When ViT Outperform ResNets

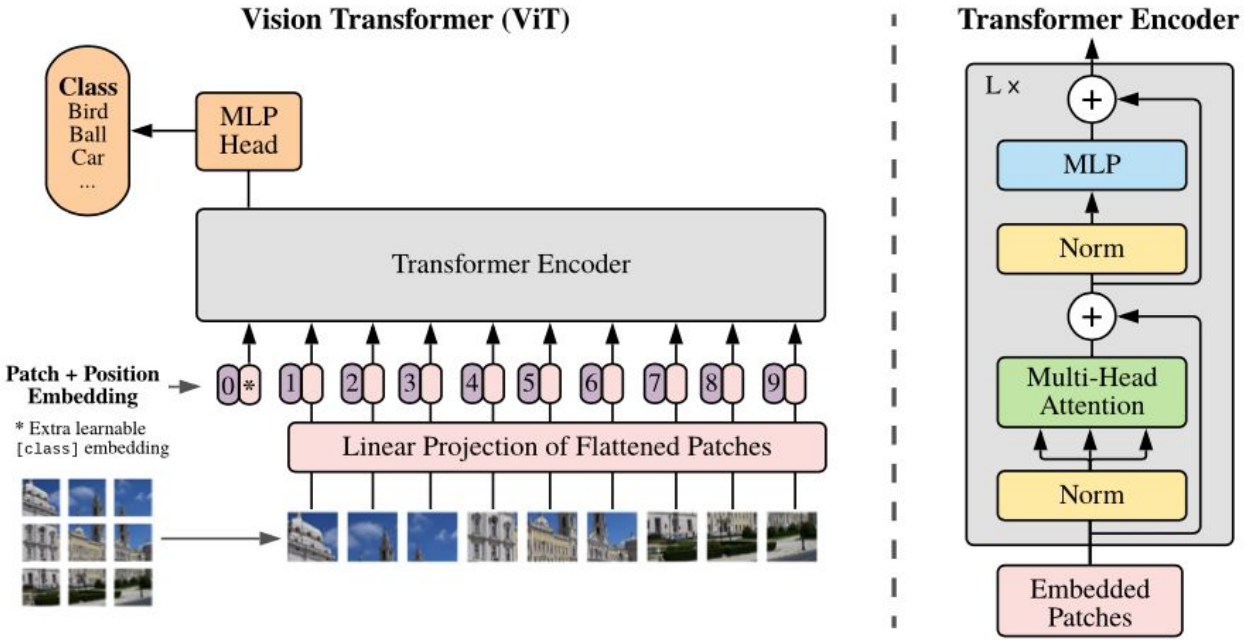
without Pre-training or Strong Data Augmentations

Докладчик: Константин Матвеев
Рецензент: Никита Андреев
Исследователь: Александра Сендерович
Хакер: Сергей Петрович

План:

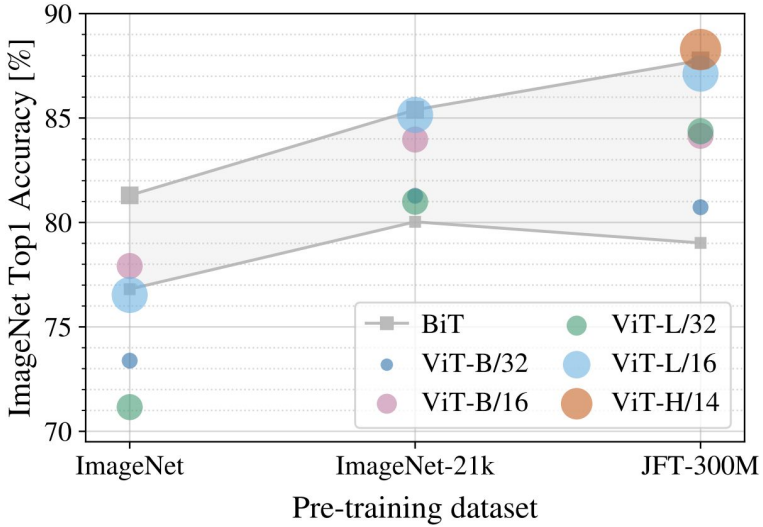
1. ViT: преимущества и проблемы
2. Оптимизатор SAM
3. Результаты ViT-SAM
4. Свойства ViT-SAM

Vision Transformer



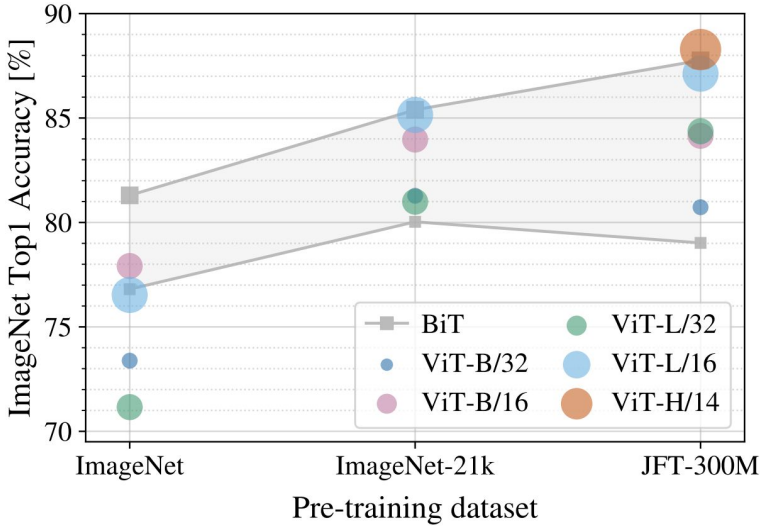
Проблема ViT

Ему нужно либо много данных,



Проблема ViT

Ему нужно либо много данных,

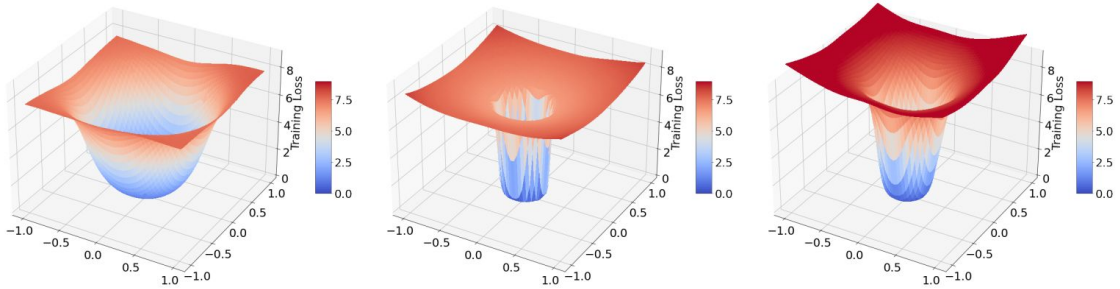


либо сложные аугментации

Pre-training	Fine-tuning	Rand-Augment	AutoAug	Mixup	CutMix	Erasing	Stoch. Depth	Repeated Aug.	Dropout	Exp. Moving Avg.
adamw	adamw	✓	✗	✓	✓	✓	✓	✓	✗	✗

Проблема ViT

- Функция потерь “острая”
- Из-за этого нужно или очень много данных, или сильные аугментации
- Высокая чувствительность к гиперпараметрам и инициализации



(a) ResNet (b) ViT (c) Mixer

План:

1. ViT: преимущества и проблемы
2. **Оптимизатор SAM**
3. Результаты ViT-SAM
4. Свойства ViT-SAM

Sharpness-Aware Minimizer (SAM)

- Хотим искусственно сгладить функцию потерь
- Для этого будем вместо задачи

$$\min_w L_{train}(w)$$

решать задачу

$$\min_w \max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon),$$

Sharpness-Aware Minimizer (SAM)

- Решаем $$\min_{w} \max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon),$$

Sharpness-Aware Minimizer (SAM)

- Решаем $\min_w \max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon),$
- На текущем шаге $\epsilon^\star = \arg\max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon)$

Sharpness-Aware Minimizer (SAM)

- Решаем $$\min_w \max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon),$$
- На текущем шаге $$\epsilon^\star = \arg\max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon)$$
- Сложно; вычисляем приближение первого порядка

$$\begin{aligned}\hat{\epsilon}(w) &= \arg\max_{\|\epsilon\|_2 \leq \rho} L_{train}(w) + \epsilon^T \nabla_w L_{train}(w) \\ &= \rho \nabla_w L_{train}(w) / \|\nabla_w L_{train}(w)\|_2\end{aligned}$$

Sharpness-Aware Minimizer (SAM)

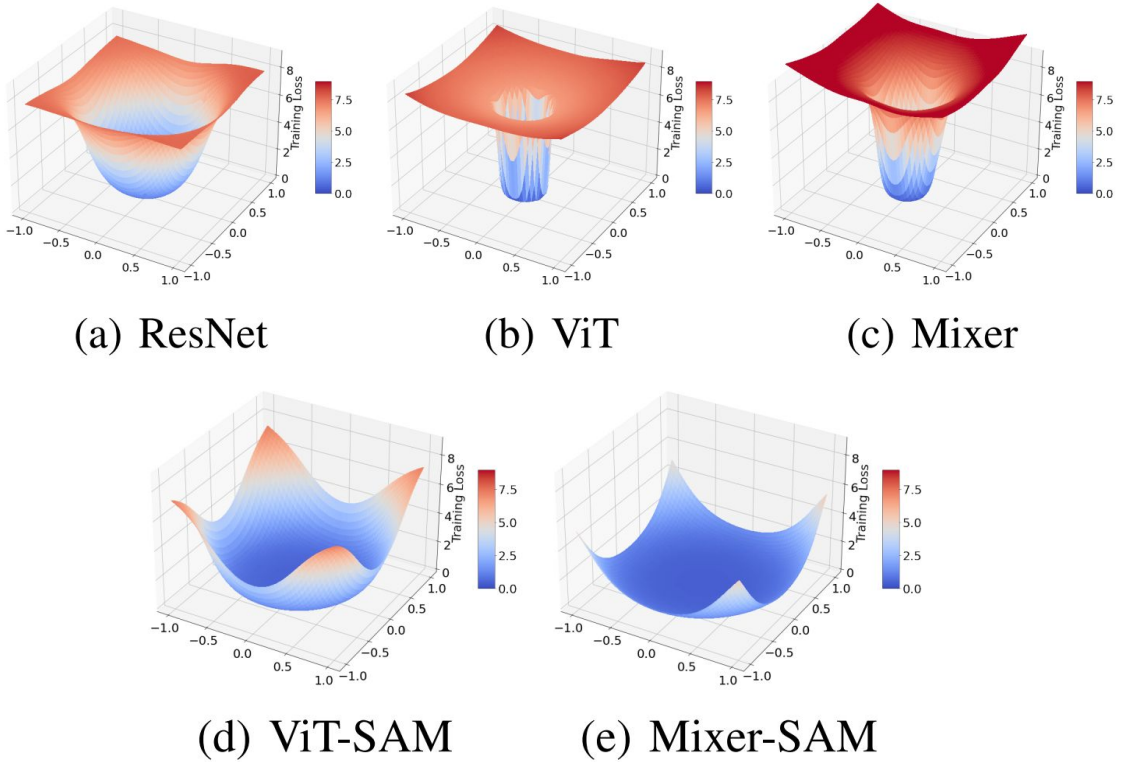
- Решаем $\min_w \max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon),$
- На текущем шаге $\epsilon^\star = \arg\max_{\|\epsilon\|_2 \leq \rho} L_{train}(w + \epsilon)$
- Сложно; вычисляем приближение первого порядка

$$\hat{\epsilon}(w) = \arg\max_{\|\epsilon\|_2 \leq \rho} L_{train}(w) + \epsilon^T \nabla_w L_{train}(w)$$

$$= \rho \nabla_w L_{train}(w) / \|\nabla_w L_{train}(w)\|_2$$

- Шаг по $\nabla_w L_{train}(w)|_{w+\hat{\epsilon}(w)}$

Sharpness-Aware Minimizer (SAM)



(a) ResNet (b) ViT (c) Mixer

(d) ViT-SAM (e) Mixer-SAM

План:

1. ViT: преимущества и проблемы
2. Оптимизатор SAM
3. **Результаты ViT-SAM**
4. Свойства ViT-SAM

Качество с SAM

- Острота лосса на порядок уменьшается
- SAM улучшает качество top-1 accuracy на ImageNet по всем моделям
- ViT превосходит ResNet и на обычной, и на corrupted выборке

	ResNet-152	ResNet-152-SAM	ViT-B/16	ViT-B/16-SAM	Mixer-B/16	Mixer-B/16-SAM
#Params	60M		87M		59M	
NTK κ †	2801.6		4205.3		14468.0	
Hessian λ_{max}	179.8	**42.0**	738.8	**20.9**	1644.4	**22.5**
ImageNet (%)	78.5	**79.3**	74.6	**79.9**	66.4	**77.4**
ImageNet-C (%)	50.0	**52.2**	46.6	**56.5**	33.8	**48.8**

Качество с SAM

- Острота лосса на порядок уменьшается
- SAM улучшает качество top-1 accuracy на ImageNet по всем моделям
- ViT превосходит ResNet и на обычной, и на corrupted выборке

	ResNet-152	ResNet-152-SAM	ViT-B/16	ViT-B/16-SAM	Mixer-B/16	Mixer-B/16-SAM
#Params	60M		87M		59M	
NTK κ †	2801.6		4205.3		14468.0	
Hessian λ_{max}	179.8	**42.0**	738.8	**20.9**	1644.4	**22.5**
ImageNet (%)	78.5	**79.3**	74.6	**79.9**	66.4	**77.4**
ImageNet-C (%)	50.0	**52.2**	46.6	**56.5**	33.8	**48.8**

Качество с SAM

- Острота лосса на порядок уменьшается
- SAM улучшает качество top-1 accuracy на ImageNet по всем моделям
- ViT превосходит ResNet и на обычной, и на corrupted выборке

	ResNet-152	ResNet-152-SAM	ViT-B/16	ViT-B/16-SAM	Mixer-B/16	Mixer-B/16-SAM
#Params	60M		87M		59M	
NTK κ †	2801.6		4205.3		14468.0	
Hessian λ_{max}	179.8	**42.0**	738.8	**20.9**	1644.4	**22.5**
ImageNet (%)	78.5	**79.3**	74.6	**79.9**	66.4	**77.4**
ImageNet-C (%)	50.0	**52.2**	46.6	**56.5**	33.8	**48.8**

Качество с SAM

TLDR: SAM улучшает все модели всех размеров на всех датасетах

Model	#params	Throughput (img/sec/core)	ImageNet	Real	V2	ImageNet-R	ImageNet-C
ResNet							
ResNet-50-SAM	25M	2161	76.7 (+0.7)	83.1 (+0.7)	64.6 (+1.0)	23.3 (+1.1)	46.5 (+1.9)
ResNet-101-SAM	44M	1334	78.6 (+0.8)	84.8 (+0.9)	66.7 (+1.4)	25.9 (+1.5)	51.3 (+2.8)
ResNet-152-SAM	60M	935	79.3 (+0.8)	84.9 (+0.7)	67.3 (+1.0)	25.7 (+0.4)	52.2 (+2.2)
ResNet-50x2-SAM	98M	891	79.6 (+1.5)	85.3 (+1.6)	67.5 (+1.7)	26.0 (+2.9)	50.7 (+3.9)
ResNet-101x2-SAM	173M	519	80.9 (+2.4)	86.4 (+2.4)	69.1 (+2.8)	27.8 (+3.2)	54.0 (+4.7)
ResNet-152x2-SAM	236M	356	81.1 (+1.8)	86.4 (+1.9)	69.6 (+2.3)	28.1 (+2.8)	55.0 (+4.2)
Vision Transformer							
ViT-S/32-SAM	23M	6888	70.5 (+2.1)	77.5 (+2.3)	56.9 (+2.6)	21.4 (+2.4)	46.2 (+2.9)
ViT-S/16-SAM	22M	2043	78.1 (+3.7)	84.1 (+3.7)	65.6 (+3.9)	24.7 (+4.7)	53.0 (+6.5)
ViT-S/14-SAM	22M	1234	78.8 (+4.0)	84.8 (+4.5)	67.2 (+5.2)	24.4 (+4.7)	54.2 (+7.0)
ViT-S/8-SAM	22M	333	81.3 (+5.3)	86.7 (+5.5)	70.4 (+6.2)	25.3 (+6.1)	55.6 (+8.5)
ViT-B/32-SAM	88M	2805	73.6 (+4.1)	80.3 (+5.1)	60.0 (+4.7)	24.0 (+4.1)	50.7 (+6.7)
ViT-B/16-SAM	87M	863	79.9 (+5.3)	85.2 (+5.4)	67.5 (+6.2)	26.4 (+6.3)	56.5 (+9.9)
MLP-Mixer							
Mixer-S/32-SAM	19M	11401	66.7 (+2.8)	73.8 (+3.5)	52.4 (+2.9)	18.6 (+2.7)	39.3 (+4.1)
Mixer-S/16-SAM	18M	4005	72.9 (+4.1)	79.8 (+4.7)	58.9 (+4.1)	20.1 (+4.2)	42.0 (+6.4)
Mixer-S/8-SAM	20M	1498	75.9 (+5.7)	82.5 (+6.3)	62.3 (+6.2)	20.5 (+5.1)	42.4 (+7.8)
Mixer-B/32-SAM	60M	4209	72.4 (+9.9)	79.0 (+10.9)	58.0 (+10.4)	22.8 (+8.2)	46.2 (12.4)
Mixer-B/16-SAM	59M	1390	77.4 (+11.0)	83.5 (+11.4)	63.9 (+13.1)	24.7 (+10.2)	48.8 (+15.0)
Mixer-B/8-SAM	64M	466	79.0 (+10.4)	84.4 (+10.1)	65.5 (+11.6)	23.5 (+9.2)	48.9 (+16.9)

Качество с SAM vs сильные аугментации

- SAM лучше аугментаций (AUG)
- Для ViT особенно заметно на меньших размерах
- Помогает их объединять, но не всегда

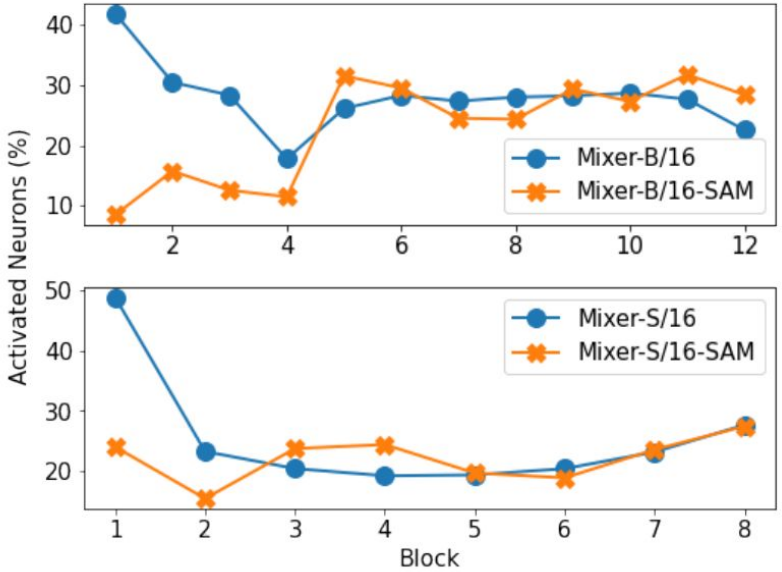
Dataset	#Images	ResNet-152				ViT-B/16				Mixer-B/16			
		Vanilla	SAM	AUG	SAM + AUG	Vanilla	SAM	AUG	SAM + AUG	Vanilla	SAM	AUG	SAM + AUG
ImageNet	1,281,167	78.5	79.3	78.8	78.9	74.6	79.9	79.6	81.5	66.4	77.4	76.5	78.1
i1k (1/2)	640,583	74.2	75.6	75.1	75.5	64.9	75.4	73.1	75.8	53.9	71.0	70.4	73.1
i1k (1/4)	320,291	68.0	70.3	70.2	70.6	52.4	66.8	63.2	65.6	37.2	62.8	61.0	65.8
i1k (1/10)	128,116	54.6	57.1	59.2	59.5	32.8	46.1	38.5	45.7	21.0	43.5	43.0	51.0

План:

1. ViT: преимущества и проблемы
2. Оптимизатор SAM
3. Результаты ViT-SAM
4. **Свойства ViT-SAM**

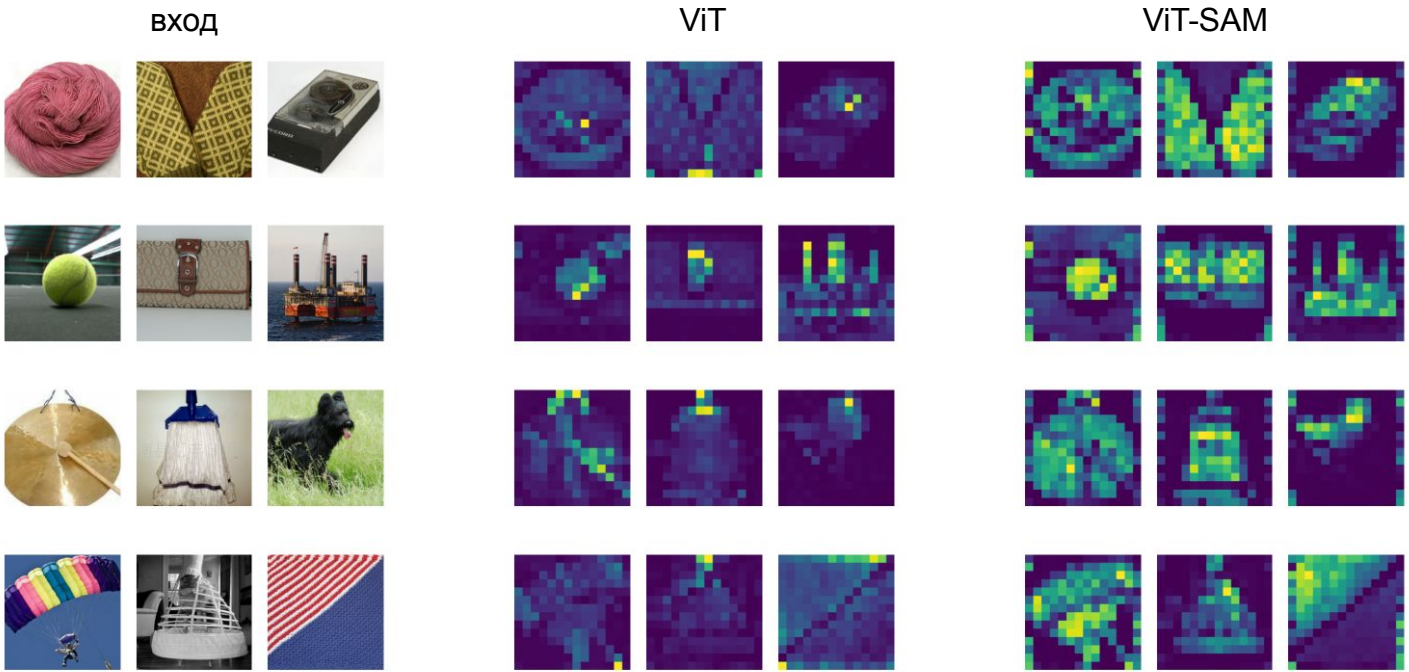
Свойства моделей с SAM: разреженность

- Модели разреженнее, особенно на начальных слоях (< 10% ненулевых нейронов)
  => Можно прунить



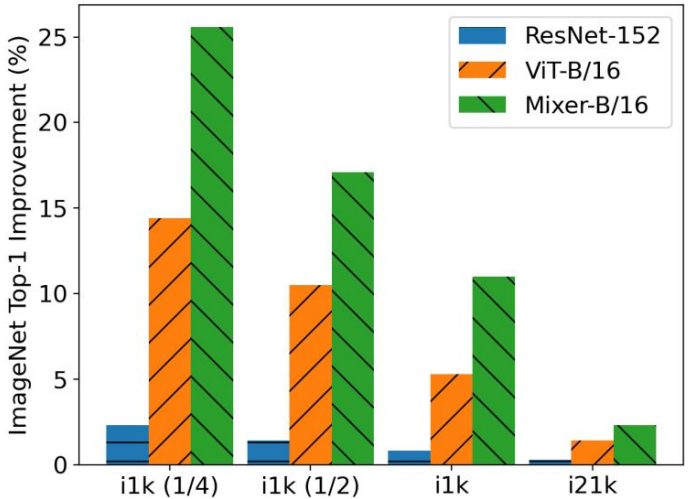
Свойства моделей с SAM: интерпретируемость

- Attention-маски классификационного токена ViT более интерпретируемы с SAM



Свойства моделей с SAM: вклад SAM

- Чем меньше обучающая выборка, тем больше вклад SAM в качество



Итого

- ViT и MLP-Mixer тяжело обучать из-за остроты функции потерь
- SAM сглаживает лосс и улучшает скоры
- ViT-SAM побеждает ResNet на ImageNet без сильных аугментаций
- Модели после SAM более разреженные и интерпретируемы

Рецензия

Сильные стороны

- Статья не очень сложная, хорошо читается
- Авторы используют SAM и объясняют почему именно он подходит для улучшения обучения моделей
- Также в статье есть теоретический обзор SAM
- Статья больше про эксперименты, поэтому их проведено большое количество, все грамотно описано, понятные таблицы
- Есть ablation study, где рассказано про contrastive learning, adversarial learning, а также изменение размера выборки и поведение моделей с SAM в этих контекстах

Слабые стороны

- В статье не предлагается ничего принципиально нового - существующий метод SAM применяется для двух архитектур, однако эксперименты по применению и есть суть статьи
- В самой статье нет результатов для больших модификаций моделей (например ViT-L, ResNet-152x4) однако авторы объясняют это ограничением ресурсов.

Основная критика на openreview - отсутствие теоретической новизны статьи

Оценки 8, 6, 5

Воспроизводимость

Авторы выложили чекпоинты на гитхаб, проблем с воспроизведением результатов исходя из статьи быть не должно.

Оценка по критериям НИПСа: 8

Уверенность: 4

Исследование

Публикация

- ICLR 2022 Spotlight
- 1 версия – 3 июня 2021 года, 2 версия – 11 октября 2021 года (на конфу)
- Оценки рецензентов: 5, 6, 8, 8
- После этого добавили немного экспериментов
- Далее была поставлена ещё одна 8

Авторы

Xiangning Chen, стажёр в Google, PhD в UCLA

Cho-Jui Hsieh, доцент, глава научной группы в UCLA

Boqing Gong, исследователь из Google

Авторы. Первый автор, Xiangning Chen

- Стажировка в Google Research, AutoML Team
- PhD в University of California, Los Angeles
- Все статьи с начала PhD – в соавторстве со вторым автором
- Статьи по теме:
  - 2 статьи про ViTs отозвал с ICLR 2022 из-за плохих оценок:
    - Can Vision Transformers Perform Convolution?
    - Sharpness-Aware Minimization in Large-Batch Training: Training Vision Transformer In Minutes
  - В 2020 году – статья про Neural Architecture Search со сглаживанием лосса
- Статьи не по теме:
  - Ещё 2 статьи принято на ICLR 2022: распределённое adversarial обучение, AutoML
  - Ранее статьи по NAS + 1 статья про adversarial аугментации для компьютерного зрения в соавторстве со 2 и 3 автором

Авторы

- Второй автор, Cho-Jui Hsieh:
  - Доцент, глава UCLA Computational Machine Learning Group
  - Группа занимается adversarial robustness, model compression
  - Последний или предпоследний автор публикаций своей группы
  - Научный руководитель первого автора
  - В статье использовался оптимизатор LAMB из его статьи 2020 года

- Третий автор, Boqing Gong:
  - Исследователь из Google Research
  - Занимается компьютерным зрением, adversarial robustness
  - Много публикаций последним или предпоследним автором

Ссылки

4 основных идеи:

- Vision Transformer (Dosovitsky, 2020)
- MLP Mixer (Tolstikhin, 2021)
- Sharpness-Aware Minimization (Foret, 2021)
- ResNet (He, 2015)

1 нестандартная метрика: NTK condition number из (Xiao, 2020)

Цитирования

- SemanticScholar: 26 цитирований, из них 17 – в обзоре литературы
- В основном – оптимизируют ViT и сравниваются:
  - Bootstrapping ViTs: Towards Liberating Vision Transformers from Pre-training
- Продолжение – применяют SAM к текстовым трансформерам:
  - Sharpness-Aware Minimization Improves Language Model Generalization
  - “Encouraged by wins in the vision domain, we ask whether SAM can deliver similar gains in the language domain”

Дальнейшие исследования и приложения

- Прунинг, сжатие Vision Transformers
  - В статье говорилось, что при обучении с SAM мало ненулевых активаций
- Эксперименты с corrupted labels, как в статье про mixup
  - Corrupted labels – в обучающем наборе данных есть неправильные таргеты
- Сделать многофункциональный трансформер
  - Работающий с картинками, текстом, звуком