MLP-Mixer: An all-MLP Architecture for Vision[1]

Докладчик - Коля Карташев, 181
Рецензент - Артём Щербинин, 181
Практик-исследователь - Алексей Цеховой, 181

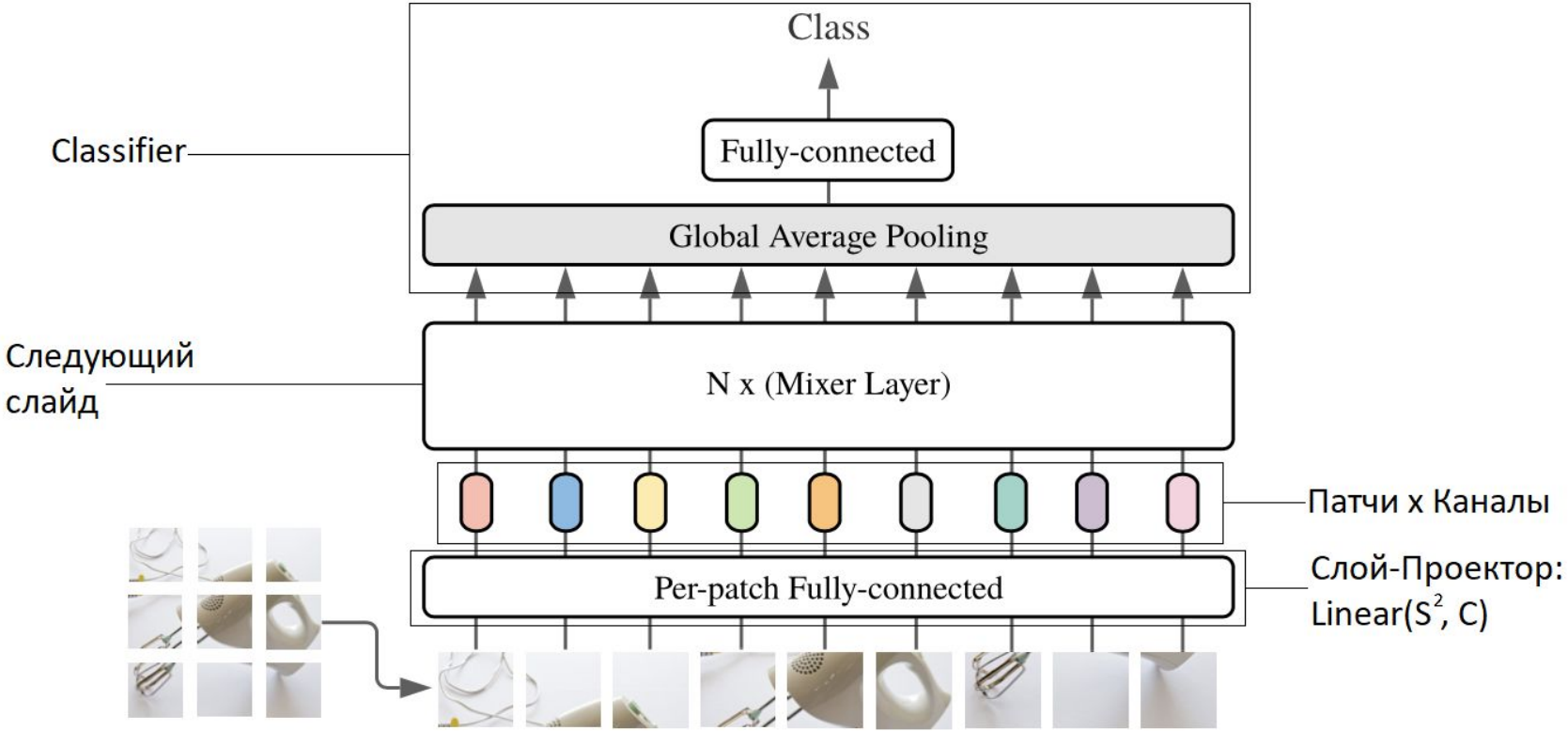
[1] https://arxiv.org/abs/2105.01601

MLP-Mixer: идея

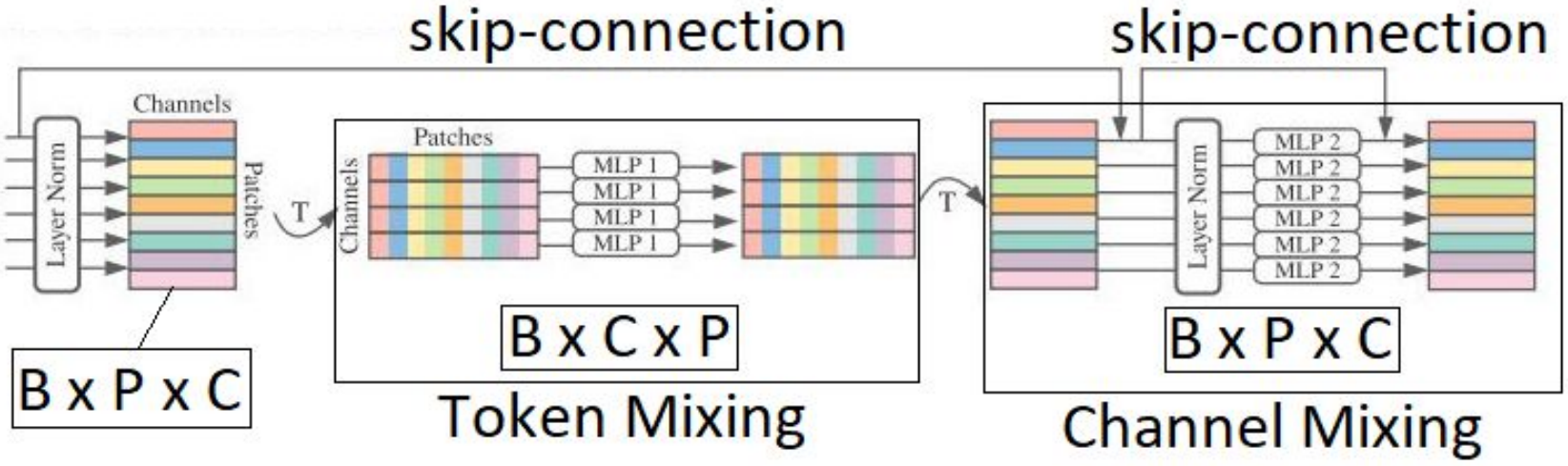
Цель: Достичь близких к SoTA результатов, используя только MLP в архитектуре модели.

План: Предобучать на огромных наборах данных, использовать современные методы регуляризации/

MLP-Mixer: архитектура



“Mixer” слой



MLP слой: линейный + GeLU + линейный

Цели Сверточной Нейронной Сети

1. Скомпозировать разные признаки в одной локации
2. Скомпозировать информацию из разных локаций

CNN выполняет обе задачи одновременно
Mixer Layer выполняет каждую задачу поочередно

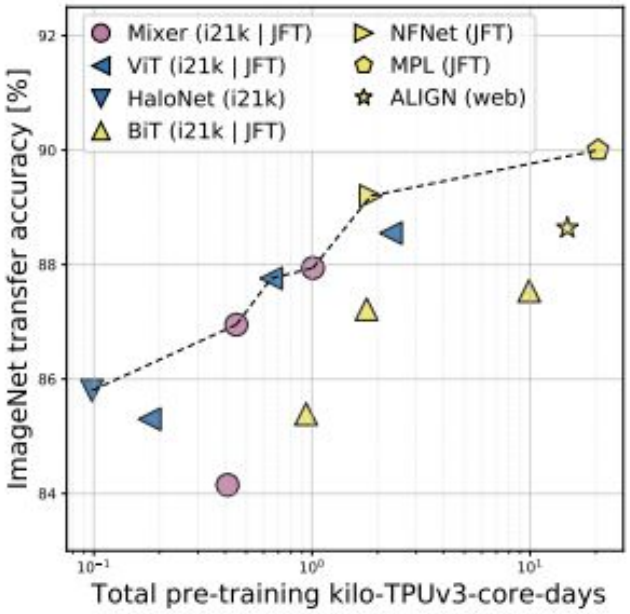
Конфигурации модели

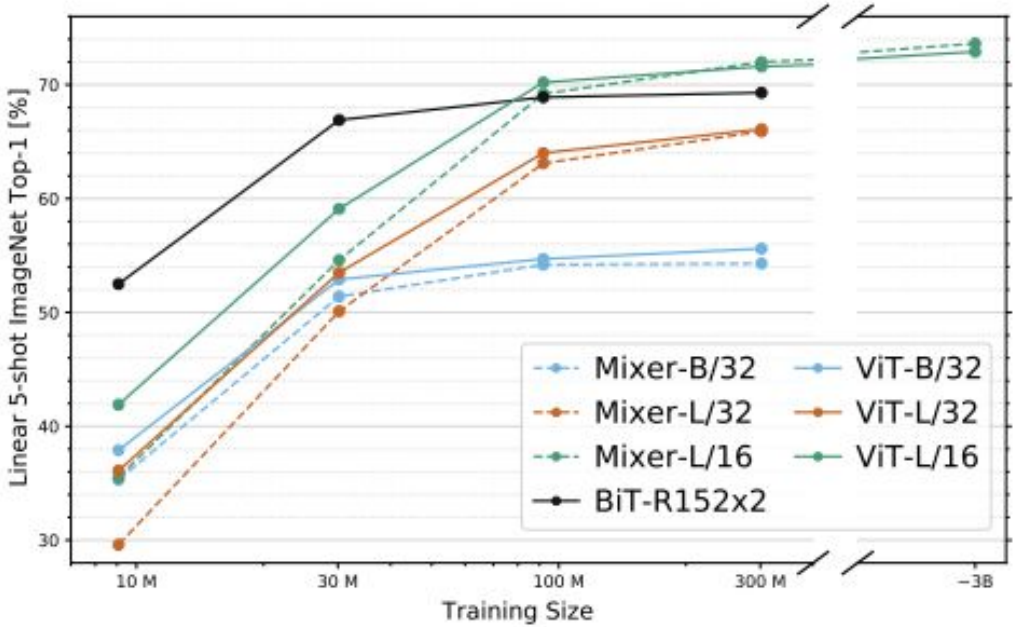
Конфигурации	S32	S16	B32	B16	L32	L16	H14
№ Слоев	8	8	12	12	24	24	32
Размер патча	32×32	16×16	32×32	16×16	32×32	16×16	14×14
№ Каналов	512	512	768	768	1024	1024	1280
№ Патчей	49	196	49	196	49	196	256
Hid_S	256	256	384	384	512	512	640
Hid_C	2048	2048	3072	3072	4096	4096	5120
Параметры (M)	19	18	60	59	206	207	431

Результаты экспериментов

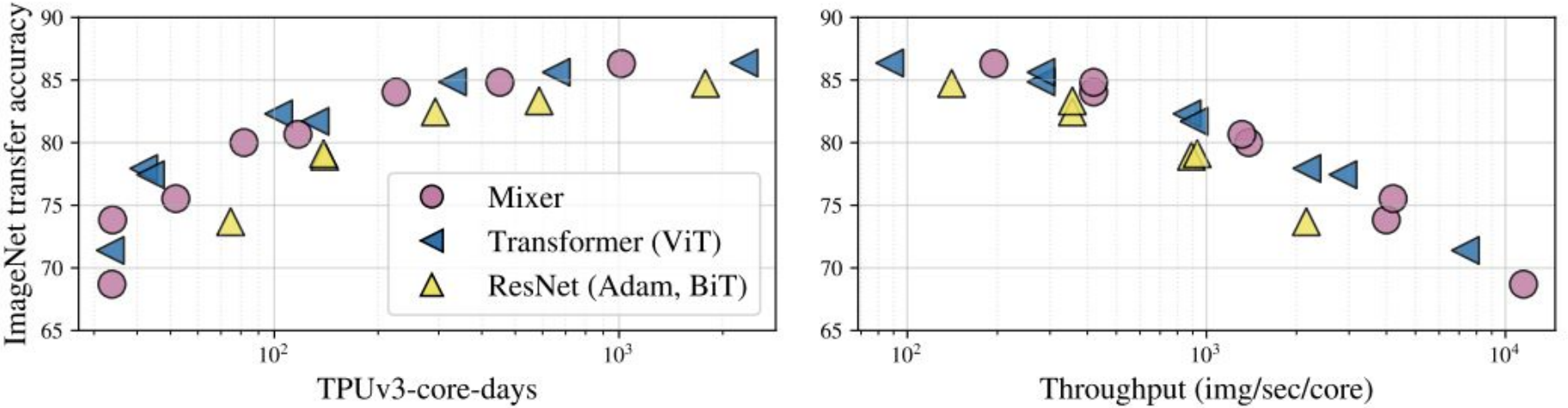
	ImNet top-1	ReaL top-1	Avg 5 top-1	VTAB-1k 19 tasks	Throughput img/sec/core	TPUv3 core-days
Pre-trained on ImageNet-21k (public)						
● HaloNet [51]	85.8	—	—	—	120	0.10k
● Mixer-L/16	84.15	87.86	93.91	74.95	105	0.41k
● ViT-L/16 [14]	85.30	88.62	94.39	72.72	32	0.18k
● BiT-R152x4 [22]	85.39	—	94.04	70.64	26	0.94k
Pre-trained on JFT-300M (proprietary)						
● NFNet-F4+ [7]	89.2	—	—	—	46	1.86k
● Mixer-H/14	87.94	90.18	95.71	75.33	40	1.01k
● BiT-R152x4 [22]	87.54	90.54	95.33	76.29	26	9.90k
● ViT-H/14 [14]	88.55	90.72	95.97	77.63	15	2.30k
Pre-trained on unlabelled or weakly labelled data (proprietary)						
● MPL [34]	90.0	91.12	—	—	—	20.48k
● ALIGN [21]	88.64	—	—	79.99	15	14.82k

Влияние размера датасета на качество





Влияние конфигурации модели на качество



Mixer - чем меньше, тем хуже по сравнению с аналогичными в размере моделями

	Image size	Pre-Train Epochs	ImNet top-1	ReaL top-1	Avg. 5 top-1	Throughput (img/sec/core)	TPUv3 core-days
Pre-trained on ImageNet (with extra regularization)							
● Mixer-B/16	224	300	76.44	82.36	88.33	1384	0.01k[‡]
● ViT-B/16 (☎)	224	300	79.67	84.97	90.79	861	0.02k[‡]
● Mixer-L/16	224	300	71.76	77.08	87.25	419	0.04k[‡]
● ViT-L/16 (☎)	224	300	76.11	80.93	89.66	280	0.05k[‡]
Pre-trained on ImageNet-21k (with extra regularization)							
● Mixer-B/16	224	300	80.64	85.80	92.50	1384	0.15k[‡]
● ViT-B/16 (☎)	224	300	84.59	88.93	94.16	861	0.18k[‡]
● Mixer-L/16	224	300	82.89	87.54	93.63	419	0.41k[‡]
● ViT-L/16 (☎)	224	300	84.46	88.35	94.49	280	0.55k[‡]
● Mixer-L/16	448	300	83.91	87.75	93.86	105	0.41k[‡]
Pre-trained on JFT-300M							
● Mixer-S/32	224	5	68.70	75.83	87.13	11489	0.01k
● Mixer-B/32	224	7	75.53	81.94	90.99	4208	0.05k
● Mixer-S/16	224	5	73.83	80.60	89.50	3994	0.03k
● BiT-R50x1	224	7	73.69	81.92	—	2159	0.08k
● Mixer-B/16	224	7	80.00	85.56	92.60	1384	0.08k
● Mixer-L/32	224	7	80.67	85.62	93.24	1314	0.12k
● BiT-R152x1	224	7	79.12	86.12	—	932	0.14k
● BiT-R50x2	224	7	78.92	86.06	—	890	0.14k
● BiT-R152x2	224	14	83.34	88.90	—	356	0.58k
● Mixer-L/16	224	7	84.05	88.14	94.51	419	0.23k
● Mixer-L/16	224	14	84.82	88.48	94.77	419	0.45k
● ViT-L/16	224	14	85.63	89.16	95.21	280	0.65k
● Mixer-H/14	224	14	86.32	89.14	95.49	194	1.01k
● BiT-R200x3	224	14	84.73	89.58	—	141	1.78k
● Mixer-L/16	448	14	86.78	89.72	95.13	105	0.45k
● ViT-H/14	224	14	86.65	89.56	95.57	87	2.30k
● ViT-L/16 [14]	512	14	87.76	90.54	95.63	32	0.65k

Инвариантность к пиксельным перестановкам

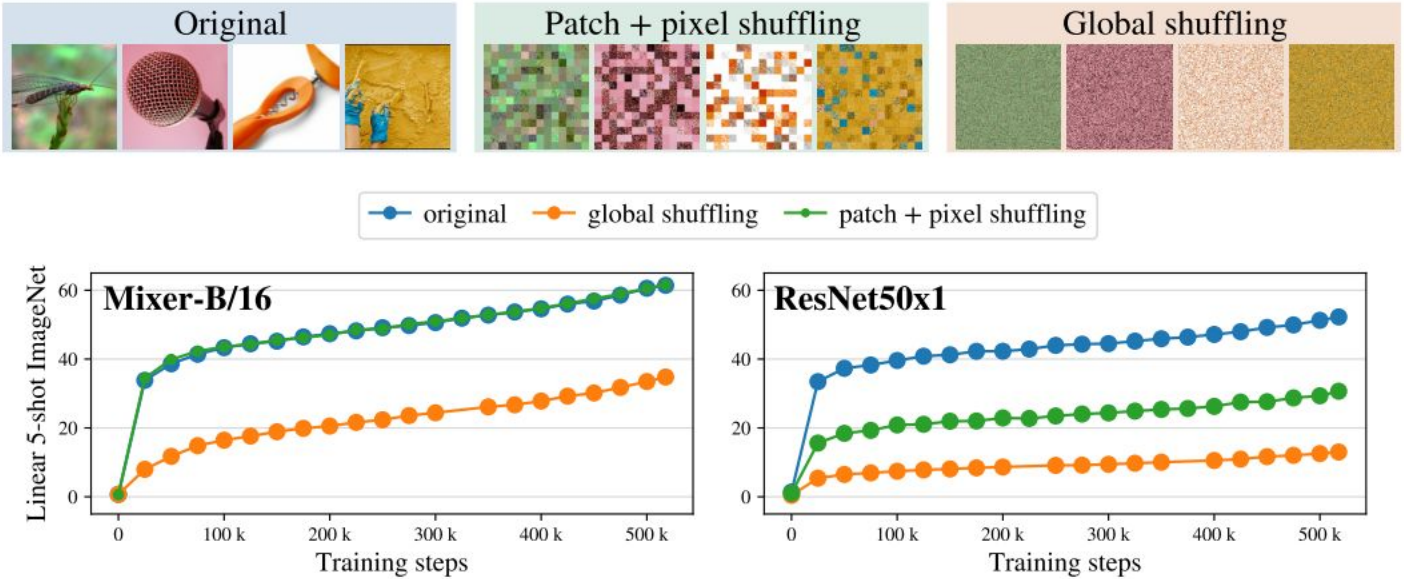


Figure 4: **Top:** Input examples from ImageNet before permuting the contents (left); after shuffling the 16 × 16 patches and pixels within the patches (center); after shuffling pixels globally (right). **Bottom:** Mixer-B/16 (left) and ResNet50x1 (right) trained with three corresponding input pipelines.

Выводы

1) Авторы придумали архитектуру, не подходящую ни под один из двух существующих архетипов
2) Достигнутые результаты совсем немного, но ощутимо отстают от текущего state-of-the-art

Направления развития исследования

1) Теоретический анализ обучаемых представлений в MLP-Mixer.
2) Большее количество статей посвященным нестандартным архитектурам.

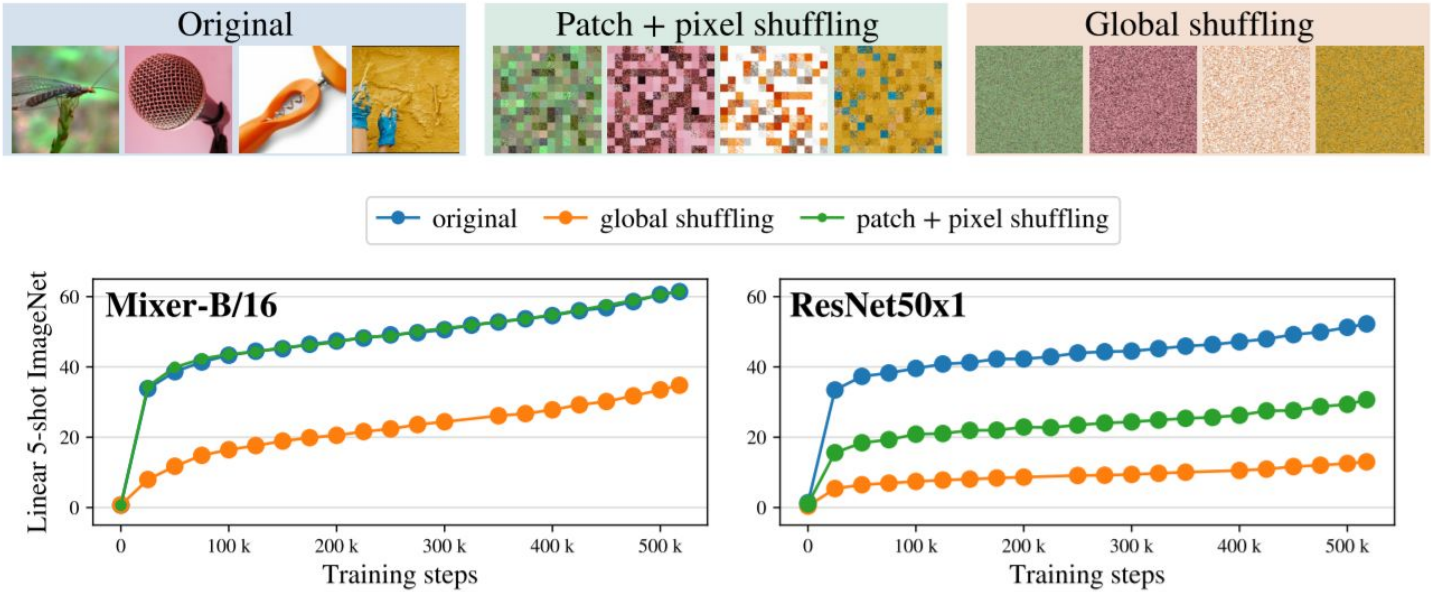
Рецензия

1) 4 различных набора данных
2) SOTA модели convolution-based и attention-based
3) Дополнительные исследования помимо качества
4) Рассмотрены различные метрики и параметры для разных моделей
5) Визуализированы сравнения и некоторые веса
6) Спецификация описана
7) Помимо github в конце статьи есть полноценный код на Jax

Сравнения с первой версией на arxiv

1) Рассмотрен ещё один набор данных JFT-3B

2) Проведено исследование на инвариантность моделей при перемешивании пикселей.



Контекст работы

1) Google Brain, основной вклад от I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer
2) N. Houlsby, A. Kolesnikov, L. Beyer - в числе авторов архитектуры ViT[2] (#2 на ImageNet, отказ от сверток)

[2] A. Dosovitskiy, et al. - "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", 2021.

Контекст работы

3) Архитектурные решения под влиянием сверточных сетей[3] и трансформеров[4]
4) Частный случай вариации Synthesizer[5]; Схожести с самими свертками при некоторых параметрах

[3] A. Krizhevsky, et al. - "ImageNet Classification with Deep Convolutional Neural Networks", 2012.
[4] A. Vaswani, et al. - "Attention Is All You Need", 2017.
[5] Y. Tay, et al. - "Synthesizer: Rethinking Self-Attention in Transformer Models", 2021.

Контекст работы

5) Другие задачи зрения или меньшие задачи распознавания не рассмотрены
6) Хороший компромисс качества производительности, массовый продукт