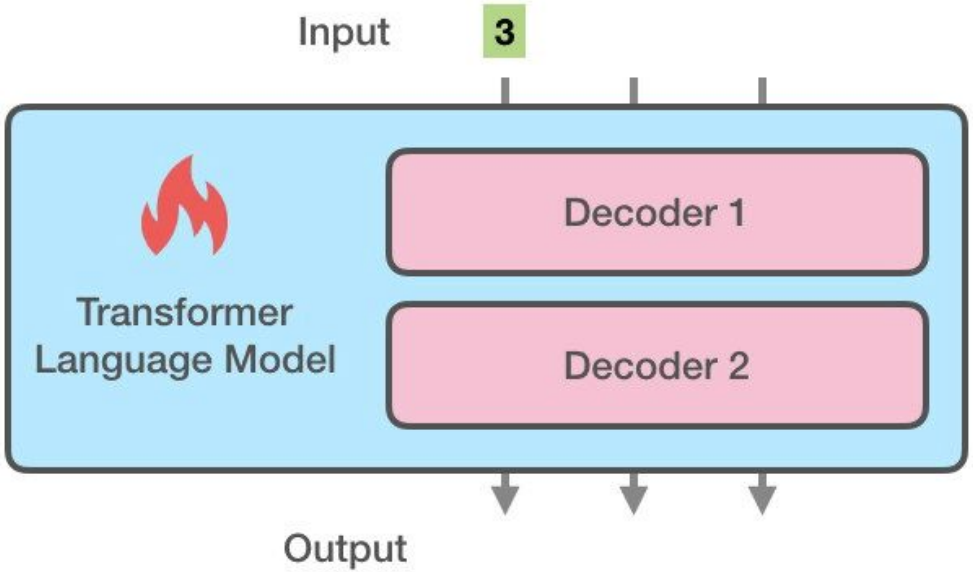
Image GPT

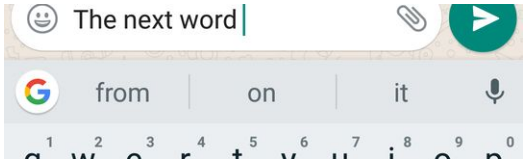
Generative Pretraining from Pixels

Szymon Tworkowski

Koło Naukowe Uczenia Maszynowego UW
18 grudnia 2020

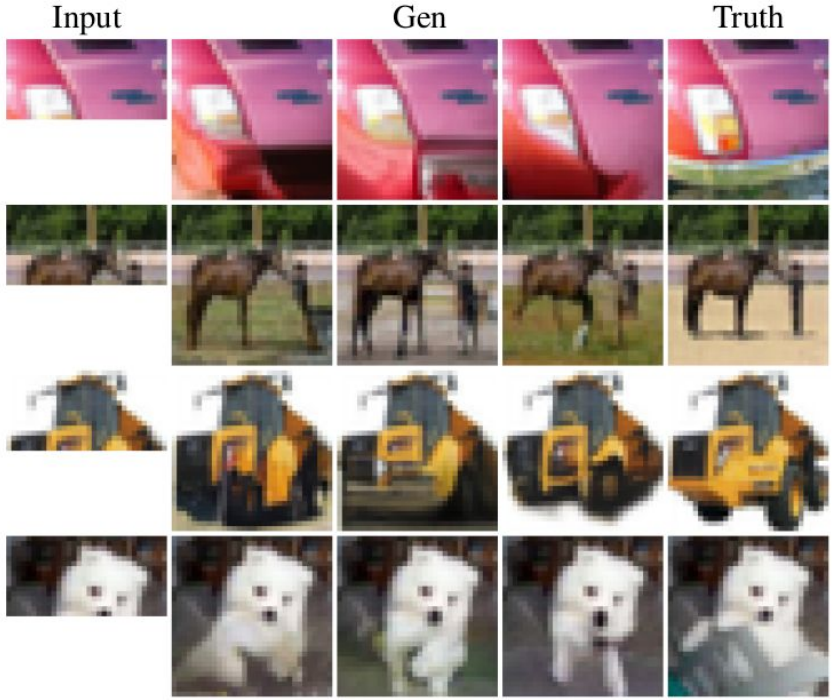
Language Model - Recap





Related Work

- PixelRNN/CNN (2016)
- Image Transformer (2018)



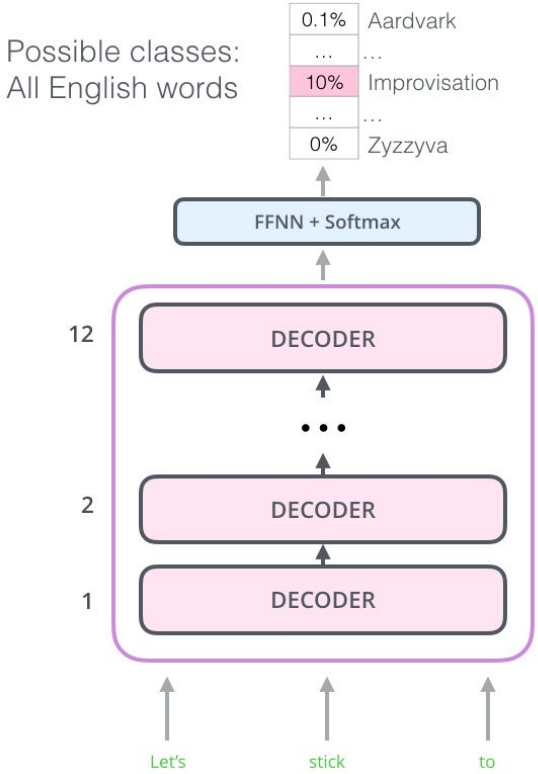
example completions of the Image Transformer

Model

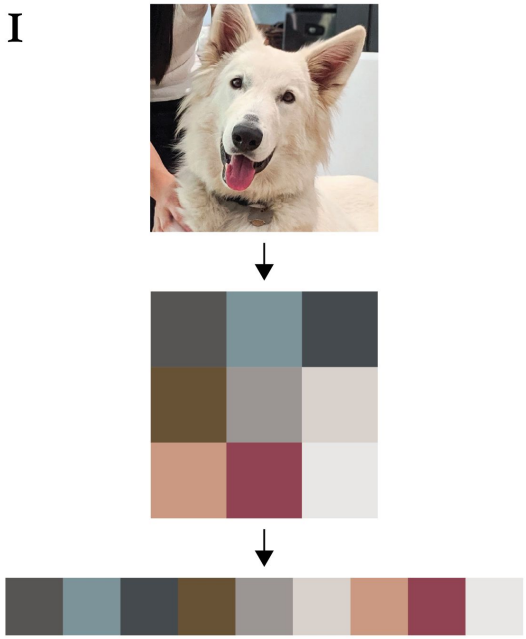
Architecture: nearly identical to the GPT-2 transformer decoder:

Sizes:

- **iGPT-XL** **60** layers, d_model = **3072**, **6.8B** params
- iGPT-L 48 layers, d_model = 1536, 1.4B params
- iGPT-M 36 layers, d_model = 1024, 455M params
- iGPT-S 24 layers, d_model = 512, 76M params

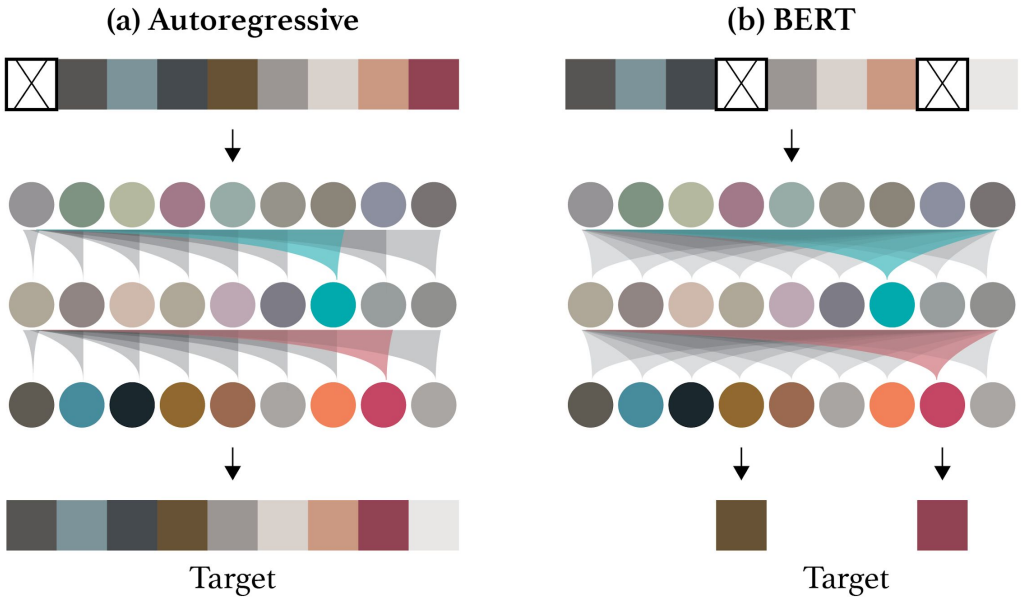


Input and Dataset

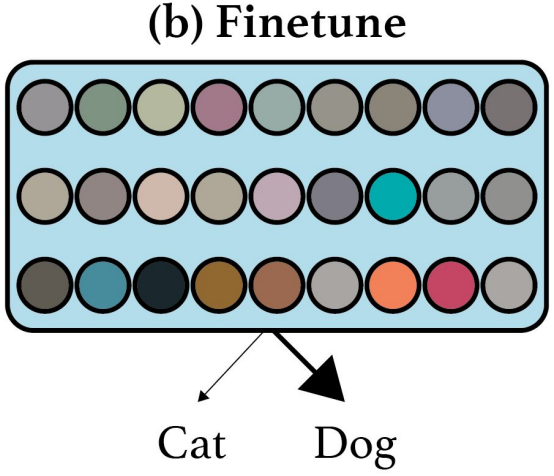


- Pre-training: unlabeled ImageNet (ILSVRC 2012) - random 224x224 crops, resized to 64x64
- Fine-tuning: CIFAR-10, CIFAR-100, ImageNet (labeled)
- Custom 9-bit color palette created by clustering (R,G,B) pixel values using k-means with k=512 - to remove the channel dimension. Now we have 3 times shorter input sequence and model resolution

Pre-training



Fine-tuning



2.3. Fine-tuning

When fine-tuning, we average pool n^L across the sequence dimension to extract a d-dimensional vector of features per example:

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

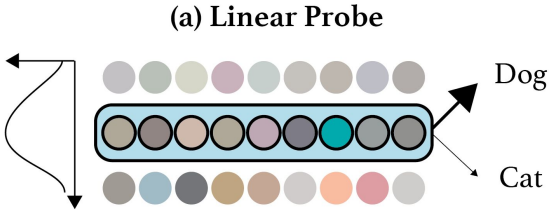
We learn a projection from f^L to class logits, which we use to minimize a cross entropy loss L_{CLF}.

While fine-tuning on L_{CLF} yields reasonable downstream performance, we find empirically that the joint objective

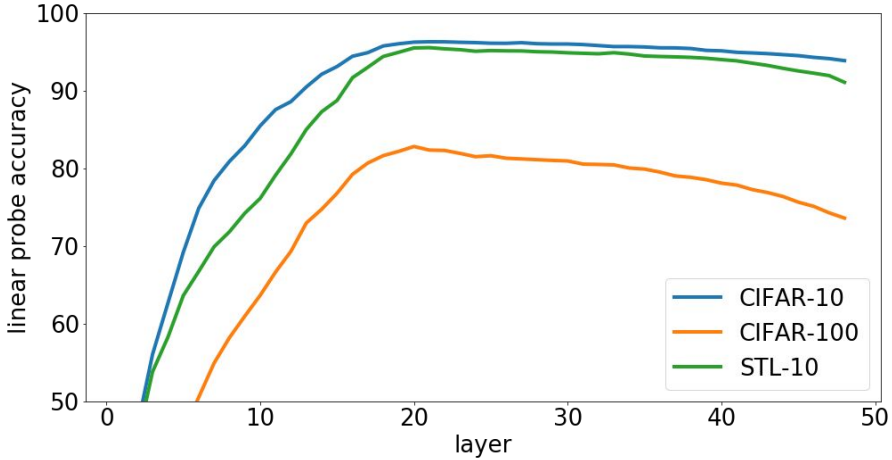
$$L_{GEN} + L_{CLF}$$

$L_{GEN} \in \{L_{AR}, L_{BERT}\}$ works even better. Similar findings were reported by Radford et al. (2018).

Linear Probes



Feature quality depends heavily on the layer we choose to evaluate. In contrast with supervised models, the best features for these generative models lie in the middle of the network.



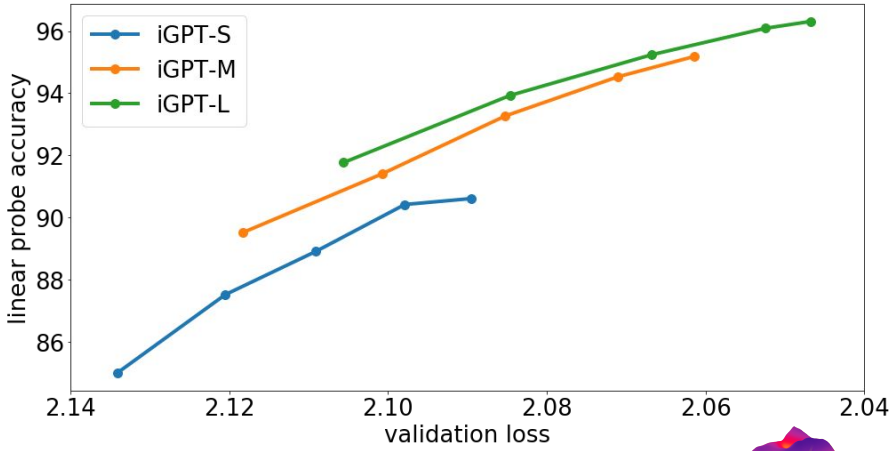


Image Classification Results

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

ImageNet linear probe accuracies

(compared to other self-supervised models)

iGPT-M achieves 54.5% accuracy, iGPT-S - 41.9%

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		√

CIFAR fine-tuning accuracies compared to models utilizing (un)supervised ImageNet transfer

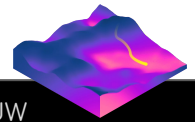
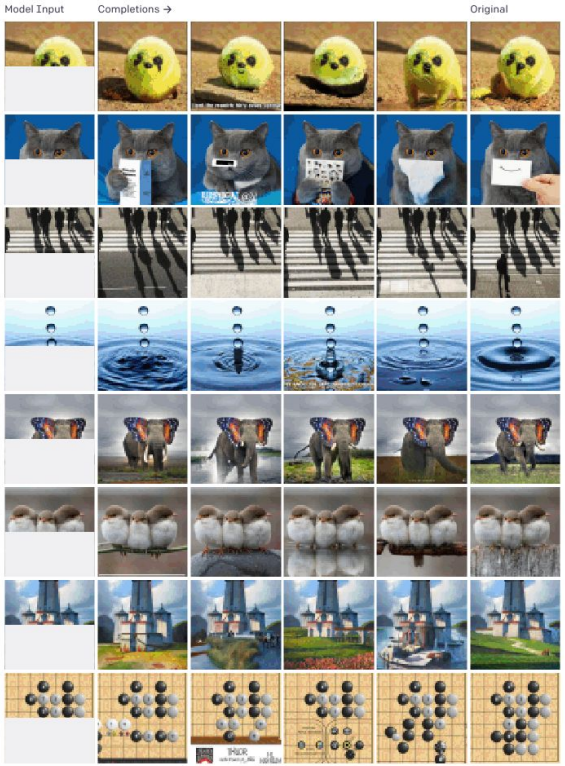
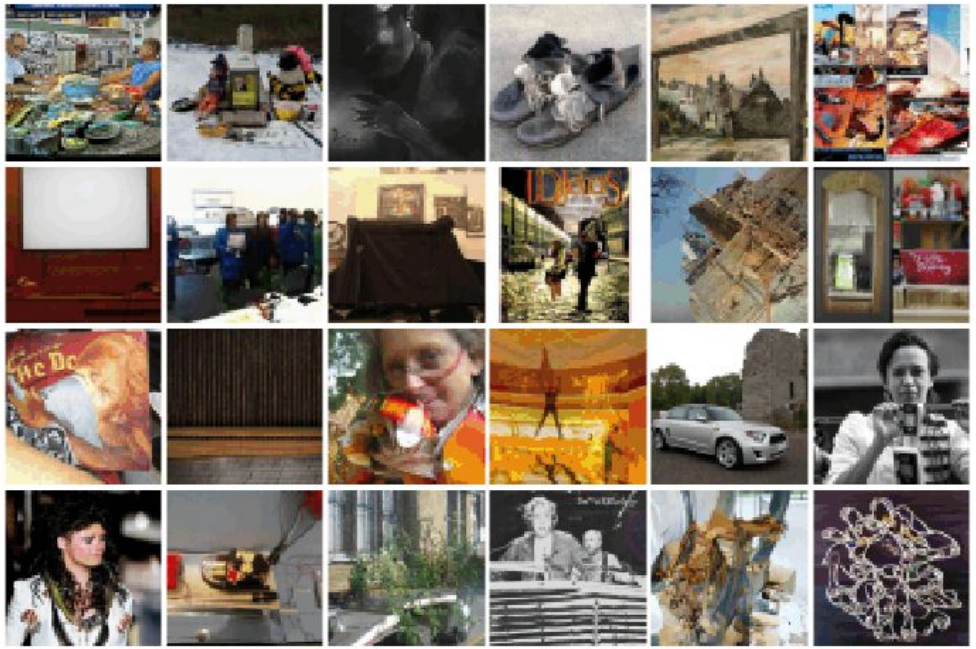


Image Generation Results

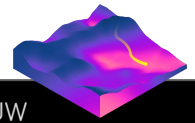


Samples



Bibliography

- https://openai.com/blog/image-gpt
- Generative Pretraining From Pixels V2 (paper)
- Image Transformer (paper)
- The Illustrated GPT-2 (TransformerLM visualization)



Thanks!