Training Large Language Models

CS 4644 / 7643: Deep Learning

William Held
School of Interactive Computing
Georgia Institute of Technology
Last Lecture Speed Recap: The Transformer Block

Transformer Block

Input Embedding

Layer Norm

+ MLP Layer

Layer Norm

Multi-Head Attention
Last Lecture Speed Recap: Attention is “All” You Need

Encoder

- Embedding Layer
- Hidden States
- Stacked Transformer Blocks
- Input Text

Decoder

- Output Text
- Embedding Layer
- Concat
- Stacked Transformer Blocks
- Linear
- Softmax
- Next Token Odds
How do we go from purpose driven models to LLMs?

https://github.com/Mooler0410/LLMsPracticalGuide
How do we go from purpose driven models to LLMs?

**Self-Supervised Learning**
How do we most effectively turn raw text into meaningful loss?

https://github.com/Mooler0410/LLMsPracticalGuide
How do we go from purpose driven models to LLMs?

**Self-Supervised Learning**
How do we most effectively turn raw text into meaningful loss?

**Covered Today**
- Encoder Only
- Decoder Only
- Encoder-Decoder

[https://github.com/Mooler0410/LLMsPracticalGuide](https://github.com/Mooler0410/LLMsPracticalGuide)
How do we go from purpose driven models to LLMs?

Data Scaling
How do we source and train on high-quality data at scale?

https://github.com/Mooler0410/LLMsPracticalGuide
How do we go from purpose driven models to LLMs?

Data Scaling
How do we source and train on high-quality data at scale?

Covered Today
- Data Curation Over Time
- Distributed Training

https://github.com/Mooler0410/LLMsPracticalGuide
LLM Advancements have been driven primarily by these two:

- **Self-Supervised Learning**
  How do we most effectively turn raw text into meaningful loss?

- **Data Scaling**
  How do we source and train on high-quality data at scale?
SSL | From raw text to loss!

Input | Masking
---|---
Hello | 1
World | 0
! | 1
[PAD] | 1

Masked Language Model

Devlin et al. 2018 (BERT)
What is the “Mask” in a Masked Language Model?

Masked Language Model

Devlin et al. 2018 (BERT)
What is the “Mask” in a Masked Language Model?

Recall

Similarities: \( E = \frac{QXT}{\sqrt{DQ}} \)

Attention Matrix: \( A = \text{softmax}(E, \text{dim}=1) \)

Output vectors: \( Y = AX \)

\[ Y_i = \sum_j A_{i,j} X \]

Masked Language Model

Devlin et al. 2018 (BERT)
SSL | What is the “Mask” in a Masked Language Model?

Masked Attention

Similarities: \( E = \frac{QXT}{\sqrt{DQ}} \times \text{MASK} \)
Attention Matrix: \( A = \text{softmax}(E, \text{dim}=1) \)
Output vectors: \( Y = AX \)
\( Y_i = \sum_j A_{i,j} X \)

Masked Language Model

Devlin et al. 2018 (BERT)
What is the “Mask” in a Masked Language Model?

**Intuition**

If $\text{MASK}_i = 0$, then $\mathbf{Y}_i = \sum_{j, j \neq i} A_{i,j} \mathbf{X}$

a.k.a the representation of the masked token is created purely from context

masked_language_model_diagram.png

**Masked Language Model**

Devlin et al. 2018 (BERT)
SSL | Masked Token Prediction

Hello
[MASK]
!
[PAD]

Masked Language Model

Devlin et al. 2018 (BERT)
SSL | Masked Token Prediction

Input: Hello [MASK] ! [PAD]

Masking: 1 0 1 1

Transformer: Encoder

Mask Prediction: World

Masked Language Model

Devlin et al. 2018 (BERT)
SSL | Masked Token Prediction

Optimize Negative Log Likelihood

\[ \text{loss} = -\log(P(\text{"World"} \mid \text{Context}) \]
Optimize Negative Log Likelihood

\[ \text{loss} = -\log(P(\text{"World"} \mid \text{Context}) \]

Equivalent to the Cross-Entropy Loss term from Lecture 3!
Side Note | Tokens v.s. Words

Languages have a lot of words!

If $V =$ Number of Words:
$O(V)$ Memory Scaling
Side Note | Tokens v.s. Words

Languages have a lot of words!

If $V =$ Number of Words:
- $O(V)$ Memory Scaling
- $O(V)$ Runtime Scaling
Languages have a lot of words!
If $V = \text{Number of Words}$:
- $O(V)$ Memory Scaling
- $O(V)$ Runtime Scaling

This limits our vocabulary size a lot.

Tokenizers:
Pre-processing to split words into smaller chunks called “Tokens” so that we can cover all words with smaller $V$
Side Note | Tokens v.s. Words

Languages have a lot of words!
If $V = \text{Number of Words}$:
- $O(V)$ Memory Scaling
- $O(V)$ Runtime Scaling

This limits our vocabulary size a lot.

Tokenizers:
- Pre-processing to split words into smaller chunks called "Tokens" so that we can cover all words with smaller $V$

Important but outside of Course Scope

HuggingFace Tokenizer Summary
Data | BERT used existing curation!

**BERT Corpus**
English Wikipedia + BooksCorpus

**Size**
~3 Billion Tokens

**Quality**
High quality text, Broad “Academic” Knowledge, Limited Diversity

Devlin et al. 2018 (BERT)
Applications | Encoders as “Foundation” Language Models

[CLS] I hate you

[CLS] Hidden State
MLP Layer
Softmax
Sentiment Label

Devlin et al. 2018 (BERT)
Applications | Encoders as “Foundation” Language Models

Devlin et al. 2018 (BERT)
Applications | Encoders as “Foundation” Language Models

[Diagram showing the architecture of a pre-trained encoder and its interaction with a transformer to produce [CLS] hidden states for candidate retrieval.]

Held et al. 2021
Masked Language Model

Input  | Masking  | Transformer  | Mask Prediction
---|---|---|---
Hello  | 1 1 0 1  | 1 0 0 1 | World
[MASK] |  |  |  
! |  |  |  
[PAD] |  |  |  

Questions?
SSL | “How does GPT work?”

Radford et al. 2019 (GPT-2)
SSL | Autoregressive Language Modeling

Radford et al. 2019 (GPT-2)
SSL | Autoregressive Language Modeling

Masking

Causal Mask

Hello
1 0 0 0

World
1 1 0 0

!

[PAD]

Radford et al. 2019 (GPT-2)
SSL | Autoregressive Language Modeling

Radford et al. 2019 (GPT-2)
SSL | Autoregressive Language Modeling

Masked Attention Again!

Similarities: \( E = \frac{QXT}{\sqrt{DQ}} \times \text{MASK} \)

Attention Matrix: \( A = \text{softmax}(E, \text{dim}=1) \)

Output vectors: \( Y = AX \)

\( Y_i = \sum_j A_{i,j} X \)

Tokens only affected by preceding tokens

Hi
World
[
PAD]
SSL | First successful GPT Model, Purely Autoregressive

Input Masking Transformer Next Token Prediction

Hello World

World !

! [EOS]

[PAD]

Causal Mask

Decoder

Radford et al. 2019 (GPT-2)
SSL | First successful GPT Model, Purely Autoregressive

Optimize Negative Log Likelihood of Whole Sequence

\[
\text{loss} = -(\log(P(\text{“World”} \mid \text{“Hello”}) + \log(P(\text{“!”} \mid \text{“Hello World”})) + \\
\log(P(\text{“[EOS]”} \mid \text{“Hello World!”})))
\]

Radford et al. 2019 (GPT-2)
Data | Increasing Token Count via Human Curation Heuristics

**GPT-2 Corpus**
All Reddit Outbound links with at least 3 karma

**Size**
~10 Billion Tokens

**Quality**
High quality text, Broad Knowledge, Improved Diversity

<table>
<thead>
<tr>
<th>URL Domain</th>
<th># Docs</th>
<th>% of Total Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>bbc.co.uk</td>
<td>116K</td>
<td>1.50%</td>
</tr>
<tr>
<td>theguardian.com</td>
<td>115K</td>
<td>1.50%</td>
</tr>
<tr>
<td>washingtonpost.com</td>
<td>89K</td>
<td>1.20%</td>
</tr>
<tr>
<td>nytimes.com</td>
<td>88K</td>
<td>1.10%</td>
</tr>
<tr>
<td>reuters.com</td>
<td>79K</td>
<td>1.10%</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
<td>72K</td>
<td>0.96%</td>
</tr>
<tr>
<td>cnn.com</td>
<td>70K</td>
<td>0.93%</td>
</tr>
<tr>
<td>cbc.ca</td>
<td>67K</td>
<td>0.89%</td>
</tr>
<tr>
<td>dailymail.co.uk</td>
<td>58K</td>
<td>0.77%</td>
</tr>
<tr>
<td>go.com</td>
<td>48K</td>
<td>0.63%</td>
</tr>
</tbody>
</table>

Radford et al. 2019 (GPT-2)
Ok, but what should I use?
### SSL | Classification Comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>MNLI</th>
<th>CoLA</th>
<th>SST-2</th>
<th>MRPC</th>
<th>STS-B</th>
<th>QQP</th>
<th>QNLI</th>
<th>RTE</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-2-original</td>
<td>85.9/85.6</td>
<td>54.8</td>
<td>94.5</td>
<td>86.9/82.2</td>
<td>86.3/85.2</td>
<td>72.5/89.3</td>
<td>91.2</td>
<td>69.8</td>
<td>80.9</td>
</tr>
<tr>
<td>GPT-2-finetuned</td>
<td>85.8/85.5</td>
<td>40.9</td>
<td>94.5</td>
<td>87.0/81.0</td>
<td>85.6/84.3</td>
<td>71.4/88.5</td>
<td>91.5</td>
<td>69.0</td>
<td>78.8</td>
</tr>
<tr>
<td>RoBERTa-large</td>
<td>90.1/89.7</td>
<td>63.8</td>
<td>96.1</td>
<td>91.2/88.3</td>
<td>90.9/90.7</td>
<td>72.5/89.6</td>
<td>94.5</td>
<td>85.9</td>
<td>86.5</td>
</tr>
</tbody>
</table>

*He et al. 2021*
SSL | Pretrained Retrieval Comparison

GPT-2 separates into two clusters

Bert consists of multiple small clusters

https://bert-vs-gpt2.dbvis.de/
SSL | Generative Comparison

Encoders can’t generate!
SSL | Encoder-Only vs. Decoder-Only

**Encoder**
- + Retrieval
- + Classification
- - No Generative Abilities

**Decoder**
- + Generative Abilities
- - Retrieval
- - Classification

Wang et al. 2022
## SSL | Encoder-Only vs. Decoder-Only

<table>
<thead>
<tr>
<th>Encoder</th>
<th>Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ Retrieval</td>
<td>+ Generative Abilities</td>
</tr>
<tr>
<td>+ Classification</td>
<td></td>
</tr>
<tr>
<td>- No Generative Abilities</td>
<td>- Retrieval</td>
</tr>
<tr>
<td></td>
<td>- Classification</td>
</tr>
</tbody>
</table>

This is pretty essential!

Wang et al. 2022
Questions?

Input  Masking  Transformer  Next Token Prediction

Hello  World  !  [EOS]

Causal Mask

Decoder

Autoregressive Language Model
SSL | Encoder-Only vs. Decoder-Only

**Encoder**
- + Retrieval
- + Classification
- - No Generative Abilities

**Decoder**
- + Generative Abilities
- - Retrieval
- - Classification

How to keep this?

This is pretty essential!
SSL | Encoder-Decoder Returns

Encoder

- Hidden States
- Stacked Transformer Blocks
- Embedding Layer
- Input Text

Decoder

- Next Token Odds
- Softmax
- Linear
- Stacked Transformer Blocks
- Concat
- Embedding Layer
- Output Text
SSL | Universal Text-to-Text

Text

Paris
is
the
capitol
of
france
!

Noising

Raffel et al. 2019
SSL | Universal Text-to-Text

Original text:
Thank you for inviting me to your party last week.

Inputs:
Thank you <X> me to your party <Y> week.

Targets:
<X> for inviting <Y> last <Z>

Raffel et al. 2019
Paris is the capital of France!

Raffel et al. 2019
SSL | Universal Text-to-Text

Text:
- Paris
- is
- the
- capitol
- of
- france
- !

Noising:
- Token Masking
- Sentence Permutation
- Document Rotation
- Token Deletion
- Text Infilling

Lewis et al. 2020
Paris is the capital of France!

Tay et al. 2023
Regardless of noise, Loss Function remains the same still!

Continue using Negative Log Likelihood

\[ \text{loss} = - \log(P(\text{Denoised Sequence} \mid \text{Noised Sequence})) \]
SSL | Universal Text-to-Text Is Architecture Agnostic

Tay et al. 2023
Paris is the capital of France!
Moving to truly Large Language Models

Today’s LLMs are driven data and model scaling

Kaplan et al. 2020
We could get a lot more data from CommonCrawl!
Data | Moving to truly Large Language Models

We could get a lot more data from CommonCrawl!
A lot of it is spam though...
We could get a lot more data from CommonCrawl! A lot of it is spam though...
How do we get “useful” data?
Data | C4 - First Scaling of Data Via Common Crawl

**T5 Corpus (AKA C4)**
All Common Crawl Text Which Meets Heuristics

**Size**
~350 Billion Tokens

**Quality**
Varying quality text, Broad Knowledge, Improved Diversity

- We only retained lines that ended in a terminal punctuation mark (i.e. a period, exclamation mark, question mark, or end quotation mark).
- We discarded any page with fewer than 3 sentences and only retained lines that contained at least 5 words.
- We removed any page that contained any word on the “List of Dirty, Naughty, Obscene or Otherwise Bad Words”.  
- Many of the scraped pages contained warnings stating that Javascript should be enabled so we removed any line with the word Javascript.
- Some pages had placeholder “lorem ipsum” text; we removed any page where the phrase “lorem ipsum” appeared.
- Some pages inadvertently contained code. Since the curly bracket “{” appears in many programming languages (such as Javascript, widely used on the web) but not in natural text, we removed any pages that contained a curly bracket.
- Since some of the scraped pages were sourced from Wikipedia and had citation markers (e.g. [1], [citation needed], etc.), we removed any such markers.
- Many pages had boilerplate policy notices, so we removed any lines containing the strings “terms of use”, “privacy policy”, “cookie policy”, “uses cookies”, “use of cookies”, or “use cookies”.
- To deduplicate the data set, we discarded all but one of any three-sentence span occurring more than once in the data set.

Raffel et al. 2019
Data | GPT-3 - Increased Scaling Via Curation

Training

Bag-Of-Words Classifier

Distinguish High and Low Quality

Low-Quality, High Volume

High Quality, Medium Volume

<table>
<thead>
<tr>
<th>URL Domain</th>
<th># Docs</th>
<th>% of Total Docs</th>
</tr>
</thead>
<tbody>
<tr>
<td>bbc.co.uk</td>
<td>116K</td>
<td>1.50%</td>
</tr>
<tr>
<td>theguardian.com</td>
<td>115K</td>
<td>1.50%</td>
</tr>
<tr>
<td>washingtonpost.com</td>
<td>89K</td>
<td>1.20%</td>
</tr>
<tr>
<td>nytimes.com</td>
<td>88K</td>
<td>1.10%</td>
</tr>
<tr>
<td>reuters.com</td>
<td>79K</td>
<td>1.10%</td>
</tr>
<tr>
<td>huffingtonpost.com</td>
<td>72K</td>
<td>0.96%</td>
</tr>
<tr>
<td>cmm.com</td>
<td>70K</td>
<td>0.93%</td>
</tr>
<tr>
<td>cbc.ca</td>
<td>67K</td>
<td>0.89%</td>
</tr>
<tr>
<td>dailymail.co.uk</td>
<td>58K</td>
<td>0.77%</td>
</tr>
<tr>
<td>go.com</td>
<td>48K</td>
<td>0.63%</td>
</tr>
</tbody>
</table>

Brown et al. 2020
Data | GPT-3 - Increased Scaling Via Curation

Filtering

Bag-Of-Words Classifier

Keep “False” Positives

“False” positive \( \sim \) High Quality

Brown et al. 2020
Data | GPT-2 to Original GPT-3 was mostly data scaling

**GPT-3 Corpus**
Common-Crawl Filtered using GPT-2 Training Data

**Size**
~400 Billion Tokens

**Quality**
High-ish quality text, Broad Knowledge, Web-scale Diversity

Brown et al. 2020
Recent Open Source models focus heavily on data scaling.

### Llama 1 Corpus

**Size**
~1.4 Trillion Tokens

**Quality**
Varying quality text, Broad Knowledge, Web-scale Diversity, Includes Code!

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Sampling prop.</th>
<th>Epochs</th>
<th>Disk size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonCrawl1</td>
<td>67.0%</td>
<td>1.10</td>
<td>3.3 TB</td>
</tr>
<tr>
<td>C4</td>
<td>15.0%</td>
<td>1.06</td>
<td>783 GB</td>
</tr>
<tr>
<td>Github</td>
<td>4.5%</td>
<td>0.64</td>
<td>328 GB</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>4.5%</td>
<td>2.45</td>
<td>83 GB</td>
</tr>
<tr>
<td>Books</td>
<td>4.5%</td>
<td>2.23</td>
<td>85 GB</td>
</tr>
<tr>
<td>ArXiv</td>
<td>2.5%</td>
<td>1.06</td>
<td>92 GB</td>
</tr>
<tr>
<td>StackExchange</td>
<td>2.0%</td>
<td>1.03</td>
<td>78 GB</td>
</tr>
</tbody>
</table>

Touvron et al. 2023
**Data** Recent Open Source models focus heavily on data scaling

**Falcon Refined Web Corpus**

**Size**
5 Trillion Tokens

**Quality**
Varying quality text, Broad Knowledge, Web-scale Diversity, Includes Code

Brown et al. 2020
Data Mixture has become the biggest “secret”

- **Llama 2 Corpus**
  - Size: > 2 Trillion Tokens
  - Quality: Minimal details known

- **PALM-2 Corpus**
  - Size: > 3.6 Trillion Tokens
  - Quality: No details known

- **GPT-4 Corpus**
  - Size: Unknown (Est. 11T Tokens)
  - Quality: No details known

---

Touvron et al. 2023 (b)  
Anil et al. 2023  
OpenAI 2023
# Questions?

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Size</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Llama 2 Corpus</td>
<td>&gt; 2 Trillion Tokens</td>
<td>Minimal details known</td>
</tr>
<tr>
<td>PALM-2 Corpus</td>
<td>&gt; 3.6 Trillion Tokens</td>
<td>No details known</td>
</tr>
<tr>
<td>GPT-4 Corpus</td>
<td>Unknown (Est. 11T Tokens)</td>
<td>No details known</td>
</tr>
</tbody>
</table>

- **Llama 2 Corpus**
  - Size: > 2 Trillion Tokens
  - Quality: Minimal details known

- **PALM-2 Corpus**
  - Size: > 3.6 Trillion Tokens
  - Quality: No details known

- **GPT-4 Corpus**
  - Size: Unknown (Est. 11T Tokens)
  - Quality: No details known

---

Touvron et al. 2023 (b)  
Anil et al. 2023  
OpenAI 2023
Scaling Parameters | Data Parallel Training

https://engineering.fb.com/2021/07/15/open-source/fsdp/
Scaling Parameters | Data Parallel Training

Total memory increases linearly with shards
Scaling Parameters | Data Parallel Training

Max memory constrains model size
Scaling Parameters | *Fully* Sharded Data Parallel Training

https://engineering.fb.com/2021/07/15/open-source/fsdp/
Scaling Parameters | *Fully* Sharded Data Parallel Training

Total memory is constant
Scaling Parameters | *Fully* Sharded Data Parallel Training

Max single GPU memory constrains layer size
Scaling Parameters | Tensor Parallel Training

https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism
Scaling Parameters | Tensor Parallel Training

https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism
Scaling Parameters | Tensor Parallel Training

Don’t need to sync gradients!

https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism
Scaling Parameters | Tensor Parallel Training

Don’t need to sync gradients!
Max GPU memory constrains a layer shard

https://huggingface.co/docs/transformers/v4.15.0/parallelism#tensor-parallelism
Scaling Parameters | FSDP + TP = ~Limitless Scaling

1 Trillion Parameter Model with Tensor Parallelism and FSDP
Final Questions?

Fill out my anonymous feedback form