DAY 3: TOWARDS SCALABLE DEEP LEARNING

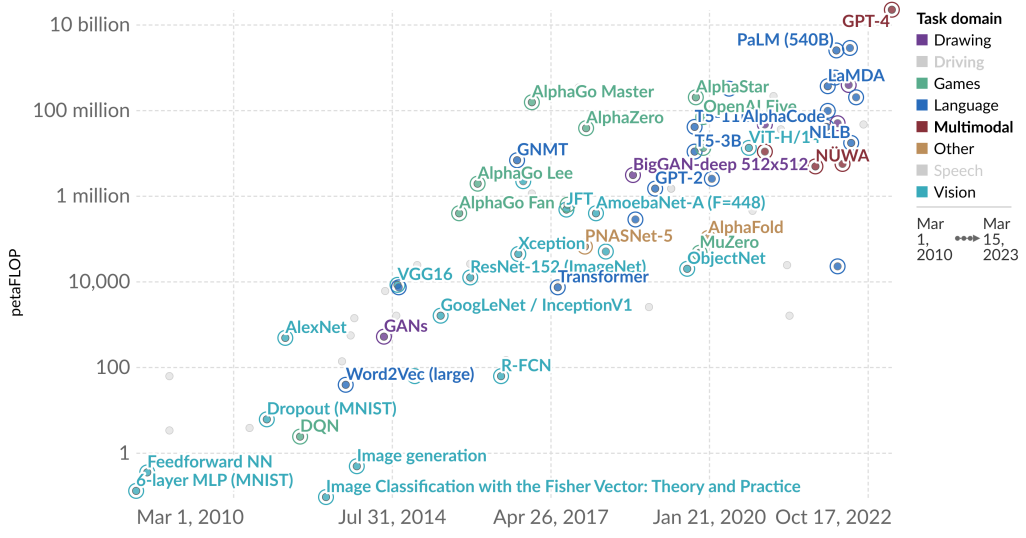
Scaling Laws and Training with Large Data

2023-04-10 | Jenia Jitsev | Scalable Learning & Multi-Purpose AI Lab, Helmholtz AI, LAION @ JSC

JÜLICH
Forschungszentrum

LARGE NETWORKS, LARGE DATASETS

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

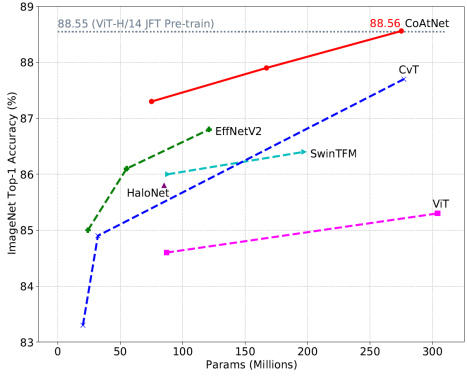


Sevilla et al., 2023; OurWorldInData.org/artificial-intelligence; CC BY Source

LARGE NETWORKS, LARGE DATASETS

- Networks : large models, many layers, large number of parameters (weights)
  - Vision: Convolutional, Transformer and Hybrid networks
  - hundreds of layers, hundred millions of parameters (currently up to 20B)

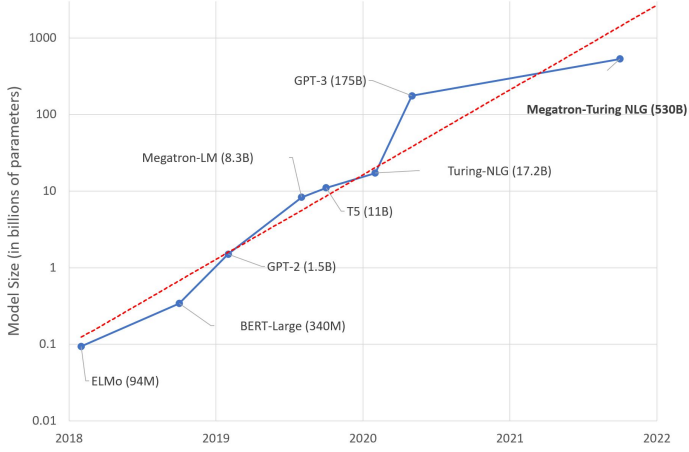
Models	Eval Size	#Params	#FLOPs	TPUv3-core-days	Top-1 Accuracy
ResNet + ViT-L/16	384^2	330M	-	-	87.12
ViT-L/16	512^2	307M	364B	0.68K	87.76
ViT-H/14	518^2	632M	1021B	2.5K	88.55
NFNet-F4+	512^2	527M	367B	1.86K	89.2
CoAtNet-3†	384^2	168M	114B	0.58K	88.52
CoAtNet-3†	512^2	168M	214B	0.58K	88.81
CoAtNet-4	512^2	275M	361B	0.95K	89.11
CoAtNet-5	512^2	688M	812B	1.82K	89.77
ViT-G/14	518^2	1.84B	5160B	>30K°	90.45
CoAtNet-6	512^2	1.47B	1521B	6.6K	90.45
CoAtNet-7	512^2	2.44B	2586B	20.1K	**90.88**



Dai et al, NeurIPS, 2021

LARGE NETWORKS, LARGE DATASETS

- Networks : large models, many layers, large number of parameters (weights)
  - Language: Transformer networks
  - hundreds of layers, billions of parameters (GPT-3: 175 Billion)



Narayanan et al, 2021

LARGE NETWORKS, LARGE DATASETS

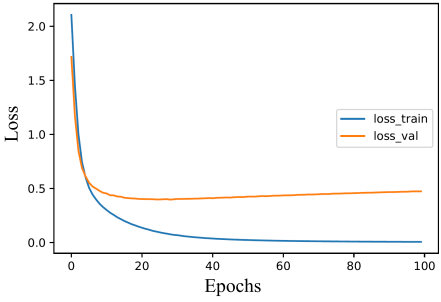
- Millions, Billions of network parameters: training demands data
- Most breakthroughs happened on large data; datasets for model training get larger and larger
  - Vision, Supervised, Self-Supervised: ImageNet-1k (1.4 M images); ImageNet-21k (14 M images, $\approx$ 1.4 TB compressed); JFT-300M/4B (300M/4B images); YouTube-8M (8 Million videos, 300 TB)
  - Language-Vision, Self-Supervised
    - CLIP trained on WIT-400M (400M image-text pairs); openCLIP on LAION-400M/5B (open data, 400M/5B image-text pairs, 11TB/240TB)
    - Stable Diffusion trained on LAION-5B
  - Language, Self-supervised
    - GPT-3 trained on 300-400 Billion word tokens
    - LLaMA, RedPajama (open) : 5 TB uncompressed text, ca. 1.2 trillion tokens

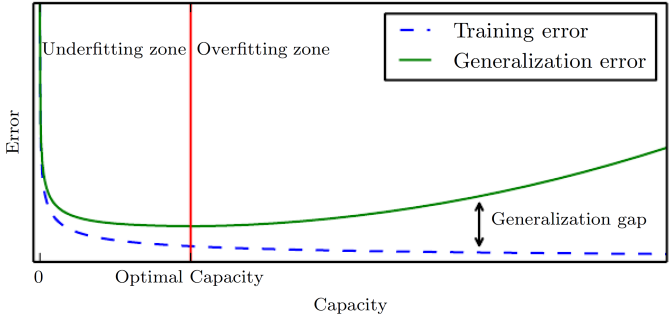
modern era

Data Set	Type	Task	Samples	Memory
MNIST (1998)	Image	Classification	60K	12MB
CIFAR-10,100 (2009)	Image	Classification	60K	160MB
ImageNet-1k/21k	Image	Classification	1.4M/14M	0.13TB/1.3TB
Pile	Text	Language modeling	$\sim$186B	0.8TB
LAION-400M/5B	Image-Text	Multi-Modal Learning	400M/5B	11TB/230TB
LLaMA/RedPajama	Text	Language modeling	$\sim$1.2T	5TB

RECONCILING LARGE MODELS AND GENERALIZATION

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

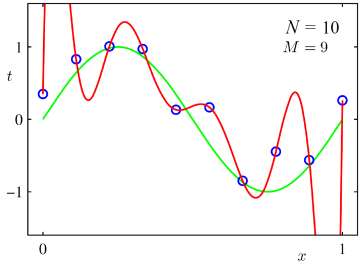


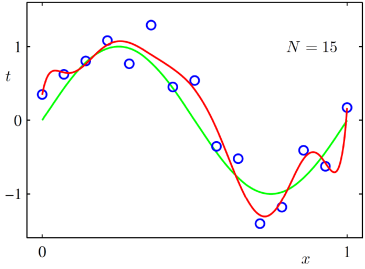


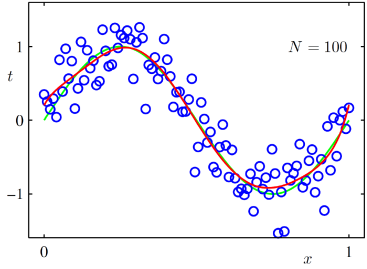
Goodfellow et al, 2016

RECONCILING LARGE MODELS AND GENERALIZATION

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



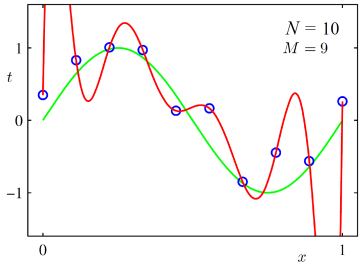


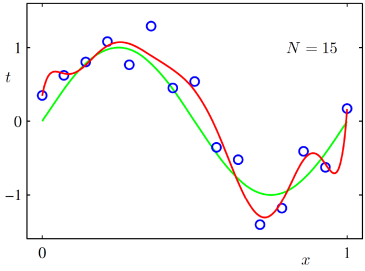


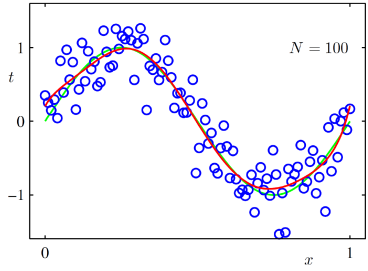
Bishop, 2006

RECONCILING LARGE MODELS AND GENERALIZATION

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



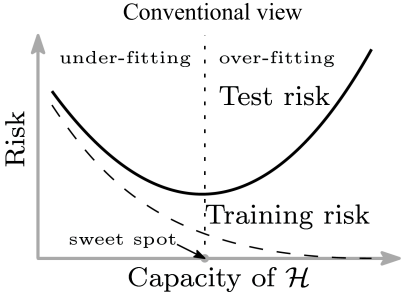


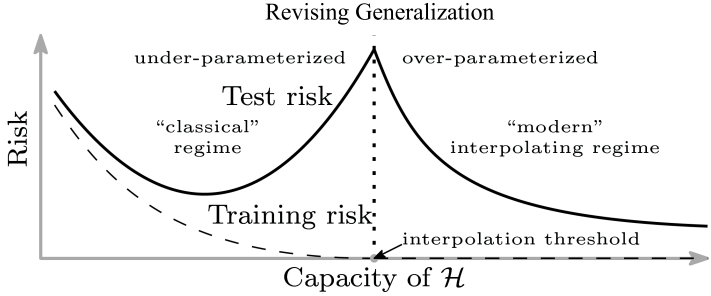


Bishop, 2006

RECONCILING LARGE MODELS AND GENERALIZATION

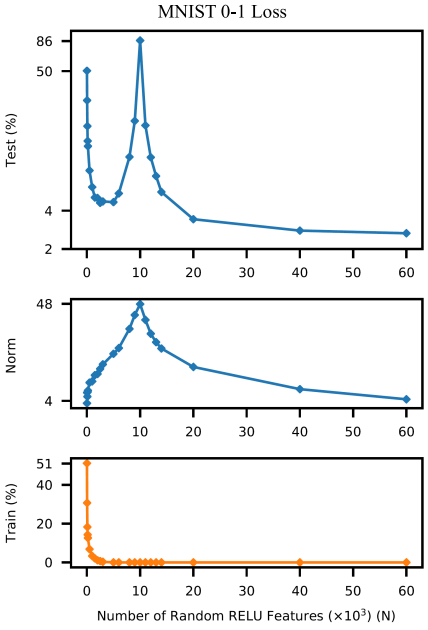
- A (very recent) complex view - larger models, better generalization
  - **Double descent** test error curve, going beyond **interpolation threshold**
  - Greatly increasing number of model parameters **reduces** generalization gap

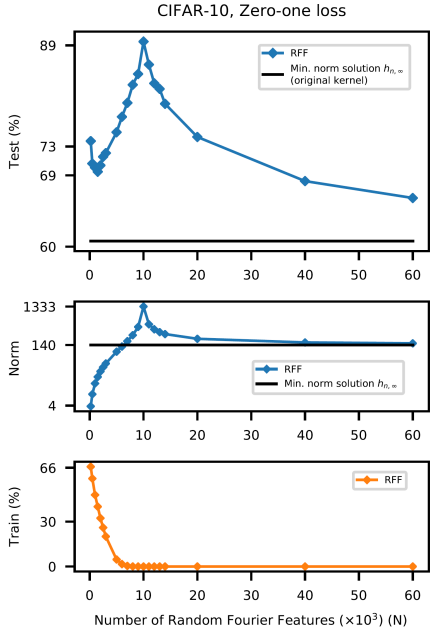




Belkin et al, PNAS, 2019

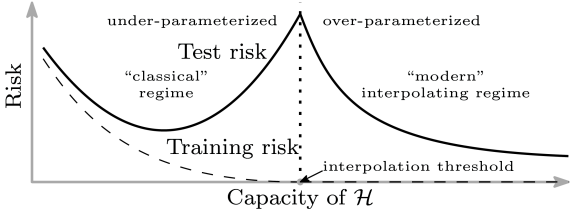
RECONCILING GENERALIZATION GAP

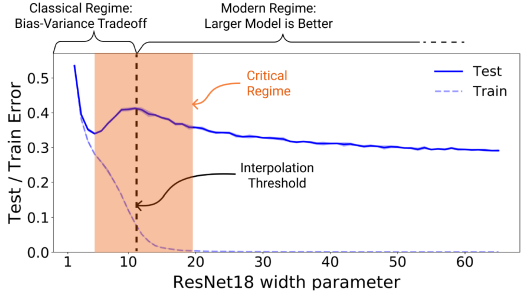




RECONCILING LARGE MODELS AND GENERALIZATION

- Larger models generalize better
  - Reconciling generalization - large, overparameterized models generalize strongly
  - Greatly increasing number of model parameters **reduces** generalization gap
  - **Double descent** test error curves

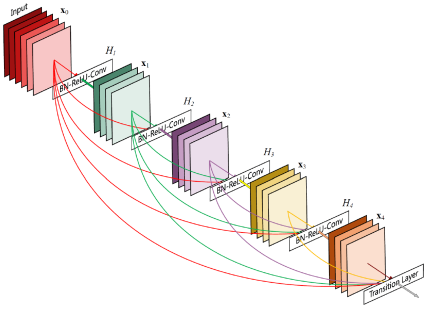


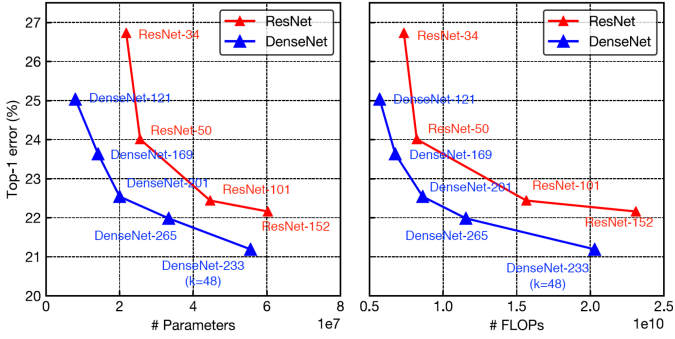


Belkin et al, PNAS, 2019, Nakkiran et al, ICLR, 2020

LARGE MODELS AND GENERALIZATION

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

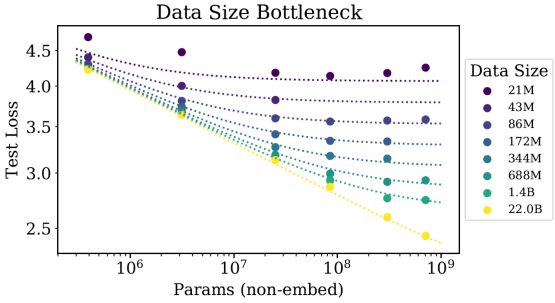


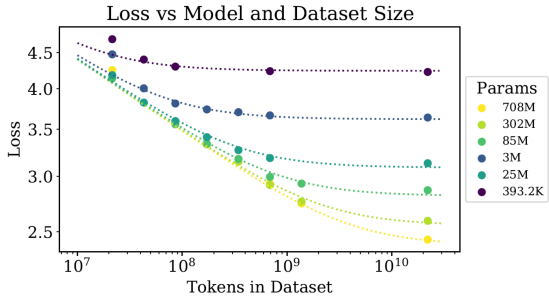


Huang et al, CVPR, 2017

LARGE MODELS AND GENERALIZATION

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

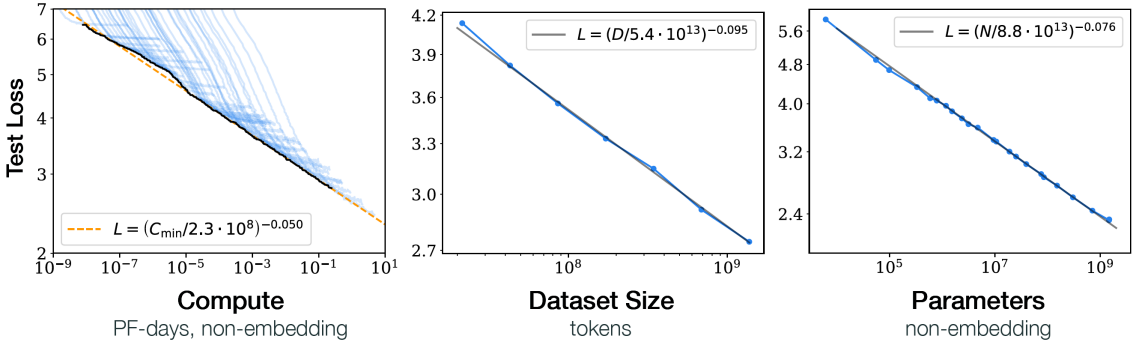




Kaplan et al, 2020; Brown et al, NeurIPS, 2020

LARGE MODELS AND LARGE DATA

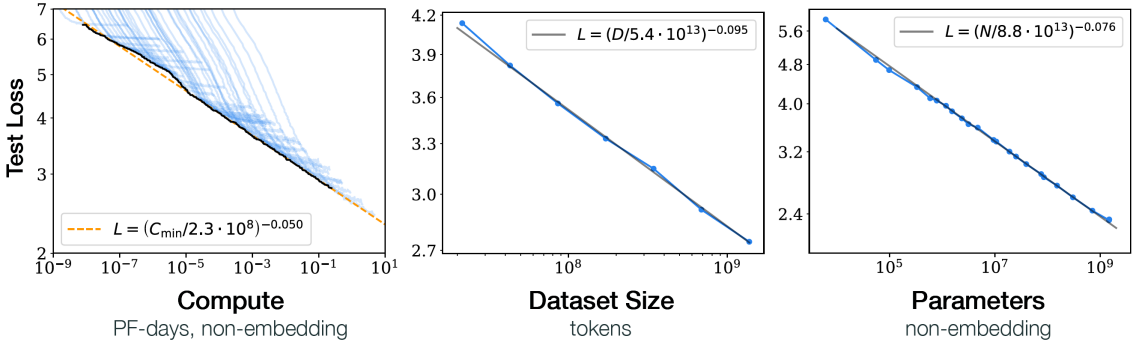
- Scaling Laws: given sufficient compute budget, increasing both model size and data size is the way to further strongly boost generalization



Kaplan et al, 2020

LARGE MODELS AND LARGE DATA

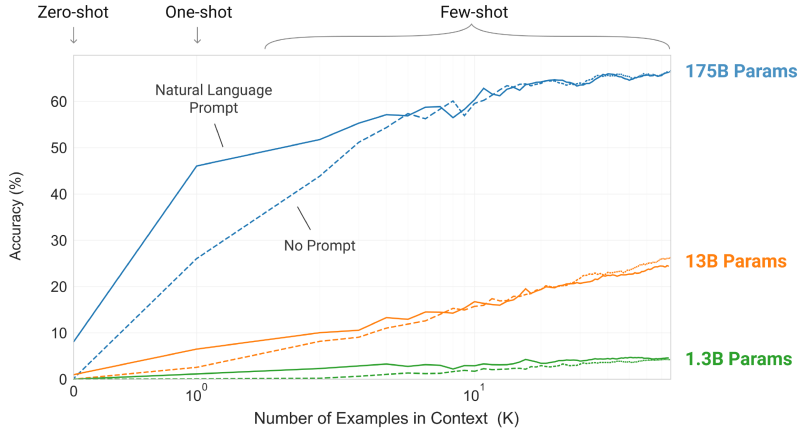
- Increasing model size is **good** idea, provided enough compute and data



Kaplan et al, 2020

LARGE MODELS, DATA AND GENERALIZATION

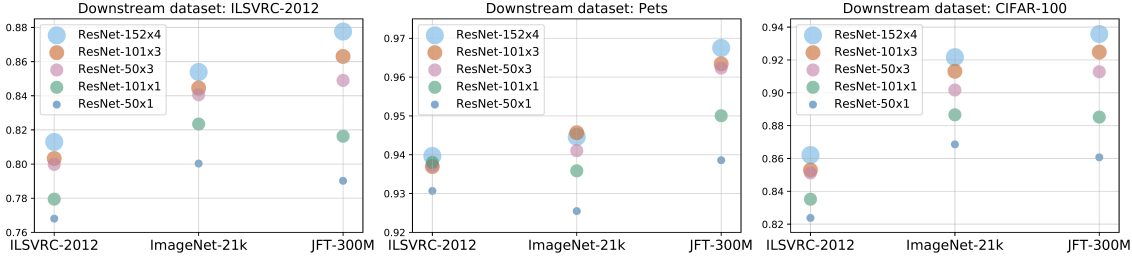
- Language Modeling : very large models, very large data, generic self-supervised pre-training (autoregressive generative sequence models)
  - GPT-3, trained on Common Crawl & co (300-400B word token samples)
  - Strong few-show and zero-shot transfer at largest scale (175B params)



Brown et al, NeurIPS, 2020

LARGE MODELS, DATA AND GENERALIZATION

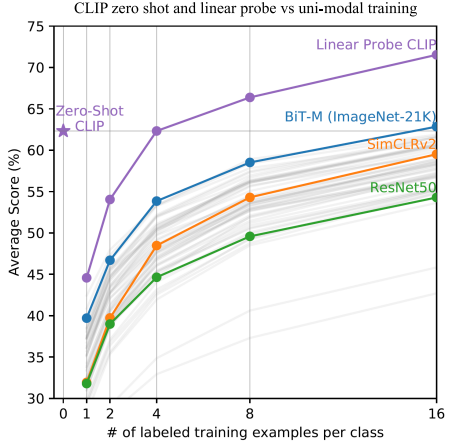
- Larger models transfer better
  - Evidence across different large scale training scenarios
  - Using large models (BiT - Big Transfer, ResNet-152x4: 928M params), large data
    - ImageNet-21k, $\sim$ 14M images (instead of standard ImageNet-1k, $\sim$ 1.4M)
    - JFT-300M : $\approx$ 18K classes, noisy labels, 300x larger than ImageNet-1k
  - Pre-training a single large model: **81 hours** with **256 A100** GPUs (20k GPU hours; ImageNet-21k, JUWELS Booster)

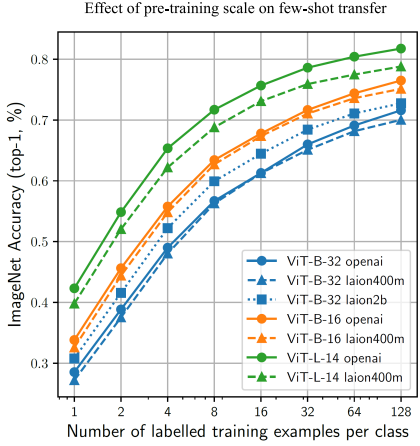


BiT - Big Transfer, Kolesnikov et al, ECCV, 2020

LARGE MODELS, DATA AND GENERALIZATION

- Self-supervised language-vision pre-training: GPT for multi-modal image-text data
  - CLIP: very strong zero- and few-shot transfer across various targets
  - very large data for pre-training (eg. open LAION-400M/5B, image-text pairs)
- Larger model and data scale - better zero- and few-shot transfer

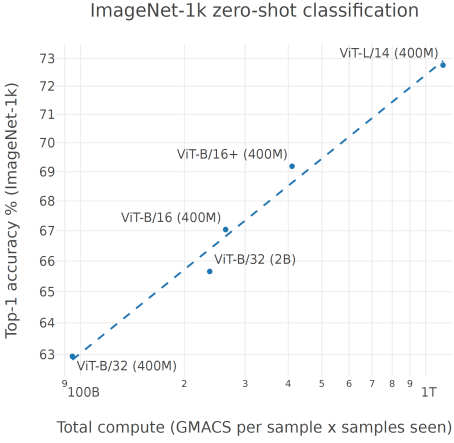


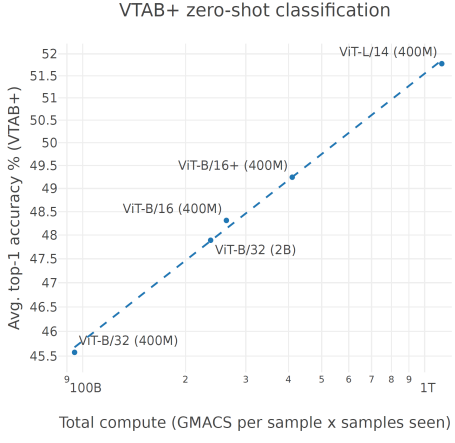


Radford, ICML, 2021; Schumann et al, NeurIPS 2022

LARGE MODELS, DATA AND GENERALIZATION

- Larger model and data scale - better zero-shot transfer
  - **88 hours** with **400 A100** (50K GPU hours) for training of ViT L/14 openCLIP on LAION-400M (JUWELS Booster)





Schumann et al, NeurIPS, 2022

LARGE MODELS, DATA AND GENERALIZATION

Summary

- Theoretical insights suggest revision of model generalization at larger scales
  - generalization can improve with larger model scales
- Scaling laws suggest that larger scales may be one key to strong generalization, model robustness and transferability
- Major breakthroughs in model transferability and robustness in language (GPT) and vision (ViT, CLIP) when using very large model, data and compute scale
- Experiments involving strongly transferable models at larger scale are extremely compute intensive
  - ten or hundred thousands of GPU hours

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

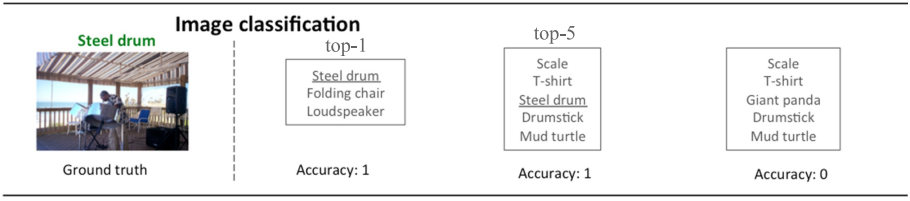
DISTRIBUTED TRAINING ON IMAGENET

- 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

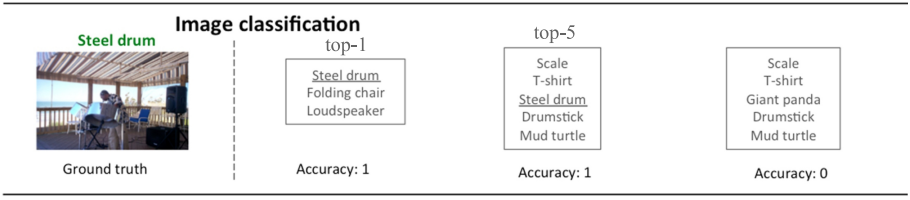
- Full dataset (ImageNet-21k) : 14M images, 21k classes labeled
- ImageNet-1k : dataset for ILSVRC competition (2010 - 2017), 1k classes
  - 1.28M Training, 100k Test, 50k Validation sets
  - usual image resolution used for training: 224x224
  - current accuracies : $>$ 88% top-1, $>$ 97% top-5



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)

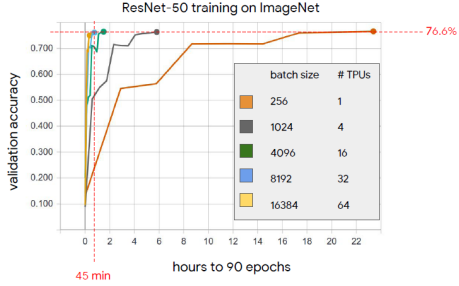


Russakovsky et al, IJCV, 2015

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

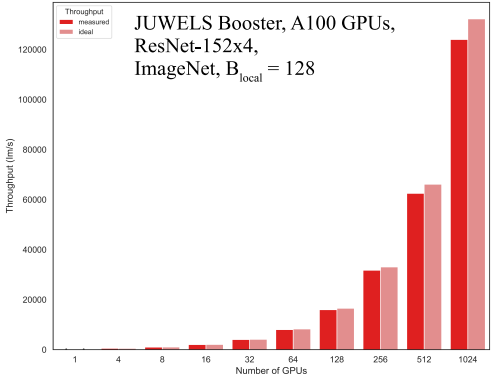
	Batch Size	Processor	DL Library	Time	Accuracy
He et al. [1]	256	Tesla P100 × 8	Caffe	29 hours	75.3 %
Goyal et al. [2]	8,192	Tesla P100 × 256	Caffe2	1 hour	76.3 %
Smith et al. [3]	8,192 → 16,384	full TPU Pod	TensorFlow	30 mins	76.1 %
Akiba et al. [4]	32,768	Tesla P100 × 1,024	Chainer	15 mins	74.9 %
Jia 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%**

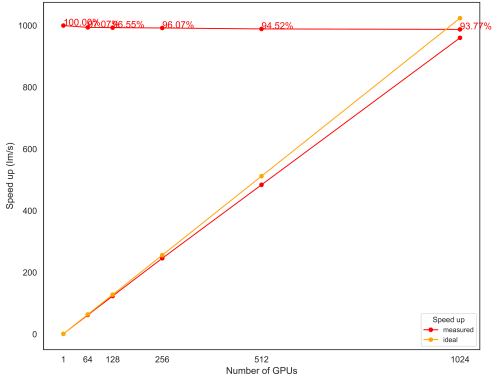


Yamazoto et al, 2019; Ying, 2018

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

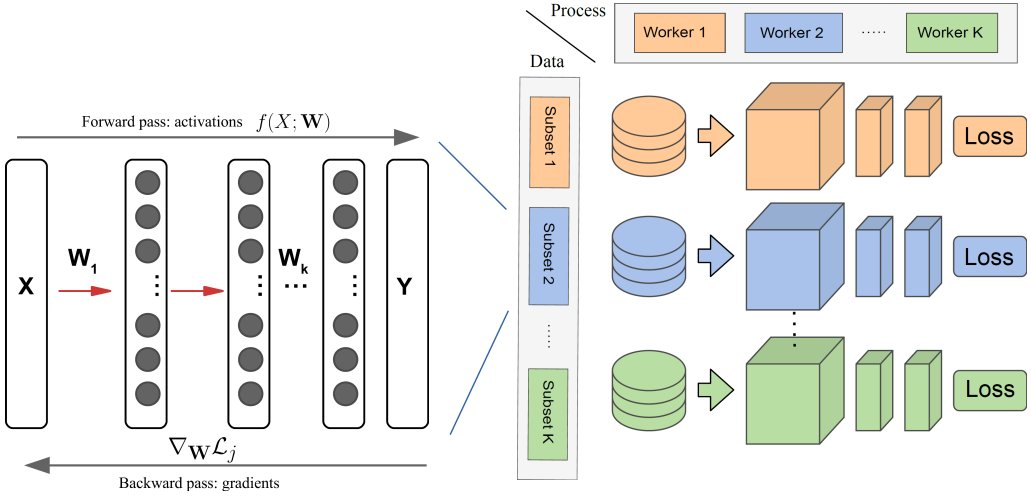




Cherti and Jitsev, arXiv:2106.00116, MedNeurIPS Workshop, 2021

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, WebDataset
- Efficient Data pipeline
  - eg tf.data : interleave, cache, prefetch, . . .
  - avoid GPU starvation

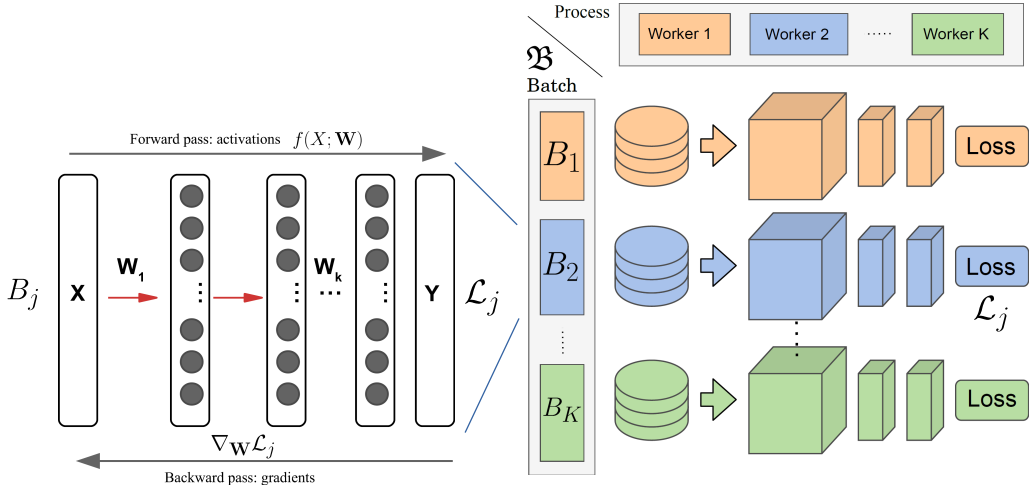
```
...

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

SGD Optimization

- 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
- 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

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!
  - necessary, but not sufficient condition for speeding up model training

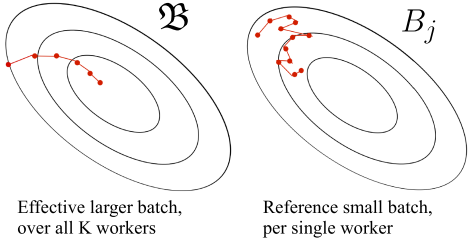
	Batch Size	Processor	DL Library	Time	Accuracy
He et al. [1]	256	Tesla P100 × 8	Caffe	29 hours	75.3 %
Goyal et al. [2]	8,192	Tesla P100 × 256	Caffe2	1 hour	76.3 %
Smith et al. [3]	8,192 → 16,384	full TPU Pod	TensorFlow	30 mins	76.1 %
Akiba et al. [4]	32,768	Tesla P100 × 1,024	Chainer	15 mins	74.9 %
Jia 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%**

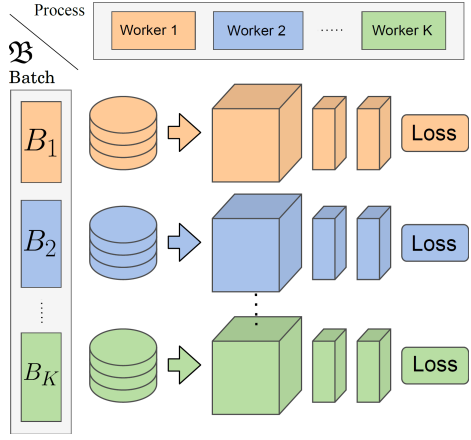
Yamazoto et al, 2019

DISTRIBUTED TRAINING ON IMAGENET

SGD Optimization

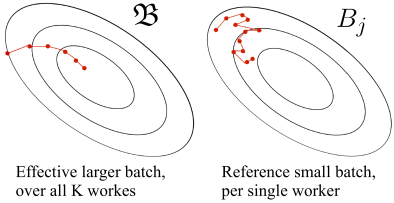
- Large effective batch size $|\mathfrak{B}|$ may require hyperparameter retuning
  - Reminder: Large effective batch sizes alter optimization

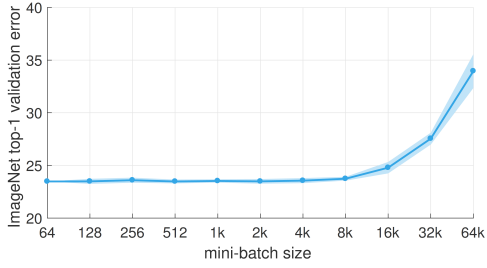




DISTRIBUTED TRAINING ON IMAGENET

- 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!

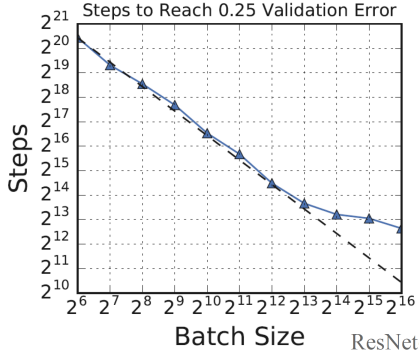


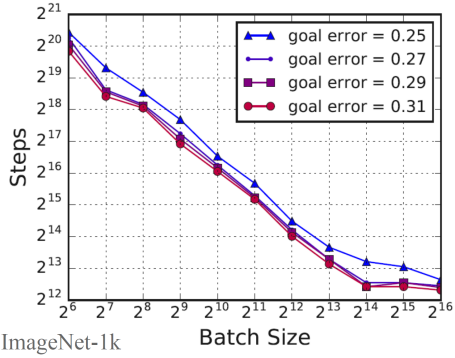


Goyal et al, 2017

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, high bandwidth & low latency (eg InfiniBand), large batch sizes pave the way
- Implementation of efficient distributed training: Horovod, PyTorch DDP, . . .
- Measures to stabilize training with large batches - upcoming lectures