Internship proposal

Title

Explainability methods for deep learning

Keywords

Artificial Intelligence, Deep learning, explainability

Location

Université Côte d'Azur Laboratoire I3S, UMR CNRS 7271 2000 route des Lucioles 06900 Sophia Antipolis France

Supervisors

Lucile Sassatelli, Maîtresse de conférences HDR, IUF, https://www.i3s.univ-cotedazur.fr/~sassatelli/ Frédéric Precioso, Professeur des Universités, https://www.i3s.univ-cotedazur.fr/~precioso/

Contact

lucile.sassatelli@univ-cotedazur.fr , 0489154347 frederic.precioso@univ-cotedazur.fr

Salary

French internship gratification : ~600€/month

Expected availability and duration

Starting in March 2022, 4 to 6 months

Description

This internship lies in the framework of ANR TRACTIVE, which is a national-funded project that regroups researchers from computer science, media studies, linguistics, and gender studies for the understanding of gender representation in visual media such as film. We integrate AI, linguistics, and qualitative media analysis in an iterative approach that aims to pinpoint the multimodal discourse patterns of gender in film, and quantitatively reveal their prevalence.

Deep learning is a branch of machine learning that has allowed crucial progress in several application domains, notably image classification. At the core are artificial Deep Neural Networks acting as parametric function approximators made of compositions of tunable functions, providing DNN with strong flexibility. Such flexibility however comes at the cost of a difficulty to explain the causes leading to the DNN's output. This is a crucial problem to prevent algorithmic bias that notably can socially impact prejudiced groups (depending on, e.g., gender, skin color and other attributes [Buo18, Rud19, Ger20]), or lead to error in medical diagnosis or car crash. That is why so-called explainability is currently a main research challenge in deep learning, with existing works for image and text classification [Rib16] or drug discovery [Jim20].

The objective of this internship is to conduct a comprehensive review and evaluation of recent explainability methods, applied to visual and textual data. The project will be carried out in 3 steps:

1/ conduct a comprehensive review of the current state of the art on explainability and interpretability in deep learning, focusing notably on a - feature attribution explainability methods, such as LIME [Rib16] or Anchors [Rib18]. They provide information on which part of the input has been most important in the decision, that is to post hoc explanations of decisions obtained with a "black box model" in the sense of [Rud19, Mel18]. b - self-explainable methods where a model's inner functioning is transparent enough such that the user can be sure of why and how the DNN makes a decision [Min19, Mel18].

2/ analyze the existing works and available code repositories and make a sub-selection of libraries or tools to evaluate,

3/ hands-on testing of selected tools and library on image and text data from the literature. In a second step, the different explainability methods will be tested on other text data, including political discourse [Van18] and film scripts.

4/ if time permits, identification of their drawbacks and proposal of potential solutions to improve the state of the art.

Pre-requisites

Mandatory: - A background in machine learning is mandatory. - Python programming proficiency Highly recommended - Knowledge of deep learning (theory and basic programming) is highly recommended. Appreciated: - Strong background in statistics

References

[Buo18] Buolamwini, J. and Gebru, T.. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender

Classification. In Proc. the 1st Conference on Fairness, Accountability and Transparency (FAT), Proc. of Machine Learning Research 81:77-91.
[Rud19] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1, 206–215, 2019.
[Ger20] Gerards, J. and Xenidis, R.. (2020). Algorithmic discrimination in Europe Challenges and opportunities for gender equality and non-discrimination law. European Commission special report, available at: https://op.europa.eu/en/publication-detail/-/publication/082f1dbc-821d-11eb-9ac9-01aa75ed71a1
[Rib16] Ribeiro, M. T., Singh, S. & Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 1135–1144 (ACM, 2016).
[Rib18] M. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In AAAI Conference on Artificial Intelligence, 2018.
[Jim20] Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. Nature Machine Intelligence, 2, 573–584 (2020).
[Mel18] D. A. Melis and T. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. NeurIPS 2018.
[Min19] Y. Ming et al.. Interpretable and Steerable Sequence Learning via Prototypes. ACM SIGKDD, 2019.
[Van18] Vanni, L., Ducoffe, M., Aguilar, C., Precioso, F., & Mayaffre, D. (2018). Textual Deconvolution Saliency (TDS): a deep tool box for linguistic analysis. In Proc. of the 56th Annual Meeting of the Association for Computational Linguistics.