ATTENTION IS ALL YOU NEED - TRANSFORMER

OCTOBER 27, 2021

PREPARED BY
RUIJING ZENG & TONGZI WU

UNIVERSITY OF TORONTO

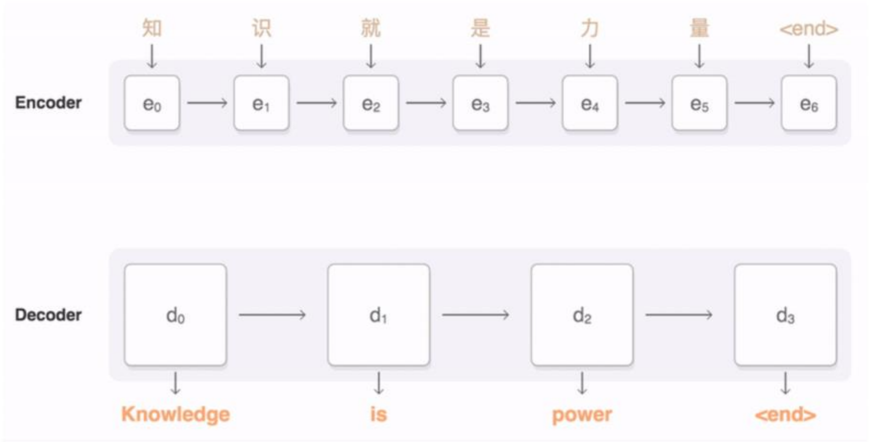
AGENDA

- Background & Related Work
- Model Architecture
- Experiment
- Medical Applications
- Conclusion

BACKGROUND

Motivation & Definition

- Motivation – 2 fold
  - Improve the performance of the machine translation model.
  - Reduce sequential computation so that allow for more parallelization and higher training speed.
- Definition of Machine Translation
  - The task of translating a sentence in a source language to a different target language.
- How to encode words?
  - One-hot: of high dimensions, too sparse
  - **Embedding**: a representation of words in a relatively low-dimensional space
  - Words => Embedding (Word2Vec)



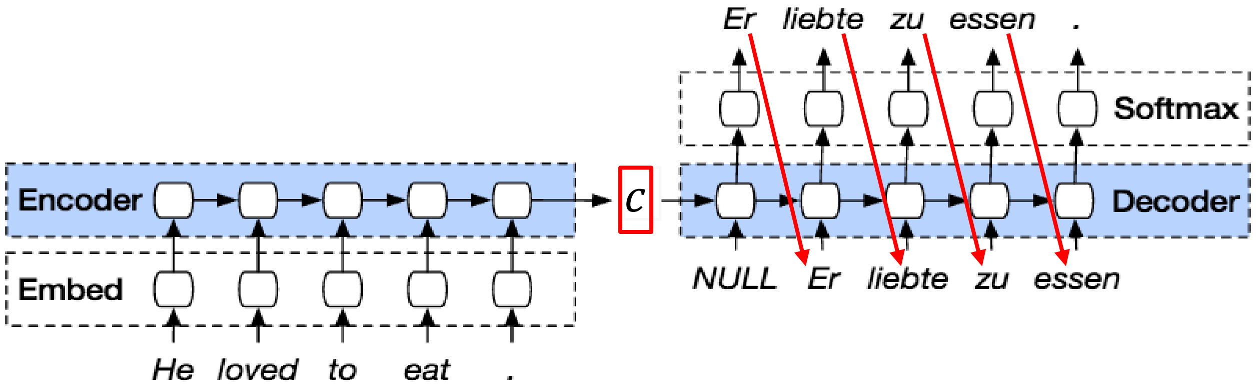
1000 words

One-hot vector representation	Embedding representation
DISORDER : [1, 0, 0, ...]	DISORDER : [-1.160, 0.343, -0.555, ...]
PROBLEM : [0, 1, 0 , ...]	PROBLEM : [0.324, -1.294, -0.608, ...]
PROCEDURE : [0, 0, 1 , ...]	PROCEDURE : [-0.484, 0.348, -0.846, ...]
1000 dimensions	16 or less dimensions

RELATED WORK

Encoder-Decoder Seq2Seq Model

- Architecture:
  - Both the encoder and decoder are Recurrent Neural Network (RNN).
  - A single context vector c is generated at the end of the encoder.
  - The decoder uses the context vector to yield the output.
- Limitations:
  - The context vector is “overloaded” with information.
  - Parallelization is precluded.

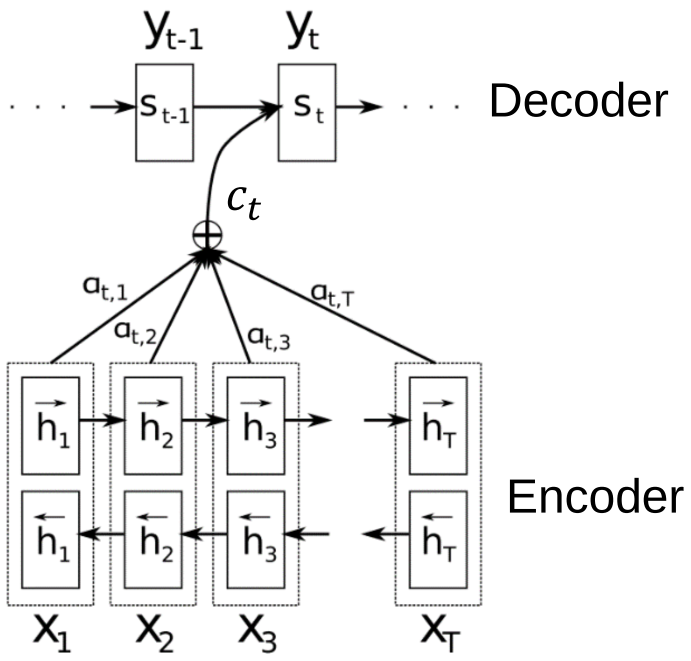


UNIVERSITY OF TORONTO

RELATED WORK

Seq2Seq with Attention Mechanism

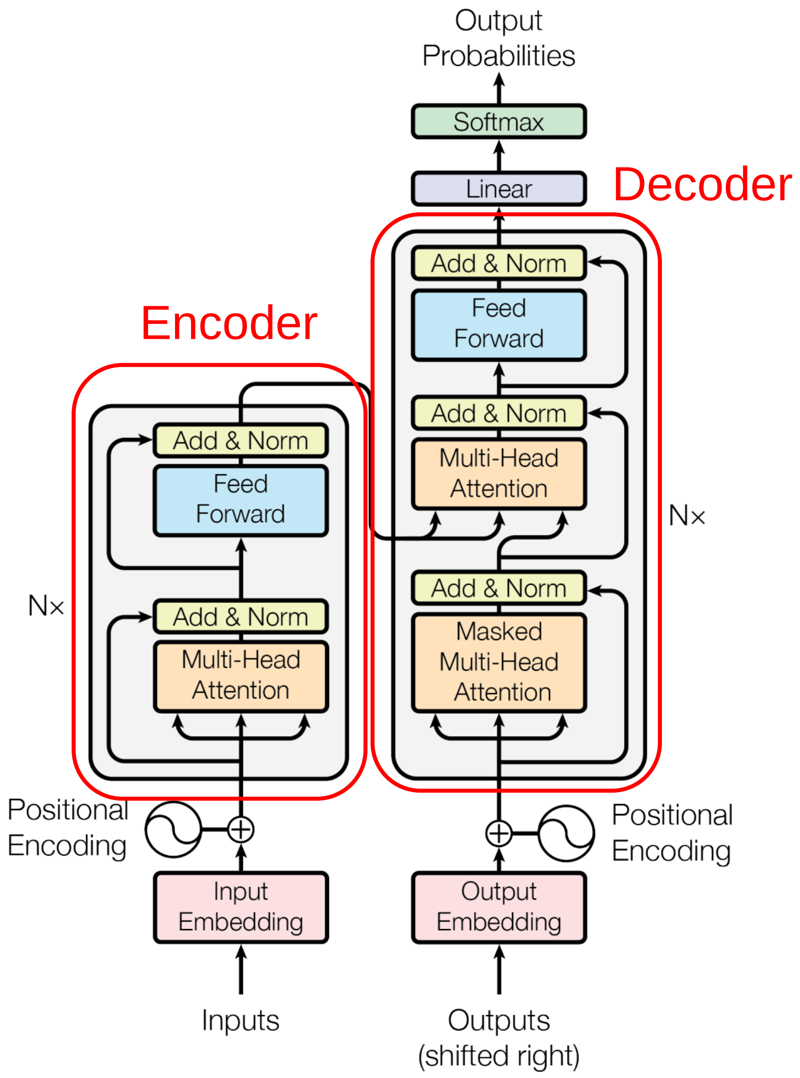
- Architecture:
  - Defines various context vector c_t for each hidden state s_t in decoder.
  - c_t is dependent on s_{t-1} and all the hidden states in the encoder.
- Strength:
  - Solved the “overloaded context vector” problem.
- Limitation:
  - The problem of parallelization remains.



MODEL ARCHITECTURE

Transformer Architecture - Parallelizable

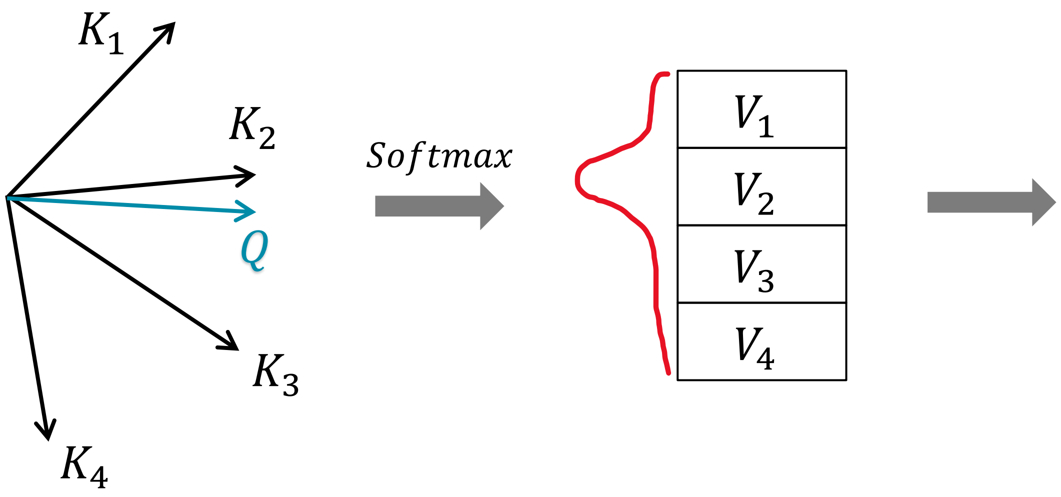
- How does it work?
    - Based solely and entirely on attention mechanisms.
    - Completely dispense with recurrence and convolutions.
- Remains the encoder-decoder structure.



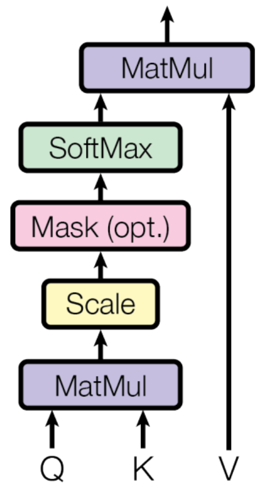
MODEL ARCHITECTURE

Attention - "Scaled Dot-Product Attention"

- What is attention?
  - Mapping a query (Q) and a set of key-value (K-V) pairs to an output
  - Similarity



weighted sum of the values



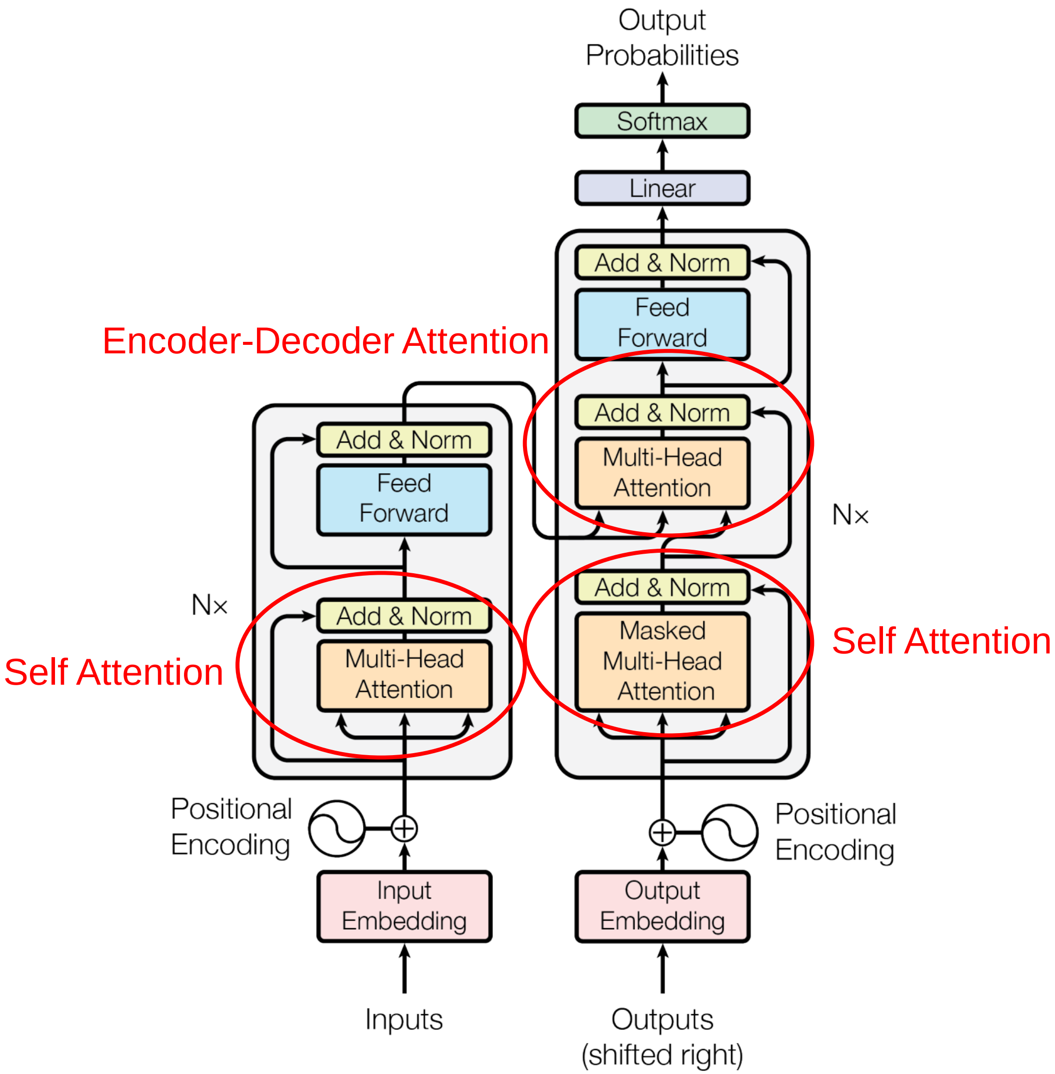
$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

UNIVERSITY OF TORONTO

MODEL ARCHITECTURE

Attention Modules

- Self Attention:
  - Computing representations of the sequence.
  - Query, key, value are the same.
- Encoder-Decoder Attention:
  - Mapping query from decoder to key-value pairs in encoder.
  - Key and value are from encoder, query is from decoder.

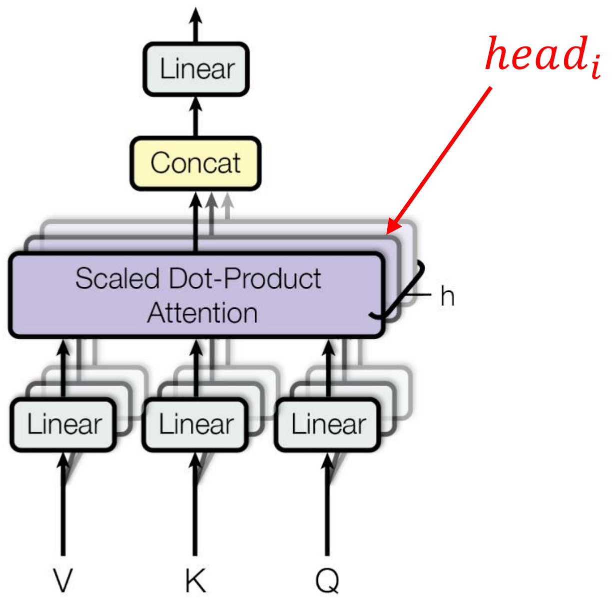


UNIVERSITY OF TORONTO

MODEL ARCHITECTURE

Multi-Head Attention

- Architecture:
    - Project Q, K and V into different subspaces
    - Perform attention to get $head_i$
    - Concatenated and projected to get final output.
- Why Multi-Head?
    - Allows the model to jointly learn the representation from different subspaces at different positions.
    - Heads – subspaces – different sentence structures
    - Higher performances with similar cost.



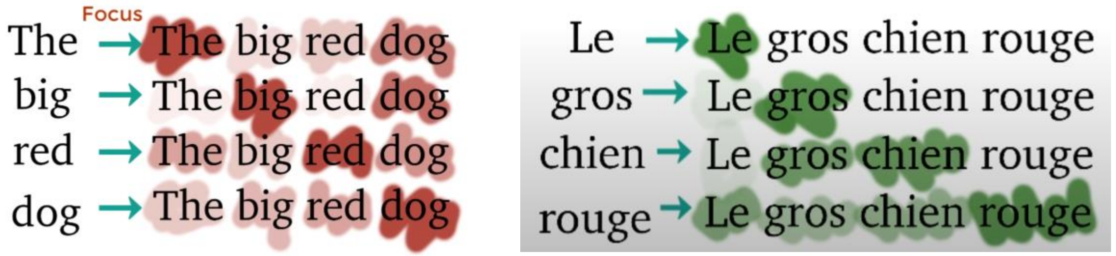
$$\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\mathrm{head_1},...,\mathrm{head_h})W^O$$
$$\text{where } \mathrm{head_i} = \mathrm{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

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

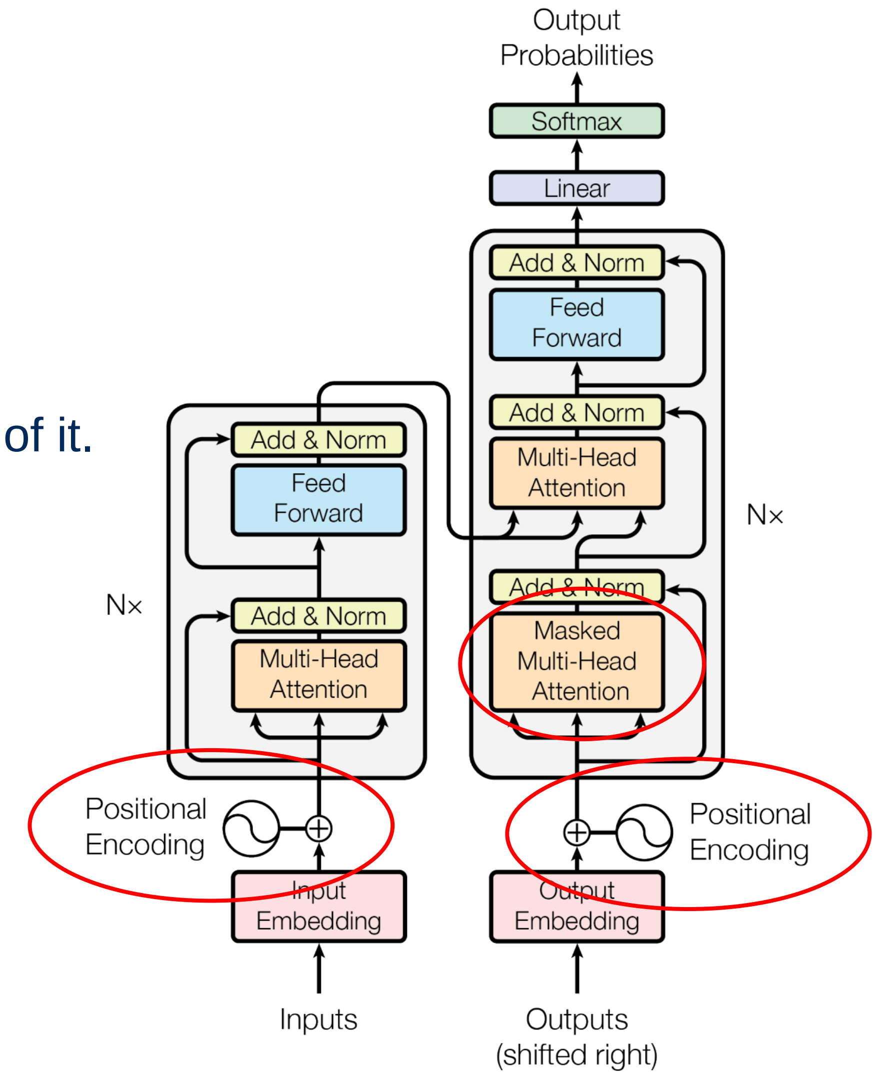
MODEL ARCHITECTURE

Another Crucial Things

- Positional Encoding
  - No recurrence, not aware of the position.
  - Add **“positional encodings”** to make the model aware of it.
- Masked Self-Attention



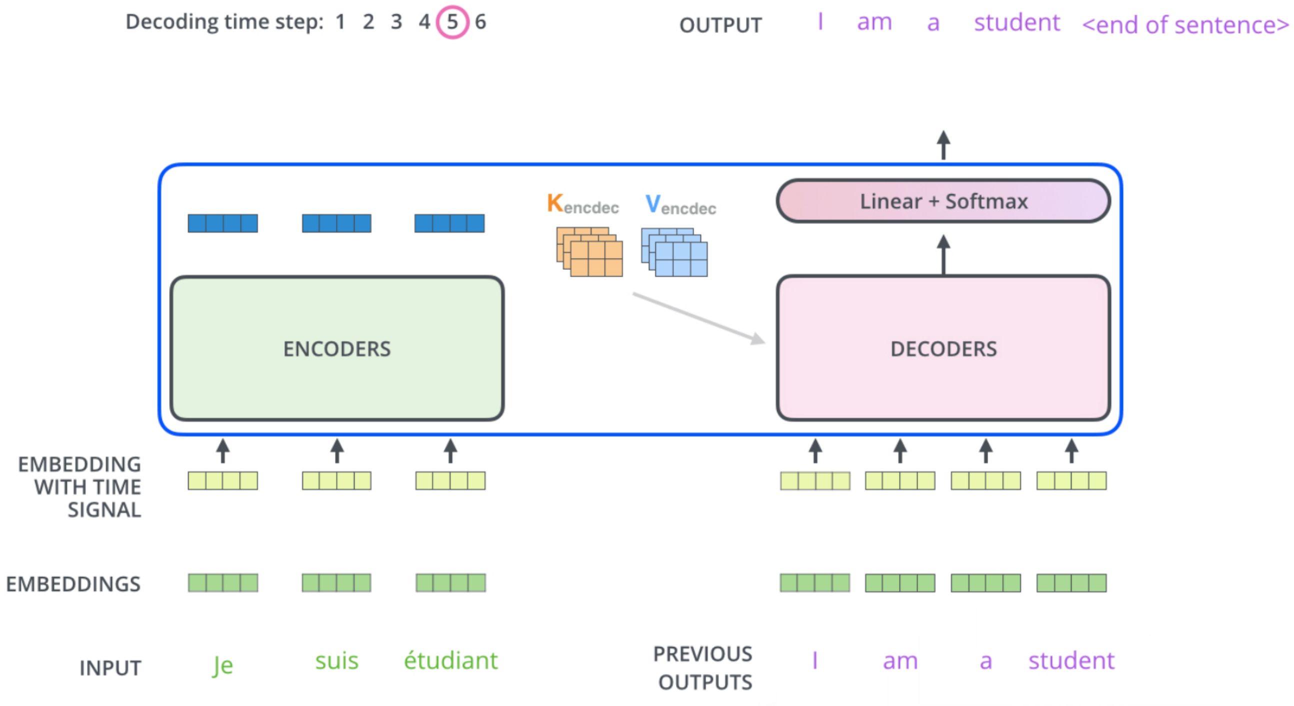
Without Mask With Mask

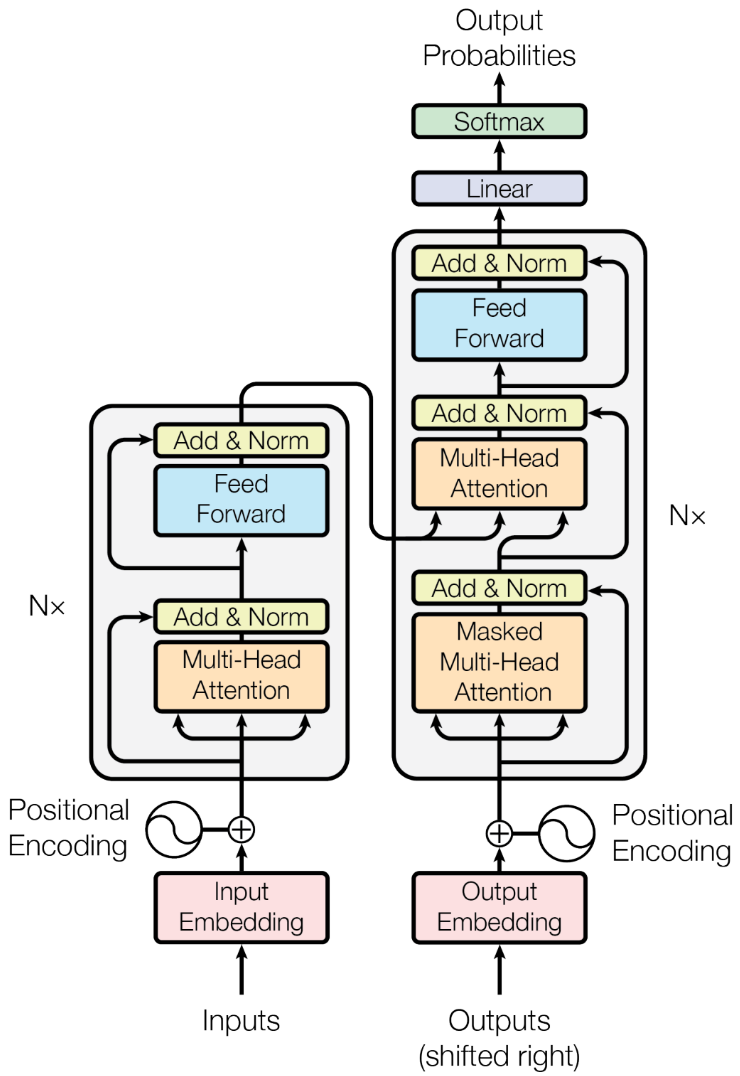


UNIVERSITY OF TORONTO

MODEL ARCHITECTURE

Workflow





UNIVERSITY OF TORONTO

EXPERIMENT

MACHINE TRANSLATION DATASET

- **WMT 2014** is a collection of datasets used in news translation, quality estimation, metrics and medical text translation tasks of the Ninth Workshop on Statistical Machine Translation.

Dataset	Sentence Pairs	Tokens
WMT 2014 English-to-German	4.5M	37,000
WMT 2014 English-to-French	36M	32,000

EXPERIMENT

BLEU SCORE

- BLEU (BiLingual Evaluation Understudy) is a metric for automatically evaluating machine-translated text.
- [0, 1], measuring the **similarity** of the machine-translated text to a set of high quality reference translations.

$$\text{BLEU} = \underbrace{\min\left(1, \exp\left(1 - \frac{\text{reference-length}}{\text{output-length}}\right)\right)}_{\text{brevity penalty}} \underbrace{\left(\prod_{i=1}^{4} \mathit{precision}_i\right)^{1/4}}_{\text{n-gram overlap}}$$

EXPERIMENT

BLEU SCORE

BLEU Score	Interpretation
< 10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
> 60	Quality often better than human

The following color gradient can be used as a general scale interpretation of the BLEU score:

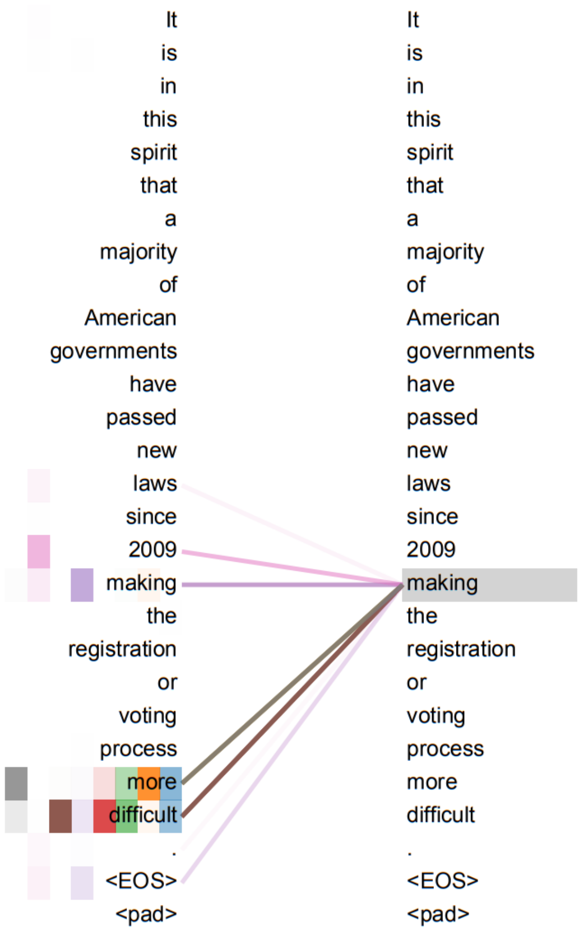
EXPERIMENT

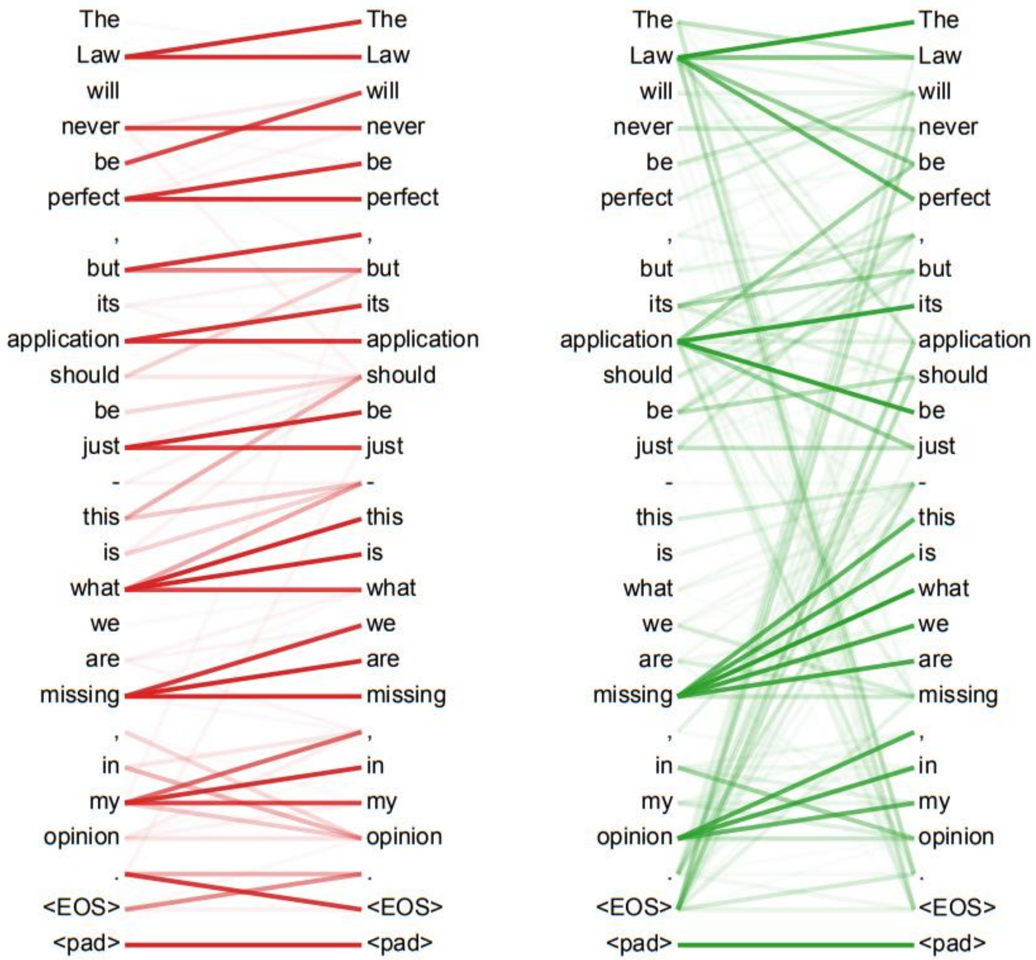
RESULT

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	$\mathbf{3.3 \cdot 10^{18}}$	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

EXPERIMENT

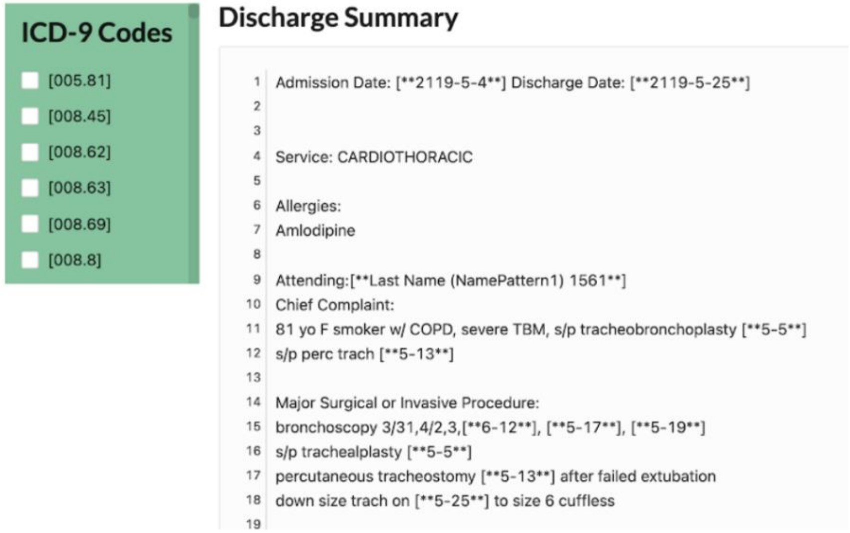
RESULT





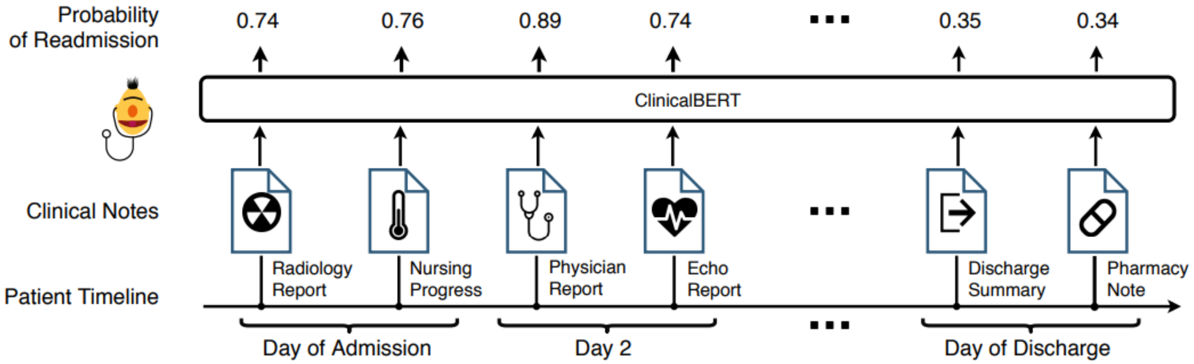
MEDICAL APPLICATIONS

MEDICAL TEXT – BERT



- ICD Coding prediction

- Readmission possibility prediction from clinical notes

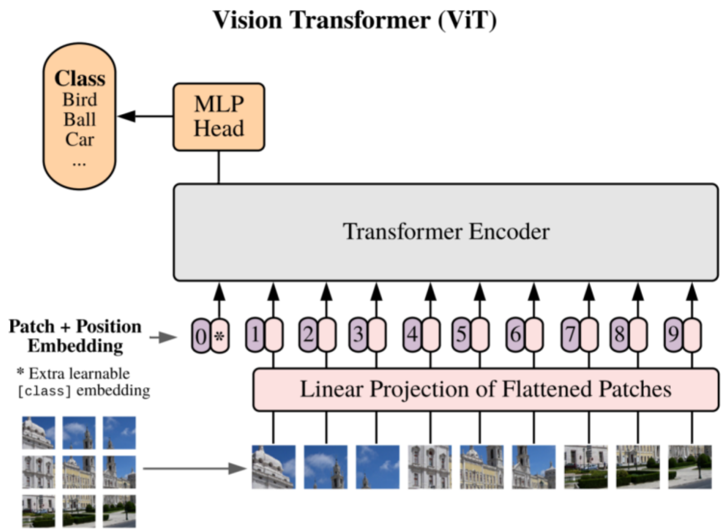


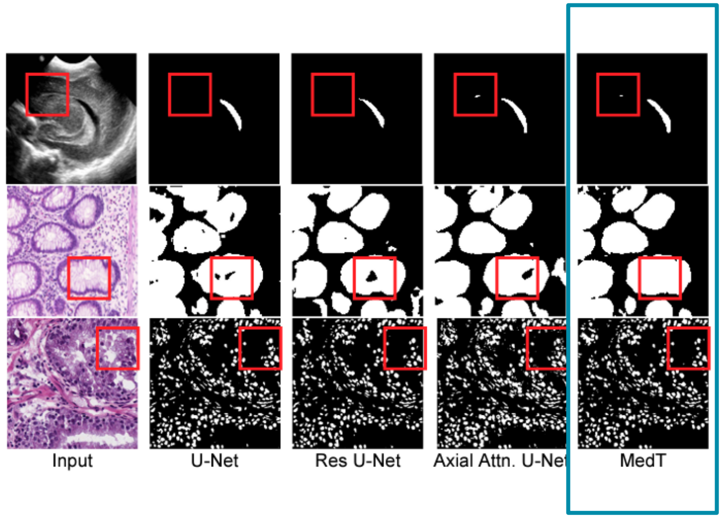
- NER, Relation Extraction, Sentence Similarity, Document Classification, Question Answering ...

MEDICAL APPLICATIONS

- **Medical Image**
  - Vision Transformer by Google 2020
  - Medical Transformer: Gated Axial-Attention for Medical Image Segmentation
  - https://arxiv.org/abs/2102.10662

- **Drug classification**
  - Toxic / Enzyme
  - Using Graph Neural Networks
  - Universal Graph Transformer Self-Attention Networks
  - https://arxiv.org/abs/1909.11855





UNIVERSITY OF TORONTO

CONCLUSION

- Transformer is the first sequence transduction model based **entirely** on attention, replacing the recurrent layers most commonly used in encoder-decoder architectures with multi-headed self-attention.

- **Strengths**
  - Multi-head attention allows the model to jointly attend to information from different representation **subspaces** at different positions.
  - Superior in quality while being more parallelizable and requiring significantly less time to train.

- **Limitations**
  - Attention can only deal with **fixed-length** text strings. The text has to be split into a certain number of segments or chunks before being fed into the system as input, which causes **context fragmentation**.
  - Attention has a **quadratic complexity** in input length, meaning attention doesn't scale well over long distances.

Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$

THANK YOU

QUESTIONS?