A Closer Look at Codistillation for Distributed Training

Shagun Sodhani

FACEBOOK AI

Collaborators

Olivier

Mido

Koustuv

Nicolas

Mike

FACEBOOK AI

Agenda

What is codistillation?

Why should we care?

Does it work?

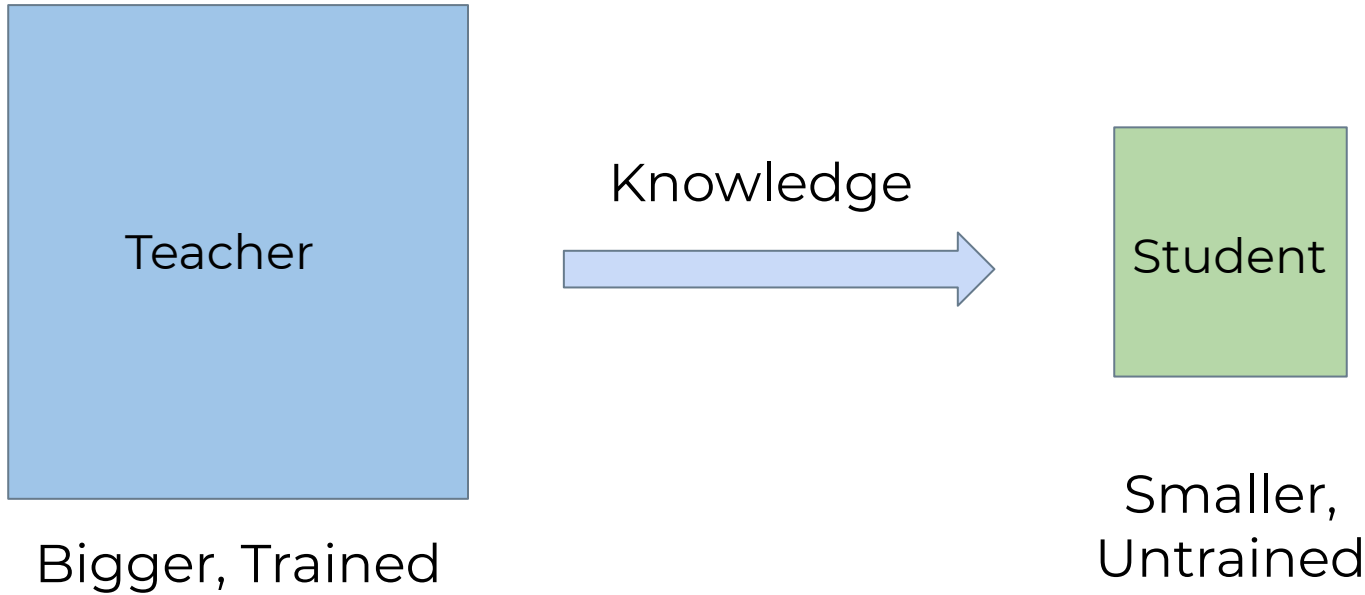
What's next?

Questions are welcome at all times :)

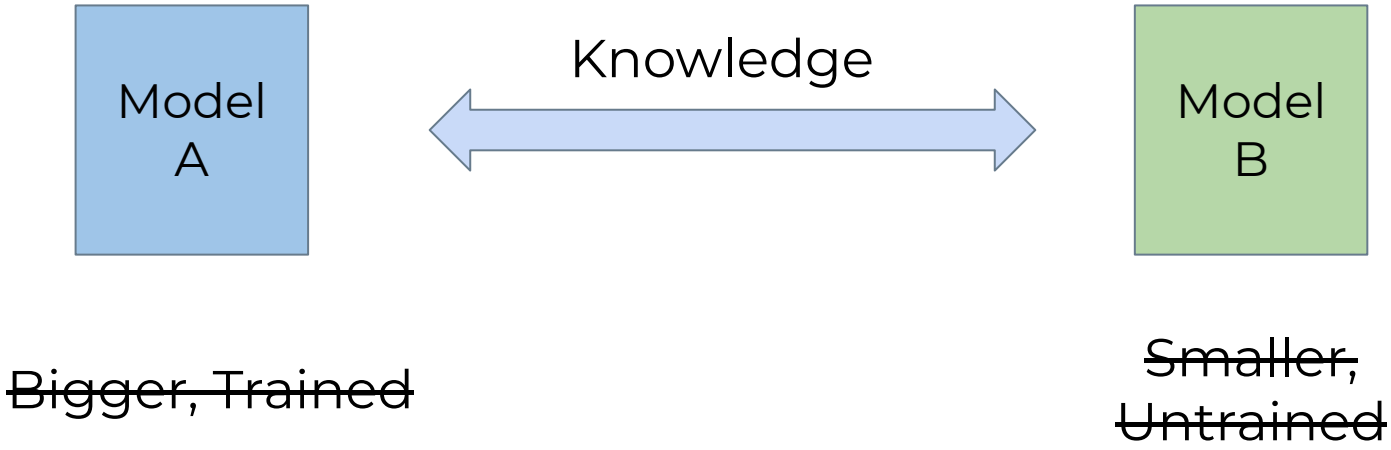
Codistillation is like ... distillation*
(*conditions apply)

Zhang et al., Deep mutual learning, CVPR 2018
Anil et al., Large scale distributed neural network training through online distillation, ICLR 2018

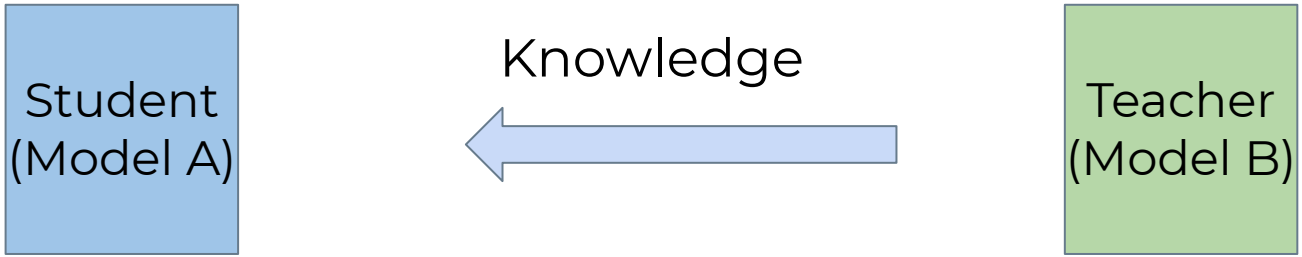
Distillation



Codistillation



PoV of Model A

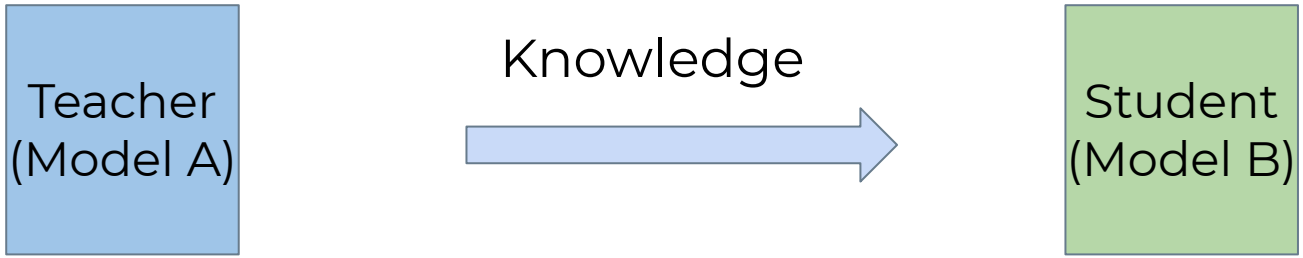


$$\theta_A^{k+1} = \theta_A^k - \eta \nabla_{\theta_A} \left(L(y, f_{\theta_A^k}(x)) + D\left(f_{\theta_B^k}(x), f_{\theta_A^k}(x)\right) \right)$$

Supervised Learning Loss

Distillation Loss

PoV of Model B



$$\theta_B^{k+1} = \theta_B^k - \eta \nabla_{\theta_B} \left(L(y, f_{\theta_B^k}(x)) + D\left(f_{\theta_A^k}(x), f_{\theta_B^k}(x)\right) \right)$$

Supervised Learning Loss

Distillation Loss

Agenda

What is codistillation?

Why should we care?

Does it work?

What's next?

Questions are welcome at all times :)

Previous work: Codistillation behaves like ensembling

- Model trained with codistillation performs similar to ensemble of independently trained models.
- Codistillation can scale training to large batch sizes.

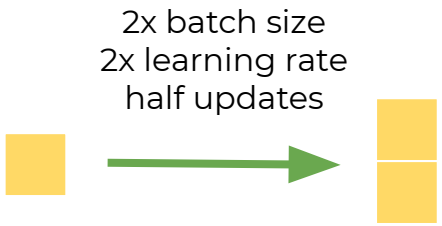
Zhang et al., Deep mutual learning, CVPR 2018
Anil et al., Large scale distributed neural network training through online distillation, ICLR 2018

Scale to large batch sizes

Scale to large batch sizes

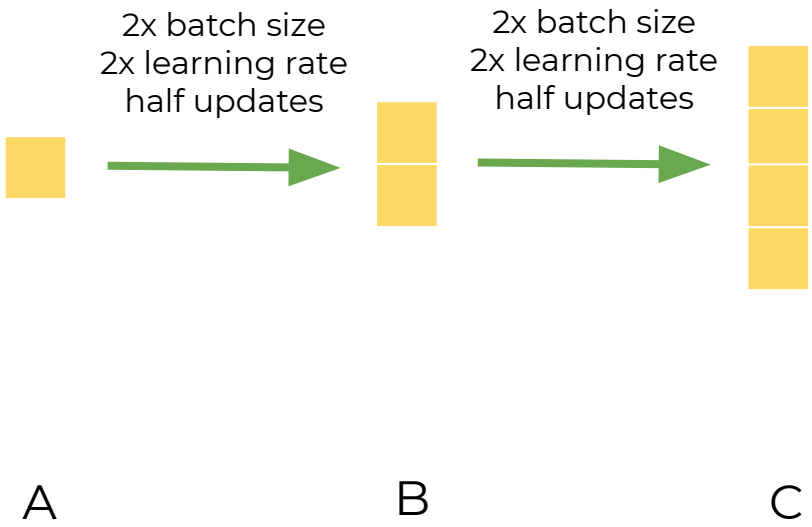
A

Scale to large batch sizes

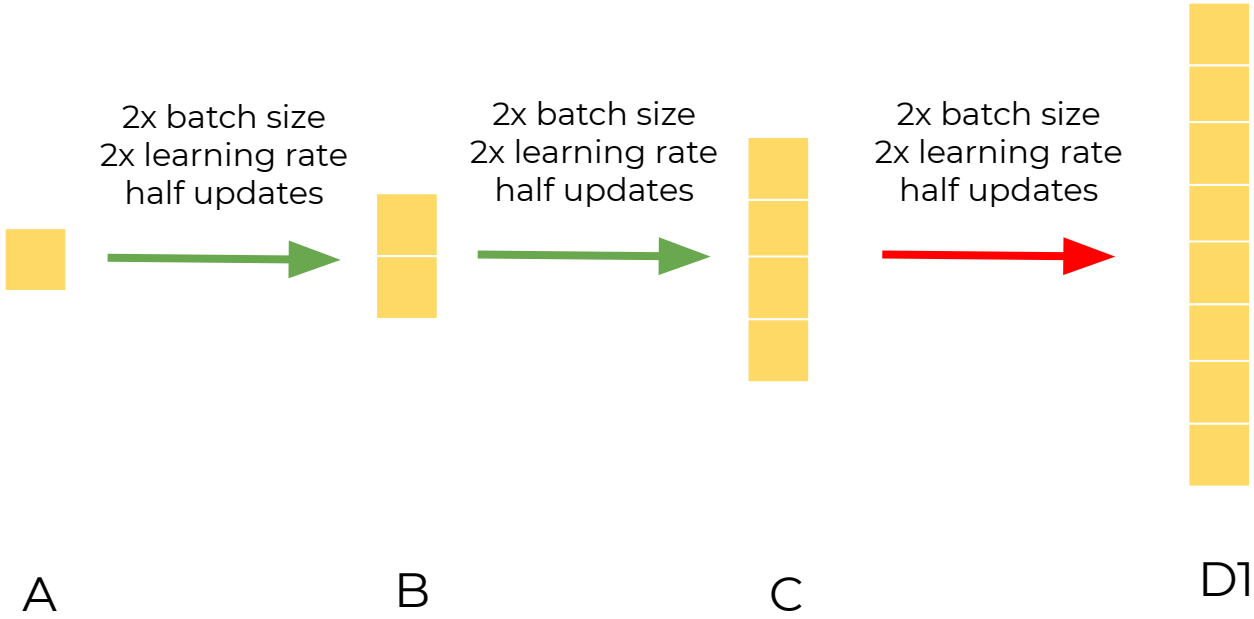


A B

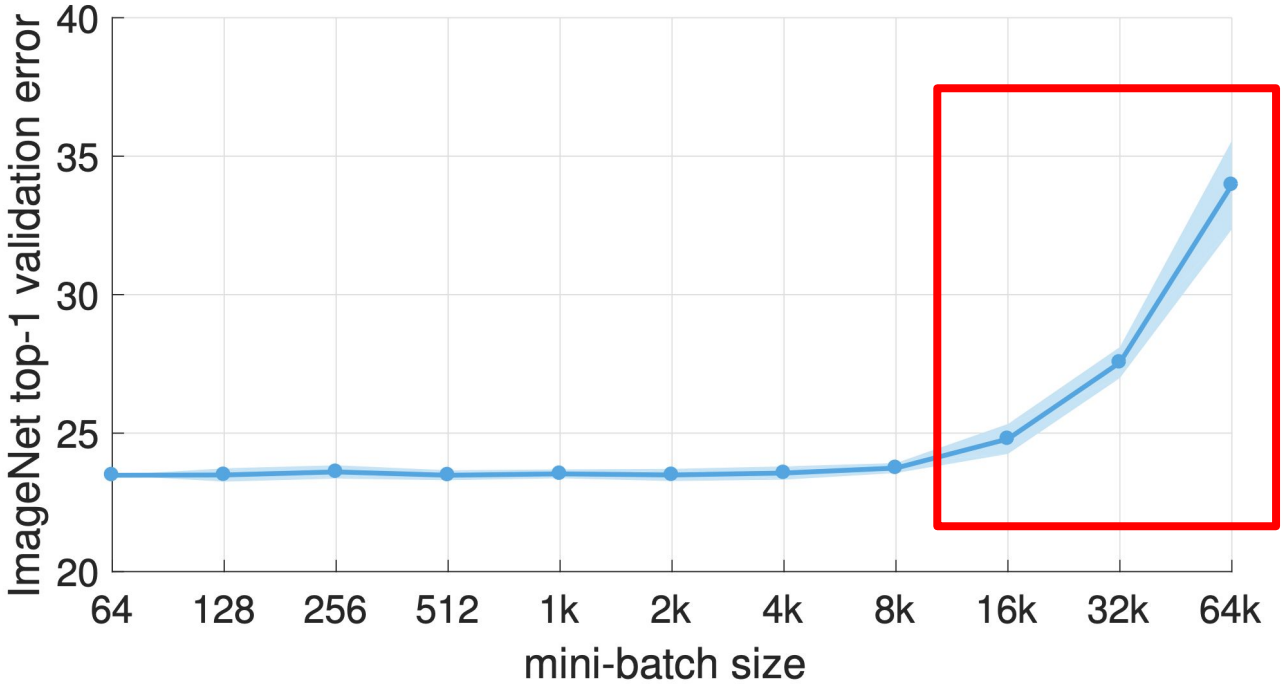
Scale to large batch sizes



Scale to large batch sizes

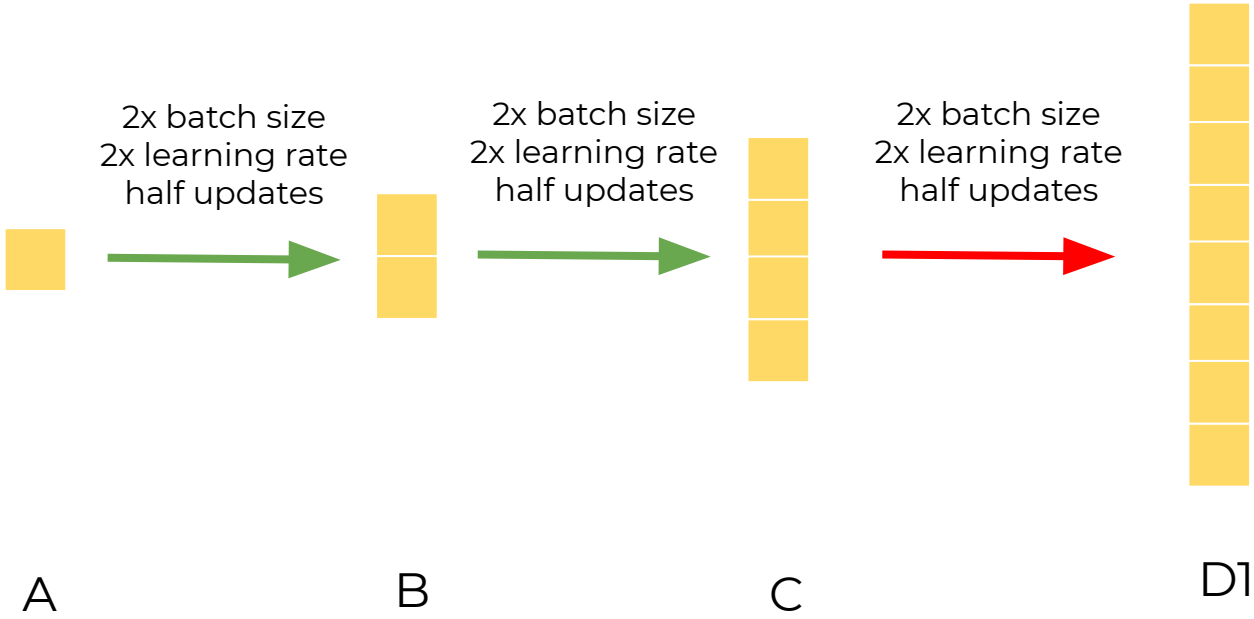


Scale to large batch sizes

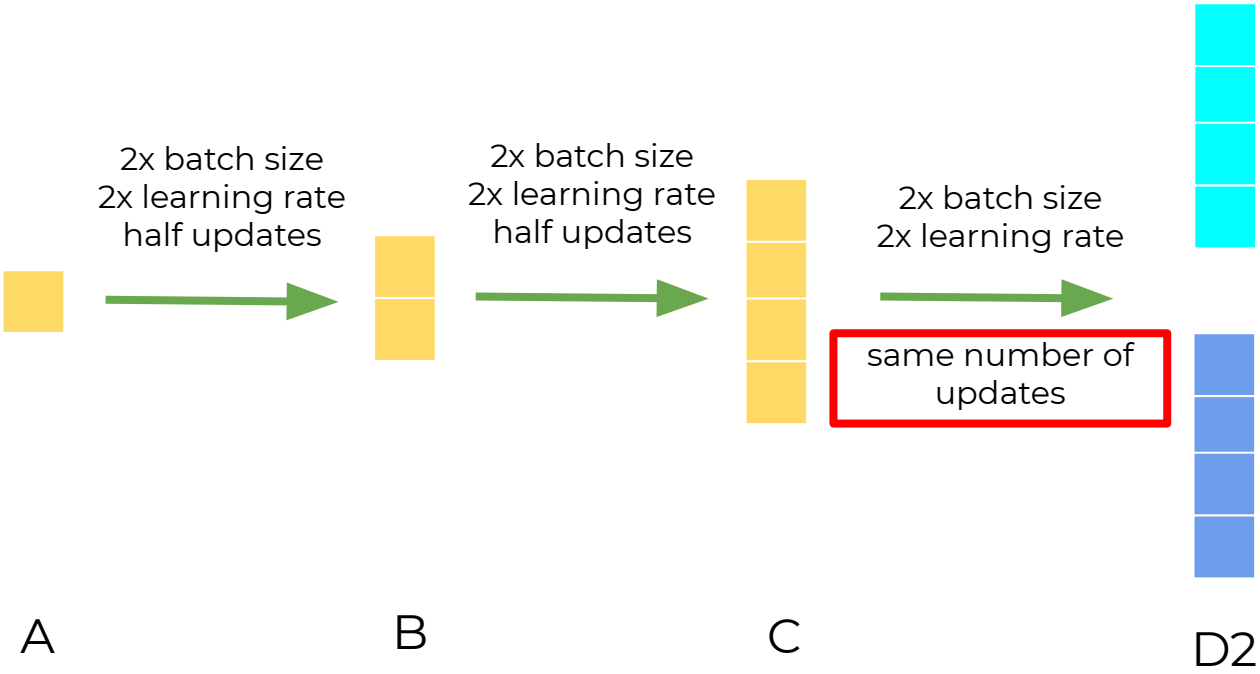


Goyal et al., Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, 2017

Scale to large batch sizes

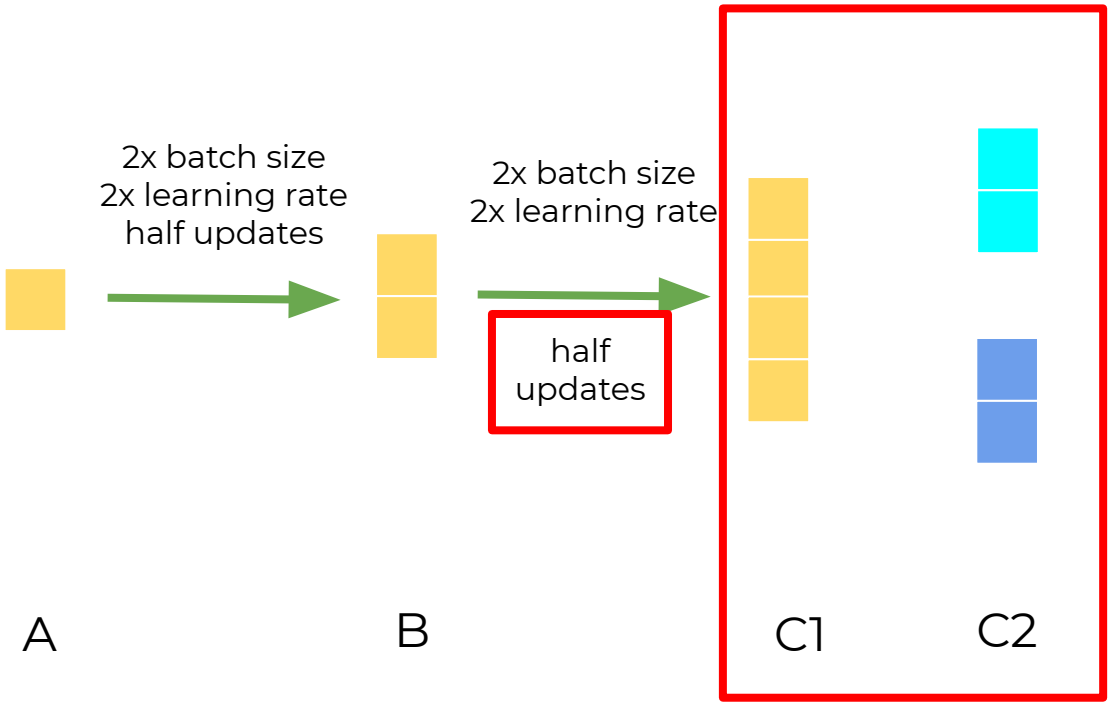


Previous Work: Scale to large batch sizes

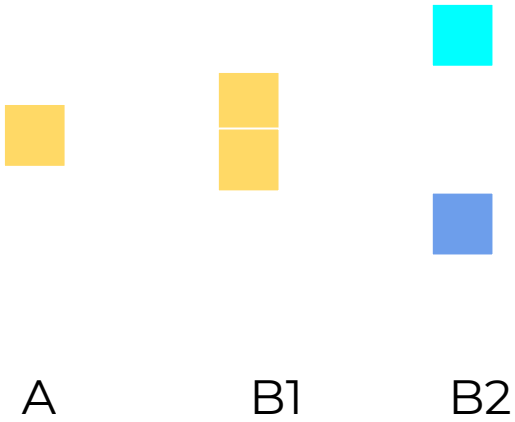


Anil et al., Large scale distributed neural network training through online distillation, ICLR 2018

Our Focus: Scale to large batch sizes

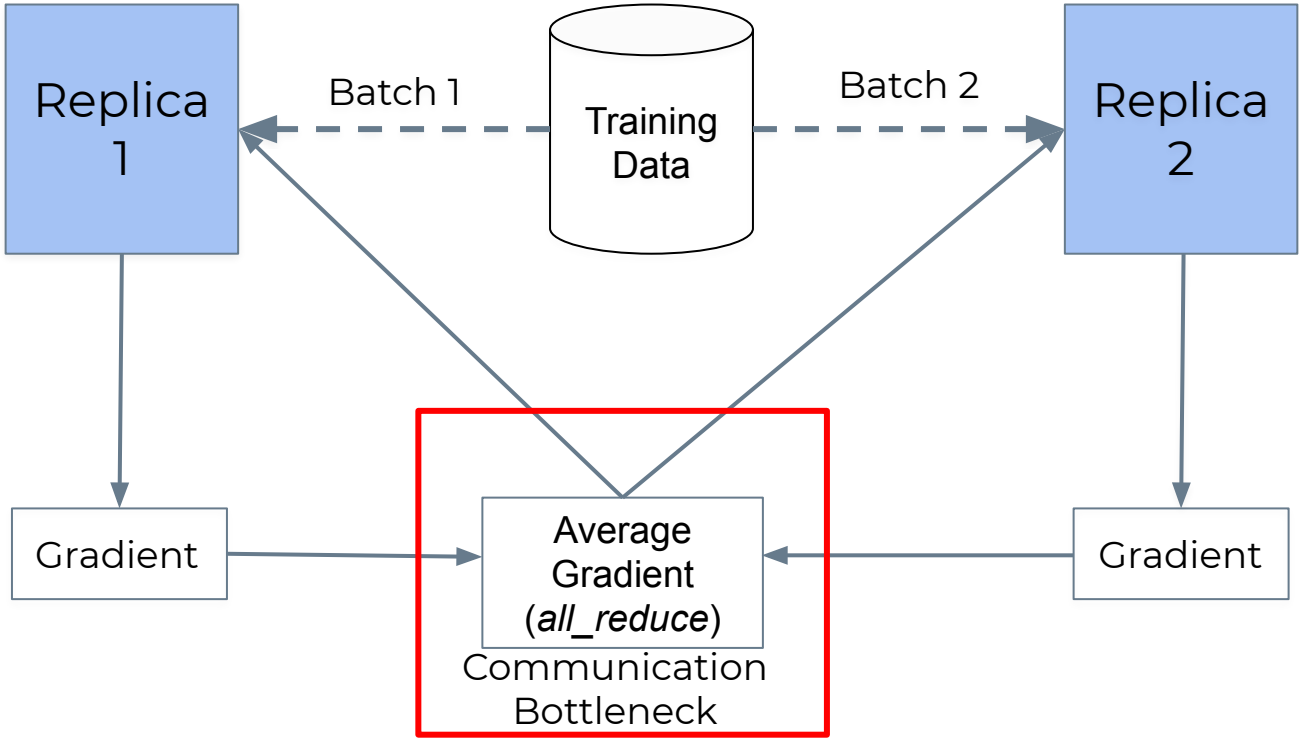


Train in parallel to reduce wall-clock time



Setup	Number of GPUs	Number of updates per GPU	Total number of updates
A	N	M	N * M
B1	2 * N	M / 2	N * M
Previous Work (B2)	2 * N	M	2 * N * M
Our Work (B2)	2 * N	M / 2	N * M

Distributed Data-Parallel Training

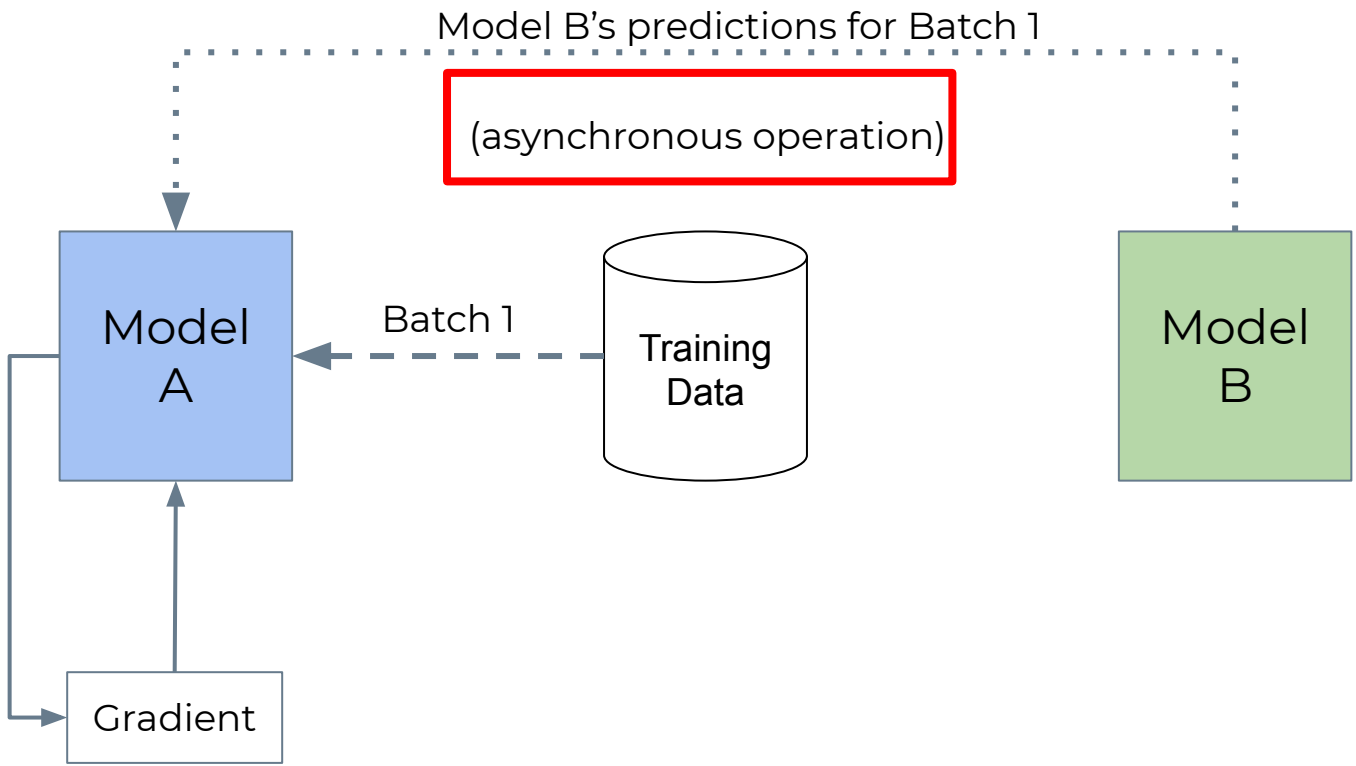


Distributed Data-Parallel Training

Number of Nodes (8 GPUs each)	Time for forward + backward + communication (in ms)
1	412
2	956
4	1568
8	1593

For Transformer-Large with 229M parameters
(Ethernet Interconnect)

Codistillation



Agenda

What is codistillation?

Why should we care?

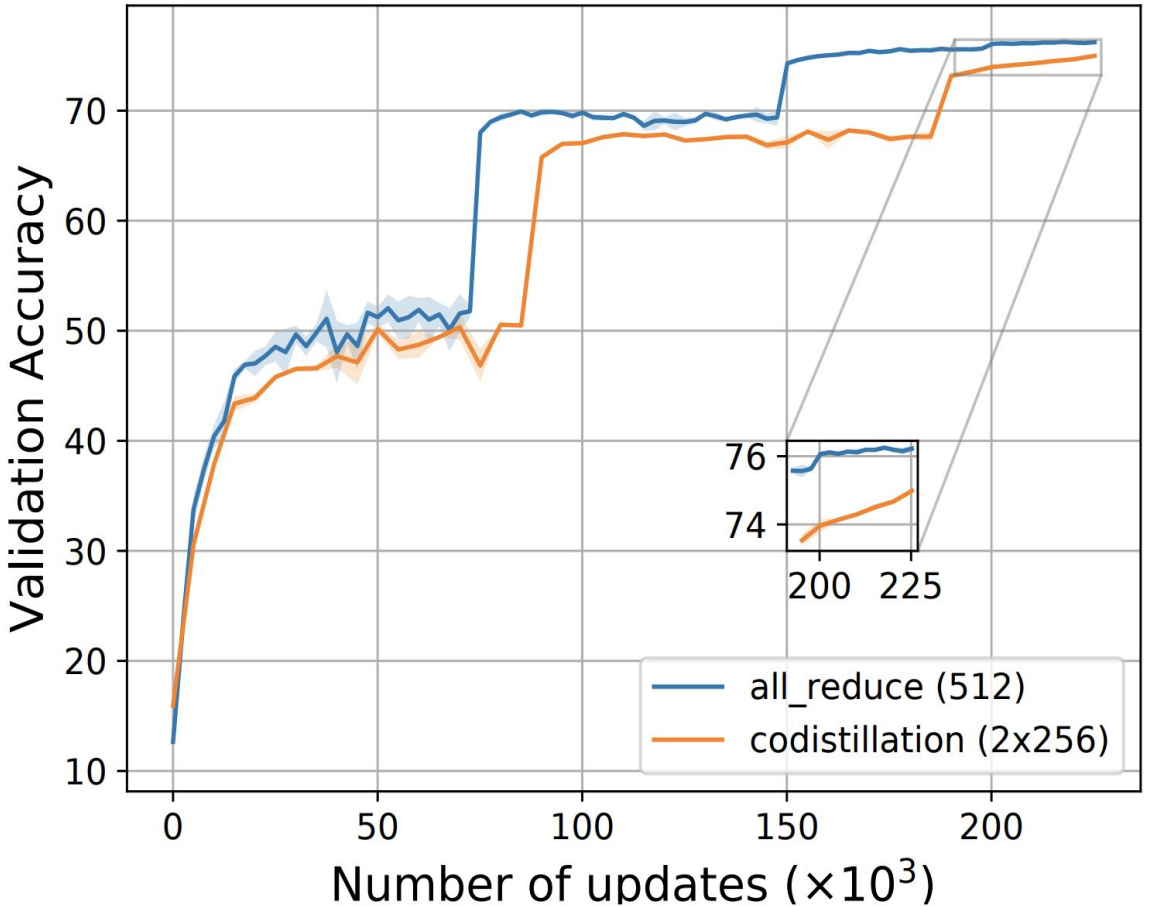
Does it work?

What's next?

Questions are welcome at all times :)

Lets try it

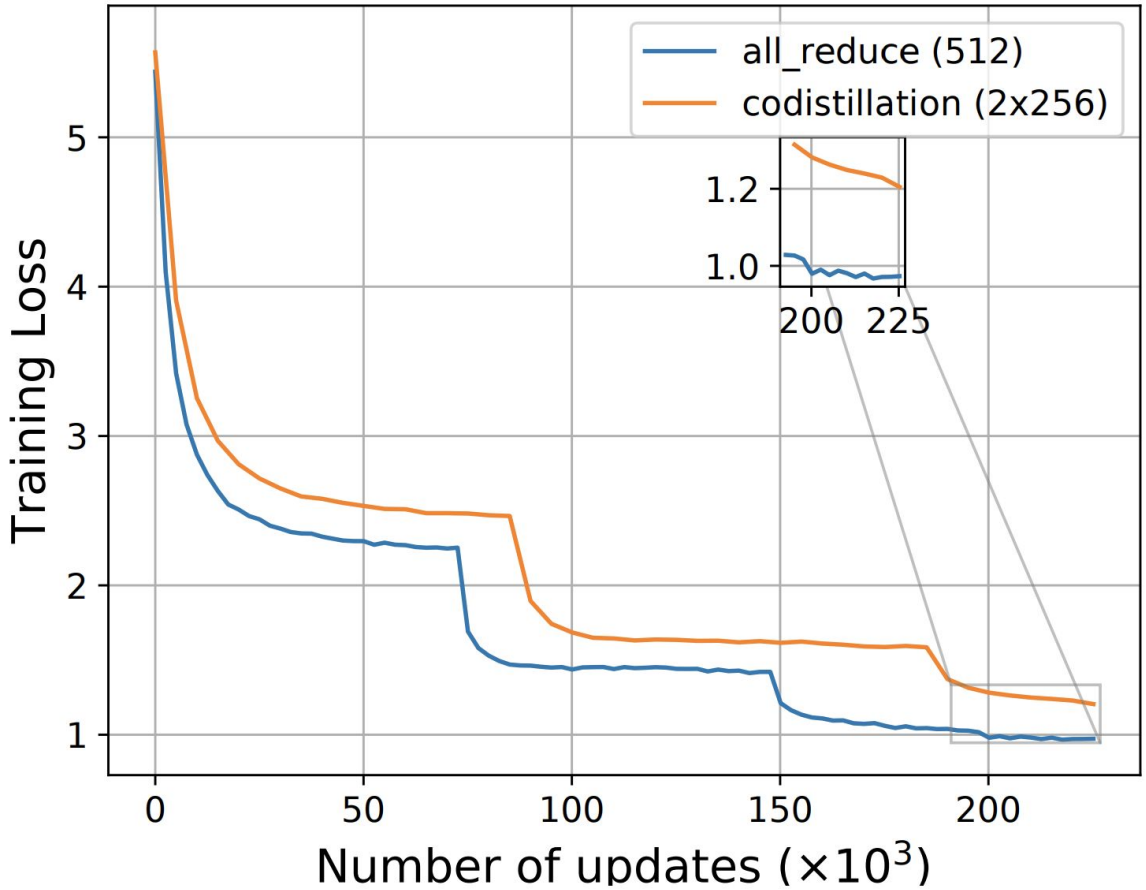
- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.



Goyal et al., Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, 2017

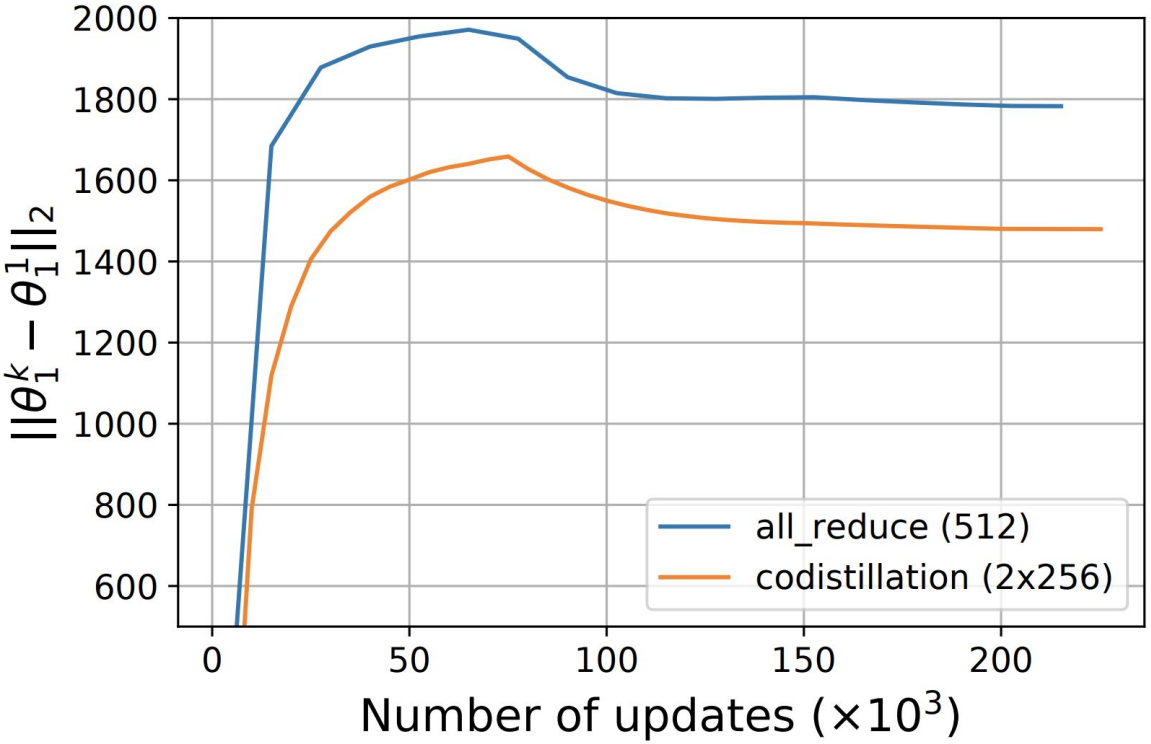
Lets debug it

- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.



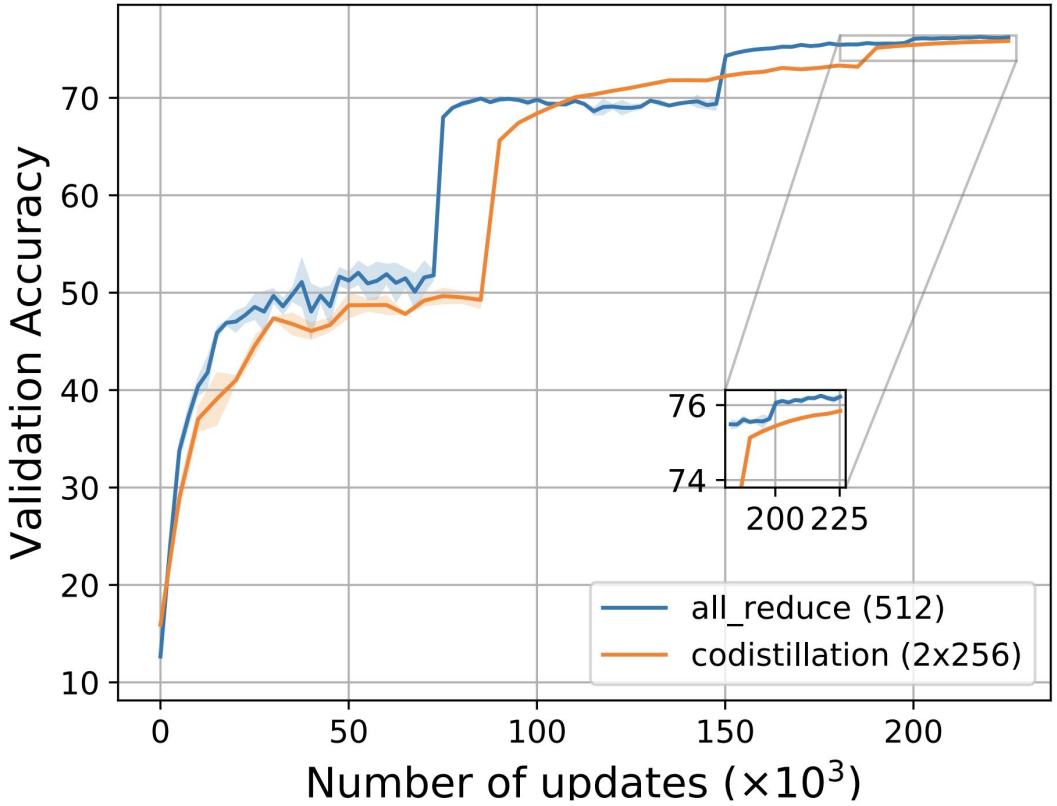
Codistillation “out of the box” can over-regularize

- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.



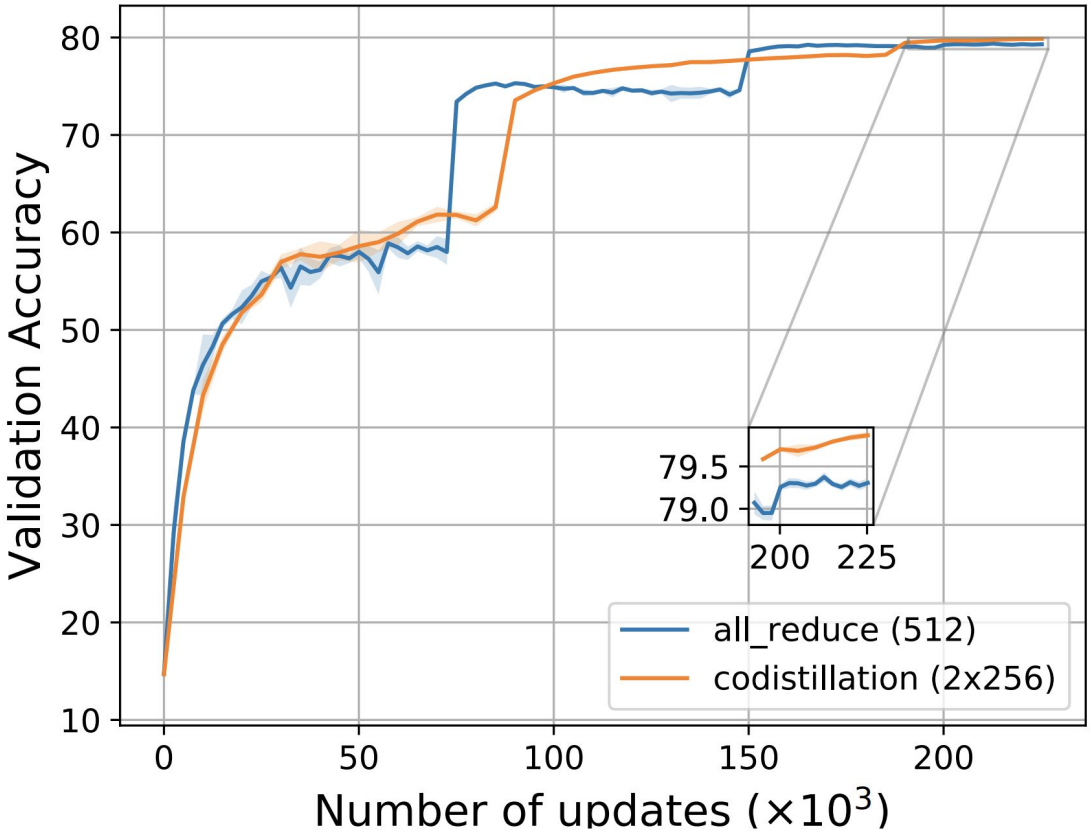
Bridging the gap

- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.
- Reduce L2 regularization



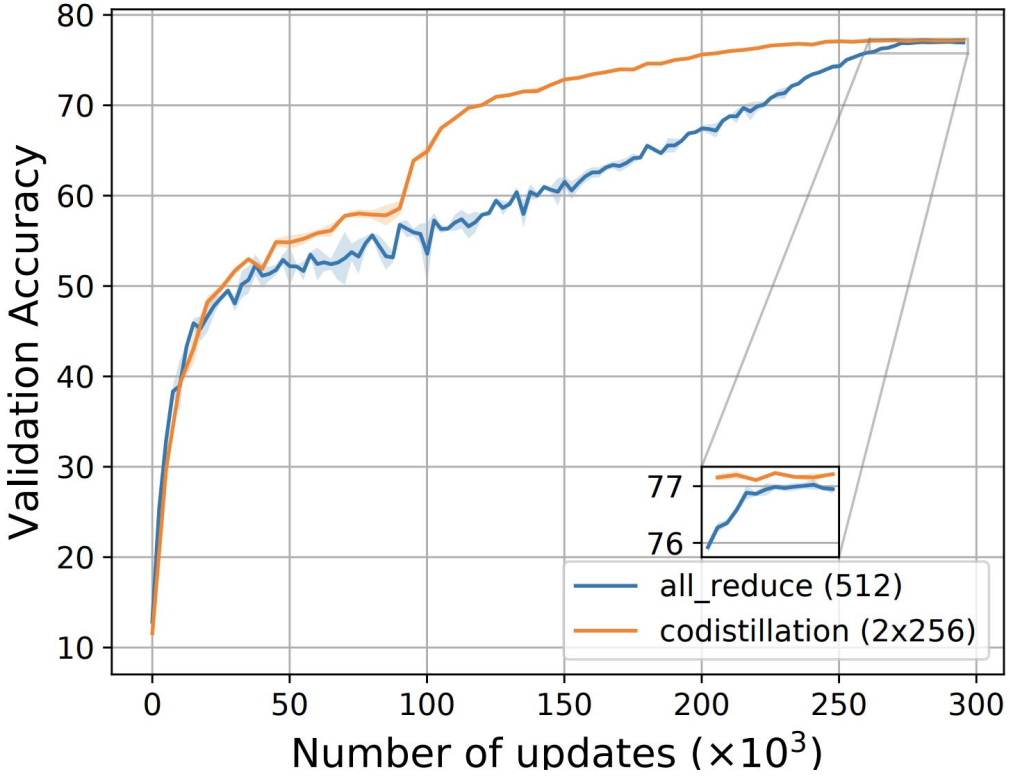
Bridging the gap

- ImageNet dataset
- ResNeXt101 model
- Based on setup from Goyal et al.



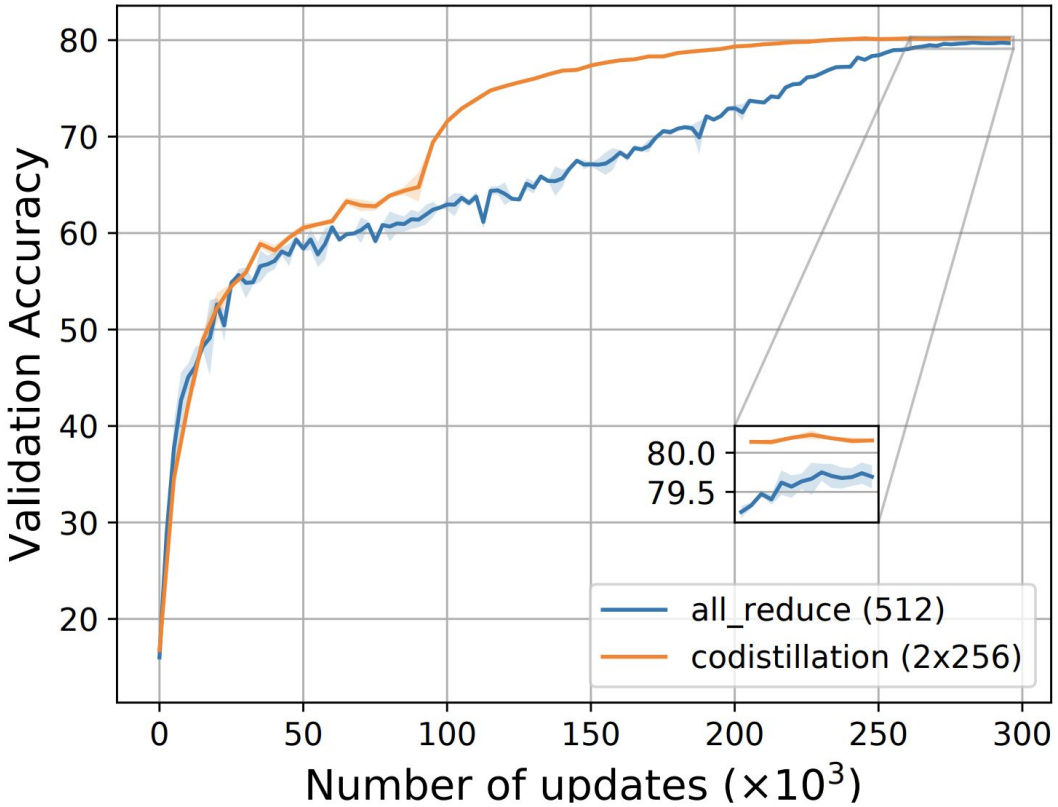
Learning rate schedule?

- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.
- Use cosine learning rate schedule



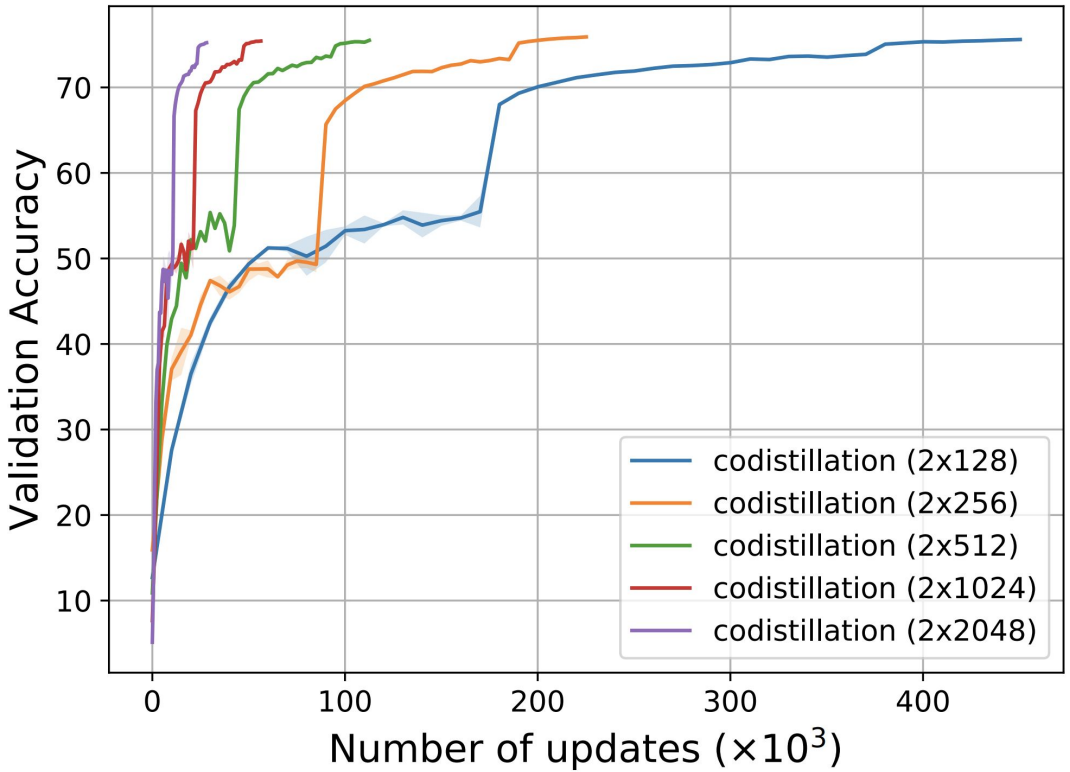
Learning rate schedule?

- ImageNet dataset
- ResNeXt101 model
- Based on setup from Goyal et al.
- Use cosine learning rate schedule



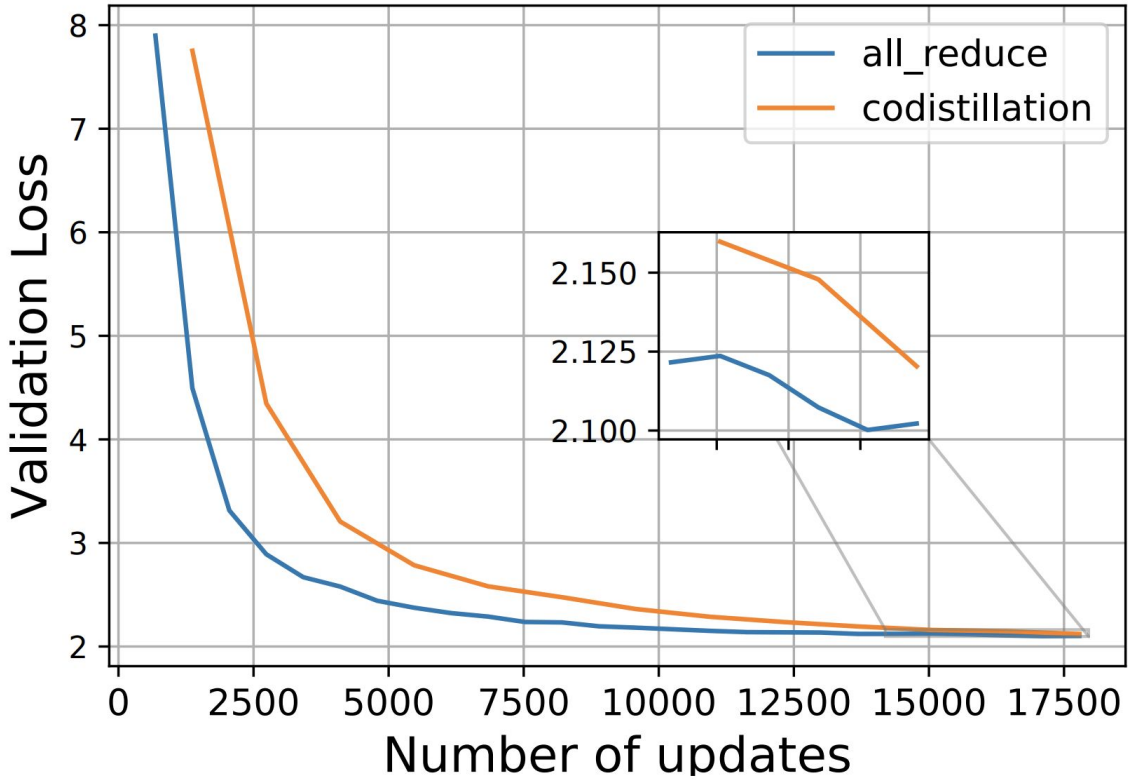
Scaling codistillation with more gpus?

- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.
- Increase number of gpus per model



What about NLP?

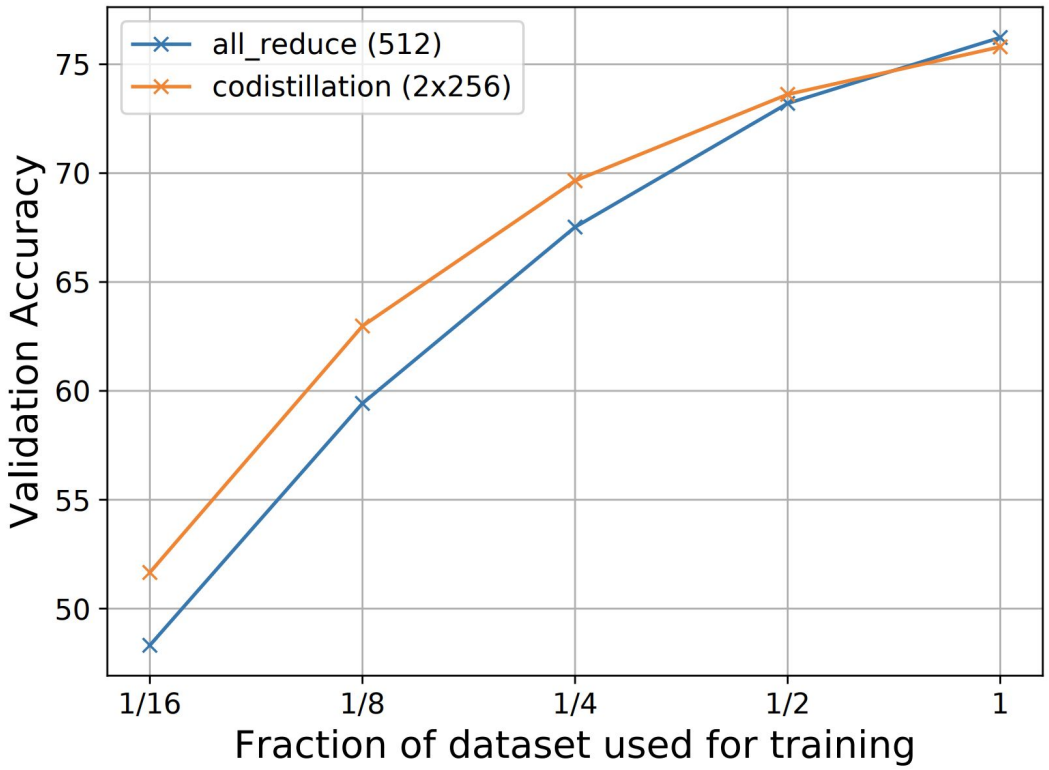
- WMT'16 En-De Translation dataset
- Transformer Large model
- Based on setup from Ott et al.



Ott et al., Scaling Neural Machine Translation, WMT 2018

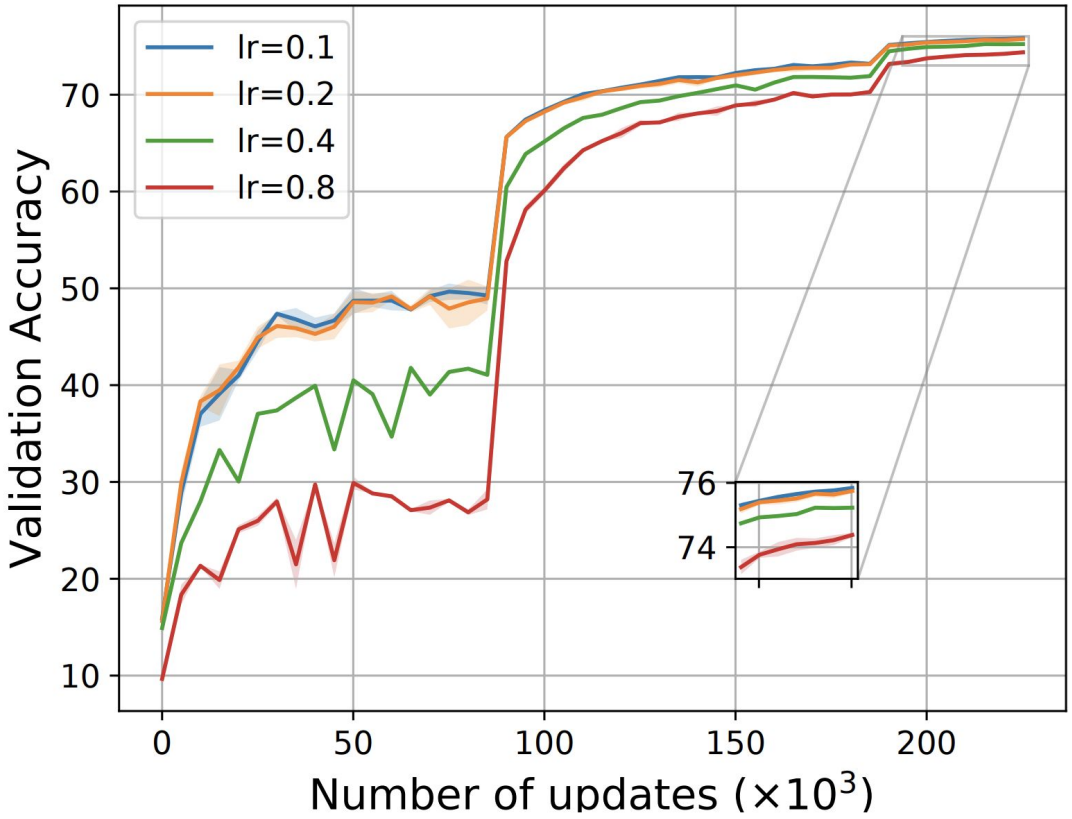
Vary number of training data points.

- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.



Robustness to various learning rates

- ImageNet dataset
- ResNet50 model
- Based on setup from Goyal et al.



Agenda

What is codistillation?

Why should we care?

Does it work?

What's next?

Questions are welcome at all times :)

Extensions

- n-way codistillation
- codistillation between differently trained models.

Thank you

@shagunsodhani

FACEBOOK AI

References

@inproceedings{zhang2018deep,
  title={Deep mutual learning},
  author={Zhang, Ying and Xiang, Tao and Hospedales, Timothy M and Lu, Huchuan},
  booktitle={Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition},
  pages={4320--4328},
  year={2018}
}

@article{anil2018large,
  title={Large scale distributed neural network training through online distillation},
  author={Anil, Rohan and Pereyra, Gabriel and Passos, Alexandre and Ormandi, Robert and Dahl, George E and Hinton, Geoffrey E},
  journal={arXiv preprint arXiv:1804.03235},
  year={2018}
}

References

@article{goyal2017accurate,
  title={Accurate, large minibatch sgd: Training imagenet in 1 hour},
  author={Goyal, Priya and Doll{\'a}r, Piotr and Girshick, Ross and Noordhuis, Pieter and Wesolowski, Lukasz and Kyrola, Aapo and Tulloch, Andrew and Jia, Yangqing and He, Kaiming},
  journal={arXiv preprint arXiv:1706.02677},
  year={2017}
}

@article{ott2018scaling,
  title={Scaling neural machine translation},
  author={Ott, Myle and Edunov, Sergey and Grangier, David and Auli, Michael},
  journal={arXiv preprint arXiv:1806.00187},
  year={2018}
}