CS 224n : Project Proposal

Abhishek Goswami
Microsoft
Redmond, WA 98052
agoswami@microsoft.com

Abstract

This document describes a proposal for the final project of Course CS224N, Winter 2019.

1 Introduction

For the final project of CS224n, we chose to do the default project. The author is a SCPD student in a single person team. There are no external collaborators. We are looking forward to a mentor being assigned to us, since we have no particular mentor. We are are not sharing this project with any other class.

2 Paper Summary

In this section we review the paper **A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task** by Chen *et al* [1] et al.

In this paper, the authors look into the task of reading comprehension (RC). Developing AI systems for reading comprehension is a complex task. It involves interpretation of the text and also making complex inferences on it.

2.1 Problem Statement

The authors address the following problem : We are given a passage p and a question q. From q select one entity (a) and replace it with a placeholder. The goal is to infer the missing entity (answer a) from all possible entities which appear in the passage.

2.2 Dataset

For this problem, the authors leverage two data sets *CNN* and *Daily Mail*. They note that these two datasets were previously used by researchers at *DeepMind* [4] as well, and present a clever automated way of creating supervised data for RC tasks.

2.3 Objectives

The authors set out to achieve the following objectives:

1. **Understand what level of natural language understanding is needed to do well on the task above.**
   To this end, the authors do a thorough analysis of the two datasets, and do a hand-analysis of a subset of (passage, question) pairs. They provide interesting insights on the level of difficulty presented by these two datasets. The authors also go on to do a thorough diagnosis of what was learned by the trained model and the kind of errors produced by the model.

Preprint. Work in progress.

2. **Explore the performance of two NLP systems for this task.**
   For this the authors present two systems:
   (a) Entity-Centric Classifier. This is a conventional feature-based classifier.
   (b) Neural Network Classifier. This is a neural network system based on the AttentiveReader model proposed by Hermann et al [4]

2.4 Evaluation Metrics

In this paper the authors use accuracy as the evaluation metric. This seems to be reasonable choice for them – the goal (as defined in Section 2.1) was to infer the missing entity (answer a) that should be used for the placeholder. It was interesting to note that the feature-based classifier trained on boosted decision trees [7] did impressively well on both datasets.

2.5 Reason for choosing this paper

My reasons for choosing this paper are as follows:

1. In the final project for CS224N, I plan to work on a similar problem – to build a deep learning question answering systems. This paper also addresses a similar problem, and provided clear explanations of how the authors went about building an end-to-end neural network system based on the *AttentiveReader* model [4] propsed earlier in literature.
2. I feel the *AttentiveReader* model used in the paper can serve as a good baseline for the work I plan to do in the final project.
3. The neural network model described in the paper was extended to build larger end-to-end systems in later work by Chen et al [2]. In particular, the model used in the *Document Reader* submodule in [2] is an interesting extension of the neural network model used in the paper, extended to select a span of words from the given passage as an answer to the question.

3 Project Description

In this section we lay out the plan for the project.

3.1 Main goals(s) of the project

The *question answering* task can be formulated as follows : As input, we are given a paragraph and a question about that paragraph. The output is a span of words from the paragraph that answers the question correctly.

The goal of the project is to build and evaluate question answering systems. Over the last couple of years there has been a lot of research on question answering and reading comprehension tasks. The systems have grown in complexity over time.

To that end, the goals of the project are as follows:

1. Study the difference between 'simple' models (e.g. the *AttentiveReader* model [4] and its variants [1], [2]) versus more 'advanced' techniques proposed recently (e.g. ELMo [5] and BERT [3]) . We want to do a thorough evaluation of these systems both from a quantitative and qualitative perspective.
2. Explore ways to combine the best of both worlds so we can improve the state of the art in question answering tasks.
3. As a stretch goal, one of the things we want to explore is how to extend these systems to *generate* answers. We understand the problem of *generating* answers is different from the problem of *selecting* a span of words as an answer. Hence this is a stretch goal. We are not sure if this has been explored before in literature and/or what datasets might be suitable for this task. Any feedback on this idea would be highly appreciated.

3.2 NLP task(s) being addressed

The project aims to address the question answering task using the SQuAD 2.0 dataset.

3.3 Dataset

We plan to use the SQuAD 2.0 dataset for this project.

3.4 Neural methods being used

Besides the baseline models mentioned below in Section 3.5 we plan to explore several neural methods that have been shown to perform well on question answering tasks.

3.5 Baselines for evaluation

The de-facto baseline model for the default project is based on BiDAF [6] without the character level embedding layer. In particular, the de-facto code implements a BiDAF variant proposed by Yu et al [8]. Another baseline we want to try out is the *AttentiveReader* model [4] used successfully within the *Document Reader* submodule of the DrQA system [2]

3.6 Evaluation metrics

We will use two metrics: Exact Match (EM) score and F1 score as our evaluation metrics for this project.

References

[1] D. Chen, J. Bolton, and C. D. Manning. A thorough examination of the CNN/Daily Mail reading comprehension task. In *Association for Computational Linguistics (ACL)*, 2016.

[2] D. Chen, A. Fisch, J. Weston, and A. Bordes. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*, 2017.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

[5] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

[6] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*, 2016.

[7] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.

[8] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.