Ongoing Research Projects on Responsible and Trustworthy AI at Università di Torino

Rossana Damiano, Antonio Lieto, Gian Luca Pozzato, Carlo Alberto Barbano, Enzo Tartaglione, Marco Grangetto, Faisal Imran, Rosa Meo
Dipartimento di Informatica, Università degli Studi di Torino
firstname.lastname@unito.it

Abstract

In this paper we summarize some research projects at Dipartimento di Informatica, Università di Torino on the topic of Responsible and Trustworthy Artificial Intelligence.

1 Introduction

The topic of Responsible and Trustworthy Artificial Intelligence is broad, including privacy-preserving data management and analysis, corrections of AI algorithms for guaranteeing fairness and respect of the ethical values: non discrimination of minority groups, transparency, non-maleficence and responsibility. Some humans decisions might have negative effects because of the outcomes of Machine Learning tasks that arise in presence of bias in the training data or in the mechanisms of the learning algorithms. We revise some of these topics in the following sections. In Section 2 we overview the project on AI for social inclusion. In Section 3 we present the tasks of de-biasing for learning in a privacy-preserving way in deep learning. Finally in Section 4 we present how randomisation protocols can be used to collect data with a guarantee of privacy with local-differential and in federated learning.

2 AI for social inclusion

Following the FARO convention signed in 2005[1], the role of cultural heritage has undergone a huge transformation from an instrument of individual education and elevation to a driver of inclusion and reflection in society. The SPICE project (Social Cohesion, Participation and Inclusion through Cultural Engagement)[2] aims at putting cultural heritage, and museums in particular, at the center of social inclusion processes, with the goal of creating bonds between individual and groups through art. In order to do so, SPICE engages museum visitors, online and offline, by involving them in collective curation processes, according to the paradigm of *citizen curation*. Within the framework of SPICE, the Unito group has developed a new generation of logic-based, affective, explainable and diversity seeking recommendation systems aiming at overcoming the limitations of traditional approaches, which rely on textual interpretations of artworks. In order to do so, it exploits the emotion labels attached by the users to the artworks as part of the collective curation processes put in practice by SPICE. By leveraging the more inclusive language of emotions to connect people to art, this approach goes beyond words, opening to a wider audience of people who don't have to access to written texts (due to literacy, sensory issues, etc.).

Based and the TCL logic [Lieto e Pozzato, 2020] and on the ArsEmotica ontology, our affective art recommender [Lieto *et al.*, 2021] exploits the opposition and similarity relations between emotions to suggest artworks associated with similar and opposite emotions. As a result, the user may receive not only recommendations of artworks that have been labeled with the same emotion, but also recommendations of artworks with similar and opposite emotion, leading to a more comprehensive and serendipitous exploration of collections. In addition, the systems allows the users to receive explanations about the attribution of emotions to artworks, by leveraging the association between user-generated labels and emotions.

3 Debiasing and privacy for deep learning

Trustworthiness, fairness and ethics have become increasingly important topics in deep learning. In this section we describe the latest research activities in the field of fairness and privacy for deep learning at the EIDOS lab research group [EidosLab, 2021]. The lab is also a member of the Italian Association for Computer Vision, Pattern Recognition and Machine Learning [CVPL, 2021].

Debiasing Artificial neural networks achieve state-of-the-art performance in wide variety of tasks, and nowadays they are used in many day-to-day applications. However, the generalization capabilities of this model is often questioned by problems such as the presence of biases in the training data. This is often the result of the data collection process, as it often is a non-trivial task. To address this issue, we have developed a debiasing technique, named EnD [Tartaglione *et al.*, 2021]. The aim of this techinque is to prevent deep models from learning unwanted biases by applying a regularization term which encourages the selection of unbiased features, and tries to suppress the selection of bias-related fea-

[1] http://conventions.coe.int/Treaty/EN/Treaties/Html/199.htm
[2] https://spice-h2020.eu

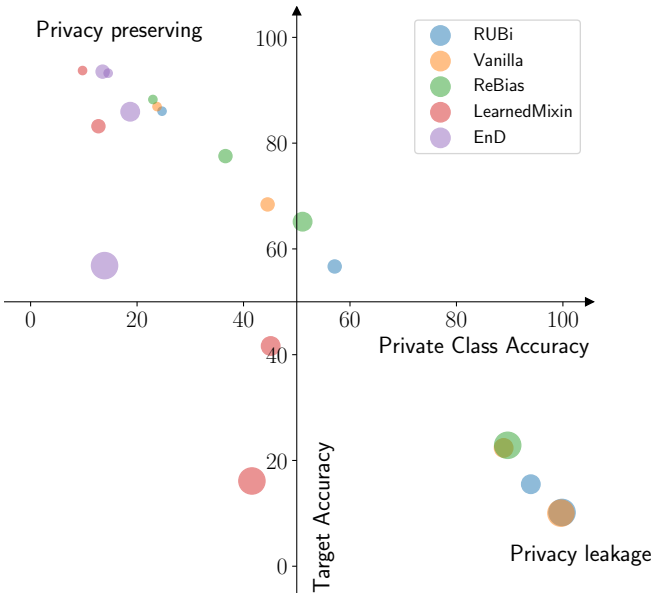

Figura 1: **Analysis of privacy leakage in different debiasing techniques.** Larger markers indicate higher correlation among private features and target class (in other words, the strenght of the bias) [Barbano *et al.*, 2021].

tures. EnD achieves state-of-the art performance in different real world scenarios.

Privacy in deep learning Besides biases issues, another important problem has risen: privacy preservation of sensitive attributes [Barbano *et al.*, 2021]. There are many similarities between debiasing and privacy preserving approaches: how far apart are these two worlds, when the private information overlaps the bias? We investigated the possibility of utilizing debiasing technique also to prevent privacy leakage. In this context, we are interested in recovering some private attribute of the data, starting from the model outputs or embeddings. These kind of private attributes can be, in the example of natural or facial images, age, gender, race, etc. We observed that, under certain conditions, some of the debiasing approaches are also suitable for privacy preservation. We discovered the determining condition to be the capability of effectively suppressing the bias related information inside of the model, rather than simply re-weighting it. Fig. 1 illustrates this concept: debiasing techniques can be used for privacy preservation purposes when they allow to retain a high accuracy on the target class, while making it harder to determine the private attributes.

4 The simplicity and the power of randomisation protocols

Some applications are data-driven and require an intense phase of data collection from patients or end-users: in healthcare, by service providers and policy-makers. In these fields, there is an increasing tension between the need to collect the data and the need to protect the individuals' sensitive information. Quite often, the data are collected from smart devices, like mobile devices, specialised sensors monitoring health-care conditions or smart home equipment and sensors connected to Internet, such as in the Internet of Things (IoT). If analysed the outcomes of data analysis might be beneficial for the society and the population like in applications for the early monitoring of emergencies, the support to olderly people from remote and for improving the services of organizations having a large user base, such as Google, Amazon, and Samsung. The data collected from the personal devices is intrinsically private and should be collected through a mechanism that guarantees the persons' privacy to make sure that privacy breaches do not occur. Some mechanisms like in Federated Learning [Li *et al.*, 2021] and in Local Differential Privacy [Bebensee, 2019] solve the privacy problems by collecting randomized responses from the users. In the version of the protocol we suggest, we do not need and are not willing to rely on a trusted data curator because this one is also a single point of failure in the mechanism of data collection and of the machine learning model creation. The advantages of randomised response techniques are that the production of the noisy outcome can be done locally, where users' data reside without any possibility of compromission of the users' privacy. According to the solution we propose, the curator can still build reliable prediction models on the collected amount of randomized data. Our approach utilizes the randomized response technique in a novel manner: it provides privacy-guarantees to users during the data collection and at the same time preserves the high-utility of the analysis [Imran, 2022]. Our proposed method can be seen as a special case of the generation of synthetic data by the production of noisy contingency tables (marginals) in a privacy-preserving mechanism. We describe our randomized response techniques and discuss the motivating applications domains with the details of the properties of differential privacy. We characterise the notion of utility [Lopuhaä-Zwakenberg *et al.*, 2019], both theoretically and by means of experimental analysis in which we compare our protocol with traditional privacy-preserving mechanisms such as differential privacy with Laplace distributed noise and geometric noise [Kacem e Palamidessi, 2018]. The advantages of randomised response techniques are that the production of the noisy outcomes can be done locally, directly where users' data reside, without any possibility of compromising the users' privacy. In addition, the protocol is light and can be executed even embedded on the sensors equipment.

Riferimenti bibliografici

[Barbano *et al.*, 2021] Carlo Alberto Barbano, Enzo Tartaglione, e Marco Grangetto. Bridging the gap between debiasing and privacy for deep learning. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3799–3808. IEEE, 2021.

[Bebensee, 2019] Björn Bebensee. Local differential privacy: a tutorial, 2019.

[CVPL, 2021] CVPL. Italian Association for Computer Vision, Pattern Recognition and Machine Learning. http://www.cvpl.it, 2021.

[EidosLab, 2021] EidosLab. Image proecessing, computer vision and virtual reality. http://eidos.di.unito.it, 2021.

[Imran, 2022] Faisal Imran. *Randomisation Protocols for Privacy Preserving Data Mining*. PhD thesis, Department of Computer Science, University of Torino, 2022.

[Kacem e Palamidessi, 2018] Lefki Kacem e Catuscia Palamidessi. Geometric noise for locally private counting queries. In *Proceedings of the 13th Workshop on Programming Languages and Analysis for Security (PLAS 2018), Toronto, Canada*, pages 13–16, 2018.

[Li *et al.*, 2021] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, e Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, page 1–1, 2021.

[Lieto *et al.*, 2021] Antonio Lieto, Gian Luca Pozzato, Stefano Zoia, Viviana Patti, e Rossana Damiano. A commonsense reasoning framework for explanatory emotion attribution, generation and re-classification. *Knowl. Based Syst.*, 227:107166, 2021.

[Lieto e Pozzato, 2020] Antonio Lieto e Gian Luca Pozzato. A description logic framework for commonsense conceptual combination integrating typicality, probabilities and cognitive heuristics. *Journal of Experimental and Theoretical Artificial Intelligence*, 32(5):769–804, 2020.

[Lopuhaä-Zwakenberg *et al.*, 2019] Milan Lopuhaä-Zwakenberg, Boris Skoric, e Ninghui Li. Information-theoretic metrics for local differential privacy protocols. *CoRR*, abs/1910.07826, 2019.

[Tartaglione *et al.*, 2021] Enzo Tartaglione, Carlo Alberto Barbano, e Marco Grangetto. End: Entangling and disentangling deep representations for bias correction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13512, 2021.