SQuAD

The Stanford Question Answering Dataset

CSCI 8945 Advanced Representation Learning

Instructor : Dr. Sheng Li

Sumer Singh

Aashish Yadavally

Introduction

- **S**tanford **Q**uestion **A**nswering **D**ataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles.

- The current iteration of the dataset has gone up in size from the previous one by ~140% - with the main distinction being the addition of negative examples, i.e, questions which do not have an answer in the given document.

- Thus, the interesting thing about SQuAD 2.0 is that it is a large-scale dataset that forces models to understand when a question cannot be answered given the context.

Problem Statement

- To test the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph.

This problem can broadly be classified as a **N**atural **L**anguage **U**nderstanding (NLU) task!

Why NLU?
Motivation to address this problem?

- Natural Language Understanding is a very important area of research under Natural Language Processing (NLP), which in turn, is an important AI area of research.
- Question-Answering is an essential NLP task.
- More importantly, using limited context to answer questions mimics real life behavior.

Literature Review

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova

BERT is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT representations can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference,without substantial task-specific architecture modifications

SG-Net: Syntax-Guided Machine Reading Comprehension - Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, Rui Wang

SG-NET uses syntax to guide the text modeling of both passages and questions by incorporating explicit syntactic constraints into attention mechanism for better linguistically motivated word representations.

Literature Review

XLNet: Generalized Autoregressive Pretraining for Language Understanding - Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le

XLNet, is a generalized autoregressive pretraining method that enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and overcomes the limitations of BERT thanks to its autoregressive formulation. Furthermore, XLNet integrates ideas from Transformer-XL, the state-of-the-art autoregressive model, into pretraining.

Proposed Solution

1) Focus on one of the State of the Art models (SG-NET, BERT, XLNet, etc).
2) Fine tune performance.
3) Time permitting, experiment with different SOTA models / create ensembles.
4) Tasks:
    a) Beat baseline score of 63.4% EM and 66.3% F1
    b) Attempt to beat human performance of 86.831% EM and 89.452% F1
    c) Attempt to beat current top score of 89.731% EM and 92.215% F1.

Additional Task

- Test our model on Reading Comprehensions (RCs) present in the GRE exam.
- This task is out of curiosity. No baseline or previous work exists.
- We aim to test our model on at least 20 RCs.

Technical Plan

- Python3 will be used as the primary language.
- Pytorch / Keras will be used as the deep learning package.
- Google Cloud servers for GPUs / TPUs.

Evaluation Plan

- Dataset - SQuAD
  - 100,000+ question-answer pairs on 500+ articles.
  - 50,000+ unanswerable questions written adversarially by crowdworkers to look similar to answerable ones.

- Dataset - GRE RCs

  Create our own dataset of at least 20 Reading Comprehensions which appear in the GRE exam and test our model(s) on them.

Evaluation Plan

- Evaluation Metrics
  - The common metrics used for this problem are: Exact Match (EM), F1 Score.
  - Human performance on these metrics are 86.3% and 89.0 respectively (on SQuAD 2.0).

Tentative Project Timeline

Milestone	Proposed Deadline
Literature Review	October 12
Achieving Baseline Results	October 25
Project Status Update	**October 29**
Proposed Solution	November 5
Improving Results	November 15
Final Presentation	**November 19**
Final Project Report	**December 4**

THANK YOU !

Any additional suggestions from any of you who’s worked on this problem before, which we could explore, so as to culminate this project successfully?

References

[1] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know What You Don't Know: Unanswerable Questions for SQuAD." *arXiv preprint arXiv:1806.03822* (2018).

[2] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova

[3] SG-Net: Syntax-Guided Machine Reading Comprehension - Zhuosheng Zhang, Yuwei Wu, Junru Zhou, Sufeng Duan, Hai Zhao, Rui Wang

[4] XLNet: Generalized Autoregressive Pretraining for Language Understanding - Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le