GROUP NO-6

29 November 2022

- TA assigned: Suraj
- Contributions: Transformer: Ayush, Ayushi, BART : Chetan, Charvi GPT : Asmita, Aurko

1 Datasets Used for Training

Transformer: Uses Standard WMT 2014 English-German dataset, 4.5 M words and WMT 2014 English-French dataset, 36 M words for training . Evaluation is done on the same datasets and the components of the transformer were evaluated on the newstest2013 dataset.

GPT : BooksCorpus dataset for training. It is a corpus of over 7000 unique and unpublished books from a range of genres. The long contiguous texts present in the corpus helps the model trained for generative tasks to capture context over a long. range of dependency.

It does evaluation on a number of tasks ,some of which are natural language inference : SNLI , MultiNLI ,Question NL, RTE , SciTail ,Question Answering : RACE , Story Cloze , Sentence similarity : MSR Paraphrase Corpus , Quora Question Pairs , STS Benchmark ,Classification : Stanford Sentiment Treebank-2 , CoLA

BART : Training done on SQuAD,MNLI,ELI5,XSum,ConvAI2,CNN/DM . Evaluation done on the following tasks,Discriminative task: SQuAD, GLUE,Summarization: CNN/DailyMail, XSum,Dialogue: CONVAI2,AbstractiveQA: ELI5,Translation: WMT16 Romanian-English

No novel dataset was introduced in any of the work and no data augmentation

2 Model Architectures Comparison

	Transformer	**BART**	**GPT**
Improvement/Modification	Self-Attention, which captured the semantic meaning of each word of the sequence with all the other words by learning focus words in the entire sequence, eliminated the need of RNN's (recurrent units) to capture sequential dependencies, thus achieving parallelization.	ReLU activation functions modified to GeLU after GPT Each layer of decoder performs cross-attention over final hidden layer Use of additional feed-forward network before word-prediction	Task agonistic models that is able to outperform models trained specifically for the task. Decoder only architecture outperforms encoder decoder architectures in GLUE tasks.
Encoder-decoder setting	Encoder-6 layered; multi-head self-attention followed by fully connected feed forward network. Decoder-same as above; a third layer, performs multi-head attention over the encoder's output.	Bidirectional encoder, left-to-right autoregressive decoder	Autoregressive Decoder only (T-DMCA) T-DMCA : Transformer Decoder with memory compressed attention.
Attention Mechanism	scaled dot-product self-attention.	Self-Attention	Memory compressed attention. In order to reduce the number of dot products,

			the value and the key vectors are passed via convolution layers and then subsequent matrix multiplication is done.
Training Strategy	a) create positional embeddings vector w.r.t every word b) create 3 vectors Query Q, Key K, Value V for each word and learn focus words by taking dot product of a word's Q with every words' K. c)then normalize, apply softmax to get the scalar and multiply it with the value V of each word; take a weighted sum. This creates attention for that particular word. d)This is done irrespective of recurrent units and without compromising long-term dependency, hence achieving parallelisation.	Trained in the following manner: a) Documents are corrupted. b) Optimized using cross-entropy between decoder's output and original document.	2 stage : a) Unsupervised language modeling: Trained with the standard language modeling objective using T-DMCA on a large corpus of unlabelled text. b) Supervised fine tuning **:** Fine tuning on a variety of tasks with the additional task of language modeling. The decoder consists of two heads one for the current fine tuning task and the other for the language modeling task."Task aware input transformations" are done to the input to simultaneously get the output for language modeling and the fi ne tuning task.

3 Downstream Tasks

The transformer is evaluated on a machine translation task where it achieves a BLEU score of 28.4 and beats the previous SOTA by 2 BLUE score.

BART is evaluated on Generation tasks namely Summarization and Dialogue Generation where it is able to beat the previous best performance by 6 ROUGE and by 2 metric respectively.

Both BART and GPT are evaluated on Natural Language Inference Tasks. BART outperforms GPT by a lot in NLI task with the RTE dataset. GPT does poorly on the RTE dataset by only being able to achieve an accuracy of 56% which is well below the baseline performance, however BART does well by achieving an accuracy of 87.0%. On MNLI-m and MNLI-mm too BART outperforms GPT.

For Question Answering, BART is evaluated on the EL!5 dataset where it outperforms previous works by 1.2 Rouge-L . GPT does very well on Story Cloze test showing that the model does well on common sense reasoning tasks

BART outperforms GPT on semantic similarity task on Quora Question Pair dataset with accuracy of 92.5 and GPT accuracy of 70.3%

In Sentence Similarity Task GPT achieves accuracy of 91.3% while BART achieves accuracy of 96.6%

BART is also evaluated on the machine translation task on WMT16 RO-EN and it outperforms baseline and achieves a ROUGE score of 37.96.