Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.1: Multimodal Transformers (Part 1)

Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Administrative Stuff

Second Project Assignment (Due Sunday 10/8)

Main goals:

1. Help clarify and expand your research ideas
   - Build qualitative intuitions by directly studying the original data
   - Perform analyses on your dataset, relevant to your research ideas
2. Understand the structure in your data and modalities
   - Perform analyses and visualizations to understand each modality
   - Study representations from language and visual modalities

Two types of analyses:

- Idea-oriented analyses
- Modality-oriented analyses

Second Project Assignment (Due Sunday 10/8)

Examples of **idea-oriented** analyses:

- What external knowledge is needed when performing the task?
- How often multimodal information is needed? How is it integrated?
- What biases may be present in the data? Which modalities?

Examples of **modality-oriented** analyses:

- What are the different verbs used in the VQA questions?
- What objects do not get detected? Are they important?
- Visualize face embeddings with respect of emotion labels

Second Project Assignment (Due Sunday 10/8)

Idea-oriented analyses:

- **Human simulations:** Instead of a computer, try to do the same task as a human. Gather notes on how you perform the task.
- **Data analysis:** study the multimodal data (e.g., using statistical methods) to clarify your hypotheses related to your research ideas

Modality oriented analyses:

- **Language modality:** explore the language structure in your dataset. You can compare word-level and sentence-level embeddings.
- **Visual modality:** study visual representations for your dataset. You visualize how your visual features successfully model your labels.

Second Project Assignment (Due Sunday 10/8)

Number of analyses:

- Teams of 3 students: 2 analyses (4 pages)
- Teams of 4 students: 3 analyses (5 pages)
- Teams of 5 students: 4 analyses (6 pages)

- ➢ You can mix and match between idea-oriented and modality-oriented
- ➢ Be sure to talk with your TA about formalizing your analysis plan
- ➢ Each analysis need a separate discussion section

Detailed instructions on Piazza (Resources section)

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.1: Multimodal Transformers (Part 1)

Louis-Philippe Morency

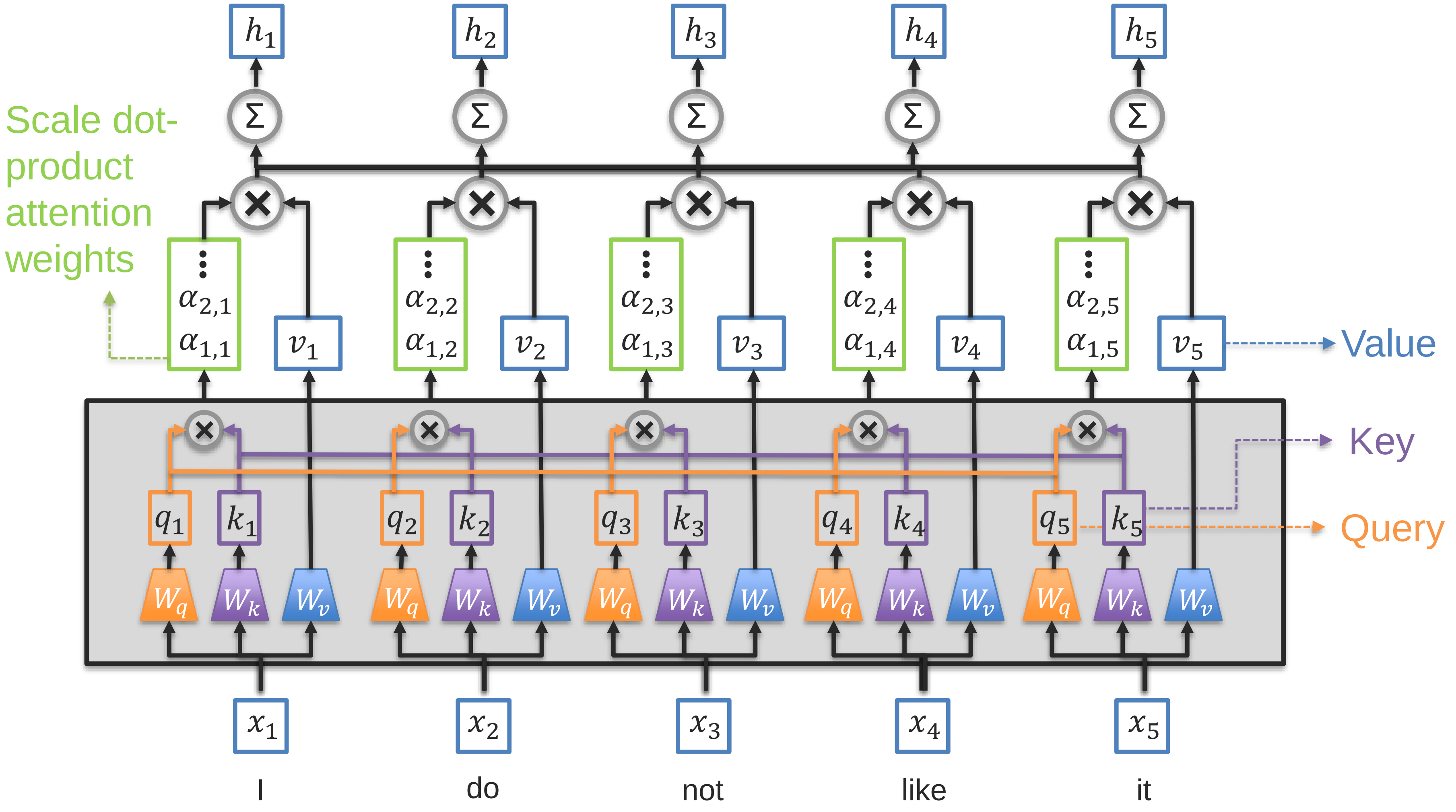
** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Objectives of today's class

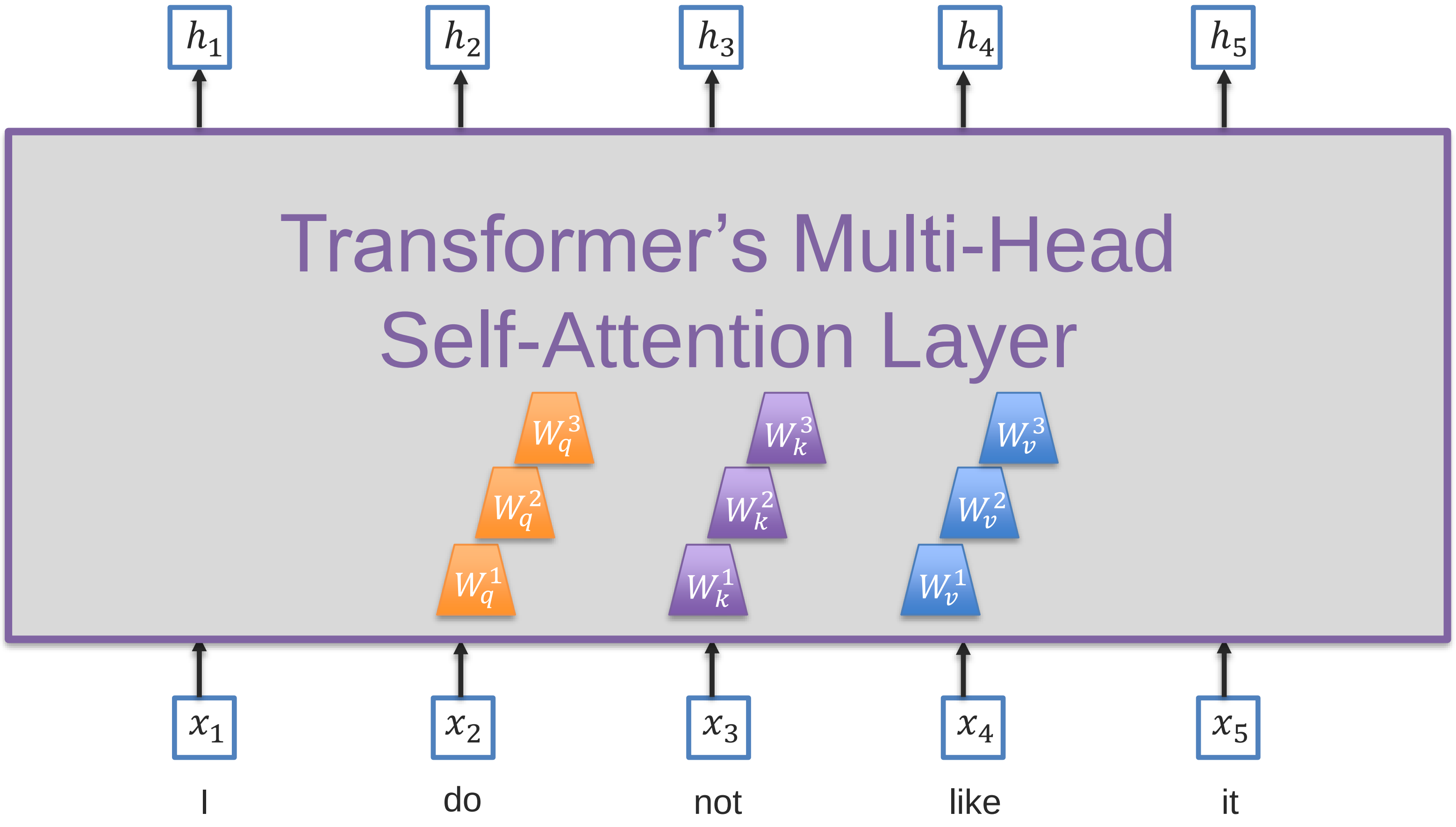
- Positional embeddings
- Language pre-training
  - BERT: Bidirectional Encoder Representations from Transformers
- Multimodal transformers (Image and language)
  - Concatenated transformers (VisualBERT, Uniter)
  - Crossmodal transformers (ViLBERT, LXMERT
  - Modality-shift transformer (MAG-BERT)
- Sequence-to-sequence modeling with Transformers

Positional Embeddings

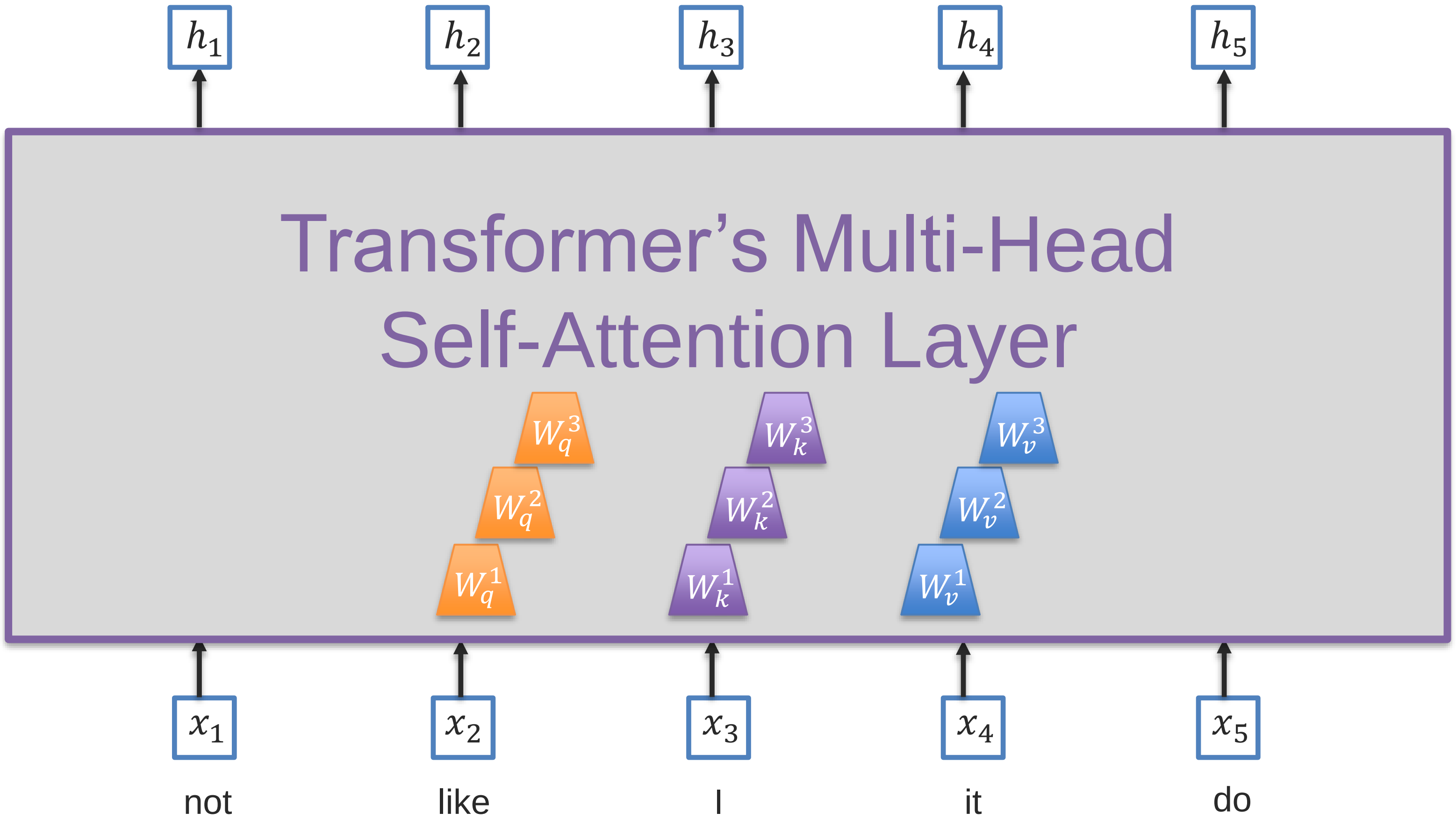
Transformer Self-Attention



Transformer Multi-Head Self-Attention



Transformer Multi-Head Self-Attention



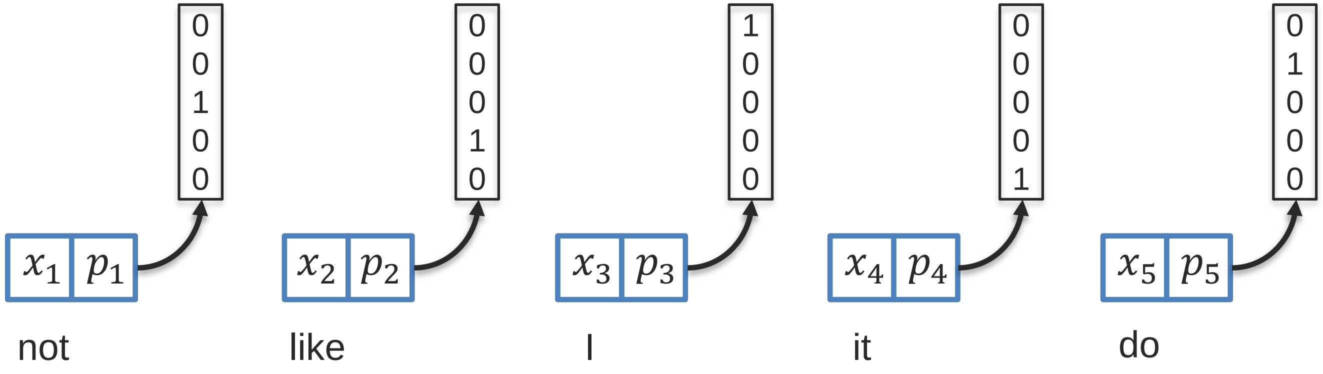
What happens if the words are shuffled?

Position embeddings

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

How can we encode position information?

Simple approach: one-hot encoding

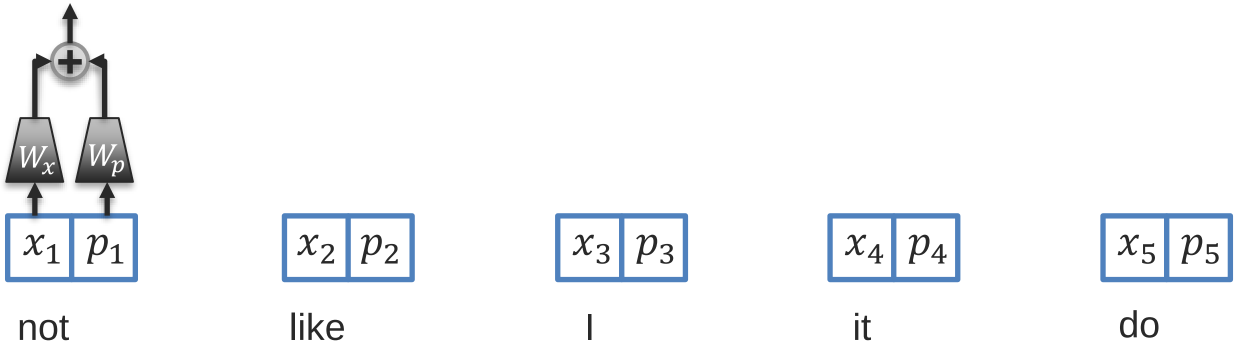


Position embeddings

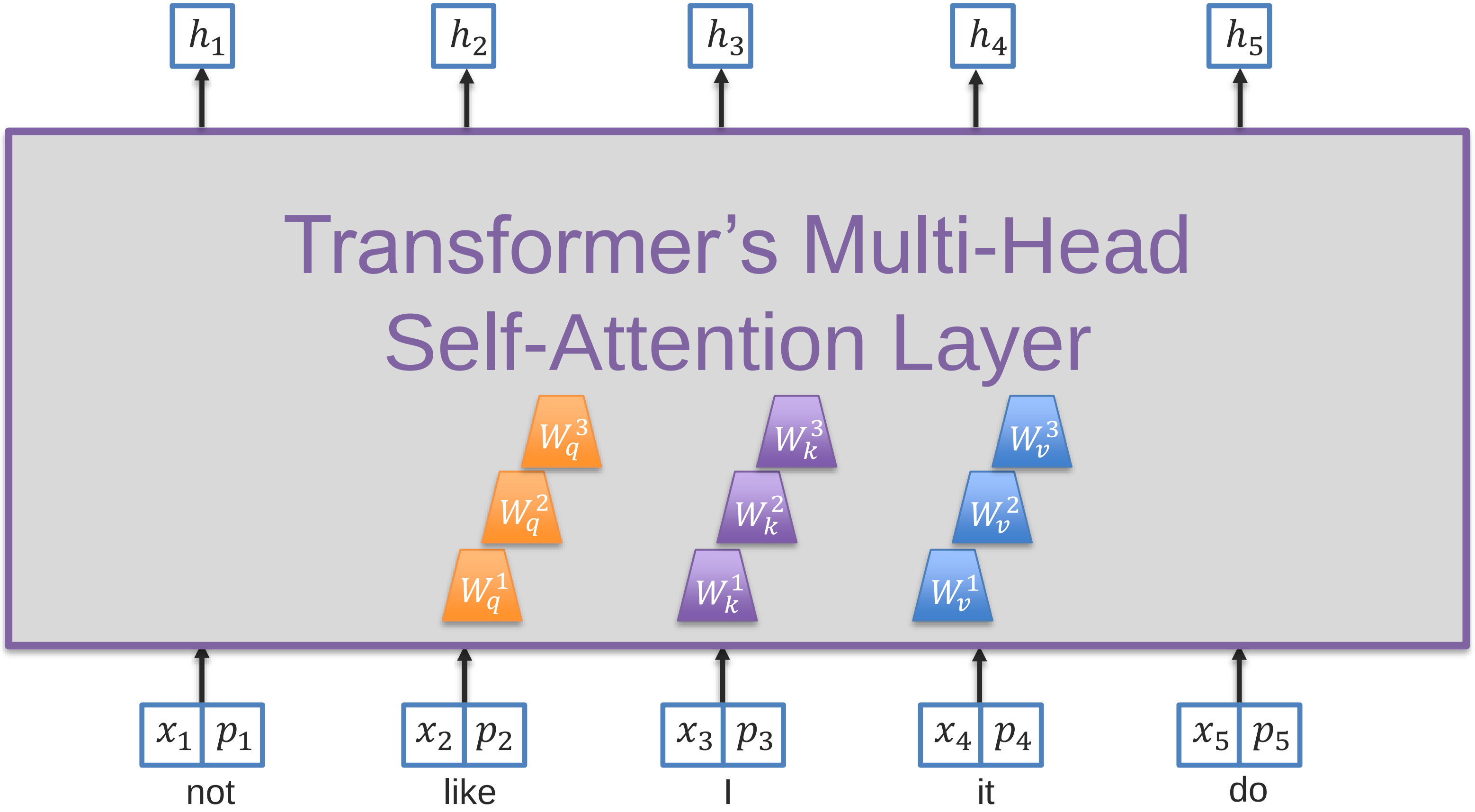
- Position information is not encoded in a self-attention module

How can we encode position information?

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

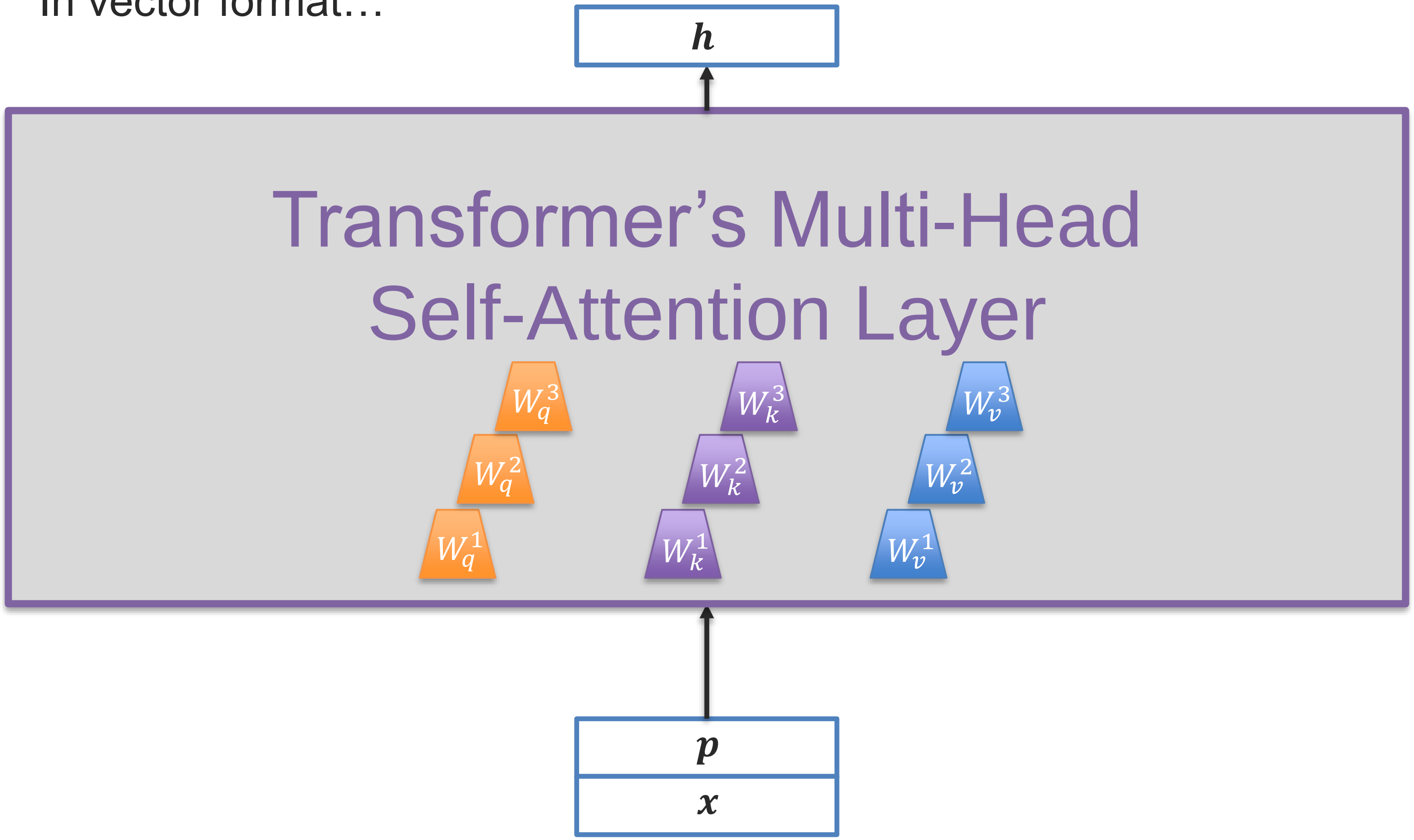


Transformer Multi-Head Self-Attention

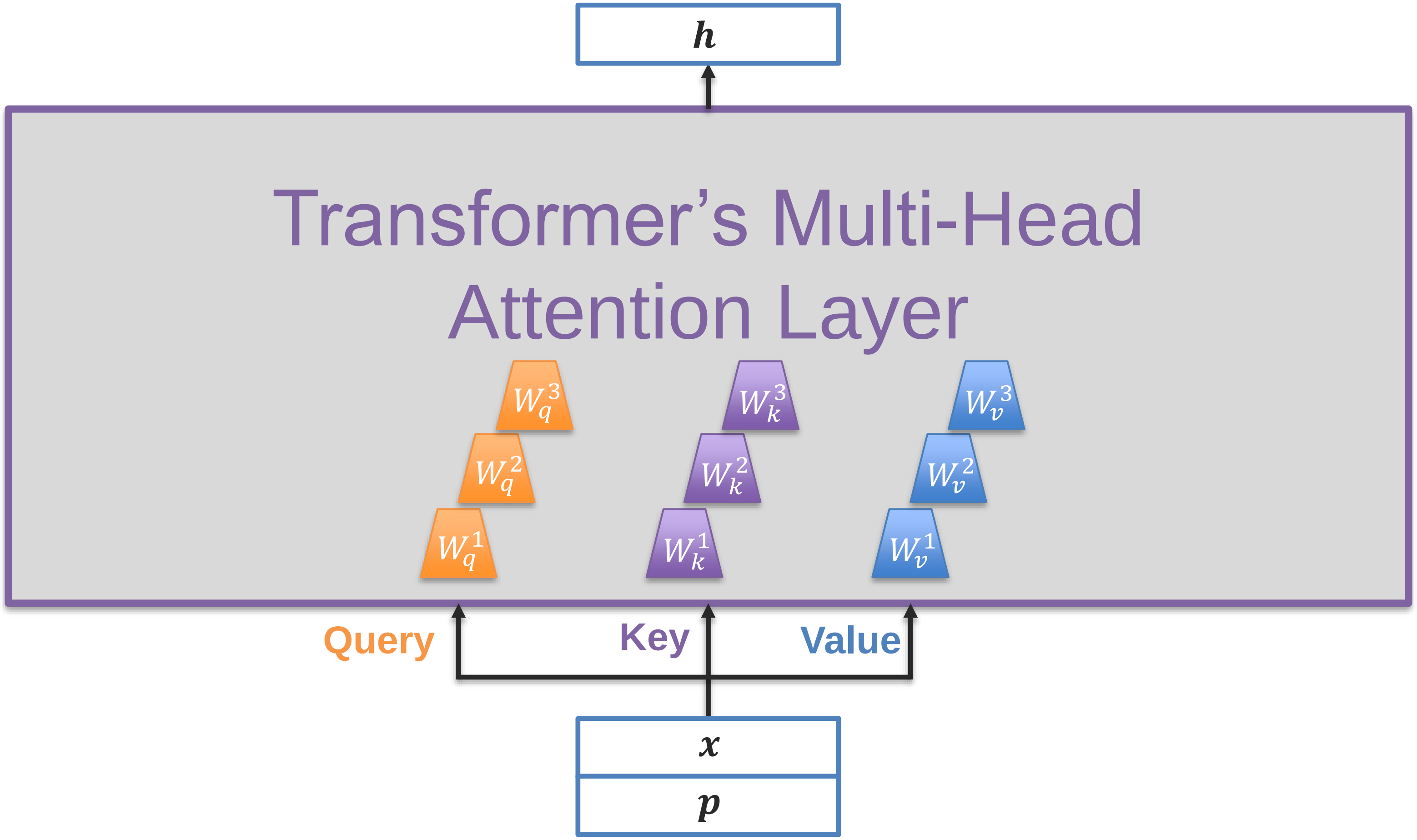


Transformer Multi-Head Self-Attention

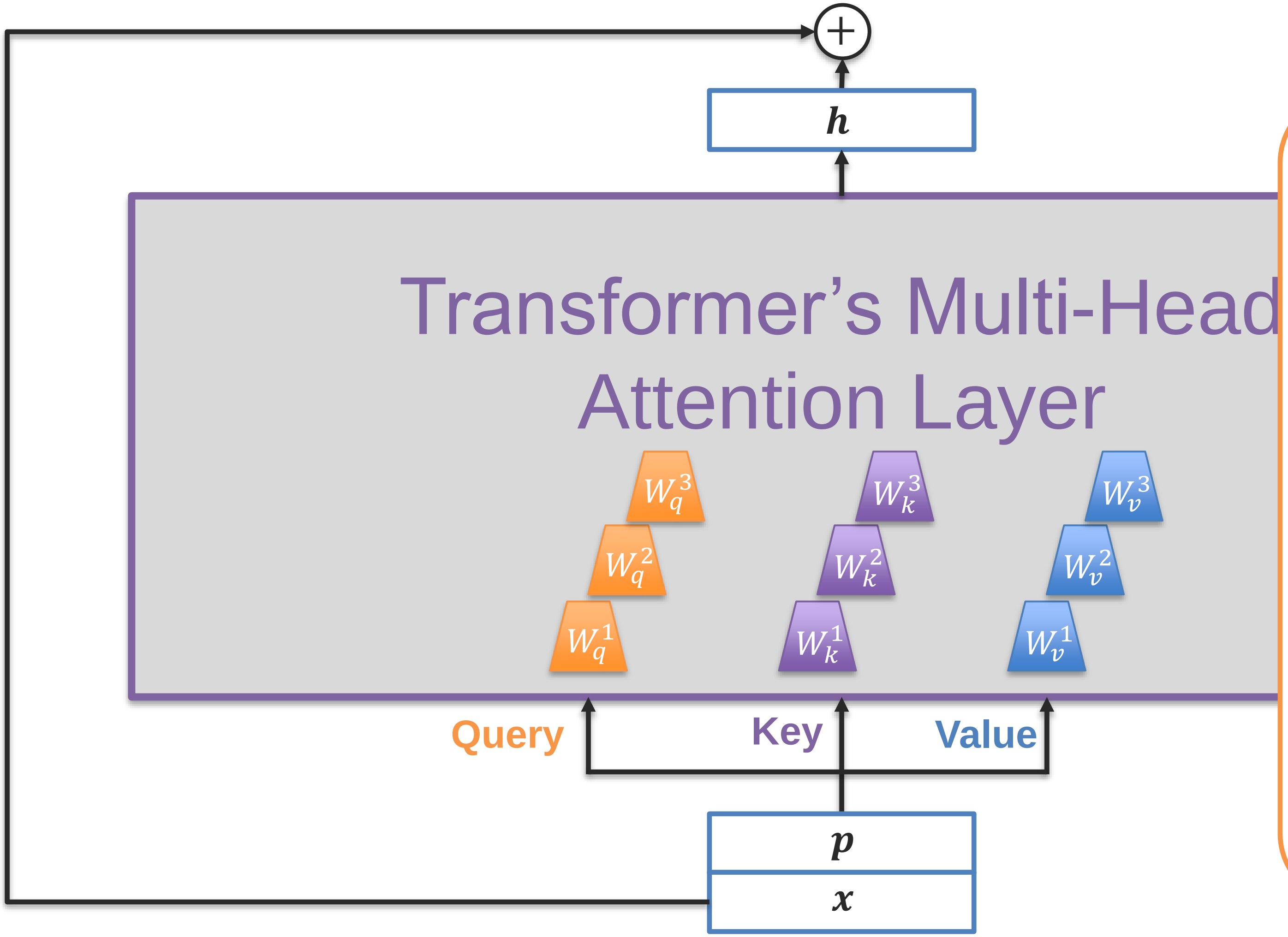
In vector format…

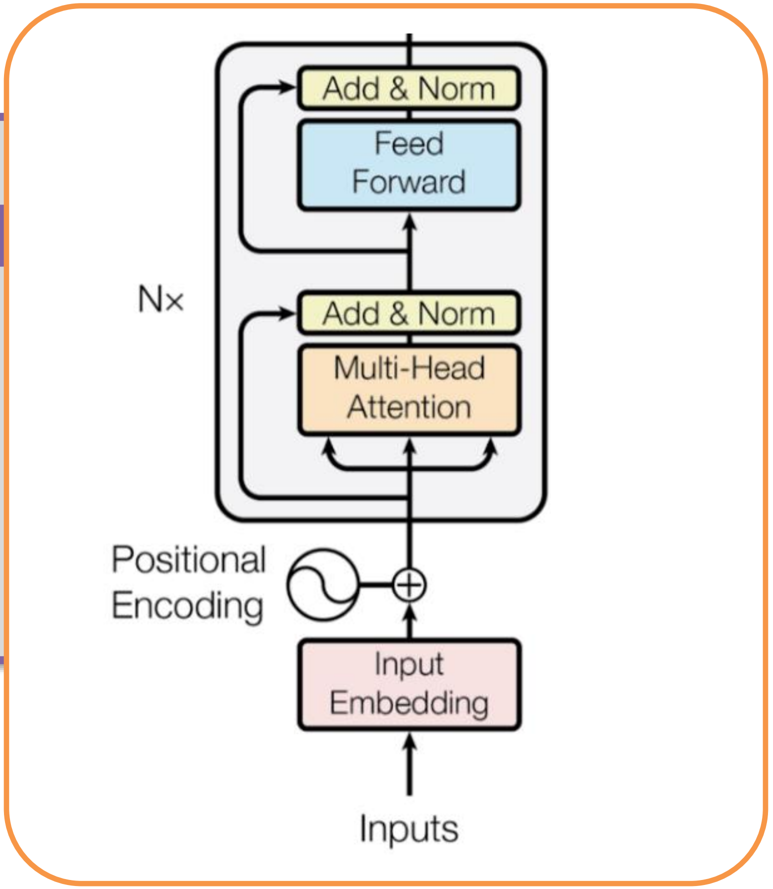


Transformer Multi-Head Attention



Transformer – Residual Connection

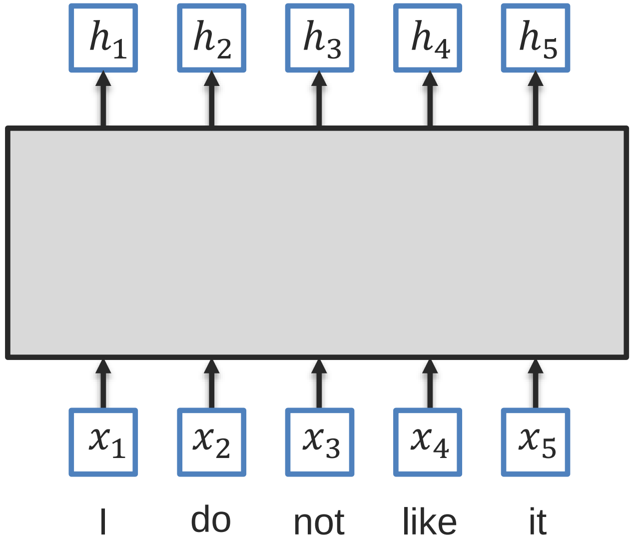




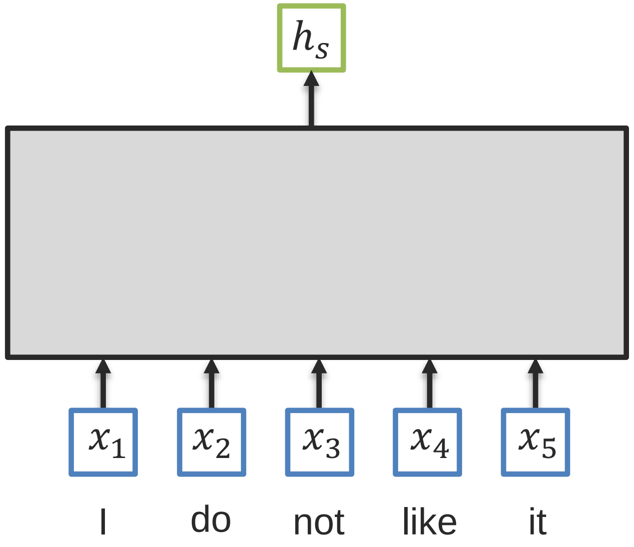
Language Pre-training

Token-level and Sentence-level Embeddings

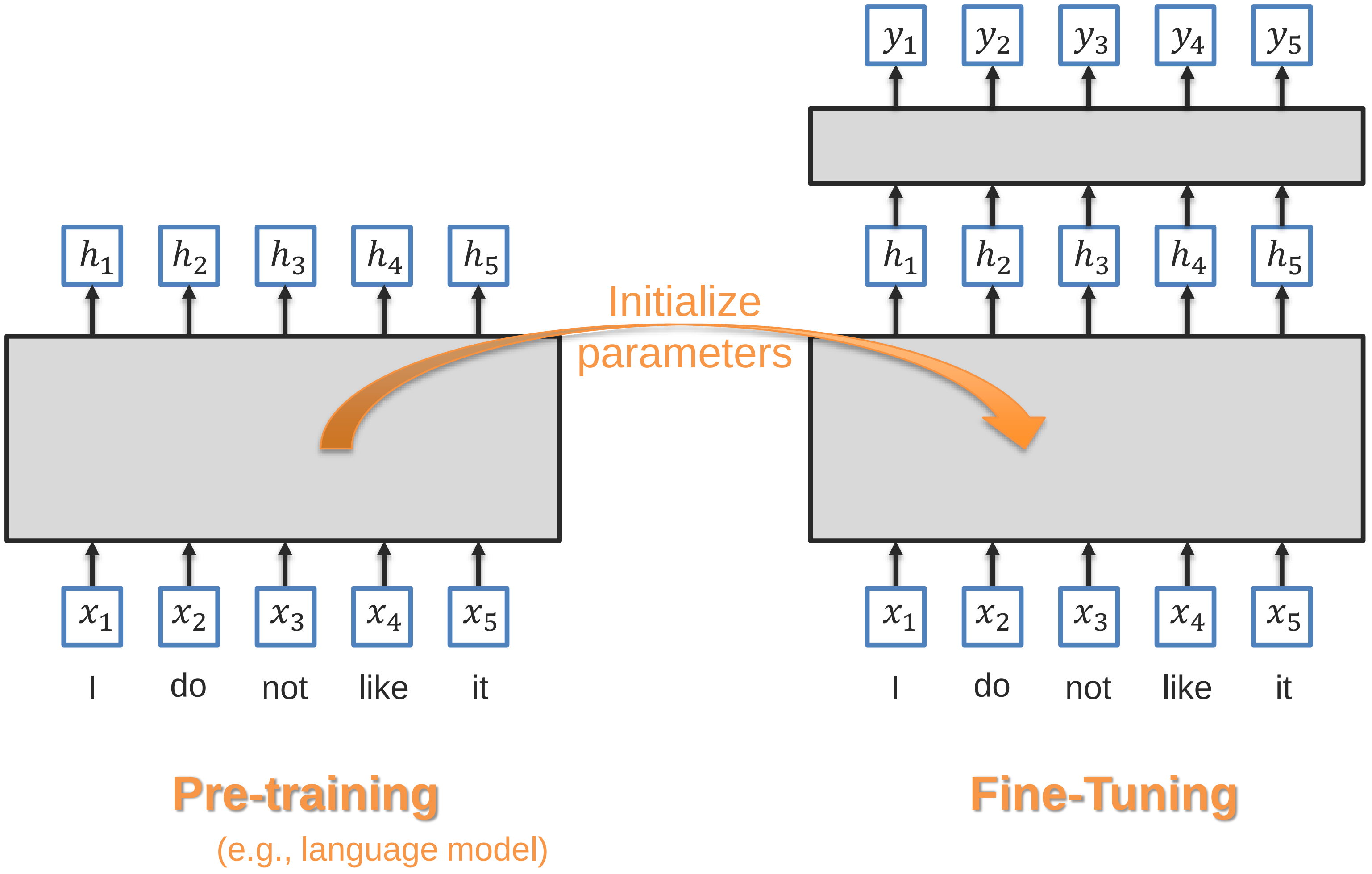
Token-level embeddings



Sentence-level embedding



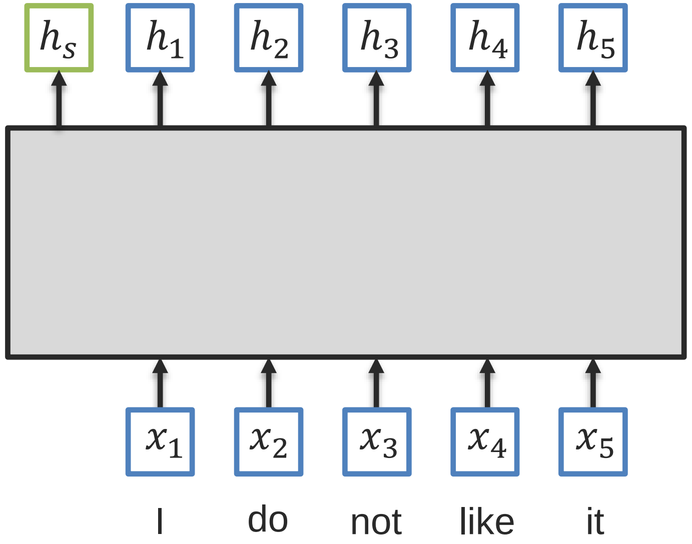
Pre-Training and Fine-Tuning



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

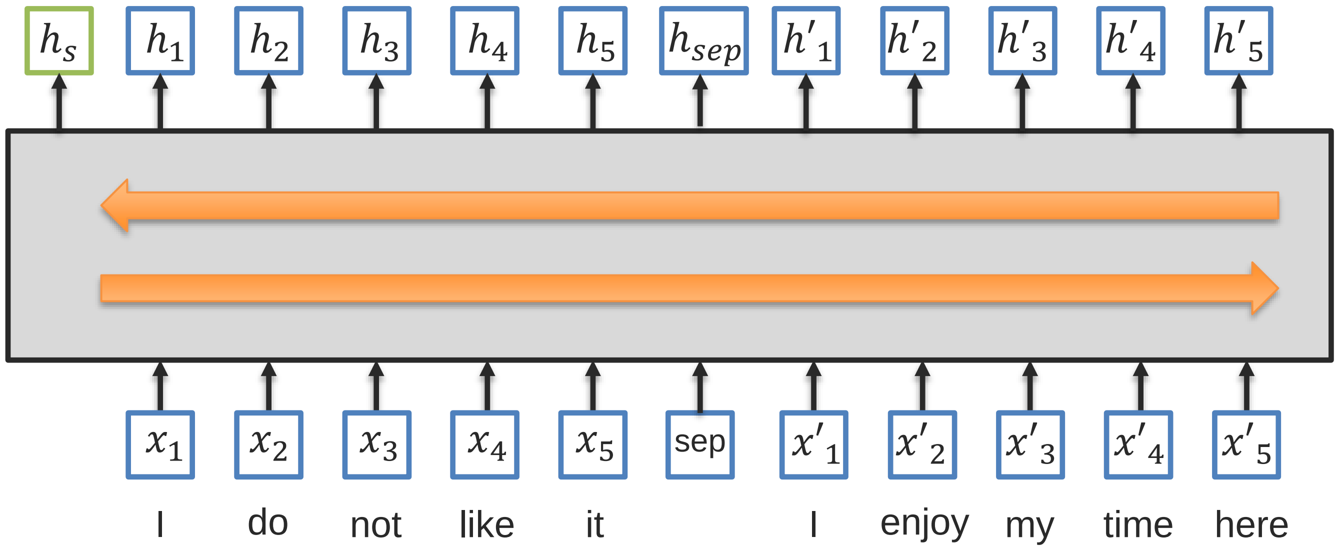


BERT: Bidirectional Encoder Representations from Transformers

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

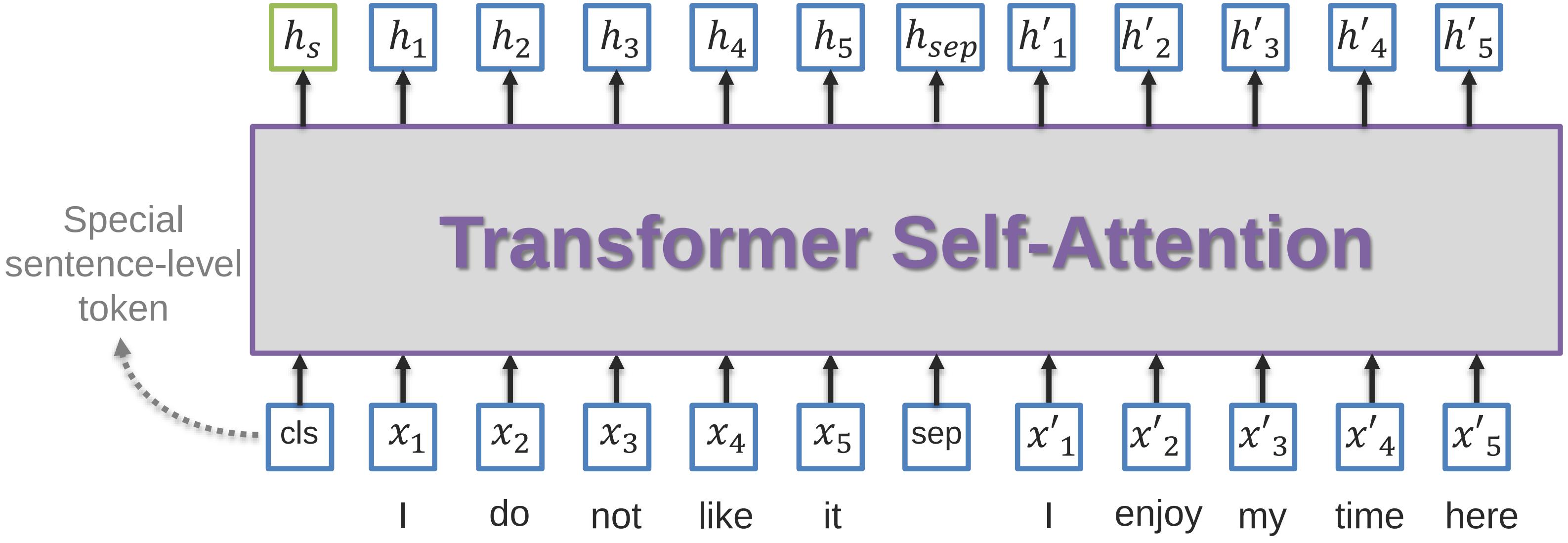
How can we do all this?



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



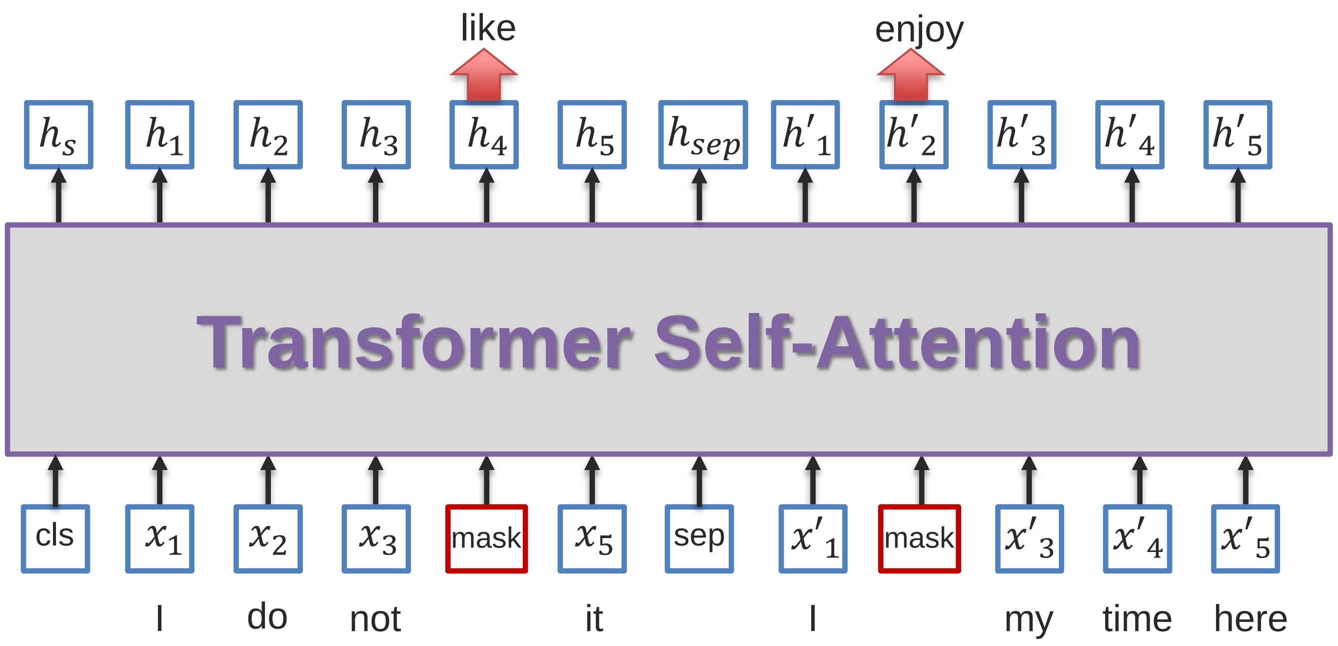
But how to train self-supervised?

Pre-training BERT Model

1 **Masked Language Model**

Randomly mask input tokens and then try to predict them

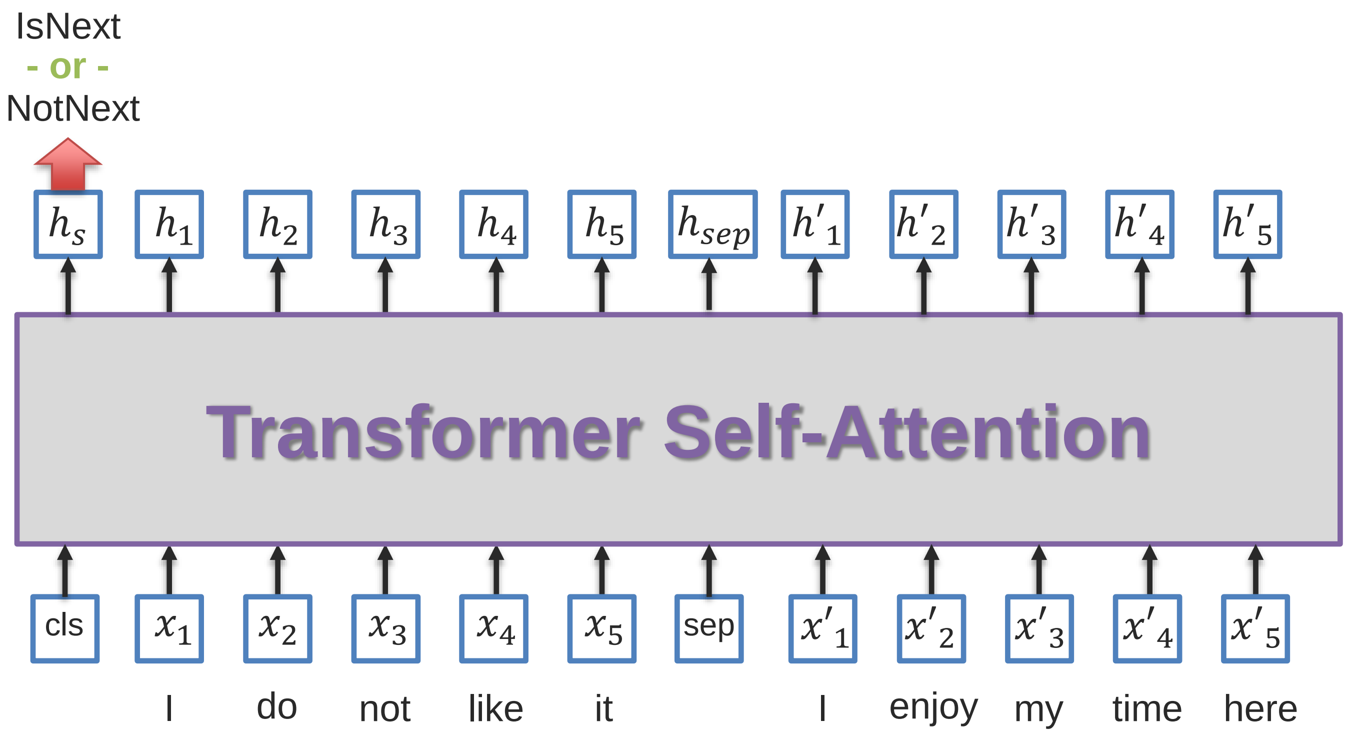
What is the loss function?



Pre-training BERT Model

2 **Next Sentence Prediction**

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

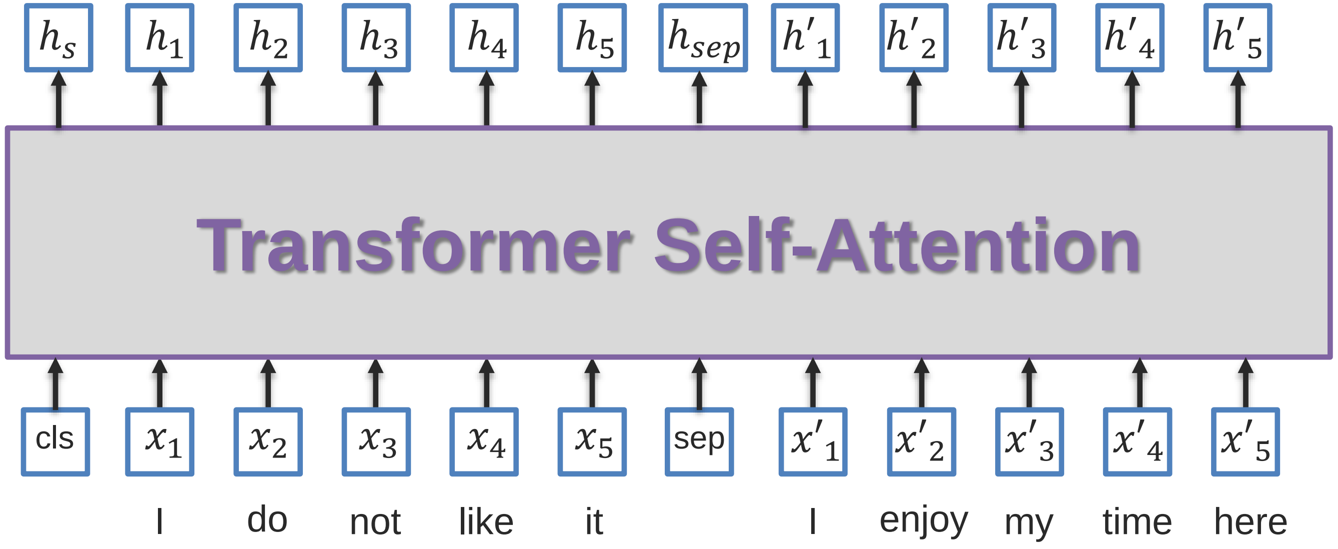


Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

How?

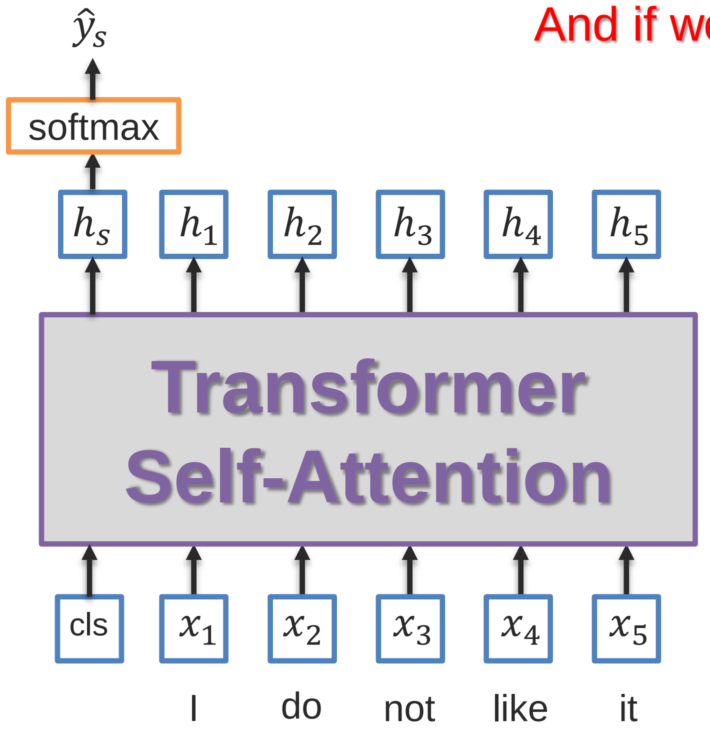


Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

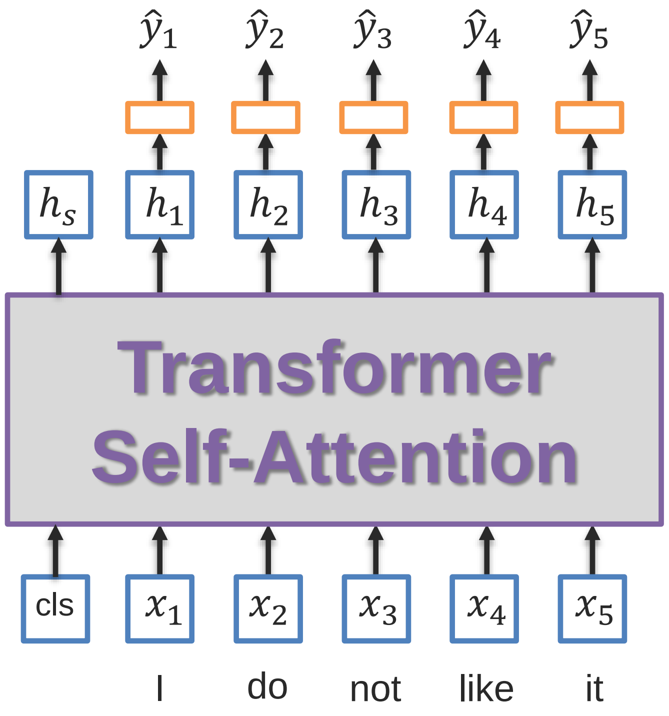
And if we have a label for each token?



Fine-Tuning BERT

2 Token-level classification for only one sentence

Examples: part-of-speech tagging, slot filling



How to compare two sentences?

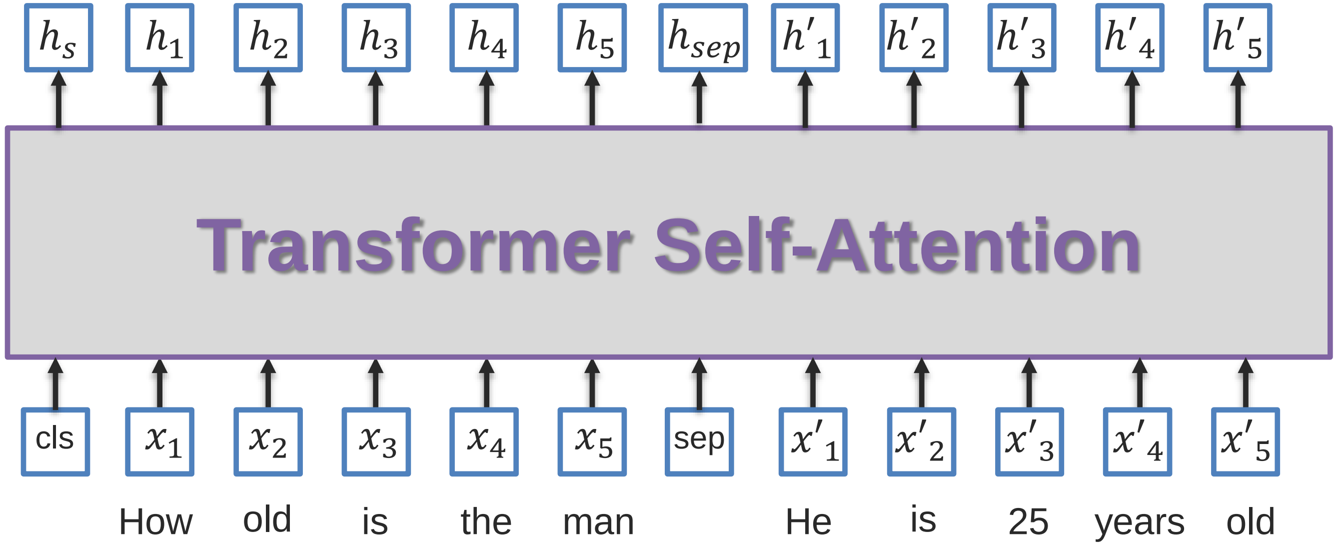
Fine-Tuning BERT

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*

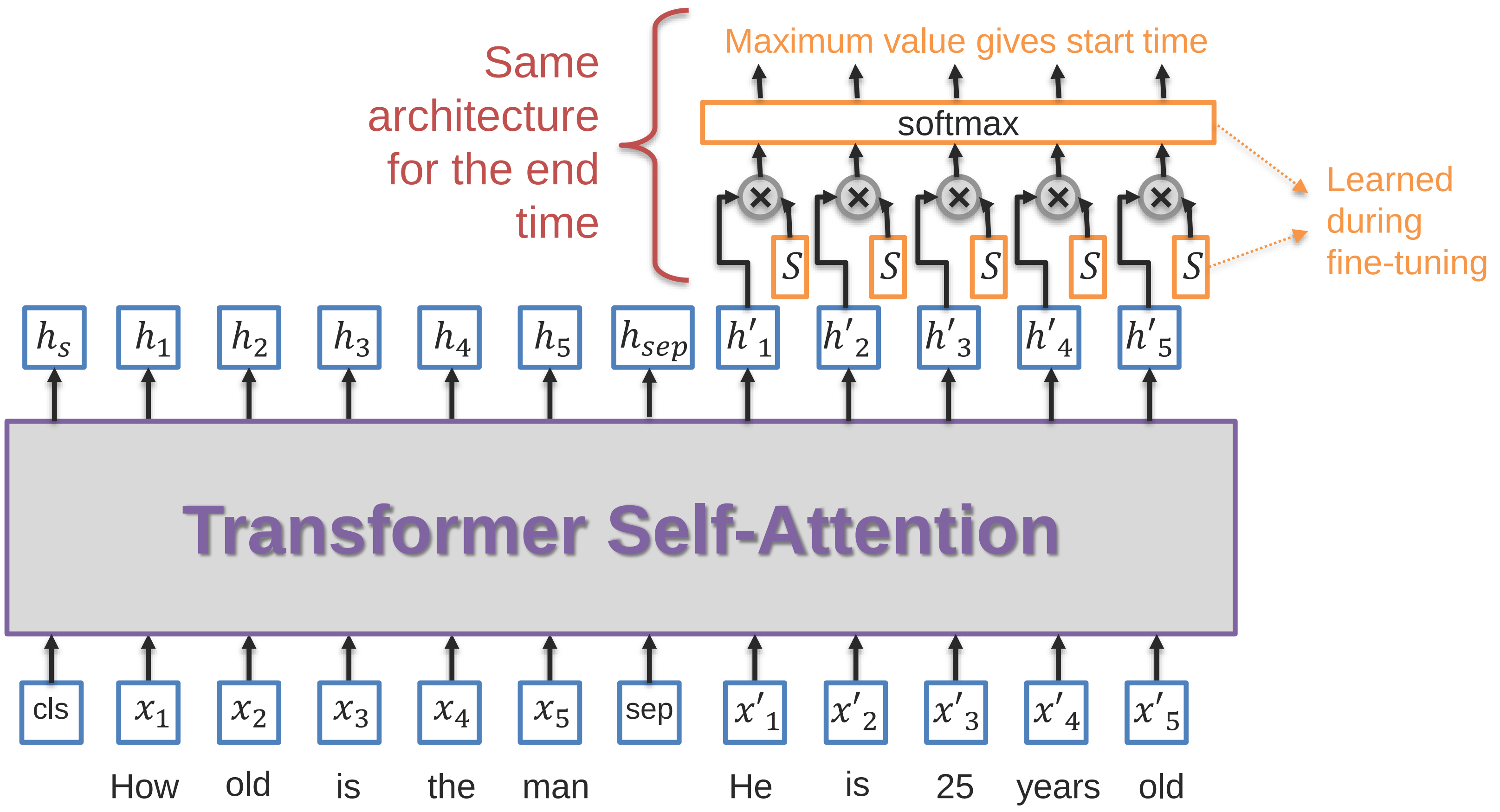
Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*



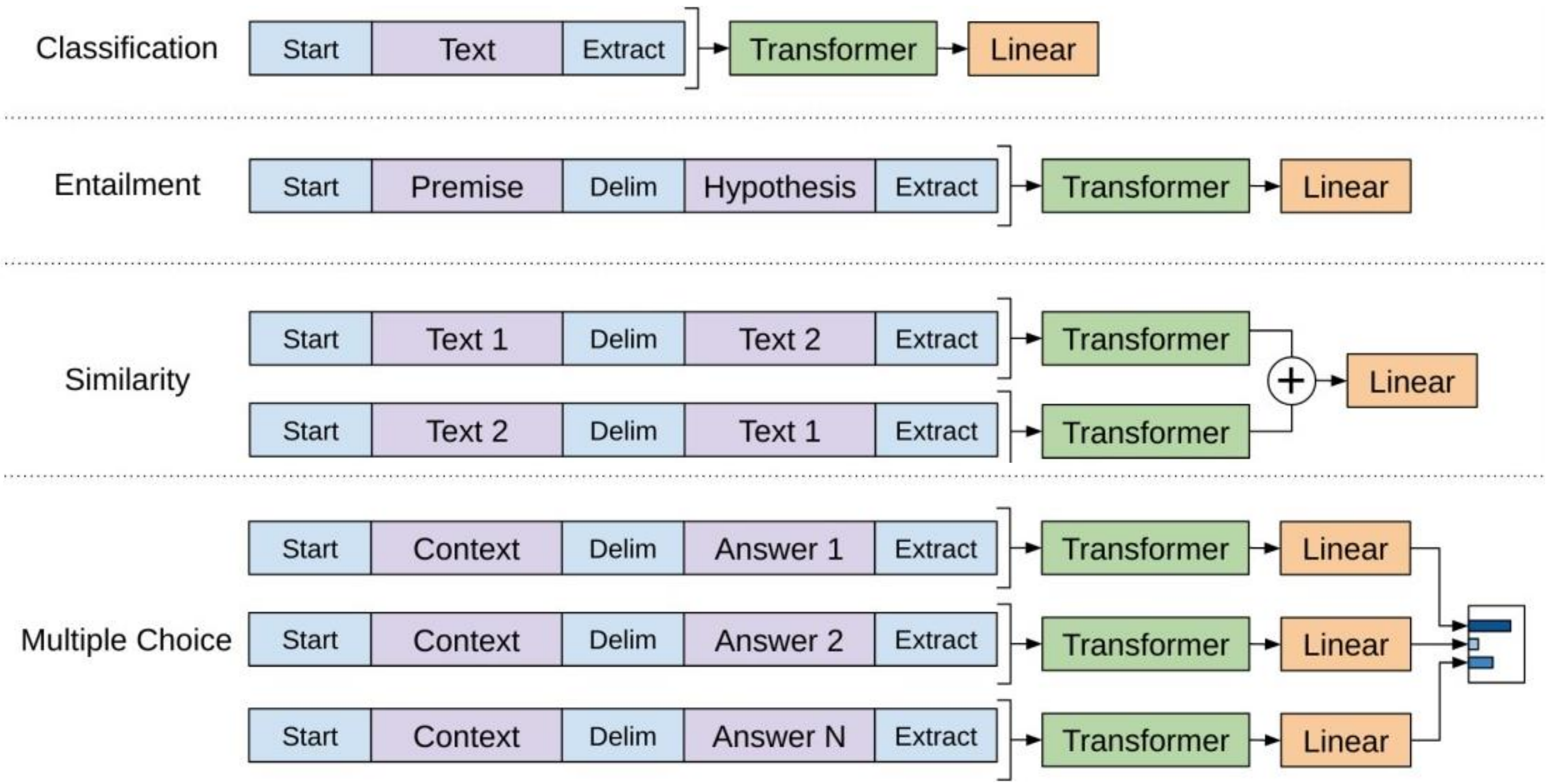
How?

Fine-Tuning BERT

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



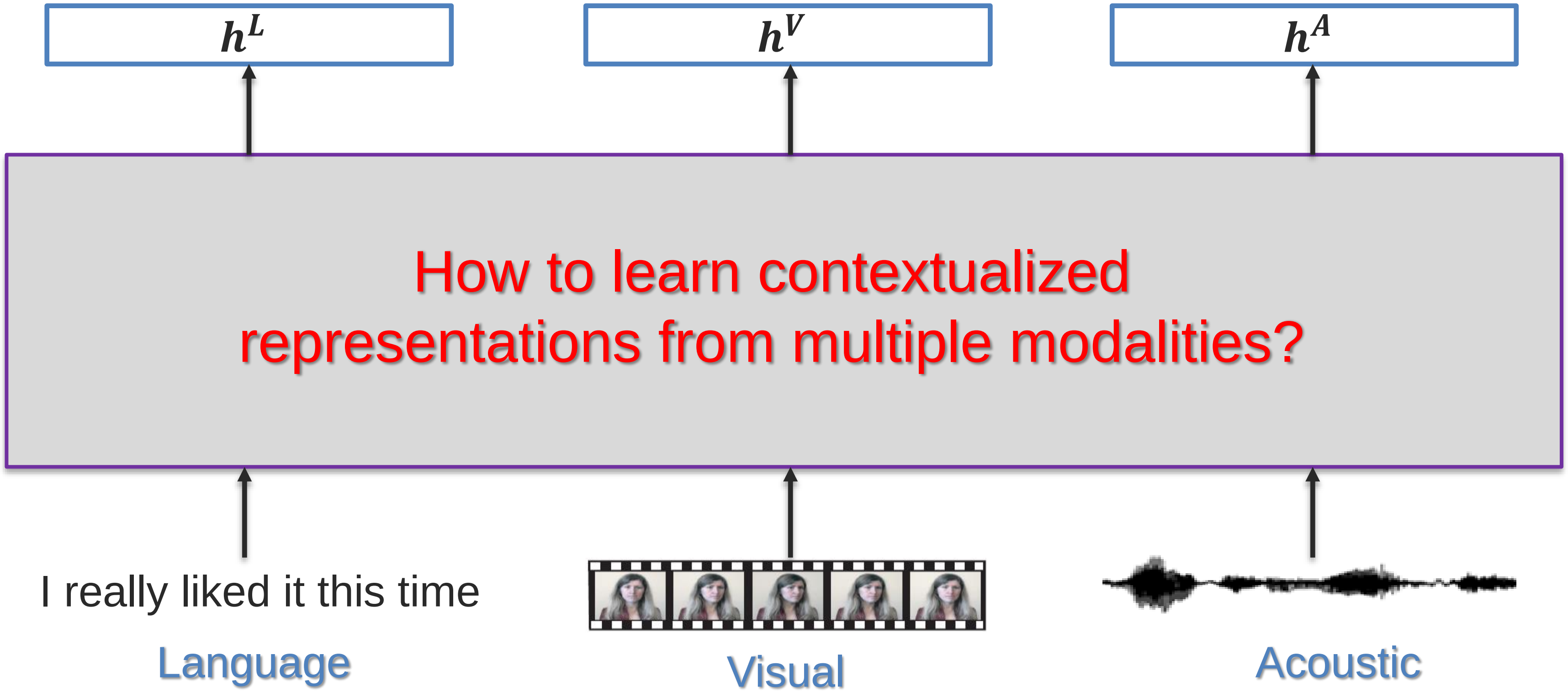
Other Fine-tuning Approaches



https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

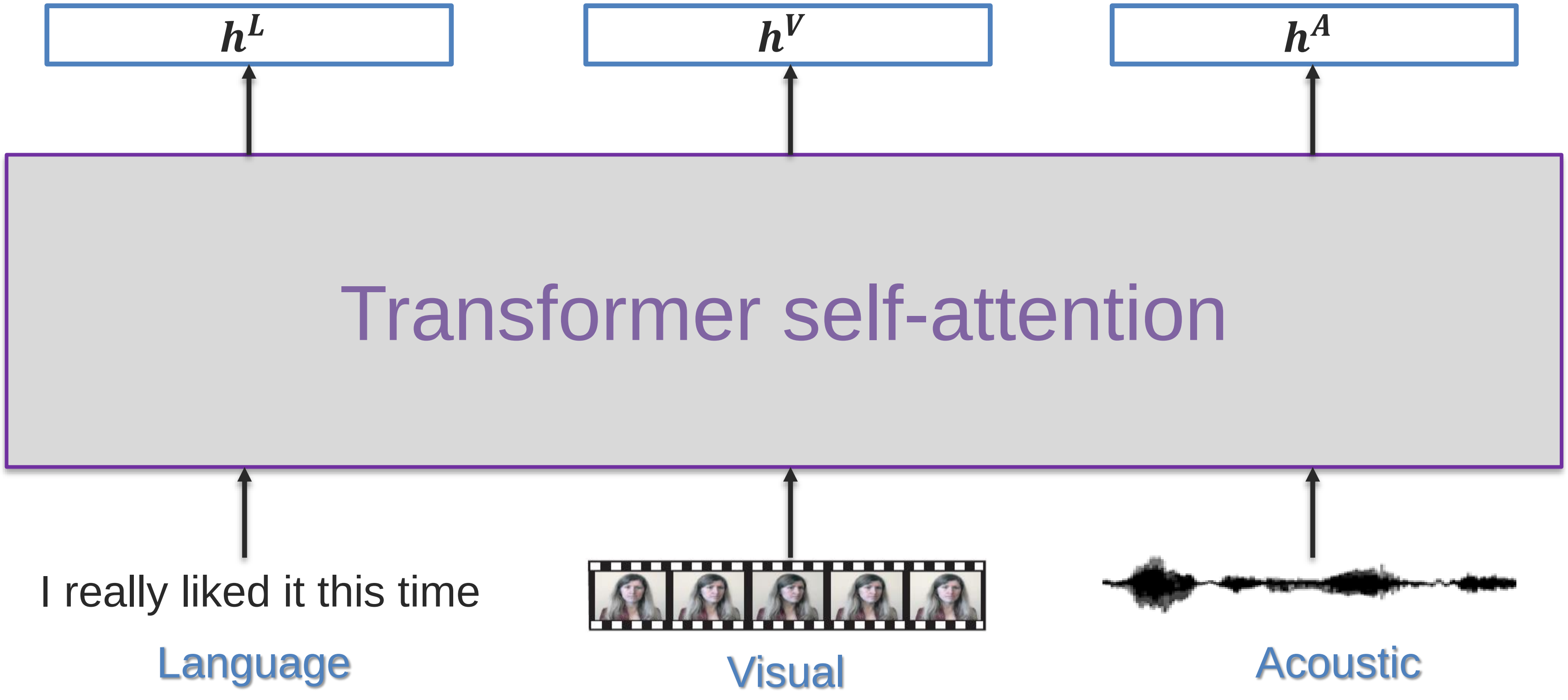
Language-Vision Transformers

Multimodal Embeddings

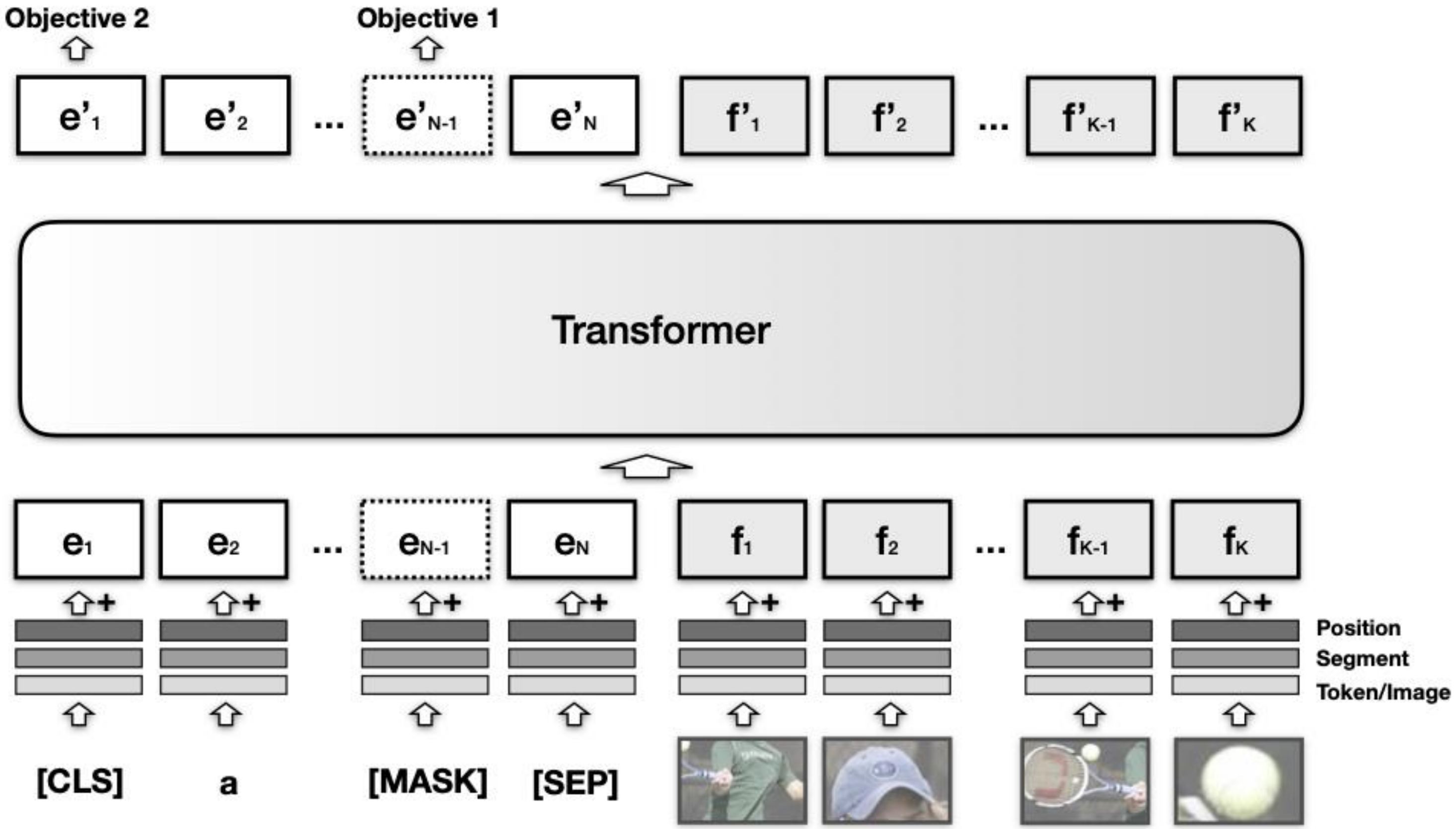


Option 1: Concatenate modalities and learn BERT transformer

Simple Solution: Contextualized Multimodal Embeddings



VisualBERT

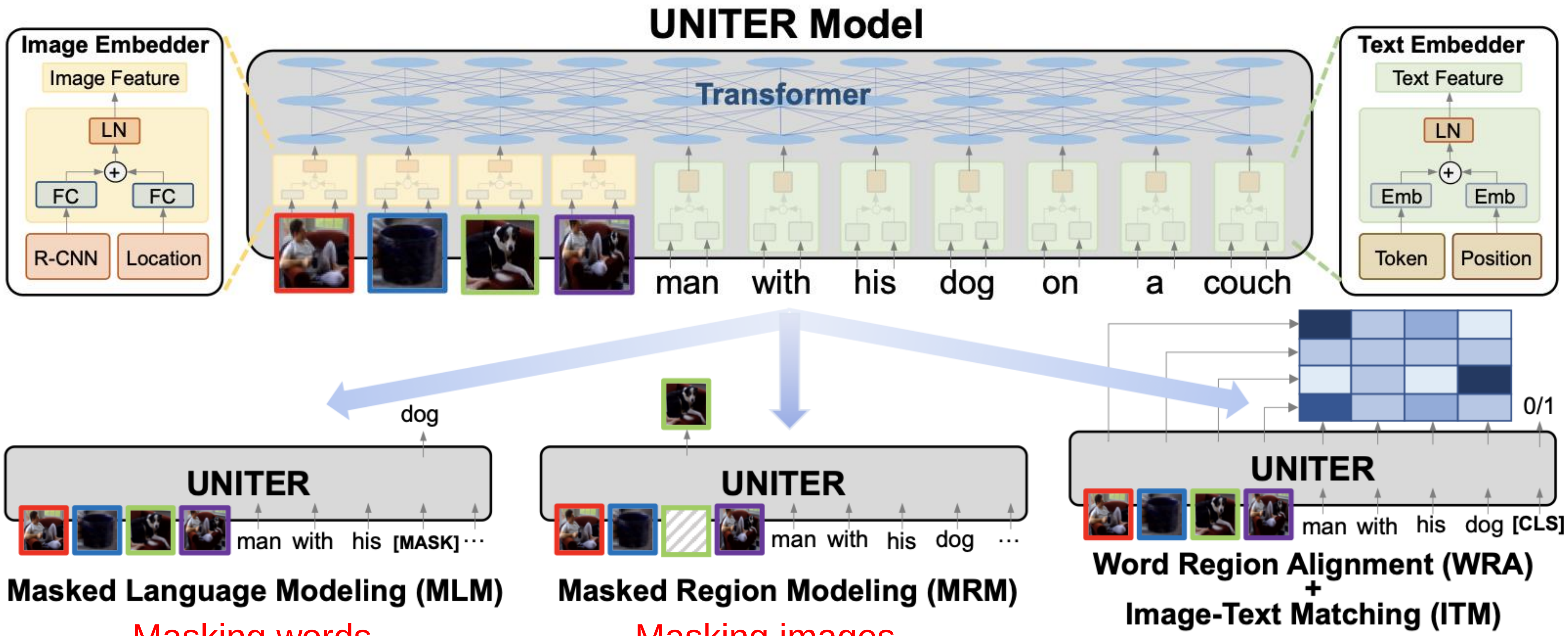


A person hits a ball with a tennis racket

Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." *arXiv* (2019).

UNITER

Similar Transformer architecture to BERT and VisualBERT… but with slightly different optimization



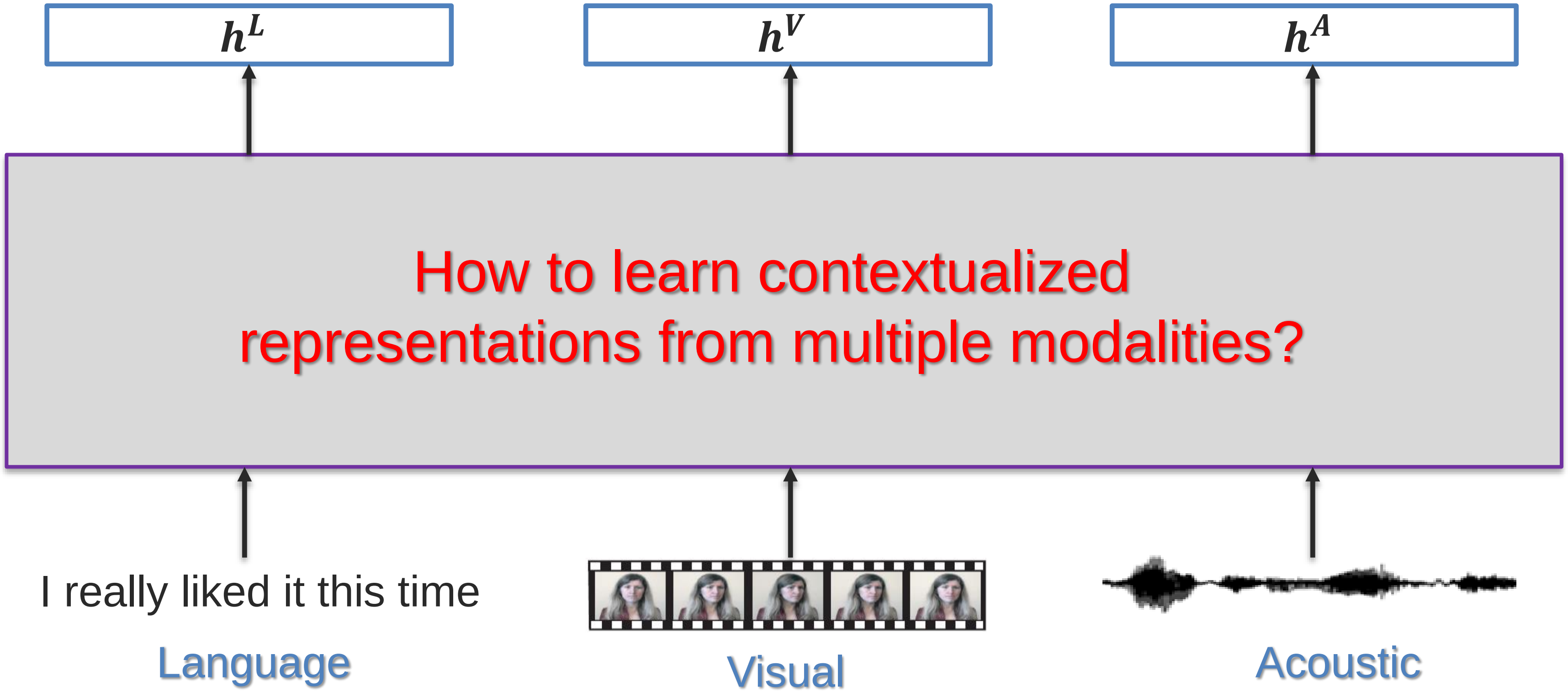
Masking words

Masking images

Is the sentence matches the image?

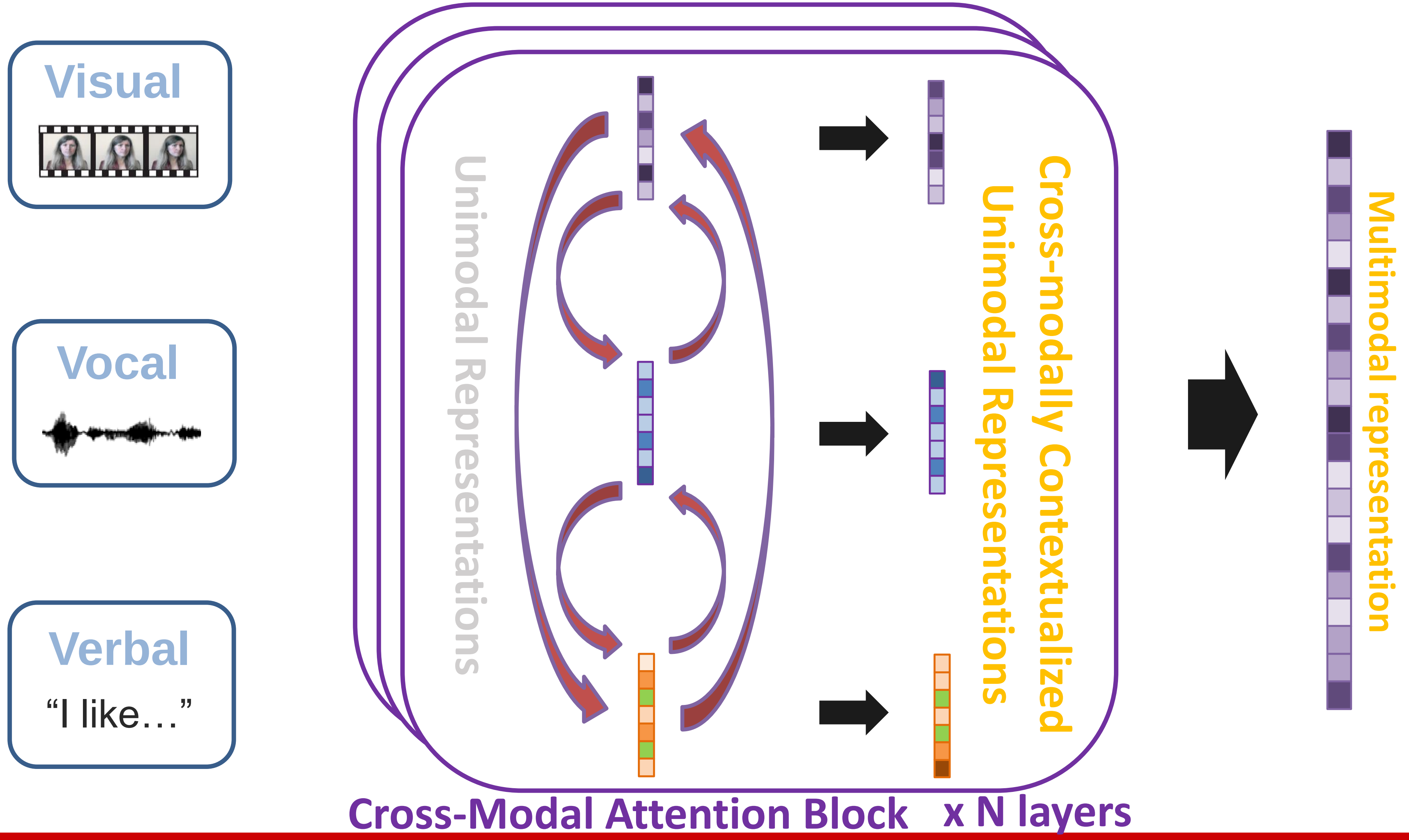
Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." *European conference on computer vision*. 2020.

Multimodal Embeddings

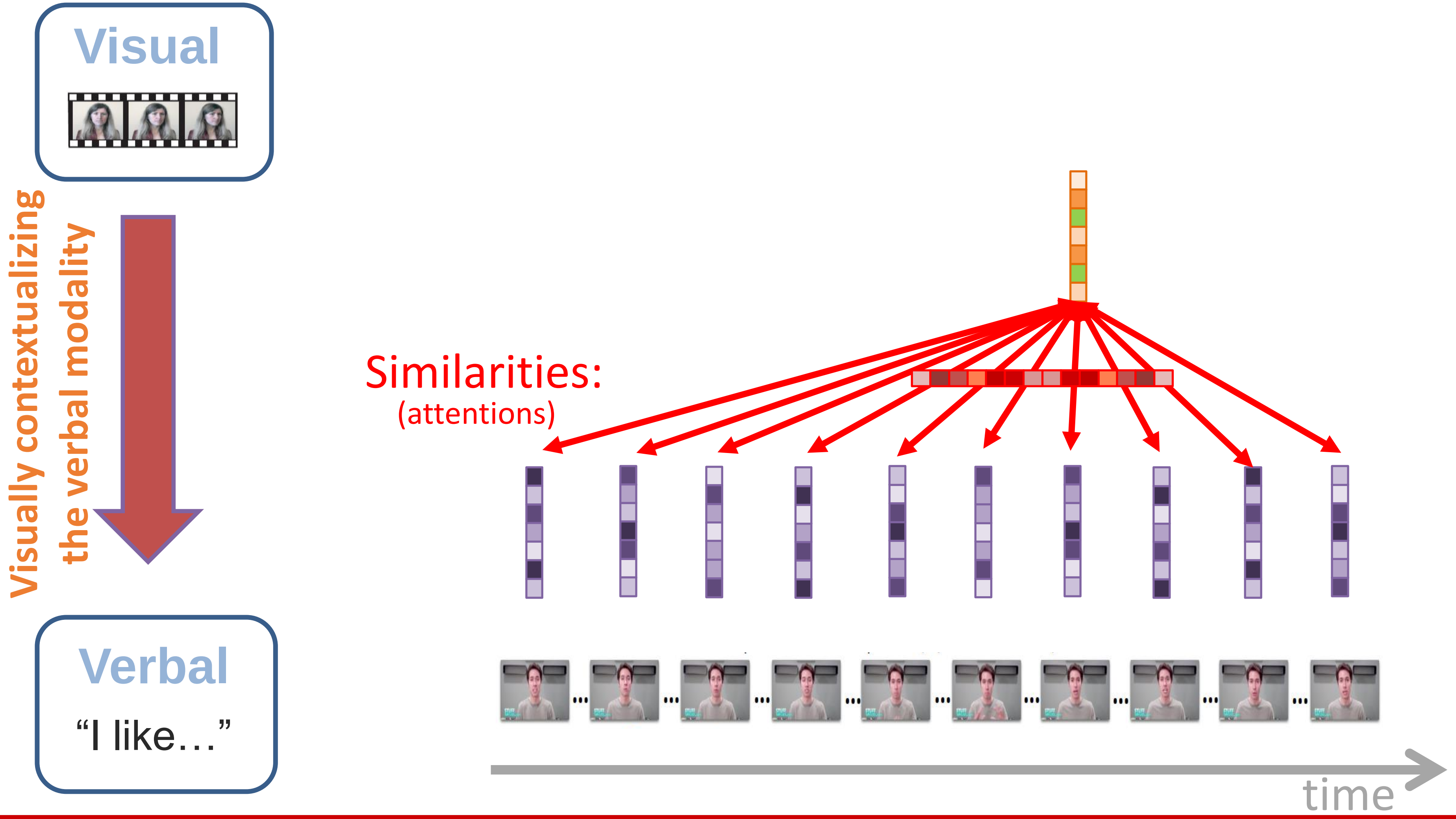


Option 2: Look at pairwise interactions between modalities

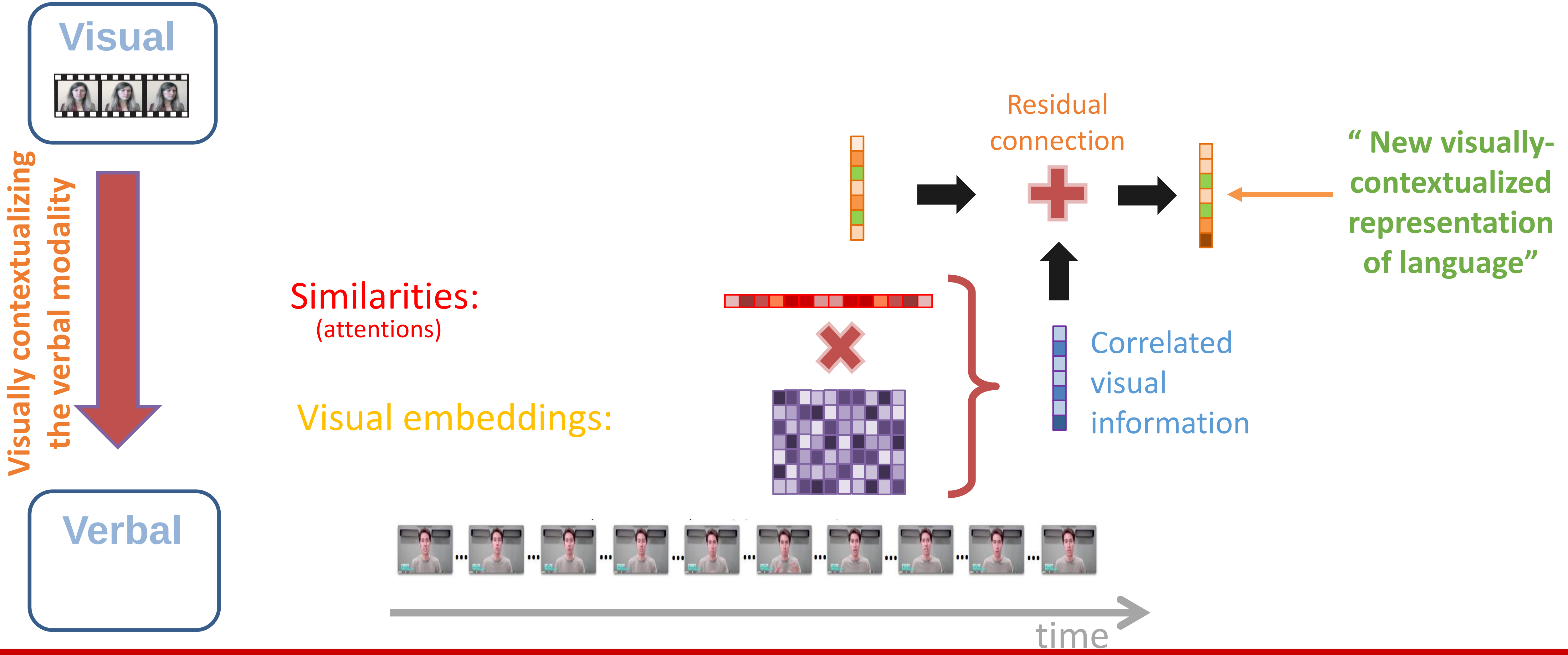
Multimodal Transformer – Pairwise Cross-Modal



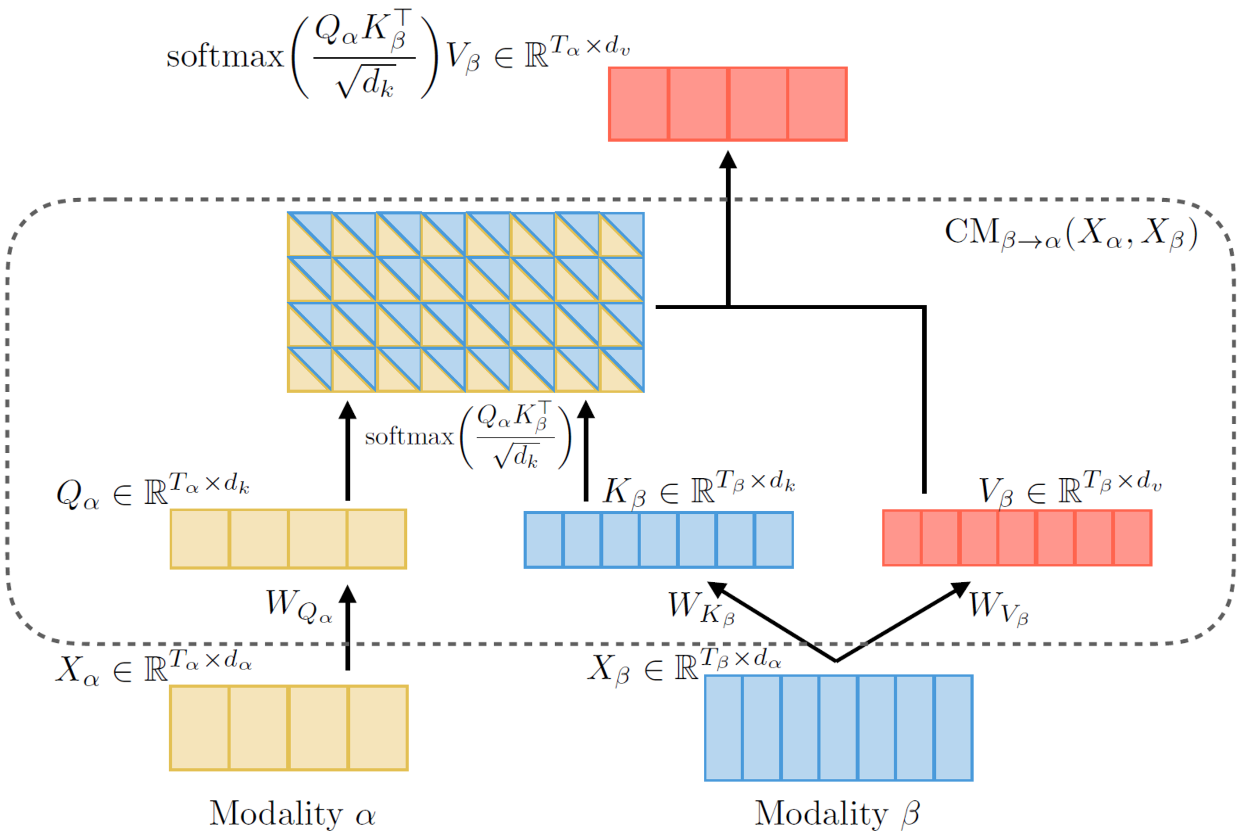
Cross-Modal Transformer Module ($V \rightarrow L$)



Cross-Modal Transformer Module ($V \rightarrow L$)

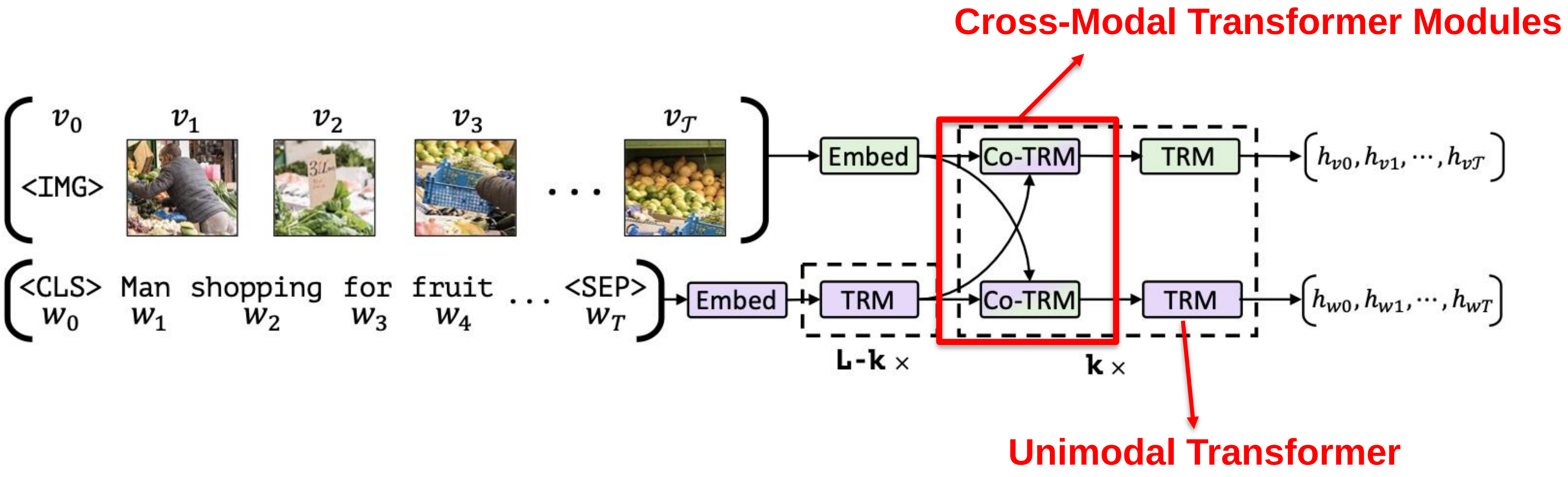


Cross-Modal Transformer Module ($\beta \rightarrow \alpha$)



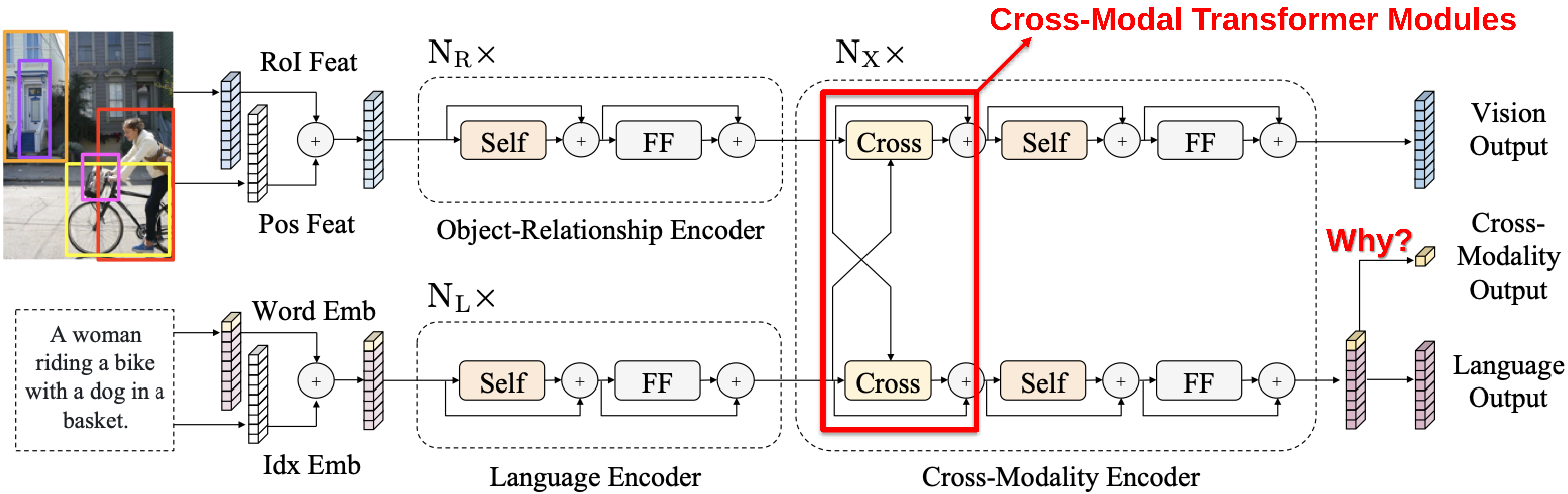
Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

ViLBERT



Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *arXiv* (August 6, 2019).

LXMERT

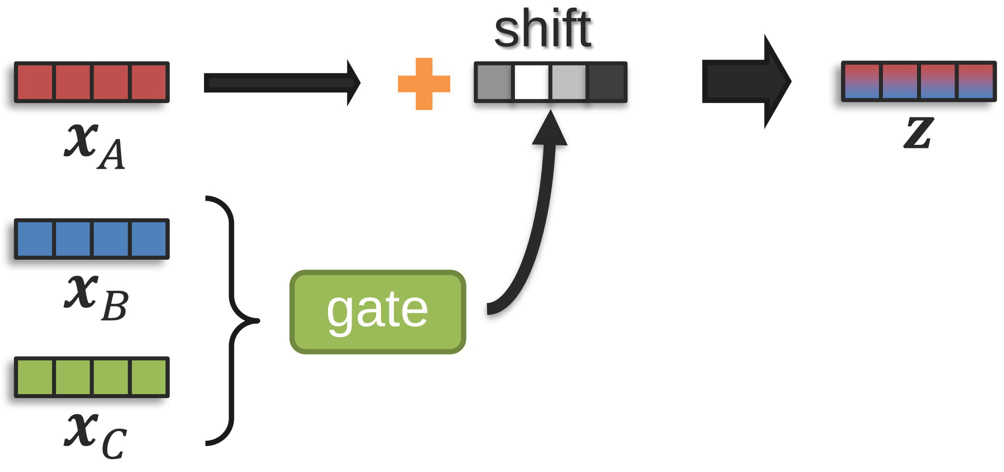


Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv* (August 20, 2019).

Reminder: Modality-Shifting Fusion

Primary modality

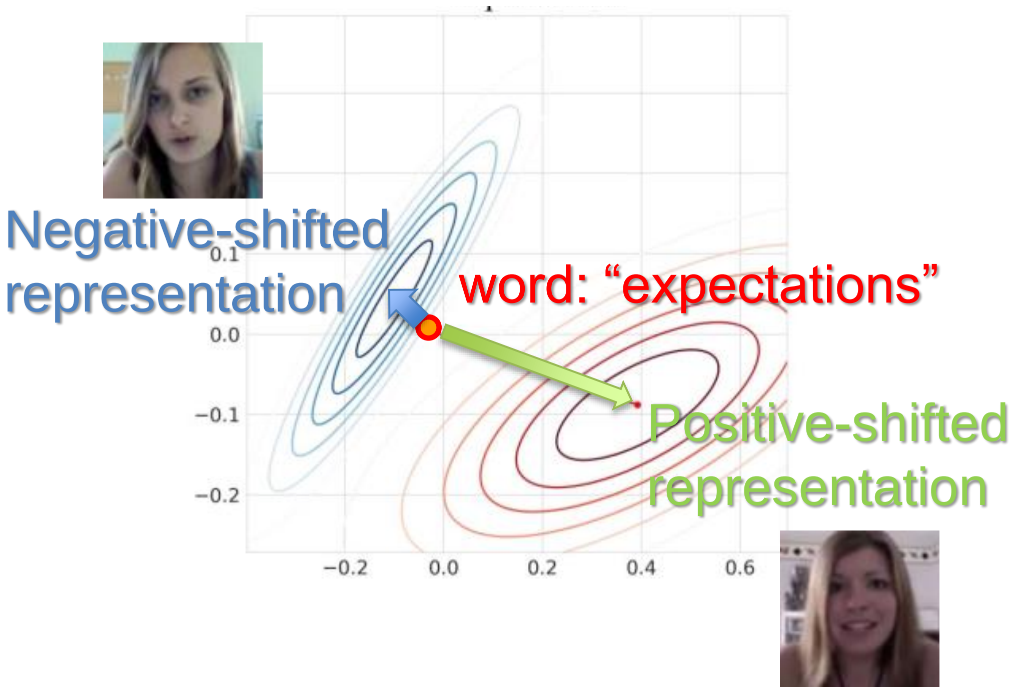
Secondary modalities



Example with language modality:

Primary modality: language

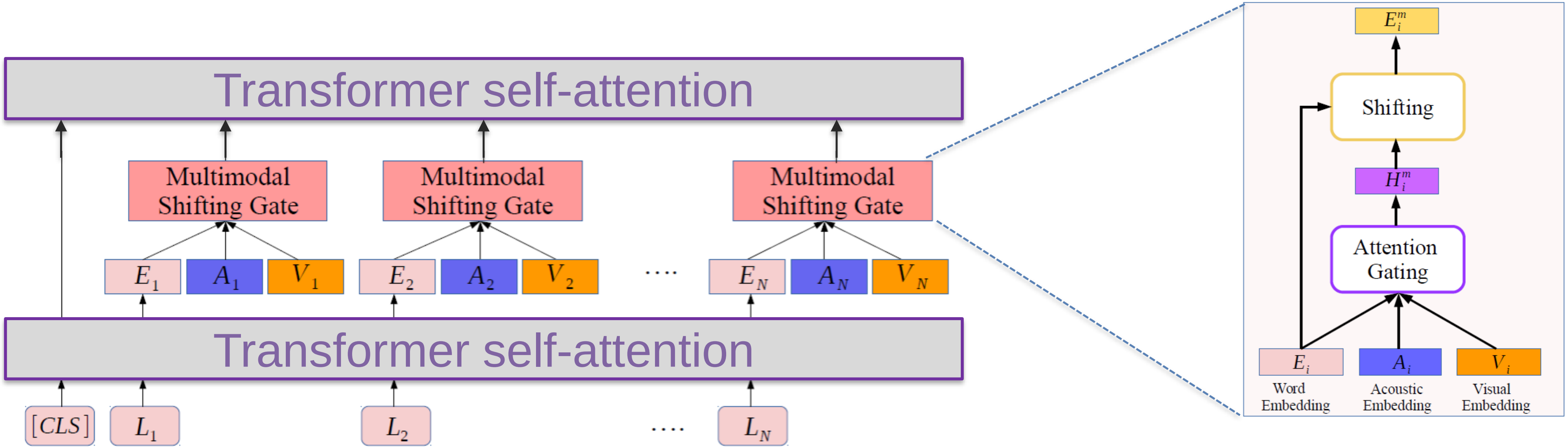
Secondary modalities: acoustic and visual



Wang et al., Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors, AAAI 2019

Modality-Shifting with Transformers

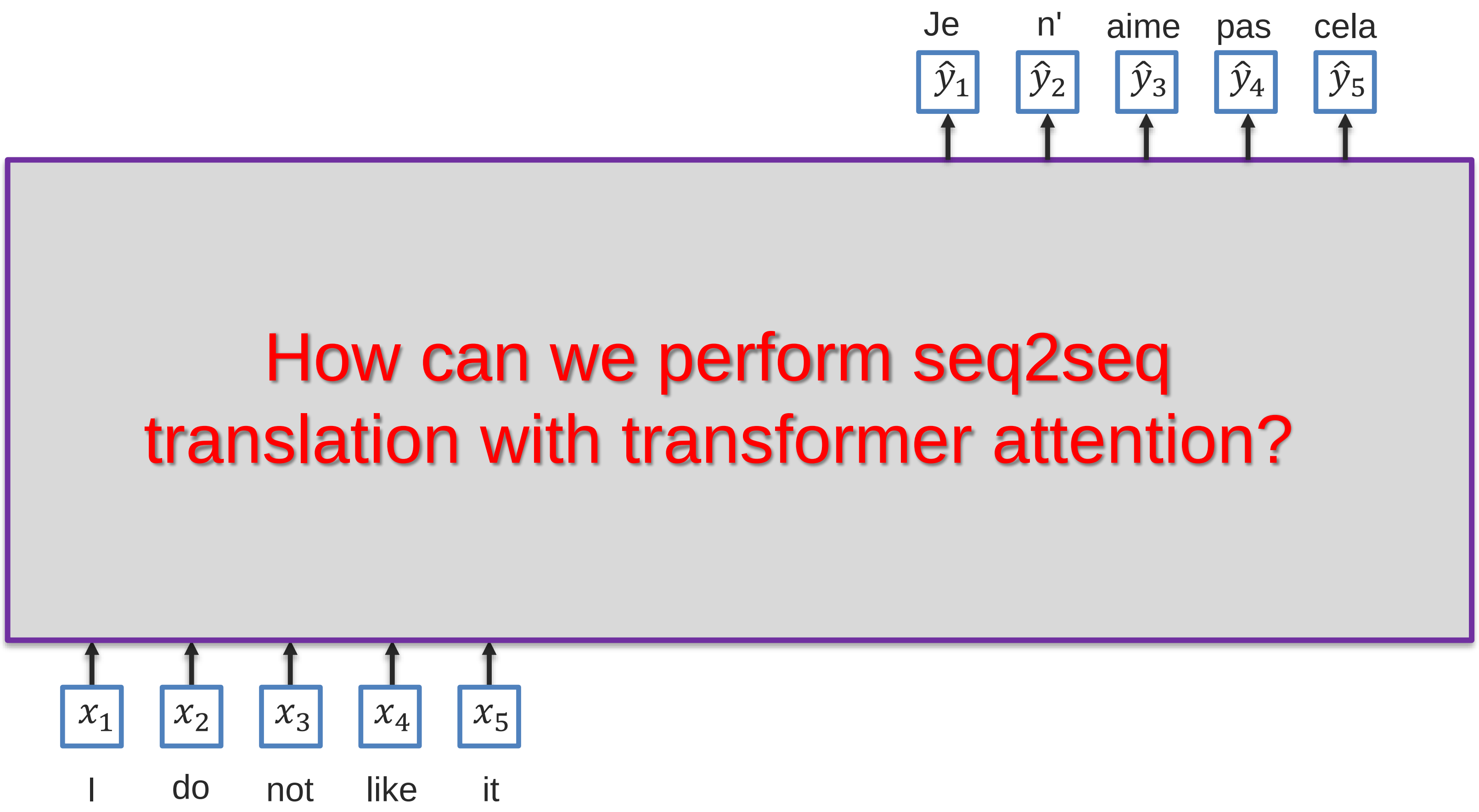
Multimodal Adaptation Gate (MAG) + BERT



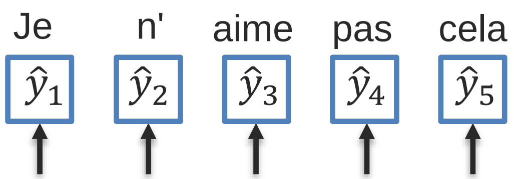
Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers, ACL 2020

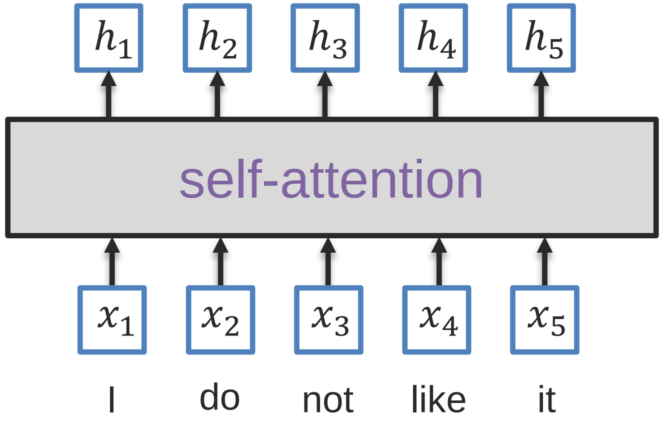
Sequence-to-Sequence Using Transformer

Sequence-to-Sequence Modeling

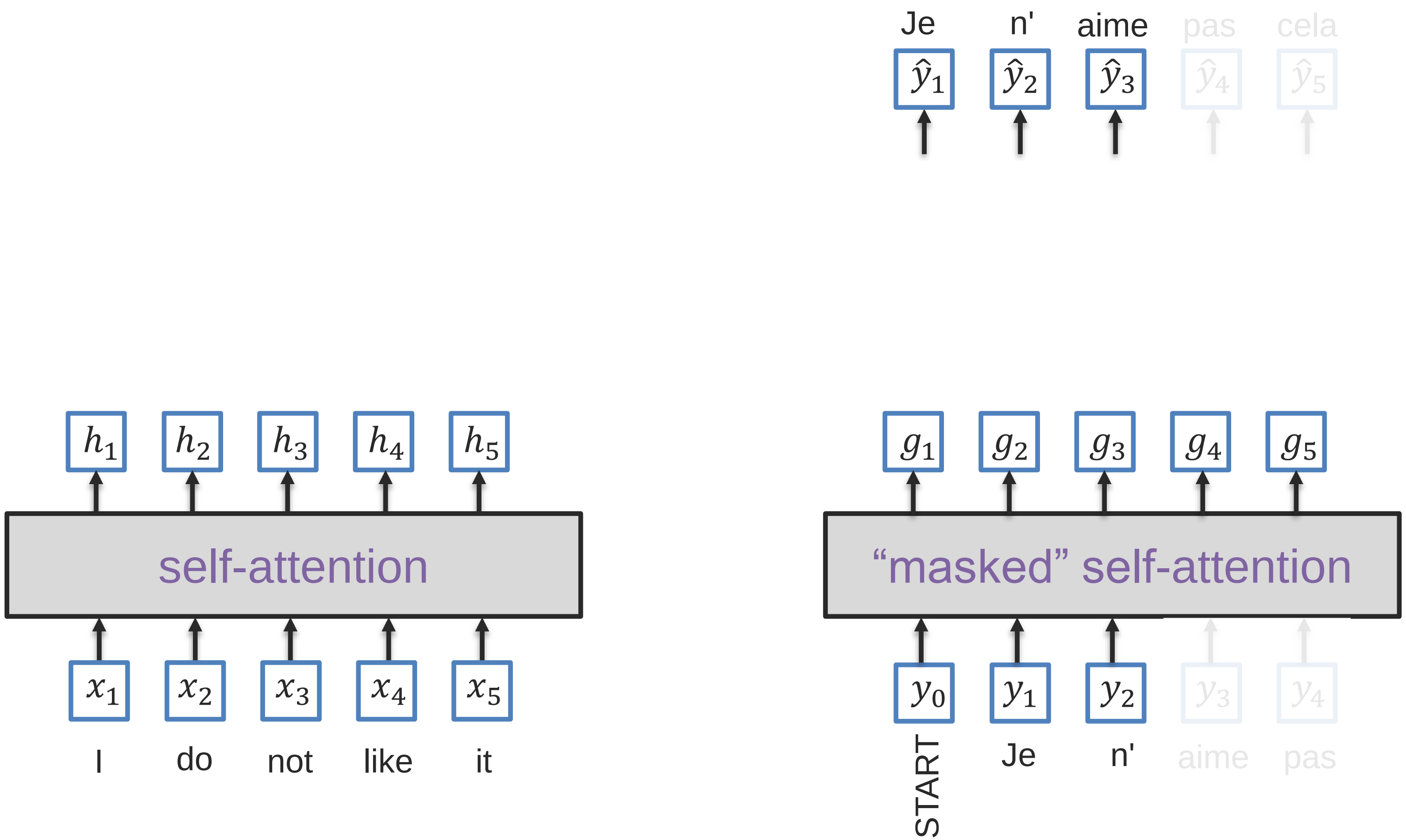


Seq2Seq with Transformer Attentions



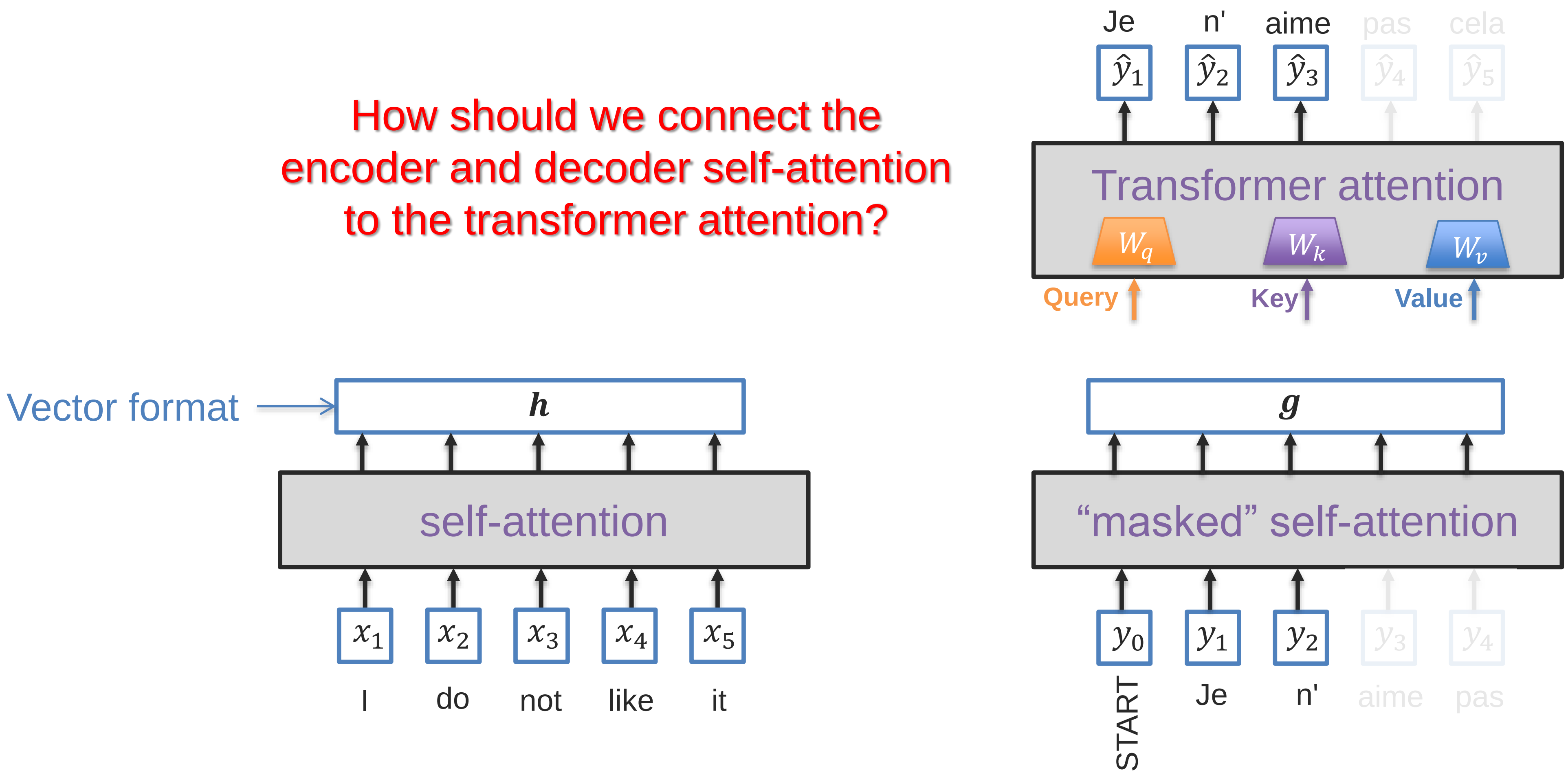


Seq2Seq with Transformer Attentions



Seq2Seq with Transformer Attentions

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



Seq2Seq with Transformer Attentions