Scaling Laws in Language and Vision Models

Machine Learning Reading Group Fall 2022

Betty Shea

2022-12-14

University of British Columbia

MLRG Fall 2022 theme: Transformers

THE UNIVERSITY OF BRITISH COLUMBIA

Papers

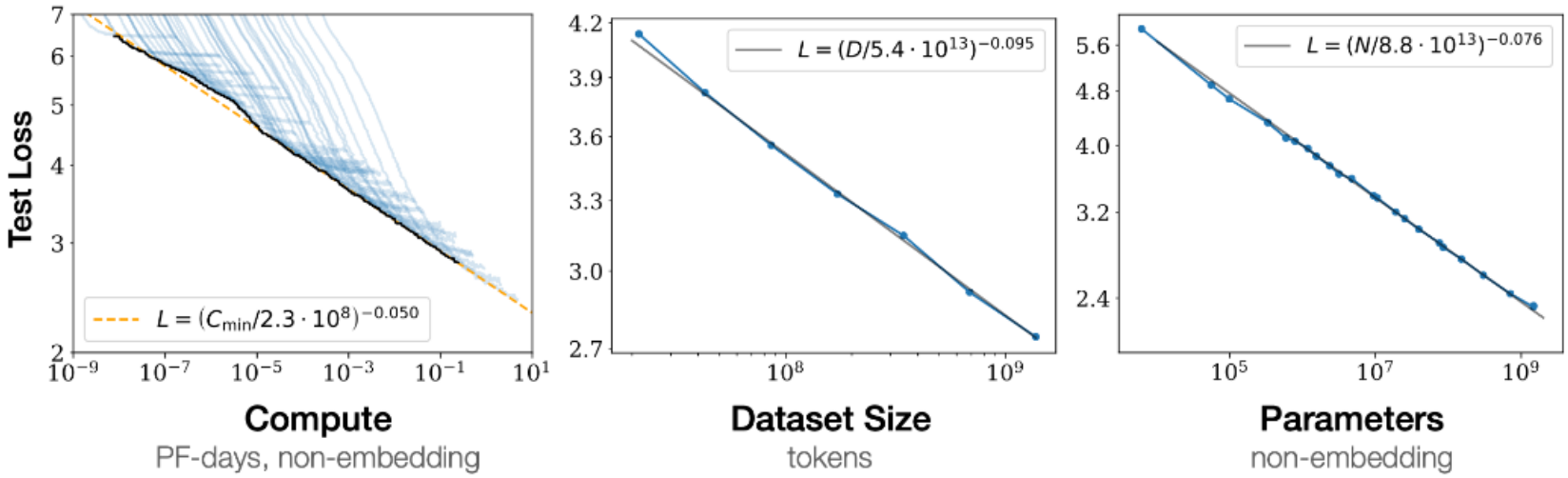
- Kaplan et al. (2020) Scaling Laws for Neural Language Models
- Abnar et al. (2021) Exploring the Limits of Large Scale Pre-training

MLRG Fall 2022 theme: Transformers

Previously

- Transformers and attention mechanisms
- Language models (GPT-3)
- Image recognition
- Chain of thought prompting
- CLIP

Transformers at Scale



MLRG Fall 2022 theme: Transformers

Today

- Bigger $\implies$ better?
- Limitations
- Generalization

Scale and transfer learning

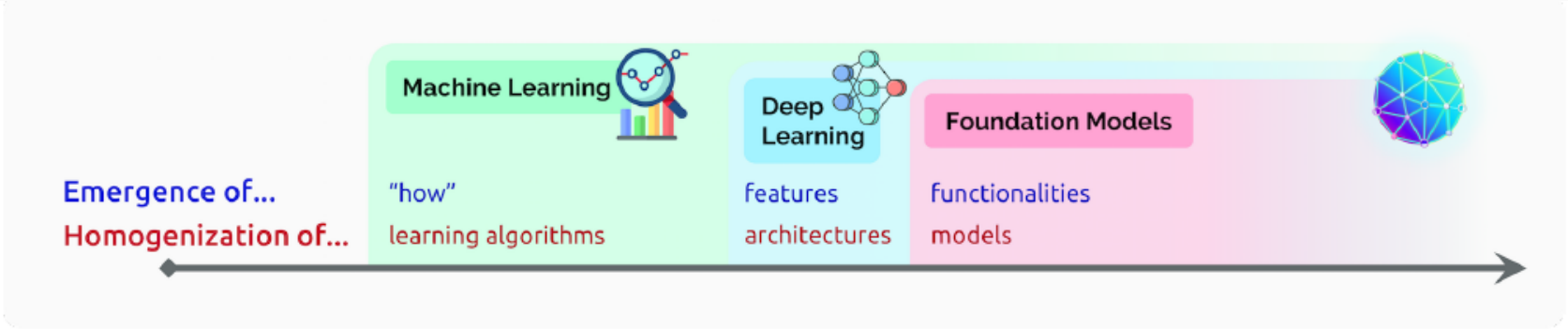


Fig. 1. The story of AI has been one of increasing *emergence* and *homogenization*. With the introduction of machine learning, *how* a task is performed emerges (is inferred automatically) from examples; with deep learning, the high-level features used for prediction emerge; and with foundation models, even advanced functionalities such as in-context learning emerge. At the same time, machine learning homogenizes learning algorithms (e.g., logistic regression), deep learning homogenizes model architectures (e.g., Convolutional Neural Networks), and foundation models homogenizes the model itself (e.g., GPT-3).

Bommasani et al. 2021

Foundation models

"A foundation model is any model that is trained on broad data ... that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks; current examples include BERT [Devlin et al. 2019], GPT-3 [Brown et al. 2020], and CLIP [Radford et al. 2021]."

Bommasani et al. 2021

Foundation models

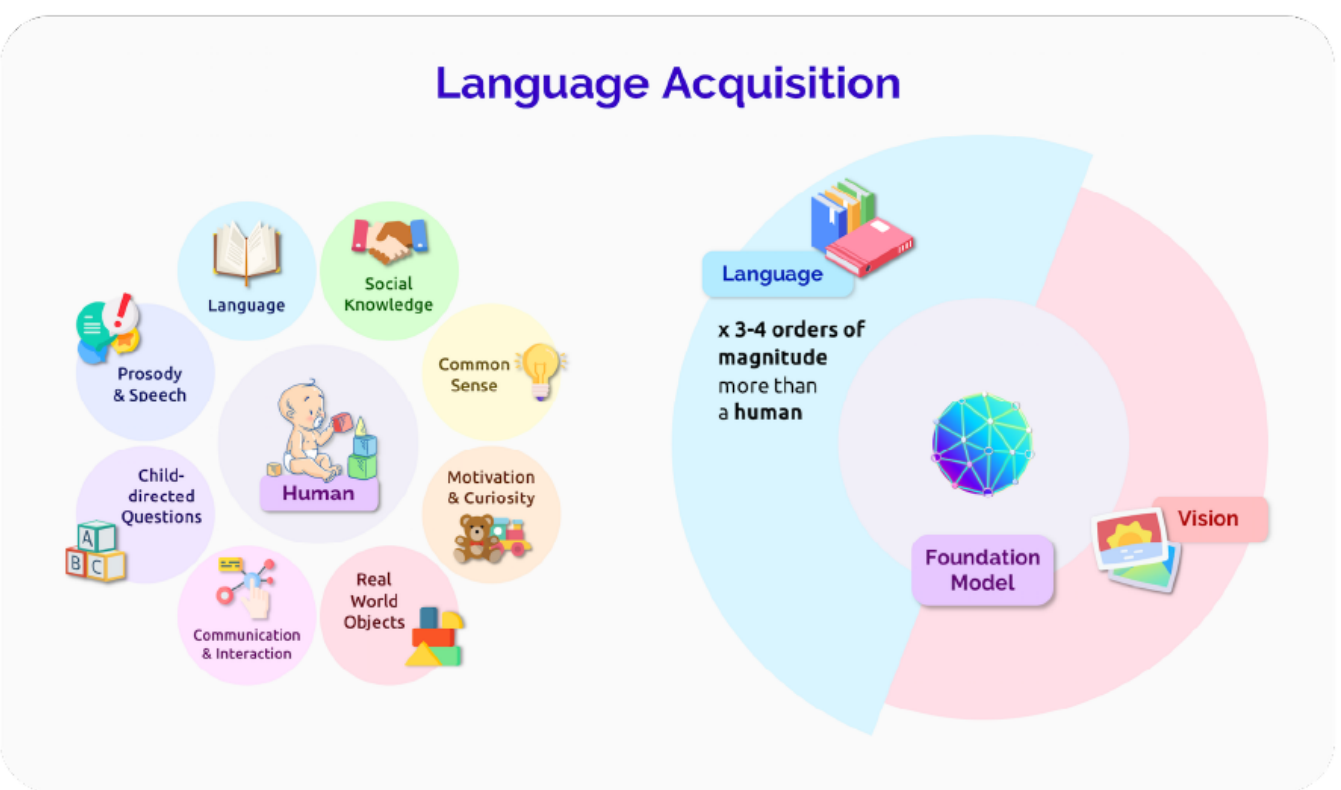


Fig. 6. Language Acquisition for humans and foundation models. While there are certainly different inductive biases between the human brain and foundation models, the ways that they learn language are also very different. Most saliently, humans interact with a physical and social world in which they have varied needs and desires, while foundation models mostly observe and model data produced by others.

Bommasani et al. 2021

Generalizability

"the feature of foundation models that has been most impactful in NLP is not their raw generation abilities but their surprising *generality and adaptability*: a single foundation model can be adapted in different ways in order to achieve many linguistic tasks"

Bommasani et al. 2021

Generalizability

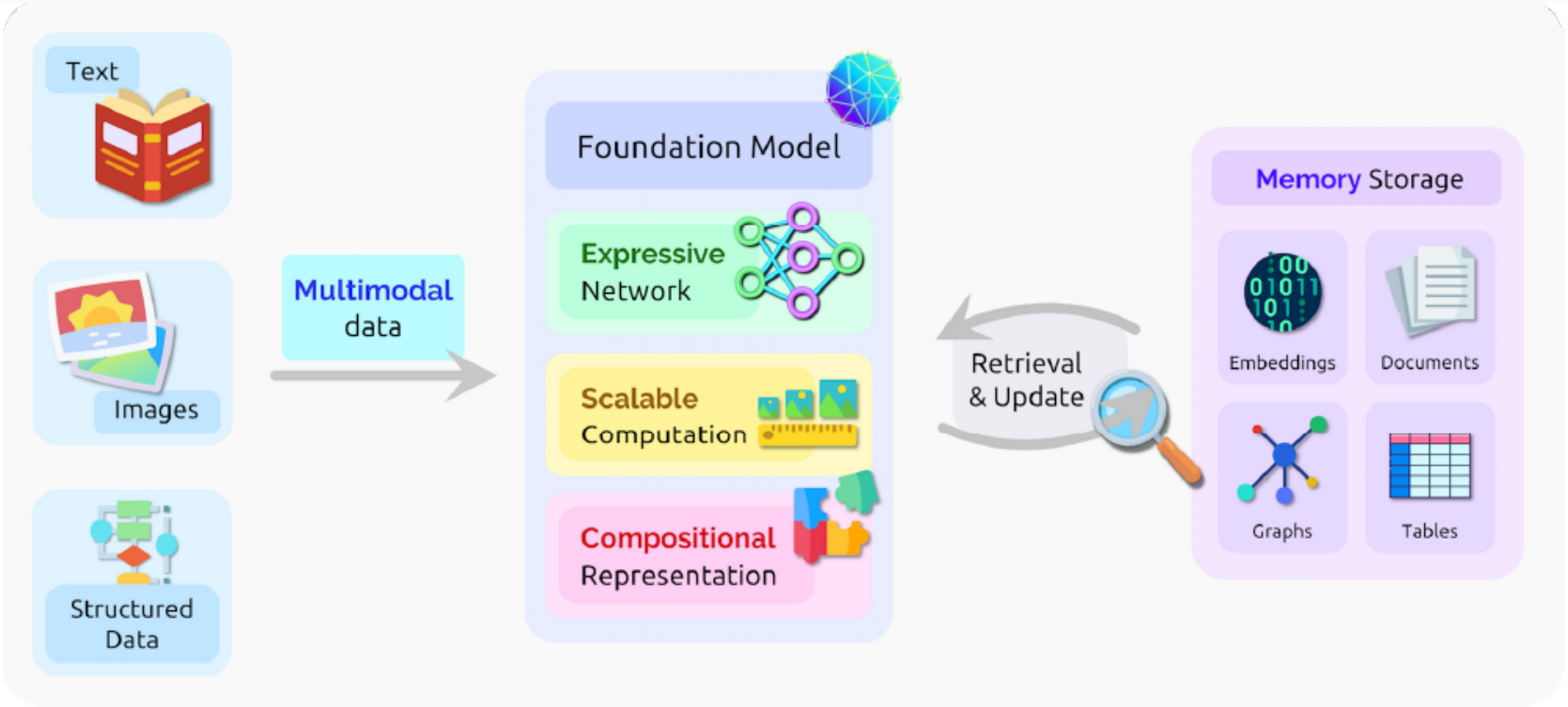


Fig. 17. The five key properties of a foundation model: *expressivity* — to flexibly capture and represent rich information; *scalability* — to efficiently consume large quantities of data; *multimodality* — to connect together various modalities and domains; *memory capacity* — to store the vast amount of accumulated knowledge; and *compositionality* — to generalize to new contexts, tasks and environments.

Bommasani et al. 2021

Scaling

- Expressivity
- **Scalability**
- Multimodality
- Memory
- Compositionality

Bommasani et al. 2021

Scaling

"For foundation models to effectively fit the complex and high-dimensional distribution of images or text, they should thereby be *scalable* across all dimensions: including both models' depth and width as well as their training time, number of parameters, and the amount of data they could process."

Bommasani et al. 2021

Neural scaling laws

Resources

- N. Number of model parameters (excluding all vocabulary and positional embeddings)
- D. Dataset size in terms of number of tokens.
- C. Amount of compute in training (PF days)

Also looks at "critical batch size" B_{crit}

Neural scaling laws

Proposed model

$$L(X) \propto \left(\frac{1}{X}\right)^{\alpha_X}$$

where X is C, D or N; α_X is its power-law exponent; L is cross entropy loss.

Experimentally solve for X_C and α_X

$$L(X) = (X_C/X)^{\alpha_X}$$

where X_C and α_X are constants.

Kaplan et al. 2020

Neural scaling laws

Power Law	Scale (tokenization-dependent)
$\alpha_N = 0.076$	$N_c = 8.8 \times 10^{13}$ params (non-embed)
$\alpha_D = 0.095$	$D_c = 5.4 \times 10^{13}$ tokens
$\alpha_C = 0.057$	$C_c = 1.6 \times 10^{7}$ PF-days
$\alpha_C^{\min} = 0.050$	$C_c^{\min} = 3.1 \times 10^{8}$ PF-days
$\alpha_B = 0.21$	$B_* = 2.1 \times 10^{8}$ tokens
$\alpha_S = 0.76$	$S_c = 2.1 \times 10^{3}$ steps

Table 5

No fundamental interpretation because they change with a change in vocabulary/ language.

Kaplan et al. 2020

Neural scaling laws

Some findings

- "Smooth power laws"
- "Performance depends strongly on scale, weakly on model shape"
- "Large models are more sample-efficient than small models"
- "Transfer improves with test performance"

Smooth power laws

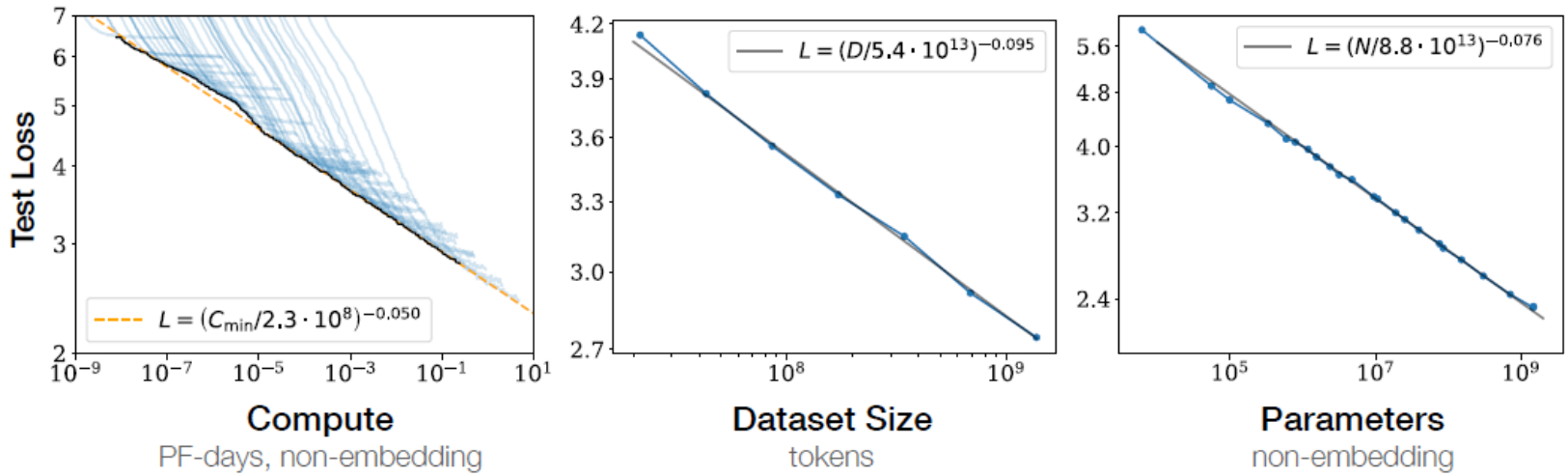


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan et al. 2020

Performance depends weakly on model shape

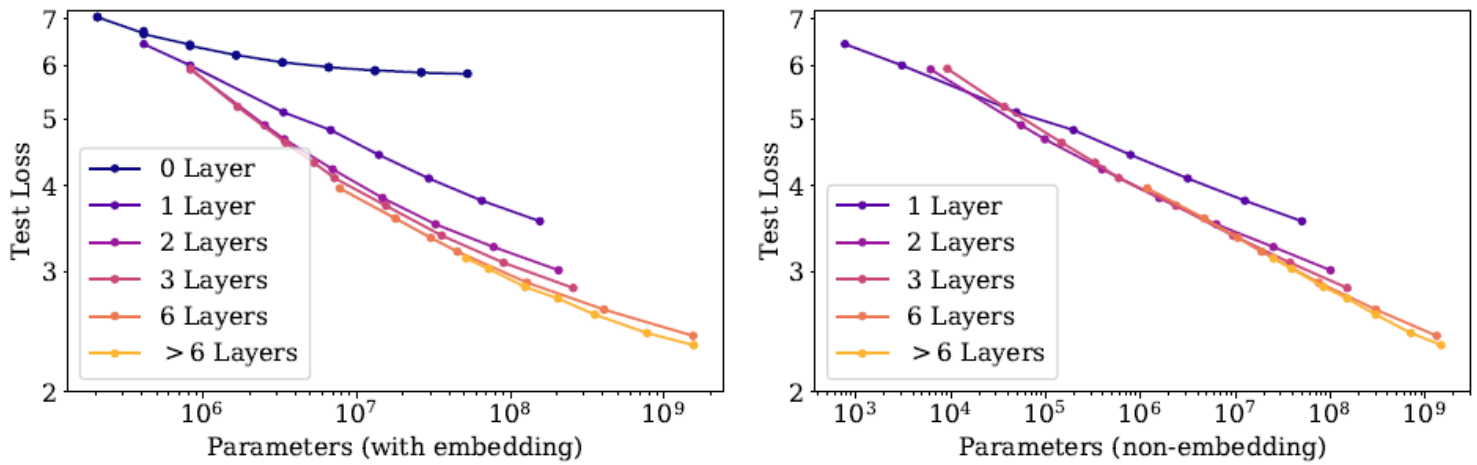


Figure 6 **Left:** When we include embedding parameters, performance appears to depend strongly on the number of layers in addition to the number of parameters. **Right:** When we exclude embedding parameters, the performance of models with different depths converge to a single trend. Only models with fewer than 2 layers or with extreme depth-to-width ratios deviate significantly from the trend.

Kaplan et al. 2020

Large models are more sample-efficient

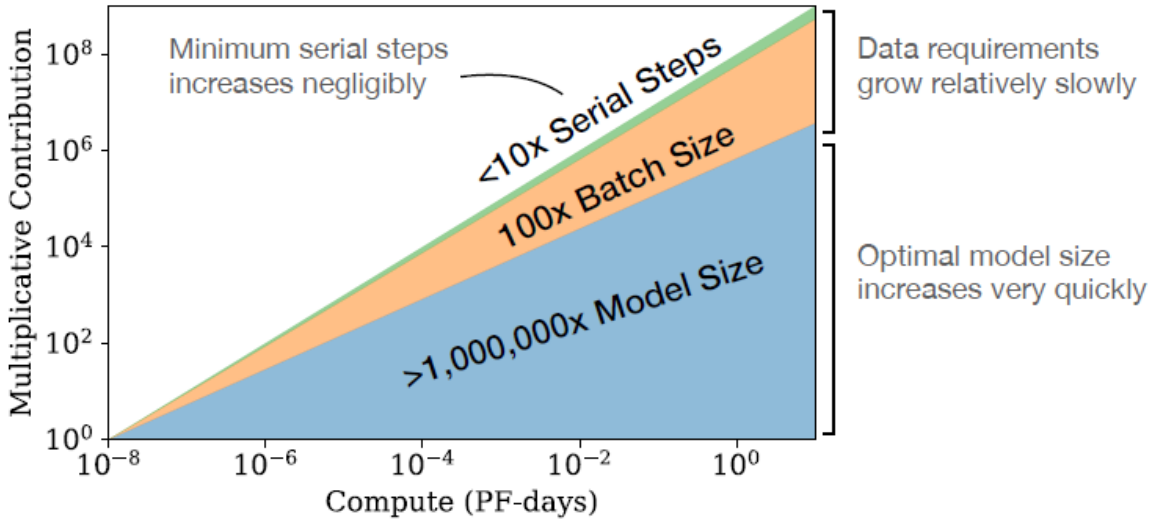
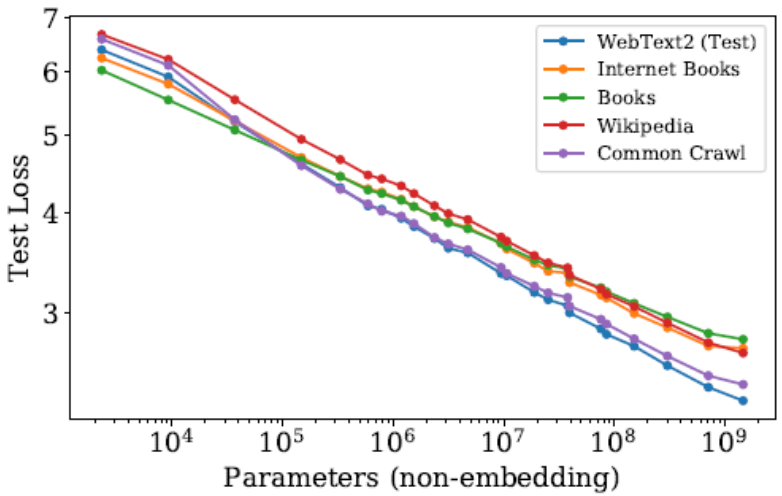


Figure 3 As more compute becomes available, we can choose how much to allocate towards training larger models, using larger batches, and training for more steps. We illustrate this for a billion-fold increase in compute. For optimally compute-efficient training, most of the increase should go towards increased model size. A relatively small increase in data is needed to avoid reuse. Of the increase in data, most can be used to increase parallelism through larger batch sizes, with only a very small increase in serial training time required.

Kaplan et al. 2020

Transfer improves with test performance



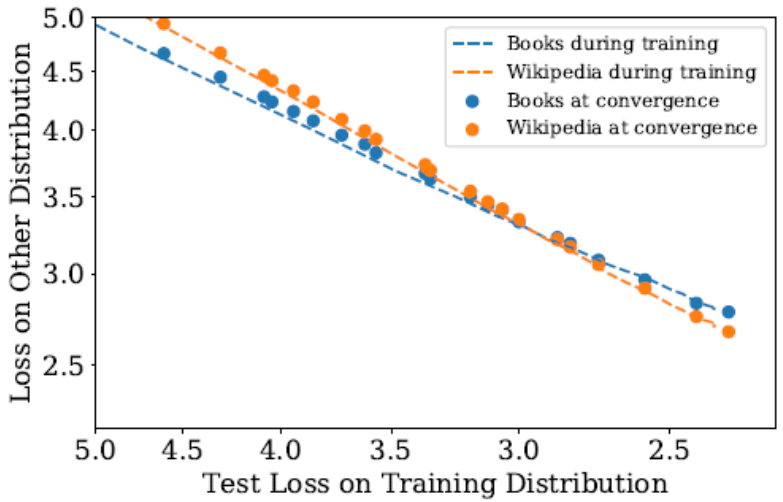


Figure 8 **Left:** Generalization performance to other data distributions improves smoothly with model size, with only a small and very slowly growing offset from the WebText2 training distribution. **Right:** Generalization performance depends only on training distribution performance, and not on the phase of training. We compare generalization of converged models (points) to that of a single large model (dashed curves) as it trains.

Kaplan et al. 2020

Neural scaling laws

"We observe no signs of deviation from straight power-law trends at large values of compute, data, or model size. Our trends must eventually level off, though, since natural language has non–zero entropy."

Kaplan et al. 2020

Exploring the limits of large scale pre-training

"Recent impressive progress on transfer and few-shot learning suggests an emerging direction that scaling up models and training them on a huge corpus of data is the main obstacle towards better performance on downstream tasks with less or no data."

- vision transformers [Zhai et al. 2021]
- transfer learning [Hernandez et al. 2021]

Abnar et al. 2021

Exploring the limits of large scale pre-training

- 4800 (imported) experiments with different configurations
- controlled experiments: increase data size, model size and training time to explore DS-vs-US accuracy
- most experiments not trained for the purpose of this paper
- aggregate different vision transformer, MLP mixer and ResNet models from different researchers in a meta-study
- focuses on downstream vs upstream accuracy instead of directly on the impact of scaling

---

Abnar et al. 2021

Exploring the limits of large scale pre-training

Proposed model inspired by Kaplan et al.

$$e_{DS} = k(e_{US})^{\alpha} + e_{IR}$$

where e_{DS}, e_{US}, e_{IR} refers to DS, US and irreducible errors respectively. k and α are constants.

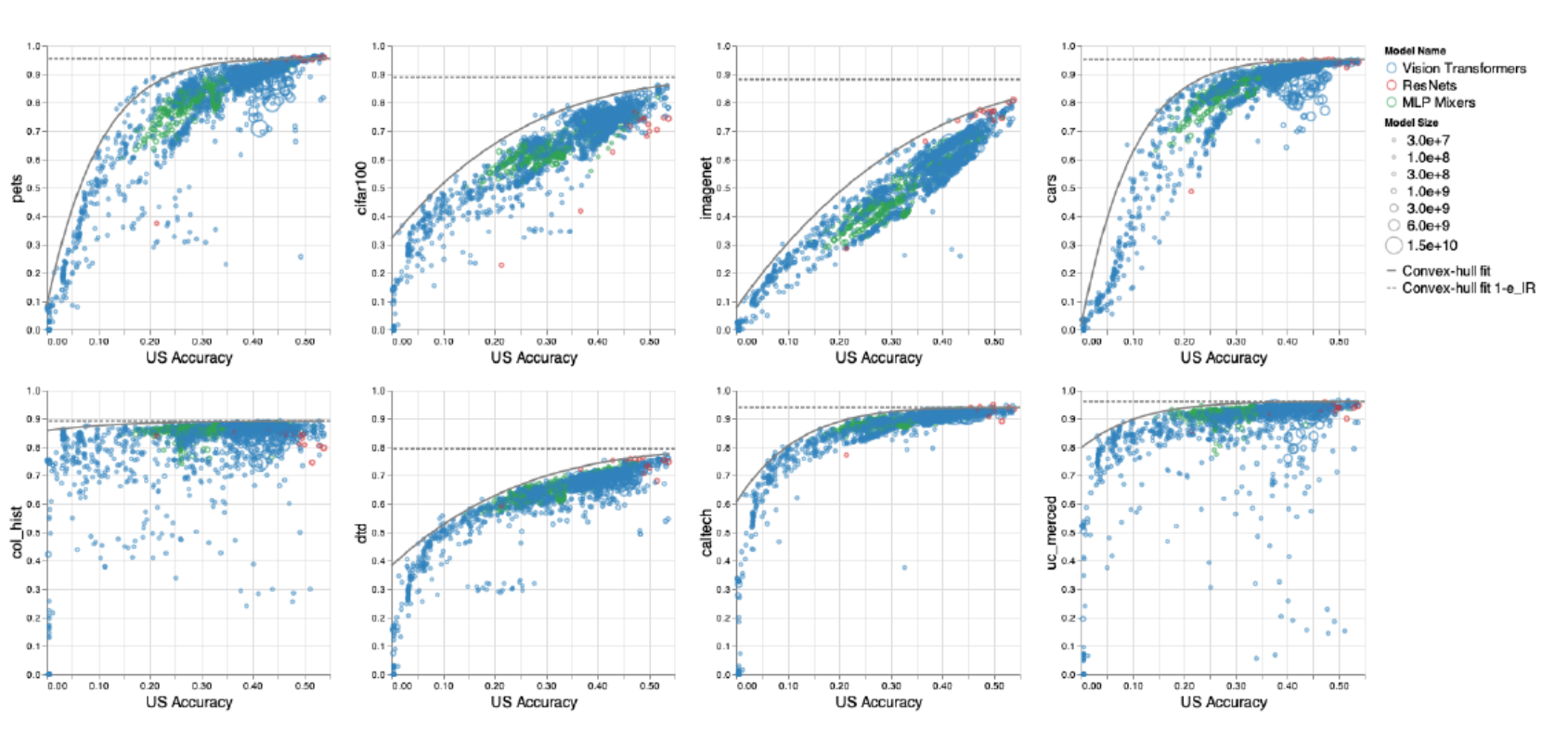
Plotting in log scaling gives a straight line only when e_{IR} is zero.

Error is 1 - accuracy.

Exploring the limits of large scale pre-training

THE UNIVERSITY OF BRITISH COLUMBIA

1500 vision transformers, 1400 MLP-mixers, 16 ResNets

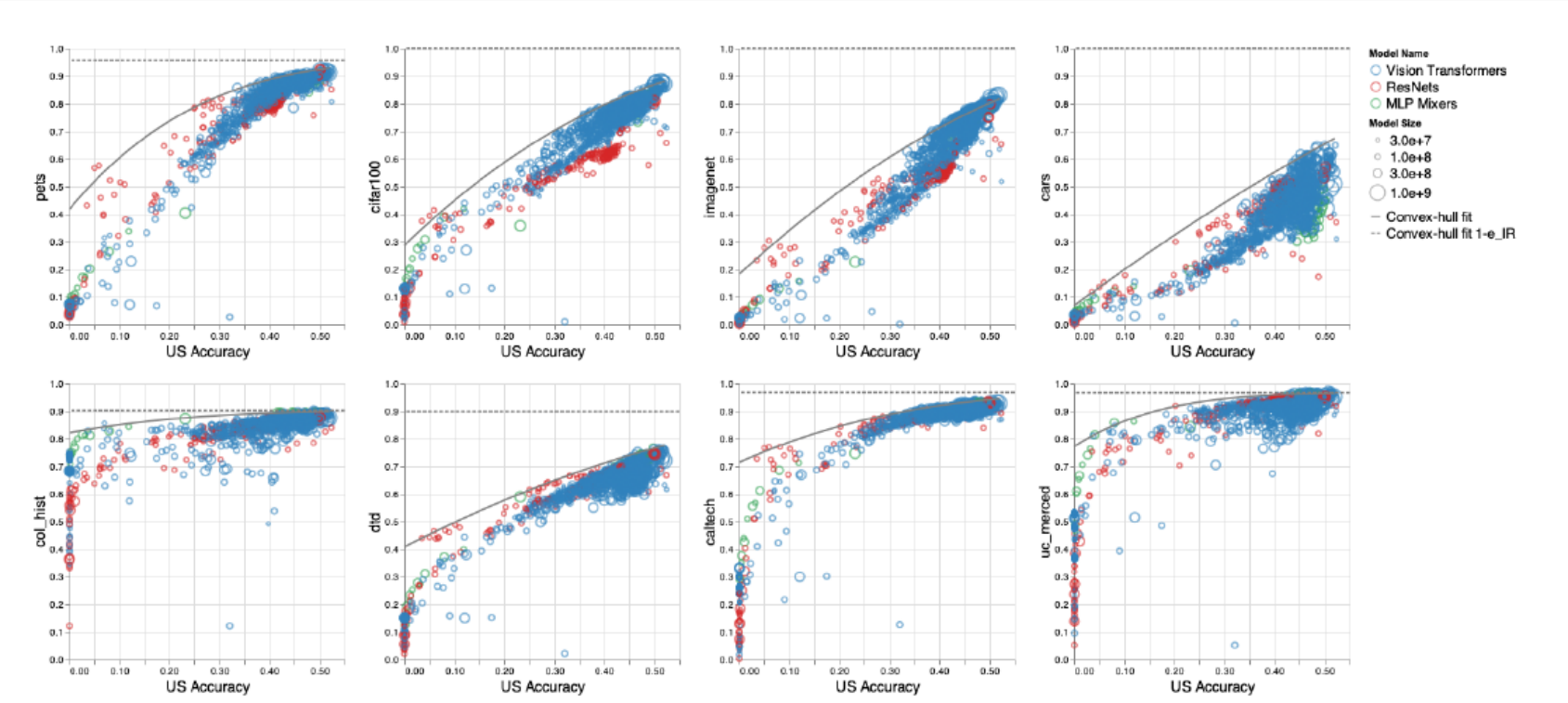


Abnar et al. 2021

Exploring the limits of large scale pre-training

THE UNIVERSITY OF BRITISH COLUMBIA

1400 vision transformers, 90 MLP mixers, 233 ResNets

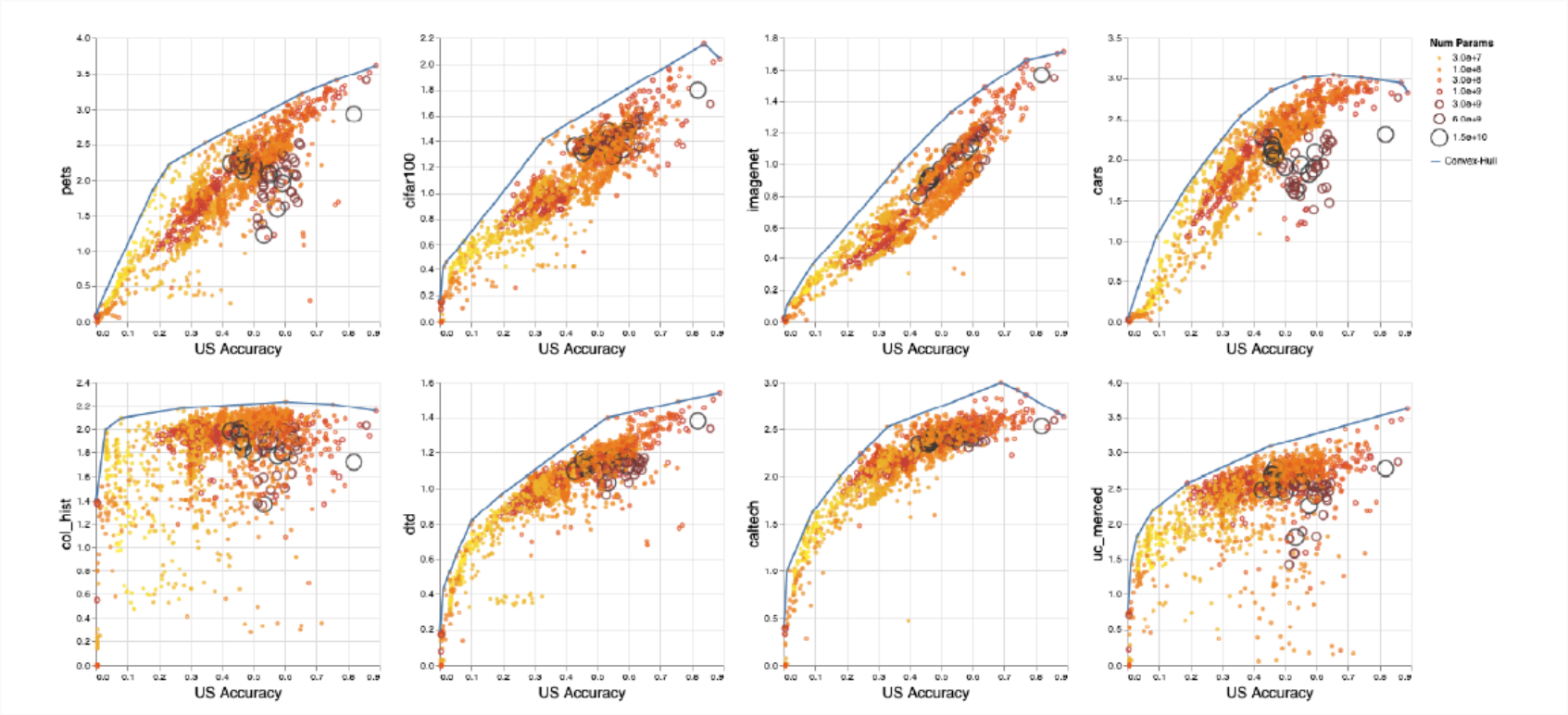


Abnar et al. 2021

Exploring the limits of large scale pre-training

THE UNIVERSITY OF BRITISH COLUMBIA

More than 3000 vision transformers, logit scaling for downstream accuracy



Abnar et al. 2021

Exploring the limits of large scale pre-training

Saturation point

- "the value of downstream accuracy as upstream accuracy reaches 1.0"
- where "it is not worth scaling up data size, compute or model size to improve US accuracy as the effect on DS accuracy is negligible"
- increasing pre-training effort may lead to downstream performance reaching a saturation point below its Bayes error

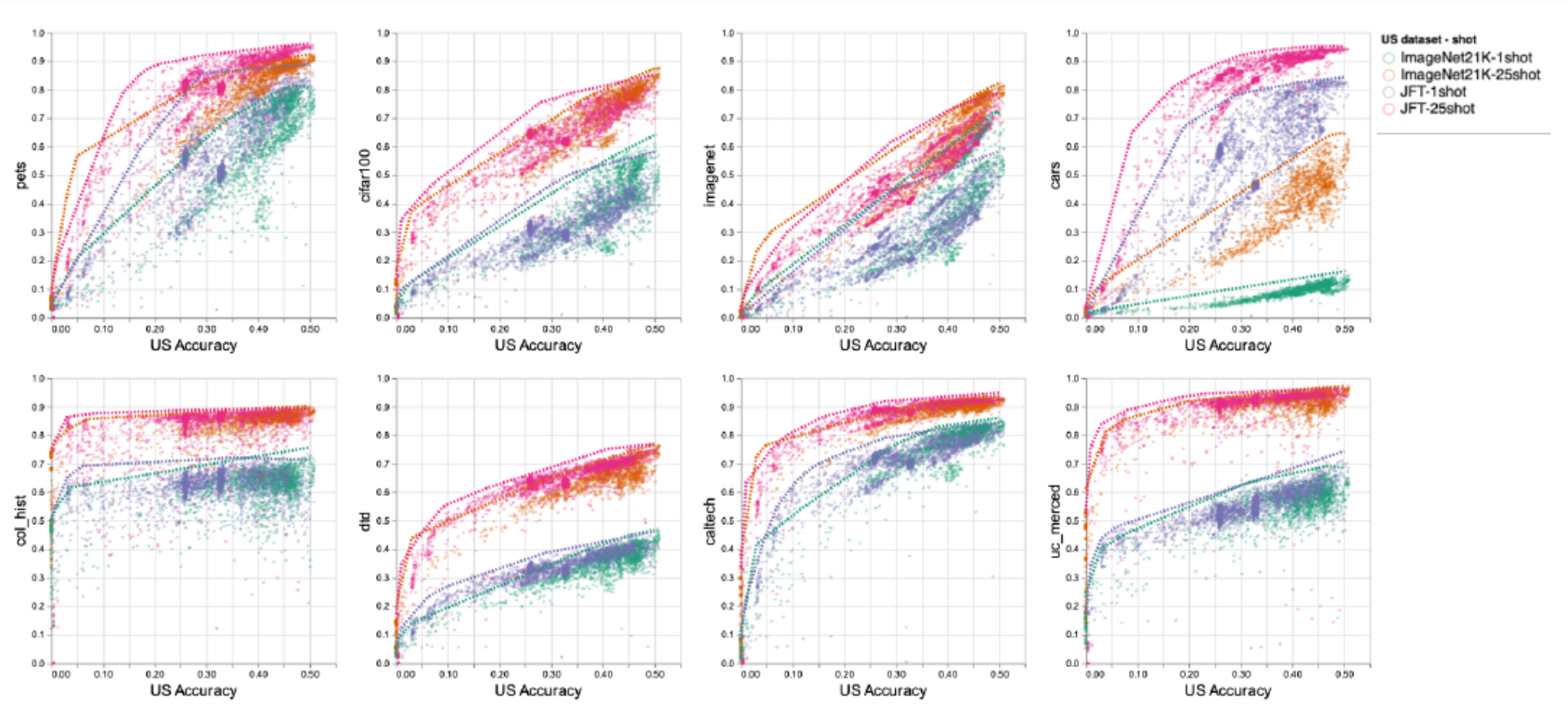
---

Abnar et al. 2021

Exploring the limits of large scale pre-training

THE UNIVERSITY OF BRITISH COLUMBIA

"[Saturation] is about the relationship between the US and DS tasks"



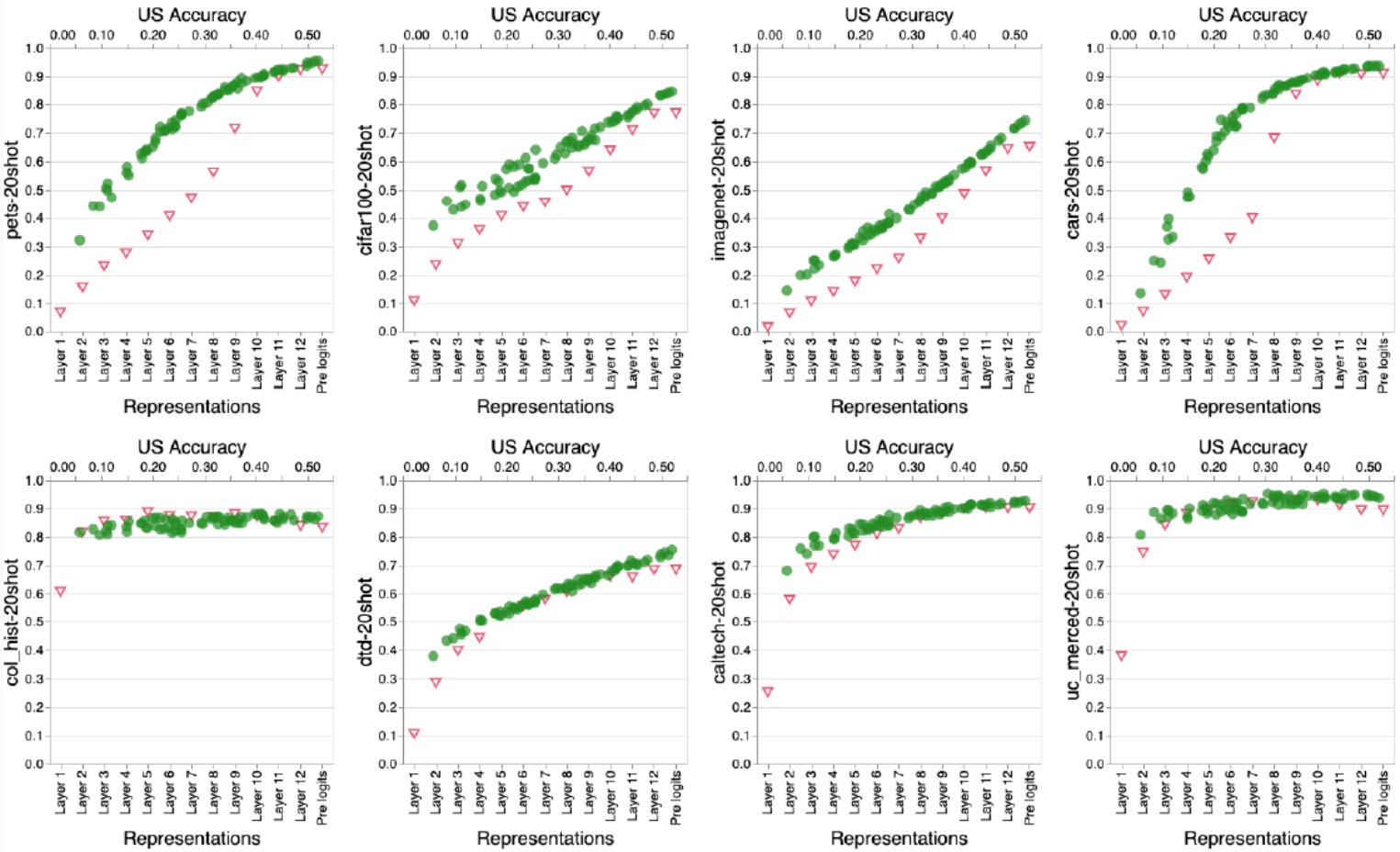
Abnar et al. 2021

Exploring the limits of large scale pre-training

- "optimal layer is not the last one"
- "pre-trained network lacks the fine-grained features required to perform well on DS"

Abnar et al. 2021

Exploring the limits of large scale pre-training



Abnar et al. 2021

Exploring the limits of large scale pre-training

"contrary to the common narrative, scaling does not lead to a one-model-fits-all solution"

"when investing in scaling in terms of data, model parameters and compute, we should think of an additional axis which is *data diversity*."

Abnar et al. 2021

References

- Samira Abnar, Mostafa Dehghani, Behnam Neyshabur and Hanie Sedghi. 2021 *Exploring the Limits of Large Scale Pre-training*
- Bommasani et al. 2021. *On the Opportunities and Risks of Foundation Models*
- Hernandez et al. 2021. *Scaling Laws for Transfer*
- Kaplan et al. 2020. *Scaling Laws for Neural Language Models*
- Zhai et al. 2021. *Scaling Vision Transformers*