Attention & Transformers

Ivo Verhoeven | Advanced Topics in Computational Semantics

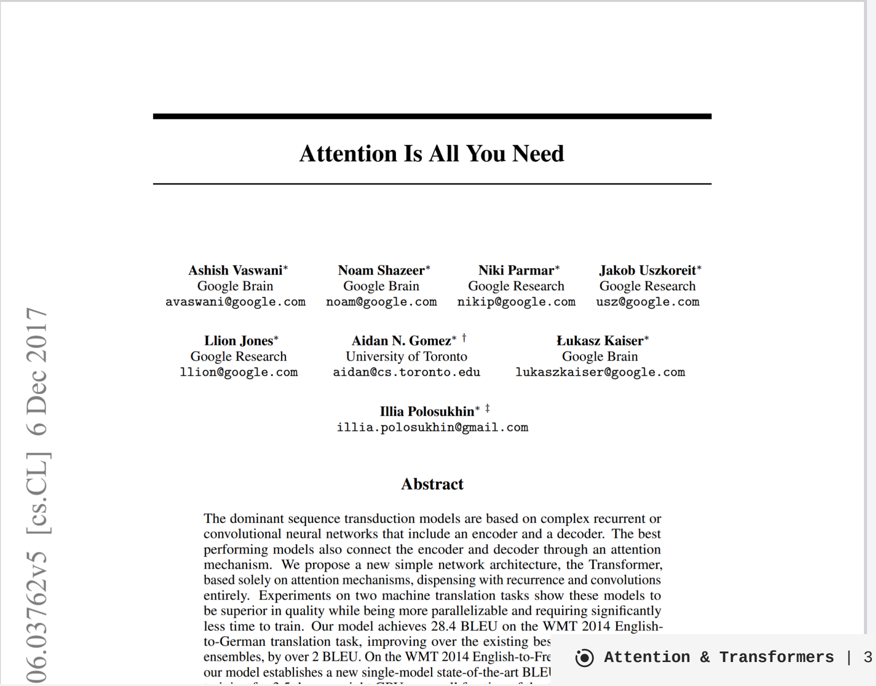
About Me

- 2017 - 2020: BSc. Liberal Arts & Sciences
- 2020 – 2022: MSc. AI at University of Amsterdam
  - Thesis on meta-learning, morphology and NMT
  - Took ATCS in 2021
- 2022 - ???: PhD at ILLC
  - Misinformation detection and generalisation with Katia Shutova

Vaswani et al.: Attention is All You Need

- Introduces the Transformer architecture in late 2017
  - Google Brain/Google Research collab

`Vaswani et al. (2017). Attention is all you need. Advances`
`neural information processing systems, 30.`



06.03762v5 [cs.CL] 6 Dec 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

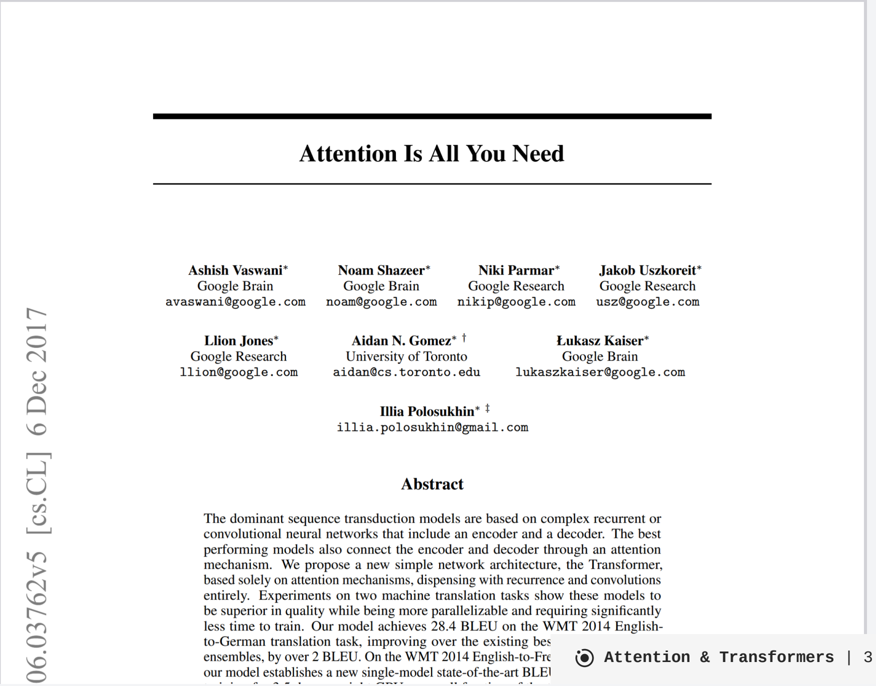
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing bes ensembles, by over 2 BLEU. On the WMT 2014 English-to-Fre our model establishes a new single-model state-of-the-art BLE

Vaswani et al.: Attention is All You Need

- Introduces the Transformer architecture in late 2017
  - Google Brain/Google Research collab

- Paper currently has **169 248** citations
  - Or **~64 citations a day**

`Vaswani et al. (2017). Attention is all you need. Advances`
`neural information processing systems, 30.`



06.03762v5 [cs.CL] 6 Dec 2017

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

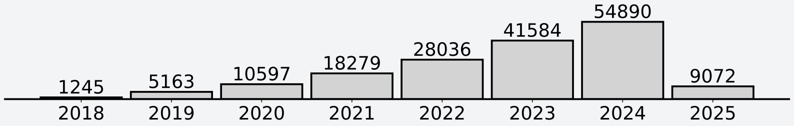
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing bes ensembles, by over 2 BLEU. On the WMT 2014 English-to-Fre our model establishes a new single-model state-of-the-art BLE

Vaswani et al.: Attention is All You Need

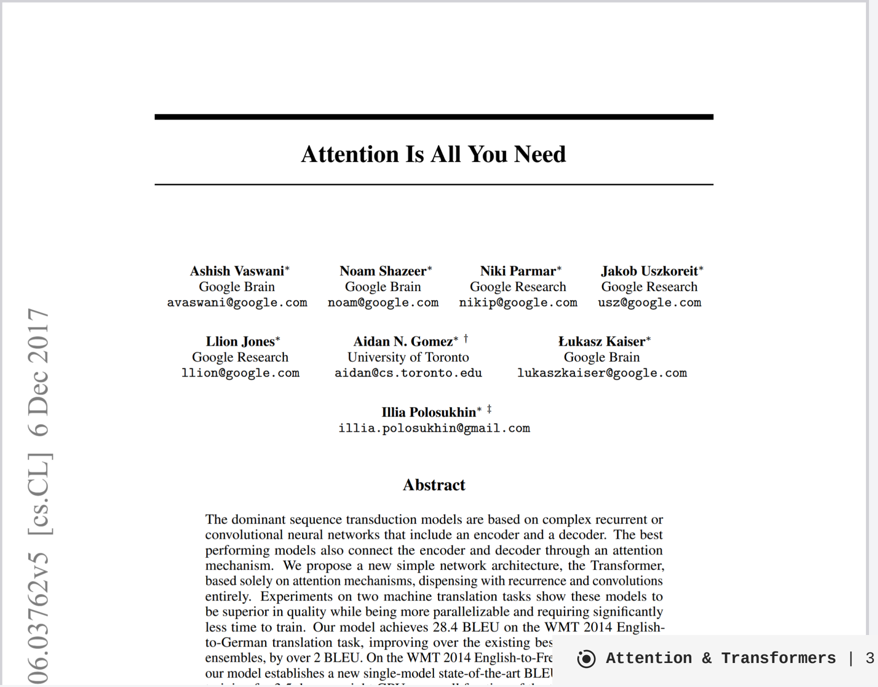
- Introduces the Transformer architecture in late 2017
  - Google Brain/Google Research collab

- Paper currently has **169 248** citations
  - Or **~64 citations a day**

- Number of citations is only accelerating



Vaswani et al. (2017). Attention is all you need. Advances
neural information processing systems, 30.



06.03762v5 [cs.CL] 6 Dec 2017

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

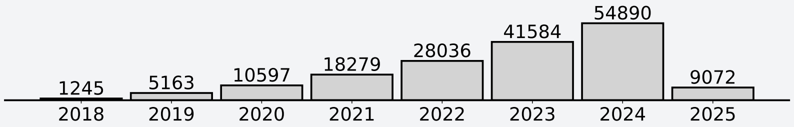
Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing bes ensembles, by over 2 BLEU. On the WMT 2014 English-to-Fre our model establishes a new single-model state-of-the-art BLE

Vaswani et al.: Attention is All You Need

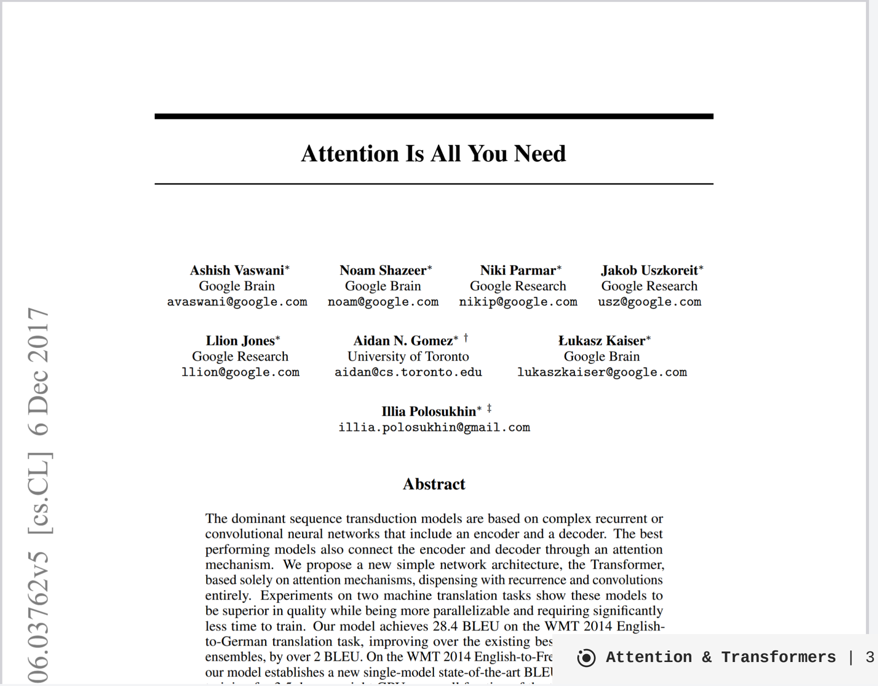
- Introduces the Transformer architecture in late 2017
  - Google Brain/Google Research collab
- Paper currently has **169 248** citations
  - Or **~64 citations a day**
- Number of citations is only accelerating



- Most cited paper ever has **233 829** citations

  `Lowry et al. (1951) Protein measurement with the folin phenol reagent.`

`Vaswani et al. (2017). Attention is all you need. Advances neural information processing systems, 30.`



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

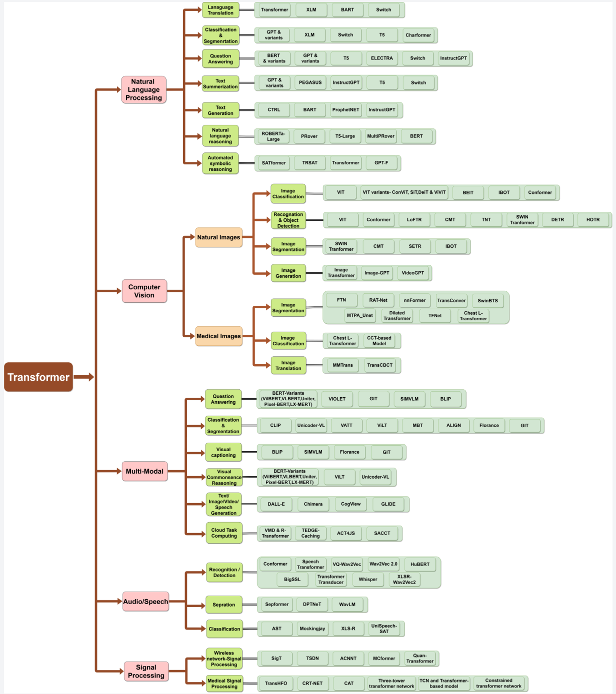
06.03762v5 [cs.CL] 6 Dec 2017

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing bes
ensembles, by over 2 BLEU. On the WMT 2014 English-to-Fre
our model establishes a new single-model state-of-the-art BLE

Vaswani et al.: Attention is All You Need

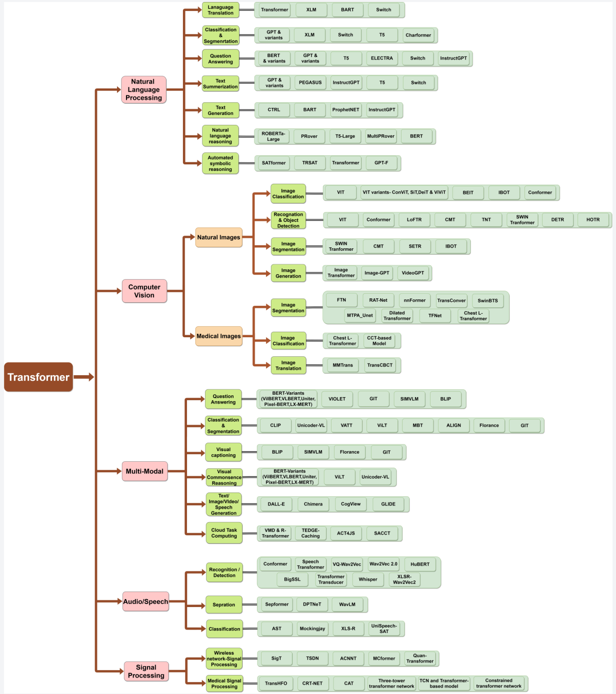
- It’s hard to think of an AI area that hasn’t been affected by the Transformer



Islam, et al. (2023). A Comprehensive Survey on Application Transformers for Deep Learning Tasks. arXiv:2306.07303.

Vaswani et al.: Attention is All You Need

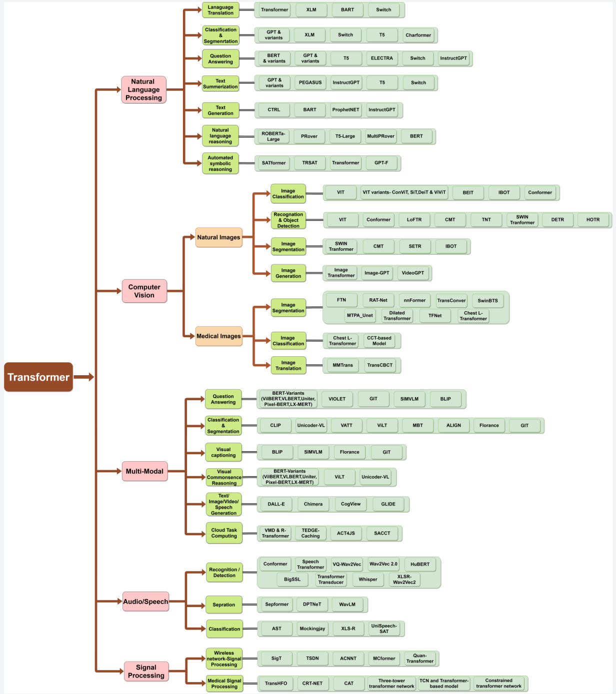
- It's hard to think of an AI area that hasn't been affected by the Transformer
  - **NLP:** Transformer > RNN
    - Seq-to-seq: what it was designed for
    - Classification: encoder-only transformers
    - Generation: decoder-only transformers



Islam, et al. (2023). A Comprehensive Survey on Application
Transformers for Deep Learning Tasks. arXiv:2306.07303.

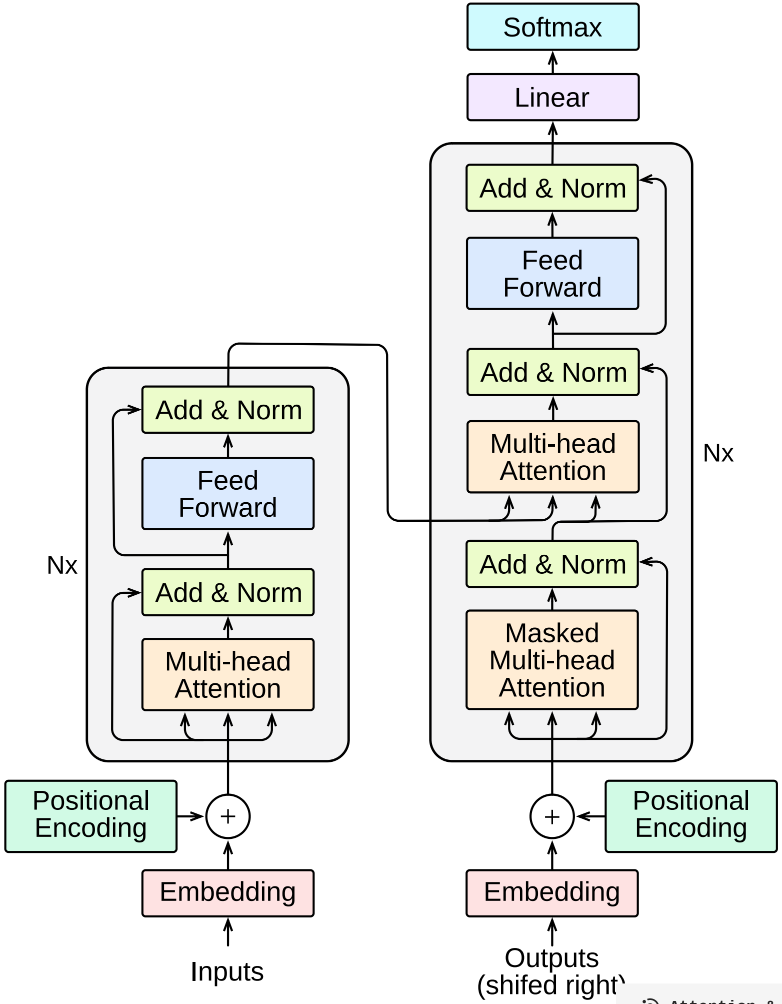
Vaswani et al.: Attention is All You Need

- It’s hard to think of an AI area that hasn’t been affected by the Transformer
  - **NLP:** Transformer > RNN
    - Seq-to-seq: what it was designed for
    - Classification: encoder-only transformers
    - Generation: decoder-only transformers
  - **CV:** ViT > CNN
  - **Multi-modal:** Transformer > different architectures
  - **Speech:** Transformer > CNN
  - **Graphs:** Transformer/Attention > GCN

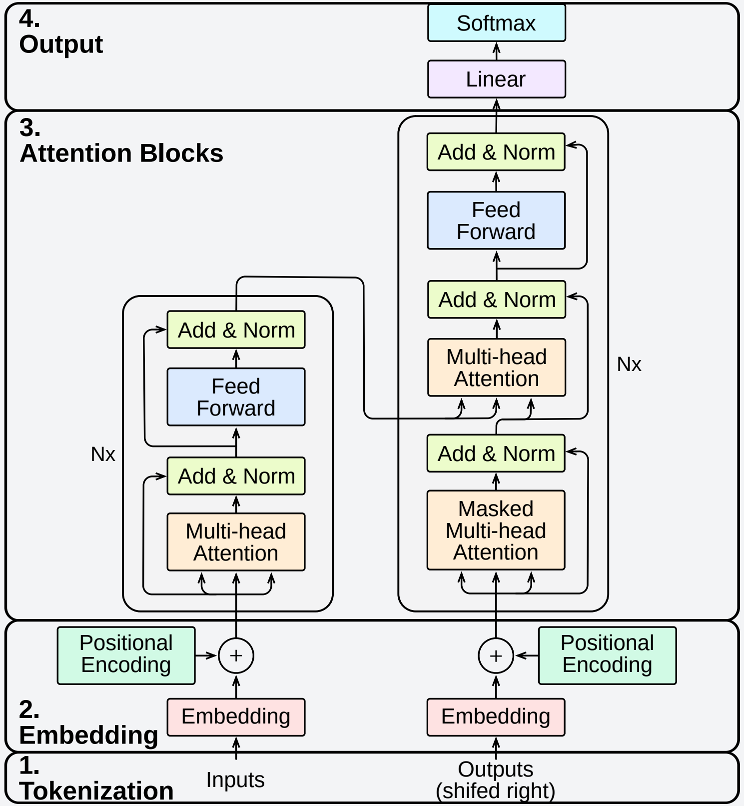


Islam, et al. (2023). A Comprehensive Survey on Application
Transformers for Deep Learning Tasks. arXiv:2306.07303.

The Transformer

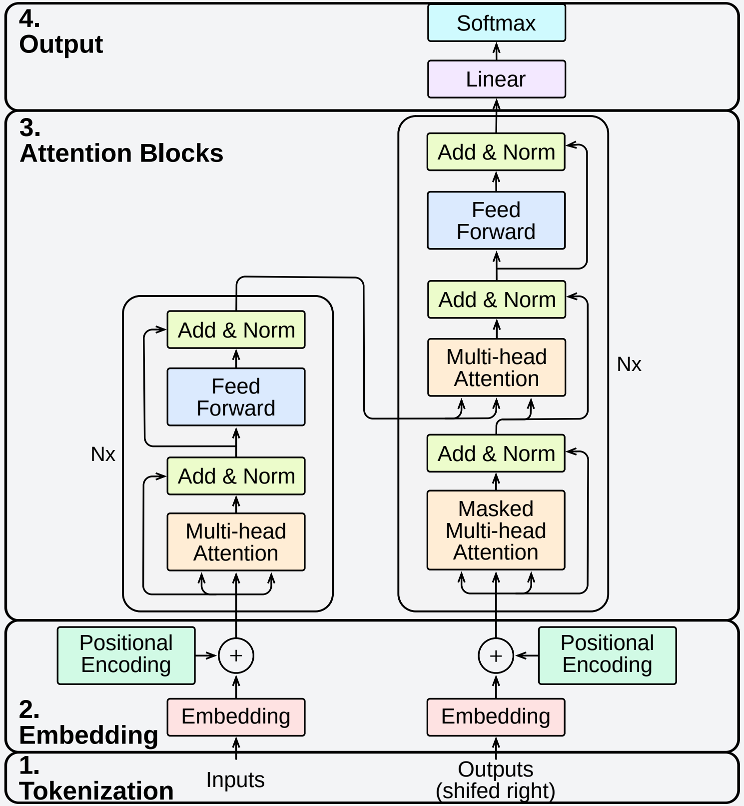


Breaking the Transformer into modules



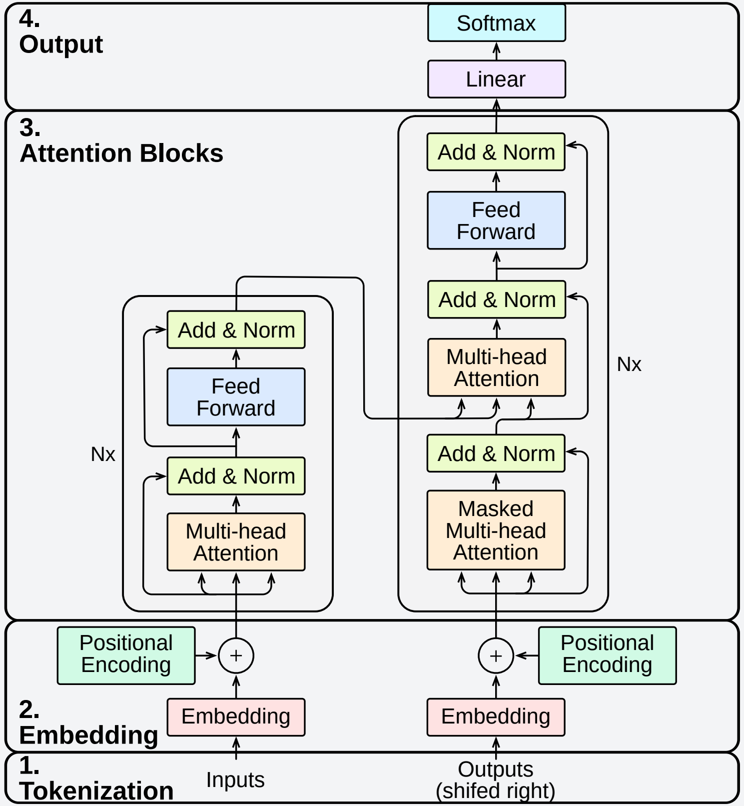
Breaking the Transformer into modules

4. Output
   - Softmax
   - Linear



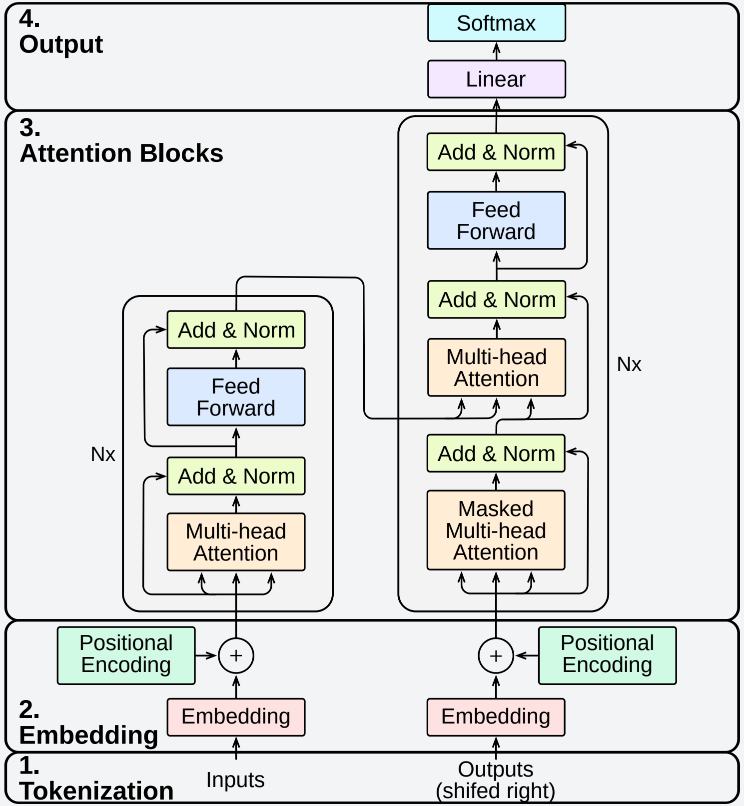
Breaking the Transformer into modules

4. Output
   - Softmax
   - Linear
3. Attention Blocks
   - Multi-head Attention
   - Add & Norm
   - Feed Forward



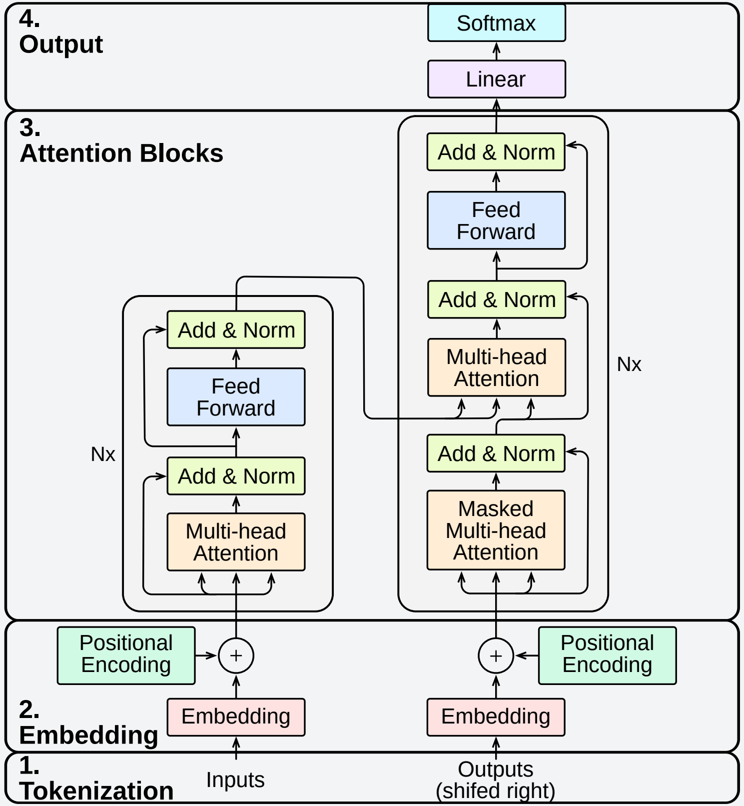
Breaking the Transformer into modules

4. Output
   - Softmax
   - Linear
3. Attention Blocks
   - Multi-head Attention
   - Add & Norm
   - Feed Forward
2. Embedding
   - Token Embedding
   - Positional Encoding



Breaking the Transformer into modules

4. Output
   - Softmax
   - Linear
3. Attention Blocks
   - Multi-head Attention
   - Add & Norm
   - Feed Forward
2. Embedding
   - Token Embedding
   - Positional Encoding
1. Tokenization
   - (Not pictured)



Breaking the Transformer into modules

4. ~~Output~~
   - ~~Softmax~~
   - ~~Linear~~
3. Attention Blocks
   - Multi-head Attention
   - Add & Norm
   - Feed Forward
2. Embedding
   - ~~Token Embedding~~
   - Positional Encoding
1. Tokenization
   - (Not pictured)

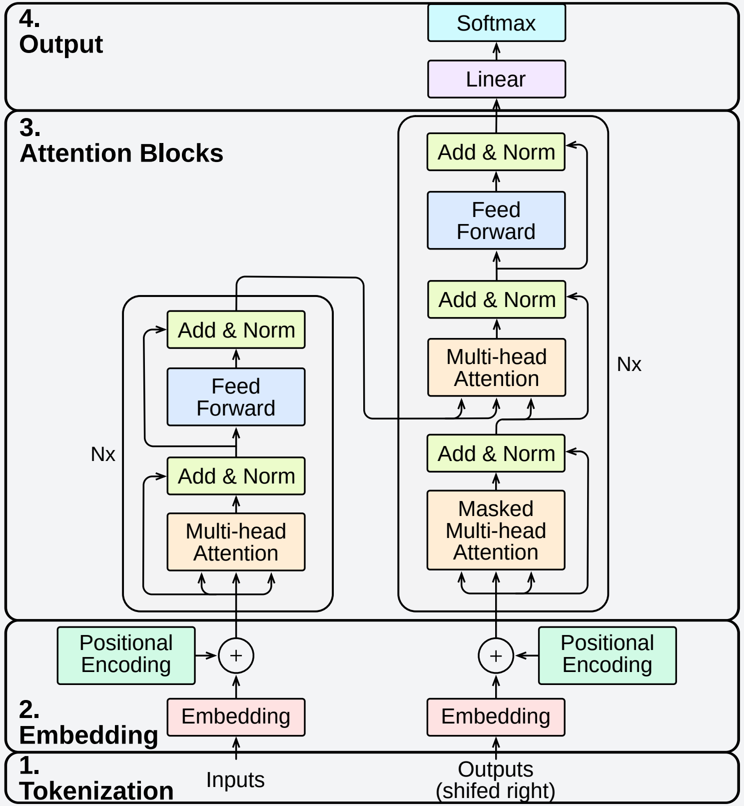
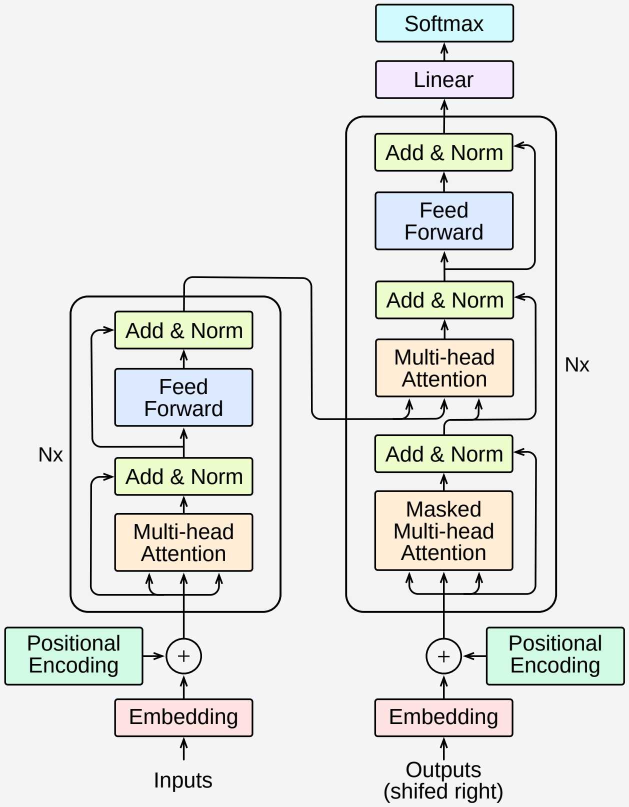


Table of Contents



Encoders & Decoders

Text comes in, text goes out

Jakob Uszkoreit (August 31, 2017). Transformer: A Novel Neural Network Architecture for Language Understanding. https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/

Attention Blocks

What makes the Transformer what it is — and where it came from

Multi-head Attention

$$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}$$

$$\nabla \cdot \vec{B} = 0$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\nabla \times \vec{B} = \mu_0 \vec{J} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$$

Multi-head Attention

$$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}$$

$$\nabla \cdot \vec{B} = 0$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\nabla \times \vec{B} = \mu_0 \vec{J} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$$

Multi-head Attention

$$\nabla \cdot \vec{E} = \frac{\rho}{\varepsilon_0}$$

$$\nabla \cdot \vec{B} = 0$$

$$\nabla \times \vec{E} = -\frac{\partial \vec{B}}{\partial t}$$

$$\nabla \times \vec{B} = \mu_0 \vec{J} + \mu_0 \varepsilon_0 \frac{\partial \vec{E}}{\partial t}$$

Definition & Properties

Multi-head Attention

Non-Transformer Examples

Multi-head Attention

Attention in Transformers

Multi-head Attention

Multi-head Attention

Multi-head Attention

Add & Norm

Residual Connections

Add & Norm

LayerNorm

Add & Norm

These are the equations

$$\begin{aligned}\mathbf{X}_l \;=\; & \texttt{LayerNorm}(\\ & \quad \mathbf{X}_{l-1} + \texttt{SubLayer}\left(\mathbf{X}_{l-1}\right) \\ &)\end{aligned}$$

Feed Forward

Embedding

Position Encoding

Tokenization

Training Transformers

The End