Keywords Inference Attack Against Language Models

Shangwei Guo, Tianwei Zhang, Guowen Xu, Han Yu, Tao Xiang, and Yang Liu

***Abstract*—**

***Index Terms*—**

I. Introduction

With the booming of intelligent healthcare, some hospitals tend to build an automatic pre-diagnosis system for more effective service flow. The system is expected to take the patient's description of the illness to predict treatment plan.

Also Sometimes airline companies survey their customers in order to improve their customer service. With the aid of advanced NLP techniques, large amounts of airline reviews in text form can be automatically processed for understanding customers' opinion. . As is widely recognized, utilizing the pre-trained language models for feature extraction can further improve the utility of many existing opinion mining systems.

II. Related Work

Recent studies have demonstrated that machine learning models are vulnerable to several privacy attacks. Both the text embedding encoded by the model and training set of the model may be attacked by adversaries. Therefore, adversaries would capture much sensitive information from this data.

A. Embedding inversion attacks

LMs are often used as embedding models to extract the embedding representation, low dimensional vector representation, of words for downstream tasks. Embedding inversion attack in the sense that the attack inverts sentence about the input from the model's output. [1] performs this attack both in white-box and black-box. In a white-box scenario, they assume that the adversary has access to the embedding model's parameters and architecture. It proposes a continuous relaxation of the sequential word input that allows more efficient optimization based on gradients to solve optimization problem. In the black-box scenario, they assume that the adversary only has query access to the embedding model and adversary observes the output embedding for a input plain text. The adversary learns an inversion model that takes a text embedding as input and outputs the set of words in the sequence.

T. Zhang, G. Xu, H. Yu, and Y. Liu are with School of Computer Science and Engineering, Nanyang Technological University, Singapore (email: {tianwei.zhang, guowen.xu, han.yu, and yangliu}@ntu.edu.sg).

B. Keyword inference attacks

The adversary in keyword inference attack is curious about the following predicate, whether certain keyword is contained in the unknown sentence. The keyword can be highly sensitive, which contains indicators for the adversary to further determine e.g., location, residence or illness history of the victim. Works in [1], [2] design attack against word embedding models.[2] train an attack model against publicly available pre-trained LMs to predict the sensitive information, given embedding vector representation of a word sequence obtained using the LM. It performs this attack in both white-box and black-box, In a white-box scenario, they assume that the adversary have a shadow corpus which is sampled from the same distribution as the unknown plain text so the adversary trains a binary classifier with the dataset to distinguish whether this keyword in plain text or not. In the black-box scenario, the adversaries has merely no prior knowledge of the plain text. So, the adversaries should obtain an external corpus from other public corpora. Due to the domain misalignment phenomenon, the attack's accuracy can sometimes be poor. Thus, they design an additional module called gradient reversal layer[3] is fundamental to learn domain-invariant representations and therefore help transfer the adversarial knowledge.

[1] performs a similar attack to infer the sensitive attributes of a word sequence using the embedding vector representation. They focus on the scenario of the adversary only having limited labeled data so as to closely match real scenarios where labeled sensitive data would be challenging to collect.

C. Membership inference attacks

The goal of membership inference is to infer whether a data point is in the training set of a given machine learning model. Membership inference attack (MIA) introduced by[4] shows that given black-box access to a classifier model, the confidence in model prediction can reveal whether a record belongs to the training data. They use shadow training technique to train an attack model to distinguish between a member vs non-member of the training data. In [5], they analyzes that MIA is closely connected to generalization where overfitted models are prone to the attacks. [1] develop simple and efficient thresholding attacks based on similarity scores as words or sentences in context used for training will be more similar to each other than them which were not used for training.

D. Data extraction attacks

Prior works have attempted to analyze leakage of sensitive information about the training data of LMs. Recent works[6],

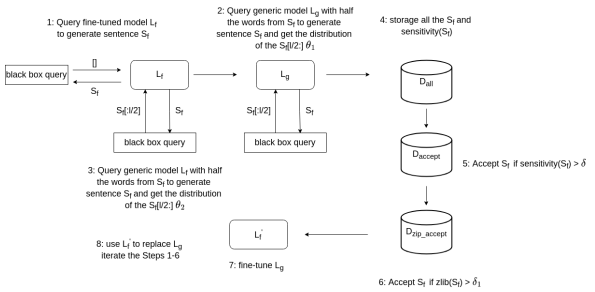

Fig. 1. Flow of the Dataset Reconstruction process

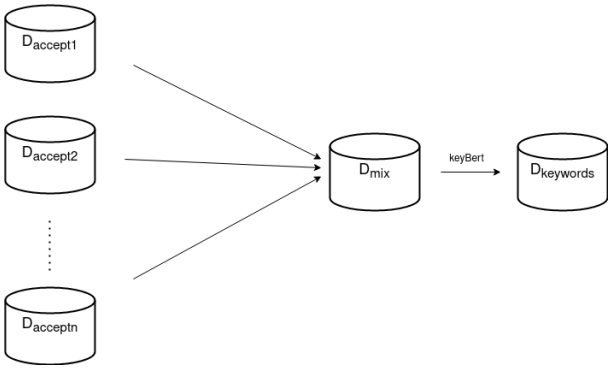

Fig. 2. Flow of the keywords extract process

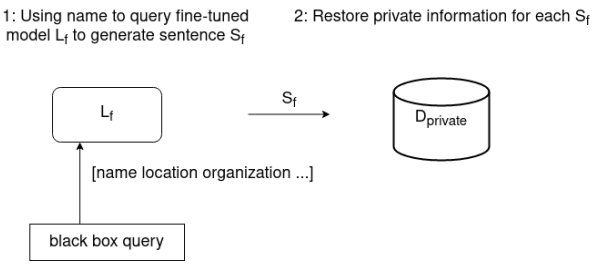

Fig. 3. Flow of the privacy information restore

[7], [8] have focused on training data extraction attack against LMs. [6] use perplexity of the generated sequences to choose the top 100 sequences as the extracted training data of publicly available pre-trained LMs. Work[7] analyses the information leakage in finetuned LMs. It proposes differential score metric to capture the difference between probabilities assigned to a word sequence by public and fine-tuned models. Finally, they rank the sequences to conclude that word sequences with higher differential score generally belong to the private data used to update the LM. Work [8] focuses on the reconstruction of the entire private dataset used to finetune the LM. The objective of the adversary is to model such behavioral changes between the generic and fine-tuned models in order to identify sentences belonging to the private dataset. The adversary's goal is to construct a representative dataset of the private dataset by iteratively constructing sentences belonging to the private dataset.Those data extraction attacks [6], [7], [8] are all black-box query access to the model and the LM returns the probability distribution over the vocabulary for the prediction of the next word.

III. Threat Model and Attack Taxonomy

IV. Domain Keywords Inference Attacks

A. Stage 1: Dataset Reconstruction

Our dataset reconstruction process exploits the observation from previous works that the behavioral change in snapshots of machine learning models can leak information about the training data used to update the model. The objective of the adversary is to model such behavioral changes between the generic(or fine-tuned model) and fine-tuned models in order to generate more sentences belonging to the private dataset. The adversary's goal is to construct a representative dataset $\mathbf{D}_{accept}$ by iteratively constructing sentences belonging to the private dataset.

We use Kullback-Leibler (KL) divergence between two distributions θ_1 and θ_2 as our sensetive($\mathcal{S}_f$) score. A lower KL divergence score indicates a higher similarity between the two distributions θ_1 and θ_2.

Fig. 1 shows the flow of the Dataset Reconstruction process.

B. Stage 2: Keywords Extraction

After reconstruct dataset, we combine all the dataset $\mathbf{D}_{accept}$ and use Bert to extract keywords sentence by sentence.

Fig. 2 shows the flow of the keywords extract process.

V. Private Keywords Inference Attacks

A. Stage 1: Dataset Reconstruction

Our dataset reconstruction process exploits the observation from previous works that the behavioral change in snapshots of machine learning models can leak information about the training data used to update the model. The objective of the adversary is to model such behavioral changes between the generic(or fine-tuned model) and fine-tuned models in order to generate more sentences belonging to the private dataset. The adversary's goal is to construct a representative dataset $\mathbf{D}_{accept}$ by iteratively constructing sentences belonging to the private dataset.

We use Kullback-Leibler (KL) divergence between two distributions θ_1 and θ_2 as our sensetive($\mathcal{S}_f$) score. A lower KL divergence score indicates a higher similarity between the two distributions θ_1 and θ_2.

B. Stage 2: Named Entity Extraction

After reconstruct dataset, we combine all the dataset $\mathbf{D}_{accept}$ and use Bert to extract Named Entity sentence by sentence.

C. Stage 3: Restore privacy information

After Named Entity Extraction, we use name which we extract as input for our target model and then we get the country, airline, date and much private information.

Fig. 3 shows the flow of the keywords extract process.

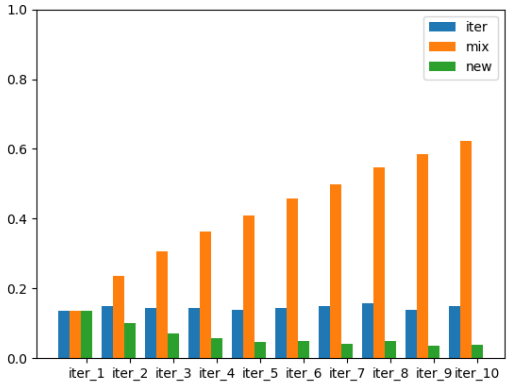

Fig. 4. keywords extraction recovery rate

VI. Experiments

A. Dataset

We use PubMed which has 200,000 abstracts of medical journals as the private dataset for finetuned a medical LM.

We also use airline review dataset from Skytrax which contains 41486 airline reviews as the private dataset for finetuned a opinion mining LM.

We also use movie review dataset from Rotten tomatoes which contains 1130000 movie reviews as the private dataset for finetuned a opinion mining LM.

B. Model

We use GPT-2 as our model, which is one of the largest LMs with 1.5 billion parameters. We download GPT-2 base model from the Internet and then fine-tune it on the private data for 200 epochs to obtain fine-tuned LMs.

C. Parameter Setting

The key parameter of our reconstruction process is the Sensitivity threshold δ , which is used to determine whether we accept or reject a sentence as a member or non-member of the private dataset respectively. After observing and fitting the sensitivity of all sentences in the D_{all}, we find it is a normal distribution. so we choose standard deviation plus mean value as our threshold.

Each iteration, We generate 10,000 sentences by querying the target model $Ł_f$ and compute their Sensitivity. Then, we use threshold to filter out the sentence we need.

D. Results

1) keywords extract: Fig. 4 shows the rate of recovery keywords is increasing with the number of iterations. The rate of new recovery Keywords is decreasing with the number of iterations.

Table. I and Table II shows in four settings, the rate of recovery keywords in each iter. Fixed generic model means we did not use finetuned model to replace our generic model. Use private data means each iteration we use part of our private data to finetuned our model. Eatch setting use all D_{accept} which we generate before to finetuned our model.

TABLE I
ITER-RECALL

Setting	iter1	iter2	iter3	iter4	iter5
fixed generic model	0.1211	0.1352	0.1184	0.1280	0.1300
our method	0.1211	0.1464	0.1561	0.1513	0.1601

TABLE II
MIX-RECALL

Setting	mix1	mix2	mix3	mix4	mix5
fixed generic model	0.1211	0.2078	0.2666	0.3200	0.3677
our method	0.1211	0.2150	0.2928	0.3536	0.4085

Table. III and Table IV shows in four settings, The proportion of the inferred keywords in the own total keywords.

Fig. 5 shows the threshold of the private training data which use to finetuned the generic model.

2) private attacks: Fig. 6 shows the airline private data we extract from the sentence which we generate.

Fig. 7 shows the rate of new recovery private information is decreasing with the number of iterations.

Fig. 8 shows the movie review private data we extract from the sentence which we generate.

Fig. 9 shows the rate of new recovery private information is decreasing with the number of iterations.

VII. Conclusion

Acknowledgements

References

[1] C. Song and A. Raghunathan, "Information leakage in embedding models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 377–390.

[2] X. Pan, M. Zhang, S. Ji, and M. Yang, "Privacy risks of general-purpose language models," in *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020, pp. 1314–1331.

[3] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand, "Domain-adversarial neural networks," *arXiv preprint arXiv:1412.4446*, 2014.

[4] 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)*. IEEE, 2017, pp. 3–18.

TABLE III
ITER-PRECISION

Setting	iter1	iter2	iter3	iter4	iter5
fixed generic model	0.5992	0.6238	0.5946	0.6109	0.5935
our method	0.5992	0.6342	0.6322	0.6361	0.6461

TABLE IV
MIX-PRECISION

Setting	mix1	mix2	mix3	mix4	mix5
fixed generic model	0.5992	0.6120	0.6065	0.6076	0.6047
our method	0.5992	0.6185	0.6236	0.6271	0.6312

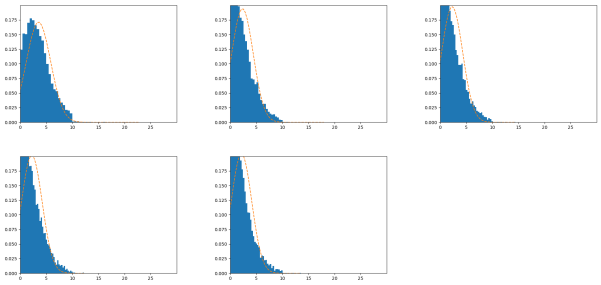

Fig. 5. threshold change

```
name,country,airline,date
S Dean Green,Australia,airasia-x,2012-04-24
T Carroll,United Kingdom,aegean-airlines,2014-08-26
D Christie,United Kingdom,american-airlines,2014-11-05
K Wong,Australia,airasia,2014-04-22
Matt A Arden,Australia,airasia-x,2012-04-24
E Johnson,United Kingdom,american-airlines,2014-10-05
S Wilson,United Kingdom,air-malta,2009-07-21
R Hopkins,United Kingdom,air-transat,2012-08-28
T Louportier Guyuca,Belgium,aeromexico,2014-04-21
C Andrews Lawrey,United Kingdom,air-transat,2013-07-29
B Kennedy,United Kingdom,american-airlines,2014-10-05
Heinland Jeff Dean Arraj,Indonesia,airasia,2014-04-22
```

Fig. 6. reconstruct data

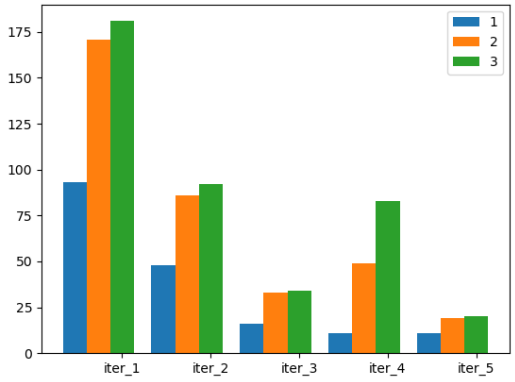

Fig. 7. airline review recovery rate

```
1 critic_name,publisher_name,review_date
2 John Beifuss,"Washington, D.C.)",2008-03-07
3 Tasha Robinson,NYCupid.com,2008-09-26
4 Simon Jordan,UK.com,2008-09-18
5 Emanuel Levy,athesteinberg.net,2008-03-07
6 Brian Orndorf,LA.com,2008-10-03
7 John Beifuss,"Washington, D.C.)",2008-03-07
8 Harvey S. Karten,Columbus,2005-10-25
9 John Beifuss,Washington DC),2008-09-26
10 Pete Hammond,Washington Blo's World,2008-03-07
11 Gina Carbone,coast Newspapers (NH/Maine),2008-03-07
12 Ross Anthony,Hollywood Report Card,2008-10-11
13 Stephen Whitty,New Haven Sentinel,2008-10-10
14 Brett McCracken,Oregon.com,2006-01-13
```

Fig. 8. reconstruct data

TABLE V
ITER-F-SCORE

Setting	iter1	iter2	iter3	iter4	iter5
fixed generic model	0.2014	0.2222	0.1975	0.2117	0.2133
our method	0.2014	0.2379	0.2503	0.2445	0.2566

TABLE VI
MIX-F-SCORE

Setting	mix1	mix2	mix3	mix4	mix5
fixed generic model	0.2014	0.3103	0.3704	0.4192	0.4573
our method	0.2014	0.3191	0.3985	0.4522	0.4960

[5] 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)*. IEEE, 2018, pp. 268–282.

[6] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson *et al.*, "Extracting training data from large language models," in *30th USENIX Security Symposium (USENIX Security 21)*, 2021, pp. 2633–2650.

[7] S. Zanella-Béguelin, L. Wutschitz, S. Tople, V. Rühle, A. Paverd, O. Ohrimenko, B. Köpf, and M. Brockschmidt, "Analyzing information leakage of updates to natural language models," in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020, pp. 363–375.

[8] R. Panchendrarajan and S. Bhoi, "Dataset reconstruction attack against language models," 2021.

TABLE VII
ITER-NEWWORD-RECALL

Setting	iter1	iter2	iter3	iter4	iter5
fixed generic model	0.1211	0.0868	0.0587	0.0534	0.0478
our method	0.1211	0.0939	0.0778	0.0608	0.0549

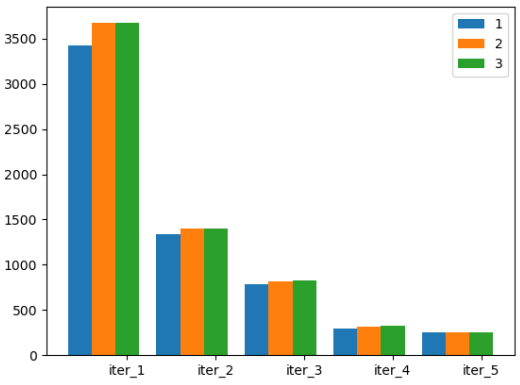

Fig. 9. movie review recovery rate