Differentially Private Model Publishing for Deep Learning

Lei Yu, Ling Liu, Calton Pu, Mehmet Emre Gursoy, Stacey Truex

Georgia Institute of Technology

2019 IEEE Symposium on Security and Privacy (SP)

Christian McDaniel

16 April 2019

Presentation Roadmap:

- Background:
  - Deep Learning and the need for Differential Privacy
- Prior Works and Justifications
  - Concentrated DP
  - Moments Accounting
- Focuses of this Paper

Background

Introduction

- **DLaaS**: Deep Learning is becoming widespread and readily available
  - e.g., Tensorflow & PyTorch | Cloud Computing & GPUs | Model Zoos & Transfer Learning | . . .

- The **complexity and flexibility** of Deep NN's mean they are potentially capable of encoding an individual's data or memorizing an exact data set

- DL models are vulnerable to **Adversarial Attack**
  - **Membership Attacks**: exploid black box access to the prediction API to infer individual instance membership
  - **Model Inversion Attacks**: Exploid prediction output and access to models to infer an input instance

Introduction

- Concentrated DP (CDP)
  - Generalization of DP targeted for algorithms with many calculations (ML); maintains strong privacy guarantees
  - Ensures the **expectation on the privacy loss** $\leq \mu$ and the **probability that the loss exceeds** μ **by** $t \cdot \tau$ **is bounded by** $-e^{\frac{-t^2}{2}}$
- Zero-Concentrated DP (z-CDP)
  - Concentrates the *privacy loss* around 0
  - Renyi Divergence $D_\alpha(\mathcal{A}(\mathcal{D})||\mathcal{A}(\mathcal{D}')) \leq \rho\alpha$
  - Single parameter ρ and its linear composition fit a privacy budget
  - Satisfies $(\epsilon, \delta) - DP$ and $\rho_i = \frac{1}{k}\rho$
  - Noise scale σ << noise scale under $(\epsilon, \delta) - DP$

Prior Work

Prior Work

- **2015**:
  - Multiple participants jointly train a model
  - Keep training data local and private; share sanitized parameters
- **2016**:
  - First DP for DL proposal
- **More recent**:
  - DP-SGD implemented; Gradient Clipping used to bound the influence of individual examples
    - Difficult to characterize max diff of model params over any 2

Current Paper

1) Refined Privacy Accountant

2) Dynamic Privacy Budget during DP-SGD

3) Privacy Preserving Parameter Selection

Current Paper

1) Refined Privacy Accountant

- Reshuffling (RF) vs. Random Sampling with Replacement (RS)
  - RS assumed by the Moments Accounting method; but RF used in batching techniques
  - RS underestimates the privacy loss
  - Distinct privacy loss for each
  - RF: lower cumulative privacy loss
    - mean (ρ_i) across disjoint partitions vs. composition on overlapping data sets
    - after E epochs, whole training satisfies ($\sum_{e=1}^{E} \rho_e$)-zCDP
  - RS:
    - CDP (useful for iterative algo's) does not capture the privacy-amplifying effect of random sampling
    - This paper relaxes CDP to $(\epsilon, \delta) - DP$
    - Composition used to determine total privacy budget

Current Paper

2) Dynamic Privacy Budget during DP-SGD

- Premise: As model accuracy converges, noise on gradients should decrease
  - Similarly applied to the learning rate
- Strategies:
  - public validation:
    - monitor a "publicly available data set" from the same sampling distribution
    - decrease noise scale whenever validation error stop improving
  - pre-defined schedule (decay function):
    - Time-Based Decay: $\sigma_t = \frac{\sigma_0}{1+kt}$
    - Exponential Decay: $\sigma_t = \sigma_0 e^{-kt}$
    - Step Decay: $\sigma_t = \sigma_0 * k^{[\frac{t}{period}]}$
    - Polynomial Decay: $\sigma_t = (\sigma_0 - \sigma_{end}) * (1 - \frac{t}{period})^k + \sigma_{end}$

Current Paper

3) Privacy Preserving Parameter Selection

- Use k-fold CV
- satisfies ϵ-DP and $\frac{1}{2}\epsilon^2$-zCDP

Current Paper

Algorithm 1. DP-SGD

- On each iteration, GD uses the average gradient on the loss formula *from a given* ***batch*** after bounding the per-example gradient via L2 norm
- Adds random noise via the Gaussian mechanism
- Updates the cumulative privacy vost c_t^{priv} depending on the batching method and terminates if c_t^{priv} > total privacy budget ρ_{total}
- Uses a schedule to dertiministically adjust the noise scale during training

Experimental Results

- Assessments:
  - Privacy Accounting Methods (RS v. RF)
  - Dynamic Privacy Budget Allocation
  - Hyper-parameter tuning under DP

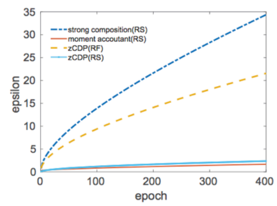
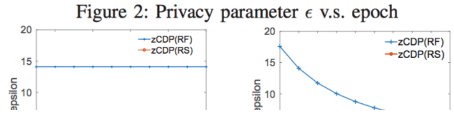
Experimental Results

Datasets	Size	Details	Model	Training
MNIST	60k Train 10k Test	28x28 grayscale images	60-dim PCA --> FF NN (single hidden, 1000ReLU units)	cross-entropy loss, 600 batch size; 0.98 acc after 100 epochs
Breast Cancer	560 Train 123 Test	11 attributes	NN with 2 hidden layers (10,20,10 ReLU units)	non-mini batch 0.96 acc after 800 epochs
CIFAR-10	40k Train 10k Test 10k Validation	10 classes, 6000 examples each	pre-trained with VGG16 CNN (Transfer Learning from non-private public dataset ImageNet)	only retrain the hidden layers with 1000 units 200 training epochs, batch size 200 0.64 training acc; 0.58 testing acc

Experimental Results

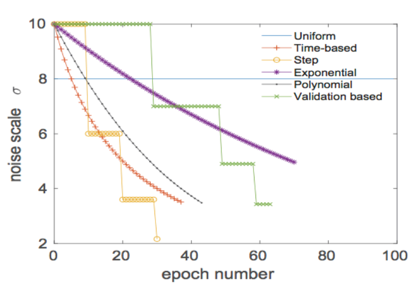
- Privacy Accounting Approaches:
  - RF and RS compared against strong composition (SC) and moments accounting (MA)

Experimental Results

Figures 2-3	–
 Figure 2: Privacy parameter ϵ v.s. epoch 	**Fig 2:** Privacy Loss growth across epochs + MA < RS << RF < SC --> RS and MA - tighter bound on p.l.; privacy amplification of RS --> increased uncertainty with RS + BUT, DL uses RF; MA underestimates real privacy loss b/c uses RS instead of RF **Fig 3a** $\sigma = 6$, varying q + sampling ratio $q = \frac{B}{N}$, B = batch size + RF invariant w.r.t q + RS p.l. increases with q

Experimental Results

- Evaluating the Dynamic Privacy Budget Allocation

Fig 4	Noise Scale
 Figure 4: The change of noise scale σ during training	+ Curves terminate at end due to depleted budget + n.s. intervals decrease over time for Validation-based

Experimental Results

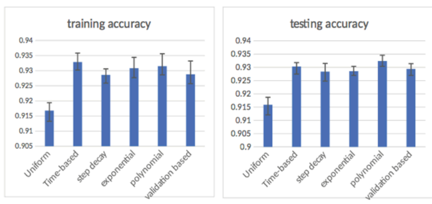
<table>
<tr><th>Fig 5-6</th><th>Decay functions</th></tr>
<tr><td>


Figure 5: The accuracy comparison of different schedules

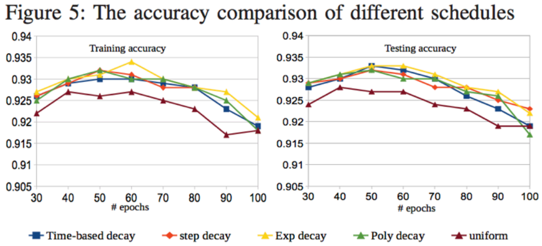


Figure 6: The accuracy under fixed training time
</td><td>

+ **Fig 5**: Accuracies when allowed to train until stopping parameter reached

+ **Fig 6**: Fixed #epochs; No clear winner amongst decay functions
</td></tr>
</table>

Experimental Results

- MNIST used for most hp tuning
  - validation-based scheduling; compare to uniform budget and non-private baseline
  - repeat all experiments 10x

Fig 7-8	Decay Rate k (how fast Noise Scale decays)

Experimental Results

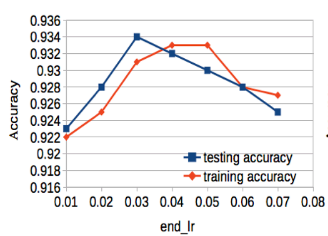
<table>
<tr><th>Fig 11</th><th>Number of Hidden Units (200-1600) / Layers (1-3)</th></tr>
<tr><td>


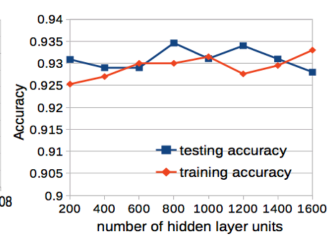
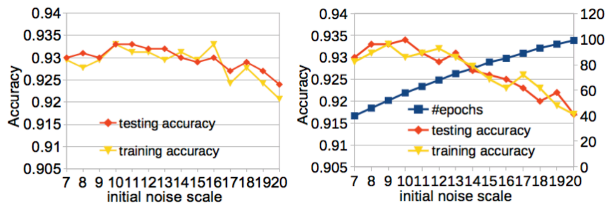
Figure 9: learning rate


Figure 10: hidden units
</td><td>
Fig 9. LR
+ Accuracy decreases when LR too small or large

Fig 10. Units
+ Increased # units = Increased sensitivity to gradient (more noise)
+ Changing units doesn't change accuracy
+ Shows the budget allocation scales w.r.t. size of NN
</td></tr>
</table>

Experimental Results

Fig 10	Noise Scale (again)
 (a) fixed training time (b) fixed decay rate	+ Accuracy is more sensitive to noise scale than to training time

Experimental Results

Experimental Results

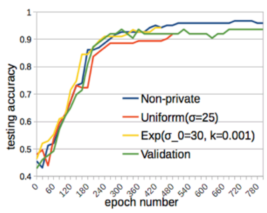
<table>
<tr><th>Fig 14-15</th><th>Other Datasets</th></tr>
<tr><td>


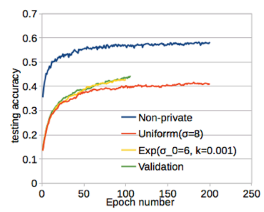
Figure 14: Accuracy (Cancer)


Figure 15: Accuracy (Cifar-10)
</td><td>
+Breast Cancer
Exponential Decay has better accuracy
Dynamic budget decreases the gap between non-private and uniform budget by 70%

CIFAR-10
Validation-based outperforms Exponential decay
</td></tr>
</table>

Conclusions and Discussion