Deep Learning and NLP: Importance of Attention Mechanism and NLP

EQ4ALL
EQual access to technology makes all even better.

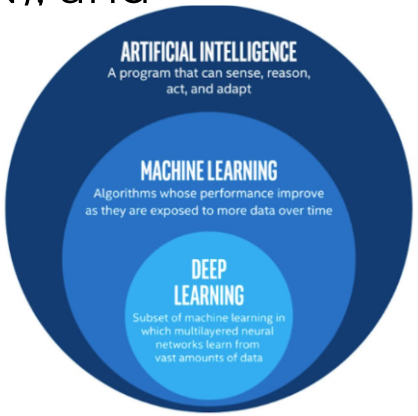
Table of Contents

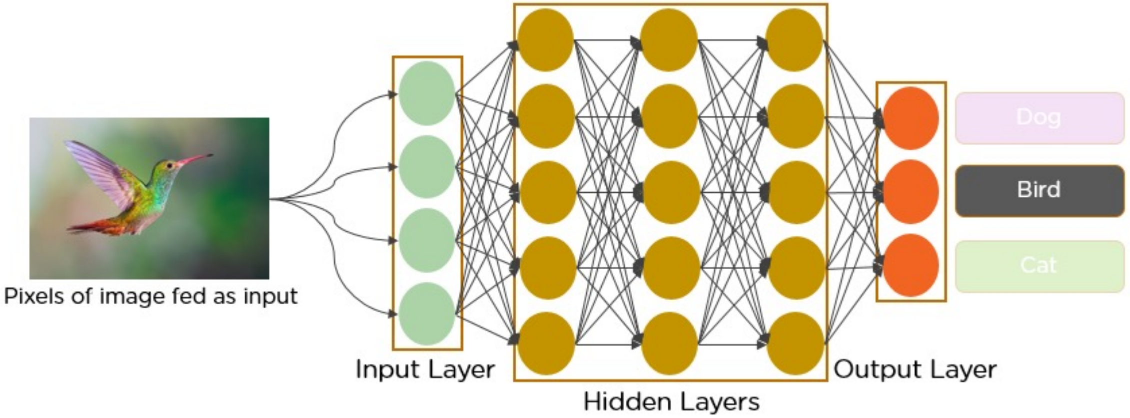
EQ4ALL
EQual access to technology makes all even better.

Deep Learning

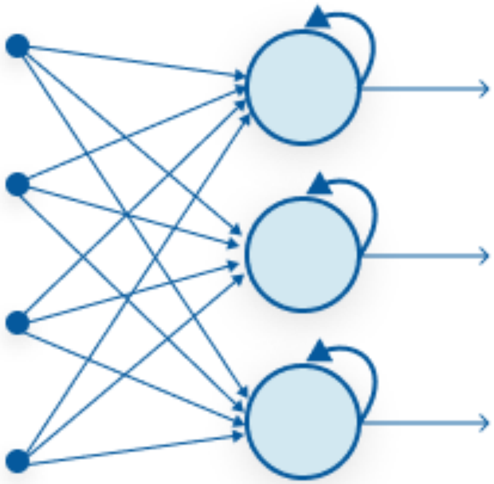
Deep Learning is a subset of machine learning that involves using artificial neural networks to l earn from data by stacking multiple layers of neurons.

- Examples: Multi-Layer Perceptron (MLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN).





Convolutional Neural Network



Recurrent Neural Network

History of Deep Learning

"The foundations for all of this artificial intelligence were laid at Cornell."

1. First Generation (1958): "Perceptron"

[Frank Rosenblatt, Cornel Aeronautical Laboratory]

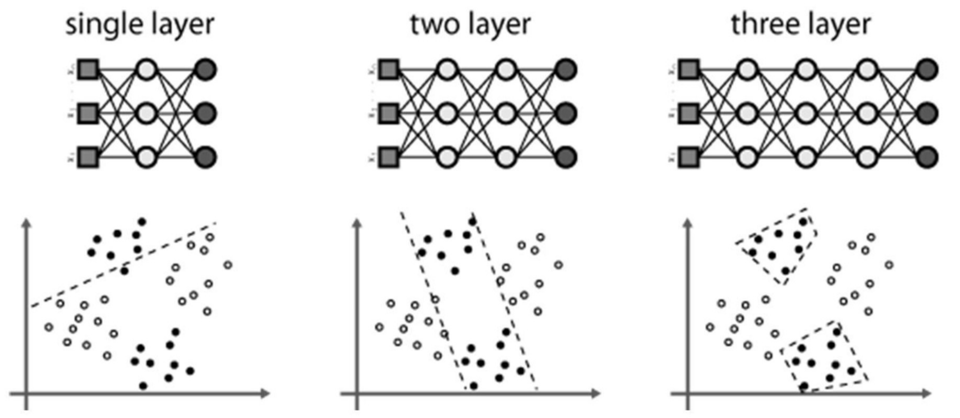
- Algorithm mimicking the structure of neurons in the brain

$$y = \varphi\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

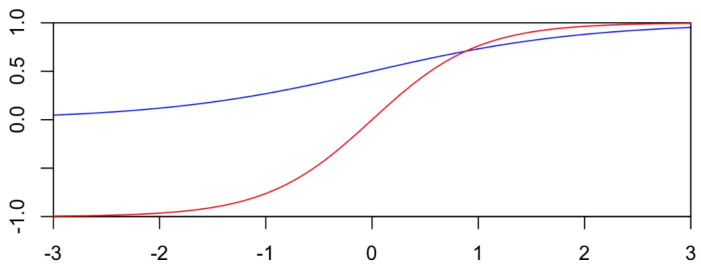
2. Second Generation: Multilayer Perceptron

- Added hidden layers

3. Third Generation: Supervised Learning – Rectified Linear Unit (ReLU), Dropout



Multilayer Perceptron



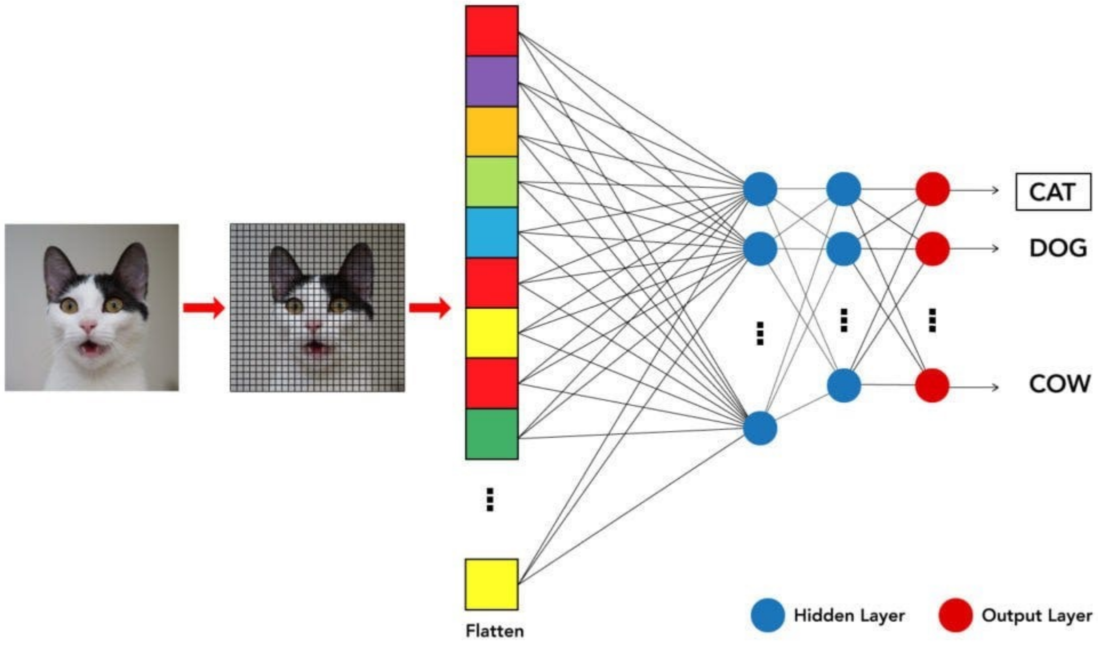
Sigmoid functions

$$\sigma(x) = \frac{1}{1+e^{-x}}$$

Key Components Of Deep Learning

Input Layer/ Hidden Layer / Output Layer

- Major application areas include image recognition, natural language processing, and autonomous driving
- The core of deep learning is its ability to learn complex patterns through multi-layered structures.
- For example, in image recognition, more complex features are gradually extracted through multiple layers.



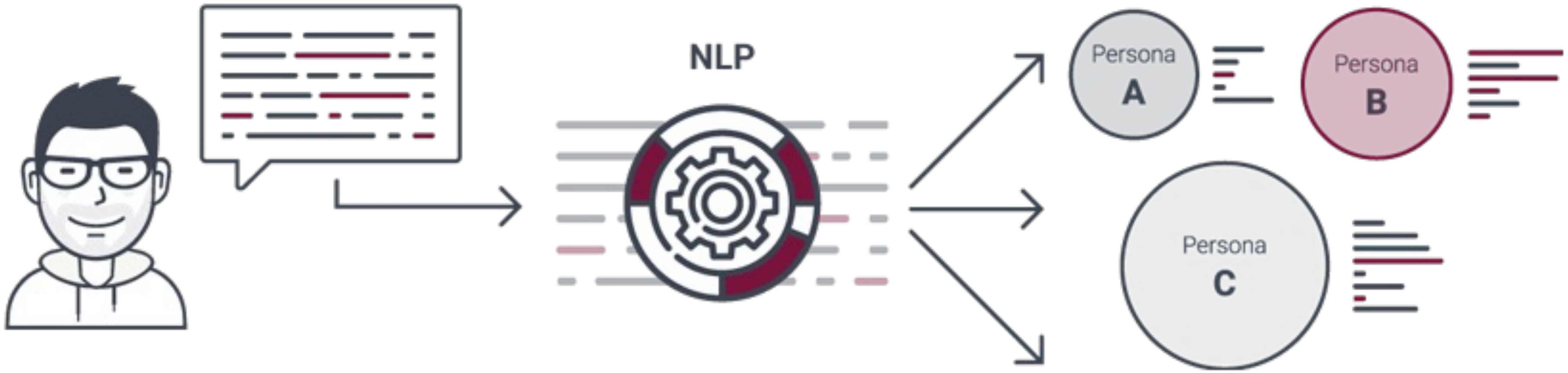
What is Natural Language Processing (NLP)?

NLP is an AI technology that enables computers to interpret, manipulate, and understand human language.

NLP Performance:

- Essential for perfectly and efficiently analyzing text and voice data
- Capable of overcoming dialects and grammatical irregularities within languages

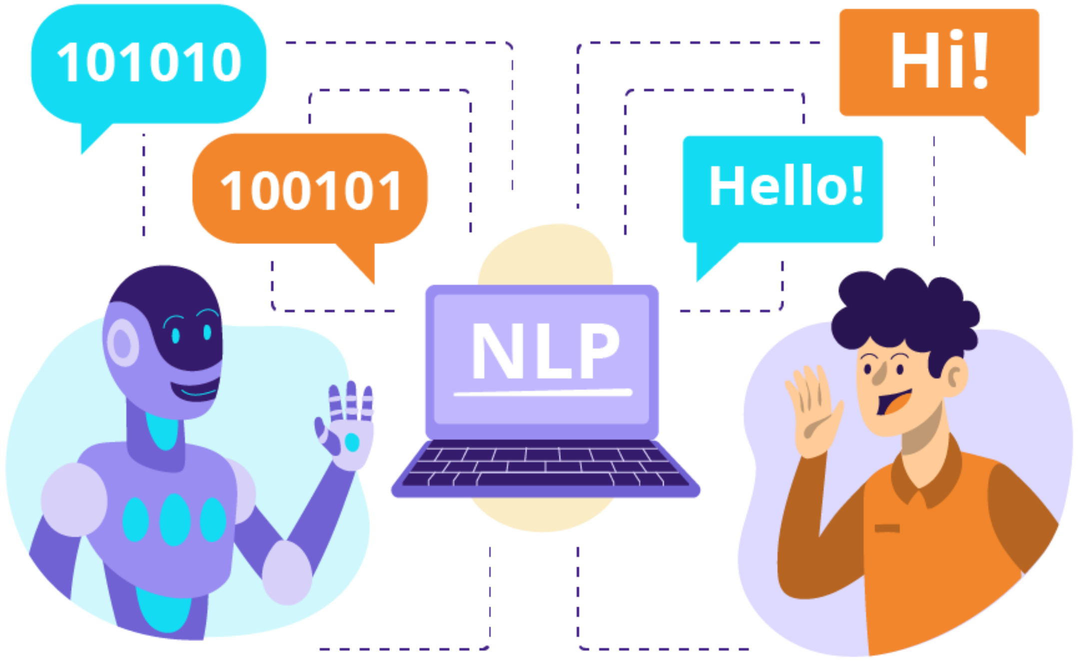
In practice, services like Google Translate use NLP to support various languages

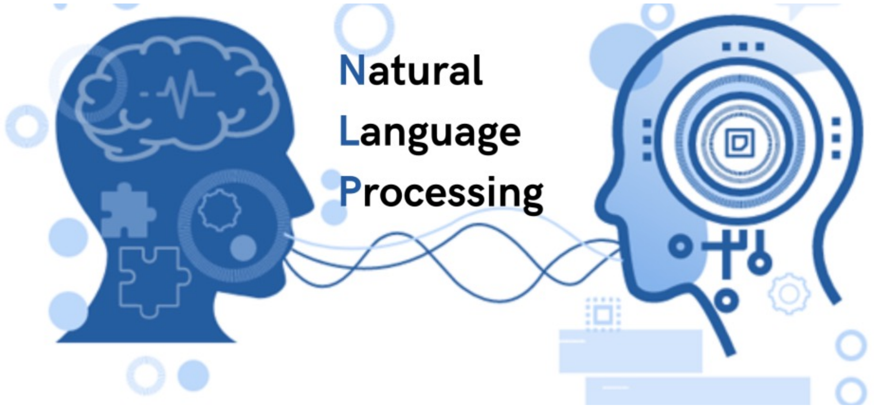


Advances in NLP

- Ultimate Goal: To make interactions between humans and AI natural and meaningful.
- Key Challenges: Processing large datasets, understanding linguistic ambiguity, grasping the context of language, and building scalable models.
  - Example: “I caught a fish at the bank.” [Then understanding “bank” as ”river or sea”]

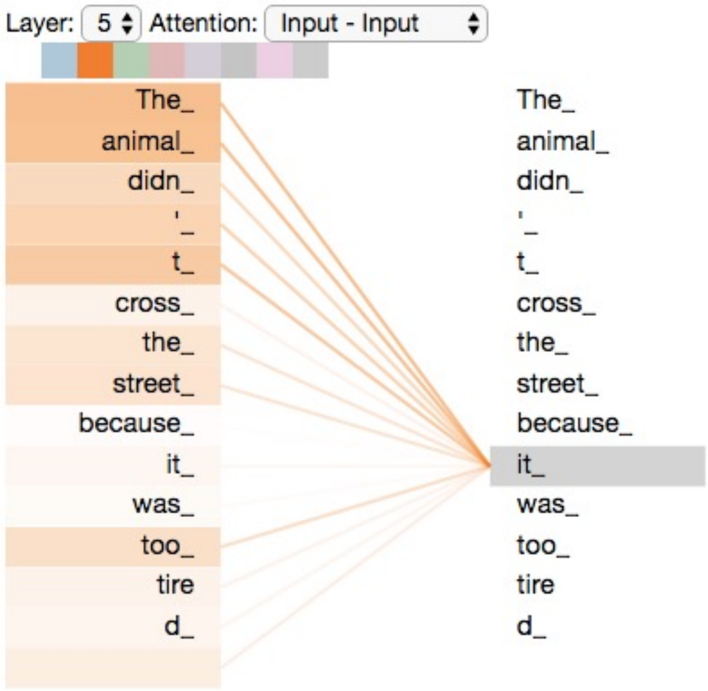
Attention mechanisms have been introduced to address these challenges.





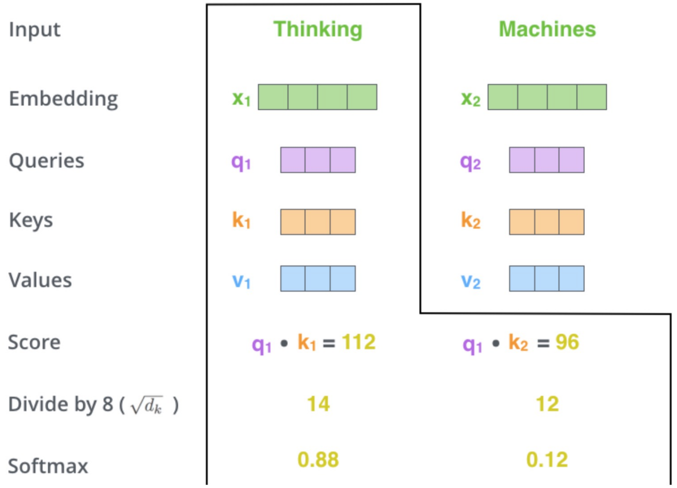
Significance of Attention Mechanism

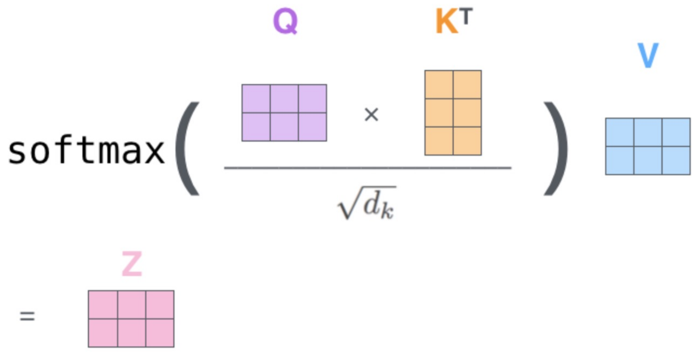
- Allows models to focus on relevant parts of the input sequence.
- Helps models assign more weight to important information, significantly aiding in understanding context in natural language processing tasks.



Working Principles of Attention Mechanisms

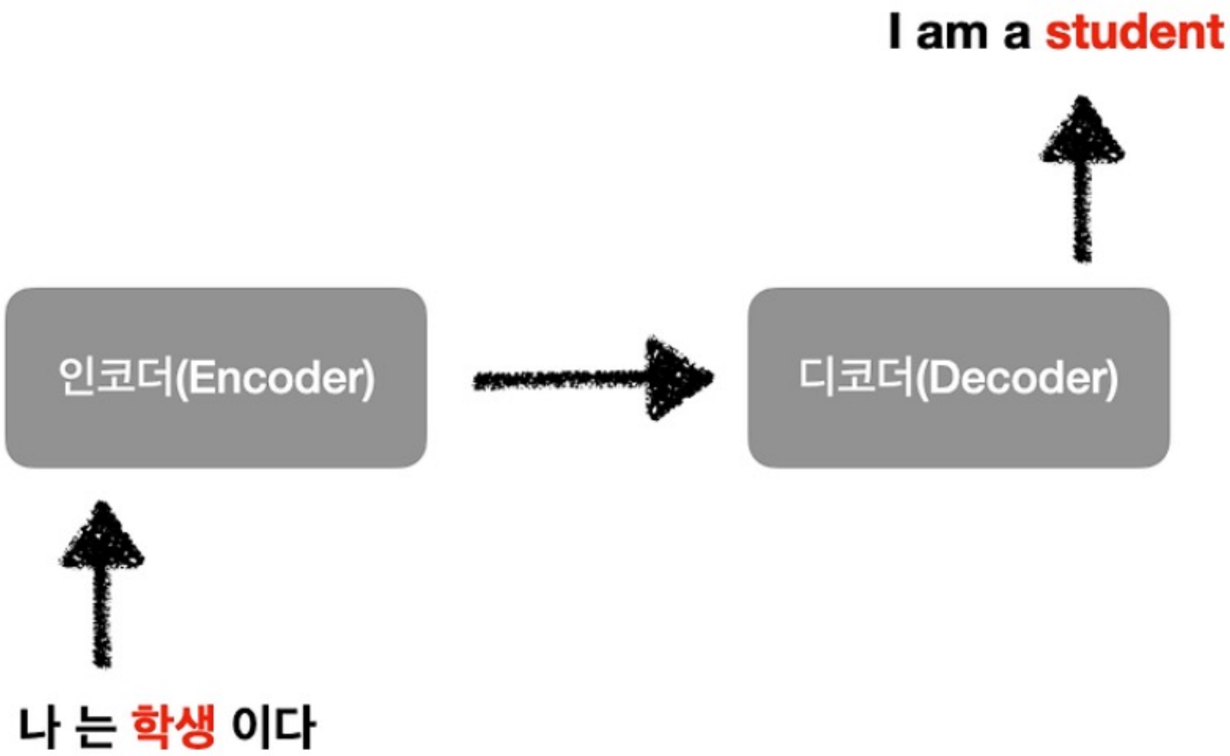
- Query, Key, Value:
  - Query: Part currently being focused on
  - Key: All input parts
  - Value: Actual values associated with keys
- Attention Score Calculation:
  - Calculated based on similarity between Query and Key
- Weight Calculation
  - Using the softmax function to calculate weights
- Weighted Sum Calculation
  - Important information is focused by calculating the weighted sum of values



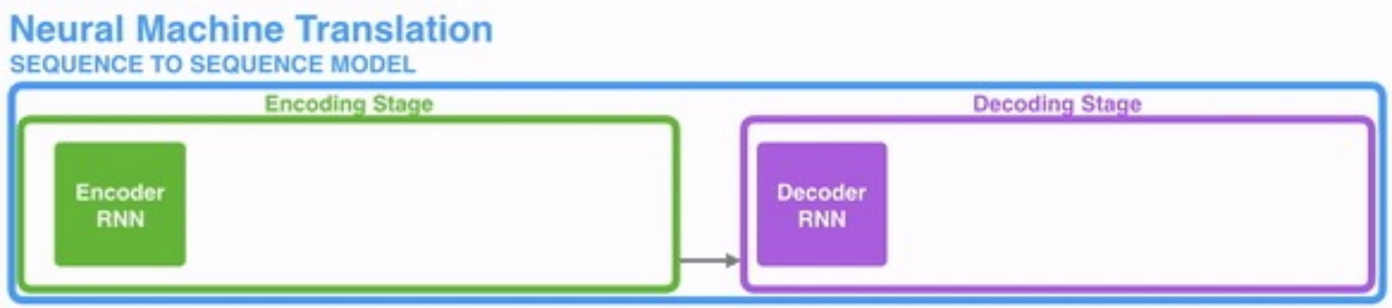


The self-attention calculation in matrix form

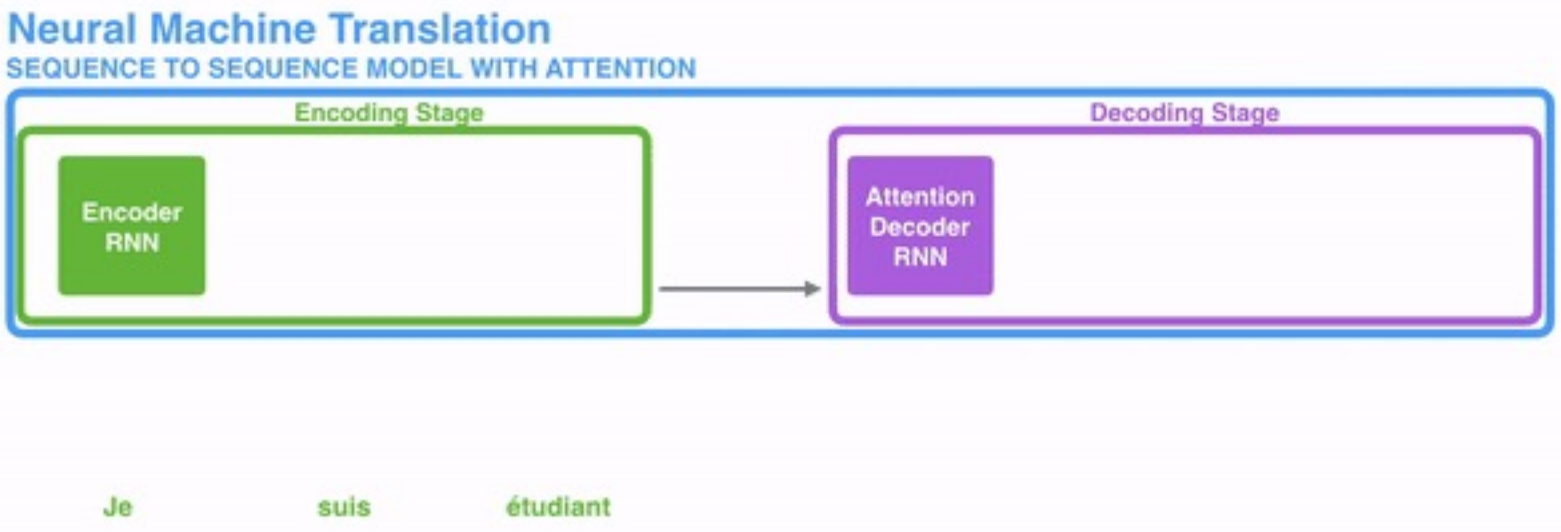
Example of Attention Mechanism Functioning



Attention Mechanism

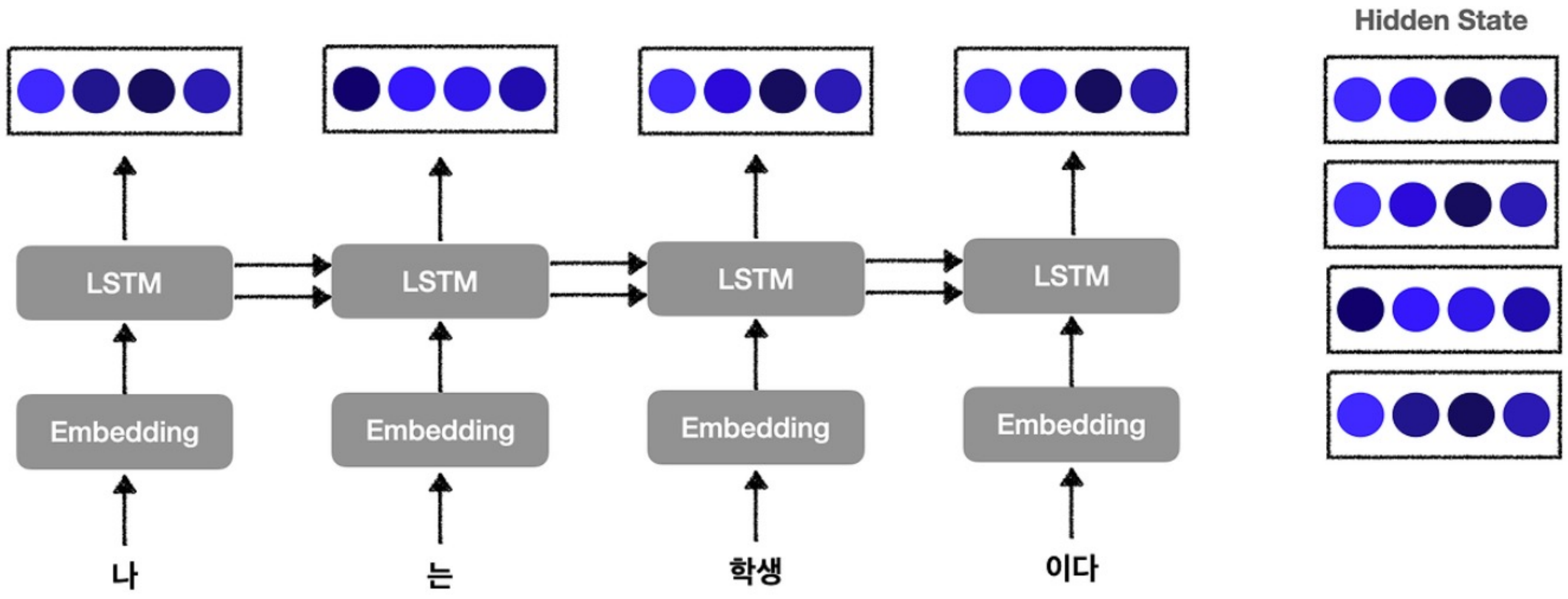


Seq2seq model

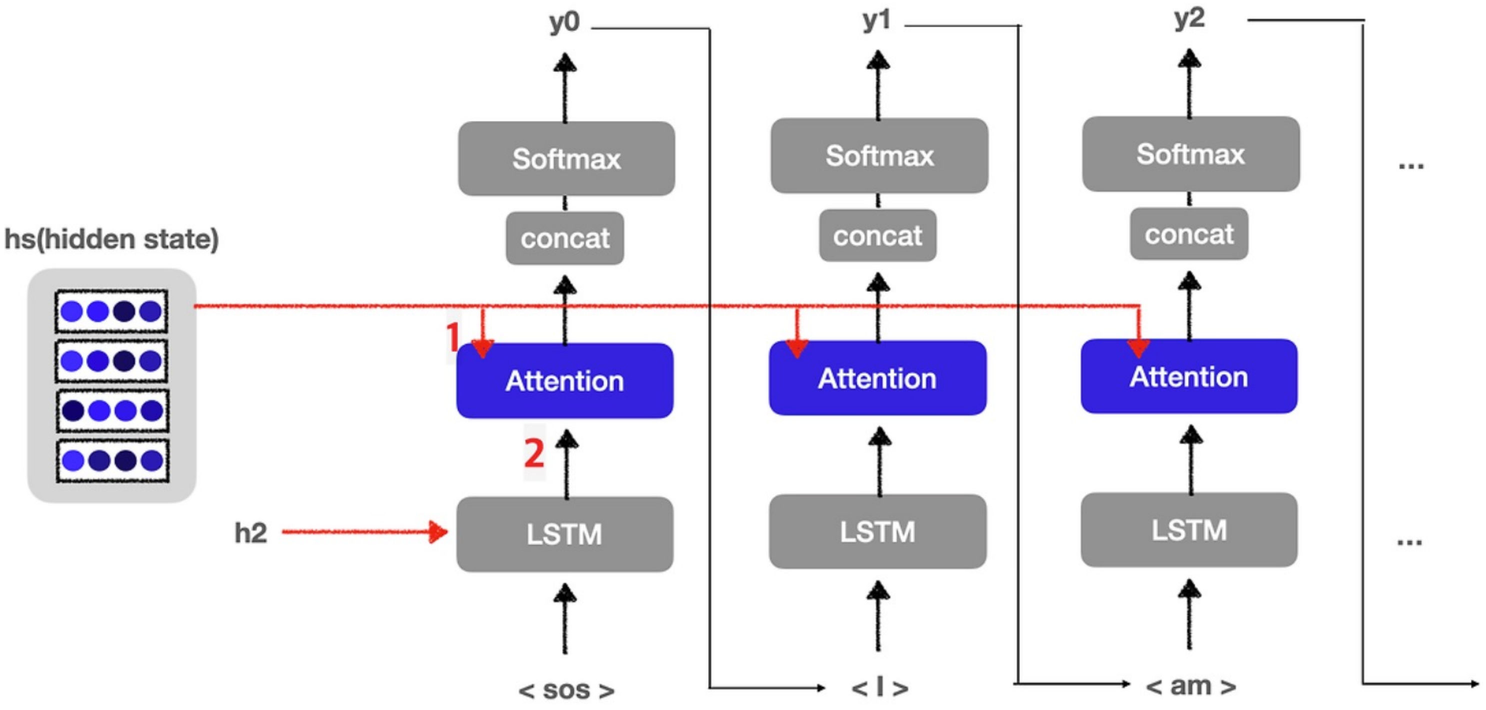


Attention Mechanism Example Continued

2-1. Attention Mechanism에서의 Encoder

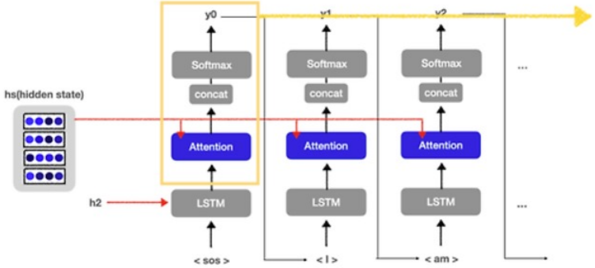


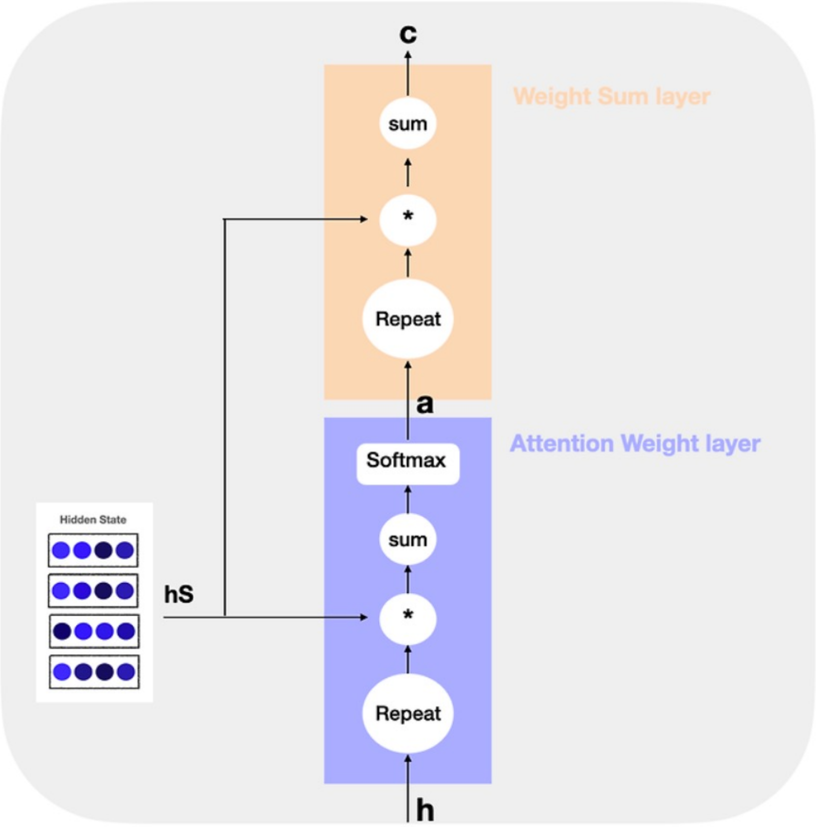
Attention Mechanism Example Continued



Attention을 포함한 decoder 구조

Attention Mechanism Example Continued





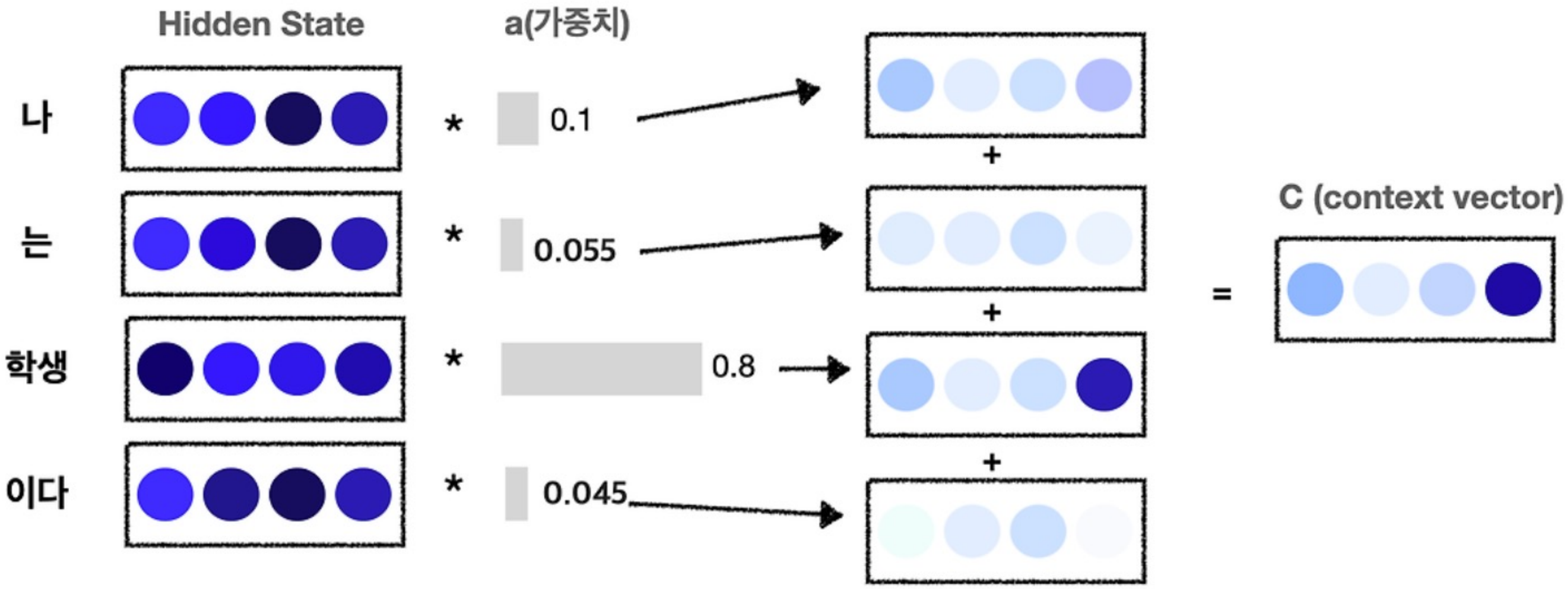
Attention layer

Attention Mechanism Example Continued

나	h0	=	1	0	0	1	*	0.1	=	0.1	0	0	0.1
는	h1	=	1	0	0	2	*	0.055	=	0.055	0	0	0.11
학생	h2	=	1	1	2	0	*	0.8	=	0.8	0.8	1.6	0
이다	h3	=	1	1	0	0	*	0.045	=	0.045	0.045	0	0

At this point, the output values from the Attention Weight layer are (0.1 0 0 0 0.1) + (0.055 0 0 0.11) + (0.8 0.8 1.6 0) + (0.045 0.045 0 0)

Weight sum layer



각 단어의 가중치와 단어 벡터의 가중합(weighted sum)

Types of Attention Mechanisms

1. Bahdanau Attention
2. Self-Attention
3. Scaled Dot-Production Attention
4. Luong Attention
5. Multi-Head Attention

$$e_{ij} = {v_a}^\top \tanh(W_a s_{i-1} + U_a h_j)$$

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

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

EQ4ALL

EQual access to technology makes all even better.

Transformer model

Introduced by Vaswani et al. in 2017, the transformer model uses attention mechanisms exclusively.

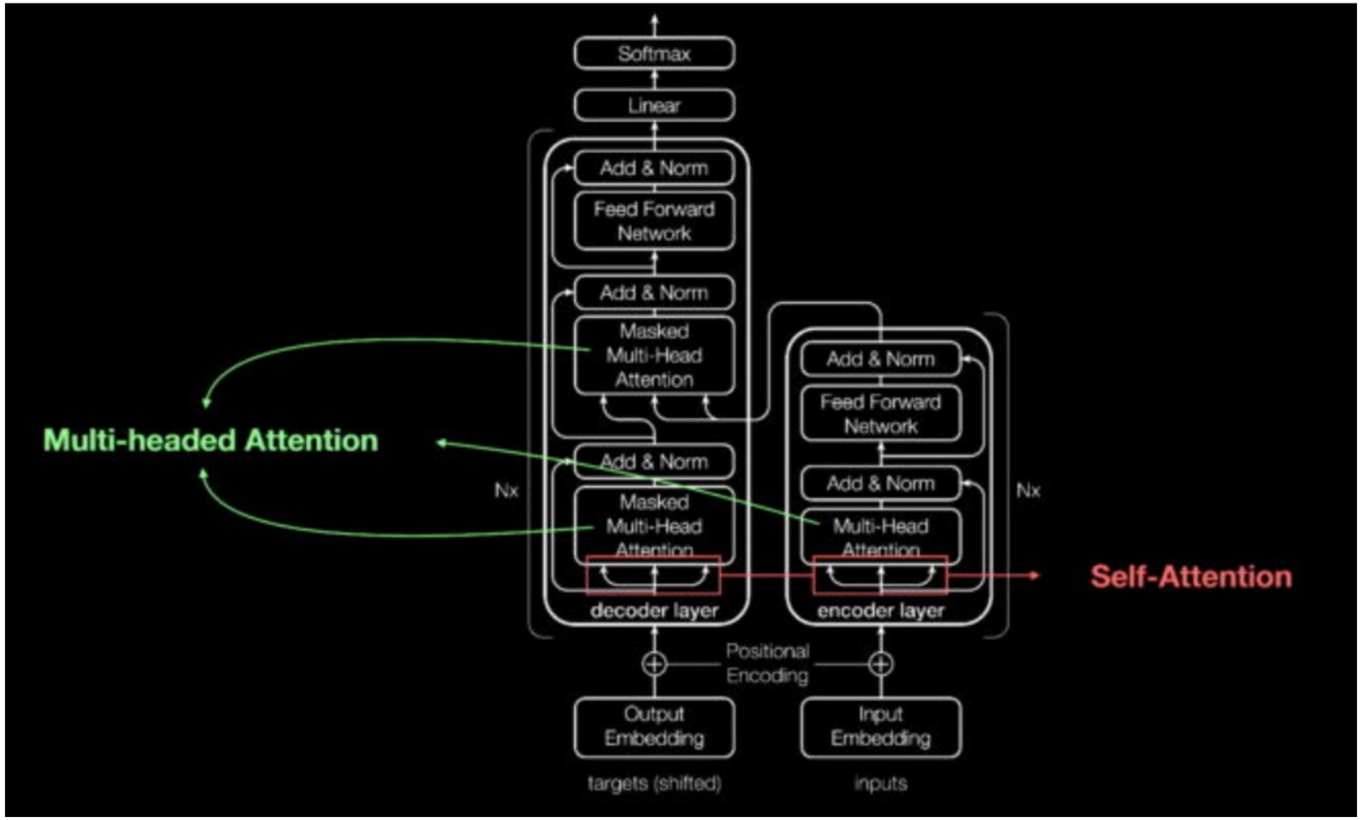
Key features: parallel processing, scalability, and efficiency

Attention is All You Need

- Transformer model is a sequence-to-sequence transformation model based on the attention mechanism
- Consists of encoder and a decoder, each composed of multiple layers of self-attention and feed-forward neural networks.
- Achieves high performance without using recurrent neural networks (RNNs) or convolutional neural networks (CNNs).
- Demonstrates excellent performance in various natural language Processing models

Continuous Research : Ongoing advancements and enhancements within the field

Broad Potential: Extensive applicability across diverse AI and machine learning domains.



EQ4ALL
EQual access to technology makes all even better.

Self-Attention in Transformers

Description: Each word in the input sequence attends to all other words, enabling the model to understand context more effectively.

Multi-Head Attention: Divides the attention mechanism into multiple parts, allowing the model to focus on different aspects of the input.

Attention Based Model

“She poured water from the kettle into the cup until it was empty”

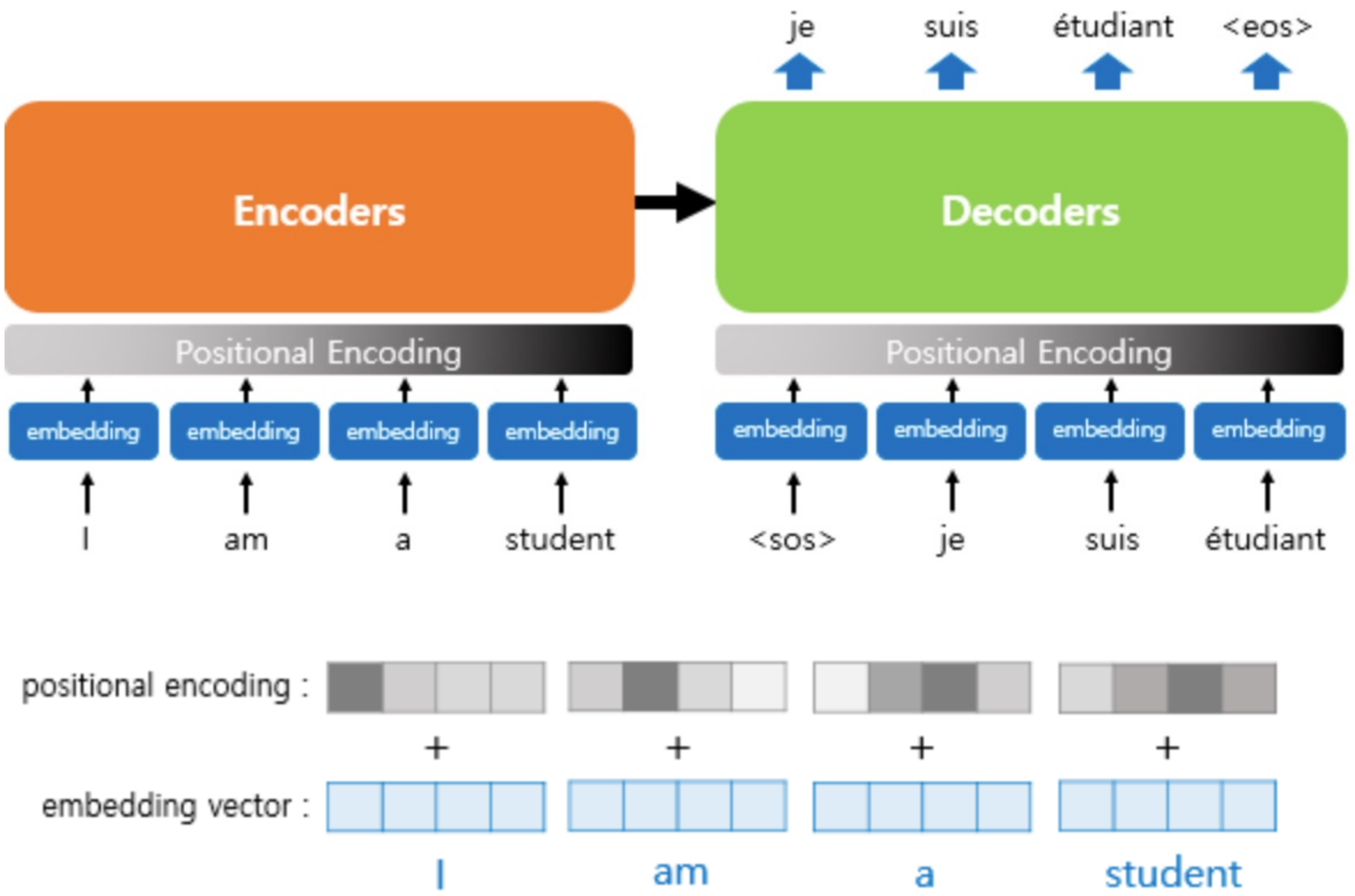
‘it’ = cup

“She poured water from the kettle into the cup until it was empty”

‘it’ = kettle

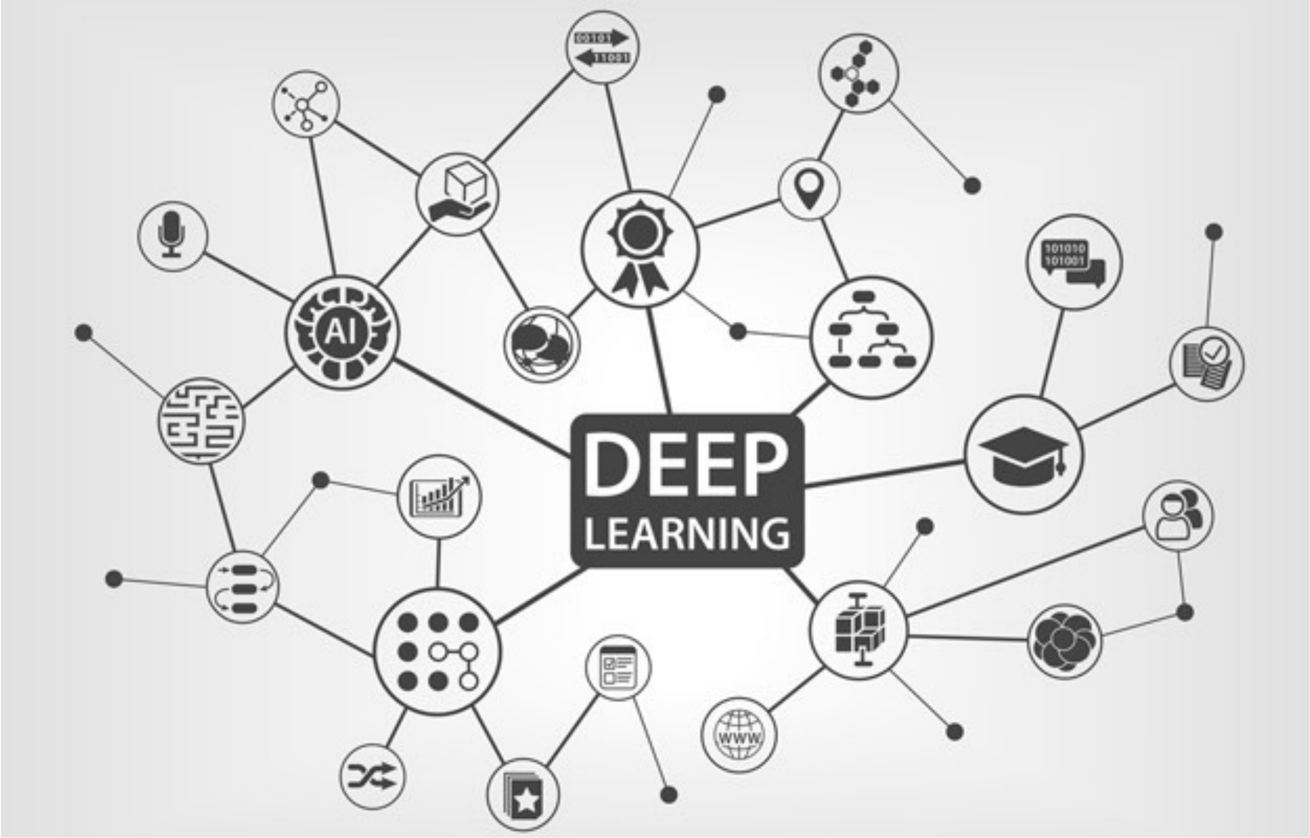
Encoder and Decoder stacks

The Transformer model generates output by attending to every position in the input sequence through position encoding, multi-head self attention, and position-wise feed-forward networks.

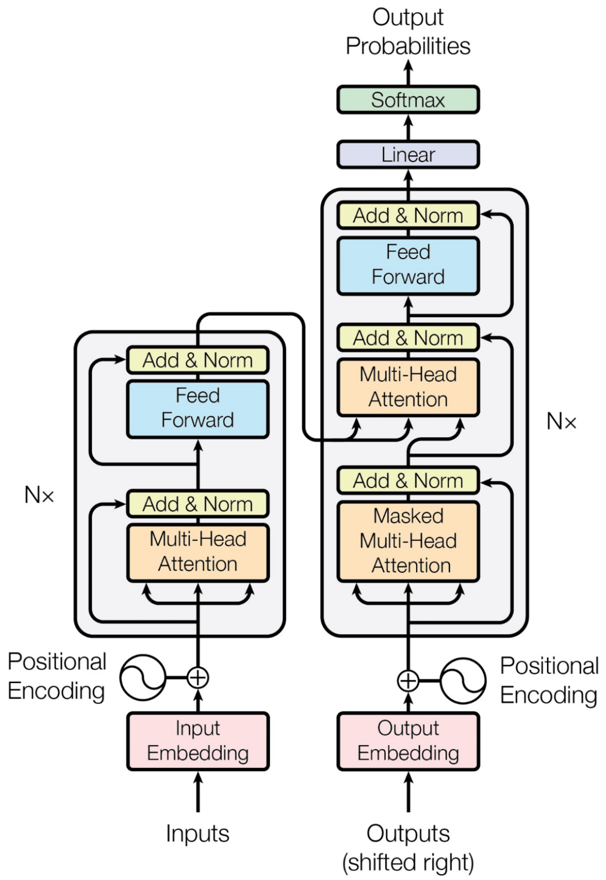


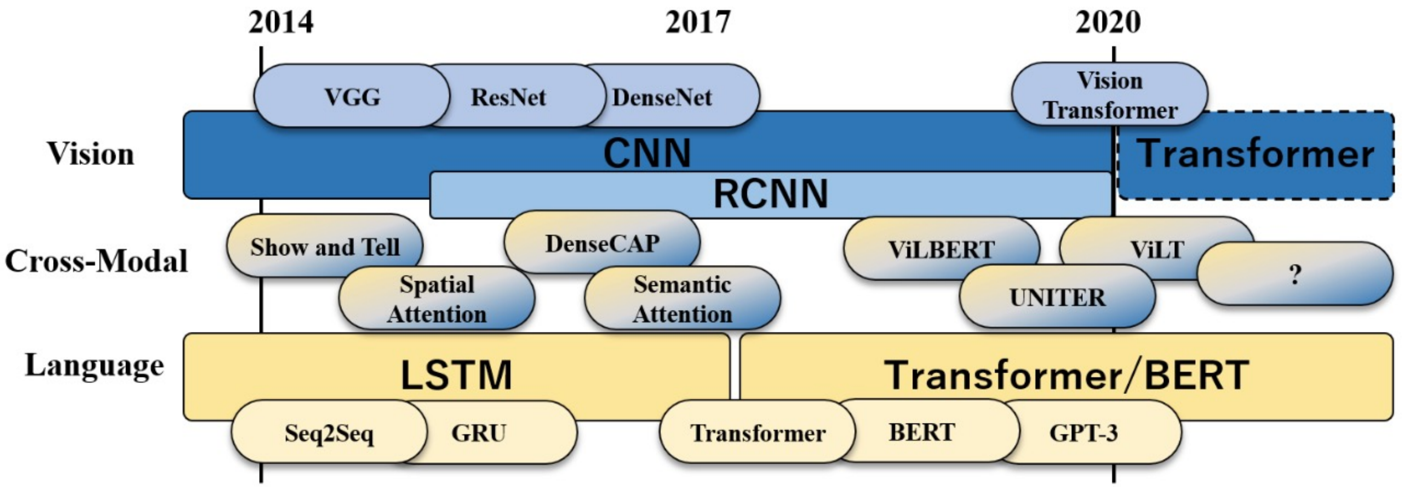
Advantages of Attention Mechanism

- Performance Improvement: Enhanced accuracy in language understanding and generation
- Context Awareness: Better handling of long-range dependencies within sequences
- Scalability: Efficient processing of large datasets.



Transformer vs Traditional Models





EQ4ALL
EQual access to technology makes all even better.

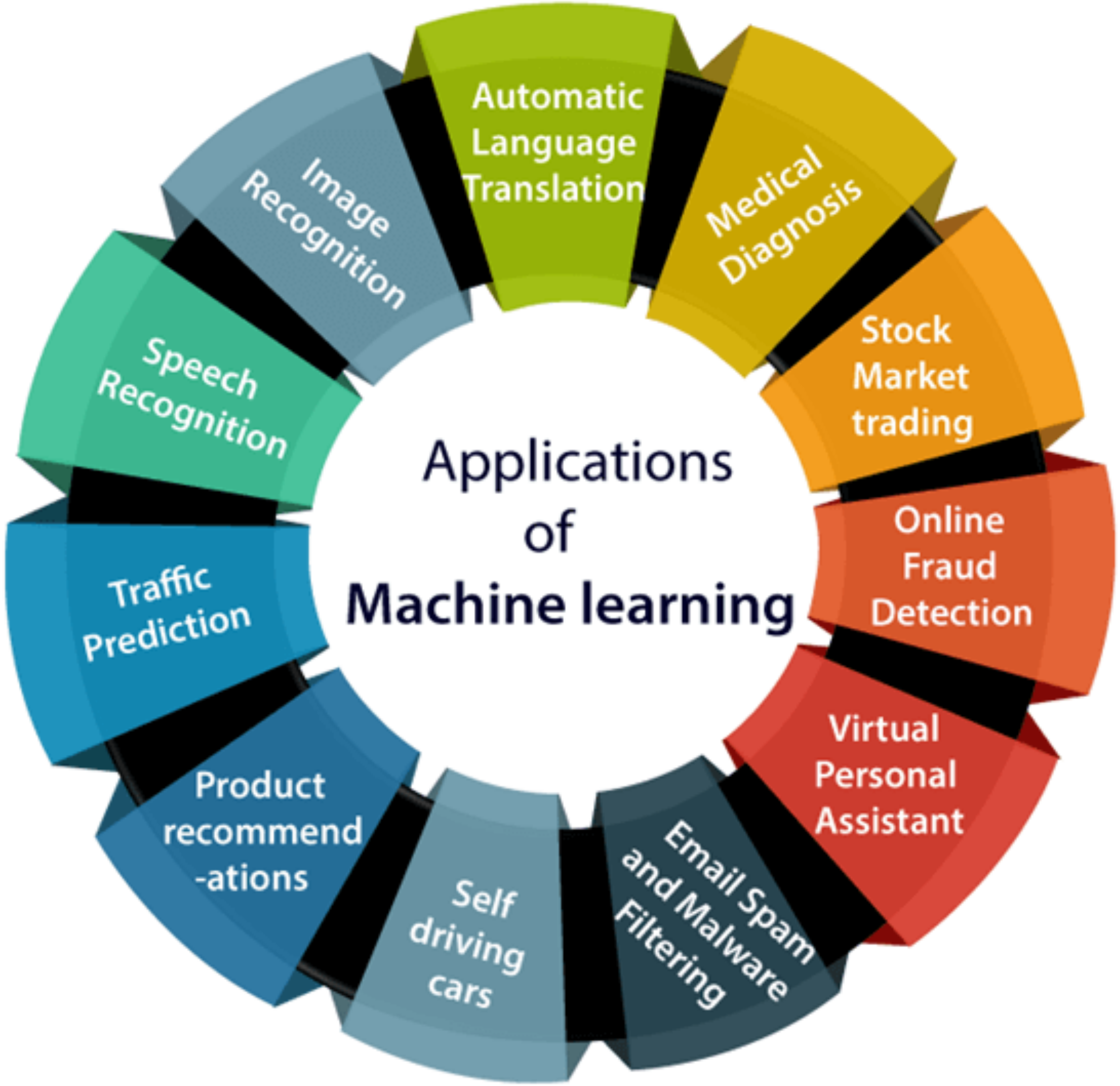
Conclusion

Significance

Transformative Impact: Emphasizing the groundbreaking influence of deep learning and attention mechanisms on NLP and other fields.

Model	BLEU		Training Cost (in FLOPS * 10^{18})	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet (Kalchbrenner et al., 2016)	23.75			
Deep-Att + PosUnk (Zhou et al., 2016)		39.2		100
GNMT + RL (Wu et al., 2016)	24.6	39.92	23	140
ConvS2S (Gehring et al., 2017)	25.16	40.46	9.6	150
MoE (Shazeer et al., 2017)	26.03	40.56	20	120
GNMT + RL Ensemble (Wu et al., 2016)	26.30	41.16	180	1100
ConvS2S Ensemble (Gehring et al., 2017)	26.36	**41.29**	77	1200
Transformer (base model)	27.3	38.1	**3.3**	
Transformer (big)	**28.4**	**41.8**	23	

Future Diretions



EQ4ALL
EQual access to technology makes all even better.