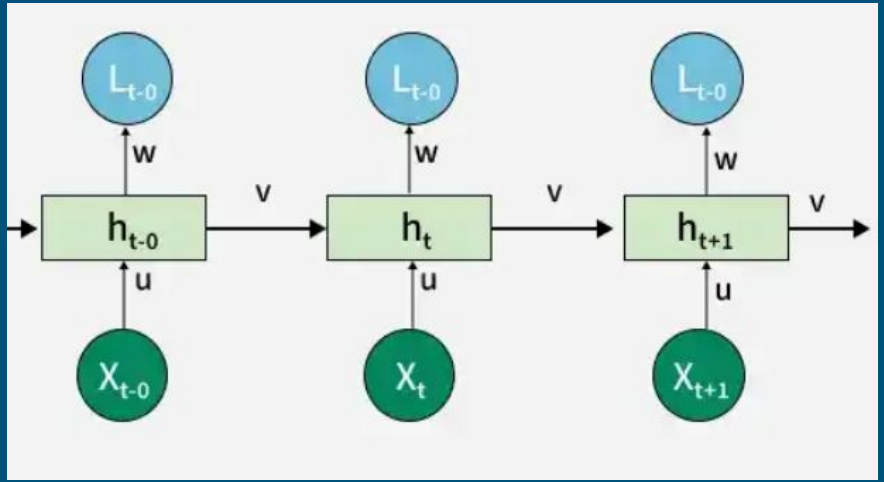
Is Attention All You Need?

Kaila Hoskins

Summary

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

Background: RNNs and LSTMs

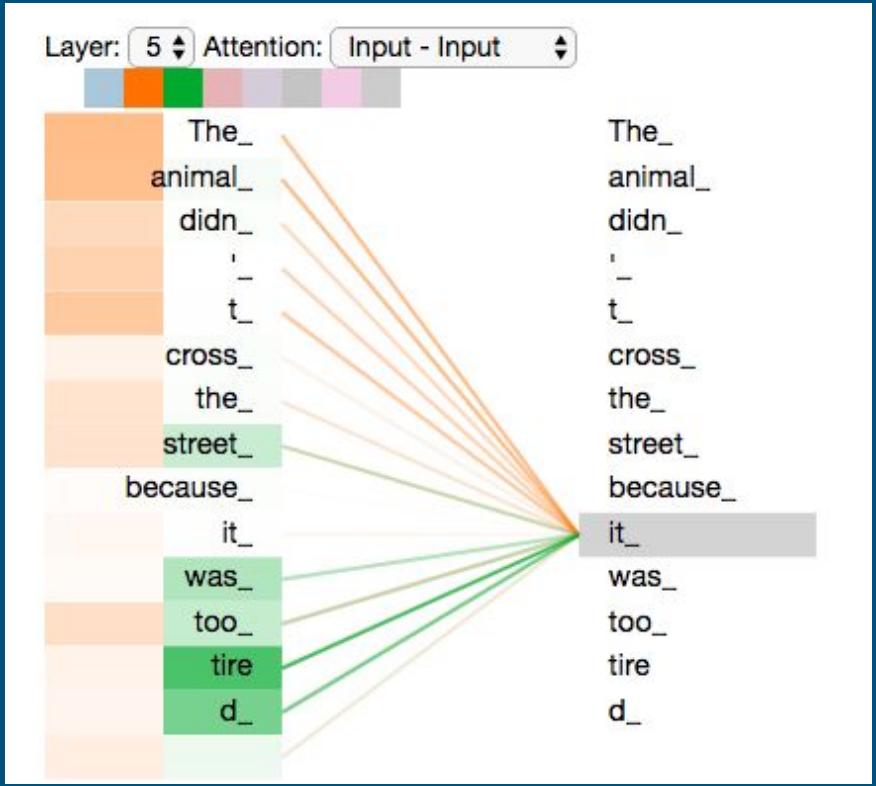


Issues:

- Not parallelable
- “Forgets”

“Attention is All You Need” proposes an idea to fix this using “Self Attention” and the “Transformer”

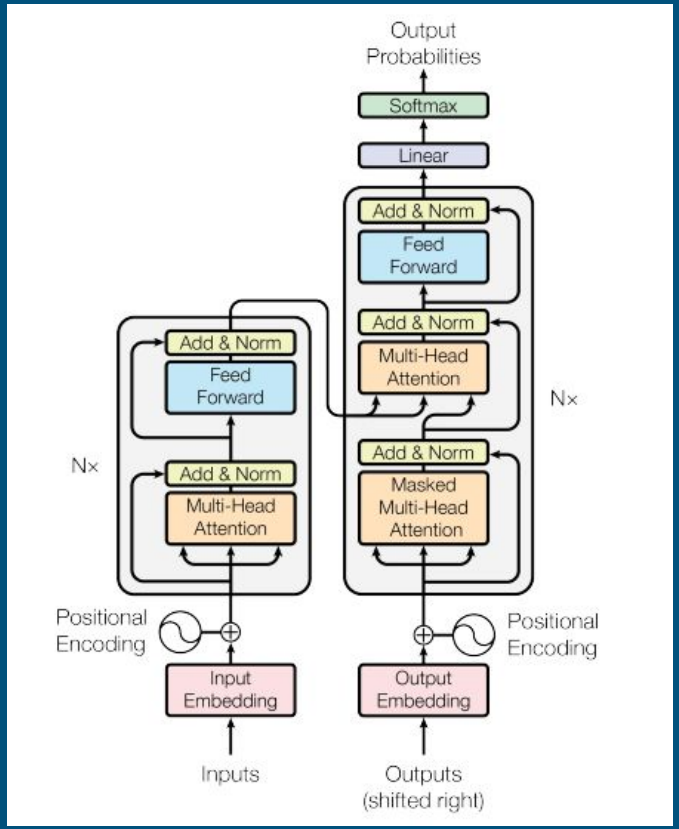
Self-Attention



- Query: Used to decide which other tokens are relevant to the current token
- Key: Used to determine how well this token matches another token's query.
- Value: The content that is actually aggregated if the token receives attention
- Dot product of Query and Key + softmax is to determine "which tokens should exchange information"

"Understanding Query, Key, Value, and Scaled Dot-Product Attention" - Manaas Gaur

The Transformer



1. Positional Embedding - encode position of each token
2. Self Attention - contextualizes token within the whole sequence
3. Add Residuals & Layer Norm - helps with training stabilization
4. Feed Forward - Refines each token individually

Strengths, Weaknesses, Adaptations

Strengths

- The transformer is the new state of the art compared to models before it
- Parallelizable - O(1) Sequential operations compared to an LSTM's O(n)
- Doesn't struggle with long term dependencies - O(1) Maximum path length compared to LSTM's O(n)

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.0**	$2.3 \cdot 10^{19}$	

Hochreiter, Sepp & Schmidhuber, Jürgen. (1997). Long Short-Term Memory. Neural Computation. 9. 1735-1780. 10.1162/neco.1997.9.8.1735.

Weaknesses

- Increases Complexity - $O(n^2*d)$ compared to an LSTM's $O(n*d^2)$
- Need to take in the whole sequence input increasing RAM need
- It is not performing actual reasoning
- The paper itself limits the scope of the transformer to translation

Layer Type	Complexity per Layer
Self-Attention	$O(n^2 \cdot d)$
Recurrent	$O(n \cdot d^2)$
Convolutional	$O(k \cdot n \cdot d^2)$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$

Dziri, Nouha & Lu, Ximing & Sclar, Melanie & Li, Xiang & Jian, Liwei & Lin, Bill & West, Peter & Bhagavatula, Chandra & Bras, Ronan & Hwang, Jena & Sanyal, Soumya & Welleck, Sean & Ren, Xiang & Ettinger, Allyson & Harchaoui, Zaid & Yejin, Choi. (2023). Faith and Fate: Limits of Transformers on Compositionality. 10.48550/arXiv.2305.18654.

Extension/ Adaptations

- The transformer can be scaled up and is very generalizable
  - Translation
  - Question Answering
  - Classification
- Bidirectional transformers are better at contextualization

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
- Devlin, Jacob & Chang, Ming-Wei & Lee, Kenton & Toutanova, Kristina. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 10.48550/arXiv.1810.04805.

Conclusion

Is attention all we need?

- It is still the current state of the art and have been for years
- However the complexity and space requirements for larger context windows make alternatives appealing
- Still doesn’t reason