Legal Judgemnet Prediction in English using Bert, Hierarchical Bert, Hierarchical Attention Networks and Longformers

Devyani Lambhate
Department of Computational and data Science, IISc

https://github.com/Devyani-Lambhate/Legal-Judgement-prediction

Abstract

Legal Judgement Prediction is the task of predicting a court case's outcome, given the case's facts. This task is a binary classification task, where we are interested in knowing if any legal articles are violated or not in a given case. The legal text describing the cases' facts is generally very long and very domain-specific, motivating to design and explore network architectures that work with long and domain-specific documents. The dataset used contains cases from the European Court of Human Rights. This prediction aims not to replace the Legal professionals but to let them better understand the biases and critical facts in a court case. Such models may assist legal practisioners. It will improve access to justice by reducing legal costs.

1 Introduction

There are many tasks associated with legal judgment like predicting the importance of a case, predicting which articles are violated in a case, court opinion generation and analysis. I will focus only on the binary legal judgment prediction task, which predicts the outcome of a case given the text describing the facts. The objective is to classify a case as positive if any human right article is violated and negative otherwise.

Very few NLP models have been tested for the Legal Judgement Prediction task because of the lack of data available and the lack of models that process long documents. A new publicly available English legal judgment prediction dataset of cases from the European Court of Human Rights(ECHR)[1] was released in 2020. Before this dataset was released, most of the models were designed and tested for Chinese datasets. The authors of [2] proposed a Hierarchical BERT model that outperforms the traditional BERT model in the Legal Judgement by dealing with the BERT's[3] length limitation. In this project, I have implemented and compared BERT, Hierarchical BERT, Hierarchical Attention Network[4], and Longformers[5].

All of these models are designed to process long text documents. BERT model can only process documents or sentences up to 512 tokens. This limitation on the token size is an impedance while working with long documents or sentences. Therefore Hierchical BERT model was proposed by[1]. Longformer is a Transformer-based model[6] that can process long sequences due to the enhanced self-attention operation that executes in o(n) time. The hierarchical nature of the document inspires Hierarchical Attention Networks. At the first level, it deals with words, and at the second level, it deals with sentences.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

2 Dataset

The European Convention of Human Rights(ECHR) dataset contains approximately 11.5k cases from ECHR's public database. For each case, the dataset provides a list of facts extracted using regular expressions from the case description. Each case is also mapped to articles of the Convention that were violated (if any). ECHR also assigns an importance score. The training and development sets contain cases from 1959 through 2013, and the test set from 2014 through 2018. The training and development set is balanced to avoid any biases towards a particular label. The train set contains 7,100 cases, and the test set contains 2998 cases. The documents are, on average, around 2500 words long.

3 Models

3.1 BERT

In recent years, researchers have been working on models based on transformers. The motivation of which comes from the requirement of transfer learning. BERT(Bidirectional Encoder Representations from Transformers) model has presented state-of-the-art results in a wide variety of NLP tasks, including Question Answering, Natural Language Inference, Neural translation and others.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.

As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional, though it would be more accurate to say that it's non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

The sentences or documents need to be truncated before sending it to the BERT model because the input's maximum size cannot be more than 512 in this model. This is a major limitation while dealing with long documents. In this Legal Judgement prediction task, the BERT model is performs very poorly.

3.2 Hierarchical Attention Networks(HAN)

The model Hierarchical Attention network is inspired by the hierarchical nature of a document. It includes the information from sentence level as well as from the word level. The model consists of several parts like a word sequence encoder, a word-level attention layer, a sentence encoder and a sentence-level attention layer. The word and sentence encoder consists of Gated Recurrent Unit(GRU)[7]. The word level attention rewards sentences that are clues to correctly classifying a sentence and the sentences level attention rewards sentences that are clues to correctly classifying a document. The complete pipeline can be described as follows: First a word level encoder is applied separately on each sentence, Then word attention is calculated and words are combined according to the attention weights. In this way the encoding of a sentence is generated. These sentence embeddings are then passed through a sentence level GRU and final embedding of a document is generated by combining the weights of sentence level attention.

3.3 Hierarchical BERT

To surpass BERT's maximum length limitation, a hierarchical version of BERT (HIER-BERT) was proposed. Firstly BERT-BASE reads the words of each fact, producing fact embeddings. Then a self-attention mechanism reads fact embeddings, producing a single case embedding that goes through a similar output layer as in HAN.

3.4 Longformers

The major drawback of transformers based models is that they cannot attend to longer sequences. The attention mechanism used in transformer model is O(n^2), which is a compute limitation for extending the transformer to larger models. To overcome this issue the Longformer combines several attention patterns like Sliding Window, Dilated Sliding Window and Global Attention (full self-attention). These attention patterns reduces the complexity from O(n^2) to O(n). Like transformers, pretrained versions of Longformers were also available. My assumption was it will perform poorly as the domain specific data is absent, but it gives best F1-score among all the four models discussed.

4 Results

Table 1: Results

Model	Precision	Recall	F1-score
BERT	45.3	45.8	39.0
HAN	85.7	87.5	85.1
Hier-BERT	90.4	79.3	82.0
Longformers	83.1	96.2	89.2

In the above experiments, I have shown Precision, Recall and F1-score because the ECHR dataset was unbalanced. It has around 60% positive examples and 40% negative examples. I achieved best recall and F1-score using Longformers model. All of these results are on test set. The Hierarchical Bert model is still training, thats why I have reported the numbers from the original paper only for this model. From the results we can say that longformers are performing best for the ECHR dataset, even though it does not encode any domain specific knowledge, but pretraining the model on a large corpus helps. It was also the second fastest model, following BERT which was the fastest in terms of training and evaluation.

5 Conclusions

Except BERT base, all the three models were performing good on the ECHR dataset. Some more experiments on different datasets are needed to check the consistency of these models. Longformers takes minimum time to run and gives best scores. Some more experiments are also needed to check the biases encoded in the model, specifically the demographic biases.

References

[1] ECHR's Data Repository (http:// hudoc.echr.coe.int/eng?i=001-193071)

[2] Chalkidis, I., Androutsopoulos, I., Aletras, N. (2019). Neural legal judgment prediction in English

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[4] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E. (2016). Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (pp. 1480–1489).

[5] Beltagy, I., Peters, M., Cohan, A. (2020). Longformer: The long-document transformerarXiv preprint arXiv:2004.05150.

[6] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, ., Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing

[7] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.