GROUP 20

Shruti Jha Shubham Agrawal Siddhant Agarwal Simran
Siruvuri Karan Raju Siya Garg

November 2022

- TA assigned: Karish Grover
- Contributions: Transformer paper- Siya Garg and Shruti Jha; XLNet paper- Siddhant Agarwal and Simran; RoBERTa paper- Siruvuri Karan Raju and Shubham Agrawal

1 Datasets Used for Training

The encoder-attention-decoder transformer model uses the two standard WMT 2014 datasets: one with 4.5 million English-German sentence pairs (having shared source-target vocabulary of 37000 tokens) and the other with 36 million English-French sentences (and split tokens into a 32000 word-piece vocabulary).

The RoBERTa model uses five English-language uncompressed corpus: BookCorpus, English Wikipedia (16GB), CC news (63M English news articles), OpenWebText (web content from Reddit URLs with at least 3 upvotes), Stories (a subset of CommonCrawl).

The XLNet model uses BookCorpus, English Wikipedia (13GB), Giga5 (16GB), ClueWeb 2012-B (19 GB) and Common Crawl (110GB) containing 2.78B, 1.09B, 4.75B, 4.30B, and 19.97B subword pieces respectively with a total of 32.89B subwords.

RoBERTa and XLNet both are trained on huge datasets (both ∼160 GB) as compared to transformers. They have BookCorpus, English Wikipedia and CommonCrawl data in common.

None of the papers mention data augmentation or introduce any new dataset for training.

2 Model Architectures Comparison

2.1 Attention-based Transformer

It is an encoder-attention-decoder model without using RNNs. It has six encoder and decoder layers each. There are two sub-layers in each encoder layer: the multi-head self-attention and fully connected feedforward neural network. In the decoder, there is an additional multi-head self-attention layer acting on the output of the encoder stack. This architecture ensures that we make use of only previous predictions and their context only.

The attention is computed as $\text{Attention(Q, K, V)} = \text{softmax}(\text{QK}^T \; /\sqrt{d_\text{k}})\text{V}$
Here Q (Query), K (Key), V (Value) are obtained through projections on input. This allows parallelisation whereas previous models like RNN, LSTM and GRU rely on sequential data.

2.2 RoBERTa

The model architecture of RoBERTa uses only an encoder, which is the same as BERT. Both use the same attention mechanism as the standard transformer, but RoBERTa has the training methodology modified in various ways compared to BERT. BERT implementation uses masking only once during data preprocessing, resulting in a single static mask. RoBERTa uses dynamic masking in which, at every iteration, a new masking pattern is generated.

The input in BERT is two concatenated document segments, which are either sampled contiguously from the same or distinct documents with equal probability. In RoBERTa the input sentences are sourced from the same documents. BERT uses Next Sentence Prediction (NSP) loss for training the model unlike RoBERTa, which helps increase performance in some downstream tasks.

BERT was trained for 1M steps with a batch size of 256 sequences, whereas RoBERTa was trained for 125K steps with a huge batch size of 2K sequences. BERT uses a character-level Byte Pair Encoding (BPE) vocabulary of size 30K, which involves preprocessing the input with heuristic tokenization rules. In contrast, in RoBERTa a larger byte-level BPE vocabulary containing 50K subword units is used without any additional preprocessing or tokenization of the input.

2.3 XLNet

During their initial years, BERT/RoBERTa gained popularity due to its autoencoding-based pre-training. They achieved a better performance than then-known approaches based on autoregressive language modeling capturing only unidirectional context. Their primary aim was to reconstruct the original test by recovering masked words ignoring explicit density estimation. The autoregressive pre-training models use the independence assumption, neglecting the relationship among the masked words. So leveraging the best of both worlds - autoencoding and autoregressive, XLNet was introduced.

The XLNet architecture is similar to a transformer encoder with some modifications. To learn the bidirectional context, it maximizes the expected log-likelihood over all permutations of the factorization order (not sequence order), which also serves as a proxy for the autoencoding training. To combat pretrain-finetune discrepancies and independence assumption, XLNet incorporated autoregressive formulation in the model. To note that the standard transformer can't implement permutation modeling and won't be aware of the target, so it is reparameterized by introducing two hidden states 'h' (content stream) and 'g' (query stream) and modified attention equations accordingly. They called it Masked Two-stream Attention.

The query stream is dropped during fine-tuning, and only the last-layer query representation is kept. To enhance the performance further, it integrated the idea of Transformer-XL and used memory in the model to account for large token sequences. It adopted the concept of using the Segment Recurrence mechanism and Relative Segment Encodings, unlike BERT/RoBERTa, which uses absolute segment encoding. It may be very tempting to use XLNet everywhere and get good results. Still, a notable disadvantage stopping XLNet from taking over this field is its considerably massive computational power compared to BERT and RoBERTa.

3 Downstream Tasks

In the Transformers paper, the Transformer model's performance is evaluated against a Machine Translation task using English-German and English-French datasets as well as an English constituency parsing task. The other two papers do not provide direct comparisons to these results as they are only encoder models and do not have decoders for generative capacity.

The RoBERTa and XLNet models have similar focus and can be readily compared directly on many tasks as listed in the table. The two models are from around the same time and perform similar in evaluation metrics with slight differences in most benchmarks. However, both models show significant improvement over the BERT model due to architectural improvements in XLNet and difference in pretraining strategies in both RoBERTa and XLNet as compared to BERT.

The tasks are encoded as follows: A: Reading Comprehension Using SQuAD2.0 Dataset (F1-score); B: Multi Genre Natural Language Inference (GLUE Benchmark); C: Question Answer Natural Language Inference (GLUE Benchmark); D: Quora Question Pair (GLUE Benchmark); E: Recognising Textual Entailment (GLUE Benchmark); F: Stanford Sentiment Tree Bank (GLUE Benchmark); G: Microsoft Research Paraphrase Corpus (GLUE Benchmark); H: Corpus of Linguistic Acceptability (GLUE Benchmark); I: Semantic Textual Similarity (GLUE Benchmark); J: Winograd Natural Language Inference (GLUE Benchmark); K: ReAding Comprehension dataset from Examinations (RACE) (Accuracy)

	A	B	C	D	E	F	G	H	I	J	K
RoBERTa	89.795	**90.8**	**98.9**	90.2	**88.2**	96.7	92.3	67.8	**92.2**	89.0	**83.2**
XLNet	**90.689**	90.2	98.6	**90.3**	86.3	**96.8**	**93.0**	67.8	91.6	**90.4**	81.7