Offre de Thèse
Vers une compression non biaisée des modèles de Deep Learning

Julien Velcin et Guillaume Metzler

Laboratoire ERIC, Université de Lyon, Université Lumière Lyon 2
julien.velcin@univ-lyon2.fr; guillaume.metzler@univ-lyon2.fr

Contexte

Cette thèse s'inscrit dans le cadre du projet ANR DIKé portant sur la **Compression des modèles de Deep Learning** dans le cadre d'application au *Traitement du Langage Naturel* (NLP). La compression des modèles de *Deep Learning* revêt un enjeu majeur notamment d'un point de vue écologique. En effet, ces modèles présentent très souvent un très grand nombre de paramètres (de l'ordre du milliard de paramètres pour le modèle de langage *GPT-3*) rendant leur temps d'apprentissage extrêmement long et coûteux à la fois en terme d'architecture mais aussi d'un point de vue énergétique (Neill, 2020).

Les travaux actuels, sur le modèle *BERT* par exemple (Sanh et al., 2019) montrent qu'il est possible de réduire drastiquement le nombre de paramètres des modèles de *Deep Learning* sans pour autant dégrader leurs performances. Malheureusement, cette compression n'est pas sans conséquence d'un point de vue *éthique* ou en terme d'*équité* et a tendance à introduire un **biais** lors de la compression des modèles (Hooker et al., 2020). Par exemple, un modèle de classification tendra à favoriser la classe majoritaire au détriment des individus/objets issus de minorités. Il pourra également avoir tendance à accentuer des biais de genre que l'on pourrait retrouver dans certaines phrases faisant références aux professions des personnes.

La distillation n'est qu'un cas particulier (et certainement la plus utilisée) des méthodes de compression. On compte aussi des techniques de compression basées sur la *quantization* ou encore sur l'*élagage*. Ces dernières continuent de faire l'objet de recherches intensives tant leur nécessité, sur le plan pratique, devient crucial avec la taille des modèles actuels (Gupta and Agrawal, 2022; Hu et al., 2021).

En revanche, l'étude des problèmes liés à la compression des modèles de Deep Learning est un sujet très récent dont les travaux ne sont pas très présents dans la littérature (Xu and Hu, 2022; Stoychev and Gunes, 2022).

L'objectif de ce projet est d'être capable d'identifier si un modèle compressé est *biaisé* à travers la création de mesures permettant d'évaluer ce biais. Il s'agira ensuite d'identifier les origines de ce biais dans la compression et d'établir de nouvelles méthodes permettant de compresser des modèles *deep* complexes sans que ces derniers ne favorisent une classe d'objets/individus plutôt qu'une autre par exemple.

Objectif

Le/la futur(e) thésard(e) devra dans un premier temps réaliser un travail bibliographique pour approfondir (si besoin) ses connaissances dans le domaine du *NLP* et notamment maîtriser l'architecture des réseaux couramment utilisés dans un tel domaine. (Vaswani et al., 2017; Devlin et al., 2018; Sanh et al., 2019) qui pourront lui servir étant donné le contexte de la thèse.
Il devra ensuite étudier les différentes techniques de compression des modèles de Deep Learning, comme le **pruning** ou encore la **distillation** (Cheng et al., 2017; Neill, 2020) ainsi que les origines du biais lors de la compression de tels modèles (Hooker et al., 2019, 2020; Hooker, 2021). Cette première phase est essentielle, elle permettra de comprendre quels sont les éléments important du réseau qui permettent d'expliquer l'apparition du biais. Il sera également important d'identifier des mesures ainsi que des techniques qui vont permettre de comprendre quels sont les mots sur lesquels se focalisent le modèle dans la tâche prédicitive.

Après ce travail préliminaire, le ou la candidat(e) recruté(e) pourra tout d'abord fournir un travail bibliographique concernant l'évaluation des biais dans les modèles de Machine Learning. Il conviendra également de déterminer les origines du biais dans les dits modèles. Pour cela, il/elle pourra étudier les différentes métriques ainsi que les méthodes de compression de modèle de Deep Learning utilisé en *NLP* sur des benchmarks trouvés lors du travail bibliographique comme celui présenté dans le papier de Nadeem et al. (2020)[1] ou encore de Zhao et al. (2018) [2].
Plusieurs pistes sont proposées afin de traiter le problème en fonction des études préliminaires conduites sur les données mais aussi sur les modèles, plusieurs pistes de recherches sont proposées dans le cadre de la thèse et qui permettront de répondre à la problématique du sujet.

- L'obtention de résultats biaisés peut très souvent s'expliquer par la présence d'un déséquilibre au niveau des classes en présence. Le modèle aura donc tendance à privilégier la classe majoritaire lors de la phase d'apprentissage. Ce phénomène est accentué lorsque l'on travaille sur des modèles compressés.
  Pour palier à cela, on peut donc décider de travailler (i) à l'échelle des données ou (ii) sur le modèle en lui même.

  (i) Une méthode classique consiste à augmenter le nombre d'exemples de la classe minoritaire pour que le modèle puisse se focaliser sur ces données, on parle de d'oversampling ou de data augmentation. Si ces techniques sont très connues dans les branches classiques, elles sont cependant peu utilisées en NLP (Feng et al., 2021; Shorten et al., 2021) et laissent un champ libre à explorer.

  (ii) Il est également possible de contraindre le modèle à se focaliser sur de telles exemples par le biais de l'apprentissage d'une représentation adéquate et l'usage d'une loss pondérée qui tient compte du déséquilibre entre les classes.

- On pourra également utiliser des approches basées sur l'équité (*fairness*) (Hardt et al., 2016) afin de contraindre les modèles compressés (ou à compresser) à avoir des performances similaires sur les différentes classes.

La résolution de la problématique pourra ensuite se faire en deux temps :

[1] le jeu de données est disponible à l'adresse suivante : https://stereoset.mit.edu

[2] le jeu de données ainsi que le code et le papier sont disponibles à l'adresse suivante : https://paperswithcode.com/dataset/winobias

1. on se concentre tout d'abord sur les modèles compressés et on cherche à réduire le biais de ces modèles via les méthodes précédemment cités. Il s'agit ici de débiaiser un modèle **après** sa compression.

2. si l'on parvient à débiaiser les modèles compressés, on pourra alors chercher à développer des techniques de compression qui n'introduisent pas de biais.

Informations pratiques

Profil et durée La thèse s'effectuera sur une période de trois ans à compter de Septembre 2022. L'étudiant recruté sera membre du projet ANR DIKé qui financera intégralement cette thèse.
Le futur candidat sera issu d'une filière informatique ou mathématiques appliquées avec une coloration prononcée pour le Machine Learning et plus précisément le Deep Learning. De bonnes connaissances dans le domaine du *NLP* et des différents modèles fondés sur le *Transformer* seront un plus, de même qu'une appétence pour la programmation en Python (Tensorflow ou PyTorch de préférence, avec une préférence pour ce dernier).

Lieu et encadrement L'encadrement sera effectué par Julien Velcin (Professeur en Informatique) et Guillaume Metzler (MCF en Informatique) et la thèse s'effectuera au Laboratoire ERIC de l'Université Lyon 2 qui se trouve sur le campus de Bron (5 Avenue Pierre Mendès France).

Candidature : il est demandé d'envoyer votre **CV**, vos relevés de **notes de Master**, une **lettre de motivation** et **au moins une lettre de recommandation** : une première de votre responsable de formation et si possible de votre encadrant si vous effectuez un stage de recherche. Ces informations seront à envoyer aux adresses suivantes ,avant le 15 juin

julien.velcin@univ-lyon2.fr et **guillaume.metzler@univ-lyon2.fr**
Les candidats seront notifiés peu de temps après pour un éventuel entretien et la réponse finale sera donnée fin juin/début juillet.

Rémunération : le/la candidat(e) sera rémunéré(e) environ 1 975 euros brut par mois.

Autres : cette thèse s'inscrivant dans le cadre d'un projet ANR impliquant entités (Laboratoire Hubert Curien de Saint-Etienne et Naverlabs à Grenoble), la personne recrutée sera vivement encouragée à travailler en étroite collaboration avec les différents membres du projet.

References

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Manish Gupta and Puneet Agrawal. Compression of deep learning models for text: A survey. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(4):1–55, 2022.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Sara Hooker. Moving beyond "algorithmic bias is a data problem". *Patterns*, 2(4):100241, 2021.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

Sara Hooker, Nyalleng Moorosi, Gregory Clark, Samy Bengio, and Emily Denton. Characterising bias in compressed models. *arXiv preprint arXiv:2010.03058*, 2020.

Peng Hu, Xi Peng, Hongyuan Zhu, Mohamed M Sabry Aly, and Jie Lin. Opq: Compressing deep neural networks with one-shot pruning-quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7780–7788, 2021.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*, 2020.

James O' Neill. An overview of neural network compression. *arXiv preprint arXiv:2006.03669*, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34, 2021.

Samuil Stoychev and Hatice Gunes. The effect of model compression on fairness in facial expression recognition. *arXiv preprint arXiv:2201.01709*, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

Guangxuan Xu and Qingyuan Hu. Can model compression improve nlp fairness. *arXiv preprint arXiv:2201.08542*, 2022.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.