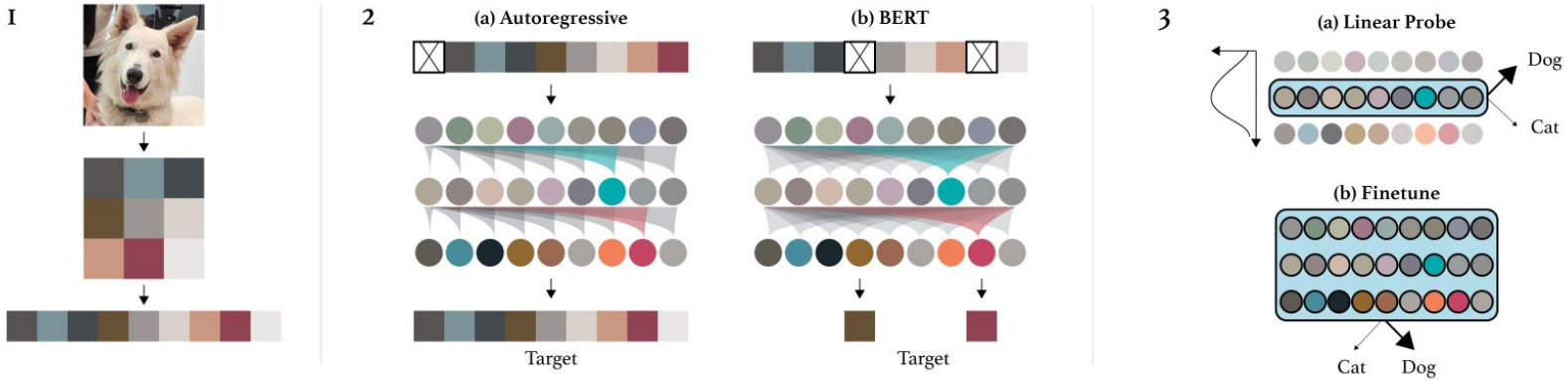
Generative Pretraining from Pixels

Сухарьков Александр, БПМИ171

Подход авторов



1. Препроцессинг картинок
2. Предобучение с авторегрессионной или BERT задачей
3. Дообучение или Linear Probe

Предобучение - авторегрессионная задача

датасет X

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

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

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

Предобучение - BERT задача

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

Архитектура

$$
\begin{aligned}
n^l &= \text{layer\_norm}(h^l) \\
a^l &= h^l + \text{multihead\_attention}(n^l) \\
h^{l+1} &= a^l + \text{mlp}(\text{layer\_norm}(a^l))
\end{aligned}
$$

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

Дообучение

Average Pooling n^L

$$f^L = \langle n_i^L \rangle_i$$

Кросс-энтропия L_{CLF}

$L_{GEN} + L_{CLF}$, где $L_{GEN} \in \{L_{AR}, L_{BERT}\}$

Датасеты

ImageNet ILSVRC 2012, 4% валидация

Аугментации: изменения размера, чтобы минимальная размерность была 224, center crop 224x224

CIFAR-10, CIFAR-100, STL-10 - 10% валидация

Аугментации: 4 пикселя отражения добавляются с каждой стороны, crop 32x32 случайно из полученного изображения или его горизонтального отображения

Размерности

$224^2 \times 3$ - даже один слой не влезет в GPU

Input Resolution (IR) - $32^2 \times 3, 48^2 \times 3, 64^2 \times 3$

Model Resolution (MR) - $32^2, 48^2 \; 64^2$

Модели

iGPT-XL: 60 слоев, размер эмбеддинга - 3072, 6.8B параметров

iGPT-L: 48 слоев, размер эмбеддинга - 1536, 1.4M параметров

iGPT-M: 36 слоев, размер эмбеддинга - 1024, 455M параметров

iGPT-S: 24 слоя, размер эмбеддинга - 512, 76M параметров

Обучение

Предобучение iGPT-XL - batch size = 64, n_iterations = 2M

Предобучение других моделей - batch size = 128, n_iterations = 1M

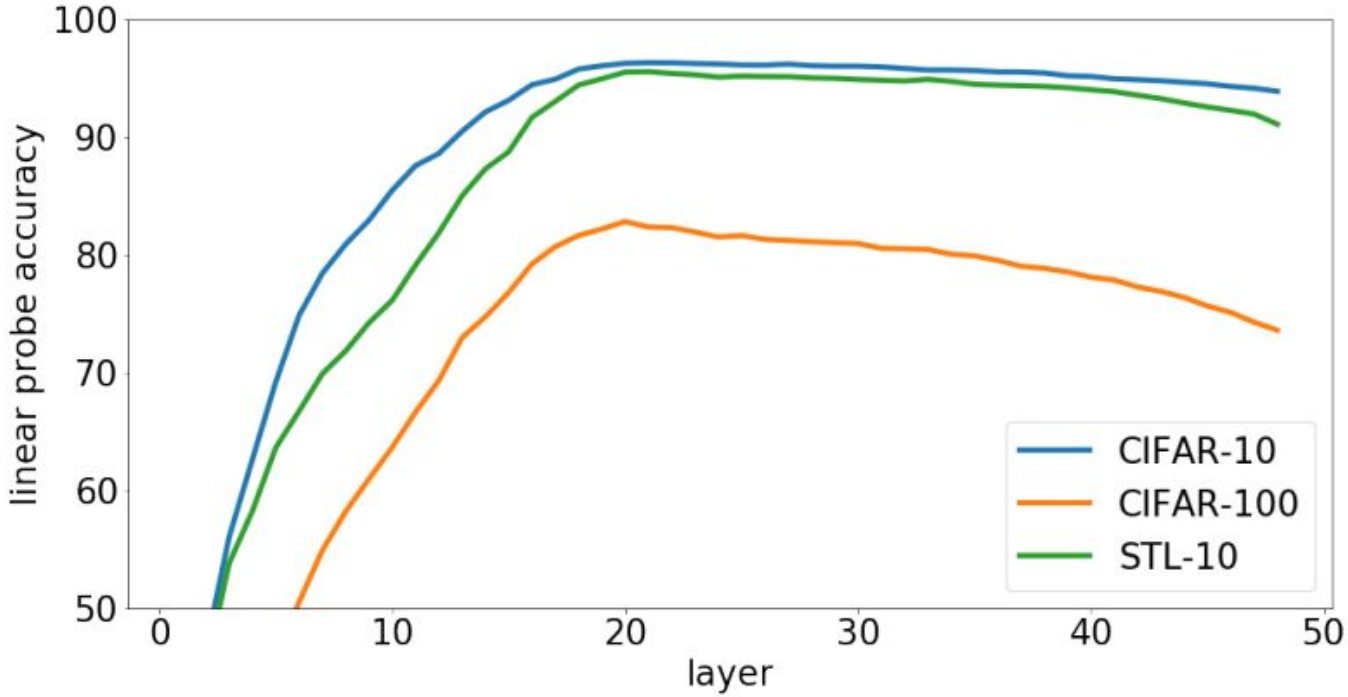
Adam с $\beta_1 = 0.9$ и $\beta_2 = 0.95$

Для дообучения тот же batch size и тот же Adam

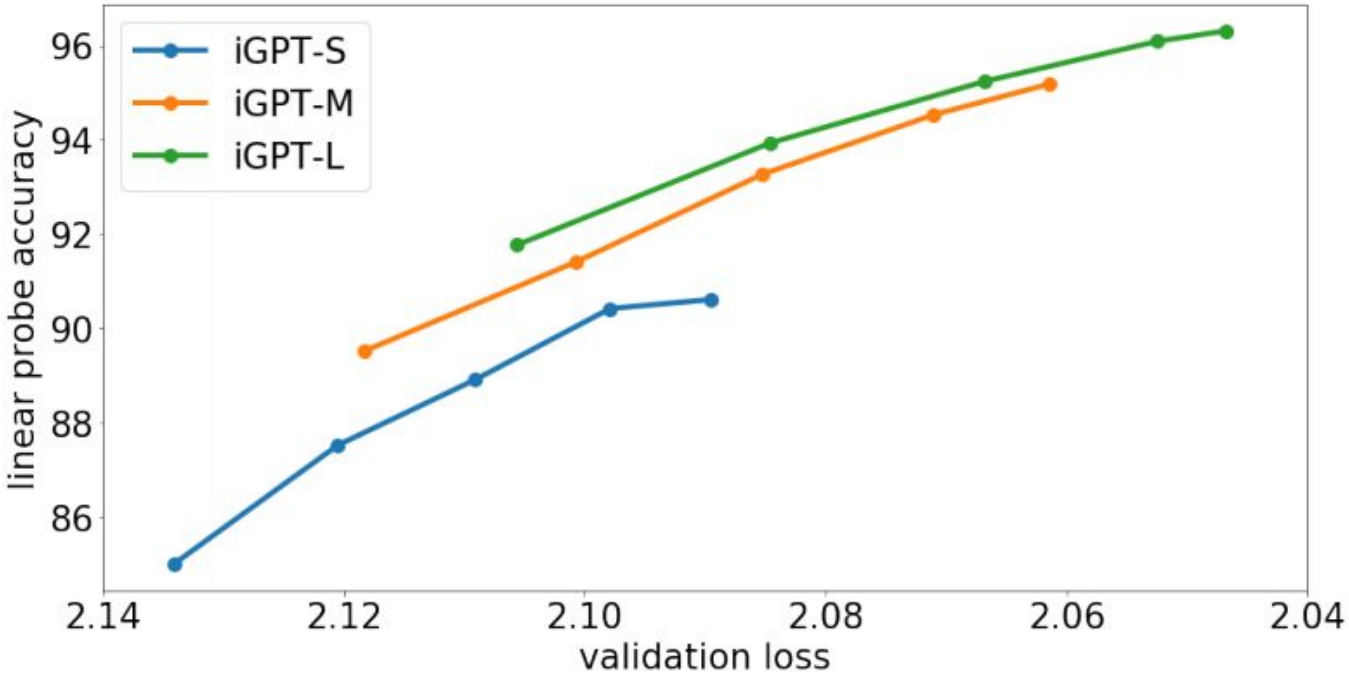
Для linear probing для ImageNet SGD с моментумом = 0.9 и высоким lr (30, 10, 3, …)

Для CIFAR и STL - L-BFGS

Эксперименты



Эксперименты



Эксперименты

Model	Acc	Unsup Transfer	Sup Transfer
CIFAR-10			
ResNet-152	94		√
SimCLR	95.3	√	
iGPT-L	96.3	√	
CIFAR-100			
ResNet-152	78.0		√
SimCLR	80.2	√	
iGPT-L	82.8	√	
STL-10			
AMDIM-L	94.2	√	
iGPT-L	95.5	√	

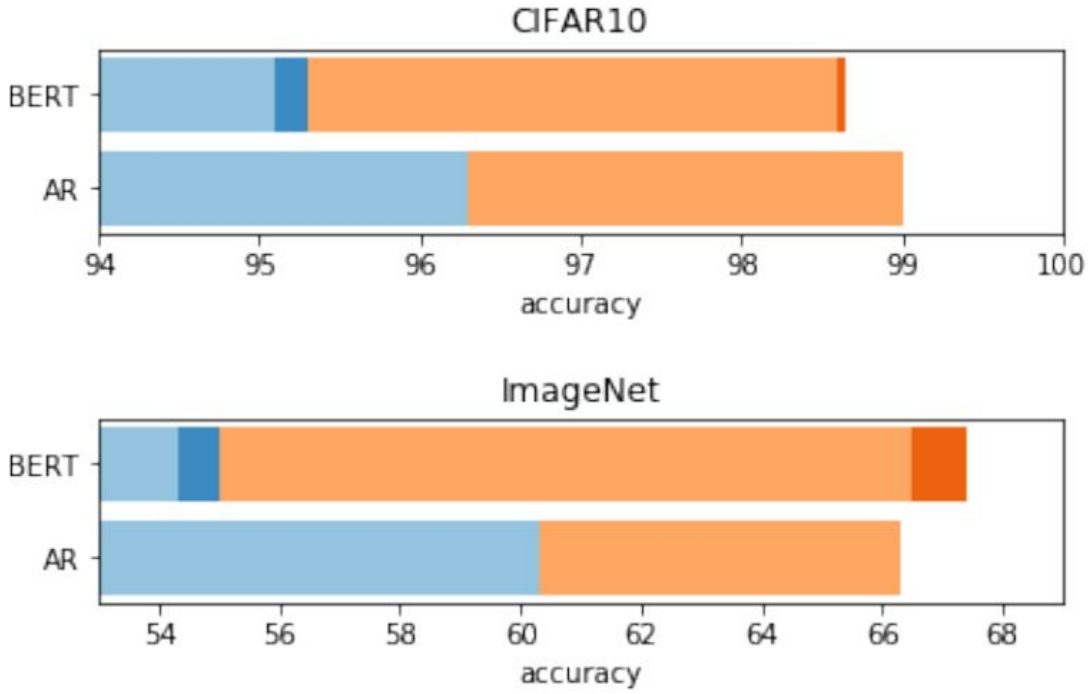
Эксперименты

Method	IR	Params (M)	Features	Acc
Rotation	orig.	86	8192	55.4
iGPT-L	$32^2 \cdot 3$	1362	1536	60.3
BigBiGAN	orig.	86	8192	61.3
iGPT-L	$48^2 \cdot 3$	1362	1536	65.2
AMDIM	orig.	626	8192	68.1
MoCo	orig.	375	8192	68.6
iGPT-XL	$64^2 \cdot 3$	6801	3072	68.7
SimCLR	orig.	24	2048	69.3
CPC v2	orig.	303	8192	71.5
iGPT-XL	$64^2 \cdot 3$	6801	15360	72.0
SimCLR	orig.	375	8192	76.5

Эксперименты

Model	Acc	Unsup Transfer	Sup Transfer
CIFAR-10			
AutoAugment	98.5		
SimCLR	98.6	√	
GPipe	99.0		√
iGPT-L	99.0	√	
CIFAR-100			
iGPT-L	88.5	√	
SimCLR	89.0	√	
AutoAugment	89.3		
EfficientNet	91.7		√

Эксперименты



Выводы

Подход авторов конкурентоспособен, также доказывает, что такие методы имеют место быть

НО

очень ресурсозатратный;

использует слишком маленькие картинки из-за этого.

Вопросы

1. Какие две разные постановки задачи используются на этапе предобучения? Какие функции могут использоваться для минимизации?

2. Как выглядит архитектура декодера трансформера?

3. Что происходит на этапе дообучения и какая функция минимизируется?

Список источников

https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf