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.

Related Work

RACE: Large-scale ReAding Comprehension Dataset From Examinations

RACE, a is a dataset for benchmark evaluation of methods in the reading comprehension task. Collected from the English exams for middle and high school Chinese students in the age range between 12 to 18, RACE consists of near 28,000 passages and near 100,000 questions generated by human experts (English instructors), and covers a variety of topics which are carefully designed for evaluating the students' ability in understanding and reasoning.

How does RACE differ from Squad?
Answer’s in RACE are in the form of multiple choices. They are not a span of the original data.

Related Work

DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

DistilBERT leverages knowledge distillation during the pre-training phase and reduces the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster.

Knowledge distillation is a compression technique in which a compact model - the student - is trained to reproduce the behaviour of a larger model - the teacher - or an ensemble of models.

Recap of ALBERT

ALBERT: A LITE BERT FOR SELF - SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS

- Albert is a followup to the highly effective pre-training technique BERT. It mainly focuses on reducing model size. Albert introduces two parameter reduction techniques.
- The first technique is factorized embedding parameterization. The authors decompose the large vocabulary embedding matrix into two smaller ones. This separation makes it easier to increase the hidden layers, without adding a significant cost.
- The second technique is cross-layer parameter sharing. The technique prevents the parameters from growing as the depth of the networks increases.
- These techniques have the added benefit of having a regularization effect, thus reducing overfitting.
- An ALBERT configuration similar to BERT-large has 18x fewer parameters and can be trained about 1.7x faster.

Proposed Approach

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.

Experiments

Model	# Parameters	Performance (EM / F1)	Speedup
Bert-base (Baseline)	110M	69.61 / 73.02	1.00x (3.5 hours)
Distilbert-base	66M	57.08 / 60.69	1.3x
Distilbert-base-distilled-squad	66M	57.52 / 61.23	1.3x
XLNet-base	112M	36.07 / 41.29	0.7x
Albert-base	12M	76.1 / 79.0	1.4x
Albert-L	18M	78.4 / 81.3	0.3x
Albert-XL-v1	60M	82.1 / 85.1	0.13x (BERT-base) 2.4x (BERT-xlarge)
Albert-XL-v2	60M	83.05 / 86.3	0.13x (BERT-base) 2.4x (BERT-xlarge)

Hyperparameter Tuning

Considering the time taken for training the models, we decided to perform hyperparameter tuning on only the base-versions of the models. These hyperpameters include:

- doc_stride (when splitting up a long document into chunks, how much stride to take between chunks)
- max_sequence_length (the maximum total input sequence length after WordPiece tokenization)
- learning rate
- dropout rate

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.
    - As Squad answers are supposed to be a span of paragraph, this task is incompatible with Squad.
    - Can use RACE dataset to train the models for fine-tuning.

Future Direction

- More training. Currently our training is limited to 2 epochs due to time constraints.
- Use a bigger model and Ensembles.

Technology Used

- Python3 used as the primary language.
- Pytorch and Keras used as the deep learning package.
- Google Cloud servers / UGA AI LAB for GPUs

Challenges

- Lots of State of the Art options - there is always something better. Need to make an informed decision before commiting.
- Training models extremely time consuming.
- Out of memory issues.

Discussions

- Are bigger models always better in NLU tasks?
- How to make pretrained vectors accessible to mobile devices?
- Why is speedup not proportional to reduction is number of parameters?

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 !

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] DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter - Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF

[3] RACE: Large-scale ReAding Comprehension Dataset From Examinations - Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, Eduard Hovy