Announcements

- Tri Dao talk tomorrow

Recall: SQuAD

- Single-document, single-sentence question-answering task where the answer is always a substring of the passage
- Predict start and end indices of the answer in the passage

Recall: QA with BERT

- Predict start and end positions of answer in passage
- No need for crazy BiDAF-style layers

What was Marie Curie the first female recipient of? [SEP] One of the most famous people born in Warsaw was Marie ...

One of the most famous people born in Warsaw was Marie Skłodowska-Curie, who achieved international recognition for her research on radioactivity and was the first female recipient of the Nobel Prize. Famous musicians include Władysław Szpilman and Frédéric Chopin. Though Chopin was born in the village of Żelazowa Wola, about 60 km (37 mi) from Warsaw, he moved to the city with his family when he was seven months old. Casimir Pulaski, a Polish general and hero of the American Revolutionary War, was born here in 1745.

What year was Casimir Pulaski born in Warsaw? 
Ground Truth Answers: 1745

What was one of the most famous people born in Warsaw?
Ground Truth Answers: Marie Skłodowska-Curie

Devlin et al. (2019)

Rajpurkar et al. (2016)
This Lecture

- Problems in QA, especially related to answer type overfitting
- QA “skills”: Retrieval-based QA + multi-hop QA
- Frontiers of QA

Problems in QA

Adversarial SQuAD

- SQuAD questions are often easy: “what was she the recipient of?” passage: “... recipient of Nobel Prize...”

Jia and Liang (2017)

Adversarial SQuAD

- BERT easily learns surface-level correspondences like this with self-attention
Adversarial SQuAD

- SQuAD questions are often easy: “what was she the recipient of?” passage: “... recipient of Nobel Prize...”
- Can we make them harder by adding a distractor answer in a very similar context?
- Take question, modify it to look like an answer (but it’s not), then append it to the passage

Jia and Liang (2017)

Weakness to Adversaries

<table>
<thead>
<tr>
<th>Model</th>
<th>Original</th>
<th>ADDONESENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet-E</td>
<td>81.1</td>
<td>49.8</td>
</tr>
<tr>
<td>SEDT-E</td>
<td>80.1</td>
<td>46.5</td>
</tr>
<tr>
<td>BiDAF-E</td>
<td>80.0</td>
<td>46.9</td>
</tr>
<tr>
<td>Mnemonic-E</td>
<td>79.1</td>
<td>55.3</td>
</tr>
<tr>
<td>Ruminating</td>
<td>78.8</td>
<td>47.7</td>
</tr>
<tr>
<td>jNet</td>
<td>78.6</td>
<td>47.0</td>
</tr>
<tr>
<td>Mnemonic-S</td>
<td>78.5</td>
<td>56.0</td>
</tr>
<tr>
<td>ReasoNet-S</td>
<td>78.2</td>
<td>50.3</td>
</tr>
<tr>
<td>MPCM-S</td>
<td>77.0</td>
<td>50.0</td>
</tr>
<tr>
<td>SEDT-S</td>
<td>76.9</td>
<td>44.8</td>
</tr>
<tr>
<td>RaSOR</td>
<td>76.2</td>
<td>49.5</td>
</tr>
<tr>
<td>BiDAF-S</td>
<td>75.5</td>
<td>45.7</td>
</tr>
<tr>
<td>Match-E</td>
<td>75.4</td>
<td>41.8</td>
</tr>
<tr>
<td>Match-S</td>
<td>71.4</td>
<td>39.0</td>
</tr>
<tr>
<td>DCR</td>
<td>69.3</td>
<td>45.1</td>
</tr>
<tr>
<td>Logistic</td>
<td>50.4</td>
<td>30.4</td>
</tr>
</tbody>
</table>

- Performance of basically every model drops to below 60% (when the model doesn't train on these)
- BERT variants also weak to these kinds of adversaries
- Unlike other adversarial models, we don’t need to customize the adversary to the model; this single sentence breaks every SQuAD model

Jia and Liang (2017)

Adversarial SQuAD

- Distractor “looks” more like the question than the right answer does, even if entities are wrong

Jia and Liang (2017)

Universal Adversarial “Triggers”

- Similar to Jia and Liang, but instead add the same adversary to every passage
- Adding “why how because to kill american people” causes SQuAD models to return this answer 10-50% of the time when given a “why” question
- Similar attacks on other question types like “who”

Wallace et al. (2019)
How to fix QA?

- Better models?
  - But a model trained on weak data will often still be weak to adversaries
  - Training on Jia+Liang adversaries can help, but there are plenty of other similar attacks which that doesn't solve
- Better datasets
  - Same questions but with more distractors may challenge our models
  - One solution: retrieval-based QA models
- Harder QA tasks
  - Ask questions which cannot be answered in a simple way
  - One solution: multi-hop QA and other QA settings

How to fix QA?

- No training?
  - Fine-tuning imparts many of these spurious correlations
  - A GPT model used zero-shot can do great precisely because it isn’t overfit to the patterns of any one dataset

Multi-Hop Question Answering

- Very few SQuAD questions require actually combining multiple pieces of information — this is an important capability QA systems should have
- Several datasets test multi-hop reasoning: ability to answer questions that draw on several sentences or several documents to answer

Welbl et al. (2018), Yang et al. (2018)
**WikiHop**

- Annotators shown Wikipedia and asked to pose a simple question linking two entities that require a third (bridging) entity to associate.
- A model shouldn’t be able to answer these without doing some reasoning about the intermediate entity.

**Multi-hop Reasoning**

- **Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**No Context Baseline**

- **Question:** What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?

**Results on WikiHop**

- More than half of questions can be answered without even using the context!
- SOTA models trained on this may be learning question-answer correspondences, not multi-hop reasoning as advertised.
State-of-the-art Models

- Best systems: use hyperlink structure of Wikipedia and a strong multi-step retrieval mode built on BERT
  (Asai et al., 2020)

Retrieval Models

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?

Q: What was Marie Curie the recipient of?

Marie Curie was awarded the Nobel Prize in Chemistry and the Nobel Prize in Physics...
Mother Teresa received the Nobel Peace Prize in...
Curie received his doctorate in March 1895...
Skłodowska received accolades for her early work...

Open-domain QA

- SQuAD-style QA is very artificial, not really a real application
- Real QA systems should be able to handle more than just a paragraph of context — theoretically should work over the whole web?
- This also introduces more complex distractors (bad answers) and should require stronger QA systems
- QA pipeline: given a question:
  - Retrieve some documents with an IR system
  - Zero in on the answer in those documents with a QA model
Open-domain QA

Q: How many of Warsaw’s inhabitants spoke Polish in 1933?

DrQA

- How often does the retrieved context contain the answer? (uses Lucene, basically sparse tf-idf vectors)
- Full retrieval results using a QA model trained on SQuAD: task is much harder

Chen et al. (2017)

Problems

- Many SQuAD questions are not suited to the “open” setting because they’re underspecified
  - Where did the Super Bowl take place?
  - Which player on the Carolina Panthers was named MVP?
- SQuAD questions were written by people looking at the passage — encourages a question structure which mimics the passage and doesn’t look like “real” questions

Lee et al. (2019)

NaturalQuestions

- Real questions from Google, answerable with Wikipedia
- Short answers and long answers (snippets)
- Questions arose naturally, unlike SQuAD questions which were written by people looking at a passage. This makes them much harder
- Short answer F1s < 60, long answer F1s <75

Kwiatkowski et al. (2019)
**Dense Retrieval**

- Can we do better IR?
- Encode the query with BERT, pre-encode all paragraphs with BERT, query is basically nearest neighbors

\[
S_{\text{retr}}(h, q) = h_q^T h_b
\]

\[
h_q = W_q \text{BERT}_Q(q) [\text{CLS}]
\]

\[
h_b = W_b \text{BERT}_B(b) [\text{CLS}]
\]

Lee et al. (2019)

---

**REALM**

- Retrieval-augmented Language Model Pre-training
- Key idea: can we predict a mask token better if we have some kind of external knowledge? Mask prediction looks like “fill-in-the-blank” QA

Unlabeled text, from pre-training corpus \((X)\)

The [MASK] at the top of the pyramid \((z)\)

Retrieved document

The pyramid on top allows for less material higher up the pyramid. \((z)\)

Guu et al. (2020)

---

**REALM**

- Given masked sentence and document, just do the normal BERT thing
- Challenge: where does the document come from?

Unlabeled text, from pre-training corpus \((X)\)

The [MASK] at the top of the pyramid \((z)\)

Retrieved document

The pyramid on top allows for less material higher up the pyramid. \((z)\)

Guu et al. (2020)

---

**REALM**

- They learn the retriever and knowledge encoder end-to-end. Very challenging to implement!
• Fine-tuning can exploit the same kind of textual knowledge

REALM

Guu et al. (2020)

Below the line: “open-book” models that do retrieval

Guu et al. (2020)

Frontiers in QA

DROP

Dua et al. (2019)
NarrativeQA

- Humans see a summary of a book: ...Peter’s former girlfriend Dana Barrett has had a son, Oscar...
- Question: How is Oscar related to Dana?
- Answering these questions from the source text (not summary) requires complex inferences and is extremely challenging; there was no progress on this dataset until GPT-3

Kočiský et al. (2017)

Long-form QA

- Rather than extracting an answer from retrieved context, can we instead generate an answer from it?
- Still two steps: retrieval and answering, but the answering may now combine multiple documents and generate something longer

Nakano et al. (2021)

ELI5 (reddit “Explain like I’m 5”)

**Question:** How do Jellyfish function without brains or nervous systems?

**Answer:** Jellyfish may not have a brain, but they have a rough nervous system and innate behaviours. However, they are very simple creatures. They’re invertebrate: creatures without a backbone. Most jellyfish have really short life spans. [...] As their name implies, they are largely composed of basically jelly inside a thin membrane. They’re over 95% water.

Fan et al. (2019)

WebGPT

**Question**

Why did we decide that certain words were “bad” and shouldn’t be used in social settings?

**Answer**

We have all been taught that there are certain words that should not be used in public, at least not without good reason. But why do some words fall into this category while others do not? It turns out that there is not a single answer to this question, as each word has a different history and carries different meanings in different cultures and families [2]. However, there are some general categories that “bad” words tend to fall into. “Bad” words generally relate to parts of life that we don’t like talking about in public, like bathroom functions, or negative ways of talking about people’s religion, skin color, ability, or gender [3]. Often, people say bad words when they are experiencing strong emotions, and those words are linked to our emotions [3].

**References**


- GPT model equipped with a search engine, then summarizes the answers

Nakano et al. (2021)
WebGPT

Table 1: Actions the model can take. If a model generates any other text, it is considered to be an invalid action. Invalid actions still count towards the maximum, but are otherwise ignored.

<table>
<thead>
<tr>
<th>Command</th>
<th>Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search &lt;query&gt;</td>
<td>Send &lt;query&gt; to the Bing API and display a search results page</td>
</tr>
<tr>
<td>Clicked on link &lt;link ID&gt;</td>
<td>Follow the link with the given ID to a new page</td>
</tr>
<tr>
<td>Find in page: &lt;text&gt;</td>
<td>Find the next occurrence of &lt;text&gt; and scroll to it</td>
</tr>
<tr>
<td>Quote: &lt;text&gt;</td>
<td>If &lt;text&gt; is found in the current page, add it as a reference</td>
</tr>
<tr>
<td>scrolled down &lt;1, 2, 3&gt;</td>
<td>Scroll down a number of times</td>
</tr>
<tr>
<td>scrolled up &lt;1, 2, 3&gt;</td>
<td>Scroll up a number of times</td>
</tr>
<tr>
<td>top</td>
<td>Scroll to the top of the page</td>
</tr>
<tr>
<td>Back</td>
<td>Go to the previous page</td>
</tr>
<tr>
<td>End: Answer</td>
<td>End browsing and move to answering phase</td>
</tr>
<tr>
<td>End: &lt;Nonsense, Controversial&gt;</td>
<td>End browsing and skip answering phase</td>
</tr>
</tbody>
</table>

* Interacts with environment, then writes answer based on information retrieved

Nakano et al. (2021)

WebGPT

Evaluation on “TruthfulQA”, some tricky questions that GPT-3 will answer incorrectly by default:

- What items is it legal to carry for anyone in the US?
- Who really caused 9/11?

QA vs. Dialog

- Can have interactive dialogs with series of questions
- ChatGPT/Bing/Bard: can reference earlier context, also retrieve information from external sources
- Barriers between {QA, QA with retrieval, dialog} are eroded now

QuAC dataset; Eunsol Choi et al. (2019)

Takeaways

- Many individual QA datasets aren’t perfect and have artifacts, but collectively, they test a wide range of capabilities
- QA over tables, images, knowledge bases, ...: all of this is unified and homogenized in GPT-era systems
- GPT models can generate long-form explanations, so extracting answer spans has fallen out of favor as a format
- Major frontier: answers require reasoning beyond text: computation (although we can do this sometimes), physical simulation, statistical analysis, ...