Selfie: Self-supervised Pretraining for Image Embedding

An Overview

Yuriy Gabuev

Skoltech

October 9, 2019

Trinh, T. H., Luong, M. T., & Le, Q. V. (2019). Selfie: Self-supervised Pretraining for Image Embedding. arXiv preprint arXiv:1906.02940.

Motivation

- We want to use data-efficient methods for pretraining feature extractors
- Unsupervised pretraining has revolutionized NLP (BERT, TransformerXL, GPT), while its success is still limited in other fields
- Context prediction methods, such as Masked Language Modeling, cannot be naively applied to continuous domains, such as images or audio
- **Idea**: solve masked prediction task in latent space

Background: BERT

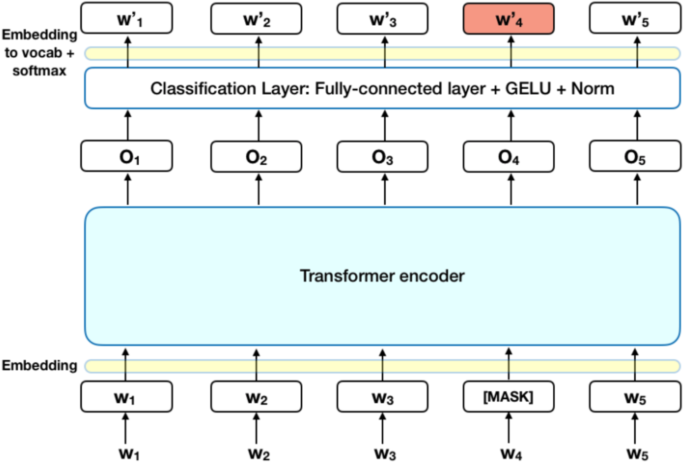

Figure 1: BERT architecture and pretraining objective

Method

Pretraining:

- Split a picture into non-overlapping patches
- Encode patches via a patch-processing convolutional network P
- Mask a fraction of the patches
- Pool non-masked patches via a pooler network A
- Use this pooled output to predict the spatial location of masked patches

Finetuning:

- Instantiate a full encoder partially or fully with P
- Finetune on the target task

Method

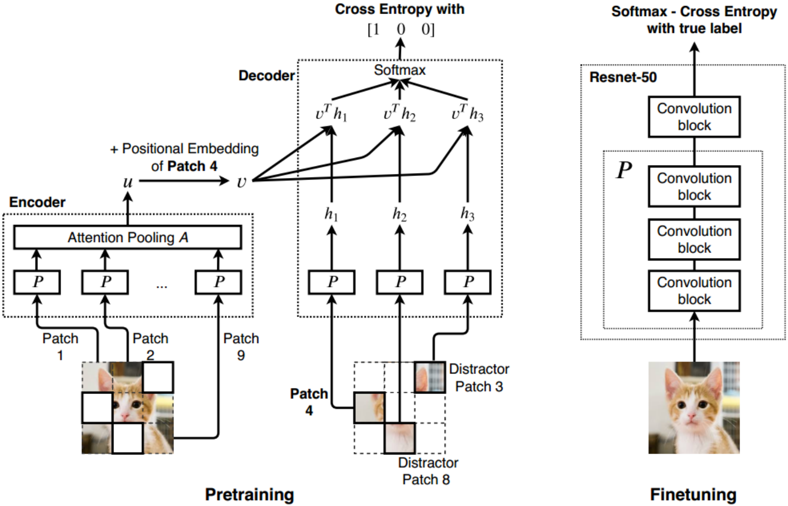

Figure 2: Overview of Selfie

Details

Data:

- CIFAR-10 (32×32) and ImageNet (32×32 and 224×224)
- Patch size: 8×8 for small images and 32×32 for large images
- Fraction of masked patches: 25% or 50%

Architecture:

- Patch-processing network P: ResNet-36
- Pooler A: Transformer with 3 layers, 32 heads, hidden size 1024, intermediate size 640
- Positional encoding: learned, decoupled across width and height dimensions
- Full encoder-1: ResNet-50, the first three blocks are initialized with P
- Full encoder-2: the same as during pretraining

Full encoder variants

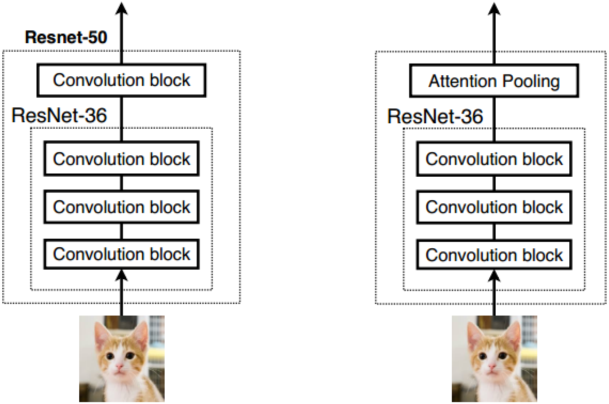

Figure 3: (Left) ResNet-50 architecture. (Right) ResNet-36 + attention pooling architecture.

Results

		Labeled Data Percentage			
		5%	*8%*	*20%*	*100%*
CIFAR-10	Supervised	75.9 ± 0.7	79.3 ± 1.0	88.3 ± 0.3	95.5 ± 0.2
	Selfie Pretrained	75.9 ± 0.4	80.3 ± 0.3	89.1 ± 0.5	95.7 ± 0.1
	Δ	0.0	**+1.0**	**+0.8**	**+0.2**
		5%	*10%*	*20%*	*100%*
ImageNet 32×32	Supervised	13.1 ± 0.8	25.9 ± 0.5	32.7 ± 0.4	55.7 ± 0.6
	Selfie Pretrained	18.3 ± 0.1	30.2 ± 0.5	33.5 ± 0.2	56.4 ± 0.6
	Δ	**+5.2**	**+4.3**	**+0.8**	**+0.7**
ImageNet 224×224	Supervised	35.6 ± 0.7	59.6 ± 0.2	65.7 ± 0.2	76.9 ± 0.2
	Selfie Pretrained	46.7 ± 0.4	61.9 ± 0.2	67.1 ± 0.2	77.0 ± 0.1
	Δ	**+11.1**	**+2.3**	**+1.4**	**+0.1**

Figure 4: Test accuracy (%) of ResNet-50 with and without pretraining

Results

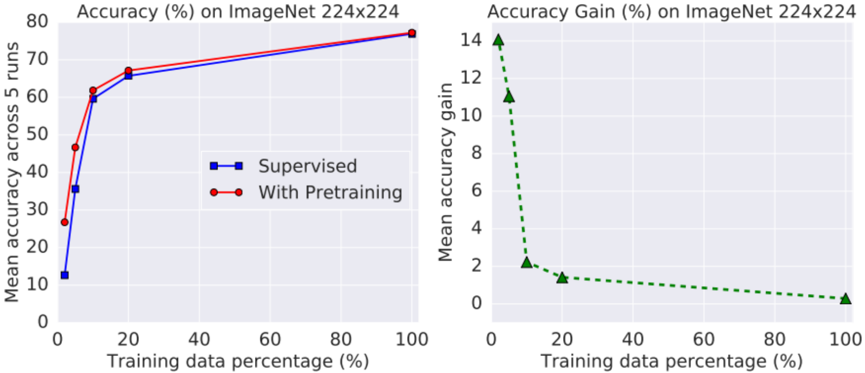

Figure 5: Accuracy gain from pretraining

Results

Method	ResNet-50	ResNet-36 + attention pooling	Δ
CIFAR-10 8%	80.3 ± 0.3	81.3 ± 0.1	**+1.0**
ImageNet 10%	61.8 ± 0.2	62.1 ± 0.2	**+0.3**
CIFAR-10 100%	95.7 ± 0.1	95.4 ± 0.2	-0.3
ImageNet 100%	77.0 ± 0.1	77.5 ± 0.1	**+0.5**

Figure 6: Comparison of two version of the final encoder: ResNet-50 and ResNet-36 + attention pooling

Results

- Selfie-pretraining provides a solid initialization for a feature extractor
- The effects of pretraining diminish with the amount of data for the target task
- A hybrid ConvNet with attention might work better than a pure ConvNet
- Pretraining on one of the datasets does not transfer well to other ones

Problems

- No fair comparison with recent methods without finetuning
- Pretrained subnet (ResNet-36) works in different regimes during pretraining and finetuning stages
- No ablation studies apropos encoder and pooler architectures, patch sampling methods
- No open-source code for reproduction

Thank You!

Further Reading

Philip Bachman, R Devon Hjelm, and William Buchwalter. "Learning representations by maximizing mutual information across views". In: *arXiv preprint arXiv:1906.00910* (2019).

Priya Goyal et al. "Scaling and benchmarking self-supervised visual representation learning". In: *arXiv preprint arXiv:1905.01235* (2019).

R Devon Hjelm et al. "Learning deep representations by mutual information estimation and maximization". In: *arXiv preprint arXiv:1808.06670* (2018).

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. "Revisiting self-supervised visual representation learning". In: *arXiv preprint arXiv:1901.09005* (2019).

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (2018).

Yonglong Tian, Dilip Krishnan, and Phillip Isola. "Contrastive Multiview Coding". In: *arXiv preprint arXiv:1906.05849* (2019).