Megatron: Improving Non-PCE Question-Answering Systems Using Transformer, Character-Embedding and Residual Bi-Directional Attention Flow (Res-BiDAF)

Matt Linker, Zhilin Jerry Jiang, Zheng Nie

Department of Computer Science, Stanford University

Problem

In this project, we explore building and improving a neural model for solving the question-answering tasks defined by SQuAD 2.0[1], in order to contribute to this popular NLP research topic, and to better understand techniques of applying neural models to NLP tasks in general.

Data/Task

Task & Dataset: The Stanford Question Answering Dataset (SQuAD)[1]

- Input:
  - Context (a text passage)
  - Question (reading comprehension)

  Output:
  - Answer to the question (plain text)
- SQuAD versions
  - SQuAD 1.1:
    All questions answerable
  - SQuAD 2.0:
    Some (~50%) questions have no answer, where the model should output N/A (as an empty string)
- Evaluation Metrics:
  - Exact-Match (EM) Score

$$\mathrm{EM} = \frac{\text{exactly matching answers}}{\text{total evaluated questions}} \times 100$$

  - F1 Score

$$P = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

$$R = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

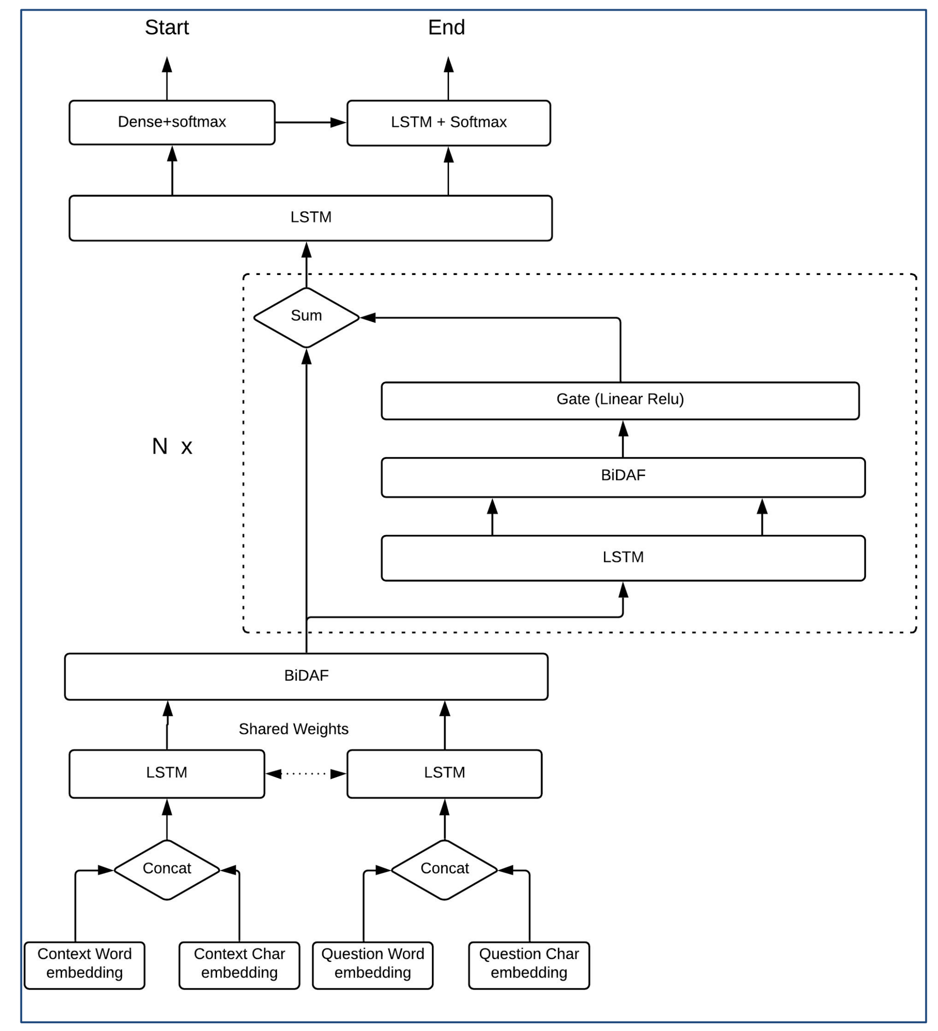
$$\mathrm{F1} = 2\frac{PR}{P+R}$$

  - AvNA (Answer vs. No-Answer)
    Accuracy only considering whether the model predicts some answer or predicts no-answer (N/A)

Approach

Models with major performance improvements:

- Transformer[2] (Single / Double)
- Char-level Embedding
- Res-BiDAF (Single Block)
- Res-BiDAF (Double BLock)



Self-Attended Residual Bi-Directional Attention Flow (Res-BiDAF)

- Inspired by ResNet[3] and Highway[4]
- Char-level embedding concat with word embeddings
- **Gated Residual Blocks** (LSTM + BiDAF + ReLU Gate, summed with pass-through)

Post-prediction TF-IDF filtering

- TF-IDF: Relevance metric, higher score for rare word appearances
- We length-normalized TF-IDF to measure prediction relevance
- Manually mark "no answer" if a prediction has low TF-IDF score (i.e. irrelevant / not interesting)

Analysis

Q&A Example:

Question: Economy, Energy and Tourism is one of the what?

Context: Subject Committees are established at the beginning of each parliamentary session, and again the members on each committee reflect the balance of parties across Parliament. Typically each committee corresponds with one (or more) of the departments (or ministries) of the Scottish Government. The current Subject Committees in the fourth Session are: Economy, Energy and Tourism; Education and Culture; Health and Sport; Justice; Local Government and Regeneration; Rural Affairs, Climate Change and Environment; Welfare Reform; and Infrastructure and Capital Investment.

Correct Answer:	current Subject Committees
Prediction (Baseline):	N/A
Prediction (Single Transformer):	The current Subject Committees
Prediction (Char-Embedding):	current Subject Committees
Prediction (Res-BiDAF, single block):	Subject Committees
Prediction (Res-BiDAF, 2 blocks):	current Subject Committees in the fourth Session

Reference

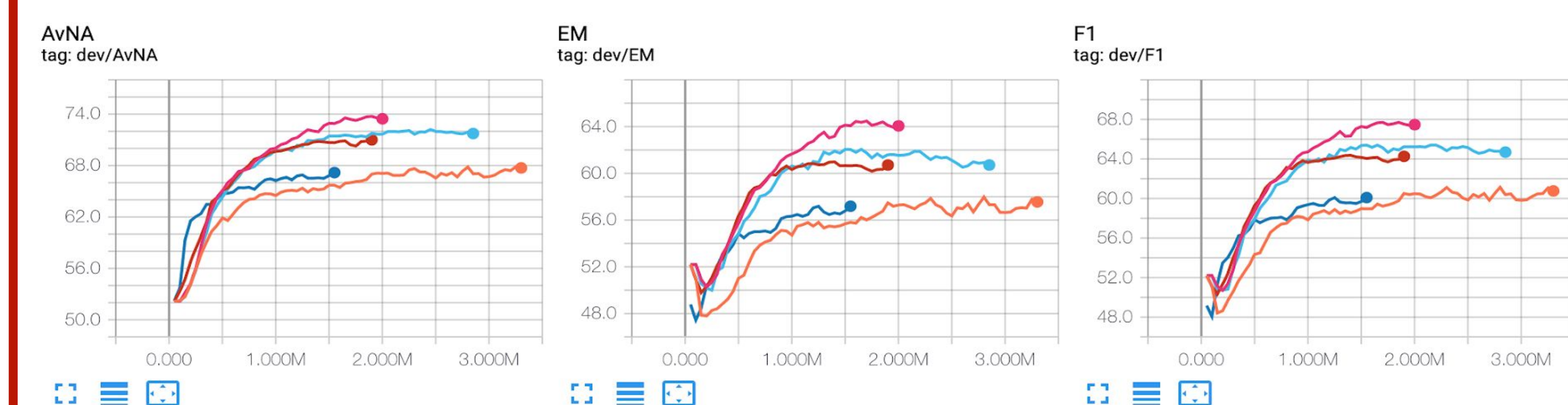
[1] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. arXiv:1806.03822, 2018.
[2] Ashish Vaswani, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
[4] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. arXiv preprint arXiv:1505.00387, 2015.

Results

Alternative models brought significant performance improvements

Model	F1 Score	EM Score
Baseline	58.98	55.81
Single Transformer	59.70	57.18
Char-Embedding	64.36	60.98
Res-BiDAF (single block)	65.40	62.06
Res-BiDAF (2 blocks)	67.69	64.48

Dev set AvNA / EM / F1 score curves



Non-PCE Test Leaderboard submissions

- Res-BiDAF (single block) EM: 58.698 F1: 62.333
- Res-BiDAF (2 blocks) EM: 62.992 F1: 66.370

TF-IDF post-filtering failed to produce satisfactory results

Model	F1 Score	EM Score	AvNA Score
Baseline (Trained), no TF-IDF	61.13	57.97	67.84
Baseline (Trained), with TF-IDF	60.06	57.57	65.74
Char-Embedding, no TF-IDF	64.36	60.98	70.81
Char-Embedding, with TF-IDF	62.86	60.12	68.53
Res-BiDAF (2 blocks), no TF-IDF	67.69	64.48	73.40
Res-BiDAF (2 blocks), with TF-IDF	65.90	63.35	70.81

Conclusion / Future Work

- Our proposed Res-BiDAF reached significant performance improvements over baseline non-PCE model
- Still far away from catching up with PCE models (ELMo & BERT)
- Potential future work:
  - Train Res-BiDAF with higher # of Gated Residual Blocks
  - Combine Res-BiDAF with other non-PCE techniques
  - Experiment applying Res-BiDAF to PCE models