Sequences Part II: Attention & Transformers

COMP 551 Applied Machine Learning

Isabeau Prémont-Schwarz
School of Computer Science
McGill University

Fall 2024

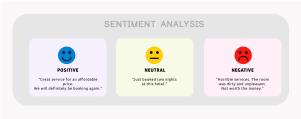
McGill
School of Computer Science

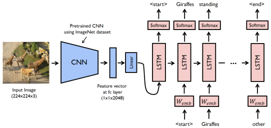
Learning Objectives

- Attention Mechanism
  - Motivation
  - How it works
  - Self-Attention
  - Multihead attention
- Positional Encoding
- Transformers
  - Encoder
  - Decoder
- Fine-tuning

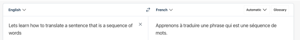
Types of Tasks with Sequences

- Seq2Vec (sequence classification)

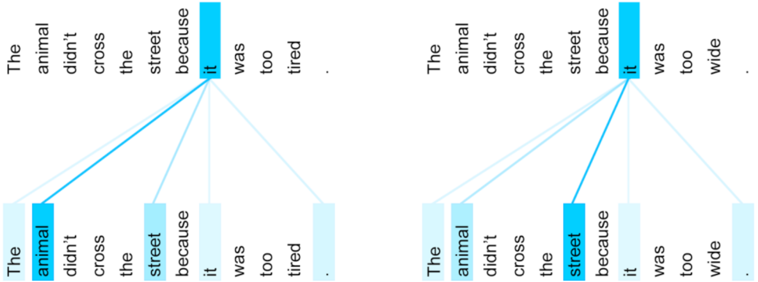




- Vec2Seq (sequence generation)
- Seq2Seq (sequence translation)

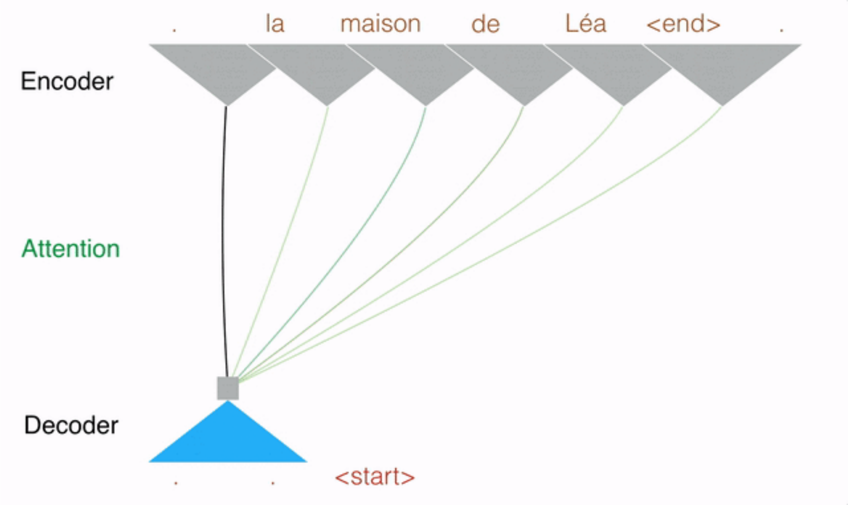


The Idea



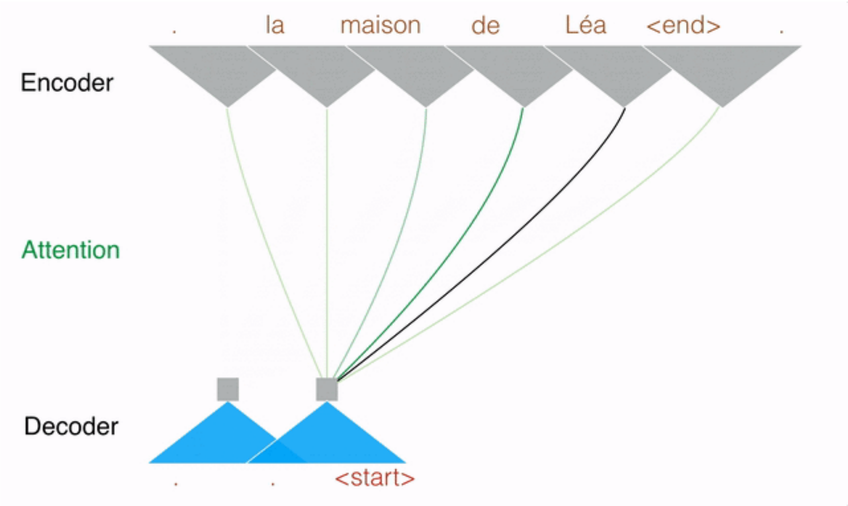
source

The Idea



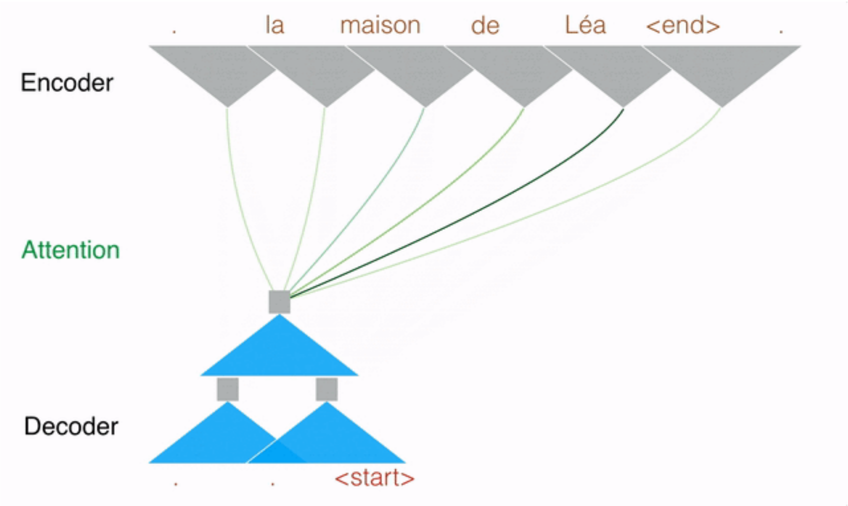
source

The Idea



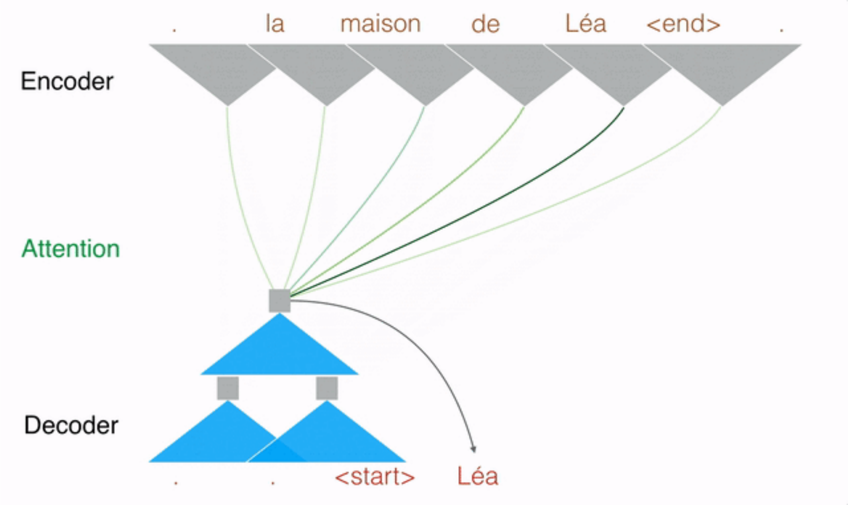
source

The Idea



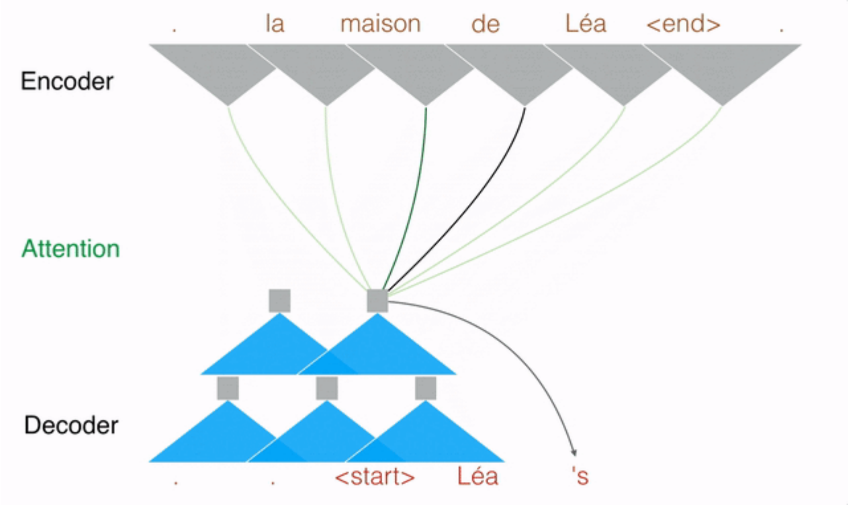
source

The Idea



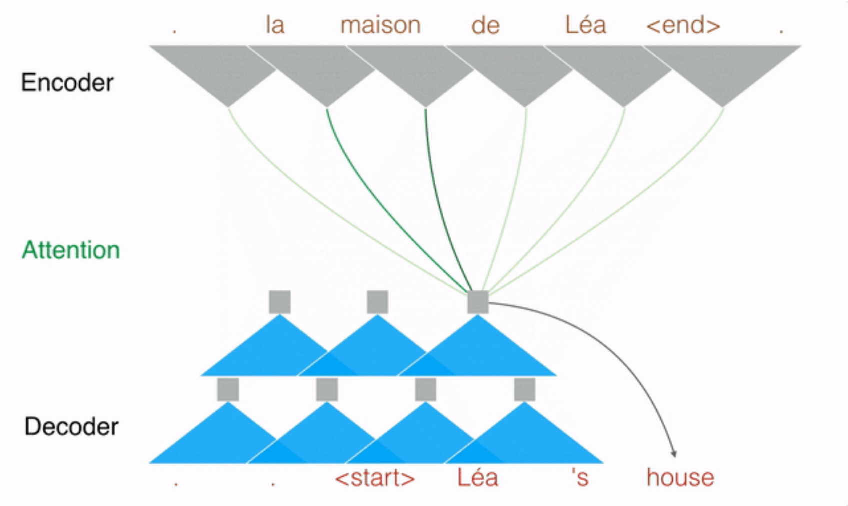
source

The Idea



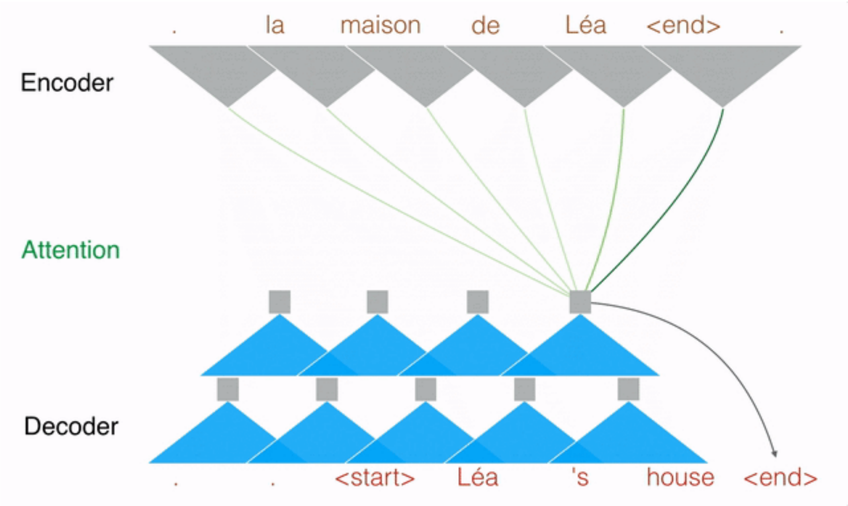
source

The Idea



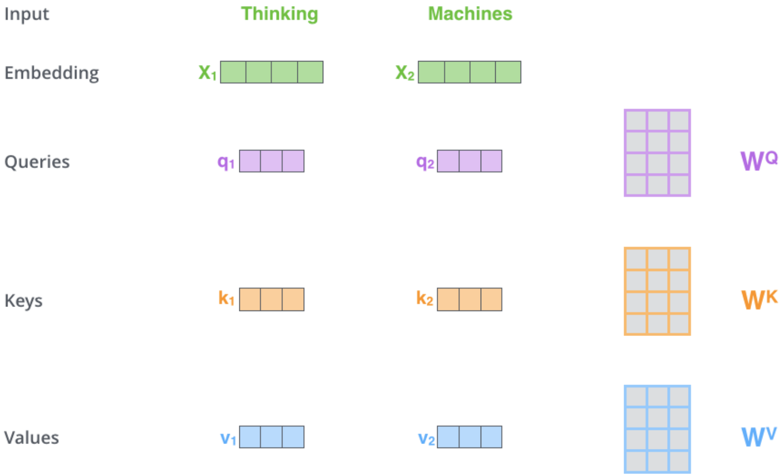
source

The Idea



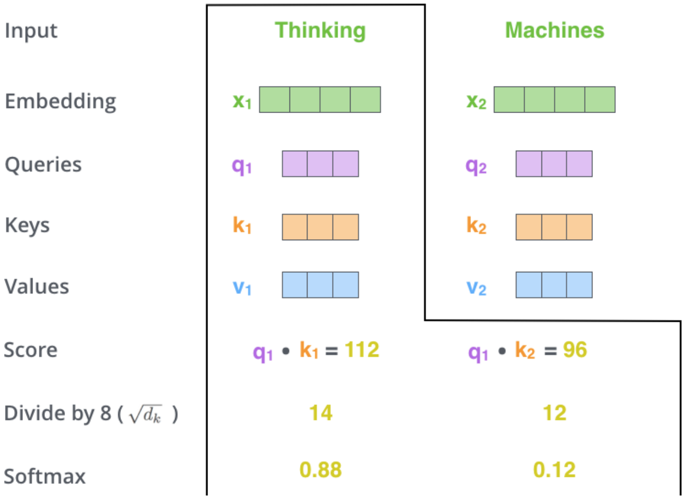
source

Attention Mechanism – High Level



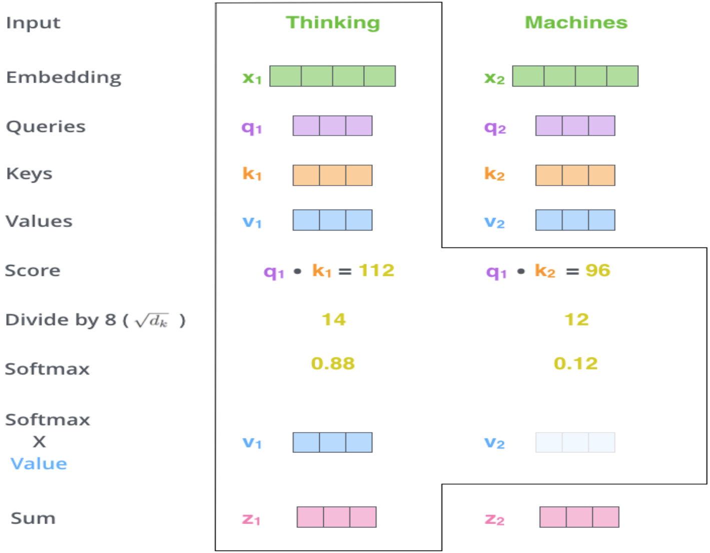
source

Attention Mechanism – High Level



source

Attention Mechanism – High Level



Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \ldots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \ldots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Query vector: $q_s = \underbrace{W^Q}_{d \times m} \cdot y_s \in \mathbb{R}^d$.

Key vector: $k_t = \underbrace{W^K}_{d \times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r \times n} \cdot x_t \in \mathbb{R}^r$.

Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Query vector: $q_s = \underbrace{W^Q}_{d \times m} \cdot y_s \in \mathbb{R}^d$.

Key vector: $k_t = \underbrace{W^K}_{d \times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r \times n} \cdot x_t \in \mathbb{R}^r$.

The more q_s is aligned with k_t, the more the token at time s pays attention to the token at time t. The attention paid by the token at time s to the token at time t is

$$a_{st} = \frac{\exp(q_s \cdot k_t/\sqrt{d})}{\sum_{i=1}^{T} \exp(q_s \cdot k_t/\sqrt{d})}$$

Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \ldots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \ldots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

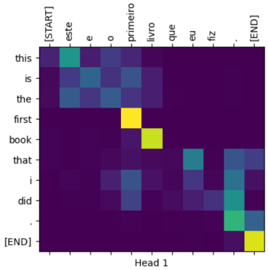
Query vector: $q_s = \underbrace{W^Q}_{d\times m} \cdot y_s \in \mathbb{R}^d$.

Key vector: $k_t = \underbrace{W^K}_{d\times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r\times n} \cdot x_t \in \mathbb{R}^r$.

$$a_{st} = \frac{\exp(q_s \cdot k_t / \sqrt{d})}{\sum_{i=1}^{T} \exp(q_s \cdot k_t / \sqrt{d})}$$

The more q_s is aligned with k_t, the more the token at time s pays attention to the token at time t. The attention paid by the token at time s to the token at time t is



Attention Mechanism – Math

Let $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ be a sequence of vectors being attended to, and $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ being a sequence of vectors doing the attending. Then we have that:

Query vector: $q_s = \underbrace{W^Q}_{d \times m} \cdot y_s \in \mathbb{R}^d$.

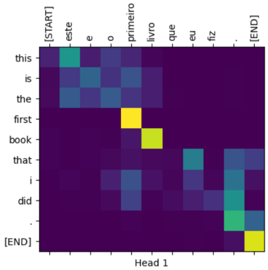
Key vector: $k_t = \underbrace{W^K}_{d \times n} \cdot x_t \in \mathbb{R}^d$.

Value vector: $v_t = \underbrace{W^V}_{r \times n} \cdot x_t \in \mathbb{R}^r$.

$$a_{st} = \frac{\exp(q_s \cdot k_t / \sqrt{d})}{\sum_{i=1}^{T} \exp(q_s \cdot k_t / \sqrt{d})}$$

$$z_s = \sum_{t=1}^{T} a_{st} v_t$$

The more q_s is aligned with k_t, the more the token at time s pays attention to the token at time t. The attention paid by the token at time s to the token at time t is



Attention Mechanism – Math

$$\underbrace{X}_{T \times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \qquad \underbrace{Y}_{S \times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \qquad \underbrace{Z}_{S \times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

Attention Mechanism – Math

$$\underbrace{X}_{T\times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \qquad \underbrace{Y}_{S\times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \qquad \underbrace{Z}_{S\times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

$$\underbrace{Q}_{S\times d} = Y \cdot {W^Q}^\top = \begin{bmatrix} q_1 \\ \cdots \\ q_S \end{bmatrix}, \qquad \underbrace{K}_{T\times d} = X \cdot {W^K}^\top = \begin{bmatrix} k_1 \\ \cdots \\ k_T \end{bmatrix}, \quad \underbrace{V}_{T\times r} = X \cdot {W^V}^\top = \begin{bmatrix} v_1 \\ \cdots \\ v_T \end{bmatrix}$$

Attention Mechanism – Math

$$\underbrace{X}_{T\times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \qquad \underbrace{Y}_{S\times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \qquad \underbrace{Z}_{S\times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

$$\underbrace{Q}_{S\times d} = Y \cdot {W^Q}^\top = \begin{bmatrix} q_1 \\ \cdots \\ q_S \end{bmatrix}, \qquad \underbrace{K}_{T\times d} = X \cdot {W^K}^\top = \begin{bmatrix} k_1 \\ \cdots \\ k_T \end{bmatrix}, \quad \underbrace{V}_{T\times r} = X \cdot {W^V}^\top = \begin{bmatrix} v_1 \\ \cdots \\ v_T \end{bmatrix}$$

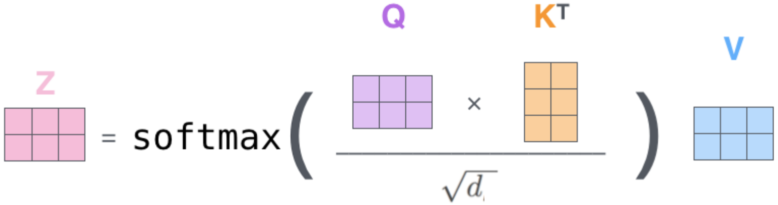
$$\underbrace{A}_{S\times T} = \text{softmax}_{\text{dim}=\text{T}}(Q \cdot K^\top / \sqrt{d}), \qquad \underbrace{Z}_{S\times r} = A \cdot V$$

Attention Mechanism – Math

$$\underbrace{X}_{T\times n} = \begin{bmatrix} x_1 \\ \cdots \\ x_T \end{bmatrix}, \quad \underbrace{Y}_{S\times m} = \begin{bmatrix} y_1 \\ \cdots \\ y_S \end{bmatrix}, \quad \underbrace{Z}_{S\times r} = \begin{bmatrix} z_1 \\ \cdots \\ z_S \end{bmatrix},$$

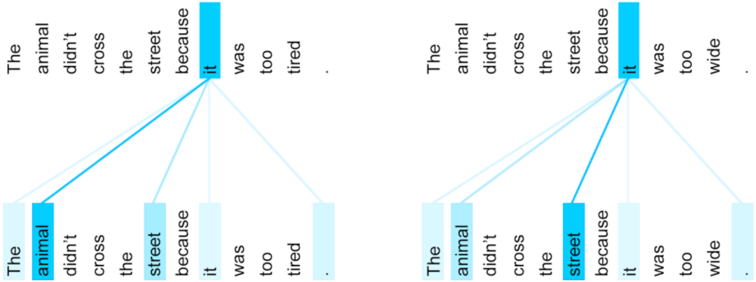
$$\underbrace{Q}_{S\times d} = Y \cdot {W^Q}^\top = \begin{bmatrix} q_1 \\ \cdots \\ q_S \end{bmatrix}, \quad \underbrace{K}_{T\times d} = X \cdot {W^K}^\top = \begin{bmatrix} k_1 \\ \cdots \\ k_T \end{bmatrix}, \quad \underbrace{V}_{T\times r} = X \cdot {W^V}^\top = \begin{bmatrix} v_1 \\ \cdots \\ v_T \end{bmatrix}$$

$$\underbrace{A}_{S\times T} = \mathrm{softmax}_{\mathrm{dim}=\mathrm{T}}(Q \cdot K^\top / \sqrt{d}), \quad \underbrace{Z}_{S\times r} = A \cdot V$$



Self-Attention

If the sequence $\{x_t \in \mathbb{R}^n | t \in \{1, \dots T\}\}$ being attended to, is the same as the sequence $\{y_s \in \mathbb{R}^m | s \in \{1, \dots S\}\}$ doing the attending, i.e. if $X = Y$, we call it **SELF-ATTENTION.**



Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$W^K = \mathbb{I}, \quad , W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$z =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$W^K = \mathbb{I}, \quad , W^V = \begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 1 & 0\end{bmatrix}, \quad W^Q = \begin{bmatrix}0 & 1 & 0\\1 & 0 & 0\\0 & 0 & 1\\1 & 0 & -1\end{bmatrix}.$$

$$k_t = \left\{ \mathbb{I}\begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \mathbb{I}\begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \mathbb{I}\begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\} = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad W^V = \begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 1 & 0\end{bmatrix}, \quad W^Q = \begin{bmatrix}0 & 1 & 0\\1 & 0 & 0\\0 & 0 & 1\\1 & 0 & -1\end{bmatrix}.$$

$$v_t =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t = \left\{ W^V \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad W^V \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \quad \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad W^V = \begin{bmatrix}1 & 0 & 0 & 0\\0 & 1 & 1 & 0\end{bmatrix}, \quad W^Q = \begin{bmatrix}0 & 1 & 0\\1 & 0 & 0\\0 & 0 & 1\\1 & 0 & -1\end{bmatrix}.$$

$$v_t = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \quad \begin{bmatrix}0\\1\end{bmatrix}, \quad W^V \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad W^V = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix} \right\}, \quad W^Q = \begin{bmatrix}0 & 1 & 0\\1 & 0 & 0\\0 & 0 & 1\\1 & 0 & -1\end{bmatrix}.$$

$$q =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \quad W^Q = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & -1 \end{bmatrix}.$$

$$q = W^Q y = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad y = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\}, \quad q = \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}.$$

$$q \cdot k_t =$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix} \right\}, \quad q = \begin{bmatrix}-1\\1\\0\\1\end{bmatrix}.$$

$$q \cdot k_t = [-1, 1, -1]$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix} \right\}, \quad q = \begin{bmatrix}-1\\1\\0\\1\end{bmatrix}.$$

$$q \cdot k_t = [-1, 1, -1], \quad \text{softmax}(q \cdot k_t/\sqrt{d}) = \frac{[e^{-1/2}, e^{+1/2}, e^{-1/2}]}{e^{-1/2} + e^{+1/2} + e^{-1/2}}$$

Attention Mechanism – Example

$$x_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad y = \begin{bmatrix}1\\-1\\0\end{bmatrix}.$$

$$k_t = \left\{ \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \begin{bmatrix}0\\1\\0\\0\end{bmatrix}, \begin{bmatrix}1\\0\\0\\0\end{bmatrix} \right\}, \quad v_t = \left\{ \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix} \right\}, \quad q = \begin{bmatrix}-1\\1\\0\\1\end{bmatrix}.$$

$$\text{softmax}(q \cdot k_t/\sqrt{d}) = \frac{[e^{-1/2}, e^{+1/2}, e^{-1/2}]}{e^{-1/2} + e^{+1/2} + e^{-1/2}}$$

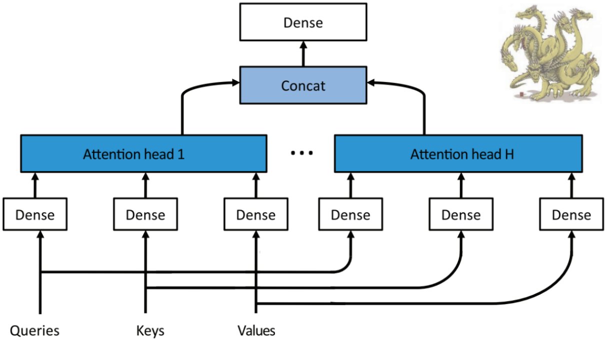
$$z = \frac{e^{-1/2}\begin{bmatrix}1\\0\end{bmatrix} + e^{+1/2}\begin{bmatrix}0\\1\end{bmatrix} + e^{-1/2}\begin{bmatrix}1\\0\end{bmatrix}}{e^{-1/2} + e^{+1/2} + e^{-1/2}} = \begin{bmatrix}2e^{-1/2}\\e^{+1/2}\end{bmatrix} /(2e^{-1/2} + e^{+1/2})$$

Attention Mechanism – Multi-Head Attention

Besides going deep, we can also go wide by having MHA with each head indexed by $i \in \{1, \ldots, H\}$:

$$\text{Query:} \quad \mathbf{Q}_i = \mathbf{X}\mathbf{W}_i^{(q)}; \quad \text{Key:} \quad \mathbf{K}_i = \mathbf{X}\mathbf{W}_i^{(k)}; \quad \text{Value:} \quad \mathbf{V}_i = \mathbf{X}\mathbf{W}_i^{(v)}$$

$$\text{Self-attention:} \quad \mathbf{A}_i = \mathrm{Attn}(\mathbf{Q}_i, \mathbf{K}_i); \quad \text{Word embedding:} \quad \mathbf{Z}_i = \mathbf{A}_i\mathbf{V}_i$$



We then concatenate the outputs from all H attention heads and feed them into a dense layer (i.e., a full-connected layer) to get the final embedding:

$$\mathbf{Z} = [\mathbf{Z}_1, \ldots, \mathbf{Z}_H]\mathbf{W}_o \in \mathbb{R}^{N \times K}$$

where $\mathbf{W}_o$ is a $(\sum_{i=1}^{H} K_i) \times K$ matrix.

Attention Mechanism – Multi-Head Attention

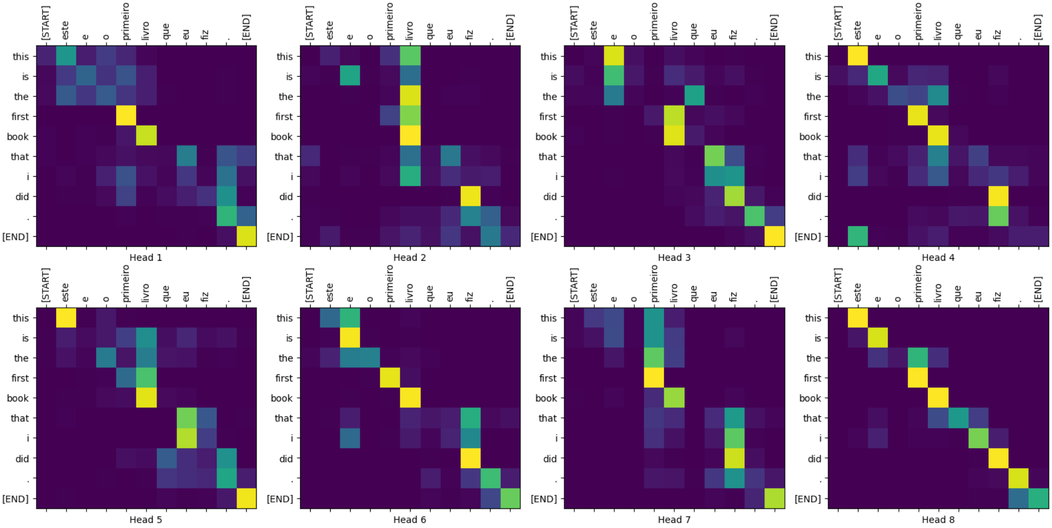


image source

Positional Encoding

- Attention is permutation invariant and therefore ignores the input word ordering. To make it aware of the word position we can use positional encoding (PE) to represent the sentence with N tokens with $P \in \mathbb{R}^{N \times D}$.

Positional Encoding

- Attention is permutation invariant and therefore ignores the input word ordering. To make it aware of the word position we can use positional encoding (PE) to represent the sentence with N tokens with $P \in \mathbb{R}^{N \times D}$.
- A commonly used PE is a set of sinusoidal basis functions. To encode the i$^t h$ word position using the j$^t h$ encoding dimension for $j \in \{0, \dots, D/2\}$, we have $p_{i,2j} = \sin\left(\frac{i}{C^{2j/D}}\right), \qquad p_{i,2j+1} = \cos\left(\frac{i}{C^{2j/D}}\right)$, where C (e.g., C $= 10, 000$) is some large constant that is not important here.

Positional Encoding

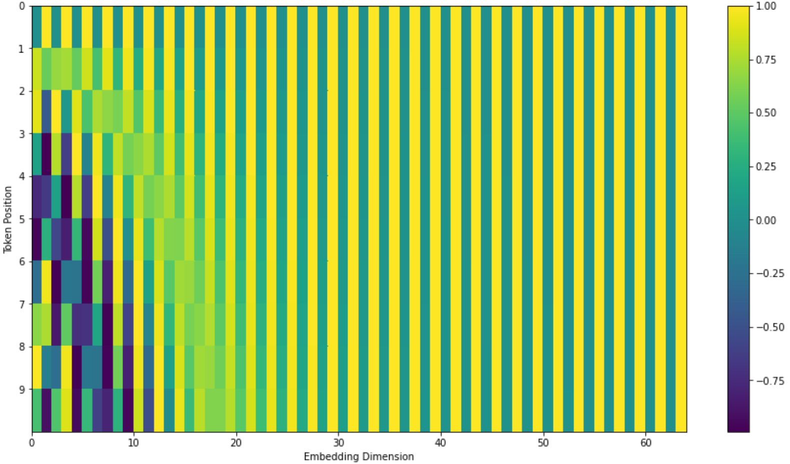
- Attention is permutation invariant and therefore ignores the input word ordering. To make it aware of the word position we can use positional encoding (PE) to represent the sentence with N tokens with $P \in \mathbb{R}^{N \times D}$.
- A commonly used PE is a set of sinusoidal basis functions. To encode the ith word position using the jth encoding dimension for $j \in \{0, \ldots, D/2\}$, we have $p_{i,2j} = \sin\left(\frac{i}{C^{2j/D}}\right), \qquad p_{i,2j+1} = \cos\left(\frac{i}{C^{2j/D}}\right)$, where C (e.g., C $=$ 10, 000) is some large constant that is not important here.
- For example, if D $=$ 4, the encoding for the ith token is: $p_i = \left[\sin\left(\frac{i}{C^{0/4}}\right), \cos\left(\frac{i}{C^{0/4}}\right), \sin\left(\frac{i}{C^{2/4}}\right), \cos\left(\frac{i}{C^{2/4}}\right)\right]$

Positional Encoding

- Attention is permutation invariant and therefore ignores the input word ordering. To make it aware of the word position we can use positional encoding (PE) to represent the sentence with N tokens with $P \in \mathbb{R}^{N \times D}$.
- A commonly used PE is a set of sinusoidal basis functions. To encode the i$^t h$ word position using the j$^t h$ encoding dimension for $j \in \{0, \ldots, D/2\}$, we have $p_{i,2j} = \sin\left(\frac{i}{C^{2j/D}}\right), \qquad p_{i,2j+1} = \cos\left(\frac{i}{C^{2j/D}}\right)$, where C (e.g., C = 10, 000) is some large constant that is not important here.
- For example, if D = 4, the encoding for the i$^t h$ token is: $p_i = \left[\sin\left(\frac{i}{C^{0/4}}\right), \cos\left(\frac{i}{C^{0/4}}\right), \sin\left(\frac{i}{C^{2/4}}\right), \cos\left(\frac{i}{C^{2/4}}\right)\right]$
- We can then combine the PE $P \in \mathbb{R}^{N \times D}$ with the original word embedding $Z \in \mathbb{R}^{N \times K}$ by either concatenating them $Z^* = [P, Z]$ or adding them $Z^* = P + Z$. For the latter, we will need to make sure $K = D$.

Positional Encoding

Positional encoding for 10 tokens and $D = 64$:



Transformers

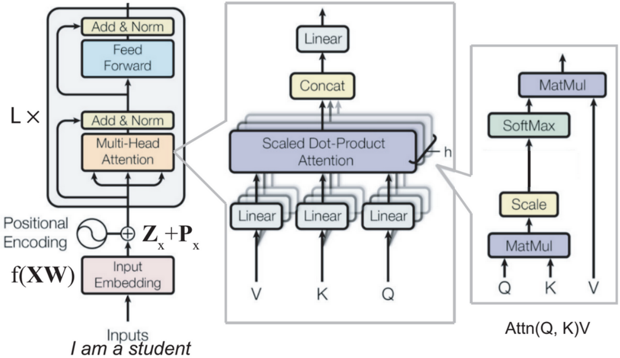
Attention Is All You Need

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

PDF of article: https://arxiv.org/pdf/1706.03762

Unless otherwise specified, images in this section are taken from the above paper.

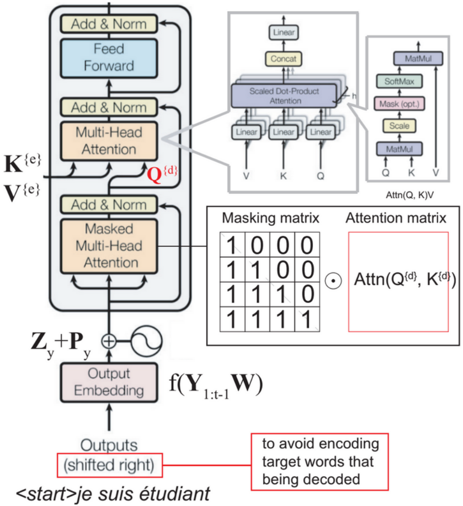
Encoder




```python
def EncoderBlock(X):
    Ze = LayerNorm(MultiHeadAttn(Q=X,K=X,V=X)
    ↪   + X)
    E = LayerNorm(FeedForward(Ze) + Ze)
    return E
```

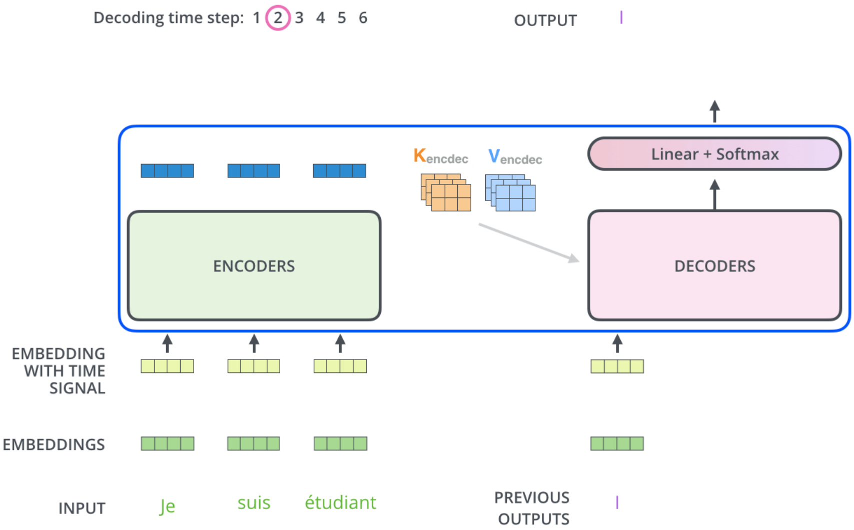
```python
def Encoder(X, L):
    Ze = POS(Embed(X))
    for l in range(L):
        Ze = EncoderBlock(Ze)
return Ze
```

Decoder



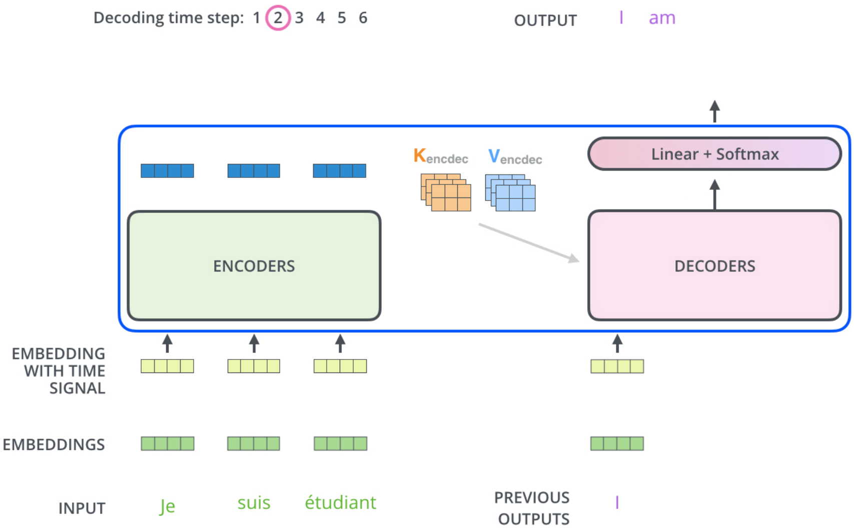

```
def DecoderBlock(Y, Ke, Ve):
    Zd = LayerNorm(
        MaskedMultiHeadAttn(Q=Y,K=Y,V=Y) + Y)
    Zd = LayerNorm(
        MultiHeadAttn(Q=Zd, K=Ke, V=Ve) + Zd)
    Zd = LayerNorm(FeedForward(Zd) + Zd)
    return Zd

def Decoder(Y, Ke, Ve, L):
    Zd = POS(Embed(Y))
    for l in range(L):
        Zd = DecoderBlock(Zd, Ke, Ve)
    return Zd
```

Decoder



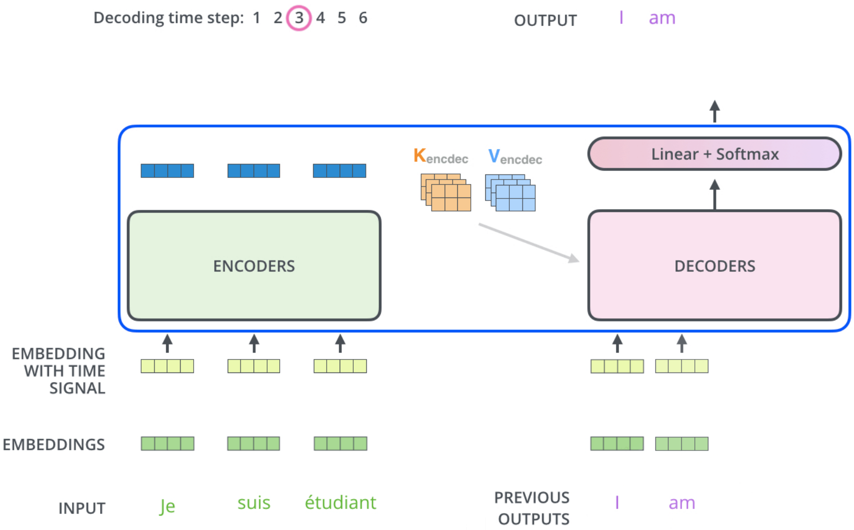
source: Jalammar Blog

Decoder



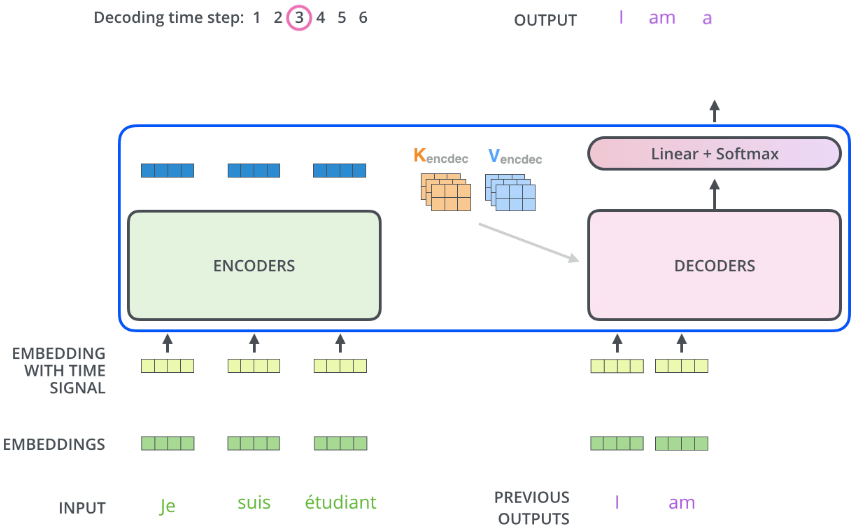
source: Jalammar Blog

Decoder



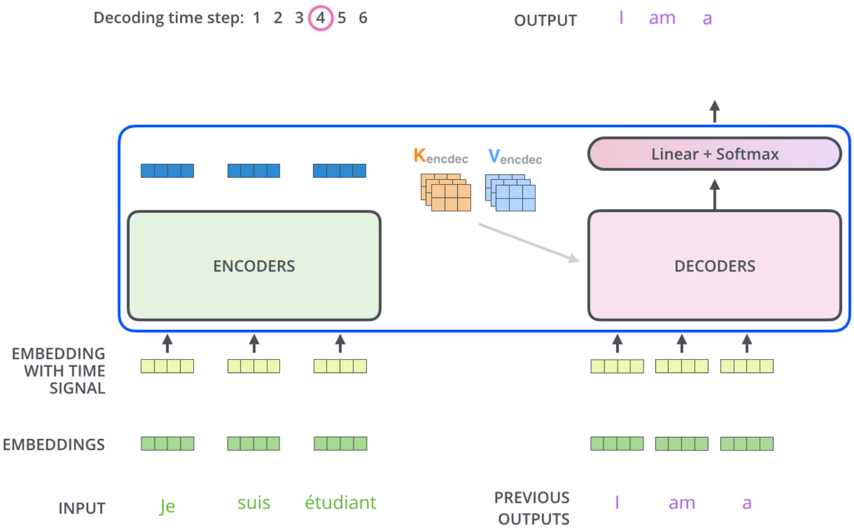
source: Jalammar Blog

Decoder



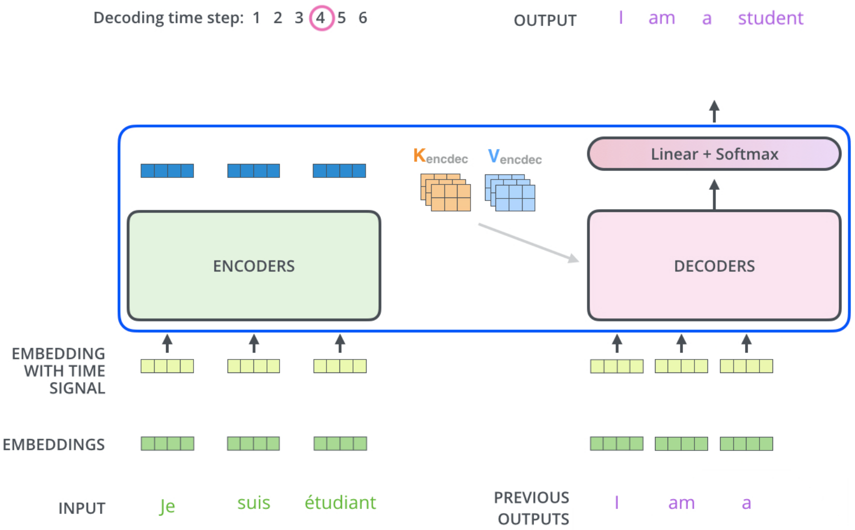
source: Jalammar Blog

Decoder



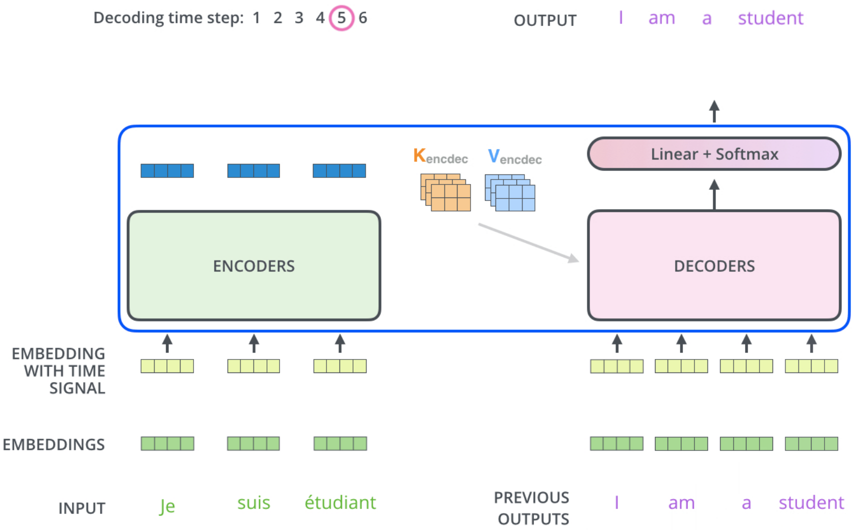
source: Jalammar Blog

Decoder



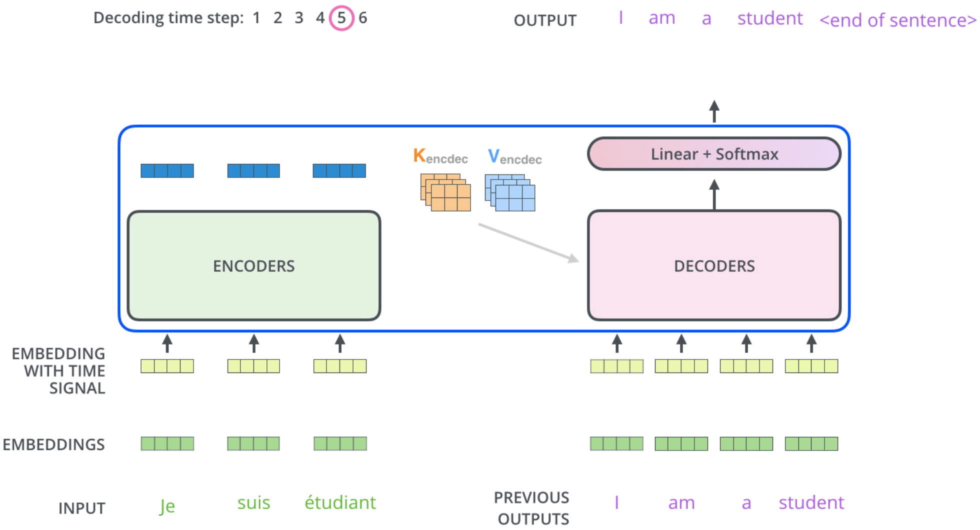
source: Jalammar Blog

Decoder



source: Jalammar Blog

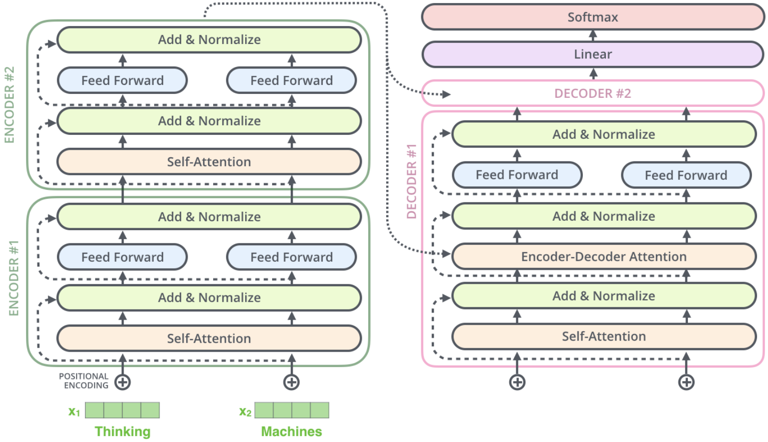
Decoder



source: Jalammar Blog

Vec2Seq and Seq2Vec ?

With a Transformer we saw how to do Seq2Seq. How can we do Vec2Seq or Seq2Vec?



Open AI Generative Pre-Trained Transformer (i.e. ChatGPT)

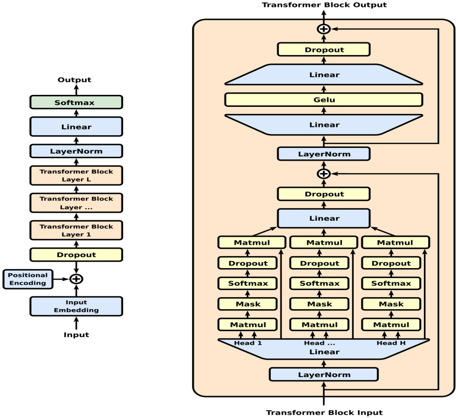


Figure: Open AI GPT architecture (source: Wikipedia)

Trained on next word(s) prediction.

Open AI GPT (i.e. ChatGPT)

Model	Architecture	Parameters	Training data	Release date	Training cost
GPT-1	12-level, 12-headed Transformer decoder (no encoder).	117 million	BookCorpus: 4.5 GB of text, from 7000 unpublished books.	June 11, 2018	30 days on 8 P600 GPUs, or 1 petaFLOP/s-day.
GPT-2	GPT-1, but with modified normalization	1.5 billion	WebText: 40 GB of text, 8 million documents, from 45 million webpages upvoted on Reddit.	February 14, 2019 (initial/limited version) and November 5, 2019 (full version)	"tens of petaflop/s-day", or 1.5e21 FLOP.
GPT-3	GPT-2, but with modification to allow larger scaling	175 billion	499 billion tokens of CommonCrawl (570 GB), WebText, English Wikipedia, book corpora (Books1 and Books2).	May 28, 2020	3640 petaflop/s-day.
GPT-3.5	Undisclosed	175 billion	Undisclosed	March 15, 2022	Undisclosed
GPT-4	Trained with both text prediction and RLHF.	Estimated 1.7 trillion.	Undisclosed	March 14, 2023	Undisclosed. Estimated 2.1 $\times$ 10 FLOP.

source: wiki

LLMs

- ELMo (Embeddings from Language Model)
  - RNN based, trained unsupervised to minimize the negative log likelihood of the input sentence, i.e. $y_t = x_{t-1}$
- BERT (Bidirectional Encoder Representations from Transformers)
  - Transformer-based: map a modified version of a sequence back to the unmodified form and compute the loss at the masked locations: fill-in-the-blank :

```
Let's make [MASK] chicken! [SEP] It [MASK] great with orange sauce
```

- GPT (Generative Pre-training Transformer)
  - uses a masked transformer as the decoder, see an open-source model here (20 billion parameters)

How to fine-tune?

Summary

- Attention Mechanism
  - $Z = \text{softmax}(Q \cdot K^\top / \sqrt{d}) \cdot V$
  - Self-Attention
  - Multihead attention
- Positional Encoding
- Transformers
  - Encoder (self-attention)
  - Decoder (self-attention to previous output + attention to encoder output)
- Fine-tuning