Internship Offer for M2 Students
Toward an Ethical Compression of Deep Learning Models

Guillaume Metzler et Julien Velcin

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

Context

This internship is part of the ANR DIKé project on the **compression of Deep Learning** models in the context of Natural Language Processing (NLP). The compression of Deep Learning models is a major issue, especially from an ecological point of view. Indeed, they often have a very large number of parameters (the order of a billion parameters for the GPT-3 language model) making their learning time extremely long and costly both in terms of architecture and energy (Neill, 2020). Current works, on the BERT model for instance (Sanh et al., 2019) shows that it is possible to drastically reduce the number of parameters of Deep Learning models without degrading their performance. Unfortunately, this compression is not without consequences from an ethical or fairness point of view and tends to introduce a **bias** (Hooker et al., 2020). For example, a classification model will tend to favor the majority class at the expense of minority individuals/objects. The goal of this project is to be able to identify whether a compressed model is biased through the creation of measures to assess this bias. Then, we want to identify the origins of this bias in the compression and to establish new methods to compress deep complex models without favouring one class of objects/individuals over another for example.

Purpose

The selected student will first have to carry out a bibliographical work to extend his knowledge (if needed) in the field of NLP (Vaswani et al., 2017; Devlin et al., 2018; Sanh et al., 2019) which will be useful given the context of this internship. He will then have to study the different compression techniques of Deep Learning models, such as **pruning** or **distillation** (Cheng et al., 2017; Neill, 2020) as well as the origins of bias when compressing such models (Hooker et al., 2019, 2020; Hooker, 2021). After having mine it the litterature, and if time allows, the recruited candidate will be able to participate in the proposal of an unbiased approach to compressing Deep Learning models used in NLP on benchmarks found during the bibliographic work such as the one presented in the paper by Nadeem et al. (2020) or by Zhao et al. (2018). Different approaches can be considered, such as fairness-based approaches (Hardt et al., 2016) or through methods used in an unbalanced context.

The work done during this internship should serve as a foundation for the thesis proposed later in the ANR project.

Information

Duration and Candidate This is at least a 4 month internship (ideally 6 months
The duration of the internship will ideally be 6 months (at least 4 months) starting in March. The future candidate will come from a computer science or applied mathematics background with a strong emphasis on Machine Learning. Good knowledge in the field of NLP will be a plus as well as an appetence for programming in Python (Tensorflow - Pytorch for example).

Place and Supervision The supervision will be done by Julien Velcin (Professor in Computer Science) and Guillaume Metzler (MCF in Computer Science) and the internship will take place at the ERIC Laboratory of the University Lyon 2 which is located on the campus of Bron (5 Avenue Pierre Mendès France).

Candidature Send your application: CV and grades to the following addresses

julien.velcin@univ-lyon2.fr et **guillaume.metzler@univ-lyon2.fr**

Rémunération : the candidate will be paid 3.90 euros/hour or approximately 578 euros/month.

Perspectives This offer, if the recruited candidate is motivated and shows a great interest for the subject by carrying out a quality internship, could lead to a thesis on the same subject and financed by the ANR.

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.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

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.

Offre de Stage M2
Vers une compression éthique des modèles de Deep Learning

Guillaume Metzler et Julien Velcin

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

Contexte

Ce stage 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.

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

L'étudiant en stage devra dans un premier temps réaliser un travail bibliographique pour approfondir ses connaissances (si besoin) dans le domaine du *NLP* (Vaswani et al., 2017; Devlin et al., 2018; Sanh et al., 2019) qui pourront lui servir étant donné le contexte de ce stage.
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).
Après ce travail bibliographique, et si le temps le permet, le ou la candidat(e) recruté(e) pourra participer à la proposition d'une approche non biaisée 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 envisageables pour faire cela, comme des approches basées sur l'équité (*fairness*) (Hardt et al., 2016) ou par le biais de méthodes utilisées dans un contexte déséquilibré.

Le travail effectué pendant ce stage doit servir de bases pour la thèse proposée par la suite dans le cadre du projet ANR.

Informations pratiques

Profil et durée La durée de stage sera idéalement de 6 mois (au minimum 4 mois) à partir du mois de mars. Le futur candidat sera issu d'une filière informatique ou mathématiques appliquées avec une coloration prononcée pour le Machine Learning. De bonnes connaissances dans le domaine du *NLP* seront un plus de même qu'une appétence pour la programmation en Python (Tensorflow - Pytorch par exemple).

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

Candidature Envoyer votre candidature : **CV, relevés de notes** aux adresses suivantes

julien.velcin@univ-lyon2.fr et **guillaume.metzler@univ-lyon2.fr**

Rémunération : le/la candidat(e) sera rémunéré(e) **3.90 euros/heure** soit environ **578 euros/mois**.

Perspectives Cette offre, dans le cas où le candidat recruté est motivé et montre un grand intérêt pour le sujet en effectuant de stage de qualité, pourra déboucher sur une thèse traitant du même sujet et financée par l'ANR.

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.

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.

[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

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

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.

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.