



Robust fine-tuning of zero-shot models

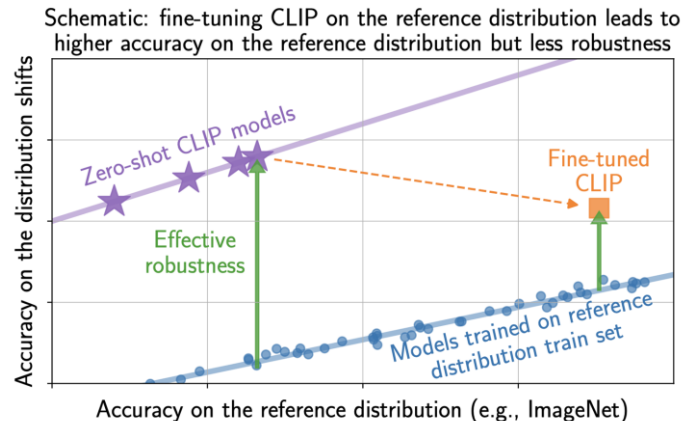
Обзор-рецензия

Основная идея

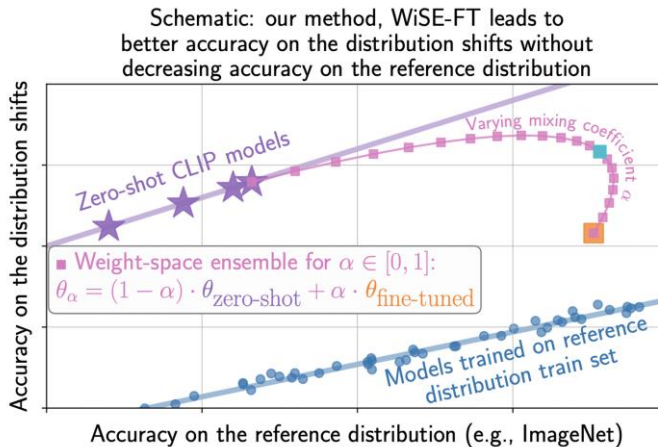
Can zero-shot models be fine-tuned without reducing accuracy under distribution shift?

- Раньше при дообучении zero-shot нейросетей их “устойчивость” (качество на других наборах данных) снижалась
- При дообучении новым способом она повышается

Было:



Стало
:



Про статью

- Первая версия – сентябрь 2021
- Финальная – июнь 2022, для конференции CVPR 2022
- Статья-финалист конкурса на лучшую работу

Robust fine-tuning of zero-shot models

Mitchell Wortsman*[†] Gabriel Ilharco*[†] Jong Wook Kim[§] Mike Li[‡]

Simon Kornblith[°] Rebecca Roelofs[°] Raphael Gontijo-Lopes[°]

Hannaneh Hajishirzi^{†°} Ali Farhadi^{†‡} Hongseok Namkoong*[‡] Ludwig Schmidt^{†△}

Abstract

Large pre-trained models such as CLIP or ALIGN offer consistent accuracy across a range of data distributions when performing zero-shot inference (i.e., without fine-tuning on a specific dataset). Although existing fine-tuning methods substantially improve accuracy on a given target distribution, they often reduce robustness to distribution shifts. We address this tension by introducing a simple and effective method for improving robustness while fine-tuning: ensembling the weights of the zero-shot and fine-tuned models (WiSE-FT). Compared to standard fine-tuning, WiSE-FT provides large accuracy improvements under distribution shift, while preserving high accuracy on the target distribution. On ImageNet and five derived distribution shifts, WiSE-FT improves accuracy under distribution shift by 4 to 6 percentage points (pp) over prior work while increasing ImageNet accuracy by 1.6 pp. WiSE-FT achieves similarly large robustness gains (2 to 23 pp) on a diverse set of six further distribution shifts, and accuracy gains of 0.8 to 3.3 pp compared to standard fine-tuning on seven commonly used transfer learning datasets. These improvements come at no additional computational cost during fine-tuning or inference.



Mitchell Wortsman



Gabriel Ilharco

Про статью: предыдущие работы авторов

Robust fine-tuning of zero-shot models

Mitchell Wortsman*, Gabriel Ilharco*, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, Ludwig Schmidt
CVPR, 2022 (oral, best paper finalist)
[arxiv](#) / [code](#)

OpenCLIP: An open source implementation of CLIP

Gabriel Ilharco*, **Mitchell Wortsman***, Ross Wightman*, Cade Gordon*, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, Ludwig Schmidt
GitHub, 2021

Learning Neural Network Subspaces

Mitchell Wortsman, Maxwell Horton, Carlos Guestrin, Ali Farhadi, Mohammad Rastegari
ICML, 2021
[arxiv](#) / [code](#)

Supermasks in Superposition

Mitchell Wortsman*, Vivek Ramanujan*, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, Ali Farhadi
NeurIPS, 2020
[arxiv](#) / [code](#)

Soft Threshold Weight Reparameterization for Learnable Sparsity

Aditya Kusupati, Raghav Somani*, Vivek Ramanujan*, **Mitchell Wortsman***, Prateek Jain, Sham Kakade, Ali Farhadi
ICML, 2020
[arxiv](#) / [code](#)

What's Hidden in a Randomly Weighted Neural Network?

Vivek Ramanujan*, **Mitchell Wortsman***, Aniruddha Kembhavi, Ali Farhadi, Mohammad Rastegari
CVPR, 2020
[arxiv](#) / [code](#)

Discovering Neural Wirings

Mitchell Wortsman, Ali Farhadi, Mohammad Rastegari
NeurIPS, 2019

- 2022.06: I'll be presenting [Robust fine-tuning of zero-shot models](#) at CVPR 2022, come say hi!
- 2022.05: I'm very thankful for the recognition of outstanding reviewer at CVPR 2022
- 2022.04: Our [open-source repository](#) for training CLIP models has reached 1000 stars!
- 2022.03: [Model soups](#) set a new state-of-the-art on ImageNet
- 2022.03: What makes zero-shot CLIP models robust? Find out [here](#)
- 2022.03: Check out our [work](#) using CLIP for zero-shot object navigation.
- 2021.10: I'm excited to be starting an internship with [Jacob Eisenstein](#) at Google Research
- 2021.09: Check out our new work [Robust fine-tuning of zero-shot models](#)!
- 2021.08: I'm very thankful for the recognition of outstanding reviewer at ACL 2021
- 2021.06: Our [paper](#) exploring the relation between visual and text representations has been accepted to NAACL
- 2021.04: Our [MultiModalQA](#) paper on complex QA over text, tables and images is out
- 2021.03: Our [paper](#) on contrastive representation learning is out
- 2020.11: We'll be presenting a [tutorial on High-Performance NLP](#) at EMNLP 2020
- 2020.06: I'm very excited to be starting an internship with [Peter Anderson](#) and [Ashish Vaswani](#) at Google Research
- 2020.05: Check out [our new preprint](#) exploring similarities between vision and language representations
- 2020.04: Our work [Evaluating NLP models via contrast sets](#) is out
- 2020.02: Check out our [new paper](#) exploring the dynamics of fine-tuning in NLP
- 2020.01: Our paper [Toward ML-Centric Cloud Platforms](#) made the cover of the Communications of the ACM
- 2019.12: Don't miss our spotlight presentation on [SDTW](#) at [ViGIL](#), NeurIPS 2019.
- 2019.11: Our [CoNLL 2019 paper](#) was awarded Honorable Mention for Best Paper in Research Inspired by Human Language Learning!

Mitchell Wortsman

Gabriel Ilharco

Про статью: предыдущие работы авторов

[Robust fine-tuning of zero-shot models](#)

Mitchell Wortsman*, Gabriel Ilharco*, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, Ludwig Schmidt
CVPR, 2022 (oral, best paper finalist)
[arxiv](#) / [code](#)

[OpenCLIP: An open source implementation of CLIP](#)

Gabriel Ilharco*, **Mitchell Wortsman***, Ross Wightman*, Cade Gordon*, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, Ludwig Schmidt
[GitHub](#), 2021

[Learning Neural Network Subspaces](#)

Mitchell Wortsman, Maxwell Horton, Carlos Guestrin, Ali Farhadi, Mohammad Rastegari
ICML, 2021
[arxiv](#) / [code](#)

[Supermasks in Superposition](#)

- 2022.06: I'll be presenting [Robust fine-tuning of zero-shot models](#) at CVPR 2022, come say hi!
- 2022.05: I'm very thankful for the recognition of outstanding reviewer at CVPR 2022
- 2022.04: Our [open-source repository](#) for training CLIP models has reached 1000 stars!
- 2022.03: [Model soups](#) set a new state-of-the-art on ImageNet
- 2022.03: What makes zero-shot CLIP models robust? Find out [here](#)
- 2022.03: Check out our [work](#) using CLIP for zero-shot object navigation.
- 2021.10: I'm excited to be starting an internship with [Jacob Eisenstein](#) at Google Research



mlfoundations / open_clip

Public

Watch 37

Fork 244

Star 2.4k

<> Code

Issues 25

Pull requests 15

Discussions

Actions

Projects

...



Про статью: предыдущие работы авторов

Learning Neural Network Subspaces

Mitchell Wortsman, Maxwell Horton, Carlos Guestrin, Ali Farhadi, Mohammad Rastegari

ICML, 2021

[arxiv](#) / [code](#)

Supermasks in Superposition

Mitchell Wortsman*, Vivek Ramanujan*, Rosanne Liu, Aniruddha Kembhavi, Mohammad Rastegari, Jason Yosinski, Ali Farhadi

NeurIPS, 2020

[arxiv](#) / [code](#)

Soft Threshold Weight Reparameterization for Learnable Sparsity

Aditya Kusupati, Raghav Somani*, Vivek Ramanujan*, **Mitchell Wortsman***, Prateek Jain, Sham Kakade, Ali Farhadi

ICML, 2020

[arxiv](#) / [code](#)

What's Hidden in a Randomly Weighted Neural Network?

Vivek Ramanujan*, **Mitchell Wortsman***, Aniruddha Kembhavi, Ali Farhadi, Mohammad Rastegari

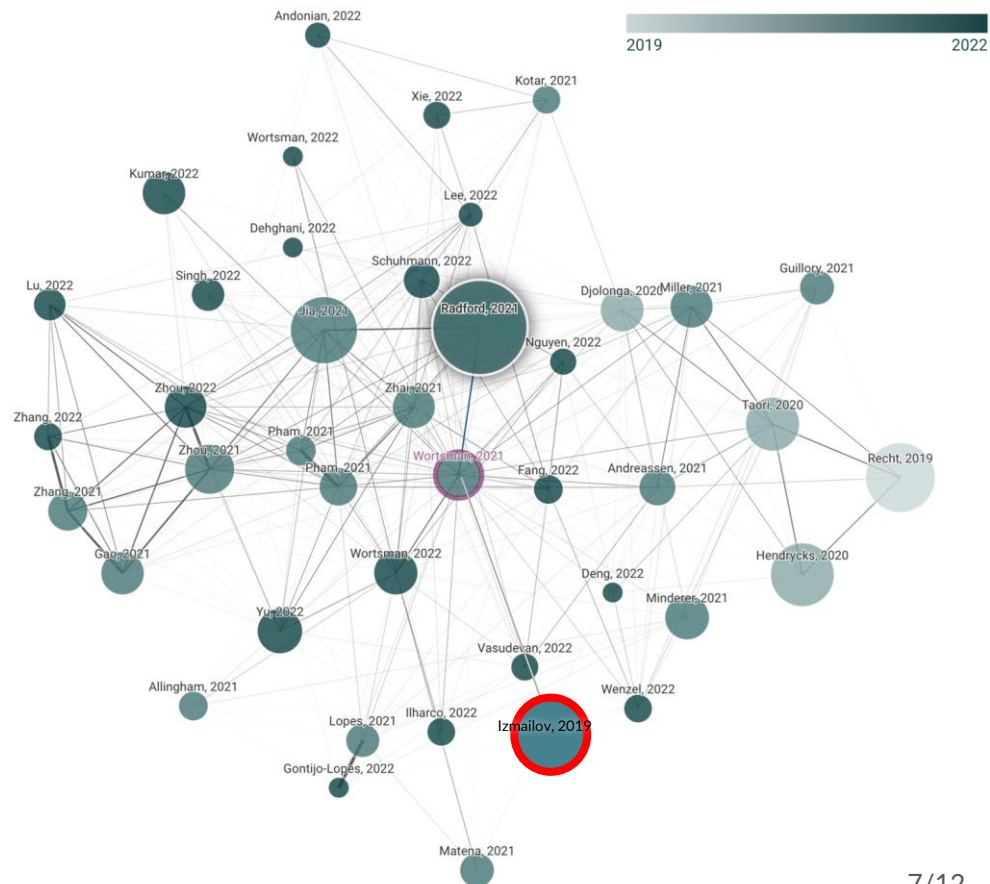
CVPR, 2020

[arxiv](#) / [code](#)

Связанные статьи

Pavel Izmailov et al., Averaging Weights Leads to Wider Optima and Better Generalization

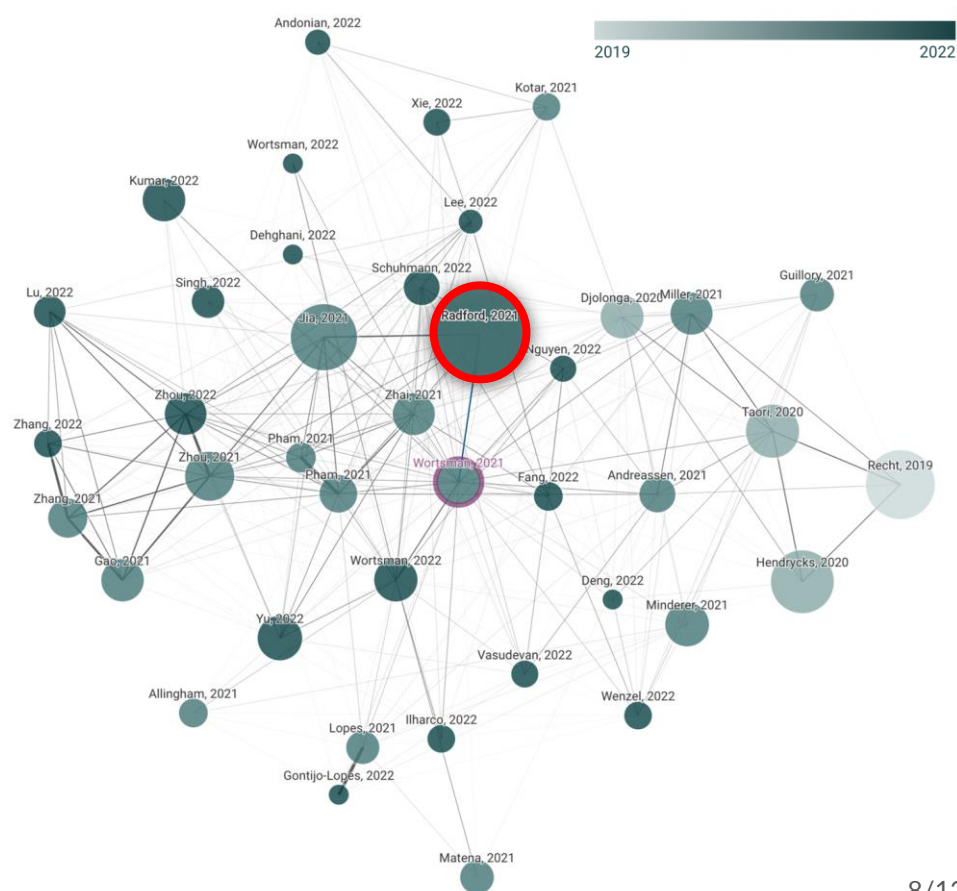
Усреднение весов чекпоинтов моделей для повышения качества



Связанные статьи

Alec Radford et al., Learning transferable visual models from natural language supervision

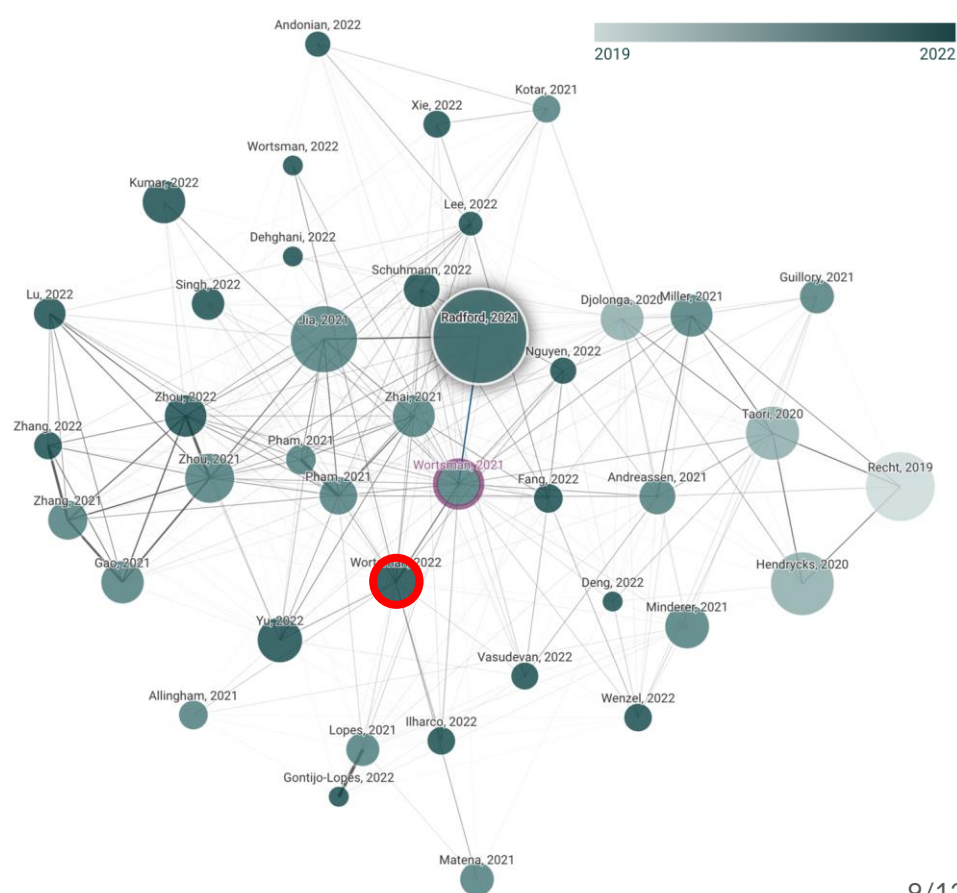
Способ повышения устойчивого дообучения предобученной модели



Связанные статьи

Mitchell Wortsman et al., Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time

Смотрите в следующих сериях!



- Проведено большое число обучений и экспериментов, много моделей
- Приложен код
- Идея несложная, но описано, как к ней пришли авторы, сама идея описана хорошо

Code

Overview

WiSE-FT can be implemented in a few lines of code. See [src/wise_ft.py](#) for more details.

```

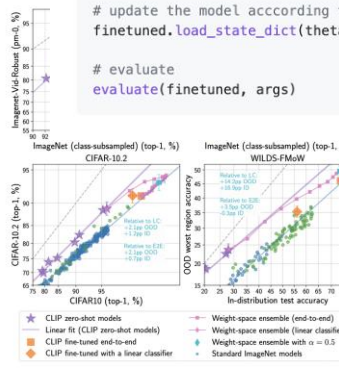
Load models
zeroshot = ImageClassifier.load(zeroshot_checkpoint)
finetuned = ImageClassifier.load(finetuned_checkpoint)

# make sure checkpoints are compatible
assert set(zeroshot.keys()) == set(finetuned.keys())

# interpolate between checkpoints with mixing coefficient alpha
theta = {
    key: (1-alpha) * zeroshot[key] + alpha * finetuned[key]
    for key in zeroshot.keys()
}

# update the model according to the new
finetuned.load_state_dict(theta)

```



F When do weight-space ensembles approximate output-space ensembles?

In practice we observe a difference between weight-space and output-space ensembling. However, it is worth noting that these two methods of ensembling are not as different as they initially appear. In certain regimes a weight-space ensemble approximates the corresponding output-space ensemble—for instance, when training is well approximated by a linear expansion, referred to as the NTK regime [44]. Fort et al. [24] find that a linear expansion becomes more accurate in the later phase of neural network training, a phase which closely resembles fine-tuning.

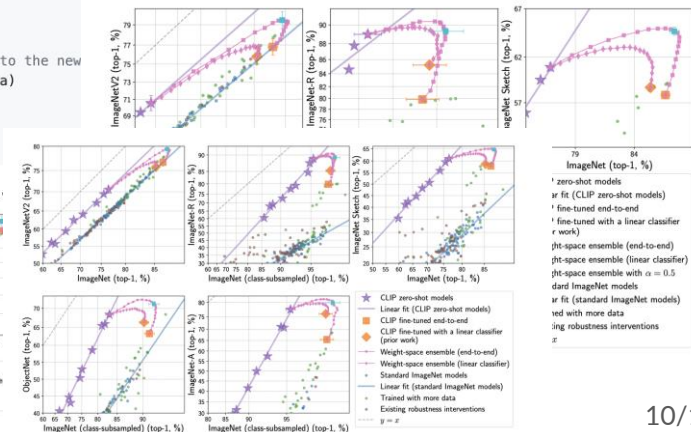
Consider the set $\Theta = \{(1 - \alpha)\theta_0 + \alpha\theta_1 : \alpha \in [0, 1]\}$ consisting of all θ which lie on the linear path between θ_0 and θ_1 .

Proposition 1. When $f(\theta) = f(\theta_0) + \nabla f(\theta_0)^\top(\theta - \theta_0)$ for all $\theta \in \Theta$, the weight- and output-space ensemble of θ_0 and θ_1 are equivalent.

Proof. We may begin with the weight-space ensemble and retrieve the output-space ensemble

$$f((1-\alpha)\theta_0 + \alpha\theta_1) \quad (12)$$

		Distribution status						Avg status	
		IN-Ver	IN-R	IN-Search	ObjectNet	IN-A		Avg status	Avg status
$f = f(h) + \nabla f(h)^T \alpha$									
$f = f(h) + \nabla f(h)^T \alpha$	WSE-FT, end-to-end								
$= (1 - \alpha) f(h)$	$\alpha=0.00$	68.3	61.9	67.8	68.2	53.0	49.8	58.1	63.3
	$\alpha=0.05$	70.7	64.0	70.8	69.6	54.5	51.5	58.8	68.2
	$\alpha=0.10$	69.7	63.7	72.9	68.1	53.7	52.3	58.1	65.0
	$\alpha=0.15$	74.0	67.2	79.9	67.1	56.6	53.5	61.8	68.3
	$\alpha=0.20$	68.7	68.1	72.5	68.1	54.2	52.2	58.1	65.0
	$\alpha=0.25$	77.9	69.9	80.1	53.1	57.4	54.6	63.0	79.4
	$\alpha=0.30$	79.9	70.9	82.1	53.6	57.5	54.6	63.0	79.4
	$\alpha=0.35$	79.7	71.5	79.9	53.9	57.6	54.3	63.4	71.5
	$\alpha=0.40$	80.5	72.1	79.6	54.1	57.7	53.8	63.3	72.9
	$\alpha=0.45$	81.2	72.4	79.3	53.9	57.9	53.2	63.0	74.0
	$\alpha=0.50$	81.7	72.8	78.7	53.9	57.3	52.2	63.3	74.0
	$\alpha=0.55$	81.0	73.0	78.0	53.8	57.6	51.1	63.0	72.9
	$\alpha=0.60$	82.4	72.9	77.2	53.4	56.2	50.9	61.9	72.2
	$\alpha=0.65$	81.7	76.3	75.0	52.0	55.5	48.9	61.0	70.7
	$\alpha=0.70$	82.6	73.2	75.2	52.4	55.5	47.4	60.6	71.6
	$\alpha=0.75$	82.6	73.1	73.9	51.8	54.3	46.0	58.8	71.2
	$\alpha=0.80$	82.5	72.8	72.7	51.0	53.5	44.6	58.1	70.7
	$\alpha=0.85$	82.3	72.4	71.1	50.0	52.7	42.9	57.8	70.0
	$\alpha=0.90$	82.1	69.5	69.5	49.9	51.7	40.9	56.8	68.5
	$\alpha=0.95$	81.7	71.5	67.7	47.6	50.7	38.5	53.5	68.5
	$\alpha=1.00$	81.3	70.9	65.6	46.3	49.6	36.7	53.8	67.5
	WSE-FT, linear classifier								
$\alpha=0.00$		68.4	62.6	67.6	68.2	53.0	50.0	58.4	63.4
$\alpha=0.05$		69.9	63.7	77.9	69.9	53.2	50.6	58.1	63.4
$\alpha=0.10$		71.3	64.8	78.2	68.5	54.7	51.0	58.6	65.4



- Новизна подхода
- Похоже на обучение на двух датасетах
- Выбор θ

Сравнить качество с обучением на двух датасетах?

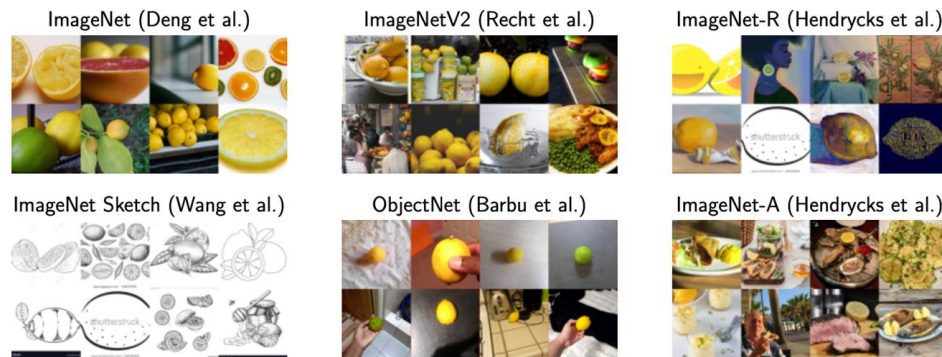


Figure 2: Samples of the class *lemon*, from the reference distribution ImageNet [17] and the derived distribution shifts considered in our main experiments: ImageNet-V2 [83], ImageNet-R [37], ImageNet Sketch [100], ObjectNet [4], and ImageNet-A [38].



Выводы

- Умеем хорошо дообучать сетки, получаем лучшее качество
- Простые идеи неплохо работают
- Ещё одно применение усреднений весов моделей
- В статье важна не только идея, но и оформление результатов