Really Paying Attention
A BERT+BiDAF Ensemble Model for Question-Answering

Introduction

What is the significance of Question Answering?

Question answering (QA) is one of the most elusive disciplines within the realm of computer science and natural language processing. Often viewed as a litmus test of artificial intelligence approaching human-level abilities, QA involves processing a corpus of information to construct answers. In a world increasingly being changed by artificial intelligence, QA has proved its importance and significance by helping augment human cognition through tools such as Wolfram Alpha.

Problem

Within the last year, the discipline of QA has experienced momentous breakthroughs with the release of Pretrained Contextual Embedding (PCE)-based approaches such as BERT. Compared to models based on classic static word embeddings, models based on PCE incorporate context in order to capture the polysemous qualities of words, among many other benefits. Interestingly, these PCE models often required only one additional layer to adapt to various NLP tasks and achieve impressive performance.

Therefore, this project seeks to answer the question of whether PCE models are poised to supplant the past era of designing and iterating on diverse, complex architectures.

Approach

<table>
<thead>
<tr>
<th>Model</th>
<th>BERT-1</th>
<th>BERT-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train Batch Size</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>3e-5</td>
<td>3e-5</td>
</tr>
<tr>
<td>Train Epochs</td>
<td>1.0</td>
<td>3.0</td>
</tr>
<tr>
<td>Max Seq Length</td>
<td>384</td>
<td>384</td>
</tr>
<tr>
<td>Doc Stride</td>
<td>128</td>
<td>128</td>
</tr>
<tr>
<td>16-Bit Float Precision</td>
<td>True</td>
<td>True</td>
</tr>
</tbody>
</table>

BiDAF Baseline

BERT-1 72.590 75.394 72.354 74.903
BERT-2 74.136 77.450 73.542 76.848
BERT+BiDAF (0.25 threshold) 69.148 72.365 N/A N/A
BERT+BiDAF (0.10 threshold) 73.097 76.318 N/A N/A
BERT+BiDAF (0.02 threshold) 73.786 77.074 73.609 77.165

Analysis

“But what’s wrong, Ernie? What’s the matter?!”

- Bert

Increasing train batch size from 2 to 8 led to an ~2 point increase in Dev and Test scores. The raw BERT with train batch size of 32 sees another ~6 point increase.

STOCHASTIC ENSEMBLE

Though more makeshift in nature, BiDAF was able to rectify some of BERT’s errors. The problem is that randomness leads to erratic or inconsistent performance.

SELECTED EXAMPLE

Q: What Yuan policies did Muslims like?
C: Some policies of the Yuan Emperors severely discriminated against Muslims, restricting Islamic practices like circumcision.

Conclusion / Future

Improvements afforded by increasing compute dwarf the small increments afforded by the ensemble model.

- More robust and comprehensive way to ensemble (e.g., learning weighting of predictions through adding one more layer)
- Implement another ensemble for BERT (e.g., AOA)

Sources: Jay Alammar, Noun Project, BiDAF team. Much thanks to Prof. Manning and the CS224n teaching staff.