Robust fine-tuning of zero-shot models

Колодезный Александр БПМИ192

Национальный исследовательский университет
«Высшая школа экономики» (Москва)

30 ноября 2022 г.

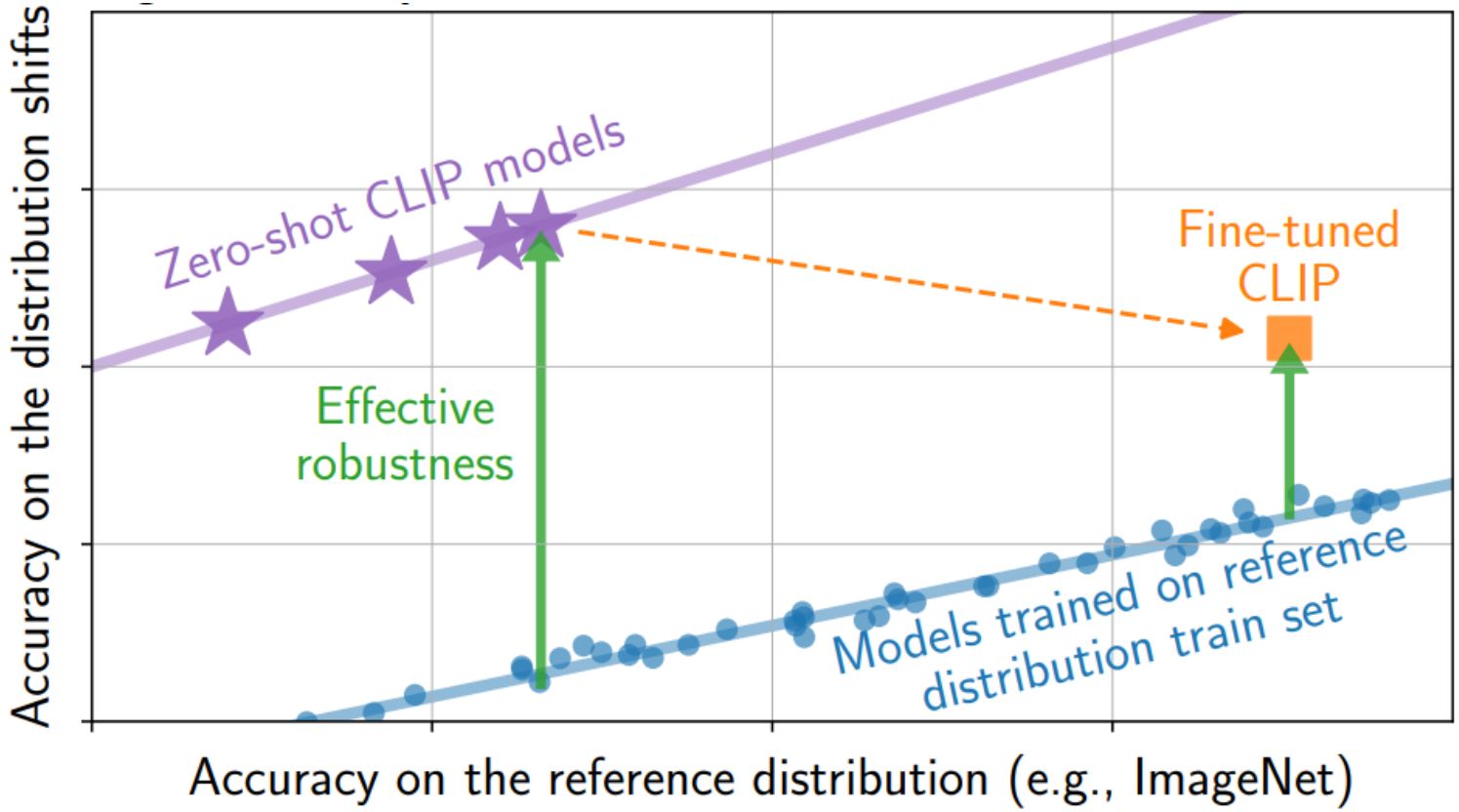
Постановка проблемы

- Есть задача классификации
- Есть целевые данные, на которых обучаем и тестируем модель
- При применении модели реальные данные могут отличаться от целевых (distribution shift)
- Хотим увеличивать качество на целевых данных, но быть устойчивыми к сдвигу распределения (robustness)

Известные модели

- Zero-shot модели (CLIP) обладают хорошей устойчивостью к сдвигам распределения, но при этом недостаточное качество на целевых данных
- Если дообучить модель - качество на целевых данных улучшится, но устойчивость упадёт

Имеющиеся модели



Используемые наборы данных

- Целевые данные — ImageNet

ImageNet (Deng et al.)

ImageNetV2 (Recht et al.)

ImageNet-R (Hendrycks et al.)

ImageNet Sketch (Wang et al.)

ObjectNet (Barbu et al.)

ImageNet-A (Hendrycks et al.)

Zero-shot CLIP

- CLIP берут предобученный на основе ViT-L
- CLIP считает соответствие картинки и текста как $\langle g(x_i), h(s_j) \rangle$, тогда классификация работает как

$$argmax_j \langle g(x_i), h(s_j) \rangle$$

- В качестве текстов используют 80 промптов
  - "a bad photo of a ..."
  - "a photo of many ..."
  - ...
- W_{zero_shot} получают усреденением этих текстовых эмбедингов

Fine-tune CLIP

- Дообучают на лосс

$$\sum_{(x,y)\in\mathcal{S}^{tr}_{ref}} CE(f(x_i,\Theta),y_i) + \lambda R(\Theta) \to \min$$

- Можно дообучать как все веса (end-to-end), так и только классификационную голову

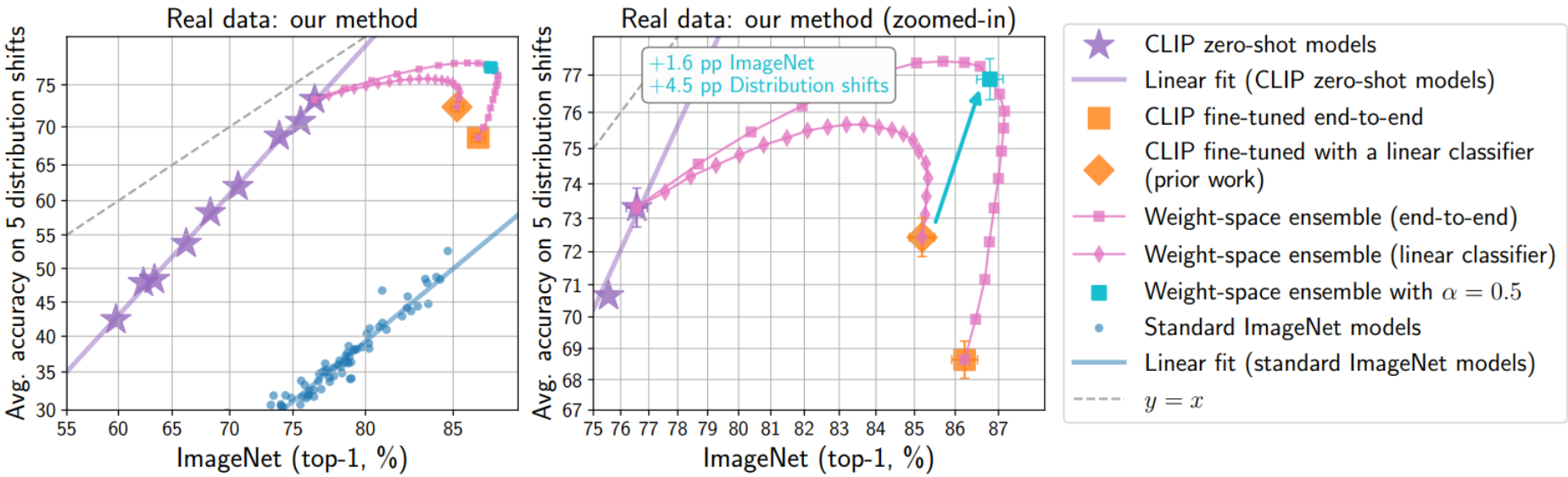
Weight-space ensambling

- Давайте усреднять веса Zero-shot и fine-tuned моделей

$$wse(x, \alpha) = f(x, (1 - \alpha)\Theta_0 + \alpha\Theta_1)$$

- По умолчанию берём $\alpha = 0.5$

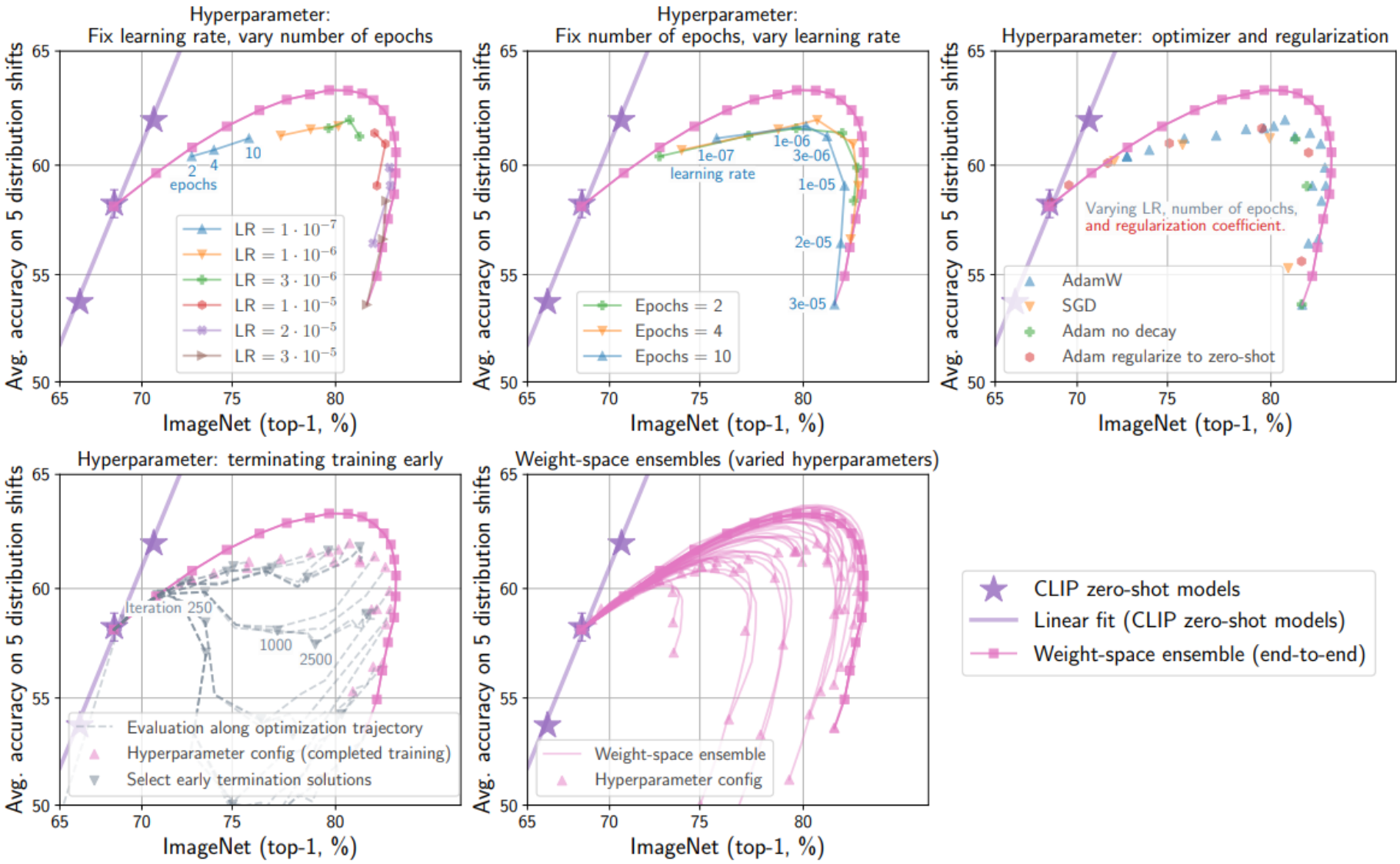
Результаты


Real data: our method
Avg. accuracy on 5 distribution shifts
ImageNet (top-1, %)
Real data: our method (zoomed-in)
+1.6 pp ImageNet
+4.5 pp Distribution shifts
CLIP zero-shot models
Linear fit (CLIP zero-shot models)
CLIP fine-tuned end-to-end
CLIP fine-tuned with a linear classifier (prior work)
Weight-space ensemble (end-to-end)
Weight-space ensemble (linear classifier)
Weight-space ensemble with $\alpha = 0.5$
Standard ImageNet models
Linear fit (standard ImageNet models)
$y = x$

Результаты

	IN (reference)	Distribution shifts					Avg shifts	Avg ref., shifts
		IN-V2	IN-R	IN-Sketch	ObjectNet*	IN-A		
CLIP `ViT-L/14@336px`								
Zero-shot [82]	76.2	70.1	88.9	60.2	70.0	77.2	73.3	74.8
Fine-tuned LC [82]	85.4	75.9	84.2	57.4	66.2	75.3	71.8	78.6
Zero-shot (PyTorch)	76.6	70.5	89.0	60.9	69.1	77.7	73.4	75.0
Fine-tuned LC (ours)	85.2	75.8	85.3	58.7	67.2	76.1	72.6	78.9
Fine-tuned E2E (ours)	86.2	76.8	79.8	57.9	63.3	65.4	68.6	77.4
WiSE-FT (ours)								
LC, α=0.5	83.7	76.3	89.6	63.0	70.7	79.7	75.9	79.8
LC, optimal α	85.3	76.9	89.8	63.0	70.7	79.7	75.9	80.2
E2E, α=0.5	86.8	**79.5**	89.4	64.7	71.1	79.9	76.9	81.8
E2E, optimal α	**87.1**	**79.5**	**90.3**	**65.0**	**72.1**	**81.0**	**77.4**	**81.9**

Сравнение с подбором гиперпараметров



Заключения

1. WiSE-FT позволяет дообучать модель с сохранинем устойчивости
2. Дообученная модель и zero-shot модель соединены линейной траекторией с низкой ошибкой
3. WiSE-FT позволяет получить ошибку ниже чем ошибка и fine-tuned, и zero-shot моделей на целевом наборе данных