Question Answering System for SQuAD 2.0

Stanford University

CS 230 Final Project | Winter 2019

Weiquan Mao {mwq@stanford.edu}

Introduction

We produce a System to answer questions correctly given paragraph context from SQuAD 2.0. The result would be the span of text or N/A if there is no answer in the paragraph. We use BiDAF as baseline, BERT-based architecture as the core, L1 regularization and other architecture changes on BERT. Ensembling method is also applied for improvement, which combines multiple models into a more robust Question Answering system.

Data

We use SQuAD 2.0 as the reading comprehension data set. Every answerable SQuAD question has three answers provided.

Dataset has been split into:

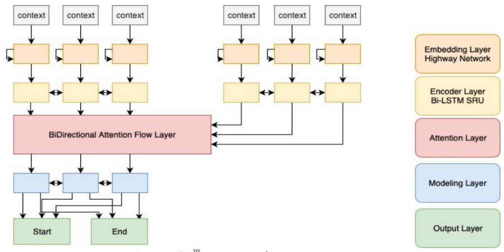
- Train Set: 129941 examples
- Dev Set: 6078 examples
- Test Set: 5291 examples

Experiment

ID	Experiment Name							Dev F1	Dev EM
lightgray	BERT Models	Pre-trained model	Number of Epoches	Learning Rate	Batch Size	Max Sequence length	Note		
1	out_maxseqlen245_bs12_lr3e-5_ep6l1	BERT-Base, Cased	6	$3e^{-5}$	12	245	$\lambda = 1e^{-4}$	77.206	73.478
2	out_maxseqlen245_bs12_lr3e-5_ep4	BERT-Base, Cased	4	$3e^{-5}$	12	245		77.166	73.643
3	out_maxseqlen140_bs24_lr3e-5_ep4_l11e-2	BERT-Base, Cased	4	$3e^{-5}$	24	140	$\lambda = 1e^{-2}$	76.76	73.955
4	out_maxseqlen140_bs24_lr3e-5_ep4_l11e-3	BERT-Base, Cased	4	$3e^{-5}$	24	140	$\lambda = 1e^{-3}$	76.666	73.824
5	out_maxseqlen245_bs12_lr3e-5_ep6	BERT-Base, Cased	6	$3e^{-5}$	12	245		75.925	72.343
6	out_maxseqlen140_bs24_lr3e-5_ep4_l11e-4_uncased	BERT-Base, Uncased	4	$3e^{-5}$	24	140	$\lambda = 1e^{-4}$	75.899	72.902
7	out_maxseqlen140_bs24_lr3e-5_ep4_l11e-4	BERT-Base, Cased	4	$3e^{-5}$	24	140	$\lambda = 1e^{-4}$	75.705	73.001
8	out_maxseqlen245_bs12_lr3e-5_ep4_l11e-4	BERT-Base, Cased	4	$3e^{-5}$	12	245	$\lambda = 1e^{-4}$	75.671	72.606
9	out_maxseqlen245_bs12_lr3e-5_ep5_l1+	BERT-Base, Cased	5	$3e^{-5}$	12	245	$\lambda = 1e^{-4}$, add one layer	75.354	71.685
10	out_maxseqlen245_bs12_lr3e-5_ep4_uncased	BERT-Base, Cased_uncased	4	$3e^{-5}$	12	245		75.071	71.372
11	out_maxseqlen140_bs24_lr3e-5_epoch4	BERT-Base, Cased	4	$3e^{-5}$	24	140		74.679	71.915
12	out_maxseqlen290_bs10_lr3e-5_epoch4	BERT-Base, Cased	4	$3e^{-5}$	10	290		74.633	71.372
13	out_maxseqlen245_bs12_lr5e-5_ep4	BERT-Base, Cased	4	$5e^{-5}$	12	245		74.546	71.092
14	out_maxseqlen200_bs12_lr3e-5_ep4	BERT-Base, Cased	4	$3e^{-5}$	12	200		74.356	71.241
15	out_maxseqlen245_bs12_lr1e-5_ep5	BERT-Base, Cased	5	$1e^{-5}$	12	245		73.885	70.829
16	out_maxseqlen400_bs6_lr3e-5_epoch4	BERT-Base, Cased	4	$3e^{-5}$	6	425		73.725	70.5
17	out_maxseqlen128_bs12_lr3e-5_ep4	BERT-Base, Cased	4	$3e^{-5}$	12	128		73.638	71.142
18	out_maxseqlen245_bs12_lr3e-5_ep4+	BERT-Base, Cased	4	$3e^{-5}$	12	245	add one layer	73.292	69.908
19	out_maxseqlen90_bs48_lr3e-5_ep4	BERT-Base, Cased	4	$3e^{-5}$	48	90		72.954	70.813
lightgray	BIDAF Models	Word Embeddings	Number of Epoches	Learning Rate	Encoder	Note			
20	baseline_sru	GloVe	30	0.5	SRU			64.08	
21	baseline	GloVe	30	0.5	LSTM	Baseline		61.508	57.99
lightgray	Ensembling Models								
21	Guidede Random Search for Weighted Average							79.944	77.081

Models

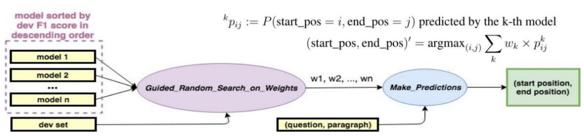
1. **Baseline: BiDAF**
2. **Baseline: SRU + BiDAF**



3. **L1 Regularization:** $J(w,b) = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^i, y^i) + \frac{\lambda}{2m}\|w\|_1$
4. **Go "deeper":** Add one more fully-connected layer to the output of BERT
5. **Ensembling**: Guided Random Search for Weighted Average Ensembling

$${}^{k}p_{ij} := P(\text{start\_pos} = i, \text{end\_pos} = j) \text{ predicted by the k-th model}$$

$$(\text{start\_pos}, \text{end\_pos})' = \text{argmax}_{(i,j)} \sum_k w_k \times p_{ij}^k$$



Results

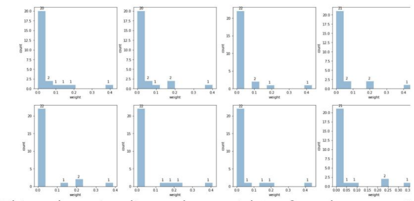
1. L1 Regularization: Train BERT with L1 regularization on weights of output classification and varies the coefficient. Increasing regularization strength helps improve the F1 and EM score.

2. Ensembling:

Ensembling Model	Dev F1	Dev EM
Guidede Random Search for Weighted Average	79.944	77.081

Regularization Coefficient	Dev F1	Dev EM
0	74.679	71.915
1e-4	75.705	73.001
1e-3	76.666	73.824
1e-2	76.76	73.955

Analysis



This plot visualizes the weights for the top 8 Ensembling models in one run(100 iters) of Guided Random Search of weights by plotting the distribution in histograms. Most of the models only bears a weight in the order of 0.001.

Conclusions

After training 28 BiDAF-based, BERT-based models, and ensemble them with two algorithms, we push test F1 score to 78.841 and Test EM to 76.010.

Future Work

For future work, we would combine the BERT and BiDAF together, which means that we replace BiDAF's GloVe word embedding with BERT last layer's output as as contextual word embedding. Hopefully we can improve our performance more with this idea.

References

Devlin, Jacob, et al. "Bert: Pre-training of deep bi-directional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).