Differentially Private Instance Encoding against Privacy Attacks

Shangyu Xie
Illinois Institute of Technology
sxie14@hawk.iit.edu

Yuan Hong
Illinois Institute of Technology
yuan.hong@iit.edu

Abstract

TextHide was proposed to protect the training data via instance encoding in the natural language domain. Due to the lack of theoretic privacy guarantee, such instance encoding scheme has been shown to be vulnerable to privacy attacks, e.g., reconstruction attacks. To address such limitation, we integrate differential privacy into the instance encoding scheme, and thus provide a provable guarantee against privacy attacks. The experimental results also show that the proposed scheme can defend against privacy attacks while ensuring learning utility (as a trade-off).

1 Introduction

Machine learning models have been widely deployed in a wide range of applications/domains, such as speech recognition (Zhang et al., 2018b), computer vision (Guo et al., 2020) and natural language processing (Chen et al., 2019; Radford et al., 2019; Brown et al., 2020). Meanwhile, the privacy issues have also aroused more and more attention as machine learning-based systems usually aggressively collect large amounts of data for better performance, which could contain user's personal information and thus jeopardize user's privacy. For instance, the hospital admission information and diagnosis report can be processed by language models to predict the readmission rate of a patient (Lehman et al., 2021). Another example is that the prediction of keyboard input would require personal users' daily input texts for better accuracy (Chen et al., 2019). This may not only lose customer trust, but also violate some data regulations or laws, e.g., GDPR (Wachter et al., 2017).

The privacy-enhancing technologies (PETs) (Gentry, 2009; Chaudhuri et al., 2011; Mohassel and Zhang, 2017; Cabrero-Holgueras and Pastrana, 2021) have been widely studied to ensure the data privacy in the machine learning, which mainly include two foundations of theory as following. First, the cryptographic protocols (Mohassel and Zhang, 2017; Mohassel and Rindal, 2018) can help to securely train the model with the private data (in encrypted format), and the privacy of data depends on the hard mathematical problems (Paillier, 1999). Although the cryptographic protocol-based schemes provide good data privacy, these also arouse high computational overheads due to the computation on encrypted data and other complicated building blocks.

Second, differential privacy (DP) (Dwork et al., 2006b, 2014) provides a lightweight way to protect the data against the adversaries with arbitrary information during the training, which can obtain quantifiable privacy guarantees. For example, the widely used DP-SGD (Bassily et al., 2014; Abadi et al., 2016) ensures the privacy of training data sample by clipping the gradients and adding DP noise (e.g., Gaussian mechanism) with the model updates. The introduction of DP noise enables the limited effect of one individual data on the trained model (and thus achieving the privacy guarantee). Additionally, another category of work is to add DP noise into the dataset following the method of DP synthetic data release and then train a model on such private data (Vaidya et al., 2013; Mohammady et al., 2020). Yet, the differential privacy-based learning schemes could cause great accuracy loss.

Alternatively, a private learning scheme called instance encoding (Huang et al., 2020a,b) has been proposed to obtain both privacy and utility for model training, which encodes the private data into "encrypted" data via mixup (Zhang et al., 2018a). While the privacy is claimed to be guaranteed by the encoding scheme, the data utility can be maintained by mixup scheme, only causing minor accuracy loss. However, it has been shown that such instance encoding scheme cannot provide strong privacy guarantee as cryptographic protocols (Mohassel and Rindal, 2018) or differential privacy (Dwork et al., 2014) against privacy attacks empiri-

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 172 - 180

cally (Carlini et al., 2020a). That is, well-designed privacy attacks (Carlini et al., 2020b; Xie and Hong, 2021) can break the instance encoding scheme to reconstruct the original data from the encoded data with high success rates. To address the privacy issue, we improve the TextHide with differential privacy and prove the improved scheme ensures theoretical privacy guarantee under the differential privacy framework. Besides, the experimental results validate the performance of proposed scheme.

2 Background & Related Work

2.1 TextHide

TextHide (Huang et al., 2020a) was proposed to protect the privacy of an individual's training data in the distributed learning by mixing up multiple raw training data. First, it utilizes a transformer encoder model, e.g., BERT (Devlin et al., 2019) as feature extractor to convert the raw training text into feature vectors. Second, TextHide designs an instance encoding method to mix up the original input feature vector with some randomly selected feature vectors from the training set (the corresponding data labels are also mixed up as well). Such mixed feature vectors with labels will be further utilized as training dataset for various down-stream language tasks, e.g., sentence classification (Cohan et al., 2019) and other natural language inference tasks (e.g., sentence similarity (Cer et al., 2017)).

More formally, we denote the language feature extractor as $\phi(\cdot)$, and the raw text data/label as x_i/y_i. Then we get the feature vector $v_i = \phi(x_i)$. Given the number of mix-up data points K, one private encoded vector $\widetilde{v}$ and corresponding mix-up label $\widetilde{y}$ can be computed as following:

$$\widetilde{v} = \sigma \circ \sum_{i=1}^{K} \lambda_i v_i, \quad \widetilde{y} = \sum_{i=1}^{K} \lambda_i y_i \qquad (1)$$

where λ_i is chosen uniformly at random such that $\sum_i^K \lambda_i = 1$, the sign-flipping mask $\sigma \in \{-1, 1\}^d$ is also chosen uniformly at random, and d denotes the dimension of the input vector. $\circ$ represents the Hardamard multiplication. For each training batch, K data points will be randomly selected to generate the private encoded vector per Equation 1. Besides, TextHide also sets another parameter m as the size of mask pool to improve the security. This formalizes the (m, K)-TextHide scheme (Algorithm 1 in (Huang et al., 2020a)). The privacy notion of TextHide was based on a *k-vector subset sum* (Abboud and Lewi, 2013) oracle with mixup, which would require $O(n^{k/2})$ efforts to break as original claim in (Huang et al., 2020a).

2.2 Privacy Attacks in ML

Privacy attacks against machine learning mainly consist of two categories: 1) membership inference attacks (MIA) (Shokri et al., 2017; Salem et al., 2018; Song and Mittal, 2021); 2) data reconstruction or extraction attacks. On the one hand, membership inference attacks (MIA) (Shokri et al., 2017; Song and Raghunathan, 2020; Hisamoto et al., 2020) have worked as state-of-the-art attack scheme due to its simpleness and effectiveness, where an attacker can determine whether a data point was used to train the ML model or not. Such MIAs have been commonly used for auditing training dataset privacy (Carlini et al., 2021).

On the other hand, as a stronger attack primitive, data reconstruction attacks (Fredrikson et al., 2015; Wu et al., 2016; Zhu et al., 2019; Carlini et al., 2020a) usually refer to the attacks that could utilize auxiliary information (e.g., background knowledge) and counter measures to reconstruct or extract the original private data. For example, model inversion attacks (Song and Raghunathan, 2020) or data extraction by memorization (Carlini et al., 2020c) could extract private information of training dataset by querying the target model without access to dataset. Another example is that the attacker can utilize gradients to recover data (Zhu et al., 2019; Geiping et al., 2020).

2.3 Privacy-Enhancing Technologies (PETs)

As data privacy risks become an emerging issue, there have been a number of research works, namely, privacy-enhancing technologies (PETs) focusing on the data protection in the machine learning (Mohassel and Rindal, 2018; Chaudhuri et al., 2011), including the two main directions as following: 1) designing secure computation protocols with cryptographic building blocks to secure the data-in-use (Bonawitz et al., 2016; Mohassel and Zhang, 2017; Mohassel and Rindal, 2018), which could achieve "perfect" secrecy but bring both extra computational and communication costs; 2) improving the privacy of machine learning algorithm with differential privacy (Vaidya et al., 2013; Abadi et al., 2016). For example, a Naïve Bayes classifier can be trained by applying Laplace noise on the dataset by computing proper sensitivity (Vaidya et al., 2013), which will be further utilized to add

Laplace noise to satisfy DP notion. Another popular but different scheme, DP-SGD (Abadi et al., 2016) applies the Gaussian noise into the gradients of a single data sample during the model training, which aims to bound the influence of such one individual data sample under the paradigm of differential privacy. It is worth noting that there have been recent works in NLP (Kerrigan et al., 2020; Yu et al., 2021; Li et al., 2021; Dupuy et al., 2022), which aim to empirically train/fine-tune language models to satisfy DP notion. We will further discuss such related literature in Section 2.4.

Both categories of privacy-enhancing schemes above can provide provable privacy guarantee for the training data. However, the instance encoding scheme may not obtain such privacy guarantee. As mentioned earlier in Section 2.1, the instance encoding scheme (Huang et al., 2020a,b) was proposed to protect the training data's privacy by mixing up input data (Zhang et al., 2018a). The paper claims that such scheme can preserve data privacy while maintaining good data utility. However, recent data reconstruction attacks (Carlini et al., 2020a) have shown that instance encoding lacks provable privacy guarantee. That is, the "indistinguishability" definition of privately encoded data is rather spurious, which does not comply with the concept of indistinguishability in either cryptography or DP. For example, the security of asymmetric encryption scheme could be theoretically proven by a security game (defined as IND-CPA (Goldreich, 2009)) where no adversary can win the game with significantly greater probability than an adversary with random guessing. Similarly, differential privacy (Dwork et al., 2006b; Abadi et al., 2016) also presents the individual data with deniability that attacker cannot differentiate it with some probability bound. Considering that TextHide fails to provide such privacy guarantee, it can be broken by the carefully designed attacks and leak the private data (Carlini et al., 2020a; Xie and Hong, 2021).

In this work, we focus on integrating the instance encoding scheme with differential privacy to address the privacy risks of the instance encoding scheme presenting with privacy attacks, which would obtain provable privacy under the paradigm of differential privacy as shown in Section 4.

2.4 Differentially Private Learning in NLP

Differentially Private Stochastic Gradient Descent (DP-SGD) (Abadi et al., 2016) has been a gold standard for preserving data privacy in machine learning. There have been various DP-related works in the language domain (Hoory et al., 2021; Yu et al., 2021; Li et al., 2021; Mireshghallah et al., 2021; Anil et al., 2021; Dupuy et al., 2022). For example, public pretraining has been shown to be helpful for the downstream DP fine-tuning (Kerrigan et al., 2020). Hoory et al. (Hoory et al., 2021) pretained a differentially private BERT model with DP optimization and identified the existence of memory issues with large batch size for high performance. Dupuy et al. (Dupuy et al., 2022) have also proposed an efficient DP-SGD training for large transformer model with GPU architecture. Mireshghallah et al. (Mireshghallah et al., 2021) utilized the adversarial and privacy regularization to ensure uniform treatment of under-represented subgroups in language model training. However, the previous works usually struggle with greatly decreased performance as the added DP noise needs to be scaled with large model parameters (resulting in high noise levels).

Recently, Li et al. (Li et al., 2021) and Yu et al. (Yu et al., 2021) have both demonstrated that the large pre-trained language models can be effectively and efficiently fine-tuned for various downstream tasks with very few privacy leakage. For example, Yu et al. proposed to use ghost clipping to reduce the memory costs of gradient clipping in DP-SGD. Besides, they also showed that there is no explicit relationship between the dimensionality of gradient updates and private fine-tuning performance (Yu et al., 2021).

It is worth noting that our work is orthogonal to all the DP-SGD-based works above in language domain in two main folds. First, the threat models are different. Specifically, DP-SGD considers a trusted authority to train on the private dataset. It aims to convert the learning algorithm with differential privacy, and thus get the trained model to defend against a "weak" adversary for "distinguishing" data, e.g., membership inference attacks (Shokri et al., 2017). In this work, we consider a stronger attack based on the scenario of instance encoding, i.e., the attacker could have access to the instance encoded data and try to reconstruct the original data by reconstruction attacks (Carlini et al., 2020b).

Second, the privacy protection methods are different. To address the risk of data reconstruction attacks, we follow the notion of conventional data publishing with differential privacy, i.e., adding

noise on the training data directly (integrated in the instance encoding scheme) while DP-SGD is to add noise on the gradient updates during the learning process (Abadi et al., 2016).

3 Preliminaries of Differential Privacy

As one main category of privacy-enhancing technologies, differential privacy (DP) (Dwork et al., 2006b, 2014) has been widely used as a de facto standard notion in protecting individual's data privacy for data collection and analysis (Dwork and Smith, 2010), especially in machine learning applications (Vaidya et al., 2013; Abadi et al., 2016).

The principle of the differential privacy (Dwork et al., 2006b, 2014) is that an individual's data point x in one dataset D will not arouse significant change to the outcome of a randomized mechanism or algorithm applied to the D. Thus, the attacker cannot make difference with such a specific data point x by observing the outputs of D by the randomized mechanism, which thus provides deniability for the existence of x (ensuring data privacy).

Formally, to define individual's privacy, we first define the neighboring datasets, i.e., D, $D' \in \mathcal{D}$ are the neighbors if they only differs in one data point, denoted as $D \sim D'$. Then we define the DP notation as following:

Definition 1 (Differential Privacy (Dwork et al., 2006b, 2014)). *For any two neighboring datasets, $D, D' \in \mathcal{D}$, a randomized mechanism $\mathcal{M}$ is said to be (ϵ, δ)-differentially private if it satisfies the following equation:*

$$Pr(\mathcal{M}(D) \in \mathcal{O}) \leq e^{\epsilon}\, Pr(\mathcal{M}(D') \in \mathcal{O}) + \delta \quad (2)$$

where $\mathcal{O}$ denote all the events in the output space of $\mathcal{M}$. If $\delta = 0$, $\mathcal{M}$ is ϵ-differentially private.

In this work, we will utilize the Laplace and Gaussian mechanisms to guarantee (ϵ, δ)-DP.

The Laplace mechanism (Dwork et al., 2006b) adds the noise from Laplace distribution with mean zero and scale parameter b, denoted as $\text{Lap}(b)$ with density function $\frac{1}{2b}\exp^{\frac{-|x|}{b}}$. Formally, we have the following theorem:

Theorem 1 (Laplace Mechanism (Dwork et al., 2006b, 2014)). *Given any function $f : \mathcal{D} \to \mathbb{R}^d$, the Laplace mechanism is defined as $\mathcal{M}_L(D, f, \epsilon) = f(D) + N$, where N is the random noise drawn from Laplace distribution $Lap(\frac{\Delta f}{\epsilon})$, and Δf is ℓ_1 sensitivity. Laplace mechanism satisfies $(\epsilon, 0)$-DP.*

Theorem 2 (Gaussian Mechanism (Dwork et al., 2006a, 2014)). *Given any function $f : \mathcal{D} \to \mathbb{R}^d$, the Gaussian mechanism is defined as $\mathcal{M}_G(D, f, \epsilon) = f(D) + N$, where N is the random noise drawn from Gaussian Distribution $\mathcal{N}(0, \sigma^2 I_d)$ with $\sigma \geq \Delta f \sqrt{2\ln(1.25/\delta)}/\epsilon$. Δf is the ℓ_2 sensitivity of function f, i.e., $\ell_2 = \sup_{D \sim D'} ||f(D) - f(D')||_2$. Guassian mechanism satisfies (ϵ, δ)-DP.*

4 DP Instance Encoding

Given a training batch of data samples of size M $\mathcal{B} = \{(x_1, y_1), (x_2, y_2), \cdots, (x_i, y_i)\}, i \in [1, M]$, which is randomly sampled from the training set. TextHide will first encode every sample into a feature vector of dimension size d by a pretrained feature extractor $\phi(\cdot)$, i.e., $v_i = \phi(x_i)$. Then we can get the corresponding batch of encoded feature vectors $\mathcal{B}_e = \{(v_1, y_1), (v_2, y_2), \cdots, (v_N, y_N)\}$. For original instance encoding, TextHide would mixup such set of size k vectors to generate private encoded vectors as training data per Equation 1. To address the privacy issue, we apply the differential private mechanism to such mixup process. Algorithm 1 demonstrates the details.

Algorithm 1: DP Instance Encoding

Input: Batch of encoded vectors $\mathcal{B}_e$,
Number of mixed data samples k,
clip bound for encoder vectors C
DP Noise $\mathcal{M}$: Laplace, Gaussian
Output: Differentially private encoded vector set $\mathcal{B}_{dp}$ of size $|\mathcal{B}_{dp}|$

Initialize DP mechanism $\mathcal{M} = \{\mathcal{M}_L, \mathcal{M}_G\}$
Randomly sample K mixup coefficients: $\Sigma_i^K \lambda_i = 1, \lambda_i \in \mathcal{N}(0, I)$
`// Instance Encoding by mixup`
Randomly sample K data samples from $\mathcal{B}_e$
for $i \to 1$ *to* $|\mathcal{B}_e|$ **do**
    `// Clip Input Vector`
    $v_i \leftarrow v_i \cdot \min(1, \frac{C}{||v_i||_2})$
if $\mathcal{M}_G$ **then**
    $N \leftarrow^s \mathcal{N}(0, \sigma^2 I_d)$
else
    $N \leftarrow^s \frac{\epsilon}{4C}\exp^{\frac{-\epsilon|x|}{2C}}$
for $j \to 1$ *to* $|\mathcal{B}_{dp}|$ **do**
    $\widetilde{v}_j \leftarrow \sum_{i=1}^K \lambda_i v_i + N$
    $\widetilde{y}_j \leftarrow \sum_{i=1}^K \lambda_i y_i$
return $|\mathcal{B}_{dp}|$ private encoded data vectors

Theorem 3. *The DP Instance Encoding revised with Laplace noise satisfies $(\epsilon, 0)$-DP, where the added noise N_L is draw from Laplace distribution as following:*

$$N_L = \frac{\epsilon}{4C}\exp^{\frac{-\epsilon|x|}{2C}} \quad (3)$$

Proof. The proof complies with the original proof of Laplace mechanism (Dwork et al., 2006b, 2014). The instance encoding scheme with clipping works as the function f. The ℓ_1 sensitivity here is $2C$ since the maximum ℓ_1 norm difference of two vectors are $2C$ (viewed as a hyper-sphere of radius C). Then replacing Δf with $2C$ in Laplace distribution, we get the Equation 3. It has shown that adding Laplace noise sampled from Eq. 3 satisfies ϵ-DP (Dwork et al., 2006b), i.e., the DP instance encoding with $\mathcal{M}_L$ satisfies $(\epsilon, 0)$-DP. □

Theorem 4. *The DP Instance Encoding revised with Gaussian noise satisfies (ϵ, δ)-DP.*

Proof. Similar to the previous proof for Laplace, we choose the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with mean zero and standard deviation $\sigma^2 = (\frac{1+\sqrt{2\log(1/\delta)}}{\epsilon})^2 C^2$, where the ℓ_2 sensitivity is C. Note that the input vectors are multi-dimensional, and the noise added will be drawn independently from $\mathcal{M}_G$. Then we can derive that DP instance encoding with $\mathcal{M}_G$ satisfies (ϵ, δ)-DP. □

5 Experimental Evaluation

For experiments, we would like to evaluate both utility and privacy of the proposed scheme as the following: 1) utility of the private instance encoding scheme, i.e., the performance (accuracy) of model trained on the private dataset; 2) privacy guarantee of the scheme against reconstruction attacks, i.e., the attack success rate (the percentage of reconstructed private vectors).

5.1 Experimental Setup

Dataset. We consider the sentence classification task with two popular datasets: 1) Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) (about 8500 training samples) for acceptability; 2) Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) (about 67000 samples) for sentiment analysis.

Model Implementation. We use the pre-trained BERT model (Devlin et al., 2019) as the language feature extractor to generate the text representation vectors (the dimensionality d is 768). Note that TextHide will encode such representation vectors into the training vectors for downstream tasks. For downstream task training, we follow TextHide to choose a multilayer perceptron of hidden-layer size $(768, 768, 768)$ since we take TextHide as baseline.

Utility Evaluation. We will apply our scheme (including Gaussian and Laplace mechanism, denoted as "DP-IE Gaussian" and "DP-IE Laplace", respectively) and TextHide to the two datasets during training, and then report the model accuracy, respectively. In addition, we will also demonstrate the accuracy of the raw dataset (without any privacy protection scheme) for better utility comparison.

Privacy Evaluation. To fully evaluate the proposed DP instance encoding scheme, we also utilize a privacy reconstruction attack (Xie and Hong, 2021) on instance encoding scheme. Specifically, we first construct a set of private vectors generated by our proposed scheme and TextHide (as baseline), respectively. We report the final attack success rate (the percentage of reconstructed data vectors out of the original set) by implementing reconstruction attack on the generated vectors above.

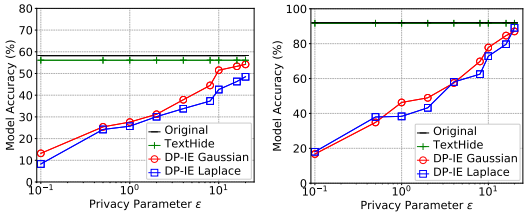
5.2 Utility Evaluation

For our proposed scheme, we set the privacy parameter $\epsilon = \{0.1, 1, 2, 4, 8, 10, 15, 20\}$. For Gaussian mechanism, we set δ to be 10^{-5}. Then we evaluate the model accuracy with varied ϵ for both Laplace and Gaussian mechanism on the two datasets as depicted above. For TextHide, we select $(m = 16, k = 4)$ as its own privacy parameters. We also evaluate the base case (without any privacy-protection scheme). We report the final model accuracy (the testing performance of trained model on the private dataset).

Figure 1 demonstrates the results. From the figure, we can observe that the model accuracy increases as the private parameter ϵ increases for both Gaussian and Laplace. This is reasonable since the privacy parameter ϵ of the DP schemes works as the privacy budget to determine the privacy-protection level for the dataset. That is, the larger the privacy budget, the smaller the noise added to the original data vectors (the privacy-protection would be weaker). As a result, the utility of the training set would not be affected too much. In addition, we can also observe that the model accuracy can approach the base case as ϵ increases, which will cause the compromise of privacy to some extent (as shown in the privacy evaluation).

5.3 Privacy Attack Evaluation

We follow the attack model setting (Carlini et al., 2020a; Xie and Hong, 2021) that the attacker could obtain the background knowledge of the private

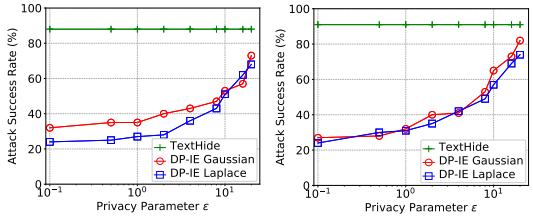


(a) CoLA (b) SST-2

Figure 1: Accuracy (learning utility) on the two datasets with DP-IE schemes.

dataset but be unaware of the specific data for training, which would utilize any auxiliary information to reconstruct the vectors (as a strong attack). We reproduce the attack scheme following the attack proposed in (Xie and Hong, 2021). More specifically, we randomly select 100 data points and generate 5000 encoded data by our DP schemes for each dataset, respectively. We measure the attack results with varying values of the privacy parameter $\epsilon = \{0.1, 1, 2, 4, 8, 10, 15, 20\}$ (referring to different levels for privacy-protection). For example, $\epsilon = 0.1$ is the strong protection and 20 is a weak protection. We repeat the same process for TextHide using the same privacy parameter as the previous utility evaluation.

Figure 2 demonstrates the final attack results. First, we can observe that the TextHide cannot ensure data privacy against privacy attacks, i.e., the privacy attack can recover around 85% of the original data vectors for both CoLA and SST-2 dataset. This also conforms to the previous works. Second, the results show that our proposed DP scheme can defend against such privacy attack from reconstructing the data. Take Figure 2(a) as an example, the overall attack success rate is lower than the baseline's. Besides, the attack success rate increases as the privacy parameter ϵ increases, which indicates that a higher privacy budget will lead weaker protection by differential privacy. Such results also validate the previous DP theorems. Again, it should be noted that DP cannot prevent leakage of the dataset completely. Instead, we would like to achieve a proper utility-privacy trade-off while applying differential privacy to the machine learning applications. For example, some privacy-sensitive applications, e.g., on-device input prediction, could require strong privacy guarantee while tolerating a fair utility loss. We can also improve our instance encoding scheme with other techniques, e.g., Federated Learning (Konečnỳ et al., 2016) or optimize the privacy budget to get a better utility accordingly.



(a) CoLA (b) SST-2

Figure 2: Attack success rate on the two datasets with DP-IE schemes.

6 Conclusion & Future Work

In this paper, we facilitate the instance encoding scheme with differential privacy. We have theoretically proven that the revised instance encoding with DP mechanism could provide good privacy guarantee under differential privacy framework. Experimental results have shown that the proposed differentially private scheme can obtain good utility for downstream learning tasks, e.g., text classification. Besides, we also evaluate the proposed DP scheme against privacy attacks and the results show that the scheme can ensure the privacy of dataset while presenting with attacks.

For the future work, we would like to further revise current DP instance encoding with another differential privacy notion, i.e., Rényi differential privacy (Mironov, 2017), which generalizes the concept of differential privacy based on the Rényi divergence. That is, revising the instance encoding scheme with Rényi DP would derive a tighter privacy bound and thus achieve better privacy-protection. Besides, another potential direction is to rescale the text vectors (generated by language feature extractor model) to a lower dimension vector by an extra MLP model or auto-encoder (Liou et al., 2014). We can utilize composition theorem (Dwork et al., 2014) in DP to theoretically find a better guarantee for various downstream tasks.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful feedback and suggestions. This work is partially supported by the NSF under the Grants No. CNS-2046335 and CNS-2034870, and the Cisco Research Award.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Amir Abboud and Kevin Lewi. 2013. Exact weight subgraphs and the k-sum conjecture. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12. Springer.

Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. 2021. Large-scale differentially private bert. *arXiv preprint arXiv:2108.01624*.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. 2016. Practical secure aggregation for federated learning on user-held data. *arXiv preprint arXiv:1611.04482*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

José Cabrero-Holgueras and Sergio Pastrana. 2021. Sok: Privacy-preserving computation techniques for deep learning. *Proceedings on Privacy Enhancing Technologies*, 2021(4):139–162.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2021. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*.

Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. 2020a. An attack on instahide: Is private learning possible with instance encoding? *arXiv preprint arXiv:2011.05315*.

Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. 2020b. Is private learning possible with instance encoding? *arXiv preprint arXiv:2011.05315*.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020c. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3).

M. Chen, B. Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Y. Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Z. Chen, Timothy Sohn, and Yonghui Wu. 2019. Gmail smart compose: Real-time assisted writing. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3693–3699, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Christophe Dupuy, Radhika Arava, Rahul Gupta, and Anna Rumshisky. 2022. An efficient dp-sgd mechanism for large scale nlu models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4118–4122. IEEE.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006a. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pages 486–503. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006b. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

Cynthia Dwork and Adam Smith. 2010. Differential privacy for statistics: What we know and what we

want to learn. *Journal of Privacy and Confidentiality*, 1(2).

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, CCS '15, page 1322–1333, New York, NY, USA. Association for Computing Machinery.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. Inverting gradients-how easy is it to break privacy in federated learning? *Advances in Neural Information Processing Systems*, 33:16937–16947.

Craig Gentry. 2009. Fully homomorphic encryption using ideal lattices. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 169–178.

Oded Goldreich. 2009. *Foundations of cryptography: volume 2, basic applications*. Cambridge university press.

Jian Guo, He He, Tong He, Leonard Lausen, Mu Li, Haibin Lin, Xingjian Shi, Chenguang Wang, Junyuan Xie, Sheng Zha, et al. 2020. Gluoncv and gluonnlp: deep learning in computer vision and natural language processing. *J. Mach. Learn. Res.*, 21(23):1–7.

Sorami Hisamoto, Matt Post, and Kevin Duh. 2020. Membership inference attacks on sequence-to-sequence models: Is my data in your machine translation system? *Transactions of the Association for Computational Linguistics*, 8:49–63.

Shlomo Hoory, Amir Feder, Avichai Tendler, Sofia Erell, Alon Peled-Cohen, Itay Laish, Hootan Nakhost, Uri Stemmer, Ayelet Benjamini, Avinatan Hassidim, and Yossi Matias. 2021. Learning and evaluating a differentially private pre-trained language model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1178–1189, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yangsibo Huang, Zhao Song, Danqi Chen, Kai Li, and Sanjeev Arora. 2020a. TextHide: Tackling data privacy in language understanding tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1368–1382, Online. Association for Computational Linguistics.

Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. 2020b. InstaHide: Instance-hiding schemes for private distributed learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4507–4518. PMLR.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 39–45, Online. Association for Computational Linguistics.

Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Eric Lehman, Sarthak Jain, Karl Pichotta, Yoav Goldberg, and Byron Wallace. 2021. Does BERT pretrained on clinical notes reveal sensitive data? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 946–959, Online. Association for Computational Linguistics.

Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*.

Cheng-Yuan Liou, Wei-Chen Cheng, Jiun-Wei Liou, and Daw-Ran Liou. 2014. Autoencoder for words. *Neurocomputing*, 139:84–96.

Fatemehsadat Mireshghallah, Huseyin Inan, Marcello Hasegawa, Victor Rühle, Taylor Berg-Kirkpatrick, and Robert Sim. 2021. Privacy regularization: Joint privacy-utility optimization in LanguageModels. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3799–3807, Online. Association for Computational Linguistics.

Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE.

Meisam Mohammady, Shangyu Xie, Yuan Hong, Mengyuan Zhang, Lingyu Wang, Makan Pourzandi, and Mourad Debbabi. 2020. R2dp: A universal and automated approach to optimizing the randomization mechanisms of differential privacy for utility metrics with no known optimal distributions. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 677–696, New York, NY, USA. Association for Computing Machinery.

Payman Mohassel and Peter Rindal. 2018. Aby3: A mixed protocol framework for machine learning. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 35–52.

Payman Mohassel and Yupeng Zhang. 2017. Secureml: A system for scalable privacy-preserving machine learning. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 19–38.

Pascal Paillier. 1999. Public-key cryptosystems based on composite degree residuosity classes. In *International conference on the theory and applications of cryptographic techniques*, pages 223–238. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2018. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 377–390.

Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2615–2632.

Jaideep Vaidya, Basit Shafiq, Anirban Basu, and Yuan Hong. 2013. Differentially private naive bayes classification. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 571–576. IEEE.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F Naughton. 2016. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pages 355–370. IEEE.

Shangyu Xie and Yuan Hong. 2021. Reconstruction attack on instance encoding for language understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2038–2044, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. 2021. Differentially private fine-tuning of language models. *arXiv preprint arXiv:2110.06500*.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018a. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zixing Zhang, Jürgen Geiger, Jouni Pohjalainen, Amr El-Desoky Mousa, Wenyu Jin, and Björn Schuller. 2018b. Deep learning for environmentally robust speech recognition: An overview of recent developments. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–28.

Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.