Language Technologies Institute

Carnegie Mellon University

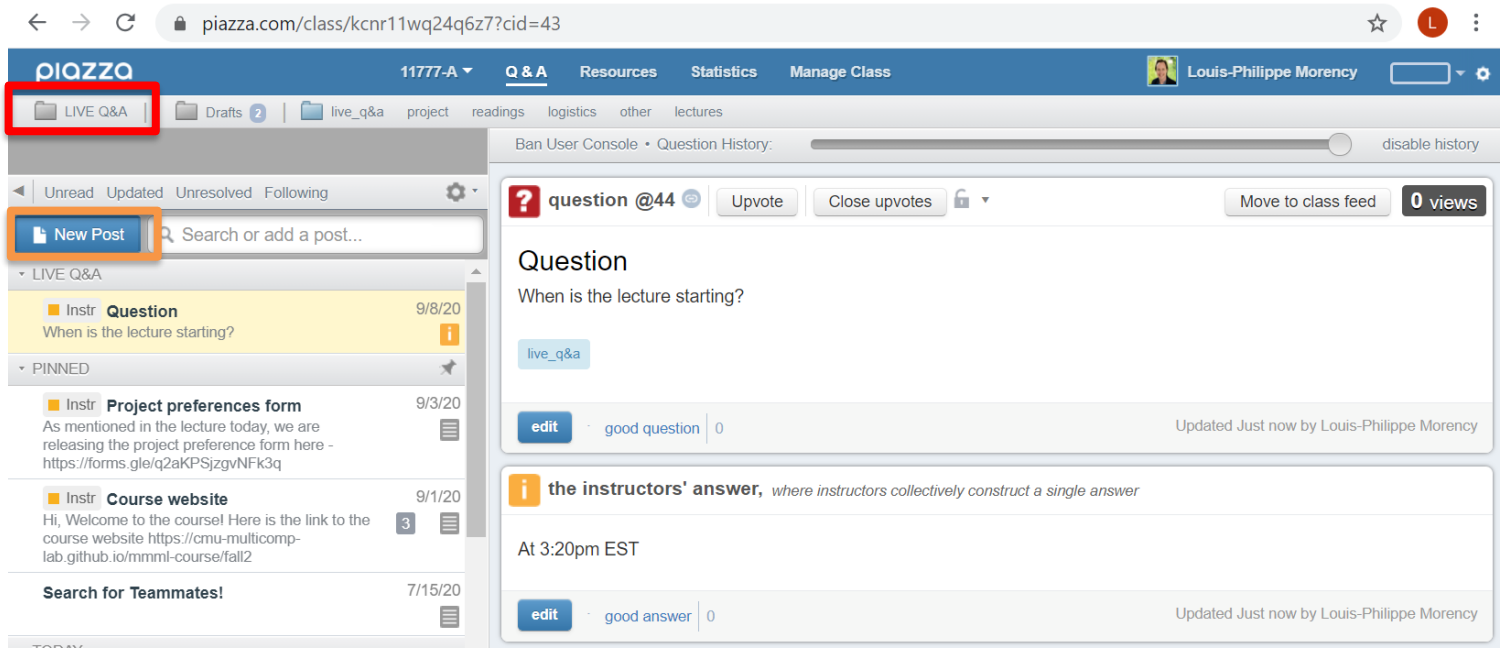
Multimodal Machine Learning

Lecture 5.2: Alignment and Representations

Louis-Philippe Morency

Administrative Stuff

Piazza Live Q&A



Please share your questions and comments on Piazza Live Q&A

- Live responses by your TAs and follow-up by the instructor after the main lecture

Next Week Schedule

Tuesday, Thursday 3:20pm: Live office hours with LP

- Use the same Zoom link (waiting room will be activated)

Friday 8pm: deadline for presentations

- Submit on Gradescope (slides) and Box (video)

Sunday 8pm: deadline for reports

- Submit on Gradescope

Friday (10/9) 8pm: Deadline for student feedback

No reading assignment for week 6

- But don’t forget to complete week 5 assignment by Monday

Reading assignment for Week 7 (starting Monday 10/5)

Student Peer Feedback

Period: Monday Oct 12th until Friday Oct 16th 8pm (week 7)

Feedback process:

- Piazza: Full list of project videos and link to feedback form
- Each student randomly matched with 6 projects
  - Feedback needs to be substantial (see form instructions)
- Feedback is anonymous, but instructors will see names
- Each team should receive feedback from ~24 students
- [optional] Share feedback for other (not matched) projects
- Your primary TA will share the feedback with your team.
  - Meeting with your primary TA during week 8 (10/19-10/23)

Share Your Thoughts! https://forms.gle/ar7BZgVKB6XoyPGq5

Course Feedback - 11777 Fall 2020

Please take a moment to share with us your feedback regarding the course Multimodal Machine Learning (11777 Fall 2020). We love to hear about how your feel related to the course structure and content, so that we can adjust the course if necessary. Thank you for your time!

* Required

How do you like the course so far? *

	Poor	Fair	Satisfactory	Very good	Excellent
Answer	◯	◯	◯	◯	◯

Course content *

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
Learning objectives were clear	◯	◯	◯	◯	◯

Deadline

Please submit your feedback about this course before this Sunday 10/4

Optional,
but greatly appreciated! ☺

Anonymous, by default.

- You can optionally share your email address if you want us to follow-up with you directly.

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 5.2: Alignment and Representations

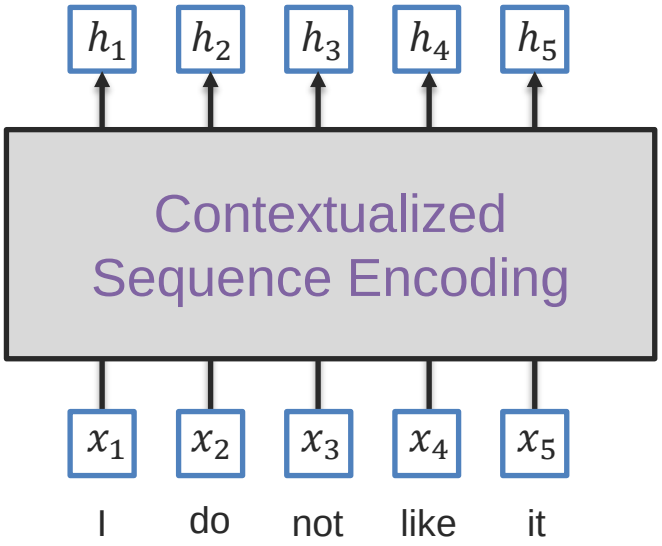
Louis-Philippe Morency

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

Contextualized Sequence Encoding

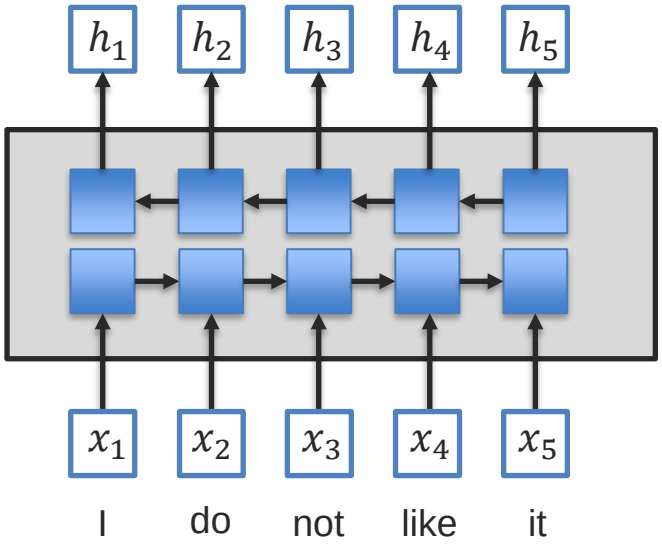
Sequence Encoding - Contextualization



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

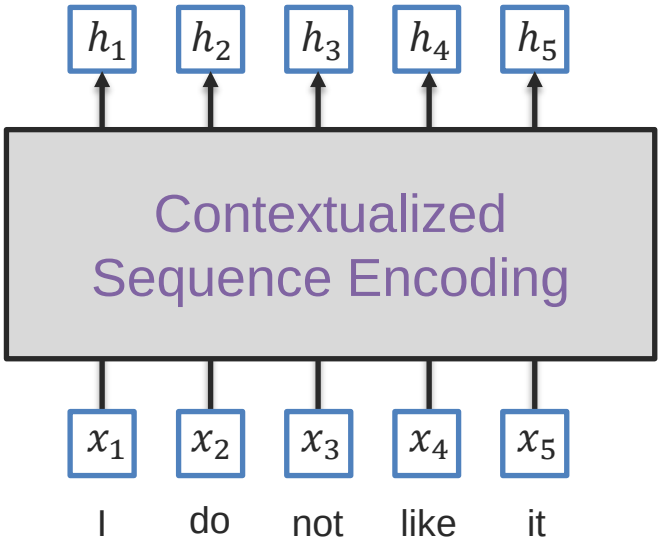
Option 1: Bi-directional LSTM:

(e.g., ELMO)

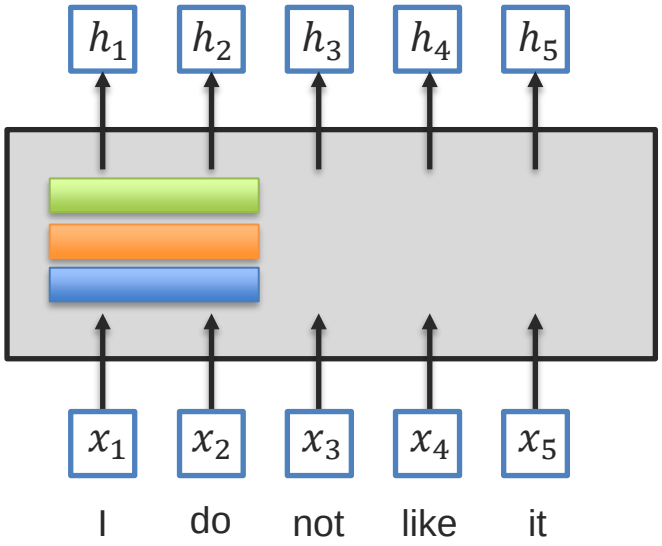


But harder to parallelize…

Sequence Encoding - Contextualization



Option 2: Convolutions

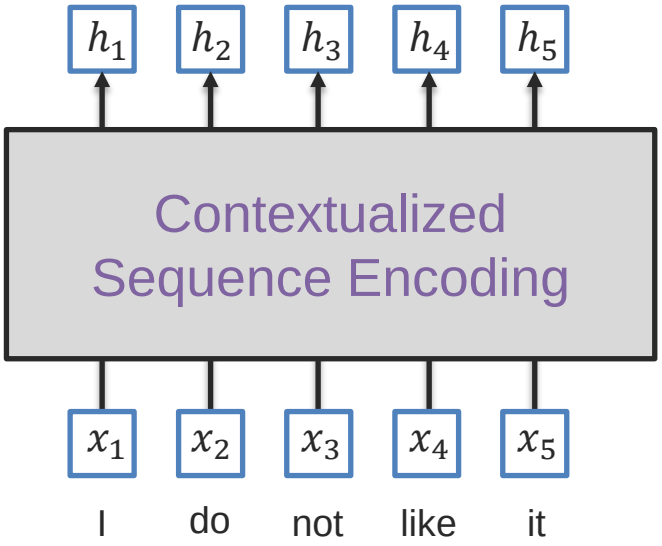


Can be parallelized!

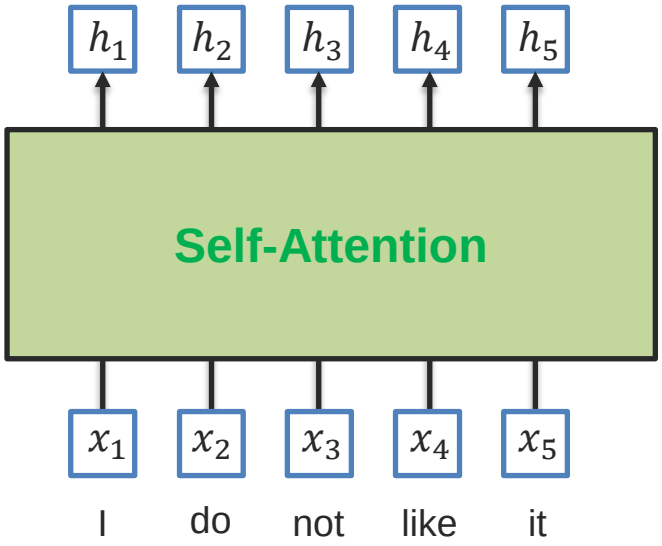
But modeling long-range dependencies require multiple layers

And convolutional kernels are static

Sequence Encoding - Contextualization



Option 3: Self-attention



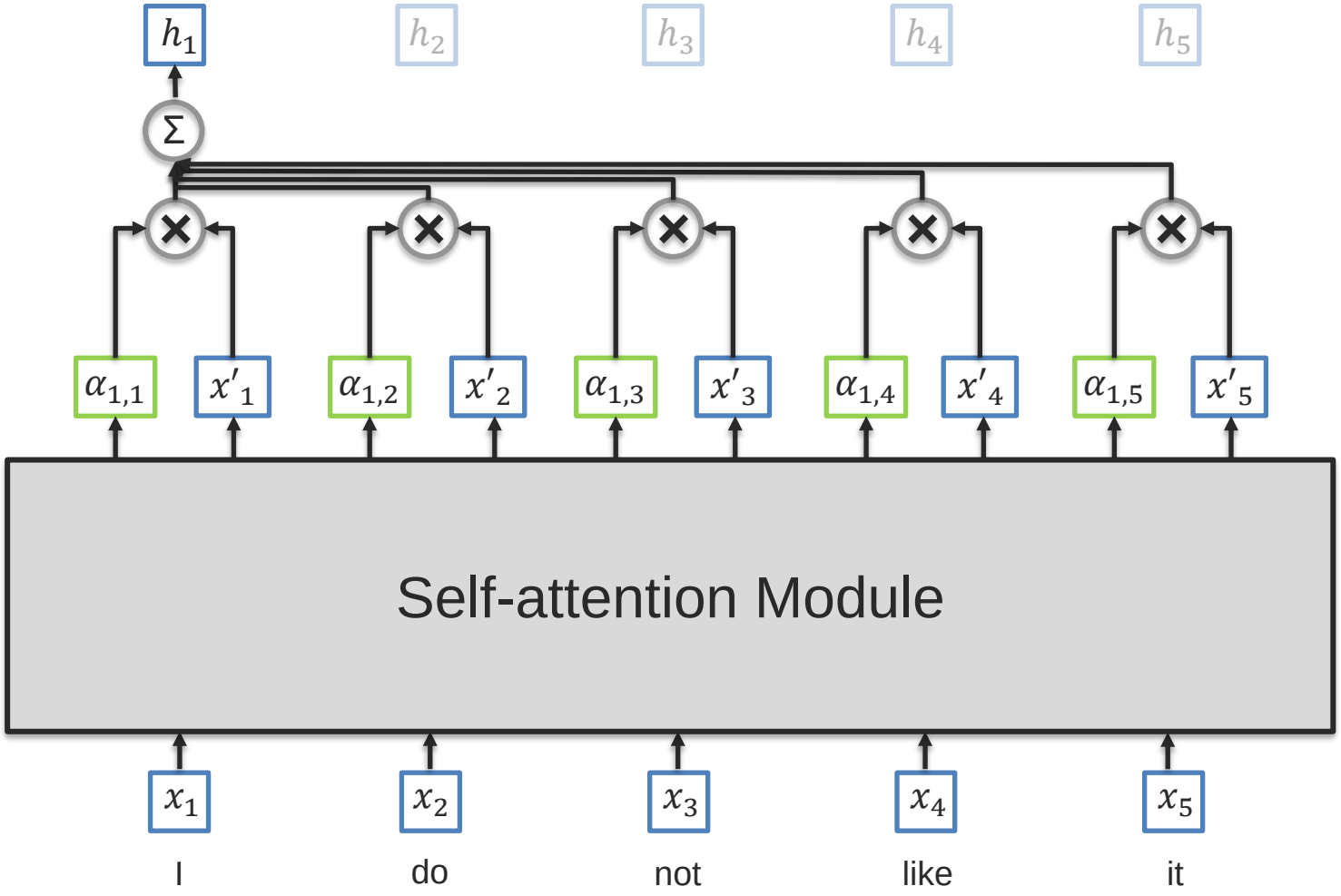
Can be parallelized!

Long-range dependencies

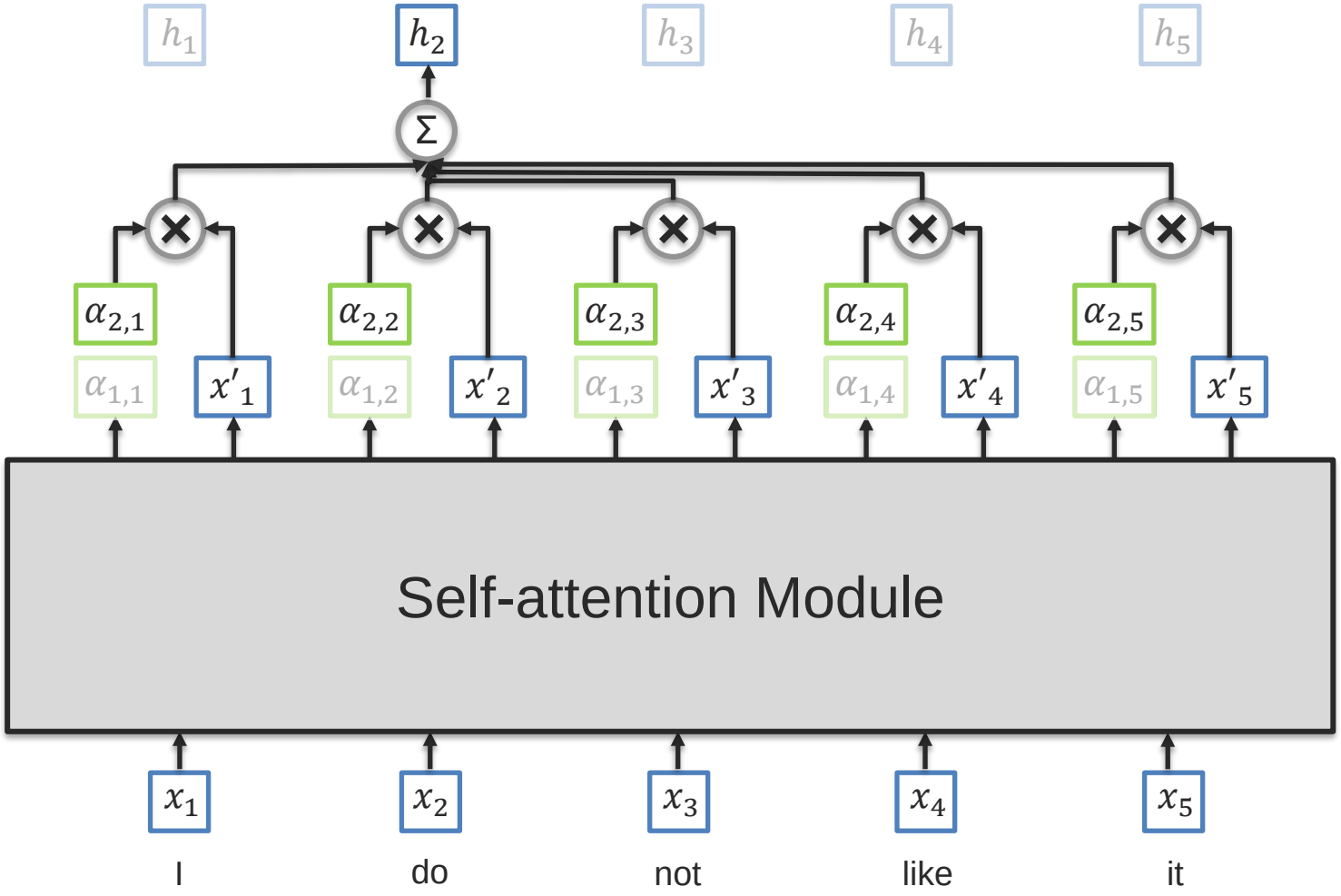
Dynamic attention weights

Self-Attention

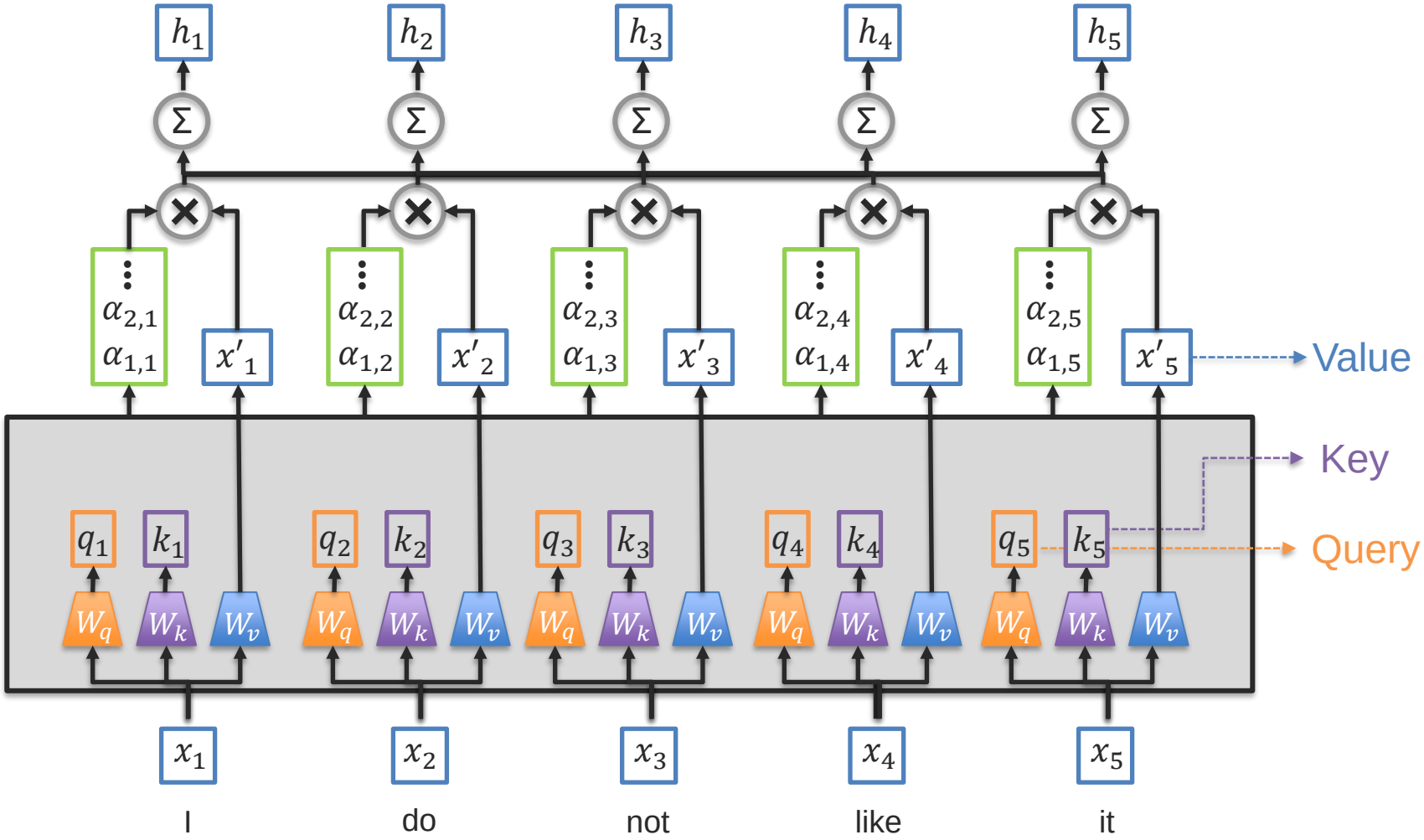
Self-Attention



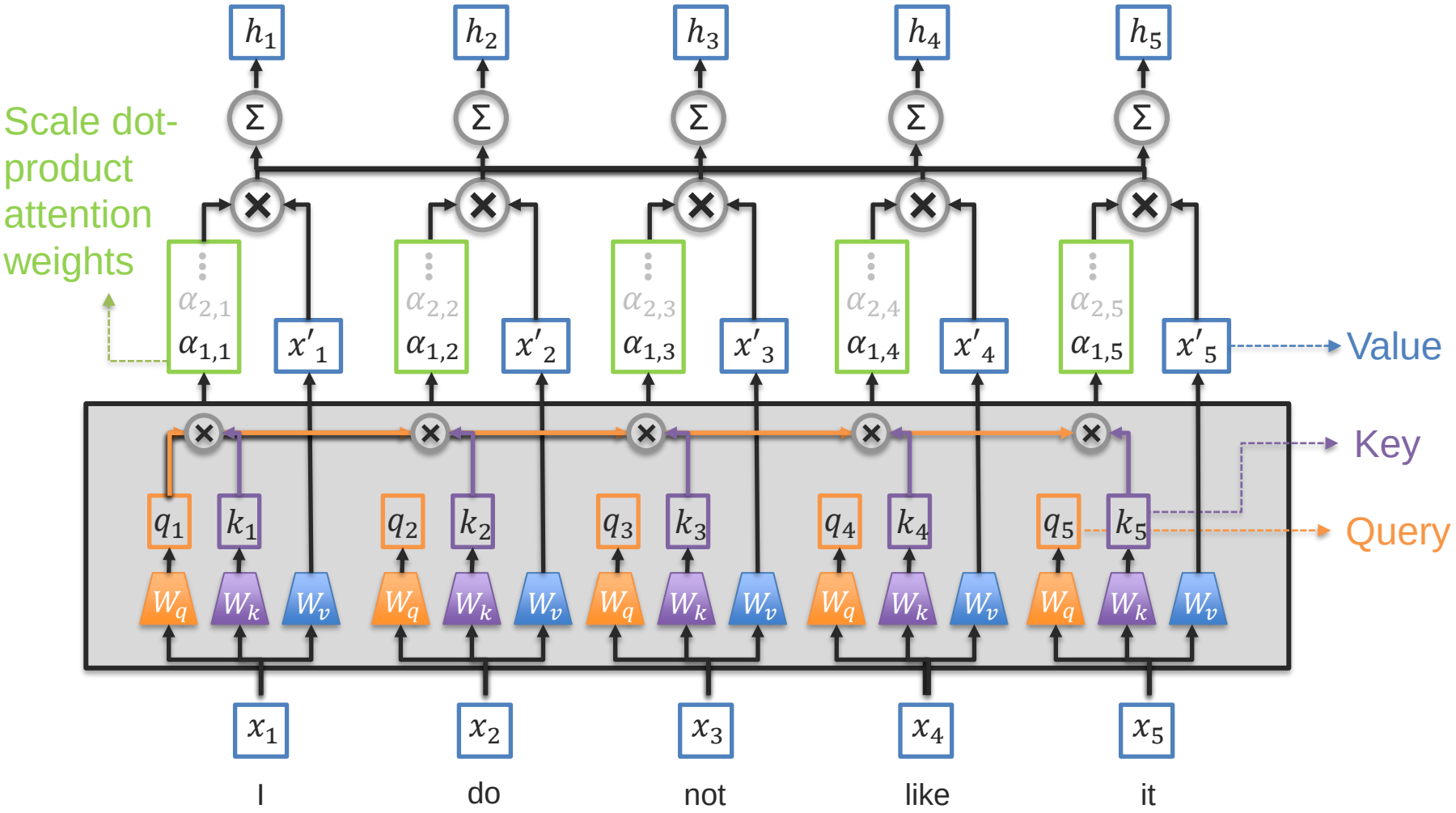
Self-Attention



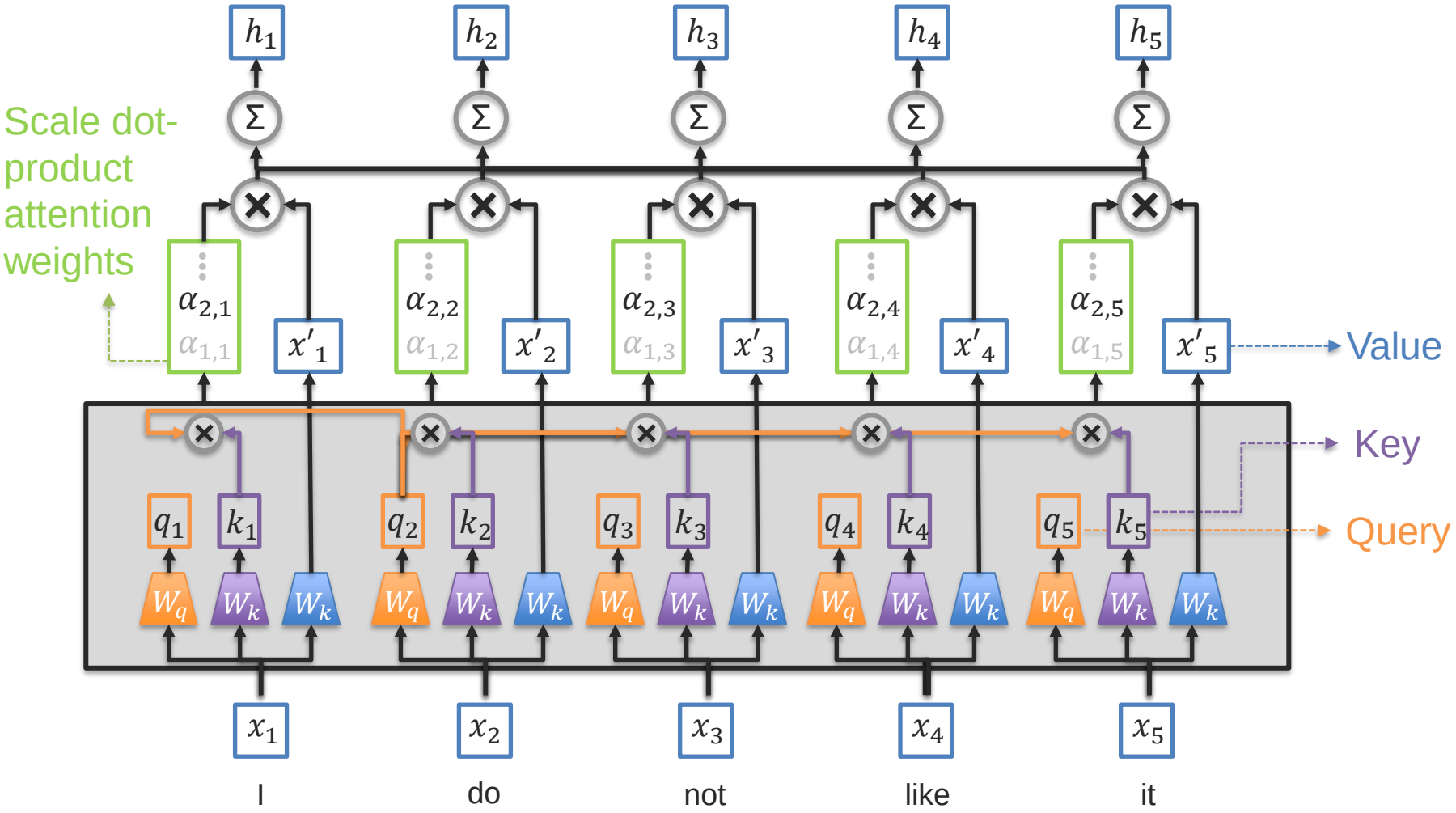
Transformer Self-Attention



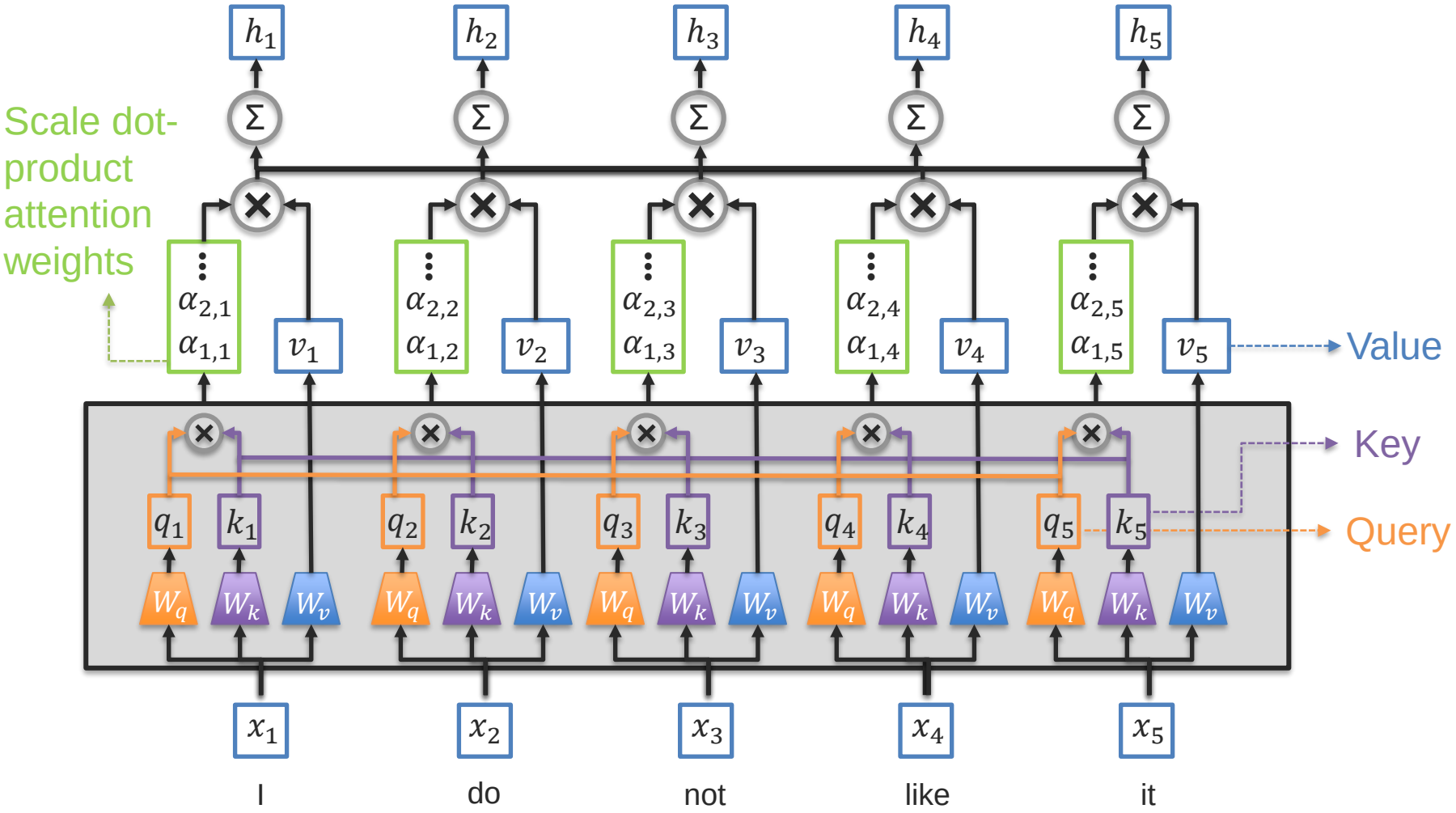
Transformer Self-Attention



Transformer Self-Attention

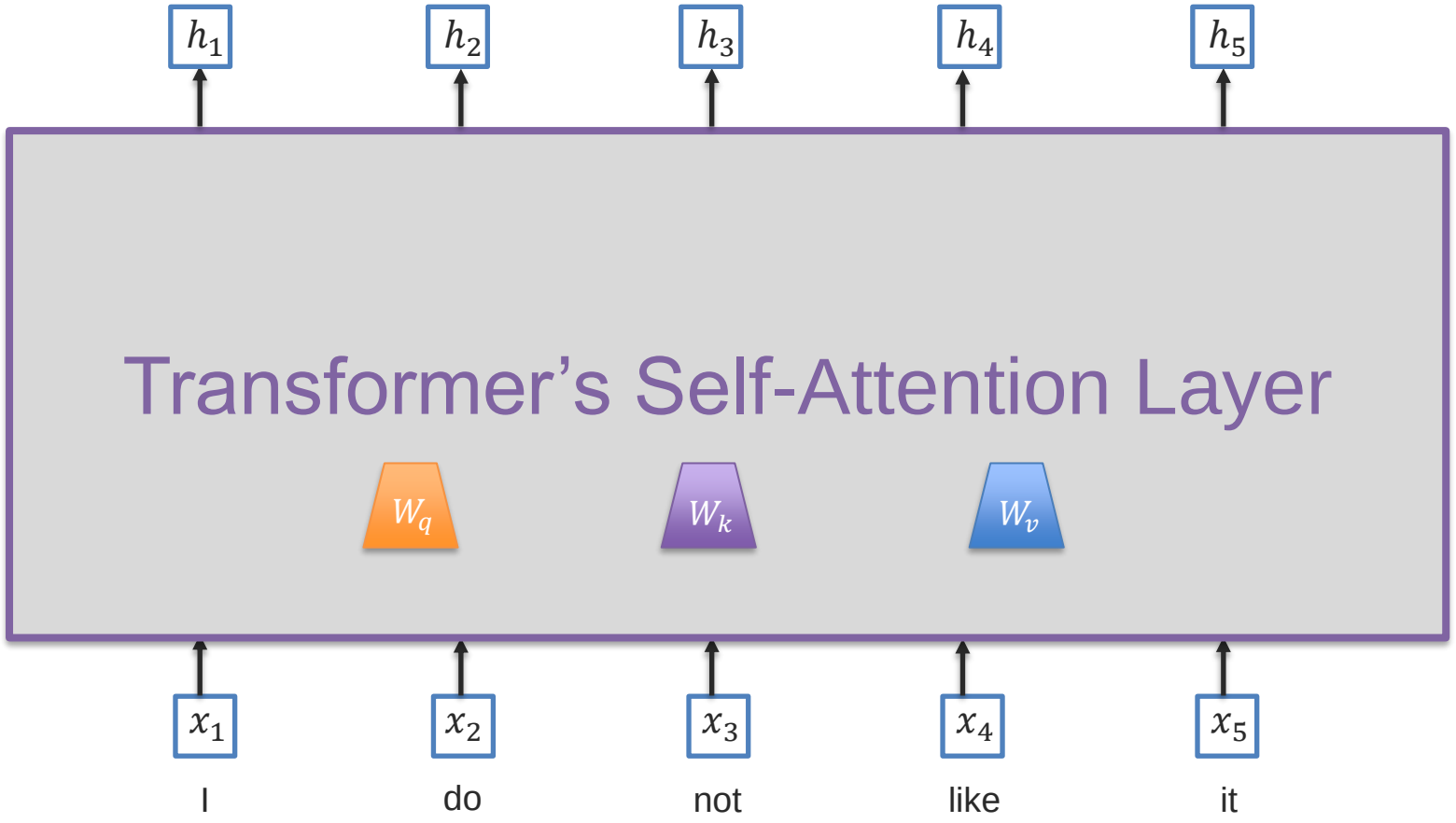


Transformer Self-Attention

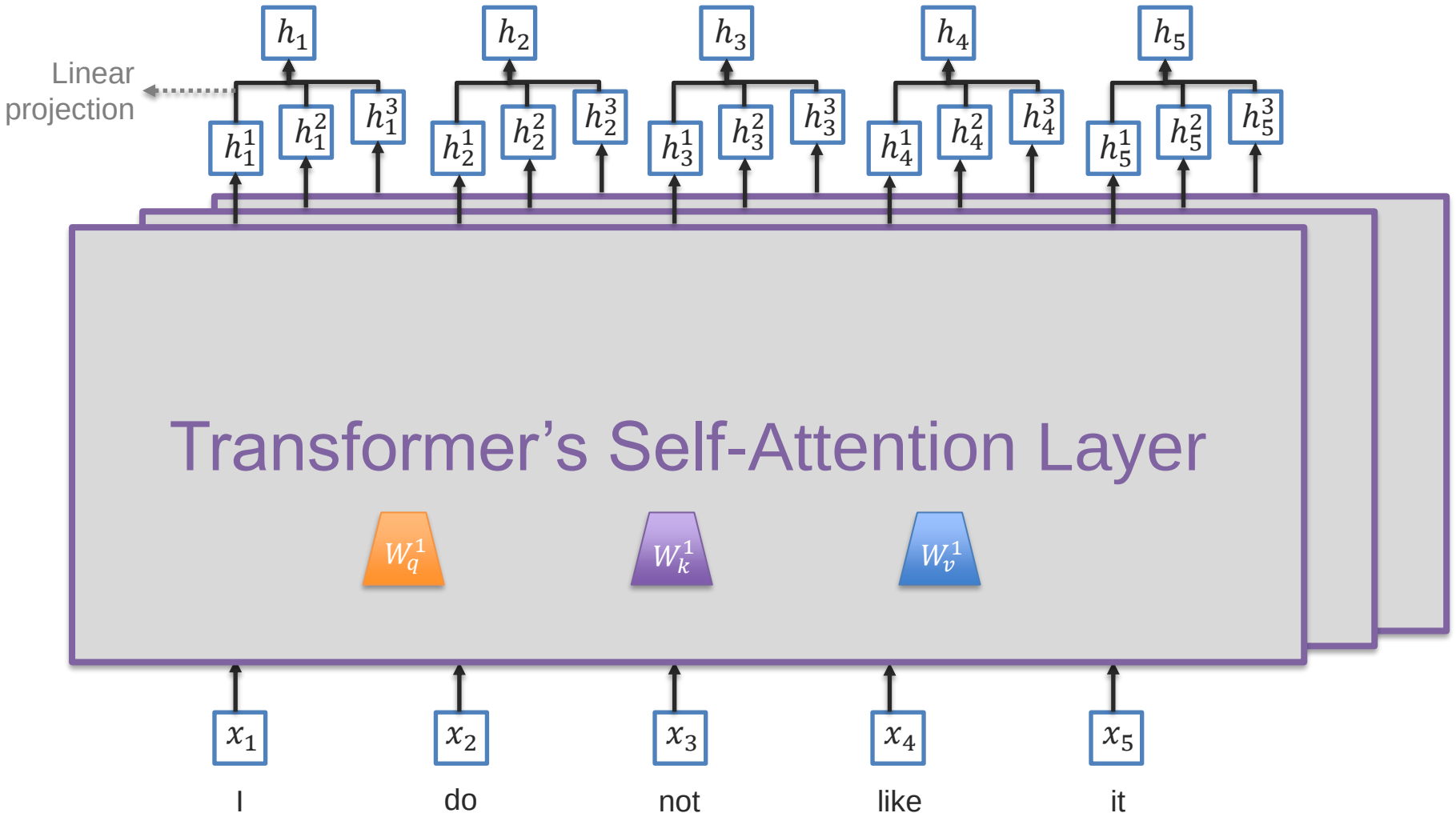


Transformer Self-Attention

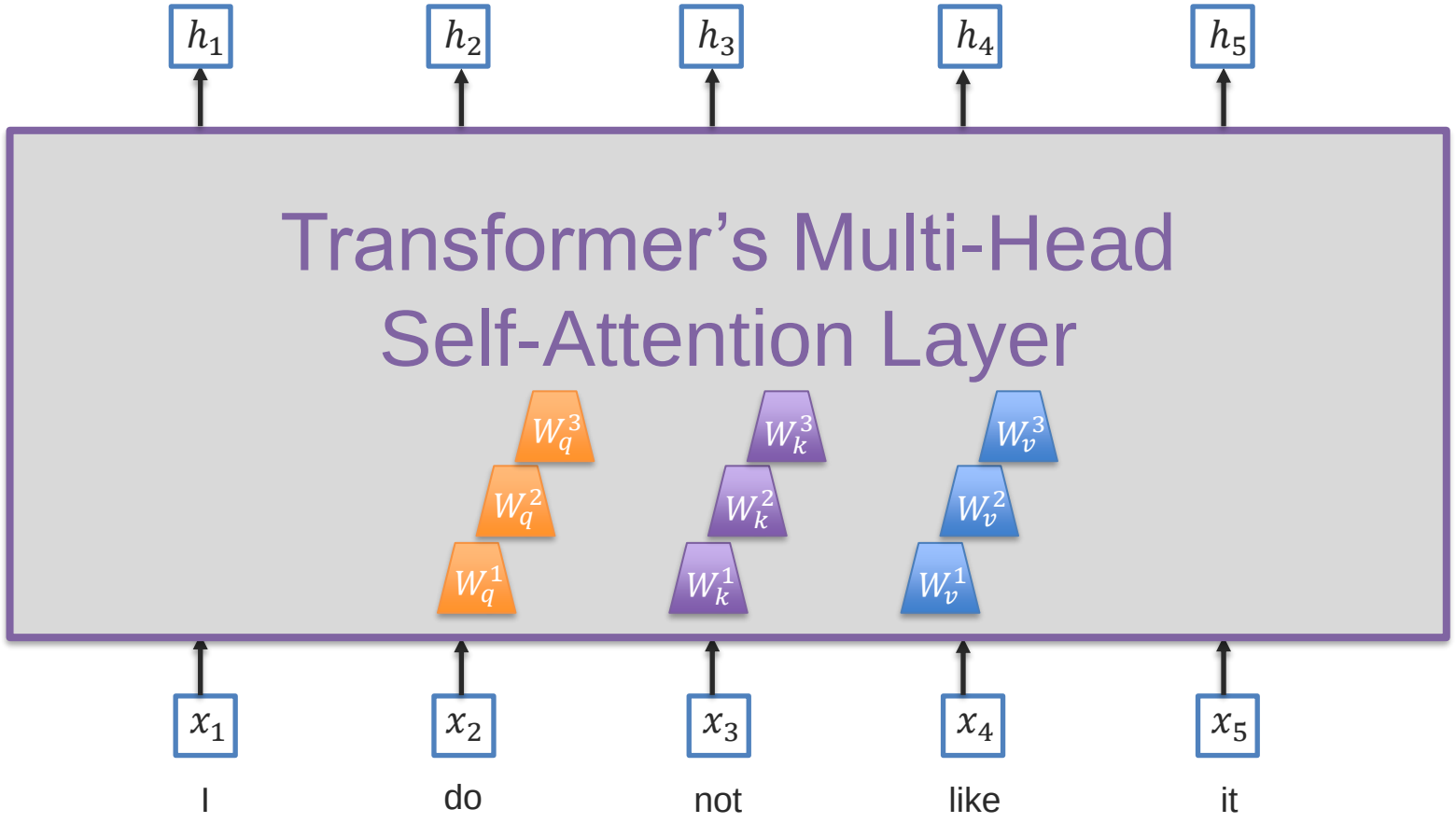
What if we want to attend simultaneously to multiple subspaces of x?



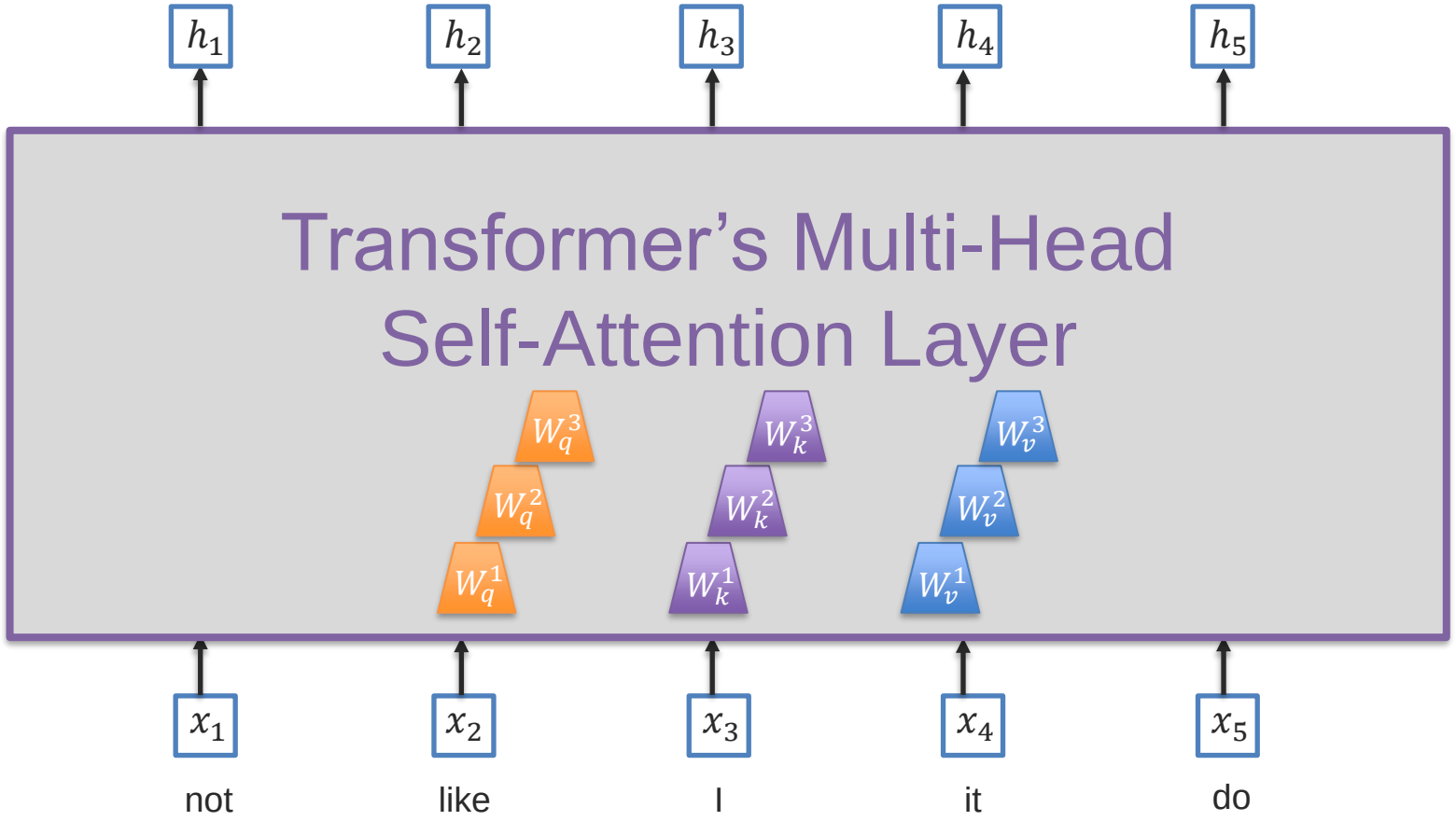
Transformer Multi-Head Self-Attention



Transformer Multi-Head Self-Attention



Transformer Multi-Head Self-Attention



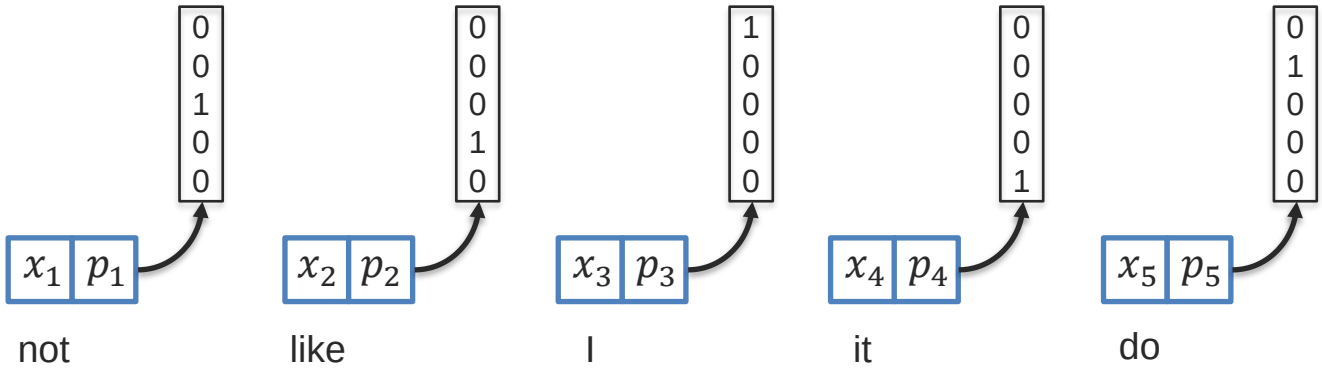
What happens if the words are shuffled?

Position embeddings

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

How can we encode position information?

Simple approach: one-hot encoding

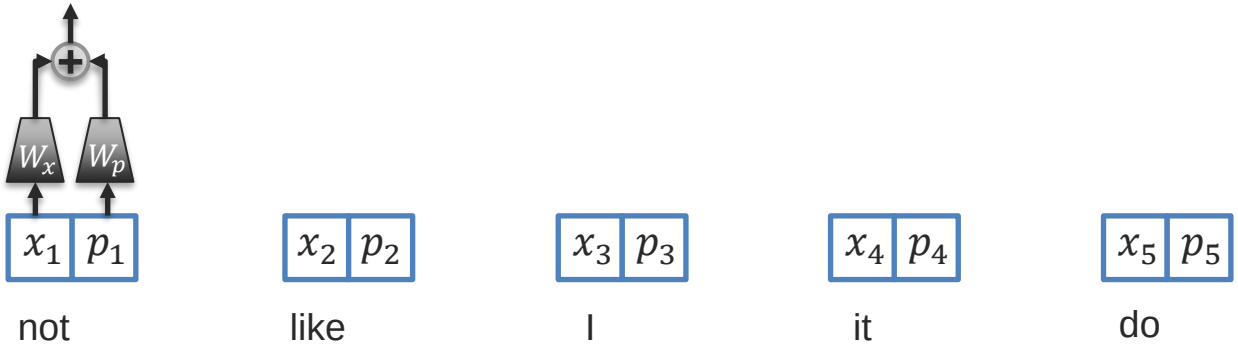


Position embeddings

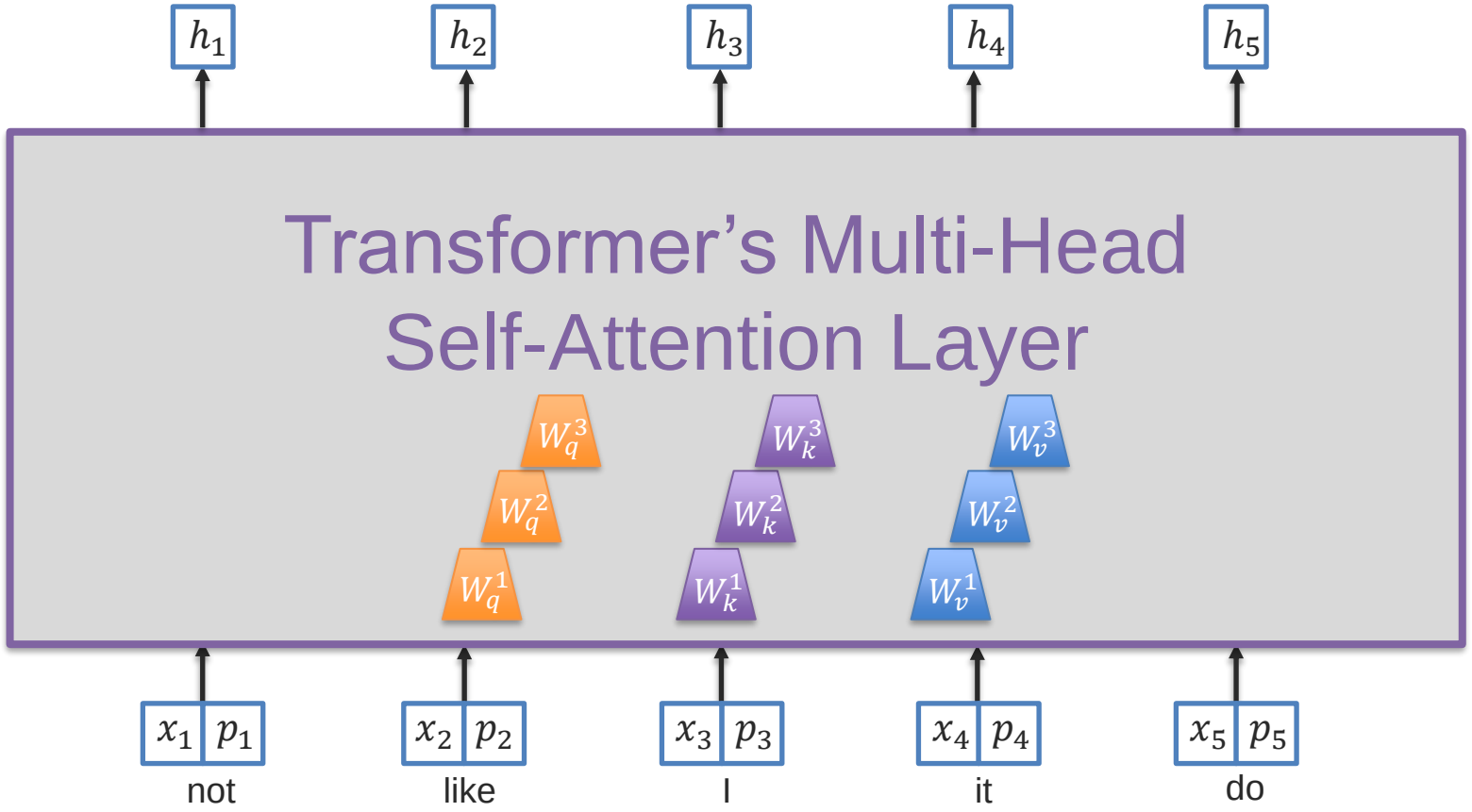
- Position information is not encoded in a self-attention module

How can we encode position information?

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

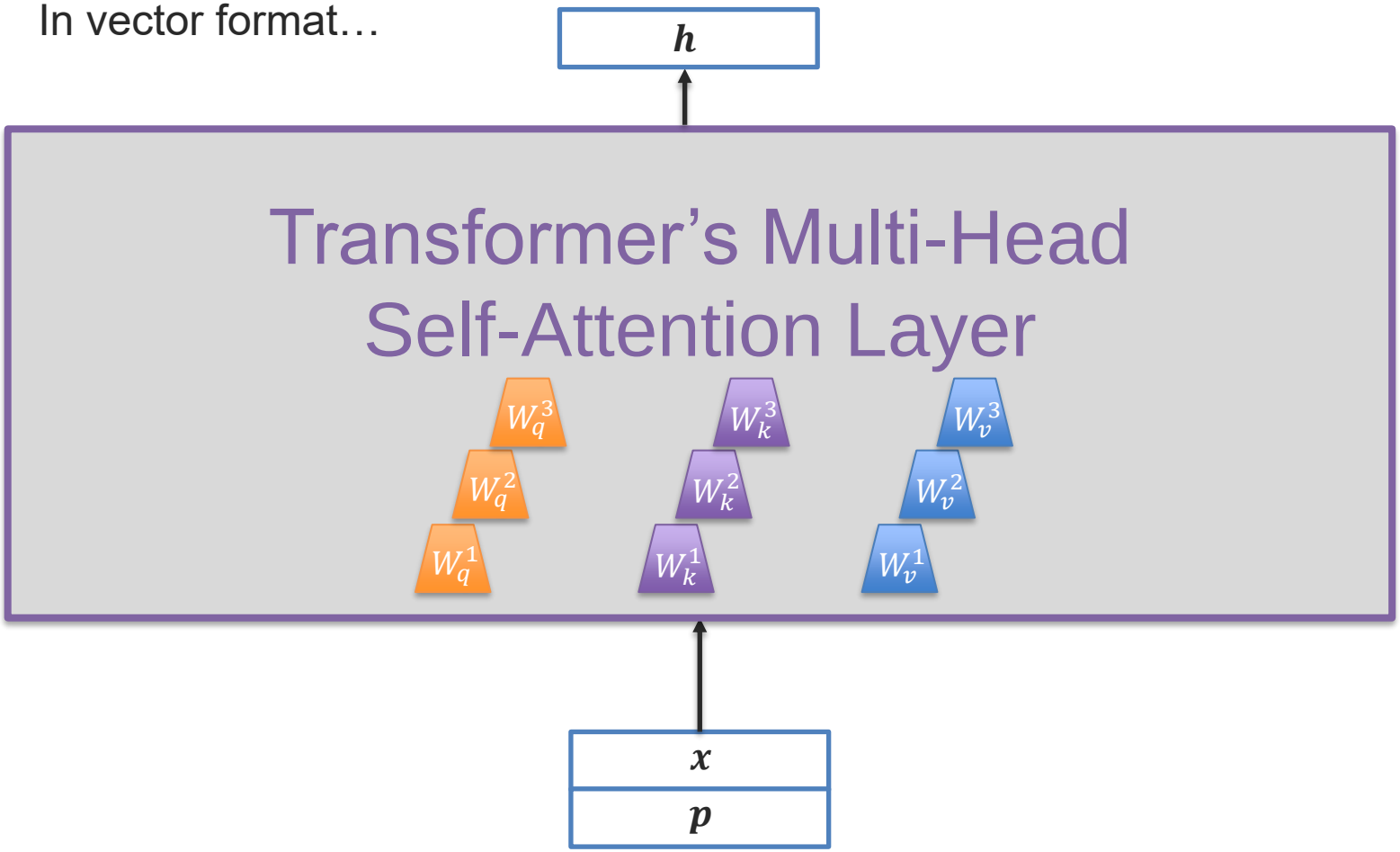


Transformer Multi-Head Self-Attention

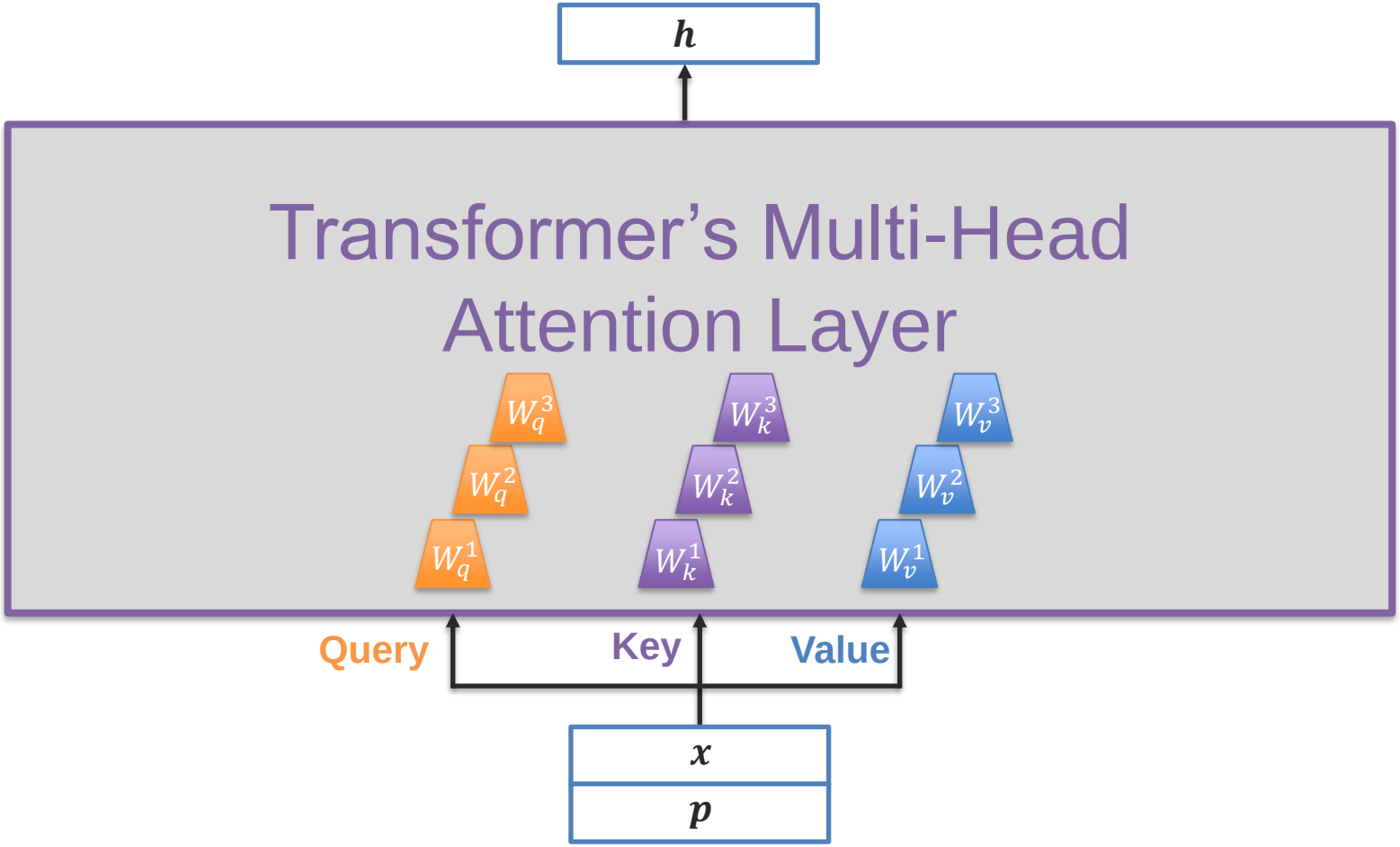


Transformer Multi-Head Self-Attention

In vector format…

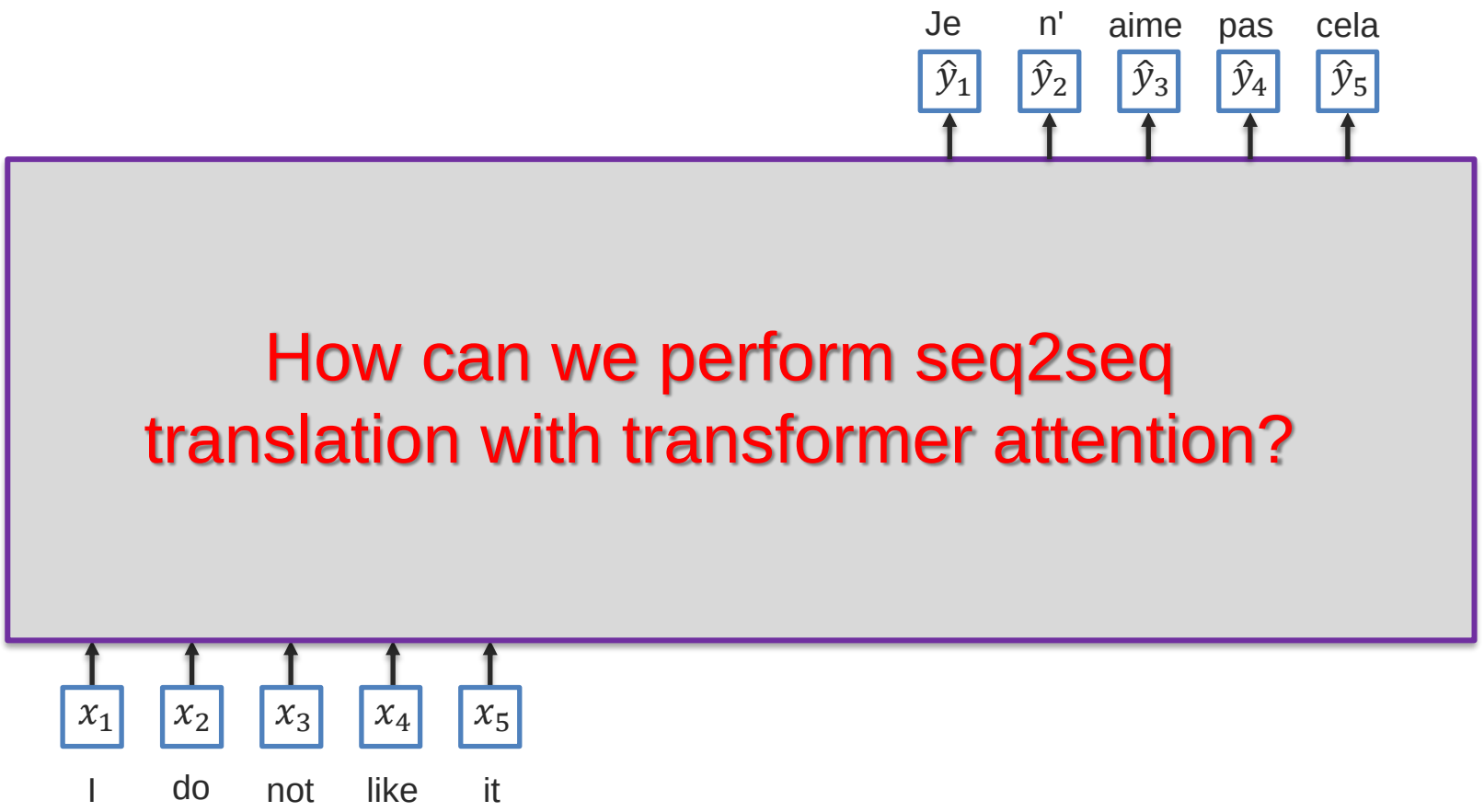


Transformer Multi-Head Attention

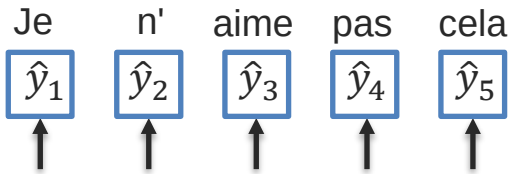


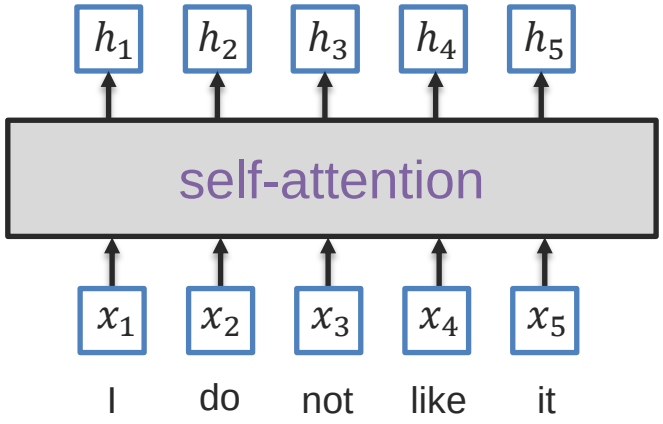
Sequence-to-Sequence Using Transformer

Sequence-to-Sequence Modeling

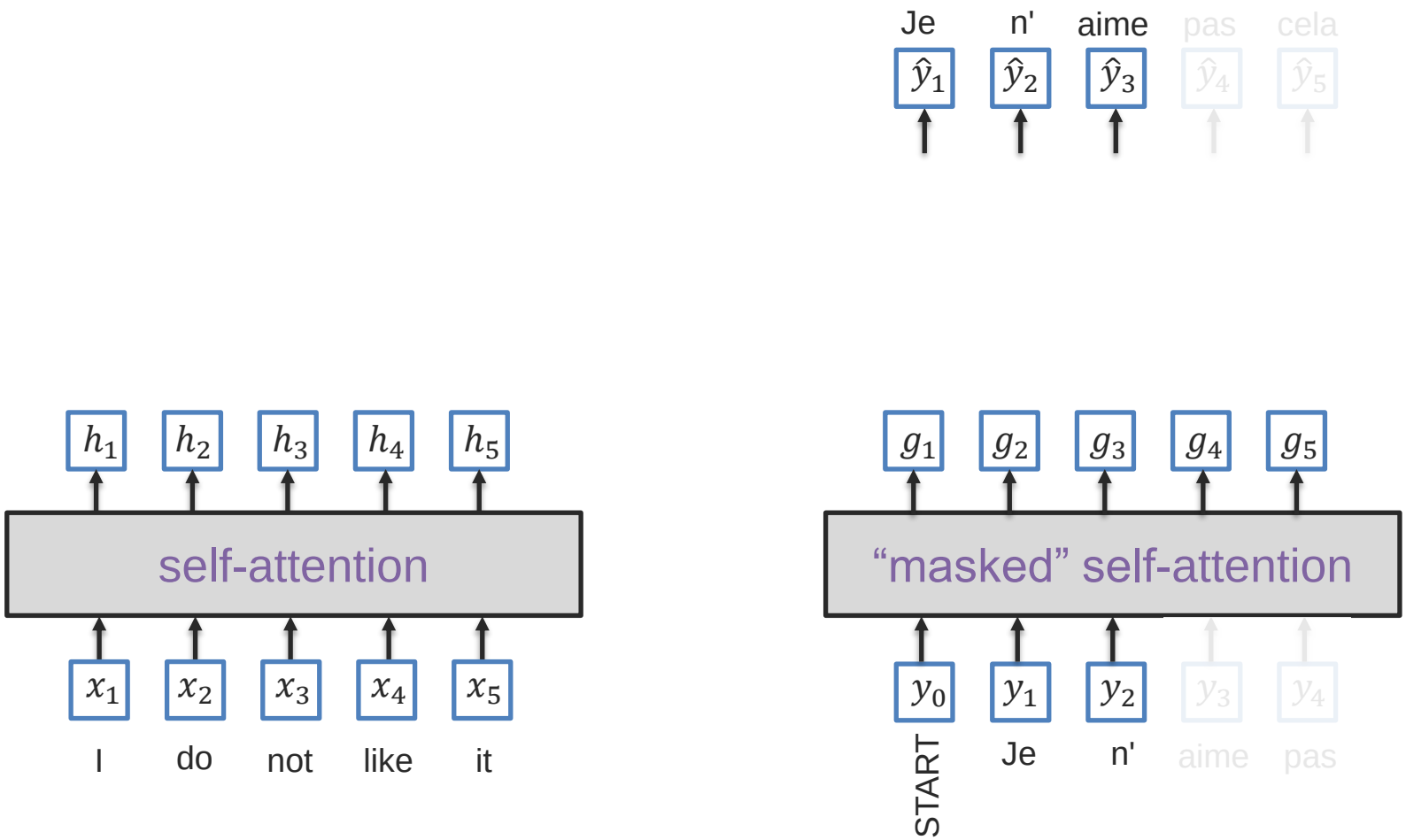


Seq2Seq with Transformer Attentions



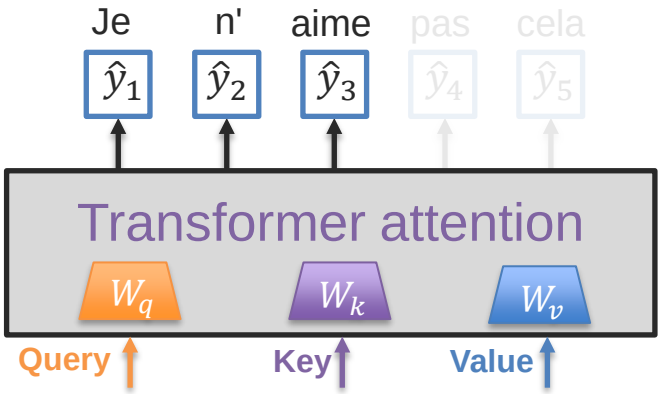


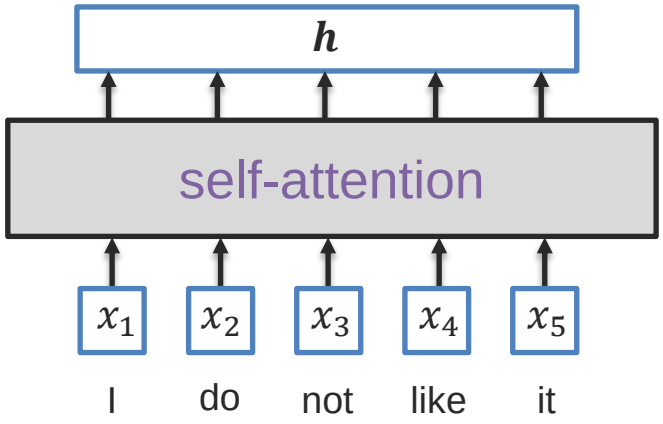
Seq2Seq with Transformer Attentions

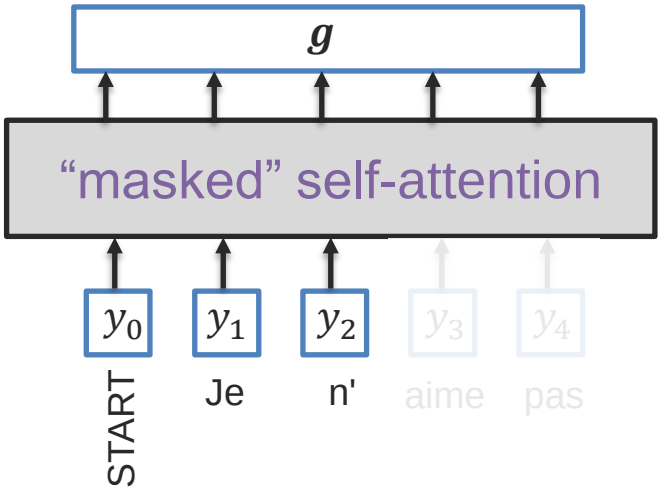


Seq2Seq with Transformer Attentions

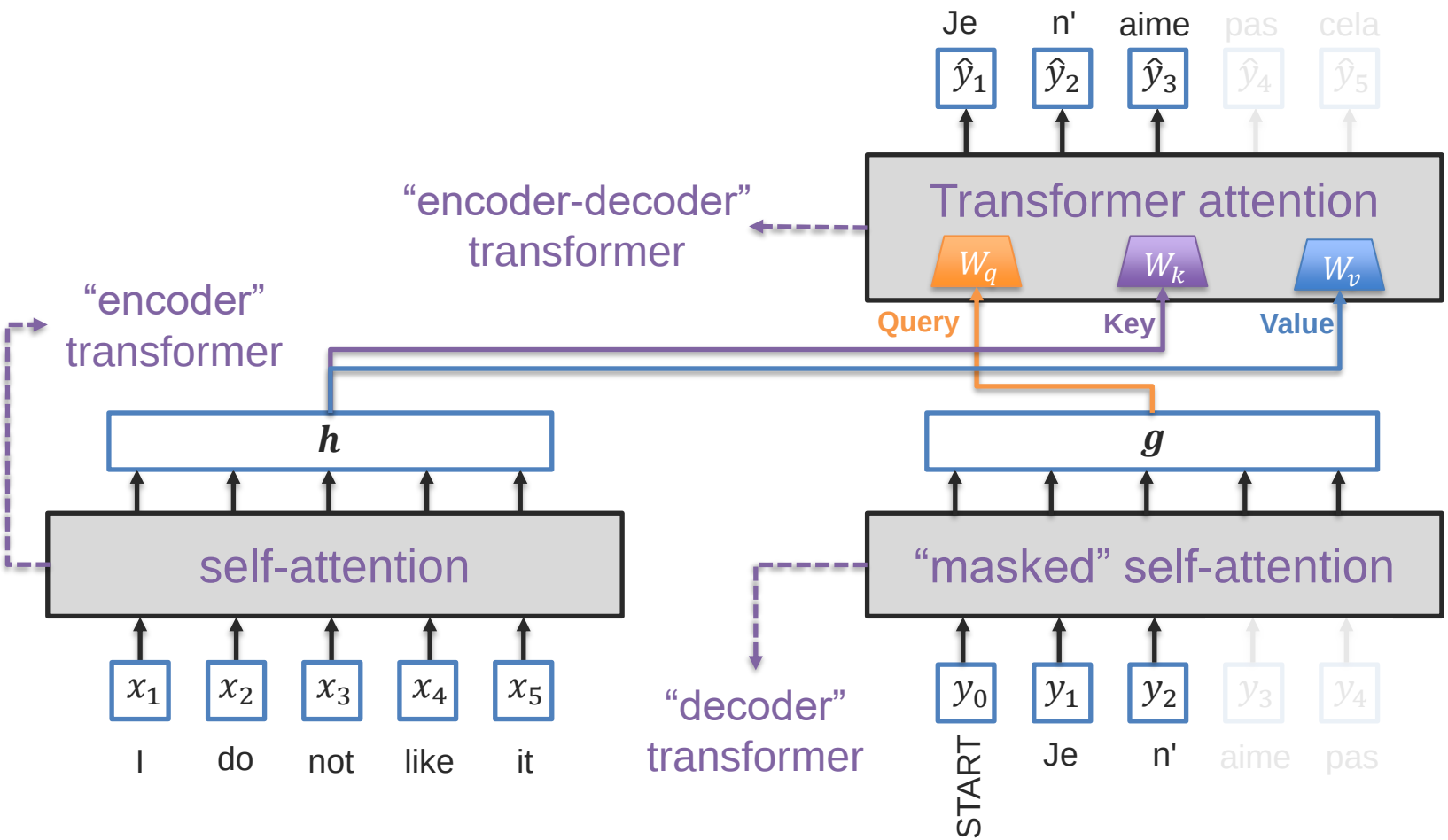
How should we connect the encoder and decoder self-attention to the transformer attention?





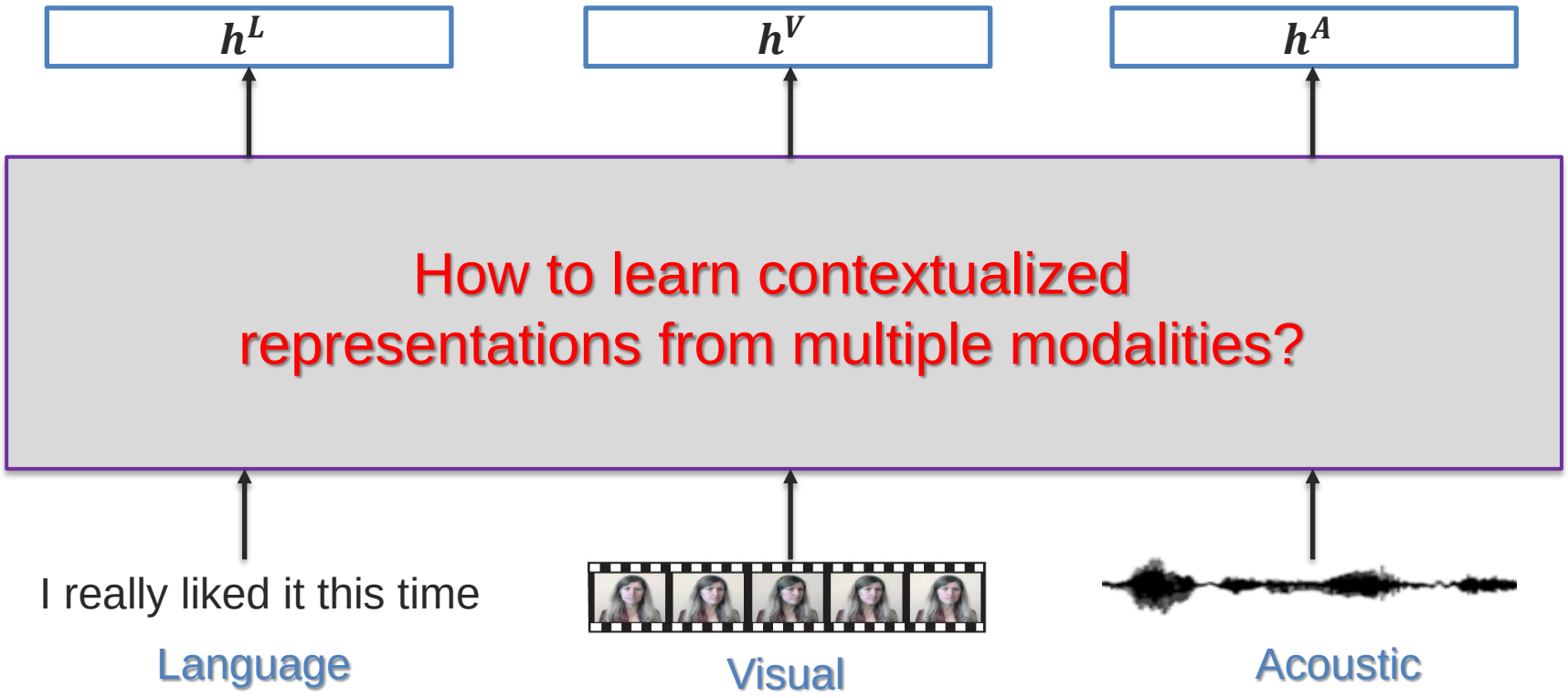


Seq2Seq with Transformer Attentions

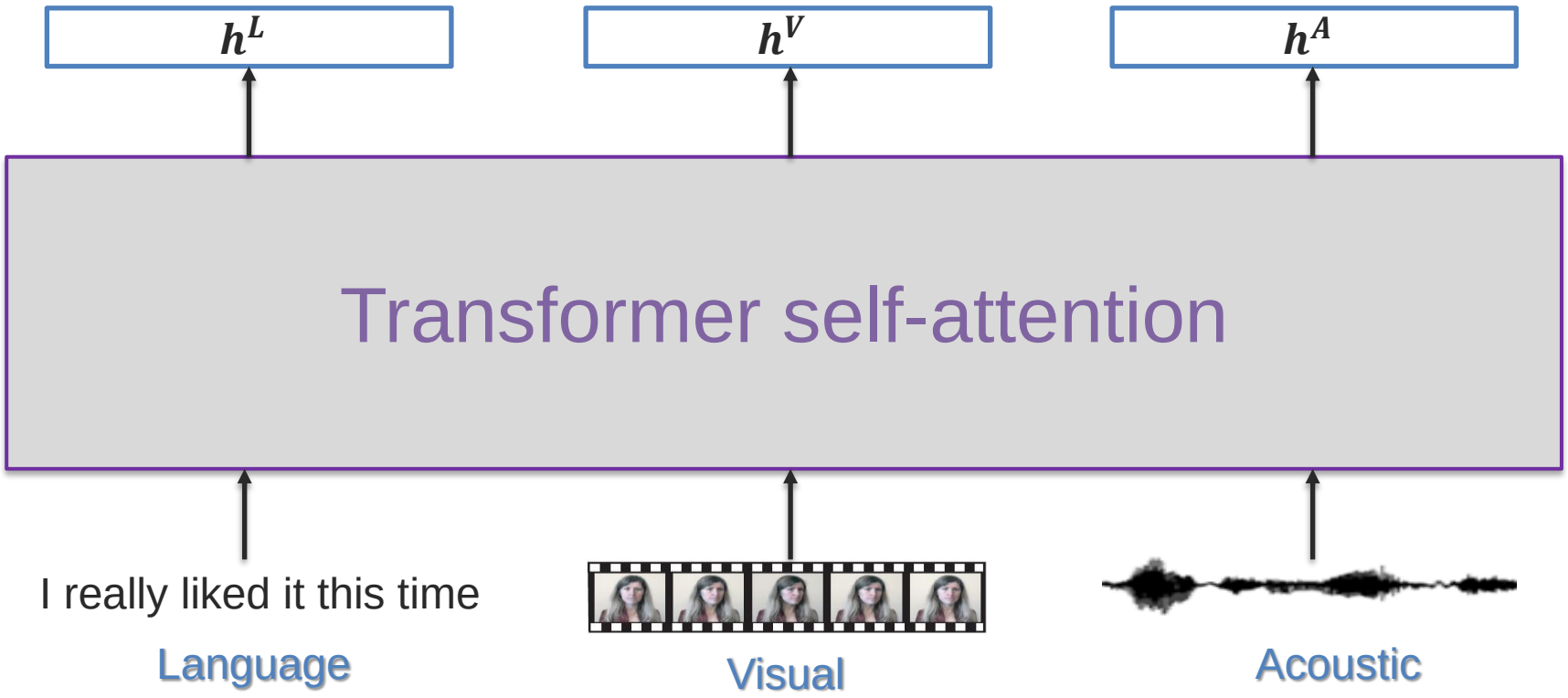


Contextualized Multimodal Embedding

Multimodal Embeddings

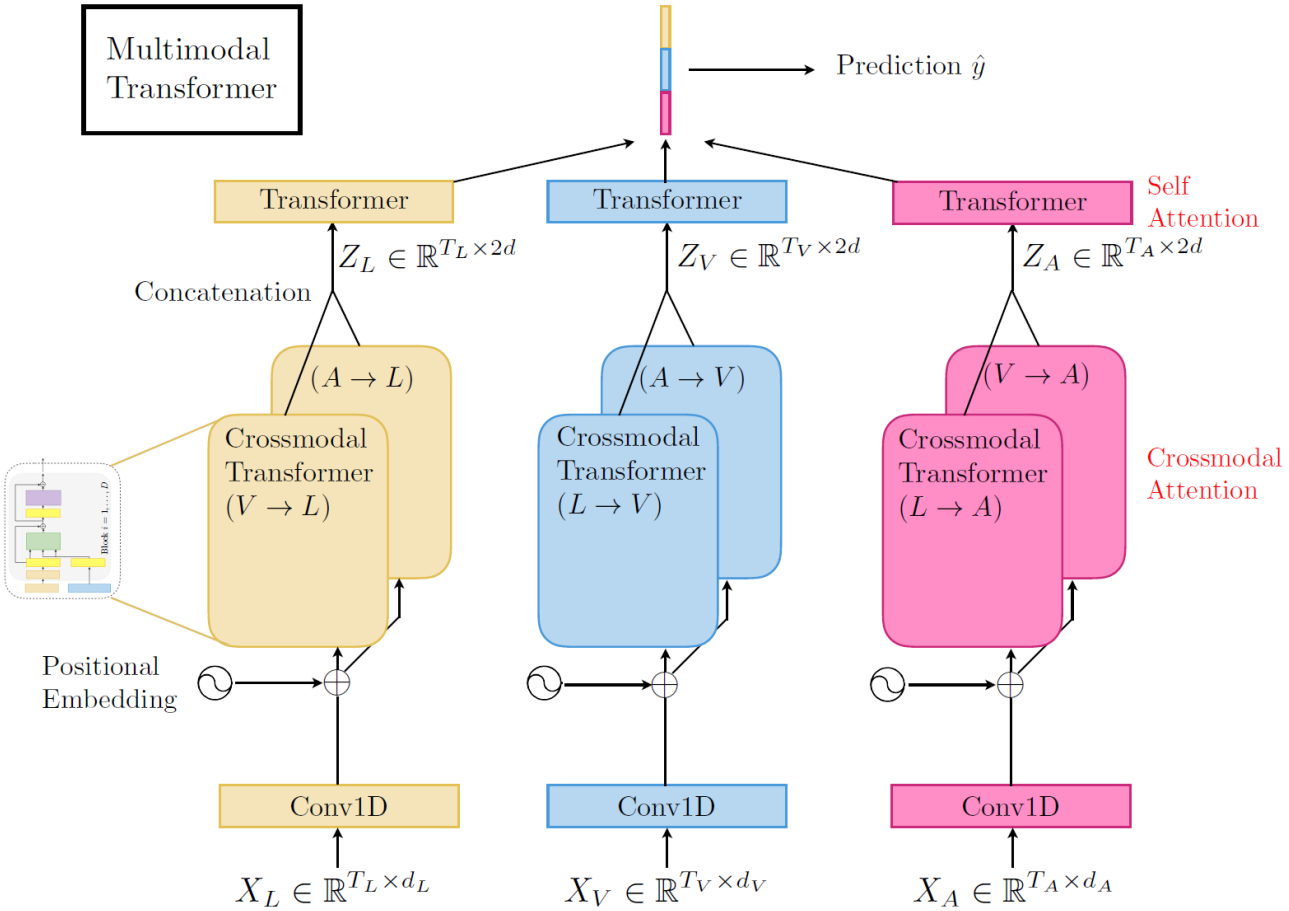


Contextualized Multimodal Embeddings



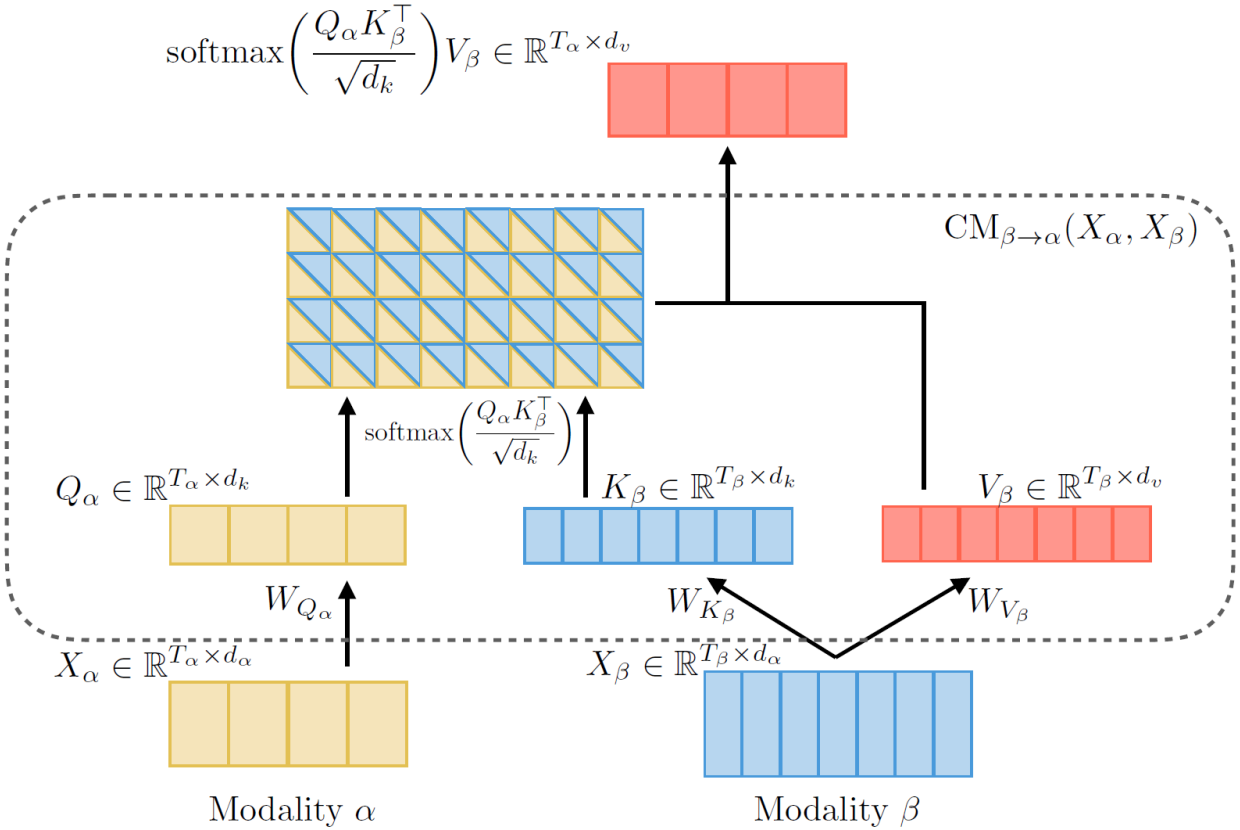
Any other approach?

Multimodal Transformer



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

Cross-Modal Transformer

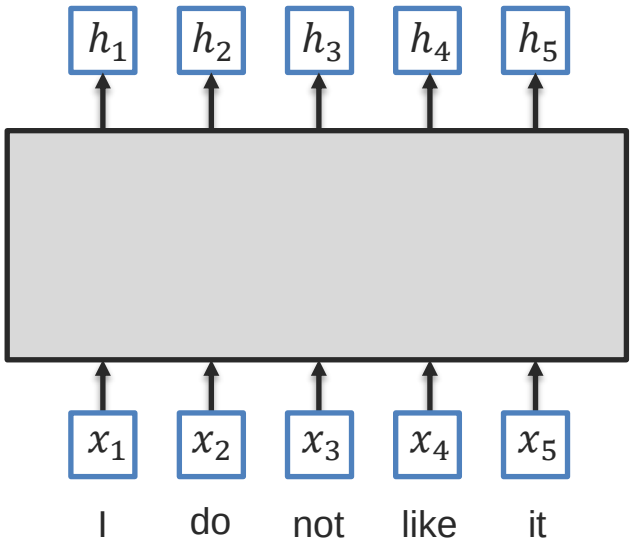


Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

Language Pre-training

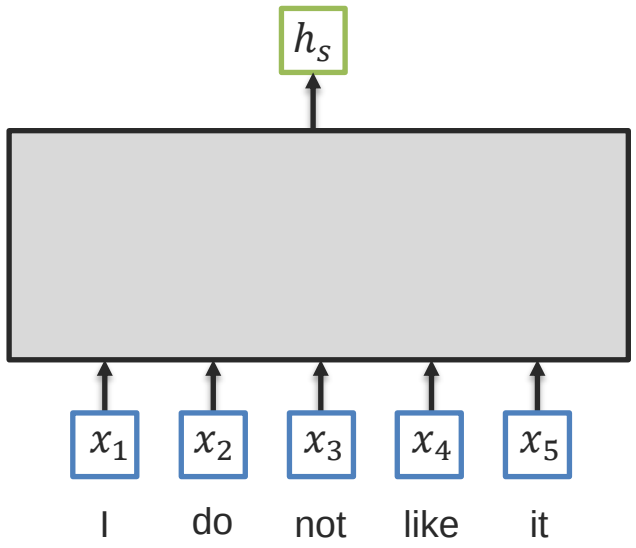
Token-level and Sentence-level Embeddings

Token-level embeddings



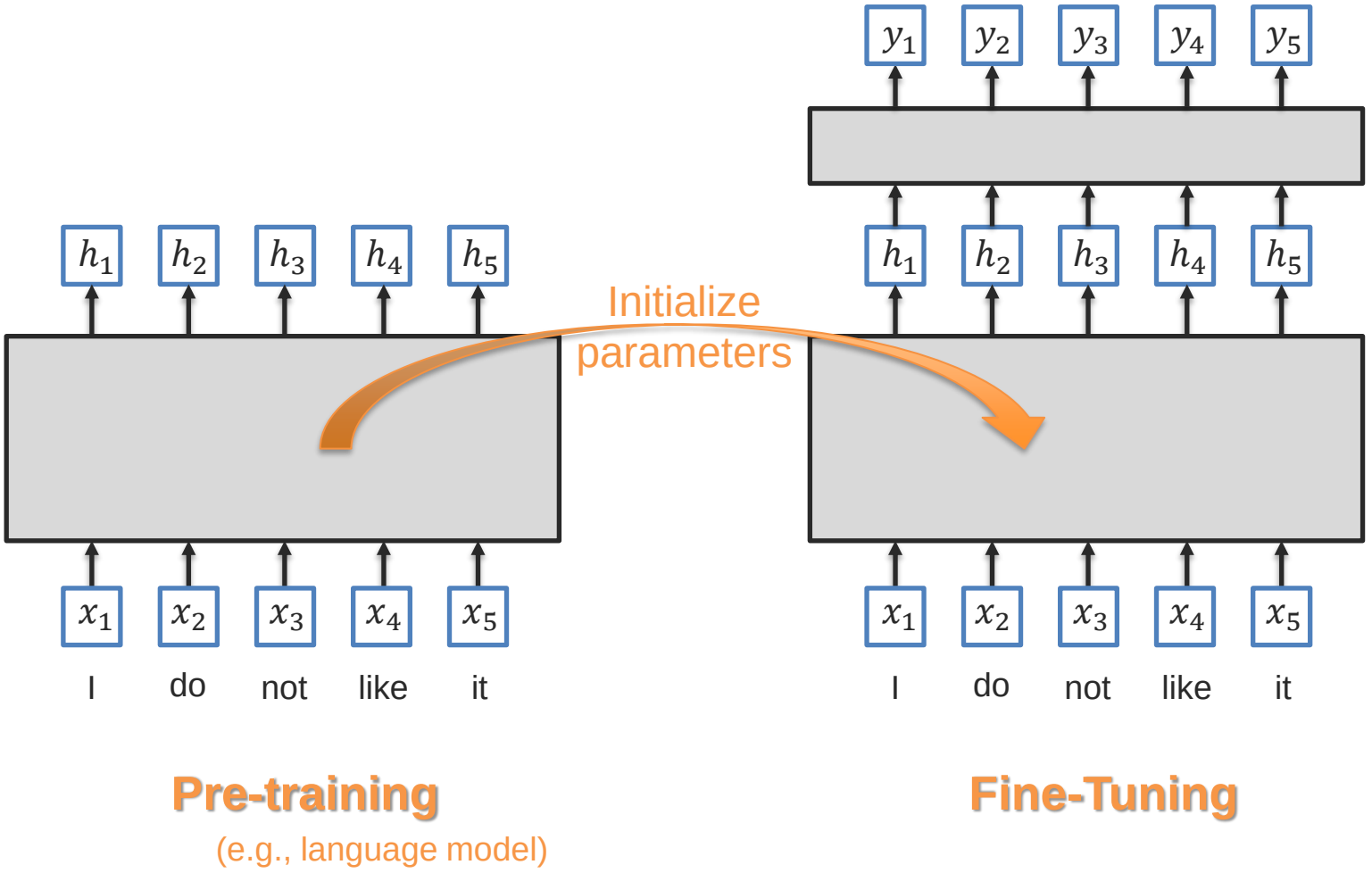
Which tasks?

Sentence-level embedding



Which tasks?

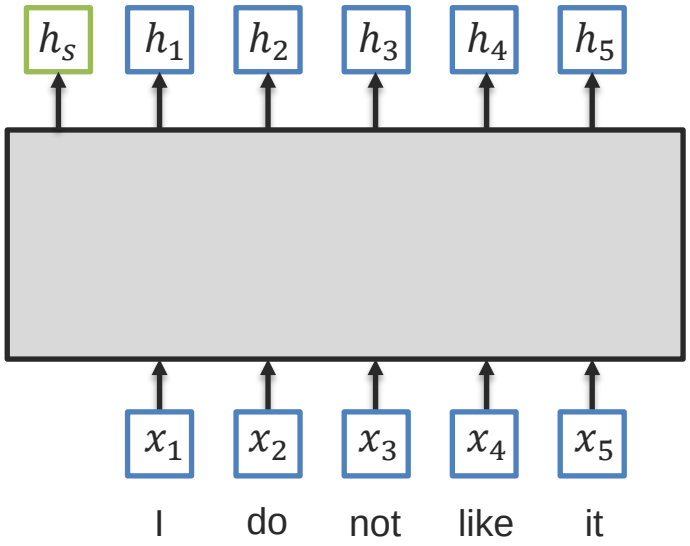
Pre-Training and Fine-Tuning



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

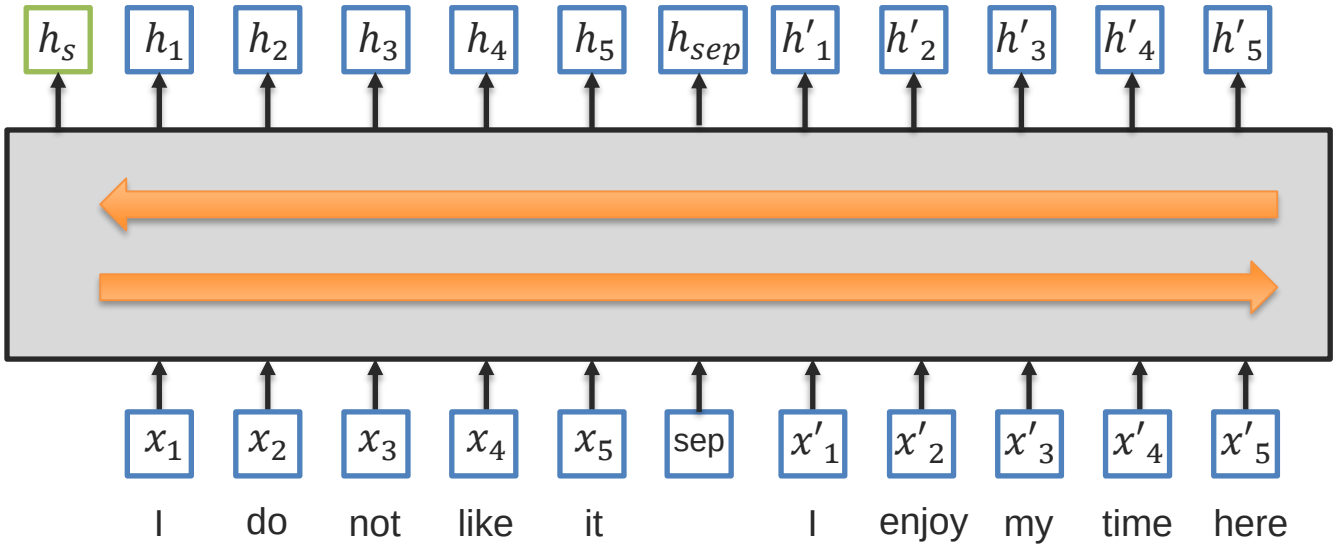


BERT:
Bidirectional Encoder Representations from Transformers

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

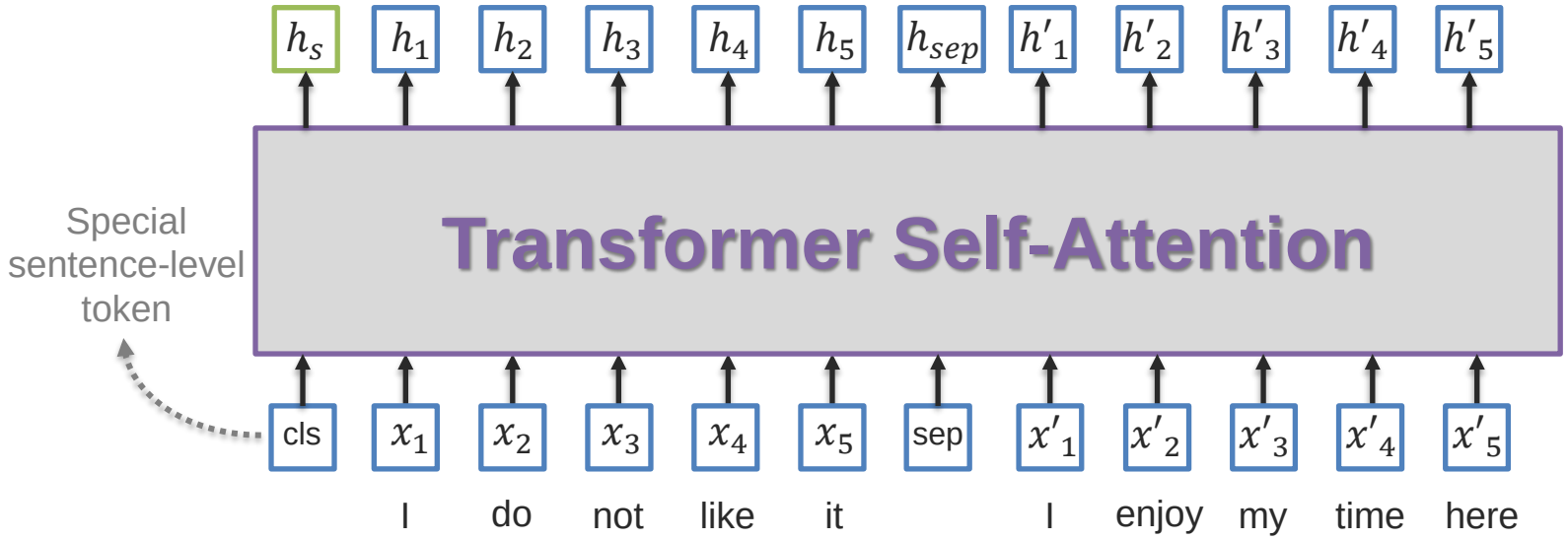
How can we do all this?



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



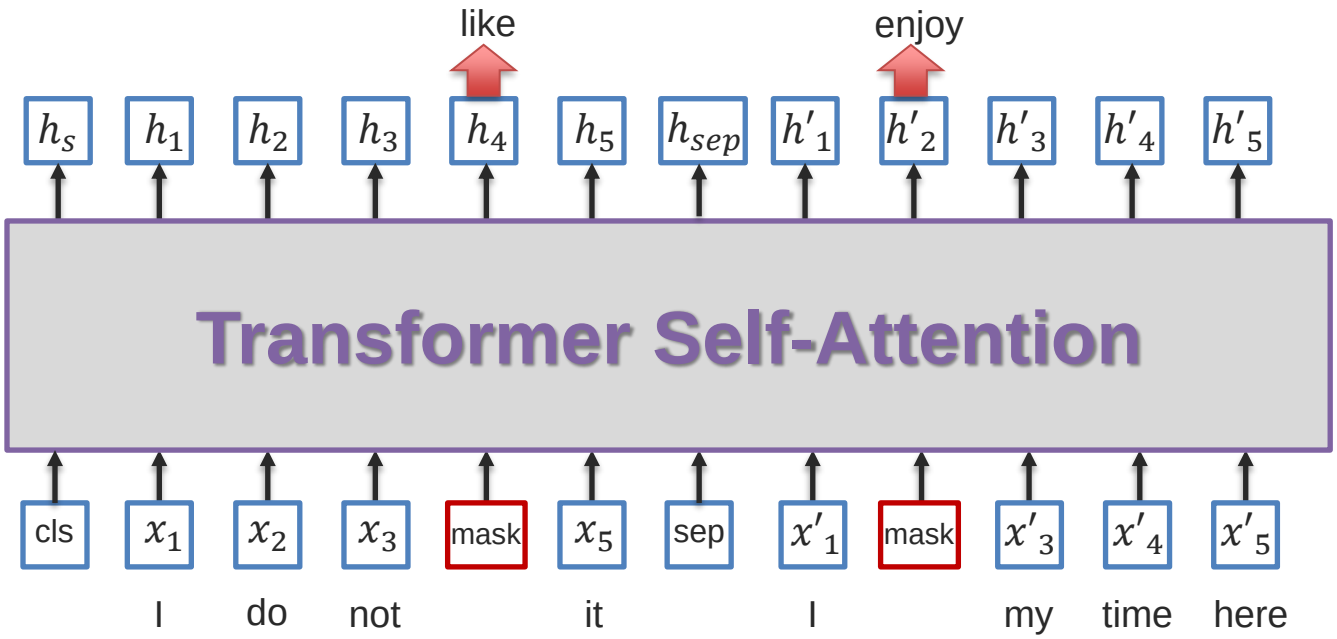
But how to train unsupervised?

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

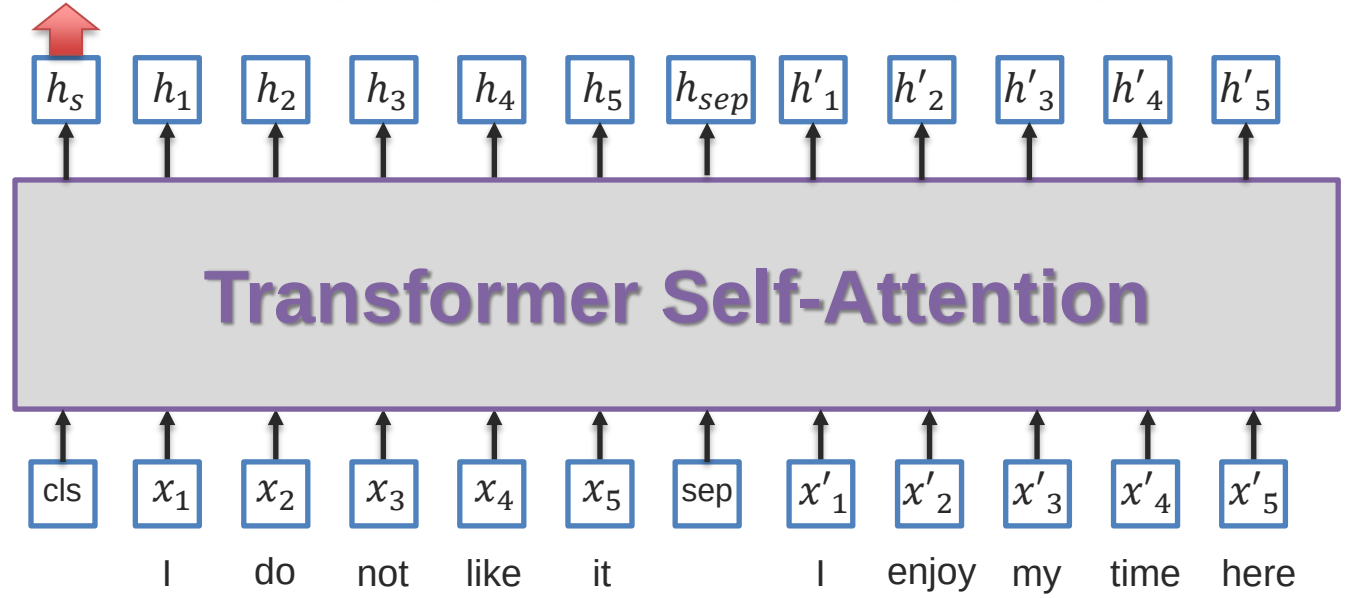
2 **Next Sentence Prediction**

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

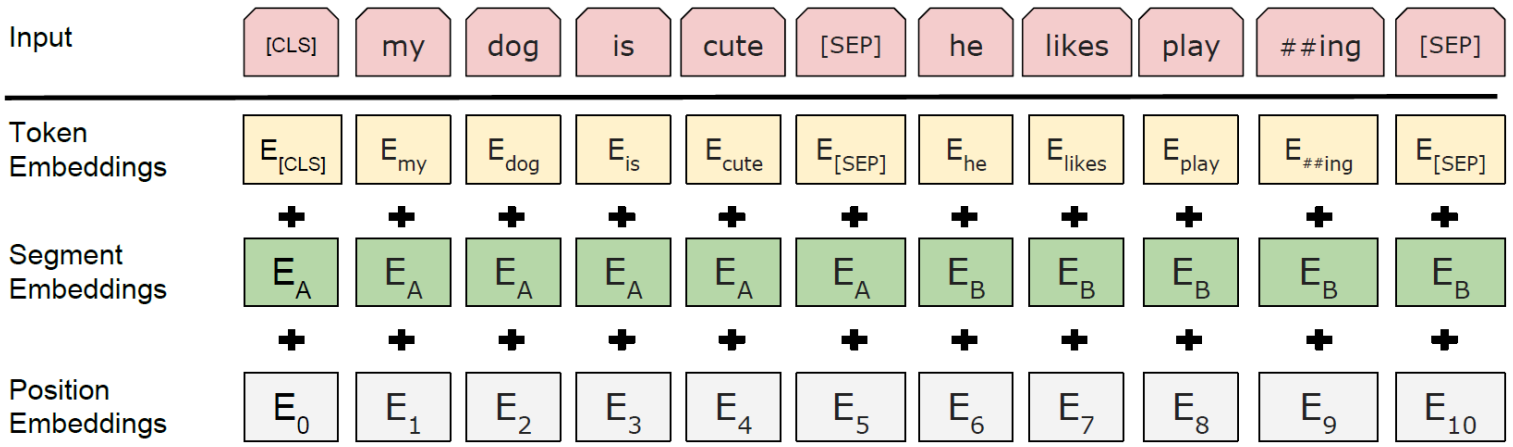
What is the loss function?

Where can we find training data?

How can BERT know the difference between both sentences?



Three Embeddings: Token + Position + Sentence

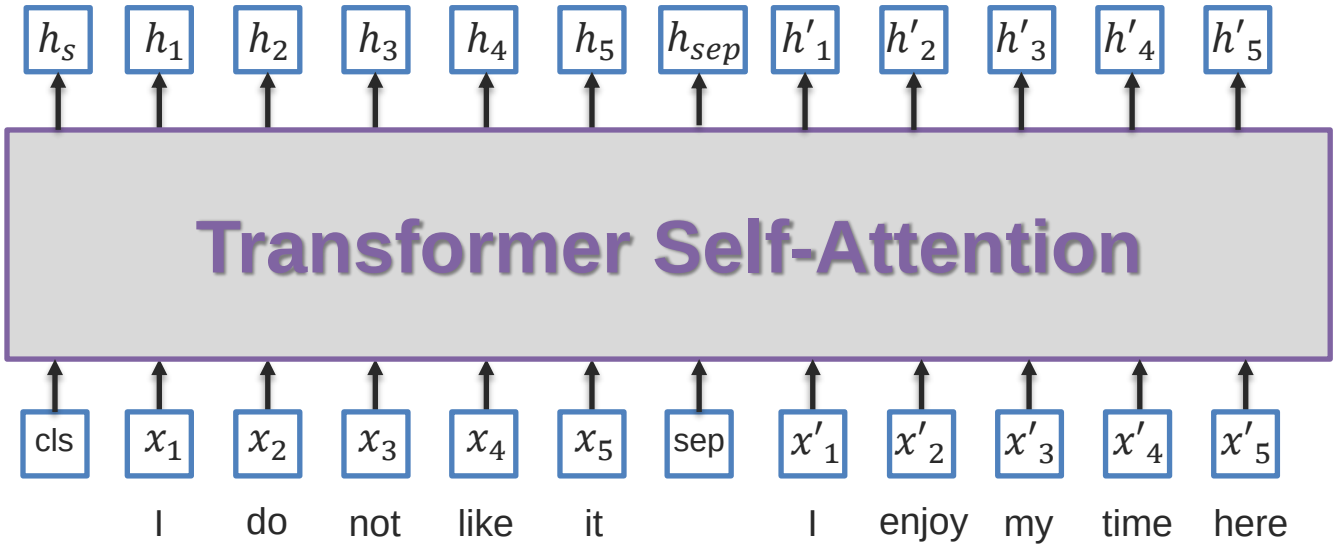


Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

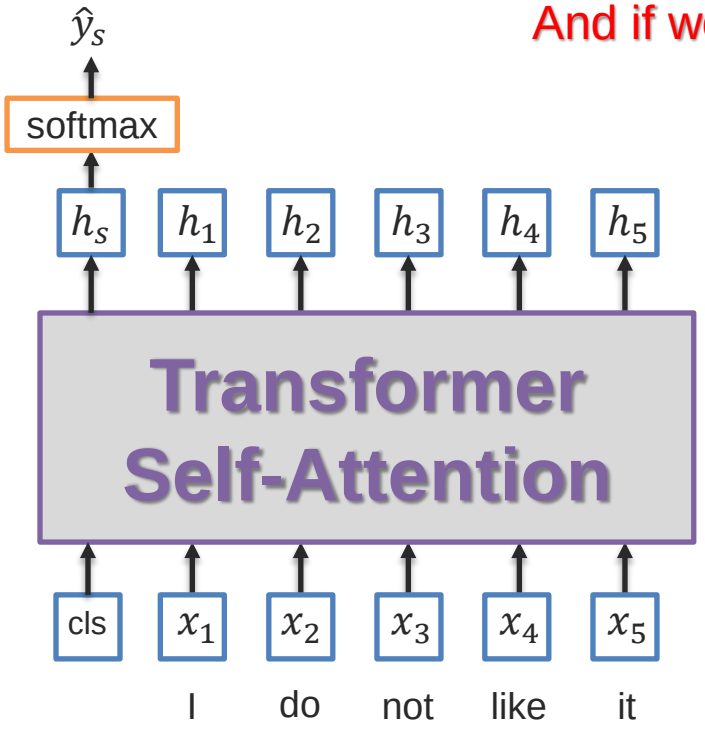
How?



Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

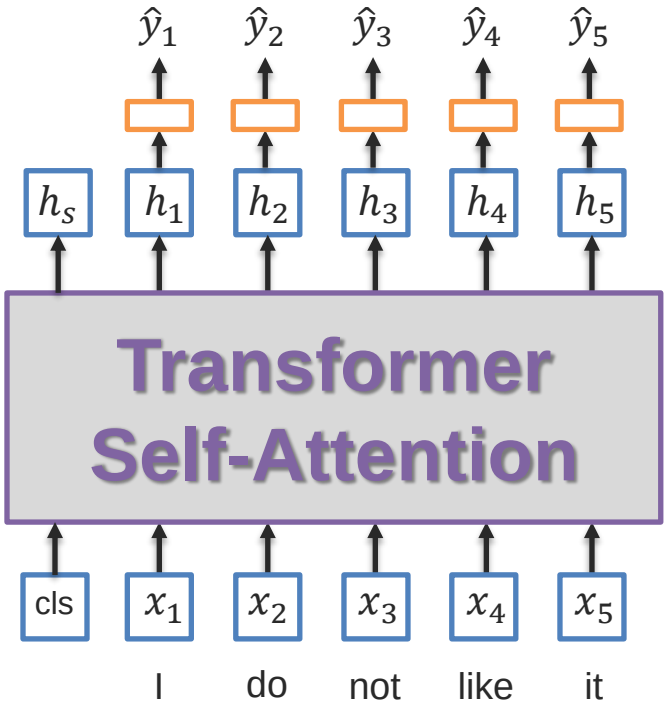


And if we have a label for each token?

Fine-Tuning BERT

2 Token-level classification for only one sentence

Examples: part-of-speech tagging, slot filling

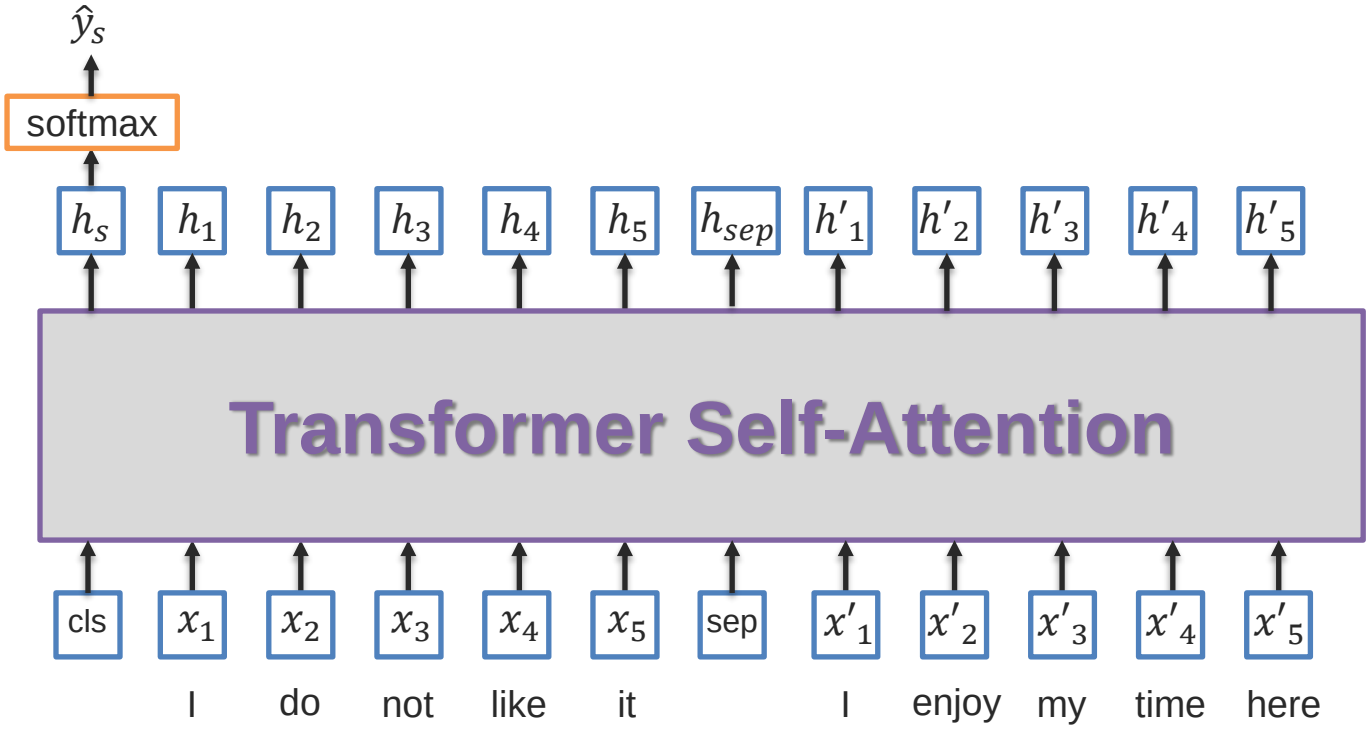


How to compare two sentences?

Fine-Tuning BERT

3 Sentence-level classification for two sentences

Examples: natural language inference



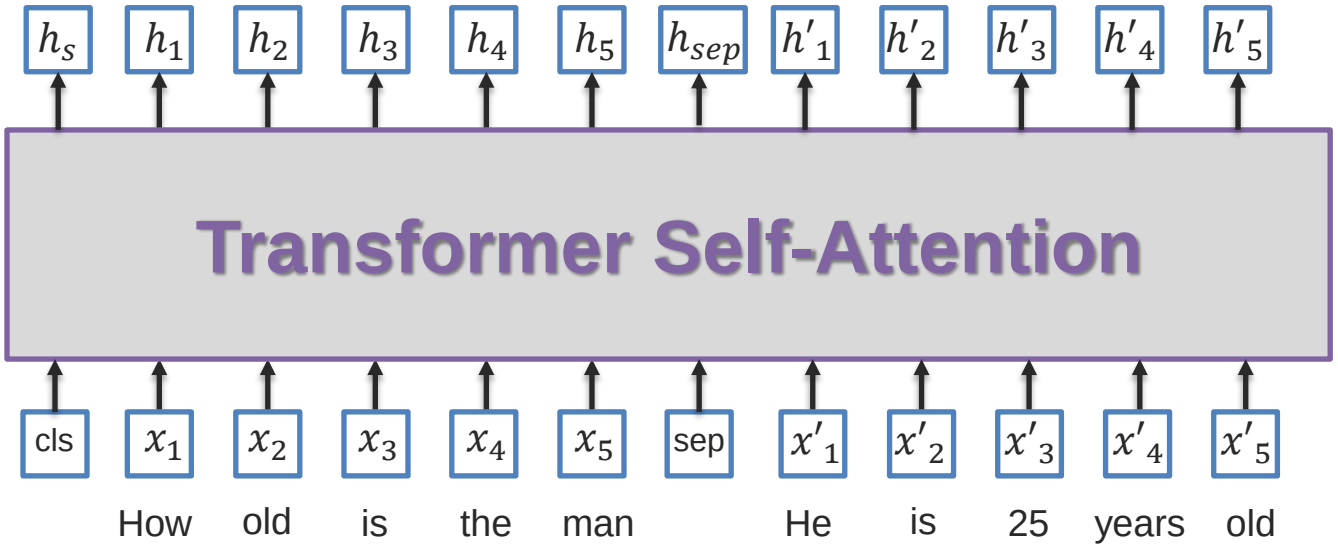
Fine-Tuning BERT

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*

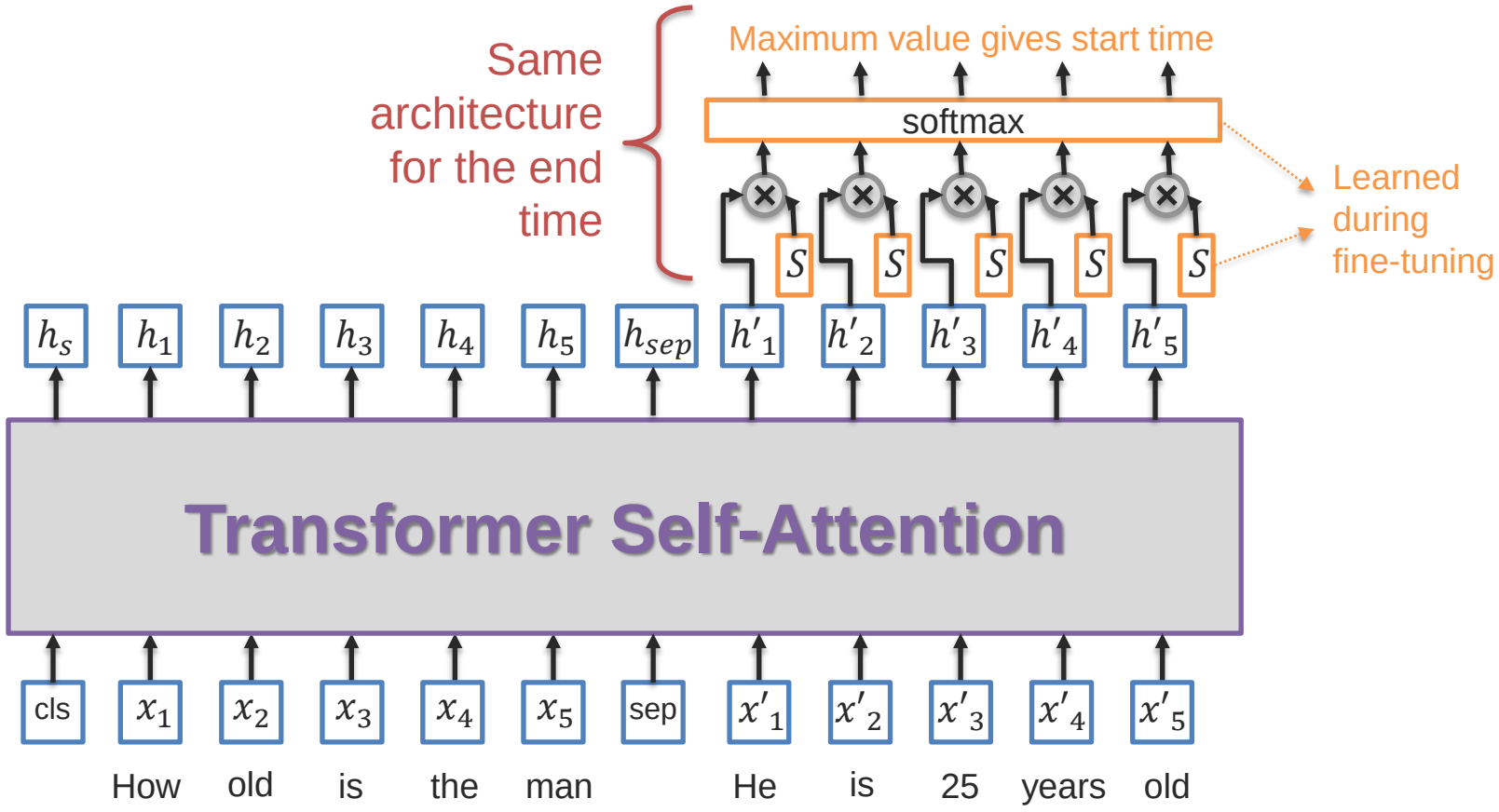
Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*



How?

Fine-Tuning BERT

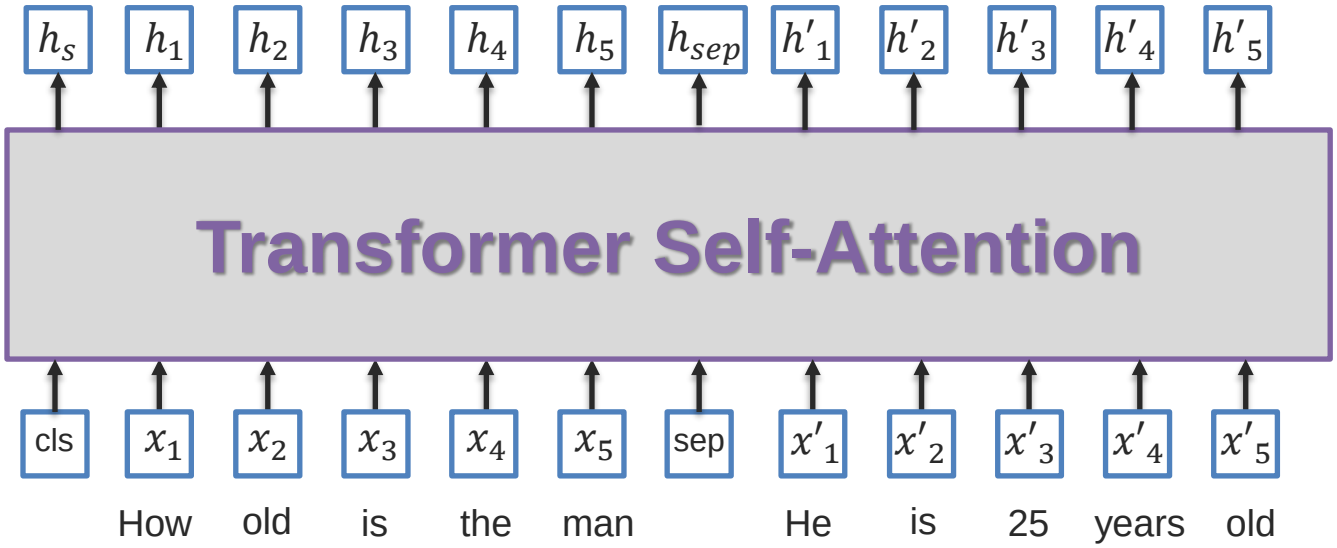
4 Question-answering: find start/end of the answer in the document



Multimodal Pre-training

Multimodal Pre-Training

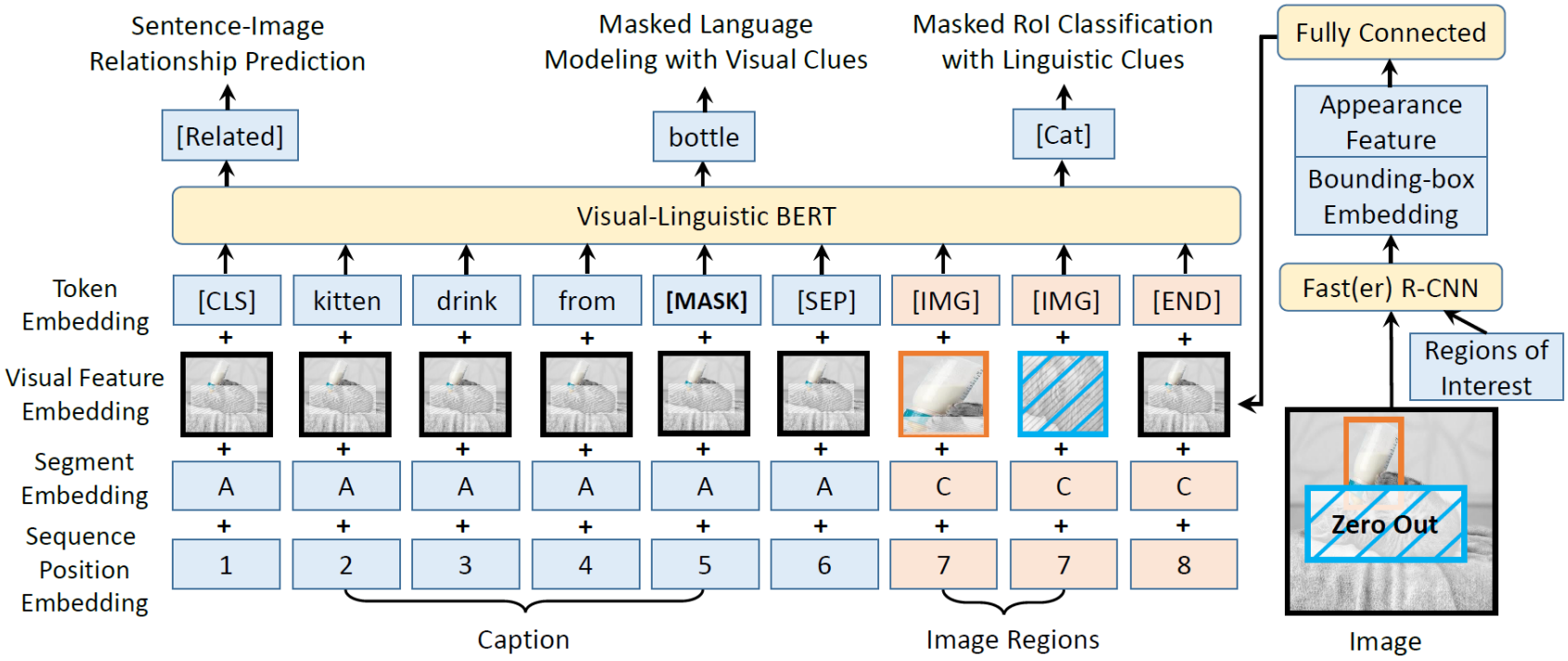
How to extend to multimodal modalities?



VL-BERT

How to extend to multimodal modalities?

Option 1: Simply concatenate tokens from different modalities

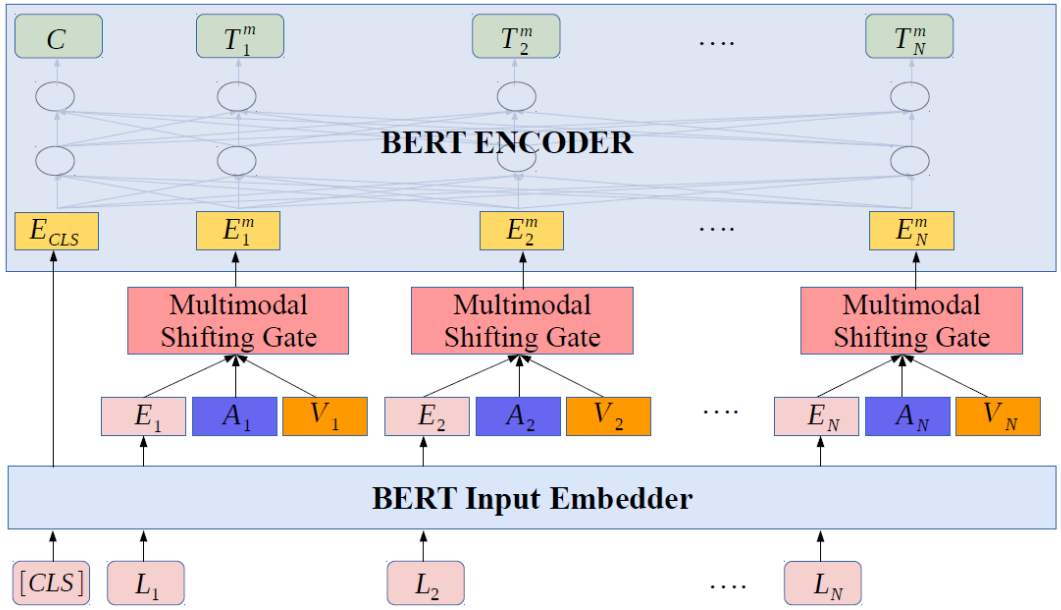


https://arxiv.org/pdf/1908.08530.pdf

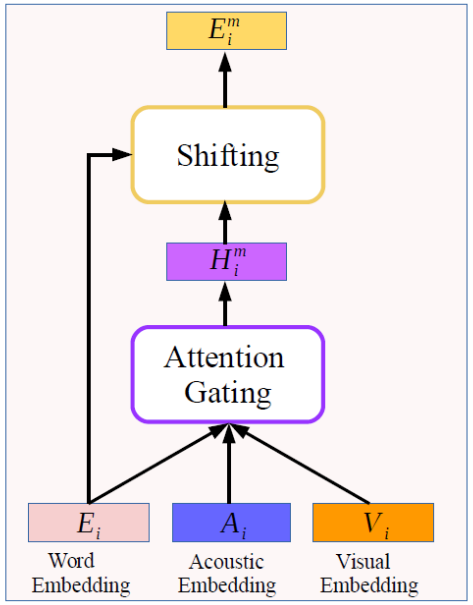
M-BERT

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