Research Internship
Compression and Robustness of artificial neural networks

9 décembre 2019

1 Context

Artificial neural networks, particularly in the context of deep learning [1], are the state of the art in most automatic machine learning problems. Because they contain a very large number of driveable parameters, they are able to capture important statistical characteristics to meet the given tasks. On the other hand, this dependence on a very large number of parameters has two harmful effects : a) networks are expensive to train and implement [2], 2) they tend to capture information specific to the training data, making them less robust to changes in data acquisition processes.

2 Subject

The objective of this Research project is to focus on compression techniques that have been introduced in the scientific literature in recent years, with the aim of quantifying their impact on the robustness of embedded driven networks. Concretely, techniques of quantification[3], pruning[4], grouping[5] or distillation[6] have shown that it is possible to considerably reduce the size of architectures while maintaining the level of performance. We will therefore first try to evaluate the impact of compression techniques on the robustness of architectures as part of this internship. We will be interested in the use of these methods for both machine learning and neuroscience neural networks[7]. Then, we will focus on proposing compression techniques that increase robustness. Finally, we will evaluate the cost of implementing these different techniques on FPGA hardware circuits. The project will rely on the automatic generators of HW architectures for CNN and SNN, available at the laboratory[8].

3 Internship

The internship will take place in the eBRAIN research group of LEAT laboratory (University Cote d'Azur). The LEAT is located in Sophia-Antipolis, near Nice, France. The duration is 6 months and the salary is 529 euros/month. Profile : machine learning, artificial neural networks, embedded systems, python

The intership will be supervised by Vincent Gripon (IMT Atlantique) and Pr. Benoît Miramond (University Cote d'Azur).

Proposition d'un stage de recherche
Robustesse des architectures profondes compressées

4 Contexte

Les réseaux de neurones artificiels, notamment dans le cadre de l'apprentissage profond [1], sont l'état de l'art dans la plupart des problèmes de l'apprentissage automatique. Du fait qu'ils contiennent un très grand nombre de paramètres entraînables, ils parviennent à capturer des caractéristiques statistiques importantes pour répondre aux tâches données. D'un autre côté, cette dépendance à un très grand nombre de paramètres a deux effets délétères : a) les réseaux sont coûteux à entraîner et à implémenter [2], 2) ils ont tendance à capturer des informations spécifiques aux données d'entraînement, les rendant peu robustes à des changements dans les processus d'acquisition des données [9].

5 Sujet de stage

L'objectif de ce stage de Master est de s'intéresser aux techniques de compression ayant été introduites dans la littérature scientifique ces dernières années, en ayant pour objectif de quantifier leur conséquence sur la robustesse des réseaux entraînés embarqués. Concrètement, des techniques de quantification [3], d'élagage [4], de regroupement [5] ou encore de distillation [6] ont montré qu'il était possible de considérablement réduire la taille des architectures tout en maintenant le niveau de performance. Nous chercherons donc dans un premier temps dans le cadre de ce stage à évaluer l'impact des techniques de compression sur la robustesse des architectures. Nous nous intéresserons à l'utilisation de ces méthodes la fois pour des réseaux de neurones issues du machine learning et des neurosciences [7]. Ensuite, nous nous attacherons à proposer des techniques de compression ayant pour effet d'augmenter la robustesse. Enfin nous évalurons le coût d'implémentation de ces différentes techniques sur circuit matériel FPGA. Le stagiaire s'appuiera pour cela sur les générateurs automatiques d'architectures CNN et SNN, disponibles au laboratoire [8].

6 Environnement

Ce stage de Master se déroulera au sein du groupe eBRAIN du laboratoire LEAT du mois de mars à aout. Une rémunération est prévue pour une durée maximale de 6 mois pour une gratification de 529 euros/mois. Il sera co-encadré par Vincent Gripon, chargé de recherche à l'IMT Atlantique et chercheur invité au LEAT et Benoît Miramond, professeur des universités au LEAT.

Références

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] G. B. Hacene, C. Lassance, V. Gripon, M. Courbariaux, and Y. Bengio, "Attention based pruning for shift networks," in *Arxiv Preprint*, 2019.

[3] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect : Training deep neural networks with binary weights during propagations," in *Advances in neural information processing systems*, 2015, pp. 3123–3131.

[4] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in neural information processing systems*, 2015, pp. 1135–1143.

[5] Y. Gong, L. Liu, M. Yang, and L. Bourdev, "Compressing deep convolutional networks using vector quantization," *arXiv preprint arXiv :1412.6115*, 2014.

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv :1503.02531*, 2015.

[7] L. Khacef, N. Abderrahmane, and B. Miramond, "Confronting machine-learning with neuroscience for neuromorphic architectures design," in *International Joint Conference on Neural Networks (IJCNN)*, 2018.

[8] N. Abderrahmane, E. Lemaire, and B. Miramond, "Design space exploration of hardware spiking neurons for embedded artificial intelligence," *Elsevier Journal on Neural Networks*, pp. 366–386, 2019.

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv :1312.6199*, 2013.