Deep Learning

Project Proposal

Sahil Makwane, sm9127
Sakshi Mishra, sm9268

OBJECTIVE

Implementing Hierarchical Attention Models to classify news articles.

PROBLEM STATEMENT

There are huge volumes of texts available on the internet that need to be categorised. The purpose of document classification is to manage these documents and gain valuable information from them faster and easily. For instance classification can help categorise news articles in appropriate news categories by examining the article texts. Classification can be done automatically with the help of machine learning techniques. One such technique that we are going to use in this project is Hierarchical Attention Networks.

APPROACH

We can say that a document is a sequence of sequence. Meaning, a sequence of words group together to form a sentence, and further a sequence of sentences make up a document. To examine this hierarchical structure of a document we chose to implement Hierarchical Attention Networks. Moreover the attention mechanism can be used to emphasize on those words and sentences that convey essential information in a document. Thus HANs help identify those particular words and sentences that represent the document.

We currently plan to analyse news classification datasets available on kaggle and train our HAN model. We plan to incorporate word encoding and sentence encoding as well. With the help of attention mechanism, we'll implement attention to words and sentences. We aim to classify a chunk of new articles into appropriate categories such as sports, rentals, etc. with HAN model.

REFERENCES

https://www.researchgate.net/publication/305334401_Hierarchical_Attention_Networks_for_Document_Classification
https://humboldt-wi.github.io/blog/research/information_systems_1819/group5_han/#text-classification-with-hierarchical-attention-networks