Training Vision Transformers with Synthetic Images

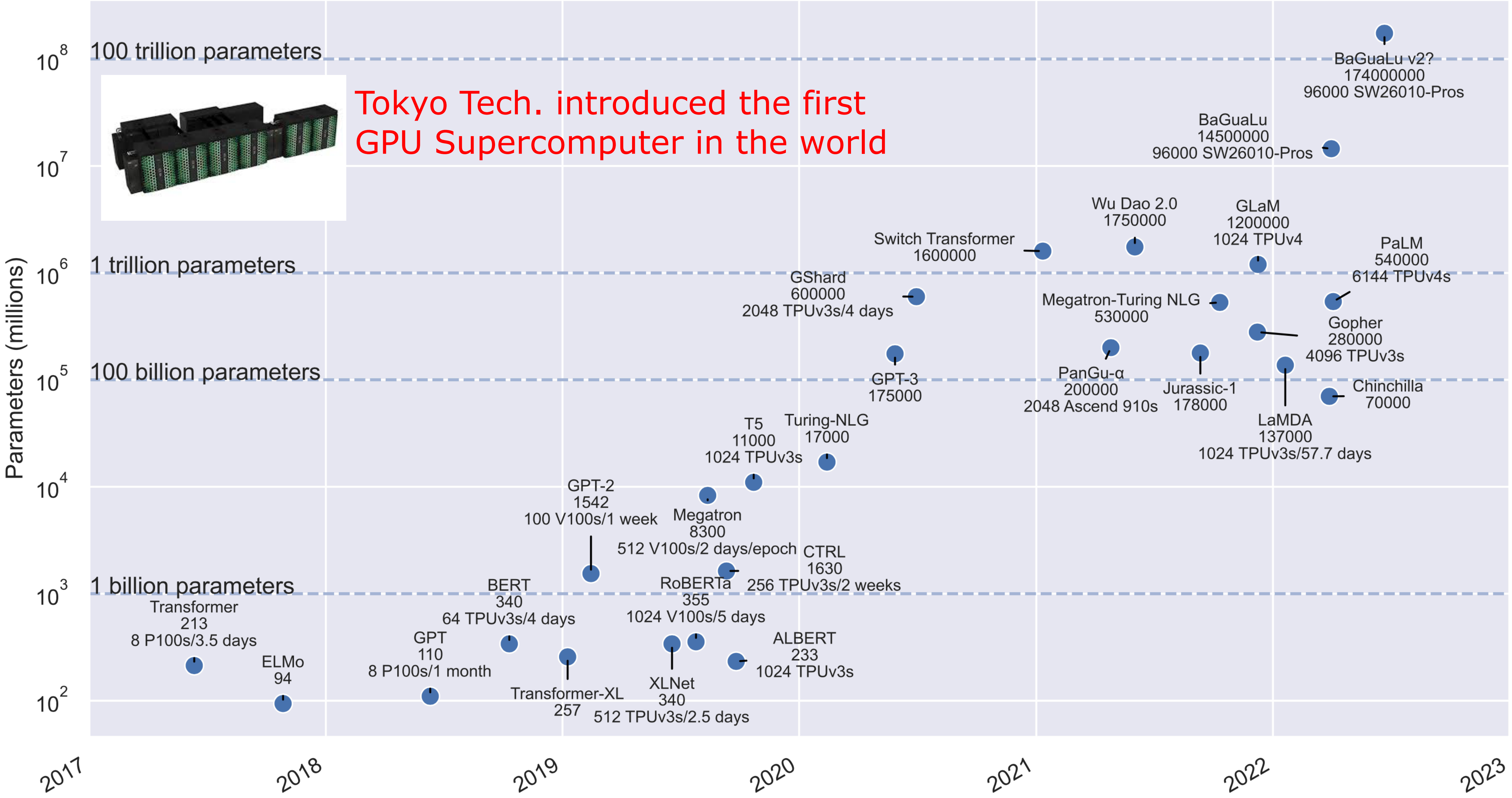
Multicore World X
13–17 Feb. 2023
Wellington, NEW ZEALAND

Tokyo Institute of Technology
Rio Yokota
rioyokota@gsic.titech.ac.jp

Deep Learning Requires GPU Supercomputers

Tokyo Tech

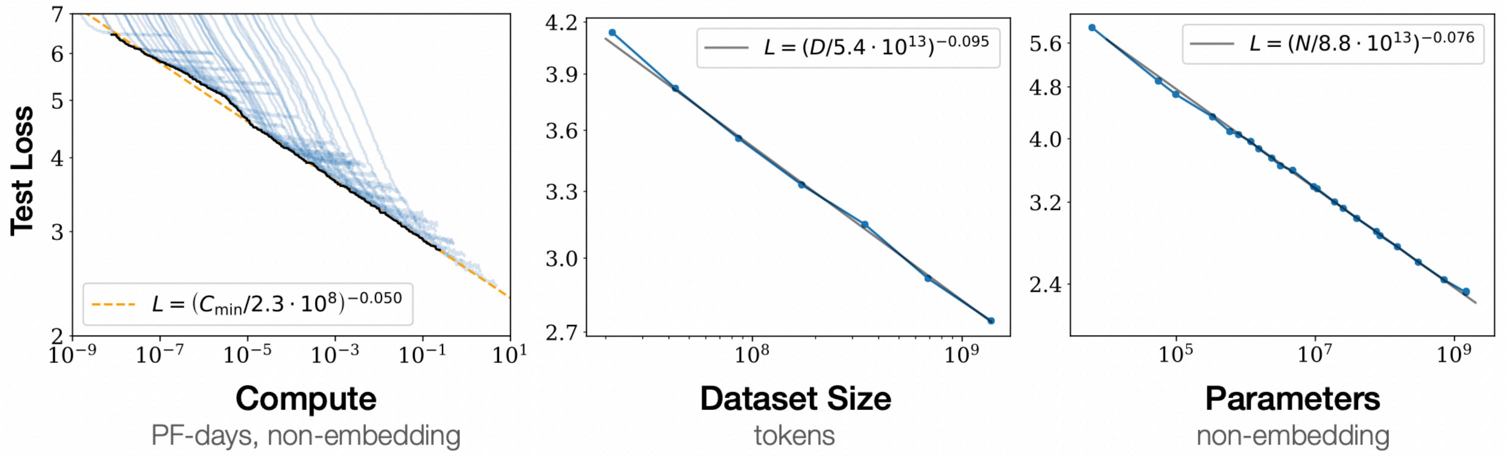
Tokyo Tech. introduced the first
GPU Supercomputer in the world



Scaling Laws of Transformers

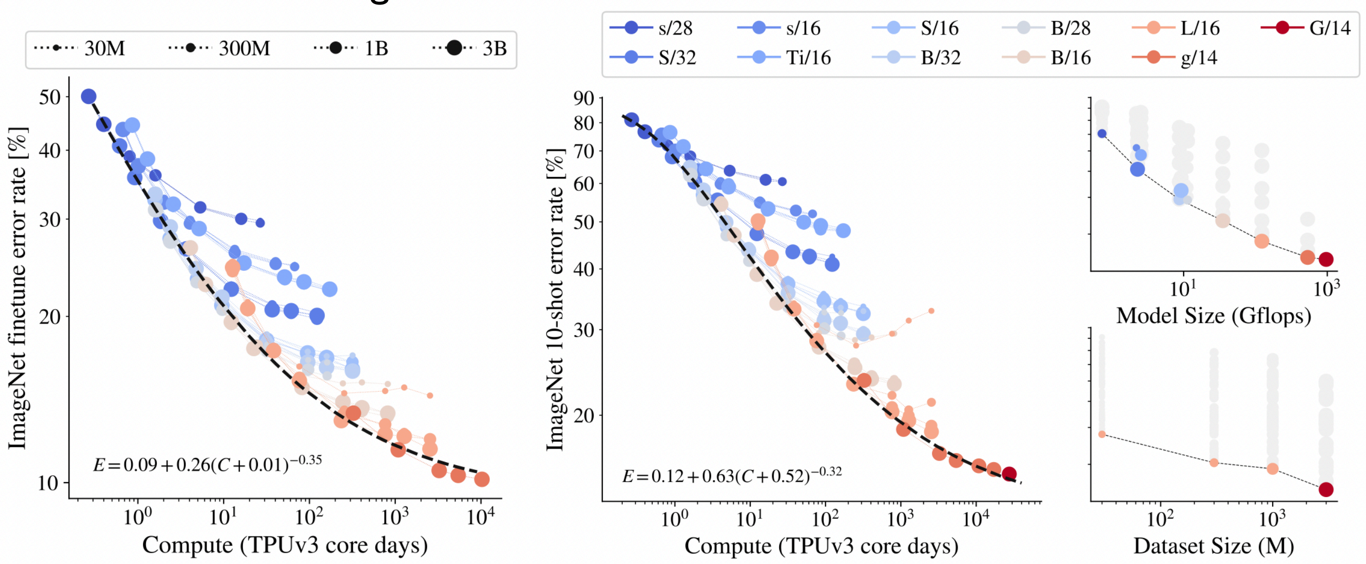
Language

Kaplan et al. "Scaling Laws for Neural Language Models", arXiv:2001.08361



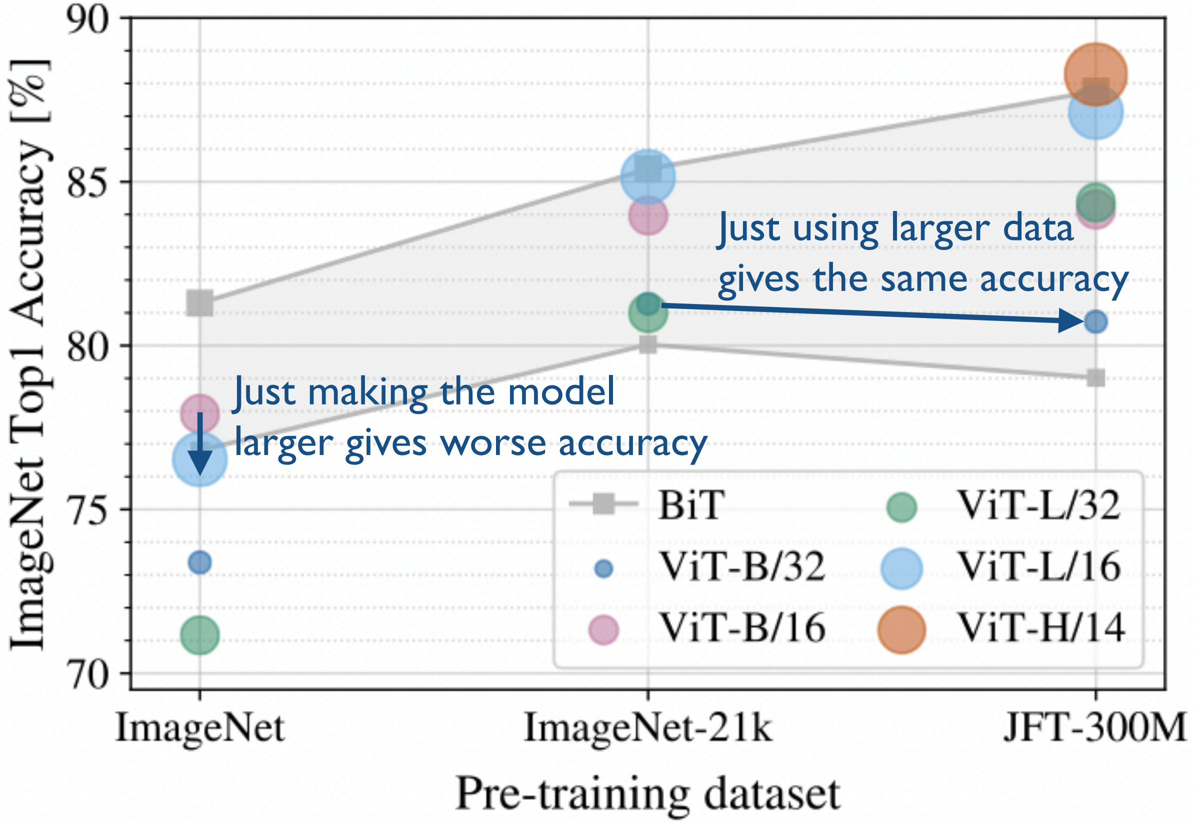
Vision

Zhai et al. "Scaling Vision Transformers", arXiv:2106.04560



Pre-training of Vision Transformers

- The pre-training of large vision transformers requires large datasets
  - Large models alone do not lead to better results
  - Large data alone do not lead to better results



- The largest dataset JFT-300M is own by Google and is not available publicly

Open-source Multi-modal Datasets

- Classifying into a million bins?
  - Classification doesn't scale
  - Hierarchical labels?
- Vision+Language models
  - Pair of text and image
  - Works at scale

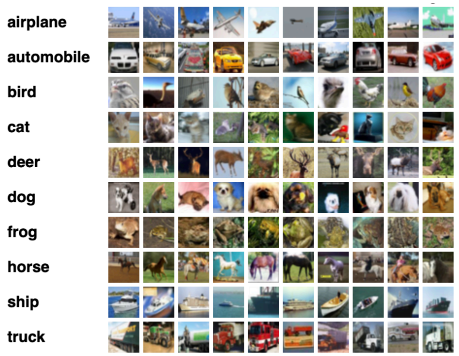
LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI-MODAL DATASETS

by: Romain Beaumont, 7 Jul, 2022

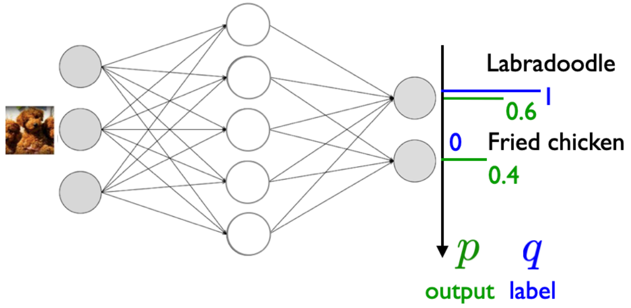
We present a dataset of 5,85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world.

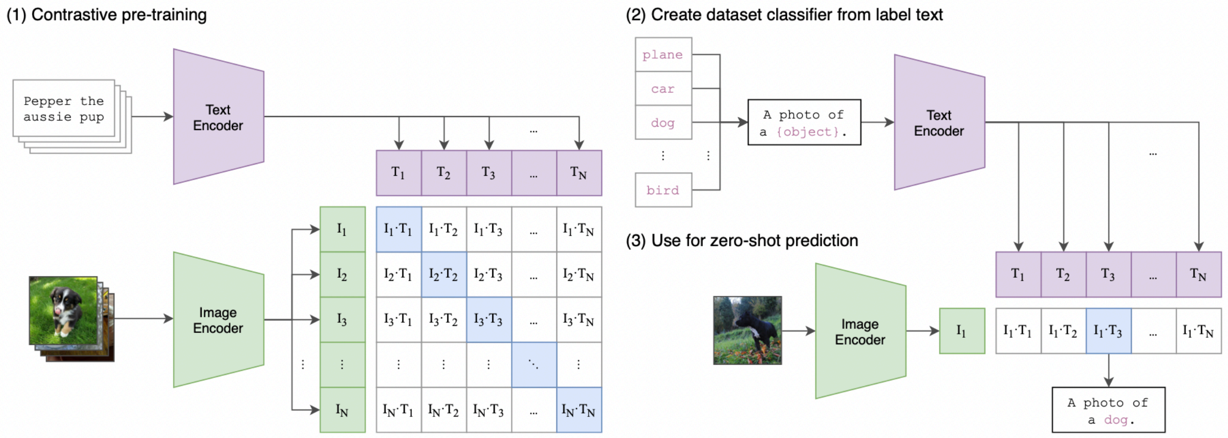
Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev

CIFAR10

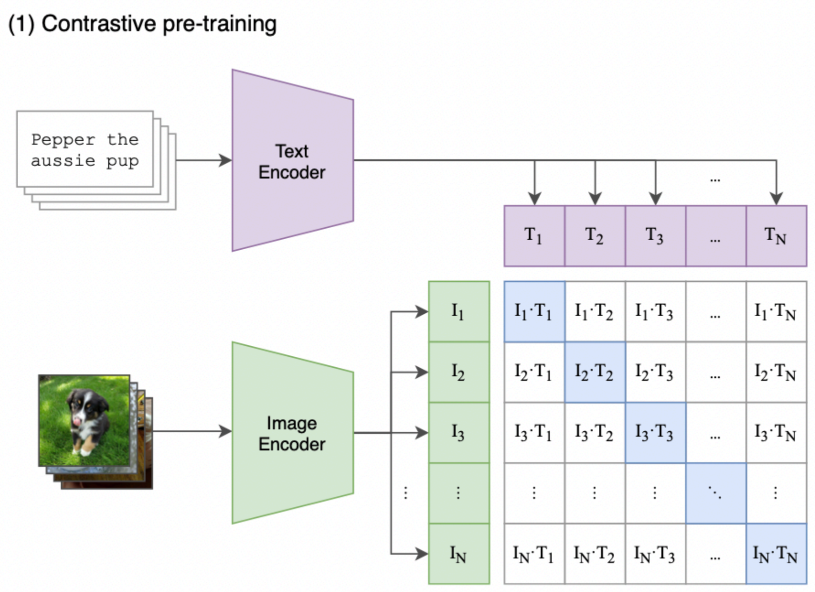


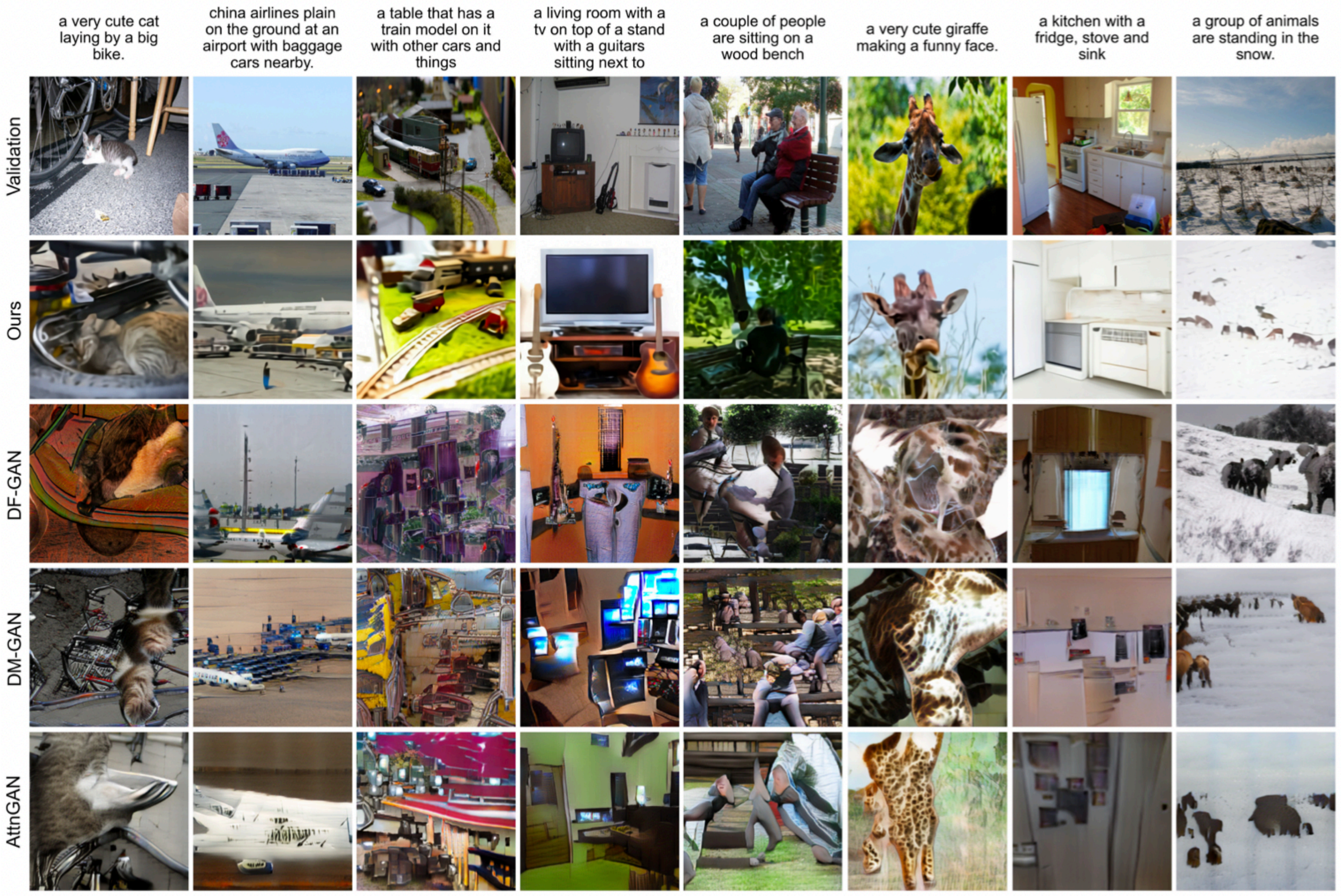
ImageNet





Vision + Language Models





Vision + Language Models

GLIDE

"a hedgehog using a calculator" | "a corgi wearing a red bowtie and a purple party hat" | "robots meditating in a vipassana retreat" | "a fall landscape with a small cottage next to a lake"

Dall-E2

a dolphin in an astronaut suit on saturn, artstation | a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese | a teddy bear on a skateboard in times square

Imagen

Teddy bears swimming at the Olympics 400m Butter-fly event. | A cute corgi lives in a house made out of sushi. | A cute sloth holding a small treasure chest. A bright golden glow is coming from the chest.

Parti

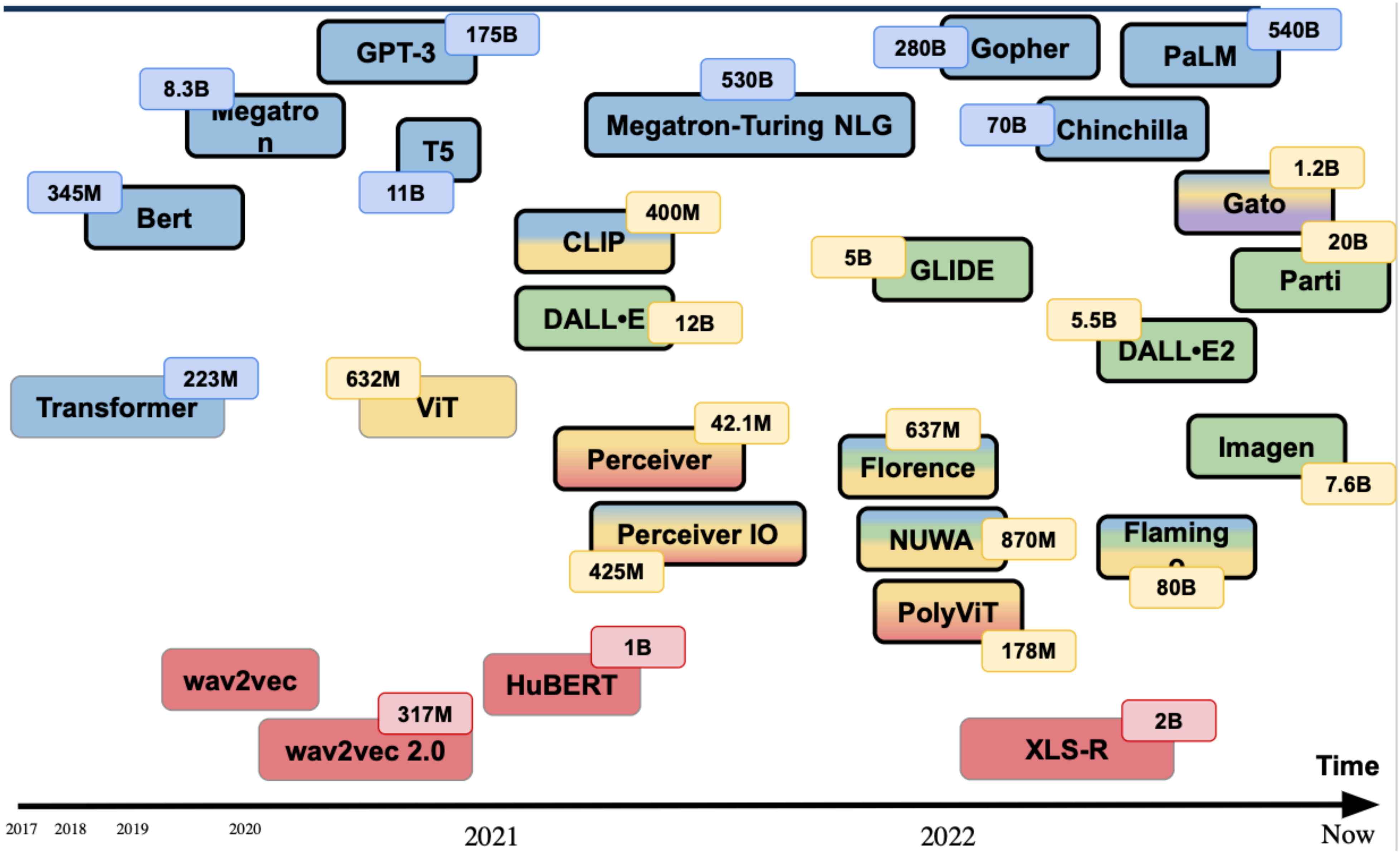
A. *A photo of a frog reading the newspaper named "Toaday" written on it. There is a frog printed on the newspaper too.*

B. *A portrait of a statue of the Egyptian god Anubis wearing aviator goggles, white t-shirt and leather jacket. The city of Los Angeles is in the background. Hi-res DSLR photograph.*

C. *A high-contrast photo of a panda riding a horse. The panda is wearing a wizard hat and is reading a book. The horse is standing on a street against a gray concrete wall. Colorful flowers and the word "PEACE" are painted on the wall. Green grass grows from cracks in the street. DSLR photograph. daytime lighting.*

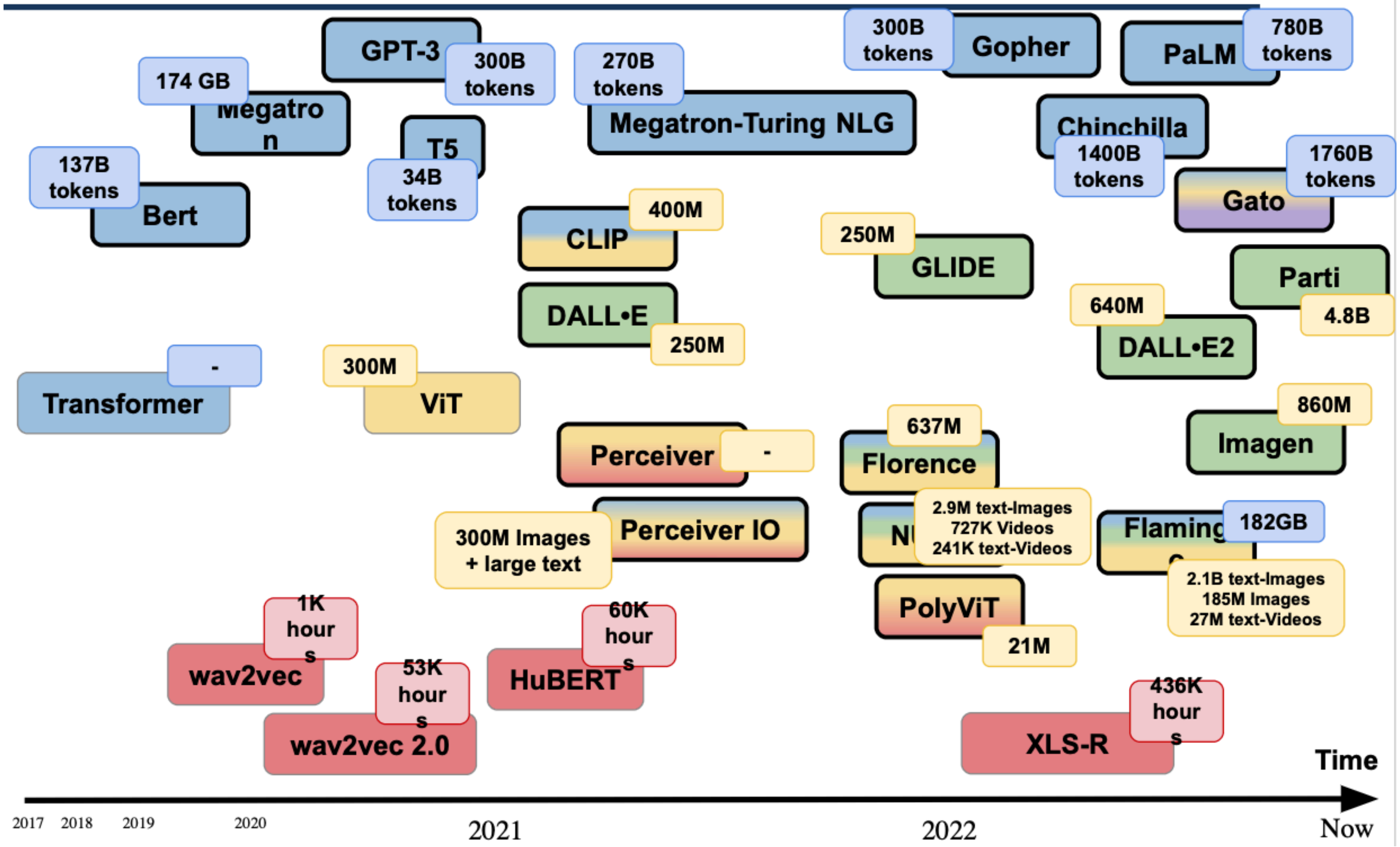
- Mix of modalities
  - Text encoder
  - Vision encoder
  - Diffusion model
- Diffusion models are extremely good at generating high res. images
- Using a good language model (T5-XLL) is important

Model size

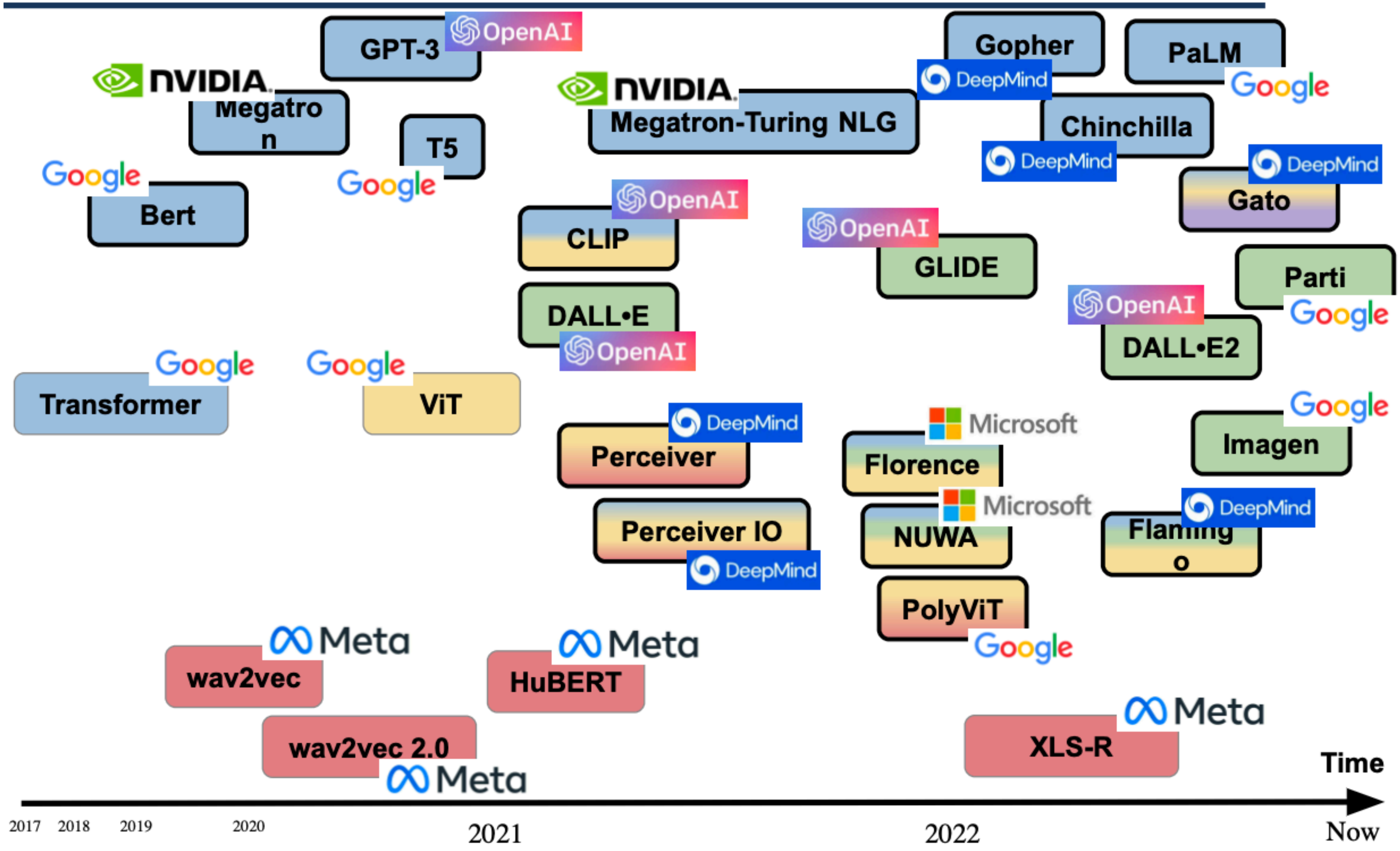


GPT-3
175B
8.3B
Megatron
n
T5
11B
345M
Bert
530B
Megatron-Turing NLG
280B
Gopher
PaLM
540B
70B
Chinchilla
1.2B
Gato
400M
CLIP
DALL•E
12B
5B
GLIDE
20B
Parti
5.5B
DALL•E2
223M
Transformer
632M
ViT
42.1M
Perceiver
637M
Florence
Imagen
7.6B
Perceiver IO
425M
NUWA
870M
Flaming
o
80B
PolyViT
178M
wav2vec
317M
wav2vec 2.0
1B
HuBERT
2B
XLS-R
Time
2017
2018
2019
2020
2021
2022
Now

Dataset size



Affiliation



INCITE Project

U.S. DEPARTMENT OF ENERGY
INCITE
LEADERSHIP COMPUTING




Type: New
Title: "Scalable Foundational Models for Transferable Generalist AI"

Principal Investigator: Irina Rish, University of Montreal, Mila - Quebec AI Institute

- Project Members
  - Irina Rish (University of Montreal)
  - Sergey Panitkin (University of Montreal)
  - Guillaume Dumas (University of Montreal)
  - Stella Biderman (EleutherAI)
  - Jenia Jitsev (Jülich Supercomputing Centre)
  - Mehdi Cherti (Jülich Supercomputing Centre)
  - Quentin Anthony (Ohio State University)
  - Guillermo Cecchi (IBM Research)
  - Rio Yokota (Tokyo Institute of Technology)





INCITE Project

Tokyo Tech

- Research Item 1: GPT-NeoX

Parameters	FLOPs	V100-hours	Node-hours
400 Million	1.9×10^{19}	176	30
1 Billion	1.2×10^{20}	1,120	186
11 Billion	1.2×10^{22}	113,888	18,981
30 Billion	1.2×10^{23}	1,120,370	186,728
40 Billion	2.2×10^{23}	2,046,296	341,049
Total	7.0×10^{23}	3,281,844	546,974

Data Strong Scalability

# Nodes	# GPUs	samples/sec	tokens/sec	Efficiency
2	12	5.9	96,666	100.0%
4	24	11.2	183,501	94.9%
8	48	21.9	358,810	92.8%
16	96	42.9	702,874	90.9%
32	192	82.8	1,356,595	87.7%
64	384	168.3	2,757,427	89.1%
128	768	320	5,242,880	84.7%
256	1536	637.4	1,0443,162	84.4%

- Research Item 2: OpenCLIP

Pre-training	samples [count]	runs [count]	resources #nodes	node-h
CLIP extended (ViT L/14, LAION-400m)	$\sim 4 \cdot 10^8$	10	64 – 1024	10,000
CLIP (ViT L/14, LAION-400m)	$\sim 4 \cdot 10^8$	10	64 – 1024	100,000
CLIP (ViT B/32, LAION-400m)	$\sim 4 \cdot 10^8$	10	64 – 1024	10,000
CLIP, CLIP extended (ViT L/14, X-Ray superset)	$\sim 0.9 \cdot 10^6$	20	32 – 1024	5,000
CLIP, CLIP extended (ViT H/14, X-Ray superset)	$\sim 0.9 \cdot 10^6$	20	32 – 1024	30,000
CLIP, CLIP extended (ViT g/14, X-Ray superset)	$\sim 0.9 \cdot 10^6$	20	32 – 1024	180,000
Total		90		425,000

Data Strong Scalability

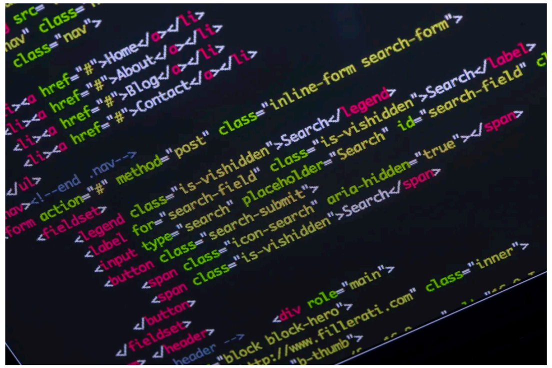
# Nodes	# GPUs	Samples/sec	Efficiency (%)
1	6	1342	100.00%
2	12	2455	91.47%
4	24	4210	78.43%
8	48	7471	69.59%
16	96	14292	66.56%
32	192	28328	65.96%
64	384	54082	62.97%
128	768	103978	60.53%

Issues with Large Datasets

Tokyo Tech

- Impossible to clean
  - Privacy
  - Copyright
  - Racial/gender bias
- Impossible to download
  - Petabytes of data
  - Scientific data is even larger
  - Bandwidth not keeping up
- Synthetic datasets
  - No need to clean
  - No need to download

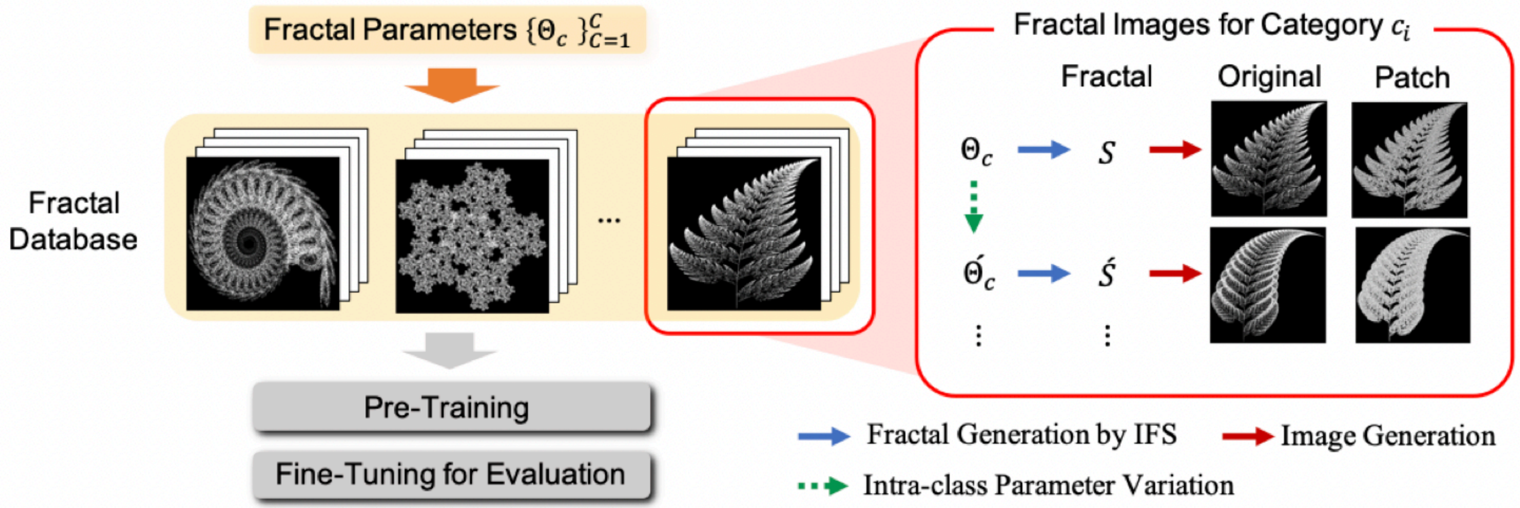
Microsoft, GitHub, and OpenAI ask court to throw out AI copyright lawsuit

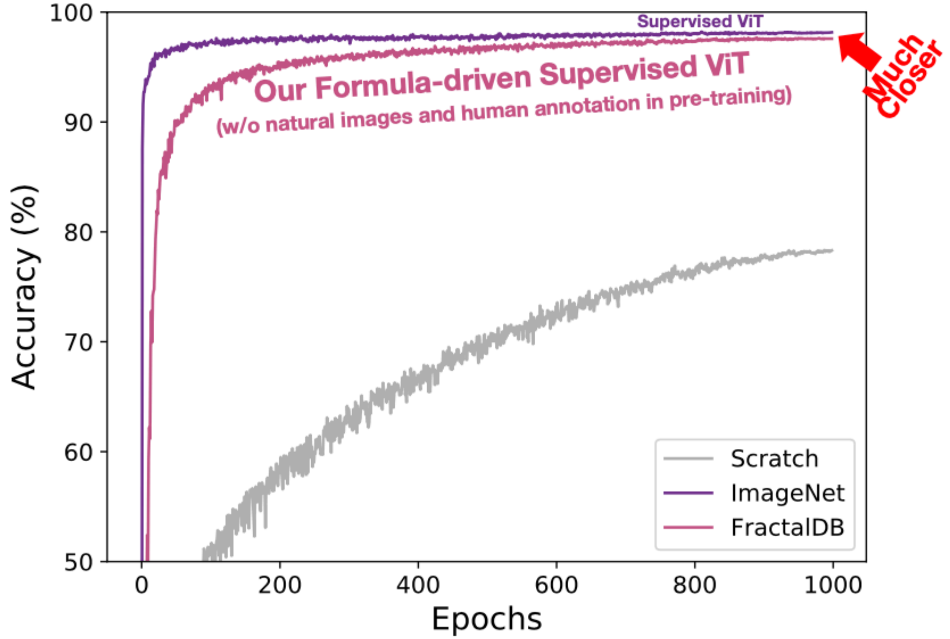


Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

Training with Synthetic Images from Fractals

Tokyo Tech





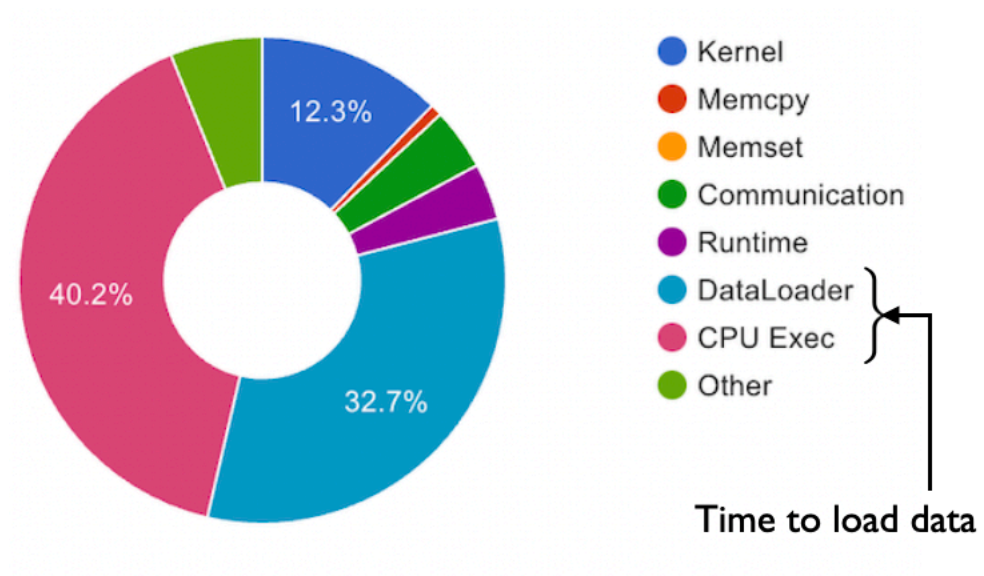
ImageNet ⇒ 1million images

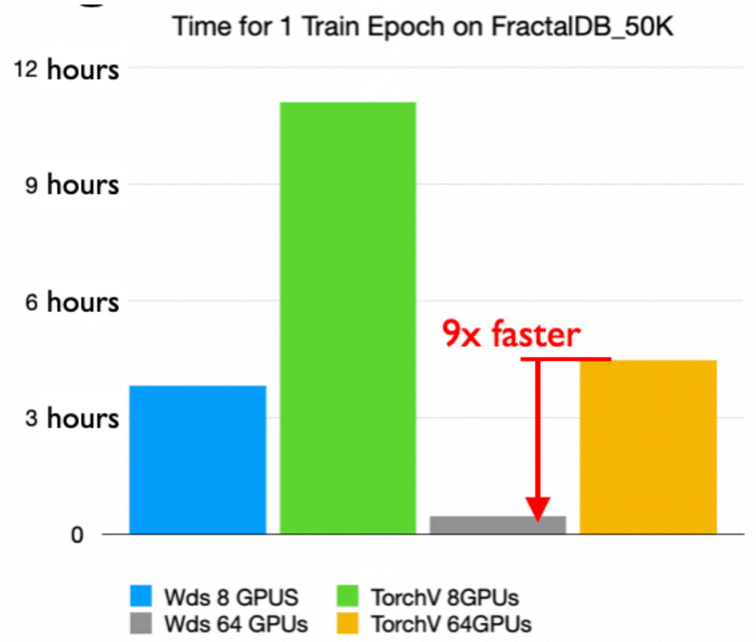
JFT-3B ⇒ 3 billion images

3000x

K. Nakashima, H. Kataoka, A. Matsumoto, K. Iwata, N. Inoue, Can Vision Transformers Learn without Natural Images? arXiv:2103.13023

Challenges when Handling Large Datasets





PyTorch DataLoader

```
train/n01440764/n01440764_0.png
                n01440764_1.png
                n01440764_2.png
                n01440764_3.png
                       :
     /n01443537/*
           :
     /n01507514/*
```

Save png files in folders for each class
The large amount of meta-data access
puts a huge strain on the network file system
For 50 million images, the training did not start
even after 3 hours

WebDataset

```
train/shard00.tar
     /shard01.tar
          :
     /shard99.tar
```

```
shard00.tar: n01440764_4.png
             n01440764_4.cls
             n01443537_2.png
             n01443537_2.cls
             n01445623_7.png
             n01445623_7.cls
                   :
```

Put image+label pairs into a large tar file
and train from the tar file directly

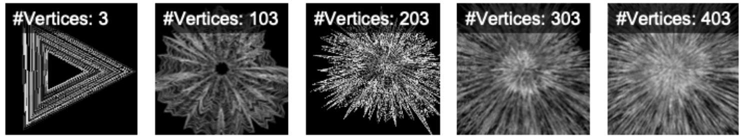
Testing Various Synthetic Datasets

Different synthetic datasets

Pre-training	C10	C100	Cars	Flowers
Scratch	78.3	57.7	11.6	77.1
Perlin Noise [21]	95.0	78.4	70.6	96.1
Dead Leaves [3]	95.9	79.6	72.8	96.9
Bezier Curves [21]	96.7	80.3	82.8	98.5
RCDB	**96.8**	**81.6**	84.2	**98.7**
FractalDB [27]	**96.8**	**81.6**	**86.0**	98.3

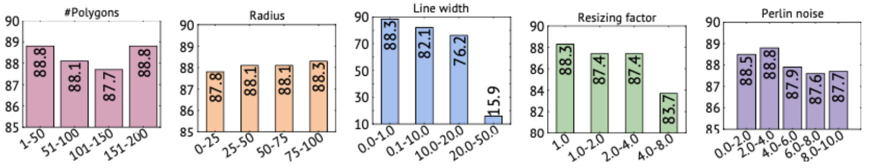
Different # vertices in RCDB

#Vertices	C10	C100	Cars	Flowers
3–102	**95.5**	**79.4**	**78.4**	**96.4**
103–202	94.2	76.3	55.8	95.9
203–302	71.3	46.9	4.9	49.8
303–402	59.4	33.9	2.5	26.8
403–502	40.1	13.6	0.8	5.3
3–502	**96.4**	**80.7**	**83.0**	**98.5**



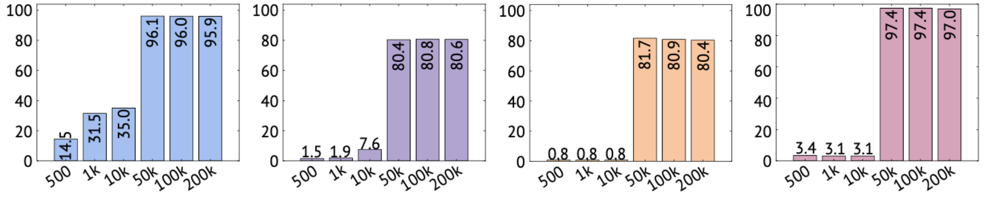
Increasing the #parameters of the dataset

Pre-training	C10	C100	Cars	Flowers
BC	96.9 (0.2)	81.4 (1.1)	85.9 (3.1)	97.9 (-0.6)
RCDB	97.0 (0.2)	**82.2** (0.6)	86.5 (2.4)	**98.9** (0.2)
ExFractalDB	**97.2** (0.4)	81.8 (0.2)	**87.0** (1.0)	**98.9** (0.6)

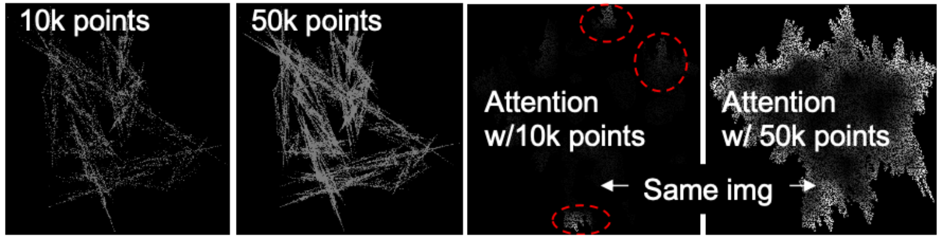


(a) N (b) r (c) l_w (d) $\mathbf{o}$ (e) λ

Different #points to render



(a) C10 (b) C100 (c) Cars (d) Flowers



Synthetic Data is Competitive with Real Data

Tokyo Tech

Classification (ImageNet-1k)

Pre-training	Img	Type	ViT-Ti	ViT-B
Scratch	–	–	72.6	79.8
ImageNet-21k	Real	SL	**74.1**	81.8
FractalDB-21k	Synth	FDSL	73.0	81.8
FractalDB-50k	Synth	FDSL	73.4	82.1
ExFractalDB-21k	Synth	FDSL	73.6	**82.7**
ExFractalDB-50k	Synth	FDSL	**73.7**	82.5
RCDB-21k	Synth	FDSL	73.1	82.4
RCDB-50k	Synth	FDSL	73.4	**82.6**

Detection, Segmentation (COCO)

Pre-training	COCO Det AP_{50} / AP / AP_{75}	COCO Inst AP_{50} / AP /
Scratch	63.7 / 42.2 / 46.1	60.7 / 38.5 /
ImageNet-1k	69.2 / 48.2 / 53.0	66.6 / 43.1 /
ImageNet-21k	**70.7 / 48.8 / 53.2**	**67.7 / 43.6 /**
ExFractalDB-1k	69.1 / **48.0 / 52.8**	66.3 / **42.8** /
ExFractalDB-21k	**69.2 / 48.0** / 52.6	**66.4 / 42.8** /
RCDB-1k	68.3 / 47.4 / 51.9	65.7 / 42.2 /
RCDB-21k	67.7 / 46.6 / 51.2	64.8 / 41.6 /

- ImageNet-21k is one of the largest open datasets
- Our synthetic dataset gives slightly better accuracy

H. Kataoka, R. Hayamizu, R. Yamada, K. Nakashima, S. Takashima, X. Zhang, E. J. Martinez-Noriega, N. Inoue, R. Yokota, "Replacing Labeled Real-image Datasets with Auto-generated Contours", CVPR2022.

Pre-training on Synthetic Data at Scale

ViT-B: epoch, effect of fine-tuning image size

pre-dataset	paper	shard	N	batch_size	lr	epochs	acc
mvf50k	deit	✔	224	8192	1.0e-3	40	**82.5** (main)
mvf50k	deit	✔	**384**	8192	1.0e-3	40	83.6
mvf50k	deit	✔	224	8192	1.0e-3	**90**	82.5
mvf50k	deit	✔	**384**	8192	1.0e-3	**90**	83.4
-	-	-	-	-	-	-	-
rc50k	deit	✔	224	8192	1.0e-3	40	**82.6** (main)
rc50k	deit	✔	**384**	8192	1.0e-3	40	83.6
rc50k	deit	✔	224	8192	1.0e-3	**90**	82.5
rc50k	deit	✔	**384**	8192	1.0e-3	**90**	83.7

Pretraining with 10M images increases the downstream accuracy

Accuracy with i21k is 81.8、JFT-300M is 84.2

We achieve 83.7, which is close to JFT-300M

ViT-B vs ViT-L

pre-dataset	shard	model	N	batch_size	lr	epochs	acc
i21k	✔	base	224	8192	1.0e-3	90	81.8
i21k	✔	large	224	2048	5.0e-4	90	79.8
mvf21k	✖	base	224	8192	1.0e-3	90	82.7
mvf50k	✔	base	224	8192	1.0e-3	40	82.5
mvf100k	✔	base	224	8192	1.0e-3	20	82.7
mvf21k	✔	large	224	2048	5.0e-4	90	80.4
mvf50k	✔	large	224	2048	5.0e-4	40	81.0

Something is not working for ViT-L

Even ImageNet-21k results are bad for ViT-L

We need to find the optimal hyperparameters for ViT-L

Why does this work so well?

- We are only pre-training
  - The model is fine-tuned on real data
  - Your visual cortex is ready to learn when you are born
  - Lower layers only learn local features
  - Semantics are dealt with at higher layers

- Advantage of synthetic datasets
  - They can be improved continuously
  - Real datasets are static and can only increase in quantity
  - They can be investigated systematically
  - What makes a good vision dataset?

Tokyo Tech

Thank You