Hush-hush Gradients: A Review of Differential Privacy for Deep Learning

Søren Winkel Holm

June 10, 2022

Introduction

The field of Deep Learning (DL) is for many subfields moving towards a setup where large multi-purpose foundation models are developed and trained at major companies or research instutitions, and then released for engineers to adapt to specific applications [Bom+21, pp. 3]. This application of the open-source principle to pre-trained models improves scientific reproduction ability [HO20, pp. 3] and technology accessibility [Bom+21, pp. 139]. One risk, however, is an adversarial actor exploiting a property of DL models: Parts of training data is generally recoverable from model weights [NSH19; Sho+17]. This might expose proprietary data or the private data of individuals as exemplified for Natural Language Processing (nlp) language models in Figure 1. As large-scale data sets are here to stay [Sun+17], algorithmic methods for improving the privacy of foundation models are required. The methods of Differential Privacy (DP) are suitable for this task and the relevant concepts, algorithms and problems will here be reviewed.

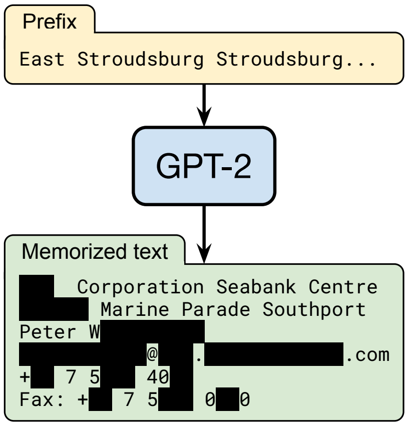
Prefix
East Stroudsburg Stroudsburg...
GPT-2
Memorized text
[redacted] Corporation Seabank Centre
[redacted] Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

Figure 1: The extraction attack performed on GPT-2 [Car+21, Fig. 1] (private data redacted).

Fundamental Concepts

Achieving DP corresponds to making a promise of hiding information about individuals when publishing quantitative patterns about groups [DR14, pp. 5]. This general problem is historically faced in releases of statistical data analyses by e.g. official organizations [Dal77; Wik22]. An algorithm is thus differentially private if a third party observer cannot extract individual information from its' computation. In this context, a Machine Learning (ML) model $f(x|w) = \hat{y} \approx y$ trained on a data set $\mathcal{D}$ exposes data patterns when either its' parametrization w or predictions $(x, \hat{y})$ are released.

Let $\mathcal{D}_i \in \mathbb{D}$ be a data set containing the private information of individual i and $\mathcal{D}_{\hat{i}} \in \mathbb{D}$ be identical except excluding this private information. For most approaches, the goal is to maximise DP by minimising the impact of individual data on computation which can be be quantified by measures such as ℓ_1-sensitivity [DR14, pp.31] Δg of a numeric statistic $g : \mathbb{D} \to \mathbb{R}^k$,

$$\Delta g = \max_{i,\hat{i}\in\mathbb{D}} ||g(x) - g(y)||_1. \qquad (1)$$

To achieve this, additive noise mechanisms can be used. For Δg, this can be achieved by adding Gaussian noise to outputs [DR14, p. D 3.3]

$$f(x) + (Y_1, \ldots, Y_k), Y_i \sim \mathcal{N}(0, \sigma^2_{\Delta g}) \quad (2)$$

The end goal of such DP mechanisms on random algorithms $\mathcal{F}(\mathcal{D})$ outputting $w \in \mathrm{Im}(\mathcal{F})$ with probability $p_{\mathcal{F}}(w|\mathcal{D})$, is to guarantee (ε, δ)-privacy [DR14, Def. 2.4] requiring $\forall S \in \mathrm{Im}(\mathcal{F})$ that

$$P\left(\mathcal{F}(\mathcal{D}_i) \in S\right) \leq \exp(\varepsilon) P\left(\mathcal{F}(\mathcal{D}_{\hat{i}}) \in S\right) + \delta. \qquad (3)$$

Thus, for $1 - \delta$ of the probability density over algorithmic randomness, it is promised that adding your private data to $\mathcal{D}$ does not raise your risk of harm by more than $\exp(\varepsilon)$ [DR14, pp. 21]. Often, $\delta = 0$, requring the stronger ε-privacy [Wik22]. For the Gaussian additive noise mechanism (2), (ε, δ)-privacy is achieved when $\sigma^2_{\Delta g} = \Delta g \ln(1/\delta)\varepsilon^{-1}$ [DR14, App. A].

State of the Art

Open Problems

References

[Bom+21] Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models". In: *ArXiv* abs/2108.07258 (2021).

[Car+21] Nicholas Carlini et al. "Extracting Training Data from Large Language Models". In: *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Aug. 2021, pp. 2633–2650. ISBN: 978-1-939133-24-3. URL: https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

[Dal77] Tore Dalenius. "Towards a methodology for statistical disclosure control". In: *Statistisk Tidsskrift* 15 (1977), pp. 429–444.

[DR14] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3–4 (Aug. 2014), pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: https://doi.org/10.1561/0400000042.

[HO20] Matthew Hartley and Tjelvar S.G. Olsson. "dtoolAI: Reproducibility for Deep Learning". In: *Patterns* 1.5 (2020), p. 100073. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2020.100073. URL: https://www.sciencedirect.com/science/article/pii/S2666389920300933.

[NSH19] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning". In: Mar. 2019. DOI: 10.1109/SP.2019.00065.

[Sho+17] R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. "Membership Inference Attacks Against Machine Learning Models". In: *2017 IEEE Symposium on Security and Privacy (SP)* (2017), pp. 3–18.

[Sun+17] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 843–852. DOI: 10.1109/ICCV.2017.97.

[Wik22] Wikipedia contributors. *Differential privacy — Wikipedia, The Free Encyclopedia*. https://en.wikipedia.org/w/index.php?title=Differential_privacy&oldid=1091066967. [Online; accessed 2-June-2022]. 2022.