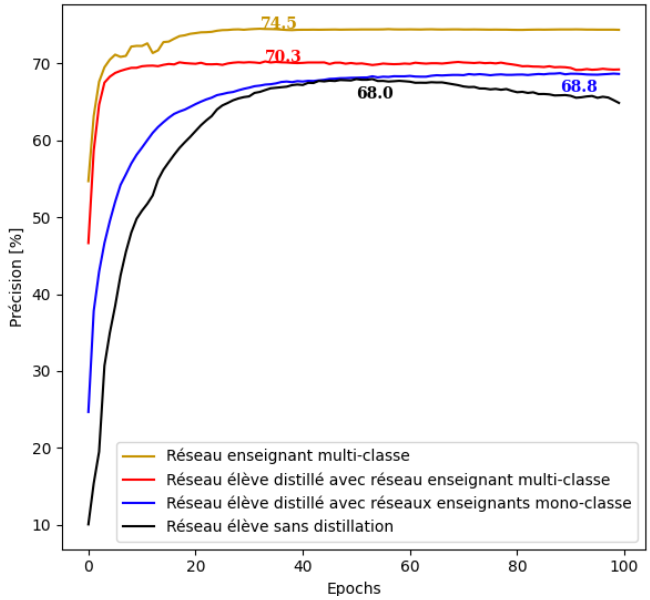

Figure 3 : Evolution de la précision lors de l'entrainement du réseau élève, du réseau enseignant multi-classe, du réseau élève distillé avec le réseau enseignant multi-classe et du réseau élève distillé avec les réseaux mono-classes sur CIFAR10

3.4 Analyse des résultats

D'après les expérimentations, la combinaison des fonctions de coûts MSE (distillation) et CE (vérité terrain) semble donner les meilleurs résultats à la fois pour la distillation classique et pour la méthode proposée.

La méthode proposée utilisant des réseaux enseignants mono-classe permet d'obtenir un gain de précision inférieur à une distillation classique (0,8% vs 2,3%) mais a l'avantage d'être adaptée à un contexte purement Edge Computing et ne souffre pas d'overfitting. En effet, un léger phénomène d'overfitting est observable pour le réseau élève sans distillation et le réseau élève distillé avec le réseau enseignant multi-classe ce qui est contre intuitif mais qui peut être dû à la topologie très compacte du réseau élève ou bien aux hyper paramètres d'entrainement.

4 Conclusion

Dans cet article, nous proposons une méthode de distillation offline multi-enseignants pouvant être utilisée dans un contexte de Edge Computing. La méthode proposée est une alternative aux méthodes de distillation classiques qui utilisent des réseaux enseignants de tailles largement supérieures au réseau élève et nécessitent des ressources externes pour pouvoir être exécutées. Cette méthode se base sur l'utilisation de plusieurs réseaux enseignants mono-classes ayant des architectures similaires au réseau élève. Les tests réalisés sur la base de données CIFAR10 permettent d'obtenir un gain de précision de 0,8% par rapport à un entrainement classique sans distillation. Notre méthode permet d'effectuer de la distillation de connaissance offline dans le Edge sans l'utilisation de ressources de calcul externes performantes. Nos perspectives se portent sur une réduction des bases de données nécessaires à l'entrainement des réseaux mono-classe afin répondre aux faibles capacités de stockage disponibles dans le Edge.

Références

[1] J. H. Cho and B. Hariharan, "On the Efficacy of Knowledge Distillation," 2019, pp. 4794–4802. Accessed: Jun. 22, 2023. [Online]. Available: https://openaccess.thecvf.com/content_ICCV_2019/html/Cho_On_the_Efficacy_of_Knowledge_Distillation_ICCV_2019_paper.html

[2] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge Distillation: A Survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Mar. 2021, doi: 10.1007/s11263-021-01453-z.

[3] H. Zhao, X. Sun, J. Dong, C. Chen, and Z. Dong, "Highlight Every Step: Knowledge Distillation via Collaborative Teaching." arXiv, Jul. 22, 2019. doi: 10.48550/arXiv.1907.09643.

[4] L. Zhang, Y. Shi, Z. Shi, K. Ma, and C. Bao, "Task-Oriented Feature Distillation," in *Advances in Neural Information Processing Systems*, 2020, vol. 33, pp. 14759–14771. Accessed: Apr. 03, 2023. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/a96b65a721e561e1e3de768ac819ffbb-Abstract.html

[5] J. Yang, B. Martinez, A. Bulat, and G. Tzimiropoulos, "Knowledge distillation via adaptive instance normalization." arXiv, Mar. 09, 2020. doi: 10.48550/arXiv.2003.04289.

[6] K. Xu, L. Rui, Y. Li, and L. Gu, "Feature Normalized Knowledge Distillation for Image Classification," in *Computer Vision – ECCV 2020*, Cham, 2020, pp. 664–680. doi: 10.1007/978-3-030-58595-2_40.