facebook Artificial Intelligence

A Closer Look at Cross-modal Transfer of Pretrained Transformers

Danielle Rothermel[1], Margaret Li[2], Tim Rocktäschel[1,3], Jakob Foerster[1]

[1]Facebook AI Research, [2] University of Washington, [2] University College London

Introduction

Background

The transfer of transformer language models, pretrained on increasingly large language datasets and then briefly finetuned on a wide variety of tasks with even very small datasets, has revolutionized NLP in recent years. As a natural follow-up question, there has been increasing interest in the possibility of cross-modal transfer, with Lu et al. providing an impressive first proof of concept for transferring from the language domain to bit manipulation, vision, math, and protein homology classification tasks.

However, in our initial experiments on cross-modal transfer to other tasks, we found that it was crucial to tune hyperparameters separately between finetuning from pretrained models vs. training from scratch. We investigate the impact of this tuning on the transfer tasks Lu et al. propose.

Contribution

We compare transfer of frozen and unfrozen pretrained language models against training from scratch on four of the tasks considered by Lu et al: ListOps, MNIST, CIFAR10 and CIFAR10-LRA. We carefully match the architectures between the settings and demonstrate that the choice of learning rate impacts the final performance and ordering between variants.

We verify the conclusion of Lu et al., that finetuning from natural language pretraining improves performance on the tasks identified. However, our results contradict their findings about the benefits of a Frozen Pretrained Transformer (FPT) model, instead finding that FPTs significantly underperform unfrozen models, but that Unfrozen Pretrained transformers match or outperform training from scratch.

Tasks & Tokenizations

All tasks considered are classification tasks.

ListOps

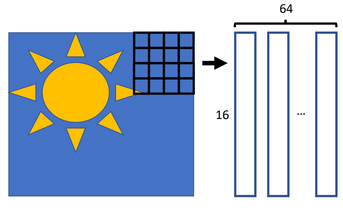
From the Long Range Arena, sequences of up to 2000 operators where each token is a 16 dimension one-hot.

INPUT: `[MAX 4 3 [MIN 2 3 ] 1 0 [MEDIAN 1 5 8 9, 2]]`

OUTPUT: 5

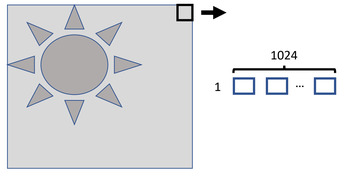
MNIST and CIFAR10

Images from the MNIST & CIFAR10 datasets are tokenized by taking 4 x 4 image patches and flattening them, producing sequences of 64 tokens of dimension 16.



CIFAR10-LRA

The CIFAR10 images in grayscale are instead directly flattened, with each token being a single pixel, producing a 1024 token sequence with 1 dimensional tokens.



Method

Architecture Variants

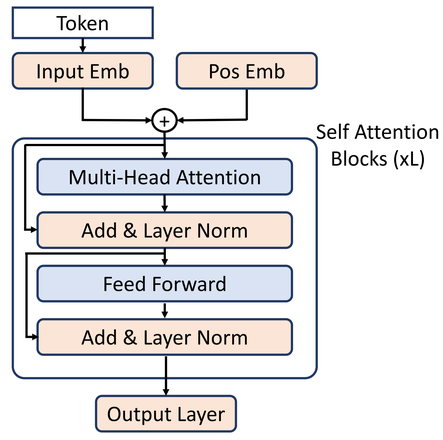
The full sized GPT2 architecture is used (12 layers, 12 heads, 768 embed dim), with 4 different variants:

Frozen Pretrained (FPT): The transformer is initialized with the GPT2 pretrained LM and the blue shaded components below are frozen while the orange shaded components are finetuned on the task.

Frozen Random: The transformer is initialized with random weights and the blue shaded components are frozen while the orange shaded components are finetuned.

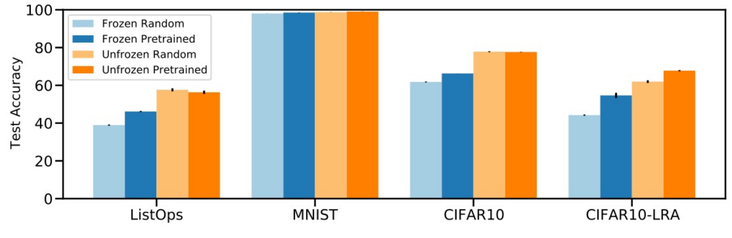
Unfrozen Pretrained: The transformer is initialized with the GPT2 pretrained LM and all components are finetuned.

Unfrozen Random: The transformer is initialized with random weights and all components are finetuned.



For all settings we use the Adam optimizer and sweep the learning rate logarithmically from $1x10^{-6}$ to$1x10^{-2}$, selecting the best LR from the validation accuracy and reporting the test accuracy and error over 3 runs.

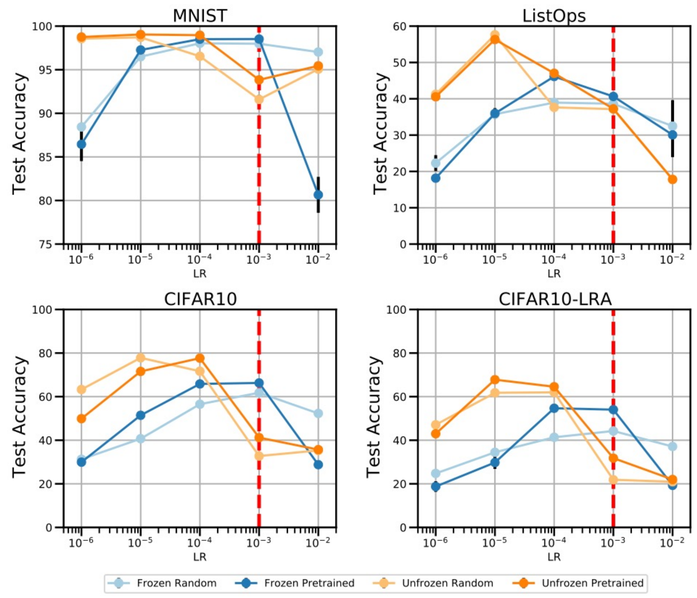
Results



	ListOps	MNIST	CIFAR10	CIFAR10-LRA
Frozen Random	38.9 ± 0.3	98.0 ± 0.0	61.8 ± 0.2	44.2 ± 0.3
Frozen Pretrained	46.1 ± 0.3	98.5 ± 0.1	66.3 ± 0.0	54.7 ± 1.4
Unfrozen Random	$\mathbf{57.6 \pm 0.8}$	98.7 ± 0.0	$\mathbf{77.8 \pm 0.2}$	62.0 ± 0.7
Unfrozen Pretrained	$\mathbf{56.3 \pm 0.9}$	$\mathbf{99.0 \pm 0.0}$	77.7 ± 0.1	$\mathbf{67.8 \pm 0.3}$

In all cases the Unfrozen variants outperform the Frozen variants significantly. For the CIFAR10-LRA task, Unfrozen Pretrained outperforms the rest by a large margin, for MNIST by a small margin, and it matches on the remaining tasks.

Performance Across Learning Rates



Here we show how the test accuracy varies across learning rates for each setting. The dashed red line shows the learning rate used across experiments by Lu et al. (though some of their Unfrozen Random results came from the Long Range Arena benchmark where the hyperparameters were tuned separately).

Note that the optimal learning rate for the Unfrozen variants is generally substantially lower than that of the Frozen variants. If choosing a single learning rate to evaluate on, the conclusions drawn depend on which learning rate is chosen. And, as seen in the plots, choosing a lower learning rate would have inverted their results.

Computational Efficiency

	ListOps	MNIST	CIFAR10	CIFAR10-LRA
Unfrozen Random	$\mathbf{1.0 \times 10^5}$	1.1×10^5	7.0×10^4	1.8×10^5
Frozen Random	-	-	-	-
Unfrozen Pretrained	1.9×10^5	$\mathbf{4.6 \times 10^4}$	$\mathbf{4.3 \times 10^4}$	$\mathbf{5.3 \times 10^4}$
Frozen Pretrained	1.9×10^5	2.4×10^5	3.8×10^5	2.4×10^5

The number of gradient steps required to match the best performance of the Frozen Pretrained variant.

Effects of Scaling Capacity

		MNIST	CIFAR10	CIFAR10-LRA
Frozen Random	DistilGPT2	$\mathbf{98.0 \pm 0.1}$	60.1 ± 0.1	$\mathbf{45.0 \pm 0.1}$
	GPT2	$\mathbf{98.0 \pm 0.0}$	$\mathbf{61.8 \pm 0.2}$	44.2 ± 0.3
Frozen Pretrained	DistilGPT2	$\mathbf{98.5 \pm 0.1}$	65.2 ± 0.5	51.1 ± 0.4
	GPT2	$\mathbf{98.5 \pm 0.1}$	$\mathbf{66.3 \pm 0.0}$	$\mathbf{54.7 \pm 1.4}$
Unfrozen Random	DistilGPT2	$\mathbf{98.6 \pm 0.1}$	$\mathbf{77.5 \pm 0.1}$	59.7 ± 0.2
	GPT2	$\mathbf{98.7 \pm 0.0}$	$\mathbf{77.8 \pm 0.2}$	$\mathbf{62.0 \pm 0.7}$
Unfrozen Pretrained	DistilGPT2	98.9 ± 0.0	76.8 ± 0.1	65.5 ± 0.5
	GPT2	$\mathbf{99.0 \pm 0.0}$	$\mathbf{77.7 \pm 0.1}$	$\mathbf{67.8 \pm 0.3}$

The performance of different model sizes, with DistilGPT2 at 6 layers as compared to GPT2 at 12.