Benjamin LeBrun
benjamin.lebrun@mail.mcgill.ca

Data Selection Proposal
MAIS 202

31/01/20

Natural language inference (NLI) is the task of determining whether a hypothesis is true (entailment), false (contradiction), or undetermined (neutral) given a premise. For instance, the following sequence is tagged as *entailment*.[1]

(1) *Premise: A soccer game with multiple males playing.*
Hypothesis: Some men are playing a sport.

1 Dataset

For the purpose of this project, I will be using the MNLI dataset.[2] I've chosen this dataset in particular as it is used for most benchmark performance calculations. Additionally, I will be using HANS,[3] an NLI dataset that tests specific hypotheses about invalid heuristics that NLI models may adopt.[4]

2 Methodology

2.1 Data Preprocessing

To pre-process my data, I will need to lowercase all text, strip out accent markers, and convert white space to characters. Then, I will split punctuation on both sides. And finally, I will apply WordPiece tokenization to all text.

2.2 Machine learning model

In this project, I intend on implementing two popular models for NLI tasks: BERT, a model developed by Google, and RoBERTa, a model based on BERT developed by Facebook. I am choosing these models since they are the current state-of-the-art for NLI tasks. Both have code available on GitHub.[5][6]

2.3 Final conceptualization

For this project, I intend on preparing a poster presentation in which I will attempt to reproduce (to the best of my abilities) some of the results obtained in this paper: `https://arxiv.org/pdf/1902.01007v4.pdf`. In it, the authors implement a series of NLI models using the MNLI dataset. Then, they test these same models on HANS. They observe a drastic drop in performance, suggesting these models, most notably BERT, adopts a series of heuristics (outlined in the paper).

For the purpose of my project, I will attempt to reproduce the results obtained using BERT, and carry out a similar task on RoBERTa, which was not released at the time of the paper. More precisely, my project will consist of:

1. Implementing BERT and RoBERTa using the MNLI dataset.
2. Testing BERT and RoBERTa on the HANS dataset
3. Re-training (or fine-tuning) BERT and RoBERTa using examples from HANS.
4. Re-testing the models to see if performance improves

Additionally, depending on the complexity of implementing these models, I can carry out a similar analysis on a recurrent neural network, which seem to be suitable for tasks in which context within a sequence is useful.

The metric I will use to assess performance is the percentage of premise/hypothesis pairs the model correctly predicts. The current state-of-the-art is 90.8%, which was obtained by RoBERTa.[7]

[1] http://nlpprogress.com/english/natural_language_inference.html
[2] https://www.nyu.edu/projects/bowman/multinli/
[3] https://github.com/hansanon/hans
[4] https://arxiv.org/pdf/1902.01007v4.pdf
[5] https://github.com/google-research/bert#fine-tuning-with-bert
[6] https://github.com/pytorch/fairseq/blob/master/examples/roberta/README.md
[7] Ibid 1