Тестирование моделей

ВШЭ ФКН, Методы предобучения без учителя

Шабалин Александр

Probing

- Нужен для того, чтобы узнать, какая скрытая информация содержится в представлениях обученной модели

- Используется чаще для текстовых моделей, потому что текст интереснее

king - queen = man - woman

Probing для текстов

3 : first + 1 : second + 1 : third + 6 : fourth

3 : fourth-to-last + 1 : third-to-last + 1 : second-to-last + 6 : last

3 : (0,3) + 1 : (1,2) + 1 : (2,1) + 6 : (3,0)

3 : #_1 + 1 : 3_1 + 1 : 1_6 + 6 : 1_#

3 : L + 1 : RLL + 1 : RLR + 6 : RR

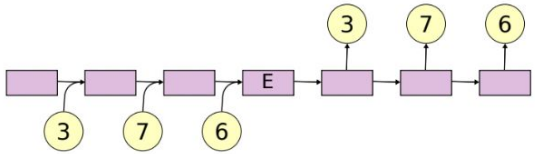
3 : r_0 + 1 : r_0 + 1 : r_0 + 6 : r_0

	3	1	1	6	5	2	3	1	9	7
Left-to-right	0	1	2	3	0	1	2	3	4	5
Right-to-left	3	2	1	0	5	4	3	2	1	0
Bidirectional	(0, 3)	(1, 2)	(2, 1)	(3, 0)	(0, 5)	(1, 4)	(2, 3)	(3, 2)	(4, 1)	(5, 0)
Wickelroles	#_1	3_1	1_6	1_#	#_2	5_3	2_1	3_9	1_7	9_#
Tree	L	RLL	RLR	RR	LL	LRLL	LRLR	LRRL	LRRR	R
Bag of words	r_0	r_0	r_0	r_0	r_0	r_0	r_0	r_0	r_0	r_0

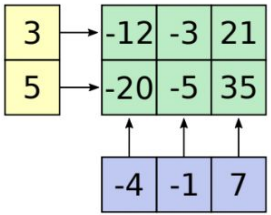
https://arxiv.org/abs/1812.08718

Probing для текстов

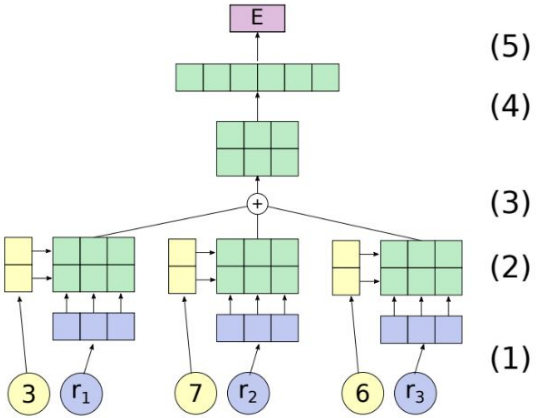
3 : first + 7 : second + 6 : third



(a)

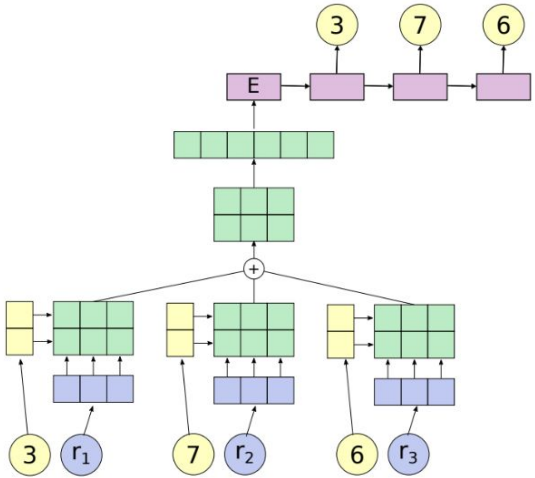


(b)



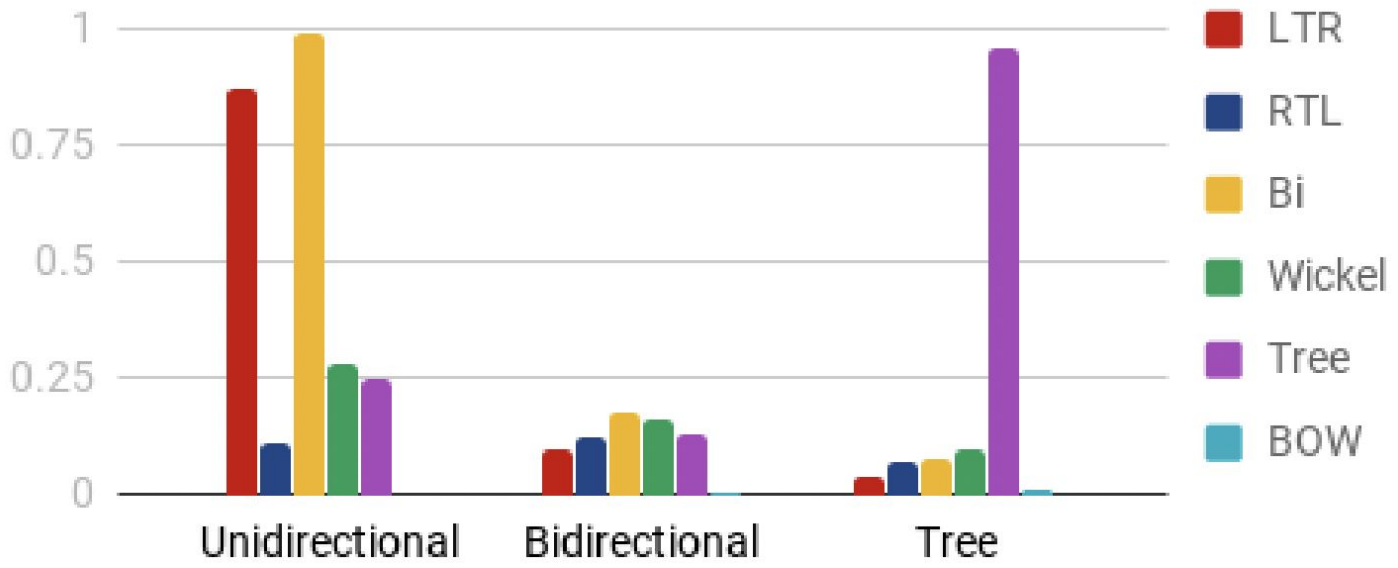
(5)
(4)
(3)
(2)
(1)

(c)



(d)

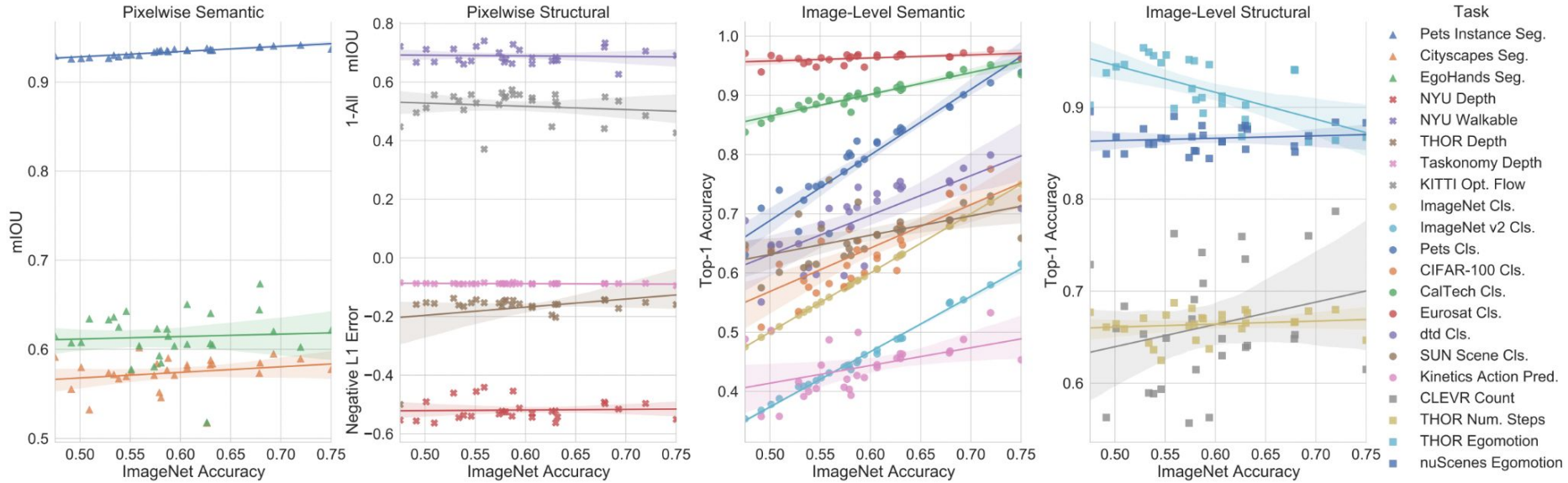
Probing для текстов



Как измерить качество модели?

- Попробовать поатаковать модель (добавление шума, маскирование и тд)
- Обучить голову на ImageNet и замерить точность
- Обучить голову на **подмножестве** ImageNet и замерить точность
- Обучить на downstream задачу (transfer learning)

Почему мерить качество на ImageNet не лучшая идея



https://arxiv.org/abs/2103.14005

Почему мерить качество на ImageNet не лучшая идея

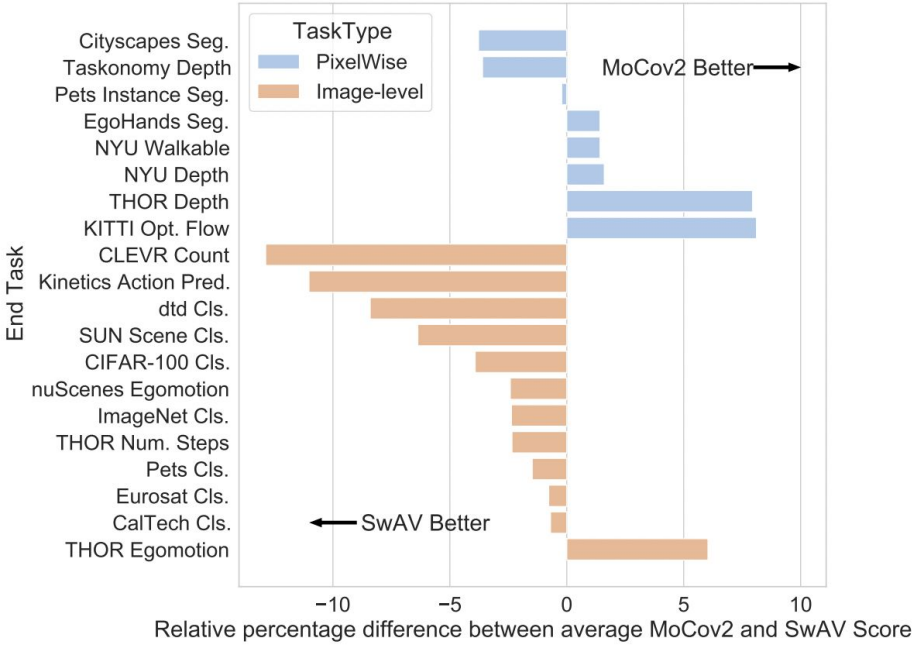
Table 3: **Training in small batch setting.** Top-1 accuracy on ImageNet with a linear classifier trained on top of frozen features from a ResNet-50. All methods are trained with a batch size of 256. We also report the number of stored features, the type of cropping used and the number of epochs.

Method	Mom. Encoder	Stored Features	`multi-crop`	epoch	batch	Top-1
SimCLR		0	2×224	200	256	61.9
MoCov2	✓	65,536	2×224	200	256	67.5
MoCov2	✓	65,536	2×224	800	256	71.1
SwAV		3,840	$2\times160+4\times96$	200	256	72.0
SwAV		3,840	$2\times224+6\times96$	200	256	72.7
SwAV		3,840	$2\times224+6\times96$	400	256	**74.3**

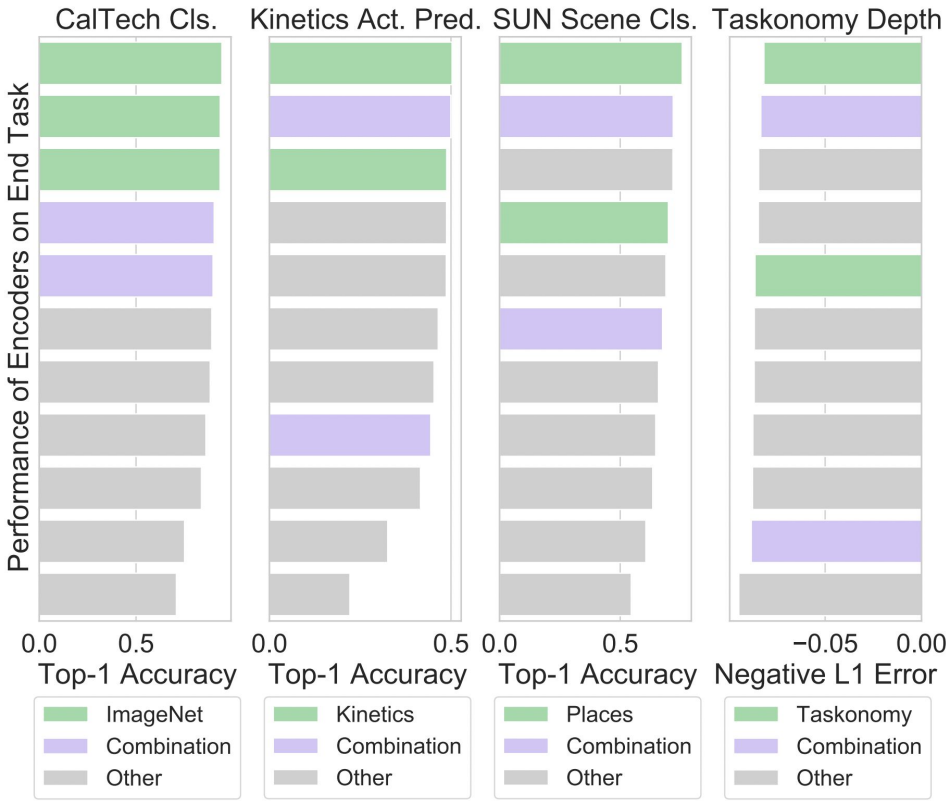
Почему мерить качество на ImageNet не лучшая идея

Table 3: **Training in small batch setting.** Top-1 accuracy on ImageNet with a linear classifier trained on top of frozen features from a ResNet-50. All methods are trained with a batch size of 256. We also report the number of stored features, the type of cropping used and the number of epochs.

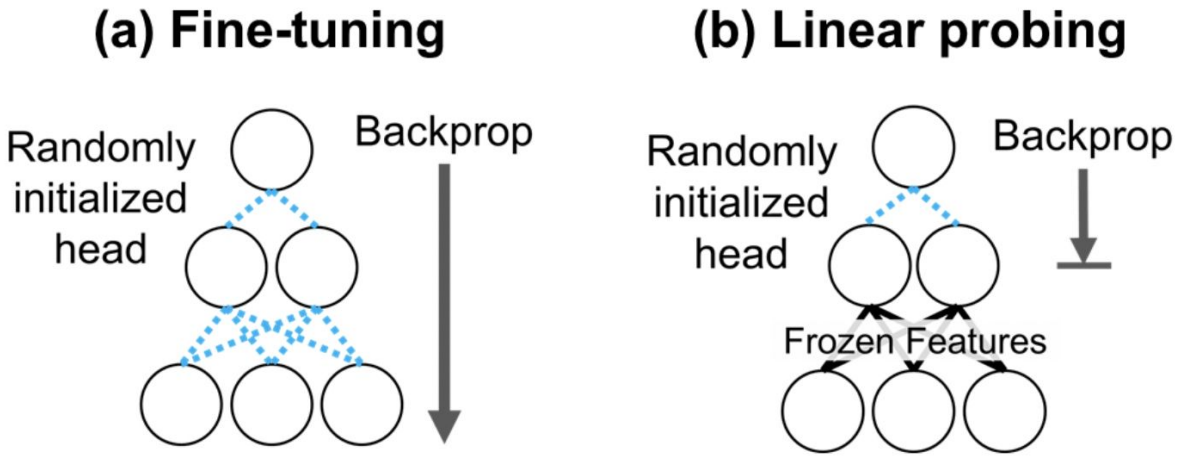
Method	Mom.	Encoder	Stored Features	`multi-crop`	epoch	batch	Top-1
SimCLR			0	2×224	200	256	61.9
MoCov2	✓		65,536	2×224	200	256	67.5
MoCov2	✓		65,536	2×224	800	256	71.1
SwAV			3,840	$2\times160+4\times96$	200	256	72.0
SwAV			3,840	$2\times224+6\times96$	200	256	72.7
SwAV			3,840	$2\times224+6\times96$	400	256	**74.3**



Учиться надо на наиболее похожем датасете



Как дообучать модели?



Как дообучать модели?

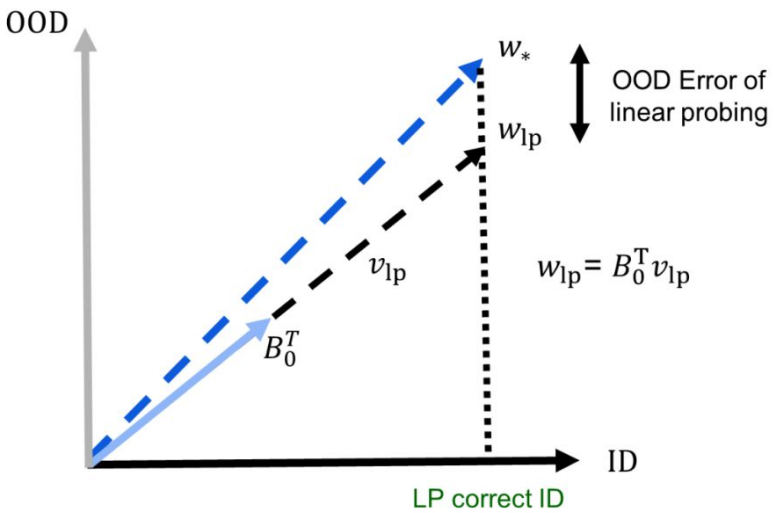
$$\widehat{L}(v,B) = \|XB^\top v - Y\|_2^2$$

$$\nabla_B \widehat{L}(v,B) = 2v(Y - XB^\top v)^\top X$$

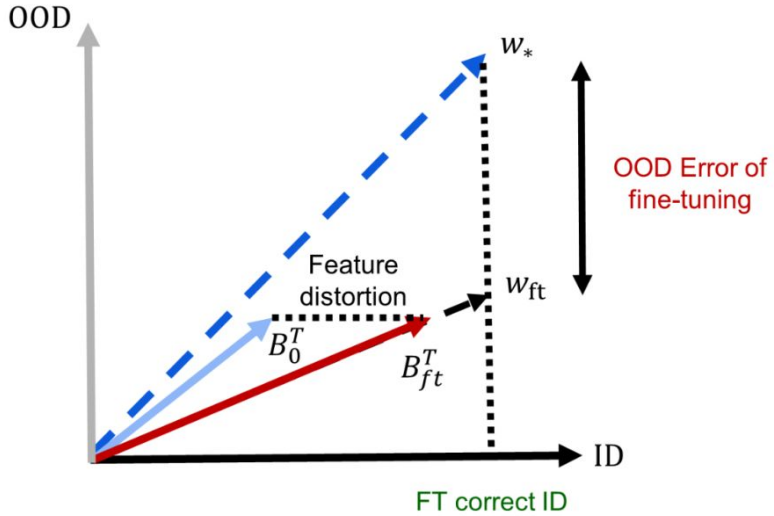
$$\nabla_B \widehat{L}(v,B)u = 0$$

$$B^{k+1}u = \left(B^k - \eta \nabla_{B^k} \hat{L}(v, B^k)\right)u = B^k u$$

Linear Probing vs Fine-tuning



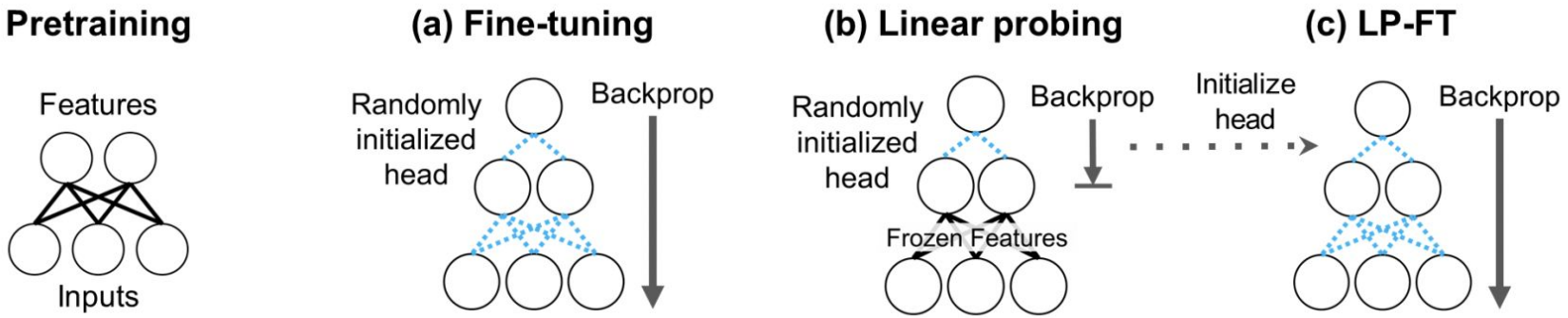
(a) Toy example (Linear probing)



(b) Toy example (fine-tuning)

https://arxiv.org/abs/2202.10054

Linear Probing vs Fine-tuning



ID test	85.1%	82.9%	**85.7%**
OOD test	59.3%	66.2%	**68.9%**

Average accuracies (10 distribution shifts)

https://arxiv.org/abs/2202.10054

Table 1: **ID accuracies** with 90% confidence intervals over 3 runs—fine-tuning does better than linear probing on all datasets except DomainNet (which could be because the version of the DomainNet training dataset from Tan et al. (2020) is fairly small, with around 20K examples). LP-FT does the best on all except FMoW where it is in between linear probing and fine-tuning.

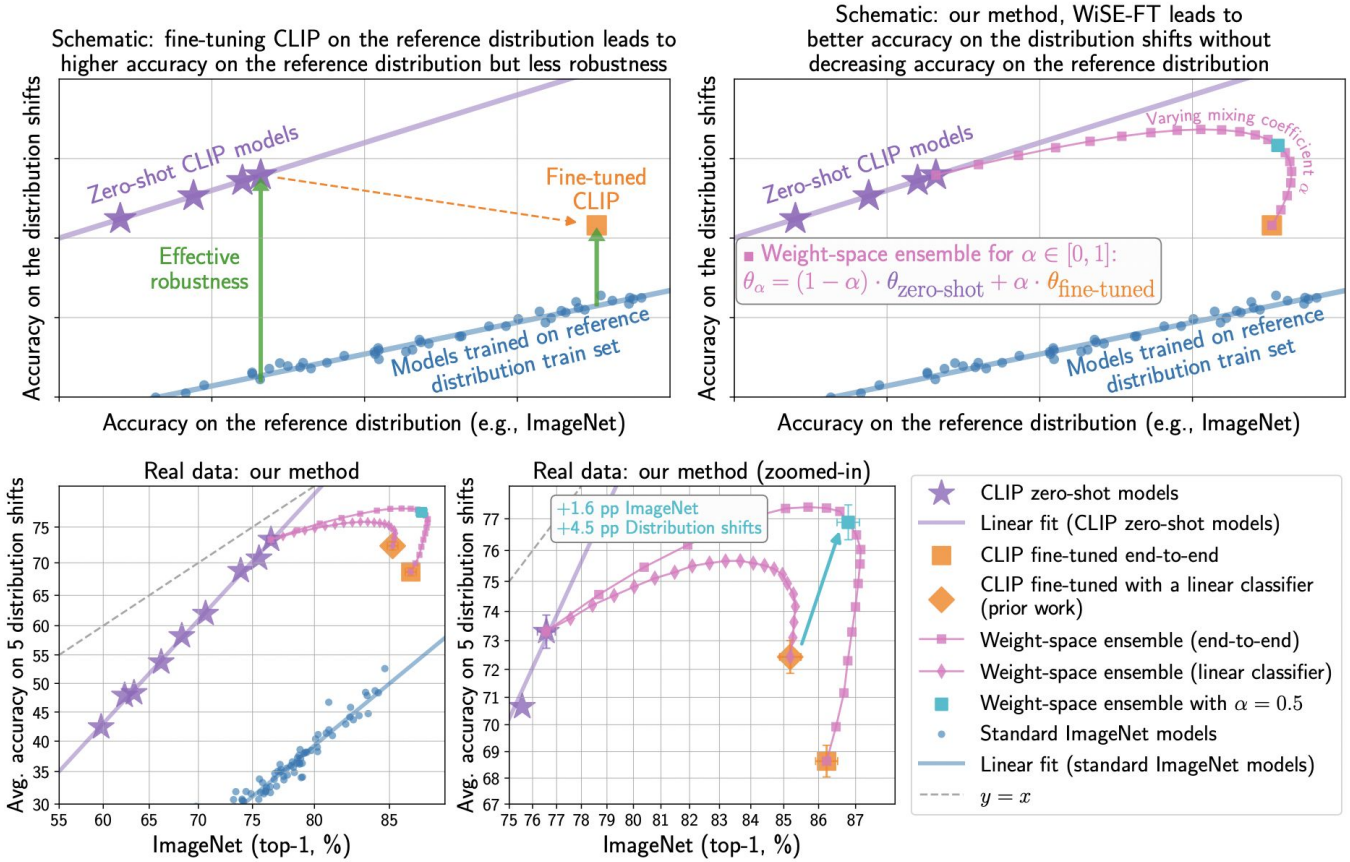
	CIFAR-10	Ent-30	Liv-17	DomainNet	FMoW	ImageNet	Average
FT	**97.3 (0.2)**	**93.6 (0.2)**	97.1 (0.2)	84.5 (0.6)	**56.5 (0.3)**	**81.7 (-)**	85.1
LP	91.8 (0.0)	90.6 (0.2)	96.5 (0.2)	89.4 (0.1)	49.1 (0.0)	79.7 (-)	82.9
LP-FT	**97.5 (0.1)**	**93.7 (0.1)**	**97.8 (0.2)**	**91.6 (0.0)**	51.8 (0.2)	**81.7 (-)**	**85.7**

Table 2: **OOD accuracies** with 90% confidence intervals over 3 runs. Linear probing does better than fine-tuning on all datasets except CIFAR-10.1 and ImageNetV2, where the ID and OOD are similar (consistent with our theory). LP-FT does the best on all 10 datasets.

	STL	CIFAR-10.1	Ent-30	Liv-17	DomainNet	FMoW
FT	82.4 (0.4)	92.3 (0.4)	60.7 (0.2)	77.8 (0.7)	55.5 (2.2)	32.0 (3.5)
LP	85.1 (0.2)	82.7 (0.2)	**63.2 (1.3)**	82.2 (0.2)	79.7 (0.6)	**36.6 (0.0)**
LP-FT	**90.7 (0.3)**	**93.5 (0.1)**	**62.3 (0.9)**	**82.6 (0.3)**	**80.7 (0.9)**	**36.8 (1.3)**

	ImNetV2	ImNet-R	ImNet-Sk	ImNet-A	Average
FT	**71.5 (-)**	52.4 (-)	40.5 (-)	27.8 (-)	59.3
LP	69.7 (-)	70.6 (-)	46.4 (-)	45.7 (-)	66.2
LP-FT	**71.6 (-)**	**72.9 (-)**	**48.4 (-)**	**49.1 (-)**	**68.9**

Как дообучить лучше?



https://arxiv.org/abs/2109.01903