COMPRESSING TRANSFORMER-BASED SEQUENCE TO SEQUENCE MODELS WITH PRE-TRAINED AUTOENCODERS FOR TEXT SUMMARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We proposed a technique to reduce the decoder's number of parameters in a sequence-to-sequence (seq2seq) architecture for automatic text summarization. This approach uses a pre-trained AutoEncoder (AE) trained on top of a pre-trained encoder to reduce the encoder's output dimension and allow to significantly reduce the size of the decoder. The ROUGE score is used to measure the effectiveness of this method by comparing the four different compression rates: 96%, 66%, 50%, 44%. A few well-known frozen pre-trained encoders (BART, BERT, and DistilBERT) have been tested, paired with the respective frozen pre-trained AE's encoder to test the reduced dimension latent space's ability to train a 3-layer transformer decoder. Also, we repeated the same experiments on a small transformer model that has been trained for text summarization. This study shows an increase of the R-1 score by 5% while reducing the model size by 44% using the DistilBERT encoder, and competitive scores for all the other models.