Google DeepMind

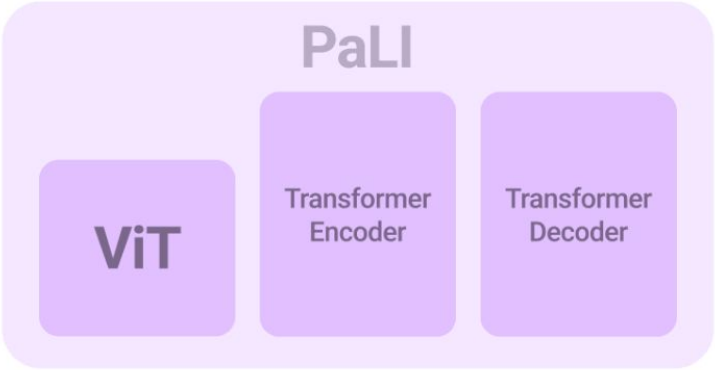
ViT-22B: Scaling Vision Transformers to 22 Billion Parameters

Piotr Padlewski

ICML - July 2023

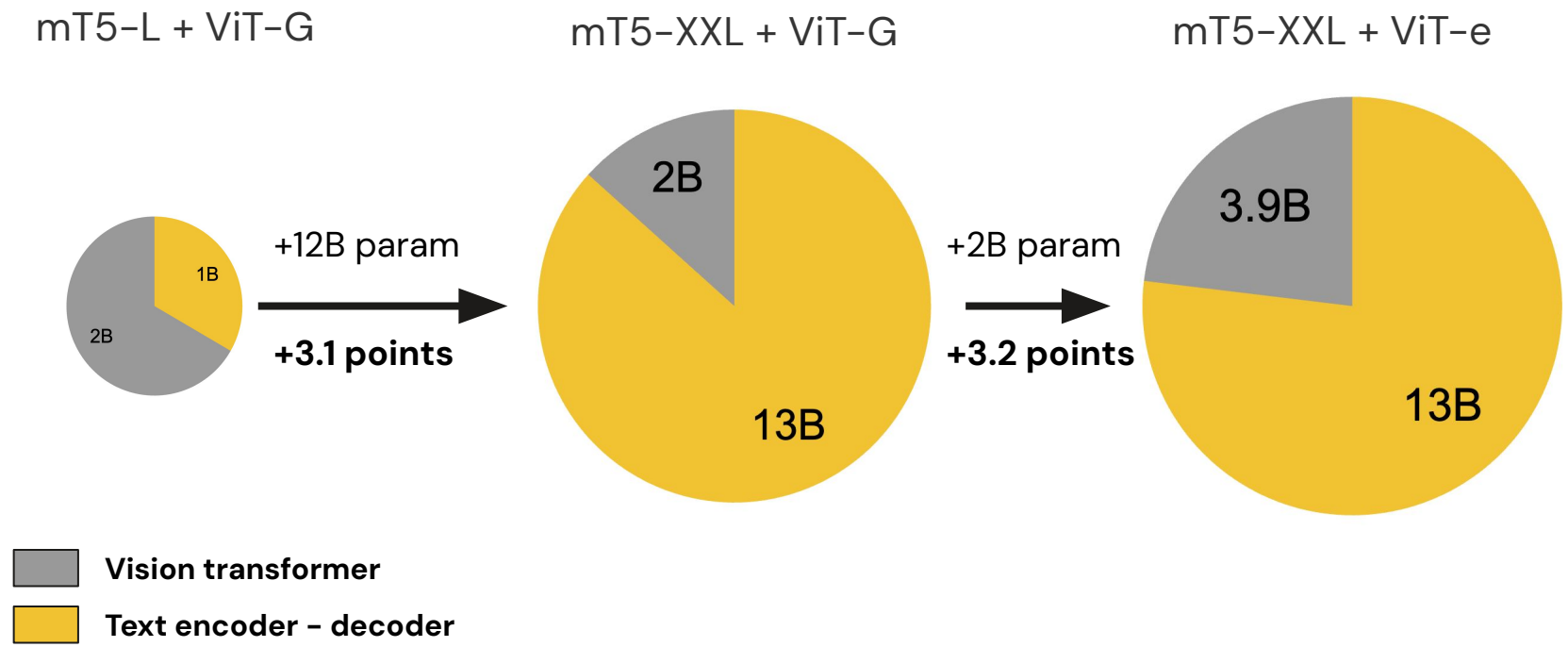
Google DeepMind

Is there headroom for scaling vision models?



Google DeepMind

Balancing model scaling and vision/text proportions

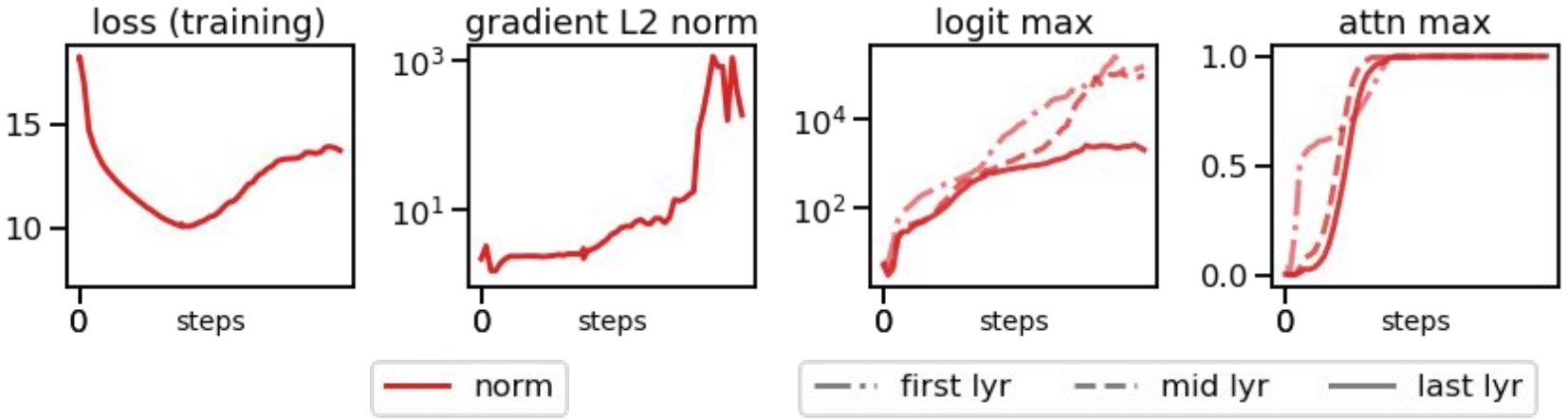


Chen et. al.: PaLI: A Jointly-Scaled Multilingual Language-Image Model, ICLR 2023

Google DeepMind

Stability: Numerical troubles (8B case)

Hit instabilities past ~8B params. The problem: uncontrolled attention logit growth.



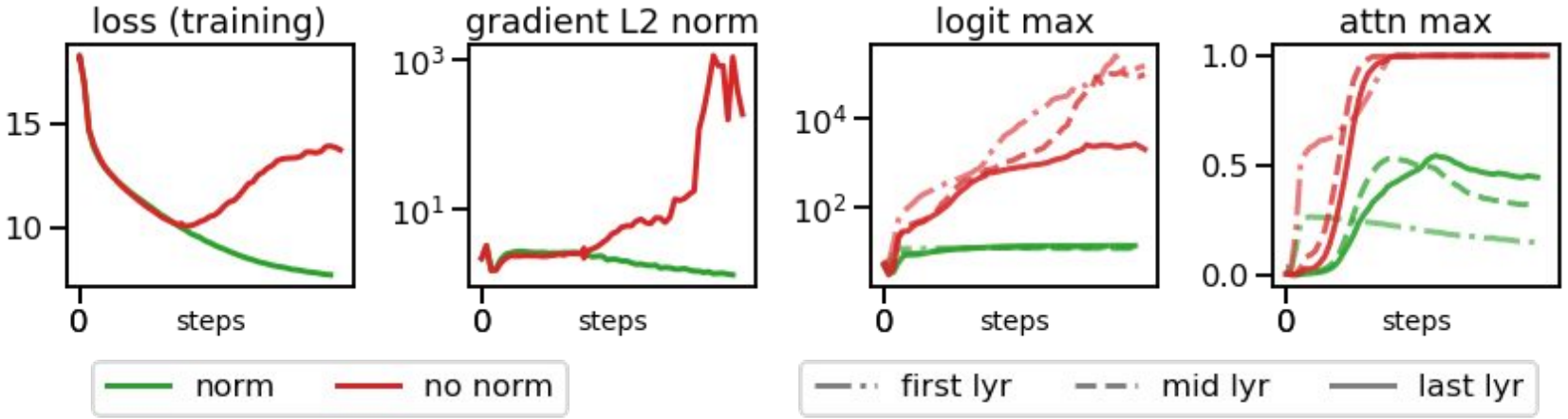
Google DeepMind

Stability: The remedy

QK Normalization prevents divergence.

Query LayerNorm Key LayerNorm

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

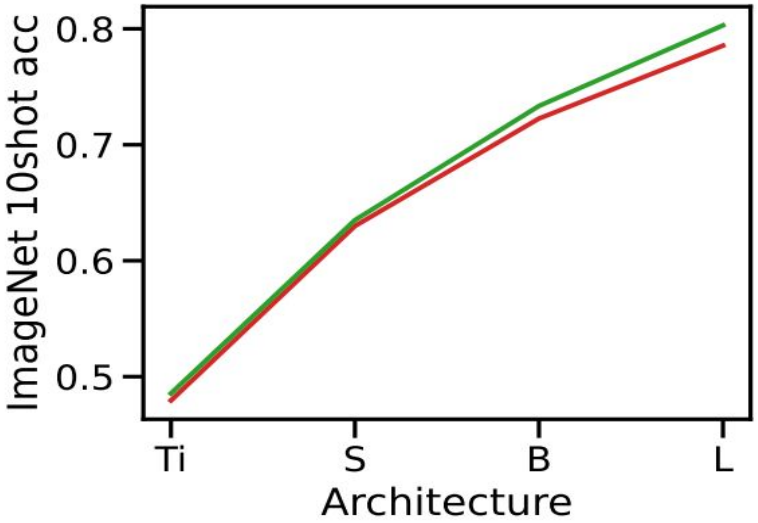


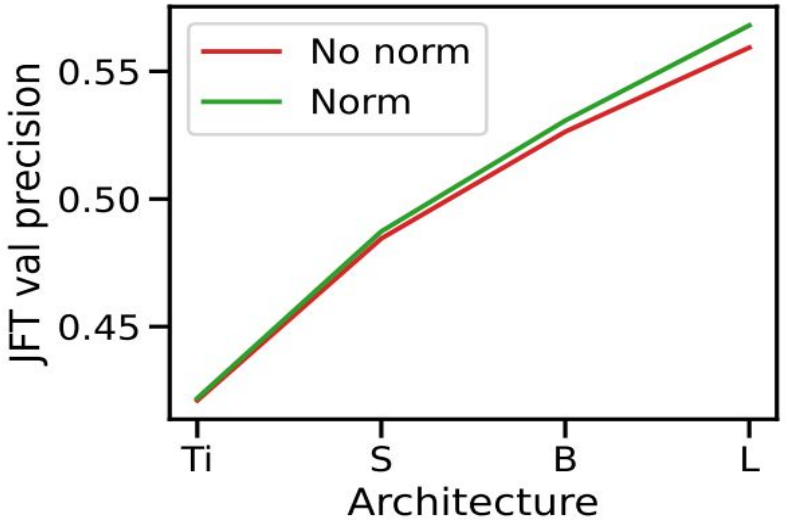
Google DeepMind

Stability: The remedy

QK Normalization prevents divergence.

Also helps at smaller scales, partly due to enabling larger learning rates

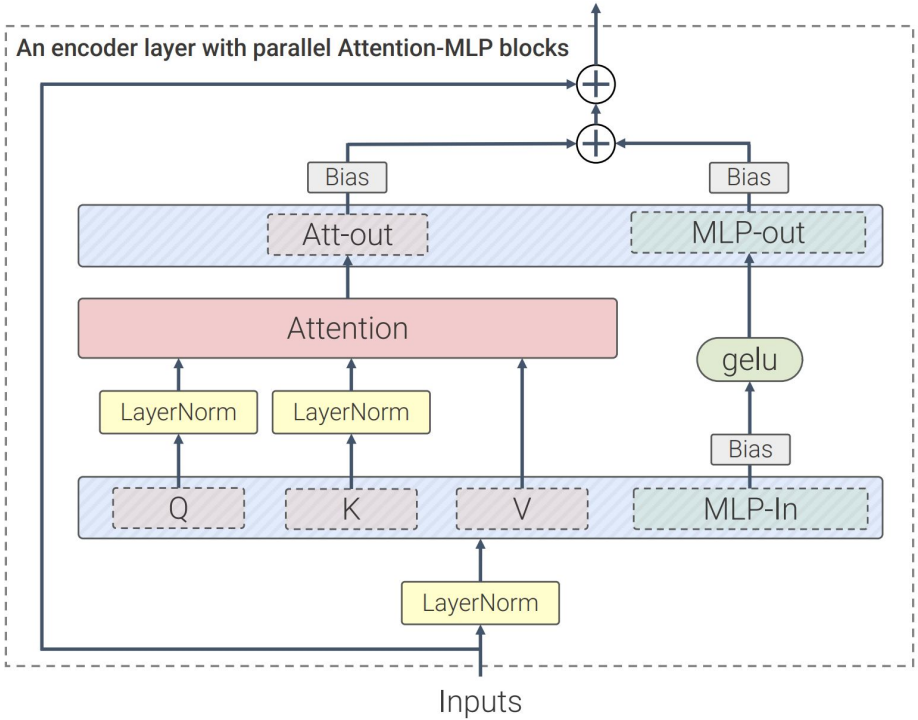




Google DeepMind

Speed: Modeling changes

- Parallel layers (like PaLM)
- No biases on QKV & LN

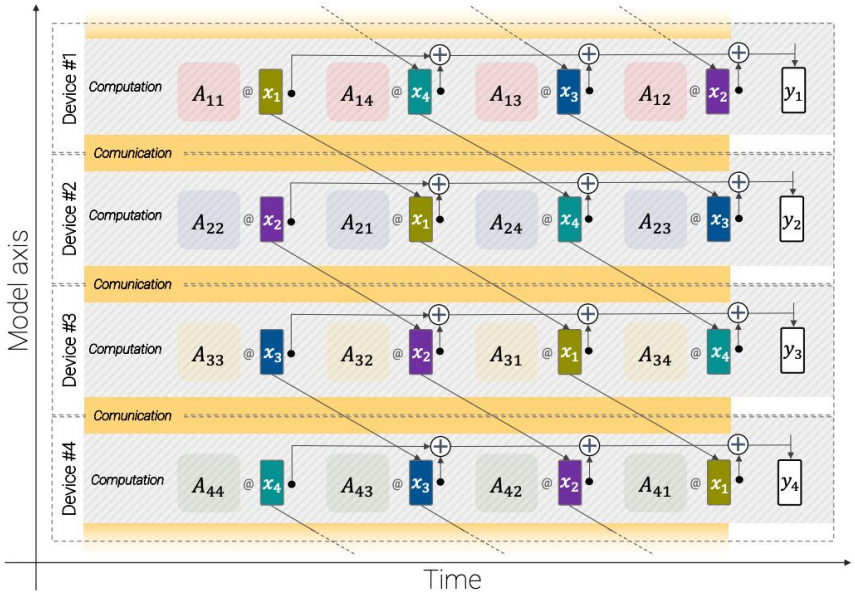


Chowdhery et. al. PaLM: Scaling Language Modeling with Pathways

Google DeepMind

Speed: Parallelism implementation details

- Optimized Scenic implementation using jax.xmap
  - *Async compute and communication*
  - *Customized for Transformer blocks*
- 1150 tokens/sec/core. 54.9% MFU
  - c.f. PaLM: 46.2%, ViT-e: 44.0%



(a) The matrix A is row-sharded across the devices.

Wang et. al.: Overlap communication with dependent computation via decomposition in large deep learning models
Chowdhery et. al.: PaLM: Scaling Language Modeling with Pathways
Chen et. al.: PaLI: A Jointly-Scaled Multilingual Language-Image Model (ViT-e)

Google DeepMind

Evaluation

Linear eval

Excellent, especially on OOD, almost matches fine-tuning (89.5% ImageNet @224x224)

Classification / OOD

ImageNet saturating, significant improvements on OOD evaluations

Video

Competitive results using per-frame representations (88% Kinetics 400)

Contrastive zero-shot

Outperforms ViT-e/g, a bit worse than SOTA on ImageNet, but better OOD (87.6% ObjectNet)

Dense prediction

Competitive (not SOTA) with frozen tower on semantic segmentation, monocular depth estimation

Fairness

Favorable trade-off between demographic parity bias & performance

Human alignment

Highest ever recorded shape bias of artificial NN (close to humans)

Calibration

Surprisingly, better calibrated than smaller equivalents (as well as more accuracy)

Reliability (PLEX)

Gains on all metrics

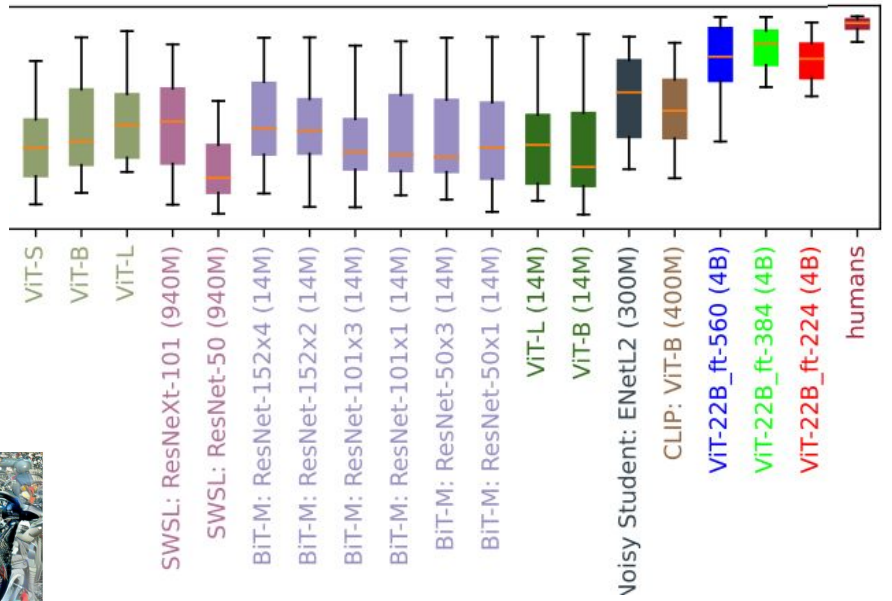
Distillation

New SOTA at smaller sizes. (ImageNet: 88.1% w/ B/16; 89.3% w/ L/16).

Google DeepMind

Alignment to human visual perception

- Humans are at 96% shape / 4% texture bias
- ViT-22B-384 is at 87% shape bias / 13% texture bias
- Other models: 20–30% shape bias / 70–80% texture bias



Geirhos et al.: ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. ICLR 2019

Google DeepMind

ViT-22B as frozen feature extractor

- Frozen representations approach full fine-tuning numbers
- Training of high-res fine-tuning of ViT-e is more **expensive** than training frozen feature extractor of ViT-22B

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	-	-

Google DeepMind

Distilled ViT-22B to ViT-B and ViT-L achieve SOTA (for that size)

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.)	**88.1**
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.)	**89.3**

ViT-B is 255x smaller (0.39%)

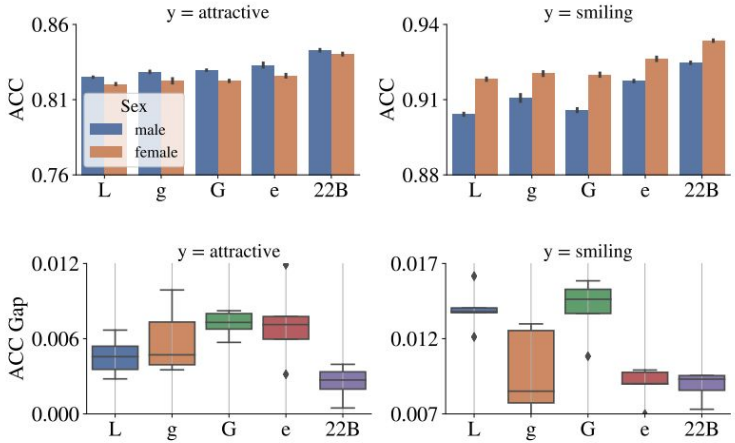
ViT-L is 71x smaller (1.39%)

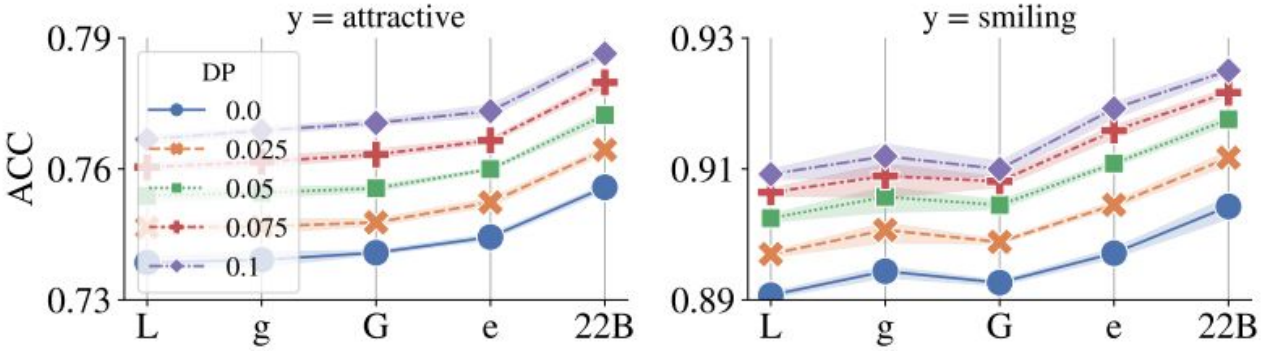
Image resolution: 384x384

Google DeepMind

Sub-group fairness

- All subgroups benefit from scale
- Subgroup disparity tends to decrease
- ViT-22B performs better at any level of demographic parity (after debiasing)





Google DeepMind

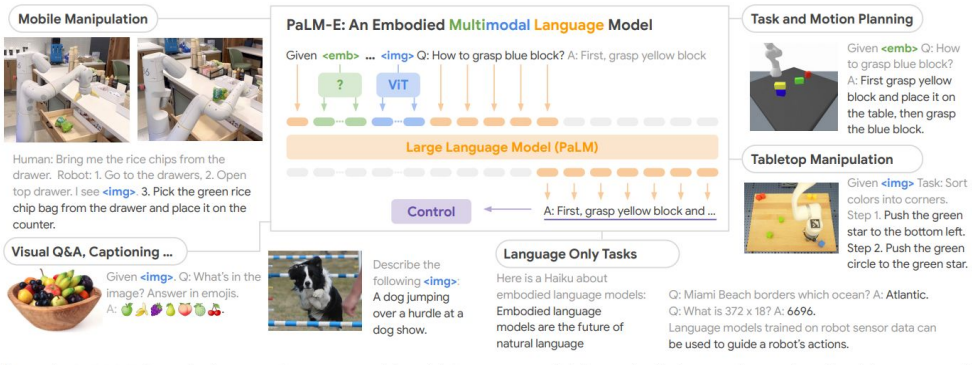
ViT-22B in the wild

PaLM-E: An Embodied Multimodal Language Model

Danny Driess [1 2] Fei Xia [1] Mehdi S. M. Sajjadi [3] Corey Lynch [1] Aakanksha Chowdhery [3]
Brian Ichter [1] Ayzaan Wahid [1] Jonathan Tompson [1] Quan Vuong [1] Tianhe Yu [1] Wenlong Huang [1]
Yevgen Chebotar [1] Pierre Sermanet [1] Daniel Duckworth [3] Sergey Levine [1] Vincent Vanhoucke [1]
Karol Hausman [1] Marc Toussaint [2] Klaus Greff [3] Andy Zeng [1] Igor Mordatch [3] Pete Florence [1]

[1]Robotics at Google [2]TU Berlin [3]Google Research

https://palm-e.github.io



PaLI-X: On Scaling up a Multilingual Vision and Language Model

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, AJ Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, Radu Soricut

Google Research
pali-communications@google.com

Abstract

We present the training recipe and results of scaling up PaLI-X, a multilingual vision and language model, both in terms of size of the components and the breadth of its training task mixture. Our model achieves new levels of performance on a wide-range of varied and complex tasks, including multiple image-based captioning and question-answering tasks, image-based document understanding and few-shot (in-context) learning, as well as object detection, video question answering, and video captioning. PaLI-X advances the state-of-the-art on most vision-and-language benchmarks considered (25+ of them). Finally, we observe emerging capabilities, such as complex counting and multilingual object detection, tasks that are not explicitly in the training mix.

Scaling vision transformers to 22 billion parameters

FRIDAY, MARCH 31, 2023

Posted by Piotr Padlewski and Josip Djolonga, Software Engineers, Google Research

Large Language Models (LLMs) like PaLM or GPT-3 showed that scaling transformers to hundreds of billions of parameters improves performance and unlocks emergent abilities. The biggest dense models for image understanding, however, have reached only 4 billion parameters, despite research indicating that promising multimodal models like PaLI continue to benefit from scaling vision models alongside their language counterparts. Motivated by this, and the results from scaling LLMs, we decided to undertake the next step in the journey of scaling the Vision Transformer.

In "Scaling Vision Transformers to 22 Billion Parameters", we introduce the biggest dense vision model, ViT-22B. It is 5.5x larger than the previous largest vision backbone, ViT-e, which has 4 billion parameters. To enable this scaling, ViT-22B incorporates ideas from scaling text models like PaLM, with improvements to both training stability (using QK normalization) and training efficiency (with a novel approach called asynchronous parallel linear operations). As a result of its modified architecture, efficient sharding recipe, and bespoke implementation, it was able to be trained on Cloud TPUs with a high hardware utilization[1]. ViT-22B advances the state of the art on many vision tasks using frozen representations, or with full fine-tuning. Further, the model has also been successfully used in PaLM-e, which showed that a large model combining ViT-22B with a language model can significantly advance the state of the art in robotics tasks.

Google DeepMind

Thanks for listening!

Scaling Vision Transformers to 22 Billion Parameters

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 A. 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*