Attention is All You Need

(2017 Paper)

Presenter: Noah Sealy

P-08 Presentation Road Map

1. **Transformers - Noah**
   - Discussing *Attention is All You Need* (2017).
   - Paper sets foundation for the BERT language models.
     - Main focus on P-08's research project for CSCI 6609!
   - *Focus on Self-Attention.*
2. BERT Model - Shakhboz
3. SBERT Model - Rakshit
4. Our Project - Bhuvaneshwari

Authors

Google Research

Asish Vaswani - Google Brain: ~70 publications & ~25k citations.

Noam Shazeer - Google: ~120 publications & ~27k citations.

Niki Parmar - Google Brain: ~30 publications & ~21k citations.

Jakob Uszkoreit - Google: ~100 publications & ~25k citations.

Llion Jones - Google Brain: ~30 publications & ~21k citations.

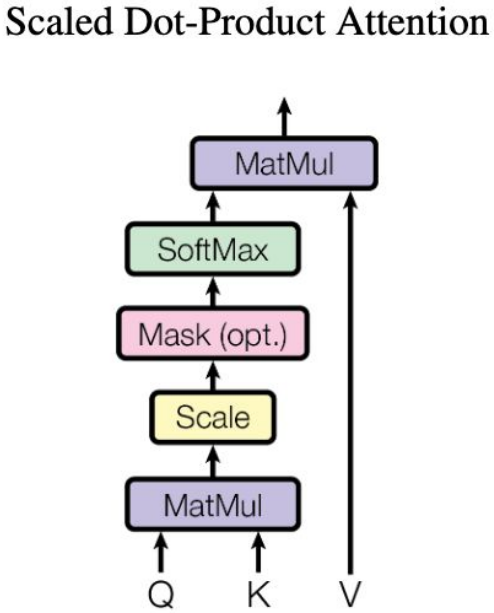
Aidan N. Gomez - University of Toronto: ~20 publications & ~21k citations.

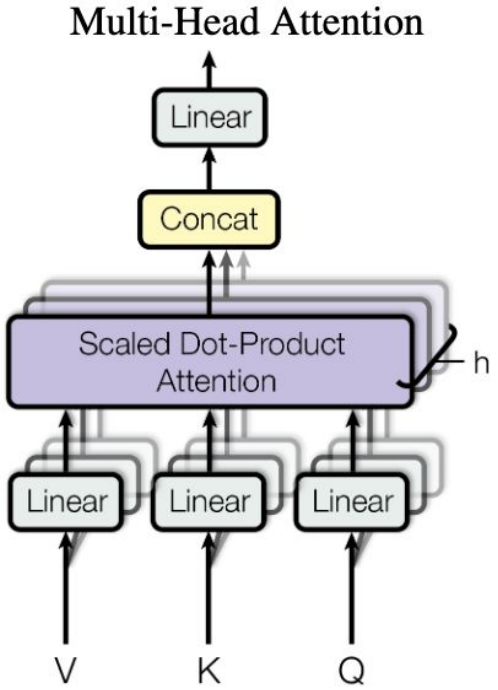
Łukasz Kaiser - Google Brain: ~130 publications & ~60k citations.

Illia Polosukhin - NEAR.AI: ~20 publications & ~20k citations.

Attention Mechanisms

- Attention refers to capturing dependencies in language without having to process a word's position in text.





Self-Attention (High Level)

1. Query
   - Projected representation of the embeddings.
   - Query is used to determine the context of the words.
   - Example, a query may a projection of the embeddings in the direction of “location”. Supporting semantic meaning of locations of the input text.
2. Keys
   - The input text.
3. Values
   - The values which the keys mean to index, usually just the keys themselves.

Self-Attention (High Level)

1. Embed Queries (Word2Vec)
2. Scaled Dot Product Attention Function

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

- Create a distribution with peaks at context words, based on query.

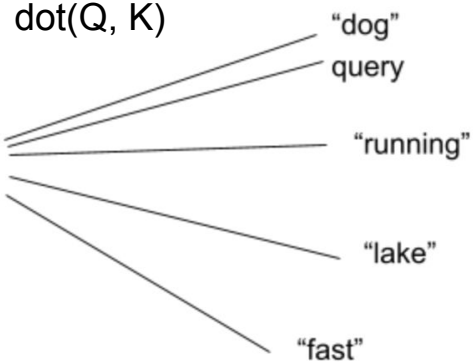
3. Generate new embedding based on combining values and distribution.
    - Derives new contextualized embeddings!
    - Embeddings will inherit traits of embeddings that relate to it.

(Simplified) Illustrated Example

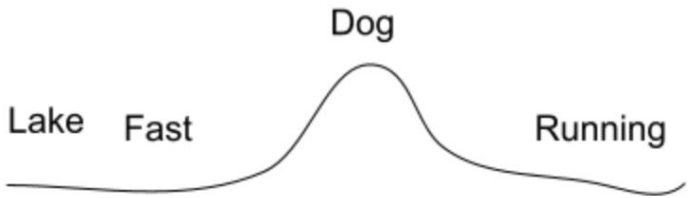
- “the dog is running to the lake, it is going fast”
- Using a query which goes in the direction of identifying pronouns.
  - In this case “it”.
- Shown on next slide...

(Simplified) Example

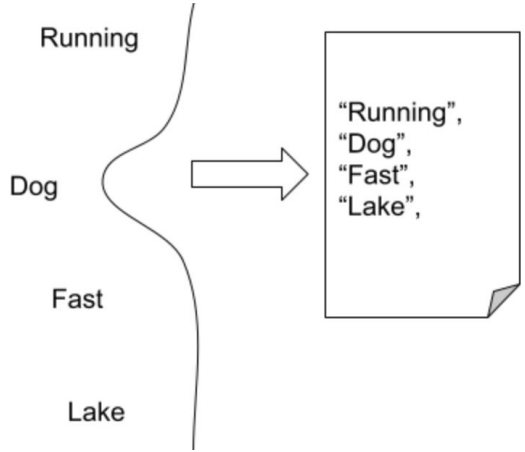
1. dot(Q, K)



2. softmax(dot(Q, K))



3. softmax(dot(Q, K)) * V



4. Contextualized Embedding!

Multi-Headed Self Attention (High Level)

- We can execute self-attention multiple times (in parallel) to investigate many different relationships, using different query projections.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

- The output is some sort of function of all of these contextual embeddings.

Study's Model

- Used for text translation.
- State-of-art language model in translation test (2017).
- 3 Multi-Head Attention layers.
- Split into encoder stack (left) and decoder stack (right).
- **Encoder stack used in BERT model to generate contextual word embeddings.**

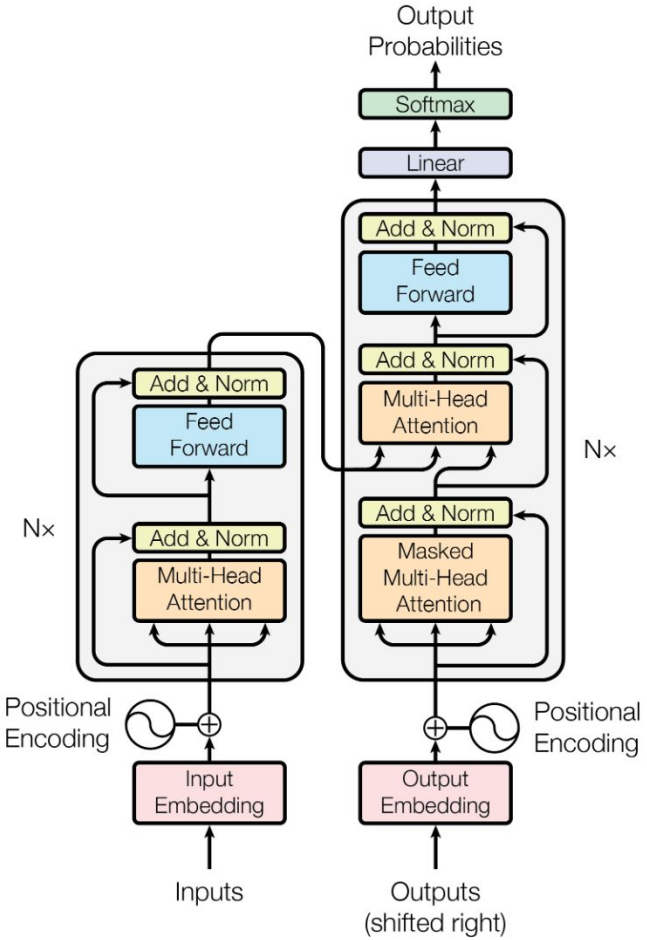


Figure 1: The Transformer - model architecture.

References & Resources

1. **A. Vaswani, N. Shazeer, N. Parmar, et al. (2017). *Attention is all you need.* Advances in neural information processing systems. Retrieved from https://arxiv.org/abs/1706.03762.**
2. Alammar, Jay. (2018) *The illustrated transformer*. [Blog post]. Retrieved from https://jalammar.github.io/illustrated-transformer/.
3. Peltarion. (2020). *How to get meaning from text with language model bert | ai explained*. [Video file]. Retrieved from https://www.youtube.com/watch?v=-9vVhYEXeyQ.
4. Dirac, Leo. (2019). *Lstm is dead. long live transformers!* [Video file]. Retrieved from https://www.youtube.com/watch?v=S27pHKBEp30&t=931s.
5. Kilcher, Yanic. (2017) *Attention is all you need.* [Video file]. Retrieved from https://www.youtube.com/watch?v=iDulhoQ2pro&t=614s.

Thanks :)

Questions?