Project Report

Overview:

Newspaper plays a significant role in our day to day life. In a news article, readers are attracted towards headline. Headline creation is very important while preparing news. Our goal is to implement text summarization by generating headline for a news body using recurrent neural networks.

Data:

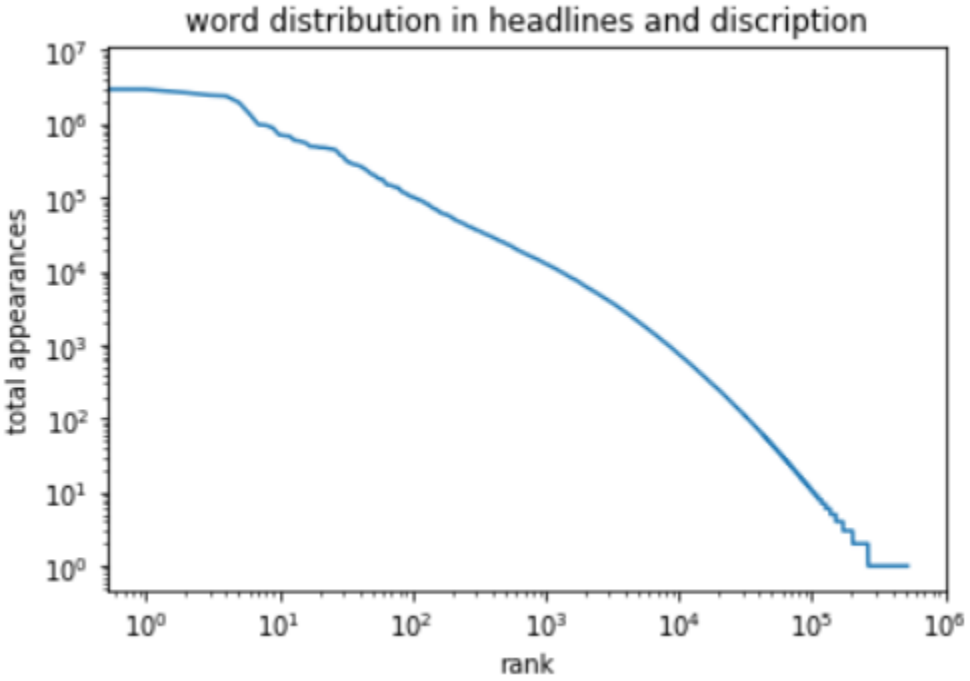
We have used "All the news dataset" from Kaggle. The data set consisted of approximately 1,50,000 articles. We extracted only title and content from the dataset.

We preprocessed our data by removing punctuations, stop words and converting the text lower case. Later, we tokenized the title, content and saved it as a pickle file.

We divided our data set into train (100000 articles), validation (30000 articles) and test validation (20000 articles) sets respectively.

We are using google news word2Vec file to map words in our data to vectors.

Below graph depicts words distribution in headlines and articles



Architecture:

We use the encoder-decoder architecture described in the figure below: The architecture consists of two parts - an encoder and a decoder - both by themselves recurrent neural networks.

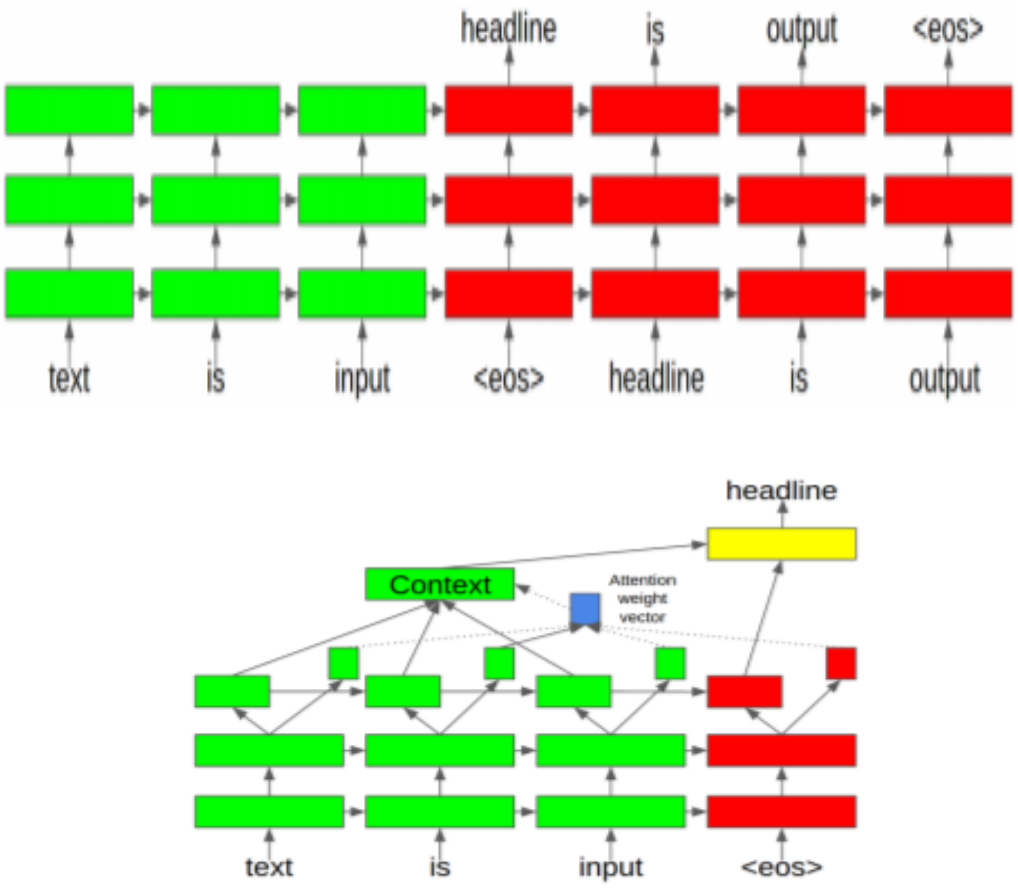


Figure 3: Simple attention

As shown in the figure news article and headline is fed into our model. At the end of headline we generate <unk> token to replace the word if it does not exist in our WordtoVec model, <eos> token to denote end of the headline and an empty token. For our decoder model, <eos> token is first fed as an input. Decoder model then uses a SoftMax layer and attention mechanism to generate each word of a headline.