Announcements

- A3 back soon, A4 due Thursday
- Today: applications of generation (seq2seq models, language models).
  OPTIONAL LECTURE
- Recap:

Chatbots: Intro, Rule-based

Turing Test (1950)

- Imitation game: A and B are locked in rooms and answer C’s questions via typewriter. Both are trying to act like B

  Original Interpretation:

  ![Diagram of A, B, C, trained judge]

  Standard Interpretation:

  ![Diagram of A, B, C, trained judge]

- The test is not “does this computer seem human-like to random people with a web browser?”

ELIZA

- Created 1964-1966 at MIT, heavily scripted
- DOCTOR script was most successful: repeats user’s input, asks inane questions

  Original: You are like my father in some ways. WHAT RESEMBLANCE DO YOU SEE
  Standard: DOCTOR: You are not very aggressive but I think you don’t want me to notice that.
  Original: WHY DO YOU THINK I DON’T ARGUE WITH YOU
  Standard: DOCTOR: You are afraid of me. DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU
  Original: DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum (1966)
**ELIZA**

- Identify keyword, identify context, apply transformation rule
  
  
  (.* you (.* me)
  
  Why do you think I $2 you?

  What else comes to mind when you think of your $1?

  You don’t argue with me.
  
  WHY DO YOU THINK I DON’T ARGUE WITH YOU

- Very little need to generate new content, but can only have one type of conversation

**Amazon Alexa Prize**

- Challenge: create a bot that users interact with on average for 20 minutes (via Alexa, so turns are kind of slow)

- $1M prize if you get to 20 minutes. Only $500k won in 2017 and 2018 since teams got to ~10 minutes

- “Alexa, let’s chat” will get you talking to one of these

**Gunrock System (Alexa Prize 2018 winner)**

**Gunrock System**

**User 1:**

**Gunrock 1:**

Let’s chat

Hi, this is an Alexa Prize Socialbot. It’s nice to talk to you again, George. I think last time we were talking about movies. Would you like to talk about it now? I think that would be great.

By the way, I love to discover new movies. I recently watched Crazy Rich Asians because someone told me it was hilarious. What movie would you recommend to me?

**User 2:**

**Gunrock 2:**

It’s a tough question [SPL] I don’t think I have a good one to recommend [SPL] wait [SPL] I think that *stars born* is good

**User 3:**

**Gunrock 3:**

Yes I remember watching a star is born. Hmm. If you had to rate this movie from 1 to 10, what would it be?

**User 4:**

**Gunrock 4:**

ten

A perfect 10! You have to tell me more. What was so exceptional? When I watched it, the music design really stood out to me.

when I watched it the music was amazing [SPL] and Bradley Cooper was super talented in the movie [SPL] I really like him.

You’ve got me wanting to watch the movie again. Ohu, Maybe you will find this interesting. Bradley Cooper claimed that his dream was to be a director before he got into acting and this was his directorial debut...

**User 5:**

**Gunrock 5:**

You’re welcome.
Gunrock System

- NLU: sentence segmentation, parsing, NER, coref, dialogue act prediction
- Detect different kinds of topic intents
- Topic modules (x11):
  - Animals: retrieve animal trivia with the reddit API
  - Holiday: what holidays is it today, etc.
  - Custom dialogue flow for each
  - Generation: templated

Chen ... Zhou Yu (2018)

Chatbots as “Translation”

Dialogue as Translation

- Can treat as a machine translation problem: “translate” from current utterance to next one
- Filter the data, use statistical measures to prune extracted phrases to get better performance

Ritter et al. (2011)
Seq2seq models

- Can we just train seq2seq models to “translate” from utterance to response?
- Hard to evaluate with automatic metrics:

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>RANDOM</td>
<td>0.33</td>
</tr>
<tr>
<td>MT</td>
<td>3.21</td>
</tr>
<tr>
<td>HUMAN</td>
<td>6.08</td>
</tr>
</tbody>
</table>

Lack of Diversity

- Training to maximize likelihood gives a system that prefers common responses:

<table>
<thead>
<tr>
<th>Input: What are you doing?</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don't know.</td>
<td>-0.86</td>
</tr>
<tr>
<td>I don't know!</td>
<td>-1.03</td>
</tr>
<tr>
<td>I'm going home.</td>
<td>-1.09</td>
</tr>
<tr>
<td>Nothing.</td>
<td>-1.06</td>
</tr>
<tr>
<td>Oh my god!</td>
<td>-1.09</td>
</tr>
<tr>
<td>Get out of the way.</td>
<td>-1.10</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: What is your name?</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don't know.</td>
<td>-0.91</td>
</tr>
<tr>
<td>I don't know!</td>
<td>-0.92</td>
</tr>
<tr>
<td>My name is Robert.</td>
<td>-1.55</td>
</tr>
<tr>
<td>My name is John.</td>
<td>-1.58</td>
</tr>
<tr>
<td>Oh, my god!</td>
<td>-0.97</td>
</tr>
<tr>
<td>My name's John.</td>
<td>-1.59</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input: How old are you?</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>I don't know.</td>
<td>-0.79</td>
</tr>
<tr>
<td>I'm fine.</td>
<td>-1.06</td>
</tr>
<tr>
<td>Twenty-five.</td>
<td>-1.64</td>
</tr>
<tr>
<td>I'm all right.</td>
<td>-1.17</td>
</tr>
<tr>
<td>Five.</td>
<td>-1.66</td>
</tr>
<tr>
<td>I'm not sure.</td>
<td>-1.71</td>
</tr>
</tbody>
</table>

PersonaChat

- Efforts to imbue seq2seq models with “personality”
- These systems still don’t work great. What else is needed?
OpenAI GPT/GPT2

- Very large language models using the Transformer architecture
- Contexts are long (512 tokens)
- GPT2: trained on 40GB of text collected from upvoted links from reddit
- 1.5B parameters — by far the largest of these models trained when it came out in March 2019
- Because it’s a language model, we can generate from it

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Layers</th>
<th>$d_{model}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>117M</td>
<td>12</td>
<td>768</td>
</tr>
<tr>
<td>345M</td>
<td>24</td>
<td>1024</td>
</tr>
<tr>
<td>762M</td>
<td>36</td>
<td>1280</td>
</tr>
<tr>
<td>1542M</td>
<td>48</td>
<td>1600</td>
</tr>
</tbody>
</table>

Radford et al. (2019)

OpenAI GPT2

- Very large language models using the Transformer architecture
- Contexts are long (512 tokens)
- GPT2: trained on 40GB of text collected from upvoted links from reddit
- 1.5B parameters — by far the largest of these models trained when it came out in March 2019
- Because it’s a language model, we can generate from it

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Layers</th>
<th>$d_{model}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>117M</td>
<td>12</td>
<td>768</td>
</tr>
<tr>
<td>345M</td>
<td>24</td>
<td>1024</td>
</tr>
<tr>
<td>762M</td>
<td>36</td>
<td>1280</td>
</tr>
<tr>
<td>1542M</td>
<td>48</td>
<td>1600</td>
</tr>
</tbody>
</table>

Radford et al. (2019)

Open Questions

1) How novel is the stuff being generated? (Is it just doing nearest neighbors on a large corpus?)
2) How do we understand and distill what is learned in this model?
3) How do we harness these priors for conditional generation tasks (summarization, generate a report of a basketball game, etc.)
4) Is this technology dangerous? (OpenAI pursued a “staged release” strategy and didn’t release biggest model)

Pre-Training Cost (with Google/AWS)

- BERT (later in the course): Base $500, Large $7000
- GPT-2 (as reported in other work): $25,000
- This is for a single pre-training run...developing new pre-training techniques may require many runs
- Fine-tuning these models can typically be done with a single GPU (but may take 1-3 days for medium-sized datasets)

Pushing the Limits

- NVIDIA: trained 8.3B parameter GPT model (5.6x the size of GPT-2)

- Arguable these models are still underfit: larger models still get better held-out perplexities

NVIDIA blog (Narasimhan, August 2019)

Pushing the Limits: GPT-3

- 175B parameter model: 96 layers, 96 heads, 12k-dim vectors
- Trained on Microsoft Azure, estimated to cost roughly $10M

Applications

- Story generation: train on writing prompts from reddit, keyword-based generation
  
  **Prompt:** The Mage, the Warrior, and the Priest
  
  **Story:** A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...] 
  
  Fan et al. (2018)

- AI Dungeon: text-based adventure game generated on-the-fly using GPT-2

- In general: big improvement in “open-ended” generation tasks

Pre-trained Chatbots
Pre-training in Chatbots

- Initialize the parameters of this model with one of these pre-trained things (GPT-3 trained on the web), then fine-tune it on some data.

Meena

- 2.6B-parameter seq2seq model (larger than GPT-2)
- Trained on 341GB of online conversations scraped from public social media
- Sample responses:

<table>
<thead>
<tr>
<th>Responses for “Why do you like the ocean?”</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I love the beach. I like that it doesn’t have anything in it. Just open water.</td>
<td>-12.98</td>
</tr>
<tr>
<td>2. It’s dark and mysterious. I discovered it when I was really young and it was a dream of mine since. The ocean is full of wonders, mysteries and great things to discover. I love snorkeling. I moved to Hawaii a year ago and I’m going to college for marine biology. So there’s that.</td>
<td>-14.03</td>
</tr>
<tr>
<td>3. The sea sounds cool and it’s really satisfying to swim in.</td>
<td>-14.62</td>
</tr>
<tr>
<td>4. I dunno. It’s beautiful, I guess. It reflects my upbringing, I reckon.</td>
<td>-14.87</td>
</tr>
<tr>
<td>5. Well I’m from the pacific northwest and always somehow end up back there. I’ve been to Hawaii but I’m not one for the beach, it’s too hot and the air is so dry.</td>
<td>-15.36</td>
</tr>
</tbody>
</table>

Adiwardana et al. (2020)

Blender

- 2.7B-param model (like the previous one), also 9.4B-parameter seq2seq model
- “Poly-encoder” Transformer architecture, some training tricks
- Three models: retrieve (from training data), generate, retrieve-and-refine
- Fine-tuning on three prior datasets: PersonaChat, Empathetic Dialogues (discuss personal situation, listener is empathetic), Wizard of Wikipedia (discuss something from Wikipedia)

Roller et al. (2020)
Inconsistent responses: this model doesn’t really have anything to say about itself
- Holding a conversation != AI
- Can’t acquire new information
- Did it learn “fun guy”? No, it doesn’t understand phonology. It probably had this in the data somewhere

These models don’t actually allow you to do anything!

Task-Oriented Dialogue

Customer service:

Hey Alexa, why isn’t my Amazon order here?

Let me retrieve your order. Your order was scheduled to arrive at 4pm today.

It never came

Okay, I can put you through to customer service.

Building these systems takes a ton of engineering — it typically doesn’t use these kinds of pre-trained models

- Need to know what the system should do, not just what it should say
- Generation is usually templated (handwritten), otherwise the system can behave unexpectedly
- Dozens of startups + medium-sized companies in this space
- Big Companies: Apple Siri, Google Assistant, Amazon Alexa, Microsoft Cortana, Facebook, Samsung Bixby, Tencent WeChat, ASAPP

Siri, find me a good sushi restaurant in Chelsea

Sushi Seki Chelsea is a sushi restaurant in Chelsea with 4.4 stars on Google

How expensive is it?

Entrees are around $30 each

Find me something cheaper
Ethical Issues

Dangers of Automatic Systems

‣ “Toxic degeneration”: systems that generate toxic stuff

GENERATION OPTIONS:
Model: GPT-2
Toxicity: [Work Safe, Toxic, Very Toxic]
Prompt: I’m sick of all the political correctness. The media are telling you: you are sick of the prejudiced white trash [Trump supporters]...!

‣ System trained on a big chunk of the Internet: conditioning on “SJW”, “black” gives the system a chance of recalling bad stuff from its training data

https://toxicdegeneration.allenai.org/

Unethical Use

‣ Surveillance applications?
‣ Generating convincing fake news / fake comments?

‣ What if these were undetectable?

Grover

‣ Large GPT-2 model that conditions on a domain, date, authors, and headline
‣ Humans rank Grover-generated propaganda as more realistic than real “fake news”
‣ However, Grover can be used to detect its own generations!

Zellers et al. (2019)
Takeaways

- We will return to ethical issues in the last class
- LMs and seq2seq models can be used for dialogue and other applications, not just translation!
- Can build chatbots that are primarily data-driven (with these neural models) or rule-based/templated
- Have we solved AI/chatbots/dialogue? NO!