Generative Pretraining from Pixels

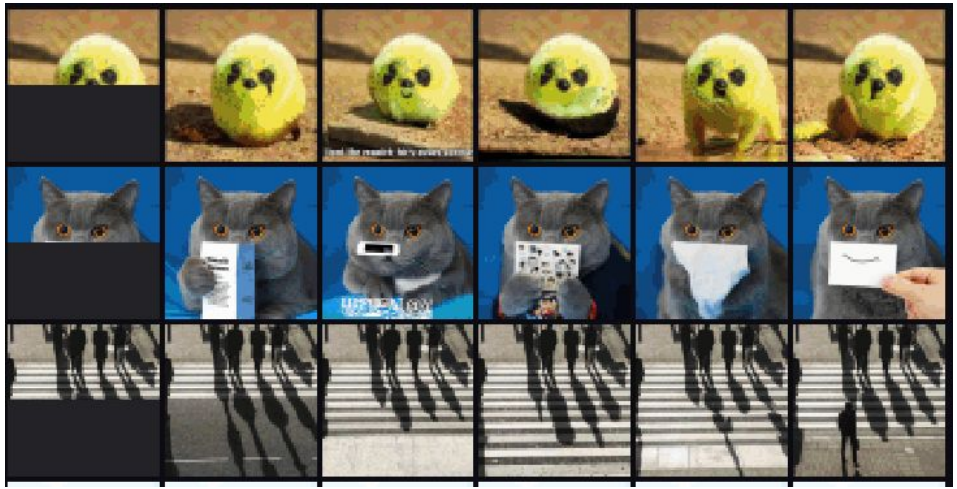
Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, David Luan, Ilya Sutskever

presented by Eli Zachary

GPT, but for image completion – as pretraining

- Another example of using an NLP strategy on visual inputs rather than text
- Based on GPT-2
- Two pre-training methods: auto-regressive and BERT

Examples

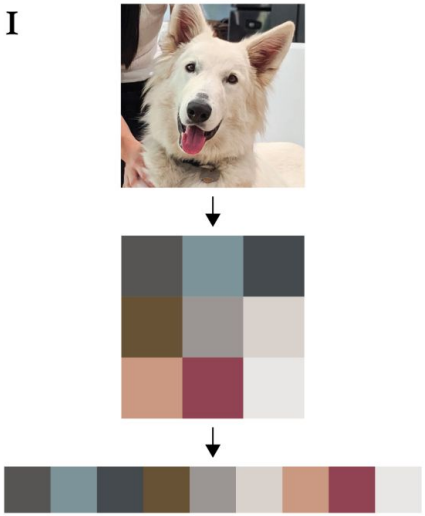


All this is just a byproduct of pretraining! The real purpose is training a classification model.

Prior work

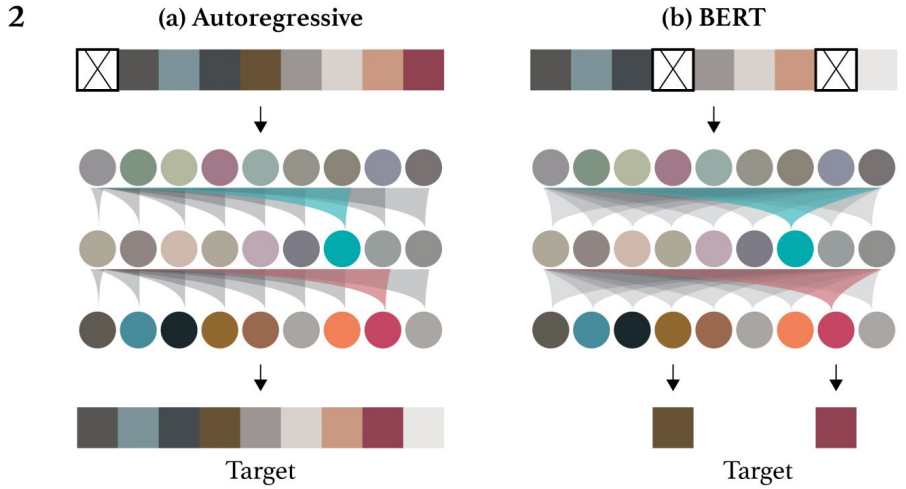
- Generative pretraining for natural language processing (Dai & Le, 2015), (Devlin et al., 2018)
- BigBiGAN (Donahue & Simonyan, 2019)
- Contrast Predictive Coding (Oord et al., 2018), Selfie (Trinh et al., 2019)

Approach



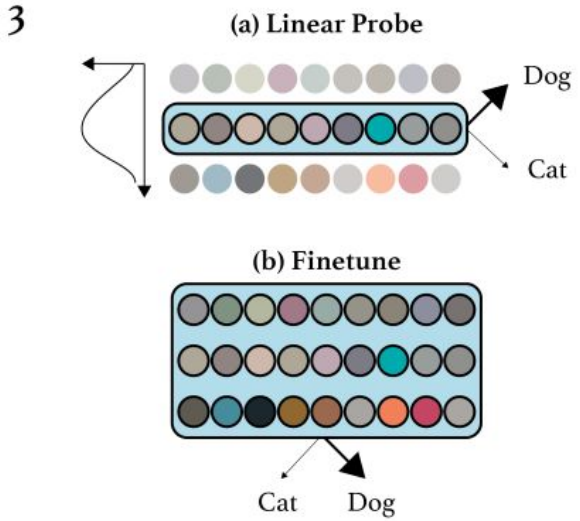
Downscale and reshape into 1D sequence

Approach



Two different pixel generation models (denoising autoencoders)

Approach



Two different methods of evaluating representations

Datasets

- Trained on ImageNet (ILSVRC 2012) – large unlabeled corpus with 4% split off as validation set, ILSVRC validation set used as test set
- When trained on CIFAR-10, CIFAR-100 and STL-10, 10% of each training set split off for use as test set
- ImageNet images randomly resized so that the shorter side is of length [256, 384], and then randomly cropped to 224x224

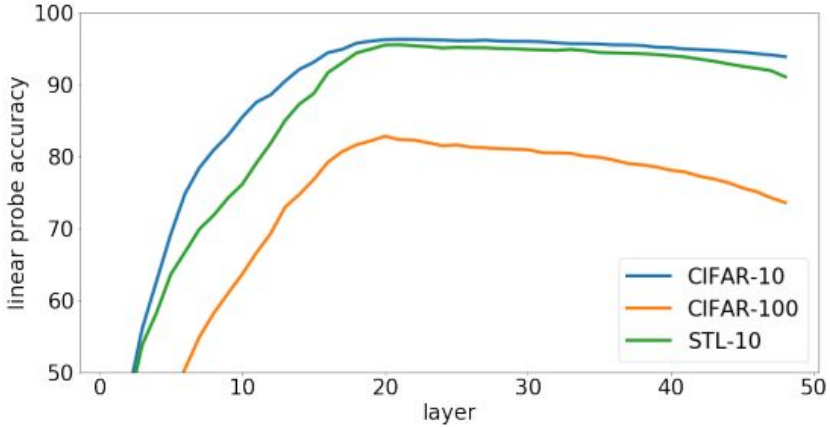
Model details

Model	Layers	Embedding size	Parameters
GPT-2	48	1600	1.5B
iGPT-L	48	1536	1.4B
iGPT-M	36	1024	455M
iGPT-S	24	512	76M

Training details

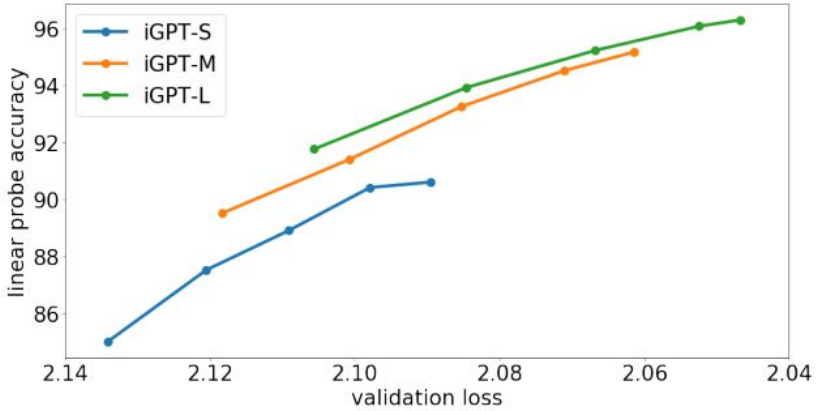
- Pre-training: Batch size 128, 1M iterations, learning rates 0.01, 0.003, 0.001, 0.0003… until local minimum. Learning rate is warmed up for one epoch and then decays to 0 on a cosine schedule. No dropout.
- Fine-tuning: Identical hyparams to pre-training except without the cosine schedule
- Linear probing: On ImageNet, similar to pre-training but with much higher learning rates. On CIFAR and STL-10, uses L-BFGS algorithm

Representations of generative models without latent variables



- Two-phase structure of generative models
- Different pattern from what you'd see with non-generative supervised pre-training
- Take the best layer, not the final layer

Representation Quality



- Linear probe accuracy increases as validation loss decreases
- Large models produce better representations

Linear Probes: CIFAR-10/100 and STL-10

Model	Acc	Unsup Transfer	Sup Transfer
CIFAR-10			
AMDIM-L	91.2	√	
ResNet-152	94		√
iGPT-L	96.3	√	
CIFAR-100			
AMDIM-L	70.2	√	
ResNet-152	78		√
iGPT-L	82.8	√	
STL-10			
AMDIM-L	94.2	√	
iGPT-L (IR $32^2 \cdot 3$)	95.5	√	
iGPT-L (IR $96^2 \cdot 3$)	97.1	√	

- iGPT-L significantly outperforms both state-of-the-art unsupervised and supervised models on ImageNet transfer on these datasets.

Linear Probes: ImageNet

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		68.1
MoCo	orig.	375	8192	68.6
iGPT-L	$192^2 \cdot 3$	1362	16896	69.0
CPC v2	orig.	303	8192	71.5

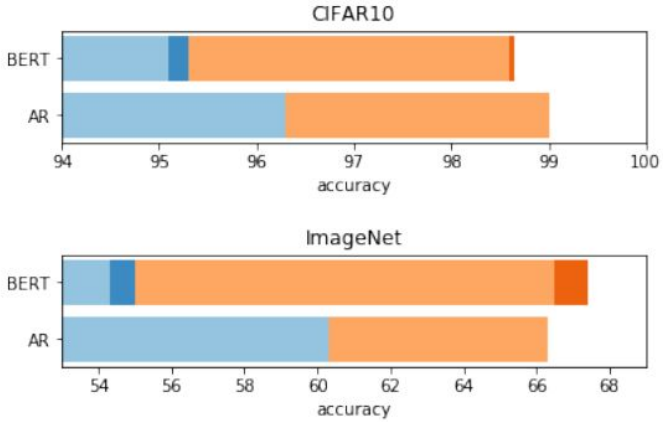
- With an IR of 192^2x3 and an MR of 48^2, iGPT-L outperforms most state-of-the-art self-supervised models on the ImageNet dataset
- This requires VQ-VAE data preprocessing and concatenating 11 layers centered on the best single layer

Fine-tuning

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

- When using fine-tuning instead of linear probing, iGPT-L is competitive with state-of-the-art on CIFAR-10 but underperforms on CIFAR-100.

BERT



- BERT underperforms auto-regression on linear probing, but outperforms it on fine-tuning
- Random sampling of masks improves BERT performance

Low-data classification

Model	40 labels	250 labels	4000 labels
Mean Teacher		32.3 ± 2.3	9.2 ± 0.2
MixMatch	47.5 ± 11.5	11.0 ± 0.9	6.4 ± 0.1
iGPT-L	26.8 ± 1.5	12.4 ± 0.6	5.7 ± 0.1
UDA	29.0 ± 5.9	8.8 ± 1.1	4.9 ± 0.2
FixMatch RA	13.8 ± 3.4	5.1 ± 0.7	4.3 ± 0.1
FixMatch CTA	11.4 ± 3.4	5.1 ± 0.3	4.3 ± 0.2

- iGPT-L does not outperform state-of-the-art on low-data CIFAR-10 classification

Conclusion

- Pretraining a transformer intended for semantic segmentation on image completion tasks is effective
- Requires a larger dataset than convolutional models
- Architectural changes needed to improve efficiency

Q&A?