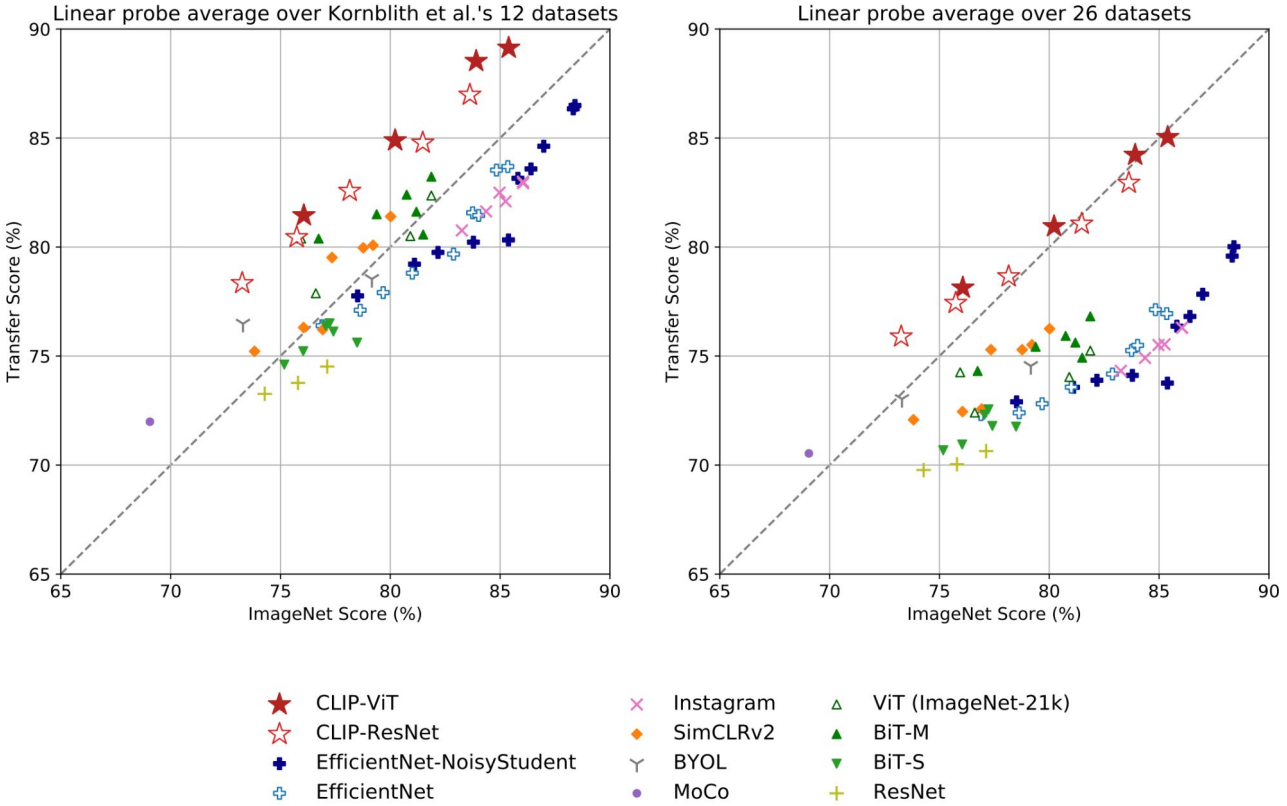
Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time

Fine-tuning

What is meant by fine-tuning of neural network?

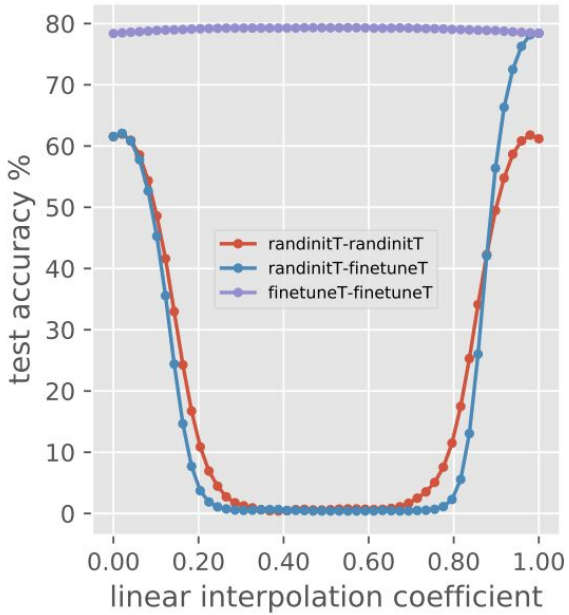
Asked 4 years, 9 months ago Modified 4 years, 9 months ago Viewed 44k times

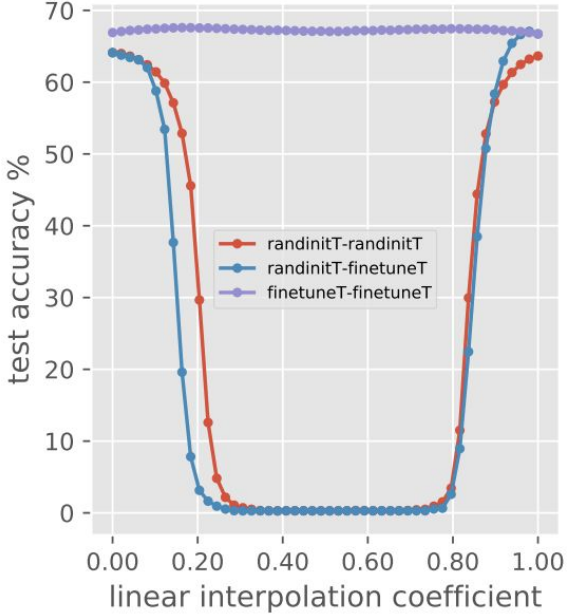
Fine-tuning. Transfer score

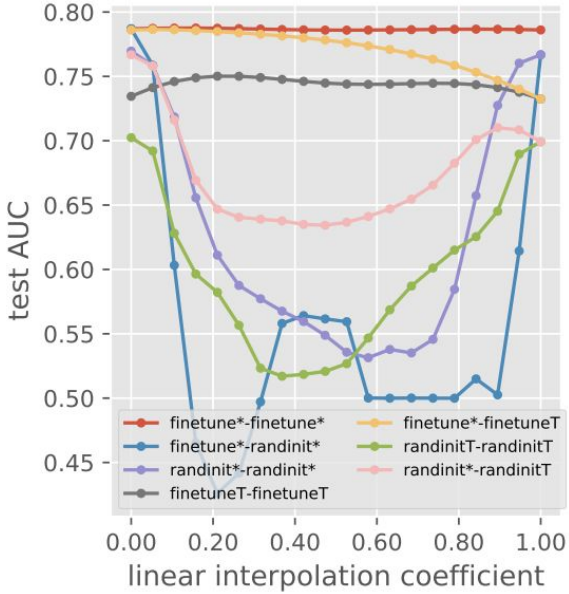


https://arxiv.org/abs/2103.00020

Fine-tuning. Flat basin

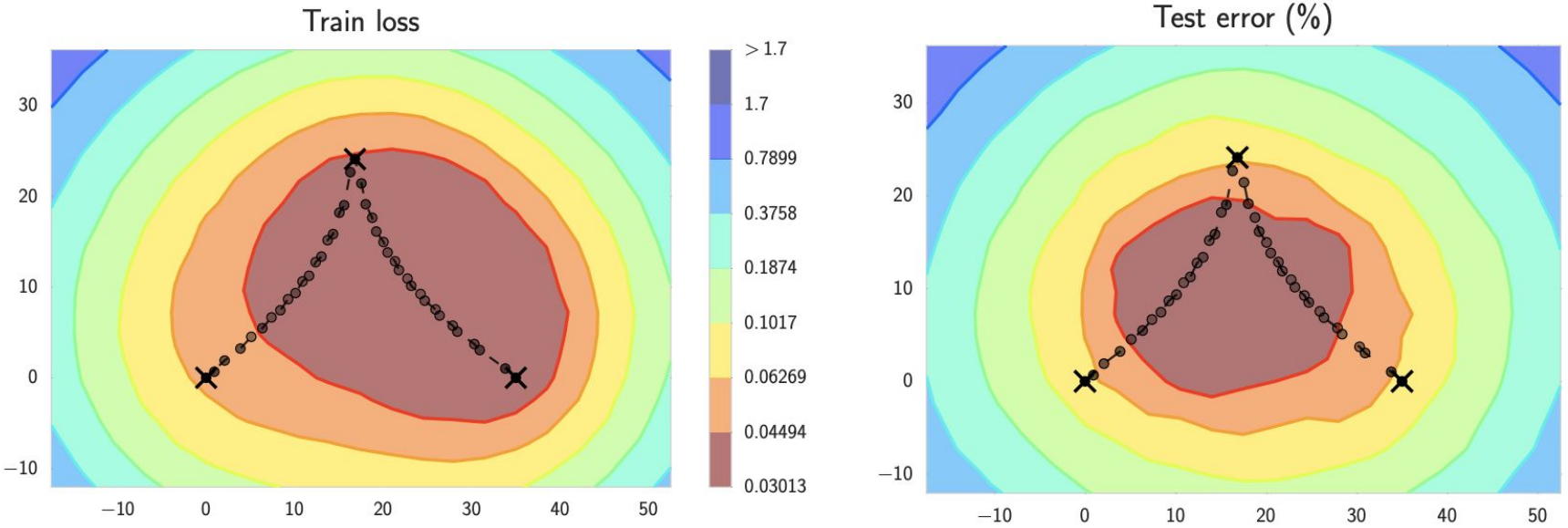






https://arxiv.org/abs/2008.11687

Fine-tuning. Stochastic Weight Averaging



$$w_{\text{SWA}} \leftarrow \frac{w_{\text{SWA}} \cdot n_{\text{models}} + w}{n_{\text{models}} + 1}$$

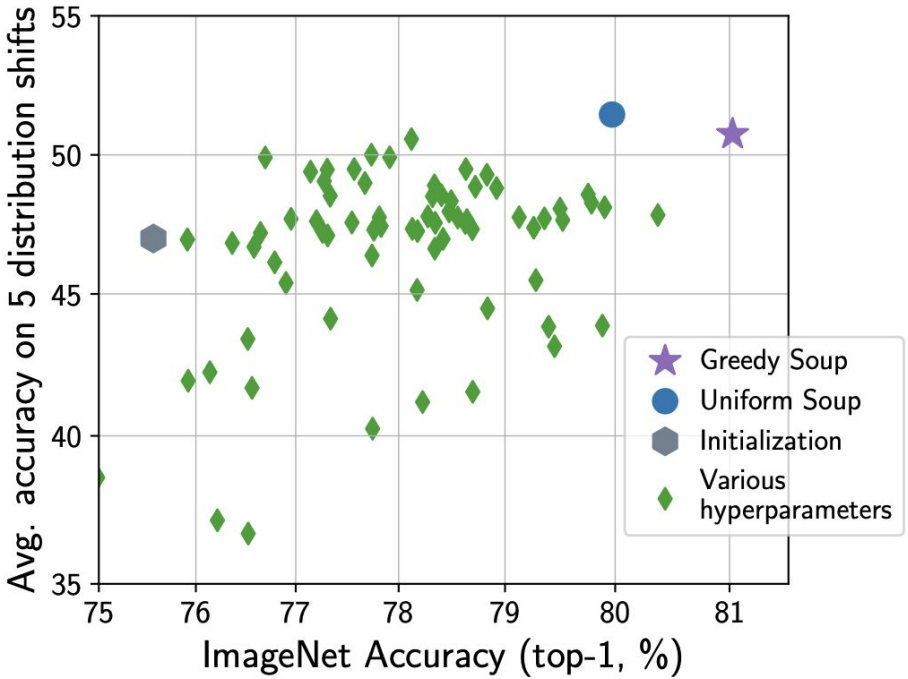
https://arxiv.org/abs/1803.05407

Model Soups

- Best on validation set
- Ensemble

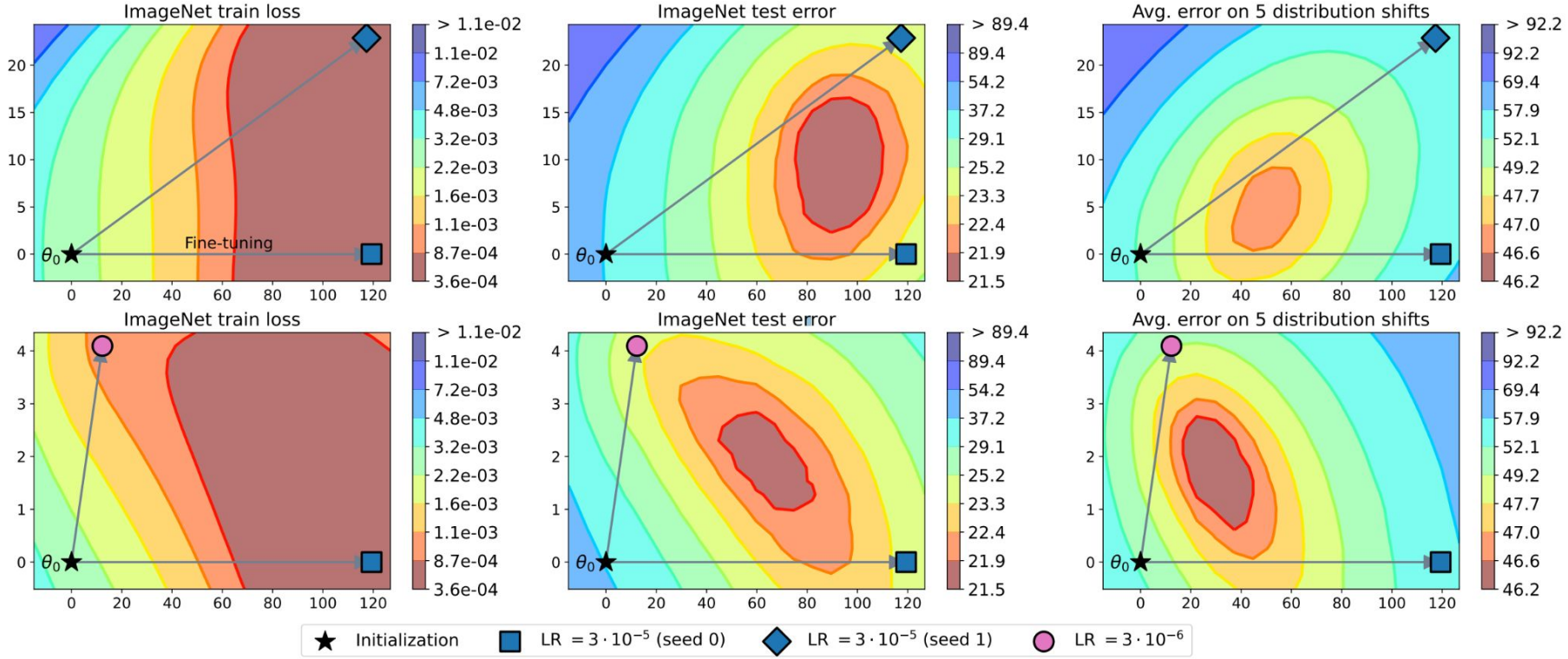
Proposed:

- Uniform soup
- Greedy soup
- Learned soup



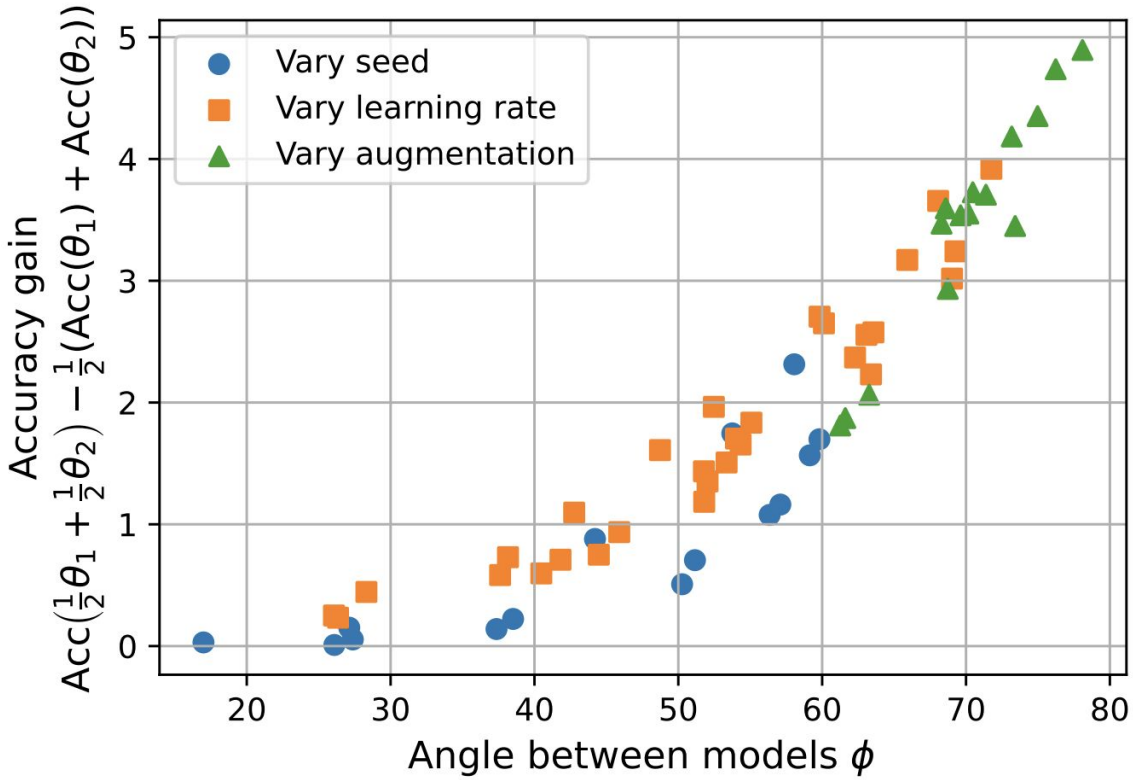
https://arxiv.org/abs/2203.05482

Intuition and motivation. Error landscapes



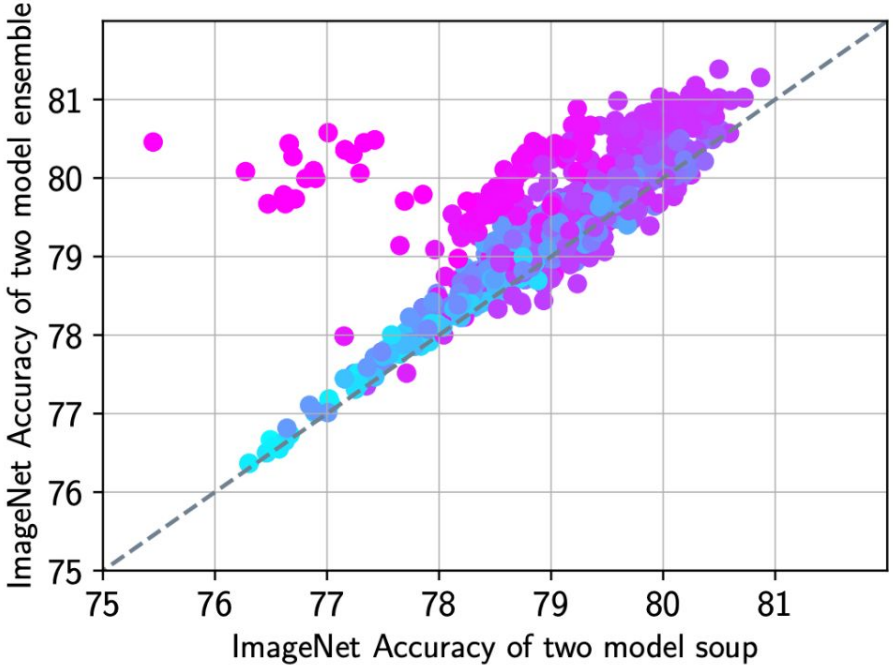
https://arxiv.org/abs/2203.05482

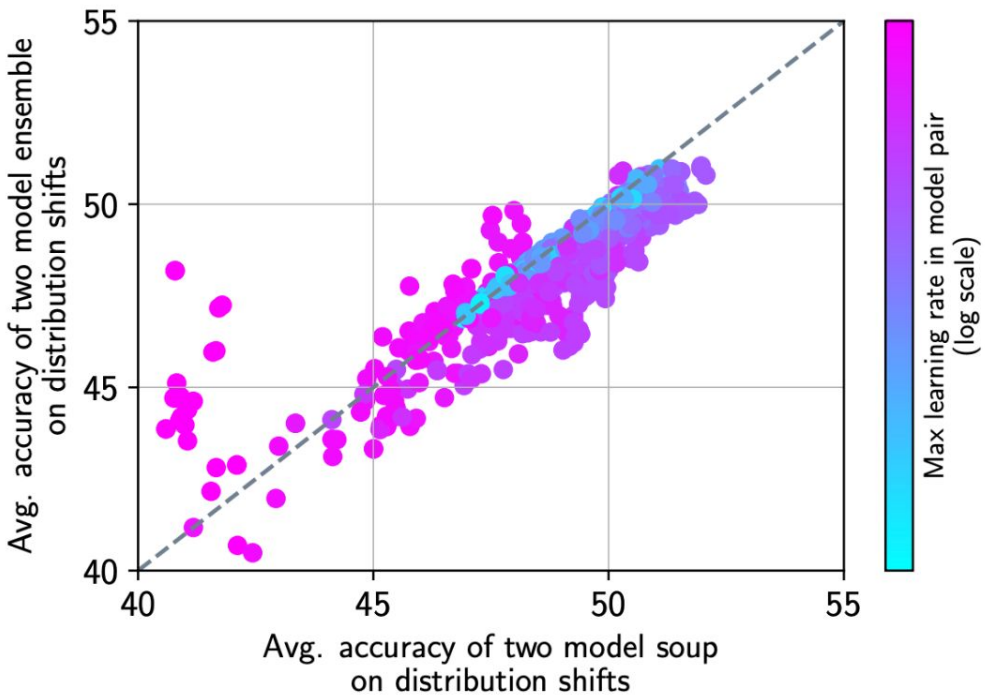
Intuition and motivation. Correlation



https://arxiv.org/abs/2203.05482

Intuition and motivation. Ensembles





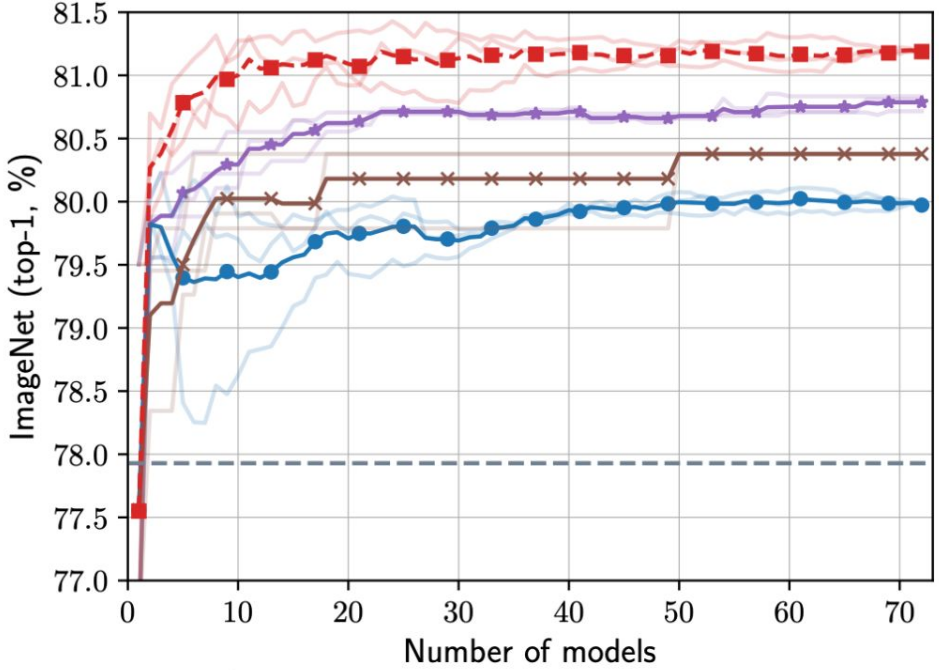
https://arxiv.org/abs/2203.05482

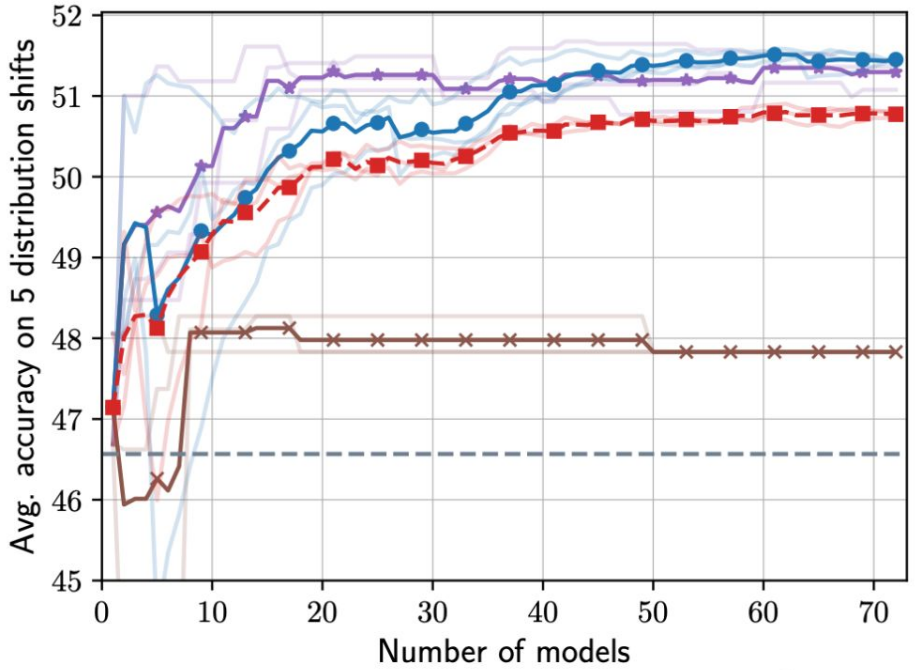
Experiments and results. State of the art

Method	ImageNet			Distribution shifts					
	Top-1	ReaL	Multilabel	IN-V2	IN-R	IN-Sketch	ObjectNet	IN-A	Avg shifts
ViT/G-14 (Zhai et al., 2021)	90.45	90.81	–	83.33	–	–	70.53	–	–
CoAtNet-7 (Dai et al., 2021)	90.88	–	–	–	–	–	–	–	–
Our models/evaluations based on ViT-G/14:									
ViT/G-14 (Zhai et al., 2021) (reevaluated)	90.47	90.86	96.89	83.39	94.38	72.37	71.16	89.00	82.06
Best model on held out val set	90.72	91.04	96.94	83.76	95.04	73.16	78.20	91.75	84.38
Best model on each test set (oracle)	90.78	**91.78**	**97.29**	**84.31**	95.04	73.73	**79.03**	92.16	84.68
Greedy ensemble	**90.93**	91.29	**97.23**	**84.14**	94.85	73.07	77.87	91.69	84.33
Greedy soup	**90.94**	91.20	**97.17**	**84.22**	**95.46**	**74.23**	78.52	**92.67**	**85.02**

https://arxiv.org/abs/2203.05482

Experiments and results





Greedy soup
Uniform soup
Best individual model on ImageNet held-out val set
Ensemble (more compute)
Avg. individual model

https://arxiv.org/abs/2203.05482

The effectiveness of model soups

- The rich literature on ensembles [Gontijo-Lopes et al. (2022)] tells us that the expected error of the ensemble is often strictly below min of errors of each model
- Whenever errors of ensemble and soup are close we expect the soup to outperform both endpoint models

$$\mathcal{L}_\alpha^{\text{soup}} - \mathcal{L}_\alpha^{\text{ens}} \approx \frac{\alpha(1-\alpha)}{2}\left(- \frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\mathcal{L}_\alpha^{\text{soup}} + \beta^2 \mathbb{E}_x \mathrm{Var}_{Y \sim p_{\text{sftmx}}(\beta f(x;\theta_\alpha))}\left[\Delta f_Y(x)\right] \right)$$

https://arxiv.org/abs/2203.05482; https://openreview.net/forum?id=BK-4qbGgIE3