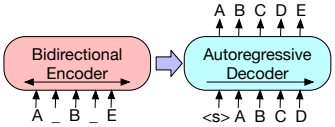
BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension

1 Contribution

The work has proposed a denoising autoencoder, named BART, for pre-training sequence-to-sequence models. BART architecture is a combination of Bidirectional Transformer model like BERT and Auto-Regressive model like GPT (please refer to Figure 1).

(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.

(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.

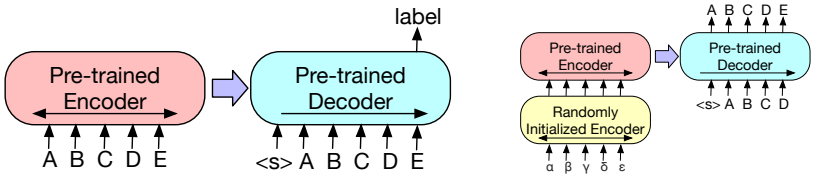


(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure 1: BART Model

2 Pre-training of BART

BART is pre-trained by corrupting documents and then optimizing a reconstruction loss – the cross-entropy between the decoder's output and the original document.



(a) To use BART for classification problems, the same input is fed into the encoder and decoder, and the representation from the final output is used.

(b) For machine translation, we learn a small additional encoder that replaces the word embeddings in BART. The new encoder can use a disjoint vocabulary.

Figure 2: Fine tuning BART for classification and translation

3 Fine-tuning of BART

The fine-tuning of BART for several downstream tasks are elaborated below:

Sequence Classification For sequence classification tasks, the same input is fed into the encoder and decoder, and the final hidden state of the last decoder token is fed into new multi-class linear classifier.

Token Classification For token classification tasks, such as answer endpoint classification for SQuAD, the complete document is fed into the encoder and decoder, and the top hidden state of the decoder is used as a representation for each word. This representation is used to classify the token.

Sequence Generation The encoder input is the input sequence, and the decoder generates outputs autoregressively.

The schematic of the fine-tuning of text classification and machine translation task is shown in Figure 2

4 Advantage of BART

Existing denoising autoencoders like BERT are tailored to specific noising schemes. BART allows to apply any types of document corruptions like token masking, token deletion, text infilling, sentence permutation etc. during pre-training.

5 Results

The work has provided results showing that BART matches the performance of RoBERTa with comparable training resources on GLUE and SQuAD, achieves new state-of-the-art results on a range of abstractive dialogue, question answering, and summarization tasks, with gains of up to 6 ROUGE. BART also provides a 1.1 BLEU increase over a back-translation system for machine translation, with only target language pre-training.