Article Classification using Hierarchical Attention Networks

Deep Learning Project Plan

Sahil Makwane (sm9127)
Sakshi Mishra (sm9268)

Problem Statement

Large volumes of scattered and uncategorized articles present on the internet can cause difficulty in accessing desired information. This issue engenders the need of classification of articles in groups with tags so that the user can quickly retrieve articles they wish for. Classification of text can be done manually but it is an arduous task. Thus, we introduce machine learning techniques for text classification so that text classification can be done quickly for large amounts of textual data.

Literature Survey

Articles are text documents that can be considered as a sequence of sequence. Meaning, sequences of words makeup a sentence, and sequences of sentences make up an article. There are unsupervised techniques for text classification such as Bag of Words, TF-IDF and N-grams which represent the text in terms of frequency of occurrence of words. However, these do not account for the contexts of words and sentences with much accuracy.

On the other hand, Hierarchical Attention Networks or HANs consider the hierarchical structure of the document. They also account for the context of the words and sentences that appear in the document. For instance, the context of the word 'book ' in the sentences, "I read a book" and "I need to book flight tickets" are different. HANs also takes into consideration that not all words and sentences equally contribute towards the representation of the article.

We shall be implementing attention mechanism both on word level and sentence level. The attention mechanism sets weights on words and sentences depending upon their context. The network uses word encodings of each sentence and then applies attention mechanism on the words. This results in a sentence vector. Now these sentence vectors are encoded, and attention mechanism is applied to each sentence vector. The network congregates such sentence vectors to represent the document.

Model

First the text in our datasets will be preprocessed by lemmatizing and tokenization techniques. This preprocessed data is further passed as in input to the HAN model. Input is a 3D Tensor with dimensions in terms of (samples, steps, features). The output is a 2D tensor with dimensions in terms of (samples, features).

The Hierarchical Attention Networks model consists of the following parts:

1. Embedding Layer: This layer shall convert words into a vector of real numbers.
2. Word Encoder Layer: To get a rich description of words, bi-directional GRU as used at the word level

3. Word Attention layer: To get the important information from each word in a particular sentence, attention mechanism is applied at the word level.
4. Sentence Encoder Layer: To get a rich description of sentences, bi-directional GRU as used at the sentence level
5. Word Attention layer: To get the important information from a sentence, attention mechanism is applied at the sentence level.
6. Fully Connected layer cascaded with a softmax layer will predict the probabilities for the document for each category.

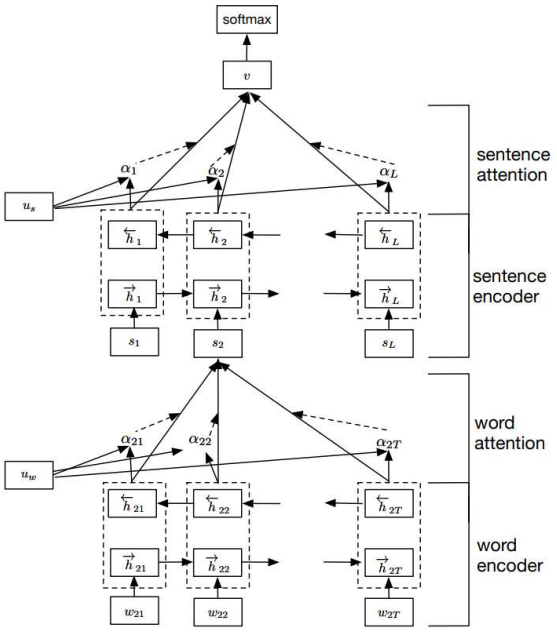

Fig 1: Hierarchical Attention Network

Reference:https://www.cs.cmu.edu/~./hovy/papers/16HLT-hierarchical-attention-networks.pdf

Dataset

Following datasets will used to implement the model:

1. **IMDB Movie Reviews**: Has labeled data of 50,000 IMDB movie used for sentiment analysis. For reviews having rating <5 the sentiment score is set to 0 and for rating >=7 sentiment score is 1.
2. **News Category Dataset**: Consists of 200,000 news headlines for the years 2012-2018.
3. **Yelp Reviews**: Consists of yelp reviews for different businesses. It has ratings from 1 to 5, 5 being the highest.

Conclusion

With the help of HANs we shall be able to classify the reviews and the news articles in the datasets to appropriate categories in terms of sentiments for reviews and news category for news articles. Few challenges that we need to take care of are that we need to account for words for multiple languages.

References

1. https://www.cs.cmu.edu/~./hovy/papers/16HLT-hierarchical-attention-networks.pdf
2. https://humboldt-wi.github.io/blog/research/information_systems_1819/group5_han/
3. https://github.com/arunarn2/HierarchicalAttentionNetworks