Compression des réseaux de neurones adaptée aux cibles matérielles spécifiques

contact
114 Boulevard Malesherbes, 75017, Paris

kb@datakalab.com

ad@datakalab.com

ey@datakalab.com

lf@datakalab.com

Mots clés: Budgeted Networks, Pruning, Quantization, Computer Vision

Environnement

Datakalab est une startup basée à Paris (17ème arrondissement) 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 recherche ont donné lieu à des publications dans les meilleures conférences et journaux du domaine (T-PAMI, NeurIPS, ICCV, CVPR, AAAI)

Le stage sera encadré par Kévin Bailly directeur de la recherche de Datakalab et Maître de conférences, HDR, à l'ISIR, Edouard Yvinec et Arnaud Dapogny, chercheurs en IA à Datakalab.

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…). Ce stage 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.

Il existe de très nombreuses approches pour compresser un réseau telles 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].

Objectifs du stage

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 du stage nous nous intéresserons plus particulièrement aux méthodes par descente de gradient qui transforment un problème d'optimisation combinatoire en un problème d'optimisation continu et dérivable (eg. [5,6]) s'appuyant sur des scores empiriques de coût de calcul. De manière similaire, l'équipe avait développé une technique basée sur la mesure de l'importance des neurones [10]

La relaxation continue du problème d'adaptation au support matériel peut se faire par bien des manières qui devront être traités lors du stage [7]. En particulier, lors d'opérations séquentielles, utiliser le coût maximal en mémoire supporté par la machine cible et s'assurer de tenir dans des

temps raisonnables 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].

Concrètement, le travail consistera incorporer les mesures empiriques des coûts d'inférence obtenues par l'équipe hardware de Datakalab dans les méthodes développées pour le pruning et/ou la quantification.

Profil et compétences recherchées

Etudiant de Master ou Grande École. Compétences requises :

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

Modalités de candidature

Pour postuler à ce stage, le candidat est invité à communiquer par mail (cf. liste des contacts associés à cette fiche de stage) :

- Son CV
- Ses résultats académiques des deux dernières années universitaires
- (optionnel) Un lien vers un des ces projets en machine learning (lien GitHub / GitLab ou Colab)

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

[10] Yvinec, Edouard and Dapogny, Arnaud and Cord, Matthieu and Bailly, Kévin SInGE: Sparsity via Integrated Gradients Estimation of Neuron Relevance, NeurIPS 2022