Investigating Model Combination to Improve Performance on SQuAD

by, Matt Riedman
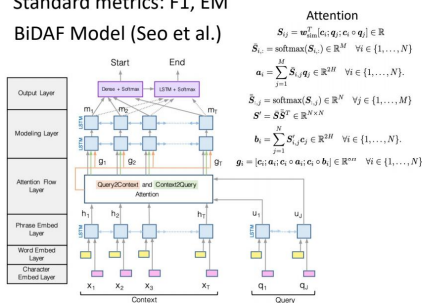
Stanford Department of Computer Science

Introduction

- Question answering is a useful tool to get info from large databases (e.g. Wikipedia, libraries)
- Present challenge is ensuring a model can efficiently identify most relevant words
- In LSTM-based models, attention layer is key to finding relevant context words

Background

- SQuAD: Widely used Q/A dataset
    - 100,000 pairs of context and queries
    - Includes unanswerable questions (mimics real life)
- Standard metrics: F1, EM
- BiDAF Model (Seo et al.)

Attention

$$S_{ij} = w_{sim}^T [c_i; q_j; c_i \circ q_j] \in \mathbb{R}$$
$$\bar{S}_{i,:} = \text{softmax}(S_{i,:}) \in \mathbb{R}^M \quad \forall i \in \{1, \dots, N\}$$
$$a_i = \sum_{j=1}^{M} \bar{S}_{i,j} q_j \in \mathbb{R}^{2H} \quad \forall i \in \{1, \dots, N\}.$$
$$\bar{\bar{S}}_{:,j} = \text{softmax}(S_{:,j}) \in \mathbb{R}^N \quad \forall j \in \{1, \dots, M\}$$
$$S' = \bar{S} \bar{\bar{S}}^T \in \mathbb{R}^{N \times N}$$
$$b_i = \sum_{j=1}^{N} S'_{i,j} c_j \in \mathbb{R}^{2H} \quad \forall i \in \{1, \dots, N\}.$$
$$g_i = [c_i; a_i; c_i \circ a_i; c_i \circ b_i] \in \mathbb{R}^{8H} \quad \forall i \in \{1, \dots, N\}$$

(Seo et al., 2017): Bidirectional Attention Flow for Machine Comprehension

Methods

- Train two similar established models with different attention layers
    - BiDAF, Dynamic Coattention
- Try to leverage likelihoods of start and end position from each to produce better predictions than either alone
- General approach: Train models separately, then create hybrid model using originals up to the softmax layer
- Provides framework for leveraging strengths of different models

Experiments

- First: assume softmax output represents actual probability, use normalized elementwise product (i.e., take product of probabilities from each)
- Second: run through training set to see "actual" probabilities given softmax values, use those in first method
- Third: concatenate hidden layer prior to softmax and use nonlinearity then softmax
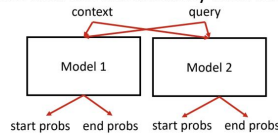
Results and Analysis

- BiDAF EM: 57.60, F1: 67.95
- Coattention EM: 63.20, F1: 73.56
- Strangely, Experiments 1 and 2 both provide lower EM, F1 metrics than either of the original models
- I suspect in cases where one model is wrong, softmax probability drags down the other's predictions
- Experiment 3 results pending

Conclusions

- Combining models is harder than anticipated
- Rather than producing a summative effect where models combine their predictive capability, maive approaches seem to collide

References

- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+
- questions for machine comprehension of text. CoRR, abs/1606.05250, 2016
- Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional
- attention flow for machine comprehension. arXiv preprint arXiv:1611.01603, 2016.
- Caiming Xiong, Victor Zhong, and Richard Socher. Dynamic coattention networks for question
- answering. arXiv preprint arXiv:1611.01604, 2016.