Lossy Compression for Privacy Preserving Representations

Nasib Naimi
Master's Thesis Proposal | ETH Zurich

January 4, 2023

1 Introduction

Artificial intelligence is permeating every field, with over 5% of peer-reviewed research publications being on AI [1]. Not only in academics, AI or statistical learning systems are increasingly being used across industries such as manufacturing [2], agriculture [3], education [4], and many others. The growing adoption follows the increasingly automated data collection, but in the era of big data, privacy of users and individuals remains the key concern [5]. Data typically involves a party generating and curating the data, a data owner, and a party wishing to use the data: the data user. In some cases, a data owner may need to release data to downstream users without having any prior knowledge of the intended uses or tasks that will be performed with the data [6]. But doing so safely, without revealing sensitive, identifying information about a user, remains a challenge. In order to facilitate the safe dissemination of visual data, we propose learning a lossy compression that yields a representation that is invariant to sensitive information, thereby anonymizing it and irrecoverably destroying it in the process.

2 Related Work

This thesis aims to build on previous work in representation learning and compression to derive provable private representations that allow for good predictive performance downstream.

Learning Controllable Fair Representation [7] presents an information theoretic objective for learning expressive, yet fair representations. The authors assume knowledge of sensitive attributes $\mathbf{u} \in \mathcal{U}$ and attempt to transform data points $(\mathbf{x}, \mathbf{u})$ into a new representation $z \in \mathcal{Z}$ which is *transferable*, i.e. useable inplace of $(\mathbf{x}, \mathbf{u})$ and **fair**, i.e. any decisions made by downstream classifier over z should be independent of the sensitive attributes u. The optimization objective to achieve both is formulated as

$$\max_{\phi \in \Phi} I_q[X; Z|U] \text{ s.t. } I_q[Z; U] < \epsilon \tag{1}$$

Similarly, representations can be learned which contain no mutual information with attributes that threaten the privacy and anonymity of the data. However, this would require clear knowledge of the sensitive attributes, that may be easily separable in the original representation to begin with.

In a similar vein, representation learning methods for obfuscating sensitive or private attributes have been an active field of study [8] [9] [10] [11] [12] [13]. Most methods rely on the idea of adversarial training, where a minmax game is played between a classifiers preserving utility and separate classifiers that attempt to retrieve private attributes [9] [12] [8]. In [9], adversarial learning for text representations is used, where the loss is formulated as minmax game between a semantic meaning discriminator preserving the utility of the representation and a private attribute discriminator, preserving the privacy of the representation. In [8], they formulate the objective using mutual information:

$$\min_{p(y|x)} I(U; X|Y) \text{ s.t. } I(S; Y) \leq k. \tag{2}$$

That is reformulated as a unconstrained optimization objective and expressed by means of expectation and parametric neural networks. This approach formulates bounds for both the preservation of the utility and obfuscation of the private attribute, but relies on adversarially learned proxy for information. It was shown to perform well on the filtering of images while preserving the space, hence allowing the reuse of existing pipelines.

The paper *Lossy Compression for Lossless Prediction* [14] formulates a self-supervised method for obtaining compressors which produce representations that are invariant to a set of transformations but allow for the same predictive performance as would be achieved with the uncompressed counterpart. Following their derivation, we can find the compressed representation containing all necessary and invariant information by finding the distribution which minimizes both the mutual information and the Bayes risk on the standard log loss,

$$\arg\min_{p(Z|X)} I[X; Z] + \beta R[M(X)|Z] \tag{3}$$

Which can be formulated as more practical variational bounds in order to define the loss function. Similarly, we can impose guarantees on anonymity and privacy by enforcing that downstream tasks should be invariant to any sensitive information or attributes. This can then be learned through augmentations, where through another process, data is transformed to either include or not include the sensitive information to which we should be invariant.

Finally, when discussing privacy it is important that the desired type of privacy is specified. Differential privacy (DP) proposed by Dwork et al. [15] aims at protecting the privacy of the individual by

preventing an attacker to determine whether an individual was part of a dataset. Differential privacy was extended to machine learning by Abadi et al. [16] in order to train differentially private machine learning models. Nonetheless, machine learning models were shown to unsafe against membership inference attacks (MIA), which are closely related to DP [17].

In contrast, we aim to protect the privacy of the individuals' sensitive or private attributes. A data breach would then consist of any sensitive data pertaining to the individual being retrieved from the compressed representation. Hence, we are concerned about attribute inference attacks (AIA).

We note, that there are deep connections between attribute inference and membership inference [18].

3 Our Contribution

Attribute inference attacks can reveal private attributes of individuals who may not have consented to that particular attribute being accessible.

Applying this idea to computer vision: what if we do not want to retain information which reveals the identity of people that are, for example, walking through a subway or working on a factory floor? In scenarios like these, interest lies in, e.g. predicting congestion based on human traffic, monitoring production lines for defects, or learning to navigate robots in crowds. In these scenarios, individuals are not of interest and any identifying information on the person is not required for good predictive performance.

Furthermore, in many scenarios where data collection is performed on edge and memory budgets are limited, it is advantageous to store the data in compressed form.

In this work, we aim to give the theory of private representations, develop targeted attribute obfuscation representation learning methods, and a general task-specific representation learning method that obfuscates all attributes which do not contribute to the downstream task performance. The ultimate goal is to find a powerful representation learning method that is ideally both attribute obfuscating and differentially private.

3.1 Theoretical Contributions

As stated, we aim to derive a theory for private representations. Through this, we hope to relate compression to attribute obfuscation. As attribute inference is related to membership inference and membership inference attacks are related to differential privacy [17], we hope to find a relation between compression and differential privacy, ideally showing that the optimal representation for each are equivalent.

3.2 Targeted Attribute Obfuscation

In the case where private attributes to be obfuscated are known, we aim to develop a framework building on *Lossy Compression for Lossless Prediction* for learning representations that obfuscate the desired attribute. For images, this could be done by using separate detection models to find the private attribute in the sample, augment the sample via inpainting, and then learning a compressor with a discriminative loss together with the BINCE objective [14]. Any inpainting of specific features formulates the transformations we wish the compressor to be invariant to.

Following this, the compressed representations can be used downstream to train models, e.g. for gauging congestion, monitoring assembly lines, while ensuring that private attributes can no longer be retrieved. Differentially private SGD can be used to ensure that downstream models remain resilient to MIAs.

3.3 General Attribute Obfuscation

When there are no specific private attributes we wish to protect or they are not explicitly known, we wish to find the minimal representation that allows for good downstream performance. Pretrained self-supervised models were explored as generic compressor and task-specific compressors [14]. The CLIP model [19] presents an especially powerful compressor, mapping images to detailed captions. We wish to investigate whether task-specific compressors trained on CLIP representations using BINCE, image-to-text compression can provide general attribution obfuscation.

Investigating task-specific compressors as privacy preserving representation learners can open up pathways for sharing datasets, that in their current form may contain information which does not assure the privacy of individuals. The goal ultimate goal of this work is to find one powerful representation learning method that is compressed and both attribute obfuscating and differentially private.

4 Timeplan

TABLE 1 Timeline

Month 1	Literature Review, code-base setup, defining benchmarks
Month 2	Mathematical formulation of approach
Month 3	Implementation of algorithms, initial experiments
Month 4	Extensive experiments and comparison to benchmarks
Month 5	Using insights gained, extend approach and run supplementary experiments
Month 6	Writing of thesis, presentation

References

[1] Artificial intelligence for science.

[2] Rahul Rai, Manoj Kumar Tiwari, Dmitry Ivanov, and Alexandre Dolgui. Machine learning in manufacturing and industry 4.0 applications. *International Journal of Production Research*, 59(16):4773–4778, August 2021.

[3] Tanha Talaviya, Dhara Shah, Nivedita Patel, Hiteshri Yagnik, and Manan Shah. Implementation of artificial intelligence in agriculture for optimisation of irrigation and application of pesticides and herbicides. *Artificial Intelligence in Agriculture*, 4:58–73, 2020.

[4] Bernard Marr. How Is AI Used In Education – Real World Examples Of Today And A Peek Into The Future.

[5] Shui Yu. Big Privacy: Challenges and Opportunities of Privacy Study in the Age of Big Data. *IEEE Access*, 4:2751–2763, 2016.

[6] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations, October 2018. arXiv:1802.06309 [cs, stat].

[7] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning Controllable Fair Representations, March 2020. arXiv:1812.04218 [cs, stat].

[8] Martin Bertran, Natalia Martinez, Afroditi Papadaki, Qiang Qiu, Miguel Rodrigues, Galen Reeves, and Guillermo Sapiro. Adversarially Learned Representations for Information Obfuscation and Inference.

[9] Ghazaleh Beigi, Kai Shu, Ruocheng Guo, Suhang Wang, and Huan Liu. Privacy Preserving Text Representation Learning. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 275–276, Hof Germany, September 2019. ACM.

[10] Ang Li, Jiayi Guo, Huanrui Yang, Flora D. Salim, and Yiran Chen. DeepObfuscator: Obfuscating Intermediate Representations with Privacy-Preserving Adversarial Learning on Smartphones. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pages 28–39, May 2021. arXiv:1909.04126 [cs].

[11] Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning Anonymized Representations with Adversarial Neural Networks, February 2018. arXiv:1802.09386 [cs, stat].

[12] Jiawei Chen, Janusz Konrad, and Prakash Ishwar. VGAN-Based Image Representation Learning for Privacy-Preserving Facial Expression Recognition, September 2018. arXiv:1803.07100 [cs].

[13] Hui-Po Wang, Tribhuvanesh Orekondy, and Mario Fritz. InfoScrub: Towards Attribute Privacy by Targeted Obfuscation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3275–3283, Nashville, TN, USA, June 2021. IEEE.

[14] Yann Dubois, Benjamin Bloem-Reddy, Karen Ullrich, and Chris J. Maddison. Lossy Compression for Lossless Prediction, January 2022. arXiv:2106.10800 [cs, math, stat].

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis.

[16] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, Vienna Austria, October 2016. ACM.

[17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks against Machine Learning Models, March 2017. arXiv:1610.05820 [cs, stat].

[18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, May 2018. arXiv:1709.01604 [cs, stat].

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. arXiv:2103.00020 [cs].