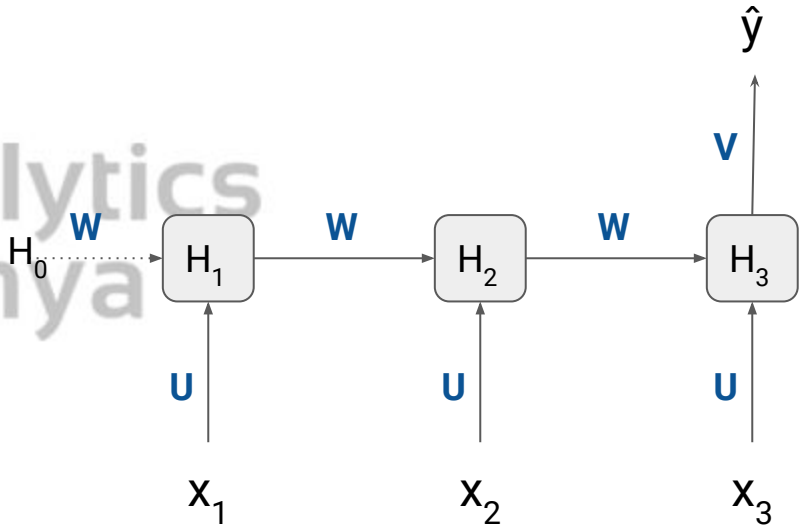
Introduction to Transformers

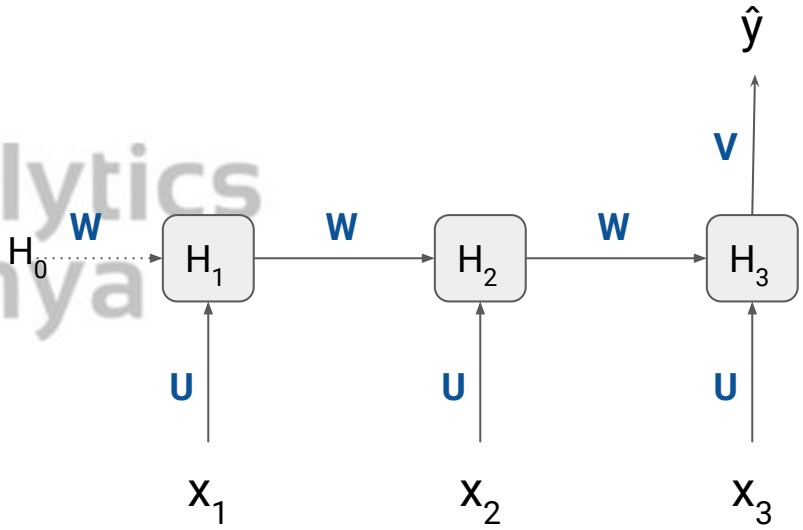
Recurrent Neural Networks

- RNN / LSTM / GRU work on data in sequential manner



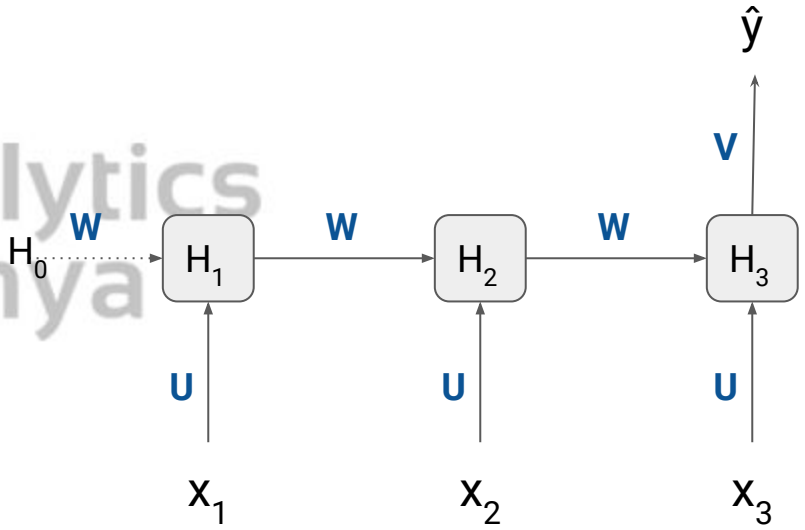
Recurrent Neural Networks

- RNN / LSTM / GRU work on data in sequential manner
- At each point they look at the current input as well as the information from the past



Recurrent Neural Networks

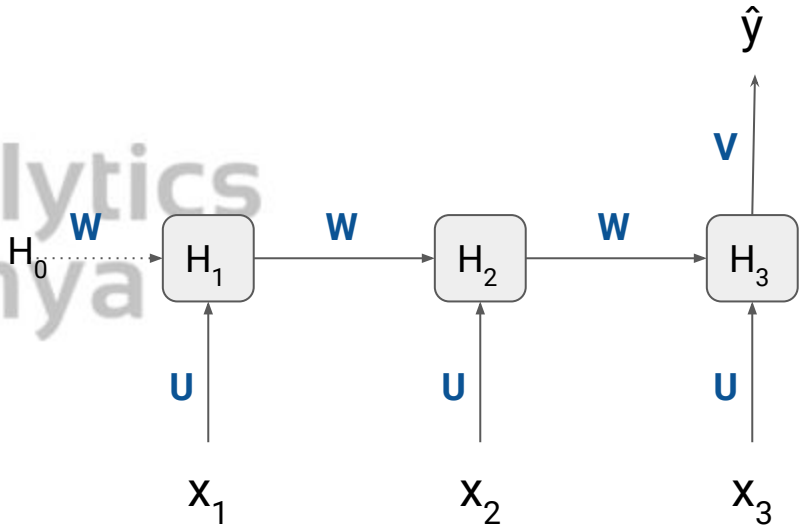
- RNN / LSTM / GRU work on data in sequential manner
- At each point they look at the current input as well as the information from the past
- Have been quite successful in NLP



Recurrent Neural Networks

Drawbacks:

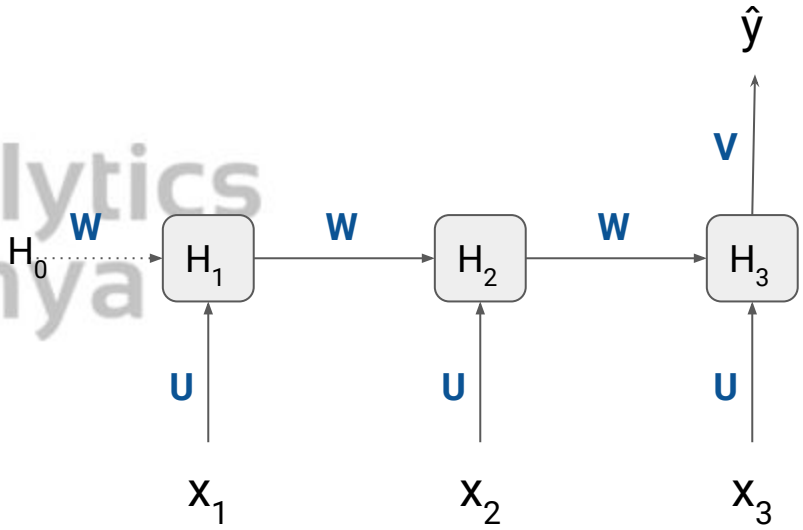
- Due to the sequential nature, it is hard to parallelize RNNs



Recurrent Neural Networks

Drawbacks:

- Due to the sequential nature, it is hard to parallelize RNNs
- Difficult to capture long range dependencies



Emergence of Transformer

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

Analytics Vidhya

Performance of Transformer

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	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**	**41.8**	$2.3 \cdot 10^{19}$	

BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of text to one or more reference translations

Key Highlights

- Since 2018, Transformer based models have taken the NLP world by storm.

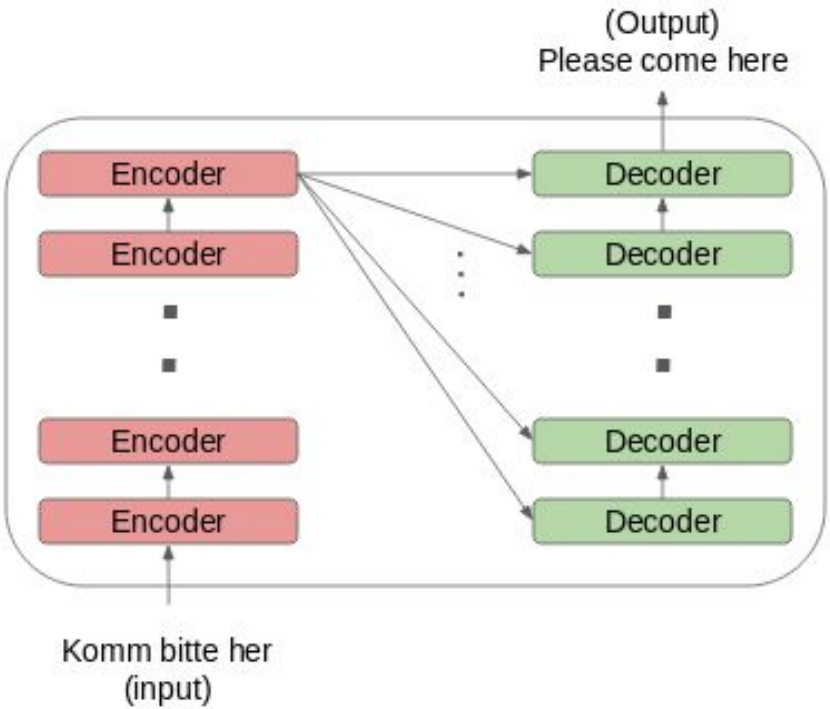
Key Highlights

- Since 2018, Transformer based models have taken the NLP world by storm.
- BERT and GPT-2 are two of most important transformer based models.

Key Highlights

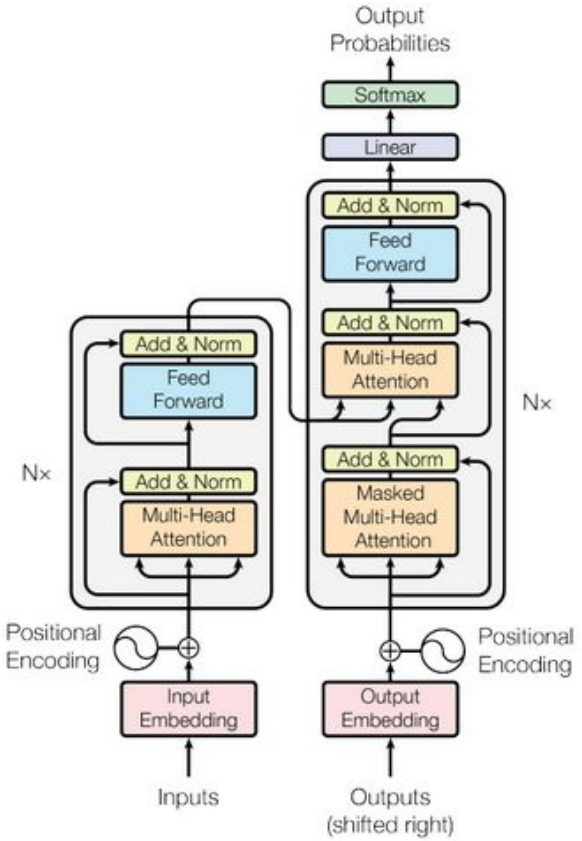
- Since 2018, Transformer based models have taken the NLP world by storm.
- BERT and GPT-2 are two of most important transformer based models.
- These models are able to understand the semantics of text to a degree never seen before.

Transformer - Architecture



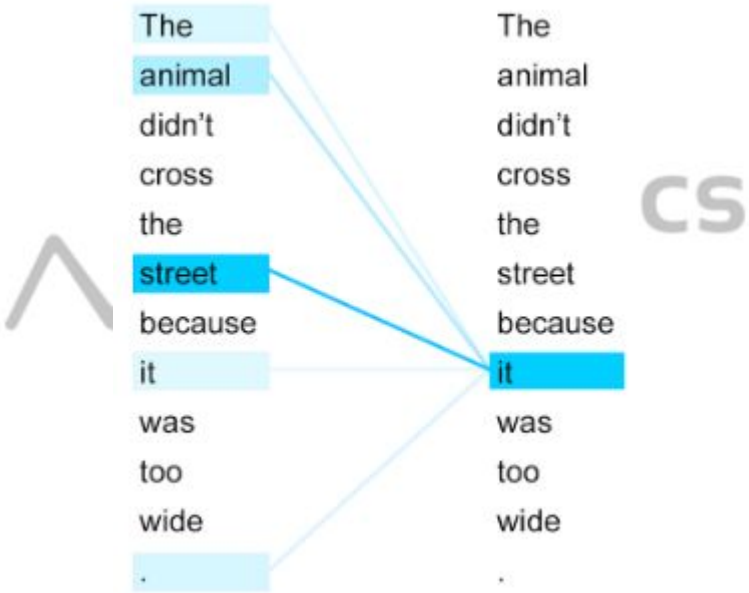
Analytics Vidhya

Transformer - Architecture



Analytics Vidhya

Transformer - Self Attention



Thank You

Analytics
Vidhya