test

Adrian Eriksen

Thesis submitted for the degree of
Master in Language Technology
60 credits

Department of Informatics
Faculty of mathematics and natural sciences

UNIVERSITY OF OSLO

Autumn 2020

test

test

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Introduction

In this essay I will explain the challenges of cross-domain sentiment analysis and how we might use transfer learning to solve it. Sentiment analysis is the process of trying to understand the sentiment behind a statement or document using machine learning. This can, among other things, be used to get information from reviews that can provide useful information. There are many different ways to formulate a sentiment. A movie review might state "The movie is not bad at all.". If we simply look for words like "bad" and classify them as negative, we will get inaccurate results.

- Explain structure of essay -

Pretraining: BERT/ELMO, task specific

LTG?

Sentiment Analysis

Sentiment analysis (SA) is the computational treatment of opinions, sentiments and subjectivity of texts. SA has a lot of different utilites, one of the bigger ones being opinion mining. There has been a lot of work using SA to mine opinions from e.g Twitter, to figure out what the general opinion on different topics are.

Pretraining

Pretraining languange models is one of the key components of natural language processing. Historically we have pretrained a model for a single task. If we wanted to make a model that can predict whether a movie review is positive or negative, we would feed the model a large amount of labelled reviews so that it may learn the differences. The challenges this raises lies in how exactly this data should be stored in the model, and figuring out what to focus on when learning. In NLP there are a lot of ways to approach text processing. The most intuitive way might be to split the words by sentences (punctuations) and whitespaces (words) and have the model learn which words and sentences are the most recurring in the different reviews, disregarding universally common words like 'the'. In addition, n-grams have been used to try and get a deeper understanding of context. Instead of only looking at one word at a time, we look at the current word as well as the n previous words. This is just one of many techniques that can be applied to improve the model. Others include POS-tagging, NP chunking and Named Entity Recognition to mention some.

There are however a couple of challenges when considering the different types of text processing for pretraining language models. A lot of text processing requires a large amount of manual labour beforehand. While splitting sentences and words in most cases are a trivial task for a program to solve, POS-tagging and NER are not. There is a fairly large amount of data that already have this, namely the Brown Corpus, but limiting language models to only learn from one source would be crippling in regards to making newer and better models. In addition, if we would feed all the additional information about every word and sentence to the model, the computational power required would be insurmountable.

The large amount of information that each word could contain, turned out to be solvable by vectorization. It was discovered that by assigning each word a vector, we could keep a lot of it's properties without having to process all of it's information. If we take the example of the words "king" and "queen", they would be placed close to eachother in the vector space, along with words like "royalty" and "palace". However, "queen" would be closer to "female", while "king" would be closer to "male". With the invention of word vectorization, it

became possible for small research groups with limited computational power to train their own language models.

Vectorization:
ELMo - Embeddings from Language Models was the next step in the evolution of word vectorization. Each token is assigned a representation that is a function of the entire input sentence. Vectors derived from a bidirectional LSTM.

Finetuning:
BERT - Bidirectional Encoder Representations from Transformers is one of the key innovations in the recent progress of contextualized representation learning. Whereas previous language representation models like OpenAI GPT had been unidirectional,

Transfer Learning

Transfer learning is a means to exract knowledge from a source setting and apply it to a different target setting