News Headline Generation

04.07.2018

Final Project:
INFO 7374 Cognitive Computing and Deep Neural Networks

Naini Shah
Vighnesh Venkatakrishnan

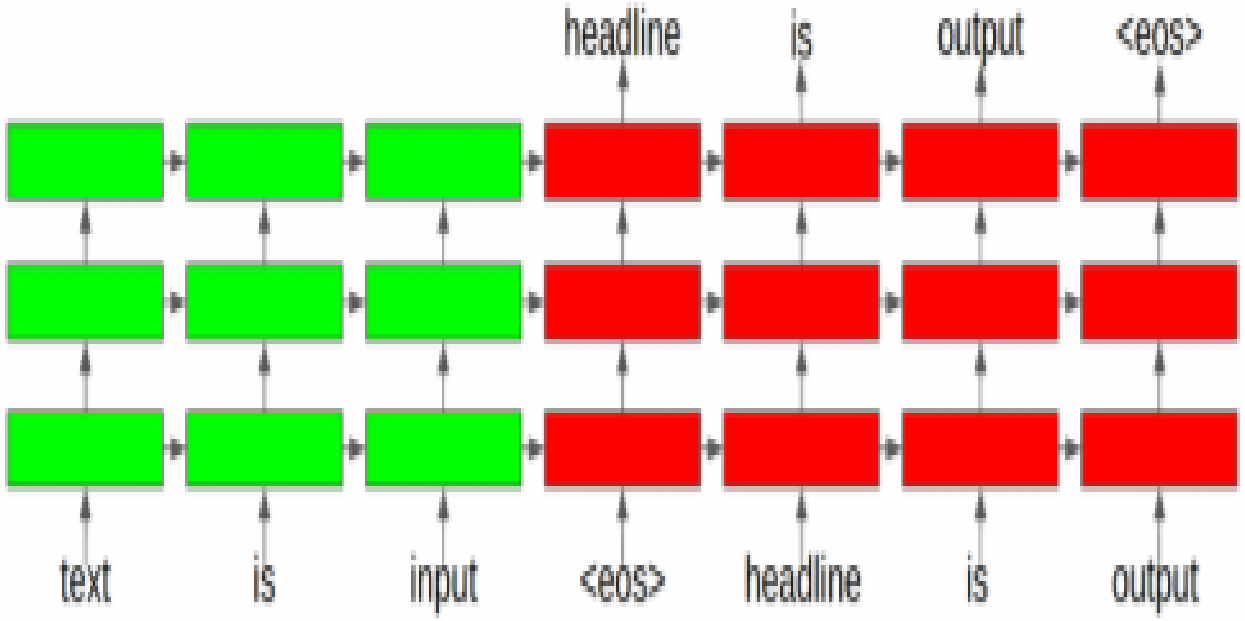
OVERVIEW

- 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.
- Our app can help journalists, bloggers quickly come up with the headline for the article they are writing
- It can also help proof readers to quickly understand the summary of an article and whether the headline describes the summary appropriately

DATA PREPROCESSING

- 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 from csv file by removing punctuations, stop words and converting the text to 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 (29999 articles) and test validation (12567 articles) sets respectively.
- We are using google news word2vec file to map words in our data to vectors and to initialize weights in the model

MODEL ARCHITECTURE



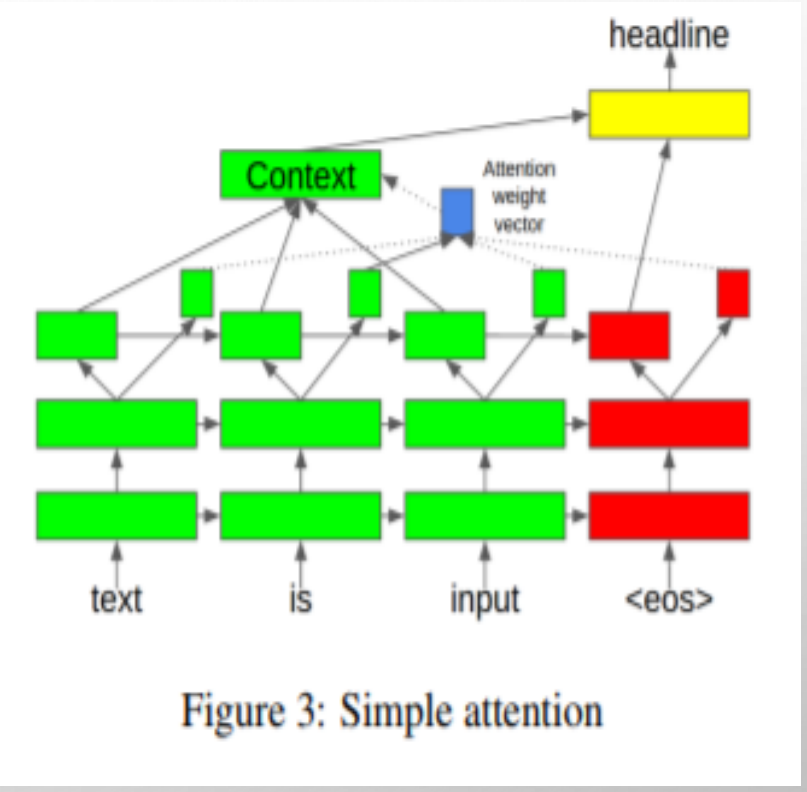


Figure 3: Simple attention

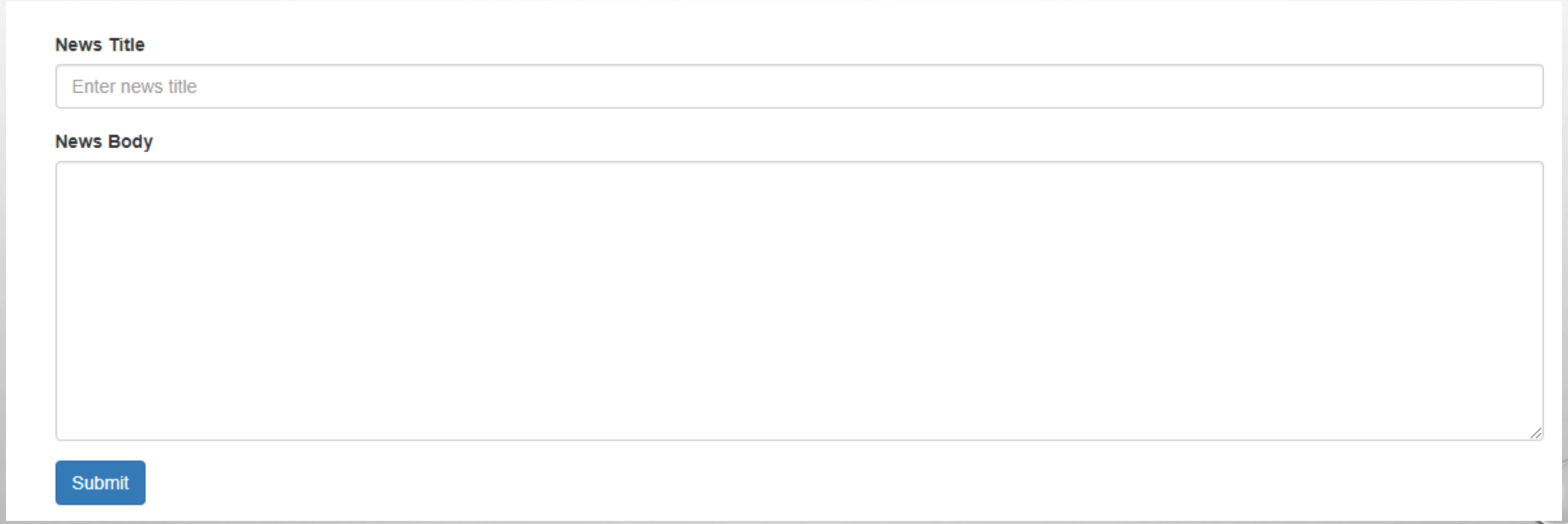
TRAINING

- In a batch size of 128 from the training pickle file, get headlines and descriptions
- Convert words to vectors less than max size
- Perform padding/trimming and add <eos> at the end of desc in X and head in y
- X contains head + description, Y has generated description (which is actual description while training)
- Flip random words in y with prev word of generated headline instead of actual headline
- Perform one hot encoding on generated headline
- Fit model and calculate bleu score, save model weights if current epoch better than previous one

TESTING

- Load saved weights
- Get headlines and descriptions from pickle file (or do real time using app)
- Perform beam search to calculate most frequently occurring words
- Convert list to string and display sentence

WEB APP (INPUTS)



OUTPUT

newstitle	Some Republicans grill Ryan over House chaplain firing
newsbody	At least two Republican House members confronted Speaker Paul Ryan on Friday morning at their caucus meeting over the firing of the House chaplain, multiple Republicans coming out of the meeting Friday told CNN. One of the members who brought up concerns was New York Republican Rep. Pete King, who said afterward that Ryan's explanation for asking the Rev. Pat Conroy, a Jesuit priest, to resign was "unsatisfactory." Both King and Ryan are Catholic.
Generated Title	firing chaplain House over Ryan grill Republicans Some The Trump is the New York on Trump The the
Cosine Simalrity	0.5773502691896258

CHALLENGES FACED

- Web scrapping/ preprocessing the data
- Resource exhaustion – large word2vec, large batch size, attention mechanism, etc
- Time taken to train model with bleu score
- Integrating with web app/ training web app on AWS
- Words like “Trump”, “New York Times” appear in every result since the model is trained heavily on Trump articles from NYT

FUTURE WORK

- One can try using gru based model instead of lstm and train for more epochs.
- Bi-directional rnn can also be considered since it provides more information to the model

CONCLUSION

We successfully implemented news headline generation app with encoder-decoder architecture along with attention mechanism. our model generated headline that has cosine similarity of up to 0.5 while comparing it with original headline.

REFERENCES

- RESEARCH PAPER: HTTPS://ARXIV.ORG/PDF/1512.01712.PDF
- REFERENCE CODE: HTTPS://GITHUB.COM/KABRAPRATIK28/DEEPNEWS
- RNN: HTTPS://MACHINELEARNINGMASTERY.COM/HOW-DOES-ATTENTION-WORK-IN-ENCODER-DECODER-RECURRENT-NEURAL-NETWORKS/
- DESC REVERSING: https://arxiv.org/pdf/1409.3215.pdf