Recurrence, Transformers, and Beam Search - Oh My!

Stanford CS224N SQuAD IID Final Project

Callan Hoskins
Department of Computer Science
Stanford University
chosk@stanford.edu

Griffin Young
Symbolic Systems
Stanford University
gcyoung@stanford.edu

Abstract

Question answering on the IID SQUAD 2.0 dataset is a proving ground for natural language processing systems. In this project, we explore recurrent and transformer-based architectures for SQuAD 2.0. We implement several improvements on the baseline BiDAF and the canonical transformer QANet. Our best model, BiDAF with character embeddings and beam search output, scores **F1 62.291** and **EM 59.493**. Finally, we suggest further directions for research in self-attention and modeling/predicting NA answers.

1 Key Information

- Mentor: John Hewitt
- Shared late days: 3
- Note: We initially planned a custom project but pivoted to this default after our project proposal.

2 Introduction

Question answering (QA) in NLP is the task of answering questions based on a context passage. Given a textual passage containing the answer to the question, a QA system aims identify and return the answer. The complexity of the task varies from reading comprehension (in which a context passage is provided) to open-domain question answering (in which no context is provided). Question answering is one of the most important areas of NLP and the one that most directly impacts the lives of consumers in their interactions with Siri, Alexa, or Google Home.

In this report, we build a high-performance neural network for answering questions on the Stanford Question Answering Dataset (SQuAD)[1]. SQuAD is the canonical dataset in natural language processing for reading comprehension-style question answering, consisting of over 100,000 examples of question, context passage, and answer pairs collected from Amazon Mechanical Turk. SQuAD 2.0, released in 2018, increases the difficulty of the SQuAD task by introducing unanswerable questions where the answer isn't in the passage[2].

The best approach to NLP problems (QA included) is constantly shifting. It was once recurrence-based models, but these have now been transplanted by transformer-style architectures. In this work, we explore the performance of an improved recurrence-based model and a classic transformer on the SQuAD 2.0 dataset.

3 Related Work

The primary inspiration for our recurrent model comes from Bidirectional Attention Flow for Machine Comprehension (BiDAF) [3]. Like most SQuAD systems, BiDAF utilizes an embedding layer,