Offre de Stage

Compression de réseaux Transformers pour la vision

Mots-clefs : Apprentissage profond, Compression, Quantization, Pruning, Transformers

contact
114 Boulevard Malesherbes, 75017, Paris

ey@datakalab.com

ad@datakalab.com

kb@datakalab.com

lf@datakalab.com

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

Dans la lignée des avancées spectaculaires dans le domaine du traitement naturel du langage (NLP), les modèles Transformers pour la vision ont donné d'excellents résultats pour de nombreuses tâches telles que la reconnaissance d'objets [1,2], la segmentation sémantique [3], la détection d'objets [4] ou la super-résolution [5]. La taille mémoire et le temps d'inférence de tels modèles en limitent toutefois fortement leur utilisation et leur déploiement à large échelle, en particulier sur des systèmes embarqués. De façon similaire aux réseaux standards utilisés jusqu'ici en vision par ordinateur (ConvNets), il existe des techniques de compression afin de permettre un déploiement moins coûteux. Parmi ces techniques, le pruning [6] et la quantification [9]. Cependant, les transformers apportent leur lot de difficultés spécifiques lorsque l'on souhaite en quantifier les valeurs. Certains articles récents, dont I-Bert [11], proposent des solutions pour la quantification des fonctions d'activation (GeLU), du softmax et des couches de layerNorm. Cependant, ces techniques restent encore insuffisantes en particulier pour les transformers les plus volumineux [12].

Objectifs du stage

L'objectif du stage sera alors d'explorer différentes stratégies de compression des réseaux Transformer avec éventuellement une mise en commun avec HuggingFace. Dans un premier temps le candidat dressera un état de l'art sur les méthodes de distillation, d'élagage et quantification appliquées aux architectures Transformer et mettra en oeuvre les stratégies les plus prometteuses. Dans un second temps et sur la base des résultats obtenus, le candidat pourra proposer des pistes d'amélioration. L'entraînement des Transformer étant trop coûteux et pas adapté aux contraintes temporelles d'un stage, on envisagera de manière privilégiée des méthodes de type post-training. Dans cette optique, le stage pourra s'appuyer sur l'expérience de l'équipe d'encadrement dans la compression de réseaux profonds par élagage [6,7] et quantification [9,10] sans données.

Profil et compétences recherchées

Edudiant 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

Références

[1] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words : Transformers for image recognition at scale. In International Conference on Learning Representations.

[2] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR.

[3] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer : Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems, 34 :12077–12090.

[4] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer.

[5] Yang, F., Yang, H., Fu, J., Lu, H., and Guo, B. (2020). Learning texture transformer network for image super-resolution. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5790–5799. IEEE Computer Society.

[6] Yvinec, E., Dapogny, A., Bailly, K., and Cord, M. (2022b). Red++ : Data-free pruning of deep neural networks via input splitting and output merging. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[7] Yvinec, E., Dapogny, A., Cord, M., and Bailly, K. (2021). Red : Looking for redundancies for data-freestructured compression of deep neural networks. Advances in Neural Information Processing Systems, 34.

[8] Yvinec, E., Dapogny, A., Cord, M., and Bailly, K. (2022c). Singe : Sparsity via integrated gradients estimation of neuron relevance. Advances in Neural Information Processing Systems.

[9] Yvinec, E., Dapogny, A., Cord, M., and Bailly, K. (2022a). Rex : Data-free residual quantization error expansion. arXiv preprint arXiv :2203.14645.

[10] Yvinec, E., Dapogny, A., Cord, M., and Bailly, K. (2022a). SPIQ : data-free per-channel static input quantization. WACV.

[11] I-bert: Integer-only bert quantization. ICML. Kim, Sehoon and Gholami, Amir and Yao, Zhewei and Mahoney, Michael W and Keutzer, Kurt. 2021.

[12] ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers. Yao, Zhewei and Aminabadi, Reza Yazdani and Zhang, Minjia and Wu, Xiaoxia and Li, Conglong and He, Yuxiong. arXiv preprint arXiv:2206.01861. 2022.s