DeepDoc: NLP with Deep Neural Networks for the American Board of Internal Medicine Certification Exam

Jonathan X. Wang, Britni Chau, Kinbert Chou
[jonwang1, britnic, kichou]@stanford.edu

Prediction Task

- No system currently exists that assists physicians through natural language queries and direct answers.
- Rapidly growing amount of literature makes it harder for physicians to find relevant information for treatment [1].
- As a first pass, can we train a neural network to answer review questions for a physician certification exam?

Data

- 3564 examples were scraped from 2012, 2015, and 2018 review questions.
- Each question is comprised of a question, accompanying context passage, and 4 or 5 answer choice selections.
- We do a time split to capture ability to generalize on future problems.
- Dev: 600 examples (1/2 of 2018).
- Test: 600 examples (1/2 of 2018).

Example of a question:

**Passage:** A 76-year-old woman is evaluated... rapid ventricular rate.

**Question:** Which of the following is the most appropriate acute treatment?

**Answer Options:**
- A. Adenosine
- B. Amiodarone
- C. Cardioversion
- D. Diltiazem
- E. Metoprolol

**Correct answer:** C. Cardioversion.

**Explanation:** This patient with atrial fibrillation is hemodynamically unstable and should undergo immediate cardioversion...or diltiazem could worsen the pulmonary edema.

Approach and Results

- DrQA used to extract relevant explanations from training set when evaluating on dev/test. Top 3 explanations are used as input.
- Models include GA, SAR adapted from RACE [2]. And a modified BiDAF baseline.

Analysis

<table>
<thead>
<tr>
<th>Group</th>
<th>Relevant Explanation (%)</th>
<th>Helpful Explanation (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 5</td>
<td>0.266</td>
<td>0.133</td>
</tr>
<tr>
<td>Bottom 5</td>
<td>0.200</td>
<td>0.066</td>
</tr>
</tbody>
</table>

**Figure 1.** Flow diagram of prediction task.

**Figure 2.** Prediction Results demonstrate strong performance of GA model for this task.

**Figure 3.** Ensembled model with correct explanations vs. with DrQA explanations show little difference, suggesting difficulties in reading comprehension or lack of signal.

**Figure 4.** ROC models for top-performing models and baselines.

**Figure 5.** Precision-Recall curve for top-performing models and baselines.

**Figure 6.** Looked at the top and bottom 5 scoring explanations (30 explanations total) and found that only 7-13% of cases had helpful explanations.

- Tuning didn’t perform well, likely due to searching too wide of a space.
- Our model could benefit from different comprehension architectures, or a better search corpus.

Conclusion & Next Steps

- Demonstrate relatively good performance of the GA model, especially compared to RACE (MC task dataset) baseline of 40%, and a 50-60% passing score on the exam.
- Next steps include:
  - Character embeddings
  - Longer hyperparameter search
  - Validation on an official released exam
  - Try DrQA on wikipedia or UpToDate

Acknowledgments & References

Thank you to Jonathan H. Chen and Yuhao Zhang for their assistance and advising of this project.
