DeepMind

Privacy in Image Classification Models

Informed Attacks and Practical Defences

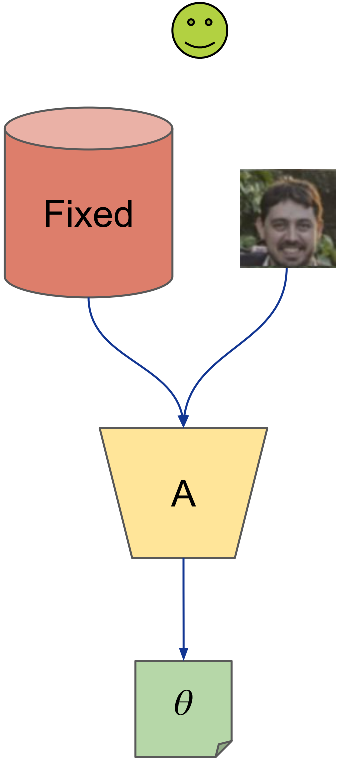
Borja Balle

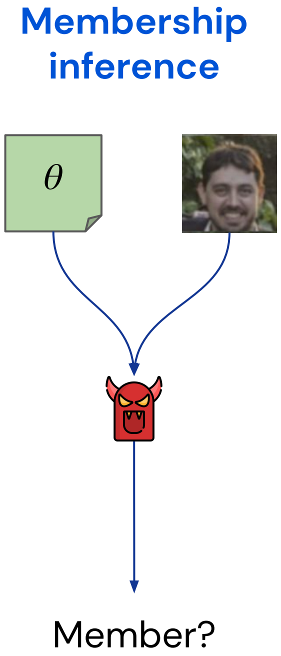
Privacy-Preserving AI @ AAAI

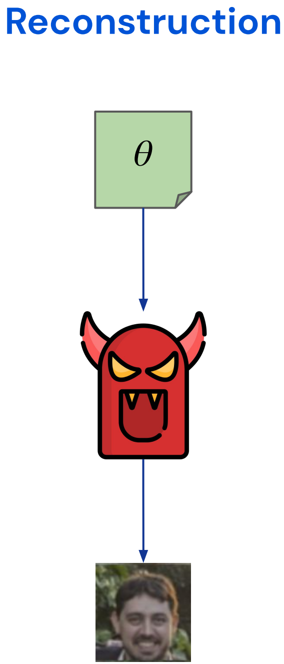
February 13, 2023

Balle, Cherubin, Hayes. "**Reconstructing Training Data with Informed Adversaries**." IEEE Security & Privacy (2022). [arxiv:2201.04845]
De, Berrada, Hayes, Smith, Balle. **"Unlocking High-Accuracy Differentially Private Image Classification through Scale."** Pre-print (2022). [arxiv:2204.13650]

Spectrum of Privacy Attacks

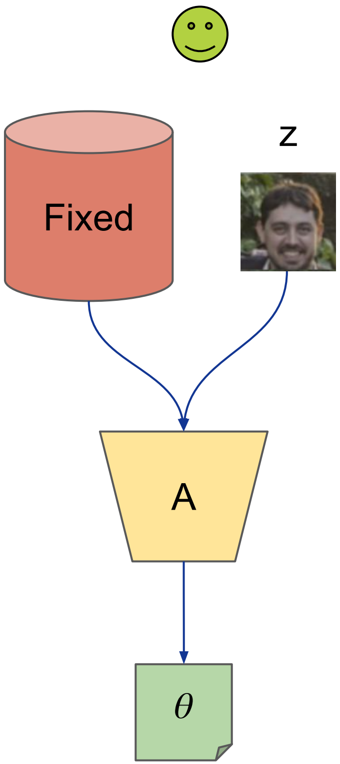


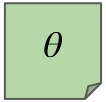




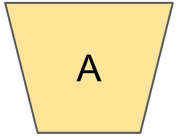
Threat Model: Informed Adversary

Blum et al.
Practical privacy: the SuLQ framework
PODS (2005)

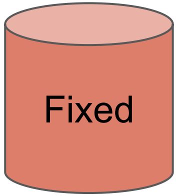




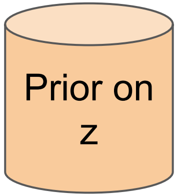
Adversary knows parameters of released model



Adversary knows training algorithm used by model developer

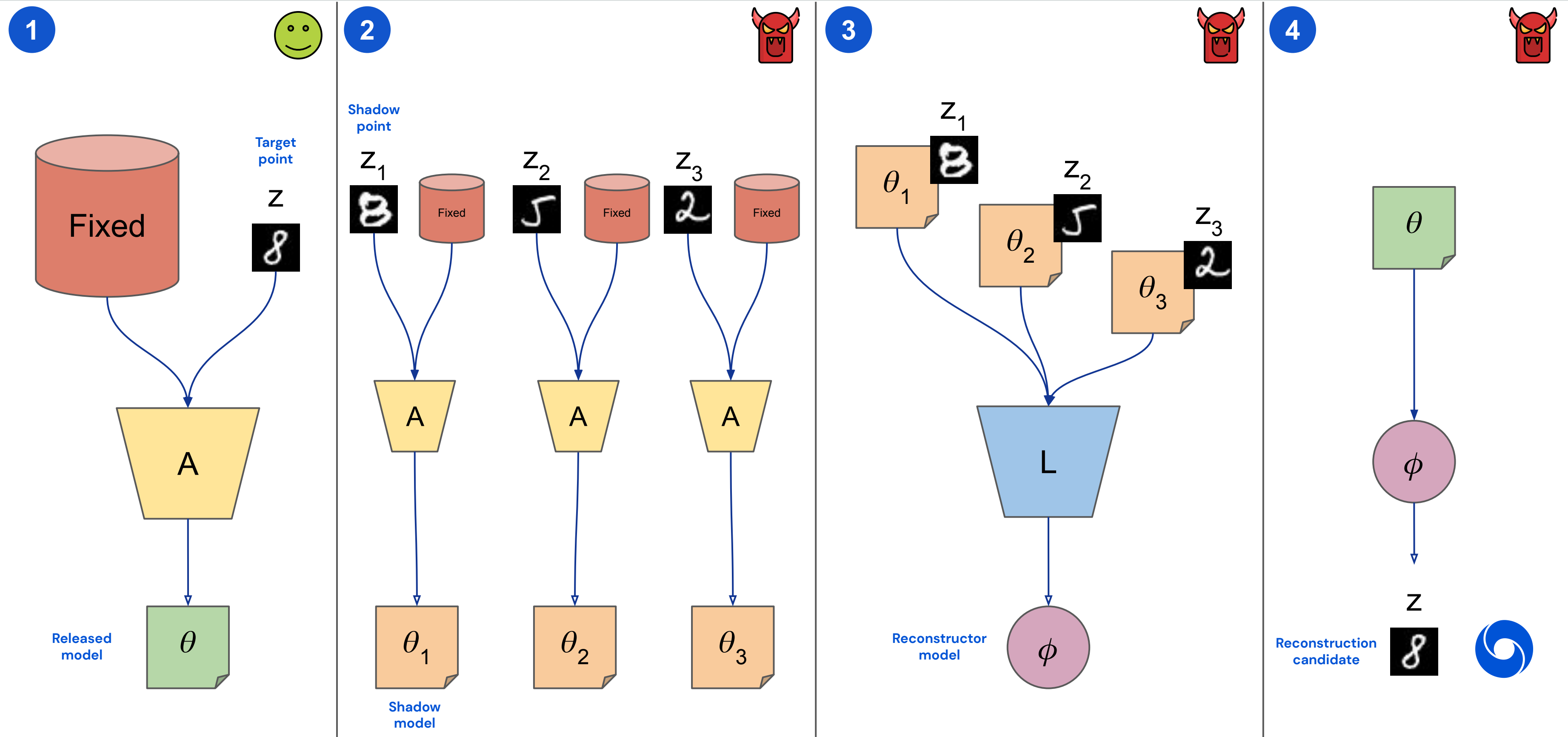


Adversary knows all data except one point

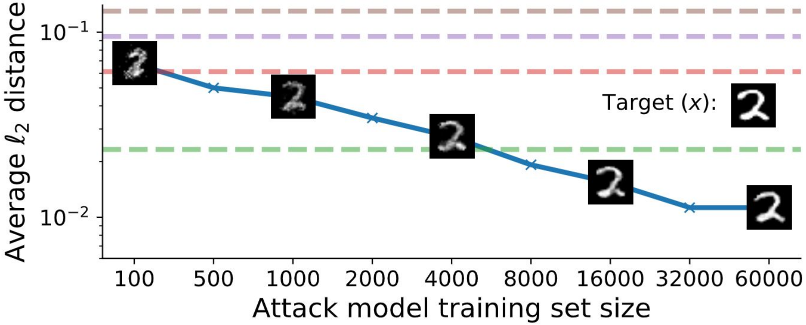


Adversary has prior knowledge of z (eg. samples from same distribution)

A Learning-Based Reconstruction Attack



Successful Reconstructions



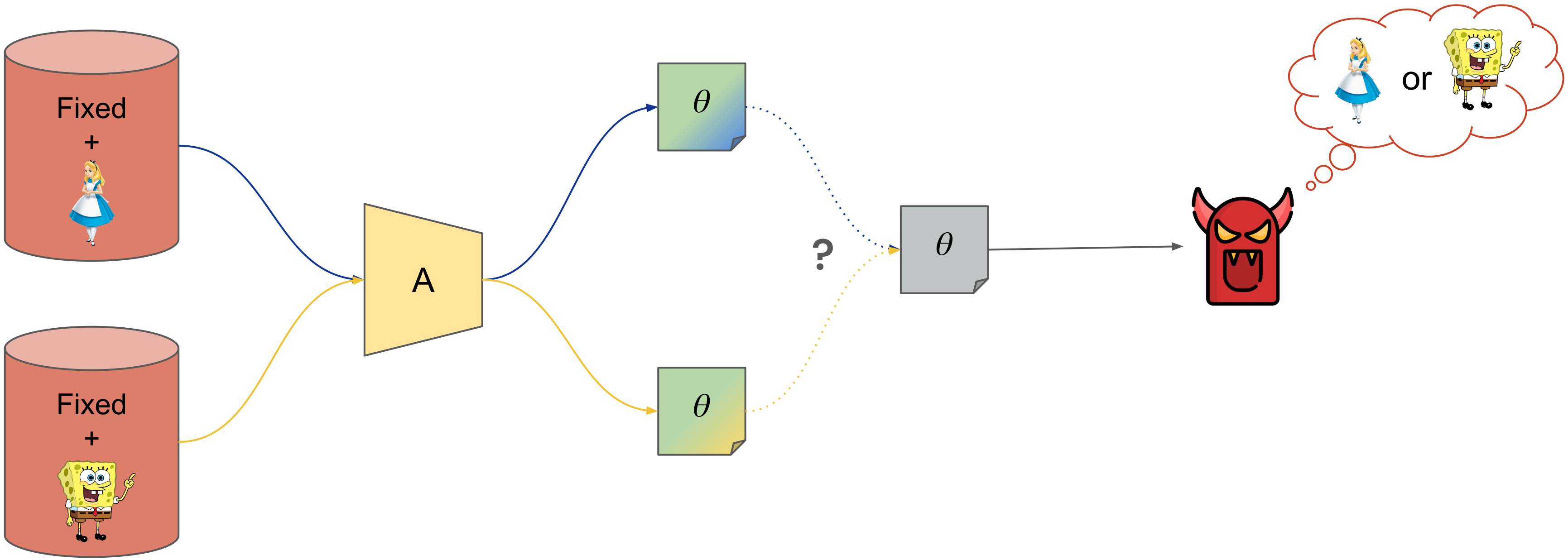
Key Takeaways

- Successfully scaled attack on fully connected and convolutional networks on MNIST and CIFAR-10 with up to 100K parameters
- Reconstructions improve as target model becomes larger
- Attack is robust to changes in training procedure (optimizer, hyper-parameters, etc)
- Reconstruction works even under mini-batch randomness
- Success is not a byproduct of overfitting
- Full access to model parameters is not necessary

Mitigations are required to safely deploy models trained on private data

Differential Privacy (In a Nutshell)

Dwork et al.
Calibrating Noise to Sensitivity in Private Data Analysis
TCC (2006)



$\forall$ Fixed, :

$$\log \frac{\mathbb{P}[\theta|\text{Alice}]}{\mathbb{P}[\theta|\text{Bob}]} \leq \varepsilon$$

Private Deep Learning with DP-SGD

Abadi et al.
Deep learning with differential privacy
CCS (2016)

$$w^{(t+1)} = w^{(t)} - \eta_t \boxed{\left(\frac{1}{|B|} \sum_{i \in B} \mathtt{clip}_C\left(\nabla l_i(w^{(t)})\right) + \frac{\sigma C}{|B|} \xi \right)}$$

Privatized mini-batch gradient

Clip gradient per sample to norm C

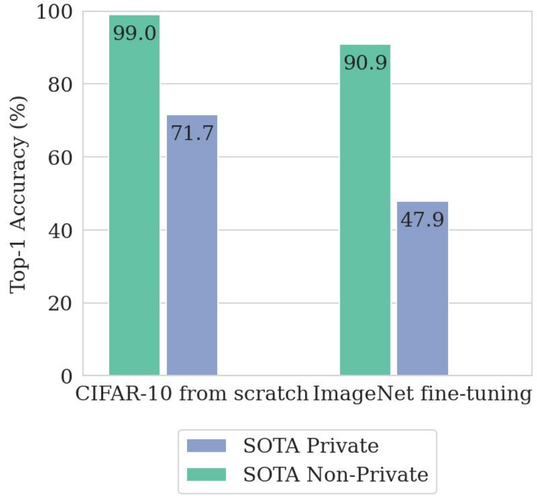
Add Gaussian noise

The total privacy loss ε of the training procedure:

- Increases with number of iterations T
- Decreases with added noise σ
- Increases with batch size |B|

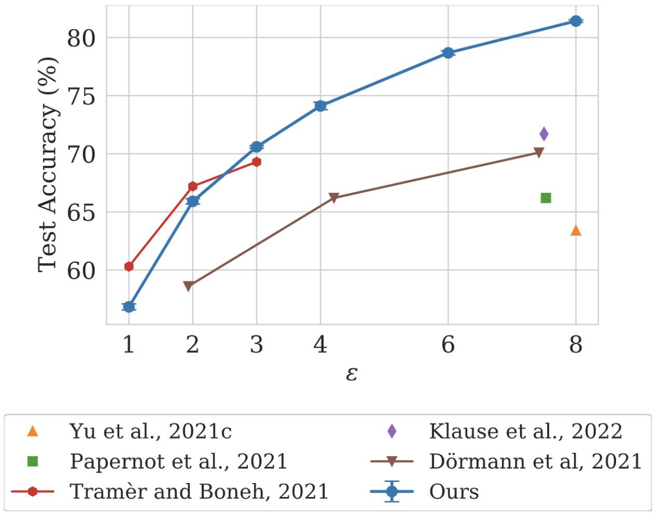
Challenges of DP-SGD

- **Bounded privacy budget ε**
  - Tradeoff between # iterations & amount of noise
  - Different hyper-parameter & regularization settings
- **Clipping per sample + Noise**
  - Privatized gradient is biased and has high variance
- **Making standard models work**
  - L2 norm of noise scales with model dimension
  - Cannot use batch normalization



Improving SOTA on CIFAR-10

CIFAR-10 classification under (8, 10^{-5})-DP	Accuracy (%)			
	Validation		Training	
Baseline (WRN-40-4 w/o batch normalization)	50.8	(0.7)	51.2	(0.7)
+ Group normalization (16 groups)	66.3	(0.6)	67.9	(0.3)
+ Larger batch size (batch size of 4096)	70.0	(0.6)	73.4	(0.9)
+ Weight standardization	71.2	(1.0)	74.7	(1.3)
+ Augmentation multiplicity (16 augmentations)	78.4	(0.9)	79.4	(0.9)
+ Parameter averaging (exponential moving average)	79.7	(0.2)	81.5	(0.2)

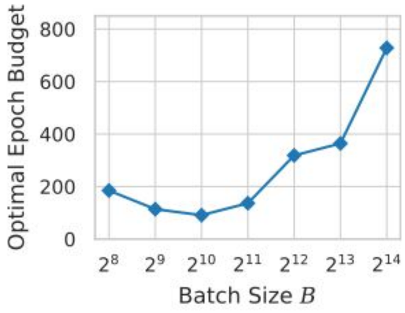


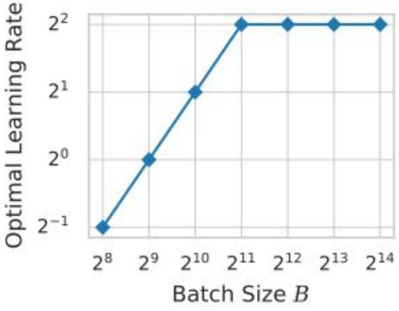
- Leverage ideas that make non-private training faster
- Improve network trainability and convergence
- Pack more compute per model update
- Careful hyper-parameter tuning

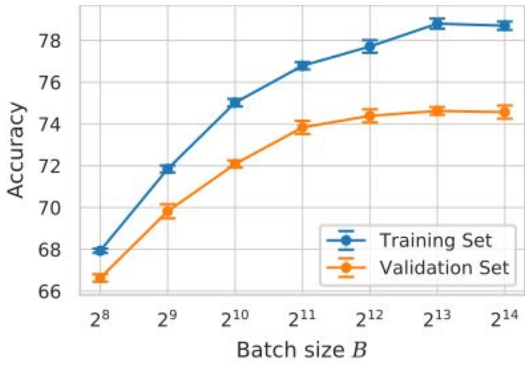
➔ **Better accuracy with larger, standard models**

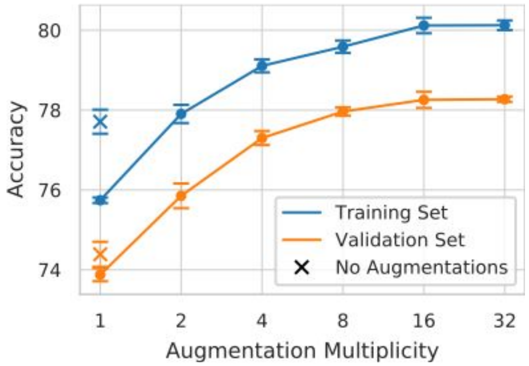
Insights Into Hyper-Parameter Tuning

- Clipping norm has little effect (eg. set C=1)
- Use constant learning rates (ie. no annealing)
- Very large batch sizes (use virtual batching)
- Add augmentation multiplicity once benefits from larger batch size saturate
- Optimal epoch budget and learning rate depend on batch size (re-tune for each batch size)

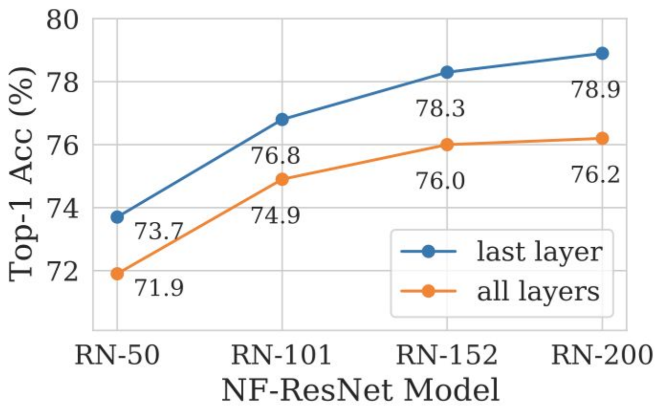


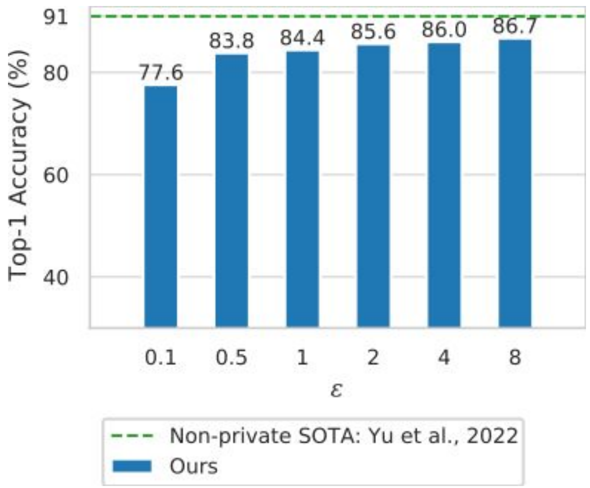






Closing the Public-Private Gap with Pre-Trained Models





- Pre-train on JFT and fine-tune with DP-SGD
- Accuracy keeps improving with model size
- Fine-tuning last layer better on ImageNet, all layers better when distribution shift is larger (eg. Places365)

➔ **Exceed accuracy of non-private ResNet-50 at ε=1**

Conclusion

1. Standard image classification models contain a "fingerprint" of each individual training example which can be extracted and used to reconstruct training examples.

2. Differential privacy provides an effective mitigation, and its accuracy degradation can be minimized by combining large models with tools to improve trainability and convergence.

https://github.com/deepmind/jax_privacy

DeepMind

Thank you!

Questions?