Privacy Protection in Transformer-based Neural Network

Jiaqi Lang*†, Linjing Li*‡, Weiyun Chen§, Daniel Zeng*‡
* The State Key Laboratory of Management and Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
† School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
‡ Shenzhen Artificial Intelligence and Data Science Institute(Longhua), Shenzhen, China
§ School of Management, Huazhong University of Science and Technology, Wuhan, China
{langjiaqi2017, linjing.li, dajun.zeng}@ia.ac.cn, chenweiyun@hust.edu.cn

***Abstract*—With the great success of neural networks, it is important to improve the information security of application systems based on them. This paper investigates a scenario where an attacker eavesdrops the intermediate representation computed by the encoder layers and tries to recover the private information of the input text. We propose a new metric to evaluate the encoder's ability to protect privacy and evaluate the Transformer-based encoder, which is the first privacy research conducted on Transformer-based neural networks. We also propose an adversarial training method to enhance the privacy of Transformer-based neural networks.**

***Keywords*—Privacy protection, Neural network, Transformer, Representation learning**

I. Introduction

Neural networks are considered to be indispensible for almost all intelligent applications. For the widely adopted encoder-decoder structure, the encoder first encodes raw input into intermediate representation, then task related decoder networks are employed on the representation to produce the final results. The encoder-decoder partition is powerful but poses great privacy risk when combined with cloud.

For the typical cloud-based architecture, the mobile devices carry encoder module which computes the representation of the raw input and send the results to the cloud. Consider the scenario illustrated by Fig.1, an attacker may steal the representation send by a mobile device and maneuver it to recover some personal information which are considered as privacy. Therefore, privacy information included in the representation must be taken into consideration when designing the neural network, especially, the encoder.

In the literature, previous privacy related works are focused mainly on LSTM-based encoder [1]–[3], as LSTM [4] was the most popular encoder in the domain of text information process. In recent years, the newly proposed Transformer [5] model achieves amazing results in translation tasks firstly, then its variant Bert (Bidirectional Encoder Representations from Transformers) [6] further presents the state-of-the-art results in a variety of NLP tasks. It is foreseeable that Transformer will be the most popular model equiped by mainstream mobile devices. Thus, in this paper, we foucs on the Transformer encoder, evaluate its ability on privacy protection and propose an adversarial training scheme to enhance it.

The contribution of this paper is threefold. First, we propose a new metric to evaluate different encoder's privacy attribute, this metric is based on the attacker's ability to recover the private information and the results of the main task of the decoder. Second, we examine the privacy attributes of Transformer networks, to the best of our knowledge, this is the first privacy related work on Transformer-based encoder. Third, we propose an adversarial training method to increase the difficulty of recovering the private information of the input from intermediate representation.

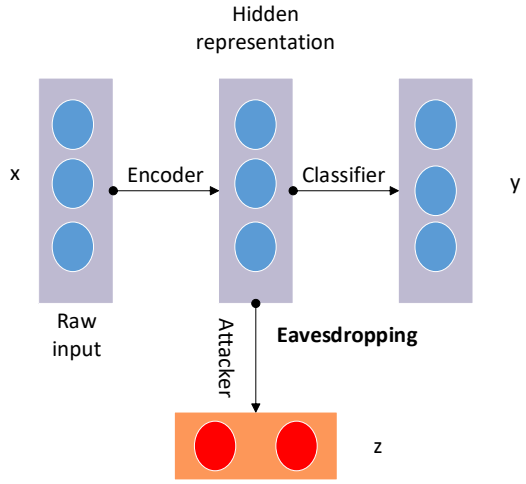

Fig. 1. According to the hidden representation, the classifier predicts a label y, while the attacker tries to recover the privacy z as much as possible.

II. Evaluation Framwork

Different encoders have different classification ability, it is impossible to directly compare the leakage of their privacy. In order to compare different encoder's privacy attribution, we propose a new metric:

$$Pr = \frac{X}{Y}, \tag{1}$$

where X is the main task accuracy, Y is the average of the accuracy of the attacker on the prediction of privacy. The bigger Pr is, the better the privacy of this encoder. In this

paper, text classification is taken as the main task, for the privacy, we consider gender and age.

When considering the encoder-decoder structure, private information in the raw text, can be further divided into two kinds: explicit and implicit. For example, demographic information about the author of a text can be predicted with above chance accuracy from the raw text [7], [8]. As to the intermediate representation, some private information may correlate with the main task, thus will be learned by the encoder and included into the representation. It is also possible that some private information be learned by the encoder accidentally. For the former, there is a tradeoff between the main task's accuracy and the encoder's privacy [1]. For the latter, adversarial training can be employed to enhance privacy. The overall Evaluation framework is consisted of the following steps:

- Train the main classifier on the (x, y) pairs and evaluate its accuracy;
- Freezing the encoder part, train the adversarial classifier on $(E(x), z)$ pairs and evaluate its accuracy, we treat this accuracy as a proxy for the privacy;
- Train a Generative Adversarial Networks (GAN [9]) on $(x, -z, y)$ pairs;
- Do the same as in the second step.

III. Building Blocks

This section briefly introduces the building blocks proposed in the overall evaluation framework.

A. Text Classification

We use Transformer-based encoder architectures, as shown in Fig.2, the encoder is composed of a stack of $N = 6$ identical layers. Each layer has two sub-layers, the first is a multi-head self-attention mechanism, the second is a position-wise fully connected feed-forward network [5]. In order to classify text, we first embed text sequences into dense vectors, then put those vectors into the encoder and get the hidden representation, a feedforward network with a softmax output activation was employed on it to predict the label. The network is trained to minimize the cross entropy of y labels:

$$Loss = \sum_{i=1}^{N} -logP(y_i|x_i), \tag{2}$$

where N is the number of the training dataset samples.

B. Attacker Classifier

An attacker tries to using the hidden representation to recover private information of the input. In its classifier, we use the frozen encoder and embedding layer of the main classifier as encoder layer, the encoder's output is fed as input to a new feedforward network. The result is treated as a proxy of this network's privacy. We use the training dataset to train the attacker, check the leakage of privacy with the development dataset, and apply the test dataset to generate the result. This network uses a sigmoid activation to compute the probabilities

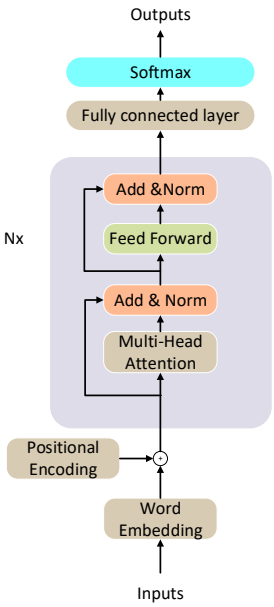

Fig. 2. The main task's architecture.

of each binary variable, it is trained on the (x_i, z_i)pairs to minimize the negative log-likelihood:

$$Loss = \sum_{i=1}^{N} -logP(z_i|x_i), \tag{3}$$

where N is the number of training samples. If the prediction accuracy of z is high, the information is easily leaked by the hidden representation. In fact, only one attacker cannot prove the hidden representation is robust to all the attackers, the best encoder is the one can defense any type of reconstruction to some extent. In this paper, we only experiment with a multilayer fully connected neural network, since it is powerful enough.

C. Adversarial Training

The attacker tries to recover the private information from the hidden representation which is similar with GAN in principle. However, the main classifier must achieve two goals: the main classification and preserving privacy. Thus, we augment the loss function to respond to the attacker:

$$Loss = -\sum_{i=1}^{N} logP(y_i|x_i) - \beta \sum_{i=1}^{N} logP(\neg z_i|x_i), \tag{4}$$

where β is a superparameter, which controls the relative importance between the main task and the privacy. The bigger β is, the model will preserve the privacy better.

IV. Experiments

We implement our model by Pytorch. For both classifier and attacker, we set epoch as 5, batch size as 32, and learning rate as 0.0002. We set the embedding dimension as 128. For adversarial training, we only set $\beta = 1$ and do not examine

other values. For Transformer encoder, we set $N = 6$ as the original paper.

A. Dataset

We use the Trustpilot(TP) dataset for text classification task. This corpus contains reviews associated with five scale sentiment score, gender and age information about the users. We use two sub-corpora corresponding to two areas (Denmark, UK). As in previous researches [10], we filter out examples containing both birth and gender of the users and bin the age of user into two categories: 'under 35' and 'over 45'. Each corpus is splited as training set (80%), development set (10%), and test set (10%), as shown in Table I.

TABLE I
STATISTICS OF THE DATASET.

Dataset	Train	Dev	Test
TP Denmark	82193	10274	10274
TP UK	48647	6080	6080

B. Results

We first conduct an experiment without adverserial training to compare LSTM and Transformer, the result is shown in Table II. All values reported in this and the following tables are accuracies apart from Pr. The result on LSTM is cited from [1] directly. We find that, in most cases, more private information can be recovered in LSTM-based network, which means Transformer is more secure. But LSTM is better when the accuracy of the main task is considered. The reason is that Transformer model has a huge number of parameters, it cannot produce comparable result in a small dataset compared with LSTM without using pretrain initialization. As to the new metric Pr we proposed, the result of main task and Pr values are shown in Table III. We find that, the values of Pr are similar between different encoders, due to the reason mentioned above, Transformer's main task accuracy is not the best as it can get. After adding adversarial training to the Transformer encoder-based network, the results shown in Table IV indicate that the effect of adversarial training is not obvious.

TABLE II
LSTM VS. TRANSFORMER.

Dataset	Most frequent class		LSTM		Transformer	
	Gender	Age	Gender	Age	Gender	Age
TP Denmark	61.6	58.4	62.0	63.4	62.6	62.7
TP UK	58.8	56.7	59.9	61.8	59.4	56.7

V. CONCLUSIONS

In this paper, we proposed a new metric Pr to gauge the privacy protection ability of encoders and examined the Transformer encoder. We also proposed a adversarial training scheme to enhance the privacy of neural networks. In the future, we shall consider more effective adversarial training methods and try to improve the main task accuracy of the Transformer encoder-based network.

TABLE III
ACCURACY AND Pr WITHOUT ADVERSARIAL TRAINING.

Dataset	LSTM		Transformer	
	Accuracy	Pr	Accuracy	Pr
TP Denmark	82.3	**1.31**	80.3	1.28
TP UK	86.9	1.43	85.0	**1.46**

TABLE IV
RESULTS WITH ADVERSARIAL TRAINING.

Dataset	Main task	Gender	Age	Pr
TP Denmark	-0.670	-0.060	-0.260	+0.002
TP UK	+0.390	+0.200	+0.400	-0.001

ACKNOWLEDGMENT

This work was supported in part by the National Key Research and Development Program of China under Grant 2016QY02D0305 and 2017YFC0820105, the National Natural Science Foundation of China under Grants 71702181, 71621002, as well as the Key Research Program of the Chinese Academy of Sciences under Grant ZDRW-XH-2017-3. Linjing Li is the corresponding author.

REFERENCES

[1] M. Coavoux, S. Narayan, and S. B. Cohen, "Privacy-preserving neural representations of text," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1–10.

[2] Y. Li, T. Baldwin, and T. Cohn, "Towards robust and privacy-preserving text representations," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2018, pp. 25–30.

[3] Y. Elazar and Y. Goldberg, "Adversarial removal of demographic attributes from text data," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 11–21.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] S. Rosenthal and K. McKeown, "Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 763–772.

[8] D. Preoţiuc-Pietro, V. Lampos, and N. Aletras, "An analysis of the user occupational class through twitter content," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 1754–1764.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] D. Hovy, "Demographic factors improve classification performance," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, vol. 1, 2015, pp. 752–762.