Team 5

Medha Vempati (2018101063)
Yallamanda Mundru (2019201029)
Shobhit Raj Gautam (2019201056)
Abhinav Gupta (2019201064)

Context Aware Hashtag Generation
Technical Report

November 18th, 2020

Problem Overview

Social media and its reach has grown exponentially over the last two decades. Everything ranging from tutorials to instantaneous news updates have been made available on social media platforms, making it highly necessary to have an efficient filtering mechanism in place. Hashtags are useful in searching for focused content in a sea of information. This project aims to generate meaningful and relevant hashtags given some target text.

Dataset

We scraped Instagram data from diverse areas of content including food, movies, fashion, travel, news, tech, blm movement, etc.
We identified multiple popular public accounts in each field, we chose the users that have significantly sized captions and use hashtags frequently to ensure better quality of results.

For each post we scraped the caption text, top 20 comments, likes, and hashtags. In the end we had roughly 17000 data points.

Database Schema:

Post ID	Text(Caption)	Hashtags	Top 20 Comments	Likes

Data Processing

1) Converted emojis to text to add meaning.

2) Stripped punctuation, stop words etc.

3) To tackle the issue of **limited text size** on Instagram we combined comments with the caption text.

4) Computed the number of likes each hashtag received over the span of all posts.

5) Converted each post into a numerical vector where we assigned one number corresponding to each word.This Numerical vector is passed into the embedding layer of models described below.

Architecture of Model

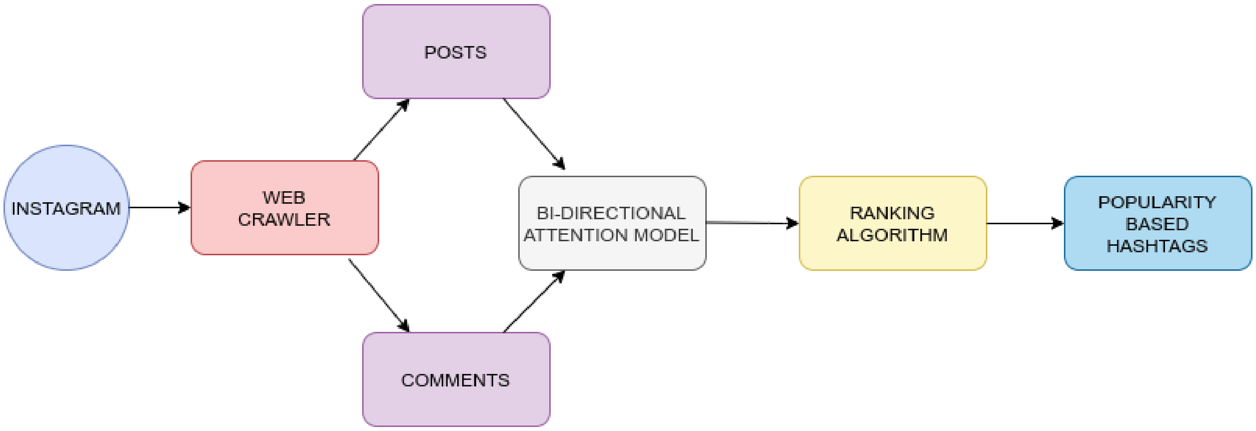


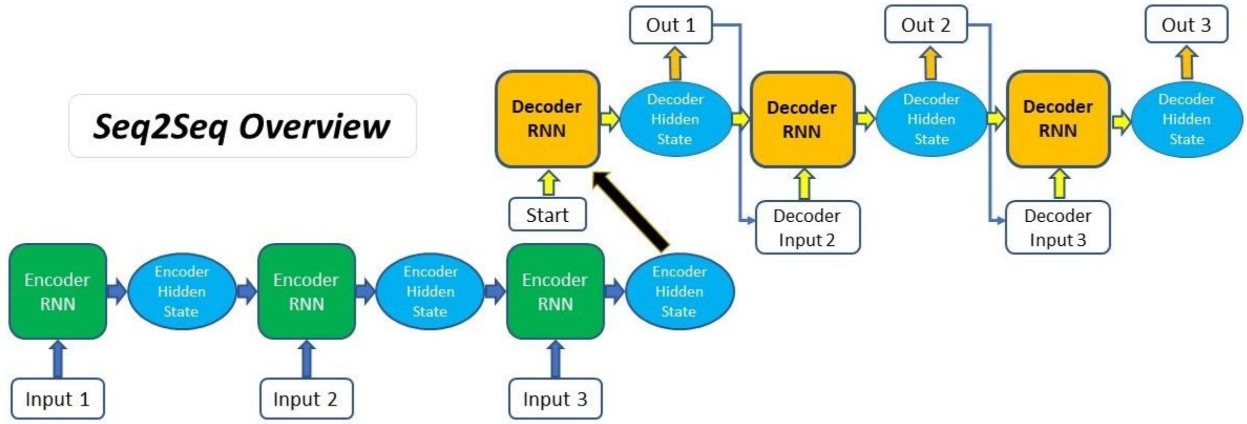
FIG: 1 High Level Overview

Models

1. Bahdanau's Encoder-Decoder model

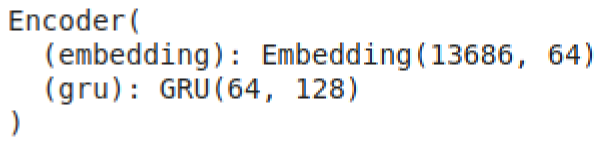
Bahdanu’s Attention model is proposed to improve sequence-to-sequence model in machine translation by aligning the decoder with the relevant input sentences and implementing Attention. The entire step-by-step process of applying Attention in Bahdanau’s paper is as follows:

1. Producing Encoder Hidden States.
2. Calculating Alignment scores.
3. Softmax on aligned scores.
4. Computing Context Vector.
5. Decoding output.
6. Repeating above step



1.1 Encoder Architecture in our model

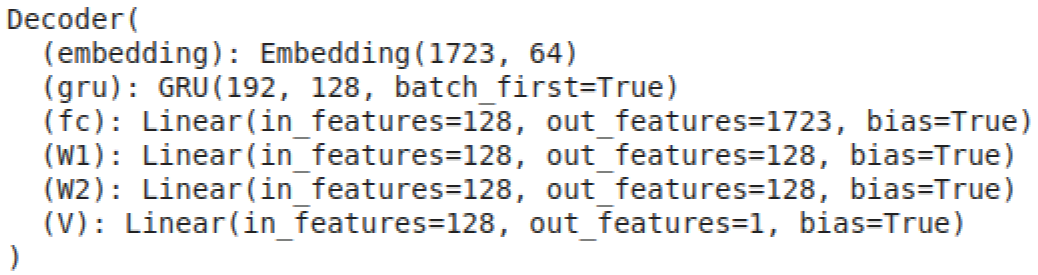
The encoder of our model has 2 layers. One is the Embedding layer which converts the number representation of each word into word vector.We chose our embeddings to be of size 128.Second layer is GRU,in our model we choose to have 64 units.Below is the dimensions of encoder model.



```
Encoder(
  (embedding): Embedding(13686, 64)
  (gru): GRU(64, 128)
)
```

1.2 Decoder Architecture of our model:

Decoder of our model contains the architecture for attention and actual decoding mechanism. Weights W1,W2 and V are used for attention mechanisms. GRU unit in decoder consists of 128 units.

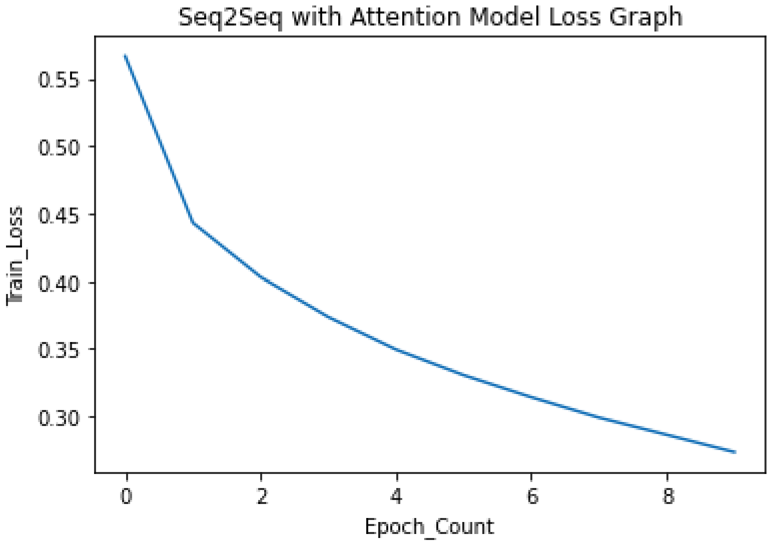


```
Decoder(
  (embedding): Embedding(1723, 64)
  (gru): GRU(192, 128, batch_first=True)
  (fc): Linear(in_features=128, out_features=1723, bias=True)
  (W1): Linear(in_features=128, out_features=128, bias=True)
  (W2): Linear(in_features=128, out_features=128, bias=True)
  (V): Linear(in_features=128, out_features=1, bias=True)
)
```

1.3 Loss Function : CrossEntropyLoss

- The CrossEntropyLoss function calculates both the log softmax as well as the negative log-likelihood of our predictions.
- Our loss function calculates the average loss per token

1.4 Evaluation:



Transformer-Based Models

After researching various existing transformer-based methods including GPT2, XLNet, Bert, RoBerta etc, we identified BertGeneration and BART to be the most suitable to our problem statement.

We used the **simple transformers** library to implement the BART model, and trained it on our hashtag data.

Also, used TRANSFORMERS library for BERT model using TfBertModel Transformer so that can work as text generation and predict hashtags.

Literary References

- *Deep Keyphrase Generation* : it attempts to capture the deep semantic meaning of the content with a deep learning method. [Rui Meng, Sanqiang Zhao, Shuguang Han, Daqing He, Peter Brusilovsky, Yu Chi] https://arxiv.org/pdf/1704.06879.pdf

- *Topic-Aware Neural Keyphrase Generation* : In this along with seq2 seq generation they also used the topic to which the post might belong to improve performance [Yue Wang, Jing Li, Hou Pong, Chan, Irwin King, Michael R. Lyu, Shuming Shi] https://arxiv.org/pdf/1906.03889.pdf

- *Microblog Hashtag Generation via Encoding Conversation :* In this along with seq2 seq generation they also used conversational contexts to address the data sparsity issue im microblogging [Yue Wang, Jing Li, Irwin King, Michael R. Lyu, Shuming Shi] https://arxiv.org/pdf/1905.07584.pdf.

- Effective Approaches to Attention-based Neural Machine Translation: Used this as reference for text generation. [Minh-Thang Luong, Hieu Pham ,Christopher D. Manning] https://arxiv.org/pdf/1508.04025.pdf

- Huggingface library for implementing BERT, BART : This helps in using library functions, https://huggingface.co/transformers/model_doc/bert.html#tfbertmodel