Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 7.1: Alignment and Representations

Louis-Philippe Morency

Objectives of today’s class

- Contextualized sentence embedding
- Transformer networks
  - Self-attention
  - Multi-head attention
  - Position embeddings
  - Sequence-to-sequence modeling
- Multimodal contextualized embeddings
- Language pre-training
  - BERT pre-training and fine-tuning
- Multimodal pre-training

Administrative Stuff

Upcoming Schedule

- First project assignment:
  - Proposal presentation (10/1 and 10/3)
  - First project report (Sunday 10/6)
- Midterm project assignment
  - Midterm presentations (11/5 and 11/7)
  - Midterm report (Sunday 11/10)
- Final project assignment
  - Final presentation (12/3 & 12/5)
  - Final report (Sunday 12/8)

Midterm Project Report Instructions

- **Goal:** Evaluate state-of-the-art models on your dataset and identify key issues through a detailed error analysis
  - It will inform the design of your new research ideas
- **Report format:** 8 pages, 2 column (ICML template)
  - The report should follow a similar structure to a research paper
- **Number of SOTA models**
  - Teams of 3 should have at least two baseline models
  - Teams of 4 or 5 should have at least three baseline models
- **Error analysis**
  - This is one of the most important part of this report. You need to understand where previous models can be improved.

Midterm Project Report Instructions

Main report sections:

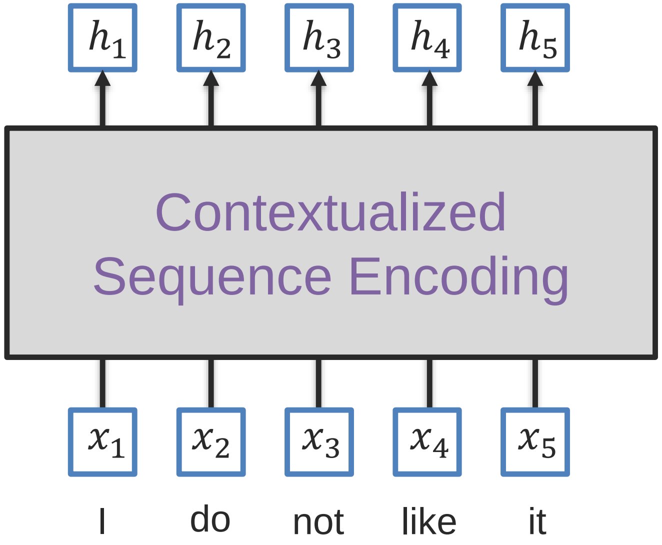
- Abstract
- Introduction
- Related work
- Problem statement
- Multimodal baseline models
- Experimental methodology
- Results and discussion
- New research ideas

Midterm Presentations

Oral Presentations OR Poster Presentations

Contextualized Sequence Encoding

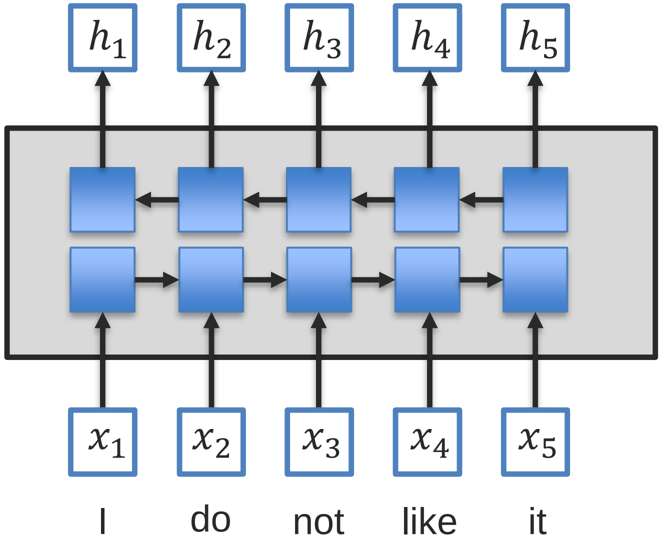
Sequence Encoding - Contextualization



How to encode this sequence while modeling the interaction between elements (e.g., words)?

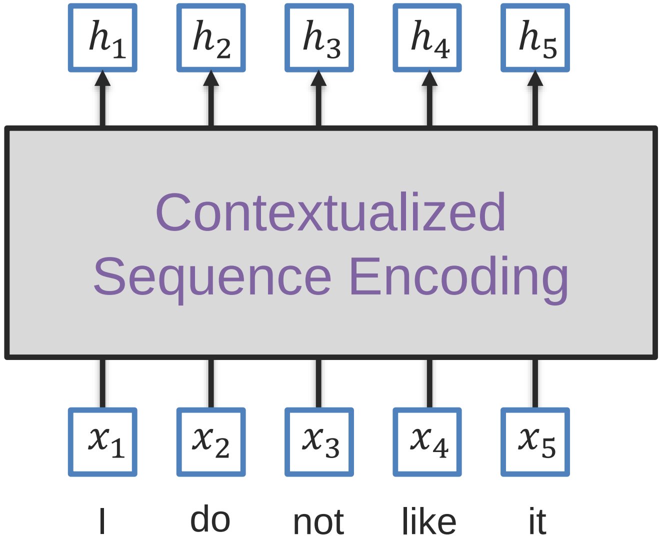
Option 1: Bi-directional LSTM:

(e.g., ELMO)

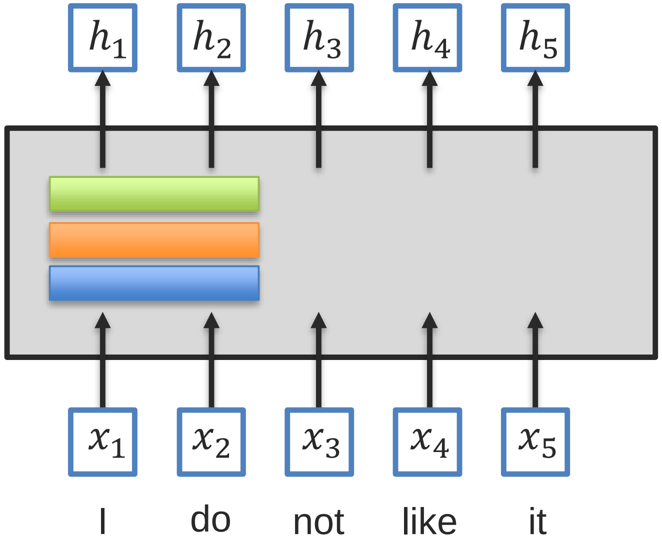


But harder to parallelize…

Sequence Encoding - Contextualization



Option 2: Convolutions



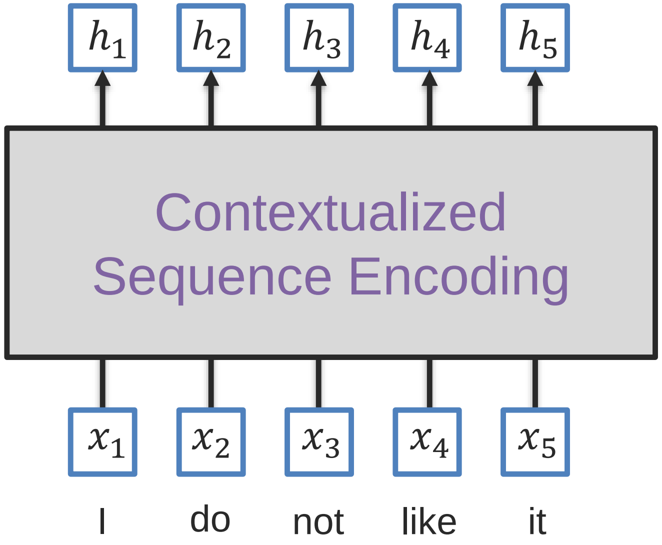
Can be parallelized!

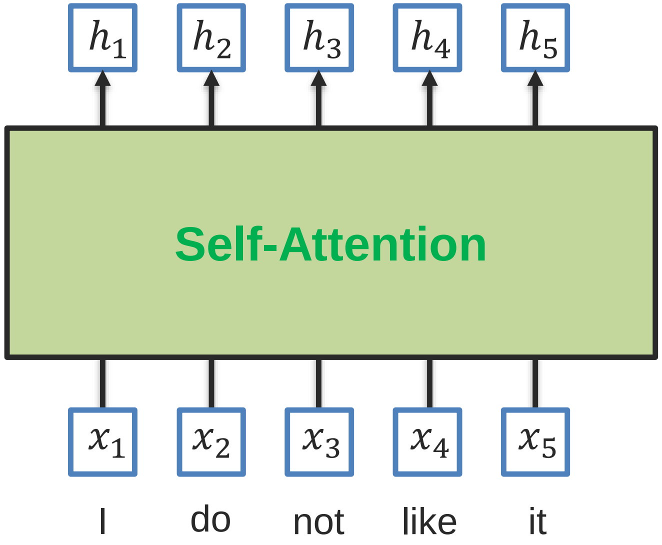
But modeling long-range dependencies require multiple layers

And convolutional kernels are static

Sequence Encoding - Contextualization

Option 3: Self-attention


h_1 h_2 h_3 h_4 h_5
Contextualized Sequence Encoding
x_1 x_2 x_3 x_4 x_5
I do not like it



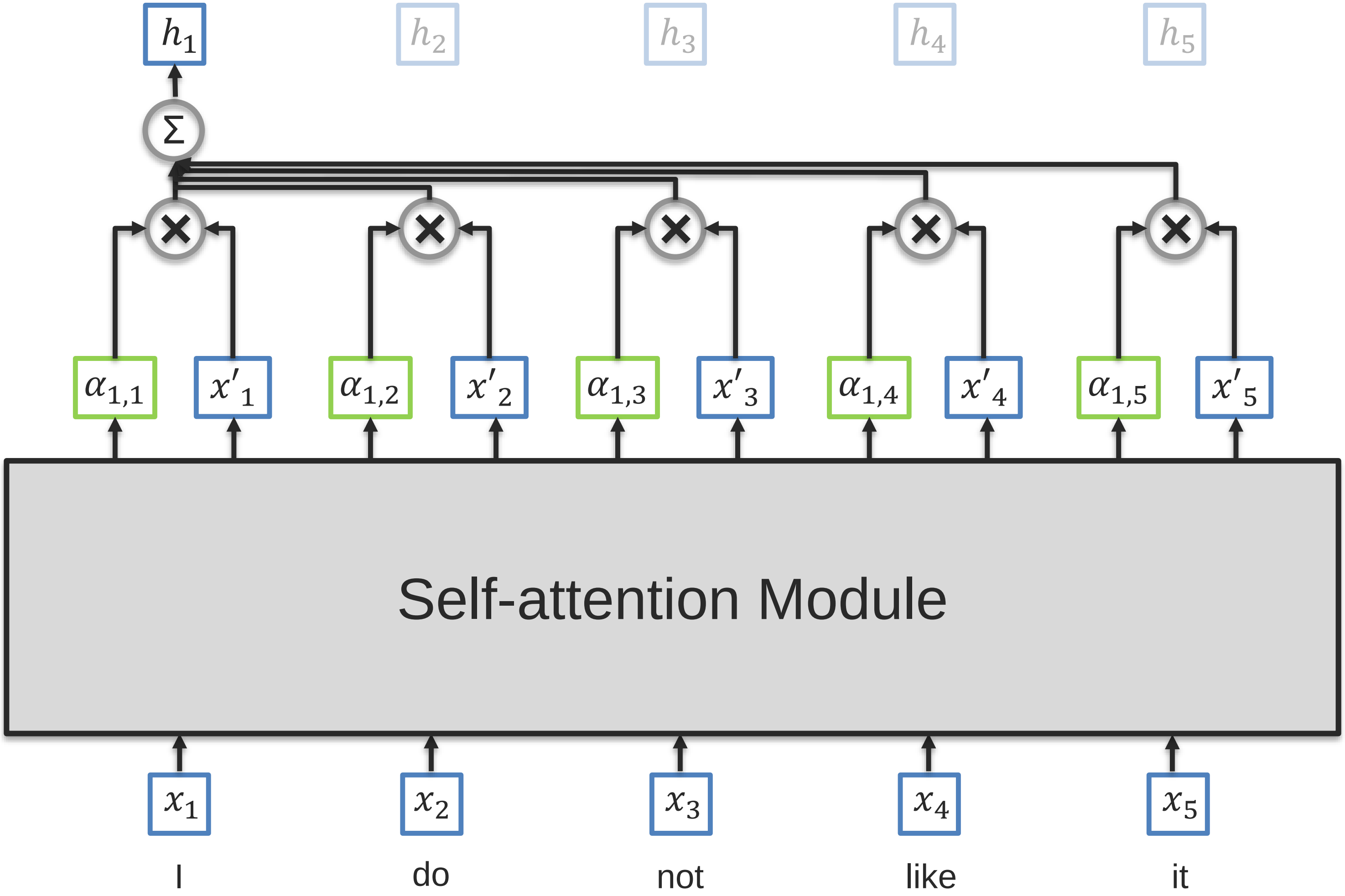
Can be parallelized!

Long-range dependencies

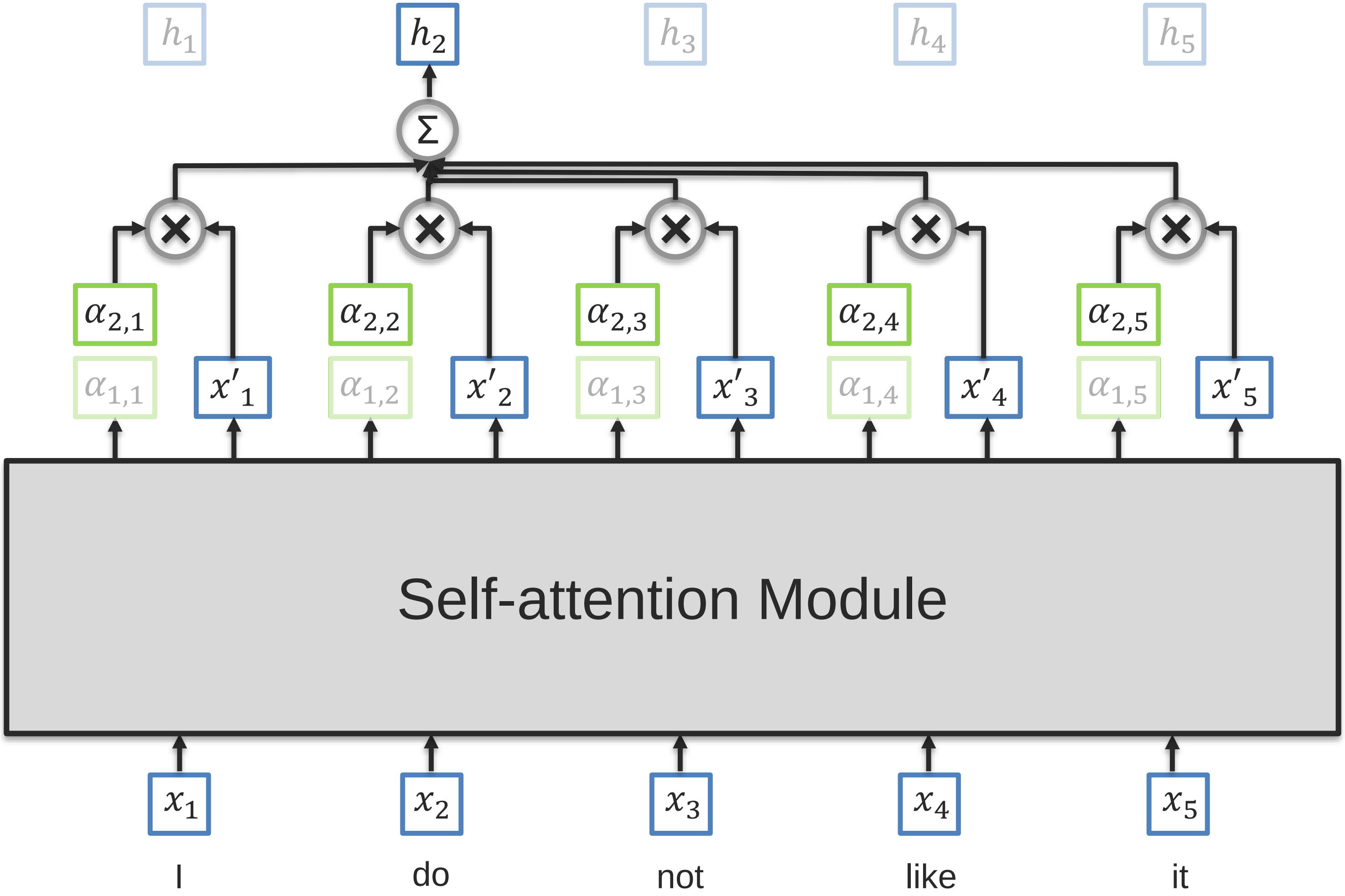
Dynamic attention weights

Self-Attention

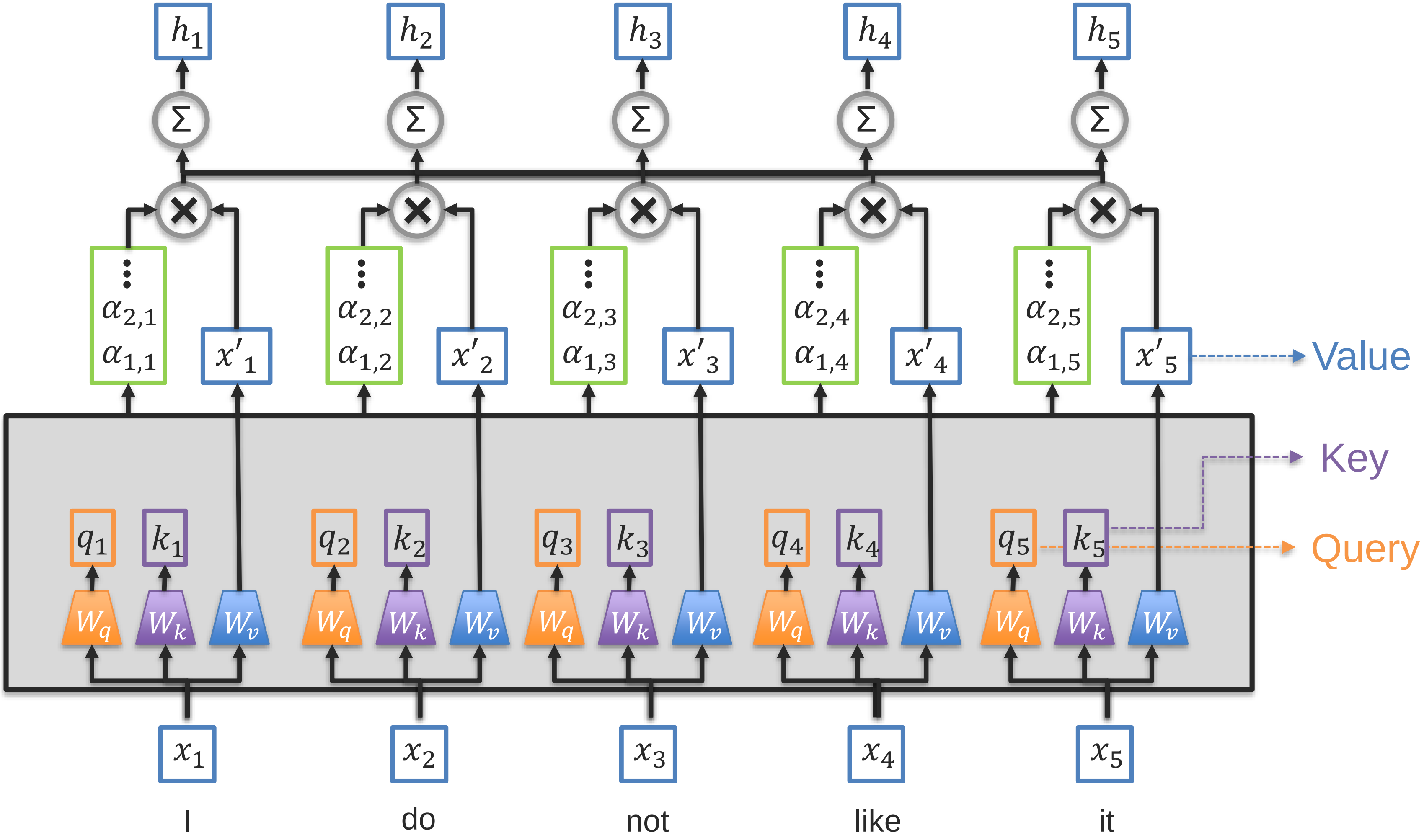
Self-Attention



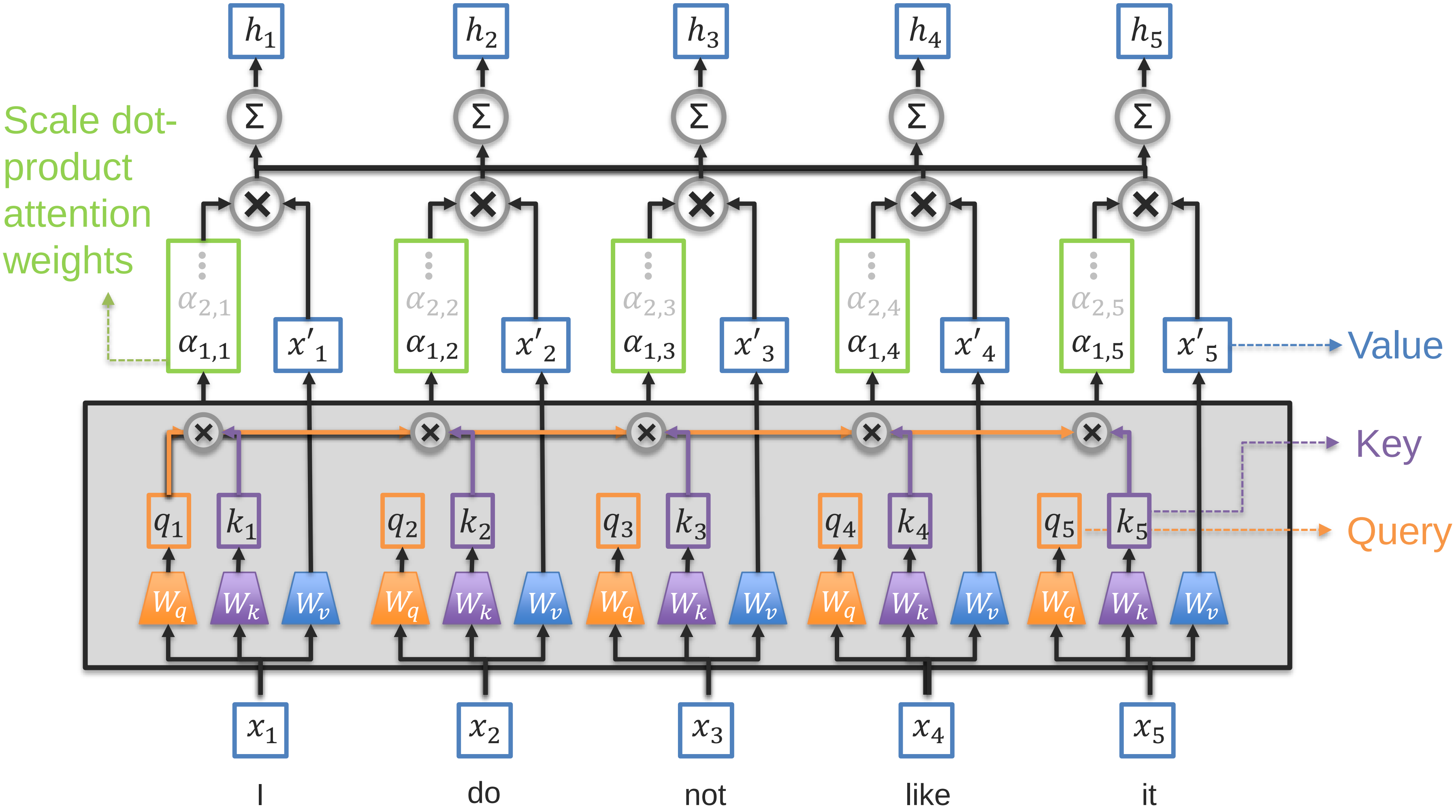
Self-Attention



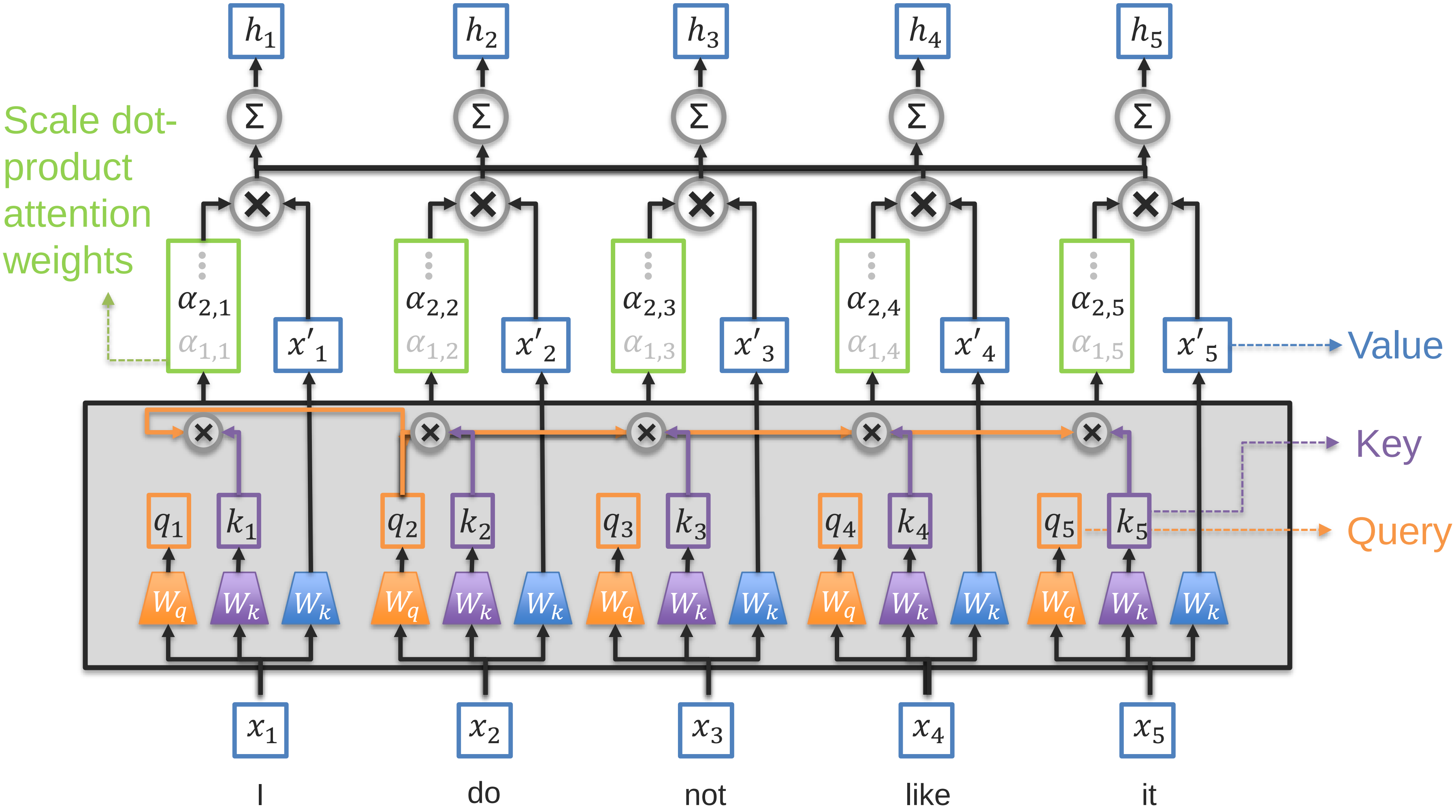
Transformer Self-Attention



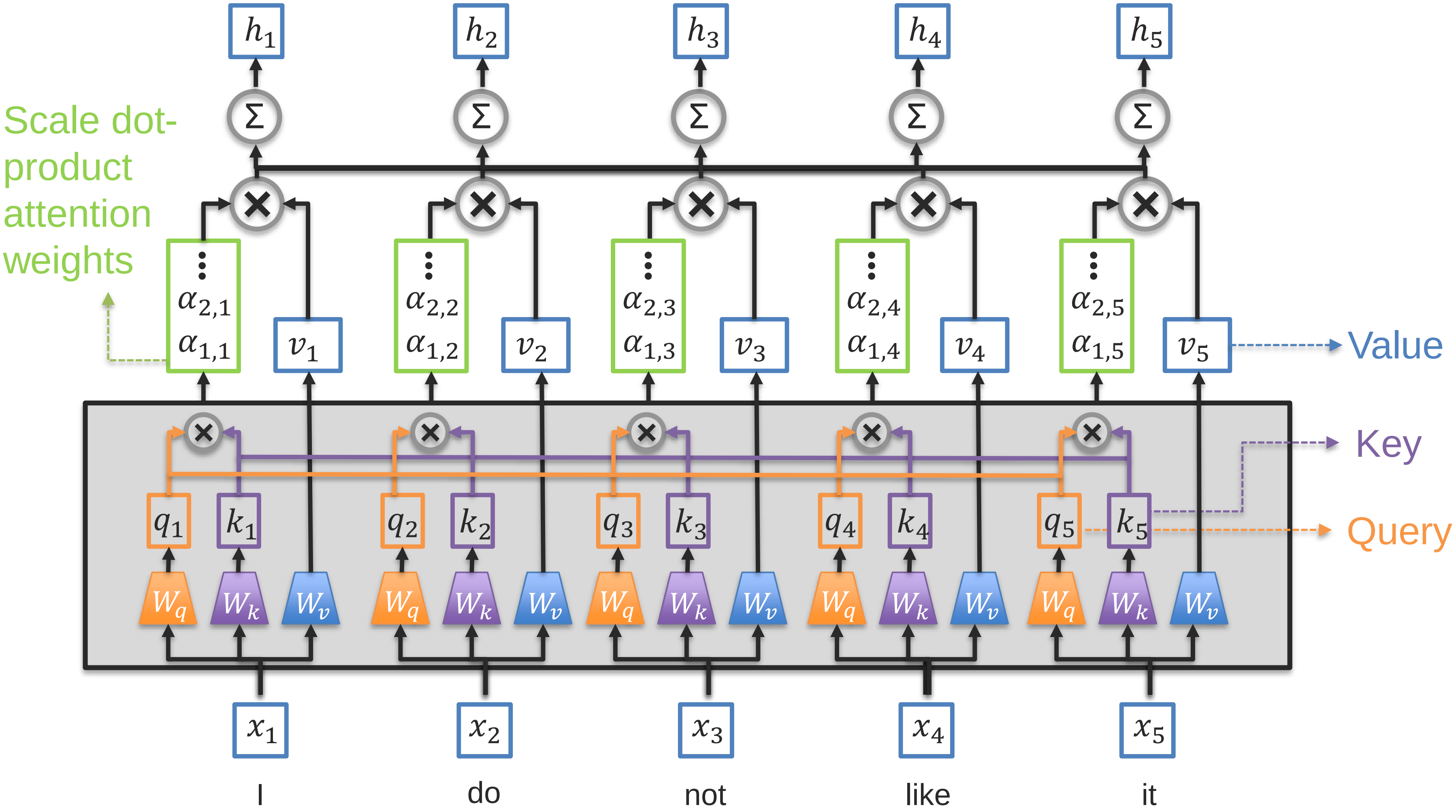
Transformer Self-Attention



Transformer Self-Attention

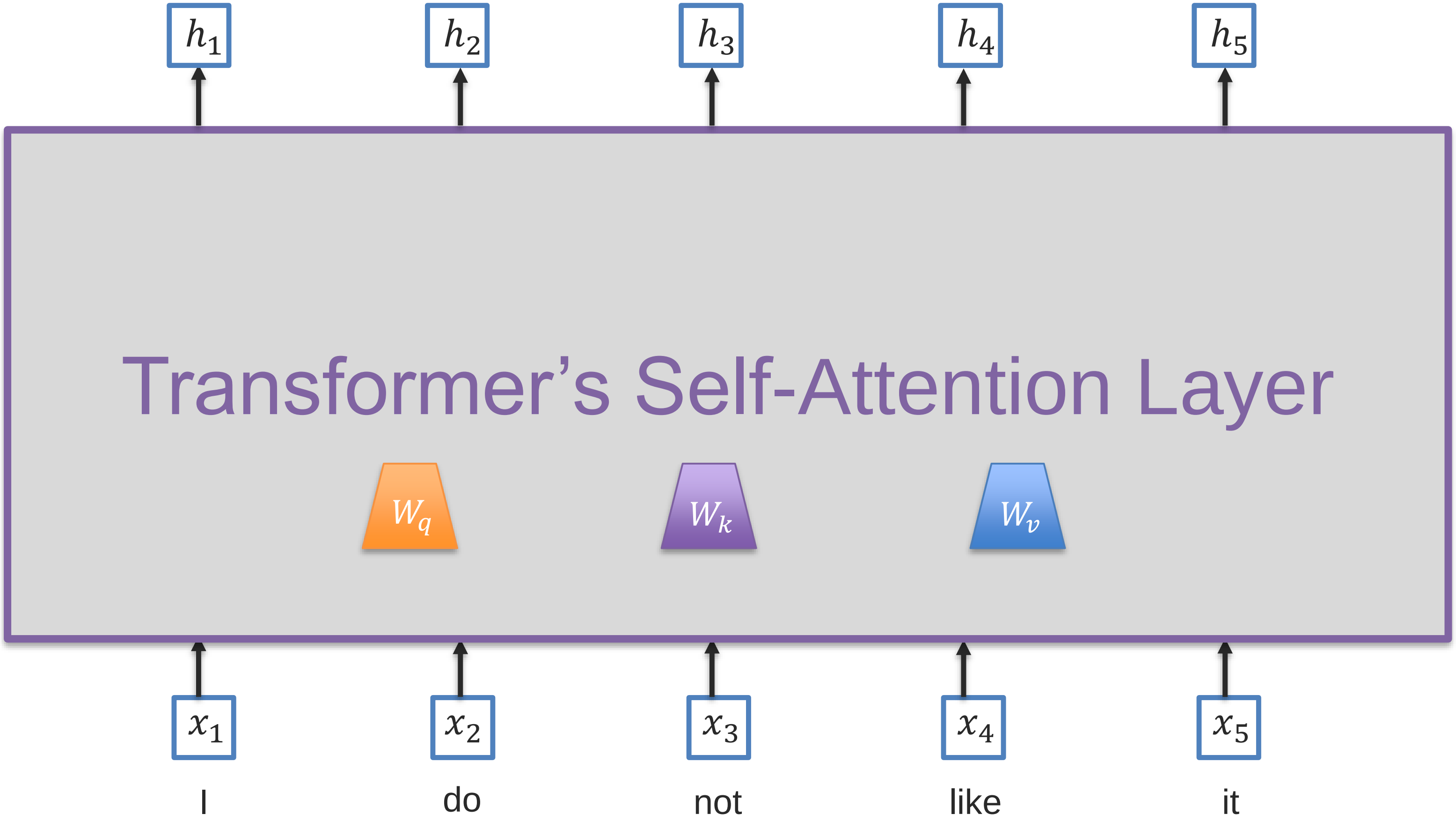


Transformer Self-Attention

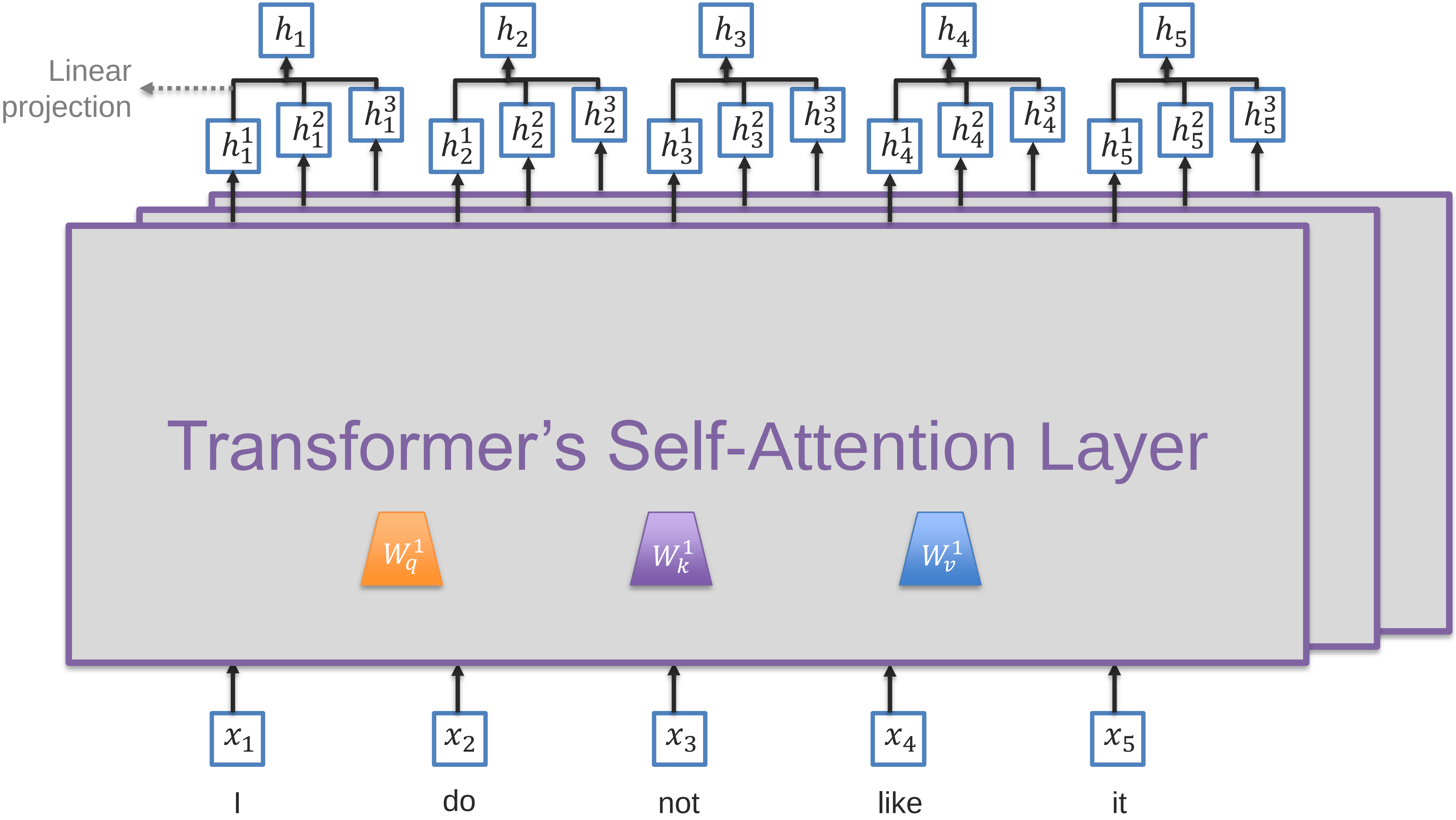


Transformer Self-Attention

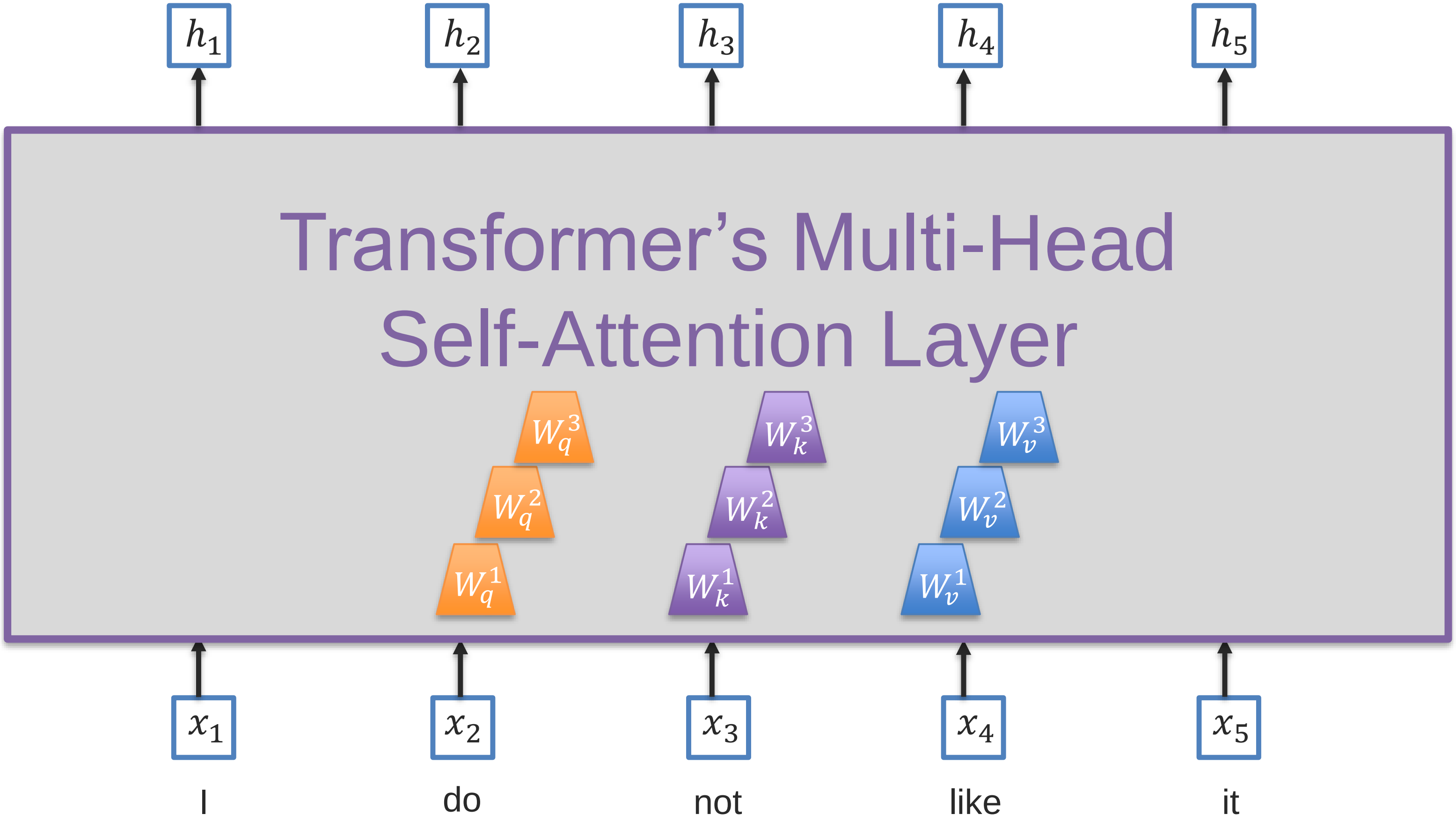
What if we want to attend simultaneously to multiple subspaces of $\boldsymbol{x}$?



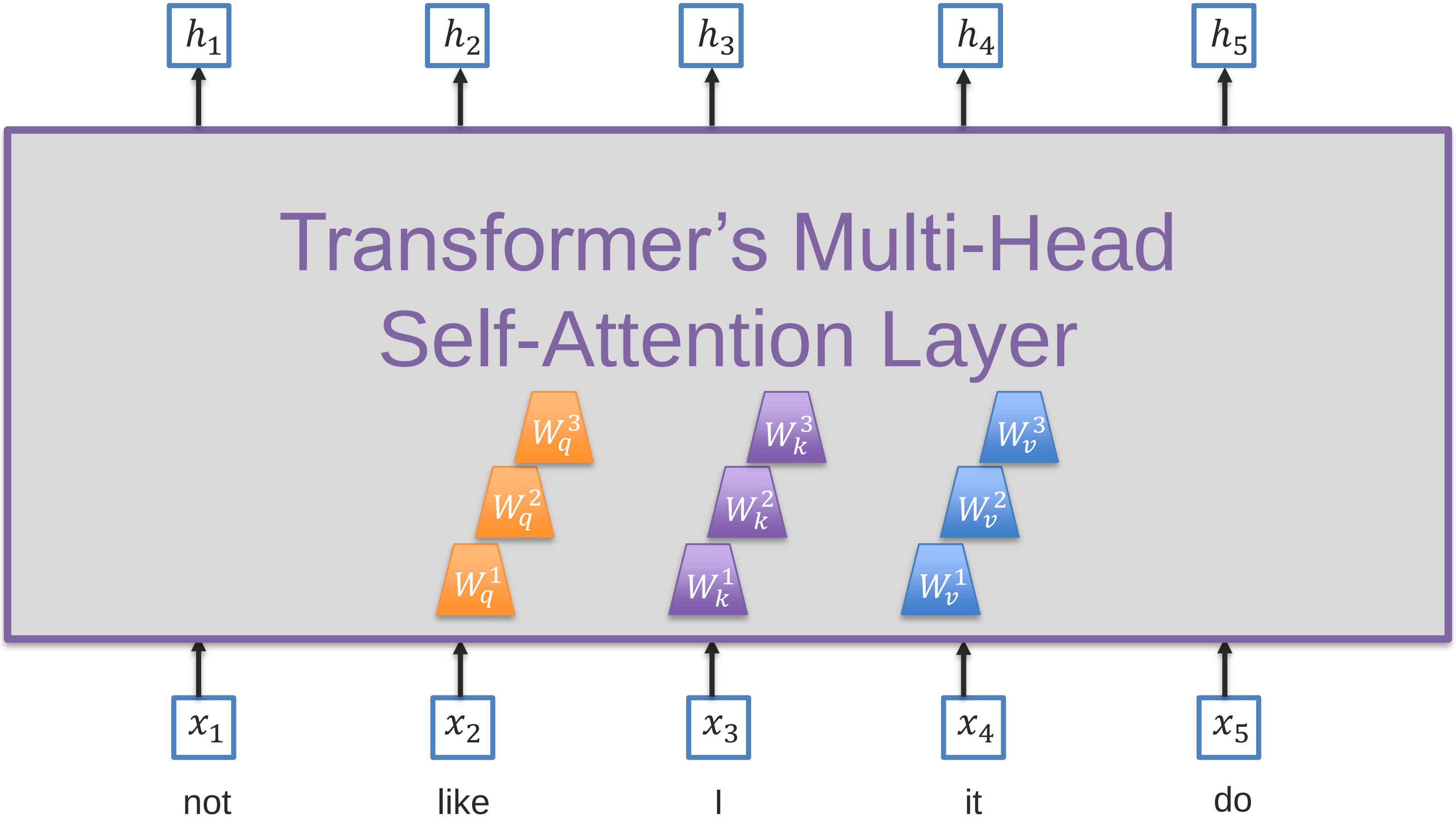
Transformer Multi-Head Self-Attention



Transformer Multi-Head Self-Attention



Transformer Multi-Head Self-Attention



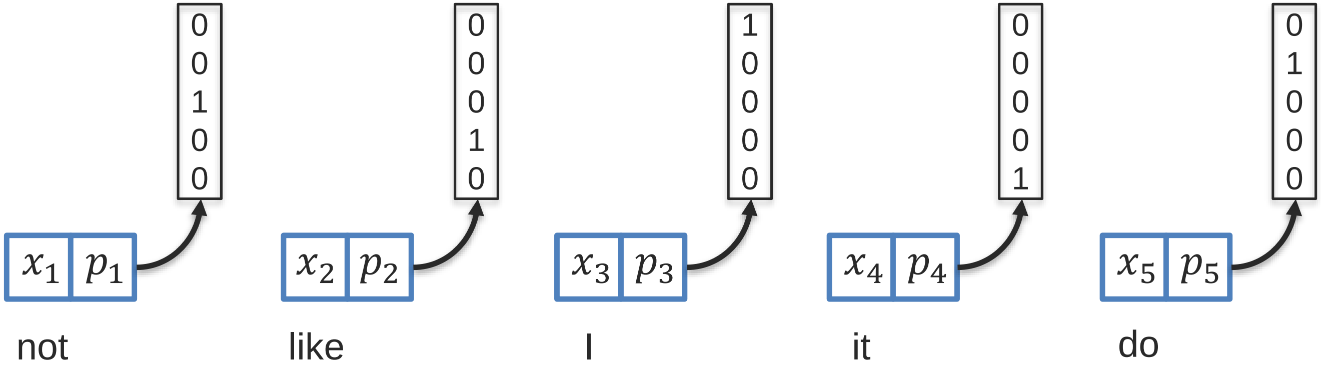
What happens if the words are shuffled?

Position embeddings

- ❑ Position information is not encoded in a self-attention module

How can we encode position information?

Simple approach: one-hot encoding

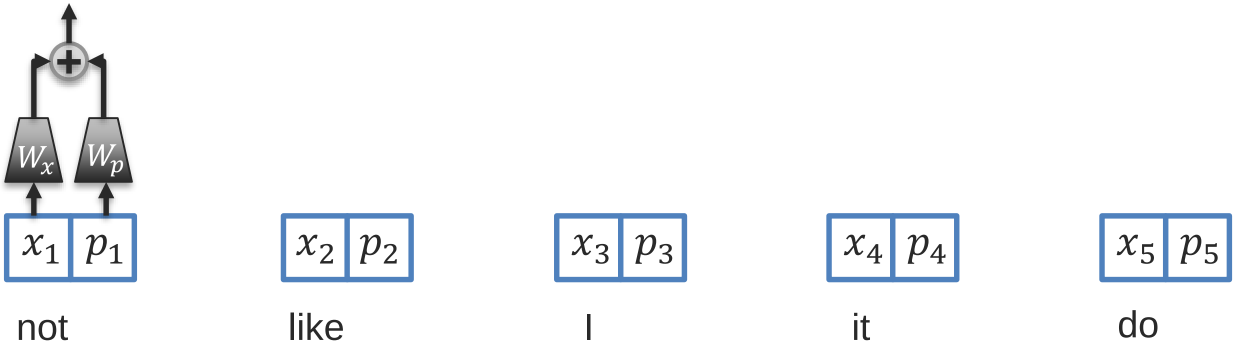


Position embeddings

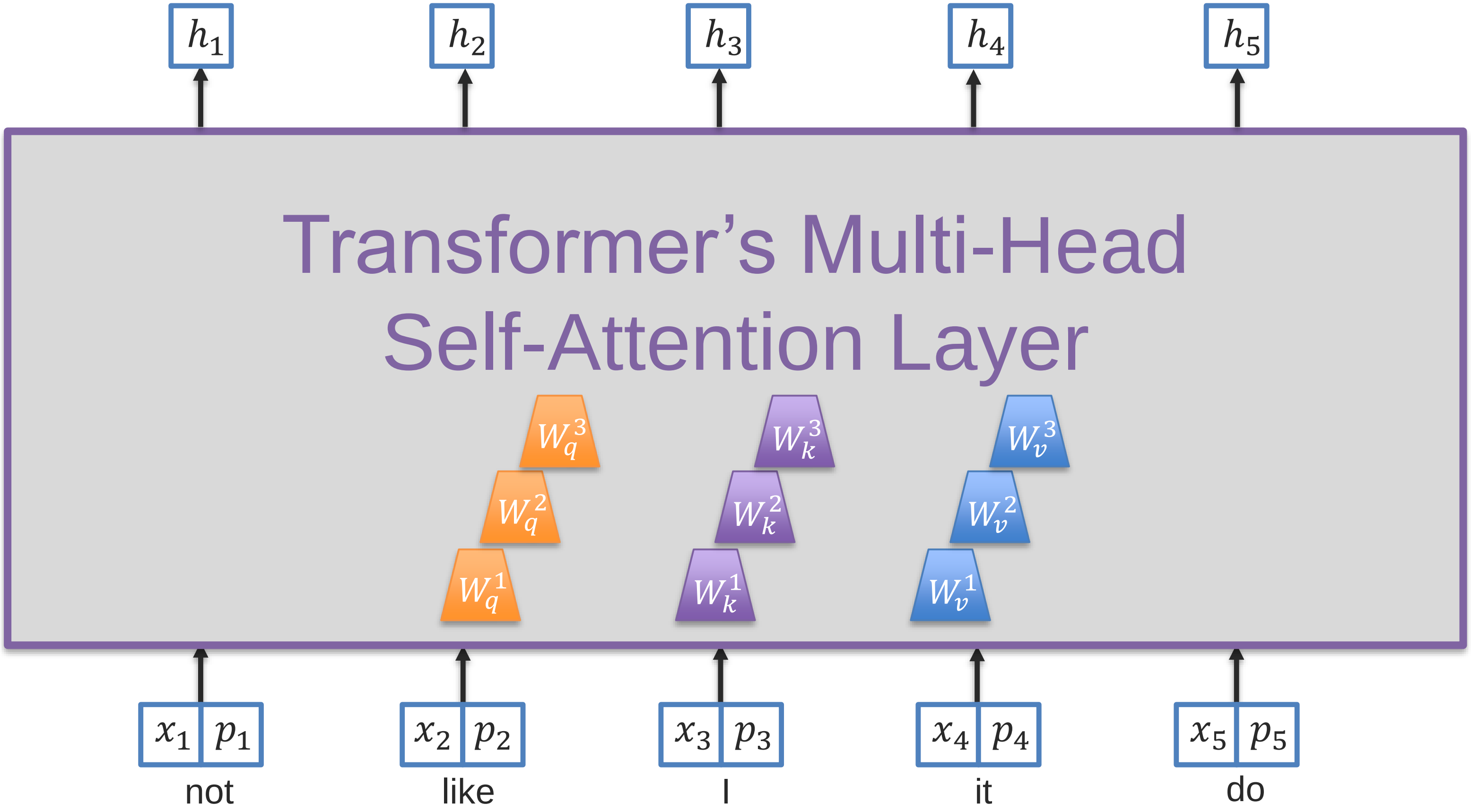
- ❑ Position information is not encoded in a self-attention module

How can we encode position information?

Simple approach: one-hot encoding + linear embeddings + { Sum - or - concat }

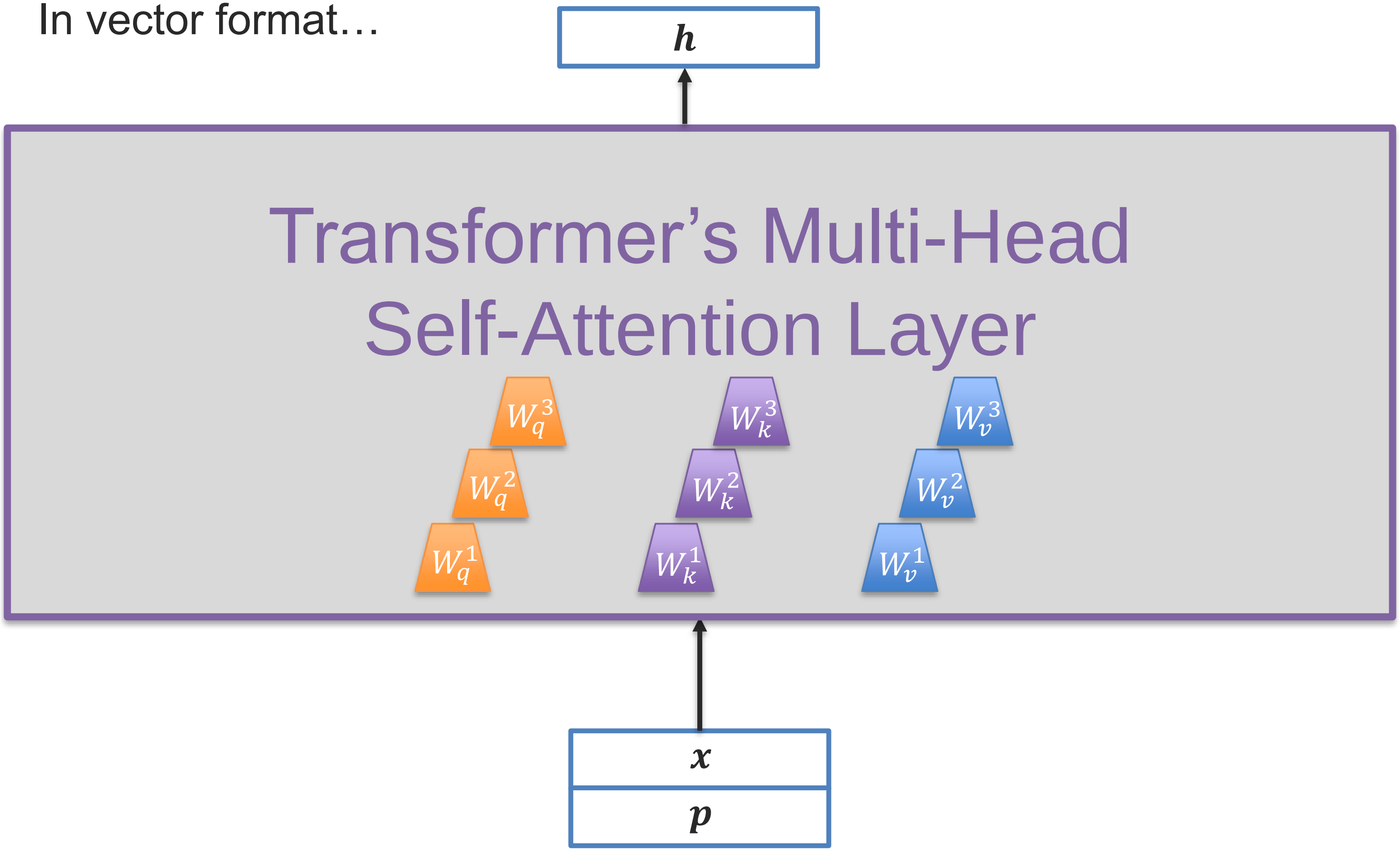


Transformer Multi-Head Self-Attention

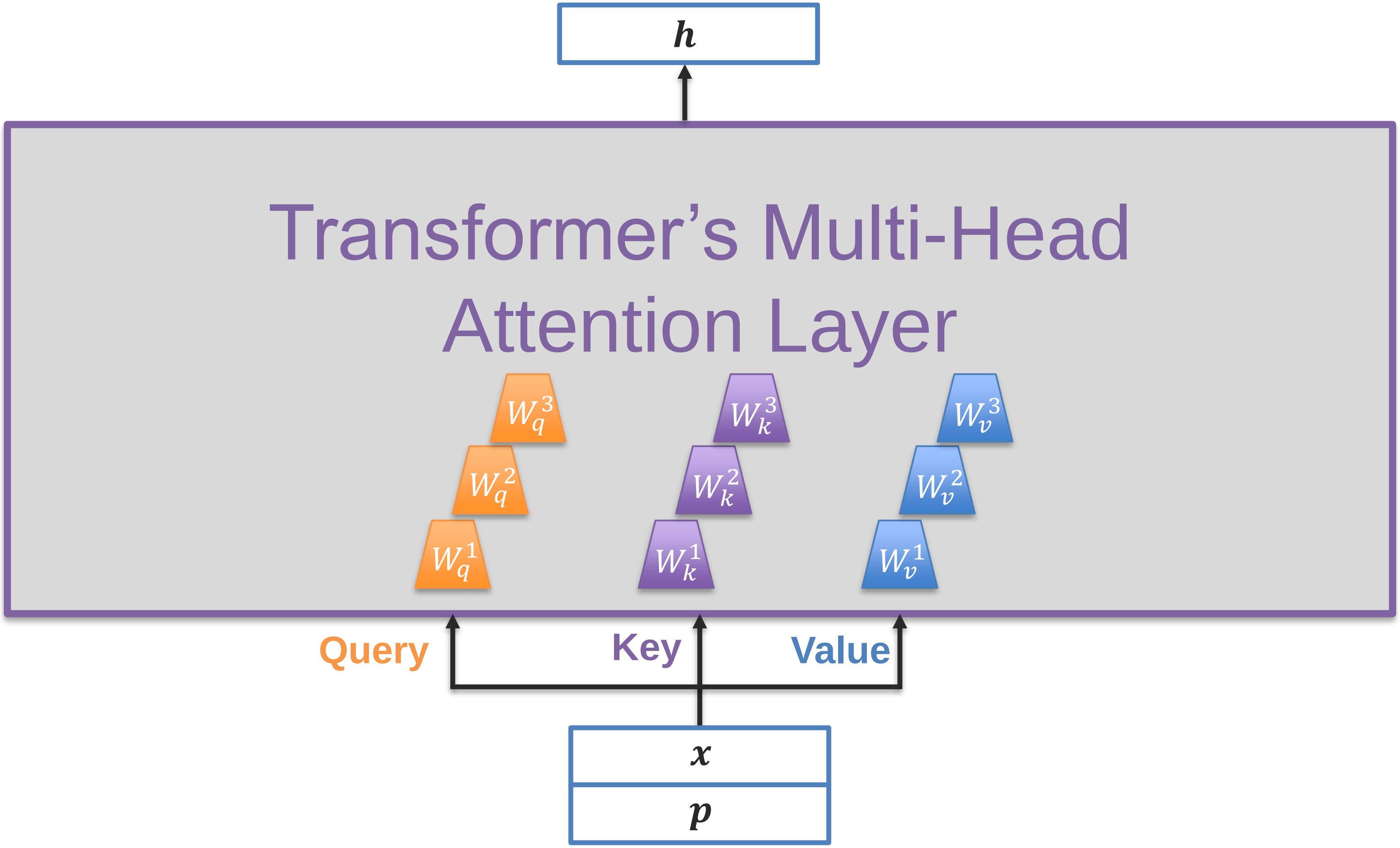


Transformer Multi-Head Self-Attention

In vector format…

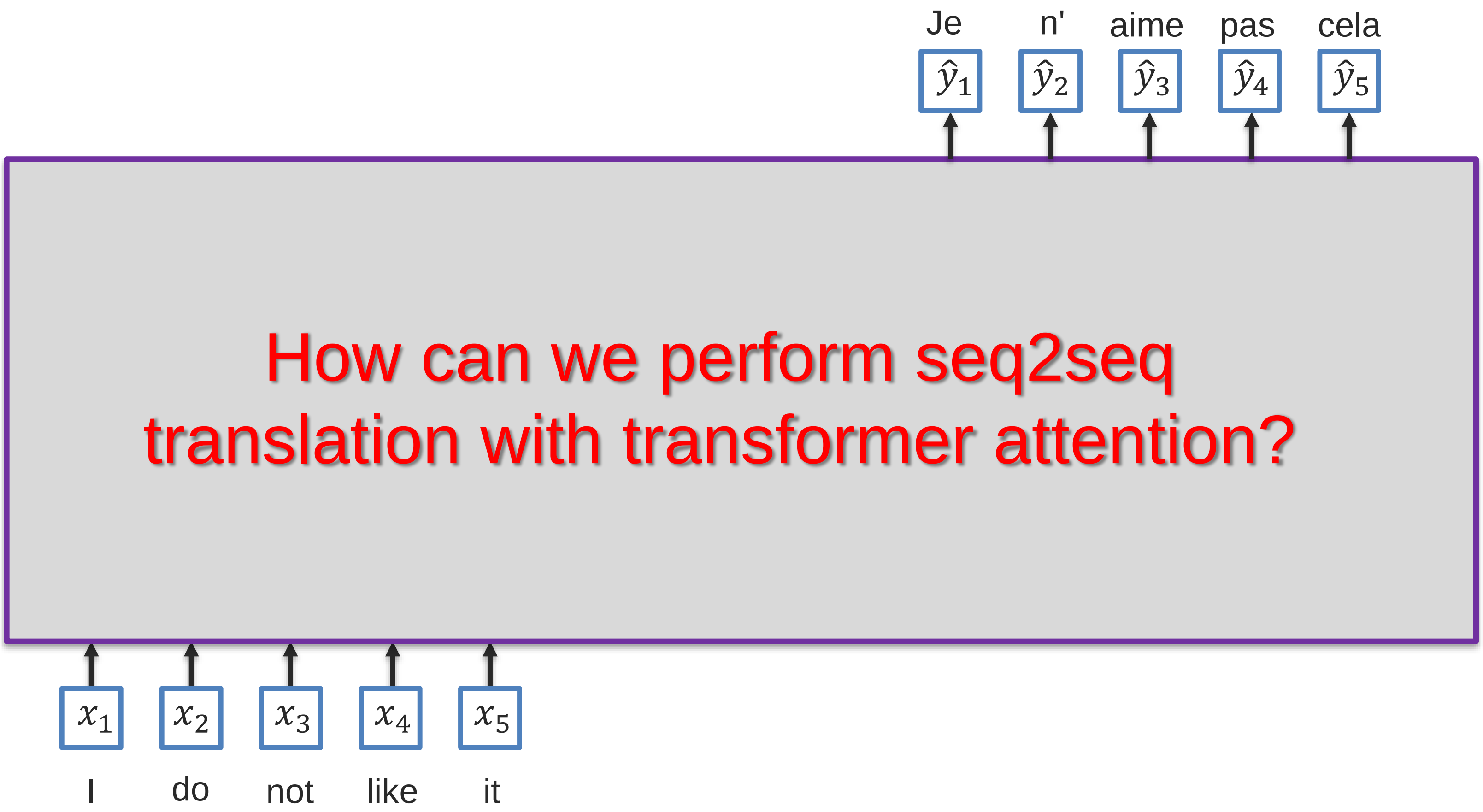


Transformer Multi-Head Attention



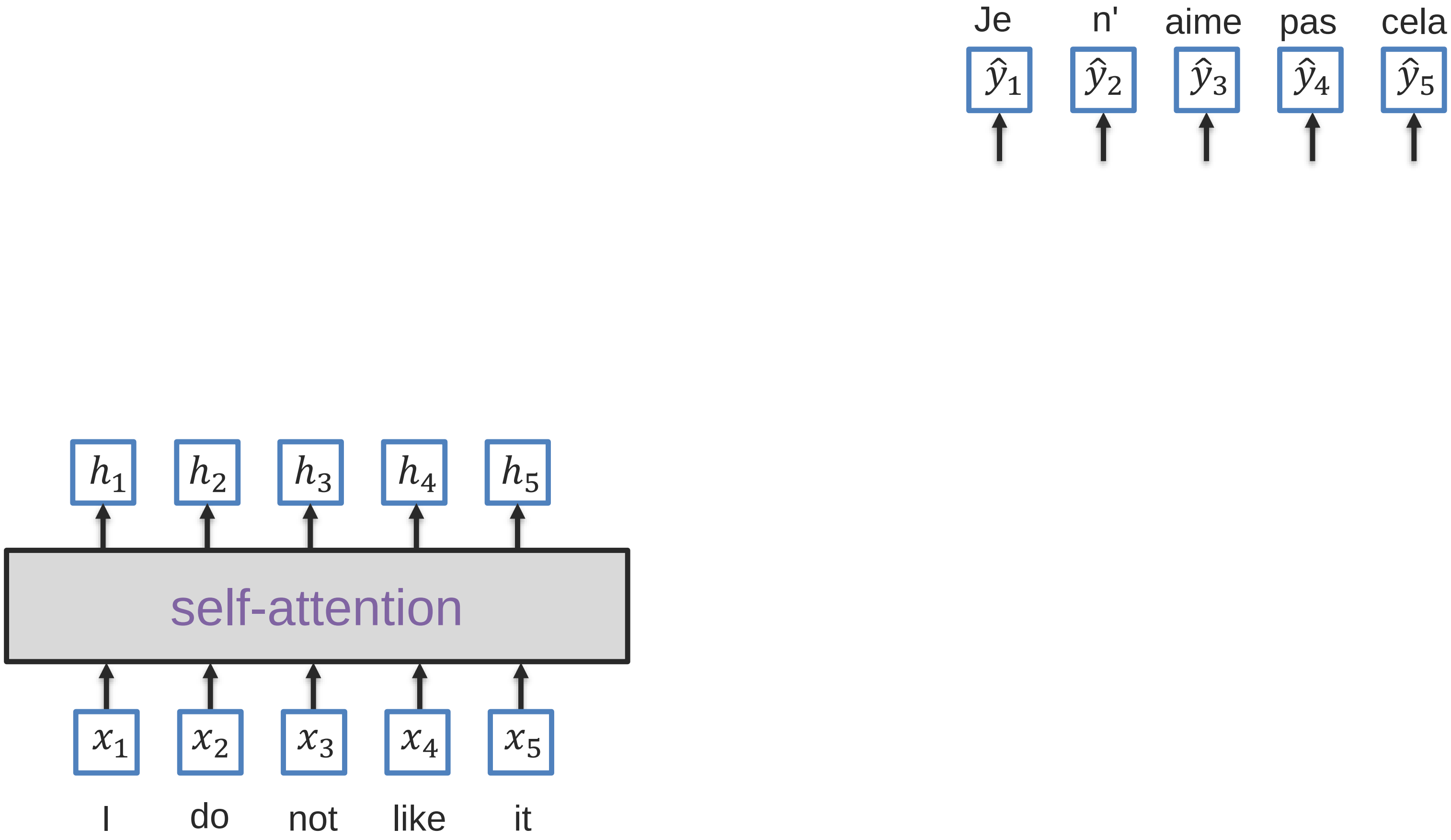
Sequence-to-Sequence Using Transformer

Sequence-to-Sequence Modeling

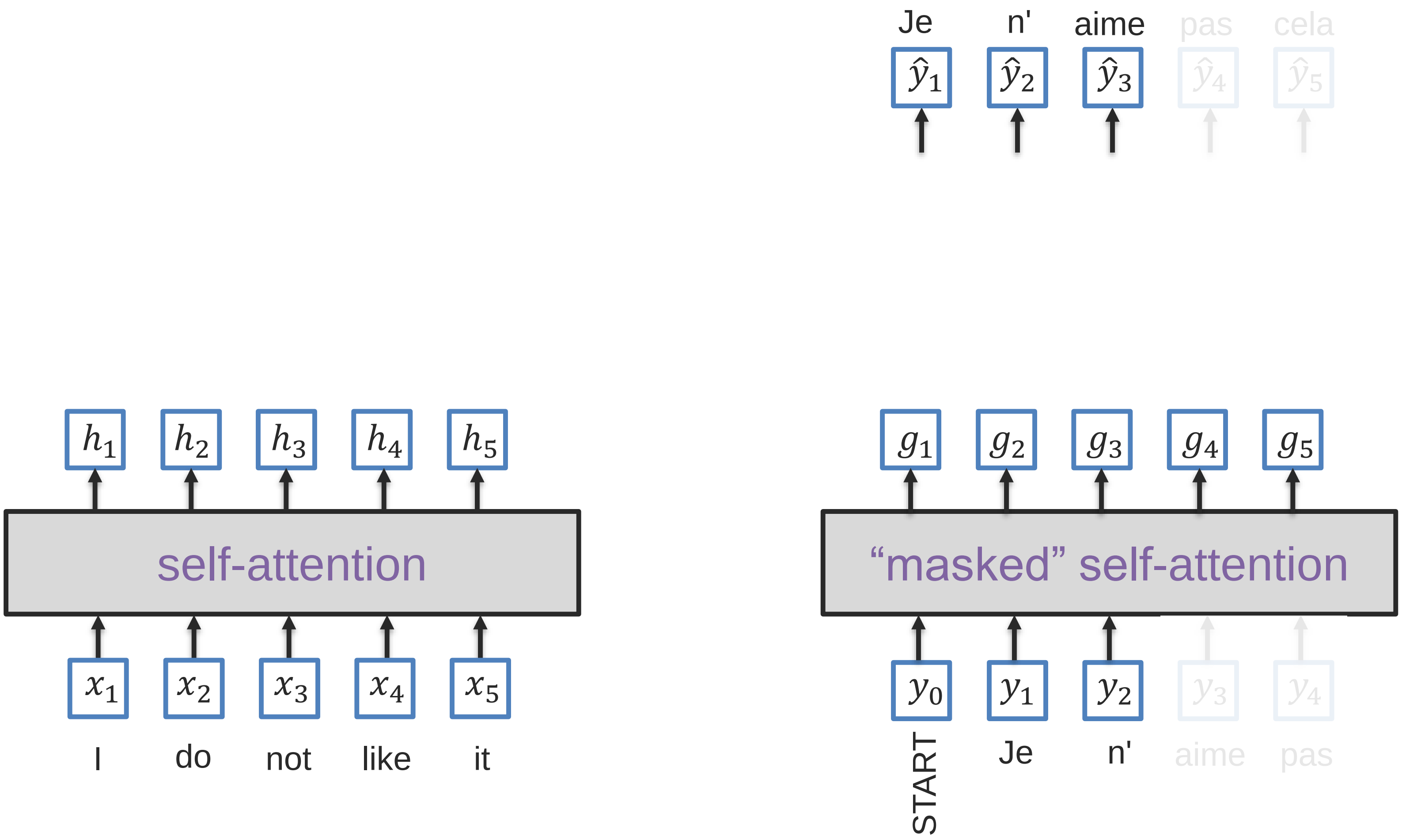


How can we perform seq2seq translation with transformer attention?

Seq2Seq with Transformer Attentions

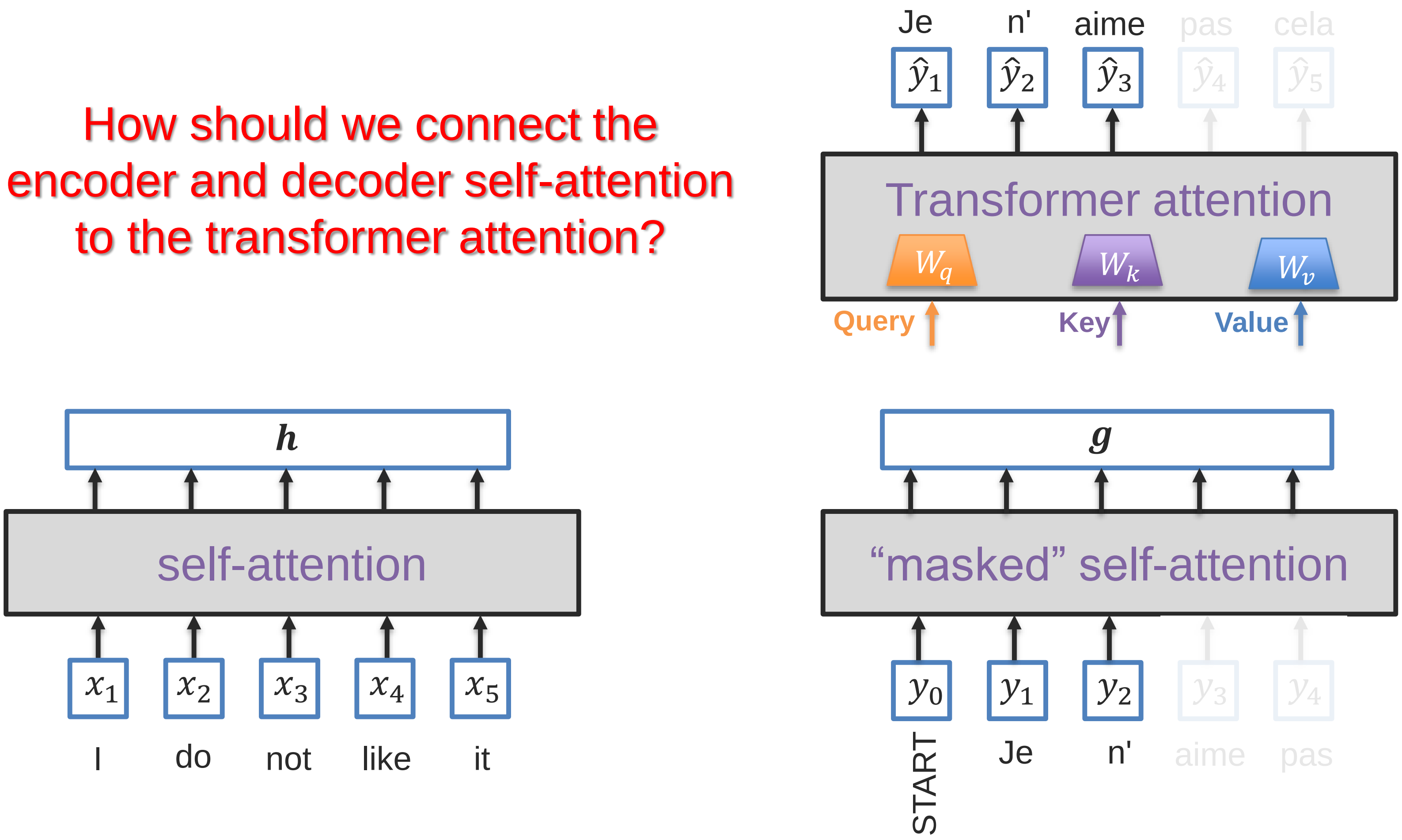


Seq2Seq with Transformer Attentions

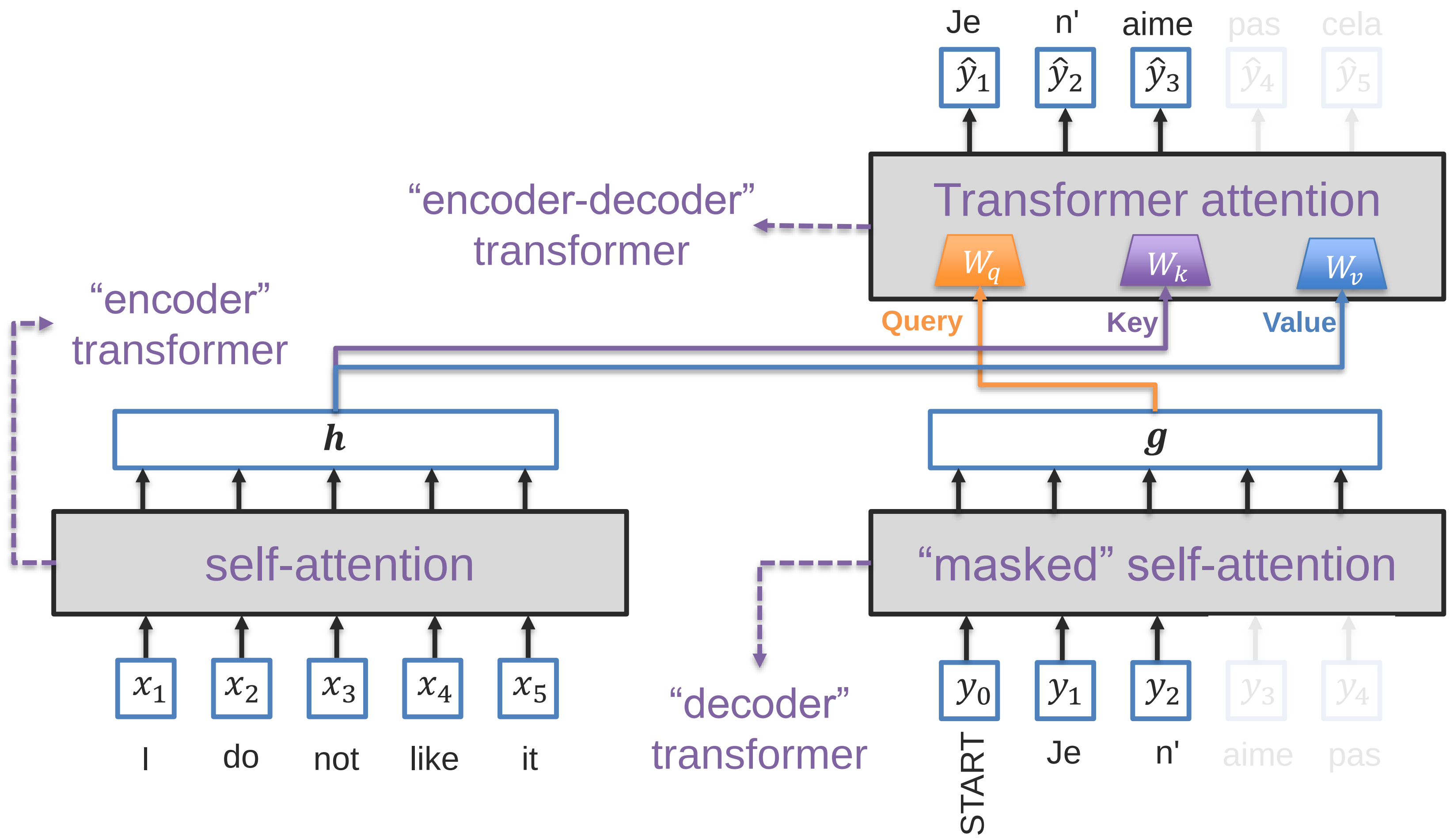


Seq2Seq with Transformer Attentions

How should we connect the encoder and decoder self-attention to the transformer attention?

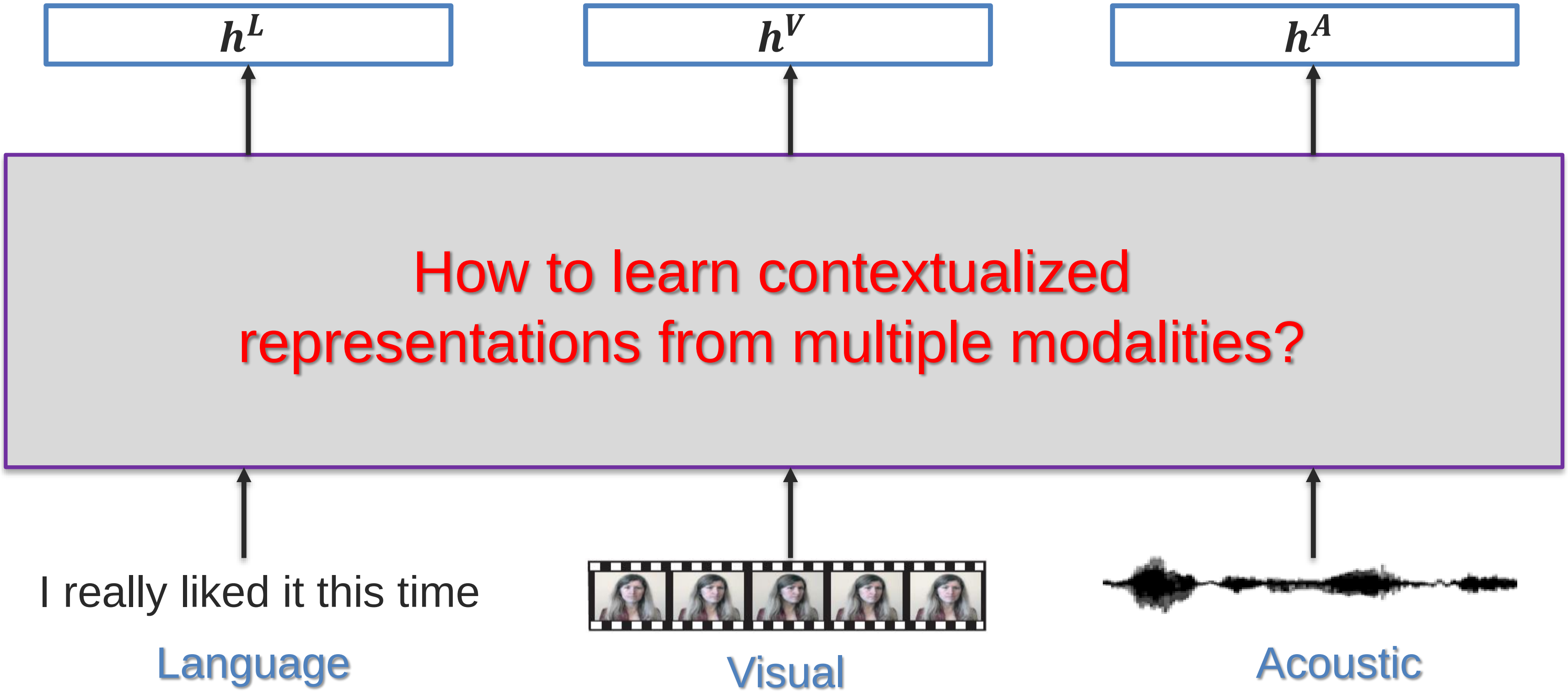


Seq2Seq with Transformer Attentions



Contextualized Multimodal Embedding

Multimodal Embeddings



h^L h^V h^A

How to learn contextualized representations from multiple modalities?

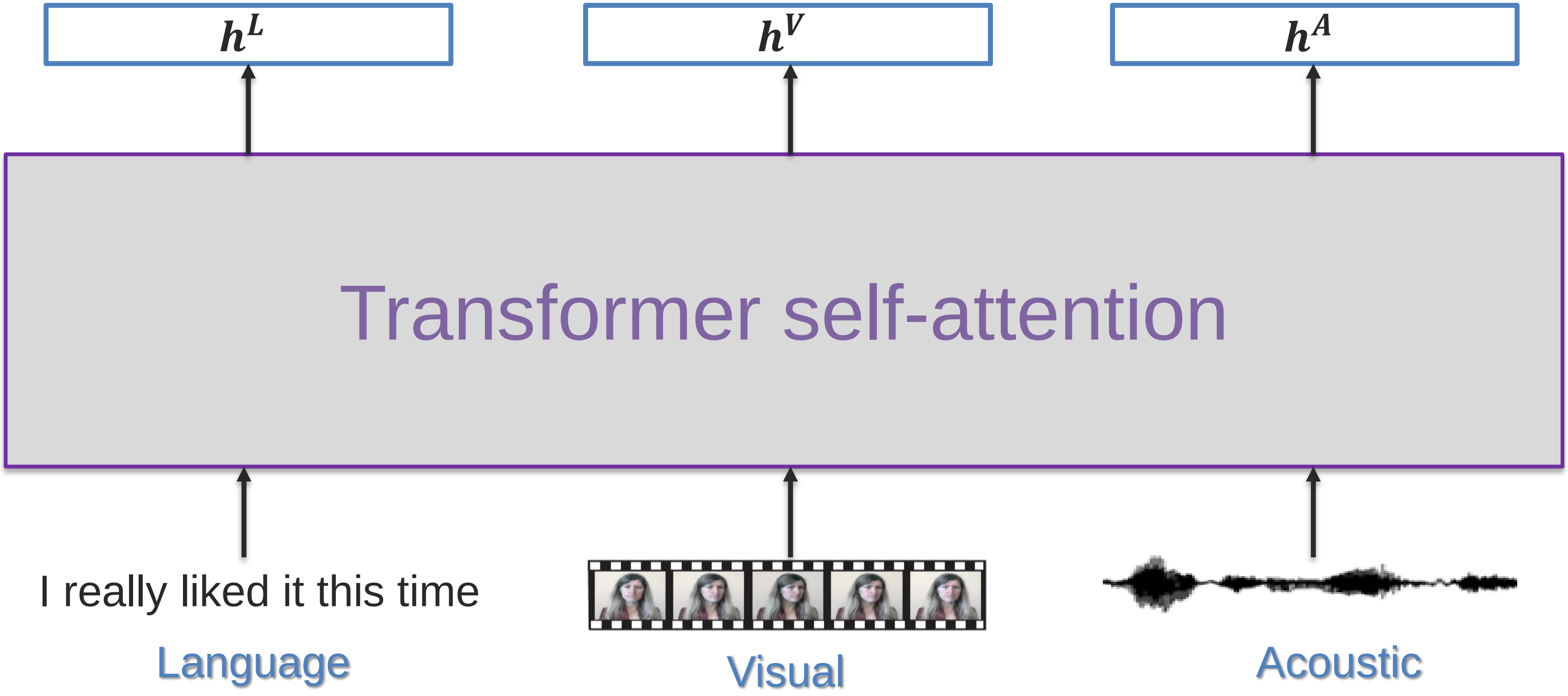
I really liked it this time

Language

Visual

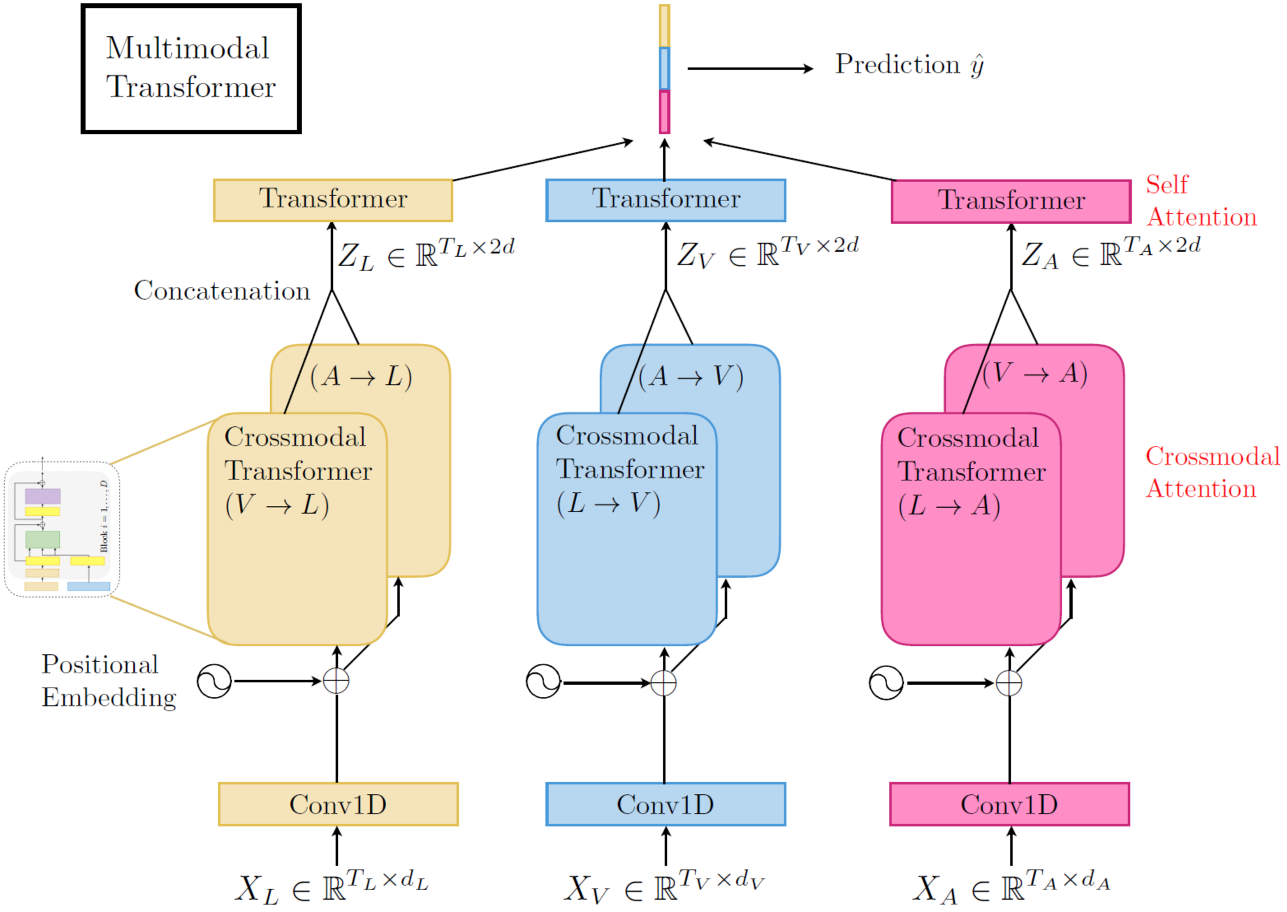
Acoustic

Contextualized Multimodal Embeddings



Any other approach?

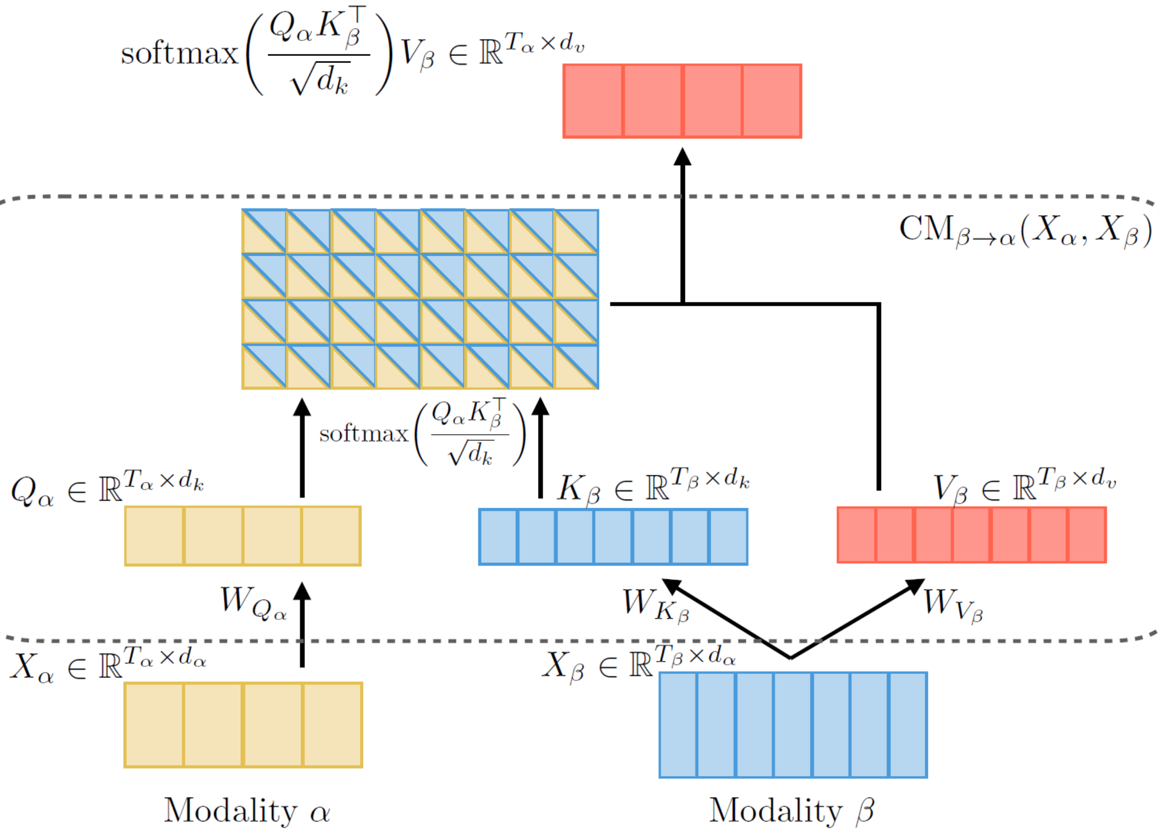
Multimodal Transformer



Multimodal Transformer
Prediction $\hat{y}$
Transformer
Transformer
Transformer
Self Attention
$Z_L \in \mathbb{R}^{T_L \times 2d}$
$Z_V \in \mathbb{R}^{T_V \times 2d}$
$Z_A \in \mathbb{R}^{T_A \times 2d}$
Concatenation
$(A \to L)$
$(A \to V)$
$(V \to A)$
Crossmodal Transformer $(V \to L)$
Crossmodal Transformer $(L \to V)$
Crossmodal Transformer $(L \to A)$
Crossmodal Attention
Positional Embedding
Conv1D
Conv1D
Conv1D
$X_L \in \mathbb{R}^{T_L \times d_L}$
$X_V \in \mathbb{R}^{T_V \times d_V}$
$X_A \in \mathbb{R}^{T_A \times d_A}$

Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

Cross-Modal Transformer

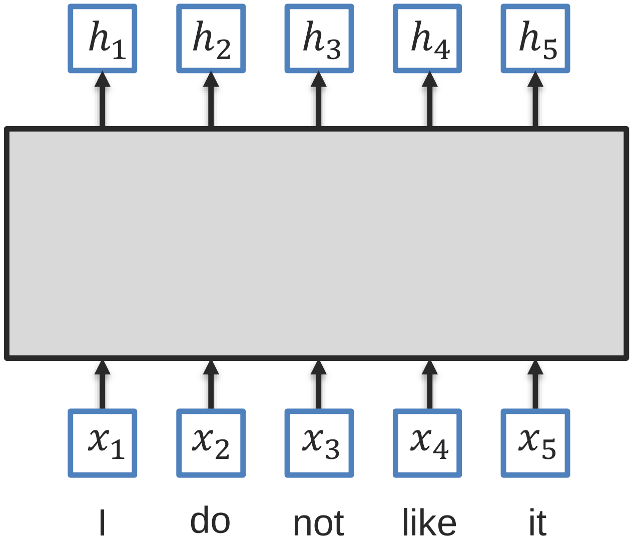

$\text{softmax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right)V_\beta \in \mathbb{R}^{T_\alpha \times d_v}$
$\text{CM}_{\beta \rightarrow \alpha}(X_\alpha, X_\beta)$
$\text{softmax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right)$
$Q_\alpha \in \mathbb{R}^{T_\alpha \times d_k}$
$K_\beta \in \mathbb{R}^{T_\beta \times d_k}$
$V_\beta \in \mathbb{R}^{T_\beta \times d_v}$
W_{Q_α}
W_{K_β}
W_{V_β}
$X_\alpha \in \mathbb{R}^{T_\alpha \times d_\alpha}$
$X_\beta \in \mathbb{R}^{T_\beta \times d_\alpha}$
Modality α
Modality β

Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

Language Pre-training

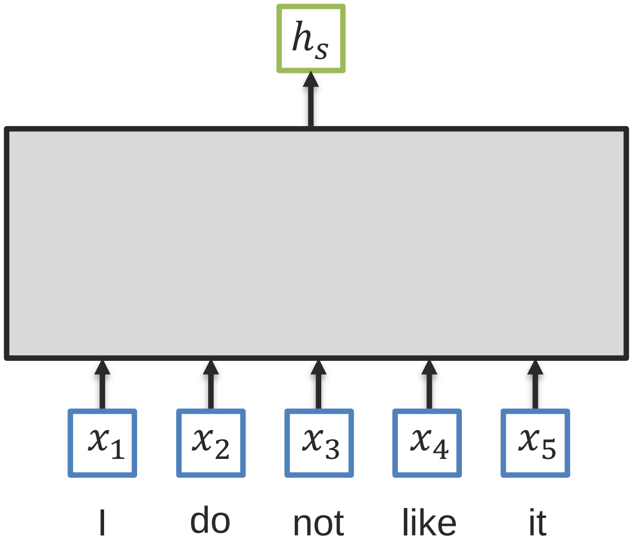
Token-level and Sentence-level Embeddings

Token-level embeddings



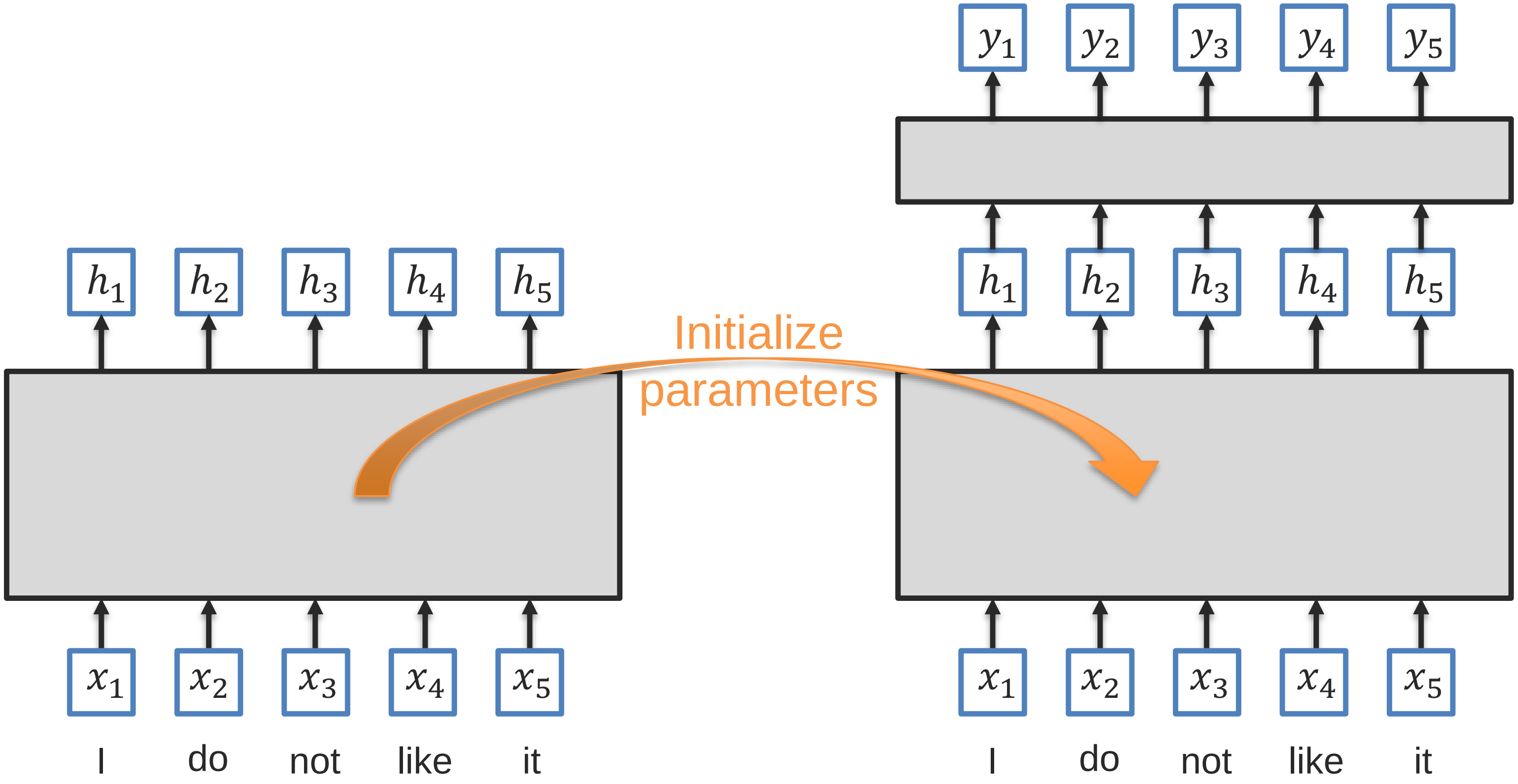
Which tasks?

Sentence-level embedding



Which tasks?

Pre-Training and Fine-Tuning



y_1 y_2 y_3 y_4 y_5
h_1 h_2 h_3 h_4 h_5
h_1 h_2 h_3 h_4 h_5
Initialize parameters
x_1 x_2 x_3 x_4 x_5
I do not like it
x_1 x_2 x_3 x_4 x_5
I do not like it

Pre-training

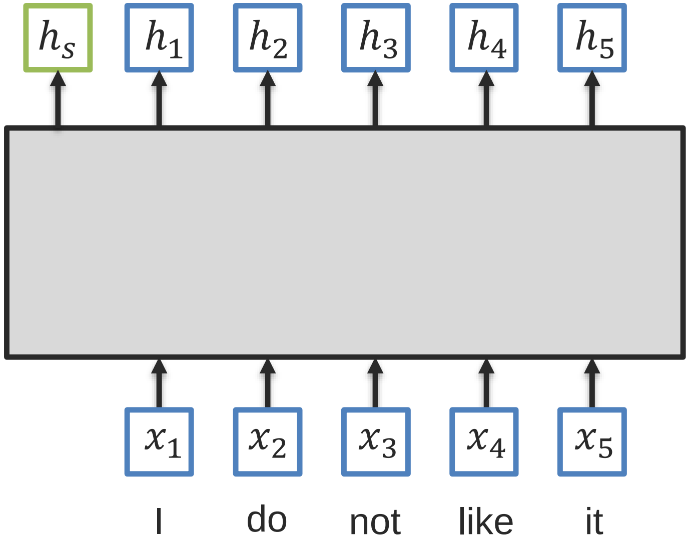
(e.g., language model)

Fine-Tuning

BERT:
Bidirectional Encoder Representations from Transformers

Advantages:

1. Jointly learn representation for token-level and sentence level
2. Same network architecture for pre-training and fine-tuning

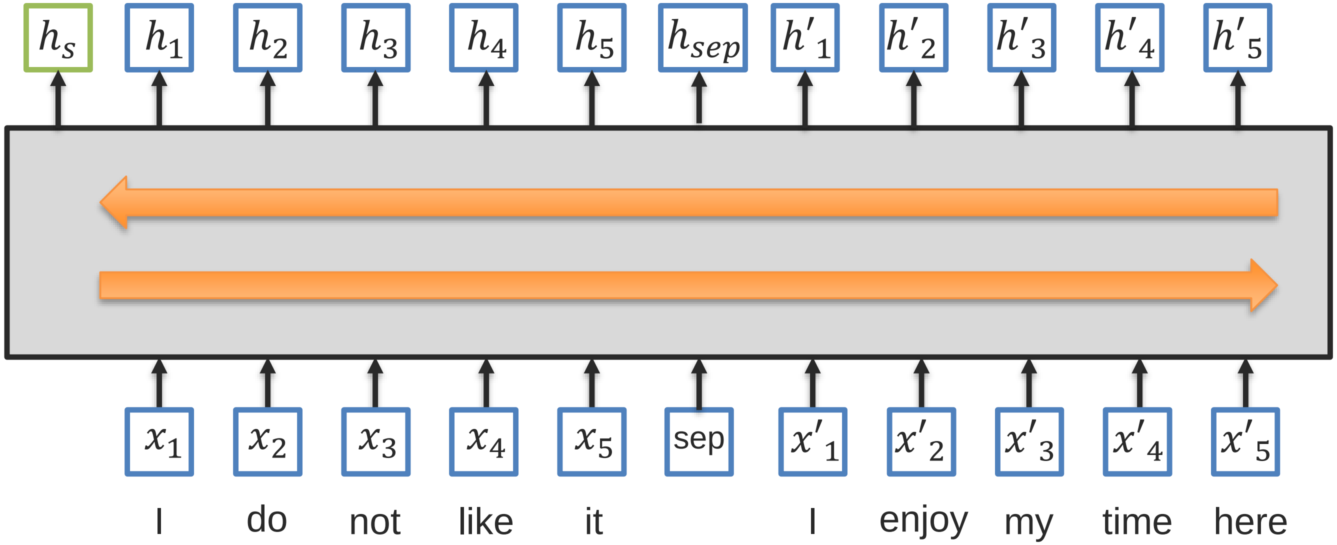


BERT: Bidirectional Encoder Representations from Transformers

Advantages:

1. Jointly learn representation for token-level and sentence level
2. Same network architecture for pre-training and fine-tuning
3. Can be used learn relationship between sentences
4. Models bidirectional and long-range interactions between tokens

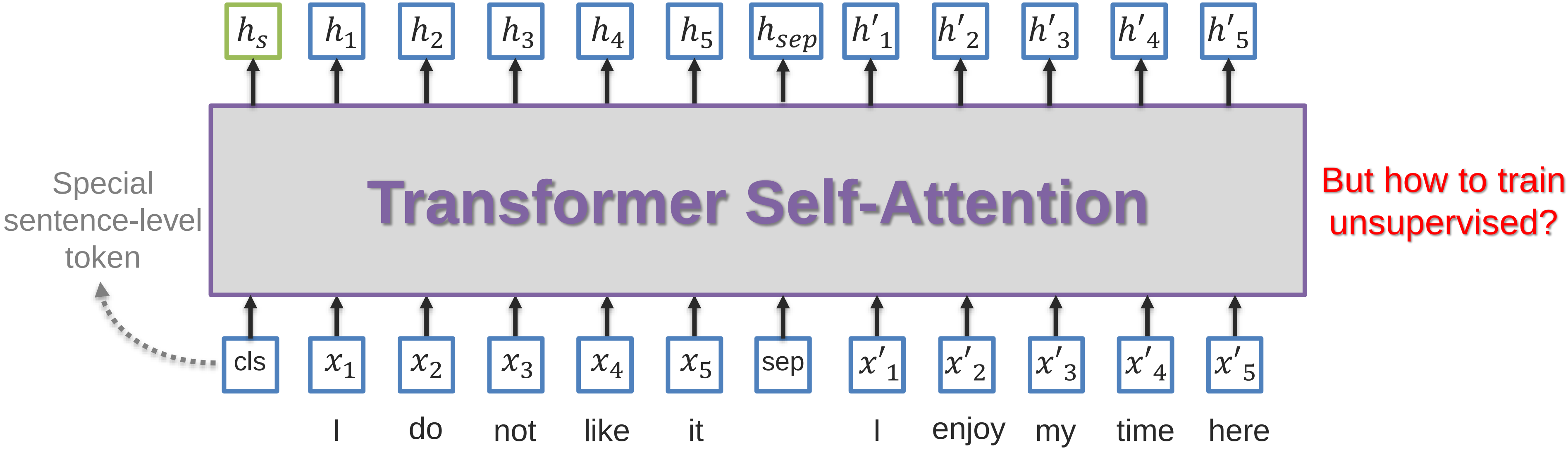
How can we do all this?



BERT:
Bidirectional Encoder Representations from Transformers

Advantages:

1. Jointly learn representation for token-level and sentence level
2. Same network architecture for pre-training and fine-tuning
3. Can be used learn relationship between sentences
4. Models bidirectional interactions between tokens



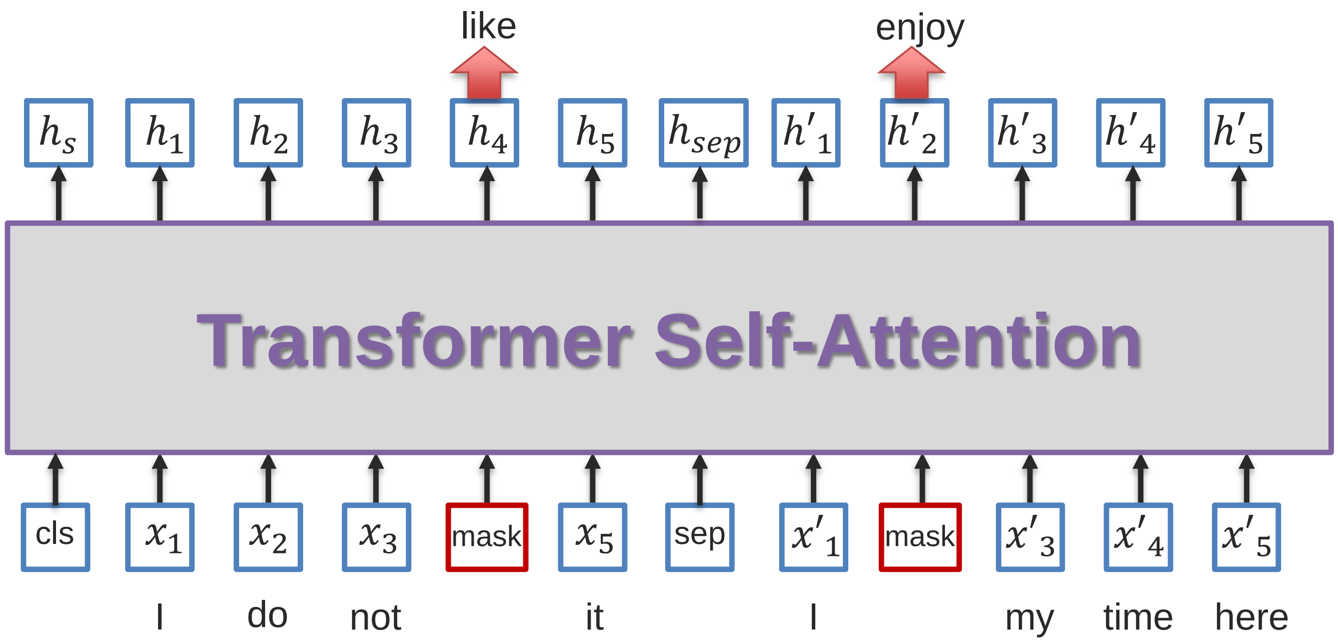
But how to train unsupervised?

Pre-training BERT Model

1 **Masked Language Model**

Randomly mask input tokens and then try to predict them

What is the loss function?



Pre-training BERT Model

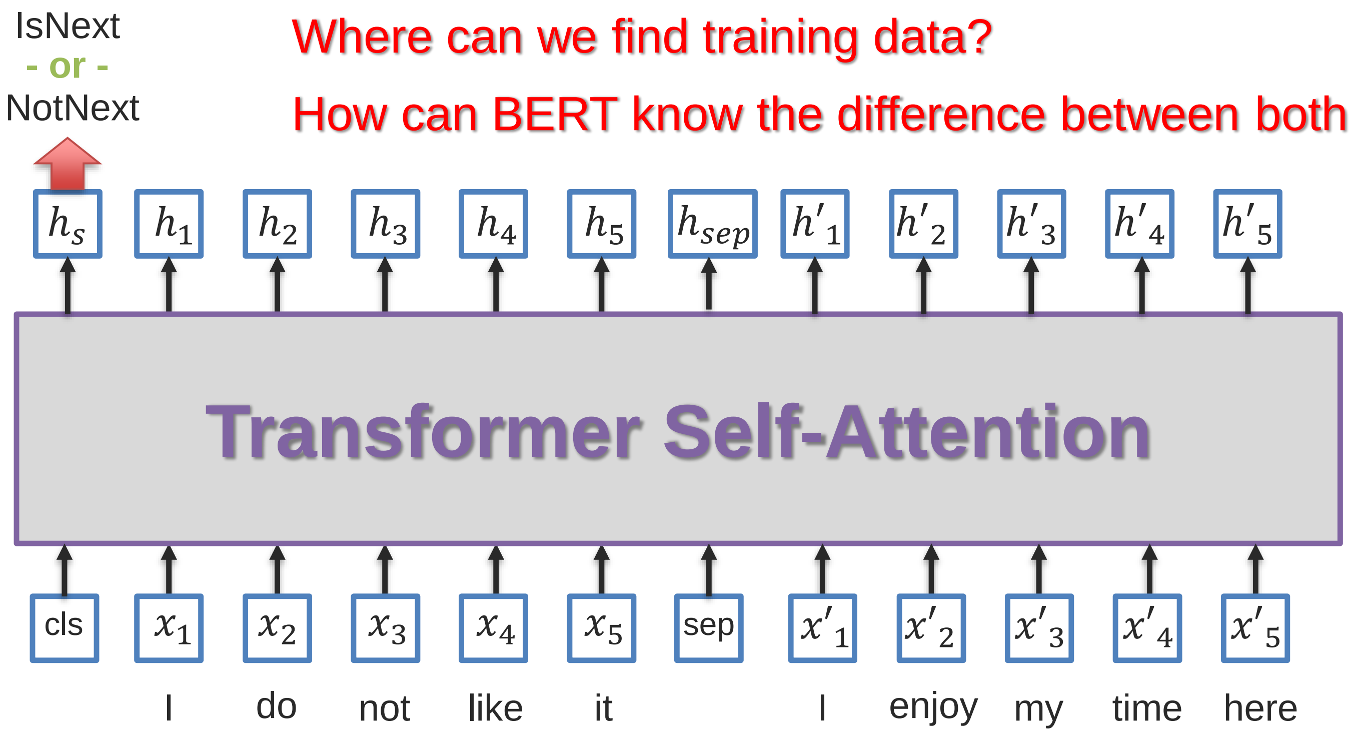
2 **Next Sentence Prediction**

Given two sentences, predict if this is the next one or not

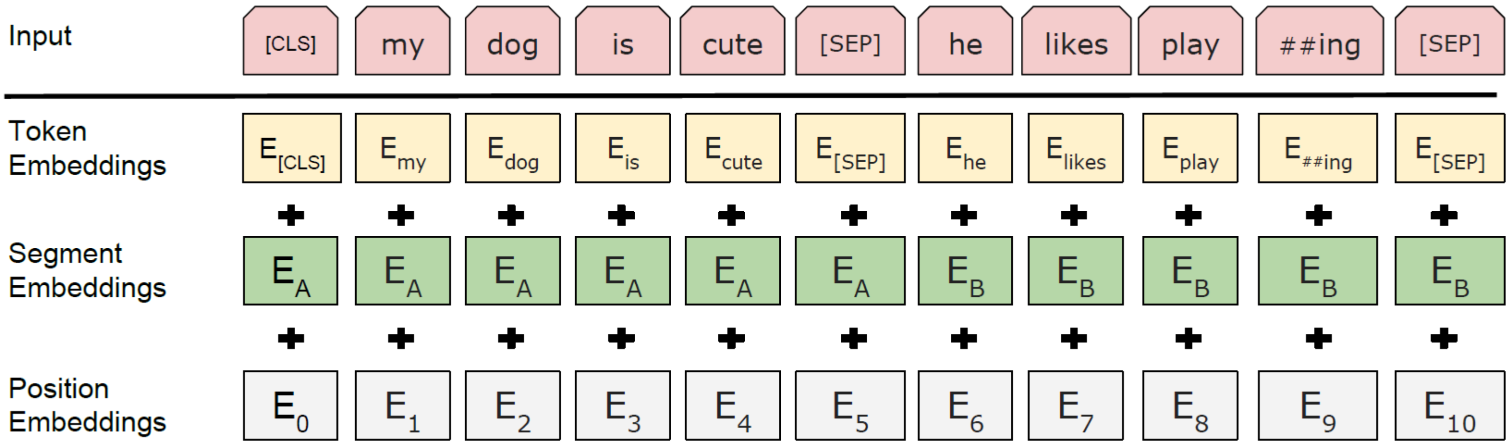
What is the loss function?

Where can we find training data?

How can BERT know the difference between both sentences?



Three Embeddings: Token + Position + Sentence



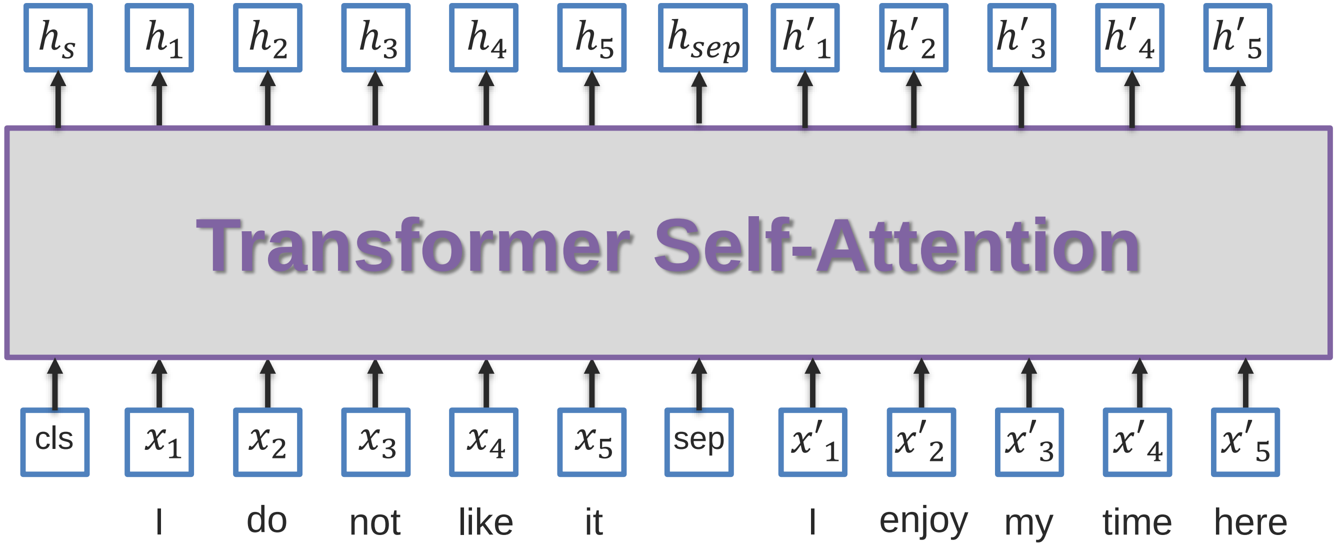
Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

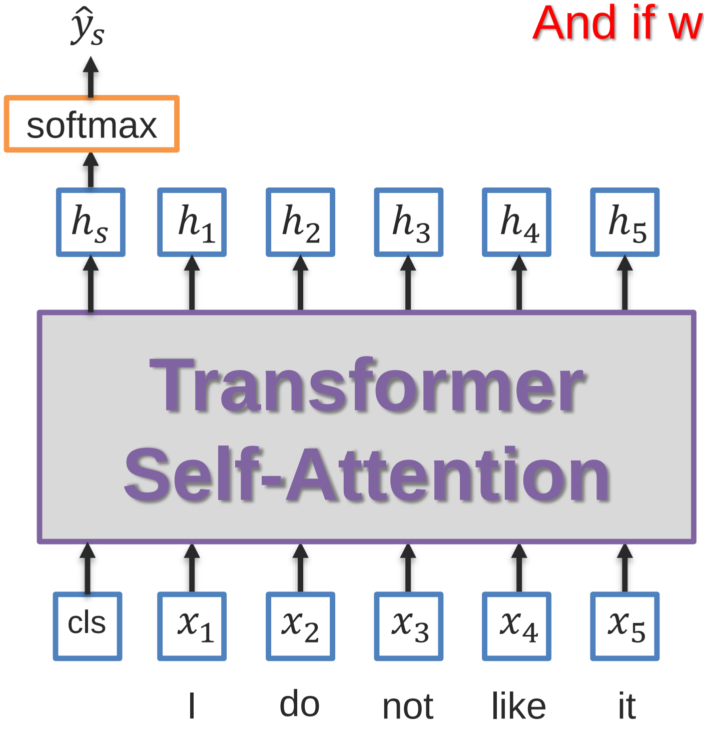
How?



Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

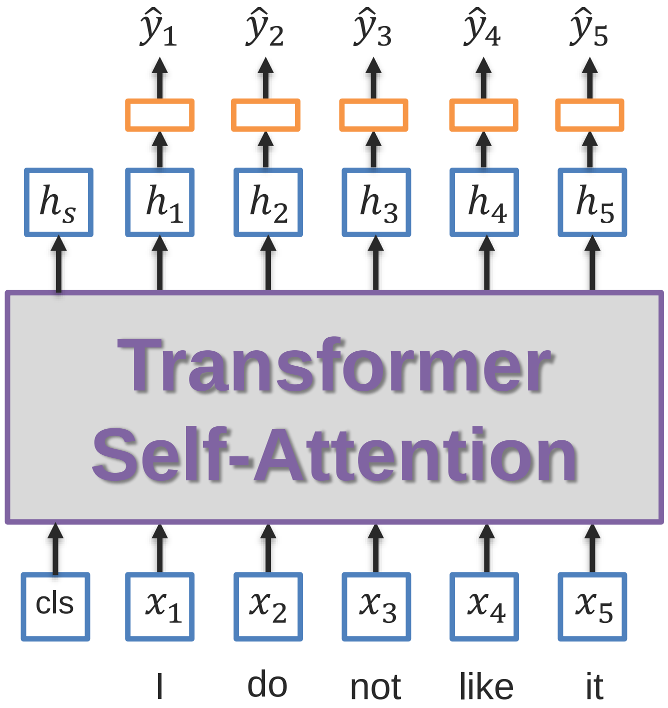

$\hat{y}_s$
softmax
h_s h_1 h_2 h_3 h_4 h_5
Transformer Self-Attention
cls x_1 x_2 x_3 x_4 x_5
I do not like it

And if we have a label for each token?

Fine-Tuning BERT

2 Token-level classification for only one sentence

Examples: part-of-speech tagging, slot filling

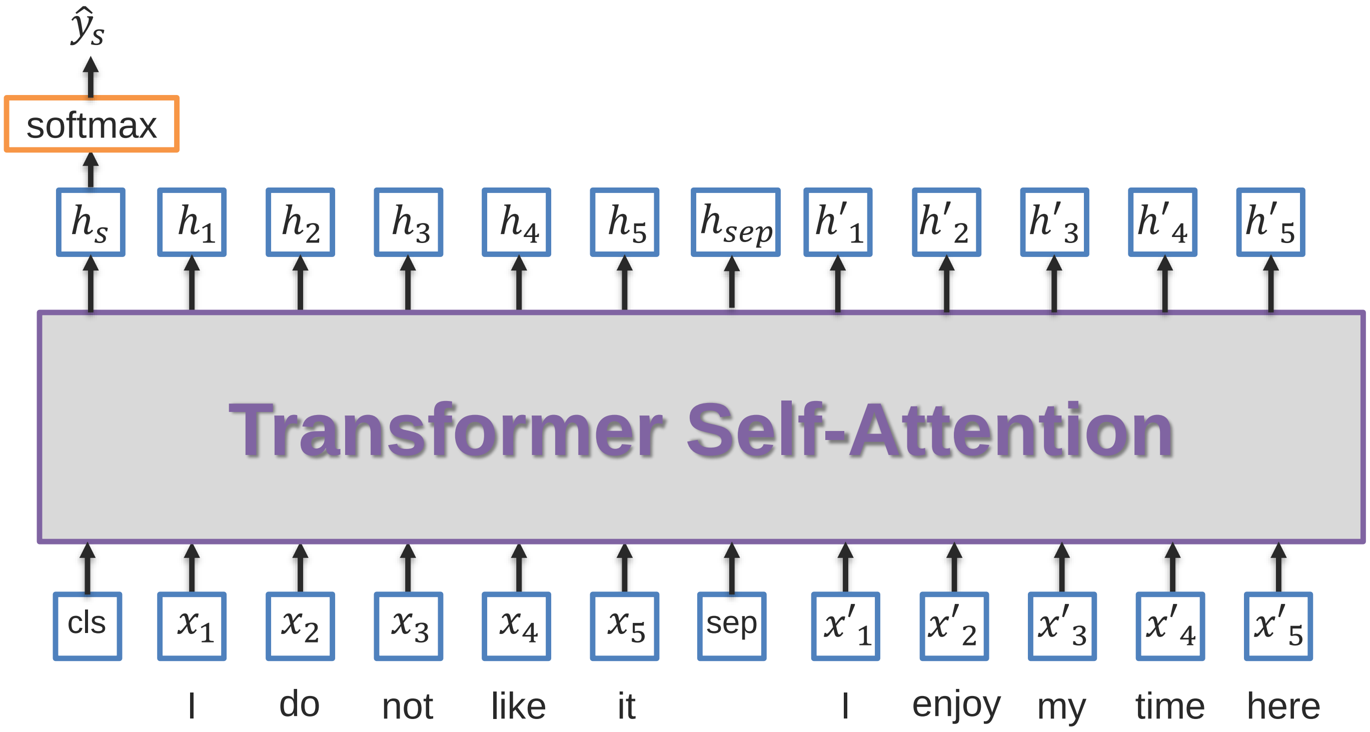


How to compare two sentences?

Fine-Tuning BERT

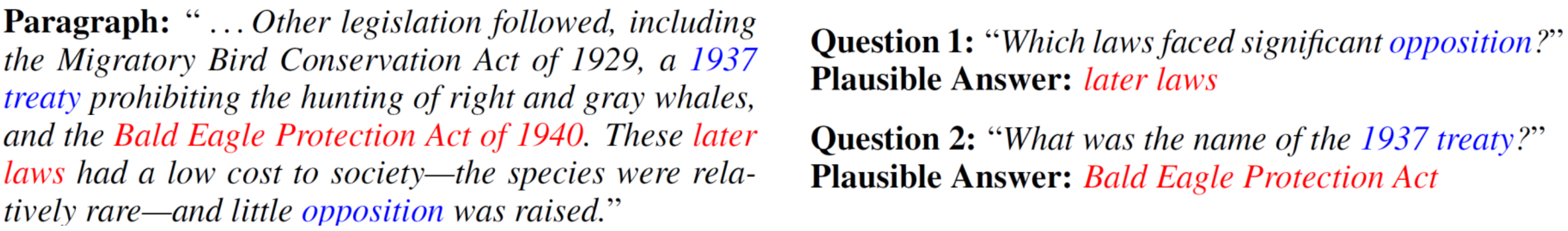
3 Sentence-level classification for two sentences

Examples: natural language inference



Fine-Tuning BERT

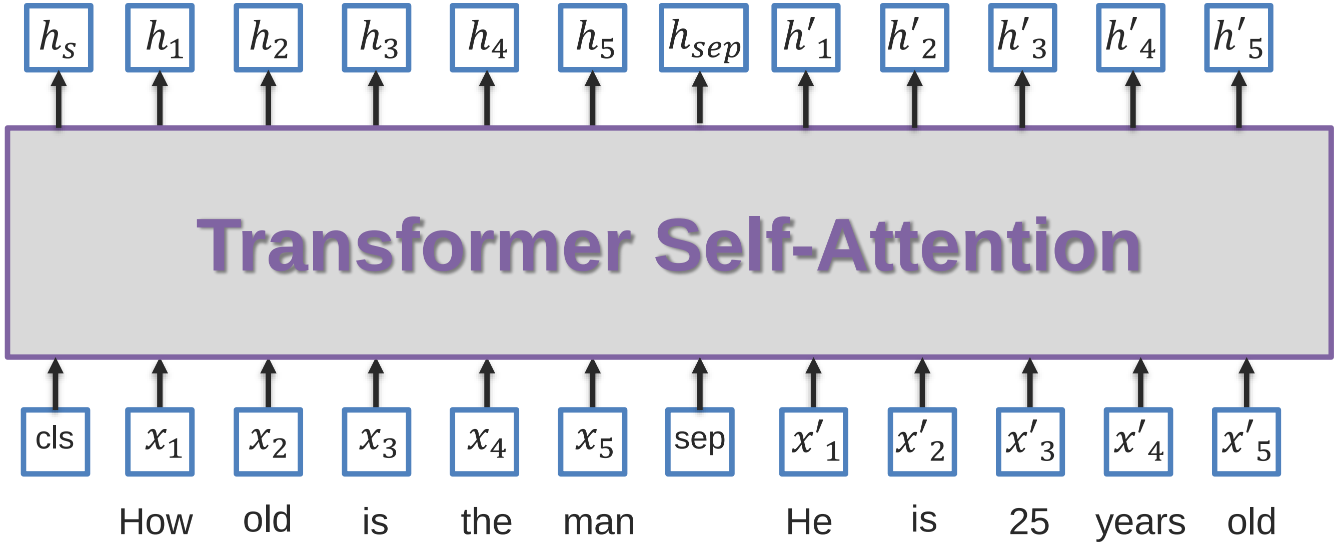
4 Question-answering: find start/end of the answer in the document



Paragraph: " *. . . Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.*"

Question 1: "*Which laws faced significant opposition?*"
Plausible Answer: *later laws*

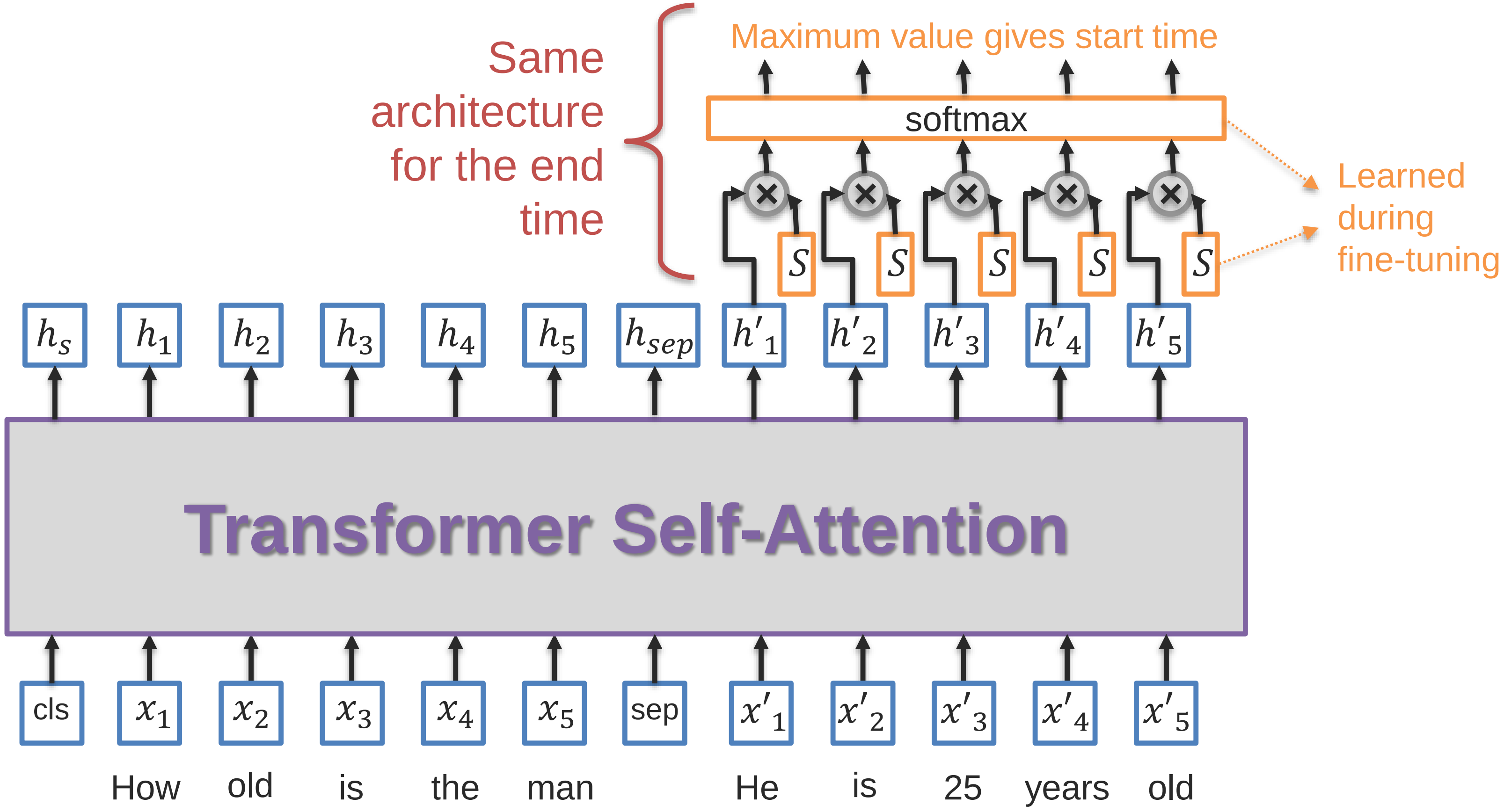
Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*



How?

Fine-Tuning BERT

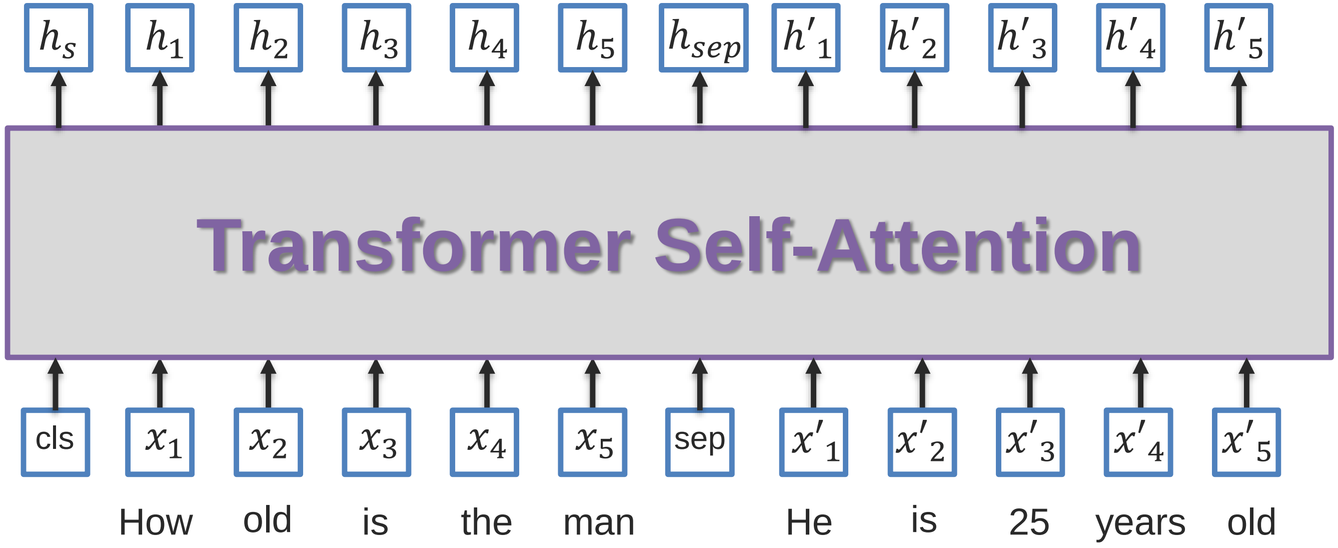
4 Question-answering: find start/end of the answer in the document



Multimodal Pre-training

Multimodal Pre-Training

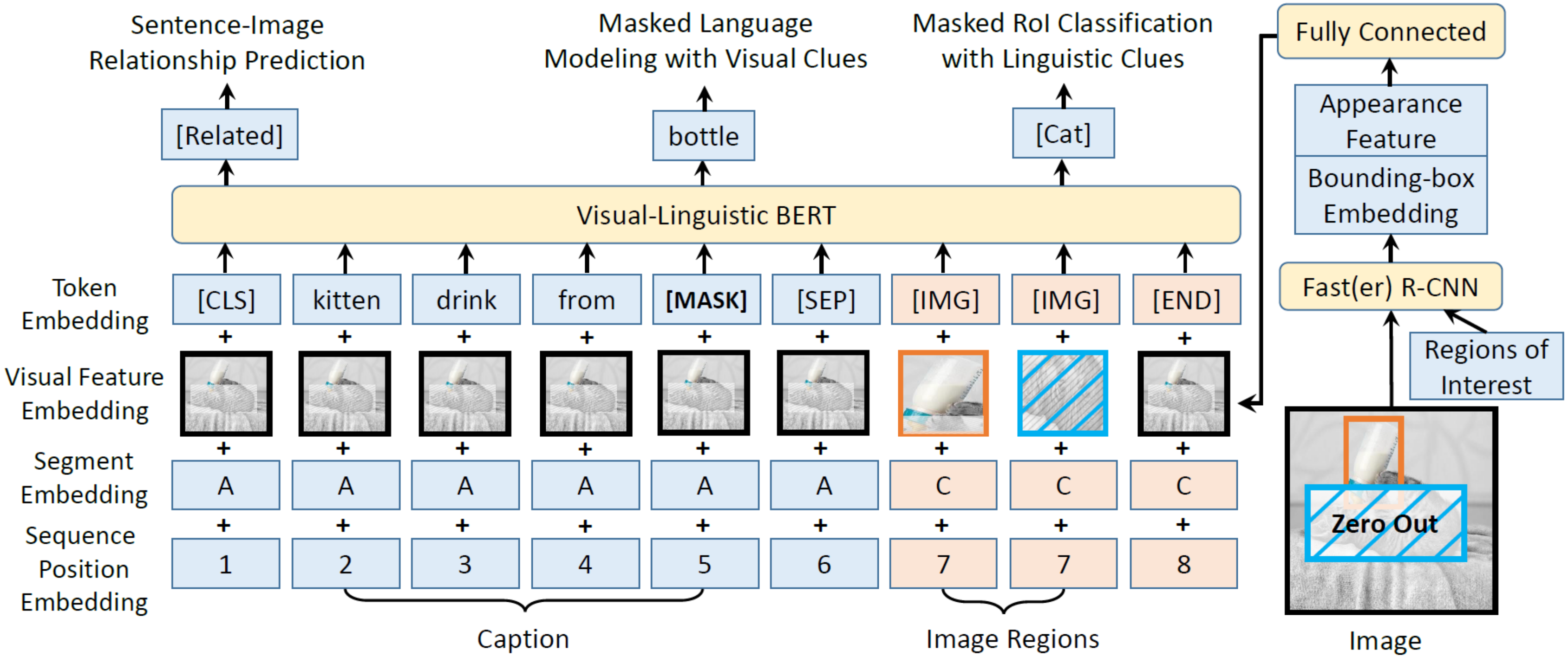
How to extend to multimodal modalities?



VL-BERT

How to extend to multimodal modalities?

Option 1: Simply concatenate tokens from different modalities

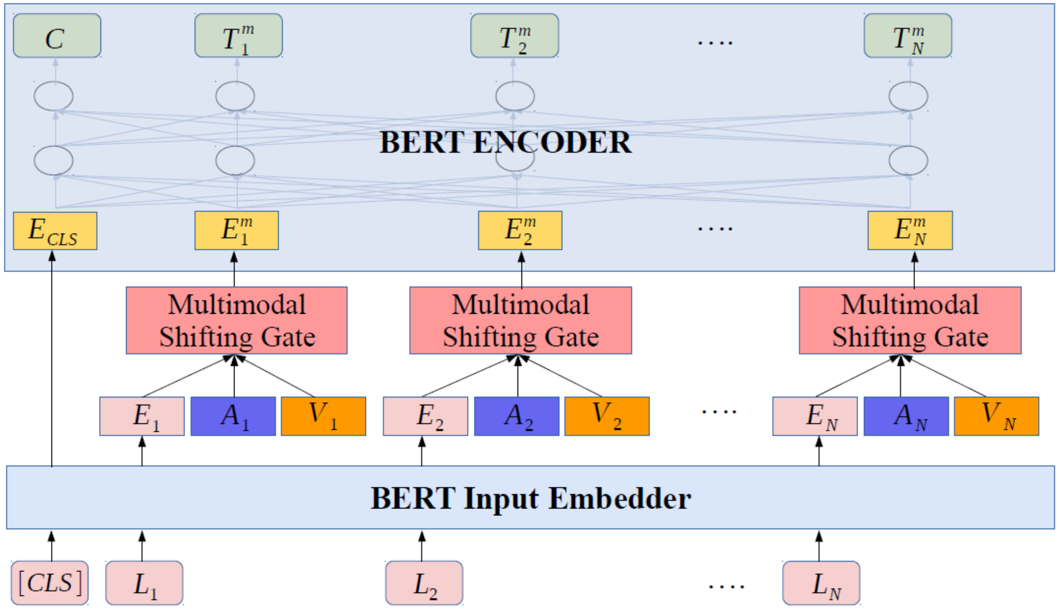


https://arxiv.org/pdf/1908.08530.pdf

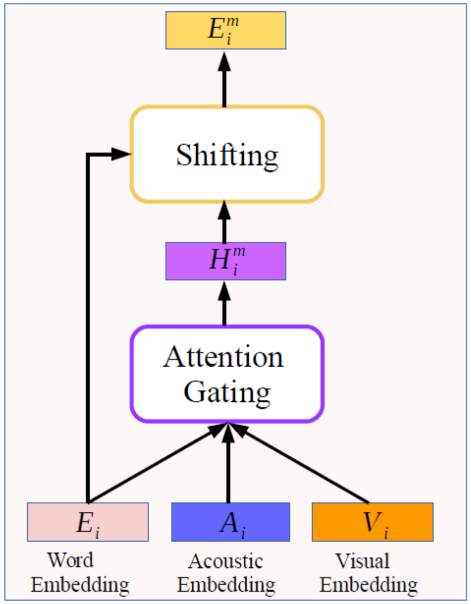
M-BERT

How to extend to multimodal modalities?

Option 2: "Shift" language representation based on the other modalities



(a) Multimodal BERT



(b) Multimodal Shifting Gate

https://arxiv.org/pdf/1908.05787.pdf