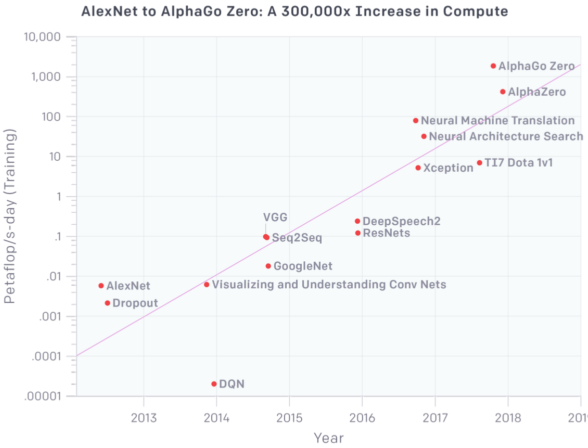
DAY 3 : TOWARDS SCALABLE DEEP LEARNING

Distributed Training with Large Data and Scaling

2021-02-03 | Jenia Jitsev | Cross Sectional Team Deep Learning, Helmholtz AI @ JSC

LARGE NETWORKS, LARGE DATASETS

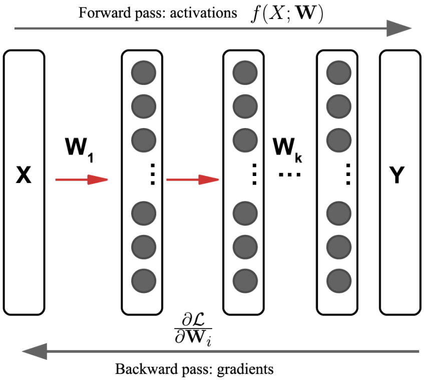
- Training models that solve complex, real world tasks requires large data

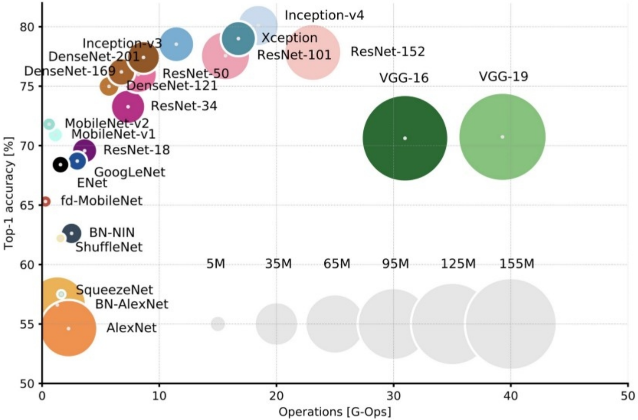


Amodei et al, 2019 (Source: https://openai.com/blog/ai-and-compute/)

LARGE NETWORKS, LARGE DATASETS

- Networks : large models, many layers, many weights
  - ResNet, DenseNet, EfficientNet, Transformers
  - hundreds of layers, hundred millions of parameters or more

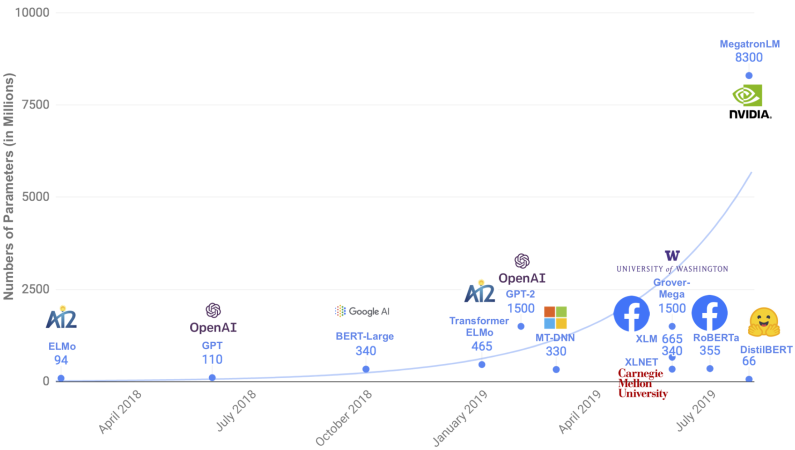




Canziani et al, 2017

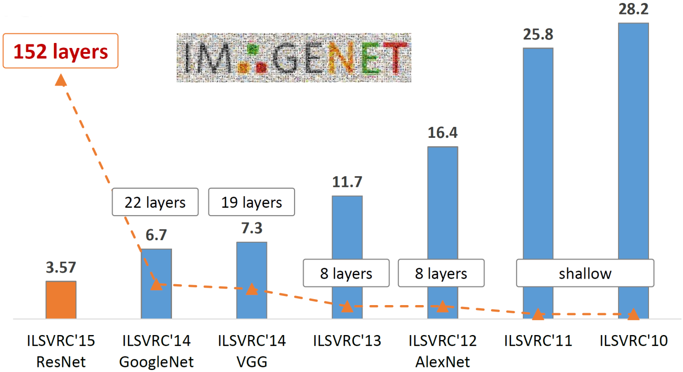
LARGE NETWORKS, LARGE DATASETS

- Networks : large models, many layers, many weights
  - ResNet, DenseNet, EfficientNet, Transformer
  - hundreds of layers, millions of parameters (GPT-3: 175 Billion)



LARGE NETWORKS, LARGE DATASETS

- Millions, even Billions of network parameters: training demands data
- Most breakthroughs happened on large data
  - Vision: ImageNet-1k (1.4 M images); ImageNet-21k (14 M images, $\approx$ 4 TB uncompressed)
  - Language: LM1B, 1 Billion Word Language Model Benchmark
- Datasets get larger and larger
  - JFT-300 (300 M images); YouTube-8M, 8 Million videos, 300 TB
  - Common Crawl dataset : 280 TB uncompressed text, ca. trillion words (as of 2020)



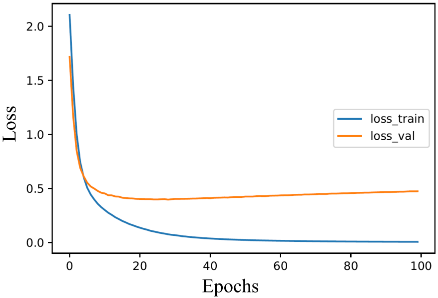
LARGE NETWORKS, LARGE DATASETS

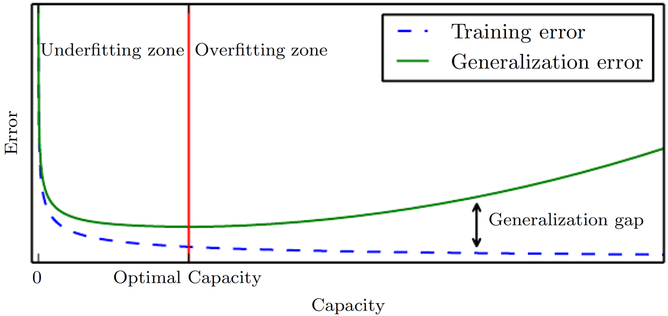
- Millions, even Billions of network parameters: training demands data
- Most breakthroughs happened on large data
- Both network models and datasets get larger and will continue to grow
    - JFT-300 (300 M images); YouTube-8M, 8 Million videos, 300 TB
    - Common Crawl: 280 TB uncompressed text, ca. trillion words;
        - GPT-3 Transformer: 175 Billion weights (350 GB required to train)

	Data Set	**Type**	**Task**	**Size**
small	MNIST	Image	Classification	55,000
	Fashion MNIST	Image	Classification	55,000
	CIFAR-10	Image	Classification	45,000
large	ImageNet	Image	Classification	1,281,167
	Open Images	Image	Classification (multi-label)	4,526,492
	LM1B	Text	Language modeling	30,301,028
	Common Crawl	Text	Language modeling	~25.8 billion

RECONCILING LARGE MODELS AND GENERALIZATION

- Both network models and datasets get larger and will continue to grow
  - Generalization: large models and the generalization gap

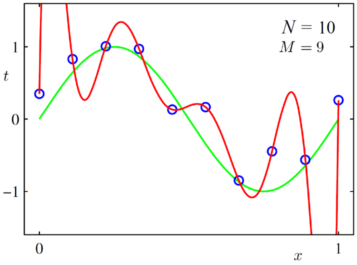


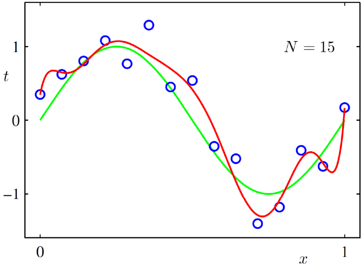


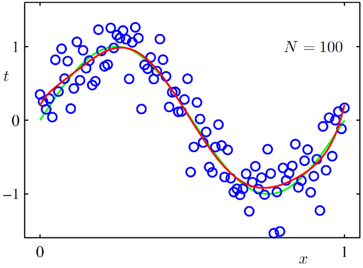
Goodfellow et al, 2016

RECONCILING LARGE MODELS AND GENERALIZATION

- A (classical) simple view - more data, better generalization



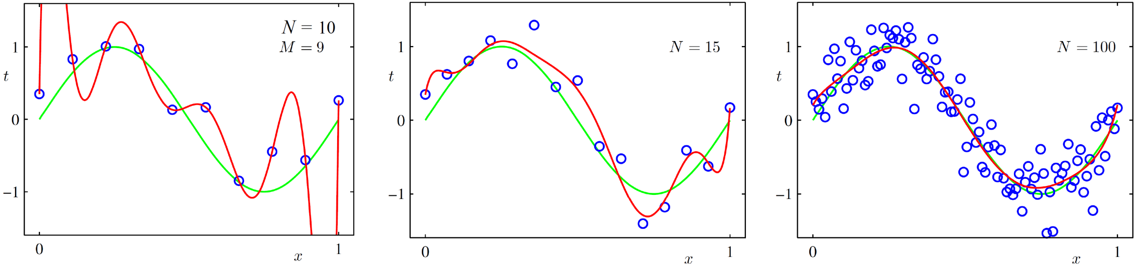




Bishop, 2006

RECONCILING LARGE MODELS AND GENERALIZATION

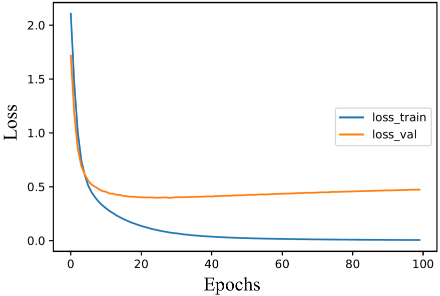
- A (classical) simple view - more data, better generalization
  - Never enough data in higher dimensions - curse of dimensionality

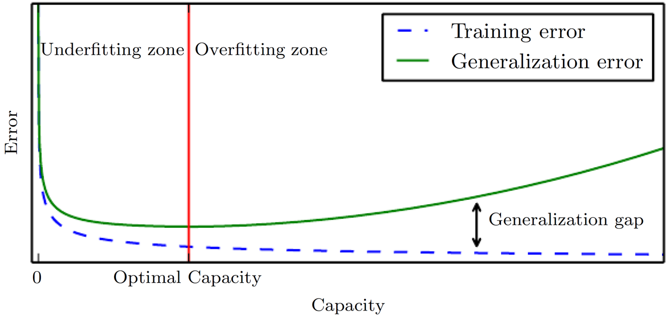


Bishop, 2006

RECONCILING LARGE MODELS AND GENERALIZATION

- A (very recent) complex view - larger models, better generalization

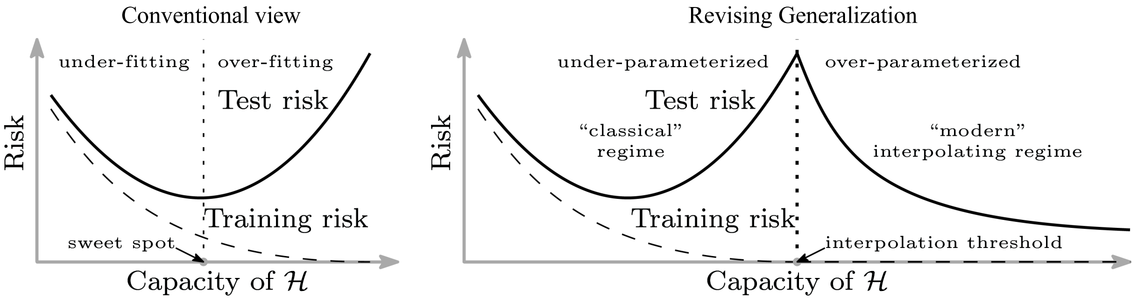




Goodfellow et al, 2016

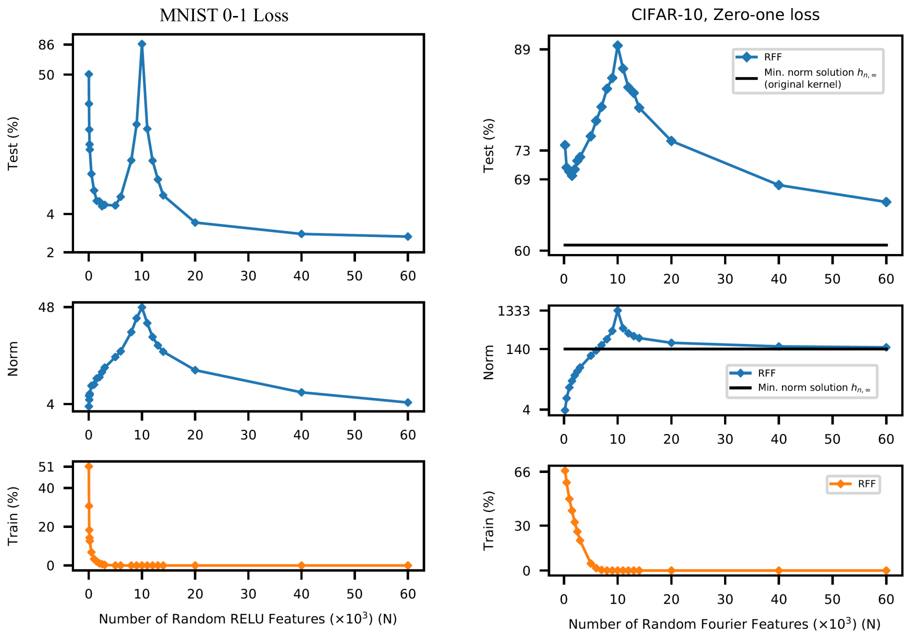
RECONCILING LARGE MODELS AND GENERALIZATION

- A (very recent) complex view - larger models, better generalization
  - **Double descent** test error curve, going beyond **interpolation threshold**
  - Greatly increasing number of model parameters **reduces** generalization gap



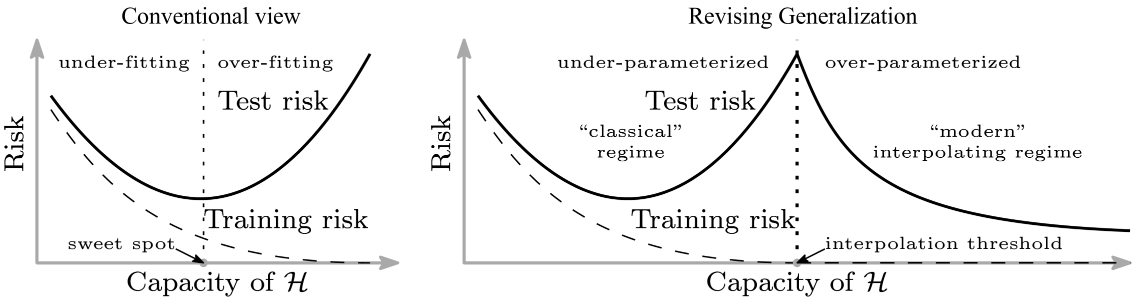
Belkin et al, PNAS, 2019

RECONCILING GENERALIZATION GAP



RECONCILING LARGE MODELS AND GENERALIZATION

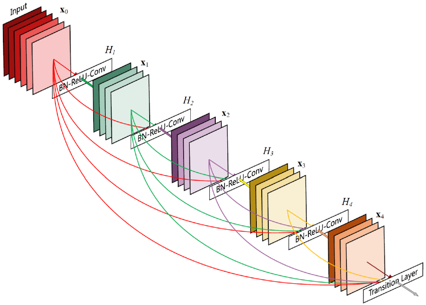
- Larger models generalize better
  - Greatly increasing number of model parameters **reduces** generalization gap

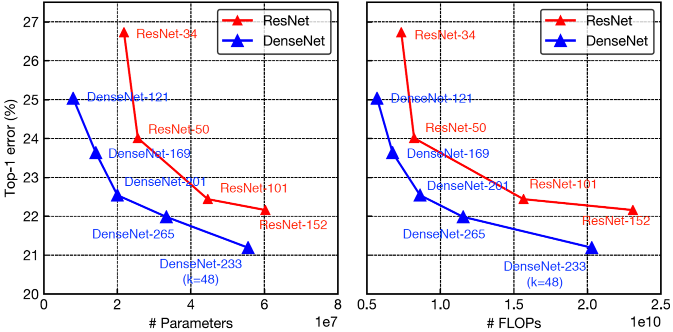


Belkin et al, PNAS, 2019

LARGE MODELS AND GENERALIZATION

- Larger models generalize better
  - Evidence across different large scale training scenarios

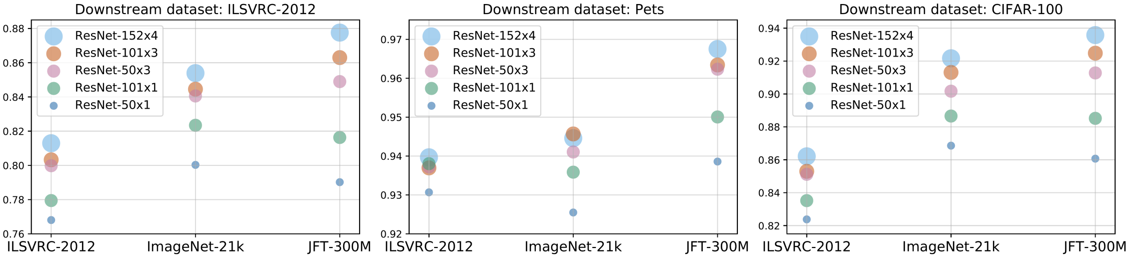




Huang et al, CVPR, 2017

LARGE MODELS AND GENERALIZATION

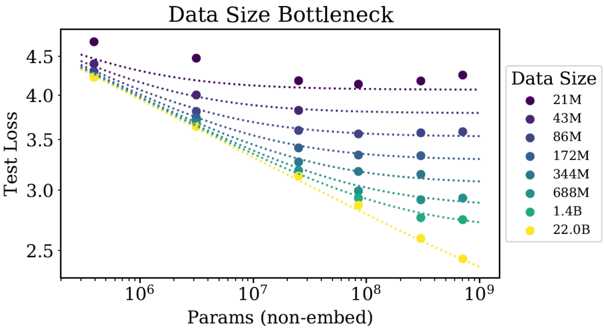
- Larger models transfer better
  - Evidence across different large scale training scenarios

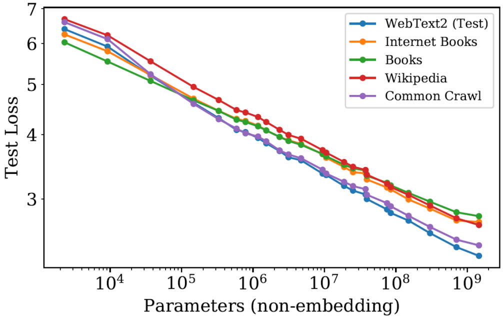


Kolesnikov et al, ECCV, 2020

LARGE MODELS AND GENERALIZATION

- Larger models generalize & transfer better
  - Evidence across different large scale training scenarios

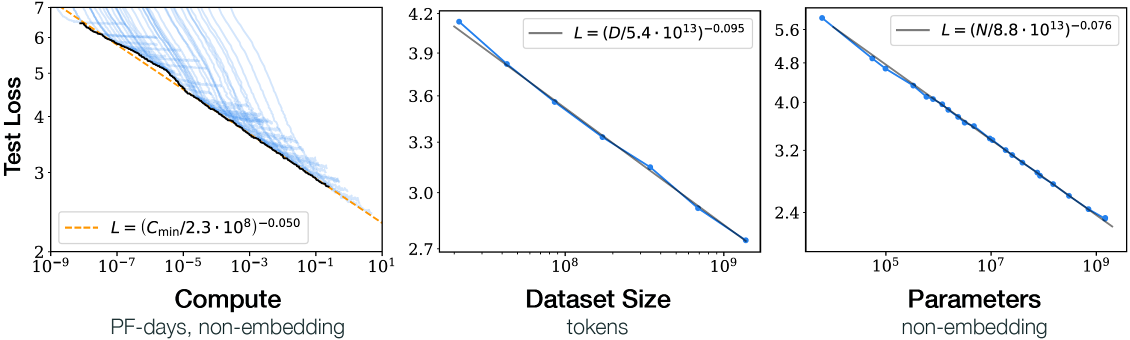




Kaplan et al, 2020; Brown et al, NeurIPS, 2020

LARGE MODELS AND LARGE DATA

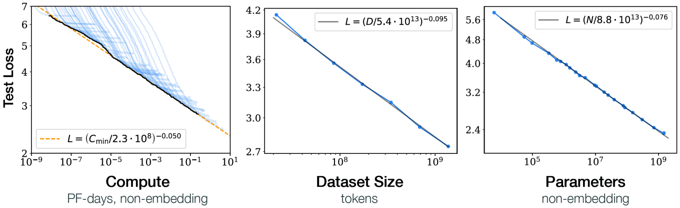
- Scaling Laws: increasing model size and data increases generalization

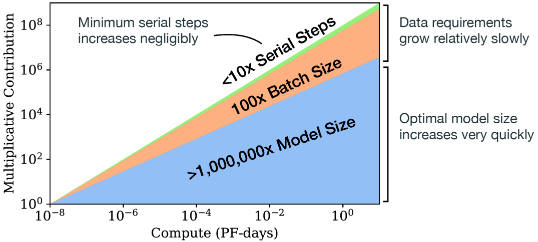


Kaplan et al, 2020

LARGE MODELS AND LARGE DATA

- Scaling Laws: given sufficient compute budget, increasing both model size and data size is the way to further strongly boost generalization

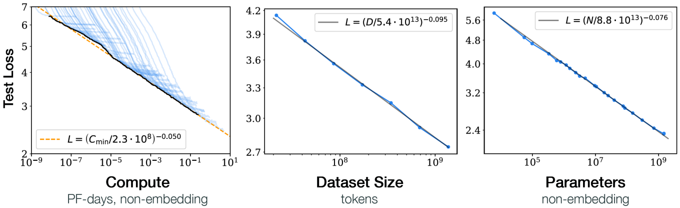


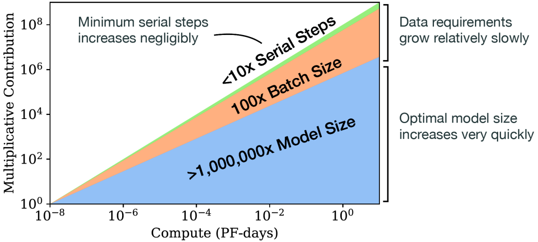


Kaplan et al, 2020

LARGE MODELS AND LARGE DATA

- Increasing model size is **good** idea, provided enough compute and data





Kaplan et al, 2020

DISTRIBUTED TRAINING WITH LARGE DATA

- ImageNet: transition to modern deep learning era;
  - outstanding effort in large data collection (Fei-Fei et al, Stanford)
  - building dataset via crowdsourcing over 4 years

MNIST, CIFAR-10/100
28x28, 32x32; 60k examples

ImageNet-1k, 21k; OpenImages, FFHQ…
224x224, 1024x1024; 1.2M examples

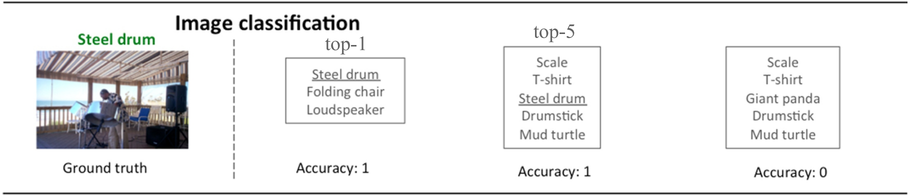
DISTRIBUTED TRAINING ON IMAGENET

- Full dataset (ImageNet-21k) : 14M images, 21k classes labeled
- ImageNet-1k : dataset for ILSVRC competition (2010 - 2017), 1k classes
  - 1.28M Training, 100k Test, 50k Validation sets
  - usual image resolution used for training: 224x224
  - current accuracies : $>$ 88% top-1, $>$ 97% top-5

Russakovsky et al, IJCV, 2015

DISTRIBUTED TRAINING ON IMAGENET

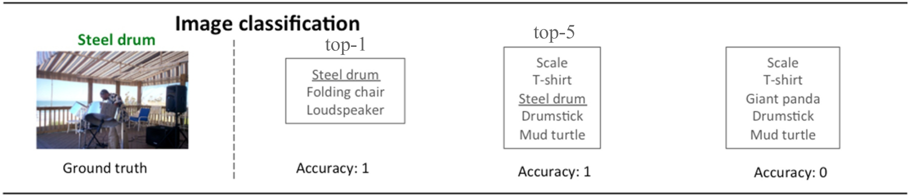
- Full dataset (ImageNet-21k) : 14M images, 21k classes labeled
- ImageNet-1k : dataset for ILSVRC competition (2010 - 2017), 1k classes
  - 1.28M Training, 100k Test, 50k Validation sets
  - usual image resolution used for training: 224x224
  - current accuracies : $>$ 88% top-1, $>$ 97% top-5



Russakovsky et al, IJCV, 2015

DISTRIBUTED TRAINING ON IMAGENET

- ImageNet-1k : still gold standard in training large visual recognition models
  - pre-trained models: transfer learning on more specific smaller datasets
- ResNet-50 : baseline model network, accuracies : $\approx$ 75% top-1, $\approx$ 94% top-5 (Winner ILSVRC 2015)

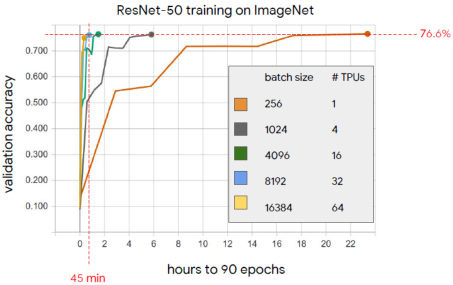


Russakovsky et al, IJCV, 2015

DISTRIBUTED TRAINING ON IMAGENET

- ResNet-50 : efficient distributed training in data parallel mode possible
  - 25M weights, 103Mb for activations, model training on 224x224 ImageNet-1k
  - $\approx$ 4 GB Memory with $B_{ref} = 64$: fits onto single GPU

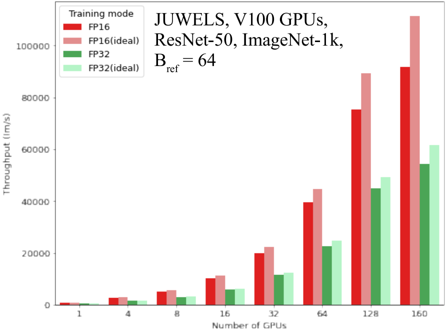
	Batch Size	Processor	DL Library	Time	Accuracy
He et al. [1]	256	Tesla P100 × 8	Caffe	29 hours	75.3 %
Goyal et al. [2]	8,192	Tesla P100 × 256	Caffe2	1 hour	76.3 %
Smith et al. [3]	8,192 → 16,384	full TPU Pod	TensorFlow	30 mins	76.1 %
Akiba et al. [4]	32,768	Tesla P100 × 1,024	Chainer	15 mins	74.9 %
Jia et al. [5]	65,536	Tesla P40 × 2,048	TensorFlow	6.6 mins	75.8 %
Ying et al. [6]	65,536	TPU v3 × 1,024	TensorFlow	1.8 mins	75.2 %
Mikami et al. [7]	55,296	Tesla V100 × 3,456	NNL	2.0 mins	75.29 %
This work	**81,920**	**Tesla V100 × 2,048**	**MXNet**	**1.2 mins**	**75.08%**

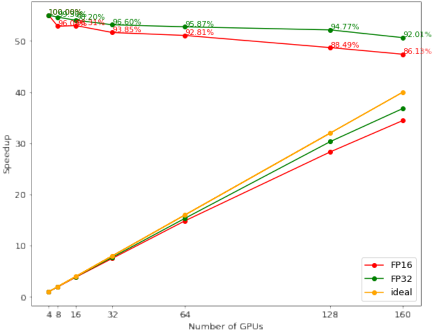


Yamazoto et al, 2019; Ying, 2018

DISTRIBUTED TRAINING ON IMAGENET

- Efficient distributed training in data parallel mode
  - requires good scaling of throughput Images/sec during training
  - image throughput during training ideally increasing as $\tau_K^* = K \cdot \tau_{ref}$ Images/sec

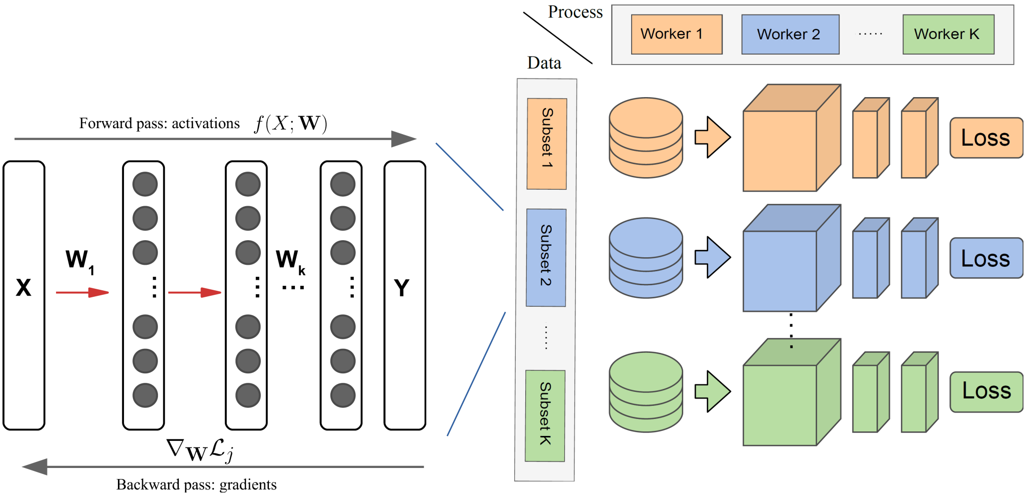




Cherti et al, 2020

DISTRIBUTED TRAINING ON IMAGENET

- Efficient distributed training in data parallel mode
  - requires good scaling of throughput Images/sec during training



DISTRIBUTED TRAINING ON IMAGENET

- Efficient distributed training in data parallel mode

Data IO

- Efficient file system, efficient data container
  - few separate large files; **sequential access**
  - LMDB, HDF5, TFRecords
- Efficient Data pipeline
  - eg tf.data : interleave, cache, prefetch, . . .
  - avoid GPU starvation

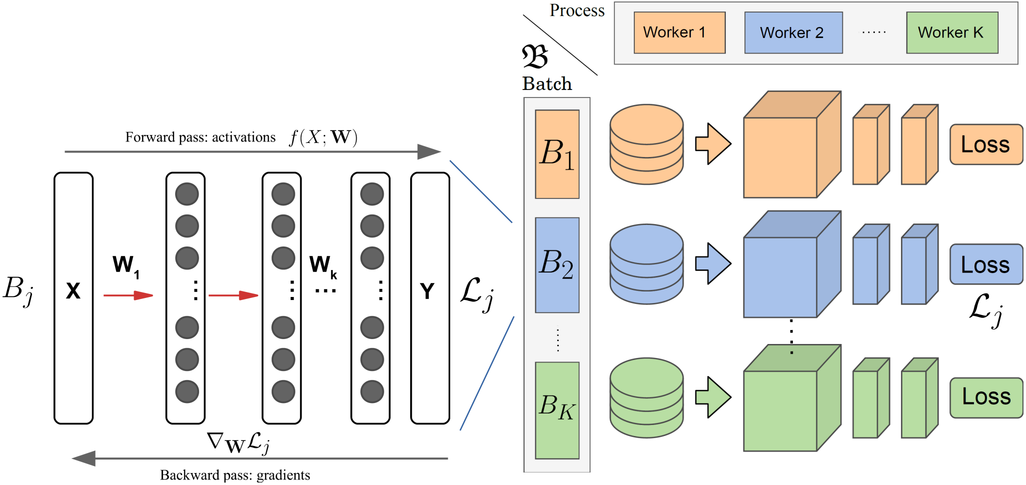
```
...

141M /p/largedata/cstdl/ImageNet/imagenet-processed/train-00171-of-01024
137M /p/largedata/cstdl/ImageNet/imagenet-processed/train-00172-of-01024
139M /p/largedata/cstdl/ImageNet/imagenet-processed/train-00173-of-01024
142M /p/largedata/cstdl/ImageNet/imagenet-processed/train-00174-of-01024

...
```

DISTRIBUTED TRAINING ON IMAGENET

- Efficient distributed training in data parallel mode
  - requires efficient balance of GPU gradient compute and communication



DISTRIBUTED TRAINING ON IMAGENET

- Efficient distributed training in data parallel mode possible

SGD Optimization

- Make sure model fits into GPU memory
  - remember: this also depends on worker's batch size $|B_{\text{ref}}|$ and input image resolution
- Avoid internode communication overhead & bottlenecks
  - Most compute for forward-backward passes
  - $|B_{\text{ref}}|$ per GPU not too small
  - High capacity network: InfiniBand
  - Horovod: additional mechanisms, eg. Tensor Fusion
- Corresponds to training single model with a larger effective batch size $|\mathfrak{B}| = K \cdot |B_{\text{ref}}|$
  - Image Throughput ideally increasing as $\tau_K = K \cdot \tau_{ref}$ Images/sec

DISTRIBUTED TRAINING ON IMAGENET

- ResNet-50 : efficient distributed training in data parallel mode on ImageNet-1k
- Ultimate aim: reducing training **time to accuracy**
  - increasing throughput Images/sec during training only intermediate station!

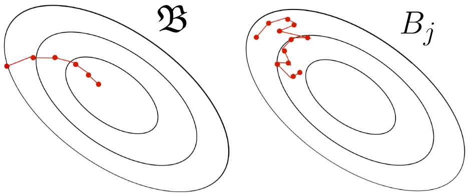
	Batch Size	Processor	DL Library	Time	Accuracy
He et al. [1]	256	Tesla P100 $\times$ 8	Caffe	29 hours	75.3 %
Goyal et al. [2]	8,192	Tesla P100 $\times$ 256	Caffe2	1 hour	76.3 %
Smith et al. [3]	8,192 $\rightarrow$ 16,384	full TPU Pod	TensorFlow	30 mins	76.1 %
Akiba et al. [4]	32,768	Tesla P100 $\times$ 1,024	Chainer	15 mins	74.9 %
Jia et al. [5]	65,536	Tesla P40 $\times$ 2,048	TensorFlow	6.6 mins	75.8 %
Ying et al. [6]	65,536	TPU v3 $\times$ 1,024	TensorFlow	1.8 mins	75.2 %
Mikami et al. [7]	55,296	Tesla V100 $\times$ 3,456	NNL	2.0 mins	75.29 %
This work	**81,920**	**Tesla V100 $\times$ 2,048**	**MXNet**	**1.2 mins**	**75.08%**

Yamazoto et al, 2019

DISTRIBUTED TRAINING ON IMAGENET

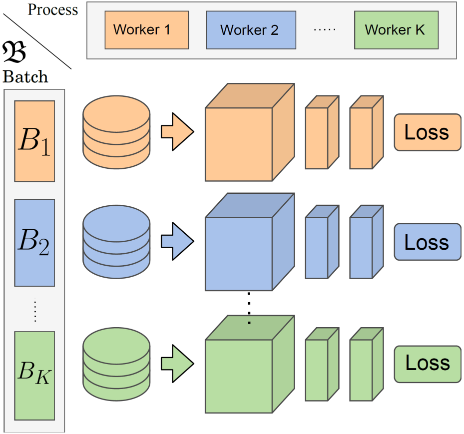
SGD Optimization

- Large effective batch size $|\mathfrak{B}|$ may require hyperparameter retuning
  - Reminder: Large effective batch sizes alter optimization



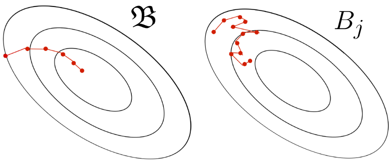
Effective larger batch, over all K workers

Reference small batch, per single worker



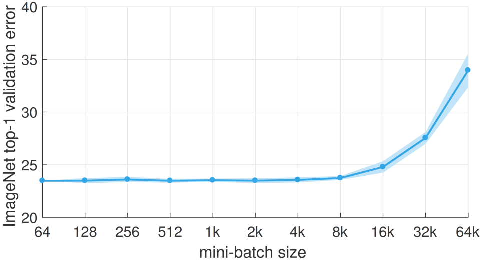
DISTRIBUTED TRAINING ON IMAGENET

- Efficient distributed training in data parallel mode
- Large effective batch sizes may require hyperparameter re-tuning
  - learning rate and schedule
  - optimizer type
- Reminder: hyperparameter tuning for a given $|\mathfrak{B}|$ - on the validation set!



Effective larger batch, over all K workes

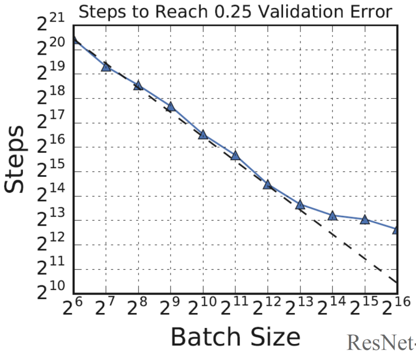
Reference small batch, per single worker

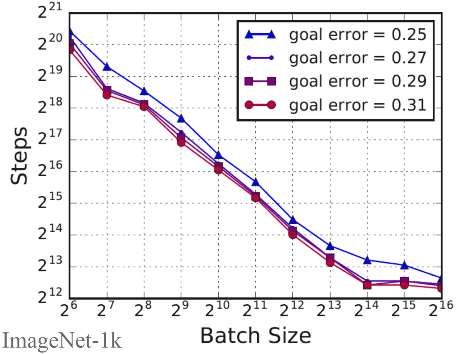


Goyal et al, 2017

DISTRIBUTED TRAINING ON IMAGENET

- Efficient distributed training in data parallel mode
  - Outlook: coping with training on large effective batch sizes
  - Reducing training **time to accuracy**





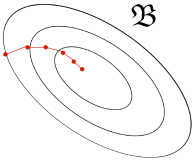
ResNet-50, ImageNet-1k

Shallue et al, JMLR, 2019

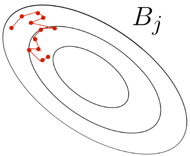
LARGE MODELS, LARGE DATA

Summary

- Reconciling generalization: large models generalize better
  - given enough data and compute to train
- Efficient data parallel training on large datasets like ImageNet-1k : possible
- Data pipelines, Horovod, InfiniBand and large batch sizes pave the way
- Measures to stabilize training with large batches - upcoming lectures



Effective larger batch,
over all K workes



Reference small batch,
per single worker