DAY 4: TOWARDS SCALABLE DEEP LEARNING

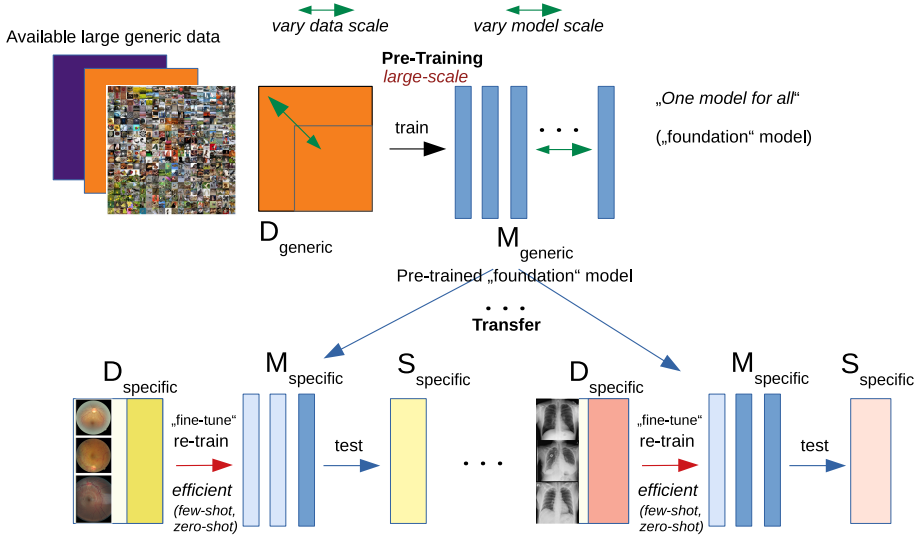
Outlook on Advanced Distributed Training

2023-04-11 | Jenia Jitsev | Scalable Learning & Multi-Purpose AI Lab, Helmholtz AI, LAION @ JSC

JÜLICH
Forschungszentrum

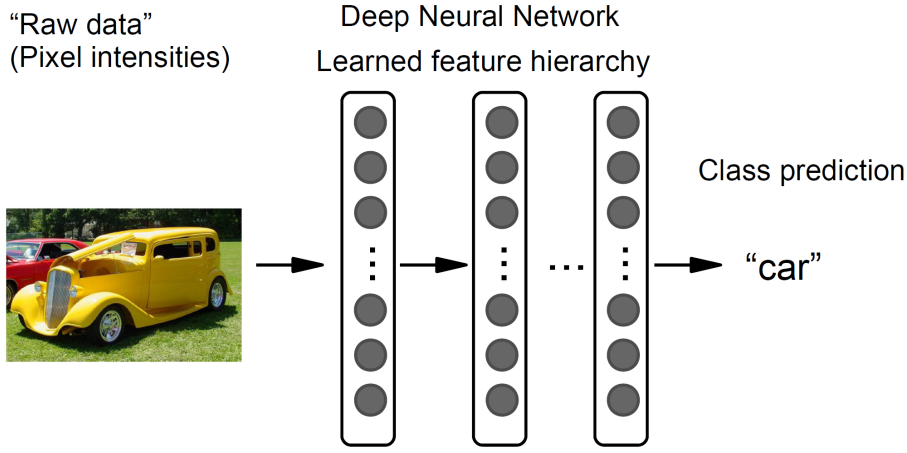
LARGE-SCALE PRE-TRAINING AND TRANSFER

- Foundation models: transferable models pre-trained on large generic data
  - transfer across domains specific smaller datasets and tasks
  - scaling laws: strong, efficient transfer - large models pre-trained on large data



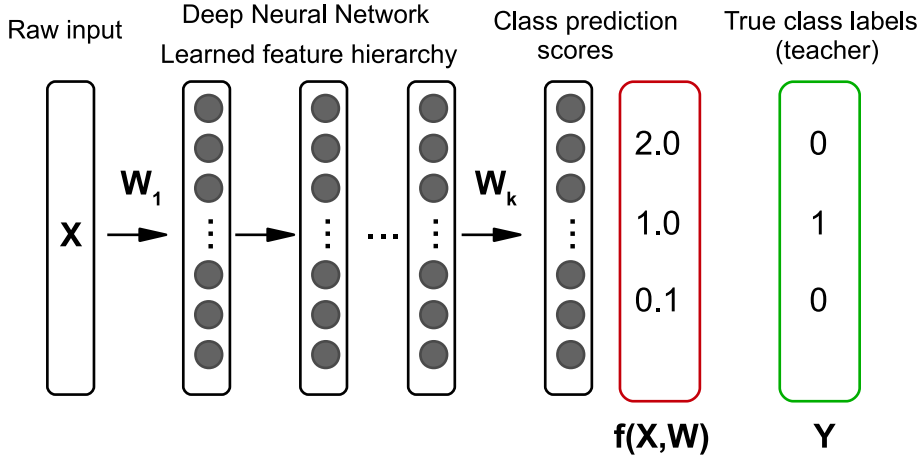
LARGE-SCALE PRE-TRAINING AND TRANSFER

- Supervised learning on generic images : relating images to low information labels



LARGE-SCALE PRE-TRAINING AND TRANSFER

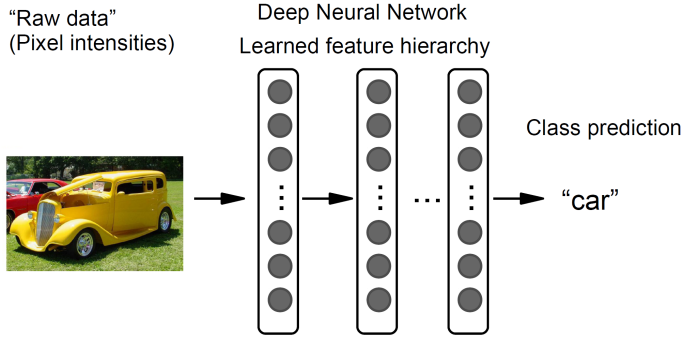
- Supervised learning : relating high information input signal to low information labels
  - fixed, rather small label “vocabulary”



$$\mathcal{L}(X; W; Y)$$

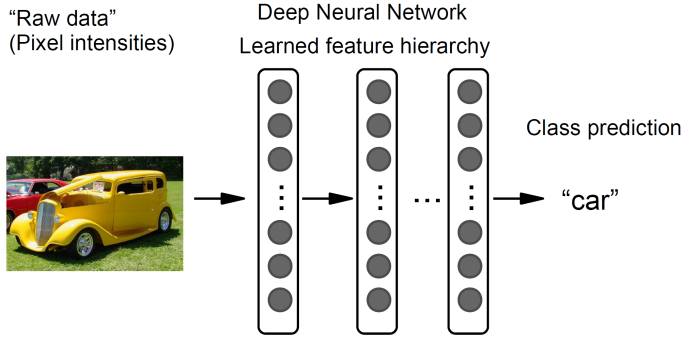
LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scaling Laws: Larger Data, Larger Models - better transfer
- Supervised learning on ImageNet-1k : pre-trained models transferable
  - pre-train on ImageNet-1k - transfer across various downstream tasks
- Problem: Human-labeled data poorly scalable
  - **ImageNet-21k**: **14x** larger
  - **JFT-300M**: **300x** larger - pseudo-labels



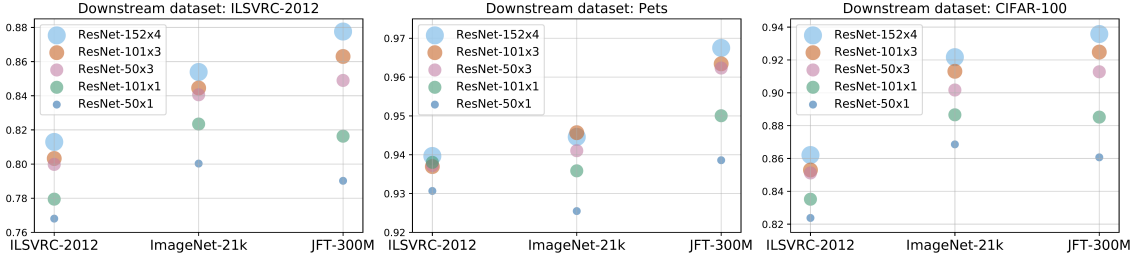
LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scaling Laws: Larger Data, Larger Models - better transfer
- Supervised learning on ImageNet-1k : pre-trained models transferable
  - pre-train on ImageNet-1k - transfer across various downstream tasks
- Problem: poor zero-shot transfer, poor robustness to data distribution shift
  - **ImageNet-21k**: **14x** larger
  - **JFT-300M**: **300x** larger - pseudo-labels



LARGE-SCALE PRE-TRAINING AND TRANSFER

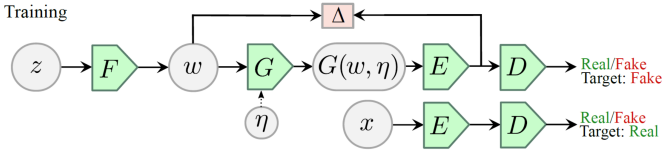
- Scaling Laws: Larger Data, Larger Models - better transfer
- Supervised learning on ImageNet-1k : pre-trained models transferable
  - pre-train on ImageNet-1k - transfer across various downstream tasks
- Problem: poor zero-shot transfer, poor robustness to data distribution shift
  - **ImageNet-21k**: **14x** larger
  - **JFT-300M**: **300x** larger - pseudo-labels

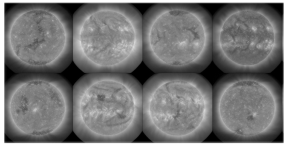


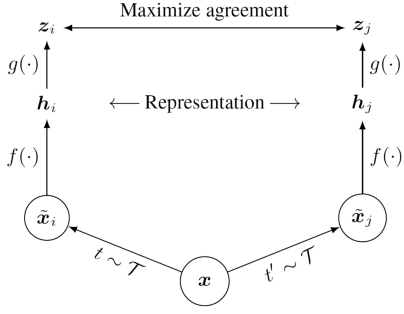
BiT - Big Transfer, Kolesnikov et al, ECCV, 2020

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scalable data: unlabeled or pseudo-labeled data
- **Unsupervised**, **Self-Supervised** learning in different flavors
  - human-made labels not required
- Often, using auxiliary tasks - self-supervised learning
  - contrastive losses (SimCLR, DINO), reconstruction based losses (eg VAEs, MAE, Diffusion models), . . .
  - adversarial losses (eg. GANs -> see Day 5 Special!)







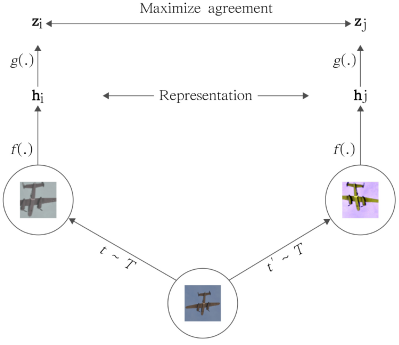
Pidhorskyi et al, 2020; Effenberger et al, 2020; Chen et al, 2020

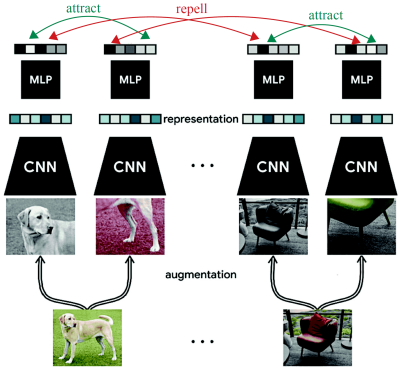
LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scalable data: unlabeled data
- Contrastive losses: construct losses from transformed pairs of inputs

$$\mathbf{z}_i = g(\mathbf{h}_i), \quad \mathbf{z}_j = g(\mathbf{h}_j), \quad \text{sim}(\mathbf{z}_i, \mathbf{z}_j) = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|}$$

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2n} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

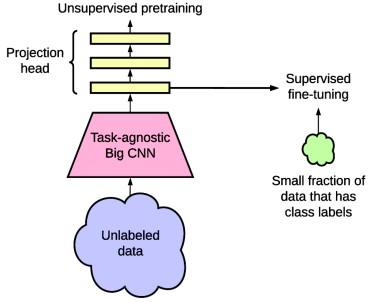


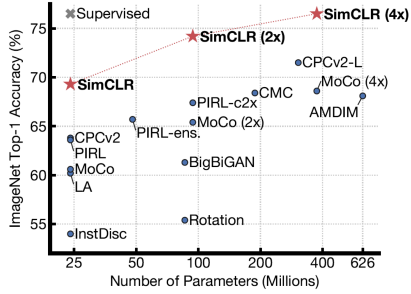


Chen et al, 2020

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scalable data: unlabeled data
- Contrastive losses: larger models - better self-supervised learning
- Evidence for better representations in larger networks after self-supervised pre-training

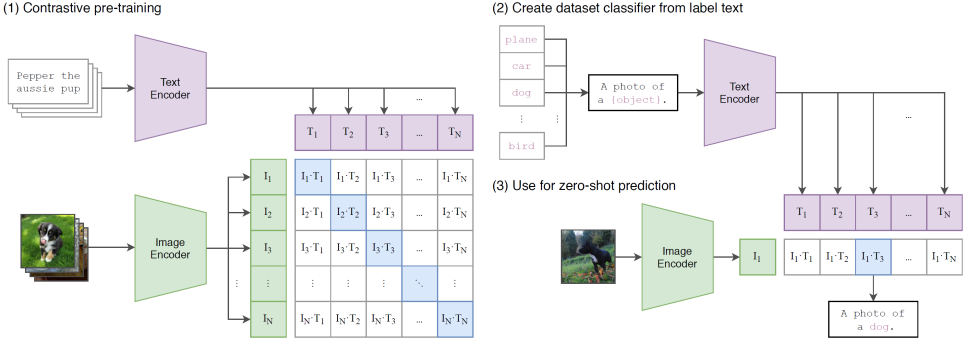




Chen et al, 2020

LARGE-SCALE PRE-TRAINING AND TRANSFER

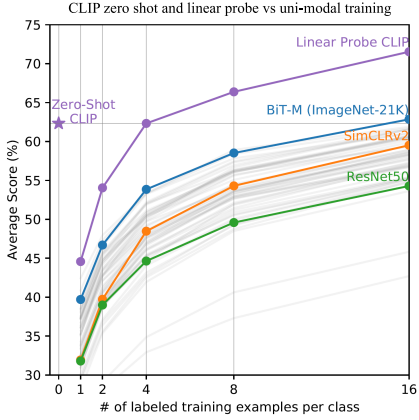
- Scalable data: weakly aligned image-text pairs from public Internet
- CLIP: foundation language-vision model for large-scale representation learning
  - **self-supervised, open vocabulary pre-training on image-text pairs**
  - very strong zero- and few-shot transfer across various downstream tasks
  - strong robustness to data distribution shift

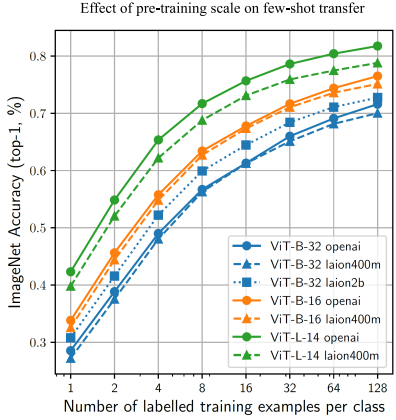


Radford, ICML, 2021

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scalable data: weakly aligned image-text pairs from public Internet
- CLIP: foundation language-vision model for large-scale representation learning
  - self-supervised, open vocabulary pre-training on image-text pairs
  - **very strong zero- and few-shot transfer across various downstream tasks**
  - strong robustness to data distribution shift

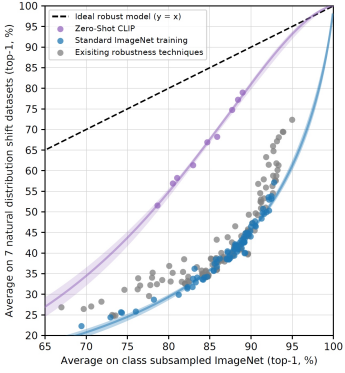


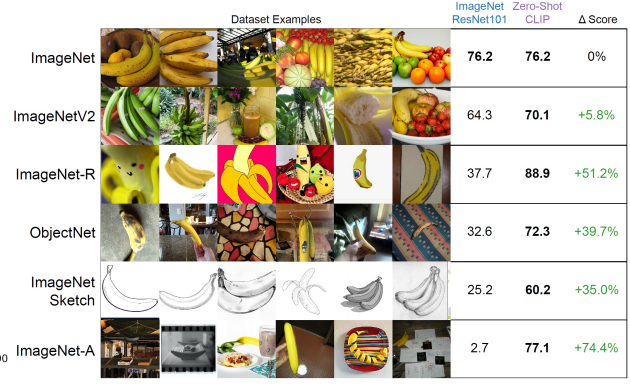


Radford, ICML, 2021; Schumann et al, NeurIPS 2022

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scalable data: weakly aligned image-text pairs from public Internet
- CLIP: foundation language-vision model for large-scale representation learning
  - self-supervised, open vocabulary pre-training on image-text pairs
  - very strong zero- and few-shot transfer across various downstream tasks
  - **strong robustness to data distribution shift**



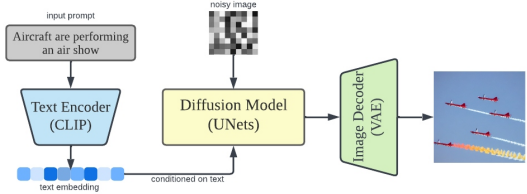


Radford, ICML, 2021

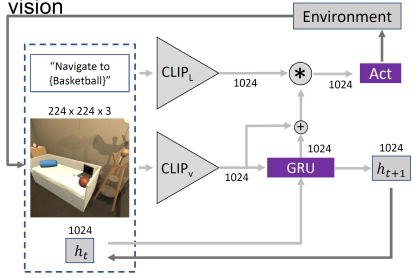
LARGE-SCALE PRE-TRAINING AND TRANSFER

- CLIP - language-vision foundation model; self-supervised language-vision learning (no labels)
- Out-of-distribution robustness & few-shot / zero-shot transfer
- Pre-trained models are **highly re-usable across various tasks & conditions**
- Generalist zero-shot function: no adaptation to new conditions / data / tasks required

Generative image models (Stable Diffusion)



General purpose open vocabulary vision



Rombach et al, CVPR, 2022; Khandelwal et al, CVPR, 2022

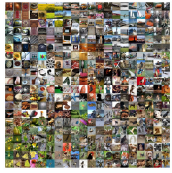
LARGE-SCALE PRE-TRAINING AND TRANSFER

- Problem - studying self-supervised foundation models is challenging: requires
  - large-scale **data** (at least 100M of samples)
  - large-scale **compute** (in order of GPU months per single experiment)
  - **expertise** in large-scale distributed training

- Solution - LAION: Large-scale Artificial Intelligence Open Network
  - **data**: LAION-400M, LAION-5B image-text datasets – **Outstanding Paper Award NeurIPS 2022**
  - **compute**: applying for publicly funded supercomputers (JUWELS, Germany, SUMMIT, USA)
  - **expertise**: strong grassroot research community skilled in large-scale experiments and distributed training
  - **Open-source** release of pre-trained models: openCLIP (work published at NeurIPS, CVPR)

LARGE-SCALE PRE-TRAINING AND TRANSFER

- LAION-400m/5B (2021): next gen datasets, 10x/100x larger than ImageNet-1k/21k, multi-modal (image-text)
  - data collection from public Internet (Common Crawl) by community effort (http://laion.ai)
- CommonPool-10B, DataComp-1B: follow-up work; systematic dataset search for pre-training strong models



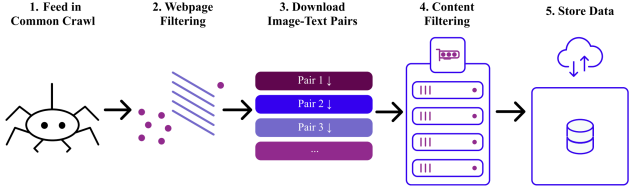
ImageNet-1k, 21k
224x224, 1024x1024; 1.4M/14M samples

LAION-400M/5B (multi-modal)
Image-text, (170M>=1024x1024); 400M/5B samples

Schumann et al, NeurIPS 2022; Gadre, Ilharco, Fang et al, arXiv:2304.14108, 2023

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Open data LAION-400M/5B, DataComp: Open sourcing data collection procedures
  - transparent dataset, open source tools & workflows, reproducible training across various scales
  - dataset of links to images in public internet, together with text captions
  - researchers can obtain full image-text dataset for experiments using open tools



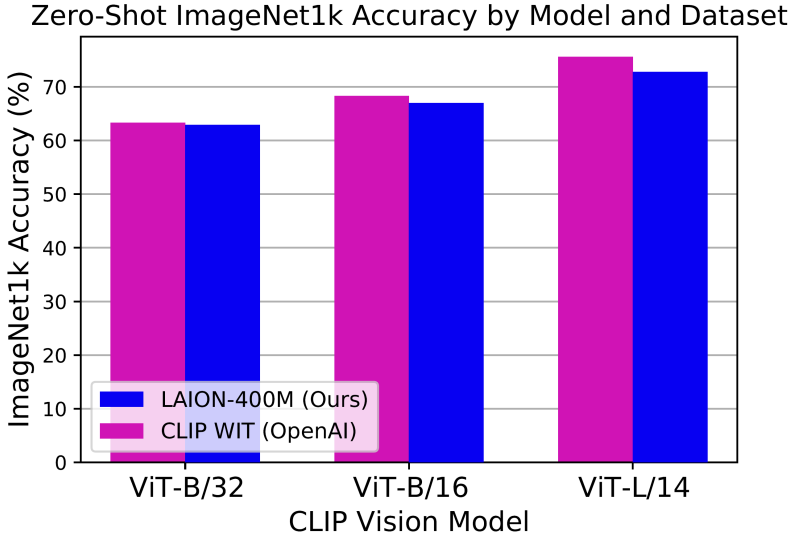
Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

Table 1: **Dataset Size.** LAION-5B is more than 20 times larger than other public English image-text datasets. We extend the analysis from Desai et al. [14] and compare the sizes of public and private image-text datasets.

Schumann et al, NeurIPS 2022; Gadre, Ilharco, Fang et al, arXiv:2304.14108, 2023

LARGE-SCALE PRE-TRAINING AND TRANSFER

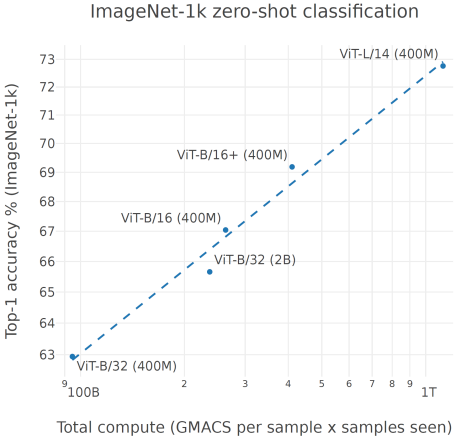
- Open-source foundation data and models - reproducible, open science

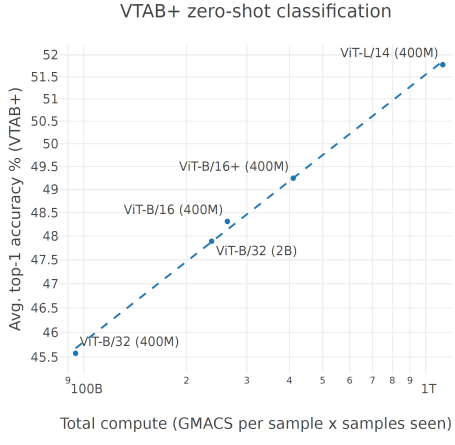


Schumann et al, NeurIPS 2022

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scaling laws for language-vision learning with LAION and openCLIP: open-source data, models and code - reproducible science





Schumann et al, NeurIPS 2022

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scaling laws for language-vision learning with LAION and openCLIP: open-source data, models and code - reproducible science

Model	Pre-training	INet	INet-v2	INet-R	INet-S	ObjNet	VTAB+
B/32	CLIP WIT	63.3	56.0	69.4	42.3	44.2	45.4
	LAION-400M	$62.9^{-0.4}$	$55.1^{-0.9}$	$73.4^{+4.0}$	$49.4^{+7.1}$	$43.9^{-0.3}$	$45.6^{+0.2}$
	LAION-2B-en	$65.7^{+2.4}$	$57.4^{+1.4}$	$75.9^{+6.5}$	$52.9^{+10.6}$	$48.7^{+4.5}$	$47.9^{+2.5}$
B/16	CLIP WIT	68.3	61.9	77.7	48.2	55.3	47.5
	LAION-400M	$67.0^{-1.3}$	$59.6^{-2.3}$	$77.9^{+0.2}$	$52.4^{+4.2}$	$51.5^{-3.8}$	$48.3^{+0.8}$
B/16+	LAION-400M	69.2	61.5	80.5	54.4	53.9	49.2
L/14	CLIP WIT	75.6	69.8	87.9	59.6	69.0	55.7
	LAION-400M	$72.8^{-2.8}$	$65.4^{-4.4}$	$84.7^{-3.2}$	59.6	$59.9^{-9.1}$	$51.8^{-3.9}$
	LAION-2B-en	$75.2^{-0.3}$	$67.7^{-2.0}$	$87.4^{-0.5}$	$63.3^{+3.7}$	$65.5^{-3.6}$	$54.6^{-1.2}$

Schumann et al, NeurIPS 2022

LARGE-SCALE PRE-TRAINING AND TRANSFER

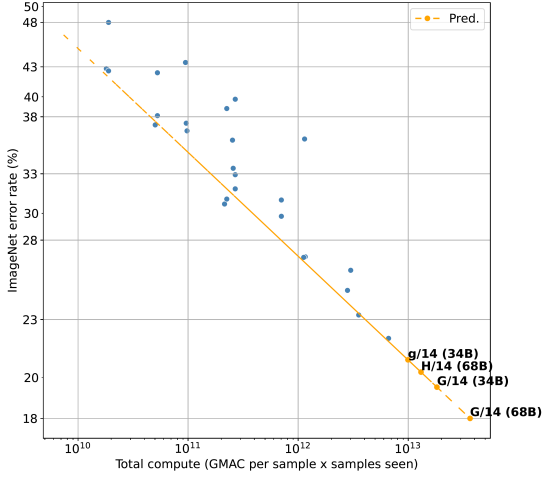
- Bottlenecks: one insufficient scale can lead to saturation if increasing others
- Larger language-vision models: stronger on larger dataset and sample seen scales

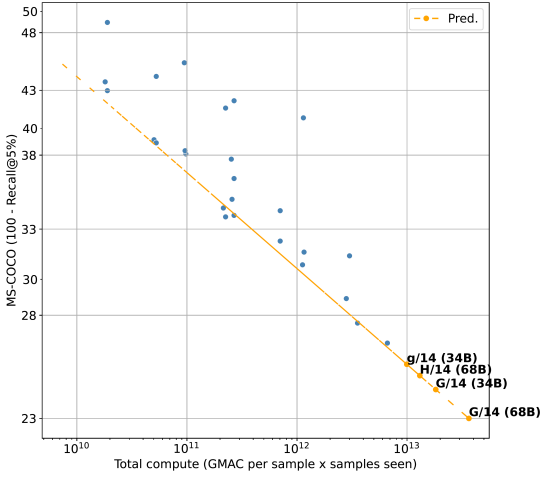
Model	Samples seen	LAION-80M	LAION-400M	LAION-2B
ViT-B/32	3B	51.94	57.12	57.36
	13B	56.46	63.23	62.53
	34B	56.43	64.06	66.47
ViT-B/16	3B	57.55	62.68	61.82
	13B	60.24	67.00	68.13
	34B	61.28	69.00	70.22
ViT-L/14	3B	61.14	69.31	68.93
	13B	63.96	73.06	73.10
	34B	64.83	73.94	75.20

Cherti et al, arXiv:2212.07143, CVPR 2023

LARGE-SCALE PRE-TRAINING AND TRANSFER

- Scaling laws for language-vision learning with LAION and openCLIP: open-source data, models and code - reproducible science
- Predicting model performance and properties on larger scales





Cherti et al, arXiv:2212.07143, CVPR 2023

LARGE-SCALE PRE-TRAINING AND TRANSFER

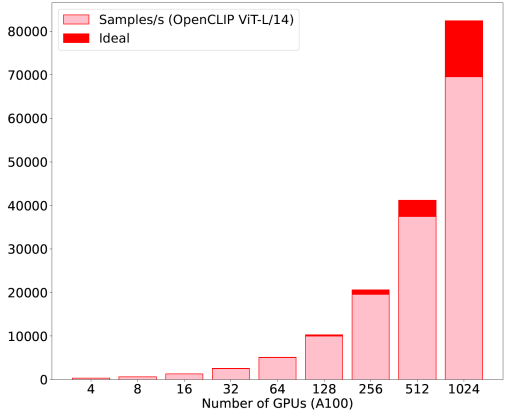
- Scaling laws for language-vision learning with LAION and openCLIP: open-source data, models and code - reproducible science
- Predicting model performance and properties on larger scales

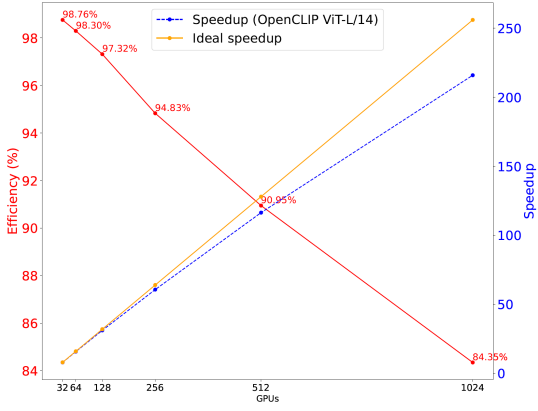
Model	ImageNet top-1 (%)	MS-COCO Recall@5 (%)
H/14 (68B)	79.73	75.03
g/14 (34B)	79.11	74.48
g/14 (68B)	80.66	75.85
G/14 (13B)	78.26	73.75
G/14 (34B)	80.47	75.68
G/14 (68B)	81.92	76.99

Cherti et al, arXiv:2212.07143, CVPR 2023

LARGE-SCALE PRE-TRAINING AND TRANSFER

- JUWELS Booster: necessary for the experiments
- 122 hours with 1024 A100 (124K GPU hours) for training of ViT L/14 openCLIP on 34B samples
- In contrast to standard supervised training: larger batch sizes beneficial for learning





Cherti et al, arXiv:2212.07143, CVPR 2023

LARGE-SCALE PRE-TRAINING AND TRANSFER

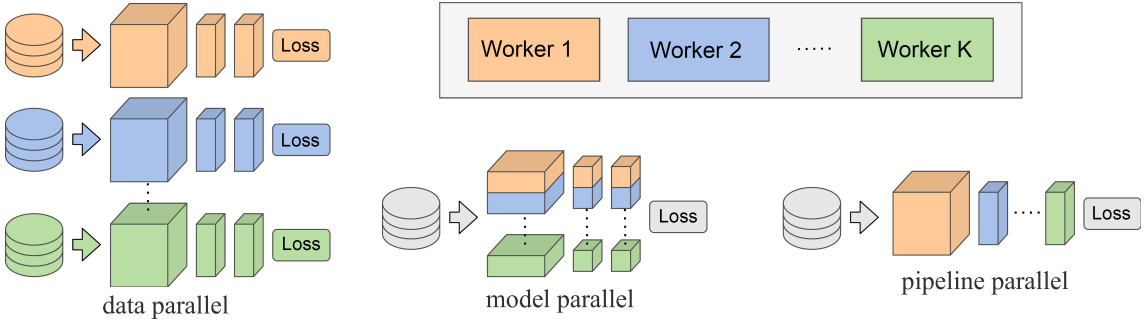
- Current language-vision models are still small scale (compared to LLMs (>100B params); PaLI (image-text-to-text) – 17B; Parti (text-to-image) - 20B params)
- Stronger transfer & robustness: aiming for larger scales
- Larger machines necessary: JUPITER Exascale upcoming at JSC

Name	Width	Emb.	Depth	Acts.	Params	GMAC
ViT-B/32	768 / 512	512	12 / 12	10 M	151 M	7.40
ViT-B/16	768 / 512	512	12 / 12	29 M	150 M	20.57
ViT-L/14	1024 / 768	768	24 / 12	97 M	428 M	87.73
ViT-H/14	1280 / 1024	1024	32 / 24	161 M	986 M	190.97
ViT-g/14	1408 / 1024	1024	40 / 24	214 M	1.37 B	290.74
ViT-G/14	1664 / 1280	1280	48 / 32	310 M	2.54 B	532.92

Cherti et al, arXiv:2212.07143, CVPR 2023

DISTRIBUTED TRAINING WITH VERY LARGE MODELS

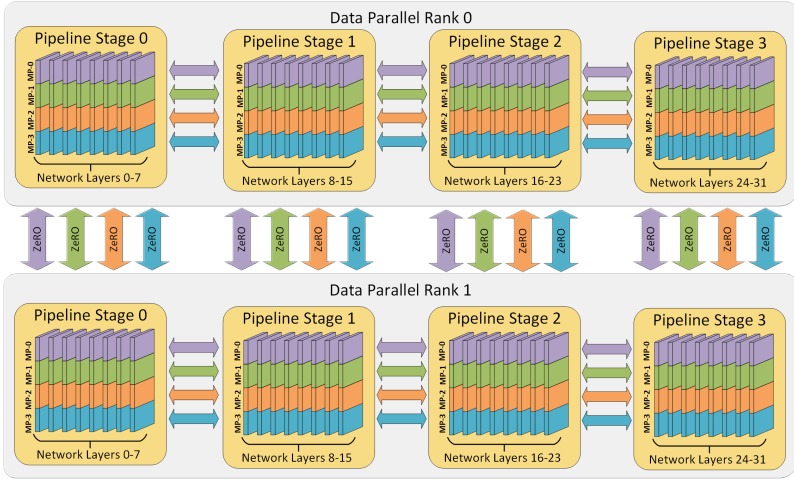
- Growing model scale: only data parallel scheme not sufficient
  - Language Modelling: GPT-3 - 175 Billion parameters; PaLM (Google) - 540B parameters
  - Vision: ViT-22B; Language-Vision: Parti - 17B params
- Model/Tensor parallelism, Pipeline Parallelism: can split a very large model across accelerators
- Different libraries: DeepSpeed (Microsoft), CollosalAI (HPC-AI), PyTorch/TensorFlow DTensor, ...



Laskin et al, 2020

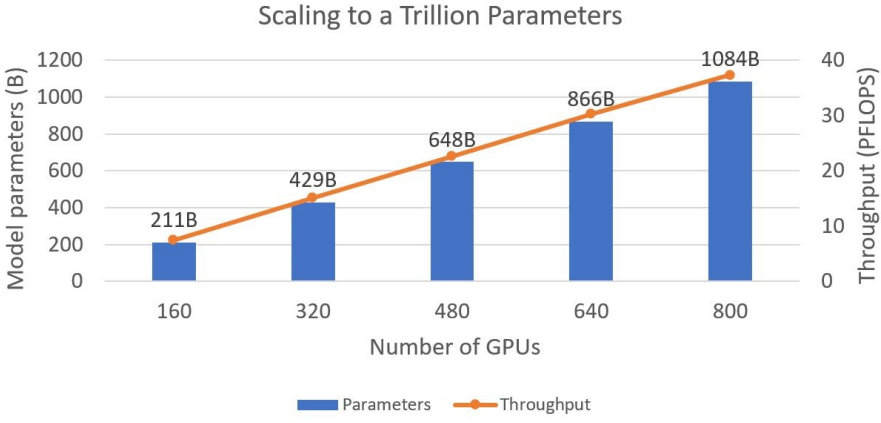
DISTRIBUTED TRAINING WITH VERY LARGE MODELS

- Hybrid parallel schemes
  - using data, model and pipeline parallelism simultaneously
- Distributed training that combines memory and compute efficiency
- DeepSpeed: supports hybrid parallelism



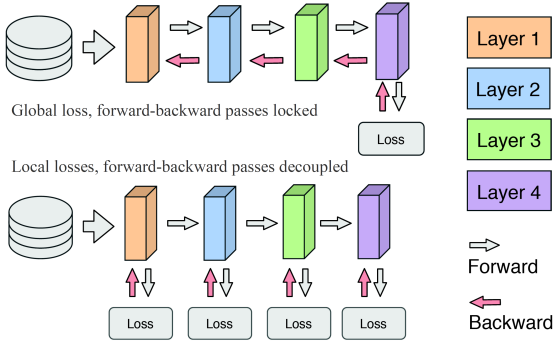
DISTRIBUTED TRAINING WITH VERY LARGE MODELS

- Hybrid parallel schemes
  - using data, model and pipeline parallelism simultaneously
- DeepSpeed: “3D Parallelism”
  - executing and speeding up a Trillion size model on 800 A100 GPUs



DISTRIBUTED TRAINING WITH VERY LARGE MODELS

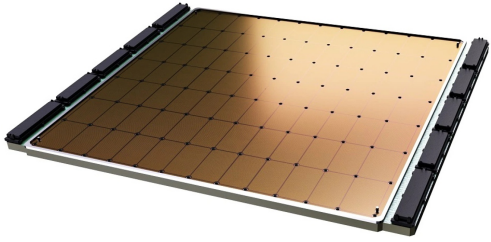
- Upcoming: local updates, decoupled gradients
- Getting rid of global forward-backward pass dependency alltogether
- Asynchronous local updates, highly beneficial for parallelization
- Towards energy-efficient in-memory computing: minimize data transfer
- New generic losses for unsupervised learning

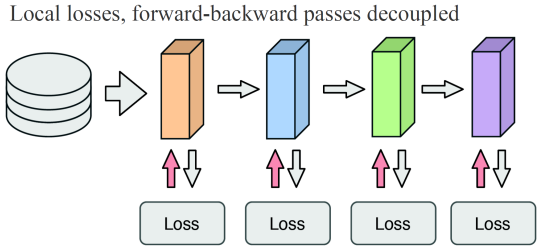


Laskin et al, 2020

DISTRIBUTED TRAINING WITH VERY LARGE MODELS

- Upcoming: local updates, decoupled gradients
- Asynchronous local updates, highly beneficial for parallelization
- Energy efficient distributed training on specialized hardware, in-memory computing
  - Graphcore IPU: Colossus Mk2
  - Cerebras : Wafer Scale Engine 2 (WSE - 850k Cores!)





LARGE-SCALE FOUNDATION GENERALIST MODELS

Outlook

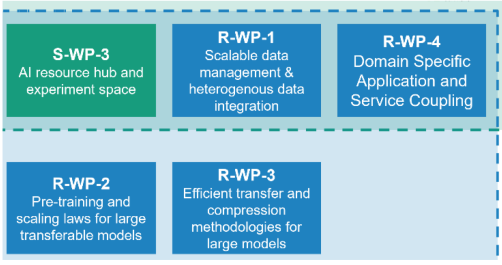
- Language-vision generalist models for strong transfer across domains and tasks
- Large model scale: hybrid parallelism required
- Large data scale: data collection and automated filtering (see DataComp)
- Systematic Search for Scalable Architectures (Project Nucleus, LAION)
- Model compression for efficient transfer and low resource deployment
- Energy efficient large scale learning with in-memory computing neuromorphic hardware

LARGE-SCALE FOUNDATION GENERALIST MODELS

- LAION: Large-Scale Artificial Intelligence Open Network (join on Discord!)
  - Scalable Learning & Multi-Purpose Lab (SLAMPAI; Jenia Jitsev, Mehdi Cherti)
  - University of Washington (Seattle), Allen AI Institute, MILA, UC Berkeley, U Tel-Aviv, Stanford, . . .
  - https://laion.ai/ - join on Discord!
- Supercomputers : JUWELS & JUWELS Booster, JUPITER to come

LARGE-SCALE FOUNDATION GENERALIST MODELS

- WestAI: AI Service Center, funded by BMBF (2022-2025, 12.4M €)
- Pillars: large-scale pre-training, scaling laws for transfer, compression, next-gen highly scalable generic, energy-efficient learning
  - SLAMPAI at JSC: scientific lead
  - U Bonn, RWTH Aachen, Fraunhofer IAIS, U Padeborn, U Dortmund



DISTRIBUTED TRAINING: ACTIVITIES AT FZJ

- Distributed Training for Hyperspectral Remote Sensing (Gabriele Cavallaro)
- Helmholtz Data Challenges : Platform for collaborative datasets and model training
  - SLAMPAI & Helmholtz AI Consultant Team: https://helmholtz-data-challenges.de/
- TOAR: Earth System Data Exploration (ESDE) Lab (Martin Schultz)
- Helmholtz AI Research Group at INM-1: deep learning for neuroimaging
- JULAIN: Juelich Artificial Intelligence Network, join in!
  - mailing list: https://lists.fz-juelich.de/mailman/listinfo/ml

LARGE-SCALE FOUNDATION GENERALIST MODELS

- LAION: Large-Scale Artificial Intelligence Open Network (join on Discord!)
  - Scalable Learning & Multi-Purpose Lab (SLAMPAI; Jenia Jitsev, Mehdi Cherti)
  - University of Washington (Seattle), Allen AI Institute, MILA, UC Berkeley, U Tel-Aviv, Stanford, . . .
  - https://laion.ai/ - join on Discord!
- Supercomputers : JUWELS & JUWELS Booster, JUPITER to come