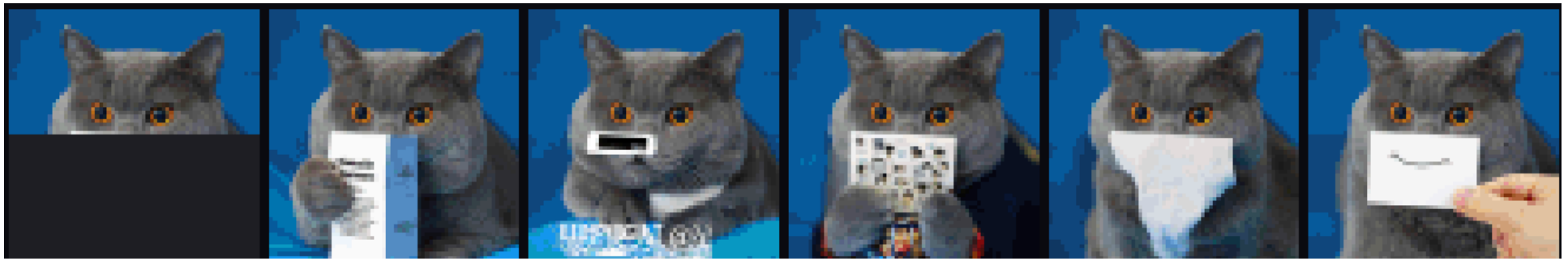
Image Pretraining from Pixels



Сеньченко Тимофей

Группа 172

Unsupervised and self-supervised learning

- Успешно используются в NLP (Авторегрессионные модели, BERT)
- Практически нет аналогичных современных применений в задачах связанных с изображениями
- Хорошие генеративные модели выучивают более хорошие представления?
- Воспользуемся domain-agnostic трансформером для авторегрессионного предсказания пикселей

Pretraining

$$x = (x_1, ..., x_n)$$

$$p(x) = \prod_{i=1}^{n} p(x_{\pi_i} | x_{\pi_1}, ..., x_{\pi_{i-1}}, \theta)$$

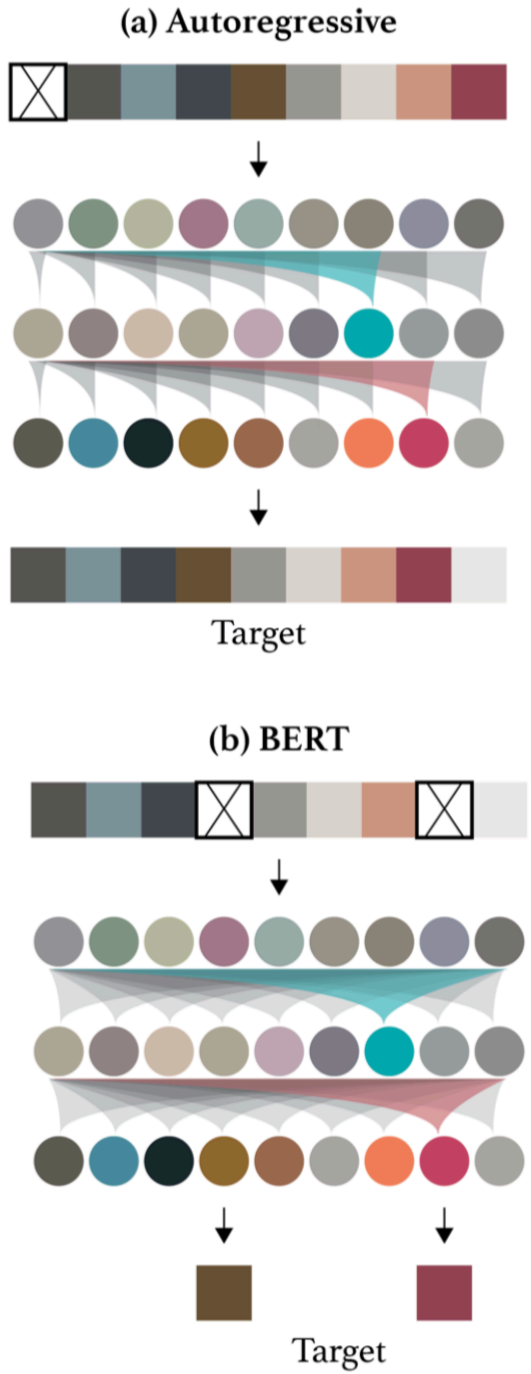
Для изображений будем минимизировать:

$$L_{AR} = \mathop{\mathbb{E}}_{x \sim X} [-\log p(x)]$$

Или:

$$L_{BERT} = \mathop{\mathbb{E}}_{x \sim X} \mathop{\mathbb{E}}_{M} \sum_{i \in M} \left[-\log p\left(x_i | x_{[1,n] \setminus M}\right)\right]$$

$$M \subset [1, n]$$



Architecture

Вход: токены $x_1, \ldots, x_n$

-> эмбеддинги размерности d для каждой позиции

Декодер состоит из L блоков, где l-й блок подает на выход промежуточные эмбеддинги $h_1^l, \ldots, h_n^l$ размерности d

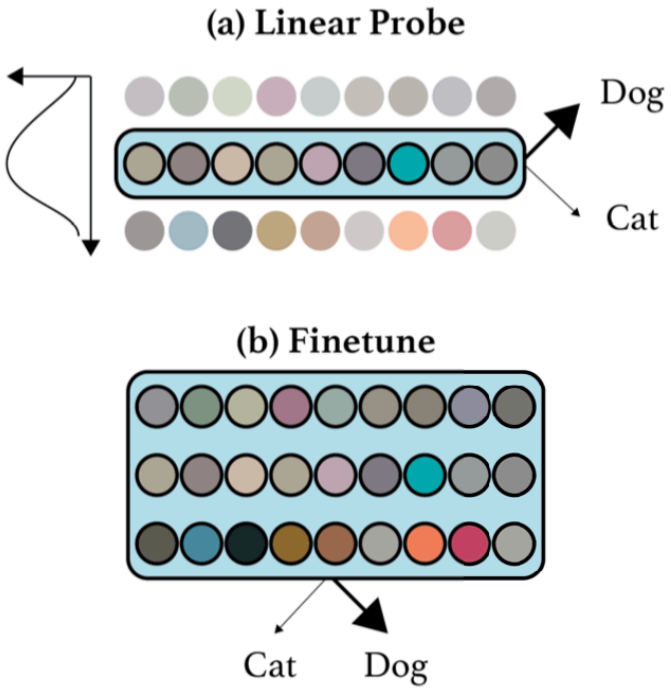
Авторы статьи используют структуру блока GPT-2: (тензор h^l - вход)

$$n^l = \text{layer\_norm}(h^l)$$

$$a^l = h^l + \text{multihead\_attention}(n^l)$$

$$h^{l+1} = a^l + \text{mlp}(\text{layer\_norm}(a^l))$$

Fine-tuning and Linear Probing



Fine-tuning:

- Добавляем голову классификатор,
- Дообучаем на лейблах, используя выученные представления
- Loss: кросс-энтропия L_{CLF}, а лучше: $L_{GEN} + L_{CLF}$

$$L_{GEN} \in \{L_{AR}, L_{BERT}\}$$

Linear Probing:

- Выбираем некоторый слой i и по его выходам учим Логистическую регрессию на лейблах

Context reduction

Входные данные значительно большей размерности, чем в NLP задачах (224^2 X 3)

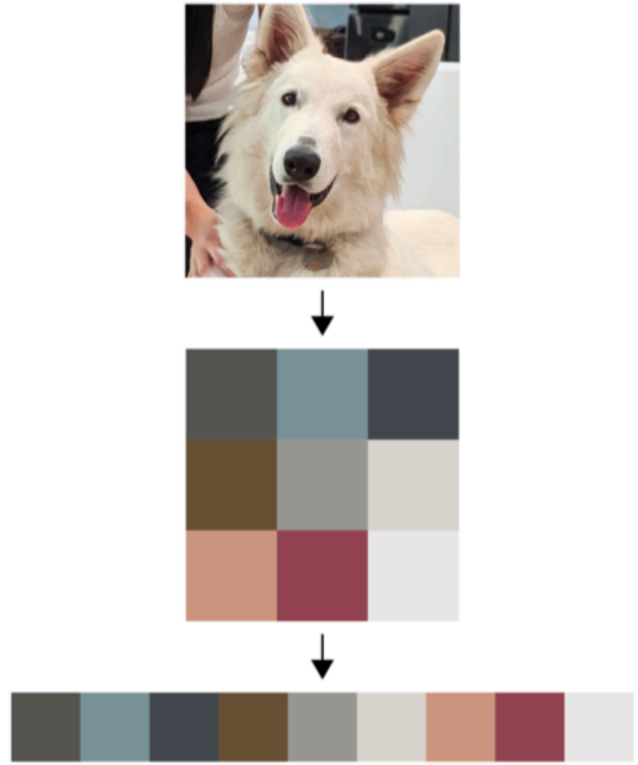
-> не можем использовать внимание с исходными данными

-> уменьшаем Image Resolution (32^2 X 3 или

48^2 X 3 или 64^2 X 3)

-> переводим картинку из палитры RGB в новую 9-ти битную палитру с помощью k-means кластеризации

-> получаем изображения значительно меньшей размерности (32^2 или 48^2 или 64^2)



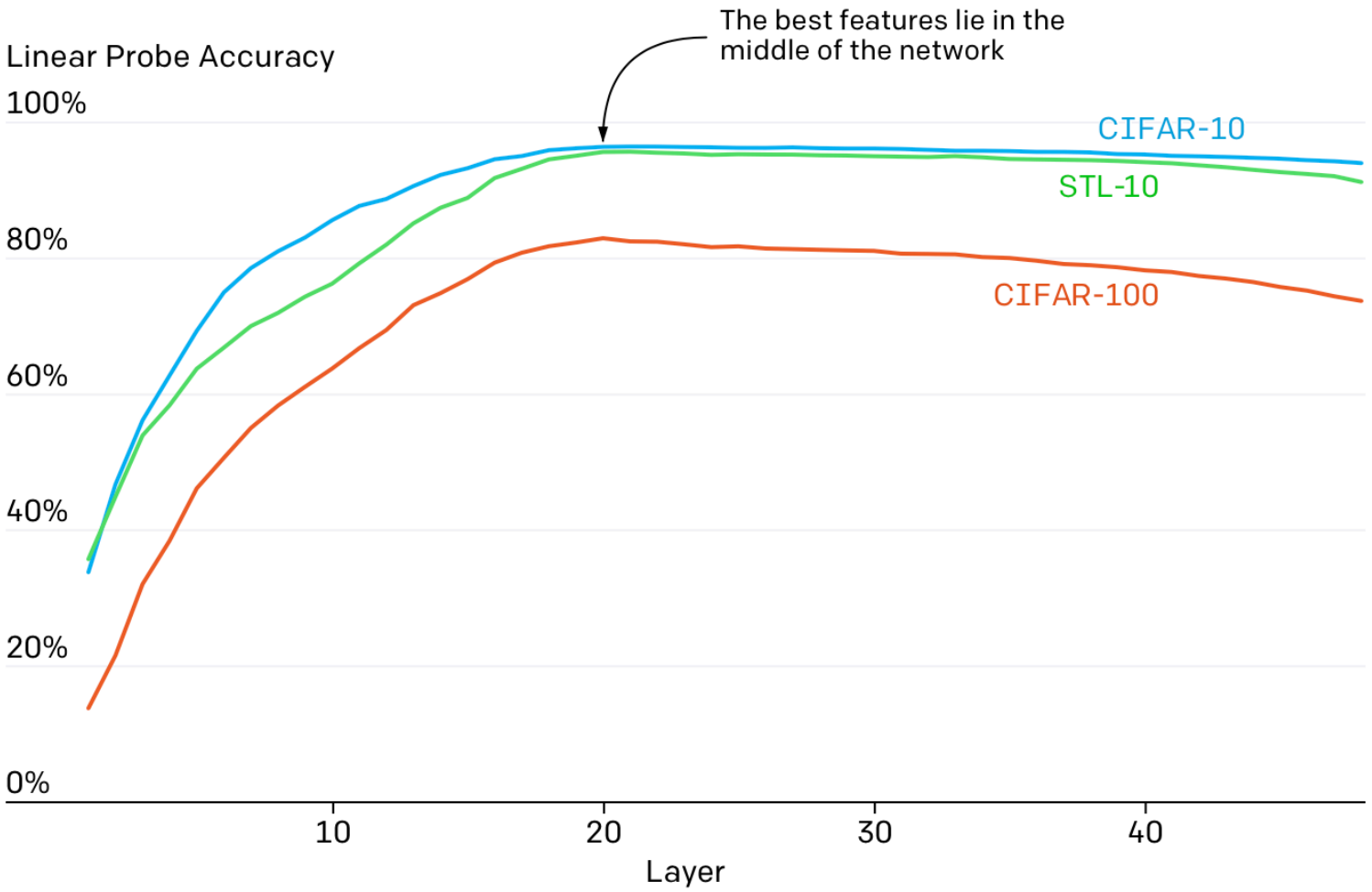
Data and models

- iGPT-XL. L = 60, d = 3072, parameters: 6.8B
- iGPT-L. L = 48, d = 1536, parameters: 1.4B
- iGPT-M. L = 36, d = 1024, parameters: 455M
- iGPT-S. L = 4, d =512, parameters: 76M

Datasets:

- CIFAR-10
- CIFAR-100
- STL-10

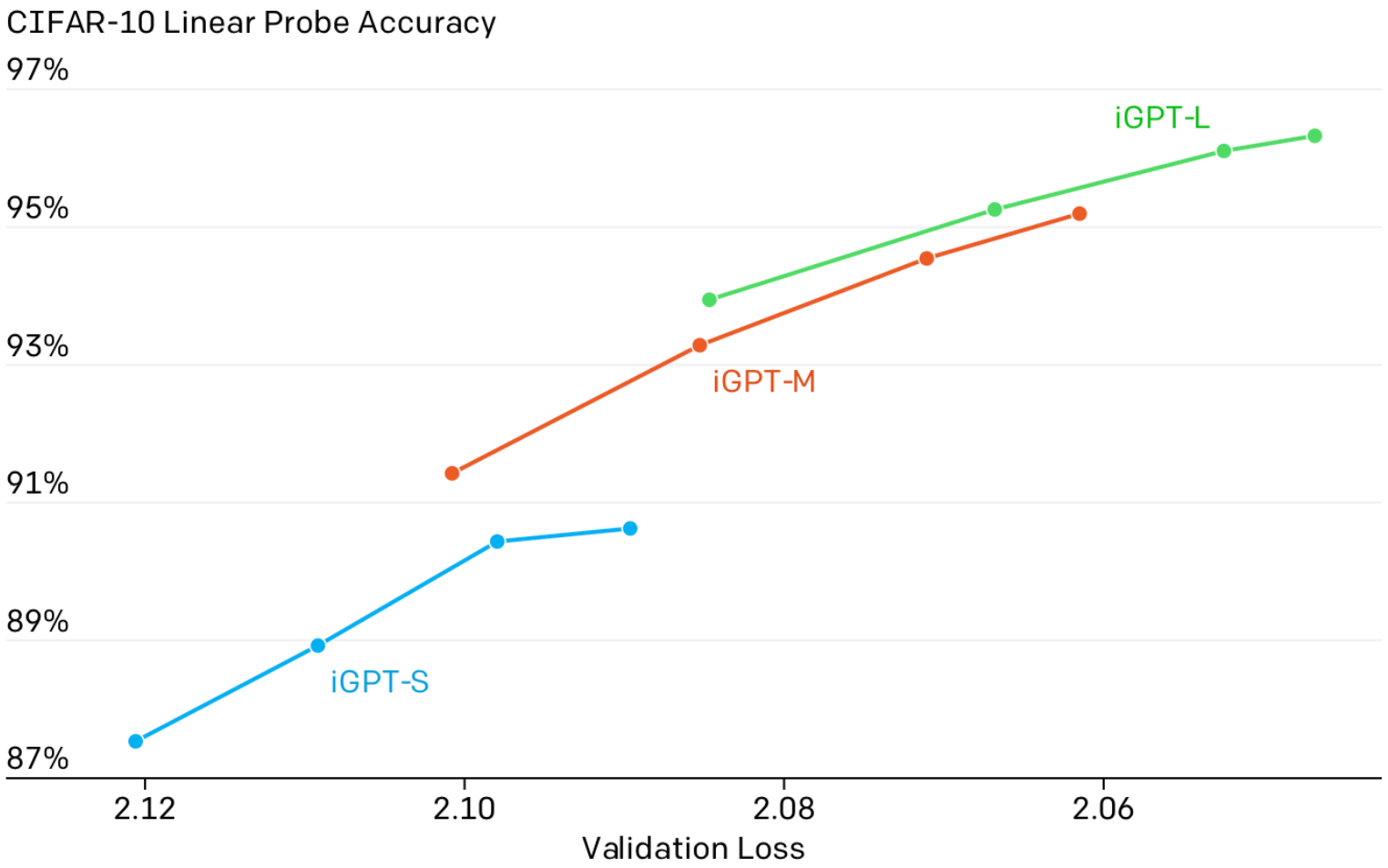
Results


Linear Probe Accuracy
100%
80%
60%
40%
20%
0%
The best features lie in the middle of the network
CIFAR-10
STL-10
CIFAR-100
10
20
30
40
Layer

Results

Steps:

- 131k
- 262k
- 524k
- 1000k



Results

Качество моделей предобученных на ImageNet

EVALUATION	MODEL	ACCURACY	PRE-TRAINED ON IMAGENET W/O LABELS	PRE-TRAINED ON IMAGENET W/ LABELS
CIFAR-10 Linear Probe	ResNet-152[50]	94.0		✓
	SimCLR[12]	95.3	✓	
	iGPT-L 32x32	**96.3**	✓	
CIFAR-100 Linear Probe	ResNet-152	78.0		✓
	SimCLR	80.2	✓	
	iGPT-L 32x32	**82.8**	✓	
STL-10 Linear Probe	AMDIM-L[13]	94.2	✓	
	iGPT-L 32x32	**95.5**	✓	
CIFAR-10 Fine-tune	AutoAugment[51]	98.5		
	SimCLR	98.6	✓	
	GPipe[15]	**99.0**		✓
	iGPT-L	**99.0**	✓	
CIFAR-100 Fine-tune	iGPT-L	88.5	✓	
	SimCLR	89.0	✓	
	AutoAugment	89.3		
	EfficientNet[52]	**91.7**		✓

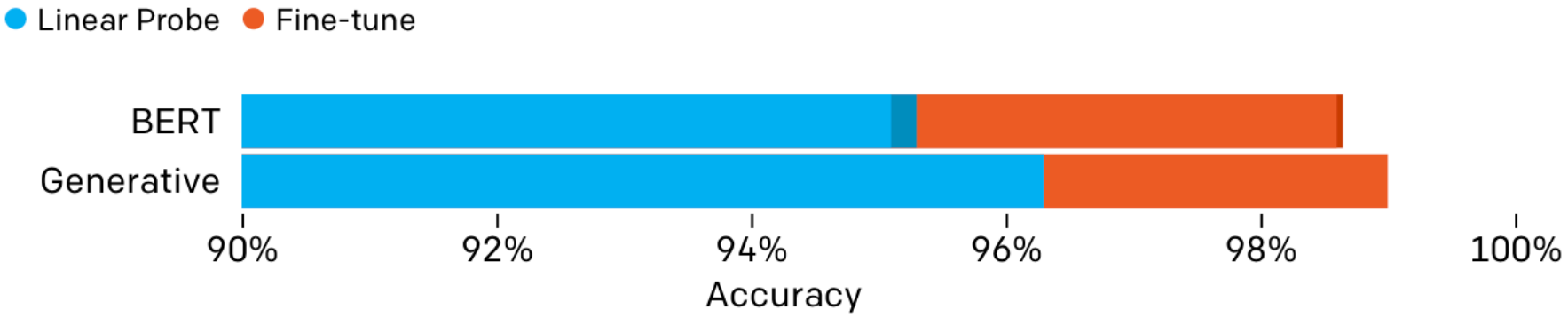
Results

Качество на ImageNet

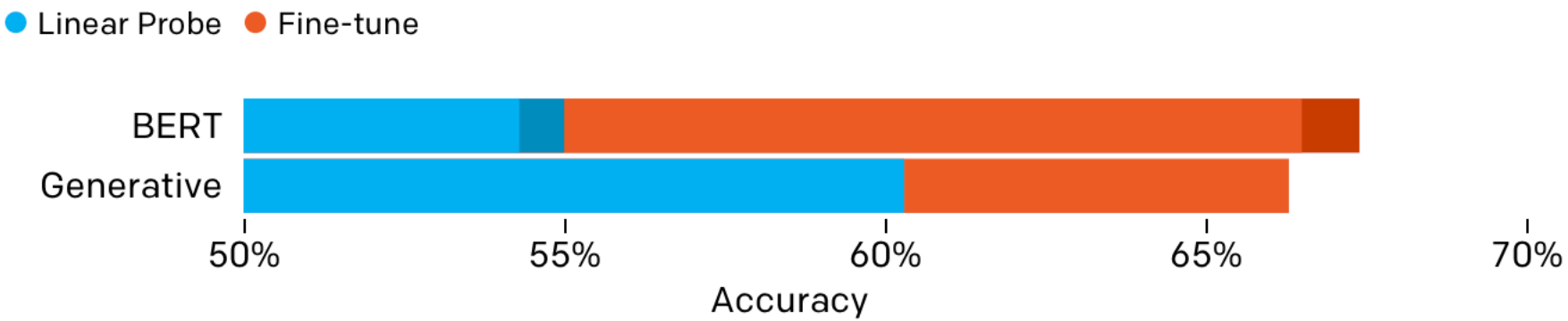
METHOD	INPUT RESOLUTION	FEATURES	PARAMETERS	ACCURACY
Rotation[53]	original	8192	86M	55.4
iGPT-L	32x32	1536	1362M	60.3
BigBiGAN[37]	original	16384	86M	61.3
iGPT-L	48x48	1536	1362M	65.2
AMDIM[13]	original	8192	626M	68.1
MoCo[24]	original	8192	375M	68.6
iGPT-XL	64x64	3072	6801M	68.7
SimCLR[12]	original	2048	24M	69.3
CPC v2[25]	original	4096	303M	71.5
iGPT-XL	64x64	3072 x 5	6801M	72.0
SimCLR	original	8192	375M	**76.5**

Results

CIFAR-10

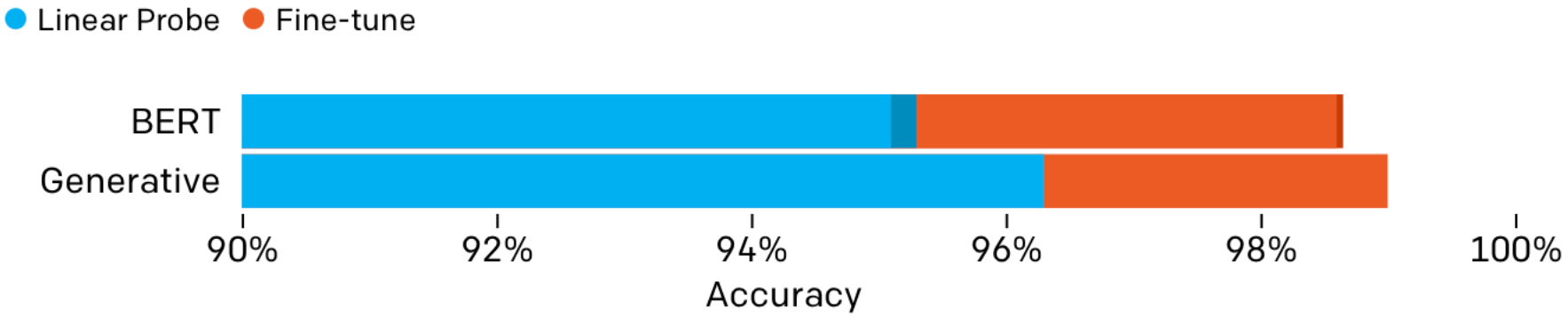


ImageNet

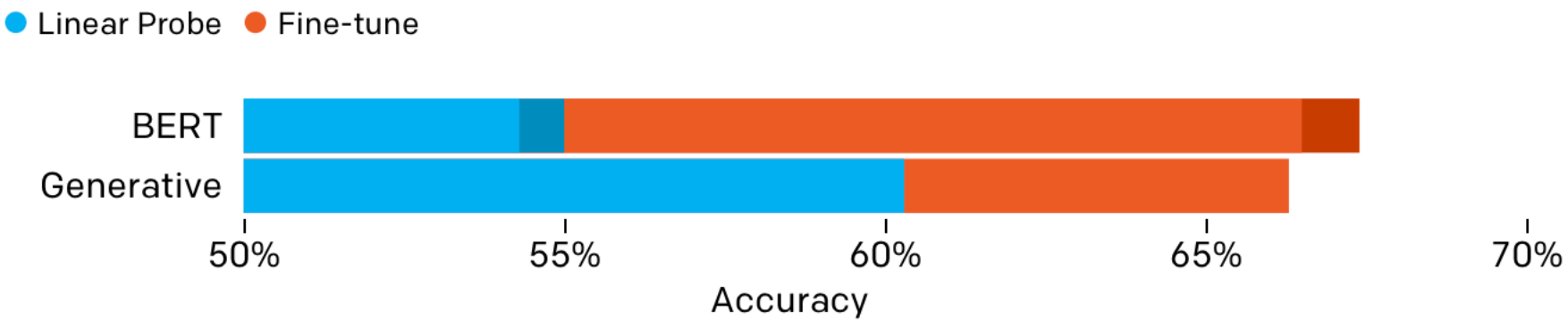


Results

CIFAR-10



ImageNet



Results

MODEL	40 LABELS	250 LABELS	4000 LABELS
Improved GAN[55]	—	—	81.4 ± 2.3
Mean Teacher[56]	—	67.7 ± 2.3	90.8 ± 0.2
MixMatch[57]	52.5 ± 11.5	89.0 ± 0.9	93.6 ± 0.1
iGPT-L	73.2 ± 1.5	87.6 ± 0.6	94.3 ± 0.1
UDA[58]	71.0 ± 5.9	91.2 ± 1.1	95.1 ± 0.2
FixMatch[59] RA	86.2 ± 3.4	94.9 ± 0.7	**95.7 ± 0.1**
FixMatch CTA	**88.6 ± 3.4**	**94.9 ± 0.3**	95.7 ± 0.2

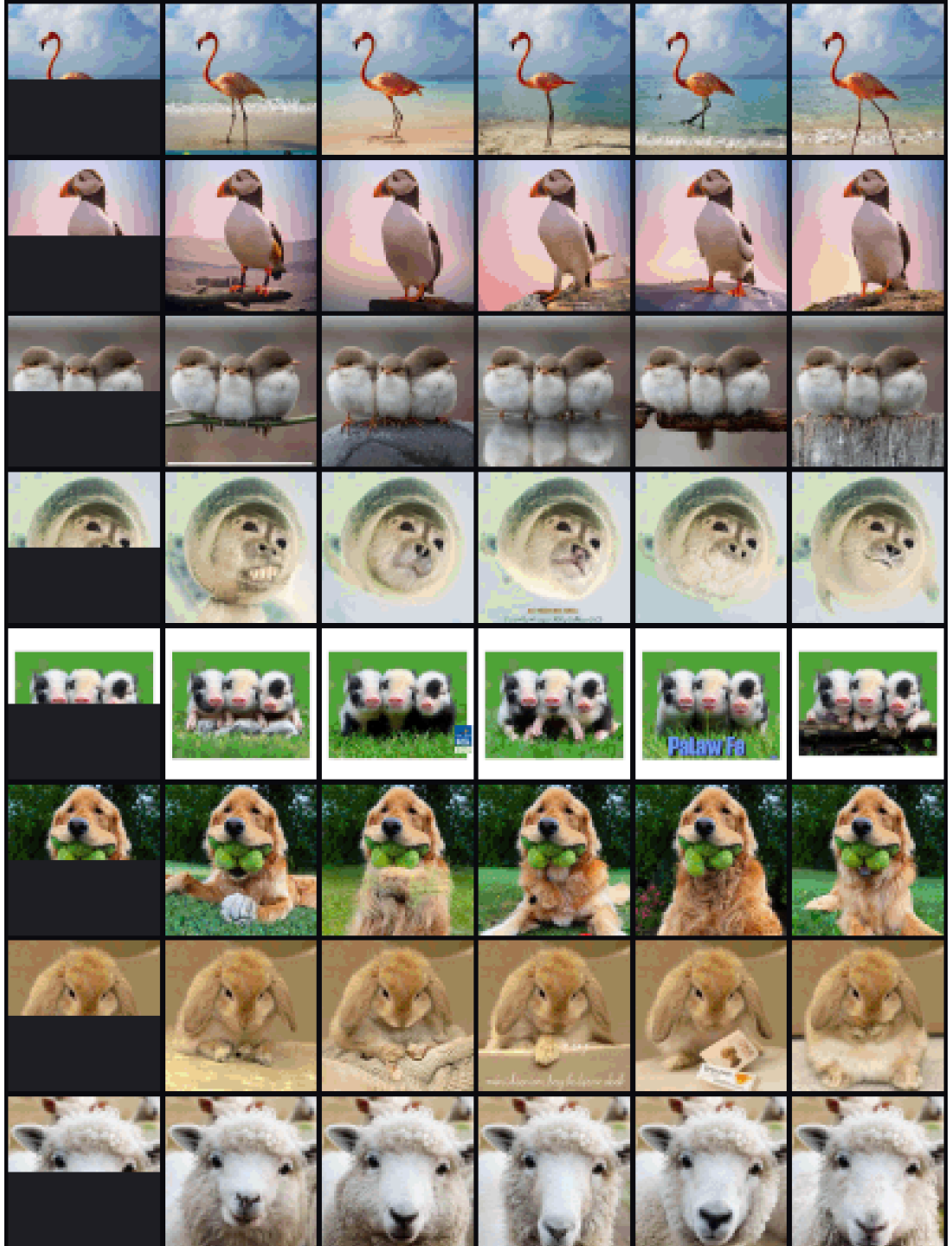
Results

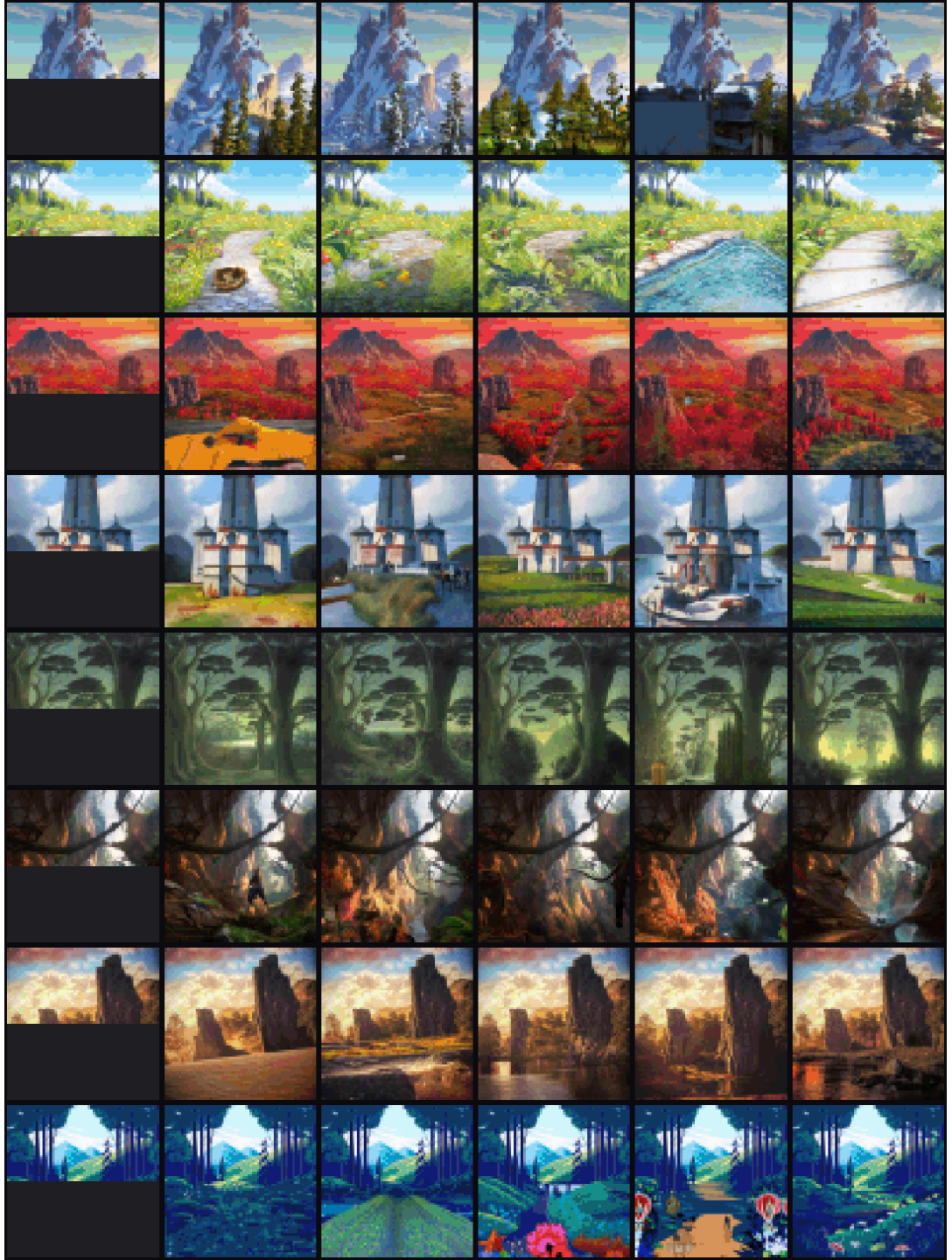
MODEL	40 LABELS	250 LABELS	4000 LABELS
Improved GAN[55]	—	—	81.4 ± 2.3
Mean Teacher[56]	—	67.7 ± 2.3	90.8 ± 0.2
MixMatch[57]	52.5 ± 11.5	89.0 ± 0.9	93.6 ± 0.1
iGPT-L	73.2 ± 1.5	87.6 ± 0.6	94.3 ± 0.1
UDA[58]	71.0 ± 5.9	91.2 ± 1.1	95.1 ± 0.2
FixMatch[59] RA	86.2 ± 3.4	94.9 ± 0.7	**95.7 ± 0.1**
FixMatch CTA	**88.6 ± 3.4**	**94.9 ± 0.3**	95.7 ± 0.2

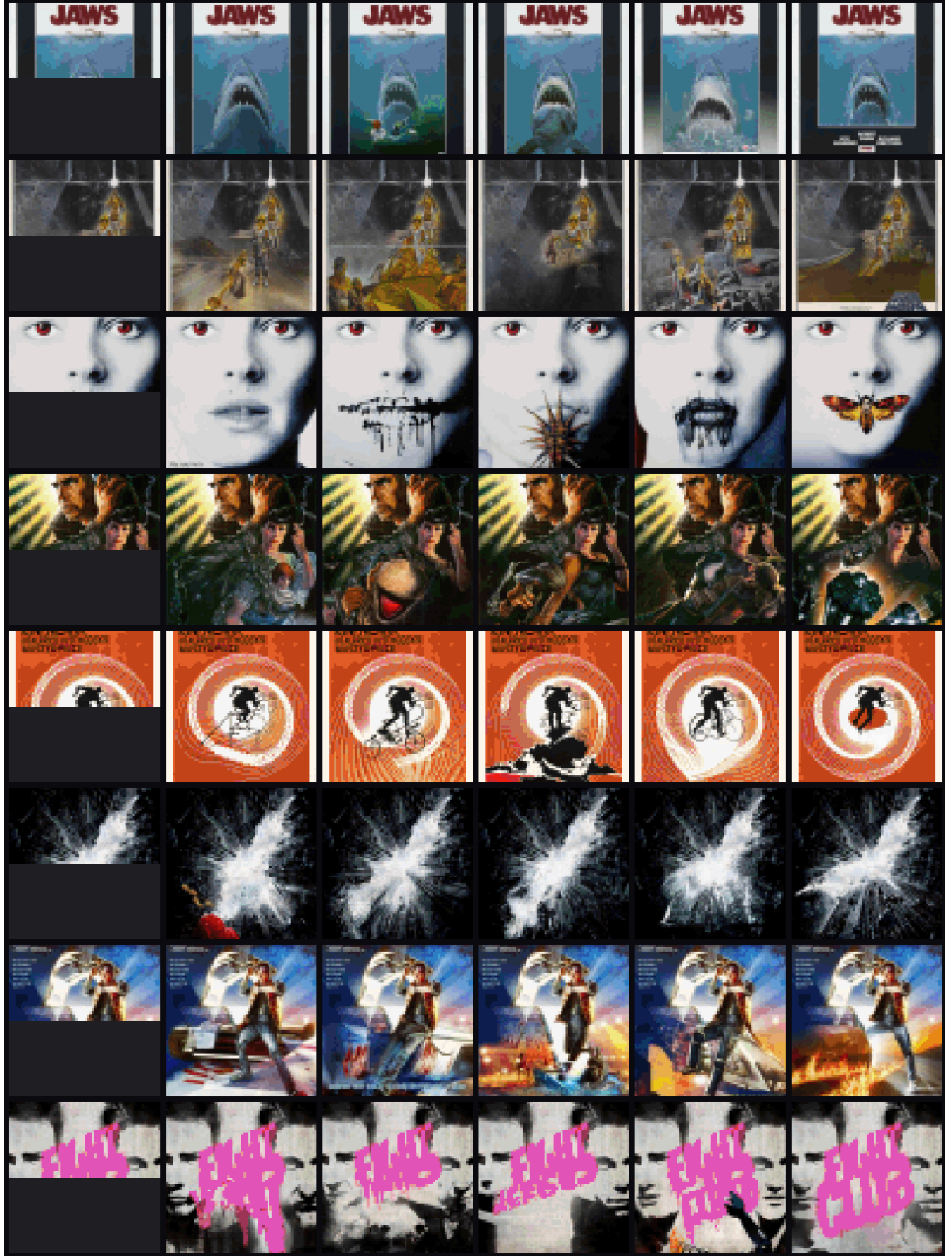
Conclusion

- Качество моделей в различных задачах не уступает сверточным сетям
- Приходится использовать данные с более низким разрешением, чем сверточные сети
- Большое количество параметров, долгое обучение

-> На данный момент достаточно непрактично для использования в реальных задачах


Palaw'Fa




JAWS
JAWS
JAWS
JAWS
JAWS
JAWS

Links

- Статья
- Блог пост