DocBERT

28 December 2023 23:05

DocBERT - https://www.researchgate.net/publication/332493790_DocBERT_BERT_for_Document_Classification

- Input to the DocBERT is larger than the usual input to a BERT model and the input contains many labels
- Knowledge is distilled from the BERT model (Large) into a LSTM (Long short term memory) and having the performance equivalent to a BERT base model
  - Why LSTM??
  - LSTM they are types RNN that are designed to handle sequential data
  - Memory : Have special **Cells** that can store information for a long period of time
  - Gates : They have **3 gates (forget, input and output)** that control the flow of information for long periods
  - This combination of memory and gates allow LSTM to learn long term dependencies in data which means they can understand relationships between distant elements in a sequence
- Uses 30 times lesser parameters
- Why is it necessary?
  - Documents have a lot of labels across many classes, which is unlike any of the tasks that BERT examines
- This comes at the price of heavy computational complexity. BERT uses hundreds of millions of parameters.
- To alleviate for the computational burden knowledge distillation is performed
- Knowledge distillation is applied to transfer knowledge from BERT(large model) to Bi-LSTM.

Approach

- Introduces a fully connected layer over the final hidden state corresponding to the [CLS] input token.
  - The token's final hidden state is considered to hold a global representation of the entire input sequence
- During fine tuning we optimize the modal end-to-end, with additional soft-max classifiers parameters.
- We reduce the cross entropy and binary cross entropy loss for single label and then multiple labels tasks
- We distil knowledge from fine-tuned BERT (Large) into much smaller LSTM (reg)
- Knowledge distillation is performed using the training examples along with minor augmentations to form the transfer set.
- We combine the two objectives of classification using target labels and distillation, using soft targets for each example of the transfer set.

Training and hyper parameters

- While fine tuning BERT, optimize the number of epochs, batch size, learning rate and maximum sequence length, the number of tokens the document is truncated to.
- The quality of the model is quite sensitive to epochs. **So it needs to be tailored for each dataset**

Hyperparameter Analysis

- Maximum Sequence length: a decrease in the MSL has minor change in the F1 score
- Epochs :
  - smaller dataset need more number of epochs to converge
  - We see a significant drop in model quality when the BERT models are fine tuned for only four epochs