Comparing Approaches to Question-Answering on SQuAD 2.0

CS224N: Natural Language Processing with Deep Learning
Ray Iyer - rri@stanford.edu
Department of Computer Science, Stanford University

Question-Answering Task

Input: question and context (i.e., paragraph) of text
Output: A correct answer to the question, where the answer is a span (i.e., excerpt of text) from the context. In some cases, the question cannot be answered using the context.

Background

Recurrent Neural Networks (RNNs)
Traditionally, the most successful models for QA utilized a recurrent neural network to encode sequential input for downstream processing.

Transformer
The Transformer has driven state-of-the-art improvements on adjacent tasks of language modeling, machine translation, etc.; here, we explore adapting its techniques of position encoding, feed-forward layers, and masked multi-head attention to the QA task.

Convolutional Neural Networks (CNNs)
CNNs are commonly used for visual analysis; among a sequence of words, they capture local textual structure.

Self-Attention
Self-attention learns the global dependencies between word pairs.

Data

Stanford Question-Answering Dataset 2.0
- Size: 129,941 train, 6,078 dev, 5,915 test examples
- Example: (context, question, answer) triple
- Three answers provided per example from different human labelers, to account for variance of reading comprehension and potential for multiple correct answers
- Train includes over 40,000 unanswerable questions

Neural Models

Bi-Directional Attention Flow (BiDAF)

Input Layer
Concat(Proj(LoVe Word Emb) + Conv2b(Char Emb)) -> Highway Network

Contextual Embedding Layer
Concat([Forward LSTM, Backward LSTM])
Model temporal interactions with words.

Modeling Layer
Two-layer bi-directional LSTM: capture context word interactions conditioned on query

Output Layer
\[ p = \text{score}(S, M) \]

Bi-Directional Attention Flow Layer

\[ S_q = \text{context}_q \]
\[ C_q = \text{context}_C \]

QANet

Input Layer
Concat(Proj(LoVe Word Emb) + Conv2b(Char Emb)) -> Highway Network

Embedding Encoder Layer
Positional Encoding -> 4 x Conv1d -> Self Attention -> Feed Forward w/ residuals

Stacked Model Encoder Blocks
BiLSTM -> 2 x Conv1d -> SelfAtt -> FF 7 blocks, applied 3x w/ shared weights

Output Layer
\[ p = \text{score}(S, M) \]

Experiments

<table>
<thead>
<tr>
<th>Model</th>
<th>Batch Size</th>
<th>Train Time</th>
<th>Dev EM</th>
<th>Dev F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline BiDAF</td>
<td>64</td>
<td>3h11m</td>
<td>57.049</td>
<td>60.686</td>
</tr>
<tr>
<td>BiDAF + Character Emb</td>
<td>64</td>
<td>4h49m</td>
<td>59.368</td>
<td>62.820</td>
</tr>
<tr>
<td>BiDAF + Character Embedding + Fusion Fm</td>
<td>64</td>
<td>4h13m</td>
<td>60.33</td>
<td>64.19</td>
</tr>
<tr>
<td>QANet (2 heads, 3 model encoder blocks)</td>
<td>64</td>
<td>3h32m</td>
<td>62.85</td>
<td>67.00</td>
</tr>
<tr>
<td>QANet (4 heads, 5 model encoder blocks)</td>
<td>32</td>
<td>6h43m</td>
<td>63.737</td>
<td>67.507</td>
</tr>
<tr>
<td>QANet (8 heads, 7 model encoder blocks)</td>
<td>16</td>
<td>13h45m</td>
<td>63.27</td>
<td>67.13</td>
</tr>
</tbody>
</table>

- Each model was trained end-to-end with hyperparameters and optimizers specified by the original papers.
- Gradient accumulation was used to counteract the training instability introduced by smaller batch sizes for the large QANet models.

References


Analysis

- QANet generally performed higher than BiDAF, as expected.
- However, the incremental benefit of adding attention heads/encoder blocks was outweighed by the steep increase in training time for larger models.
- Adding a simple MLP fusion function to post-process the BiDAF attention output significantly increased performance over the baseline.

Conclusions

- The original paper’s claim that QANet is faster to train than BiDAF is refuted in resource-constrained environments since batch size must be decreased.
- Larger model # better perf; layer dropout could have improved dev results.
- The combination of convolutions, position encoding, and self-attention in QANet is promising as an alternative to traditional RNN encoders.