UNIVERSITY OF CAMBRIDGE

Variational Depth Search in ResNets

NAS Workshop at ICLR 2020

Javier Antorán, James Urquhart Allingham, José Miguel Hernández-Lobato

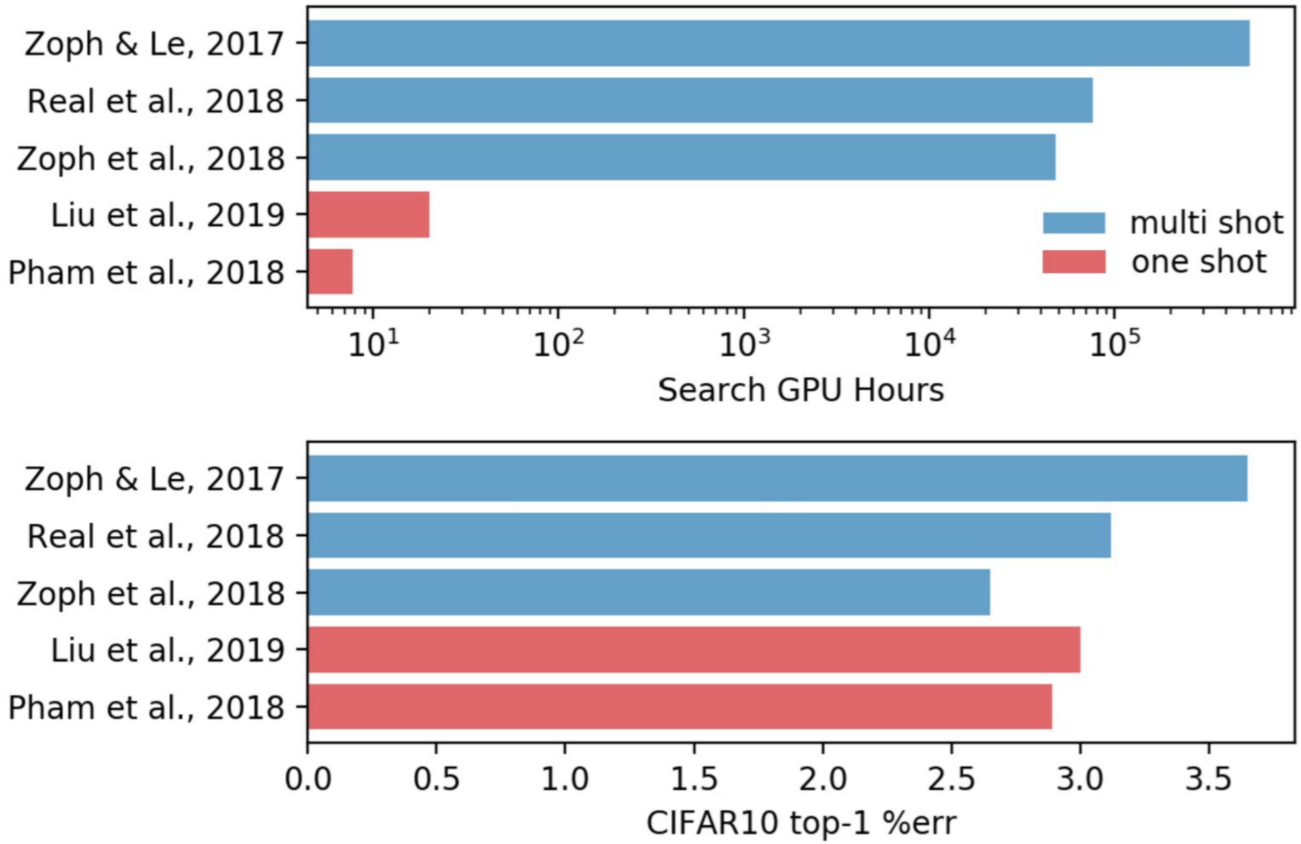
About Us

Javier Antorán
ja666@cam.ac.uk

James Urquhart Allingham
jua23@cam.ac.uk

José Miguel Hernández-Lobato
jmh233@cam.ac.uk

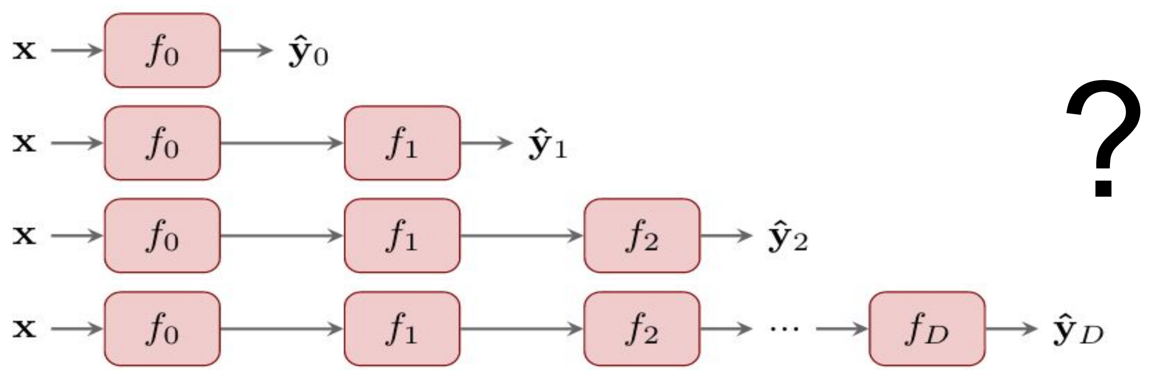
Motivation: Computationally Cheap NAS



Motivation: Computationally Cheap NAS

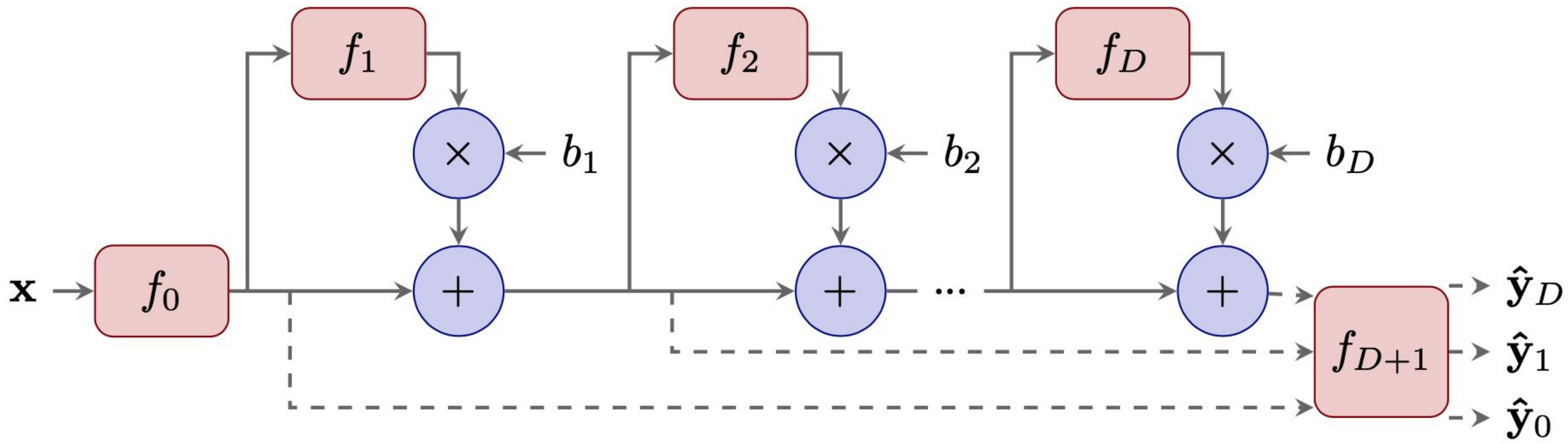
What if we Constrain the Search Space to Depth?

- Deeper is better, but how deep is best?



- If you search over depth, can re-use previous computations!

Can Evaluate All Models with Single Forward Pass

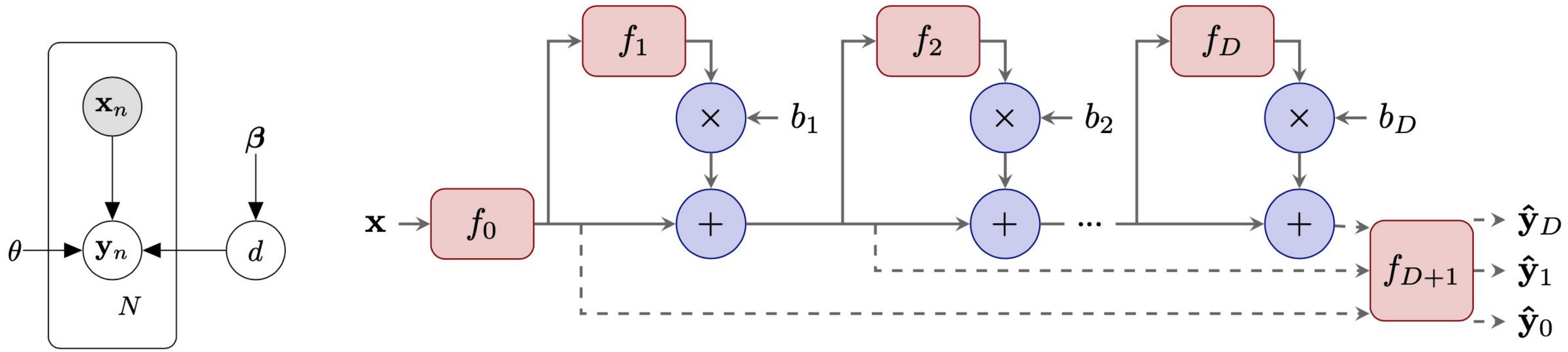


$$\mathbf{a}_i = \mathbf{a}_{i-1} + b_i \cdot f_i(\mathbf{a}_{i-1}) \qquad b_i = 1 \forall i \leq d$$

$$\hat{\mathbf{y}}_i = \mathrm{softmax}(f_{D+1}(\mathbf{a}_i))$$

* ResNets are amenable to removing layers (Huang et al., 2016)

Bayesian Model Averaging, For Free



- We obtain the Likelihood at each Depth with a Single Pass:

$$p_\theta(\mathbf{y}|\mathbf{x}, d)$$

- We define a Categorical Prior over Depth: $p_\beta(d) = Cat(d|\beta)$
- The Depth Posterior is Tractable and Cheap:

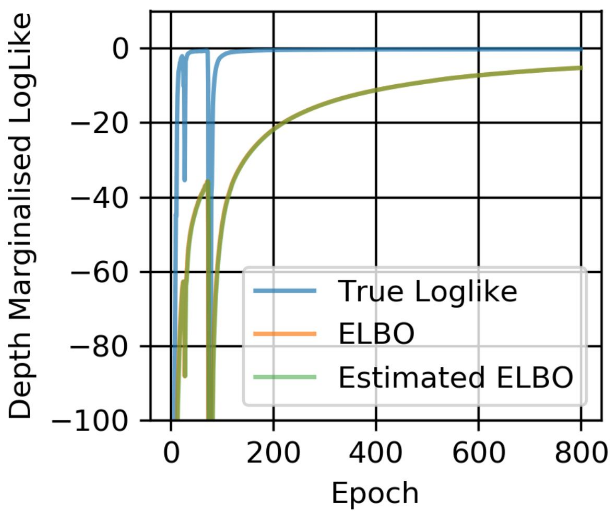
$$p_\theta(d{=}j|\mathcal{D}) = \frac{p(d{=}j)\cdot\prod_{n=1}^{N} p_\theta(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, d{=}j)}{\sum_{i=0}^{D} p(d{=}i)\cdot\prod_{n=1}^{N} p_\theta(\mathbf{y}^{(n)}|\mathbf{x}^{(n)}, d{=}i)}$$

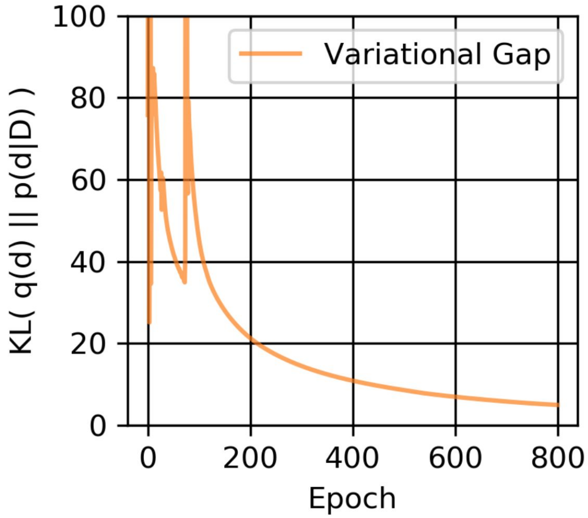
Learning with Variational Inference

- Simultaneously optimise model parameters and distribution over depth

$$\mathcal{L}(\alpha, \theta) = \sum_{n=1}^{N} \mathbb{E}_{q_\alpha(d)} \left[\log p_\theta(\mathbf{y}^{(n)} | \mathbf{x}^{(n)}, d) \right] - \mathbf{KL}(q_\alpha(d) \,\|\, p_\beta(d))$$

$$\mathcal{L}(\alpha, \theta) \approx \frac{N}{N'} \sum_{n=1}^{N'} \sum_{i=0}^{D} \left(\log p_\theta(\mathbf{y}^{(n)} | f_{D+1}(\mathbf{a}_i^{(n)})) \cdot \alpha_i \right) - \sum_{i=0}^{D} \left(\alpha_i \log \frac{\alpha_i}{\beta_i} \right)$$

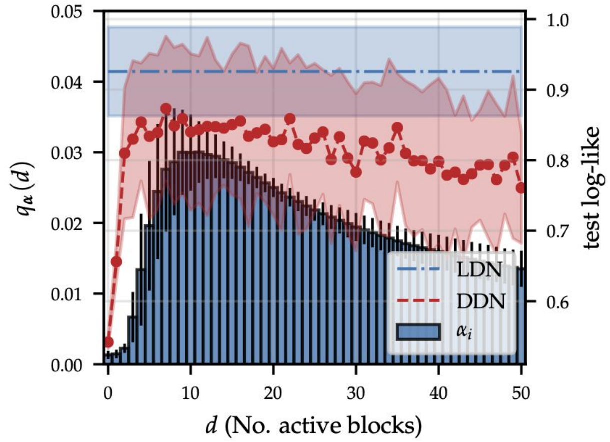




Choosing a Depth and Making Predictions

- Choosing a Depth

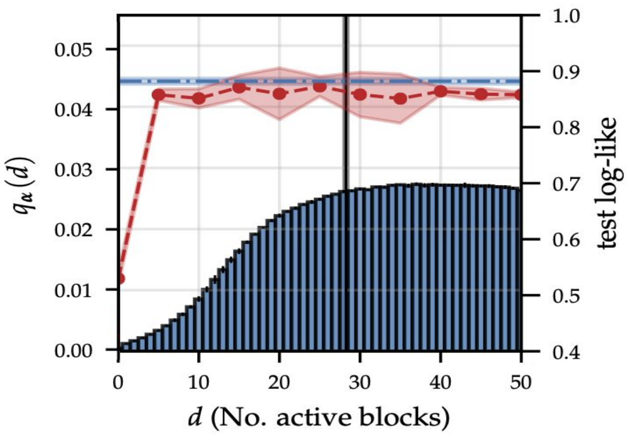
$$d_{\text{opt}} = \text{argmax}_i \, \alpha_i$$
$$d_{\text{opt}} = \min_i \{ i : \alpha_i \geq 0.95 \max_i \alpha_i \}$$



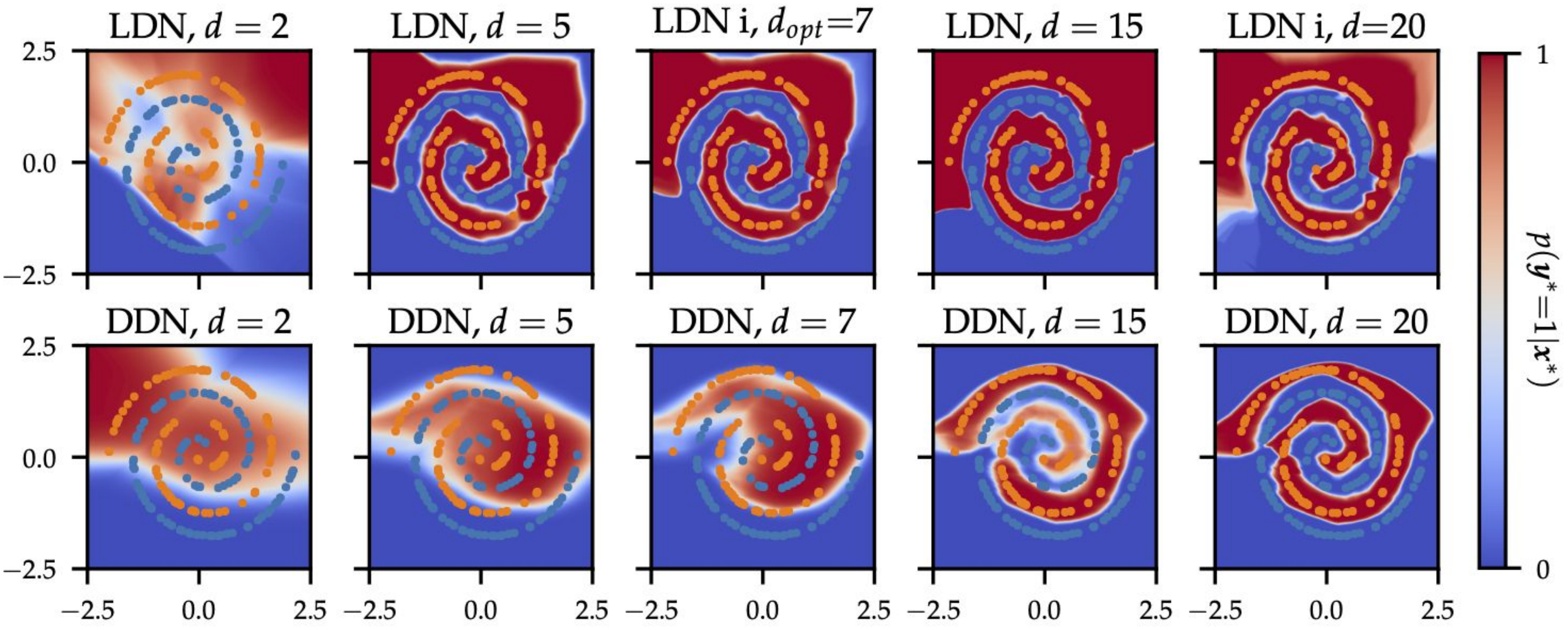
- Predicting by Marginalising

$$q_\alpha(d{=}d_{\text{opt}}) = q_\alpha(d{\geq}d_{\text{opt}})$$
$$q_\alpha(d{>}d_{\text{opt}}) = 0$$

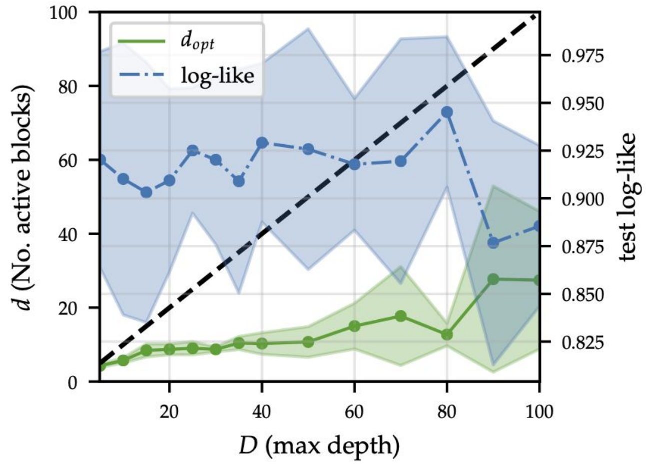
$$p(\mathbf{y}^* | \mathbf{x}^*) \approx \sum_{i=0}^{d_{\text{opt}}} p_\theta(\mathbf{y}^* | \mathbf{x}^*, d{=}i) q_\alpha(d{=}i)$$

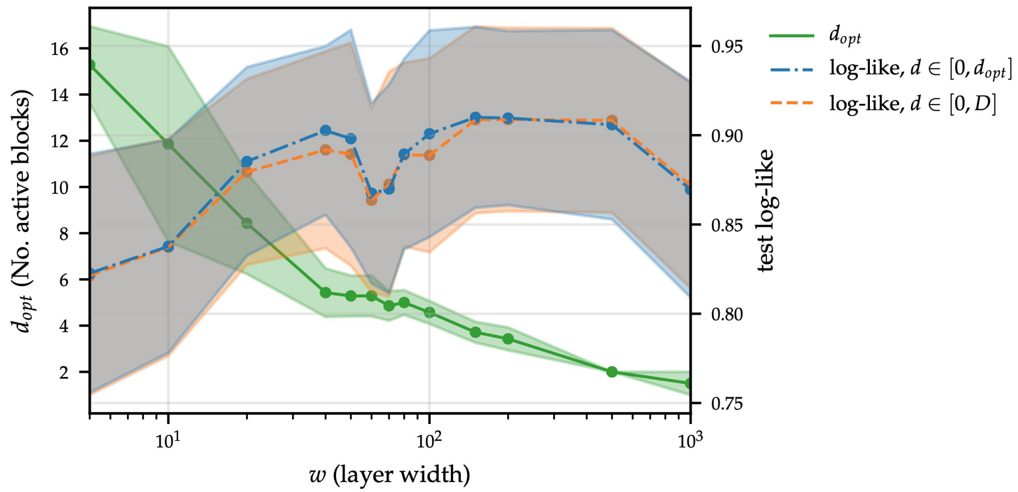


Making Efficient use of Network Layers

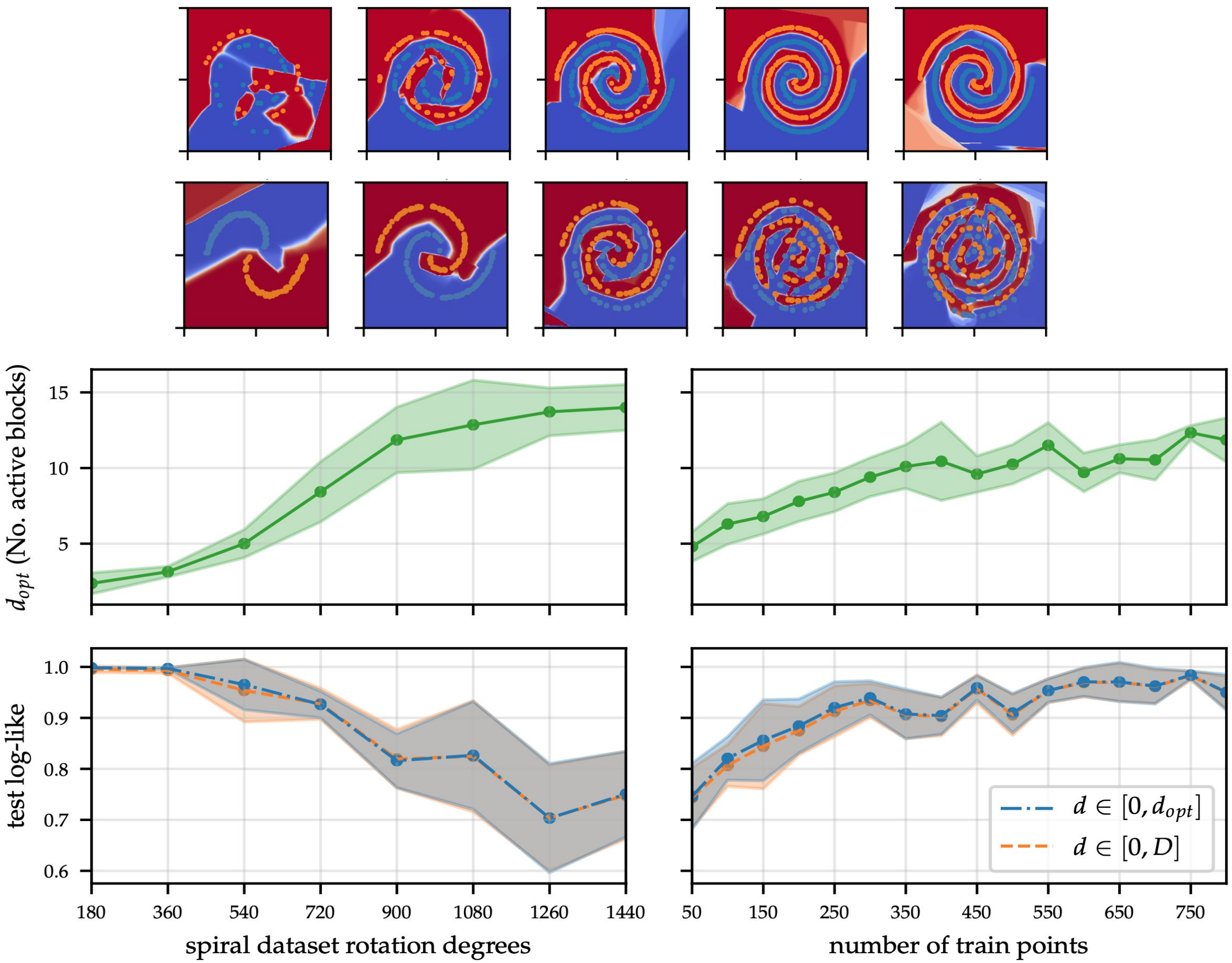


Consistent Depth Predictions

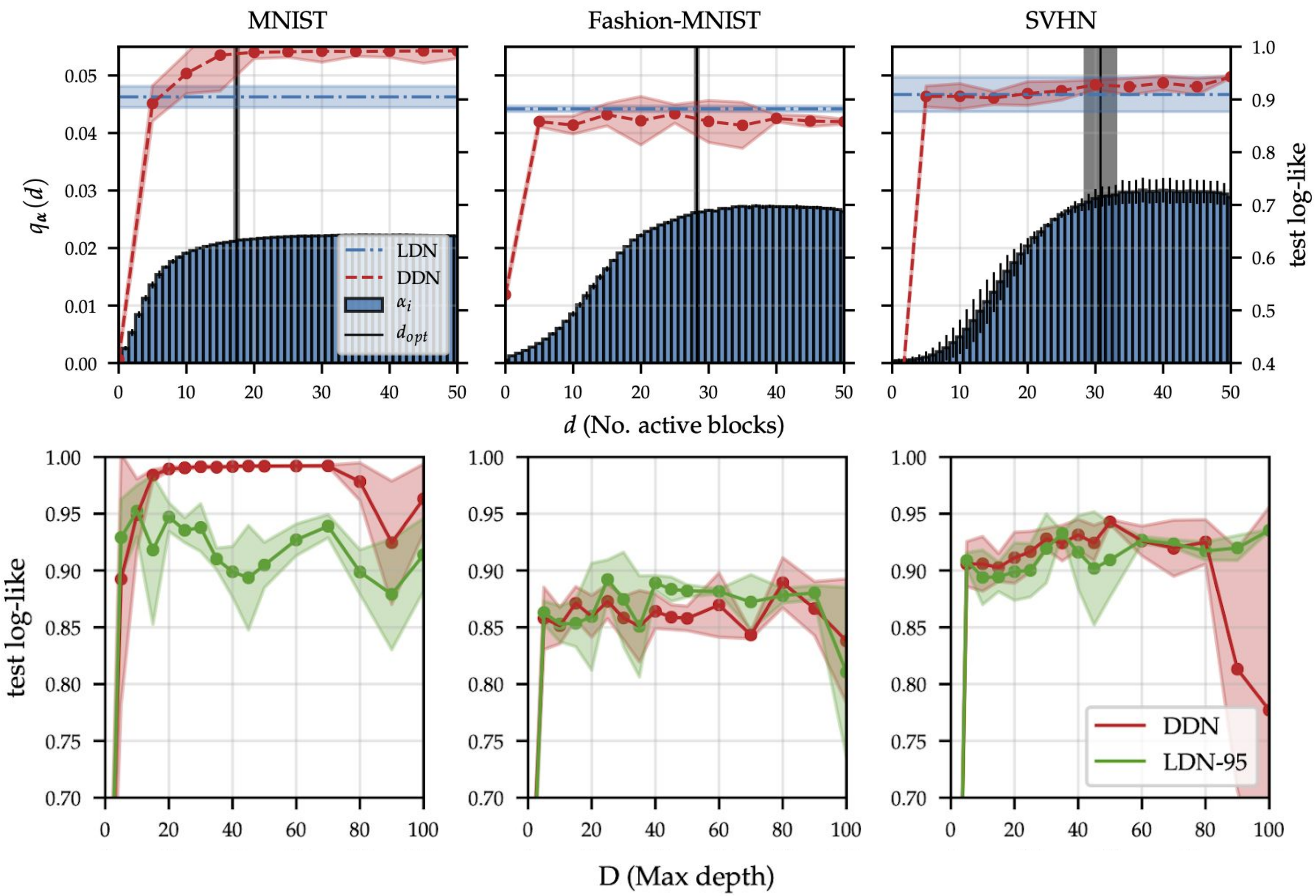




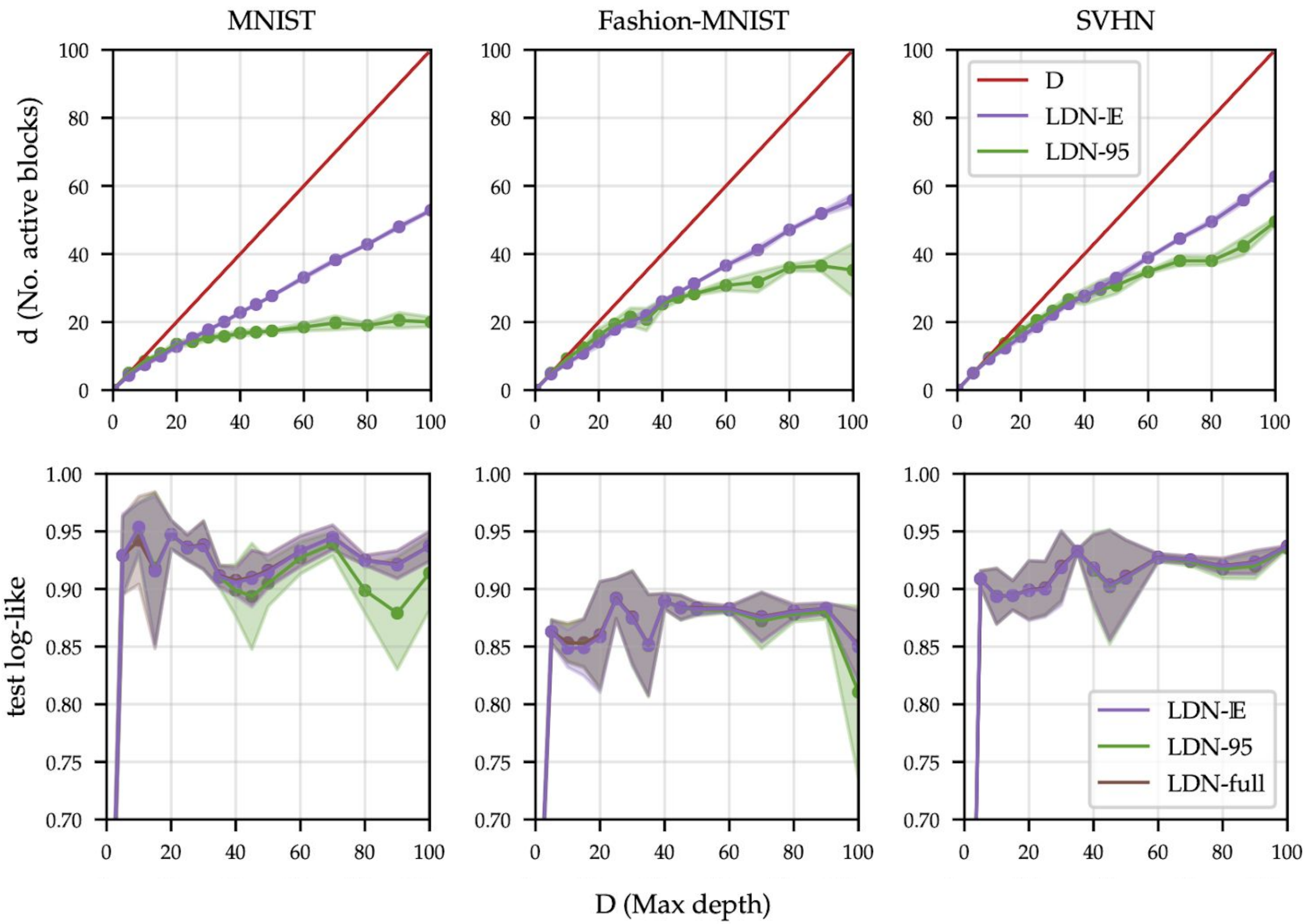
Depth Scales with Complexity



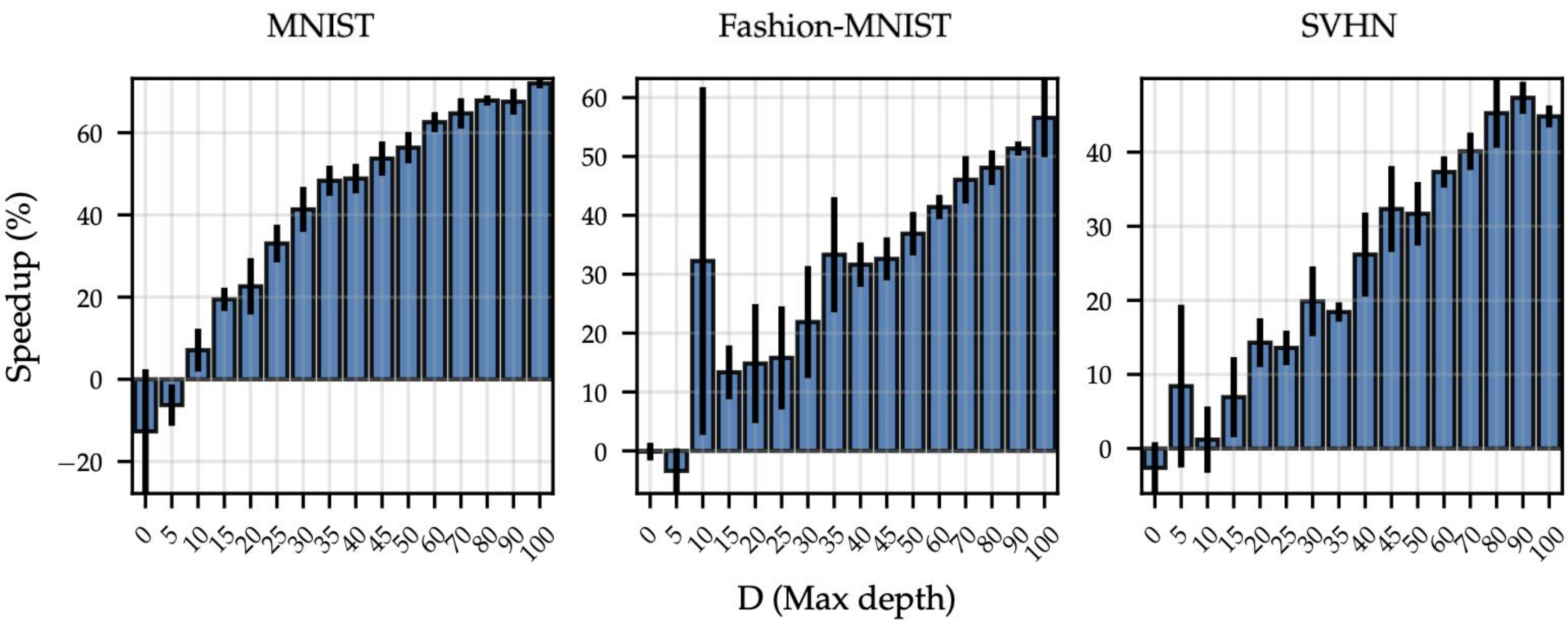
Scaling to Small Image Datasets



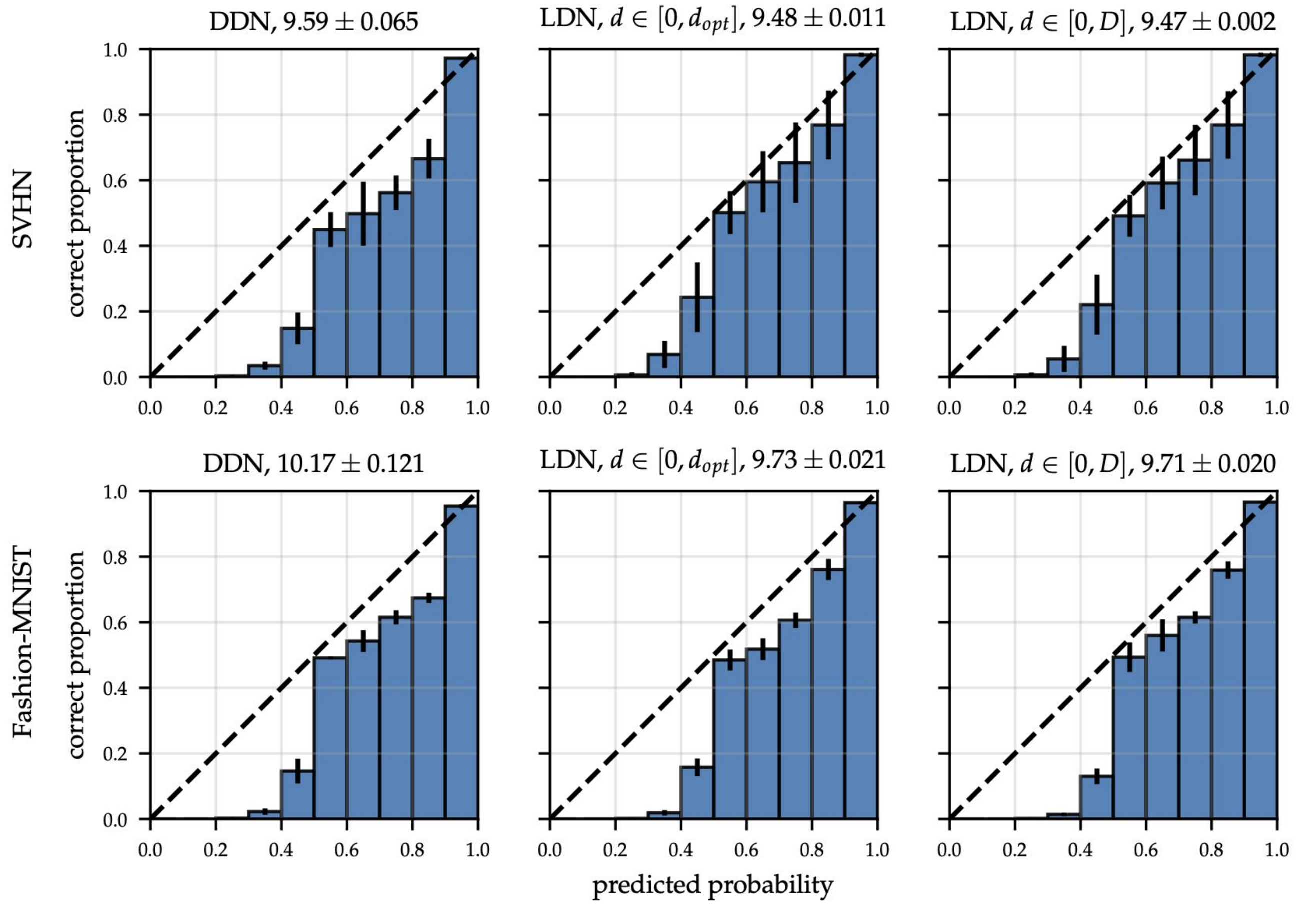
Pruning Strategies



Computational Speedup



Uncertainty Calibration



Summary

- We perform NAS over network depth at very low cost
- We find smaller, cheaper, models with no performance loss
- Procedure fits into a tractable probabilistic framework
    - Allows us to capture some model uncertainty for free

https://github.com/cambridge-mlg/arch_uncert

References

Antorán, Javier, James Urquhart Allingham, and José Miguel Hernández-Lobato. "Variational Depth Search in ResNets." *arXiv preprint arXiv:2002.02797* (2020).

Zoph, Barret, and Quoc V. Le. "Neural Architecture Search with Reinforcement Learning." (2017).

Real, Esteban, et al. "Regularized evolution for image classifier architecture search." *Proceedings of the aaai conference on artificial intelligence*. Vol. 33. 2019.

Zoph, Barret, et al. "Learning transferable architectures for scalable image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

Liu, Hanxiao, Karen Simonyan, and Yiming Yang. "Darts: Differentiable architecture search." *arXiv preprint arXiv:1806.09055* (2018).

Pham, Hieu, et al. "Efficient neural architecture search via parameter sharing." *arXiv preprint arXiv:1802.03268* (2018).

Huang, Gao, et al. "Deep networks with stochastic depth." *European conference on computer vision*. Springer, Cham, 2016.