NLU Project - MCUltra

Manoj Kumar
mks542@nyu.edu

Christina Bogdan
ceb545@nyu.edu

Summary

For our final project, we will be working on topic #2 (Story Comprehension Tests). The goal is to develop a model that can comprehend a passage and then predict missing words from query sentences summarizing the text. We want to implement a baseline and the current state of the art in Tensorflow for the Who did What dataset, and then to try to improve these models by experimenting with the query/passage embeddings, fusion of query/passage embeddings, and techniques such as multistep reasoning.

1 Task

The aim is to create a model that will be able to comprehend a passage and then answer a multiple-choice question about the text. The input is a passage and question and the output will be a probability distribution across all the answer choices. There is a lot of recent work in deep learning that tries to solve this task [1] [2], and two main datasets that are large enough to use for deep learning. We outline these datasets and other approaches below, and then explain our plans for how we will implement this work and try to improve on it.

2 Data

2.1 CNN and Daily Mail Datasets

The CNN/Daily Mail dataset [3] is most commonly used in literature relating to this task. It is made up of 287k articles taken from the CNN and Daily Mail websites. There are some issues with the construction of this dataset make it non-ideal for Q&A research. The use of NER and coreference resolution to identify labels introduces error into the model that is outside of our control. Further, there is a lot of variance in the type of questions that appear in the dataset. Some are easily solved by baselines, while others are too ambiguous to even be answered by humans. These issues are outlined in [1]. Another issue is that questions generated from author summaries may cause overfitting resulting from syntactic similarities between the question and passage.

2.2 Who did What

The Who did What (WDW) dataset [4] is made up of over 200k fill-in-the-blank questions generated from the LDC English Gigaword newswire corpus. Unlike the CNN/Daily Mail dataset, questions are generated independently of the article that they reference. The creators of the dataset also filtered out questions that are easily solved by baselines. Models that comprise the current state-of-the-art trained on the CNN/Daily Mail data performed significantly lower on WDW. We will use this dataset to evaluate our project because its careful construction allows for more accurate benchmarks.

3 Relevant Work

The current state-of-the-art for this task is the Stanford Reader [1], which achieves 73.6%, 76.6%, and 64% accuracy on the CNN, Daily Mail, and WDW datasets, respectively. The model encodes the passage and question statements separately, and assigns weights to the contextual embeddings of each word in the passage before aggregating them into one document-query representation.

The Stanford Reader is based on the Attentive Reader of Hermann et al [2]. The structure of the two models is the same, but there are slight variations in the way that the question/passage embeddings are fused in both. This paper also proposes another model, called the Inattentive Reader, that computes a representation of the document for each word in the query. With these two models, the authors achieve accuracies of 63.8%, 69%, and 53% on the CNN,Daily Mail, and WDW data sets.

4 Baselines

We would implement the following baselines for our work.

- Feed each one-hot encoded word of the query directly into each time step of the LSTM and the predicted answer would be the argmax over the softmax of the output of the last time-step. This does not involve information from the question itself.
- The second baseline would be similar to the Stanford Reader described below except that the encoding of every token in the question which is fused with the query is a one-hot encoding and not learnt.

5 Models

We will be implementing the the Stanford Reader model, which currently has the highest accuracy for this talk across all data sets. We will not be implementing the Attentive Reader because it is more complicated but has worse accuracy. The model for the Stanford Reader is given below:

- Encodes each token of the passage using a bi-directional GRU.
- Encode the entire query using a bi-directional GRU.
- Compute the weighted average of the encoding of all the tokens in the passage. Each weight is the softmax of the bi-linear term obtained by a combination of each token embedding and the query embedding.
- Multiply the document-query embedding by another weight matrix to get the predicted answer of the document-query combination.

5.1 Improvements

Generally, we would like to see if we can adapt some approaches used in VQA to this task, because the two problems are similar. Below are some specific ideas that we would like to try to improve the Stanford Reader:

1. Embeddings - The Stanford Reader uses a bi-directional GRU to create embeddings for the questions and passages. We could try different ways of creating these, like CBOW/FastText. We also want to try learning an embedding for the answer (for example, using the technique in [5])
2. Fusion - We will try different ways to combine question/passage information other than the bilinear term and TanH outlined by Chen/Hermann. If we decide to include embeddings for each of the choices we will need to figure out how to fuse this with the question/passage embedding.
3. Multistep Reasoning - Processing the document in multiple steps [6]. This is similar to the Impatient Reader in [2].

6 Timeline

- 10/19/2016: Get Data
- 10/24/2016: Implement baseline
- 11/1/2016: Implement Stanford Reader
- 11/15/2016: Experiment with different approaches
- 12/17/2016: Write Report
- 12/19/2016: Final report Due

7 Contributions

- Manoj: Baseline, Models, Timeline
- Christina: Summary, Task, Data, Relevant Work, Improvements, Timeline

References

[1] Danqi Chen, Jason Bolton, and Christopher D. Manning. A thorough examination of the cnn/daily mail reading comprehension task. *CoRR*, abs/1606.02858, 2016.

[2] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *CoRR*, abs/1506.03340, 2015.

[3] DeepMind QA Dataset. `http://cs.nyu.edu/~kcho/DMQA/`.

[4] Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David A. McAllester. Who did what: A large-scale person-centered cloze dataset. *CoRR*, abs/1608.05457, 2016.

[5] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. *CoRR*, abs/1606.08390, 2016.

[6] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. Weakly supervised memory networks. *CoRR*, abs/1503.08895, 2015.