Compression des Transformeurs pour la vision par ordinateur et le traitement automatique du langage naturel

Contexte

L'intelligence artificielle est devenue indispensable grâce à sa contribution dans la résolution de plusieurs problématiques de la vie réelle, en particulier depuis l'émergence des réseaux de neurones artificiels profonds. Parmi les réseaux de neurones les plus récents et performant, on trouve les Transformeurs [1]. Ils ont été proposés pour des tâches de traitement automatique de langage naturel telles que la traduction [1], le transfert de texte [2], le résumé automatique [3], ainsi que la compréhension de texte [4]. Ils ont été adoptés par la suite pour résoudre des problématiques de la vision par ordinateur.

Parmi ces modèles, les plus populaires sont BERT (Bidirectional Encoder Representations from Transformers) [4], GPT (Generative Pre-trained Transformer) [5], RoBERTa (Robustly Optimized BERT Pre-training) [6] et T5 (Text-to-Text Transfer Transformer) [2]. L'impact profond des modèles Transformeurs est devenu plus clair avec leur évolutivité vers des modèles de très large nombre de paramètres. Par exemple, le modèle BERT-large, avec 340 millions de paramètres, a été largement dépassé par le modèle GPT-3 qui contient 175 milliards de paramètres, tandis que le dernier Transformeur Switch [7] a atteint 1,6 trillion de paramètres. Malheureusement, malgré le grand succès des Transformeurs, ils ne trouvent actuellement que peu d'applications dans les systèmes embarqués. Leur taille monstrueuse implique un temps d'inférence important, et une occupation mémoire impossible pour de tels systèmes.

Pour autant, la compression est devenue un domaine indispensable afin de limiter la taille des réseaux de neurones volumineux tels que les Transformeurs et les réseaux convolutifs profonds. La compression permet non seulement la réduction de l'empreinte mémoire des modèles, mais aussi une inférence efficiente et une optimisation des ressources de calcul. De plus, elle permet de réduire l'empreinte carbone et le coût environnemental des modèles lors de leur apprentissage.

Etat de l'art et enjeux

Il existe actuellement plusieurs travaux qui ciblent la compression des Transformeurs et utilisent des méthodes connues de compression telles que :

- Weights sharing [8, 9, 10, 11, 12] : repose sur l'hypothèse que les réseaux à grande échelle sont sur-paramétrés [8]. Il fournit un moyen de découpler les calculs et les paramètres en réutilisant les mêmes paramètres pour de multiples calculs afin de réduire l'empreinte mémoire lors de l'inférence et le nombre de paramètres. Un exemple de ce type de compression est de partager des couches du réseaux de neurone, voire l'encodeur et le décodeur dans son ensemble.
- Low-Rank Factorization [13, 14, 15] : Les matrices de poids d'un réseau neuronal sont souvent de faible rang, ce qui indique une redondance dans les poids du modèle. Ainsi, une idée naturelle est de factoriser les matrices de poids en deux ou plusieurs matrices plus petites pour économiser le nombre de paramètres. Une technique courante de factorisation à faible rang est la décomposition en valeurs singulières (SVD).
- Pruning : [16, 17, 18] sert à supprimer les poids non importants dans un réseau neuronal afin afin de réduire le nombre de paramètres du réseau tout en préservant les performances du modèle.
- Quantization : [19, 20, 21] vise à compresser un réseau neuronal en réduisant le nombre de bits (la précision) dans les poids du modèle.
- Knowledge Distillation : [22, 23, 24] permet de transférer les connaissances d'un gros modèle vers un modèle de plus petite taille en utilisant une fonction de perte adaptée.

Quelques chercheurs ont aussi essayé de combiner plusieurs méthodes de compression [25, 26]. Cependant, plus d'effort est nécessaire afin d'obtenir un niveau de compression satisfaisant en termes de performance et nombre de paramètres.

Objectifs de la thèse

Le but de la thèse est donc de proposer des méthodes de compression de réseaux de neurones de type Transformeur pour la vision par ordinateur et/ou le traitement du langage naturel. Cela implique : (1) une étude détaillée de l'état de

l'art actuel, (2) une implémentation des approches les plus récentes afin d'en établir une base solide d'expérimentations, et (3) l'exploration des autres verrous et à la proposition de nouvelles idées par le candidat pour favoriser son autonomie nécessaire en tant que futur chercheur. Ces travaux devront donner lieu à une ou plusieurs publications dans des journaux internationaux.

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] Vaswani, Ashish, et al. *"Attention is all you need."* Advances in neural information processing systems 30 (2017).
[2] Raffel, Colin, et al. *"Exploring the limits of transfer learning with a unified text-to-text transformer."* J. Mach. Learn. Res. 21.140 (2020) : 1-67.
[3] Zhang, Jingqing, et al. *"Pegasus : Pre-training with extracted gap-sentences for abstractive summarization."* International Conference on Machine Learning. PMLR, 2020.
[4] Devlin, Jacob, et al. *"Bert : Pre-training of deep bidirectional transformers for language understanding."*. NAACL-HLT (1) 2019 : 4171-4186
[5] Brown, Tom, et al. *"Language models are few-shot learners."* Advances in neural information processing systems 33 (2020) : 1877-1901.
[6] Liu, Yinhan, et al. *"Roberta : A robustly optimized bert pretraining approach."* arXiv preprint arXiv :1907.11692 (2019).
[7] Fedus, William, Barret Zoph, and Noam Shazeer. *"Switch transformers : Scaling to trillion parameter models with simple and efficient sparsity."* (2021).
[8] Li, Zhuohan, et al. "Train big, then compress : Rethinking model size for efficient training and inference of transformers." International Conference on Machine Learning. PMLR, 2020.

[9] Xia, Yingce, et al. *"Tied transformers : Neural machine translation with shared encoder and decoder."* Proceedings of the AAAI conference on artificial intelligence. Vol. 33. No. 01. 2019.
[10] Rothe, Sascha, Shashi Narayan, and Aliaksei Severyn. *"Leveraging pre-trained checkpoints for sequence generation tasks."* Transactions of the Association for Computational Linguistics 8 (2020) : 264-280.
[11] Dabre, Raj, and Atsushi Fujita. *"Recurrent stacking of layers for compact neural machine translation models."* Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.
[12] Dehghani, Mostafa, et al. *"Universal transformers."* arXiv preprint arXiv :1807.03819 (2018).
[13] Sainath, Tara N., et al. *"Low-rank matrix factorization for deep neural network training with high-dimensional output targets."* 2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013.
[14] Grachev, Artem M., Dmitry I. Ignatov, and Andrey V. Savchenko. *"Neural networks compression for language modeling."* International Conference on Pattern Recognition and Machine Intelligence. Springer, Cham, 2017.
[15] Ma, Xindian, et al. *"A tensorized transformer for language modeling."* Advances in neural information processing systems 32 (2019).
[16] LeCun, Yann, John Denker, and Sara Solla. *"Optimal brain damage."* Advances in neural information processing systems 2 (1989).
[17] Wang, Wenhui, et al. *"Minilmv2 : Multi-head self-attention relation distillation for compressing pretrained transformers."* arXiv preprint arXiv :2012.15828 (2020).
[18] Wang, Yong, et al. *"On the sparsity of neural machine translation models."* arXiv preprint arXiv :2010.02646 (2020).
[19] Kim, Sehoon, et al. *"I-bert : Integer-only bert quantization."* International conference on machine learning. PMLR, 2021.
[20] Bengio, Yoshua, Nicholas Léonard, and Aaron Courville. *"Estimating or propagating gradients through stochastic neurons for conditional computation."* arXiv preprint arXiv :1308.3432 (2013).
[21] De Lathauwer, Lieven, Bart De Moor, and Joos Vandewalle. *"A multilinear singular value decomposition."* SIAM journal on Matrix Analysis and Applications 21.4 (2000) : 1253-1278.
[22] Wang, Wenhui, et al. *"Minilmv2 : Multi-head self-attention relation distillation for compressing pretrained transformers."* arXiv preprint arXiv :2012.15828 (2020).
[23] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. *"Distilling the knowledge in a neural network."* arXiv preprint arXiv :1503.02531 2.7 (2015).
[24] Sanh, Victor, et al. *"DistilBERT, a distilled version of BERT : smaller, faster, cheaper and lighter."* arXiv preprint arXiv :1910.01108 (2019).
[25] Kim, Young Jin, and Hany Hassan Awadalla. *"Fastformers : Highly efficient transformer models for natural language understanding."* arXiv preprint arXiv :2010.13382 (2020).
[26] Sanh, Victor, Thomas Wolf, and Alexander Rush. *"Movement pruning : Adaptive sparsity by fine-tuning."* Advances in Neural Information Processing Systems 33 (2020) : 20378-20389.