DEPN: Detecting and Editing Privacy Neurons in Pretrained Language Models

Xinwei Wu[1], Junzhuo Li[2], Minghui Xu[1], Weilong Dong[1], Shuangzhi Wu[4], Chao Bian[3,4], Deyi Xiong[1,2*]

[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]School of New Media and Communication, Tianjin University, Tianjin, China
[3]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[4]ByteDance Lark AI, Beijing, China
{wuxw2021,jzli,xuminghui,willowd,dyxiong}@tju.edu.cn,
wufurui@bytedance.com, bianc18@mails.tsinghua.edu.cn

- Inspired by model editing, the paper firstly proposed a neuron-based editing technique to the privacy issue of language models.
- The method is efficient to reduce the privacy issue while maintaining the model performance successfully.

Background

上海科技大学
ShanghaiTech University

- ➢ LLMs are pretrained on a huge amount of data;
- ➢ A variety of methods have been explored to attack LLMs for **training data extraction**

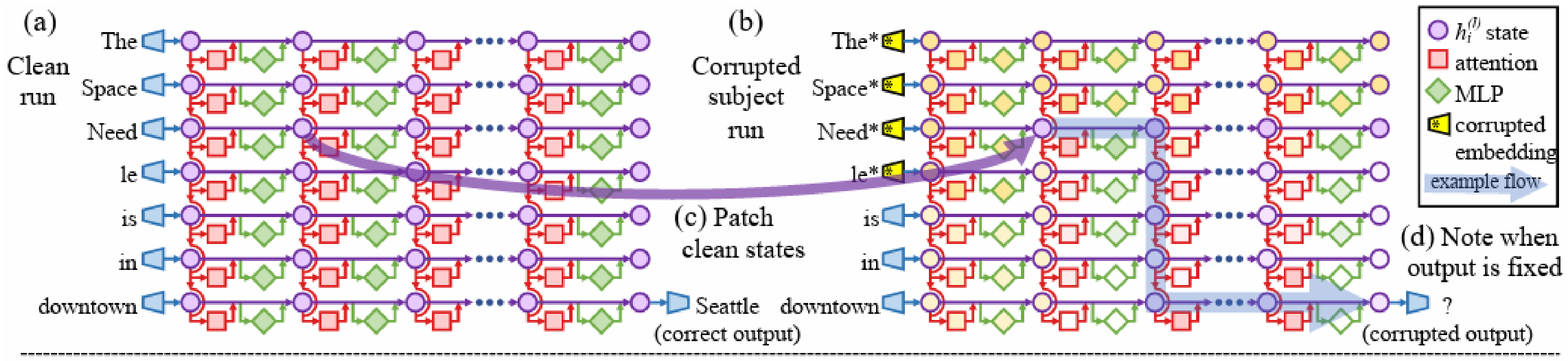
Existing methods to protect privacy:

- Data processing stage: removing sensitive information
- training stage: reducing the extent to which models memorize training data
- post-processing stage: retraining the model

立志成才 报国裕民

Background

上海科技大学
ShanghaiTech University



Factual knowledge is found to be stored in the feed-forward networks

Private information might be also encoded in specific neurons.

立志成才 报国裕民

DEPN

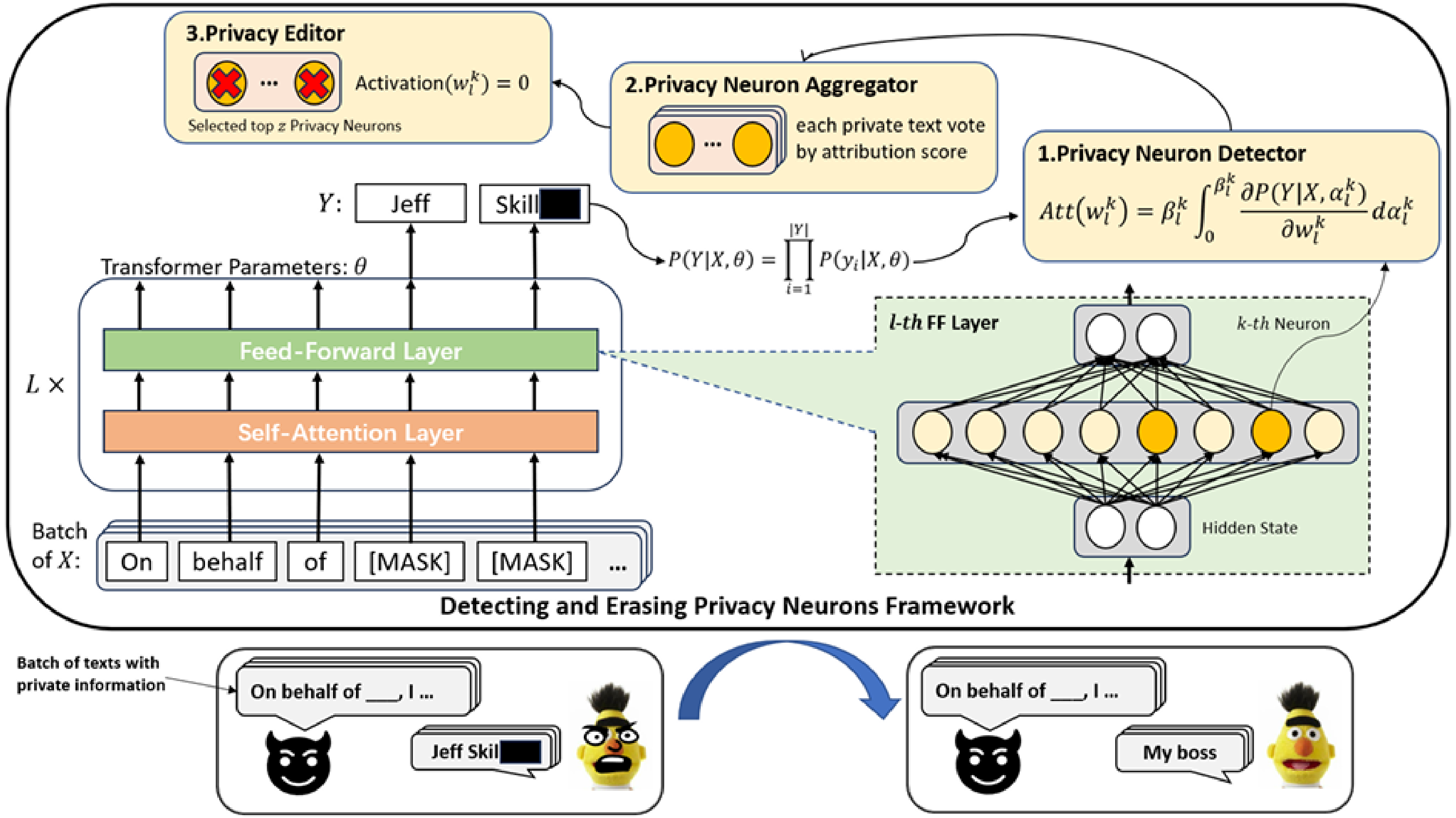


Figure 1: The diagram of DEPN. When a language model leaks privacy information, DEPN calculates privacy attribution scores using the Privacy Neuron Detector. It then selects the top z privacy neurons with the Privacy Neuron Aggregator and eliminates the model memorization of privacy information using the Privacy Editor.

1. Neuron detector
2. Neuron aggregator
3. Privacy editor

DEPN

Given a tuple $\boldsymbol{T} = \{\boldsymbol{X}, \boldsymbol{Y}\}$, let $\boldsymbol{Y} = \{y_1, ..., y_n\}$ be the sequence with private information, $\boldsymbol{X}$ be the the context of the sequence, $\boldsymbol{\theta}$ be the parameters of a language model. Given a context $\boldsymbol{X}$, the

$$P(\mathbf{Y}|\boldsymbol{X}, w_l^k) = \prod_{i=1}^{|\mathbf{Y}|} P(y_i|\boldsymbol{X}, w_l^k = \alpha_l^k) \quad (2)$$

Privacy Neuron Detector

$$\mathrm{Att}(w_l^k) = \beta_l^k \int_0^{\beta_l^k} \frac{\partial P(\mathbf{Y}|\boldsymbol{X}, \alpha_l^k)}{\partial w_l^k} \mathrm{d}\alpha_l^k \quad (3)$$

$$\downarrow$$

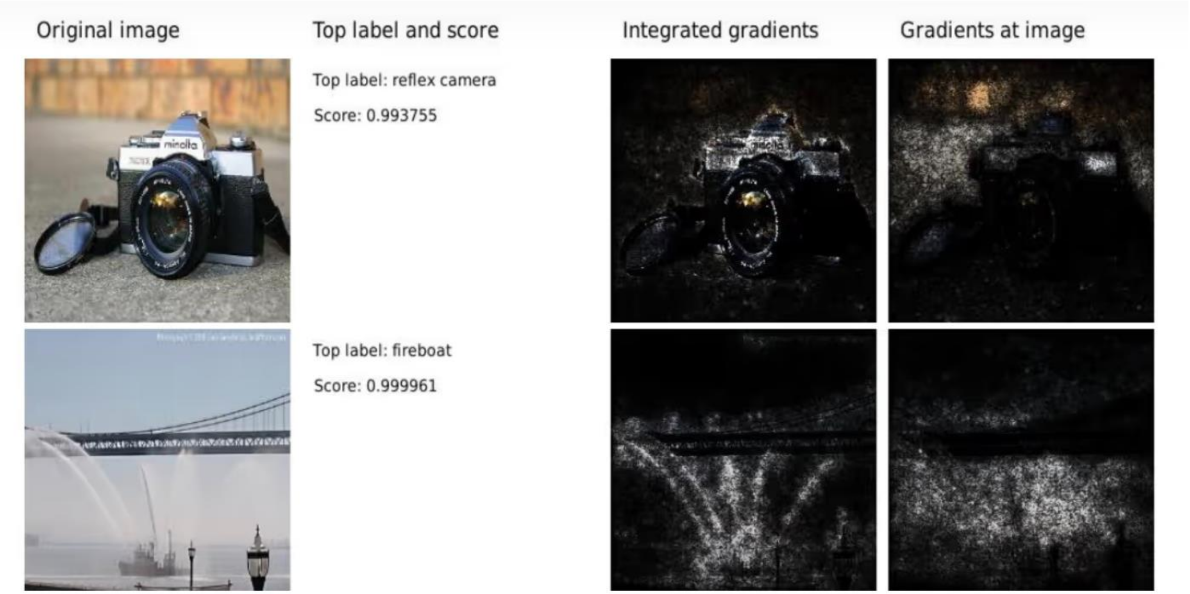
$$\mathrm{Att}(w_l^k) = \frac{\beta_l^k}{m} \sum_{j=1}^{m} \frac{\partial P(\mathbf{Y}|\boldsymbol{X}, \frac{j}{m}\beta_l^k)}{\partial w_l^k} \quad (4)$$

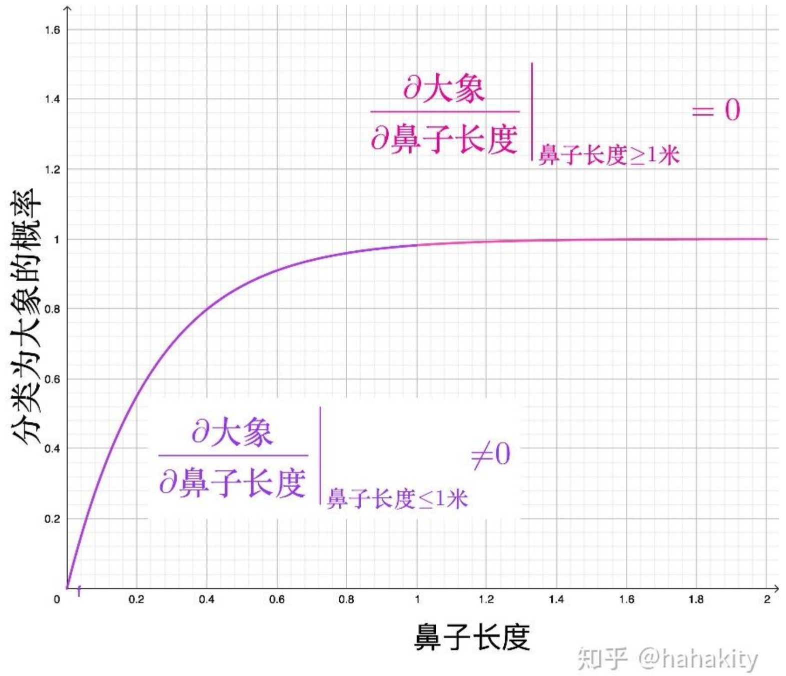
w_l^k :a neuron where l is the layer and k is the position

α_l^k :the value of w_l^k

β_l^k :the original value of α_l^k

DEPN





When a pixel or feature is enhanced to a certain extent, its contribution to network decision-making may reach saturation.

Experiment

Privacy definition: personal identity in formation, such as names, ID numbers, phone numbers and other related expressions.

Dataset: Enron, containing over 500000 emails

Private information:

1) **Names**: 20 unique names that are memorized by language models

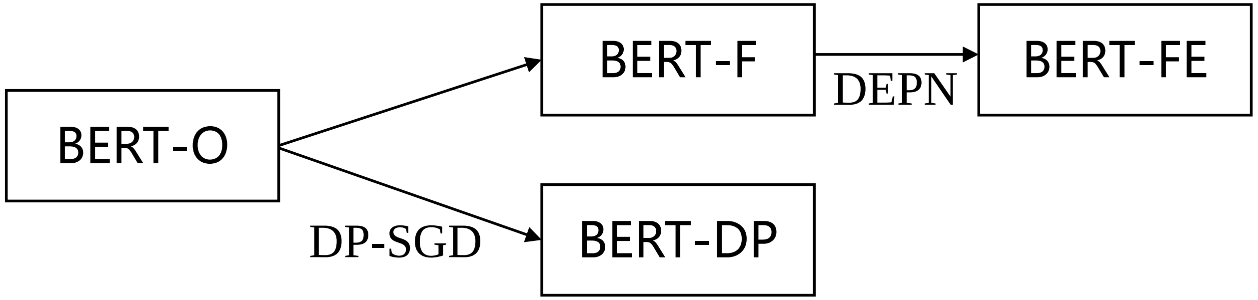
__ __ is a senior writer at ESPN.com

2) **Phone numbers**: 20 unique LM-memorized phone numbers

My phone number is 71385 _ _ _ _ _.

3) **Private texts**: 100 sentences that are not semantically overlapping with each other

Experiment



Metrics:

1) Vaild-PPL: the Perplexity of Masked Language Modeling task on the Enron validation dataset. It is used to test LLM's general performance.

3) Mean Reciprocal Rank(MRR): measure the model's memorization of names.

__ __ is a senior writer at ESPN.com

$$\frac{\sum_{i=1}^{|\boldsymbol{E}|}\frac{1}{Rank(e_i|Q)}}{|\boldsymbol{E}|}. \qquad (7)$$

4) Perplexity: measure the model's memorization of sentences.

Experiment

Main Results

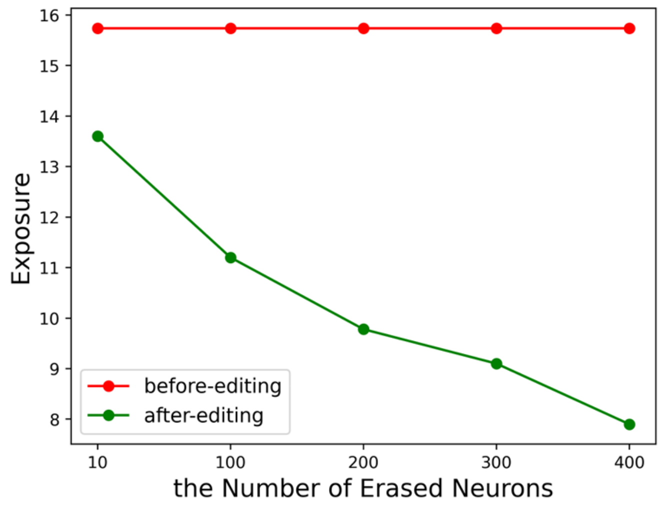
Privacy Type	Models	Time ↓	Valid-PPL ↓	Privacy Leakage Risk	
				Metric	Value
Phone Number	BERT-O	-	25.23	Exposure ↓	**1.58**
	BERT-F	100%	**3.07**		15.74
	BERT-FE	**2.4%**	3.11		9.78
	BERT-DP	181.4%	5.43		3.12
Name	BERT-O	-	25.23	MRR ↓	**0.87**
	BERT-F	100%	**3.07**		1.21
	BERT-FE	**4.4%**	3.11		1.15
	BERT-DP	181.4%	5.43		0.95
Random Text	BERT-O	-	25.23	PPL ↑	**10.05**
	BERT-F	100%	**3.07**		2.30
	BERT-FE	**4.6%**	3.11		3.67
	BERT-DP	181.4%	5.43		8.82

Table 1: Results of testing the risks of leaking private Phone Numbers, Names, and Texts on different baseline models, as well as the efficiency of protection. **Bold** and underlined results indicate the best and second best result, respectively. ↑: the higher the better. ↓: the lower the better.

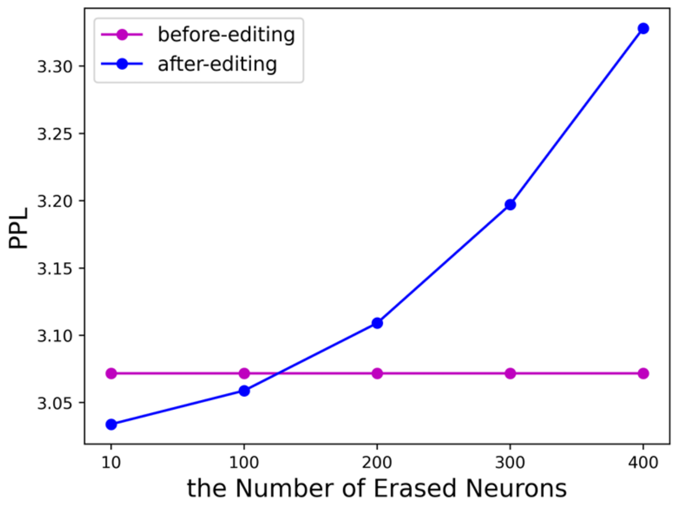
- DEPN's performance was almost similar to BERT-F and has fewest execution time cost;
- Regarding privacy leakage risk metrics, DEPN achieve the reduction of privacy leakage risk.

Experiment

Effect of the number of neurons edited



(a) Exposures with different number of edited neurons.



(b) Model performance with different number of edited neuron.

Figure 2: The performance of the model and the risk of privacy leakage with the change trend of the number of neurons edited.

- Increasing the number of edited neurons reduces the risk of privacy leakage in the model, but it also leads to a decrease in the model performance.

Experiment

Impact of Training Time on Privacy Neuron Distribution over Layers

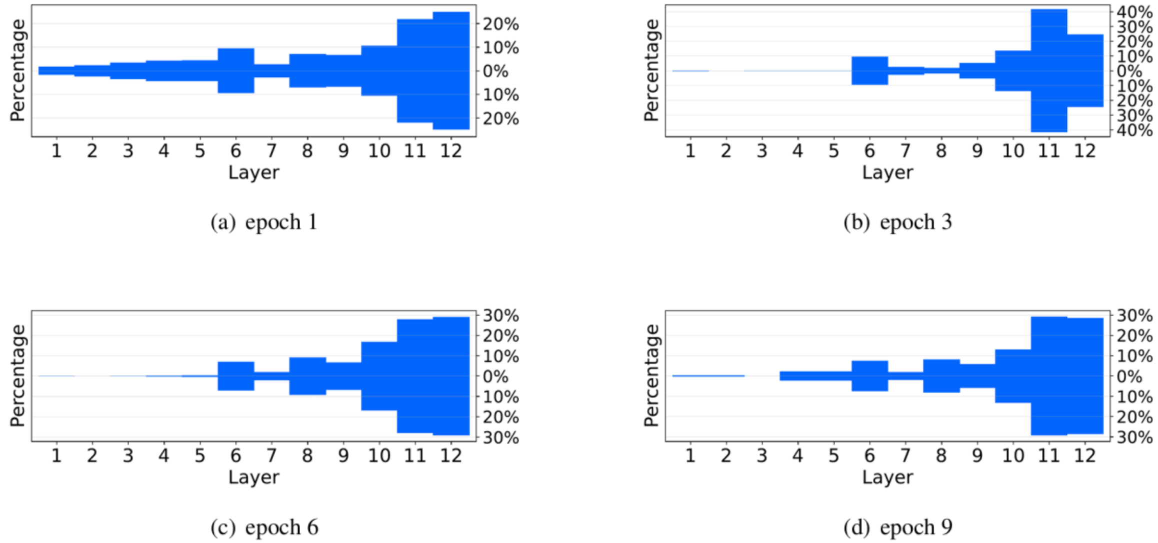


Figure 3: The distribution of privacy neurons in the bert-base model at different training epochs.

Models	# Edited Neurons	Time	Before Editing		After Editing		Reduction Rate
			Valid-PPL	Exposure	Valid-PPL	Exposure	
bert-small	100	0.26h	4.09	5.10	4.57	3.39	33.5%
bert-base	200	1.59h	3.07	15.74	3.11	9.78	37.86%
bert-large	400	7.66h	2.93	18.10	2.98	7.63	57.84%

Table 2: The privacy leakage risk reduction rate for models of different sizes.

- As the training time increases, privacy neurons of top layers increases.
- A model of a larger size has better performances on normal task. And DEPN demonstrates better performance on larger models.

Experiment

上海科技大学
ShanghaiTech University

Robustness analysis

Privacy Amount	# Edited Neurons	Time	Before Editing		After Editing	
			Valid-PPL	Exposure	Valid-PPL	Exposure
20	200	0.76h	3.07	15.74	3.11	9.78
100	500	1.59h	3.07	12.46	3.33	10.47
1000	2000	17.61h	3.07	8.32	3.81	8.03

Table 3: Analysis results on the cost-effectiveness of DEPN.

- There duction in privacy risks gradually diminishes as the amount of privacy increases.
- Privacy risk reduction across all prompts

Methods	Before Editing		After Editing	
	Valid-PPL	Exposure	Valid-PPL	Exposure
PND + Editing	3.07	15.54	3.11	**9.78**
KN + Editing	3.07	15.54	3.10	10.75
Random + Editing	3.07	15.54	3.07	12.48

Table 4: Effect of using different neuron localization methods on results.

Prompts	Original Exposure	Exposure
'Contact me at ***'	12.52	9.77 ↓
'Contact me at : ***'	11.20	9.40 ↓
'Contact me : ***'	12.50	9.68 ↓
'Call me at ***'	12.31	11.82 ↓
'My phone number is ***'	13.41	12.96 ↓
'You can call me at ***'	13.04	12.84 ↓

Table 5: Results with varying prompts during privacy attack. 'Contact me at ***' is the prefix to the private phone numbers in the training data, and the others are varying prompts used in inference.

立志成才 报国裕民

Thank you

上海科技大学
ShanghaiTech University

立志成才报国裕民