

PhD Position

Towards an unbiased compression of Deep Learning Models

Julien Velcin et Guillaume Metzler

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

Context

This thesis 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, these models often have a very large number of parameters (of the order of a billion parameters for the language model GPT-3) making their learning time extremely long and costly both in terms of architecture and from an energy point of view.

Current work, on the *BERT* model for example (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 point of view or in terms of fairness and tends to introduce a bias when compressing models. For example, a classification model will tend to favor the majority class at the expense of minority individuals/objects. It may also tend to accentuate gender biases that might be found in certain sentences referring to people's professions.

Distillation is only one particular case (and certainly the most widely used) of compression methods. There are also compression techniques based on quantization or even freezing. The latter continue to be the subject of intensive research as their necessity, on a practical level, becomes crucial with the size of the current models (Gupta and Agrawal, 2022; Hu et al., 2021).

On the other hand, the study of problems related to the compression of Deep Learning models is a very recent subject whose works are not very present in the literature (Xu and Hu, 2022; Stoychev and Gunes, 2022).

The objective of this project is to be able to identify if a compressed model is biased by creating measures to evaluate this bias. It will then be a question of identifying the origins of this bias in the compression and of establishing new methods allowing complex models to be compressed without the latter favouring one class of objects/individuals over another for example.

Purpose

The future PhD student will have to first realize a bibliographical work to deepen (if need be) his knowledge in the field of the *NLP* and in particular to master the architecture of the networks

commonly used in such a field (Vaswani et al., 2017; Devlin et al., 2018; Sanh et al., 2019) which will be able to be useful to him given the context of the thesis.

He will then have to study the different techniques of compression of Deep Learning models, such as the **pruning** or the **distillation** (Cheng et al., 2017; Neill, 2020) as well as the origins of the bias during the compression of such models (Hooker et al., 2019, 2020; Hooker, 2021). This first phase is essential, it will allow us to understand which are the important elements of the network that explain the appearance of the bias. It will also be important to identify measures as well as techniques that will allow us to understand which words the model focuses on in the predictive task. After this preliminary work, the recruited candidate will first provide a bibliographic work concerning the evaluation of biases in Machine Learning models. It will also be necessary to determine the origins of the bias in the said models. For that, he/she will be able to study the various metrics as well as the compression methods of Deep Learning model used in *NLP* on benchmarks found during the bibliographical work like the one presented in the paper of Nadeem et al. (2020)¹ or de Zhao et al. (2018)².

Several tracks are proposed in order to treat the problem according to the preliminary studies conducted on the data but also on the models, several research tracks are proposed within the framework of the thesis and which will make it possible to answer the problematic of the subject.

- The obtaining of biased results can very often be explained by the presence of an imbalance in the classes present. The model will therefore tend to favour the majority class during the learning phase. This phenomenon is accentuated when we work on compressed models. To overcome this, we can therefore decide to work (i) on the scale of the data or (ii) on the model itself.
 - (i) A classical method consists in increasing the number of examples of the minority class so that the model can focus on these data, we speak of oversampling or data augmentation. If these techniques are very well known in the classical branches, they are however little used in NLP citepfeng2021survey,shorten2021text and leave an open field to be explored.
 - (ii) It is also possible to force the model to focus on such examples by learning a suitable representation and using a weighted loss that takes into account the imbalance between classes.

fairness) (Hardt et al., 2016) approaches can also be used to force compressed (or to be compressed) models to have similar performances on the different classes.

The resolution of the problem can then be done in two steps:

1. we concentrate first of all on the compressed models and we try to reduce the bias of these models via the previously mentioned methods. The aim here is to debias a model after its compression.
2. if we manage to debias the compressed models, we can then try to develop compression techniques which do not introduce bias.

¹the dataset is available at the following address <https://stereoset.mit.edu>

²the dataset as well as the code and paper are available at: <https://paperswithcode.com/dataset/winobias>

Information

Profile and duration .The thesis will be carried out over a period of three years starting in September 2022. The recruited student will be a member of the ANR DIKé project which will fully finance this thesis. The future candidate will come from a computer science or applied mathematics background with a strong knowledge on Machine Learning and more specifically Deep Learning. Good knowledge in the field of NLP and various models based on the Transformer will be a plus, as well as an appetite for Python programming (Tensorflow or PyTorch preferably, with a preference for the latter).

Location and supervision The supervision will be carried out by Julien Velcin (Professor in Computer Science) and Guillaume Metzler (MCF in Computer Science) and the thesis will be carried out at the ERIC Laboratory of the University Lyon 2 which is located on the campus of Bron (5 Avenue Pierre Mendès France).

It is requested to send your **CV**, your **Master grades**, a **letter of motivation** and **at least one letter of recommendation**: a first one from your training supervisor and if possible from your supervisor if you are doing a research training. This information should be sent to the following addresses before **June the 15th**

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

Candidates will be notified shortly thereafter for a possible interview and the final answer will be given in late June/early July.

Salary: the candidate will be paid about 1 975 euros gross per month.

Other: as this thesis is part of an ANR project involving entities (Laboratoire Hubert Curien in Saint-Etienne and Naverlabs in Grenoble), the recruited person will be strongly encouraged to work in close collaboration with the different members of the project.

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.
- 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.
- 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.

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 importants 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 prédictive.

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 :

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

²le 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 Boriko 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.