Compression guidée par les données et le matériel cible

Contexte

Les réseaux de neurones profonds ont permis des avancées impressionnantes dans de nombreux domaines de l'intelligence artificielle tels que la vision par ordinateur, la reconnaissance de parole ou encore le traitement naturel des langues. Ce gain de performances s'est toutefois fait au prix d'une complexité accrue des calculs et donc des ressources requises en inférence (énergie, mémoire et puissance de calcul). Cette contrainte limite ainsi leur déploiement sur des dispositifs embarqués (périphériques mobiles, micro-contrôleurs...). Cette thèse s'inscrit dans le domaine émergent et très actif de la compression de réseaux de neurones, et vise à proposer des stratégies prenant en compte les spécificités du dispositif cible au moment de la conception et de l'entraînement du réseau.

Objectifs de la thèse

Il existe de très nombreuses approches pour compresser un réseau telle que, par exemple, la quantification (les opérations en arithmétique à virgule flottante sur 32 bits sont remplacées par des opérations à virgule fixe sur un plus faible nombre de bits), par élagage (des opérations sont retirées du graphe de calcul) ou par distillation (la connaissance d'un réseau est transférée dans un réseau de plus petite taille). L'équipe de recherche de Datakalab s'est principalement intéressée aux approches dites sans données (data-free) qui utilisent uniquement l'information contenue dans le réseau [1,2,3,4]. Le premier axe de recherche de la thèse sera de concevoir des méthodes de compression orientée données, afin d'améliorer la qualité de la compression (ie., soit le taux de compression, soit la précision du modèle compressé) pour les cas d'utilisation qui le permettent.

Par ailleurs, nombre des méthodes de compression sont agnostiques du matériel cible qui sera utilisé pour l'inférence. Certains micro-contrôleurs auront par exemples des limitations fortes en termes de mémoire et d'opérations supportées, alors que d'autres gammes de processeurs auront une prise en charge limitée pour des réseaux quantifiés sur un faible nombre de bits. Certains travaux, de type NAS (Network Architecture Search), visent à trouver l'architecture optimale d'un réseau de neurones parmi un ensemble d'architectures possibles. L'espace de recherche est souvent représenté par un méta-réseau pré-appris, appelé supernet, à partir duquel il est possible d'échantillonner des réseaux avec des caractéristiques variées. La recherche de l'optimum peut alors se faire par des méthodes de type génétiques ou par renforcement, au prix toutefois d'un apprentissage long et fastidieux. Dans la cadre de la thèse nous nous intéresserons plus particulièrement aux méthodes par descente de gradient qui transforment un problème d'optimisation combinatoire en une problème d'optimisation continu et dérivable (eg. [5,6]).

La relaxation continue du problème d'adaptation au support matériel peut se faire par bien des manières qui devront être traités lors de la thèse [7]. En particulier, lors d'opérations séquentielles, déterminer le coût maximal en mémoire supporté par la machine cible et s'assurer de tenir dans cette contrainte par des outils d'élagage et de quantification afin d'atteindre le meilleur compromis entre expressivité et vitesse d'inférence réelle. Ces outils seront donc optimisés opération par opération (mixed precision) [8,9]. Un autre exemple de compression hardware aware, repose sur le fait que le temps d'inférence des réseaux de neurones, selon le support physique, dépend de la nécessité de récupérer des résultats précédemment calculé (skip connections). Le travail de recherche portera alors sur la minimisation du recours à ces sous-architectures tout en préservant la précision des réseaux de neurones.

Profil et compétences recherchées

Diplôme de Master ou Grande École. Compétences requises :

- Machine Learning / Deep Learning
- Vision par ordinateur et/ou Traitement automatique du Langage Naturel (NLP)
- Programmation Python et librairie deep learning (tensorflow ou pytorch)
- Excellentes capacités relationnelles et rédactionnelles (français et anglais)

Modalités de candidature

Pour postuler à cette thèse, le candidat est invité à communiquer par mail à kb@datakalab.com, ad@datakalab.com et lf@datakalab.com :

— Le CV

— Les résultats académiques des deux dernières années universitaires

— Un lien vers un projet de machine learning (lien GitHub / GitLab ou Colab)

Environnement

La thèse se déroulera dans le cadre d'un contrat de collaboration CIFRE entre la société Datakalab et l'Institut des Systèmes Intelligents et de Robotique (ISIR) de Sorbonne Université.

Datakalab est une startup spécialisée dans des algorithmes d'apprentissage profond à faible consommation, efficaces en termes d'exécution, respectueux de la vie privée et fonctionnant entièrement en embarqué. Ses travaux de recherches ont données lieux à des publications dans les meilleures conférences et journaux du domaine (T-PAMI, NeurIPS, ICCV, CVPR, AAAI)

L'**ISIR** est une unité mixte de recherche sous la tutelle de Sorbonne Université, du Centre National de la Recherche Scientifique (CNRS), et de l'Inserm, dont la recherche pluridisciplinaire rassemble des chercheuses, chercheurs, enseignantes-chercheuses et enseignants-chercheurs relevant de différentes disciplines en robotique, apprentissage machine, sciences du vivant et sciences médicales.

La thèse sera encadrée par Kévin Bailly directeur de la recherche de Datakalab et Maître de conférences, HDR, à l'ISIR et Arnaud Dapogny chercheur en IA à Datakalab.

Références

[1] *RED : Looking for Redundancies for Data-FreeStructured Compression of Deep Neural Networks*, 2021, NeurIPS, Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[2] *RED++ : Data-Free Pruning of Deep Neural Networks via Input Splitting and Output Merging*, 2022, TPAMI, Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[3] *To Fold or Not to Fold : a Necessary and Sufficient Condition on Batch-Normalization Layers Folding*, 2022, IJCAI, Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[4] *REx : Data-Free Residual Quantization Error Expansion*, arXiv preprint arXiv :2203.14645, E Yvinec Edouard, Dapogny Arnaud, Cord Matthieu and Bailly, Kévin

[5] *DARTS : Differentiable Architecture Search*, 2019, ICLR, Hanxiao Liu, Karen Simonyan, Yiming Yang

[6] *Learning Time/Memory-Efficient Deep Architectures With Budgeted Super Networks*, 2018, CVPR, Tom Véniat, Ludovic Denoyer

[7] ProxylessNAS : Direct Neural Architecture Search on Target Task and Hardware, ICLR 2019 Han Cai, Ligeng Zhu, Song Han

[8] *Mixed Precision Quantization of ConvNets via Differentiable Neural Architecture Search*, arXiv preprint arXiv :1812.00090. Bichen Wu, Yanghan Wang, Peizhao Zhang, Yuandong Tian, Peter Vajda, Kurt Keutzer

[9] HAQ : Hardware-Aware Automated Quantization With Mixed Precision, CVPR 2019 Kuan Wang, Zhijian Liu, Yujun Lin, Ji Lin, Song Han ;