Deep Neural Network Debiasing through Neural Architecture Search

Deep Neural Networks (DNNs) have gained popularity over the past few decades, entering various fields like healthcare, education, retail, and entertainment to name a few. Their use as a tool is often followed by several concerns regarding privacy and security, transparency and **explainability**, **fairness**, reliability, and ethics.

To address these concerns, several regulations and guidelines have been proposed. In 2016, the General Data Protection Regulation (GDPR), proposed by the European Union in 2012, was approved and has been enforced since May 2018. More recently, in 2021, the European Commission proposed the Artificial Intelligence Act (AI Act) to classify and regulate artificial intelligence applications based on their **risk of causing harm**. Specifically, the AI Act remarks certain AI-related technologies as at "high risk". We urge providing warranties. ANNs are unfortunately under the threat of learning spurious relations in data, leading among others to the potential problem of bias in the ANN model's outcome. This typically is divided into two different typologies: **algorithmic bias** and social bias.

Algorithmic bias and social bias are related yet distinct concepts in the realm of artificial intelligence. Algorithmic bias, also known as technical bias, arises from factors such as biased training data or design choices during model development, leading to unintended discrimination or skewed outcomes within AI systems. In contrast, social bias is a broader societal issue, encompassing prejudiced judgments, stereotypes, and attitudes towards individuals or groups based on characteristics like race, gender, age, or socioeconomic status, often deeply rooted in cultural and historical norms and systemic inequalities. Social bias can be reflected in AI systems due to algorithmic bias but requires addressing systemic societal issues, raising awareness, promoting inclusivity, and crafting AI systems that promote fairness and equity. While mitigating algorithmic bias is crucial in combating social bias (and this will be treated in this project), these two forms of bias necessitate distinct strategies and considerations. It is for instance known that simpler models tend to learn "easier" features, which in some scenarios are represented by biases: can we have efficient models without employing massive computation? Can we restrain the computation towards the edge, such that we can also ensure privacy (preventing transmission of information to far servers)?

Within this PhD, we will try to address both the challenge of removing algorithmic bias and enhancing model efficiency, in a unique framework. More specifically, we are motivated by recent literature that a proper training strategy disallowing biases is indeed possible and that a complex architecture is not really necessary if a proper architectural configuration can be found. In such a sense, Neural Architecture Search (NAS) can come in support.

This PhD project will be within the framework of the Agence Nationale de la Recherche (ANR) project BANERA, which targets in the long run to provide a library of efficient, unbiased

(sub-)architectures, that will boost model deployability while at the same time improving model's interpretability.

Objectives

1. Characterize algorithmic bias in DNN models. For this phase, an extensive study of the state-of-the-art regarding the problem of bias within DNNs will be necessary.
2. Validate or develop a proper debiasing approach.
3. Study state-of-the-art Neural Architecture Search (NAS) algorithms and implement them with proper validation on real-scale datasets.
4. Integrate debiasing constraint(s) to the NAS learning dynamics.
5. Identify biased/debiasing architectures, aiding in the creation of an open-access library for bias-attracting/repulsing architectures.
6. Deploy all the aforementioned contributions on a set of eight use cases, which range from traditional settings for debiasing (Biased-MNIST, CelebA, ImageNet-A, Multicolor MNIST, Corrupted CIFAR-10), NAS (NAS-Bench and CIFAR-10) and more general settings, like Mental State classification from EEG signals and Video compression (VideoSet).

Methodology

- Conduct a comprehensive literature review to understand the current state-of-the-art.
- Develop mathematical models and/or effective algorithms to formalize the debiasing problem and the NAS approach.
- Design novel algorithmic approaches, potentially drawing inspiration from techniques such as network pruning, quantization, knowledge distillation, architecture optimization.
- Implement the proposed algorithms using proper languages and frameworks.

Profile searched

- Strong competencies in both computer science and machine learning/deep learning.
- Master's degree in computer science or related fields.
- Proficiency in PyTorch/Tensorflow and Python coding.
- Knowledge of the working principles of existing methods for Deep Neural Network training.
- Great commitment, passion, and enthusiasm toward research and learning.
- Advanced level in reading, writing, listening, and speaking English.
- [optional] Previous publications in Deep Learning.
- [optional] Knowledge of the French language.

FAQ

- *When is this PhD starting?* At the earliest date possible (but minimum two months after the interview for bureaucratic reasons).
- *Do I need to already have an MS title to apply?* No, but you can not start the PhD program without first providing proof of graduation.
- *Is the PhD in presence?* YES, and this point is non-negotiable.
- *Will all the instrumentation I need be provided?* Yes, we have a budget for buying a PC and accessories allowing the work development. Besides, we have both GPU-equipped PCs, an internal computing cluster counting >50 GPUs, plus an application for more computation resources will be possible. Finally, in case of acceptance of article(s) to top conferences (core A/A*), funding for travel expenses is also allocated.

How to apply?

Please send an email with your CV, Master's transcript, and motivation letter to `enzo.tartaglione@telecom-paris.fr`.