Sparse MoEs meet Efficient Ensembles

A recipe for BIG models with low compute cost and strong robustness in the world of fine-tuning

Machine Learning Efficiency Workshop @ DLI 2022

James Urquhart Allingham

26 August 2022

Sparse MoEs meet Efficient Ensembles

James Urquhart Allingham[1,*] *jua23@cam.ac.uk*

Florian Wenzel[2,†] *fln.wenzel@gmail.com*

Zelda E Mariet, Basil Mustafa *{zmariet,basilm}@google.com*

Joan Puigcerver, Neil Houlsby *{jpuigcerver,neilhoulsby}@google.com*

Ghassen Jerfel[3,†] *ghassen@google.com*

Vincent Fortuin[1,4,*] *vbf21@cam.ac.uk*

Balaji Lakshminarayanan, Jasper Snoek *{balajiln,jsnoek}@google.com*

Dustin Tran, Carlos Riquelme, Rodolphe Jenatton *{trandustin,rikel,rjenatton}@google.com*

Google Research, Brain Team; [1] *University of Cambridge;* [2] *no affiliation;* [3] *Waymo;* [4] *ETH Zürich*

Self Driving Cars

A brief case study in safety critical applications on the edge

The story so far...

- There is a need for models which have some notion of uncertainty and robustness to dataset shift.
- Many safety critical applications are at the edge.
- **Larger** models have been shown to be **more** robust and uncertainty aware[1].
- This poses a problem since we want both efficiency **and** uncertainty + robustness. 
- Additionally, many practitioners can't train such models. 
- Spoiler Alert!

- Our solution is the combination of sparse MoEs and efficient ensembles.

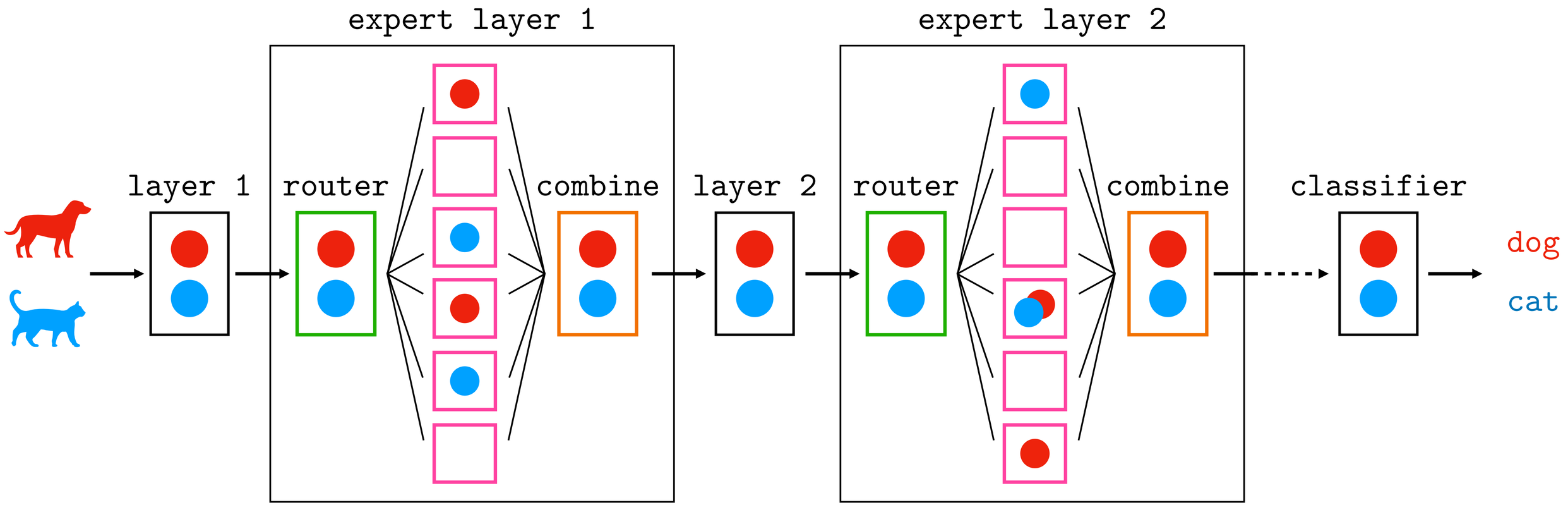
[1] "Plex: Towards Reliability using Pretrained Large Model Extensions", Tran, et al. 2022

Sparse Mixtures of Experts (Sparse MoEs)

Bigger models without bigger compute

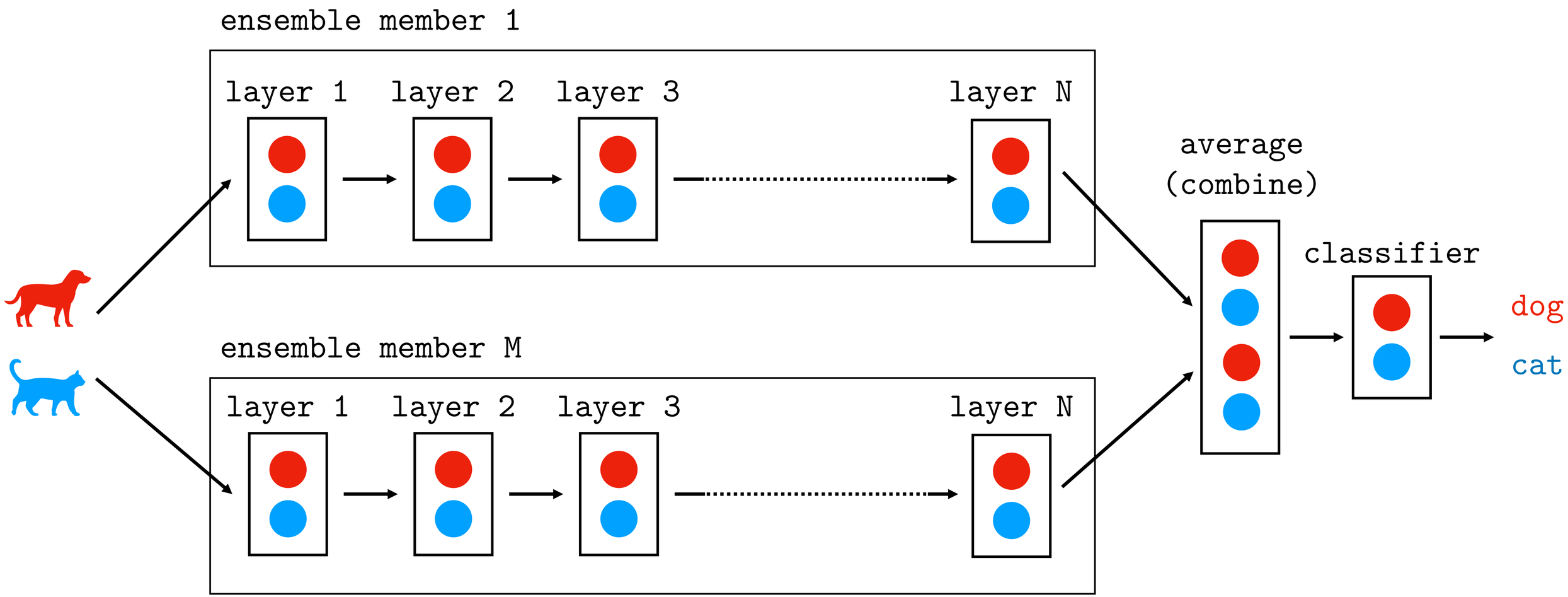
$$f(\mathbf{x}) = \sum_{j=1}^{E} r_j(\mathbf{x}) \cdot expert_j(\mathbf{x})$$

"in parallel"



Ensembles of Neural Networks

Easy robustness and uncertainty awareness

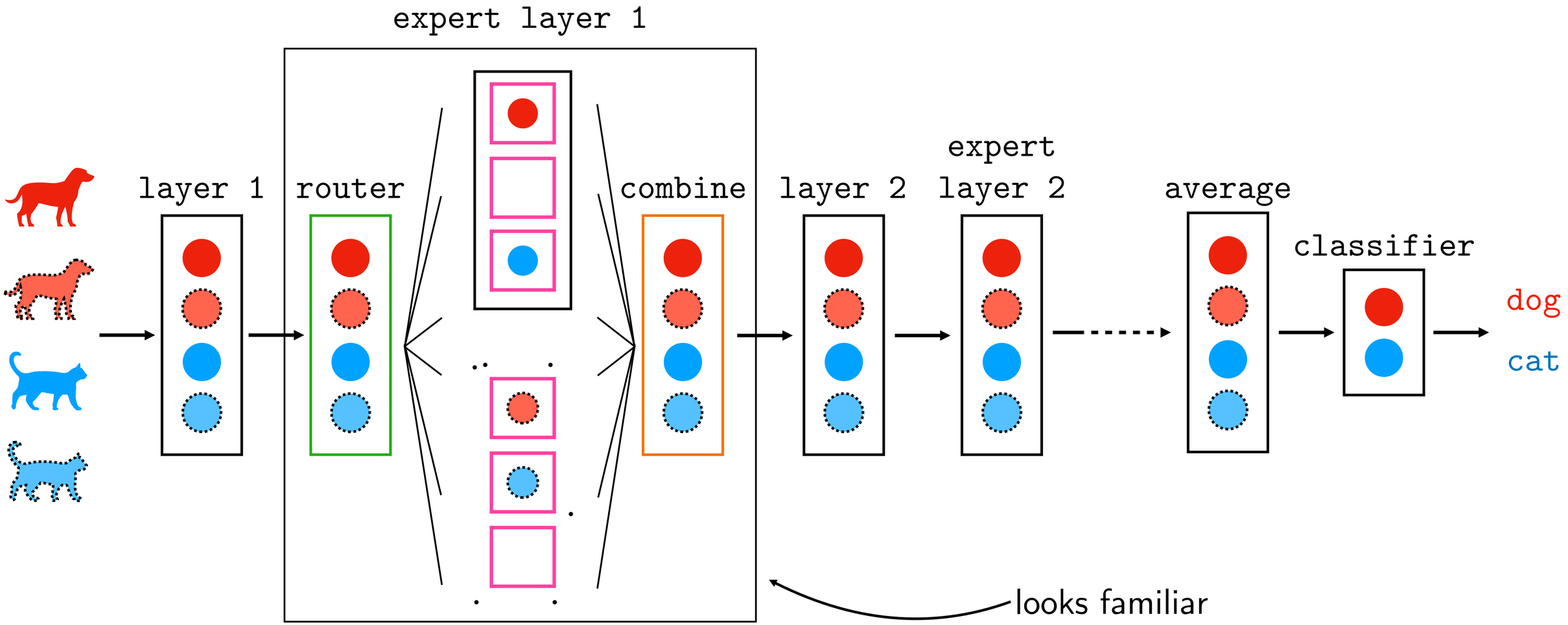


Sparse MoEs vs Ensembles

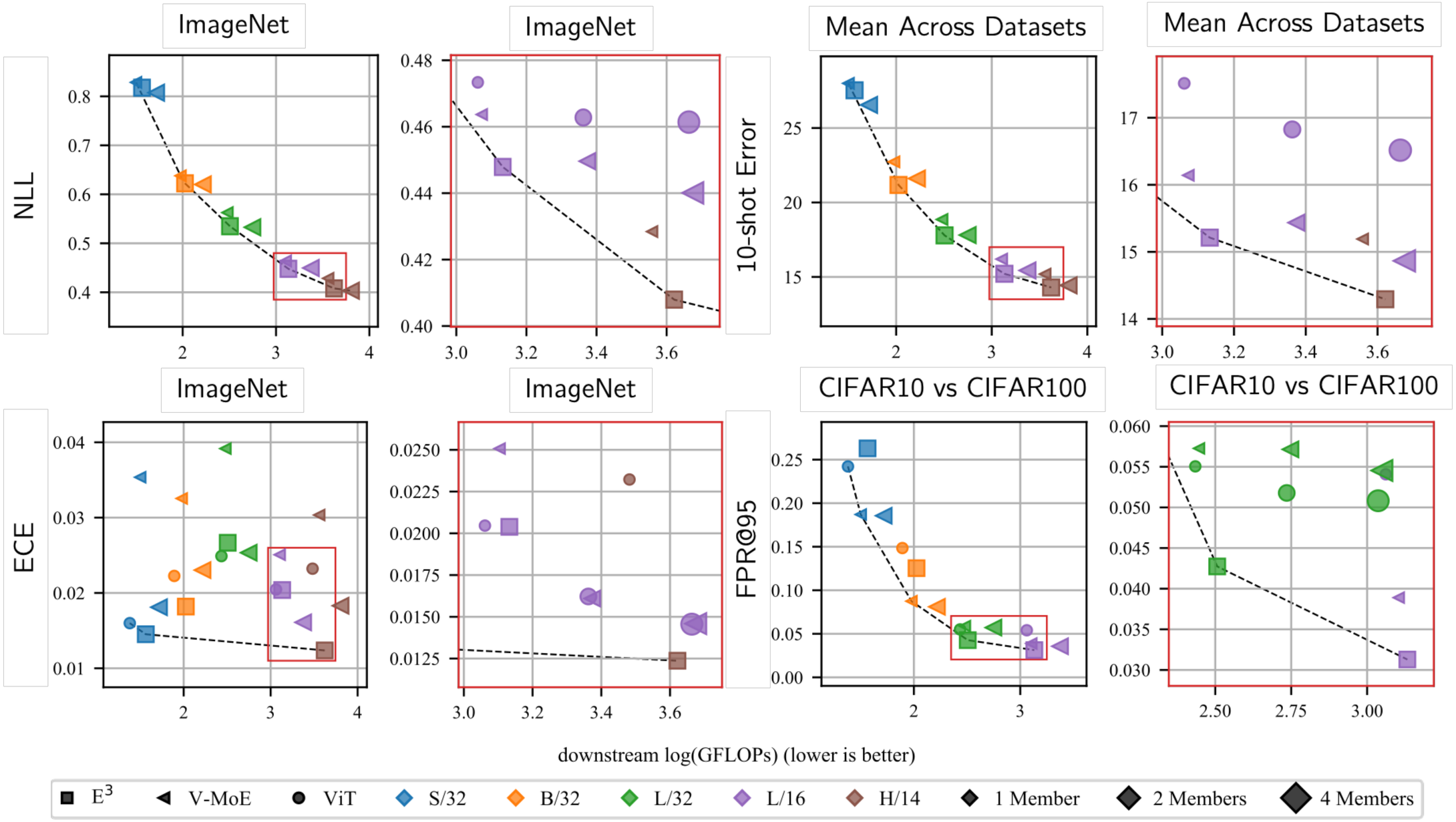
Sparse MoEs	Ensembles
Single prediction	Multiple predictions
Per-example adaptivity	Static combination
Combination at activation level	Combination at prediction level
Compute $\approx$ standard NN	Compute $\gg$ standard NN
???	Robust to distribution shift, well-calibrated uncertainty, good OOD detection

Efficient Ensemble of Experts (E^3)

Sparse MoEs meet Efficient Ensembles



Highlighted Results (lower is better)



Conclusion

- Safety critical applications on the edge need robust **and** efficient models
- This is in contrast with modern methods which can be robust but **very** inefficient
- E3 aims to fill this niche:
  - Uses the same compute as a standard NN
  - Improves on uncertainty estimation, robustness, few-shot learning
  - Can be fine-tuned from MoE check points
  - Downside: requires a lot of memory! (But, memory is cheaper than compute)