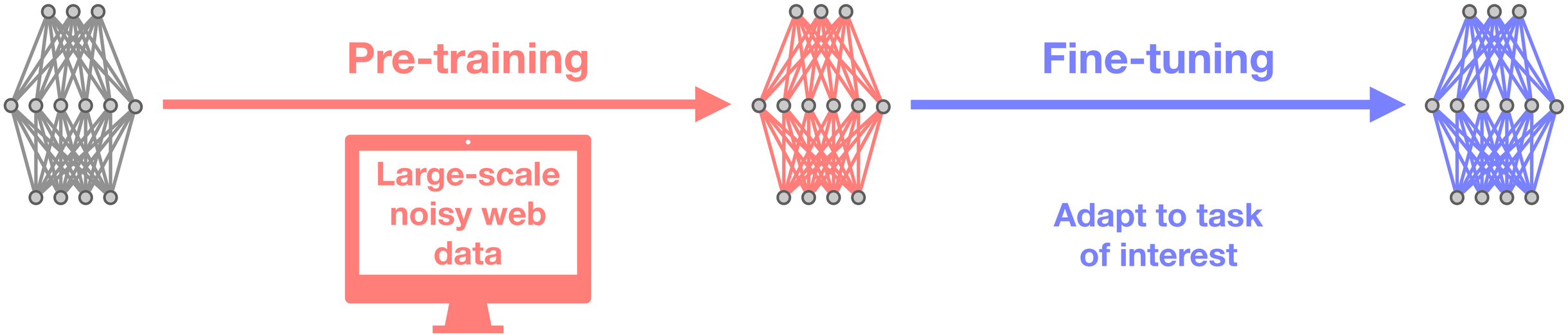
Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time

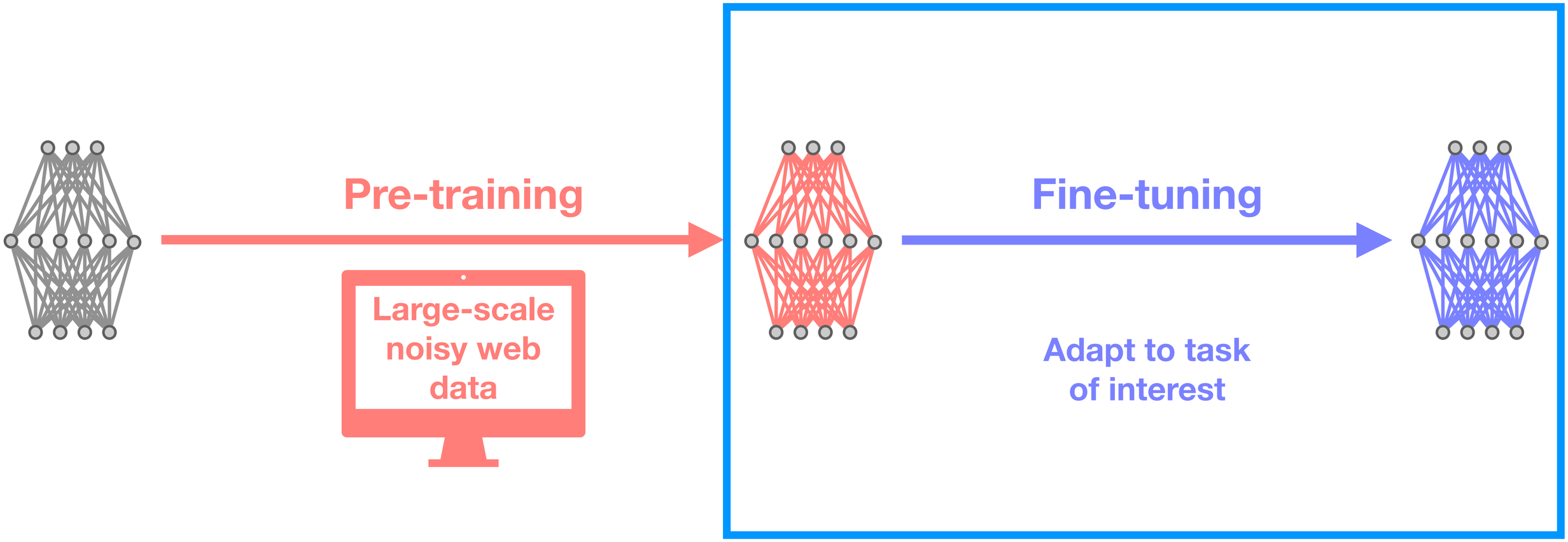
Mitchell Wortsman, Gabriel Ilharco, Samir Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes,
Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon*, Simon Kornblith*, Ludwig Schmidt*

Google Brain

Modern machine learning workflow

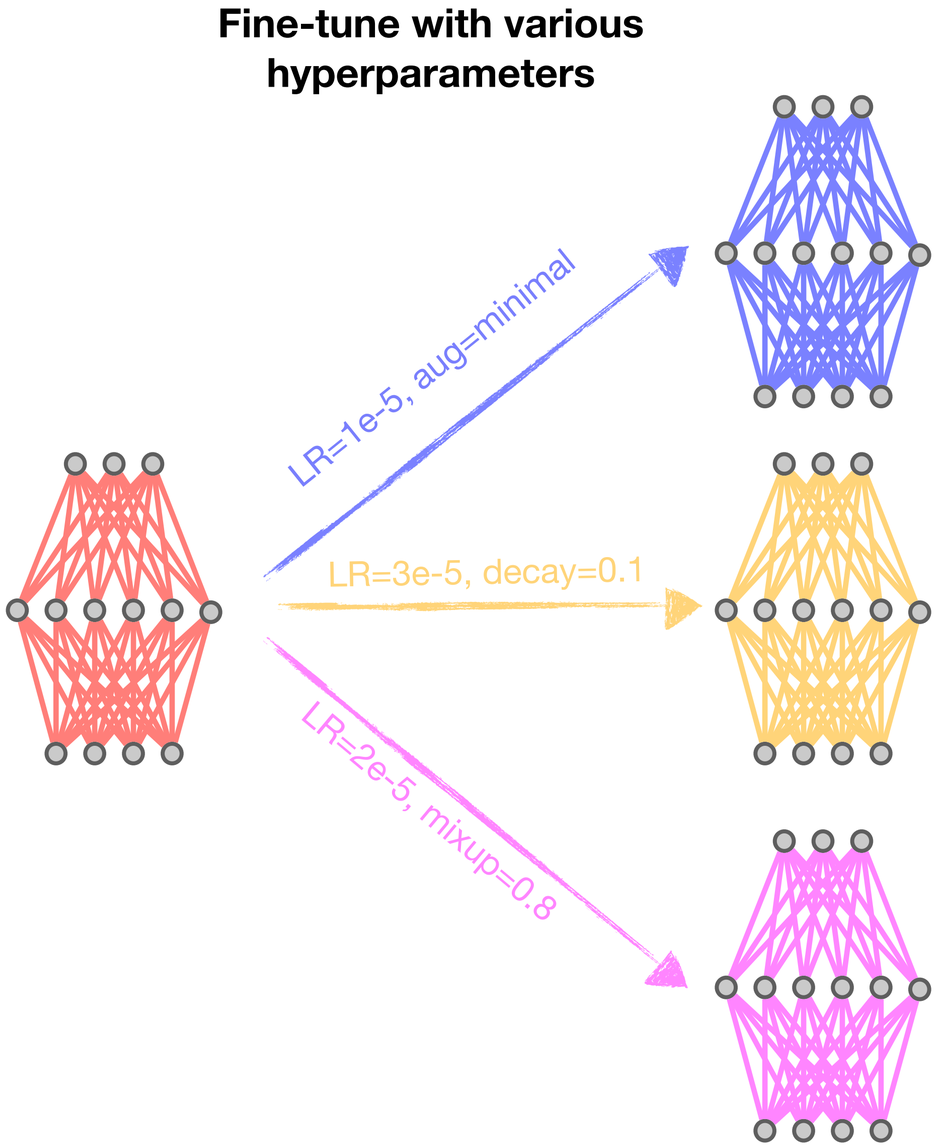


Modern machine learning workflow



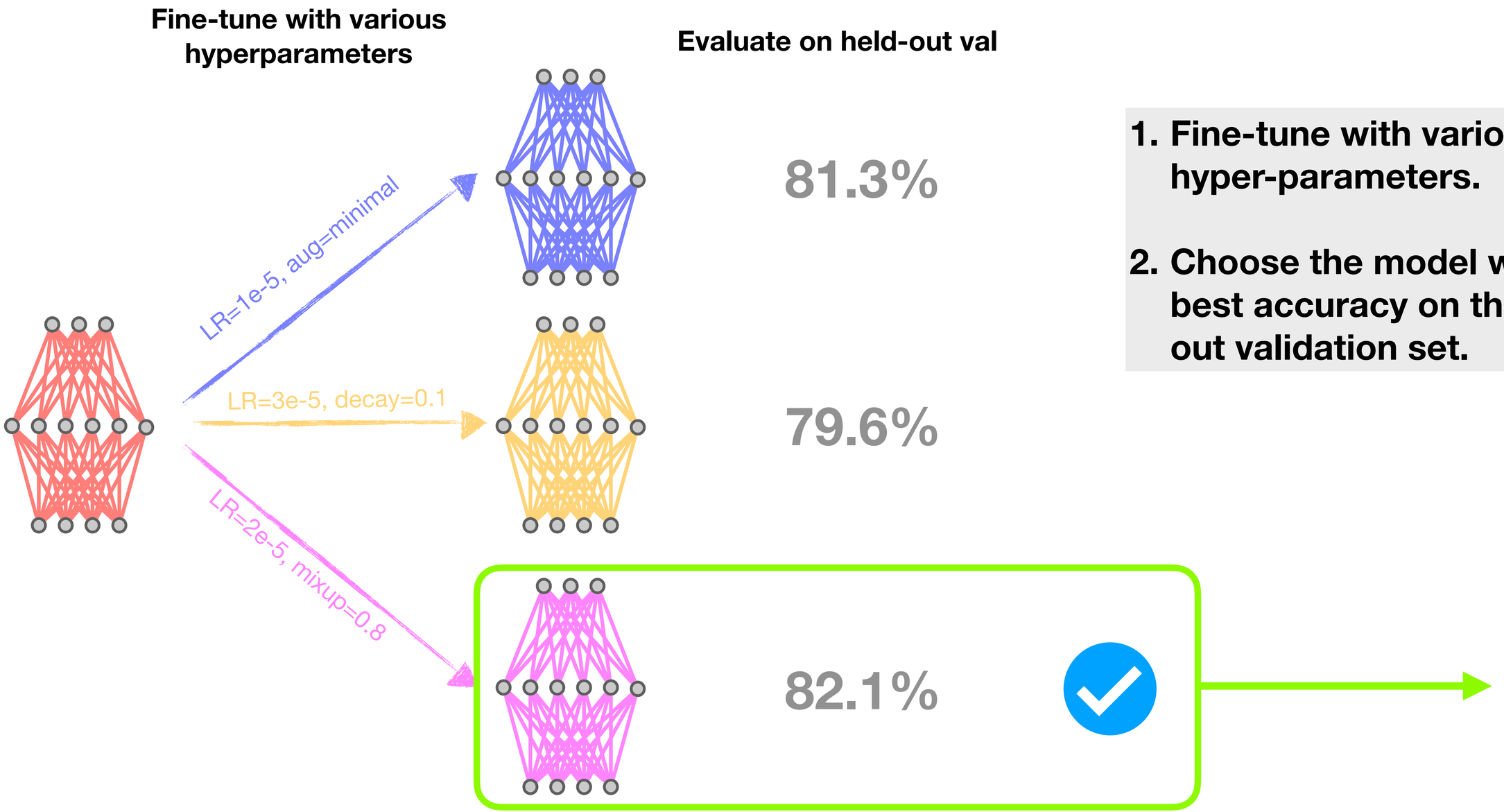
Conventional fine-tuning recipe

Conventional fine-tuning recipe



1. **Fine-tune with various hyper-parameters.**

Conventional fine-tuning recipe



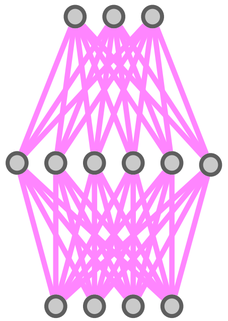
1. Fine-tune with various hyper-parameters.
2. Choose the model with the best accuracy on the held-out validation set.

Downsides of the conventional fine-tuning recipe

Downsides of the conventional fine-tuning recipe

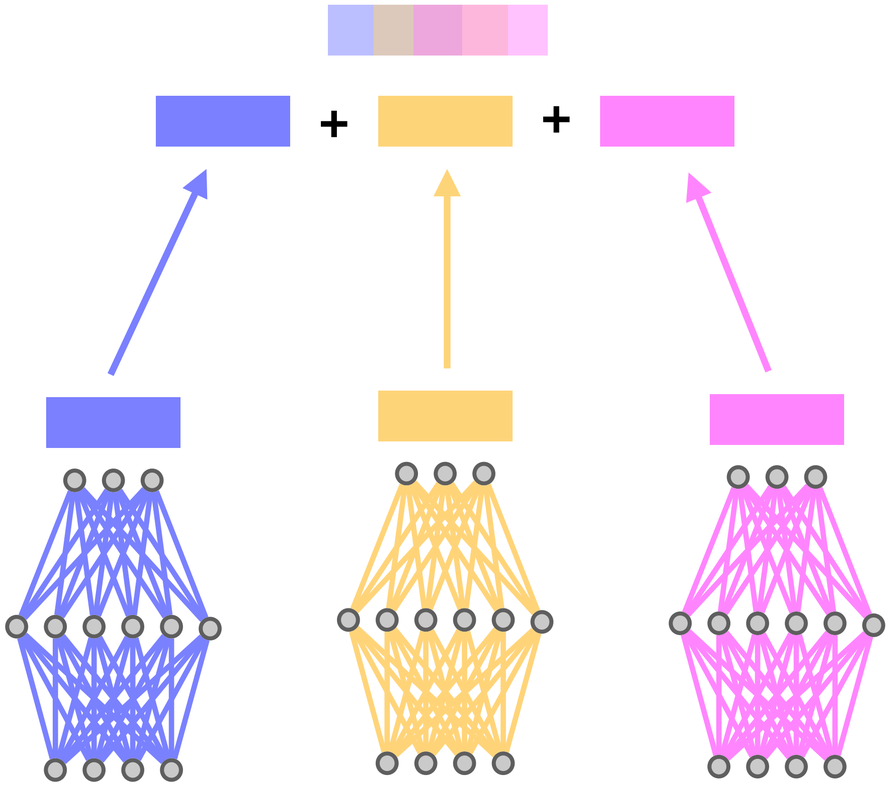
Choosing the best individual model on the held-out validation set

80.1%



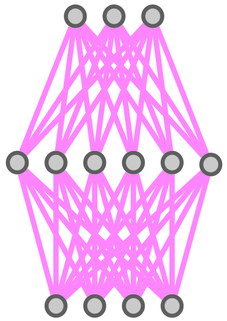
Ensemble

82.3%

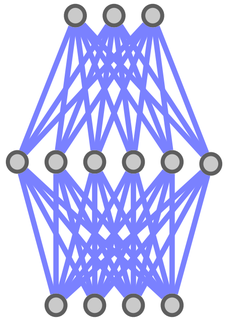


Downsides of the conventional fine-tuning recipe

Target Distribution



80.1%

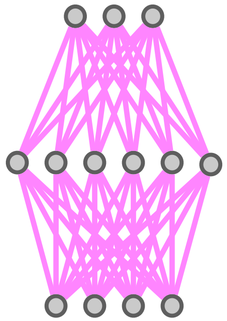


78.9%

Downsides of the conventional fine-tuning recipe

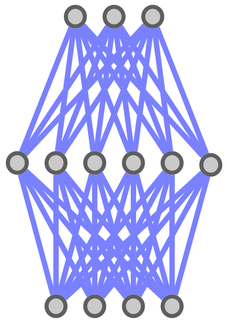
Target Distribution

Distribution shift



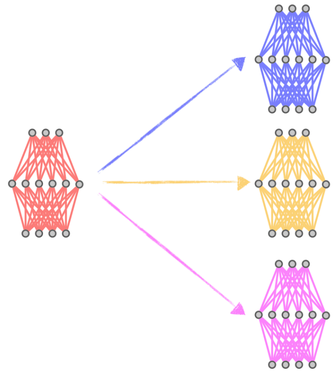
80.1%

69.8%

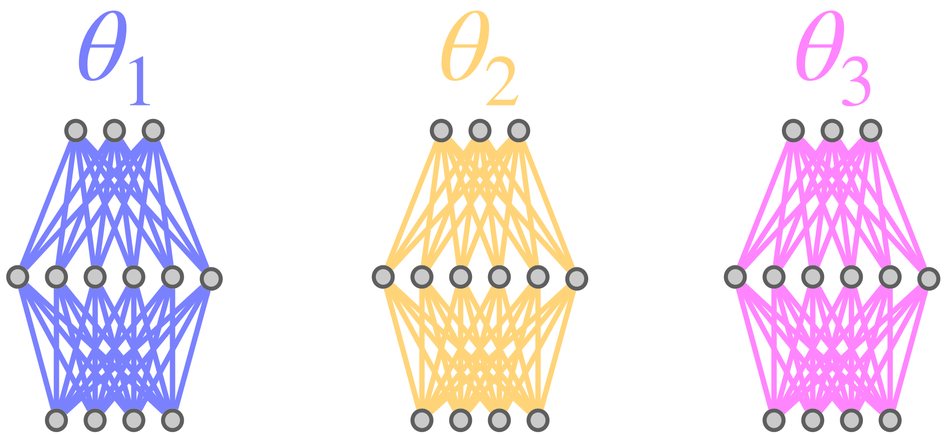


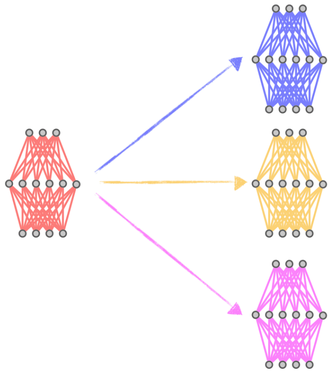
78.9%

72.0%

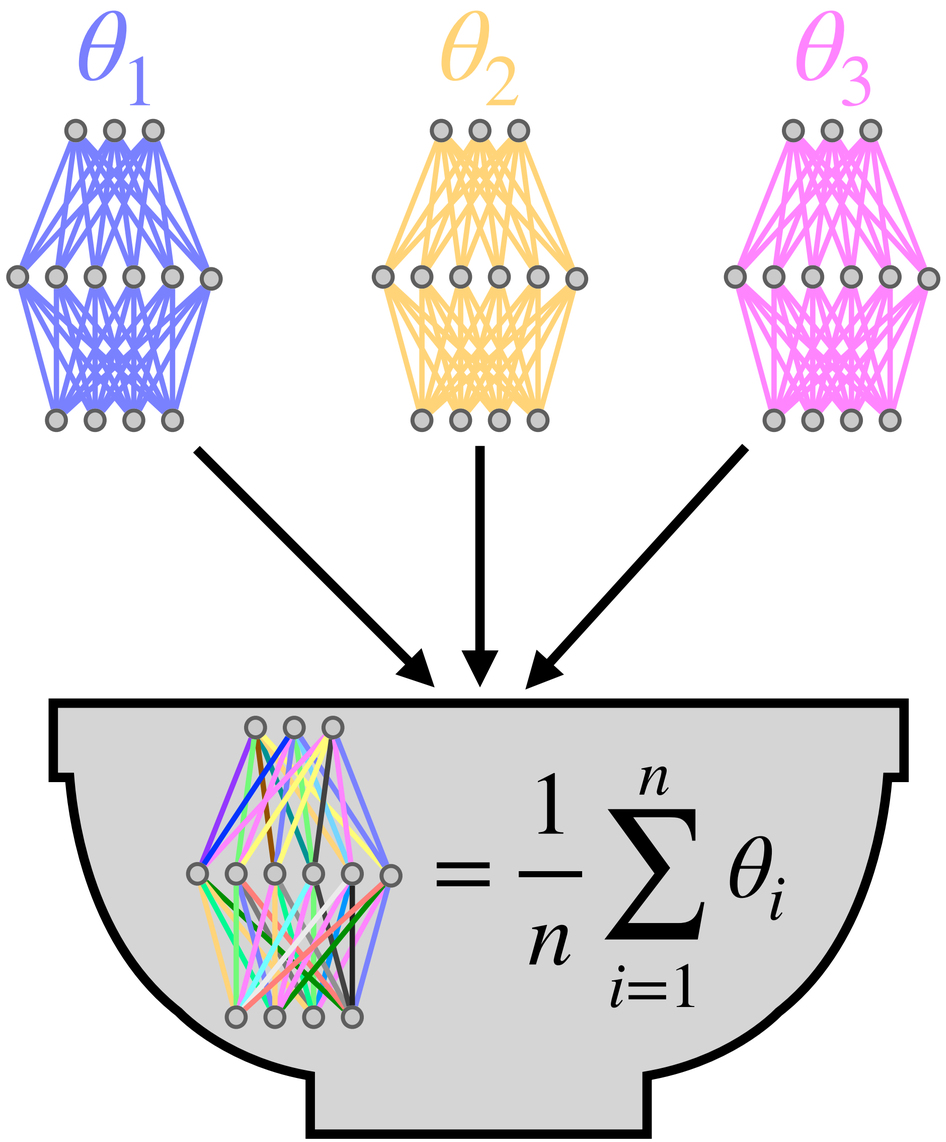


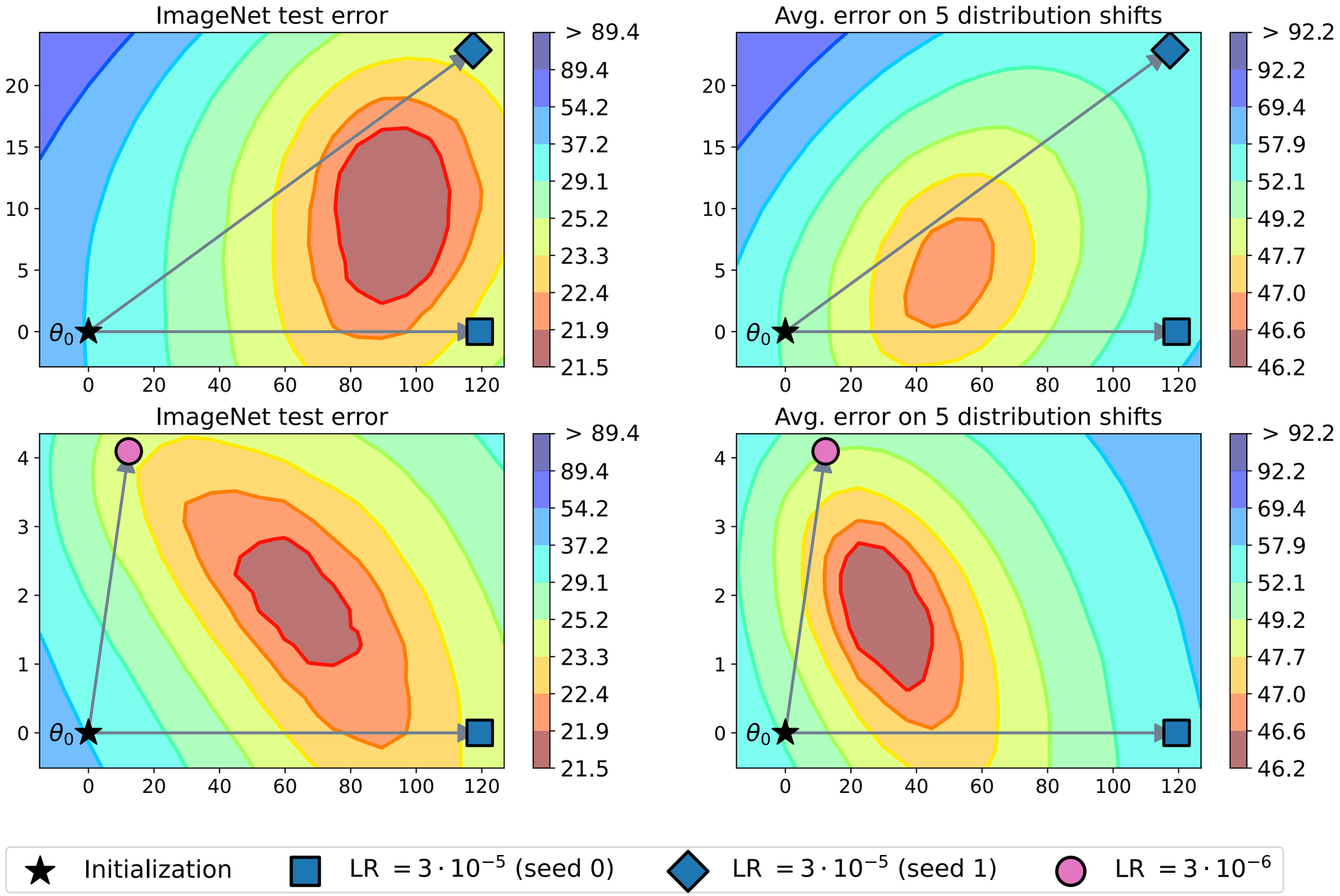
Model soups





Model soups

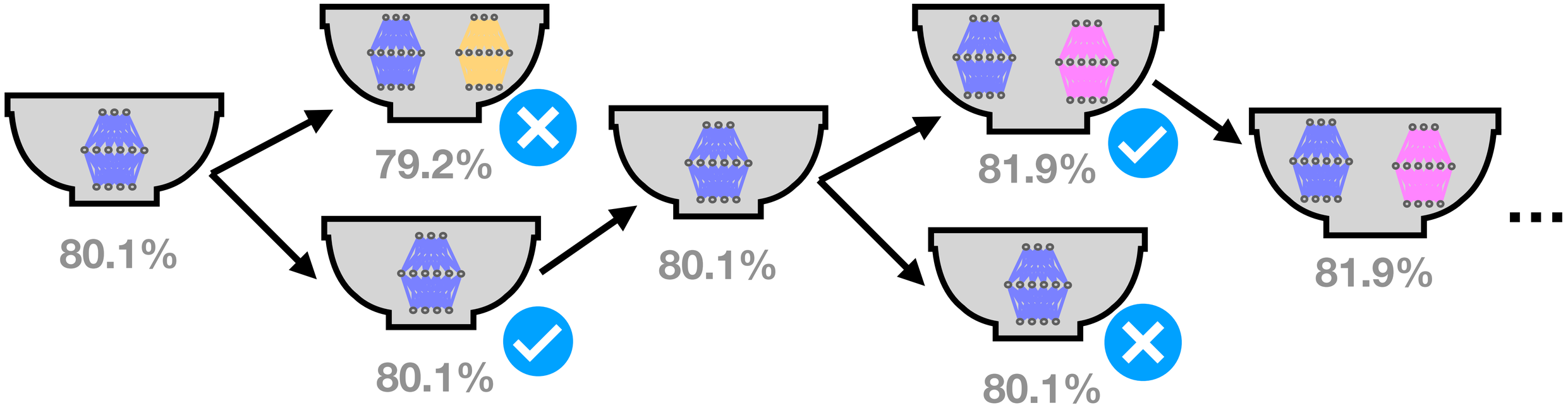


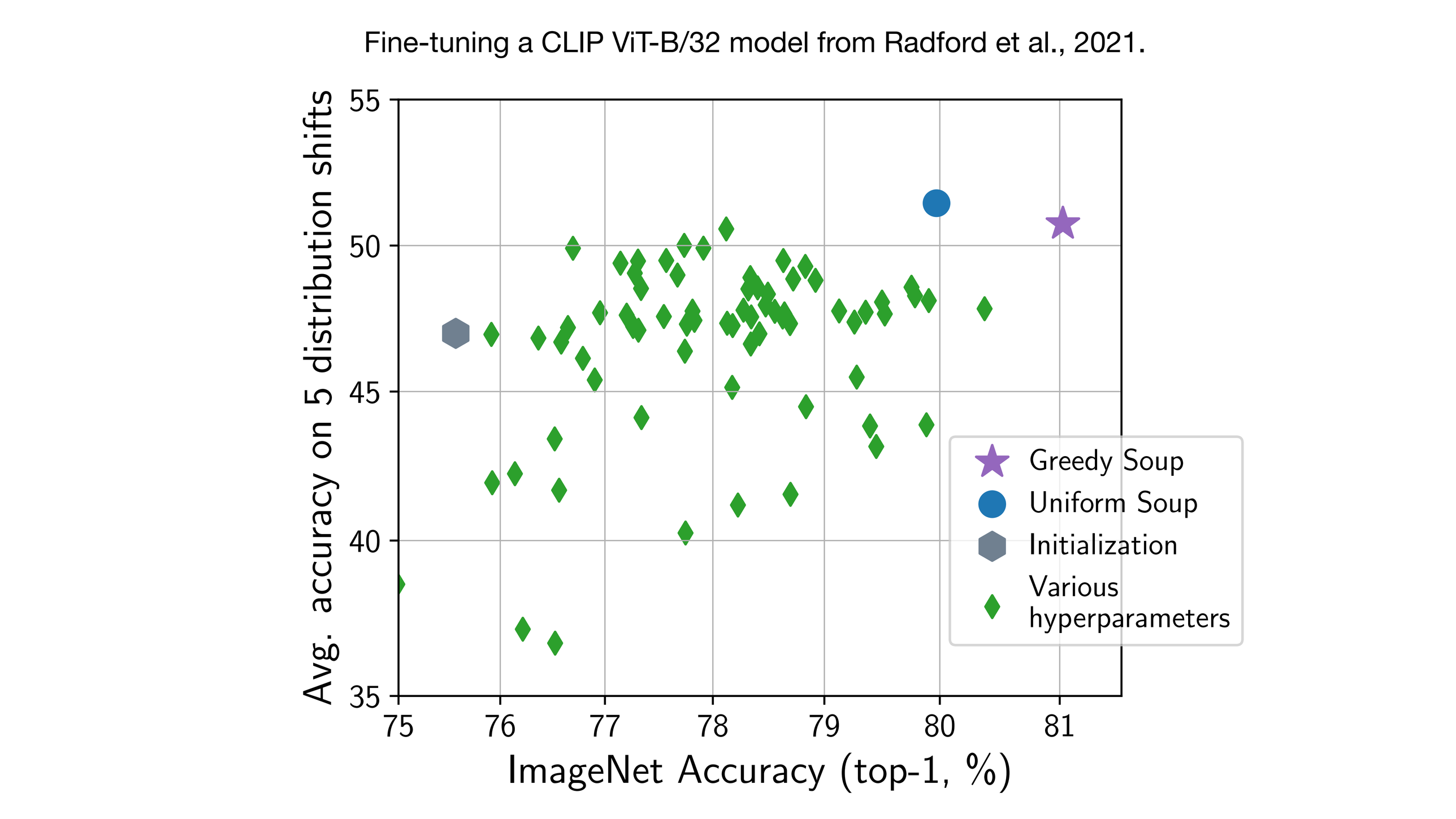

ImageNet test error
Avg. error on 5 distribution shifts
ImageNet test error
Avg. error on 5 distribution shifts
θ_0
> 89.4
89.4
54.2
37.2
29.1
25.2
23.3
22.4
21.9
21.5
> 92.2
92.2
69.4
57.9
52.1
49.2
47.7
47.0
46.6
46.2
Initialization
LR $= 3 \cdot 10^{-5}$ (seed 0)
LR $= 3 \cdot 10^{-5}$ (seed 1)
LR $= 3 \cdot 10^{-6}$

Greedy soup

1. Sort models by decreasing accuracy on the held-out validations set.

2. Add models sequentially to the soup, keeping the model in the soups if the held-out validation accuracy does not decrease




Fine-tuning a CLIP ViT-B/32 model from Radford et al., 2021.
Avg. accuracy on 5 distribution shifts
ImageNet Accuracy (top-1, %)
55
50
45
40
35
75
76
77
78
79
80
81
Greedy Soup
Uniform Soup
Initialization
Various hyperparameters

Method	ImageNet acc. (top-1, %)	Distribution shifts
ViT-G (Zhai et al., 2021)	90.45	–
CoAtNet-7 (Dai et al., 2021)	90.88	–
Our models/evaluations based on ViT-G:		
ViT-G (reevaluated)	90.47	82.06
Best model in hyperparam search	90.78	84.68
Greedy soup	**90.94**	**85.02**

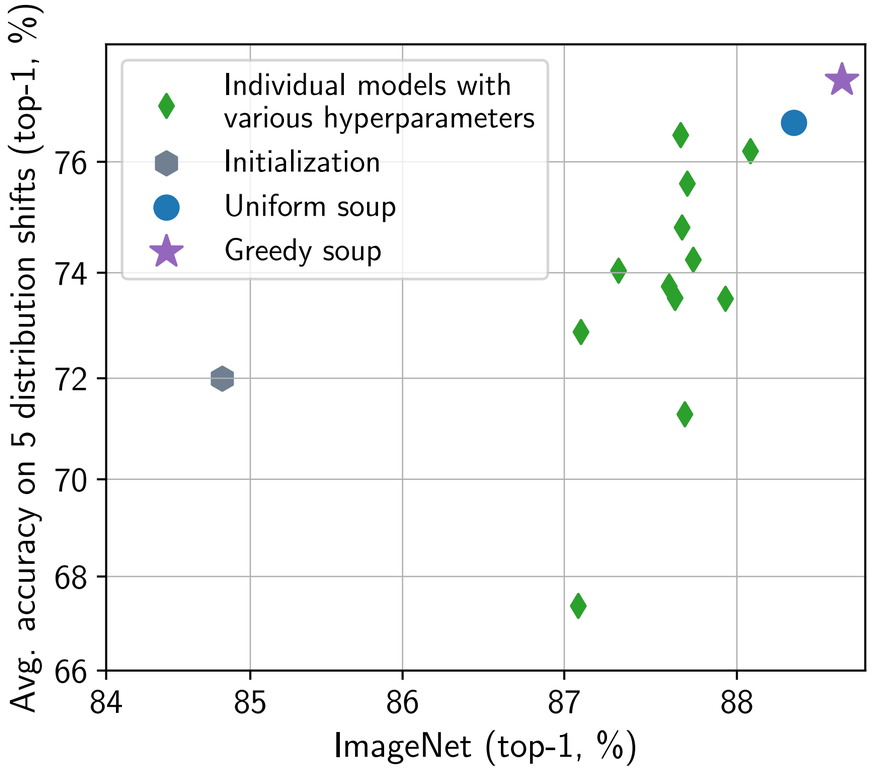
Related work

- Models remain in low loss basin during fine-tuning (Neyshabur et al., 2020; see also Frankle et al., 2020; Nagarajan et al., 2019).
- Weight averaging along an individual optimization trajectory produces high accuracy models (Izmailov et al., 2018; Szegedy et al., 2016).
- Much more (see paper!)

Koh et al., 2021

WILDS

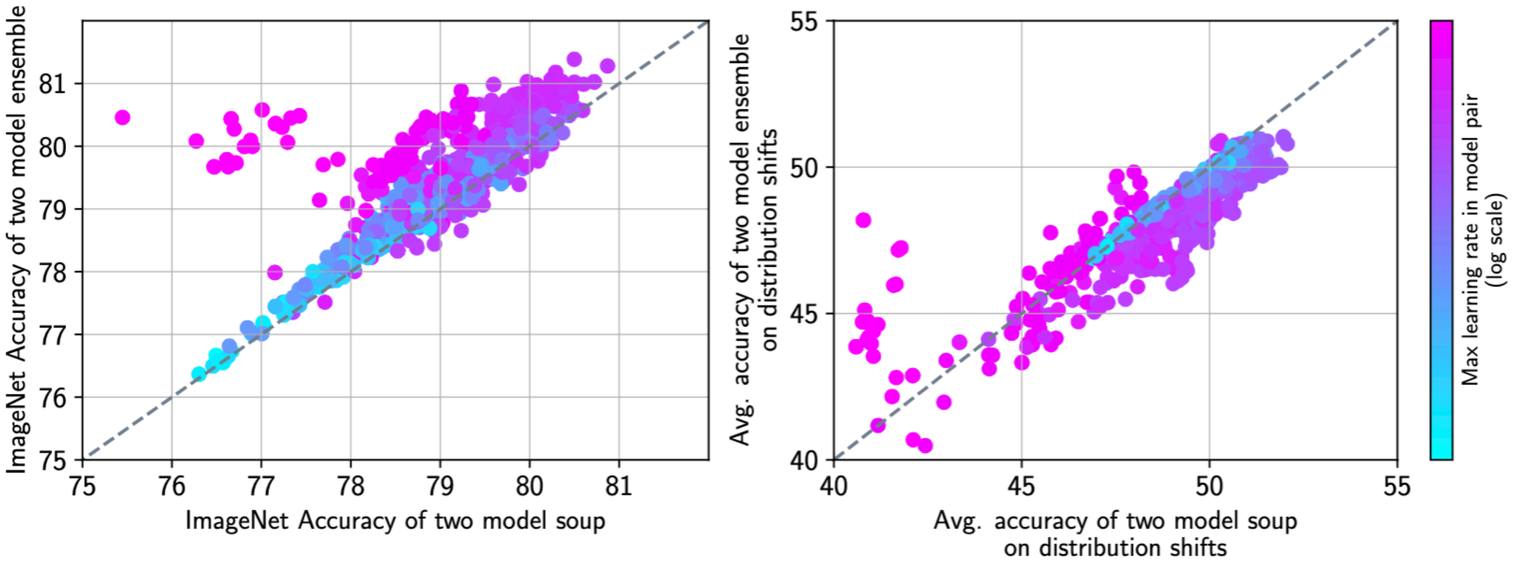
ALIGN (Jia et al., 2021)



Fine-tuning BASIC-L (Pham et al., 2022)

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	–	–	–	–	–	–	–	–
BASIC-L (zero-shot) (Pham et al., 2021)	85.7	–	–	80.6	95.7	76.1	82.3	85.6	84.06
CoCa (zero-shot) (Yu et al., 2022)	86.3	–	–	80.7	96.5	77.6	82.7	90.2	85.54
CoCa (fine-tuned) (Yu et al., 2022)	91.0	–	–	–	–	–	–	–	–
ViT-G/14 greedy soup (Table 4)	90.94	91.20	97.17	84.22	95.46	74.23	78.52	92.67	85.02
Our models/evaluations with fine-tuned BASIC-L:									
Best model on held out val set	90.83	90.84	98.16	**84.42**	95.50	76.98	78.09	93.13	85.63
Greedy ensemble	**91.02**	**91.11**	**98.46**	**84.65**	95.79	76.63	**79.91**	**94.05**	86.20
Greedy soup	**90.98**	91.03	98.37	**84.63**	**96.10**	**77.18**	**79.94**	**94.17**	**86.40**
Best model on each test set (oracle)	90.87	91.24	98.41	84.84	95.89	77.30	80.94	94.47	86.54

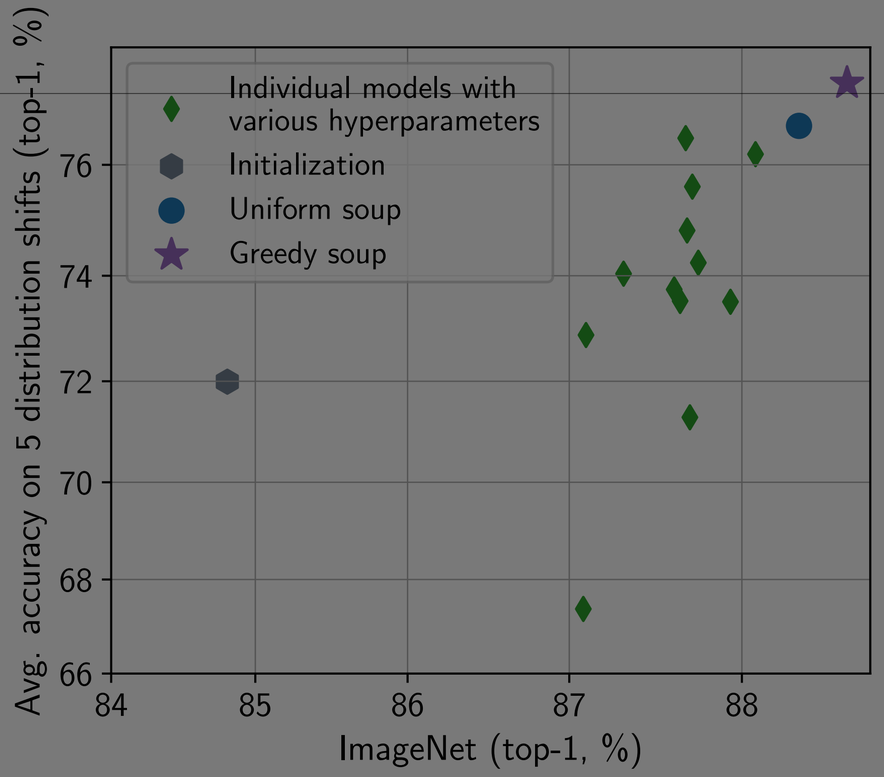
Analytically comparing model soups and ensembles



Koh et al., 2021

WILDS

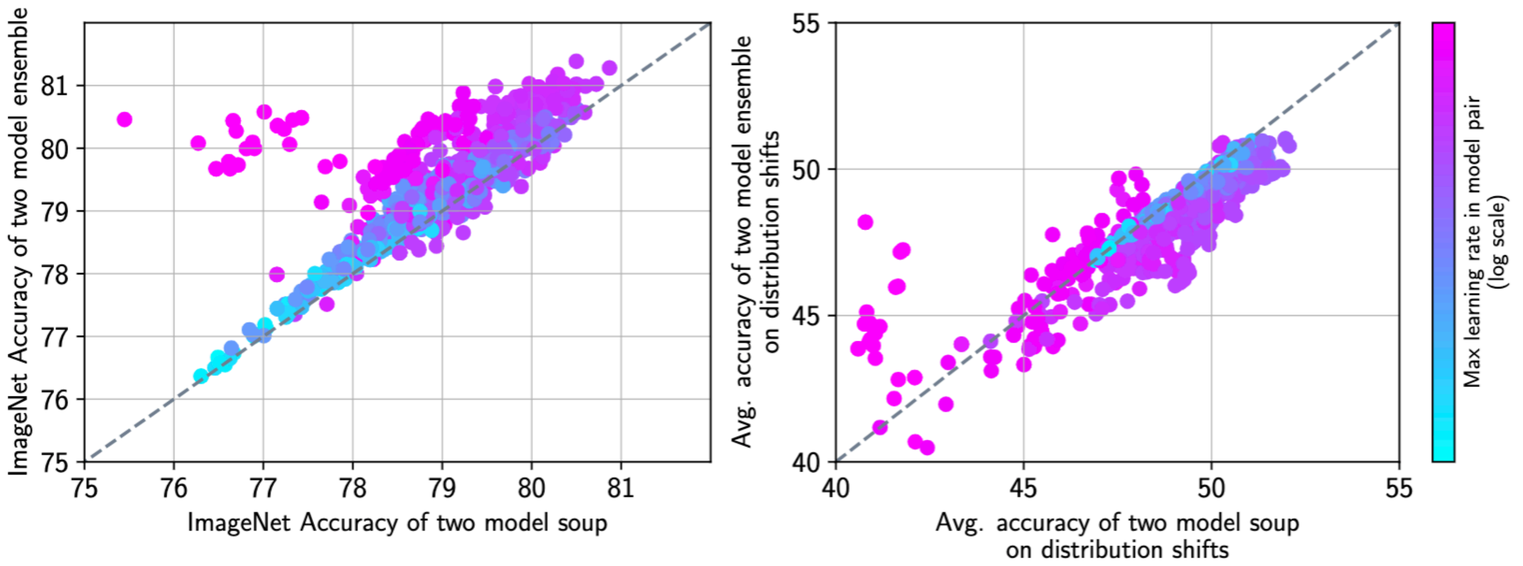
ALIGN (Jia et al., 2021)



Fine-tuning BASIC-L (Pham et al., 2022)

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	–	–	–	–	–	–	–	–
BASIC-L (zero-shot) (Pham et al., 2021)	85.7	–	–	80.6	95.7	76.1	82.3	85.6	84.06
CoCa (zero-shot) (Yu et al., 2022)	86.3	–	–	80.7	96.5	77.6	82.7	90.2	85.54
CoCa (fine-tuned) (Yu et al., 2022)	91.0	–	–	–	–	–	–	–	–
ViT-G/14 greedy soup (Table 4)	90.94	91.20	97.17	84.22	95.46	74.23	78.52	92.67	85.02
Our models/evaluations with fine-tuned BASIC-L:									
Best model on held out val set	90.83	90.84	98.16	**84.42**	95.50	76.98	78.09	93.13	85.63
Greedy ensemble	**91.02**	**91.11**	**98.46**	**84.65**	95.79	76.63	**79.91**	**94.05**	86.20
Greedy soup	**90.98**	91.03	98.37	**84.63**	**96.10**	**77.18**	**79.94**	**94.17**	**86.40**
Best model on each test set (oracle)	90.87	91.24	98.41	84.84	95.89	77.30	80.94	94.47	86.54

Analytically comparing model soups and ensembles



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

Mitchell Wortsman, Gabriel Ilharco, Samir Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes,
Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon*, Simon Kornblith*, Ludwig Schmidt*

Google Brain