The Impact of Differential Privacy on Post-Hoc Methods of Model Explanation

Shefali Singh
Northeastern University

1 INTRODUCTION

Modern parallelized computing architecture has given humans the ability to leverage massive amounts of data in ways previously unfathomable. This phenomenon, known as 'Big Data' refers not only to the amount of data, but also to the personalized and granular nature of such data [6]. Models are used to direct this data into conclusions about populations and individuals within these populations – medical records can be used to determine insurance rates; criminal history can be used to determine the length of a prison sentence. From such models have emerged serious privacy concerns. How should companies regulate their data collection and model creation practices to protect individual privacy? These concerns are not just grounded in theoretical debate but also codifed into legislation. For instance, the European Union's GDPR protects users against data which allows individuals to be identified or singled out [3]. Therefore, for models to be usable, there must exist reasonable protections against identification in the dataset. One method of reasonable protection which has grown in popularity over the last decade is known as **differential privacy**, [3] which works, roughly speaking, by intelligently adding noise to a model, thus severing the one to one tie of data collected by the model and data used by the model.

With the application of models to determine decisions as significant as medical diagnoses, credit scores, insurance premiums, and criminal sentence severity, concerns on model **transparency** have also emerged. The complexity of such models often create an opaque decision procedure commonly referred to as 'black box.' To address the issue of transparency within black-box models, methods of **post-hoc model explanation** have emerged [10] Such explanations involve using tools not a part of the model itself to analyze the model – adding natural text to elucidate how to step through a model for example, or using salience maps [10]. The GDPR underscores the significance of such explanations by granting data subjects the **right to explanation** [1].

This paper seeks to understand the impact of differential privacy on methods of post hoc-explanation. How does introducing noise into the inner workings of a model affect the accuracy of an explanation which is produced based on the output of the model? Do such explanations differ for correctly made model predictions versus incorrectly made model predictions?

2 BACKGROUND AND RELATED WORK

Differential privacy:
The effectiveness of traditional statistical disclosure limitation (SDL) techniques has been significantly impeded by advances in analytical capabilities, increases in computational power, and the expanding availability of personal data from a wide range of sources [16]. Traditional SDL methods include but are not limited to data aggregation, k-anonymization, and p-sensitivity [16].

One privacy threat which has been shown to not be properly mitigated by traditional SDL techniques is known as a **membership inference attack** (MIA) [3]. A MIA is most relevant in models where an individual being identified as part of a training dataset has negative privacy implications for that individual. For example, being part of a training dataset of those with a rare disease means that an individual being part of the dataset reveals that she has the disease.

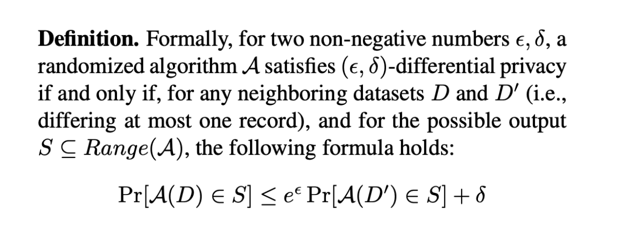
An MIA could can come from data re-linkage with the dataset and an auxiliary source [16]. An MIA could also come in the form of inferences via model overfitting – exploiting confidence values to discern which datapoints were seen during training [3]. Differential privacy is important because it provides significant protection against MIAs, and also guards against attacks unknown at the time of model deployment [16]. DP does not force secrecy around a differentially private computation or its parameters [16], and allows for publication and data releases which are not always possible with traditional SDL [16].

Differential privacy mathematically guarantees that anyone viewing the result of a differentially private analysis will essentially make the same inference about any individual's private information, whether or not that individual's private information is included in the input to the analysis [16].

One technique for DP is known as noisy stochastic gradient descent, or **noisy SGD** [3]. SGD is used to optimize an objective function by calculating the downward 'slope' of the data in batches, and converging into a global (and not batch specific) minimum. It uses random noise (hence the word 'stochastic) to help calculate the minimum [1]. Noisy SGD adds noise which is not a part of the training data in order to annoynmize the training data [1]. During training, noise can be added directly to the parameter gradients themselves, which preserves the privacy of the data. [1].

[1] For discussion on the extent to which the GDPR covers such a right, see [14]

Conference'17, July 2017, Washington, DC, USA
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn



Definition. Formally, for two non-negative numbers ϵ, δ, a randomized algorithm $\mathcal{A}$ satisfies (ϵ, δ)-differential privacy if and only if, for any neighboring datasets D and D' (i.e., differing at most one record), and for the possible output $S \subseteq Range(\mathcal{A})$, the following formula holds:

$$\Pr[\mathcal{A}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{A}(D') \in S] + \delta$$

Figure 1: Formal Def. of DP, from [3]

An important parameter for understanding the noise within a differentially private algorithm is the ϵ parameter (Figure 1) as it functions as a ’tuning knob’ for the tradeoff between privacy and accuracy. A larger ϵ value signifies a larger privacy budget (and larger noise multiplier), and thus a smaller amount of privacy but greater amount of accuracy. A smaller ϵ value signifies a smaller privacy budget (and smaller noise multiplier), and thus a greater amount of privacy but lesser amount of accuracy.

Post-hoc explanations:
One method of model agnostic post hoc explanations is LIME, or Locally Interpretable Model Agnositic Explanation [4] [12]. Lime works by perturbing a number of datapoints around an original datapoint, and then calculating weights for a new explanation model by determining the relation between the perturbed injected datapoints and the original datapoint. As the name indicates, LIME generates explanations which are locally but not globally faithful. by finding local linear explanations for feature importance. [2] LIME provides an out of the box assignment of feature importance which can apply to a variety of classification models. [4]

Differential Privacy and Post Hoc Explanations:
While there is significant amount of research on the limitations and applications of differential privacy [16] [13], and the limitations an applications of post hoc explanations [9] [5] [15], there has been less research on how the two relate to each other.

Some research does exist: Patel, Shokri, and Zick create an adaptive algorithm [11] which provides an explanation that leaks significantly less training data information while falling under a required privacy budget. They demonstrate that sparse data leads to poorer performance either in terms of privacy or accuracy of explanation [11].

This paper differs in scope than Patel et. al [11] as it focuses on analyzing the efficacy of the existing LIME algorithm explanation rather than creating a new method of explanation.

3 METHODS

Though there exist limitations in differential privacy methods and post hoc explanation methods, both methods are still heavily in use and therefore warranted examination. This section provides an outline of how impact of differential privacy on post hoc explanations was measured.

Data:
The dataset used was the CIFAR10 dataset [8], a collection of 50,000 images with 10 class labels. From this dataset, a smaller sample of 10,000 images were chosen, with roughly equal proportions of samples chosen for each of the 10 classes.

Models:
5 models were created: an untrained model, a trained non-private model, a differentially private model with an epsilon parameter of .1, a differentially private model with an epsilon parameter of 5, and a differentially private model with an epsilon parameter of 25. Those epsilon value were chosen as they represent high privacy, worthwhile but low privacy, and extremely low privacy respectively. [7] The three differentially private models were trained with non private settings and then the final layer was fine tuned with differential privacy. The models were all residual neural network models, with cross entropy loss functions and stochastic gradient descent optimizer functions.

Evaluation:
The correct and incorrect predictions of the entire 50,000 CIFAR10 training set were stored and grouped by class for each model. Three experiments were performed: An evaluation of explanation for an image which was predicted correctly for all 5 models, an evaluation of an image which was predicted incorrectly for all 5 models and an evaluation of an image which was predicted correctly for a non private model and incorrectly for a high privacy models. To combat deviation in explanation, explanations were generated 5 times each for each for each model for each image, generating a total of 25 images per experiment.

4 RESULTS

All images of explanations follow the format: From left to right: same model but a new explanation generated. From top to bottom: untrained model, trained non-private model, model with $\epsilon = .1$, $\epsilon = 5$, $\epsilon = 25$.

Rows three and four of Figure 2 demonstrates a greater variation in explanation than rows one two and five. Rows three and four were the only rows which depict models trained with reasonable levels of privacy, indicating that the variation in explanation varies more for incorrect predictions.

Images that were correctly predicted by all models (Figure 3) and incorrectly predicted by all models (Figure 4 had less variation in explanation.

5 CONCLUSION

People who receive incorrect assessments (for example, receiving an incorrect medical diagnoses based on the automated decision model which takes patient CT scans as input) are more likely to seek legal recourse through the legislation such as the GDPR, therefore

[2] Another popular method, not studied in this paper, is called SHAP, based on the game theoretic idea that a players value varies depending on whether or she is part of the team, see [4] for further discussion.

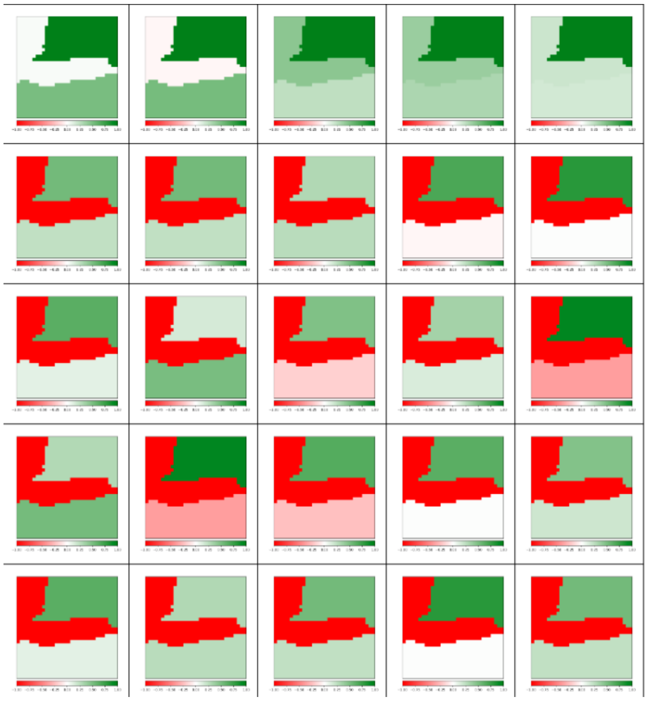


Figure 2: Correctly predicted image for non private model, incorrectly predicted image for private model explanations

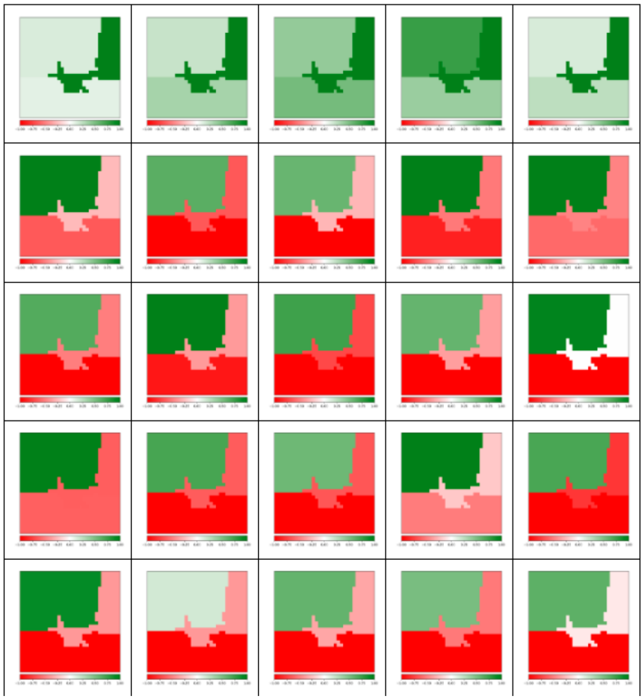


Figure 3: Correctly predicted image explanations

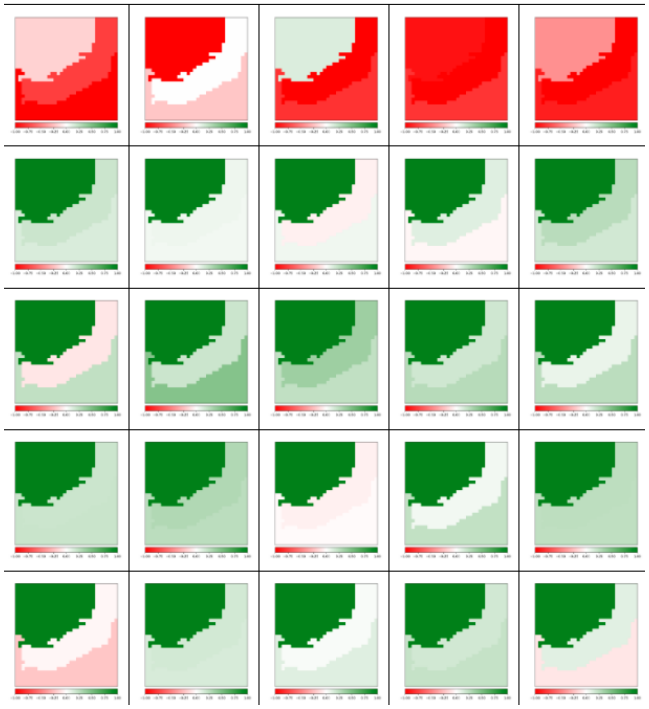


Figure 4: Incorrectly predicted image explanations

it is vital that explanatory methods possess consistency for incorrect predictions. However, the findings demonstrate that variation in private versus non private explanation is greater for incorrect predictions than in correct predictions. Thus, more work needs to be done in order to standardize explanations while maintaining a reasonable level of privacy.

There exist limitations to this research which point to the need for caution before the results of this study are generalizable, as well as the potential areas of improvement for future iterations of such research.

- **Computing power**. Models were evaluated on a small dataset of relatively low resolution (32x32 pixel) images, and trained on an even smaller dataset. Models were trained for only 5 epochs. More computing power would enable a higher resolution / larger dataset, and training for greater than 5 epochs, meaning the models would emulate the accuracy levels of models deployed in the real world.
- **Segmentation**. The LIME explanation works by using 'super pixels' for to attribute feature significance. The 'super pixels' mapping generated for the images that were analyzed were not always the most representative of the most intuitive mapping possible (See Figure 5). A future iteration could use images accompanied by preannotated masks, or develop a more precise segmentation algorithm.
- **Fine Tuning**. The differentially private models were fine tuned with differential privacy on the last layer of the non private model. A more robust approach would to be train the model on a separate dataset which emulates "public" data, and then fine tune on a separate dataset which emulates 'private data', considering the work of Yannis et. al [2] to obtain ideal hyperparameters.
- **Biased models**. The models were trained on subsets which had equal proportions of each class. An interesting extension of this research would be to create biased models which have

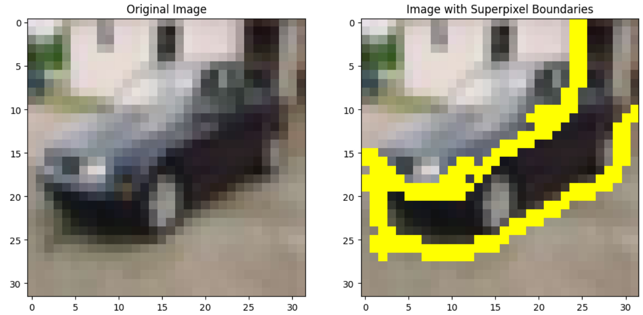


Figure 5: Image and 'superpixel' mask for incorrectly predicted image

unequal proportions of classes to study the effects of explanation variation on biased models. These biased models would need accuracy thresholds imposed or a parallel objective function in order to prevent the model from "shortcutting" and predicting every image as one class.

Ultimately the research provides a starting point for probing the complex issue of model privacy and explanation, and emphasizes the importance of explanation accuracy for differentially private models.

REFERENCES

[1] Differential Privacy Series Part 1 | DP-SGD Algorithm Explained. https://medium.com/pytorch/differential-privacy-series-part-1-dp-sgd-algorithm-explained-12512c3959a3. (????). Accessed: 2023-10-30.

[2] Yannis Cattan, Christopher A. Choquette-Choo, Nicolas Papernot, and Abhradeep Thakurta. 2022. Fine-Tuning with Differential Privacy Necessitates an Additional Hyperparameter Search. (2022). arXiv:cs.LG/2210.02156

[3] Emiliano De Cristofaro. 2020. An Overview of Privacy in Machine Learning. (2020). arXiv:cs.LG/2005.08679

[4] Avijit Ghosh. 2023. Interpretibility in Machine Learning. (2023). 2005.https://evijit.io/materials/Lecture_6_Model_Interpretability.pdf

[5] Leif Hancox-Li and I. Elizabeth Kumar. 2021. Epistemic Values in Feature Importance Methods: Lessons from Feminist Epistemology. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 817–826. https://doi.org/10.1145/3442188.3445943

[6] M. I. Jordan and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349, 6245 (2015), 255–260. https://doi.org/10.1126/science.aaa8415 arXiv:https://www.science.org/doi/pdf/10.1126/science.aaa8415

[7] Near Joseph and Darais David. 2022. Differential Privacy: Future Work Open Challenges. (2022). https://www.nist.gov/blogs/cybersecurity-insights/differential-privacy-future-work-open-challenges

[8] Alex Krizhevsky and Geoffrey Hinton. 2009. *Learning multiple layers of features from tiny images.* Technical Report 0. University of Toronto, Toronto, Ontario. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[9] I. Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. 2020. Problems with Shapley-value-based explanations as feature importance measures. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Hal Daumé III and Aarti Singh (Eds.), Vol. 119. PMLR, 5491–5500. https://proceedings.mlr.press/v119/kumar20e.html

[10] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[11] Neel Patel, Reza Shokri, and Yair Zick. 2022. Model Explanations with Differential Privacy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1895–1904. https://doi.org/10.1145/3531146.3533235

[12] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). arXiv:1602.04938 http://arxiv.org/abs/1602.04938

[13] Jayshree Sarathy. 2022. From Algorithmic to Institutional Logics: The Politics of Differential Privacy.

[14] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (12 2017), 233–242. https://doi.org/10.1093/idpl/ipx022 arXiv:https://academic.oup.com/idpl/article-pdf/7/4/233/22923065/ipx022.pdf

[15] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 180–186. https://doi.org/10.1145/3375627.3375830

[16] Alexndra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David O'Brien, Thomas Steinke, and Salil Vadhan. 2018. Differential Privacy: A Primer for a Non- Technical Audience. In *Vanderbilt Journal of Entertainment and Technology Law*. 209–276.