Scaling Vision Transformers to 22 Billion Parameters

PMLR 2023

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, Rodolphe Jenatton, Lucas Beyer, Michael Tschannen, Anurag Arnab, Xiao Wang, Carlos Riquelme, Matthias Minderer, Joan Puigcerver, Utku Evci, Manoj Kumar, Sjoerd van Steenkiste, Gamaleldin F. Elsayed, Aravindh Mahendran, Fisher Yu, Avital Oliver, Fantine Huot, Jasmijn Bastings, Mark Patrick Collier, Alexey Gritsenko, Vighnesh Birodkar, Cristina Vasconcelos, Yi Tay, Thomas Mensink, Alexander Kolesnikov, Filip Pavetić, Dustin Tran, Thomas Kipf, Mario Lučić, Xiaohua Zhai, Daniel Keysers, Jeremiah Harmsen, Neil Houlsby

Motivation

- Scaling has been one of the most important trends in the last several years.

Model	Size
GPT3 [55]	175B
PanGU- α [75]	207B
OPT [81]	175B
PaLM [56]	540B
BLOOM [69]	176B
MT-NLG [97]	530B
Gopher [59]	280B
Chinchilla [34]	70B
Galactica [35]	120B
LaMDA [63]	137B
Jurassic-1 [91]	178B
LLaMA [57]	65B
GLM-130B [83]	130B
T5 [73]	11B

a) Natural Language Processing

Motivation

- Scaling has been one of the most important trends in the last several years.

Model	Size
GPT3 [55]	175B
PanGU- α [75]	207B
OPT [81]	175B
PaLM [56]	540B
BLOOM [69]	176B
MT-NLG [97]	530B
Gopher [59]	280B
Chinchilla [34]	70B
Galactica [35]	120B
LaMDA [63]	137B
Jurassic-1 [91]	178B
LLaMA [57]	65B
GLM-130B [83]	130B
T5 [73]	11B

a) Natural Language Processing

Model	Params
ViT-Base	86M
ViT-Large	307M
ViT-Huge	632M

b) Computer Vision

Motivation

- Scaling has been one of the most important trends in the last several years.

Model	Size
GPT3 [55]	175B
PanGU- α [75]	207B
OPT [81]	175B
PaLM [56]	540B
BLOOM [69]	176B
MT-NLG [97]	530B
Gopher [59]	280B
Chinchilla [34]	70B
Galactica [35]	120B
LaMDA [63]	137B
Jurassic-1 [91]	178B
LLaMA [57]	65B
GLM-130B [83]	130B
T5 [73]	11B

a) Natural Language Processing

Model	Params
ViT-Base	86M
ViT-Large	307M
ViT-Huge	632M

b) Computer Vision

~1000x times smaller SOTA models in CV

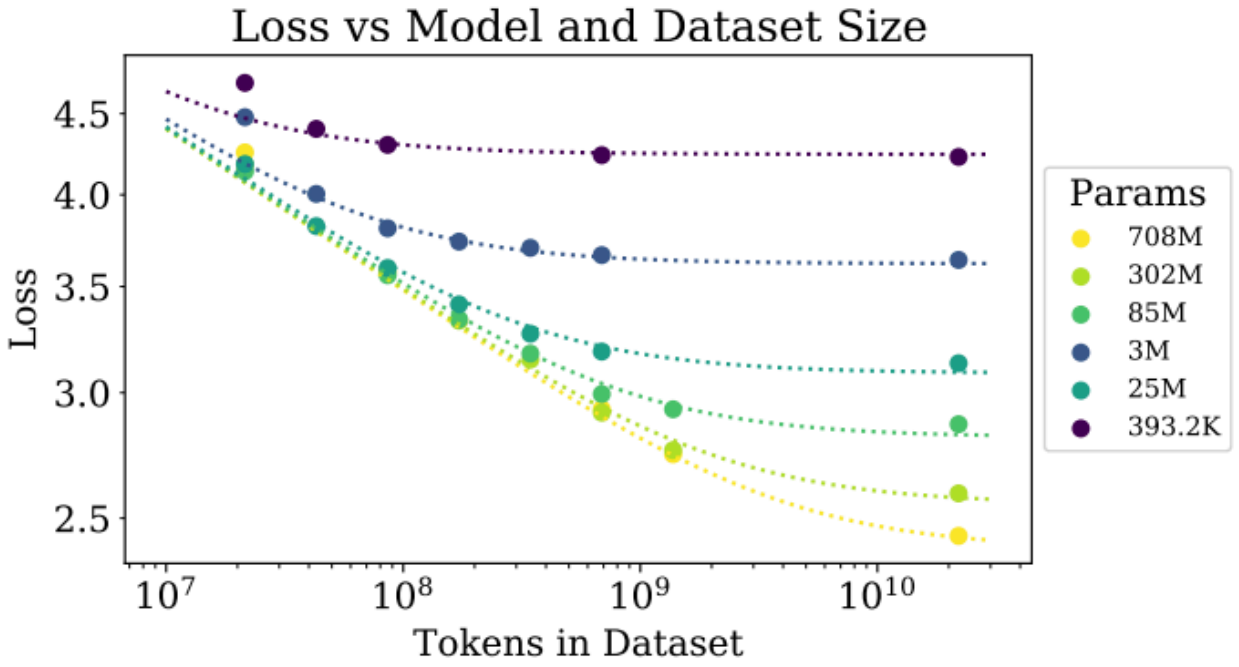
Scaling Laws for Transformers in NLP

1. Performance depends strongly on scale (model size, dataset size, the amount of compute) and weakly on model shape (model depth or width).
2. Performance improves predictably as long as we scale up the model and data in tandem.
3. Large models are more sample-efficient than small models.

Kaplan et al. “Scaling Laws for Neural Language Models,” 2020.

Loss vs Model and Data Size

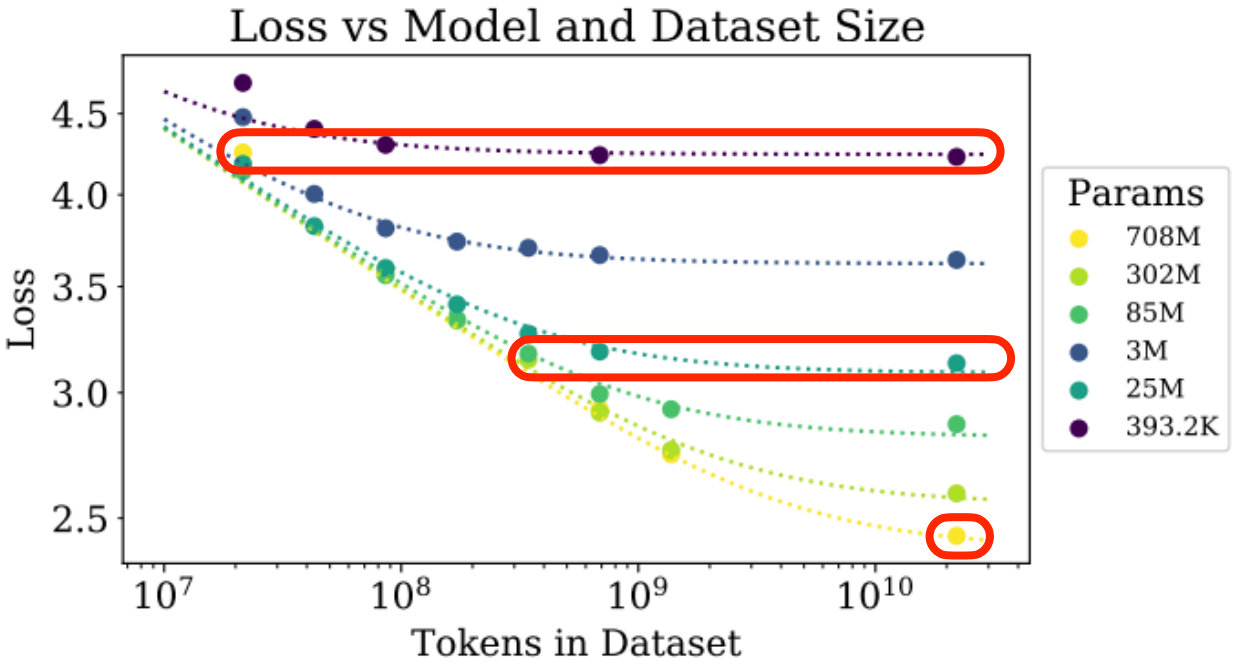
Large models are more sample-efficient than small models.



Kaplan et al. “Scaling Laws for Neural Language Models,” 2020.

Loss vs Model and Data Size

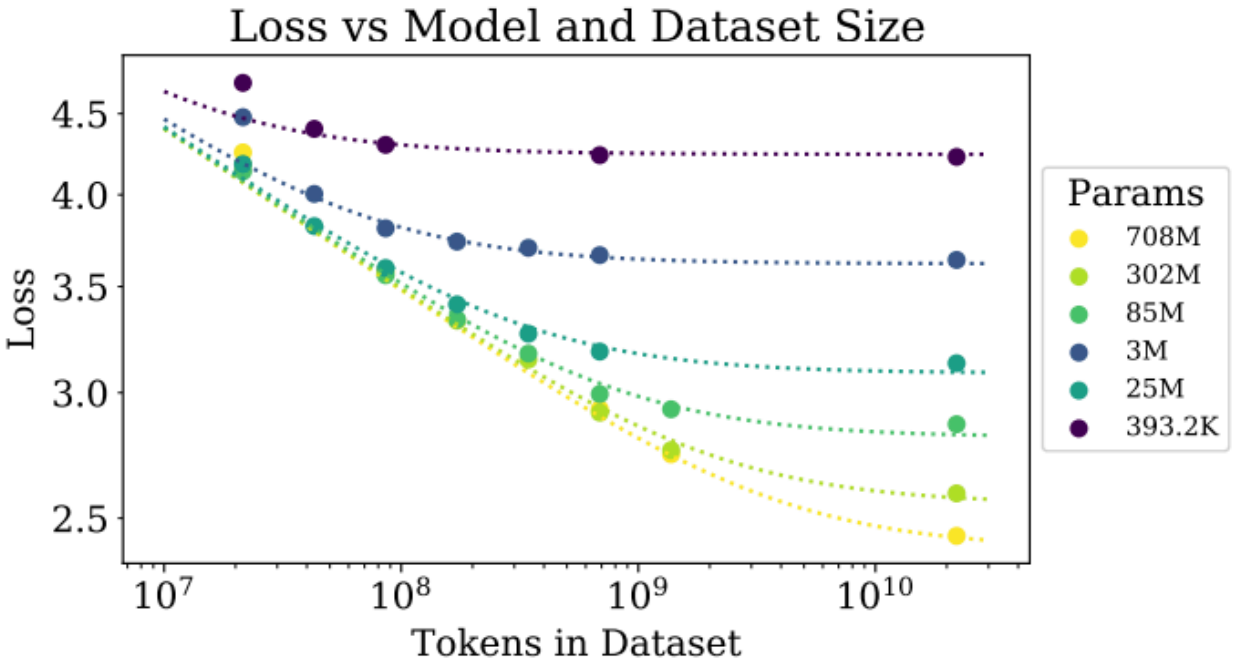
Large models are more sample-efficient than small models.



Kaplan et al. “Scaling Laws for Neural Language Models,” 2020.

Loss vs Model and Data Size

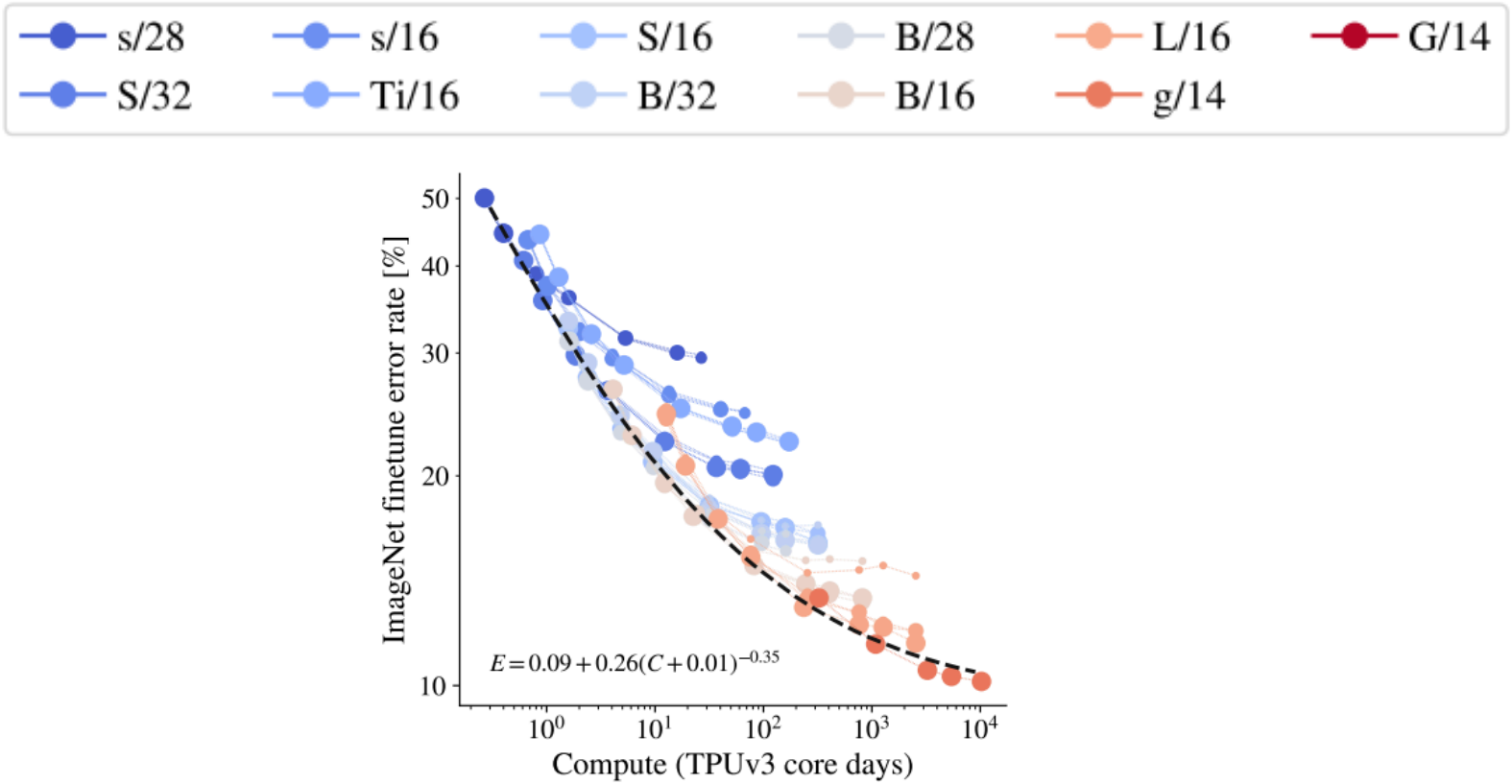
Large models are more sample-efficient than small models.



How do these findings transfer to vision domain?

Scaling in Computer Vision

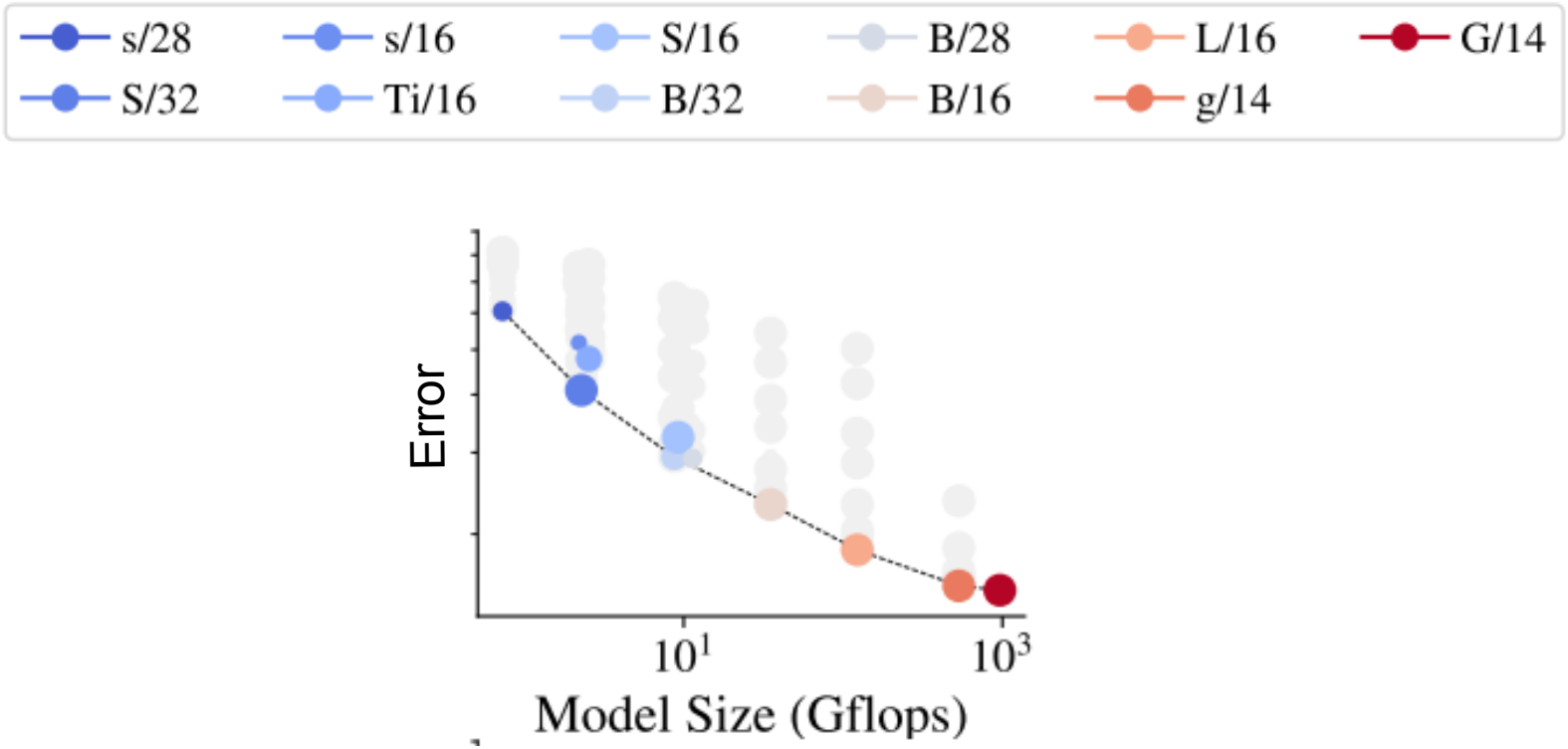
1. Scaling up compute, model and data together improves representation quality



Zhai et al. “Scaling Vision Transformers,” 2022.

Scaling in Computer Vision

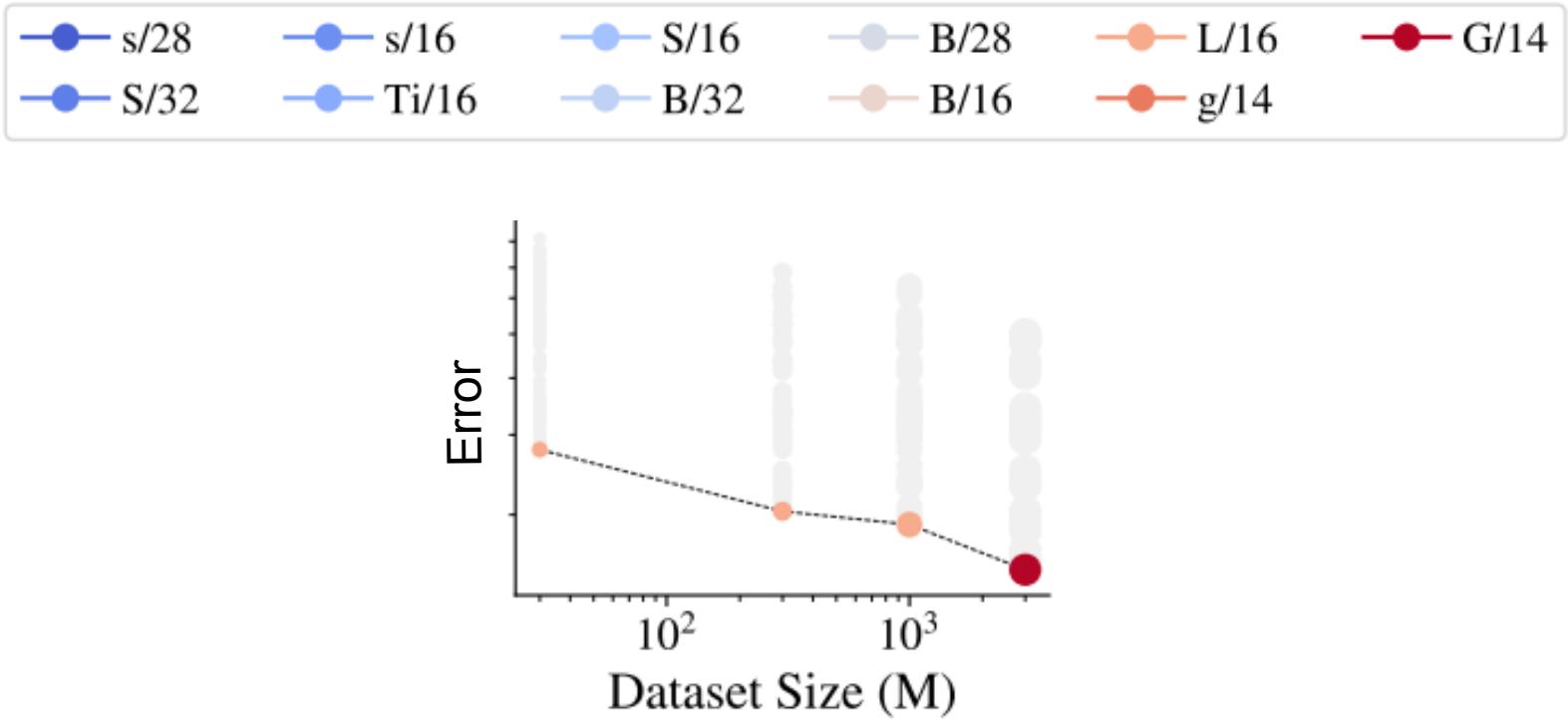
2. Representation quality can be bottlenecked by model size.



Zhai et al. “Scaling Vision Transformers,” 2022.

Scaling in Computer Vision

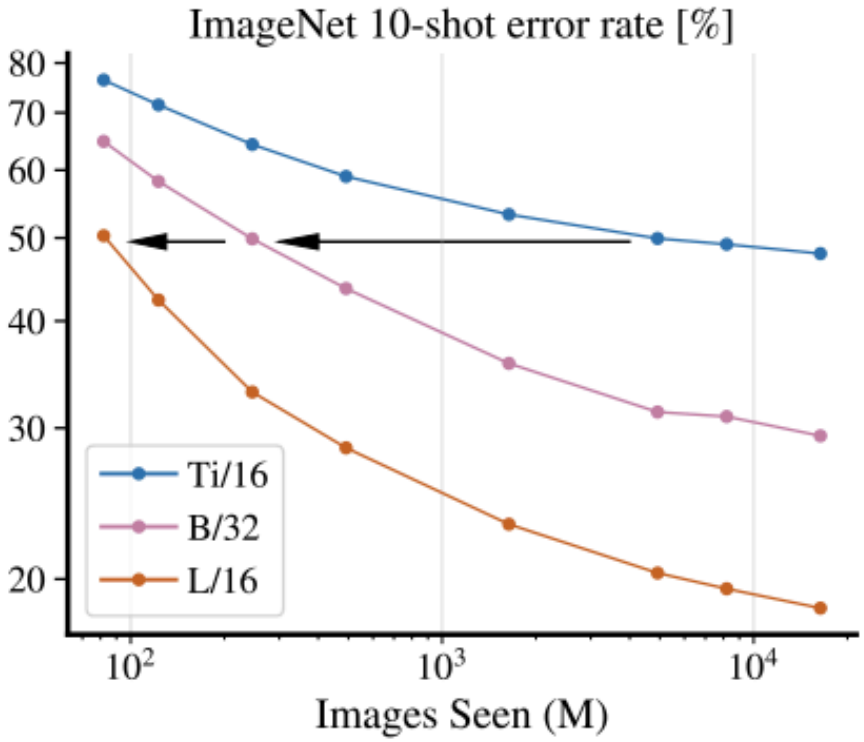
3. Large models benefit from additional data, even beyond 1B images.

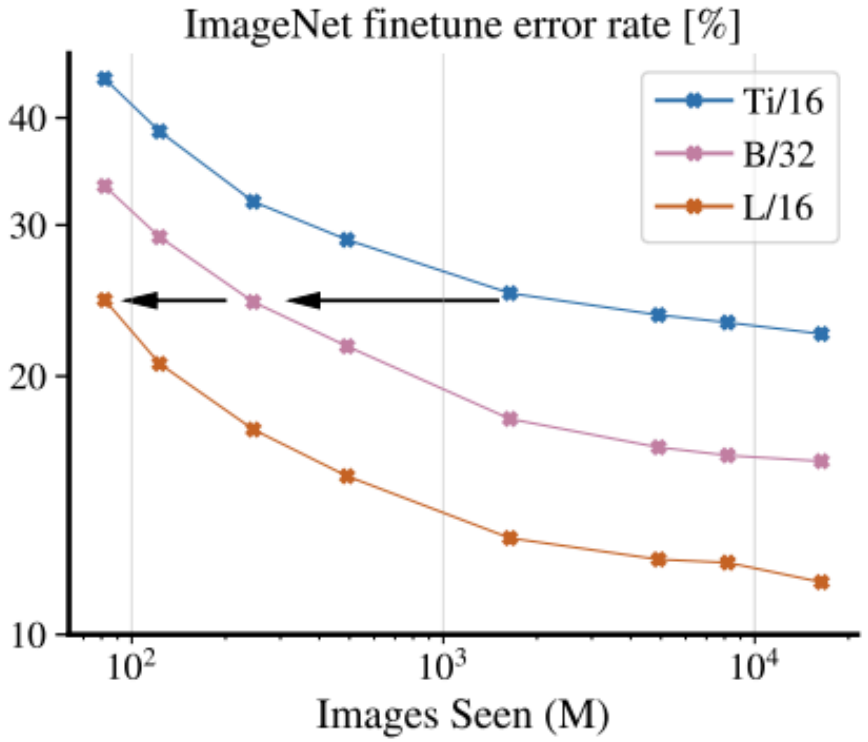


Zhai et al. “Scaling Vision Transformers,” 2022.

Scaling in Computer Vision

4. Large models are more sample efficient.

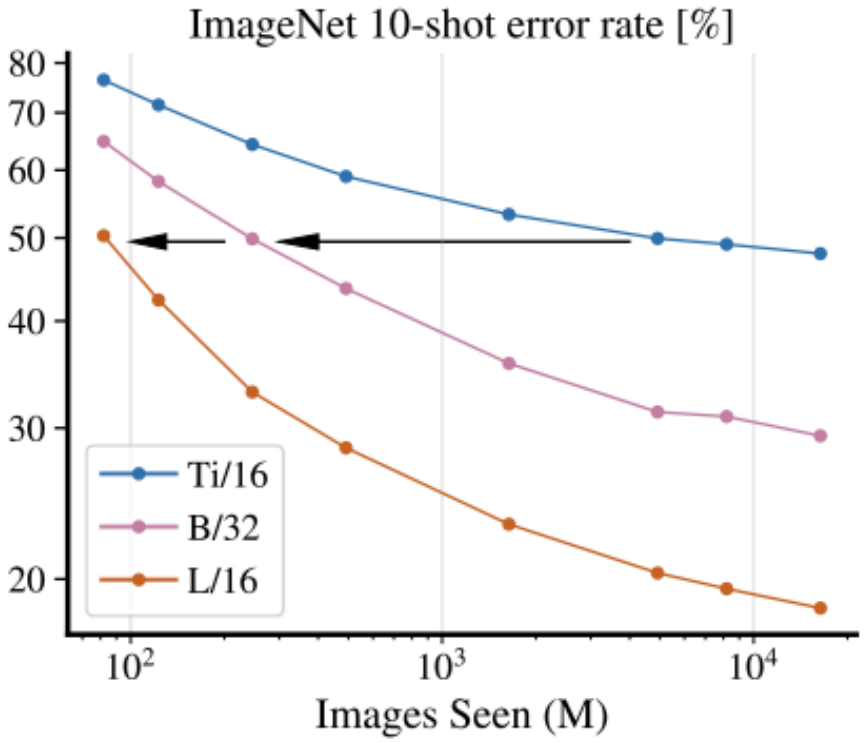


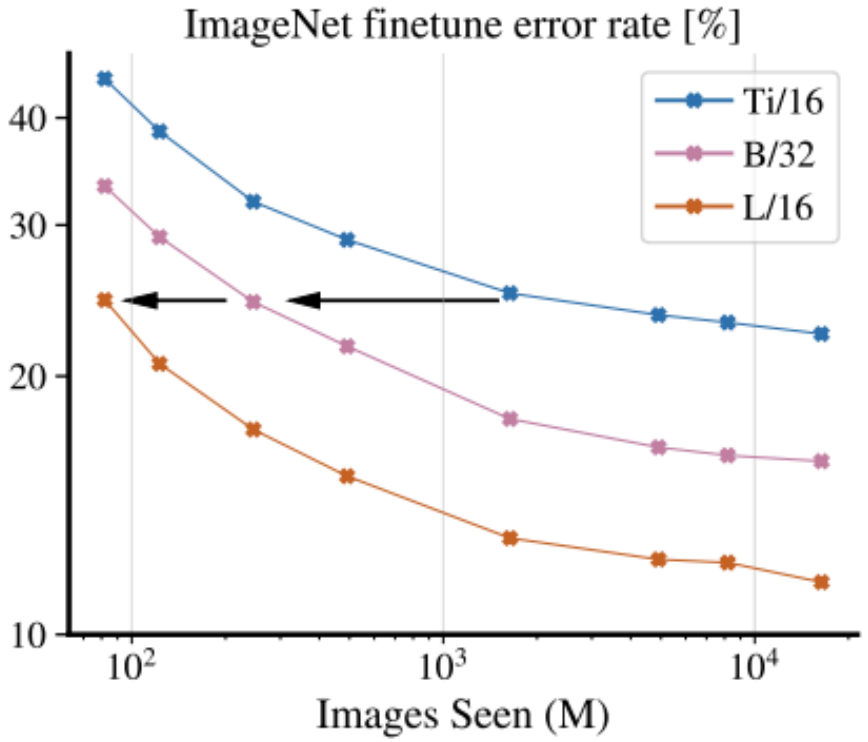


Zhai et al. “Scaling Vision Transformers,” 2022.

Scaling in Computer Vision

4. Large models are more sample efficient.





How can we efficiently/effectively scale computer vision models?

Architecture Details

Name	Width	Depth	MLP	Heads	Params [M]
ViT-G	1664	48	8192	16	1843
ViT-e	1792	56	15360	16	3926
ViT-22B	6144	48	24576	48	21743

ViT-22B Implementation Details

- The authors introduce three main modifications to improve efficiency and training stability at scale:
    1. Parallel Layers
    2. QK Normalization
    3. Omitting Bias Vectors on QKV Projections and Layer Norms

ViT-22B Implementation Details

- The authors introduce three main modifications to improve efficiency and training stability at scale:
    1. Parallel Layers
    2. QK Normalization
    3. Omitting Bias Vectors on QKV Projections and Layer Norms

Parallel Layers

- ViT-22B applies the Attention and MLP blocks in parallel, instead of sequentially as in the standard ViT.

$$y' = \text{LayerNorm}(x),$$
$$y = x + \text{MLP}(y') + \text{Attention}(y').$$

Can be computed in parallel

ViT-22B Implementation Details

- The authors introduce three main modifications to improve efficiency and training stability at scale:
    1. Parallel Layers
    2. QK Normalization
    3. Omitting Bias Vectors on QKV Projections and Layer Norms

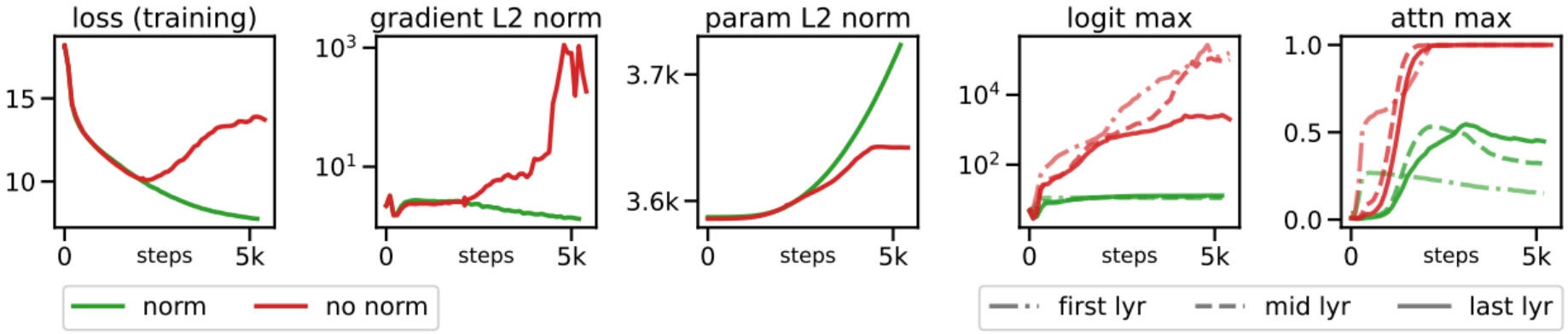
QK Normalization

- To avoid divergent training, the authors apply LayerNorm to the queries and keys before the dot product attention computation.

$$\text{softmax}\left[\frac{1}{\sqrt{d}}\text{LN}(XW^Q)(\text{LN}(XW^K))^T\right]$$

QK Normalization

- QK normalization prevents divergence due to uncontrolled attention logit growth.

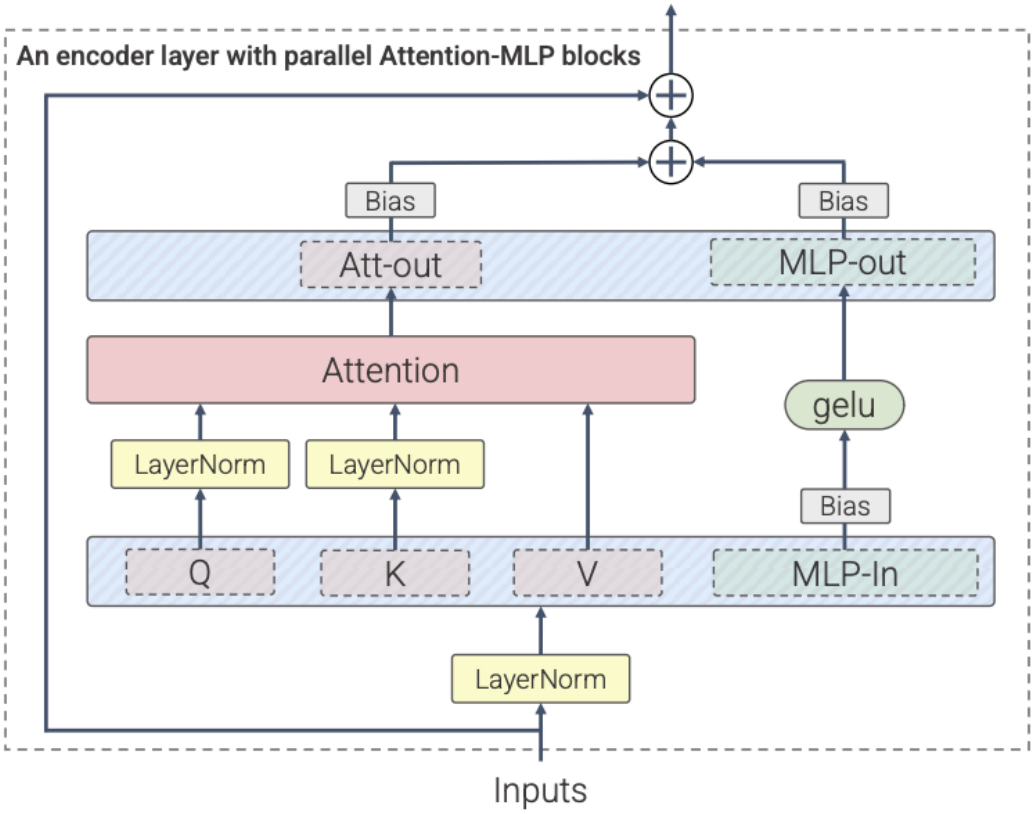


ViT-22B Implementation Details

- The authors introduce three main modifications to improve efficiency and training stability at scale:
    1. Parallel Layers
    2. QK Normalization
    3. Omitting Bias Vectors on QKV Projections and Layer Norms

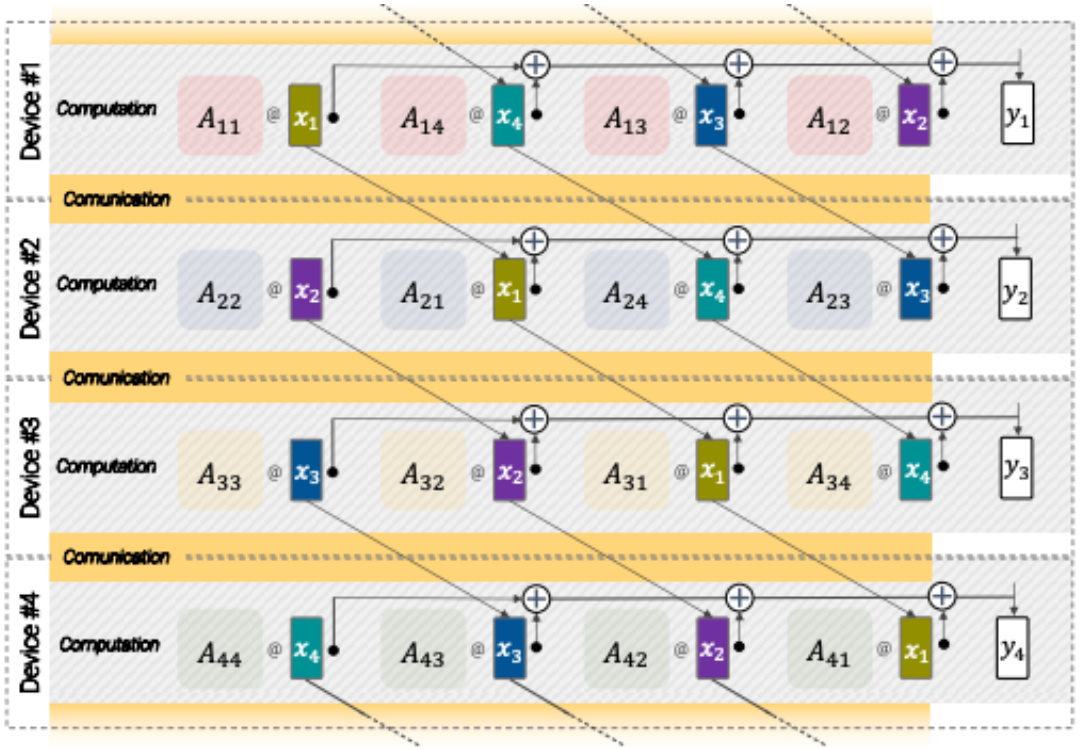
Omitting Bias Vectors

- Omitting bias vectors improves accelerator utilization (by 3%), without quality degradation



Training Infrastructure

- Training infrastructure leverages both model and data parallelism.



Other Details

- ViT-22B is trained on JFT extended to 4B images with 30K categories.
- Training is done using 1024 TPU V4 chips for 177K steps with a batch size of 65K (~3 epochs in total).

Model Card

Model Summary	
Model Architecture	Dense encoder-only model with 22 billion parameters. Transformer model architecture with variants to speed up and stabilize the training. For details, see Model Architecture (Section 2).
Input(s)	The model takes images as input.
Output(s)	The model generates a class label as output during pretraining.
Usage	
Application	The primary use is research on computer vision applications as a feature extractor that can be used in image recognition (finetuning, fewshot, linear-probing, zeroshot), dense prediction (semantic segmentation, depth estimation), video action recognition and so on. On top of that, ViT-22B is used in research that aim at understanding the impact of scaling vision transformers.
Known Caveats	When using ViT-22B, similar to any large scale model, it is difficult to understand how the model arrived at a specific decision, which could lead to lack of trust and accountability. Moreover, we demonstrated that ViT-22B is less prone to unintentional bias and enhances current vision backbones by reducing spurious correlations. However, this was done through limited studies and particular benchmarks. Besides, there is always a risk of misuse in harmful or deceitful contexts when it comes to large scale machine learning models. ViT-22B should not be used for downstream applications without a prior assessment and mitigation of the safety and fairness concerns specific to the downstream application. We recommend spending enough time and energy on mitigation the risk at the downstream application level.
System Type	
System Description	This is a standalone model.
Upstream Dependencies	None.
Downstream Dependencies	None.
Implementation Frameworks	
Hardware & Software: Training	Hardware: TPU v4 (Jouppi et al., 2020). Software: JAX (Bradbury et al., 2018), Flax (Heek et al., 2020), Scenic (Dehghani et al., 2022).
Hardware & Software: Deployment	Hardware: TPU v4 (Jouppi et al., 2020). Software: Scenic (Dehghani et al., 2022).
Compute Requirements	ViT-22B was trained on 1024 TPU V4 chips for 177K steps.

Model Card

Model Characteristics	
Model Initialization	The model is trained from a random initialization.
Model Status	This is a static model trained on an offline dataset.
Model Stats	ViT-22B model has 22 billion parameters.
Data Overview	
Training Dataset	ViT-22B is trained on a version of JFT (Sun et al., 2017), extended to contain around 4B images (Zhai et al., 2022a). See Section 4.1 for the description of datasets used to train ViT-22B.
Evaluation Dataset	We evaluate the ViT-22B on a wide variety of tasks and report the results on each individual tasks and datasets (Dehghani et al., 2021b). Specifically, we evaluate the models on: ADE20K (Zhou et al., 2017b), Berkeley Adobe Perceptual Patch Similarity (BAPPS) (Zhang et al., 2018), Birds (Wah et al., 2011), Caltech101 (Li et al., 2022), Cars (Krause et al., 2013), CelebA (Liu et al., 2015), Cifar-10 (Krizhevsky et al., 2009), Cifar-100 (Krizhevsky et al., 2009), CLEVR/count (Johnson et al., 2017), CLEVR/distance (Johnson et al., 2017), ColHist (Kather et al., 2016), DMLab (Beattie et al., 2016), dSprites/location (Matthey et al., 2017), dSprites/orientation (Matthey et al., 2017), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), Flowers102 (Nilsback and Zisserman, 2008), ImageNet (Deng et al., 2009), Inaturalist (Cui et al., 2018), ImageNet-v2 (Recht et al., 2019), ImageNet-R (Hendrycks et al., 2020), ImageNet-A (Hendrycks et al., 2021), ImageNet-C (Hendrycks and Dietterich, 2019), ImageNet-ReaL-H (Tran et al., 2022), Kinetics 400 (Kay et al., 2017), KITTI (Geiger et al., 2013), Moments in Time (Monfort et al., 2019), ObjectNet (Barbu et al., 2019), Pascal Context (Mottaghi et al., 2014), Pascal VOC (Everingham et al., 2010), Patch Camelyon (Teh and Taylor, 2019), Pets (Parkhi et al., 2012), Places365 (Zhou et al., 2017a), Resisc45 (Cheng et al., 2017), Retinopathy (Kaggle and EyePacs, 2015), SmallNORB/azimuth (LeCun et al., 2004), SmallNORB/elevation (LeCun et al., 2004), Sun397 (Xiao et al., 2010), SVHN (Netzer et al., 2011), UC Merced (Yang and Newsam, 2010), Waymo Open real-world driving dataset (Sun et al., 2020).

Experimental Setup

- After pretraining on JFT, transfer learning results are done using ViT-22B as a frozen feature extractor.

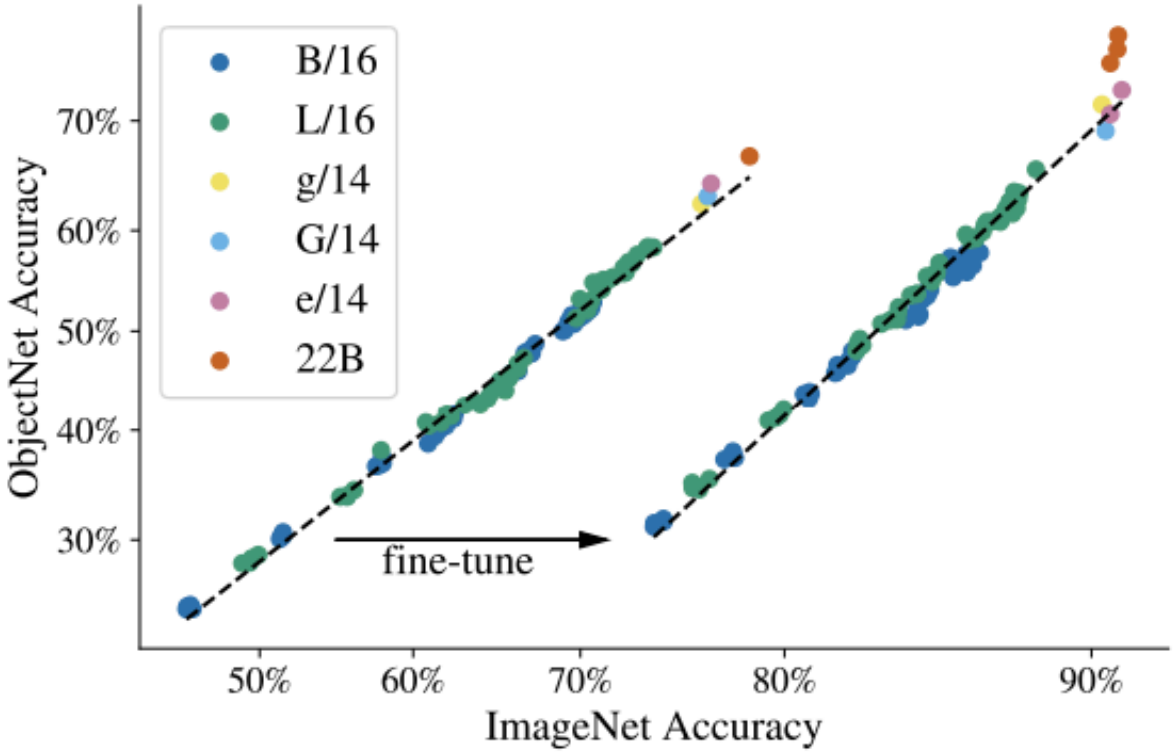
Image Classification

- Linear evaluation on Imagenet-1K with varying scale.

Model	IN	ReaL	INv2	ObjectNet	IN-R	IN-A
224px linear probe (frozen)						
B/32	80.18	86.00	69.56	46.03	75.03	31.2
B/16	84.20	88.79	75.07	56.01	82.50	52.67
ALIGN (360px)	85.5	-	-	-	-	-
L/16	86.66	90.05	78.57	63.84	89.92	67.96
g/14	88.51	90.50	81.10	68.84	92.33	77.51
G/14	88.98	90.60	81.32	69.55	91.74	78.79
e/14	89.26	90.74	82.51	71.54	**94.33**	81.56
22B	**89.51**	**90.94**	**83.15**	**74.30**	94.27	**83.80**
High-res fine-tuning						
L/16	88.5	90.4	80.4	-	-	-
FixNoisy-L2	88.5	90.9	80.8	-	-	-
ALIGN-L2	88.64	-	-	-	-	-
MaxViT-XL	89.53	-	-	-	-	-
G/14	90.45	90.81	83.33	70.53	-	-
e/14	90.9	91.1	84.3	72.0	-	-

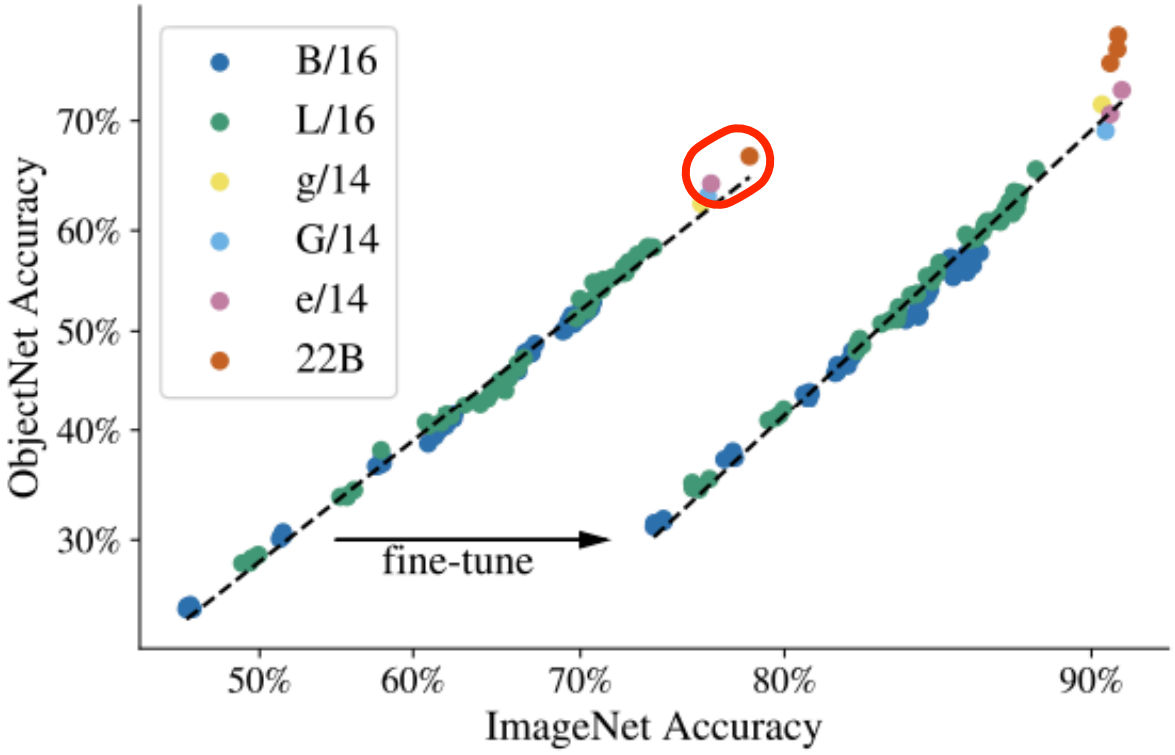
Out-of-distribution Classification

- Scaling the model increases out-of-distribution performance in line with the improvements on ImageNet



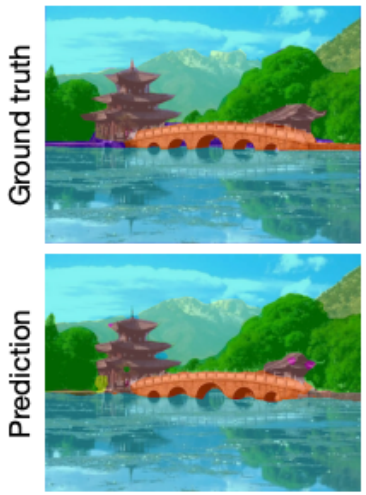
Out-of-distribution Classification

- Scaling the model increases out-of-distribution performance in line with the improvements on ImageNet

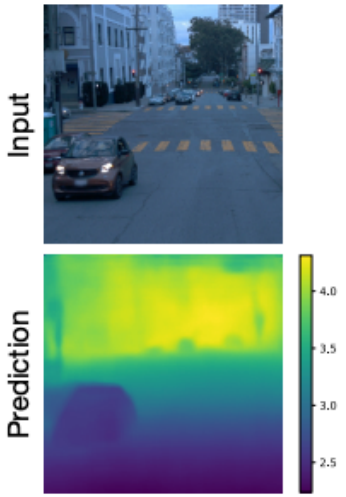


Dense Prediction Tasks

- The authors investigate transfer learning performance for dense prediction tasks.



(a) Semantic segmentation



(b) Depth estimation

a) Semantic segmentation results

Fraction of ADE20k train data	1/16	1/8	1/4	1/2	1
ViT-L (Touvron et al., 2022)	36.1	41.3	45.6	48.4	51.9
ViT-G (Zhai et al., 2022a)	42.4	47.0	50.2	52.4	**55.6**
ViT-22B (Ours)	**44.7**	**47.2**	**50.6**	**52.5**	54.9

	Model	MSE ↓	AbsRel ↓	$\delta \uparrow$ < 1.1	< 1.25	$< 1.25^2$
DPT	ViT-L	0.027	0.121	0.594	0.871	0.972
	ViT-e	0.024	0.112	0.631	0.888	0.975
	ViT-22B	**0.021**	**0.095**	**0.702**	**0.909**	**0.979**
Linear	ViT-L	0.060	0.222	0.304	0.652	0.926
	ViT-e	0.053	0.204	0.332	0.687	0.938
	ViT-22B	**0.039**	**0.166**	**0.412**	**0.779**	**0.960**

b) Depth estimation results

Video Classification

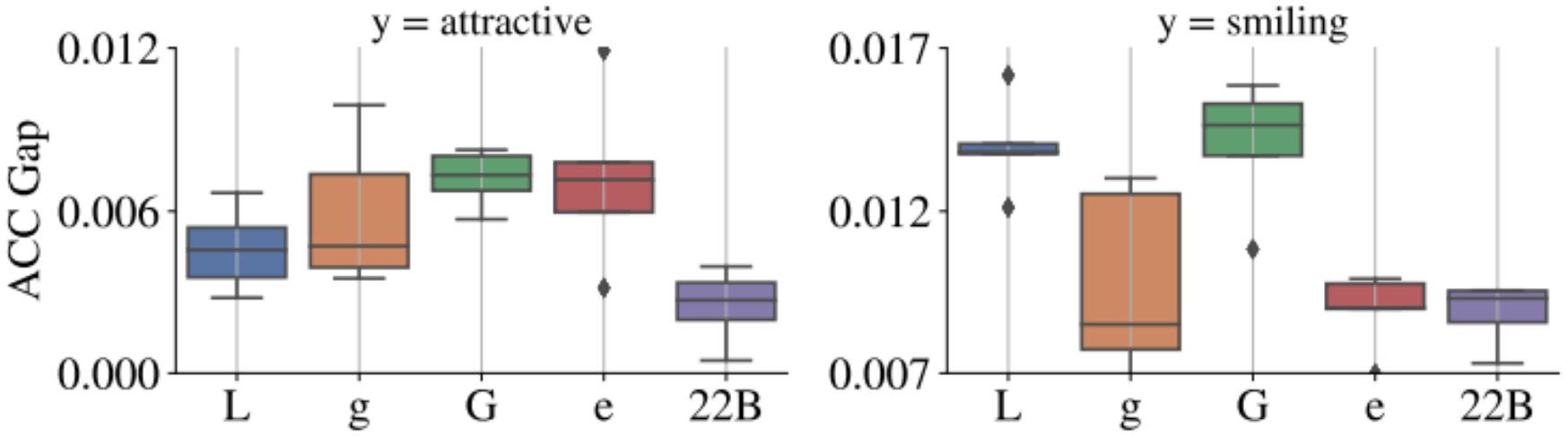
- The authors evaluate the quality of the representations learned by ViT-22B by adapting the model pretrained on images for video classification.

	Kinetics 400	Moments in Time
Frozen backbone		
CoCA*	88.0	47.4
ViT-e	86.5	43.6
ViT-22B	88.0	44.9
Fully finetuned SOTA	91.1	49.0

*Note that CoCA uses pre-pool spatial features and higher spatial resolution for both datasets. More details in Appendix F.

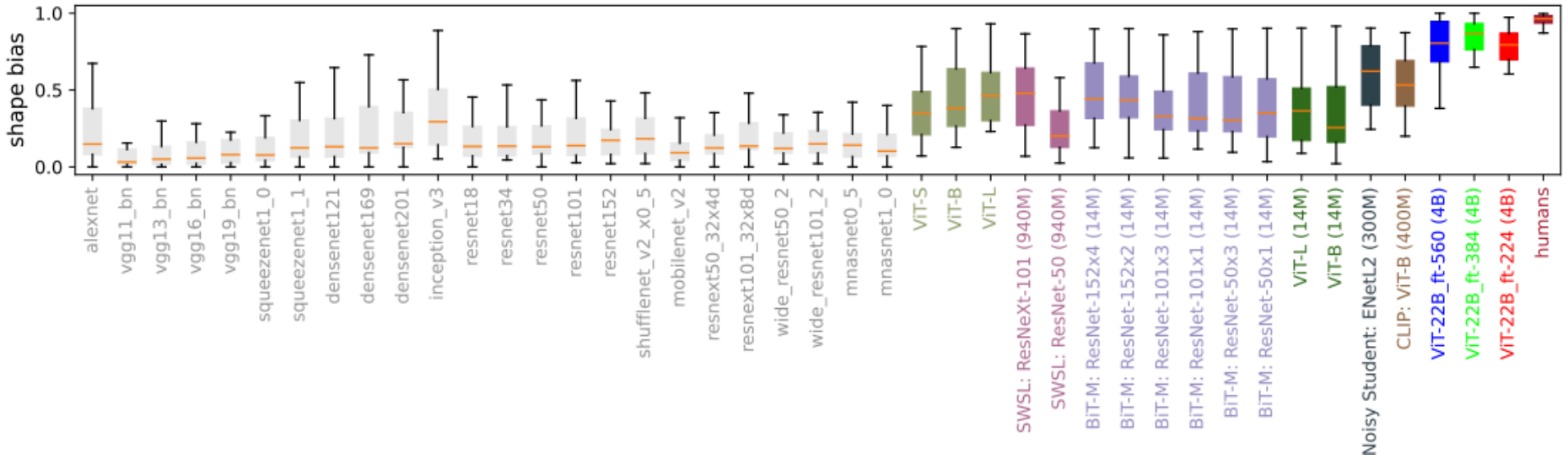
Fairness

- ViT-22B provides more equitable performance compared to smaller ViT architectures



Alignment with Human Perception

- Evaluating how well ViT-22B classification decisions align with human perception.



Distillation

- The authors perform model distillation to compress the ViT-22B into smaller, more widely usable ViTs.

Model		ImageNet1k
ViT-B/16	(Dosovitskiy et al., 2021) (JFT ckpt.)	84.2
	(Zhai et al., 2022a) (JFT ckpt.)	86.6
	(Touvron et al., 2022) (INet21k ckpt.)	86.7
	Distilled from ViT-22B (JFT ckpt.)	**88.6**
ViT-L/16	(Dosovitskiy et al., 2021) (JFT ckpt.)	87.1
	(Zhai et al., 2022a) (JFT ckpt.)	88.5
	(Touvron et al., 2022) (INet21k ckpt.)	87.7
	Distilled from ViT-22B (JFT ckpt.)	**89.6**

Zero-shot Classification