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

Søren Winkel Holm

June 2, 2022

Introduction

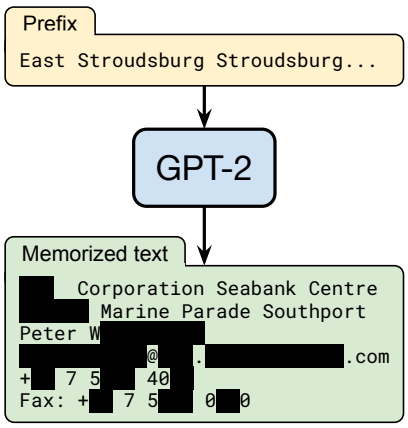

Figure 1: The extraction attack performed on GPT-2 [Car+21, Fig. 1] (private data redacted).

Fundamental Concepts

Differential Privacy (DP) is a field seeking to hide information about individuals when publishing quantitative patterns about groups. 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. The goal can has been robustly defined [DR14, pp. 5] as

> Differential privacy describes a promise, made by a data holder, or curator, to a data subject: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available."

A Machine Learning (ML) model $f(x|w) = \hat{y} \approx y$ trained on a data set $\mathcal{D}$ exposes data patterns when its' parametrization w or simply predictions $(x, \hat{y})$ are released, and these data patterns may be subject to DP concerns [DR14, Chap. 11].

State of the Art

Open Problems

References

[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.

[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.