Machine Learning Engineer Nanodegree Capstone Proposal

Detecting contradiction and entailment in multilingual text using TPUs

Matthew Soh

22 July 2021

1. Domain Background

Natural language processing (NLP) is the intersection of computer science, linguistics and machine learning. The field focuses on communication between computers and humans in natural language and NLP is all about making computers understand and generate human language.

In the past few years, there have been great advancements in the research of NLP models, particularly transformers, to tackle language tasks like question answering, text extraction, sentence generation and many other complex tasks. In this Kaggle Competition, the task is to build an NLP model that can determine the relationships between sentences, which could have profound implications for fact-checking, identifying fake news, analyzing text and much more. The nature of the task is called Natural language inference (NLI), where the model must determine whether a “hypothesis” is true (entailment), false (contradiction), or undetermined (neutral) given a “premise”. The current state of the art model for NLI tasks is the RoBERTa model (Liu et al., 2019) that was able to achieve a categorization accuracy of 90.8 on the Multi-Genre Natural Language Inference (MultiNLI) corpus.

I am particularly interested in this domain of Artificial Intelligence, as a model that can successfully determine the relationship between sentences would improve the way we conduct focus group discussions for my company’s research projects. Transcripts can be more thoroughly analyzed and we would be able to leverage our NLP models to generate deeper insights.

2. Problem Statement

Taken from Kaggle Competition Page:

Our brains process the meaning of a sentence like this rather quickly.
We're able to surmise:

- Some things to be true: "You can find the right answer through the process of elimination.”
- Others that may have truth: "Ideas that are improbable are not impossible!"

- And some claims are clearly contradictory: "Things that you have ruled out as impossible are where the truth lies.

If you have two sentences, there are three ways they could be related: one could entail the other, one could contradict the other, or they could be unrelated. Natural Language Inferencing (NLI) is a popular NLP problem that involves determining how pairs of sentences (consisting of a premise and a hypothesis) are related.

Your task is to create an NLI model that assigns labels of 0, 1, or 2 (corresponding to entailment, neutral, and contradiction) to pairs of premises and hypotheses. To make things more interesting, the train and test set include text in fifteen different languages!

3. Datasets and Inputs

3.1 Dataset Description

The dataset provided was taken from the XNLI data, which is a subset of a few thousand examples from MNLI which has been translated into 15 different languages. As with MNLI, the goal is to predict textual entailment (does sentence A imply/contradict/neither sentence B) and is a classification task (given two sentences, predict one of three labels).

The fifteen different languages, including: Arabic, Bulgarian, Chinese, German, Greek, English, Spanish, French, Hindi, Russian, Swahili, Thai, Turkish, Urdu, and Vietnamese.

3.2 Dataset Example

Taken from Kaggle Competition Page on Data Description:

Let's take a look at an example of each of these cases for the following premise:

He came, he opened the door and I remember looking back and seeing the expression on his face, and I could tell that he was disappointed.

Hypothesis 1:
Just by the look on his face when he came through the door I just knew that he was let down.

We know that this is true based on the information in the premise. So, this pair is related by **entailment**.

Hypothesis 2:
He was trying not to make us feel guilty but we knew we had caused him trouble.

This very well might be true, but we can't conclude this based on the information in the premise. So, this relationship is **neutral**.

Hypothesis 3:
He was so excited and bursting with joy that he practically knocked the door off it's frame.
We know this isn't true, because it is the complete opposite of what the premise says. So, this pair is related by **contradiction**.

3.3 Training and test set

A training set and test set was provided to train the language model.

The training set contains 12,120 entries, with the columns (i) id, (ii) premise, (iii) hypothesis, (iv) lang_abv, (v) language and (vi) label. There are 8,209 unique premises, 12,119 unique hypotheses and 3 unique labels.

The test set contains 5,195 entries with 4,336 unique premises and 5,195 unique hypotheses. The test set will be used for the competition submission.

4 Solution Statement

Given the size of the dataset, it would be unwise to train an NLP model from scratch. A proposed solution would be to fine-tune different pre-trained transformer models that have already been trained on terabytes of textual data.

I will test out the effectiveness of different encoder models built using the transformer architecture, such as BERT, DistilBERT, XLM-RoBERTa, etc, and try to understand what works best for the dataset.

5 Benchmark Model

My results can be benchmarked against other submissions made on the leaderboard for this competition page. I will use the accuracy achieved from a standard pre-trained BERT model as my benchmark, and hope to surpass the accuracy achieved.

6 Evaluation Metrics

The evaluation metric used for this competition is a simple categorization accuracy, which can be calculated with the following equation:

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative}$$

7 Project Design

I will approach this problem by conducting the following steps:

7.1 Data Analysis and Cleaning

This step would comprise the analysis of the textual data, to check for spelling mistakes and cleaning of duplicate entries.

As the data contains different languages, a possible preprocessing step would be to translate the text back to English, and perform the relevant data cleaning steps, e.g. removal of stop words, stemming of words, removal of punctuation.

As I will be using the transformer model for my analysis, the vectorisation of the sentences will be done by the model tokenizers.

7.2 Iteration of data augmentation and model building

I would go through a few steps of trial and error, to test out what works best with the dataset. This could entail:

- Data augmentation on the textual data (e.g. translation of text back to english).
- Creation of larger training dataset by translating the textual data into multiple languages.
- Supplement the current dataset with external datasets of the same task (e.g. MultiNLI corpus)
- Testing out the effectiveness of different models
- Tweaking of model hyperparameters

8 Resources

Kaggle competition page: https://www.kaggle.com/c/contradictory-my-dear-watson