Large Language Models

Michael Wollowski

Excerpts of Chapter 10 from
Speech and Language Processing,
Jurafsky and Martin, Aug. 20, 2024 draft

1

Training corpora for LLMs

- Web text is usually taken from corpora of automatically-crawled web pages like the *common crawl*.
- It is a series of snapshots of the entire web produced by the non-profit Common Crawl that each have billions of webpages.
- Various cleanups of common crawl data exist.
- One is *Colossal Clean Crawled Corpus* (C4)
- It is a corpus of 156 billion tokens of English that is filtered in various ways.
- Filtering includes:
  - Removing duplicated data,
  - removing non-natural language like code,
  - sentences with offensive words from a blocklist.

2

Training corpora for LLMs

- Large language models are mainly trained on text scraped from the web, augmented by more carefully curated data.
- Since those training corpora are so large, they are likely to contain many natural examples that can be helpful for NLP tasks:
  - question and answer pairs (for example from FAQ lists),
  - translations of sentences between various languages,
  - documents together with their summaries, and so on.

3

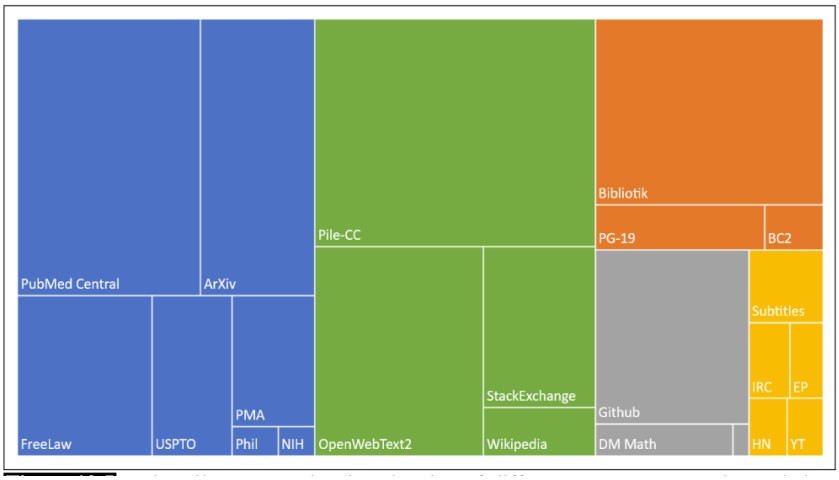
The Pile

- This C4 corpus seems to consist in large part of patent text documents, Wikipedia, and news sites
- Wikipedia plays a role in lots of language model training, as do corpora of books and code.
- The Pile contains much more varied data.

4

The Pile

- Colors:
  - Academic
  - Internet
  - Prose
  - Dialoge
  - Misc



5

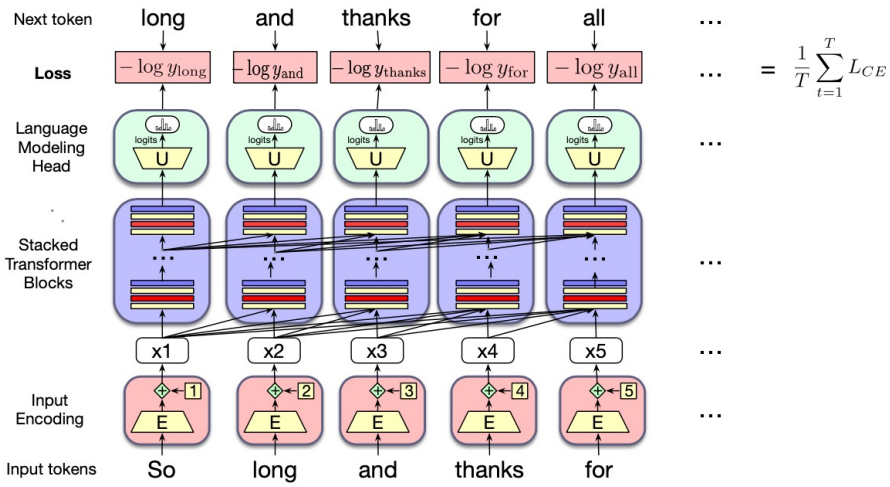
Self-supervised training algorithm

- Transformers are trained on a corpus of text.
- At each time step *t*, we ask the model to predict the next word.
- We call such a model *self-supervised,* because the natural sequence of words is its own supervision.
- We simply train the model to minimize the error in predicting the true next word in the training sequence.
- During training, the probability assigned to the correct word is used to calculate the loss for each item in the sequence.
- The weights in the network are adjusted to minimize the average loss over the training sequence via gradient descent.

6

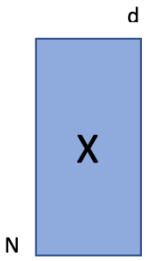
Self-supervised training algorithm for Transformers

- At each step, given all the preceding words, the final transformer layer produces an output distribution over the entire vocabulary.



7

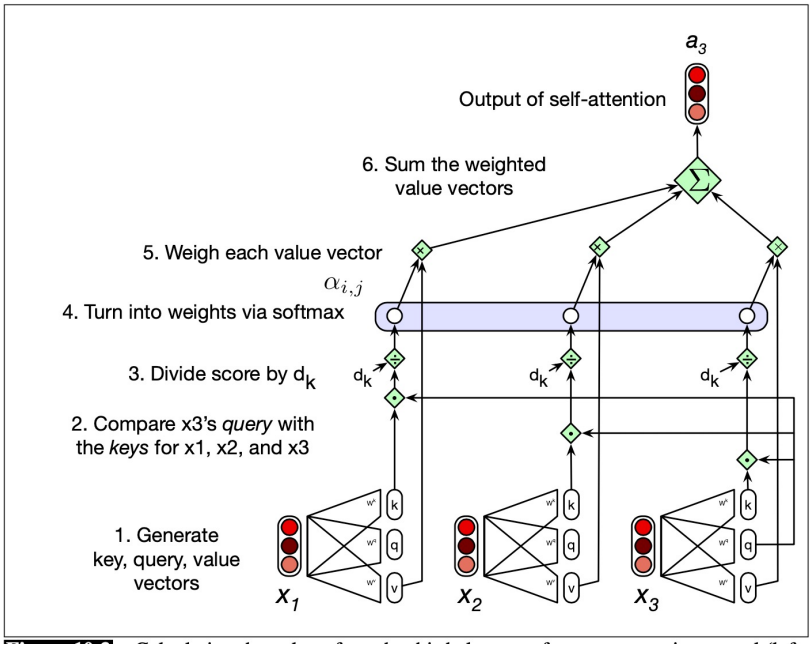
Parallelizing Self-Attention



- So far, we computed a single output at a single time step *i*.
- Each output, y_i, is computed independently.
- The calculation can be parallelized.
- We pack the input embeddings of the *N* tokens of the input sequence into a **single** matrix $\mathbf{X} \in R^{N \times d}$
- Each row of **X** is the embedding of **one** token of the input.
- Transformers for large language models can have an input length *N* = 1024, 2048, or 4096 tokens.
- **X** has between 1K and 4K rows, each of the dimensionality of the embedding *d*.

8

Reminder



9

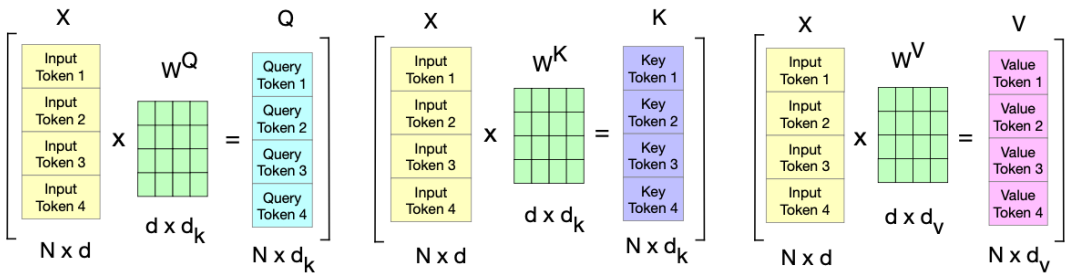
Parallelizing Self-Attention

- We multiply **X** by the key, query, and value matrices.
- They all are of size *d* x *d*.
- This produces matrices $Q \in R^{N \times d}$, $K \in R^{N \times d}$, and $V \in R^{N \times d}$
- And the query, key, and value vectors:

  $Q=XW^Q$; $K=XW^K$; $V=XW^V$

10

Computation of the Q, K, and V matrices.



11

Masking out the Future

- We can compute all the query-key comparisons simultaneously by multiplying Q and K^T in a single matrix multiplication.
- However, the calculation in QK^T results in a score for each query value to every key value, *including those that follow the query*.
- This is inappropriate in the setting of language modeling: guessing the next word is pretty simple if you already know it!
- Hence, the upper-triangle portion of the comparisons matrix set to $-\infty$.
- Softmax will turn them into zeros

N

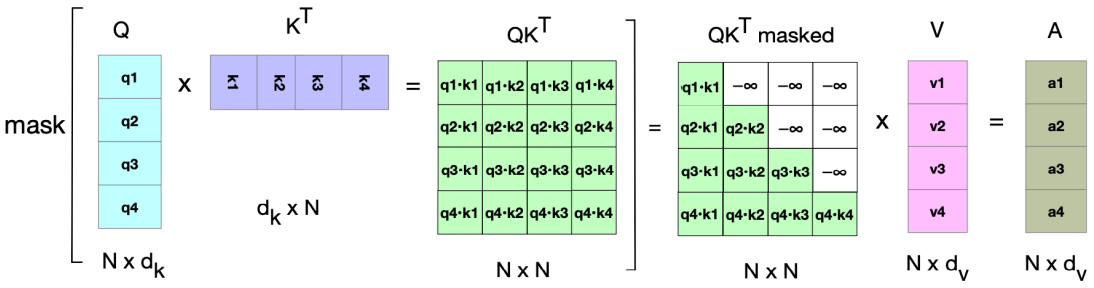
q1•k1	−∞	−∞	−∞	−∞
q2•k1	q2•k2	−∞	−∞	−∞
q3•k1	q3•k2	q3•k3	−∞	−∞
q4•k1	q4•k2	q4•k3	q4•k4	−∞
q5•k1	q5•k2	q5•k3	q5•k4	q5•k5

N

Image source: Speech and Language Processing, Jursafky and Martin, Jan. 12, 2022 draft

12

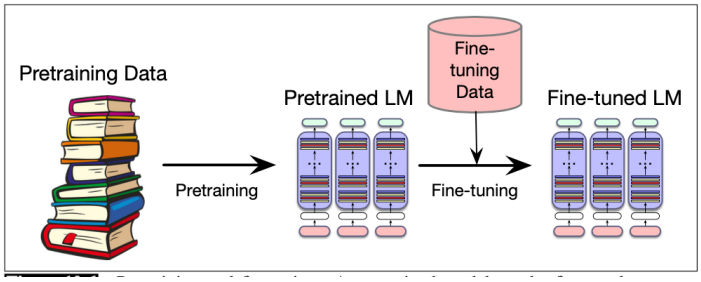
Computation of QK^T



<---------- This happens before softmax. ---------- >

13

Finetuning



- After an LLM has been trained on the large corpus, it can be used outright, such as for Claude.ai.
- However, the general nature of C4 or the Pile may not have sufficient data to apply an LLM in a specific domain or task.
- Example: a language model that's specialized to medical text
- In such a case, we can continue training the model on relevant data from the new domain or language.

14