Transformer Architecture

Michael Wollowski

1

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.

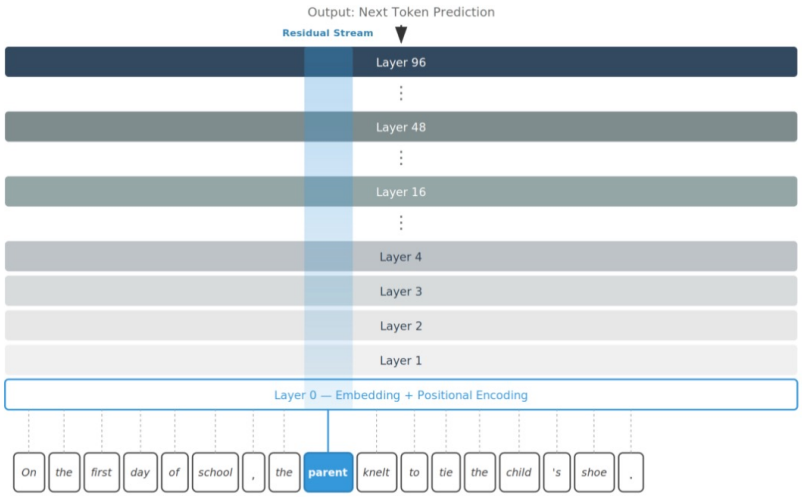
2

Transformers: The Basics

- GPT-3 (175B) has 96 layers of transformer blocks through which data flows sequentially.
- It has space for 2048 input tokens.
- Each token is represented by a 12,288-dimensional vector
- Each token is processed in its own residual stream.

3

Birds-eye view of a Transformer



4

Transformers: The Basics

- The residual streams can be thought of as a whiteboard.
- Information is added to it but never deleted.
- Adding vectors emphasizes or deemphasizes specific information.

5

Transformer Blocks

- Each block is a multilayer network, consisting of:
  - a multi-head attention layer,
  - feedforward network and
  - layer normalization steps.

6

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.

7

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.

8

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.
- Among others in the vocabulary.
- The bulk of this knowledge acquisition happens as a by-product of reading.
- It occurs as part of the rich processing and reasoning that we perform when we read.
- So, read more!

9

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.

10

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

11

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

12

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

13

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

14

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

15

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

16

Chickens!

- Consider:
  The chicken didn't cross the road because **it** …
- The referent for "it" is yet undetermined.
- It could be:
  - The chicken didn't cross the road because **it** was too tired.
  - The chicken didn't cross the road because **it** was too wide.
- A representation of the input must be such that "it" can be resolved to "chicken" or "road."

17

Chickens!

- The self-attention weight distribution 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.

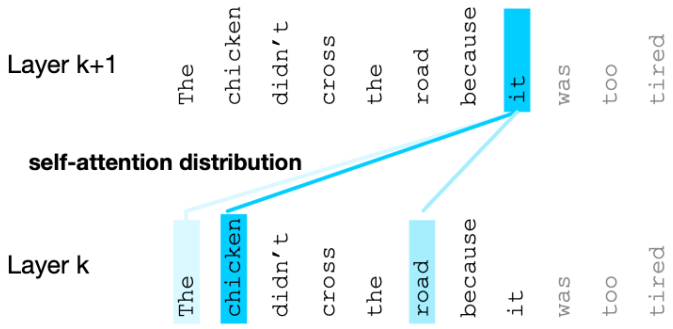
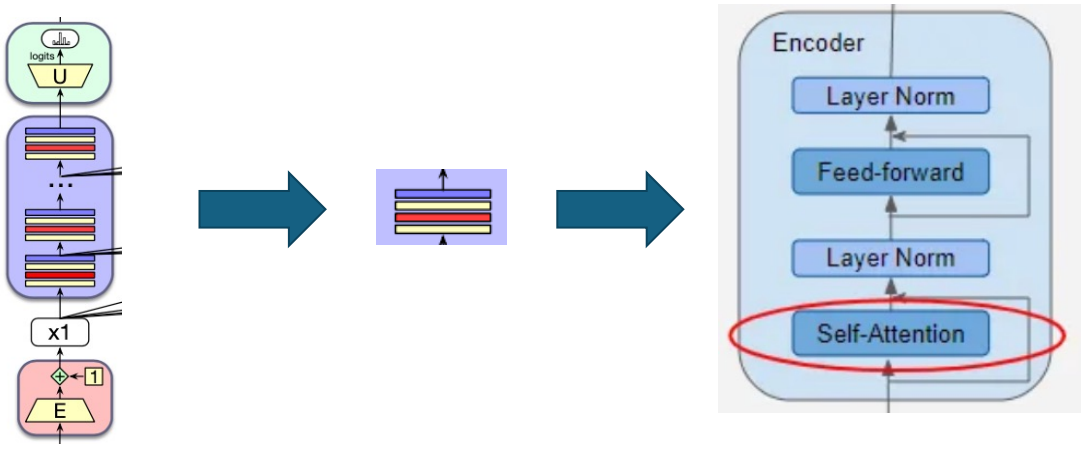


Image source: Speech and Language Processing, Jurafsky and Martin, Aug. 20, 2024 draft

18

Locating the Attention Head



19

Transformer Blocks

- The self-attention calculation lies at the core of what is called a **transformer block**.
- In addition to the self-attention layer, it includes feedforward layers, residual connections, and normalizing layers.

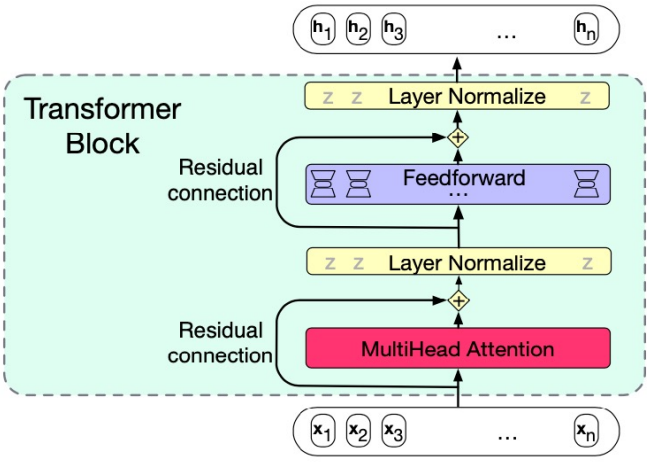
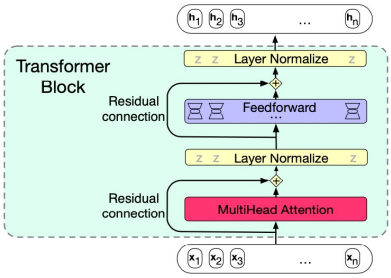


Image source: TBA

20

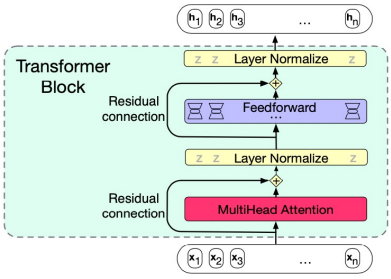
Transformer Blocks



- **Feedforward layer**: It contains *N* position-wise networks, one at each position.

21

Transformer Blocks



- **Residual connections**: They pass information from a lower layer to a higher layer without going through the intermediate layer.
- **Layer normalization (Layer norm)**.
  - Summed vectors are normalized.
  - It is used to improve training performance in deep neural networks.
  - It keeps the values of a hidden layer in a range that facilitates gradient-based training.

22

Residual Stream View of the Transformer

- Residual streams capture the processing of a transformer network.
- Hence, they give the best view of the processing.
- The feedforward layers and the multi-head attention layers are added to a stream.
- The Multi Head Attention layer reads information from the other residual streams.
- The feedforward layer processes just one token.

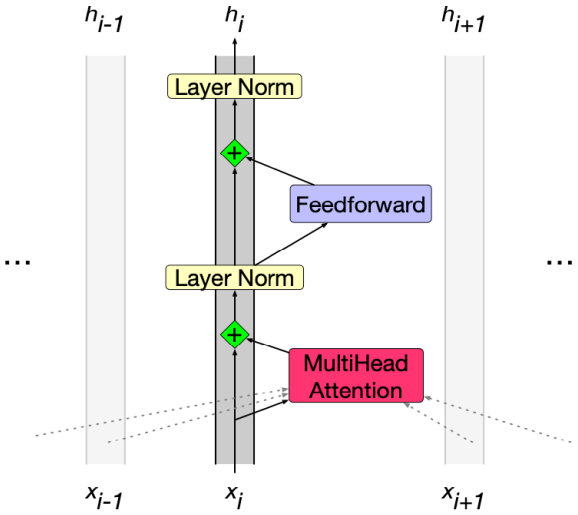
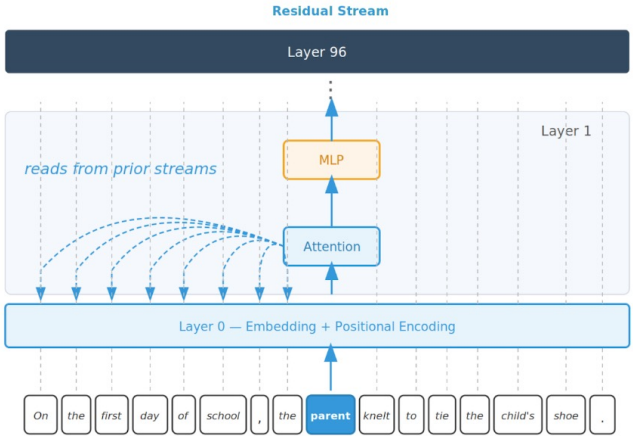


Image source: Chapter 9, Jurafsky and Martin, Aug. 20, 2024.

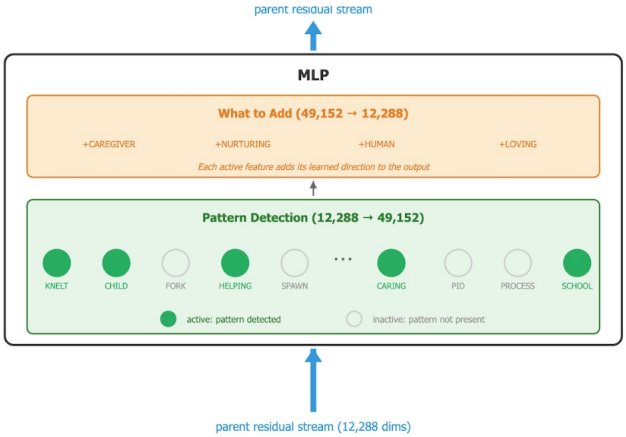
23

Attention Has Access to Prior Streams



24

Feedforward or MLP Portion: Only has access to its own Stream



25

Moving Information

- The attention head for residual stream B can move information from token A's residual stream into its own stream.

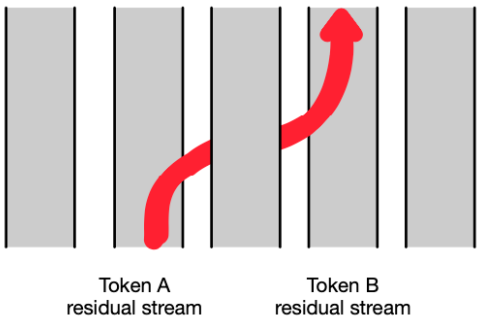


Image source: Chapter 9, Jurafsky and Martin, Aug. 20, 2024.

26