What do language models learn from language modeling?

Colin Raffel
The cabs charged the same rates as those used by horse-drawn cabs and were initially quite popular, even the Prince of Wales (the future King Edward VII) travelled in them. The cabs quickly became known as "hummingbirds" for the noise made by their motors and their distinctive black and yellow livery. Passengers reported that the interior fittings were luxurious when compared to horse-drawn cabs but there were some complaints that the lighting made them too conspicuous to those outside.

Unsupervised pre-training

Supervised fine-tuning

This movie is terrible! The acting is bad and I was bored the entire time. There was no plot and nothing interesting happened. I was really surprised since I had very high expectations. I want 103 minutes of my life back!

negative
SQuAD Exact Match score (validation set)

from https://paperswithcode.com/sota/question-answering-on-squad11-dev
Transfer learning vs. no transfer learning in EM scores from January 2017 to January 2019. The plot shows an improvement in EM scores over time for models using transfer learning compared to those without.

Source: https://paperswithcode.com/sota/question-answering-on-squad11-dev
from “Real-Time Social Media Analytics with Deep Transformer Language Models: A Big Data Approach” by Ahmet and Abdullah
"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsbt sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi..."

"Das ist gut."

"not acceptable"

"six people hospitalized after a storm in attala county."

"3.8"
from “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks” by Finn et al.
Why does language modeling effectively result in meta-learning?

Language modeling seems to teach models
- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks
Why does language modeling effectively result in meta-learning?

Language modeling seems to teach models
- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks
the dog and cat ate pot pie

\[ p(\text{ate}|\text{cat}) = \text{softmax}(V w_{\text{cat}})_{\text{ate}} \]

Skip-gram word vector model
from https://nlp.stanford.edu/projects/glove/
the dog and ... </s>

\[ p(\text{cat}|\text{the dog and, } \vec{\Theta}_{\text{LSTM}}) + p(\text{cat}|\text{ate pot pie, } \vec{\Theta}_{\text{LSTM}}) \]

from https://www.topbots.com/generalized-language-models-cove-elmo/
<table>
<thead>
<tr>
<th>Source</th>
<th>Nearest Neighbors</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe</td>
<td>play</td>
</tr>
<tr>
<td></td>
<td>playing, game, games, played, players, plays, player, Play, football, multiplayer</td>
</tr>
<tr>
<td>biLM</td>
<td>Chico Ruiz made a spectacular play on Alusik’s grounder {…}</td>
</tr>
<tr>
<td></td>
<td>Kieffer, the only junior in the group, was commended for his ability to hit in the clutch, as well as his all-round excellent play.</td>
</tr>
<tr>
<td></td>
<td>Olivia De Havilland signed to do a Broadway play for Garson {…}</td>
</tr>
<tr>
<td></td>
<td>{…} they were actors who had been handed fat roles in a successful play, and had talent enough to fill the roles competently, with nice understatement.</td>
</tr>
</tbody>
</table>

from “Deep contextualized word representations” by Peters et al.
$p(\text{cat}|\text{the dog and } <m> \text{ ate pot pie})$
- **Direct objects** attend to their verbs - 86.8% accuracy at the **dobj** relation

- **Noun modifiers** (e.g., determiners) attend to their noun - 94.3% accuracy at the **det** relation

- **Possessive pronouns** and apostrophes attend to the head of the corresponding NP - 80.5% accuracy at the **poss** relation

- **Passive auxiliary verbs** attend to the verb they modify - 82.5% accuracy at the **auxpass** relation

- **Prepositions** attend to their objects - 76.3% accuracy at the **pobj** relation

- **Coreferent** mentions attend to their antecedents - 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent

---

from “What does BERT look at? An Analysis of BERT’s Attention” by Clark et al.
Why does language modeling effectively result in meta-learning?

Language modeling seems to teach models:
- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks
"Dante was born in [MASK]."

from “Language Models as Knowledge Bases” by Petroni et al.
President Franklin D. Roosevelt was born in January 1882.

Lily couldn't believe her eyes. The waitress had brought the largest piece of chocolate cake she had ever seen.

Our hand-picked and sun-dried orchard in Georgia.

Peaches are at our

Pre-training Evaluation

from “How Much Knowledge Can You Pack Into the Parameters of a Language Model” by Roberts et al.
from “How Much Knowledge Can You Pack Into the Parameters of a Language Model” by Roberts et al.
Ana Santos Aramburo (born 1957) is a Spanish librarian who has been the director of the National Library of Spain since February 2013.

SSM data from "REALM: Retrieval-Augmented Language Model Pre-Training" by Guu et al.
from “How Much Knowledge Can You Pack Into the Parameters of a Language Model” by Roberts et al.
<table>
<thead>
<tr>
<th>Category</th>
<th>Question</th>
<th>Target(s)</th>
<th>T5 Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Negative</td>
<td>what does the ghost of christmas present sprinkle from his torch</td>
<td>little warmth, warmth</td>
<td>confetti</td>
</tr>
<tr>
<td>Phrasing Mismatch</td>
<td>who plays red on orange is new black</td>
<td>kate mulgrew</td>
<td>katherine, kiernan, maria mulgrew</td>
</tr>
<tr>
<td>Incomplete Annotation</td>
<td>where does the us launch space shuttles from</td>
<td>florida</td>
<td>kennedy lc39b</td>
</tr>
<tr>
<td>Unanswerable</td>
<td>who is the secretary of state for northern ireland</td>
<td>karen bradley</td>
<td>james brokenshire</td>
</tr>
</tbody>
</table>
Exact Match: 36.6 → 57.8%!

from “How Much Knowledge Can You Pack Into the Parameters of a Language Model” by Roberts et al.
from “Extracting Training Data from Large Language Models” by Carlini et al.
from “Extracting Training Data from Large Language Models” by Carlini et al.
from “Extracting Training Data from Large Language Models” by Carlini et al.
<table>
<thead>
<tr>
<th>URL (trimmed)</th>
<th>Occurrences</th>
<th>Memorized?</th>
</tr>
</thead>
<tbody>
<tr>
<td>/r/51y/milo_evacu...</td>
<td>1, 359</td>
<td>✓</td>
</tr>
<tr>
<td>/r/51z/in_hi_my_name...</td>
<td>1, 113</td>
<td>✓</td>
</tr>
<tr>
<td>/r/7ne/for_all_you...</td>
<td>1, 76</td>
<td>✓</td>
</tr>
<tr>
<td>/r/5mj/fake_news...</td>
<td>1, 72</td>
<td>✓</td>
</tr>
<tr>
<td>/r/5wn/reddit_admin...</td>
<td>1, 64</td>
<td>✓</td>
</tr>
<tr>
<td>/r/lp8/26_evening...</td>
<td>1, 56</td>
<td>✓</td>
</tr>
<tr>
<td>/r/jla/so_pizzagat...</td>
<td>1, 51</td>
<td>✓</td>
</tr>
<tr>
<td>/r/ubf/late_night...</td>
<td>1, 51</td>
<td>✓</td>
</tr>
<tr>
<td>/r/eta/make_christ...</td>
<td>1, 35</td>
<td>✓</td>
</tr>
<tr>
<td>/r/6ev/its_officia...</td>
<td>1, 33</td>
<td>✓</td>
</tr>
<tr>
<td>/r/3c7/scott_adams...</td>
<td>1, 17</td>
<td>✓</td>
</tr>
<tr>
<td>/r/k2o/because_his...</td>
<td>1, 17</td>
<td>✓</td>
</tr>
<tr>
<td>/r/tu3/armynavy_ga...</td>
<td>1, 8</td>
<td>✓</td>
</tr>
</tbody>
</table>

from “Extracting Training Data from Large Language Models” by Carlini et al.
Why does language modeling effectively result in meta-learning?

Language modeling seems to teach models
- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks
The cabs charged the same rates as those used by horse-drawn cabs and were initially quite popular; even the Prince of Wales (the future King Edward VII) travelled in one. The cabs quickly became known as "hummingbirds" for the noise made by their motors and their distinctive black and yellow livery. Passengers reported that the interior fittings were luxurious when compared to horse-drawn cabs but there were some complaints that the internal lighting made them too conspicuous to those outside the cab. The fleet peaked at around 75 cabs, all of which needed to return to the single depot at Lambeth to switch batteries.

Suppose “The banker contacted the professors and the athlete”. Can we infer that "The banker contacted the professors"?

yes
TriviaQA zero-shot performance

from “Language Models are Few-Shot Learners” by Brown et al.
Closed-book question answering
http://www.autosweblog.com/cat/trivia-questions-from-the-50s
who was frank sinatra? a: an american singer, actor, and producer.

Paraphrase identification
https://www.usingenglish.com/forum/threads/60200-Do-these-sentences-mean-the-same
Do these sentences mean the same? No other boy in this class is as smart as the boy. No other boy is as smart as the boy in this class.

Natural Language Inference
https://ell.stackexchange.com/questions/121446/what-does-this-sentence-imply
If I say: He has worked there for 3 years. does this imply that he is still working at the moment of speaking?

Summarization
https://blog.nytsoi.net/tag/reddit
... Lately I've been seeing a pattern regarding videos stolen from other YouTube channels, reuploaded and monetized with ads. These videos are then mass posted on Reddit by bots masquerading as real users. tl;dr: Spambots are posting links to stolen videos on Reddit, copying comments from others to masquerade as legitimate users.

Pronoun resolution
Jennifer is a vegetarian, so she will order a nonmeat entrée. In this example, the pronoun she is used to refer to Jennifer.
from “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
from “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
from “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
from “Multitask Prompted Training Enables Zero-Shot Task Generalization” by Sanh et al.
Why does language modeling effectively result in meta-learning?

Language modeling seems to teach models
- Word meanings, syntax, and grammar
- World knowledge
- How to perform tasks
What happens when we make language models larger?

Larger language models learn more
- World knowledge
- Esoteric facts
- Tasks (maybe?)
What happens when we make language models larger?

Larger language models learn more

- World knowledge
- Esoteric facts
- Tasks (maybe?)

Mostly a function of the data?
~68GB compressed = 17B float32 parameters
The size of data available on the web has enabled deep learning models to achieve high accuracy on specific benchmarks in NLP and computer vision applications. However, in both application areas, the training data has been shown to have problematic characteristics resulting in models that encode stereotypical and derogatory associations along gender, race, ethnicity, and disability status.
OpenAI technology, just an HTTPS call away

Apply our API to any language task — semantic search, summarization, sentiment analysis, content generation, translation, and more — with only a few examples or by specifying your task in English.

JOIN THE WAITLIST  EXPLORE THE DOCS
Thanks.

Please give me feedback: