MIT 6.862 Pre-Proposal

Junyi Guo
Department of Biostatistics
Harvard University
Boston, MA 02115
junyiguo@hsph.harvard.edu

Yuxin Xu
Department of Biostatistics
Harvard University
Boston, MA 02115
yuxinxu@hsph.harvard.edu

Listed in this document are three ideas that we currently have in mind. Since we are not totally sure which one sounds better/has the right level of difficulty for this project, we decide to list them all and see if the reviewer has any suggestions.

1 Idea 1: Named Entity Recognition

Abstract

Our first idea focuses on one of the hot topics in NLP related tasks, which is the Named Entity Recognition problem. This topic is very interesting to us because we are curious about how machine can learn not only the meaning of words but also their context information in the entire sentence. We intend to follow the guidance provided by Sterbak[7] and the BERT Based Named Entity Recognition (NER) Tutorial And Demo[3] to fine tune a NER model on the Annotated Corpus data set we found from Kaggle[1]. During this project, we hope to learn some useful information extracting techniques with language models. Some possible helpful models we might use in this project include pre-trained Transformers or Bert-base-NER model provided on Huggingface. We would expect the model to learn the proper word types as well as their meanings during the training process and provide reasonable tagging for each words in the sentences.

2 Idea 2: Text Summarization

Abstract

Another project idea we have is also related to the applications of NLP models, which is the text summarization task. We would like to see how deep learning models for example Transformers and Bert can simplify a sentence, extract relevant information, and reconstruct the information into a comprehensive sentence. We intend to follow the guidance provided by Licht[6] and fine tune our models on the BBC News Summary[2] data set obtained from Kaggle. Some possible helpful models we might use in this project include pre-trained transformers or BERT model provided on Huggingface. We would expect the model to not only extract important and concise information from the input paragraph, but also manage to output a human readable sentence.

Preprint. Under review.

3 Idea 3: Pneumonia Detection

Abstract

On the other hand, given the pandemic and all the emphasis on medical care, we are interested in doing some medical-related image classification problems in hope of learning about possible ways to reduce the burden on front-line healthcare professionals.

We found an image data set[5] of 5856 images on pneumonia detection containing X-rays from normal patients and pneumonia patients. We plan to explore image-data conversion and the architecture of convolutional neural networks (e.g. Google Inception, VGGNet, ResNet) from packages like keras and tensorflow and use them to train a model on our pneumonia classification problem.

[4]

References

[1] Annotated corpus for named entity recognition, . URL https://www.kaggle.com/abhinavwalia95/entity-annotated-corpus#ner_dataset.csv.

[2] Bbc news summary, . URL https://www.kaggle.com/pariza/bbc-news-summary.

[3] Bert based named entity recognition (ner) tutorial and demo. URL https://www.pragnakalp.com/bert-named-entity-recognition-ner-tutorial-demo/.

[4] Identifying medical diagnoses and treatable diseases by image-based deep learning. 2018. doi: https://doi.org/10.1016/j.cell.2018.02.010.

[5] D. S. K. et. al. Identifying medical diagnoses and treatable diseases by image-based deep learning. 172(5):1122–1131, 2018. doi: https://doi.org/10.1016/j.cell.2018.02.010.

[6] G. Licht. Extractive summarization using bert. URL https://towardsdatascience.com/extractive-summarization-using-bert-966e912f4142.

[7] T. Sterbak. Named entity recognition with bert. URL https://www.depends-on-the-definition.com/named-entity-recognition-with-bert/.