Transformers – Part 1

Summary of Chapter 10 from
Speech and Language Processing,
Jurafsky and Martin, August 20, 2024 draft
Michael Wollowski

1

Next Word Prediction

- It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a …

2

Next Word Prediction

- It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a **wife**.
  - Jane Austen: Pride and Prejudice
- In my younger and more vulnerable years my father gave me some advice that I've been turning over in my mind ever …

3

Next Word Prediction

- It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a **wife**.
  - Jane Austen: Pride and Prejudice
- In my younger and more vulnerable years my father gave me some advice that I've been turning over in my mind ever **since**.
  - F. Scott Fitzgerald, *The Great Gatsby*
- All this happened, more or …

4

Next Word Prediction

- It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a **wife**.
  - Jane Austen: Pride and Prejudice
- In my younger and more vulnerable years my father gave me some advice that I've been turning over in my mind ever **since**.
  - F. Scott Fitzgerald, *The Great Gatsby*
- All this happened, more or **less**.
  - Kurt Vonnegut*, Slaughterhouse-Five*

5

Transformers: The Basics

- The transformer is the standard architecture for building large language models.
- Left-to-right (autoregressive) language modeling:
  - Given a sequence of input tokens,
  - Predict output tokens one by one,
  - Conditioned on the prior context.
- Key component of a transformer:
  - self-attention also called multi-head attention.

6

Quick Review of Attention

- Build contextual representations of a token's meaning.
- Attending to and integrating information from surrounding tokens.
- Helping the model learn how tokens relate to each other over large spans.

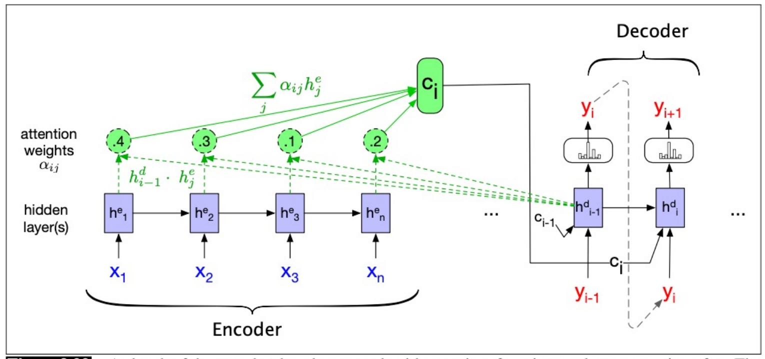


Figure 9.22 A sketch of the encoder-decoder network with attention, focusing on the computation of $\mathbf{c}_i$. The context value $\mathbf{c}_i$ is one of the inputs to the computation of $\mathbf{h}_i^d$. It is computed by taking the weighted sum of all the encoder hidden states, each weighted by their dot product with the prior decoder hidden state $\mathbf{h}_{i-1}^d$.

7

Transformers: The Basic Architecture

- Unlike an RNN, a transformer processes several tokens at once.
- This called the context window.
- The basic unit of a transformer is a block.
- A block processes the entire input sequence.
- Blocks are "stacked" i.e. they run in sequence.

8

Basic View of Transformer Architecture

- Sequences of text as I/O
- An encoder-decoder

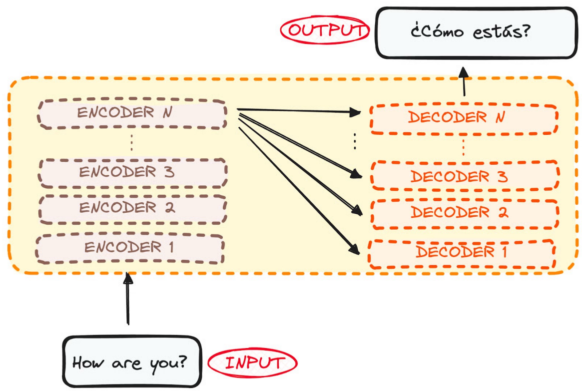


Image source: https://www.datacamp.com/tutorial/how-transformers-work

9

Transformer Blocks

- Each block is a multilayer network, consisting of:
  - a multi-head attention layer,
  - feedforward networks and
  - layer normalization steps.
- Lot's of weights!
- We will investigate those in detail.

10

More detailed view of Transformer Architecture

- A bit complex
- But! Lots of repetition.

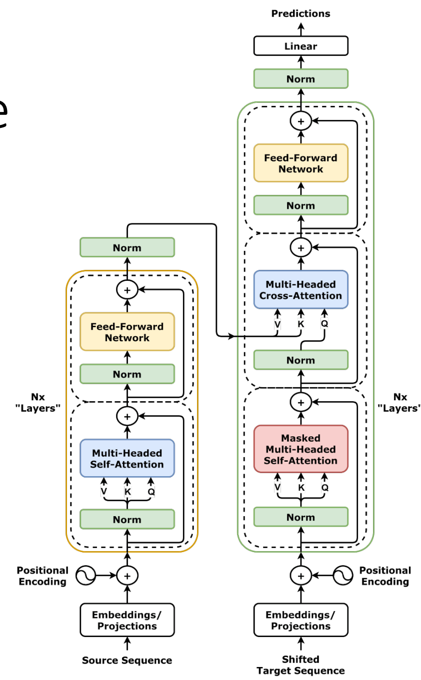


Image source: Wikipedia entry on transformers.

11

Transformers: The Basics

- **Input encoding** though embedding matrix E
- **Language modeling head** through unembedding matrix U.
- Number of stacked blocks: 12 to 96.
- GPT-4: 120 blocks

12

GPT-4

- Standard GPT-4 model offers 8,000 tokens for the context*).
- 8000 tokens amount to about 26 pages of a novel**).

*) Source: Maximum Token length in GPT-4. https://community.openai.com/t/maximum-token-length-in-gpt-4/385914
**) Assuming 250-300 words per book page. Source: https://hotghostwriter.com/blogs/blog/novel-length-how-long-is-long-enough It should be noted that the token count is typically larger than the word count.

13

GPT-4

- An extended 32,000 token context-length model is available*).
- 32000 tokens amount to about 106 pages of a novel**).
- Suddenly, next word prediction does not seem to be such a hard problem any longer.

*) Source: Maximum Token length in GPT-4. https://community.openai.com/t/maximum-token-length-in-gpt-4/385914
**) Assuming 250-300 words per book page. Source: https://hotghostwriter.com/blogs/blog/novel-length-how-long-is-long-enough It should be noted that the token count is typically larger than the word count.

14

Attention

- Consider the following examples.
  - The <u>chicken</u> didn't cross the road because **it** was too tired.
  - The chicken didn't cross the <u>road</u> because **it** was too wide.

16

Language and World Knowledge

- Fluent speakers of a language bring an enormous amount of knowledge to bear during comprehension and production.
- This knowledge is embodied in many forms, perhaps most obviously in the vocabulary.
- Most of this growth is not happening through direct vocabulary instruction in school.
- The bulk of this knowledge acquisition happens as a by-product of reading, as part of the rich processing and reasoning that we perform when we read.
- So, read more!

17

Transformers and World Knowledge

- The stacked layers in a transformer: used to build up richer and richer contextualized representations of the words in a sentence.

18

Processing Through the Layers

Early Layers (1-4)

- Focus on **syntactic and surface-level features**
- Detect basic linguistic patterns: punctuation, capitalization, common prefixes/suffixes
- Identify word types, basic grammatical categories
- Handle tokenization artifacts and positional information
- Features are relatively simple and directly interpretable

Lower-Middle Layers (5-12)

- Develop **grammatical and structural features**
- Parse sentence structure, identify parts of speech more sophisticatedly
- Detect phrase boundaries, dependency relationships
- Begin to handle basic semantic relationships (synonyms, antonyms)
- Start forming more complex compositional representations

19

Processing Through the Layers

Middle Layers (13-20)

- Form **semantic and conceptual features**
- Represent entities, relationships, and factual knowledge
- Handle more abstract concepts (emotions, themes, topics)
- Develop features for logical reasoning and inference
- Begin to integrate information across longer contexts

Upper-Middle Layers (21-28)

- Focus on **discourse and pragmatic features**
- Understand conversational context, intent, and tone
- Handle complex reasoning patterns and multi-step inference
- Develop features for different writing styles and genres
- Integrate world knowledge with current context

20

Processing Through the Layers

Final Layers (29+)

- Concentrate on **task-specific and output features**
- Transform representations toward the vocabulary space for prediction
- Handle specific formatting and response generation patterns
- Fine-tune for particular behaviors (helpfulness, safety, etc.)

Key Patterns:

- **Increasing abstraction**: From tokens → words → phrases → concepts → discourse
- **Growing receptive fields**: Later layers integrate information from much larger contexts
- **Task specialization**: Final layers become more specialized for the model's training objectives
- **Residual connections**: Information from all levels can influence final outputs, not just the last layer

21

Example of Processing Through Layers

"The restaurant was not very good."

Layer 2-3: Surface Pattern Detection
- Feature detects the token "not"
- Basic pattern: [word] + "not" + [word]
- No understanding of meaning, just recognizing the negation token

Layer 6-8: Syntactic Structure
- Feature recognizes "not" as a negation modifier
- Understands it syntactically modifies "very good"
- Pattern: negation + intensifier + adjective
- Still largely structural, not semantic

Layer 12-15: Local Semantic Negation
- Feature begins to understand that "not very good" means the opposite of "very good"
- Can flip polarity: positive adjective → negative meaning
- Operates on immediate phrase: "not very good" = negative evaluation

22

Example of Processing Through Layers

"The restaurant was not very good."

Layer 18-22: Contextual Integration
- Feature integrates negation with broader context
- Understands that "restaurant was not very good" is specifically about restaurant quality
- Can handle more complex cases like "not unhappy" (double negation)
- Considers pragmatic implications (understatement, politeness)

Layer 25-28: Discourse-Level Understanding
- Feature understands conversational implications
- "Not very good" in a review context implies disappointment, might suggest looking elsewhere
- Can generate appropriate follow-up responses
- Integrates with knowledge about restaurant reviews, social norms around criticism

Layer 30+: Output Preparation
- Feature helps generate contextually appropriate responses
- If asked "Should I go there?", uses negation understanding to suggest "probably not"
- Formats response with appropriate tone and helpfulness

23

Example of Processing Through Layers

"The restaurant was not very good."

Evolution Pattern:

- **Scope expansion**: From single token → phrase → sentence → discourse
- **Semantic depth**: From pattern matching → meaning reversal → pragmatic implications
- **Context integration**: From local → global understanding
- **Functional specialization**: From detection → reasoning → response generation

24

Back to Chickens Though

- Consider:
  The chicken didn't cross the road because **it** …
- At this point we do not yet know which thing "it" is going to end up referring to.
- A representation of the input must be such that "it" can be resolved to "chicken" or "road."

25

Back to Chickens Though

- The self-attention weight distribution α that is part of the computation of the representation for the word "it" at layer k + 1.
- In computing the representation for it, we attend differently to the various words at layer k.
- Darker shades indicate higher self-attention values.

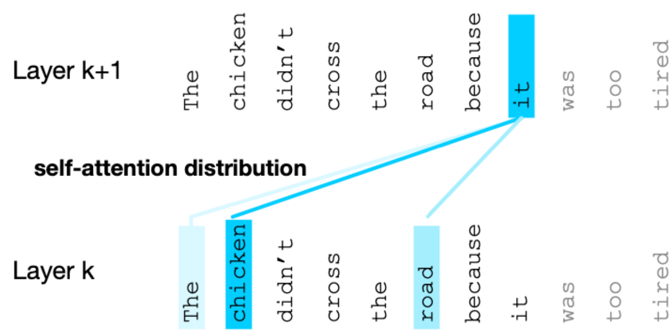
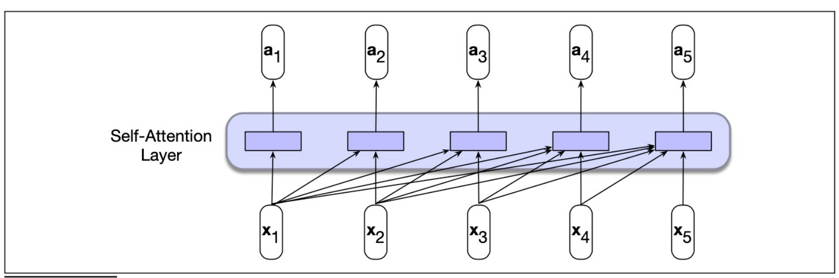


Image source: Speech and Language Processing, Jurafsky and Martin, Aug. 20, 2024 draft

26

Causal or Backward-looking Self-attention

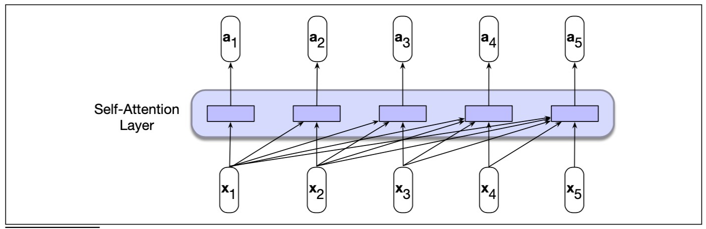


- In causal, or backward looking self-attention, the context is any of the prior words.
- In general bidirectional self-attention, the context can include future words.

Image source: Speech and Language Processing, Jurafsky and Martin, Feb. 3, 2024 draft

27

Self-attention more formally

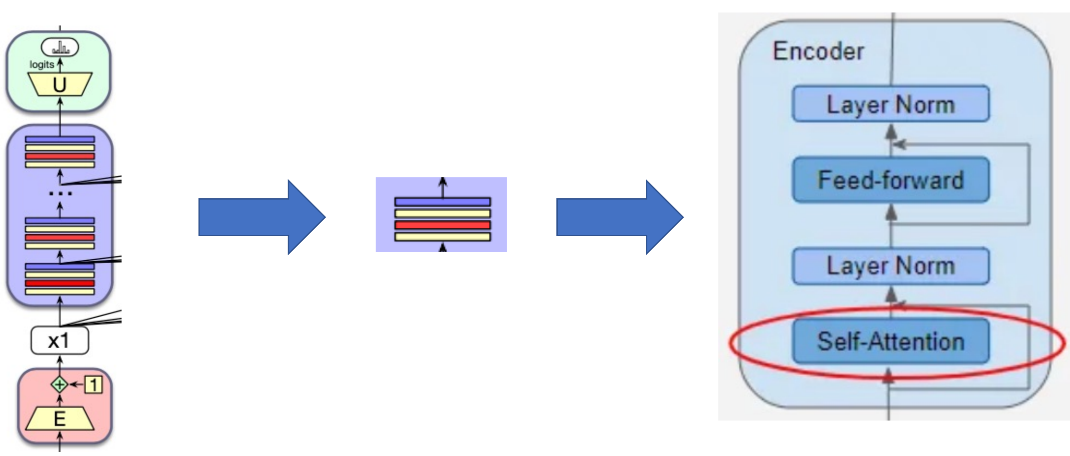


- The core intuition of attention is the idea of *comparing* an item of interest to a collection of other items in a way that reveals their relevance in the current context.
- For example, in the figure the computation of a_3 is based on a set of comparisons between the input x_3 and its preceding elements x_1 and x_2, and to x_3 itself.

Image source: Speech and Language Processing, Jurafsky and Martin, Feb. 3, 2024 draft

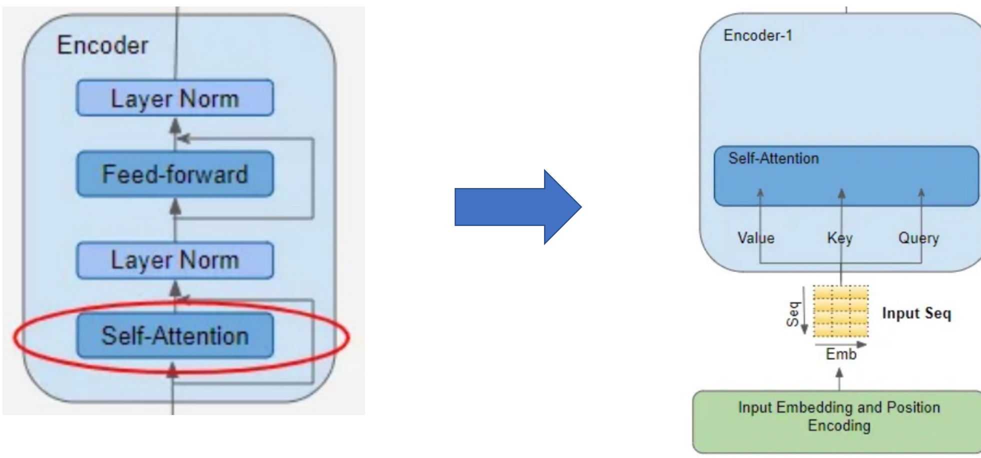
28

Detour: Locating the Attention Head



29

Detour: Locating the Attention Head



30

Less Simplified Version of Attention

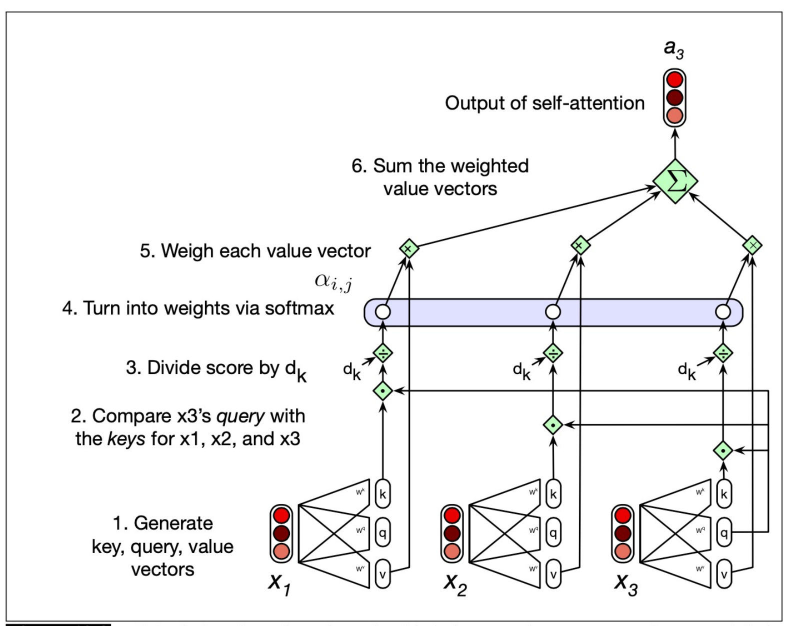


Image source: Speech and Language Processing, Jurafsky and Martin, Feb. 3, 2024 draft

31

Simplified Version of Attention

- We compute similarity scores via dot product, which maps two vectors into a scalar value ranging from $-\infty$ to ∞.
- The larger the score, the more similar the vectors that are being compared.
- We'll normalize these scores with a softmax to create the vector of weights α_{ij}, $j \leq i$.
- Simplified version:

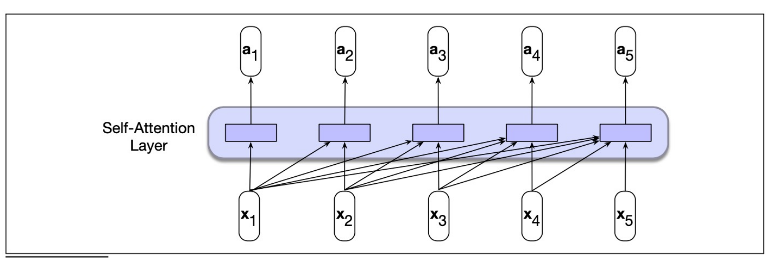
$$\text{score}(x_i, x_j) = x_i \cdot x_j$$

$$\alpha_{ij} = \text{softmax}(\text{score}(x_i, x_j)) \ \forall j \leq i$$

35

Simplified Version of Attention

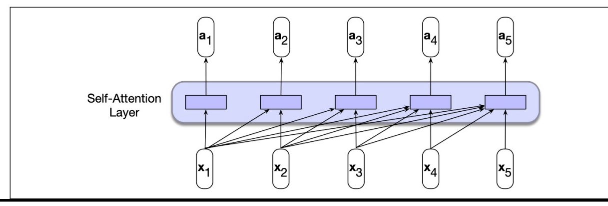
- In the example from the figure, the first step in computing a_3 would be to compute three scores:
  1. $x_3 \cdot x_1$,
  2. $x_3 \cdot x_2$,
  3. $x_3 \cdot x_3$.



36

Simplified Version of Attention

- The resulting values are treated as weights
- They indicate the proportional relevance of the prior token to the current token at position *i.*
- The softmax value will likely be highest for $\mathbf{x}_i$, since it is very similar to itself.
- However, other context words may also be similar to *i*, and softmax will also assign some weight to those words.



37

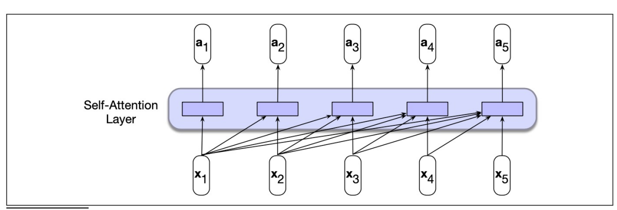
Simplified Version of Attention

- Putting everything together, we get attention $\mathbf{a}_i$:

$$\text{score}(x_i, x_j) = x_i \cdot x_j$$

$$\alpha_{ij} = \text{softmax}(\text{score}(x_i, x_j))\ \forall j \leq i$$

$$\mathbf{a}_i = \sum_{j \leq i} \alpha_{ij} x_j$$



38