Training Text-to-Text Transformers with Privacy Guarantees

Natalia Ponomareva, Jasmijn Bastings, Sergei Vassilvitskii

(nponomareva, bastings, sergeiv)@google.com

---

### Introduction

**LMs are growing in size of data and parameters**

- Modern Transformer-based Large Language Models (LLMs) like T5, GPTs, etc.
  - Are pre-trained on large amounts of data
  - Can have up to billions of parameters
  - Often released as modifiable checkpoints that can be easily fine-tuned to your task given limited amount of data
  - Extremely good at various NLP tasks

**Pre-training data is not really "public"**

- It still likely contains private information (e.g. data erroneously released to the web, copyrighted text, etc.)
  - LLMs often exhibit episodic memory (e.g. memorizing the training data and outputting it verbatim) \[1\]
  - Preserved even after fine-tuning!
  - Embeddings can also contain private data \[3\]
  - This can expose owners of pre-trained and fine-tuned models to legal risks
  - And could also be bad for generalization

**Differential Privacy (DP) to the rescue**

- DP \[2\] provides robust theoretical guarantees on information leakage
- DP can potentially fix some of the "empirical" privacy concerns like training data extraction attacks (memorization)

---

### Methods

**Fully Private T5**

The pre-training data is used twice: for the subword vocabulary and for gradient updates.

We modify both parts of T5:
- Private SentencePiece: a modification of SentencePiece that adds noise to histogram of word counts (works for any SP algorithm)
- Private Training: Modified optimization using DP Adam \[4\]

---

### Results

**Does private (pre-) training hurt performance?**

- We look at both private tokenization and private training separately, as well as their combination
  - The private tokenizer serves as a regularizer on the pre-training task, improving pre-training accuracy.
  - While private training results in a pre-training performance drop, fine-tuning is hardly affected
  - Fully private model (private tokenizer-training) is even able to recover/improve pre-train accuracy, but is not significantly better on fine-tuning tasks
  - For some tasks fine-tuning performance can be better than that of a (non-private) baseline

**Does private training prevent memorization?**

- The way pre-training objective is formulated matters!
  - Span corruption is extremely robust to a (common definition of) memorization.
  - Prefix training exhibits a lot of memorization (the baseline outputs -2% training data verbatim)
  - Fully private models are able to mitigate the effect of memorization on commonly seen data:
    - for an ε of 6.23, Full-DP-T5 models exhibit 366x less memorization
    - even very large values of ε like 320 provide 15x improvement in memorization.

---

### Conclusion

**Summary**

- DP is a theoretically justified way of providing privacy guarantees for pretraining Large Language Models
- Using T5, a Transformer-based encoder-decoder, we investigated whether differential privacy (DP) would hurt utility (i.e., pre-training accuracy) and subsequent fine-tuning performance
- Fully private pre-training of Large Language Models can preserve good pre-training performance
- Can achieve comparable final task (fine-tuning) performance
- Can also mitigate empirical privacy attacks like training data extraction
- Private training is only 25% slower than training a baseline without DP
- It can be implemented efficiently using JAX's vmap operator.
- Code: bit.ly/private_text_transformers

---

### References

- \[1\] Carlini et al.. 2020. Extracting training data from large language models.
- \[2\] Dwork and Roth. 2014. The algorithmic foundations of differential privacy.
- \[3\] Thomas et al. 2020. Investigating the impact of pre-trained word embeddings on memorization in neural networks.
- \[4\] Abadi et al. 2016. Deep learning with differential privacy.
- \[5\] Lee et al. 2021. Deduplicating training data makes language models better.