Ensemble Leaning for Stanford Question Answering Challenge

Yuzhou Liu
yuzliu@microsoft.com

Abstract

In this paper we explored techniques used in machine comprehension and question answering using the Stanford Question Answering Dataset (SQuAD). The SQuAD challenge defines a task where a context passage and a query are given and an answer to the query needs to be identified in the context passage.

Typically, this kind of task is approached using deep recurrent neural network such as Gated Recurrent Unit (GRU), Long-Short Term Memory (LSTM) and Attention mechanism. In this paper, we explored several recently published successful neural network architecture and attention mechanism such as BiDAF, Dynamic Coattention and Self Attention. In the end, this study was able to produce a F1 score of 76.7 and an EM score of 67.1 on the development dataset and a 77.65 F1, 68.1 EM score on the test dataset.

1   Introduction

Machine Comprehension (MC) is a task where an AI agent is given a query about a paragraph of context and is required to predict the answer. Such problems have gained popularity in recent years due to it's vast range of applications. The recent advancement of deep learning techniques has shown good progress in many machine learning areas including machine comprehension and question answering. To solve this problem, the machine learning system is required to understand both the context passage and the query and in most cases the connections between them to be able to accurately predict the location of the answer phrase in the context passage.

The Stanford Question Answering Dataset (SQuAD)[1] is a dataset of 100,000+ question and context pairs for machine comprehension on text. The training and development dataset are publicly available whereas the test dataset is kept secret. There is a public leaderboard recording various model's F1 and EM scores. At the time this report is written, there are many state-of-the-art models producing predictions that are close to human level performance in terms of F1 and EM scores.

Since solving this task involves the understanding and modeling of relationship between context and query, attention mechanisms are usually used in the model and in this paper, I mainly investigate several recently published attention mechanisms and some other techniques.

Specifically, we implemented and evaluated Dynamic Coattention[2], BiDAF Attention[3], Self-Attention[4] and also experimented with a new attention mechanism where the output of Coattention is fed to a Self-Matching Attention network introduced in the RNET[4] paper.