Deep Learning as a New Statistical Paradigm

Liam Hodgkinson

Workshop on Modern Statistical Perspectives in Machine Learning

What is a Neural Network?

A **neural network** is a map that is a composition of

- linear maps; and
- scalar-wise nonlinearities.

What is a Neural Network?

A **neural network** is a map that is a composition of

- linear maps; and
- scalar-wise nonlinearities.

...but not always. *You can include fixed nonlinear transformations too*.

A neural network is a big nonlinear, parameterised model

(and weakly differentiable, trained with stochastic optimisers)

Large Models are Not Scary

Gaussian processes and **kernel ridge regression** are infinite-dimensional non-parametric approaches and *perform very well*.

$$\hat{f} = \underset{f \in H}{\arg\min} \; \frac{1}{2} \| f(\boldsymbol{x}) - \boldsymbol{y} \|^2 + \lambda \| f \|_H^2$$

Problems:

1. Results are often sensitive to the choice of kernel (space H)
2. Inference is expensive: scales **quadratically** in size of dataset

Why Do Neural Networks Perform So Well?

~~Why Do Neural Networks Perform So Well?~~

When neural networks do perform well, it is because of a strong form of **(implicit) regularisation** arising from a combination of the model structure and the training procedure.

Is it **sparsity**? Is it **shrinkage**? *It is **neither**.*

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107-115.

Regimes of Learning

Yang, Yaoqing, et al. "Taxonomizing local versus global structure in neural network loss landscapes." *Advances in Neural Information Processing Systems* 34 (2021): 18722-18733.

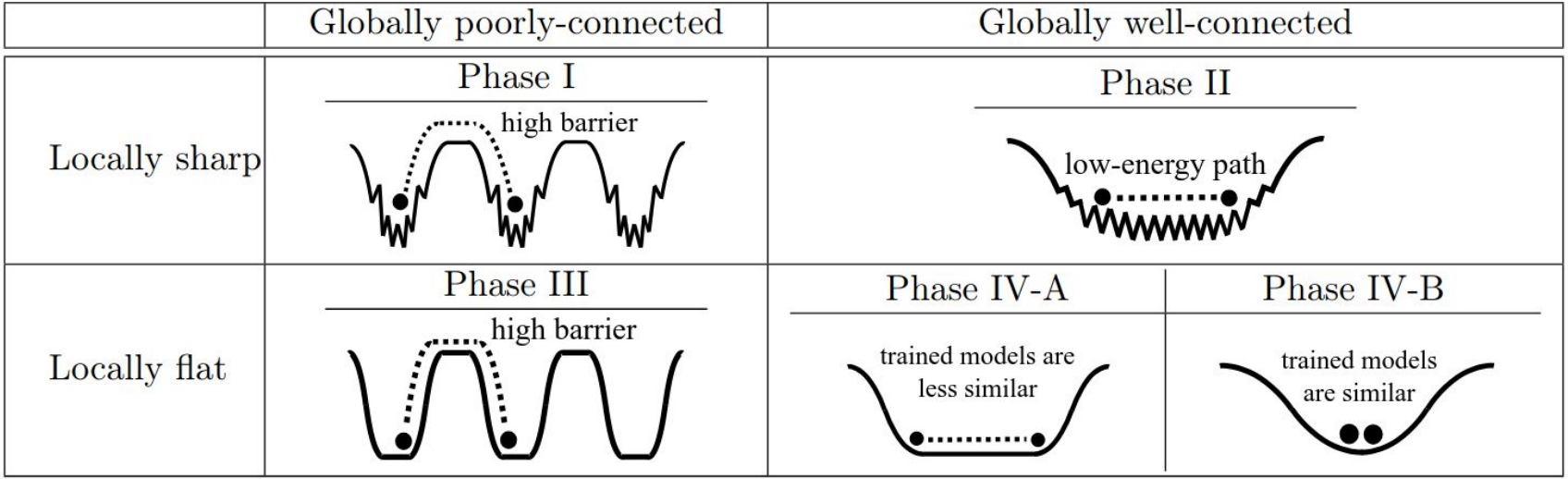


Figure 1: **(Caricature of different types of loss landscapes).** Globally well-connected versus globally poorly-connected loss landscapes; and locally sharp versus locally flat loss landscapes. Globally well-connected loss landscapes can be interpreted in terms of a global "rugged convexity"; and globally well-connected and locally flat loss landscapes can be further divided into two sub-cases, based on the similarity of trained models.

Jackpot?



Classical Paradigm:

For given **data and problem**, find the appropriate **model** (usually **small**).

New Paradigm:

For given **model** (usually **big**), use the appropriate **data and problem**.

Neural Scaling Laws (Kaplan et al., 2020; Hoffman et al., 2023)

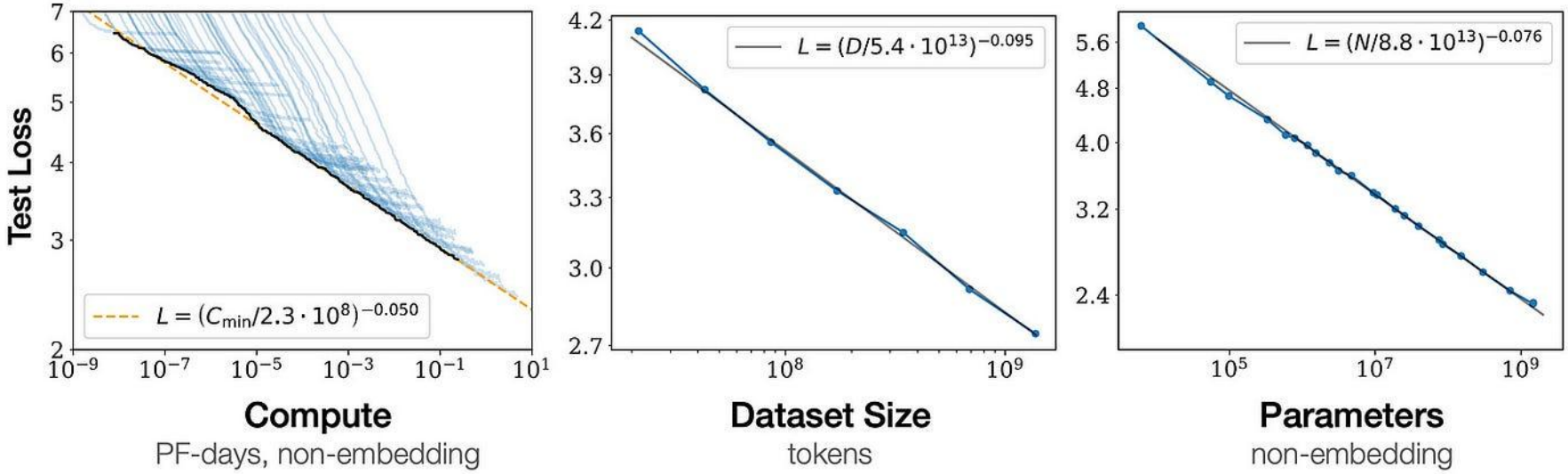


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when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

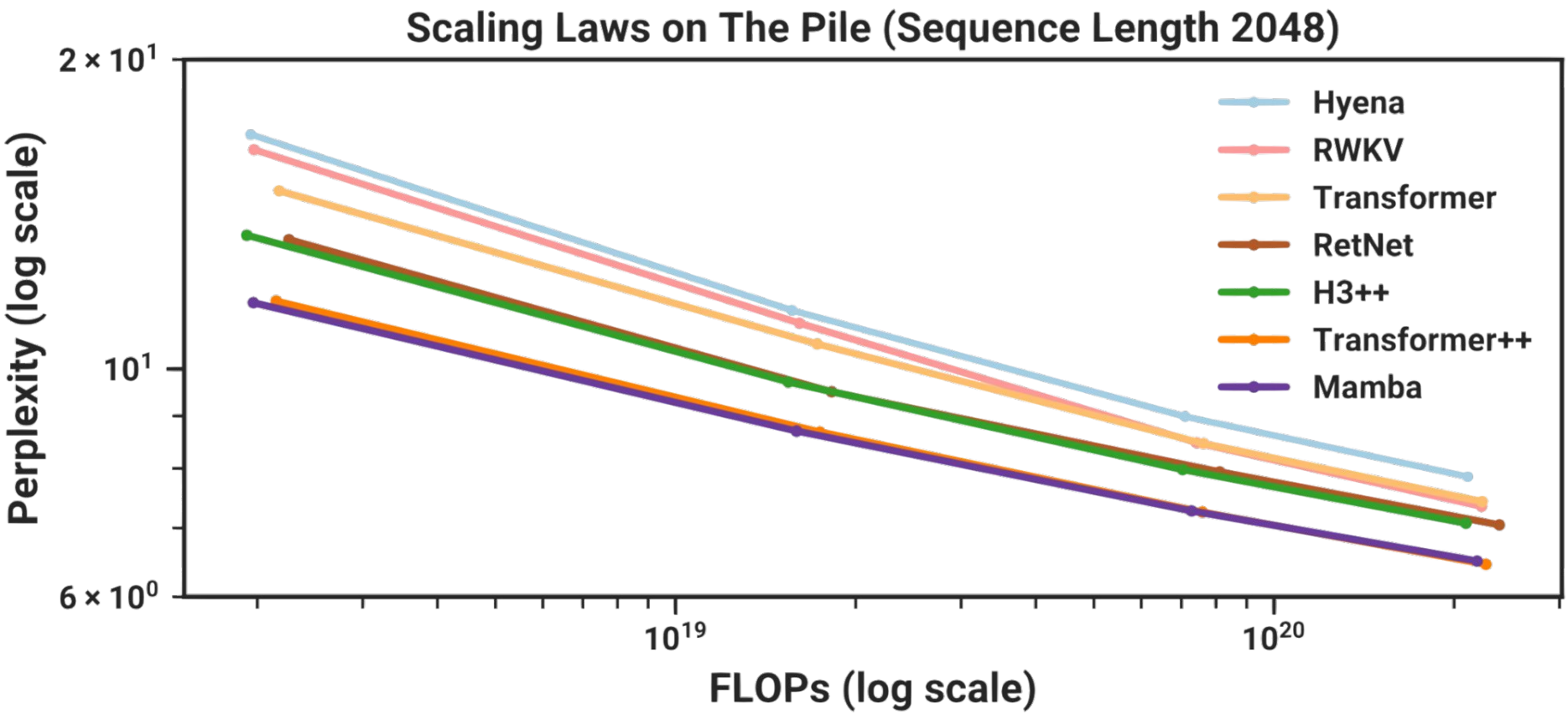
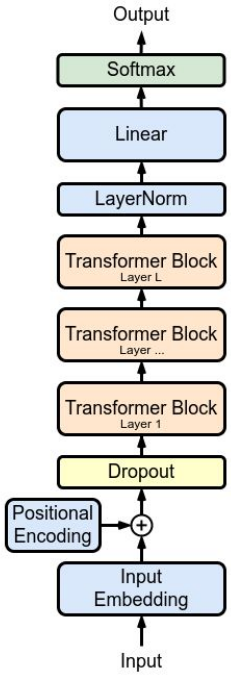
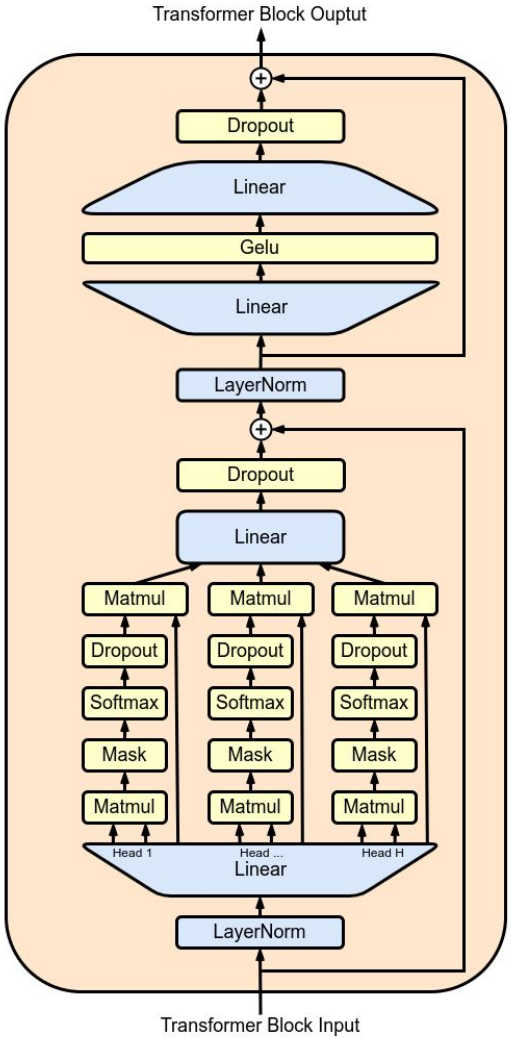


Figure (Gu & Dao, 2024): At large scale, most model classes scale similarly

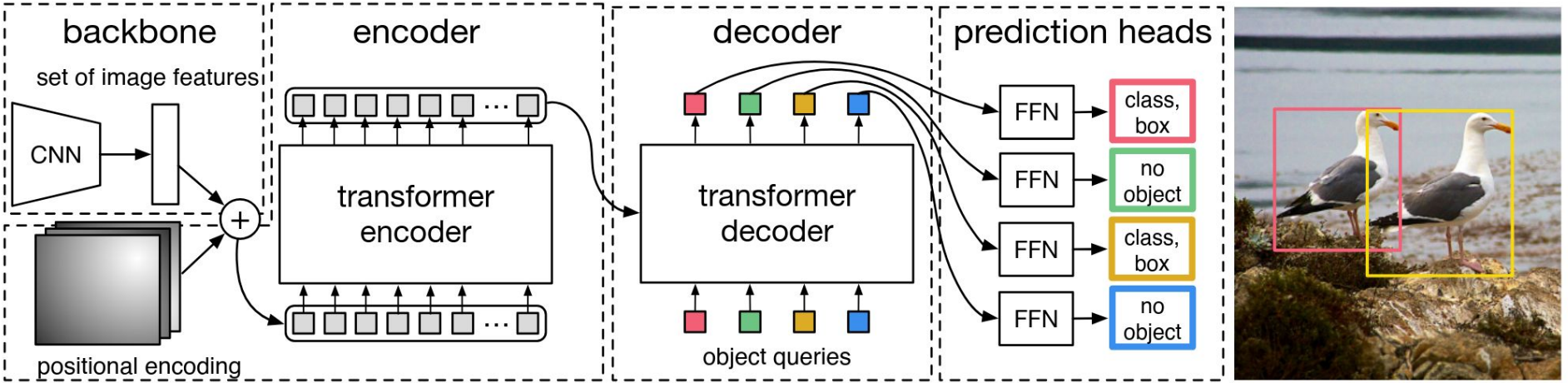
At the state-of-the-art, fine-grained alterations to the model design are rarely worth the effort.



GPT Architecture



Deep learning models are designed in a **modular** way: *stacking existing models on top of each other*



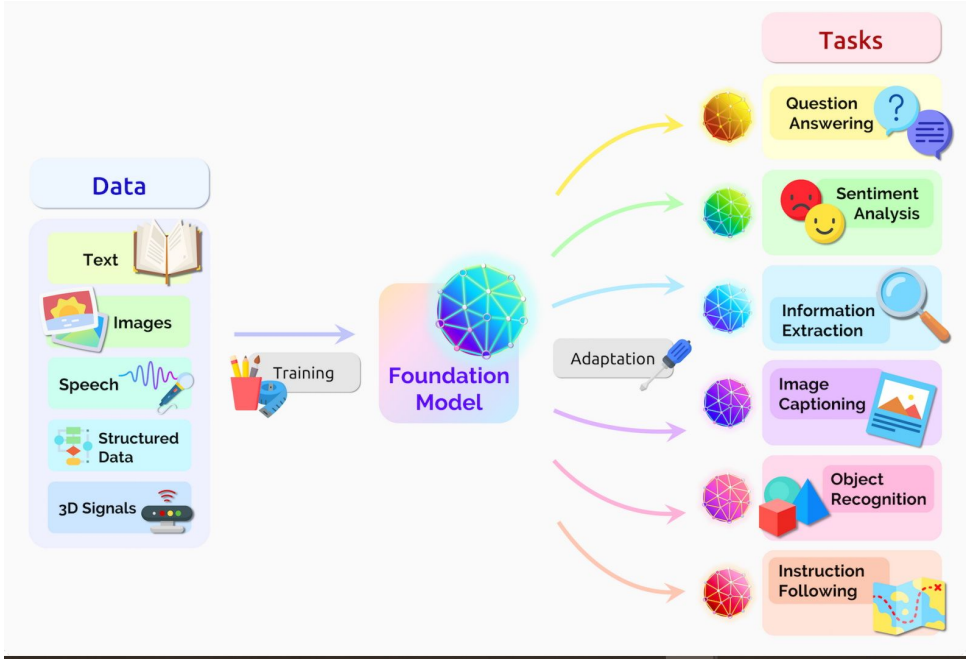
Detection Transformer (Carion et al., 2020)

...regardless of the task.

Foundation Models

Trained on **enormous** corpus of data; fine-tuned for a particular task.

General: Multimodal LLMs

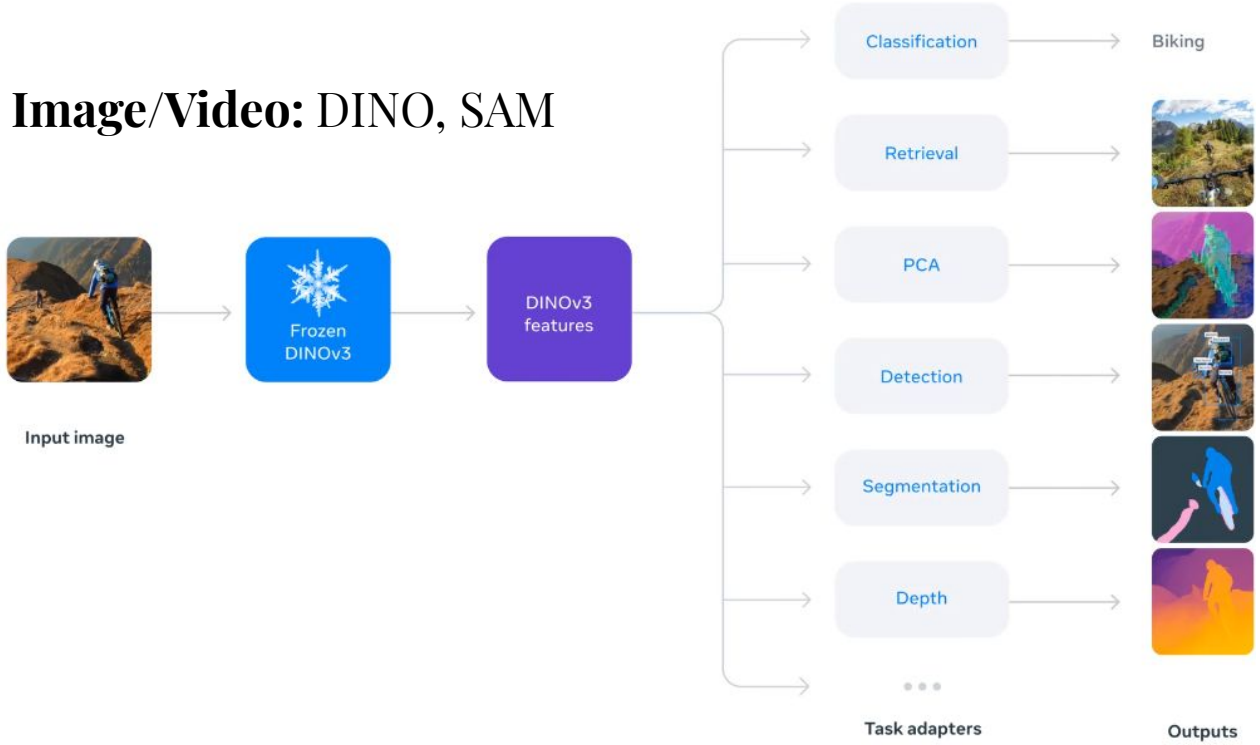


(NVIDIA, 2025)

Foundation Models

Trained on **enormous** corpus of data; fine-tuned for a particular task.

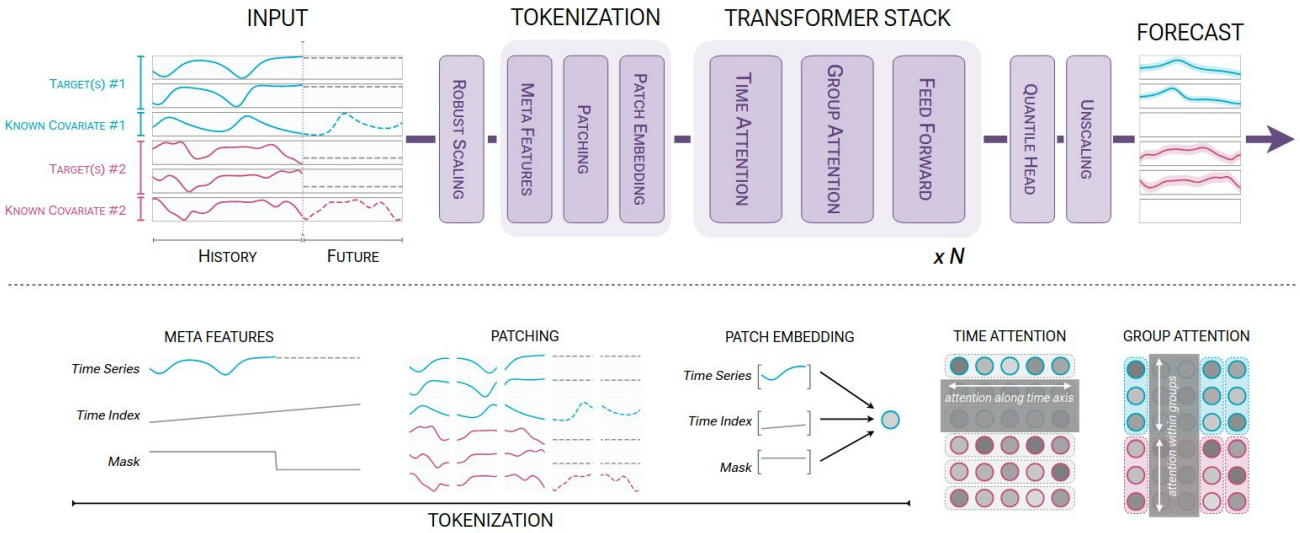
Image/Video: DINO, SAM



Foundation Models

Trained on **enormous** corpus of data; fine-tuned for a particular task.

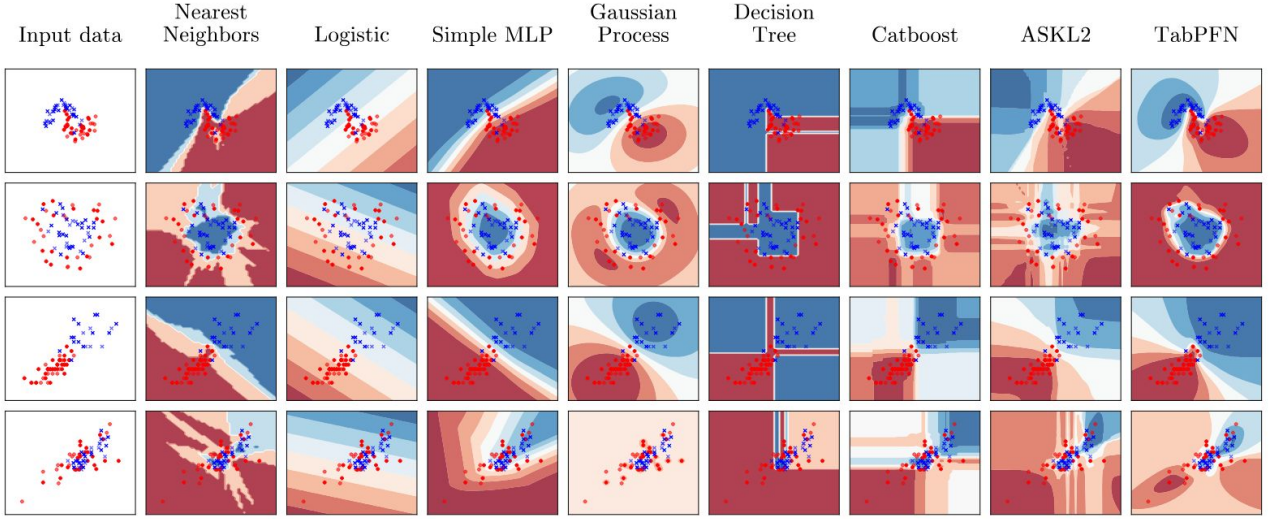
Time Series: Chronos, TimesFM



Foundation Models

Trained on **enormous** corpus of data; fine-tuned for a particular task.

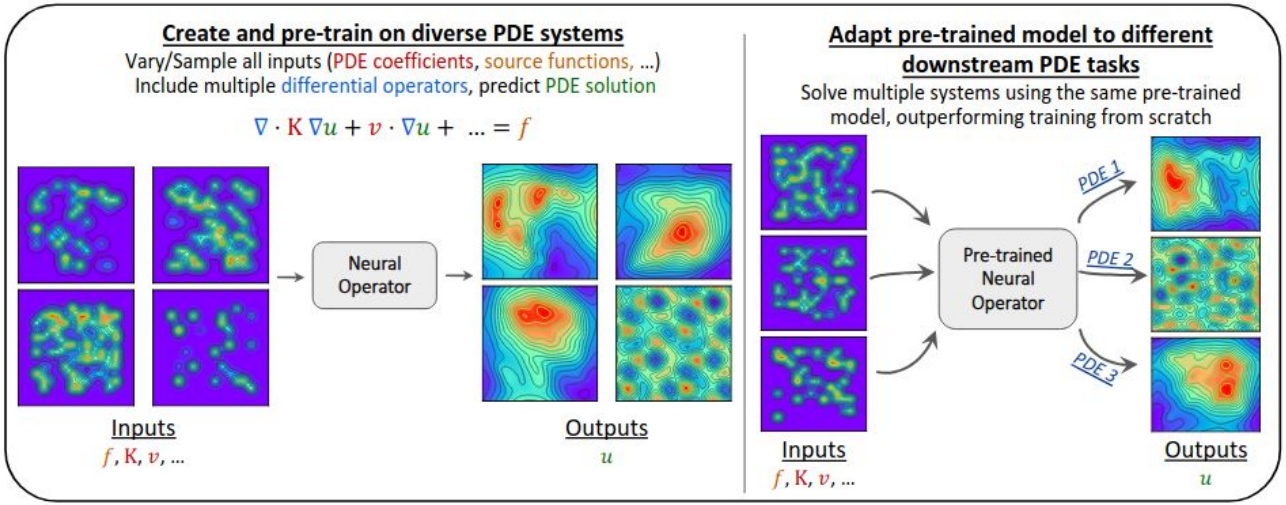
Tabular Data: TabPFN



Foundation Models

Trained on **enormous** corpus of data; fine-tuned for a particular task.

SciML (PDEs): Poseidon, PROSE-PDE



How is this possible?

Solomonoff Induction

Theorem (informal): Under common sense axioms, given an *exact* dataset, the best predictor is the one that explains the data with the smallest Kolmogorov complexity (*minimum description length*).

- **Extends to noisy data** through Bayesian formulation: $\text{prior}(p) \propto 2^{-\text{Kol}(p)}$
- **AIXI:** The optimal AI best approximates the process of Solomonoff induction
- **Hutter et al.:** Modern transformers mimic theoretical AIXI.
- Neural networks perform well because they are strong **compressors**.
- In parameter space, this often infers more regular models.
- At scale, **all models seem to approach the optimal predictor**.

Hutter, M., Quarel, D., & Catt, E. (2024). *An introduction to universal artificial intelligence*. Chapman and Hall/CRC.

Where does classical statistical theory go wrong?

Part I: Deep Learning is a Procedure

Neural networks are not simply a class of functions. They include:

1. **Initialisation Scheme** (LeCun, He)
2. **Optimiser** (SGD, Adam)
3. **Model Architecture**
4. **Loss Function / Objective / Likelihood**
5. **Data Augmentation** (Random Transformations, MixUp, CutMix)
6. **Training Strategy** (Hyperparameters, Batch Size)
7. **Explicit Regularisation** (L^p-norm regularisation)

The interplay between these factors is critical for a successful model.

The Initialisation is Extremely Important!

Machine learning packages (e.g. PyTorch) automatically initialise networks according to a standard scheme, e.g. **He initialisation**:

$$W_{ij} \sim \mathcal{N}\left(0, \frac{2}{n_{\text{in}}}\right)$$

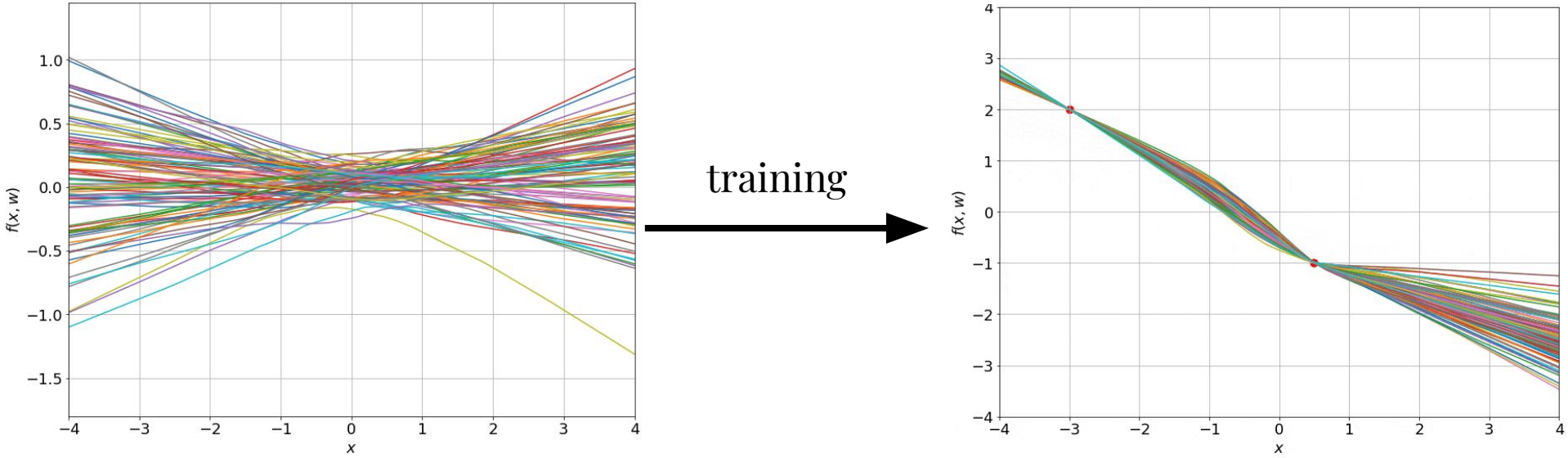
Poor choices of initialisation **will often induce a poor model**.

Hu, W., Xiao, L., & Pennington, J. (2020). *Provable benefit of orthogonal initialization in optimizing deep linear networks*. ICLR 2020.

https://www.eigentales.com/NTK/

The Large Width Limit

This particular initialisation scheme ensures that when the widths of the network are large, the neural network converges in distribution to a **Gaussian process**; regression becomes **Gaussian process regression**.



Rich vs. Lazy Training

It is tempting to construct a theoretical framework around this concept, **but we should not**.

Gaussian process regression operates on fixed basis elements via Mercer's Theorem. There is no **feature learning**; this is called *lazy training*.

Instead, real neural networks **learn an underlying basis**; this cannot be explained in the framework of Gaussian process regression.

Chizat, L., Oyallon, E., & Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, *32*.

Part II: Don't Think in Worst-Case Error

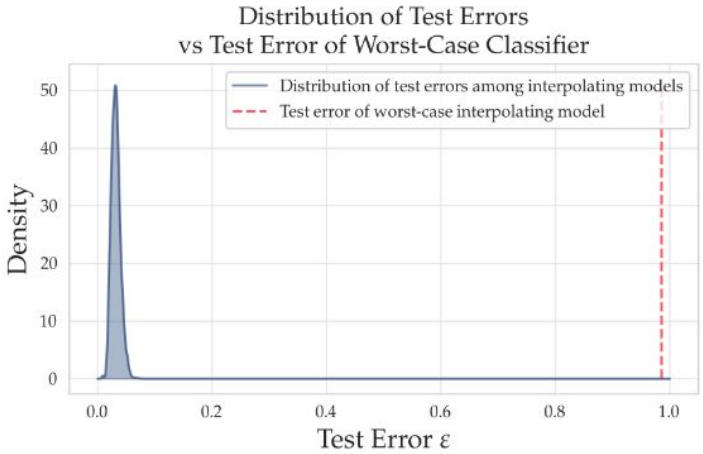


Figure 1: Test error distribution of MNIST 0 vs 1 interpolating classifiers, using $N = 1000$ random ReLU features, with $n = 500$ training samples, as well as test error of worst-case interpolating classifier. Here, for illustrative purposes, we plot the PDF (fit from a histogram using a kernel density estimate); in the remainder of the paper, we instead plot the CDFs, which can be more accurately estimated.

The PAC framework is disastrous for deep learning, as the bounds are **vacuous**.

$$\mathbb{E} \sup_{f \in \mathcal{F}} L(f) \leq \cdots$$

Good models occur in the **average-case**, not in the worst-case.

Theisen, R., Klusowski, J., & Mahoney, M. (2021, March). Good classifiers are abundant in the interpolating regime. In *International Conference on Artificial Intelligence and Statistics* (pp. 3376-3384). PMLR.

The PAC-Bayes Framework (Alquier, 2021)

Beginning with the Donsker–Varadhan inequality:

$$\mathbb{E}_{f\sim\rho} f(X) \leq \mathrm{KL}(\rho\|\pi) + \log \mathbb{E}_{f\sim\pi} e^{f(X)}$$

A single application of Markov's inequality shows that with probability $> 1-\delta$

$$\underbrace{\mathop{\mathbb{E}}_{f\sim\rho} \mathrm{acc}(f)}_{\text{test}} \leq \underbrace{\mathop{\mathbb{E}}_{f\sim\rho} \widehat{\mathrm{acc}}_n(f)}_{\text{train}} + \frac{1}{\sqrt{n}}\left[\mathrm{KL}(\rho\|\pi) + \log\left(\frac{1}{\delta}\right) + 1\right]$$

The prior π is an **arbitrary** weighting of the hypothesis class.

Non-Vacuous Generalization Bounds

The prior can be found **post-hoc** using an optimisation procedure.

Experiment	T-600	T-1200	T-300^2	T-600^2	T-1200^2	T-600^3	R-600
Train error	0.001	0.002	0.000	0.000	0.000	0.000	0.007
Test error	0.018	0.018	0.015	0.016	0.015	0.013	0.508
SNN train error	0.028	0.027	0.027	0.028	0.029	0.027	0.112
SNN test error	0.034	0.035	0.034	0.033	0.035	0.032	0.503
PAC-Bayes bound	0.161	0.179	0.170	0.186	0.223	0.201	1.352
KL divergence	5144	5977	5791	6534	8558	7861	201131
# parameters	471k	943k	326k	832k	2384k	1193k	472k
VC dimension	26m	56m	26m	66m	187m	121m	26m

Dziugaite, G. K., & Roy, D. M. (2017). Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*.

Lotfi, S., Finzi, M., Kuang, Y., Rudner, T. G., Goldblum, M., & Wilson, A. G. (2023). Non-vacuous generalization bounds for large language models. *arXiv preprint arXiv:2312.17173*.

Bayesian Duality

If ρ is the posterior formed from prior π and negative log-likelihood L_n, then

$$\mathrm{KL}(\rho \| \pi) = -n \mathop{\mathbb{E}}_{f \sim \rho} L_n(f) - \log \mathcal{Z}_n$$

where $\mathcal{Z}_n$ is the marginal likelihood.

Theorem (H. et al, 2023): For every overparameterised model, there exists a corresponding **dual** underparameterised model with the same marginal likelihood. Under mild assumptions, the dual likelihood is log-concave and the dual prior is smooth.

Hodgkinson, L., van der Heide, C., Salomone, R., Roosta, F., & Mahoney, M. W. (2023). The interpolating information criterion for overparameterized models. *arXiv preprint arXiv:2307.07785*.

The Dual Model

The **canonical likelihood** for regression typically comes from a model of the form

$$Y = f_\theta(X) + \varepsilon$$

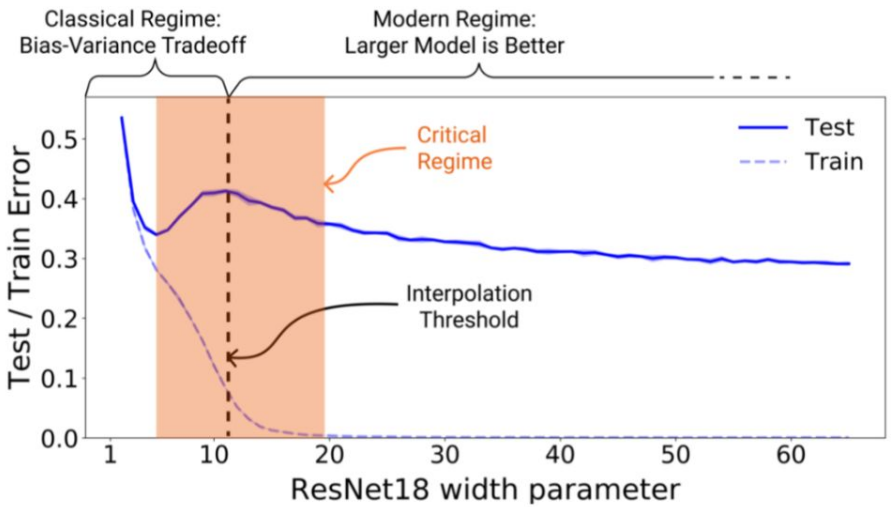
where inference is conducted in θ, given X and Y.

The **dual model** instead considers θ as the data, and inference is conducted in Y given X and θ:

$$\theta = g(X, Y, \delta)$$

The Double Descent Curve

"overfitting never happens"



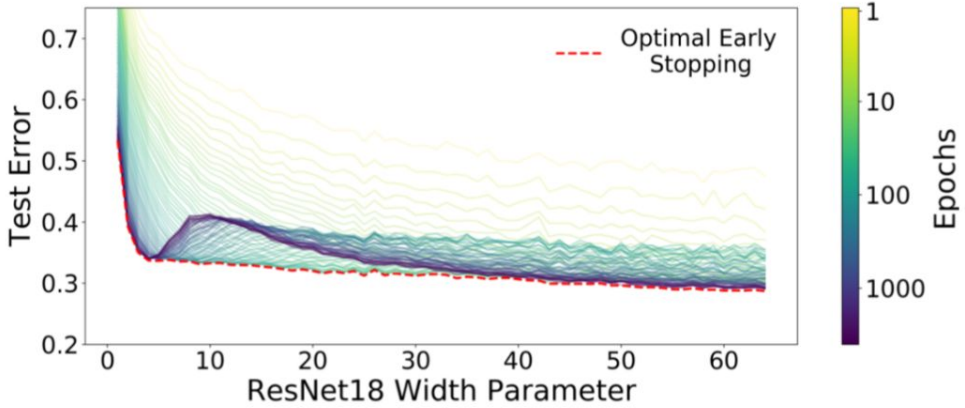


Figure 1: **Left:** Train and test error as a function of model size, for ResNet18s of varying width on CIFAR-10 with 15% label noise. **Right:** Test error, shown for varying train epochs. All models trained using Adam for 4K epochs. The largest model (width 64) corresponds to standard ResNet18.

(Nakkiran et al., 2019)

Part III: Stop Taking $n \to \infty$!

Ubiquitous in statistics is the large data limit, but this regime kills the effect of the prior by the **Bernstein-von Mises Theorem**.

The implicit prior is what makes deep learning work!

The appropriate limit, as per neural scaling laws, is $\text{model size} \propto n$

Proving convergence under this limit is **very difficult**, often requiring random matrix theory or other modern analytic tools.

Conclusion

- Deep learning offers a new landscape for theoretical statistics
  - Models are often very large
  - Powerful (but largely unknown) implicit regularisation; probably compression
  - Overfitting is not a reasonable concern at scale
  - Foundation models remove the need to compare models, just fine-tune!
  - Good performance relies on better data and more effective training strategies
- **PAC-Bayes** is a successful framework for theory in DL
- Avoid:
  - Treating neural networks only as functions
  - Taking the large-data limit
  - Considering worst-case error
- Another approach: benign overfitting

Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063-30070.

The Bitter Lesson by Rich Sutton

"We have to learn the bitter lesson that building in how we think we think does not work in the long run. The bitter lesson is based on the historical observations that

1) AI researchers have often tried to build knowledge into their agents,

2) this always helps in the short term, and is personally satisfying to the researcher, but

3) in the long run it plateaus and even inhibits further progress, and

4) breakthrough progress eventually arrives by an opposing approach based on scaling computation by search and learning.

The eventual success is tinged with bitterness, and often incompletely digested, because it is success over a favored, human-centric approach."