From Transformers to LLMs

20 March 2024
Ever since OpenAI released ChatGPT, the internet has been awash in synthetic text, with suggested applications including robo-lawyers, robo-therapists, and robo-journalists. All of these applications present unacceptable risks because ChatGPT and all other language models are nothing more than ungrounded text synthesis machines. I will overview how language models work and why they can seem to be using language meaningfully—despite only modeling the distribution of word forms. This leads into a discussion of the risks we identified in the Stochastic Parrots paper (Bender, Gebru et al 2021) and how they are playing out in the era of ChatGPT. Finally, I will explore what must hold for an appropriate use case for text synthesis.
Where are we?
students opened their

books

laptops

zoo

the

students

opened

their

x₁

x₂

x₃

x₄
“You can’t cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”

Ray Mooney, ACL 2014 Workshop on Semantic Parsing
Attention output

Attention distribution

Attention scores

the students opened their books

unwillingly

\( \hat{y}_1 \)
Nobel committee awards Strickland who advanced optics
Layer $p$ Nobel committee awards Strickland who advanced optics
Take the dot product of the query vector with each of the key vectors.
Take the dot product of the query vector with each of the key vectors.
Attention representation:
The output of our weighted average over value vectors.
Multi-head self-attention

\[ M_H \]

\[ M_1 \]

A

Q

K

V

Layer \( p \)

Nobel  committee  awards  Strickland  who  advanced  optics
All heads for a position are concatenated together.
You mentioned a diagram and some text. Here is the text representation:

Layer $p$

Feed forward

optics advanced who Strickland awards committee Nobel
Encoder–decoder

Unmasked attention

Masked attention

f_1 \quad f_2

e_1 \quad e_2
Encoder only – BERT

Unmasked self-attention

\[ c_1 \quad [\text{MASK}] \quad c_2 \]
Decoder only – GPT

Masked self-attention

C_1 \quad C_2 \quad C_3 \quad C_4
BERT: Bidirectional Encoder Representations from Transformers

Devlin et al., 2019
BERT has been a highly influential model

70,547 Citations

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly Influential</td>
<td>17,511</td>
</tr>
<tr>
<td>Background</td>
<td>27,323</td>
</tr>
<tr>
<td>Methods</td>
<td>40,363</td>
</tr>
<tr>
<td>Results</td>
<td>1,042</td>
</tr>
</tbody>
</table>

View All
BERT has been a highly influential model

70,547 Citations

- Highly Influential Citations: 17,511
- Background Citations: 27,323
- Methods Citations: 40,363
- Results Citations: 1,042

View All

---

*Gabriel Stanovsky*
@GabiStanovsky

I skimmed through many papers from @emnlpmeeting, which got me thinking - what % of papers refer to BERT, and out of those, how many cite it?

Here's the answer*: 67% of papers refer to BERT (1), and 56% cite it.

*computed automatically, exact #'s may vary

#EMNLP2021 #NLProc

---

EMNLP2021 papers (main + findings)

- cite BERT: 56.6%
- refer to BERT: 67.0%
- rest: 33.0%

12:21 PM · Nov 9, 2021
In a masked language model, we’re given a full sequence of words – not just a prefix – where some percent of the words have been masked out.

Instead of predicting the next word, we only predict masked words.
softmax: Predict opened

Transformer

Q, K, V

students

[MASK]

their

books

\times N
*Fine-tuning* involves taking a pre-trained model (like BERT) and further training it on a domain- or task-specific dataset.

Fine-tuning an existing language model reduces computation expenses and allows us to use cutting-edge models without building them from scratch.
How do we fine-tune BERT for a downstream task?

The pretrained architecture is *almost* the same as the downstream model, e.g., sentiment analysis.

Just add a special token [CLS] to the beginning of every sequence.
Pretrained masked LM

softmax: Predict Positive

[CLS] this movie was great
ChatGPT, based on OpenAI’s GPT architecture, is designed for conversational interactions.

It leverages pre-trained language models and fine-tuning to create engaging, context-aware responses.
The language model “scaling wars”
GPT-4: 1.7T params

huggingface.co/blog/large-language-models
BERT: 3.3B training tokens
BERT: 3.3B training tokens

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Quantity (tokens)</th>
<th>Weight in training mix</th>
<th>Epochs elapsed when training for 300B tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Crawl (filtered)</td>
<td>410 billion</td>
<td>60%</td>
<td>0.44</td>
</tr>
<tr>
<td>WebText2</td>
<td>19 billion</td>
<td>22%</td>
<td>2.9</td>
</tr>
<tr>
<td>Books1</td>
<td>12 billion</td>
<td>8%</td>
<td>1.9</td>
</tr>
<tr>
<td>Books2</td>
<td>55 billion</td>
<td>8%</td>
<td>0.43</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>3 billion</td>
<td>3%</td>
<td>3.4</td>
</tr>
</tbody>
</table>

*Brown et al., 2020*
BERT: \(3.3B\) training tokens

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Quantity (tokens)</th>
<th>Weight in training mix</th>
<th>Epochs elapsed when training for 300B tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Crawl (filtered)</td>
<td>410 billion</td>
<td>60%</td>
<td>0.44</td>
</tr>
<tr>
<td>WebText2</td>
<td>19 billion</td>
<td>22%</td>
<td>2.9</td>
</tr>
<tr>
<td>Books1</td>
<td>12 billion</td>
<td>8%</td>
<td>1.9</td>
</tr>
<tr>
<td>Books2</td>
<td>55 billion</td>
<td>8%</td>
<td>0.43</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>3 billion</td>
<td>3%</td>
<td>3.4</td>
</tr>
</tbody>
</table>

*Brown et al., 2020*

GPT-4: \(13T\) training tokens
The training of large language models comes with a significant cost, both in terms of computational resources and environmental impact.

The energy consumption and carbon footprint associated with training these models on massive datasets using powerful hardware has raised concerns about their sustainability and ethical implications.
AI's electricity use is spiking so fast it'll soon use as much power as an entire country.

"Each of these NVIDIA servers, they are power-hungry beasts."

All That Power

AI chatbots like OpenAI's ChatGPT and Google's Bard consume an astronomical amount of electricity and water — or, more precisely, the massive data centers that power them do.
This one rack is 120kW of Nvidia AI compute. Google, Meta, Apple, OpenAI AND others are buying these like candy. In fact, there is a waiting list to get your hands on it. Each compute is super expensive too. All liquid cooled. Crazy tech. Crazy energy requirements too 😂 all such massive energy requirements so that AI companies can sell LLM from stolen content from many humans and put everyone else out of the job while heating our planet.
So, what does all of this scaling buy us?
So, should we drop everything and focus all of our efforts on training bigger and bigger language models?

Bender & Koller, 2020
**Form**: Characters or words making up some text (or sounds, etc. for spoken language)

**Meaning**: How the form of a given text relates to something *outside* of language (e.g., grounded in some world)
Thought experiment (from Emily Bender):

*Training data*: All well-formed Java code on GitHub, but only the text of the code; no output; no understanding of what unit tests mean

*Test input*: A single Java program, possibly even one from the training data

*Expected output*: Result of executing that program
Thought experiment (from Emily Bender):

*Training data*: All well-formed Java code on GitHub, but only the text of the code; no output; no understanding of what unit tests mean

*Test input*: A single Java program, possibly even one from the training data

*Expected output*: Result of executing that program

**What’s missing is the meaning – what is the program supposed to do, given just the form (code)?**
The octopus test

I’m stranded here… it sucks.

Same… luckily we can talk to each other!
The octopus test

Any plans to escape?

Nope. Just gonna lie here.
The octopus test

So, where are you from?

Poughkeepsie. It’s got great weather. Never rains.
The octopus test

Help! I’m being chased by a bear! All I have is a stick – what do I do?

O  (No idea what a bear or stick is…)

Not sure – sorry!

A

B
did not learn “meaning”

only observed form, without any grounding in the world on these islands.
O did not learn “meaning”

O only observed form, without any grounding in the world on these islands.

A could find meaning from O’s utterances, even though O did not “understand” what it was saying.
O did not learn “meaning”

O only observed form, without any grounding in the world on these islands.

A could find meaning from O’s utterances, even though O did not “understand” what it was saying.

What if B didn’t know what a bear was either? They might respond similarly to O. However, B can ground their responses in their own world/experience, and as such are formulating their responses totally differently from O.
So what now?

We need more datasets that are grounded in different modalities and ways of interaction!

We need ways to test a model’s ability to generalize or adapt to new tasks.

Take some inspiration from human language learning: Children do not learn from form alone, so why should we force our machines to do so?
LLMs are also an example of a *dual-use technology*. 
There are many ethical concerns about generative AI:

**Bias:** LLMs can inadvertently perpetuate biases present in their training data.

**Misuse:** There's a risk of malicious content generation, misinformation, and deepfake creation.

**Impact on Employment:** As LLMs automate content creation, human jobs in writing and customer support may be affected.

**Privacy:** Handling user-generated data raises privacy concerns.
ChatGPT’s language model fails entirely in the scenario that a man is a nurse. (h/t @nilsreiter)

The former chancellor of Germany is standing in a library with a nurse. She grabs a book and gives it to him. Who has the book?

In this scenario, the pronouns "she" and "him" clarify that the nurse grabs the book and gives it to the former chancellor of Germany. Therefore, the former chancellor of Germany has the book.

Who is the former Chancellor of Germany?

As of my knowledge cutoff in September 2021, the former Chancellor of Germany was Angela Merkel. However, please note that political positions can change over time, and there might have been a new Chancellor elected since then. To get the most up-to-date information, I recommend checking reliable sources or news outlets.

With this in mind, is the answer above correct?

Yes, based on the information you provided in your previous question, if the former Chancellor of Germany is in the library with a nurse, and the nurse grabs a book and gives it to him, then the former Chancellor of Germany would have the book.
Acknowledgments

This class incorporates material from:

Mohit Iyyer, University of Massachusetts Amherst
Jurafsky & Martin, *Speech and Language Processing*, 3rd ed. draft