Inference Detection in NLP Using the MultiNLI and SNLI Datasets
Term Paper, CSC 820, Spring 2022

Chris Huber*
SFSU

ABSTRACT

This paper describes the process of classifying inference from the MultiNLI (Multi-Genre Natural Language Inference) and SNLI (Stanford Natural Language Inference) datasets, corpii of over 422,000 and 550,000 sets of paired sentences respectively. Inference involves examining pairs of sentences and determining if they are a contradiction, an entailment, or neutral. There are several models capable of this including RoBERTA and XLNet which can be fine-tuned to produce increased accuracy.

1 INTRODUCTION

Natural language inference revolves around determining whether a hypothesis is an entailment, contradiction, or neutral. An example is shown in Table 1. The MultiNLI corpus has 0.9 and 1.0 versions, both of which are examined (TBD).

To perform this task, I established a baseline accuracy using an algortihm called RoBERTa which was developed by Google in conjunction with Fairseq [3] [1], a Google-developed open-source toolkit designed for translation, data modeling, and other text related tasks.

RoBERTa is based on BERT (Bidirectional Encoder Representations from Encoders) which was developed by the Google AI team and is discussed in a paper called "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". [1] [2] In a nutshell, it works by using a Masked Language Model (MLM) which hides some of the tokens in the input and attempts to derive the missing words using the surrounding context. This also works for next-sentence prediction which is the use case for inference.

There are three baseline neural network implementations for RoBERTa which are based on CBOW (Continuous Bag of Words), bi-directional LSTM (Long Short-Term Memory), and ESIM (Enhanced Sequential Inference Model).[3] These are implementable by running the train_mlni.py script which uses TensorFlow to train the model on either MLNI data, a mix of MLNI and SLNI data, or on a single genre in the MLNI dataset. Dropout is used in all three implementations for regularization.

The inspiration for this project came from a list of tasks posted to the SuperGLUE website.[4] (The GLUE (General Language Understanding Evaluation) benchmark is a benchmark set of NLP tasks to be performed on sentence pairs which has an ongoing competition ranked by accuracy). RLE (Recognizing Textual Entailment) is a very current topic in NLP technology research and I was inspired to see what I could contribute. There have also been Kaggle competitions involving inference determination.[5]

Premise	Hypothesis	Label
A man inspects the uniform of a figure in some East Asian country	The man is sleeping.	contradiction
An older and younger man smiling.	Two men are smiling and laughing at the cats playing on the floor	neutral
A soccer game with multiple males playing	Some men are playing a sport	entailment

Table 1:

Example of evaluations of sentence pairs. Taken from http://nlpprogress.com/english/natural_language_inference.html

I propose, after getting benchmark results for one or more of the datasets, to refine the algorithm and/or try other methodologies in an effort to improve my F1 score.

2 RELATED WORK

A paper entitled "A large annotated corpus for learning natural language inference" DBLP:journals/corr/BowmanAPM15 [6] describes the development of the SNLI. It was published in August 2015 and was an effort to provide a good dataset for benchmarking since previous inference corpii were either algorithmically generated or too small and as such impeded effective analysis. It consists of pairs of image captions which were labelled using crowdsourcing via Mechanical Turk.

A paper entitled "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference" DBLP:journals/corr/WilliamsNB17 [4] [7] discusses the creation of the MultiNLI (Multi-Genre Natural Language Inference) corpus. It is comprised of 433,000 annotated examples designed to train machine learning algorithms. It offers data from 10 distinct genres of written and transcribed oral communication. It was also labelled using crowdsourcing via Mechanical Turk.

Prior to MLNI, the Stanford NLI corpus was the largest available corpus but fell short in a couple of ways. First, it was all from a single genre due to the fact that it was drawn only from image captions which lacked the robust language varieties that exist. The authors found it insufficient to provide a good benchmark for NLU (Natural Language Understanding).

The SNLI corpus contains 550,152 sets of labelled sentence pairs in its dataset. I also decided to try testing against it since it also produces measureable results. I took a 10,000 row subset of the data which takes approximately 2 hours to process since the entire set would take 1100 hours to process using my Mac laptop if it ever even completed.

*email:chrish@sfsu.edu

[1] https://aclanthology.org/N19-4009.pdf
[2] https://arxiv.org/pdf/1810.04805.pdf
[3] https://github.com/NYU-MLL/multiNLI
[4] https://super.gluebenchmark.com/tasks/
[5] https://www.kaggle.com/c/multinli-matched-open-evaluation/data
[6] https://arxiv.org/pdf/1508.05326.pdf
[7] https://arxiv.org/pdf/1704.05426.pdf

RoBERTa is a technology developed by Google described in the paper "RoBERTa: A Robustly Optimized BERT Pretraining Approach". [2] [8] It describes the limitations that their team found with BERT related to undertraining. RoBERTa's modifications are:

- training the model longer, with bigger batches, over more data
- emoving the next sentence prediction objective
- training on longer se- quences
- dynamically changing the mask- ing pattern applied to the training data

It had the best results at time of publishing on 4/9 of the GLUE tasks including MNLI and RTE.

An article entitled "Transformers: Retraining roberta-base using the RoBERTa pre-training procedure" [9] details the process of re-training RoBERTa on a custom dataset. For reference, it mentions that the Roberta-Base was "trained on 1024 V100 GPUs for 500K steps." It suggests using the TensorFlow transformers library retrain a RoBERTa neural network with new data, which is the process I will follow.

3 Implementation

I started by evaluating the data in the MLNI 0.9 dataset for the Kaggle competition, since that is what they used. It consists of 9796 sentence pairs and has a fairly even distribution of genres as shown in Table 2. It is comprised primarily of sentences under 200 characters with a few outliers with a very long length as shown in Figure 1.

I used the PyTorch framework to evaluate the data using tensors which are trained to predict

$$y = \sin(x) from - \pi to \pi.$$

by minimizing squared Euclidean distance. Tensors are similar to NumPy arrays but can be run on either the CPU or GPU. I used the PyTorch no_grad() option to disable gradient calculation data which is otherwise tracked for later calculations to optimize the code as shown in Listing 1.

```
with torch.no_grad():
for k in range(len(test_s1)):
    # Encode a pair of sentences and make a
    prediction
    tokens = roberta.encode(test_s1[k], test_s2[k
    ])
    prediction = roberta.predict('mnli', tokens).
    argmax().item()
```

Listing 1: Code to perform predictions using RoBERTa and PyTorch.

I also ran the MNLI-pretrained RoBERTa model against a train-test split of a 10,000 row subset of the SNLI train data. It came back with an F1 score of 0.8637 which is quite good for being trained on a different dataset. The most misclassifications were entailments predicted as neutral (490) and contadictions predicted as entailments (396). This reflects the robustness of the MNLI-trained RoBERTa model and suggests it is the state-of-the-art evaluator for inference.

Another ongoing competition is hosted by SuperGLUE for which entailment is one competition topic of ten total. The data is defined in a similar fashion with headings titled "premise" and "hypothesis"

[8] https://arxiv.org/pdf/1907.11692.pdf

[9] https://towardsdatascience.com/transformers-retraining-roberta-base-using-the-roberta-mlm-procedure-7422160d5764

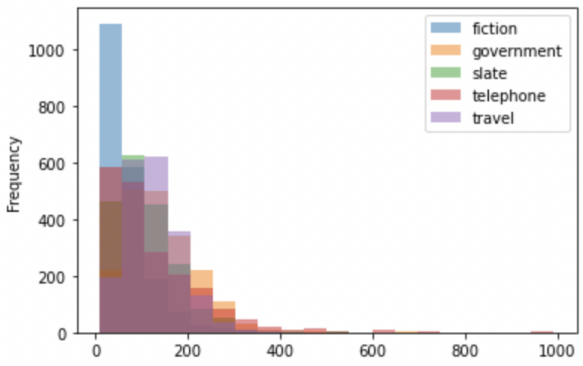

Figure 1: Distribution of sentence lengths by genre in the Kaggle competition dataset. [11]

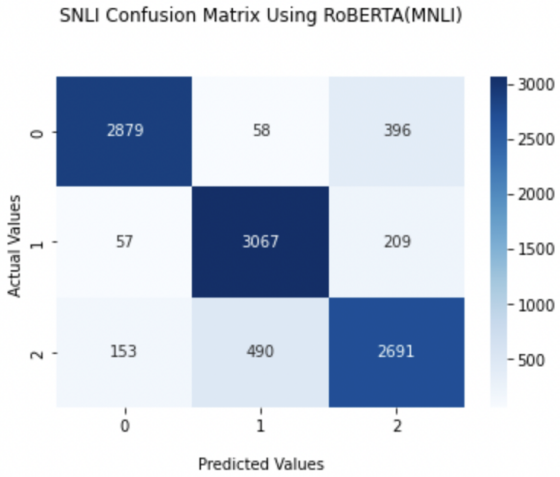

Figure 2: Results of running RoBERTa pre-trained for MLNI on the SNLI dataset.

Genre	Count
Fiction	1978
Travel	1964
Telephone	1955
Government	1953
Slate	1946

Table 2:

Genre counts for Kaggle dataset.

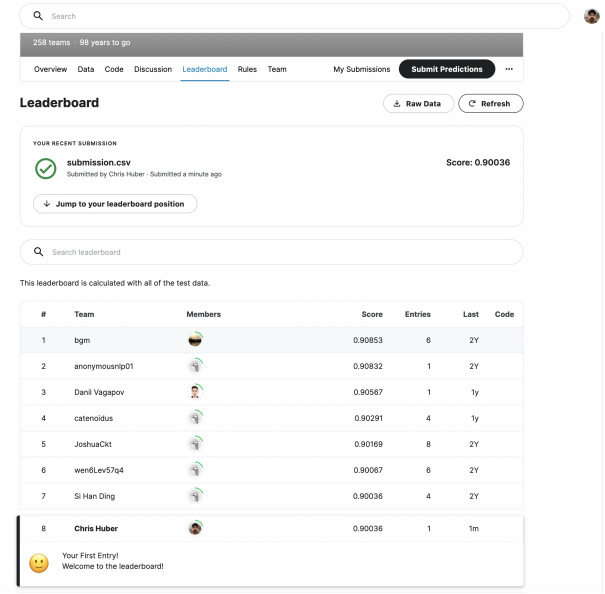

Figure 3: Results from first run of RoBERTa against the matched MultiNLI 0.9 dataset.

and a corresponding index. However, the competition only accepts two labels for entailment and non-entailment whereas the Kaggle competition required labels for contradiction, neutral, or entailment. To further illuminate the baseline performance for RoBERTa, I tried to submit its analysis of the dataset to the competition, but found that all other parts of the competition require files as well. I have sent an email to their admininstrators to see if I can get my result scored just for RTE.

I will be developing my own model based on only the supplied training data in all datasets. This involves using TensorFlow which can be very slow to process, so I will be exploring running it using CUDA to speed up the model training.

4 Evaluation

I configured and ran the RoBERTa algorithm against the MLNI corpus to establish a baseline. It performed quite well, resulting in a score of .90006 for the matched set and 0.89915 on the mismatched set which ranked as ties for 7th and 5th in the respective Kaggle competitions. The version of the RoBERTa model that I used was pre-trained on the MLNI dataset which is available as an open-source download.

I am currently investigating training the model just on the supplied train data to see how it affects the score and if I can improve upon the pretrained version.

Acknowledgments

The author wishes to thank Professor Sam Bowman from NYU who curates the SNLI and MNLI corpus .

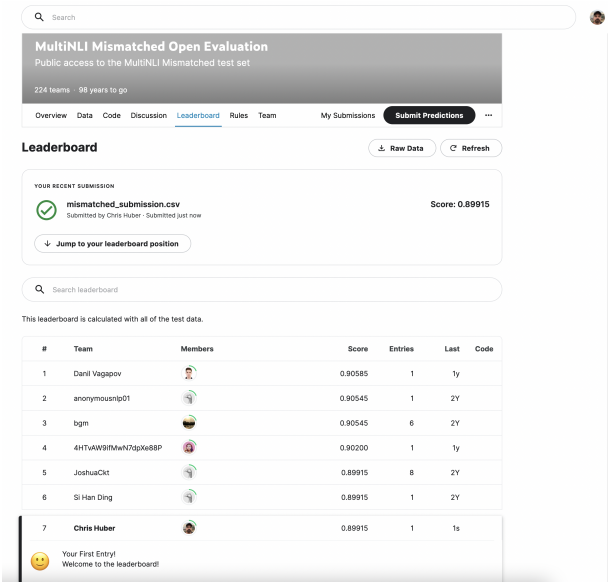

Figure 4: Results from first run of RoBERTa against the mismatched MultiNLI 0.9 dataset.

References

[1] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[3] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A fast, extensible toolkit for sequence modeling. *CoRR*, abs/1904.01038, 2019.

[4] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *CoRR*, abs/1704.05426, 2017.