Hyper-Parameter Optimization for Natural Language Processing (NLP) Models

INTRODUCTION

Battelle: Battelle is a private non-profit research organization that advances science and technology to have the greatest impact on our society and economy.

About Sematix: It can take up to 17 years to translate medical studies into practice. The Sematix NLP model will be used in Medical Literature and will save researchers much time reading through entire papers.

What is Natural Language Processing (NLP): A field of artificial intelligence that involves teaching machines to understand human language.

What is BERT: A powerful language model called to perform specific tasks like information extraction within NLP.

Our Goal: To extract crucial information from research and scholarly papers using the BERT transformer model.

Our Challenge: There are enormous amounts of medical text documents and data available. Their classification and the appropriate manifestation of their relations with each other become a necessary and tedious task. That is why we aim to fine-tune a BioBERT pre-trained language model, that has been pre-trained over large corpora of PubMed and PMC texts.

OUR MODELING PROCESS

First...

We used brat-parser to import the files and extract the named entity data to make the raw ANN files from Harvard readable by our model. Once the data was extracted, we wrote a custom script to format it into TSV files.

Then...

We applied the new TSV files to train a script based on HuggingFace BioBERT model, previously utilized on Barilla ingredient data. The script computed the prediction accuracy by using the provided data.

Finally...

We developed hyperparameter tuning scripts using the HyperBand and Population-based training algorithms from RayTune to predict the accuracy of our entity classifications.

RESULTS

Comparing Evaluation Results

- BERT
- BERT + PBT
- BERT + Hyperband

KEY TAKEAWAYS

<table>
<thead>
<tr>
<th>BERT MODEL</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large neural network that can be pre-trained to gain a general understanding of language</td>
<td>Faster convergence, Adaptive resource allocation, State-of-the-art performance</td>
<td>Limited exploration and not suitable for all problems</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BERT + Population-based optimization</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Efficient use of computing resources, Robust and improved performance</td>
<td>More complex than other techniques and has a high computational cost</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BERT + Hyperband</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust and improved performance</td>
<td>More complex than other techniques and has a high computational cost</td>
<td></td>
</tr>
</tbody>
</table>

CONCLUSION

- In the future, we will try to improve the model’s performance for every entity category, such as raising the accuracy of the ADE class by labelling more tags in the data.
- We could further expand the scope of the project to include a relation extraction model as well, which would enable us to build a knowledge graph of the entities and their relations.
- We can apply this model to other medical datasets, like those provided by n2c2 and i2b2.

ACKNOWLEDGEMENTS

A very special thank you to our TAs Ujjwal Garg and Yuhang Fang and our Battelle mentors Allen Chen and Mitch Gauthier. We would also like to thank Dr. Ward, Margaret Betz, and the entire Data Mine staff for their continuous support.

Dwijen Chawra - Vamsi Deeduvanu - Aryan Samantaray - Shilvi Agrawal - Fengxu (Kenny) Liu - Carissa Lukac - Nitin Murthy - Dave Patel