Pretrained Transformers as Universal Computation Engines

By Priyanka Sachan
priyanka_1901CS43@iitp.ac.in

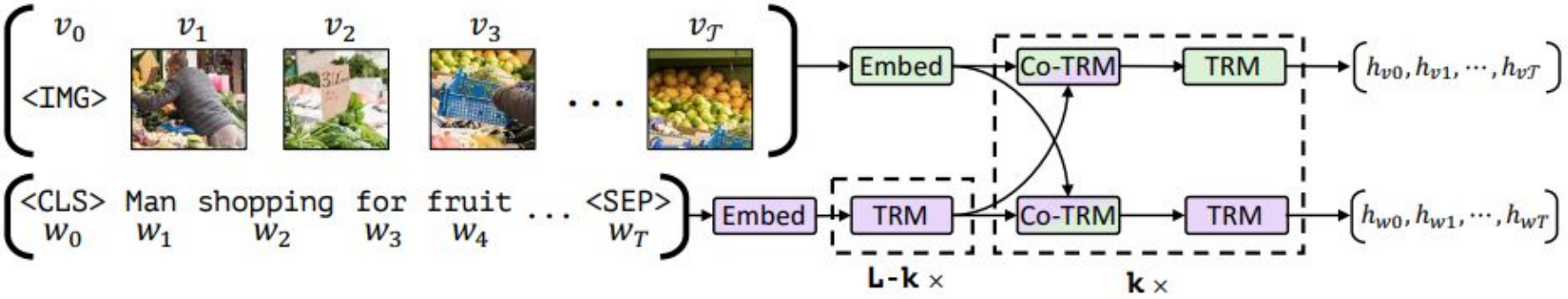
Transformers

Have shown broad success in deep learning serving as the backbone of large models.

COMMON TRAINING PATTERNS

- Training large models on unsupervised or weekly supervised objectives.
- After that, finetuning or evaluating zero-shot generalization on a downstream task.
- However, the downstream tasks that have been studied are generally restricted to the same modality as the original training set.

Transformers in multimodal settings



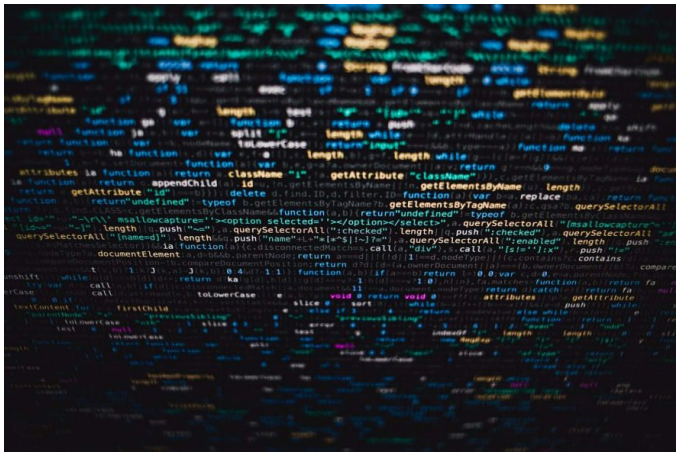
ViLBERT

Transformers in multimodal settings

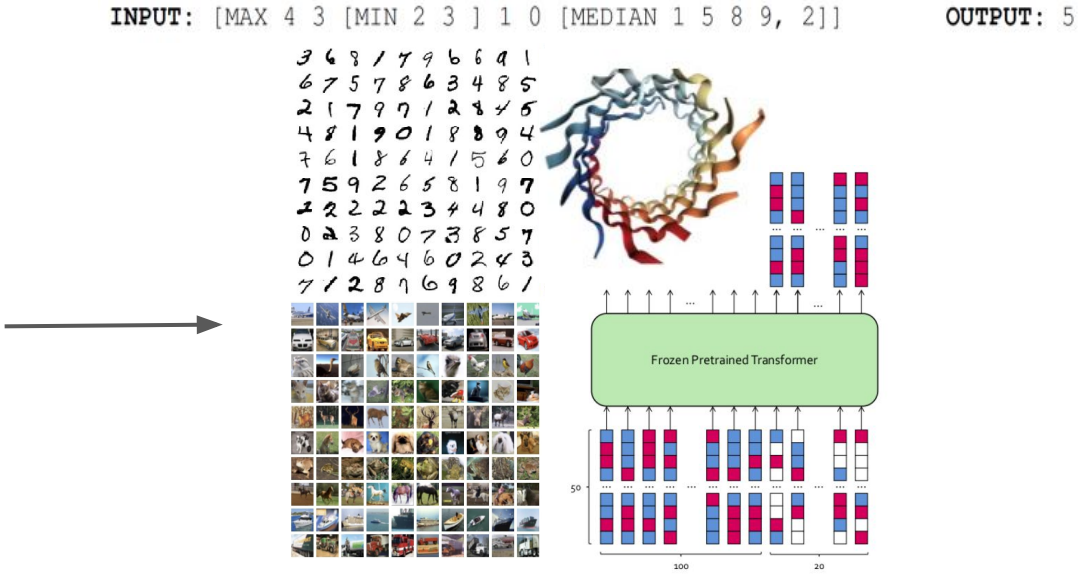
DALL-E

One-to-many modality fine-tuning regime

Exploring the generalization capabilities of a transformer in transferring from one modality to another.



TRAIN



FINETUNE

One-to-many modality fine-tuning regime

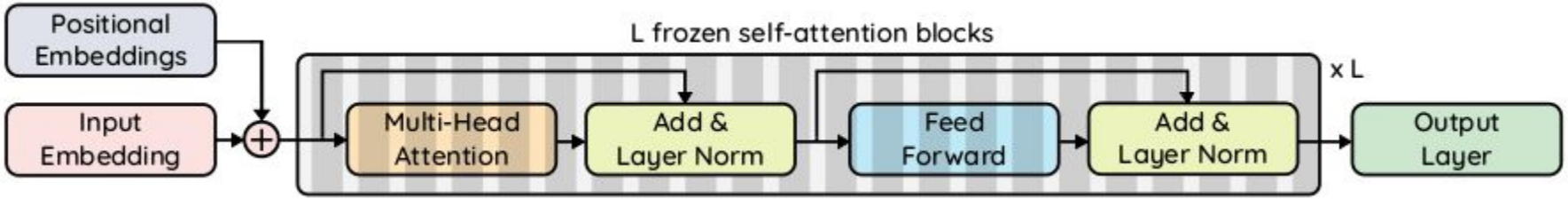
HYPOTHESIS

- Transformers, can be pretrained on a data-rich modality (i.e. where data is plentiful, such as a natural language corpus) and identify feature representations that are useful for arbitrary data sequences, enabling effective downstream transfer to different modalities without expensive finetuning of the self-attention layers.
- In particular, pretrained language models (LMs) are capable of in terms of generalizing to other modalities with sequential structure, including numerical computation, image classification, and protein fold prediction.

Architecture

Taking a transformer model pretrained on natural language data, GPT-2, and finetuning only the linear input and output layers, as well as the positional embeddings and layer norm parameters. Both self-attention and feedforward layers of the residual blocks are frozen.

This model is called a Frozen Pretrained Transformer(FPT).

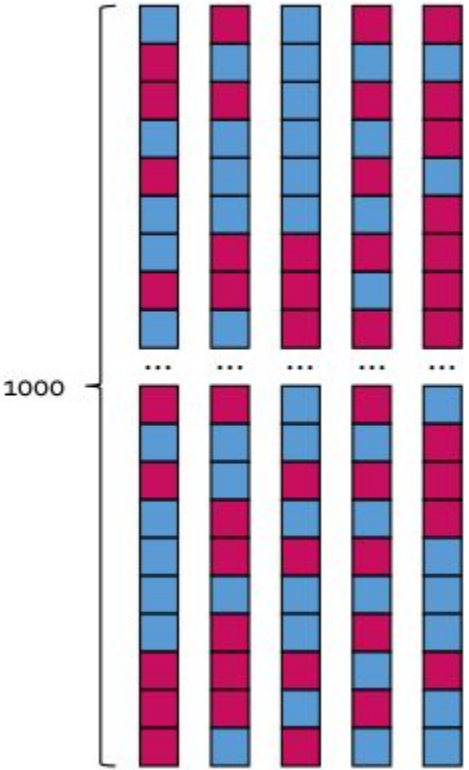


Evaluation Tasks

- Bit memory
- Bit XOR: x0 ^ x1 = y
- ListOps: [MAX 4 3 [MIN 2 3] 1 0]
- MNIST: The tokens given to the model are 4 x 4 image patches.(total 64 tokens)
- CIFAR-10: Same with MNIST
- CIFAR-10 LRA: 1 x 1 image patches (total 1024 tokens with dimension 1)
- Remote homology detection: predicting protein fold problem. (1024 tokens of dimension 25)

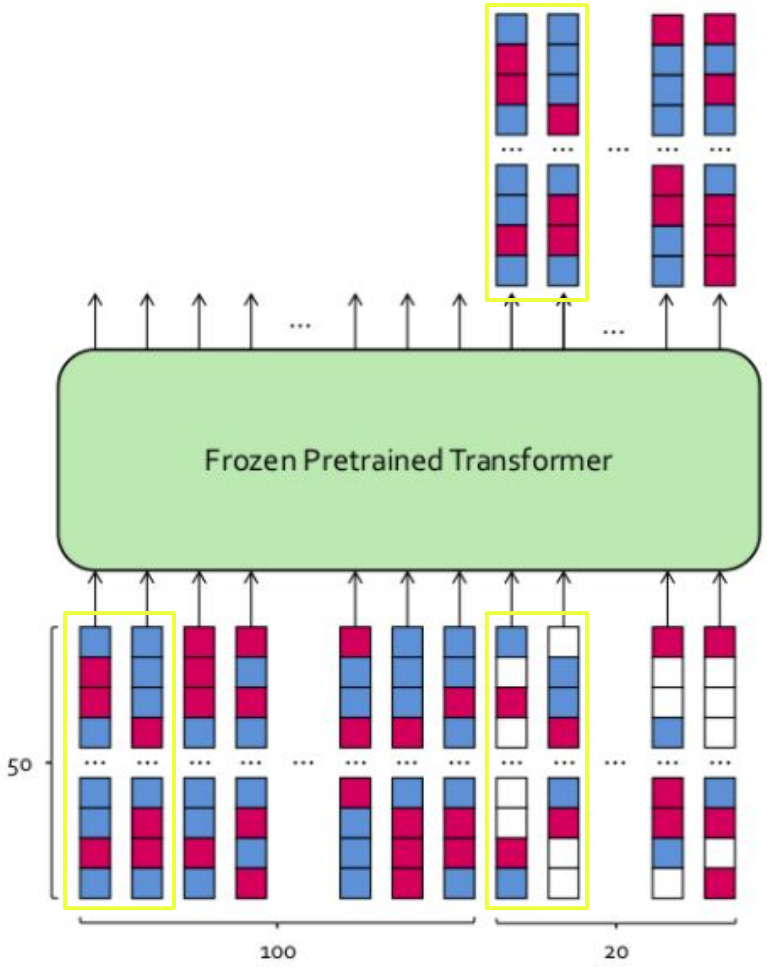
EVALUATION TASKS

Bit memory



The model is shown 5 bit strings each of length 1000. Afterwards, the model is shown a masked version of one of the bitstrings, where each bit is masked with probability 0.5, and the model is tasked with producing the original bitstring.

The bit strings are broken up into sequences of length 50, so that the models are fed 120 tokens of dimension 50.

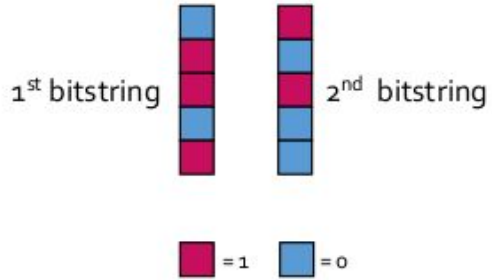


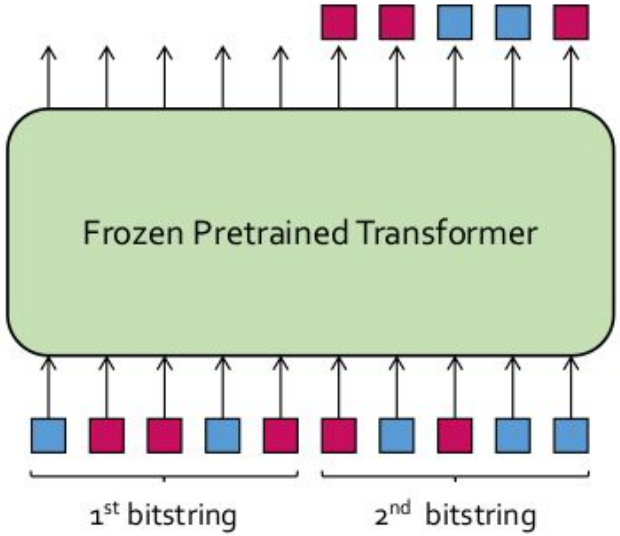
EVALUATION TASKS

Bit XOR

Similar to the bit memory task, the model is shown 2 bitstrings of length 5, where the model must predict the element-wise XOR of the two bitstrings.

The bitstrings are shown 1 bit at a time, so the models are fed 10 tokens of dimension 1.





EVALUATION TASKS

List Ops

The model is shown a sequence of list operations(MAX, MEAN, MEDIAN and SUM_MOD) and tasked with predicting the resulting output digit.

The model is shown 1 token at a time, so the models are fed 512 tokens of dimension 15.

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

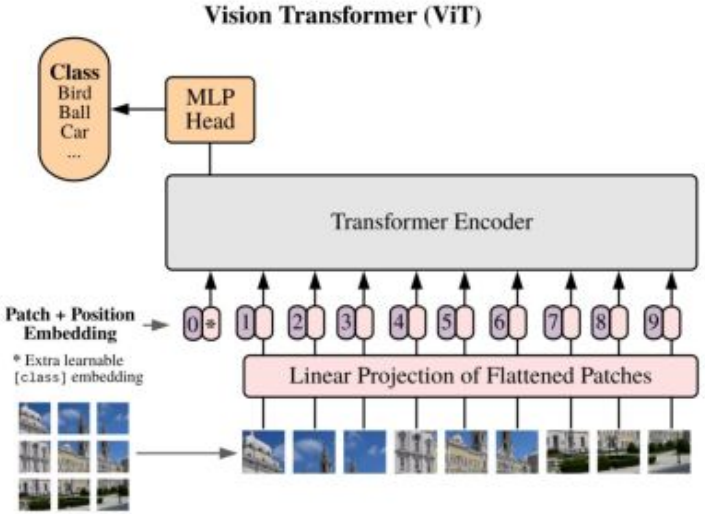
EVALUATION TASKS

MNIST

Standard MNIST benchmark, where the model must classify a handwritten digit from a 32x32 black-and-white image.

The tokens given to the model are 4x4 image patches, so the models are fed 64 tokens of dimension 16.

MNIST data example



EVALUATION TASKS

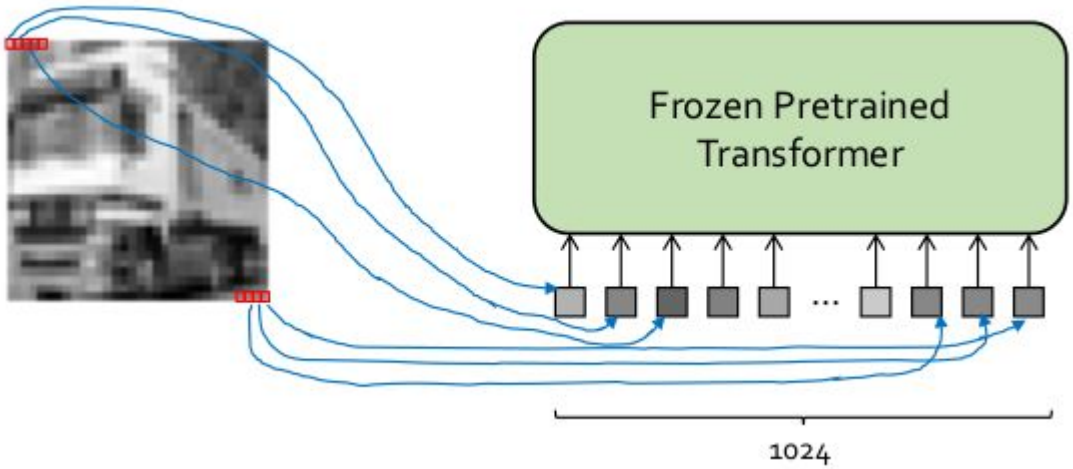
CIFAR-10

- **CIFAR-10**

Standard CIFAR-10 benchmark, where the tokens given to the model are 4x4 image patches, so the models are fed 64 tokens of dimension 16.

- **CIFAR-10 LRA**

A modified version of the above task taken from the Long Range Arena benchmark where the images are converted to grayscale and flattened with a token length of 1. As a result, the input sequence consists of 1024 tokens of dimension 1. The models must learn patterns over a significantly longer sequence length and have minimal spatial inductive bias.

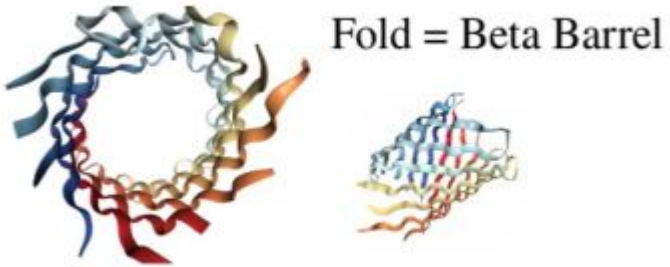


EVALUATION TASKS

Remote Homology Detection

The datasets provided by TAPE. No pretrain on Pfam, which is common in other works. There are 20 common and 5 uncommon amino acids (25 different types of inputs), and there are 1195 possible labels to predict.

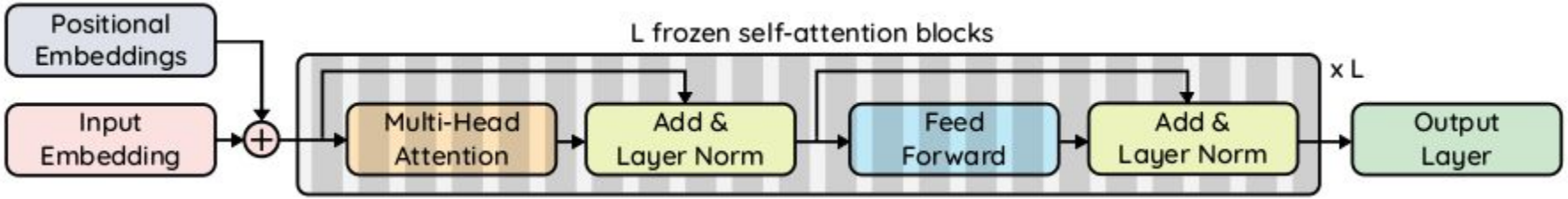
The models are fed up to 1024 tokens of dimension 25.



Architecture

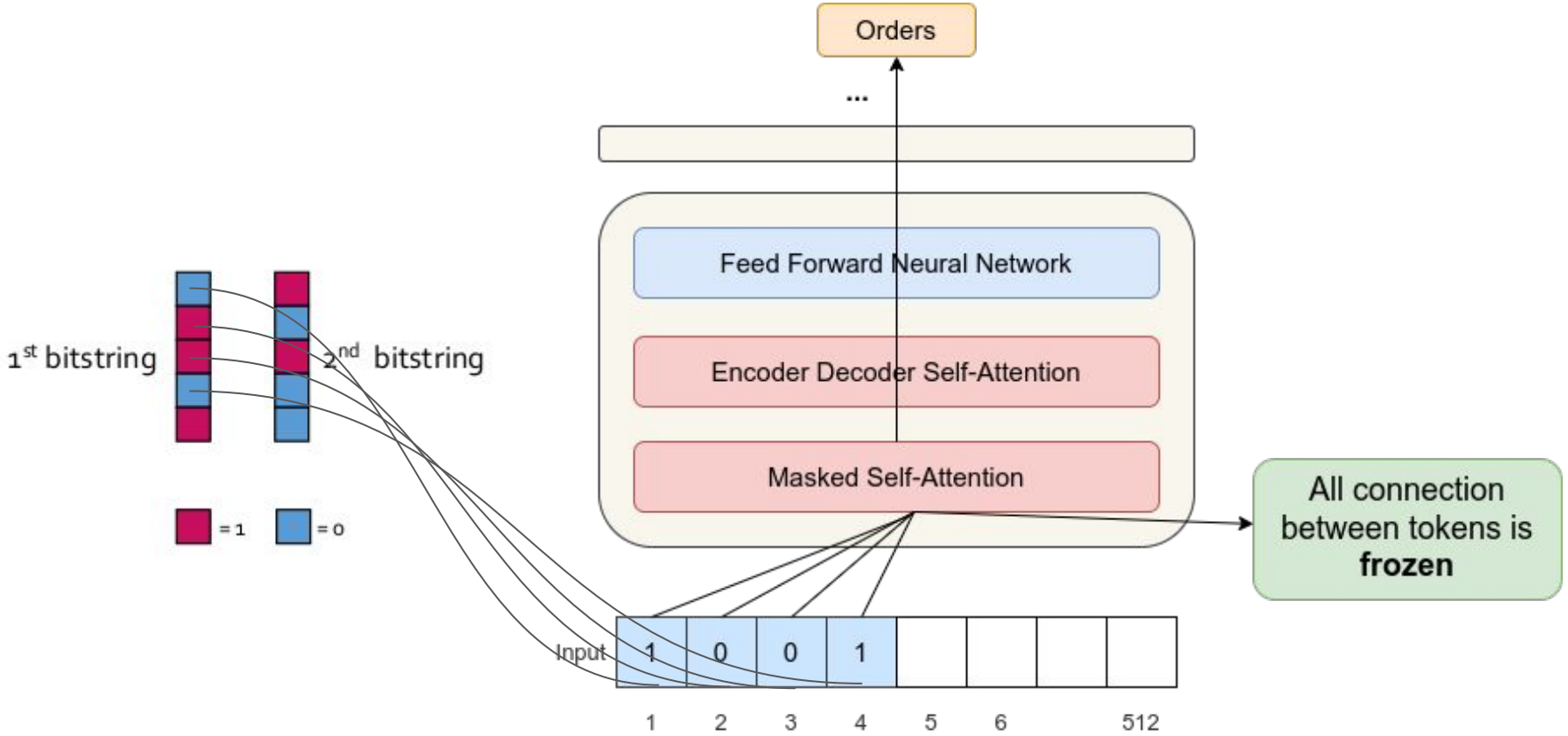
Taking a transformer model pretrained on natural language data, GPT-2, and finetuning only the linear input and output layers, as well as the positional embeddings and layer norm parameters. Both self-attention and feedforward layers of the residual blocks are frozen.

This model is called a Frozen Pretrained Transformer(FPT).



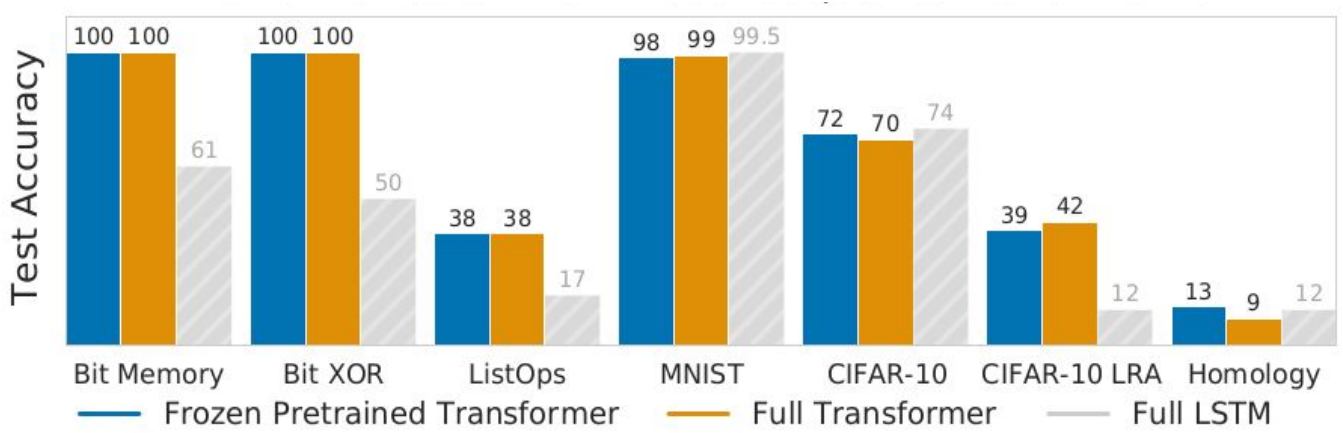
ARCHITECTURE

Frozen self-attention for arbitrary inputs



Performance on Multimodal Sequence Benchmarks

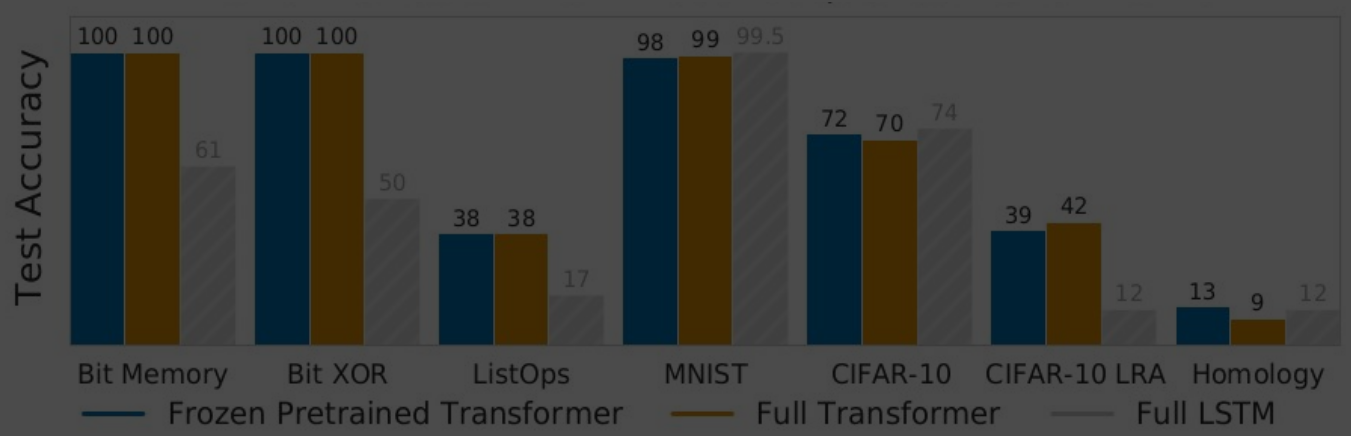
- FPT displays comparable performance to training the entire transformer or LSTM models, despite finetuning only 0.1% of the total number of parameters of the transformer model and none of the self-attention parameters.
- Because it is difficult to fully train a 12-layer transformer on small datasets, for CIFAR-10, the authors report the full transformer results for a 3-layer model.



Performance on Multimodal Sequence Benchmarks

- FPT displays comparable performance to training the entire transformer or LSTM models, despite finetuning only 0.1% of the total number of parameters of the transformer model and none of the self-attention layers.
- Because it is difficult to fully train a 12-layer transformer on small datasets, for CIFAR-10, the authors report the full transformer finetuned 3 epochs.

Frozen connections learned by language are as good as trained downstream



EMPIRICAL EVALUATION

Can pretrained language models transfer to different modalities?

- Across all seven tasks considered, FPT achieves comparable, if not marginally better performance than fully training a transformer.
- These results support the idea that these models are learning representations and performing computation that is agnostic to the modality.
- Both transformer variants significantly outperform LSTMs on some tasks, particularly ListOps and CIFAR-10 LRA, which have long sequence lengths.

Model	Bit Memory	XOR	ListOps	MNIST	CIFAR-10	C10 LRA	Homology
FPT	100%	100%	38.4%	98.0%	72.1%	38.6%	12.7%
Full	100%	100%	38%	99.1%	70.3%	42%	9%
LSTM	60.9%	50.1%	17.1%	99.5%	73.6%	11.7%	12%

EMPIRICAL EVALUATION

What is the importance of pretraining modality?

Comparing the performance of FPT to other pretraining methods for the base model sizes:

- Random initialization (Random): initialization of the frozen transformer parameters randomly using the default initialization choices for GPT-2.
- Bit memory pretraining (Bit): pretraining on the Bit Memory task and then freezing the parameters before transferring.
- Image pretraining (ViT): using a pretrained Vision Transformer (Dosovitskiy et al., 2020) pretrained on ImageNet-21k.

Model	Bit Memory	XOR	ListOps	MNIST	C10	C10 LRA	Homology
FPT	100%	100%	38.4%	98.0%	68.2%	38.6%	12.7%
Random	75.8%	100%	34.3%	91.7%	61.7%	36.1%	9.3%
Bit	100%	100%	35.4%	97.8%	62.6%	36.7%	7.8%
ViT	100%	100%	37.4%	97.8%	72.5%	43.0%	7.5%

Conclusion

- The self-attention layers learned by a language model may have properties amenable to efficient universal computation.
- Worth investigating the use of other data-rich modalities (e.g., vision) or a hybrid of multiple domains being used to provide the necessary substrate for pretraining a universal computational engine.

Thanks!