When Vision Transformers Outperform ResNets without Pre-training or Strong Data Augmentations

Сусла Диана
Терехова Юлия
Виноградова Дарья

Постановка задачи

- Современные модели глубинного обучения все еще имеют проблемы связанные с оптимизацией (например чувствительность к инициализации и скорости обучения)
- Работа направлена на исследование моделей ViT и MLP-Mixer с точки зрения геометрии ландшафта функции потерь.
- Цель: увеличить эффективность моделей при обучении и улучшить обобщающую способность при применении, при этом снизив их зависимость от предобучения на больших наборах данных или strong data augmentations
- Основная идея: используем SAM для регуляризации ViT и MLP-Mixer

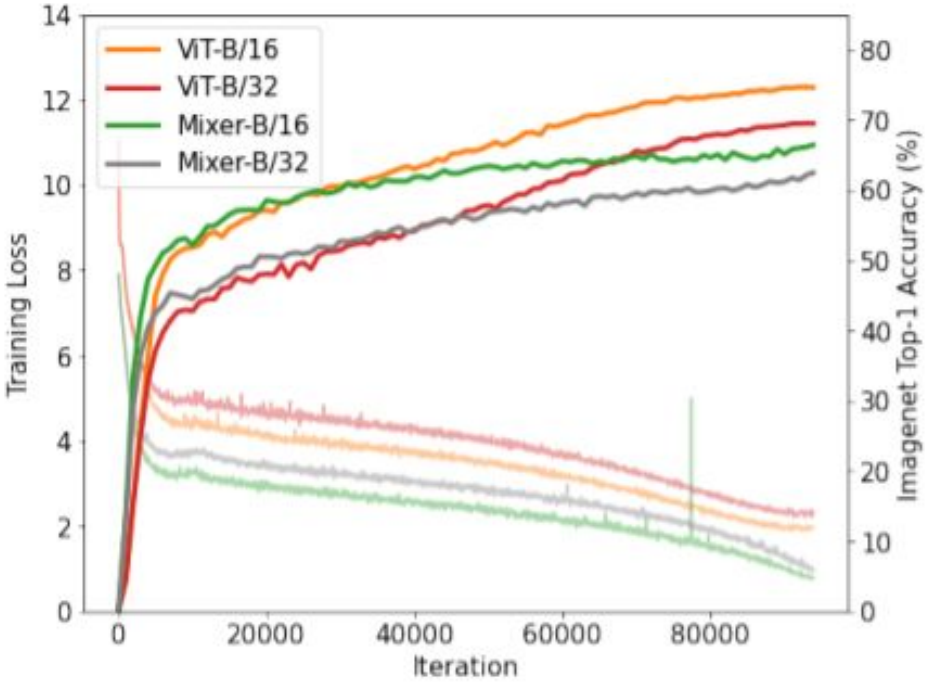
ResNet vs ViT vs MLP-Mixer

	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**

† As it is prohibitive to compute the exact NTK, we approximate the value by averaging over its sub-diagonal blocks. Please see Appendix E for details.

ResNet vs ViT vs MLP-Mixer

- Функции потерь ViT и MLP-Mixer имеют очень резкие локальные минимумы
- Все модели имеют низкую ошибку на тесте. MLP-Mixer более склонен к переобучению
- ViTs и MLP-Mixers имеют более плохую обучающую способность по сравнению с ResNet

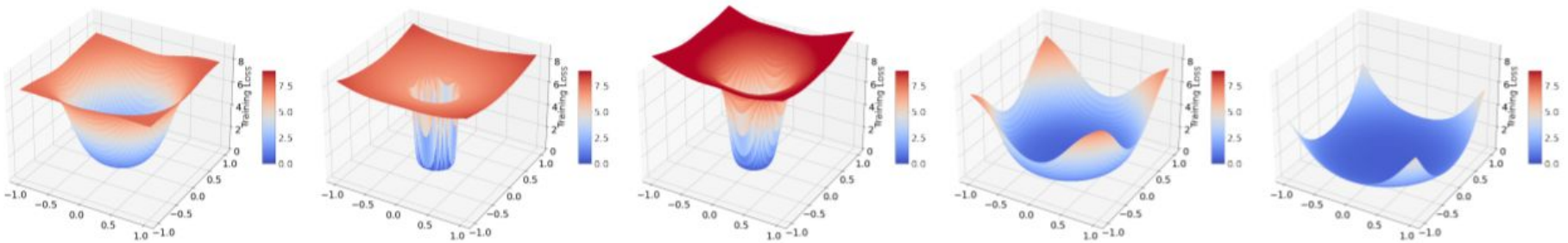


ResNet vs ViT vs MLP-Mixer

- При обучении на ImageNet трансформеры дают немного меньшую точность, чем ResNet, сопоставимого размера, обученный тем же методом
- В случае ViT и MLP-Mixer, недостаток знаний о задаче затрудняет обучение
- ResNet содержит в себе свертки, благодаря которым получается избежать “плохих” локальных минимумов при обучении

Напоминание о SAM (sharpness-aware minimization)

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



(a) ResNet (b) ViT (c) Mixer (d) ViT-SAM (e) Mixer-SAM

SAM. Улучшения ViT и MLP-Mixer

1) Область вокруг локального минимума становится более гладкой
2) Возрастает accuracy (top-1 accuracy ViT-B/16: с 74.6% до 79.9%, Mixer-B/16: с 66.4% до 77.4%)
3) Модель становится более надежной (accuracy ImageNet-C ViT-B/16: увеличилась на 9.9%, Mixer-B/16: увеличилась на 15.0%

	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**

† As it is prohibitive to compute the exact NTK, we approximate the value by averaging over its sub-diagonal blocks. Please see Appendix E for details.

Результаты

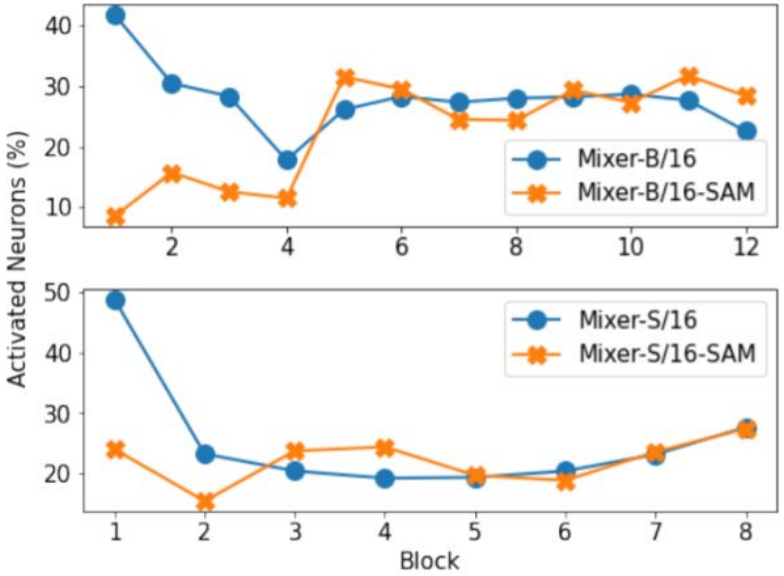
Model	#params	Throughput (img/sec/core)	ImageNet	Real	V2	ImageNet-R	ImageNet-C
ResNet							
ResNet-50-SAM	25M	2161	76.7 (+0.7)	83.1 (+0.7)	64.6 (+1.0)	23.3 (+1.1)	46.5 (+1.9)
ResNet-101-SAM	44M	1334	78.6 (+0.8)	84.8 (+0.9)	66.7 (+1.4)	25.9 (+1.5)	51.3 (+2.8)
ResNet-152-SAM	60M	935	79.3 (+0.8)	84.9 (+0.7)	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)

Внутренние изменения моделей с SAM

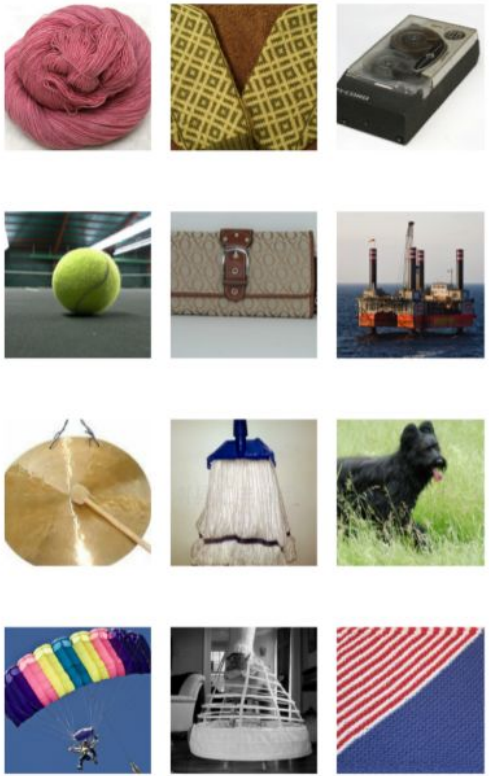
1) Более гладкий ландшафт функции потерь для каждой компоненты
2) Большая L2 норма весов

Model	λ_{max} of diagonal blocks of Hessian							$\|w\|_2$	$\|a_1\|_2$	$\|a_6\|_2$	$\|a_{12}\|_2$
	Embedding	MSA/ Token MLP	MLP/ Channel MLP	Block1	Block6	Block12	Whole				
ViT-B/16	300.4	179.8	281.4	44.4	32.4	26.9	738.8	269.3	104.9	104.3	138.1
ViT-B/16-SAM	3.8	8.5	9.6	1.7	1.7	1.5	20.9	353.8	117.0	120.3	97.2
Mixer-B/16	1042.3	95.8	417.9	239.3	41.2	5.1	1644.4	197.6	96.7	135.1	74.9
Mixer-B/16-SAM	18.2	1.4	9.5	4.0	1.1	0.3	22.5	389.9	110.9	176.0	216.1

Внутренние изменения моделей с SAM



3) Меньшее число активный нейронов на первых слоях MLP-Mixer
4) Более наглядные и содержательные маски внимания



ViT-S/16

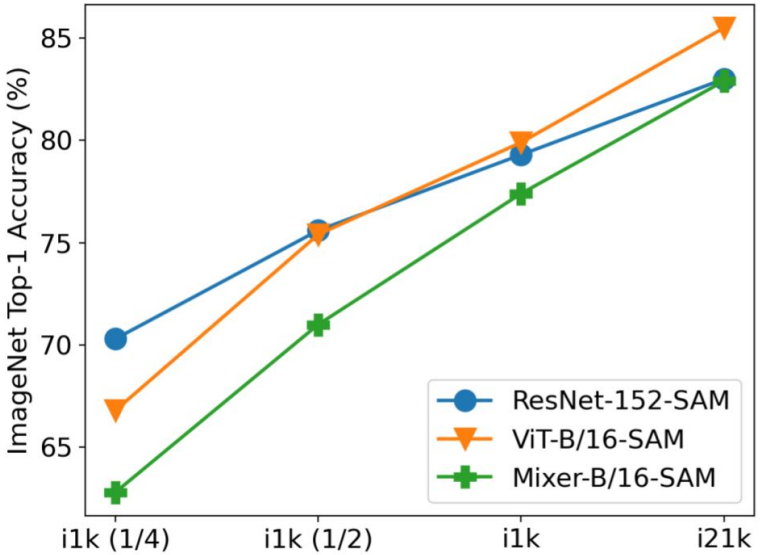
ViT-S/16-SAM

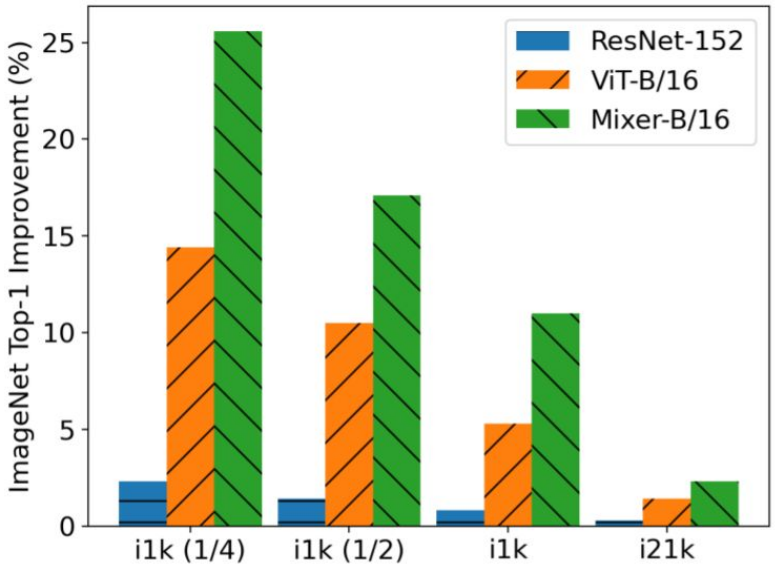
SAM vs. strong augmentations

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

Ablation studies

1) Изменение размера обучающего набора данных





Ablation studies

2) Contrastive learning (top-1 accuracy ViT-S/16: с 77.0% до 78.1%, ViT-B/16: с 77.4% до 80.0%)

3) Adverasrial обучение (10 PGD атак)

Model	#params	Throughput (img/sec/core)	ImageNet	Real	V2	PGD-10	ImageNet-R	ImageNet-C
ResNet								
ResNet-50-SAM	25M	2161	70.1 (-0.7)	77.9 (-0.3)	56.6 (-0.8)	54.1 (+0.9)	27.0 (+0.9)	42.7 (-0.1)
ResNet-101-SAM	44M	1334	73.6 (-0.4)	81.0 (+0.1)	60.4 (-0.6)	58.8 (+1.4)	29.5 (+0.6)	46.9 (+0.3)
ResNet-152-SAM	60M	935	75.1 (-0.4)	82.3 (+0.2)	62.2 (-0.4)	61.0 (+1.8)	30.8 (+1.4)	49.1 (+0.6)
Vision Transformer								
ViT-S/16-SAM	22M	2043	73.2 (+1.2)	80.7 (+1.7)	60.2 (+1.4)	58.0 (+5.2)	28.4 (+2.4)	47.5 (+1.6)
ViT-B/32-SAM	88M	2805	69.9 (+3.0)	76.9 (+3.4)	55.7 (+2.5)	54.0 (+6.4)	26.0 (+3.0)	46.4 (+3.0)
ViT-B/16-SAM	87M	863	76.7 (+3.9)	82.9 (+4.1)	63.6 (+4.3)	62.0 (+7.7)	30.0 (+4.9)	51.4 (+5.0)
MLP-Mixer								
Mixer-S/16-SAM	18M	4005	67.1 (+2.2)	74.5 (+2.3)	52.8 (+2.5)	50.1 (+4.1)	22.9 (+2.6)	37.9 (+2.5)
Mixer-B/32-SAM	60M	4209	69.3 (+9.1)	76.4 (+10.2)	54.7 (+9.4)	54.5 (+13.9)	26.3 (+8.0)	43.7 (+8.8)
Mixer-B/16-SAM	59M	1390	73.9 (+11.1)	80.8 (+11.8)	60.2 (+11.9)	59.8 (+17.3)	29.0 (+10.5)	45.9 (+12.5)

Итог

- Проведено исследование моделей ViT и MLP-Mixer с точки зрения геометрии ландшафта функции потерь, для того чтобы увеличить эффективность данных моделей при обучении и улучшить обобщающую способность при применении, при этом снизив их зависимость от предобучения на больших наборах данных или strong data augmentations
- Для решения задачи используется регуляризация при помощи SAM. С её помощью ландшафт функции потерь становится более гладким
- В результате улучшается точность, увеличивается обобщающая способность модели и ее надежность

Рецензия

Аннотация

В работе рассмотрено применение SAM к ViT и MLP-Mixer. Основная идея в том, что SAM может улучшить результаты моделей и увеличить обобщаемость применений, по сравнению с аугментациями, которые являются специфичными для различных задач.

Плюсы

+ сравнили множество моделей и их комбинаций по многим критериям
+ есть иллюстрации экспериментов, графики
+ предоставлен код с подробными указаниями по использованию
+ решает важную проблему, глубоко её рассматривает, актуальная

Рецензия

Минусы

- написано иногда трудно читаемо, много прерываний текста ссылками на референсы
- нет результатов для модели без inductive bias (напр vanilla MLP + потом SAM улучшит ли)

Воспринимаемость

Иногда использован достаточно сложный язык для простых вещей, которые в нем не нуждаются.

Оценка - 7.5
Уверенность - 4

Практик-исследователь

История статьи

- первая версия загружена на arxiv 3 июня 2021 года
- принята на ICLR-2022 (формат spotlight)
- уже 19 цитирований

Xiangning Chen

Cho-Jui Hsieh

Boqing Gong

Поподробнее про авторов

- **Xiangning Chen** - PhD из университета Калифорнии. Области интересов: рекомендательные системы, AutoML, CV
- **Cho-Jui Hsieh** - его научный руководитель. В основном работает над оптимизацией нейросетевых архитектур
- **Boqing Gong** - исследователь из гугла, руководитель стажировки Xiangning Chen. Его сфера деятельности - различные CV-задачи и изучение внутреннего устройства моделей для этих задач

Связанные работы

Начало истории

- Xinlei Chen, Saining Xie, Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers - детальное исследование предобучения трансформера с помощью Contrastive Learning

Продолжение истории

- D Bahri, H Mobahi, Y Tay. Sharpness-Aware Minimization Improves Language Model Generalization - адаптация для языковых моделей
- Y Liu, S Mai, X Chen, CJ Hsieh, Y You. Sharpness-Aware Minimization in Large-Batch Training: Training Vision Transformer In Minutes - авторы ускоряют свой метод

Что еще?

- экстраполировать SAM на задачи из других областей
- оптимизировать не только для ViT, но и для других моделей, проанализировать прирост качества
- сравнить предобученную модель и построенную с нуля с использованием SAM