# Not All Samples Are Created Equal

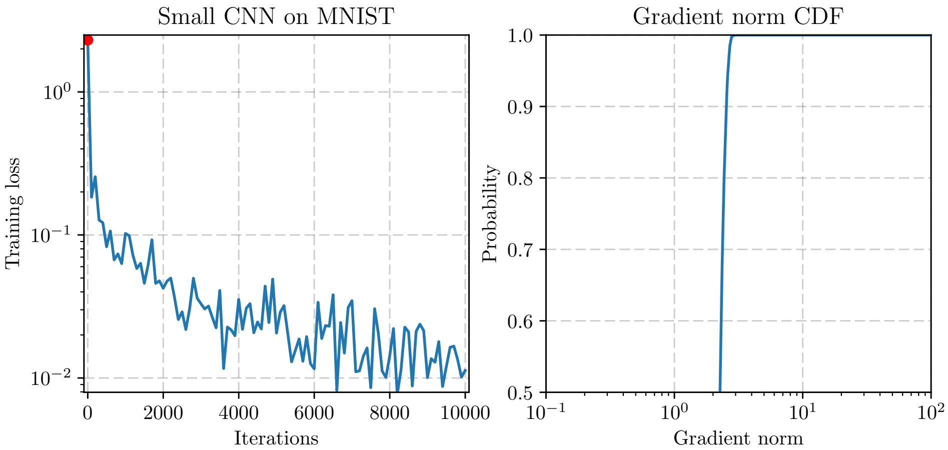
## Deep Learning with Importance Sampling

Angelos Katharopoulos & François Fleuret

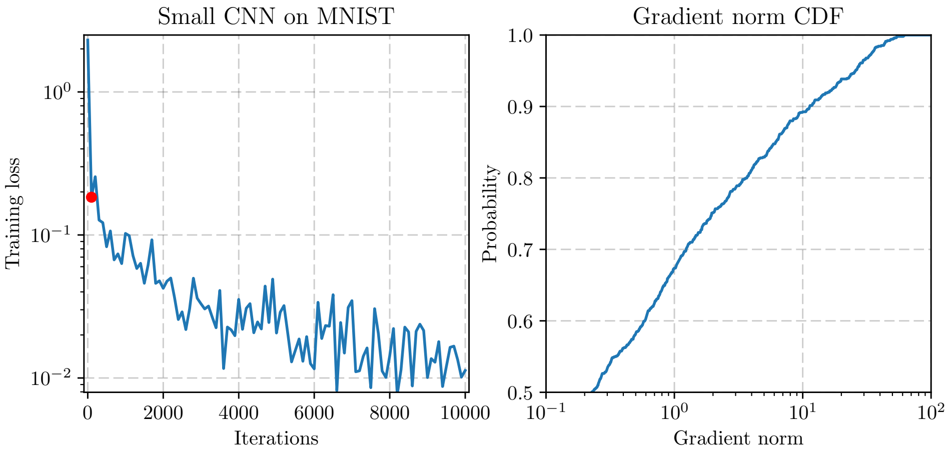
ICML, July 11, 2018

Funded by FNSNF

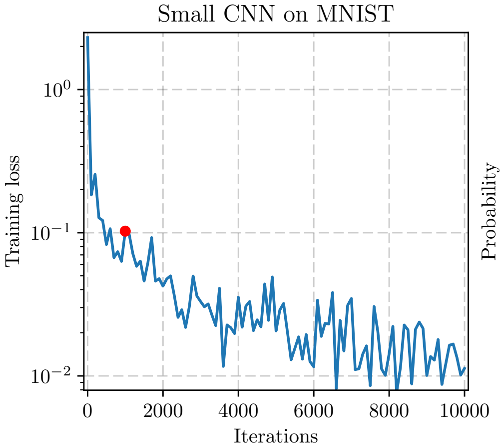
# Evolution of gradient norms during training

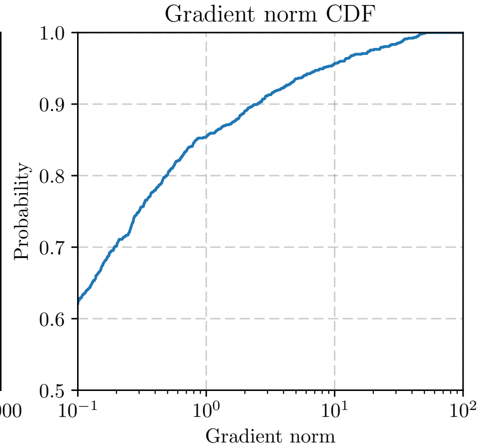


# Evolution of gradient norms during training

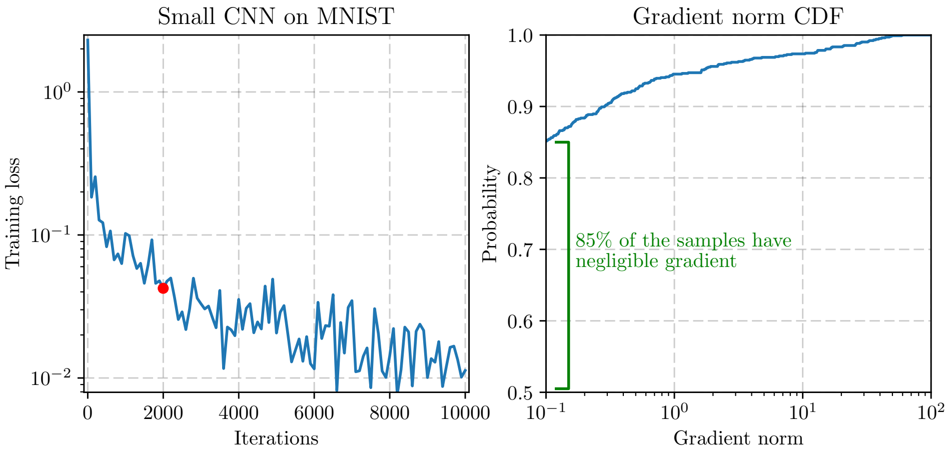


# Evolution of gradient norms during training



# Evolution of gradient norms during training



# Related work

- Sample points proportionally to the gradient norm (Needell et al., 2014; Zhao and Zhang, 2015; Alain et al., 2015)
- SVRG type methods (Johnson and Zhang, 2013; Defazio et al., 2014; Lei et al., 2017)
- Sample using the loss
  - Hard/Semi-hard sample mining (Schroff et al., 2015; Simo-Serra et al., 2015)
  - Online Batch Selection (Loshchilov and Hutter, 2015)
  - Prioritized Experience Replay (Schaul et al., 2015)

# Related work

- ~~Sample points proportionally to the gradient norm~~ (Needell et al., 2014; Zhao and Zhang, 2015; Alain et al., 2015)
- ~~SVRG type methods~~ (Johnson and Zhang, 2013; Defazio et al., 2014; Lei et al., 2017)
- Sample using the loss
  - Hard/Semi-hard sample mining (Schroff et al., 2015; Simo-Serra et al., 2015)
  - Online Batch Selection (Loshchilov and Hutter, 2015)
  - Prioritized Experience Replay (Schaul et al., 2015)

# Contributions

- Derive a fast to compute importance distribution
- Variance cannot always be reduced so start importance sampling when it is useful

# Contributions

- Derive a fast to compute importance distribution
- Variance cannot always be reduced so start importance sampling when it is useful

- Package everything in an embarassingly simple to use library BONUS

# Deriving the sampling distribution (1)

Similar to Zhao and Zhang (2015) we want to minimize the variance of the gradients.

$$P^* = \underset{P}{\arg\min}\, \mathrm{Tr}\left(\mathbb{V}_P[w_i G_i]\right) = \arg\min \mathbb{E}_P\left[w_i^2 \left\|G_i\right\|_2^2\right]$$

To simplify, we minimize an upper bound

$$\left\|G_i\right\|_2 \leq \hat{G}_i \iff \min_P \mathbb{E}_P\left[w_i^2 \left\|G_i\right\|_2^2\right] \leq \min_P \mathbb{E}_P\left[w_i^2 \hat{G}_i^{\,2}\right]$$

# Deriving the sampling distribution (1)

Similar to Zhao and Zhang (2015) we want to minimize the variance of the gradients.

$$P^* = \arg\min_P \operatorname{Tr}\left(\mathbb{V}_P[w_i G_i]\right) = \arg\min \mathbb{E}_P\left[w_i^2 \left\| G_i \right\|_2^2\right]$$

To simplify, we minimize an upper bound

$$\left\| G_i \right\|_2 \leq \hat{G}_i \iff \min_P \mathbb{E}_P\left[w_i^2 \left\| G_i \right\|_2^2\right] \leq \min_P \mathbb{E}_P\left[w_i^2 \hat{G}_i^{\,2}\right]$$

# Deriving the sampling distribution (1)

Similar to Zhao and Zhang (2015) we want to minimize the variance of the gradients.

$$P^* = \underset{P}{\arg\min} \operatorname{Tr}\left(\mathbb{V}_P[w_i G_i]\right) = \arg\min \mathbb{E}_P\left[{w_i}^2 \left\| G_i \right\|_2^2\right]$$

To simplify, we minimize an upper bound

$$\left\| G_i \right\|_2 \leq \hat{G}_i \iff \min_P \mathbb{E}_P\left[{w_i}^2 \left\| G_i \right\|_2^2\right] \leq \min_P \mathbb{E}_P\left[{w_i}^2 {\hat{G}_i}^2\right]$$

# Deriving the sampling distribution (1)

Similar to Zhao and Zhang (2015) we want to minimize the variance of the gradients.

$$P^* = \arg\min_P \operatorname{Tr}\left(\mathbb{V}_P[w_i G_i]\right) = \arg\min \mathbb{E}_P\left[w_i^2 \left\|G_i\right\|_2^2\right]$$

To simplify, we minimize an upper bound

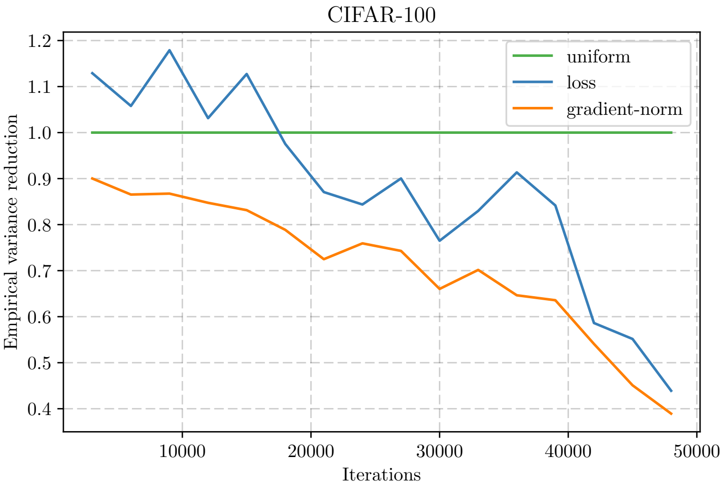
$$\left\|G_i\right\|_2 \leq \hat{G}_i \iff \min_P \mathbb{E}_P\left[w_i^2 \left\|G_i\right\|_2^2\right] \leq \min_P \mathbb{E}_P\left[w_i^2 \hat{G}_i^{\,2}\right]$$

# Deriving the sampling distribution (2)

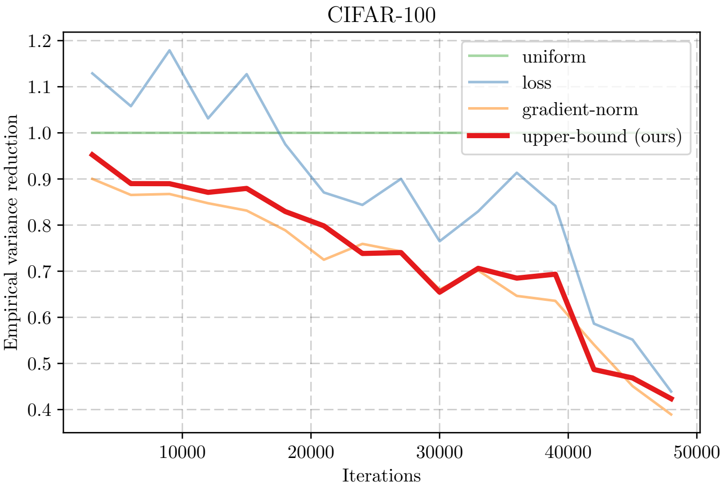
We show that we can upper bound the gradient norm of the parameters using the norm of the gradient with respect to the pre-activation outputs of the last layer.

We conjecture that batch normalization and weight initialization make it tight.

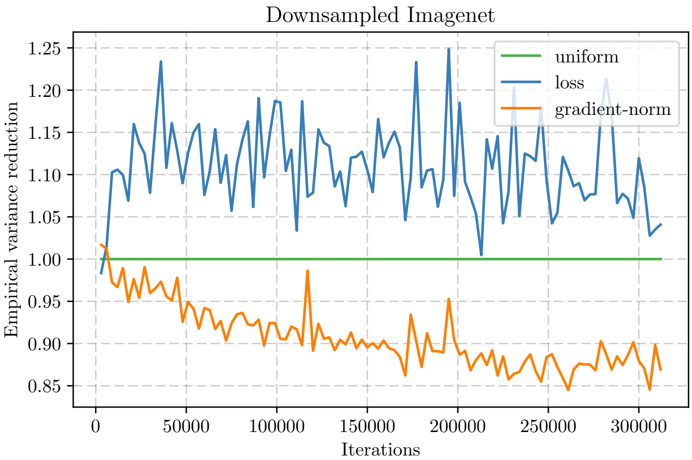
# Variance reduction achieved with our upper-bound



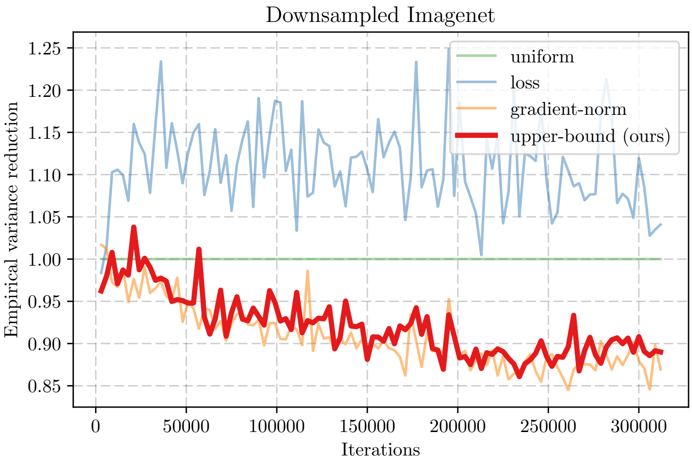
# Variance reduction achieved with our upper-bound



# Variance reduction achieved with our upper-bound



# Variance reduction achieved with our upper-bound



# Is the upper-bound enough to speed up training?

Not really, because

- a forward pass on the whole dataset is still prohibitive
- the importance distribution can be arbitrarily close to uniform

Two key ideas

- Sample a **large batch** ($B$) randomly and resample a **small batch** ($b$) with importance
- Start importance sampling when the variance will be reduced

# When do we start importance sampling?

We start importance sampling when the variance reduction is large enough

$$\mathrm{Tr}\left(\mathbb{V}_u[G_i]\right) - \mathrm{Tr}\left(\mathbb{V}_P[w_i G_i]\right) = \frac{1}{B}\sum_{i=1}^{B}\|G_i\|_2^2 \sum_{i=1}^{B}(p_i - u)^2 \propto \underbrace{\sum_{i=1}^{B}(p_i - u)^2}_{\substack{\text{distance of importance} \\ \text{distribution to uniform}}}$$

# When do we start importance sampling?

We start importance sampling when the variance reduction is large enough

$$\mathrm{Tr}\left(\mathbb{V}_u[G_i]\right) - \mathrm{Tr}\left(\mathbb{V}_P[w_i G_i]\right) = \frac{1}{B} \sum_{i=1}^{B} \|G_i\|_2^2 \sum_{i=1}^{B} (p_i - u)^2 \propto \underbrace{\sum_{i=1}^{B} (p_i - u)^2}_{\substack{\text{distance of importance} \\ \text{distribution to uniform}}}$$

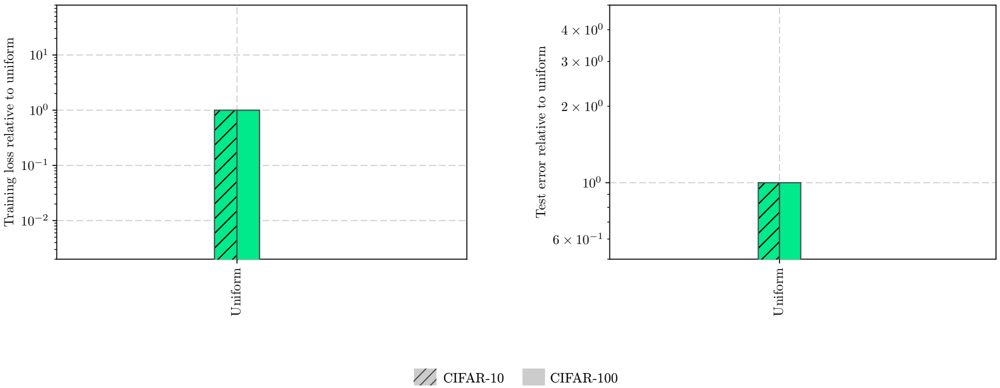
We show that the **equivalent batch increment** $\tau \geq \left(1 - \frac{\sum_i (p_i - u)^2}{\sum_i p_i^2}\right)^{-1}$ which allows us to perform importance sampling when

$$\underbrace{B t_{\text{forward}} + b\left(t_{\text{forward}} + t_{\text{backward}}\right)}_{\substack{\text{Time for } \textbf{importance} \\ \textbf{sampling iteration}}} \leq \underbrace{\tau\left(t_{\text{forward}} + t_{\text{backward}}\right) b}_{\substack{\text{Time for equivalent} \\ \textbf{uniform sampling iteration}}}$$

# Experimental setup

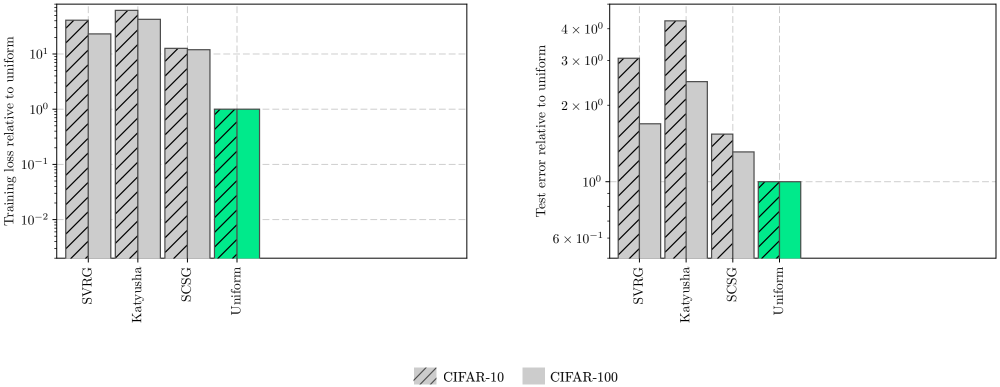
- We fix a time budget for all methods and compare the achieved training loss and test error
- We evaluate on three tasks
  1. WideResnets on CIFAR10/100 (image classification task)
  2. Pretrained ResNet50 on MIT67 (finetuning task)
  3. LSTM on permuted MNIST (sequence classification task)

# Importance sampling for image classification



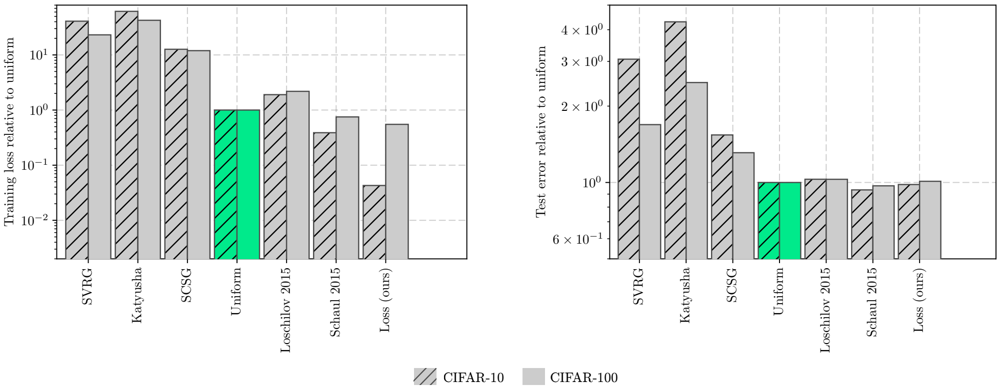
# Importance sampling for image classification

- SVRG methods do not work for Deep Learning



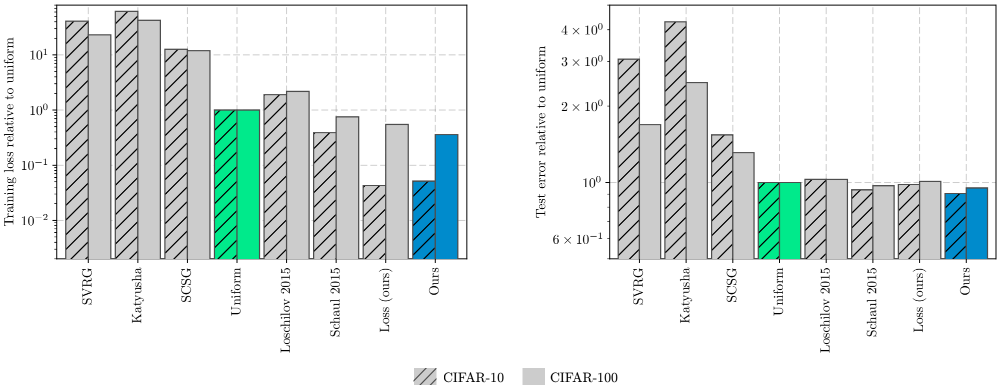
# Importance sampling for image classification

- SVRG methods do not work for Deep Learning
- Our loss-based sampling outperfoms existing loss based methods



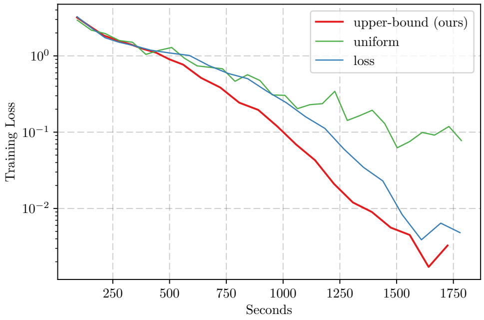
# Importance sampling for image classification

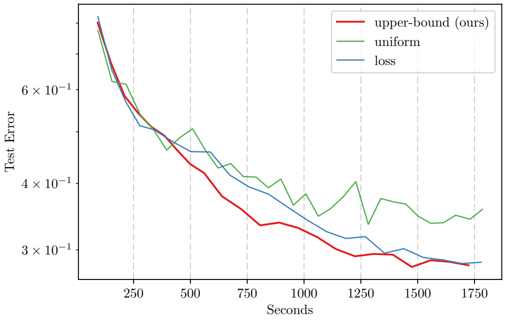
- SVRG methods do not work for Deep Learning
- Our loss-based sampling outperfoms existing loss based methods
- Improvement from **3**× **to 10**× compared to training loss with uniform sampling



# Importance sampling for finetuning

- Earlier variance reduction leads to faster convergence



# Thank you for your time!

Check out the code at http://github.com/idiap/importance-sampling .

```
from importance_sampling import ImportanceTraining
x, y = load_data()
model = load_model()
ImportanceTraining(model).fit(x, y, batch_size=128, epochs=10)
```

# References I

Deanna Needell, Rachel Ward, and Nati Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2014.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1–9, 2015.

Guillaume Alain, Alex Lamb, Chinnadhurai Sankar, Aaron Courville, and Yoshua Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481*, 2015.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

# References II


Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

Lihua Lei, Cheng Ju, Jianbo Chen, and Michael I Jordan. Non-convex finite-sum optimization via scsg methods. In *Advances in Neural Information Processing Systems*, pages 2345–2355, 2017.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Edgar Simo-Serra, Eduard Trulls, Luis Ferraz, Iasonas Kokkinos, Pascal Fua, and Francesc Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 118–126. IEEE, 2015.

# References III

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.