Training Private Models That Know What They Don't Know

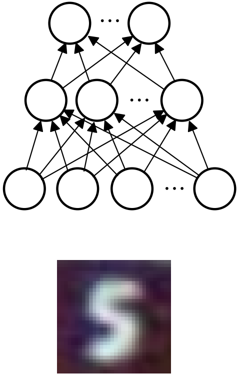
Stephan Rabanser[1,2] Anvith Thudi[1,2] Abhradeep Thakurta[3]
Krishnamurthy (Dj) Dvijotham[3] Nicolas Papernot[1,2]

[1]University of Toronto [2]Vector Institute [3]Google DeepMind

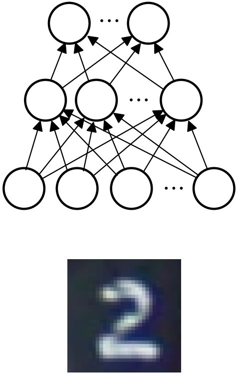
Correspondence to: stephan@cs.toronto.edu

Motivation: Input Sample Rejection

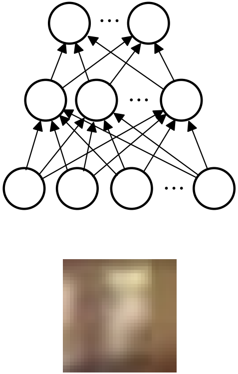
5



2



7? 1? 9?

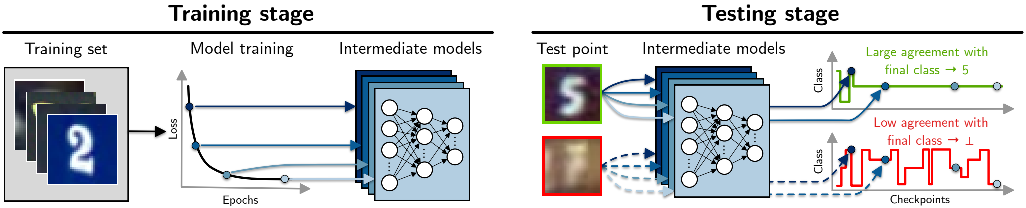


Selective Classification (SC) with Training Dynamics

Goal: Derive a selection function $g : \mathcal{X} \rightarrow \mathbb{R}$ which, given an acceptance threshold τ, determines whether a model $f : \mathcal{X} \rightarrow \mathcal{Y}$ should predict on a data point $\boldsymbol{x}$.

$$(f, g)(\boldsymbol{x}) = \begin{cases} f(\boldsymbol{x}) & g(\boldsymbol{x}) \leq \tau \\ \perp & \text{otherwise.} \end{cases}$$

New approach: **Selective Classification Training Dynamics** (SCTD)



Rabanser et al. "**Selective Classification via Neural Network Training Dynamics**." 2022.

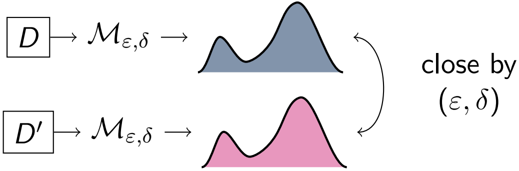
Differential Privacy

Definition: Differential Privacy

A randomized algorithm $\mathcal{M}$ *satisfies* (ε, δ) *differential privacy, if for any two datasets* $D, D' \subseteq \mathcal{D}$ *that differ in any one record and any set of outputs* S *the following inequality holds:*

$$\mathbb{P}\left[\mathcal{M}(D) \in S\right] \leq e^{\varepsilon}\mathbb{P}\left[\mathcal{M}(D') \in S\right] + \delta$$

- $\varepsilon \in \mathbb{R}_{+}$ specifies the privacy level.
- $\delta \in [0, 1]$ allows for a small violation of the bound.
- Most popular implementation for DP in DNNs is DP-SGD.



Impacts of SC on DP Guarantees: "SC $\longrightarrow$ DP"

No changes in DP guarantees.

Direct Optimization

Loss function / architecture modifications.

Post-Processing

Post-hoc modifications and training-time ensembles (`SCTD`).

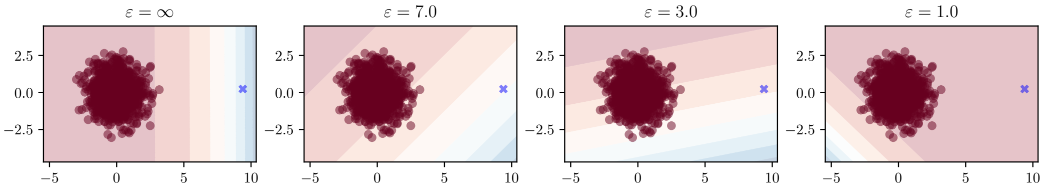
Worsened DP guarantee.

Advanced Sequential Composition

Methods iterating over the data multiple times, i.e., classical ensembling methods.

To maintain overall (ε, δ)-DP, each model needs to satisfy $\approx (\frac{\varepsilon}{\sqrt{M}}, \frac{\delta}{M})$-DP.

Impacts of DP on SC Performance: "DP $\longrightarrow$ SC"



Differential privacy degrades selective classification performance beyond a loss in overall utility!

Consistent SC Evaluations Under DP

Compare performance across SC approaches

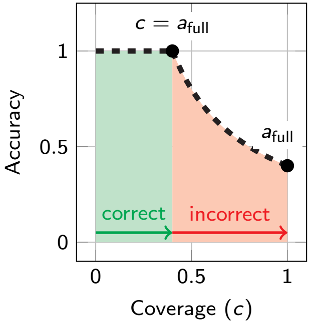
1. align SC methods to the same accuracy.
2. evaluate AUC metric for all SC methods.

But under DP:
accuracy $\overset{\approx}{\iff} \varepsilon$.

Training for less leads to expending less privacy budget.

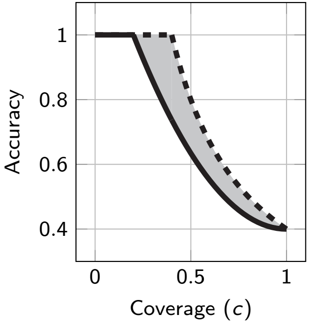
Upper bound on SC performance

$$\overline{\mathrm{acc}}(a_{\mathrm{full}}, c) = \begin{cases} 1 & 0 < c \leq a_{\mathrm{full}} \\ \frac{a_{\mathrm{full}}}{c} & a_{\mathrm{full}} < c < 1 \end{cases}$$



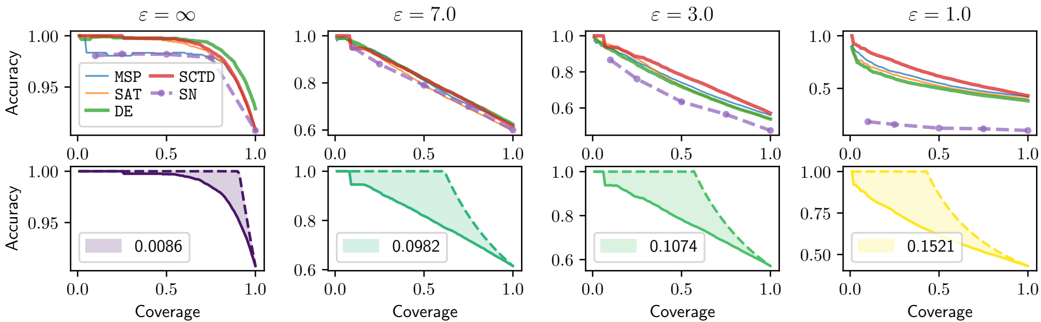
Our accuracy-normalized SC score

$$s(f,g) = \int_0^1 (\overline{\mathrm{acc}}(a_{\mathrm{full}}, c) - \mathrm{acc}_c(f,g))dc$$



UNIVERSITY OF TORONTO VECTOR INSTITUTE

Results on CIFAR-10



UNIVERSITY OF TORONTO VECTOR INSTITUTE

Training DP Models That Know What They Don't Know: Conclusion

- Analyzed how SC impacts DP guarantees and how DP impacts SC performance.
- Introduced a novel score to disentangle SC performance from baseline utility.
- SC performance degrades with stronger privacy (i.e. as $\varepsilon \to 0$).
- `SCTD` works best to quantify uncertainty under DP.

Stephan

Anvith

Abhradeep

Dj

Nicolas

Check out our paper: `https://openreview.net/forum?id=EgCjf1vjMB`