STAGE M2

Analyse de performance d'un réseau de neurones profond compressé

Contacts des encadrants :

- Rodrigo Cabral Farias, Maître de Conférences, Univ. Côte d'Azur (cabral@i3s.unice.fr)
- Lionel Fillatre, Professeur des Universités, Univ. Côte d'Azur (lionel.fillatre@i3s.unice.fr)

Laboratoire d'accueil :
- I3S (Sophia Antipolis - France)

Durée prevue :
- 4 à 5 mois

Gratification :
- Environ 550 €/mois

Contexte du stage :
Les réseaux de neurones profonds sont devenus un élément incontournable de l'état de l'art pour diverses problématiques d'inférence complexe en traitement de données telles que la détection, la classification et la segmentation d'objets dans les images et vidéos. La complexité croissante de ces réseaux rend difficile leur implantation sur un système embarqué dans un contexte temps-réel. Par conséquent, la réduction de leur complexité en termes d'empreinte mémoire et de complexité de calcul est actuellement un sujet d'intense investigation de plusieurs équipes de recherche.

Pour réduire leur empreinte mémoire, les paramètres d'un réseau profond doivent être compressés. Différentes techniques, telles que l'élagage des poids du réseau [1], la quantification [2] ou une combinaison des deux [3], ont été appliquées. Avec la méthode proposée en [3], il a été montré de manière expérimentale qu'une forte réduction de l'empreinte mémoire peut être obtenue avec une très faible perte des performances d'inférence.

Des membres de l'équipe Signal, Images et Systèmes (SIS) du laboratoires I3S s'intéressent à la compréhension théorique des effets de la compression sur les performances d'inférence d'un réseau profond, notamment, à donner une prédiction de la perte de performance en fonction du taux de compression des paramètres. Dans un cadre de classification binaire et en se focalisant sur la compression par la quantification des paramètres de la dernière couche du réseau, un travail récent de l'équipe [4] donne une approximation de la perte de justesse de classification introduite par la compression. Cette approximation est donnée en fonction des paramètres de la couche, des caractéristiques du problème de classification sous-jacent et du nombre de bits de quantification utilisé pour la compression.

Objectifs :
L'approximation obtenue en [4] n'est valable que sous certaines hypothèses de travail, notamment sur les distributions des entrées de la dernière couche du réseau et sur la distribution des erreurs de quantification des paramètres. Le but premier de ce stage est de réaliser un certain nombre d'expériences pour vérifier ces hypothèses dans un cadre pratique, *i.e.* lorsque le réseau étudié est un réseau profond utilisé en pratique (ex. : ResNet [5]) et lorsque les données du problème de classifications sont réelles (ex. : données CIFAR [6] ou ImageNet [7]). Ces expériences numériques seront réalisées en langage python et nécessiteront l'utilisation de librairies dédiées à l'apprentissage profond (pytorch [8] ou tensorflow [9]).

Selon l'avancement du stagiaire, différentes pistes théoriques pourraient ensuite être explorées : adaptation des hypothèses de travail dans le cas où elles ne sont pas exactement vérifiées en pratique, extension de l'étude [4] à la compression de plusieurs couches du réseau, ou encore, extension de [4] à la classification multi-classes.

Compétences requises :

- Formation en traitement statistique du signal ou en statistiques.
- Maîtrise du langage python.
- Connaissance des réseaux de neurones profonds et des librairies python dédiées (pytorch et/ou tensorflow).
- Ecriture de rapports scientifiques avec LaTex.

References:

[1] S. Anwar, K. Hwang, and W. Sung, “Structured pruning of deep convolutional neural networks,” JETC, vol. 13,no. 3, pp. 32:1–32:18, 2017.

[2] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. G.Howard, H. Adam, and D. Kalenichenko, “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” IEEE CVPR, pp. 2704–2713, 2018.

[3] S. Han, H. Mao, and W. J. Dally, “Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding,” 4th ICLR, Y. Bengio and Y. LeCun, Eds., 2016.

[4] D. Resmerita, R. Cabral Farias, B. D. de Dinechin, L. Fillatre, “Distortion Approximation of a Compressed Softmax Layer,” IEEE SSP, pp. 491-495, 2021.

[5] K. He, X. Zhang, S. Ren, J. Sun, “Deep residual learning for image recognition,” IEEE CVPR, pp. 770-778, 2016.

[6] A. Krizhevsky, G. Hinton. “Learning multiple layers of features from tiny images,“ 2009. https://www.cs.toronto.edu/~kriz/cifar.html

[7] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” IEEE CVPR, pp. 248-255, 2009.

[8] https://pytorch.org/

[9] https://www.tensorflow.org