Assignment: QA Baseline

11-411/11-611 NLP Teaching Staff

Due: January 31, 2023

1 Introduction

In this assignment, you will assemble, based on an existing PRETRAINED MODEL, a simple question answering system.

Hugging Face offers a variety of NLP tools and models which are pretrained on a massive amount of data and can be fine-tuned for specific tasks. They also provide an easy-to-use API for accessing these models, as well as a library of pre-built models for common NLP tasks.

A pretrained model is a collection of weights that have been trained an a particular set of (usually very general) objectives. That is, they have been trained to—as part of a neural net—do a very general task. It is then possible to TRANSFER the very general "knowledge" that the model has gained of language to more specific tasks.

The model we will be using is RoBERTa, a variant of the famous BERT model that we have already mentioned in class. RoBERTa is a language model, meaning that it is a model of the probability of sequences of words (or, equivalently, the word that is most likely to occur next given a history of words).

A pretrained model can be finetuned. This means that you continue to train the model, but on a more specific task. Often, a simple neural network is also added to the pretrained model to facilitate the more specific task. In this case, we use a version of RoBERTa that has been finetuned to answer SQuAD questions.

Your task will be to use this pretrained model to build a basic question answering system that will serve as the BASELINE for the question answering part of your project. A baseline is a system that acts as a basis for comparison. It helps you determine whether your method is an improvement over existing methods. Your QA system must be better than this QA system in some way:

- It must perform better according to some metric on the same test set (give the same training set) OR
- It must handle an important subset of questions that the baseline system does not OR
- It must be more intepretable than the baseline (that is, it is easier for a human to understand why your system makes the decisions that it does) OR
- It must be more efficient, in space or time, than the baseline

2 Learning Objectives

At the end of this component you should be able to do the following:

1. Use Hugging Face APIs to load the model for various NLP tasks.
2. Understand the basic concepts in designing an inference pipeline for QA (question answering).
3. Establish a baseline for the QA component of the project.

3 Task: Programming

Given a context paragraph and a question based on it, the task is to extract the answer from the context.

You are required to do the following things:

1. Download a pretrained model from Hugging Face Model Hub.
2. Complete the pre-processing and post-processing steps in the model pipeline.
3. Use the model to generate predictions using SQuAD 2.0 and blind dataset.

Note - We have provided the boilerplate code in the notebook with the necessary explanations.

You will complete this task by completing the TODO items in the provided notebook.
Recommendations -

1. Understand how to use Google Colab : Click here
   Credit : Bhiksha Raj (Language Technologies Institute)
2. Hugging face tutorials : Click here

4 Deliverables

We will be using Gradescope for submission of this homework.

1. Completed python notebook
2. blind_test_predictions.json file generated by executing the model on SQUAD2.0 dataset
3. squad_results.json file generated by executing the model on blind dataset

These files should be submitted as a flat ZIP archive (not in a zipped directory/folder).
In order to receive full credit in this assignment, your model should match/exceed the provided benchmarks for F1 score and exact match.

1. SQuAD2.0 Dataset — F1: 80.976, Exact Match: 77.954
2. Blind Dataset — F1: 22.440, Exact Match: 15.933

All deliverables are due by 11:59pm EST on January 31, 2023.