How Does Data Augmentation Affect Differential Privacy in Deep Learning?

Enrique David Guzman Ramirez
Department of Computer Science
University of Toronto

Jose Abraham Morales Vidales
Department of Statistical Sciences
University of Toronto

Rui Xian
Department of Statistical Sciences
University of Toronto

Abstract

Deep learning often adopts data augmentation as an essential and efficient technique to generate new training examples from existing data to improve the model robustness and generalization. Differential privacy is a technique used to preserve the privacy of individual data points while releasing statistical information about a dataset. In this work, we study the relationship between data augmentation and differential privacy in deep learning for image and text classification. We found that although data augmentation has a negative effect on the performance of models trained with differential privacy, it improves the model robustness against membership inference attacks [1].

1 Introduction

The data privacy of machine learning models refers to the security of its training data against external attacks [De Cristofaro, 2021, Strobel and Shokri, 2022]. Differential privacy (DP) [Dwork and Roth, 2013] has emerged as the leading privacy-preserving technique which provides a rigorous framework for releasing statistical information without revealing the identities of individual data points. This is achieved by introducing random noise to the data at each stage of query release. The wide-ranging applicability of DP has fueled its prompt adoption in machine learning systems [Munilla Garrido et al., 2023], thanks to the call for privacy-enhancing technologies to make models trained on sensitive personal or protected information less prone to adversarial attacks. The merging of deep learning with DP is currently under various ongoing investigations [Mireshghallah et al., 2020].

In training deep learning (DL) models, data augmentation (DA) [Shorten and Khoshgoftaar, 2019] is often used to boost the training data size and diversity, improve model robustness and generalizability, and preventing overfitting. However, its compatibility with DP is not yet known. We conjecture that DA may have a negative effect on the privacy of a private DL model. To the best of our knowledge, no existing work has quantitatively explored the trade-offs between privacy in DL models and model training with DA. Intuitively, the inclusion of DA requires splitting the privacy budget over the transformed data, therefore should make the models less private. We seek to investigate the trade-off between them, by comparing accuracy metrics of DP models with and without DA and probing the model privacy through membership inference attacks (MIAs) [Shokri et al., 2017], which measure the risk of revealing training data. Our main contributions are,

[1] The code we used to train and evaluate our models is available at https://github.com/davidguzmanr/CSC2516. All authors conducted background research. E. D. Guzman Ramirez programmed the model for image classification, J. A. Morales Vidales programmed the model for text classification, R. Xian analyzed the membership inference attacks, led the report writing with contributions from the others.

1. Applying DA improves the data privacy in trained neural network models.
2. Adding DA to differentially private neural networks decrease performance of the models.

2 Related works

DA in security and privacy Previous research have explored the idea of applying differential privacy to the deep learning framework. [Abadi et al., 2016] introduce a differntially private stochastic gradient descent (SGD) algorithm with minimal lost in model performance. It has recently been shown in [Tramèr and Boneh, 2020] that DL models suffer a significant loss of prediction accuracy when trained in a differentially private manner. [Sablayrolles et al., 2019] empirically observed that better generalization leads to worse inference success rate. The work of [Yu et al., 2020] explored the relationship between DA and traditional privacy-preserving mechanism. Their approach achieved a 70.1% success rate of MIAs, suggesting that there's still considerable room for improvement.

Data duplicates and privacy Data duplicates can be thought of as uncontrolled augmentation which exacerbate the class imbalance among training datasets and dominates the gradient during training. Their existence often hurts model performance as well as privacy due to the memorization behvaior of DL models. Recent work using deduplicated datasets to train language models show improved performance and less privacy leakage [Kandpal et al., 2022].

3 Preliminaries and background

Privacy risks in DL The characteristics of DL models make them vulnerable to privacy attacks [Mireshghallah et al., 2020]. A primary source of privacy risk is memorization. In language models, this leads to the emission of precise information verbatim, which poses severe privacy risks for personal or sensitive information [Song et al., 2017]. Another important aspect is overfitting, which creates a generalization gap can be exploited by adversaries in privacy attacks. Moreover, the privacy risk is also dependent on the model type and the diversity in the training data, etc [Hu et al., 2022].

Quantification of the privacy risk of DL models often uses the accuracy (i.e. success rate) of privacy attacks. The major types of privacy attacks on DL models include membership inference, attribute inference, and model inversion, in an increasing order of severity. Membership inference yields a binary outcome indicating whether a given instance of data is in the training set of the DL model [Shokri et al., 2017]. Attribute inference recovers certain attributes of the training data. Model inversion seeks to reconstruct entire training data from model parameters [Haim et al., 2022].

The baseline MIA [Yeom et al., 2018] (also called gap attack) uses all data from the combination of training and testing datasets. The black-box DL model is asked to answer a binary question whether the data shown is from training (1) or test (0) set. The accuracy (acc) of this baseline attack relates to the accuracy gap between training and testing and is defined as $\mathrm{acc}_{\mathrm{MIA}} = [1 + (\mathrm{acc}_{\mathrm{train}} - \mathrm{acc}_{\mathrm{test}})]/2$. The $\mathrm{acc}_{\mathrm{MIA}}$ is 1 when the training and testing accuracy is 1 (i.e. complete memorization), and 1/2 (i.e. random guessing) when the training and testing accuracies are indistinguishable.

Differential Privacy A randomized algorithm $\mathcal{A}$ is (ε, δ)-differentially private if for any E in the range of $\mathcal{A}$, and any two datasets X, X' that differ in only one element (i.e neighbouring datasets),

$$\mathbf{P}[\mathcal{A}(X) \in E] \leq e^{\varepsilon}\mathbf{P}[\mathcal{A}(X') \in E] + \delta. \tag{1}$$

The privacy parameter ε controls the trade-off between privacy leakage and accuracy of the algorithm. This requirement is guaranteed by tuning the random component in the algorithm $\mathcal{A}$. The common ways are to introduce additive noise with distributional constraints [Dwork and Roth, 2013].

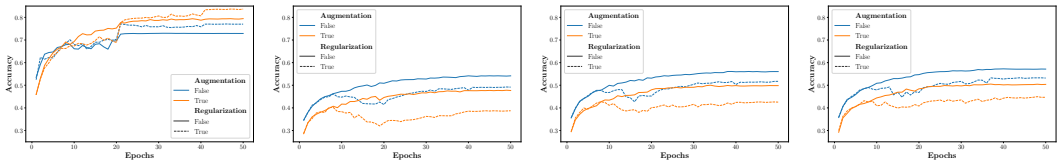
Differentially Private SGD (DP-SGD) The SGD algorithm [Bottou, 2010] is the primary workhorse in training DL models. For this work we used the (ε, δ)-DP SGD algorithm [Abadi et al., 2016], where the DP is guaranteed by the addition of Gaussian noise $\mathcal{N}(\mathbf{0}, \sigma^2 C^2 \mathbf{I})$ and gradient clipping (C as the gradient norm bound) to each optimization step. The hyperparameter $\sigma = \Omega(B/N\sqrt{T\log(1/\delta)\log(T/\delta)}/\epsilon)$ depends on the batch size (B), the sample size (N), and the number of training epochs (T). The gradient update rule, with learning rate η, for DP-SGD is,

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \frac{\eta}{|B|}\left(\sum_{i\in B}\mathrm{clip}_C(\nabla_{\boldsymbol{\theta}_t}\mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{x}_i)) + \mathcal{N}(\mathbf{0}, \sigma^2 C^2 \boldsymbol{I})\right). \tag{2}$$

4 Experimental results

Our experiments involve training image and text classification models using existing deep neural network architectures with and without data augmentation and the evaluation of their privacy using membership inference attacks. We used PyTorch-based Opacus package [Yousefpour et al., 2021] to implement differentially private model training.

Image classification. For image classification, we trained a ResNet-18 [He et al., 2016] using differentially private-SGD [Abadi et al., 2016] with the CIFAR-10 dataset (containing 50,000 training images and 10,000 test images) and compared the scenarios with or without data augmentation. In total, we used three sets of privacy parameters with (ε, δ) being $(10, 10^{-5})$, $(30, 10^{-5})$, and $(50, 10^{-5})$, respectively. The data augmentation was conducted using a Bernoulli draw with a probability of 0.5 for undergoing transformation, which include random cropping, and random horizontal flipping. These two types of DA transformation were chosen because of their better performance compared with others. In model training, we experimented with a weight decay of 10^{-3} and a multiplicative learning-rate scheduler to improve the model generalization. For each trained model, the optimization uses a minibatch size of 128 and was terminated after 50 epochs. In training with DP-SGD, the gradient norm bound C was chosen as 1.2 to optimize the validation accuracy.

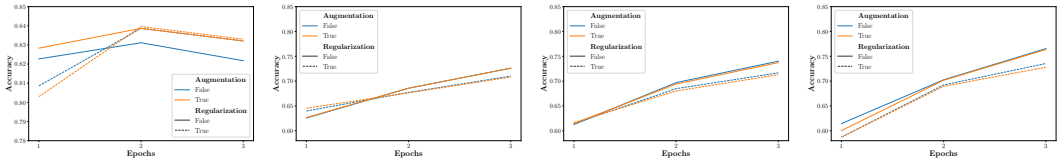


(a) $\varepsilon = \infty$ (no DP) (b) $\varepsilon = 10$ (c) $\varepsilon = 30$ (d) $\varepsilon = 50$

Figure 1: Test accuracy of the ResNet-18 model for different levels of privacy on CIFAR-10.

From the results in figure 1, we see that as the amount of noise increases (i.e., reduce ε) the accuracy of the model decreases, which is to be expected. Adding DA and weight-decay regularization improves the performance of models trained without DP. However, for private models, adding DA or weight decay actually hurts performance.

Text classification. We built a differentially private text classifier using BERT [Devlin et al., 2018] trained on a movie review dataset containing 8,530 entries for training and 1,066 for testing [Pang and Lee, 2005]. We froze most of the weights of the pre-trained model leaving us with 7,680,771 trainable parameters, which were optimized using the DP-SGD algorithm [Abadi et al., 2016]. For DA, we used the methods by [Wei and Zou, 2019], which include random swapping word positions, random word removal, insertion of synonyms in a random location, and replacement of a random word by its synonym. We used the implementation of these methods in the `textattack` package [Morris et al., 2020] to perform DA on a random sample selected from the training set. For each data entry, the DA method was chosen at random from the list described previously. We trained models with and without DA and weight-decay regularization. For models trained with differential privacy, we used three sets of privacy parameters (ε, δ): $(3, 10^{-4})$, $(5, 10^{-4})$, $(10, 10^{-4})$ and non-private. We chose the Adam optimizer [Kingma and Ba, 2014] with a batch size of 32, learning rate of 5×10^{-4}, weight decay of 10^{-2} and gradient clipping norm bound of 1.2.



(a) $\varepsilon = \infty$ (no DP) (b) $\varepsilon = 3$ (c) $\varepsilon = 5$ (d) $\varepsilon = 10$

Figure 2: Test accuracy of BERT for different levels of privacy on the Movie Review Dataset.

From the results in figure 2, we can clearly see that as we increase the amount of noise (i.e., reducing ε) the accuracy of the model decreases. As expected, the performance of non-private models

increases when DA and regularization are applied. For private models, we find that DA hurts the test accuracy. This is in agreement with the observations in the image classification task. However, in text classification, the differences in accuracy are almost negligible between models with and without DA or regularization, specially for the unregularized models with $\varepsilon = 3$ in figure 2b.

Model	No weight decay		With weight decay	
	No data augmentation	With data augmentation	No data augmentation	With data augmentation
No DP	**87.85**	**74.67**	87.08	78.40
$\varepsilon = 50$	**57.67**	50.53	52.65	**46.49**
$\varepsilon = 30$	**56.72**	49.81	51.54	**45.10**
$\varepsilon = 10$	**54.75**	48.69	49.73	**42.23**

Table 1: Accuracy of membership inference attacks on CIFAR-10. The least robust model against MIA for each setting is marked in red, while the most robust is marked in blue.

Model	No weight decay		With weight decay	
	No data augmentation	With data augmentation	No data augmentation	With data augmentation
No DP	79.04	**79.79**	78.42	**77.47**
$\varepsilon = 10$	**70.47**	70.11	68.15	**67.76**
$\varepsilon = 5$	**68.45**	68.09	66.72	**66.48**
$\varepsilon = 3$	**67.48**	67.01	66.05	**65.69**

Table 2: Accuracy of membership inference attacks on Movie Review dataset. The least robust model against MIA for each setting is marked in red, while the most robust is marked in blue.

Membership inference. We used the implementation of the MIAs within the Adversarial Robustness Toolbox [Nicolae et al., 2018] to carry out the gap attack. The outcomes are shown in Tables 1-2 are calculated using both the training and testing datasets in each classification context. Along the vertical direction in the table, we see that for both the image and text classification tasks, models trained with differential privacy show a marked drop in the MIA accuracy than those without (equivalent to $\varepsilon = \infty$), indicating the improvement in security. The attack accuracy decreases with increasing model privacy (i.e. decreasing ε).

Comparing the MIA accuracy across settings in the horizontal direction in Tables 1-2, we see that for private models, the setting with no data augmentation and no regularization is the most vulnerable to MIA (labeled in red). In general, we see consistent improvement in privacy for models trained with DP-SGD in the order: $\text{acc}_{\text{MIA}}^{\text{DA,WD}} < \{\text{acc}_{\text{MIA}}^{\text{DA,nWD}}, \text{acc}_{\text{MIA}}^{\text{nDA,WD}}\} < \text{acc}_{\text{MIA}}^{\text{nDA,nWD}}$. Here, the superscript indicates the experimental setting, with nWD meaning no weight decay. Because both regularization (such as weight decay) and DA improve model generalization by reducing overfitting, our results provide further evidence that better generalization leads to improved model privacy.

5 Conclusion

Our experiments demonstrate that while data augmentation is widely applicable to enhancing the performance of deep learning models, it should be used with caution when combined with differential privacy: the balancing effect between the factors leading to generalization (e.g. model accuracy or utility) and privacy should be accounted for. Our observations motivate the development of augmentation techniques that are more efficient at improving generalization with fewer instances [Cubuk et al., 2020], therefore compromises less the accuracy of differentially private models. Moreover, private training algorithms that take into account the data distribution, such as those that can adaptively adjust the privacy parameters [Bassily et al., 2021] in order to account for data augmentation are also a potentially promising direction in treading the balance between privacy and utility in diverse applications.

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, oct 2016. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145%2F2976749.2978318.

R. Bassily, K. Nissim, A. Smith, T. Steinke, U. Stemmer, and J. Ullman. Algorithmic Stability for Adaptive Data Analysis. *SIAM Journal on Computing*, 50(3):STOC16–377–STOC16–405, jan 2021. ISSN 0097-5397. doi: 10.1137/16M1103646. URL https://epubs.siam.org/doi/10.1137/16M1103646.

L. Bottou. Large-Scale Machine Learning with Stochastic Gradient Descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Physica-Verlag HD, Heidelberg, 2010. doi: 10.1007/978-3-7908-2604-3_16. URL http://link.springer.com/10.1007/978-3-7908-2604-3_16.

E. D. Cubuk, B. Zoph, J. Shlens, and Q. Le. RandAugment: Practical Automated Data Augmentation with a Reduced Search Space. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18613–18624. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d85b63ef0ccb114d0a3bb7b7d808028f-Paper.pdf.

E. De Cristofaro. A Critical Overview of Privacy in Machine Learning. *IEEE Security & Privacy*, 19(4):19–27, jul 2021. ISSN 1540-7993. doi: 10.1109/MSEC.2021.3076443. URL https://ieeexplore.ieee.org/document/9433648/.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

C. Dwork and A. Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407, 2013. ISSN 1551-305X. doi: 10.1561/0400000042. URL http://www.nowpublishers.com/articles/foundations-and-trends-in-theoretical-computer-science/TCS-042.

N. Haim, G. Vardi, G. Yehudai, O. Shamir, and M. Irani. Reconstructing training data from trained neural networks. *ArXiv e-prints*, June 2022. arXiv:2206.07758.

K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, jun 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang. Membership Inference Attacks on Machine Learning: A Survey. *ACM Computing Surveys*, 54(11s):1–37, jan 2022. ISSN 0360-0300. doi: 10.1145/3523273. URL https://dl.acm.org/doi/10.1145/3523273.

N. Kandpal, E. Wallace, and C. Raffel. Deduplicating Training Data Mitigates Privacy Risks in Language Models. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR, 2022. URL https://proceedings.mlr.press/v162/kandpal22a.html.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization. *arXiv*, page 1412.6980, dec 2014. URL http://arxiv.org/abs/1412.6980.

F. Mireshghallah, M. Taram, P. Vepakomma, A. Singh, R. Raskar, and H. Esmaeilzadeh. Privacy in Deep Learning: A Survey. *arXiv*, page 2004.12254, apr 2020. URL http://arxiv.org/abs/2004.12254.

J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and Y. Qi. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, 2020.

G. Munilla Garrido, X. Liu, F. Matthes, and D. Song. Lessons Learned: Surveying the Practicality of Differential Privacy in the Industry. *Proceedings on Privacy Enhancing Technologies*, 2023

(2):151–170, apr 2023. ISSN 2299-0984. doi: 10.56553/popets-2023-0045. URL https://petsymposium.org/popets/2023/popets-2023-0045.php.

M.-I. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat, M. Wistuba, V. Zantedeschi, N. Baracaldo, B. Chen, H. Ludwig, I. M. Molloy, and B. Edwards. Adversarial Robustness Toolbox v1.0.0. *arXiv*, page 1807.01069, jul 2018. URL http://arxiv.org/abs/1807.01069.

B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

A. Sablayrolles, M. Douze, Y. Ollivier, C. Schmid, and H. Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. *ArXiv e-prints*, Aug. 2019. arXiv:1908.11229.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, may 2017. ISBN 978-1-5090-5533-3. doi: 10.1109/SP.2017.41. URL http://ieeexplore.ieee.org/document/7958568/.

C. Shorten and T. M. Khoshgoftaar. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, 6(1):60, dec 2019. ISSN 2196-1115. doi: 10.1186/s40537-019-0197-0. URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0.

C. Song, T. Ristenpart, and V. Shmatikov. Machine Learning Models that Remember Too Much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 587–601, New York, NY, USA, oct 2017. ACM. ISBN 9781450349468. doi: 10.1145/3133956.3134077. URL https://dl.acm.org/doi/10.1145/3133956.3134077.

M. Strobel and R. Shokri. Data Privacy and Trustworthy Machine Learning. *IEEE Security & Privacy*, 20(5):44–49, sep 2022. ISSN 1540-7993. doi: 10.1109/MSEC.2022.3178187. URL https://ieeexplore.ieee.org/document/9802763/.

F. Tramèr and D. Boneh. Differentially private learning needs better features (or much more data). *ArXiv e-prints*, Nov. 2020. arXiv:2011.11660.

J. Wei and K. Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks, 2019.

S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, jul 2018. ISBN 978-1-5386-6680-7. doi: 10.1109/CSF.2018.00027. URL https://ieeexplore.ieee.org/document/8429311/.

A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov. Opacus: User-friendly differential privacy library in PyTorch. *ArXiv e-prints*, Sept. 2021. arXiv:2109.12298.

D. Yu, H. Zhang, W. Chen, J. Yin, and T.-Y. Liu. How does data augmentation affect privacy in machine learning? *ArXiv e-prints*, July 2020. arXiv:2007.10567.