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!

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**

Literature Review

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

BERT replaces random words in the input sentence with the special [MASK] token and attempts to predict what the original token was. In addition to this, BERT uses the powerful Transformer architecture to incorporate information from the entire input sentence.

Literature Review

RoBERTa: A Robustly Optimized BERT Pretraining Approach
(https://arxiv.org/abs/1907.11692)

- The authors replicate BERT and measure the impact of important hyperparameters.
- They discover that BERT was significantly undertrained.
- The authors modify BERT in the following ways :
  - training the model longer, with bigger batches, over more data;
    - BERT trained on BooksCorpus and English Wikipedia. RoBERTa adds 4 more corpuses.
  - Removing the next sentence prediction objective;
    - Authors find that removing the NSP loss matches or slightly improves downstream task performance.
  - training on longer sequences.
    - ROBERTa is not trained on reduced sequence length for the first 90% of updates. Only full length is used.
  - dynamically changing the masking pattern applied to the training data.
    - To avoid using the same mask each epoch, training data is duplicated 10 times and each time masked differently.

Literature Review

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.

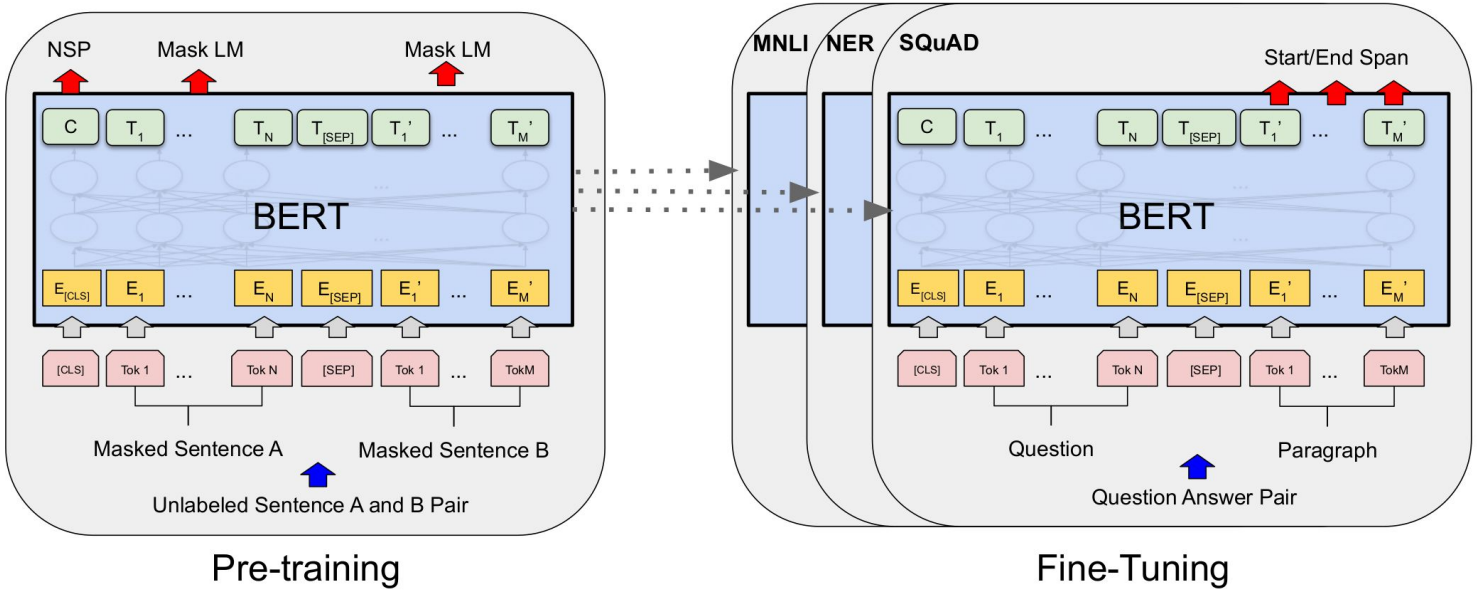
Literature Review

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

Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
	xlarge	1270M	24	2048	2048	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

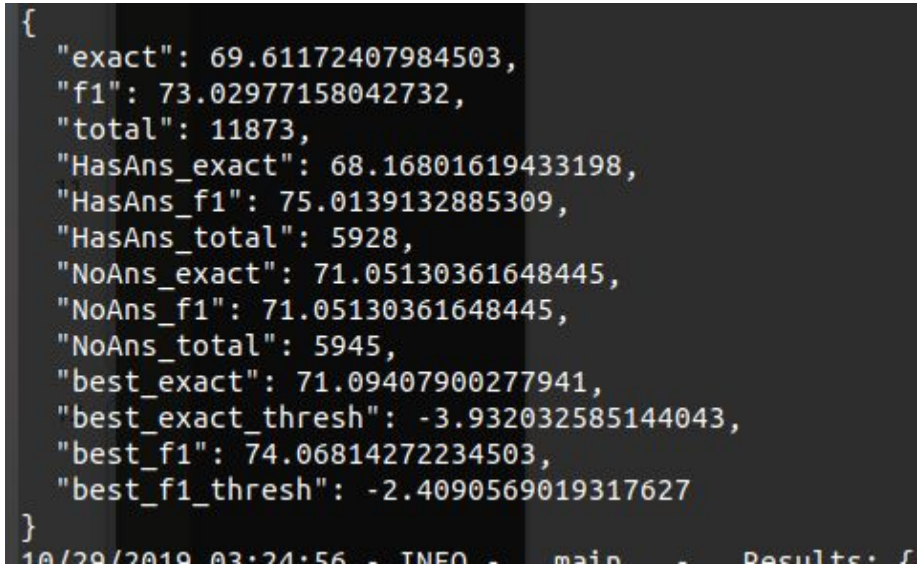
Table 2: The configurations of the main BERT and ALBERT models analyzed in this paper.

Baseline



Source - BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Baseline Performance



Leaderboard

SQuAD2.0 tests 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.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	**89.731**	**92.215**
2 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
2 Sep 16, 2019	ALBERT (single model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	88.107	90.902
2 Jul 26, 2019	UPM (ensemble) *Anonymous*	88.231	90.713
3 Aug 04, 2019	XLNet + SG-Net Verifier (ensemble) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	88.174	90.702
4 Aug 04, 2019	XLNet + SG-Net Verifier++ (single model) *Shanghai Jiao Tong University & CloudWalk* https://arxiv.org/abs/1908.05147	87.238	90.071
5 Jul 26, 2019	UPM (single model) *Anonymous*	87.193	89.934

Issues faced

- Out of memory with large batch sizes.
- Slow training with small batch sizes.
- Albert seems promising because of better memory requirements.

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.

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

[5] Keitakurita - Paper Dissected: “XLNet: Generalized Autoregressive Pretraining for Language Understanding” Explained