Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 6.2: Multimodal Aligned Representations

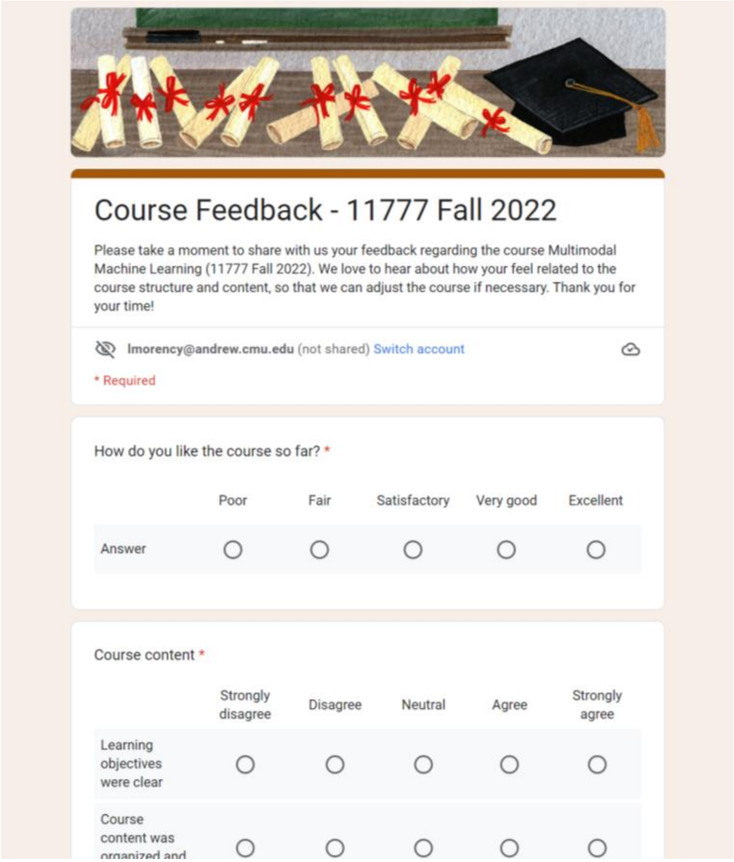
Louis-Philippe Morency

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

Administrative Stuff

Share Your Thoughts!

https://forms.gle/8vmWa7PxBfkGC2i69



Course Feedback - 11777 Fall 2022

Please take a moment to share with us your feedback regarding the course Multimodal Machine Learning (11777 Fall 2022). 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!

lmorency@andrew.cmu.edu (not shared) Switch account

* 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	○	○	○	○	○
Course content was organized and	○	○	○	○	○

Deadline

Please submit your feedback about this course before this Wednesday 10/5

Optional,
but greatly appreciated! ☺

Anonymous, by default.

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

Second Project Assignment (Due Monday 10/10)

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 CNNs, word2vec, BERT, …

Two types of analyses:

- Idea-oriented analyses
- Modality-oriented analyses

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 6.2: Alignment and Representation

Louis-Philippe Morency

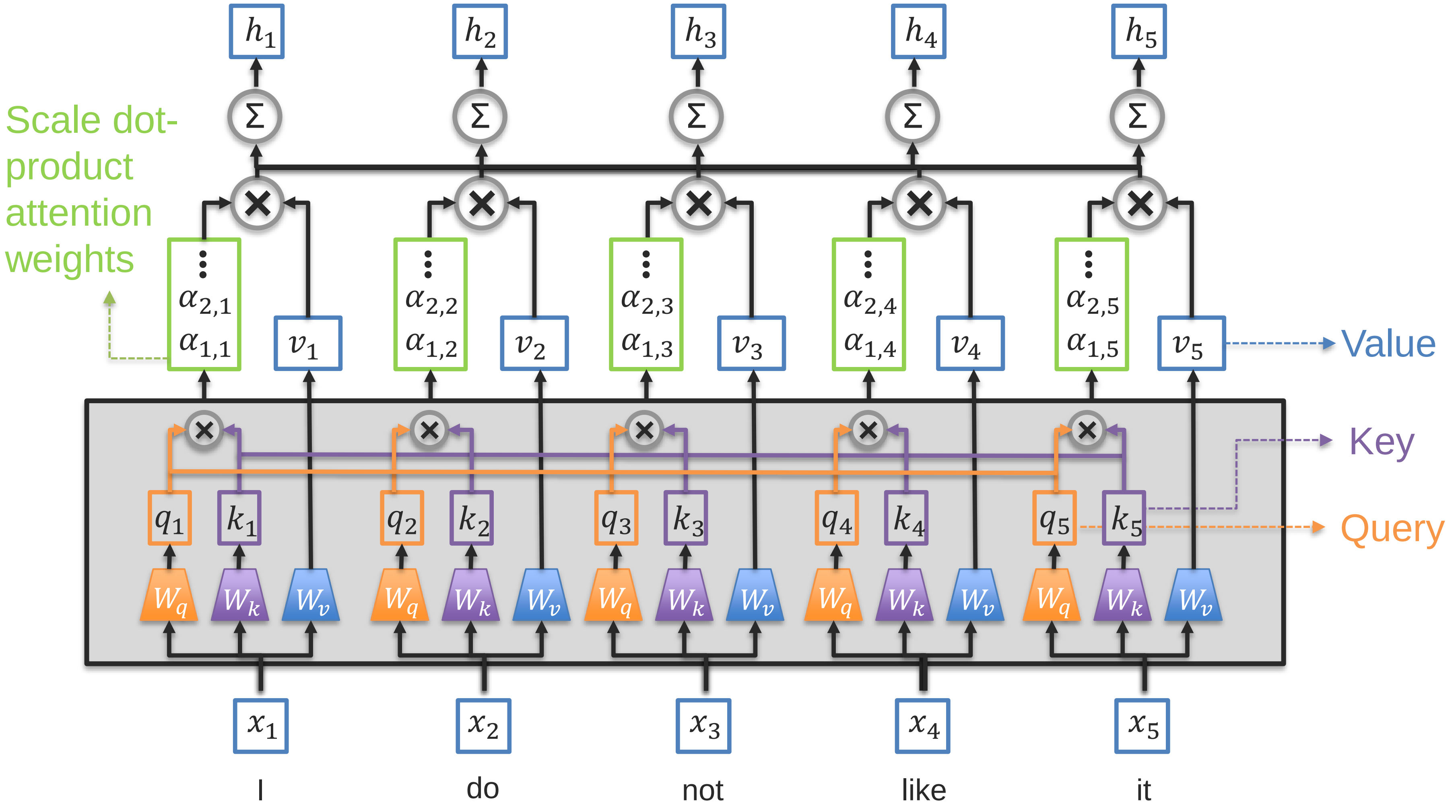
** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Objectives of today’s class

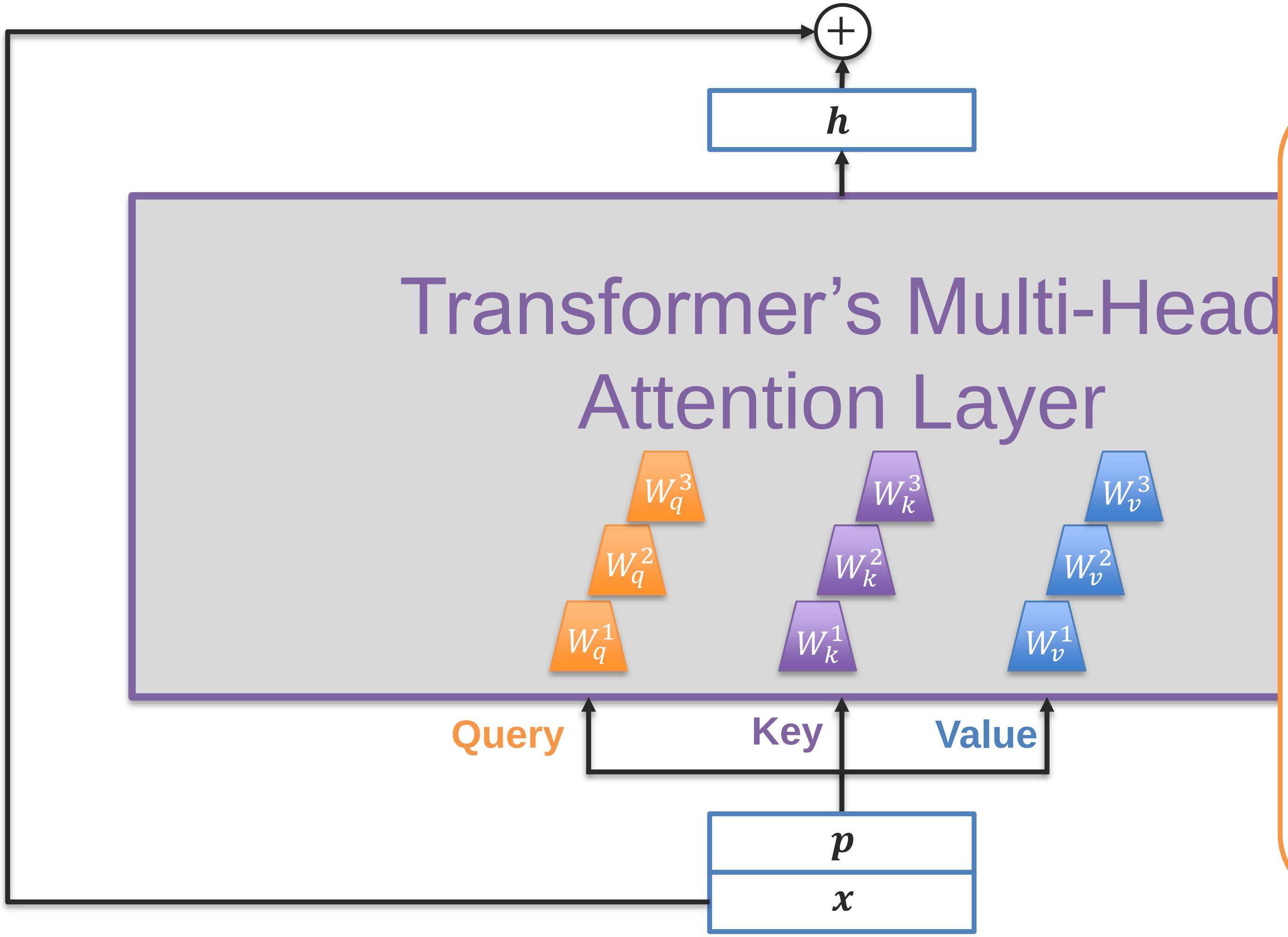
- Transformer pre-training
  - BERT: Bidirectional Encoder Representations from Transformers
- Multimodal transformers (Image and language)
  - Concatenated transformers (VisualBERT, Uniter)
  - Crossmodal transformers (ViLBERT, LXMERT
  - Modality-shift transformer (MAG-BERT)
  - Video and language transformers (VideoBERT, ActBERT)
- Visual transformers
  - Vision transformer, Masked Auto-Encoder
  - Visual-and-language transformer (ViLT, ALBEF)

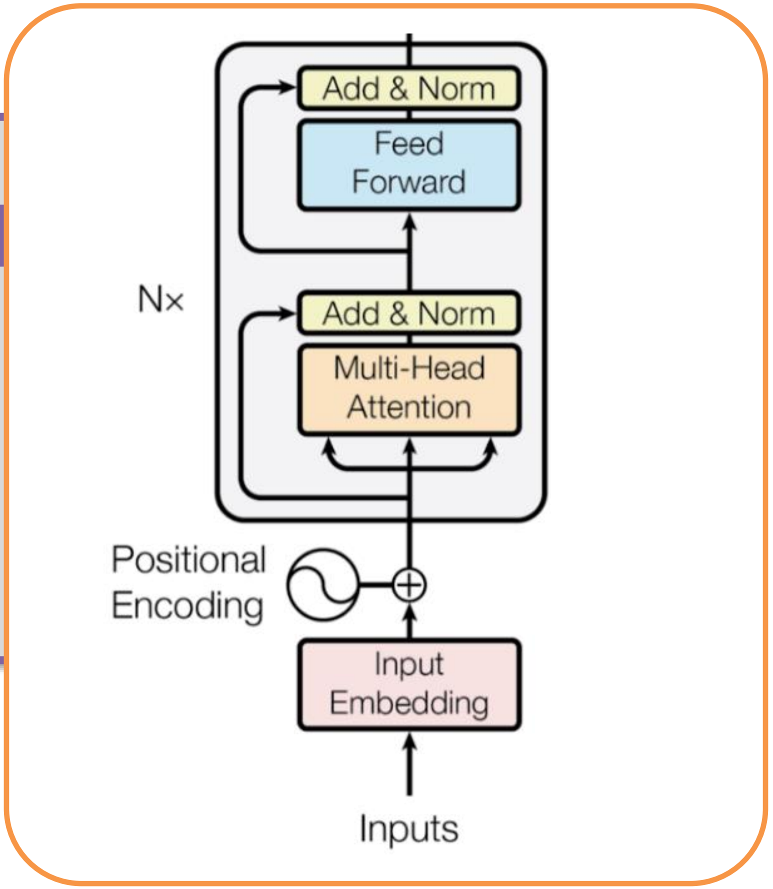
BERT:
Transformer Pre-training

Transformer Self-Attention



Transformer – Residual Connection

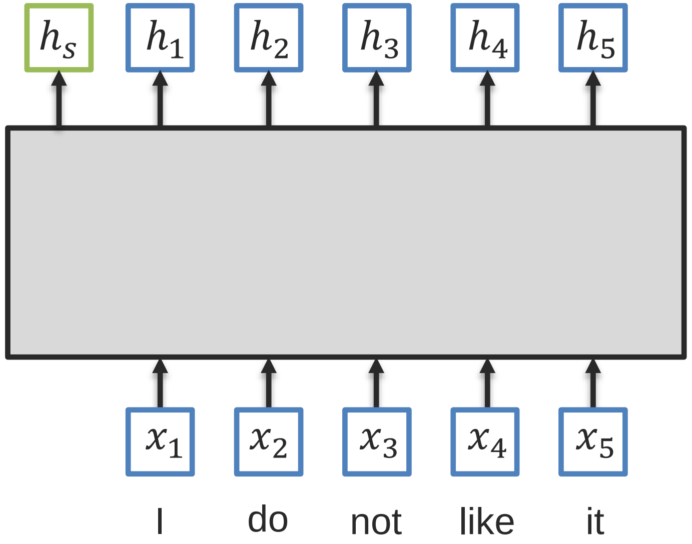




BERT: Bidirectional Encoder Representations from Transformers

Advantages:

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

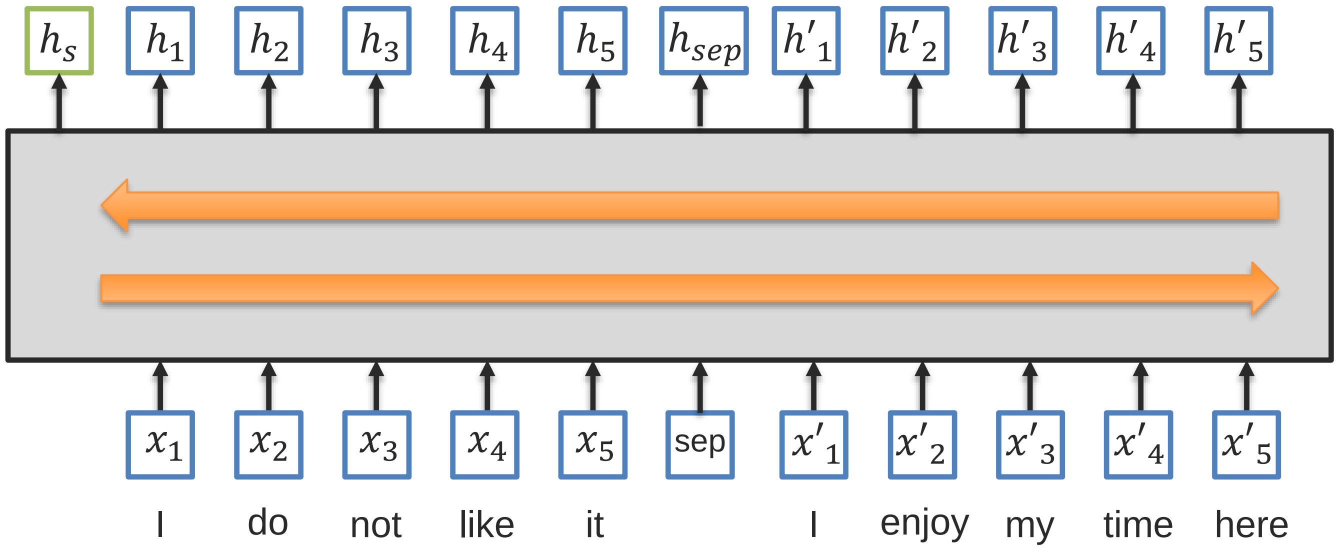


BERT: Bidirectional Encoder Representations from Transformers

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

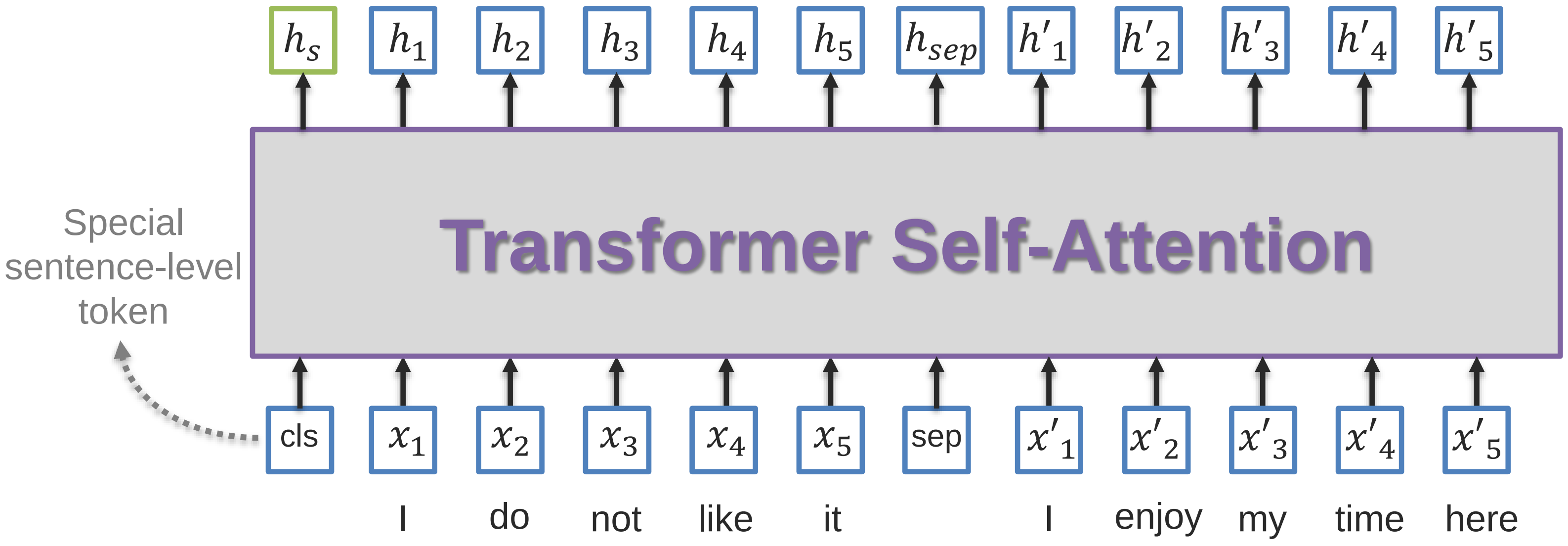
How can we do all this?



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



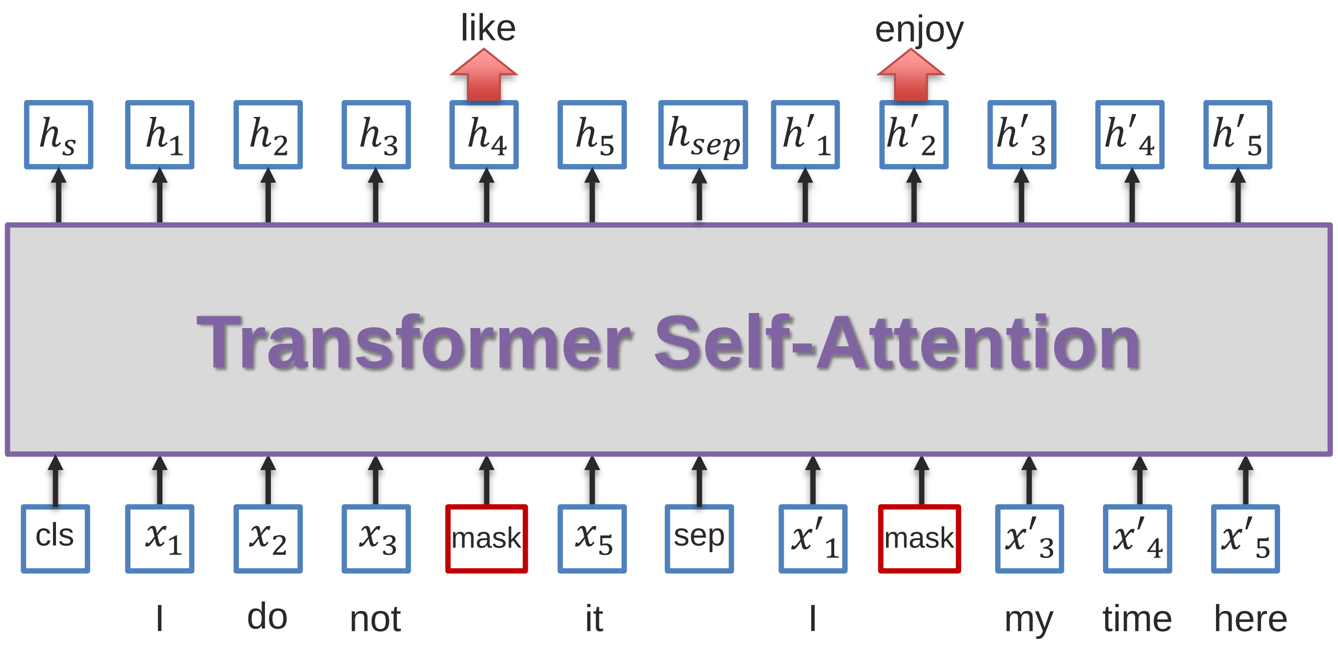
But how to train self-supervised?

Pre-training BERT Model

1 **Masked Language Model**

Randomly mask input tokens and then try to predict them

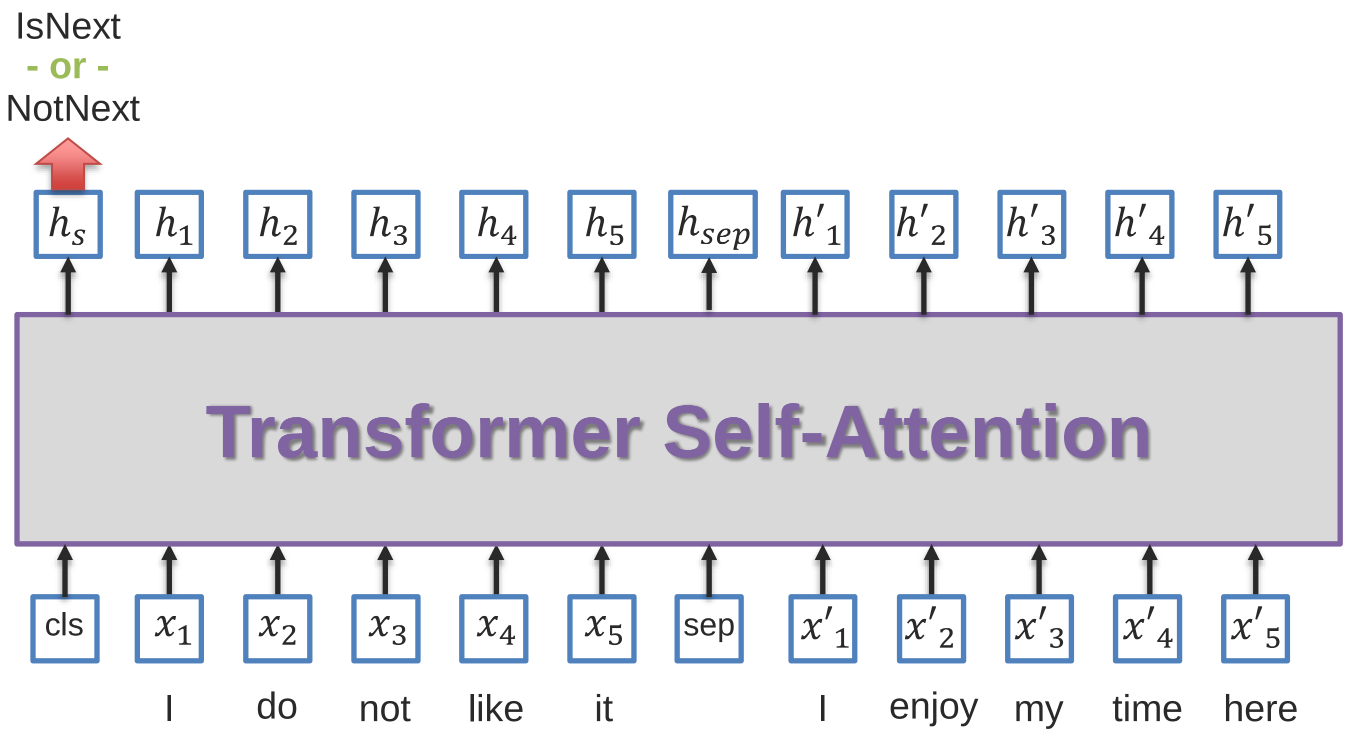
What is the loss function?



Pre-training BERT Model

2 **Next Sentence Prediction**

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

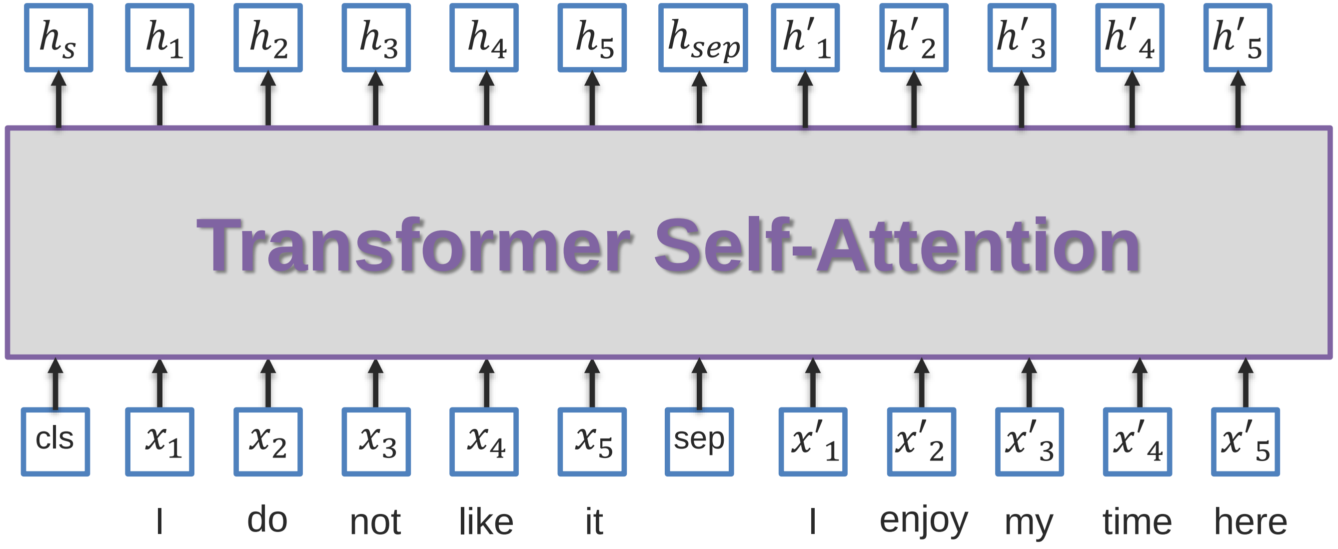


Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

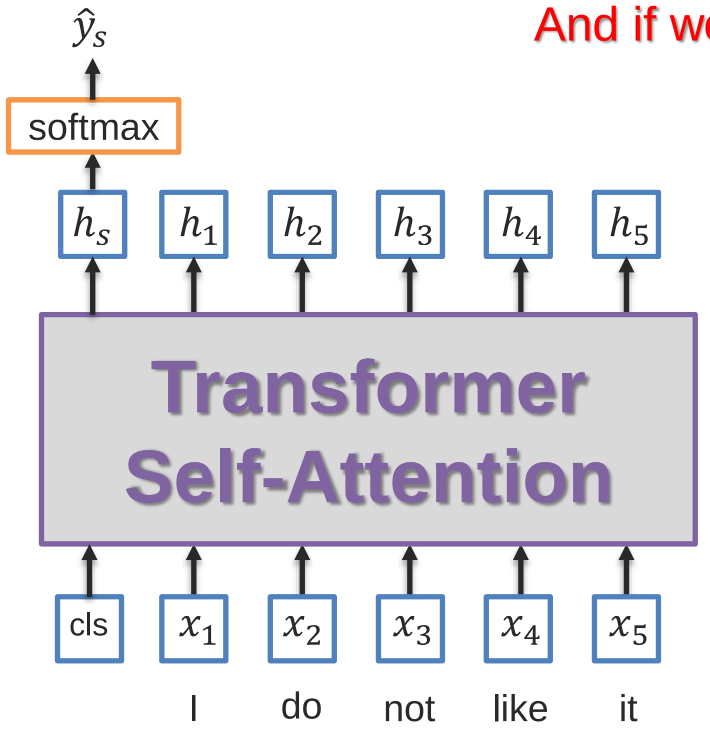
How?



Fine-Tuning BERT

1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

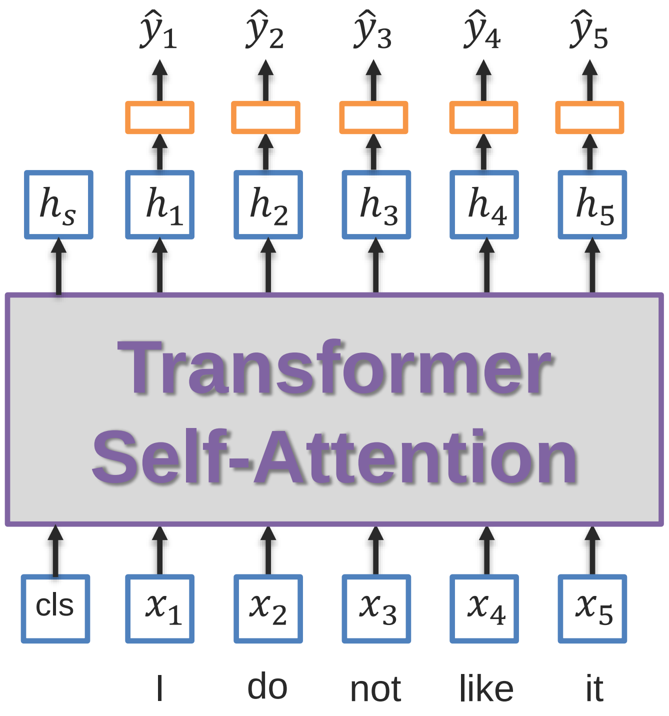


And if we have a label for each token?

Fine-Tuning BERT

2 Token-level classification for only one sentence

Examples: part-of-speech tagging, slot filling



How to compare two sentences?

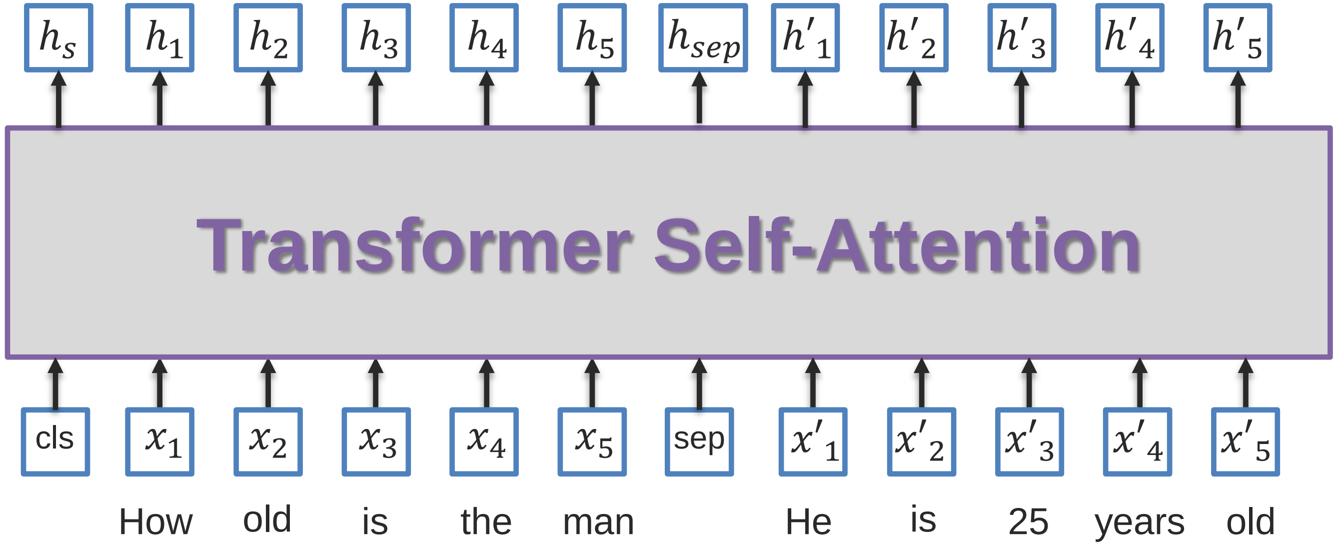
Fine-Tuning BERT

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*

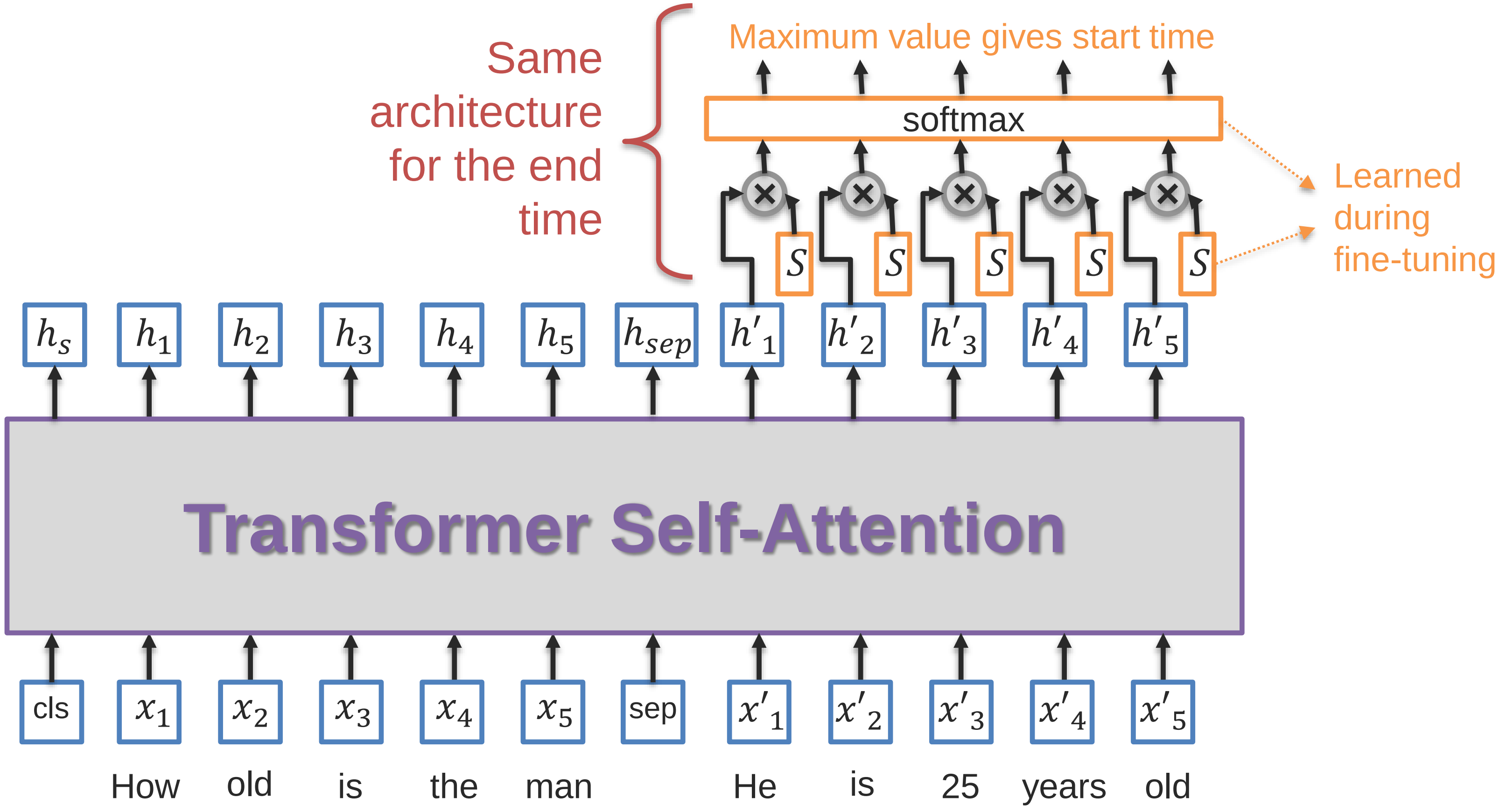
Question 2: "*What was the name of the 1937 treaty?*"
Plausible Answer: *Bald Eagle Protection Act*



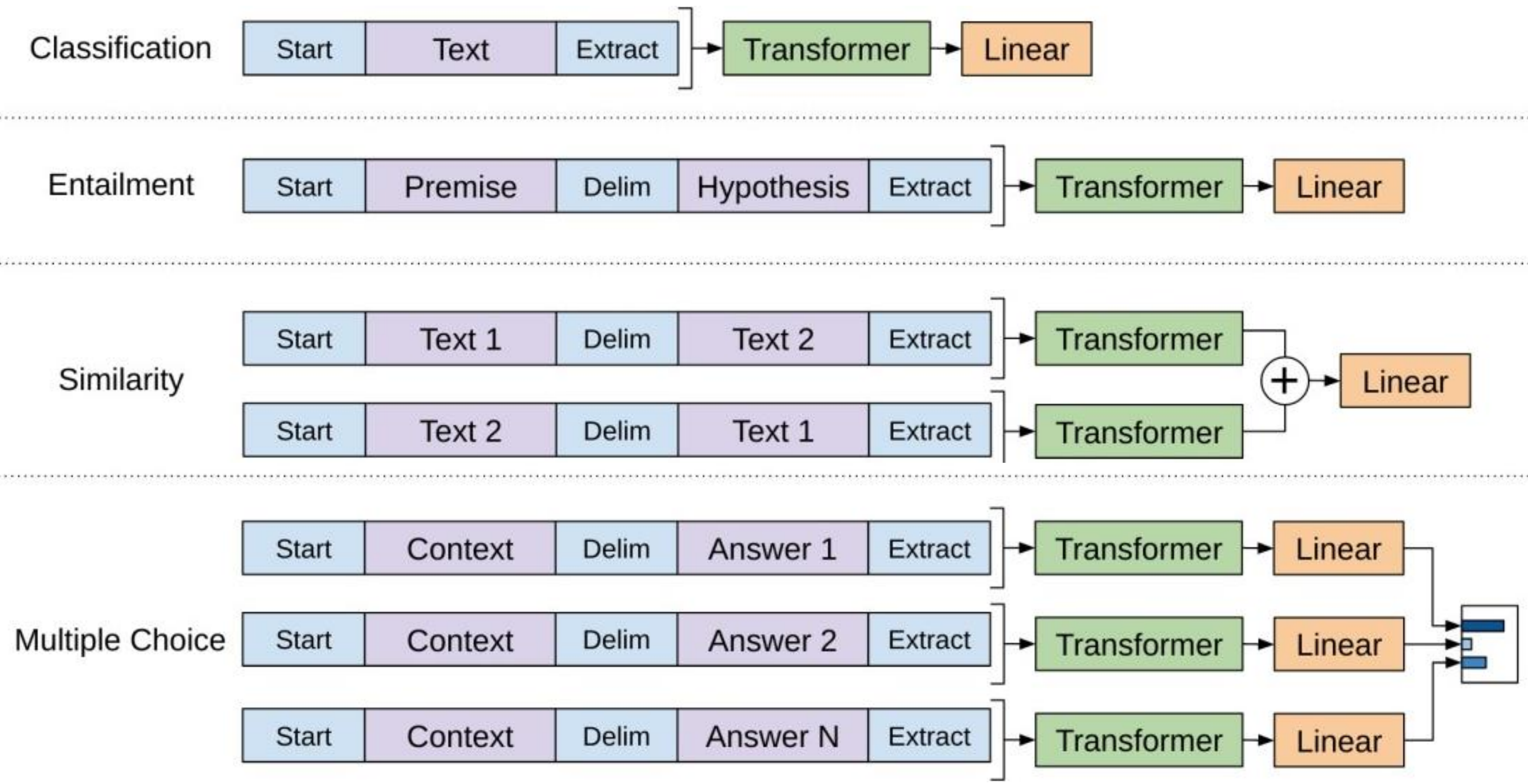
How?

Fine-Tuning BERT

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



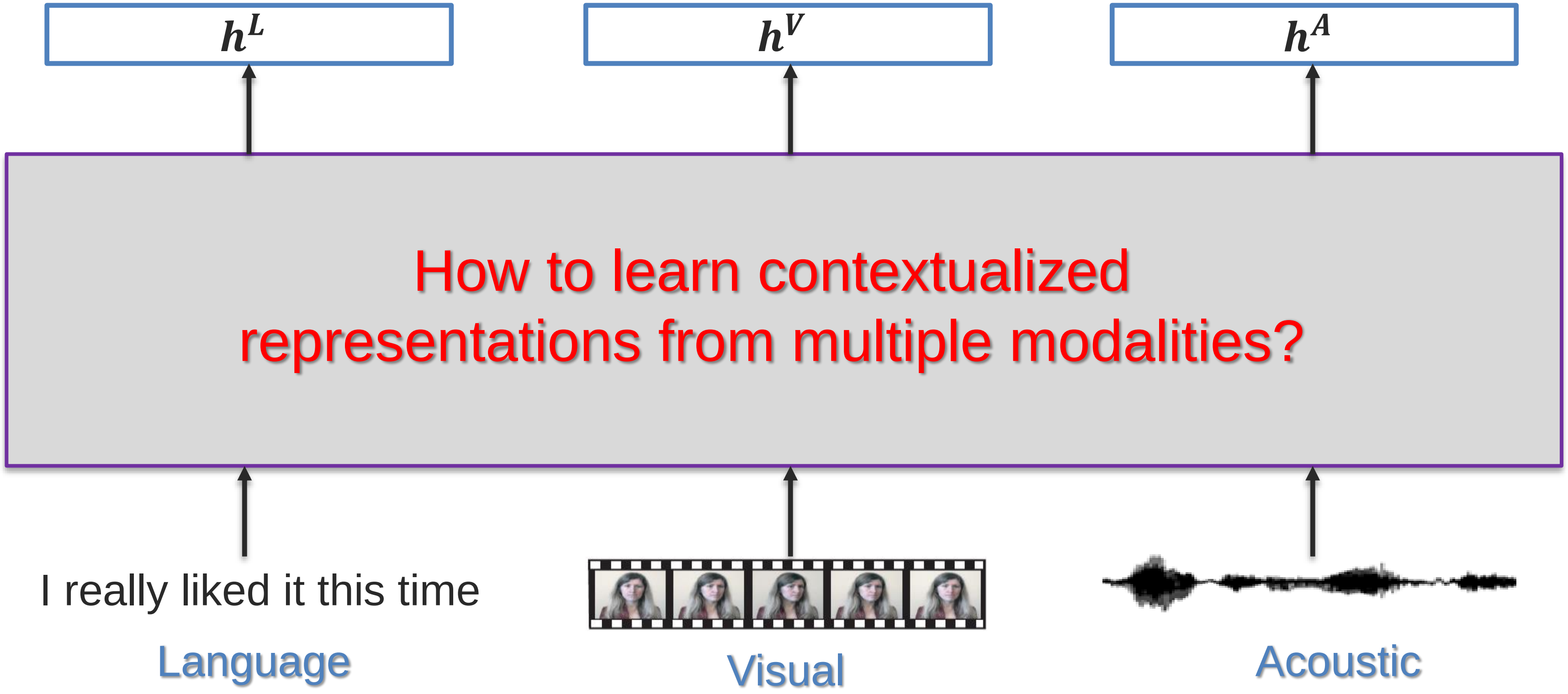
Other Fine-tuning Approaches



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

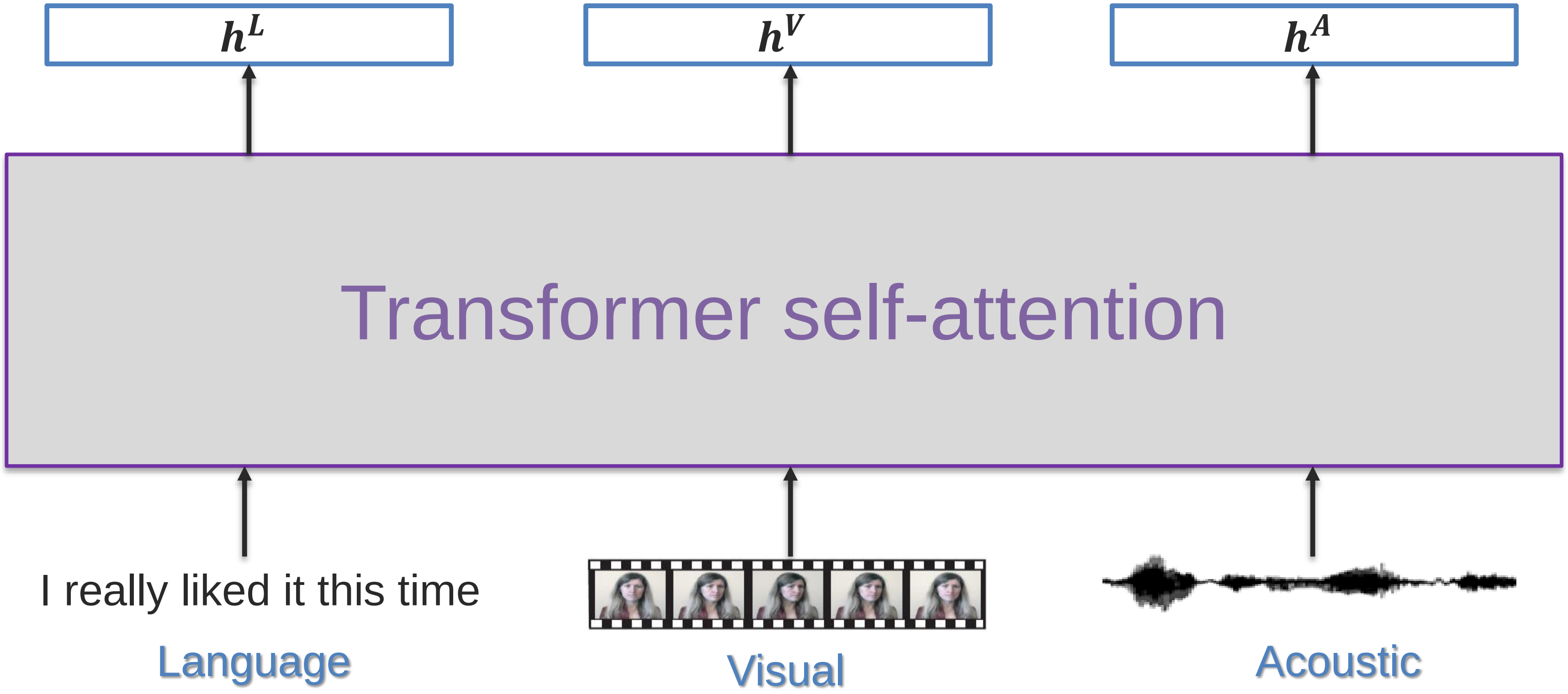
Multimodal Transformers

Multimodal Embeddings



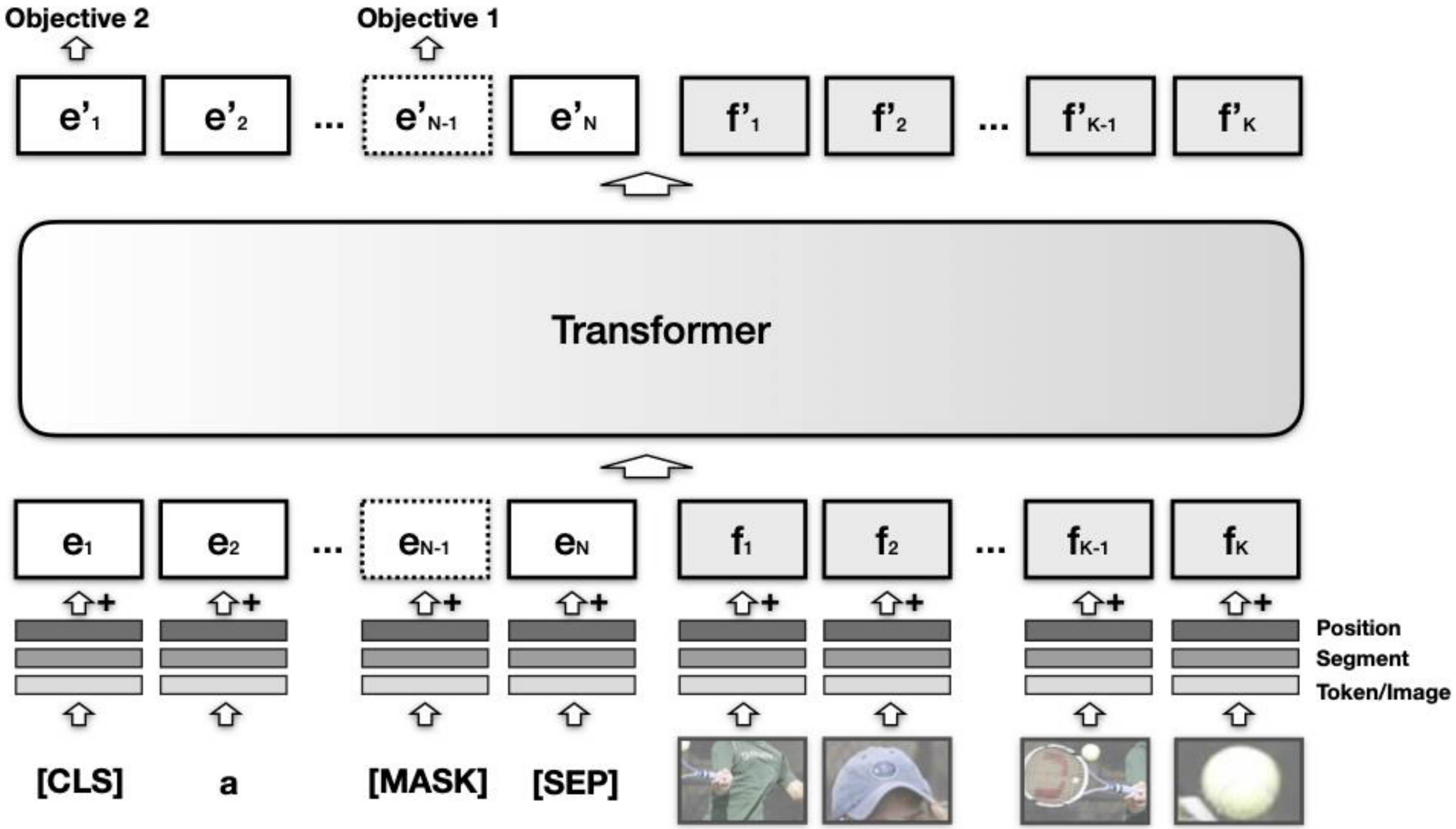
Option 1: Concatenate modalities and learn BERT transformer

Simple Solution: Contextualized Multimodal Embeddings



VisualBERT

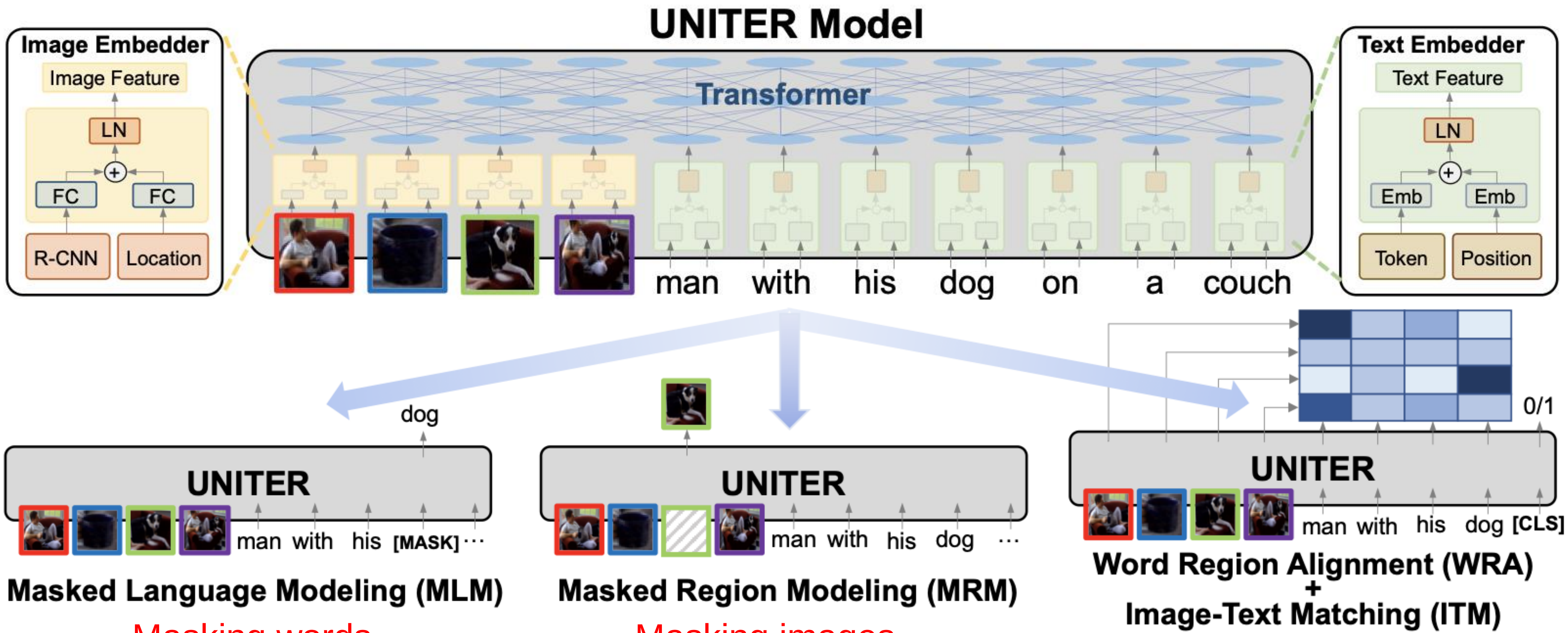
A person hits a ball with a tennis racket



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



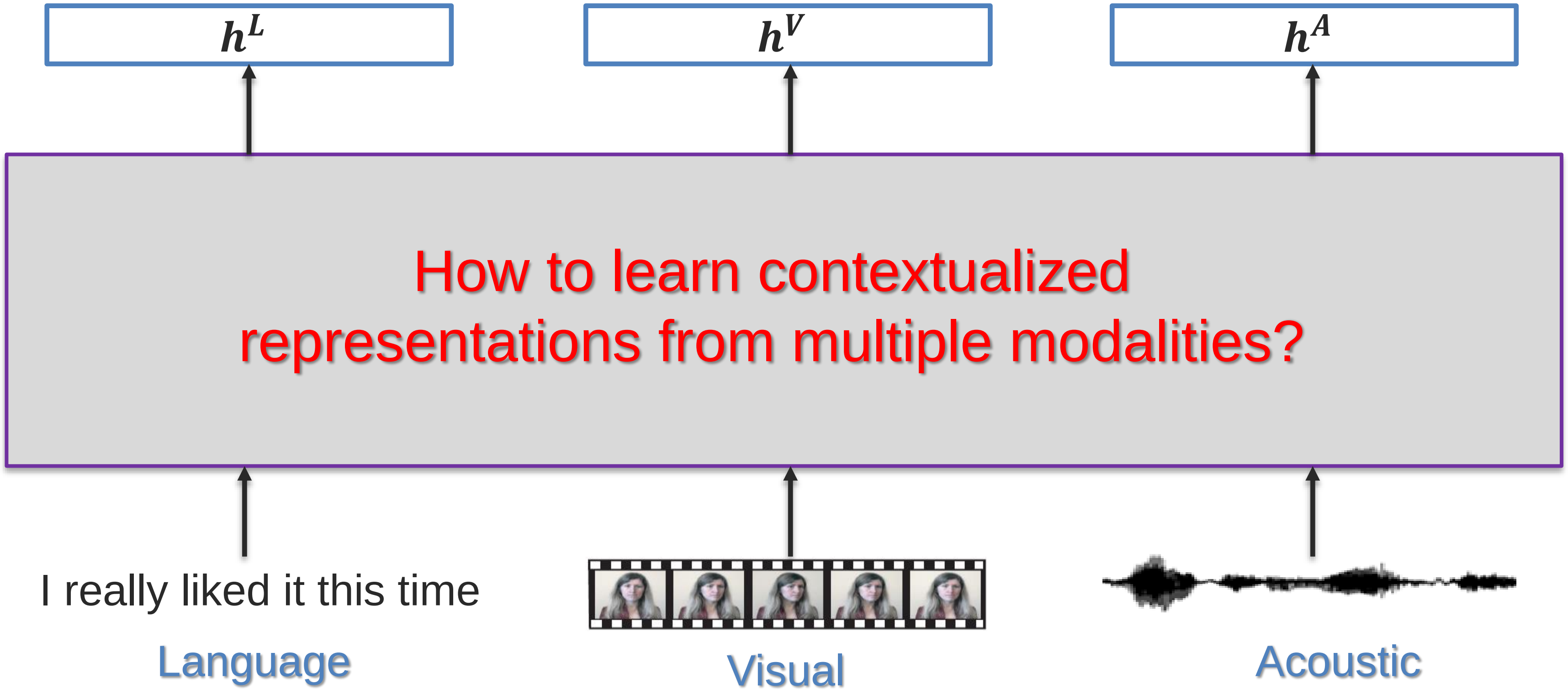
Masking words

Masking images

Is the sentence matches the image?

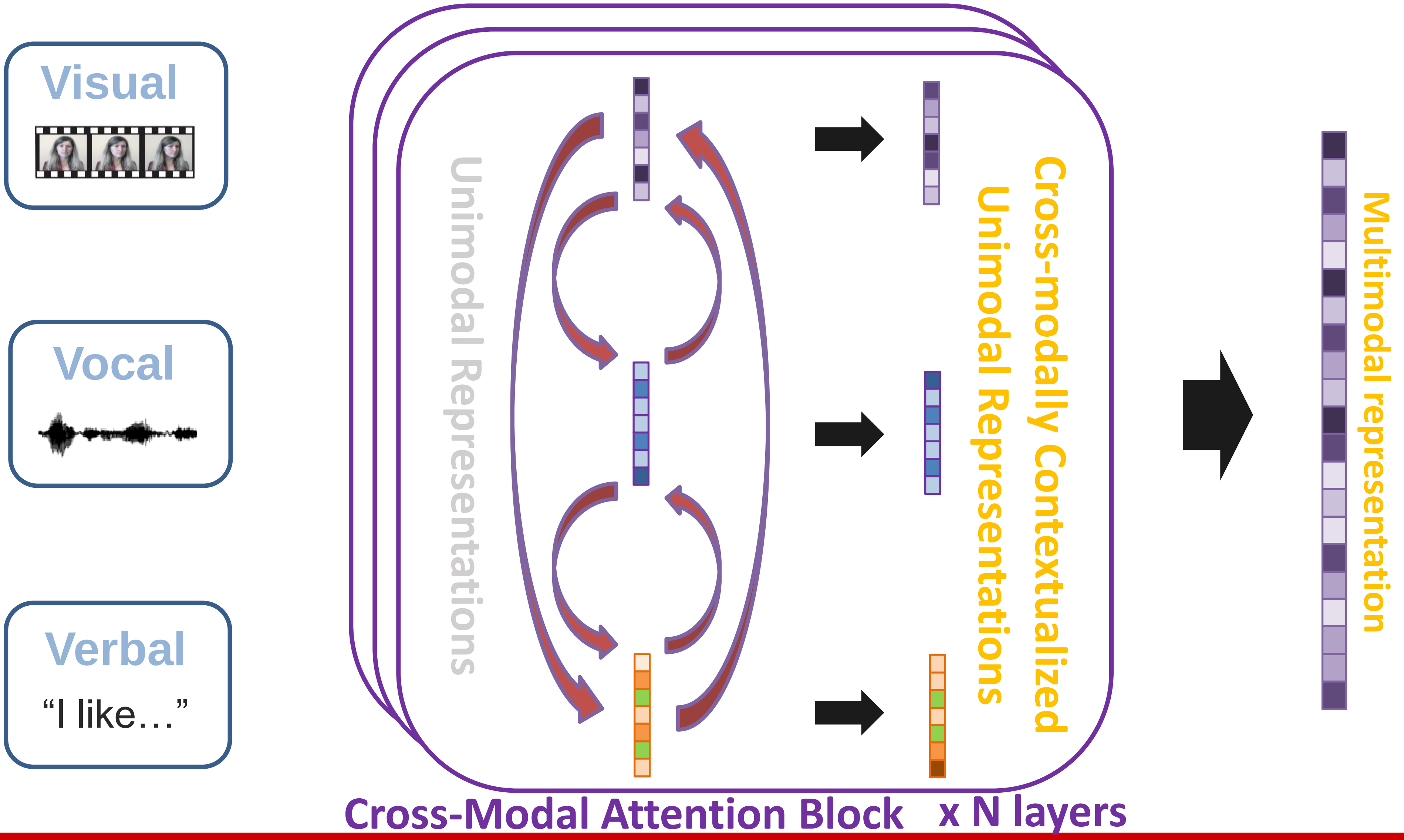
Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." *European conference on computer vision*. 2020.

Multimodal Embeddings

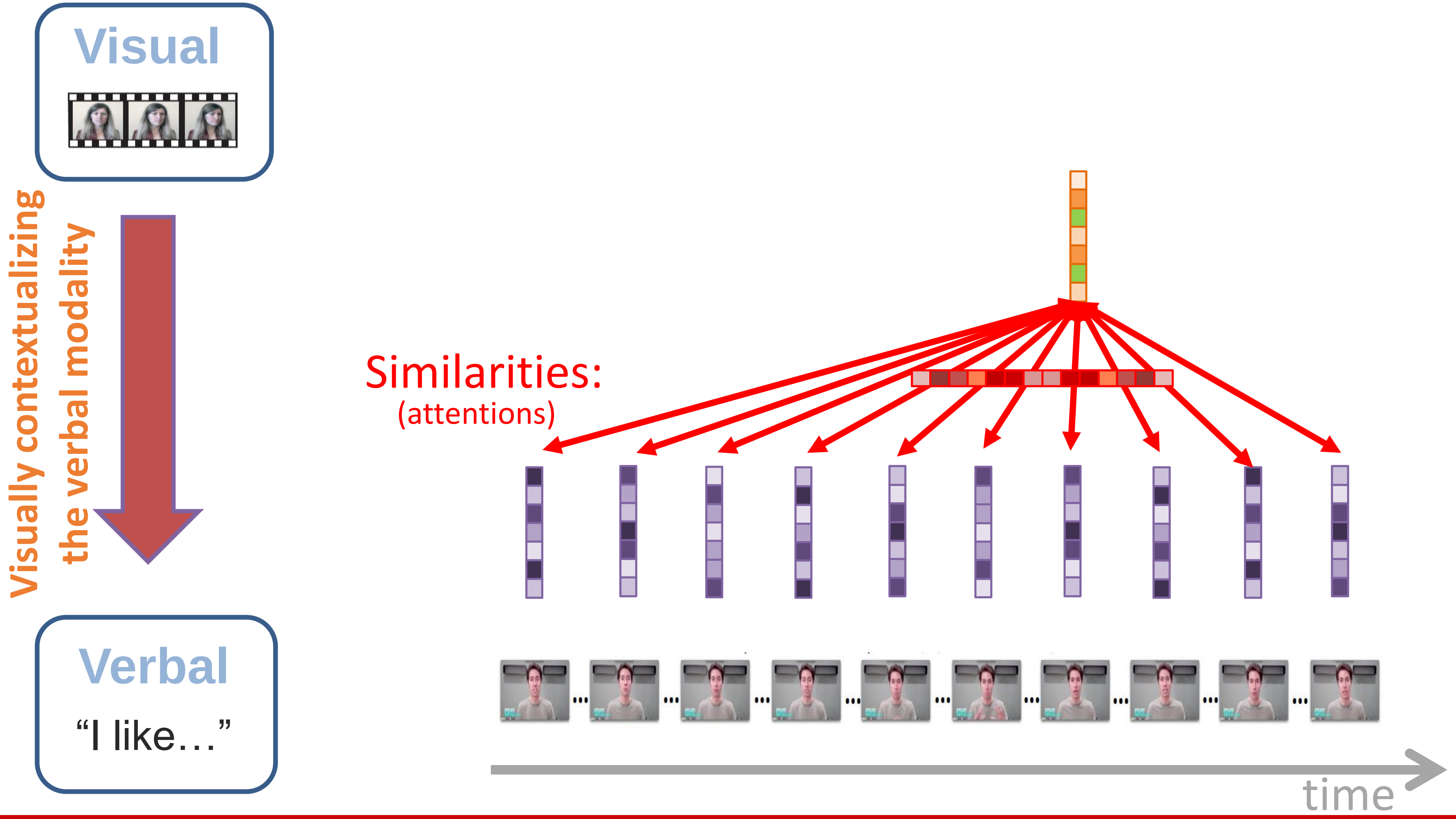


Option 2: Look at pairwise interactions between modalities

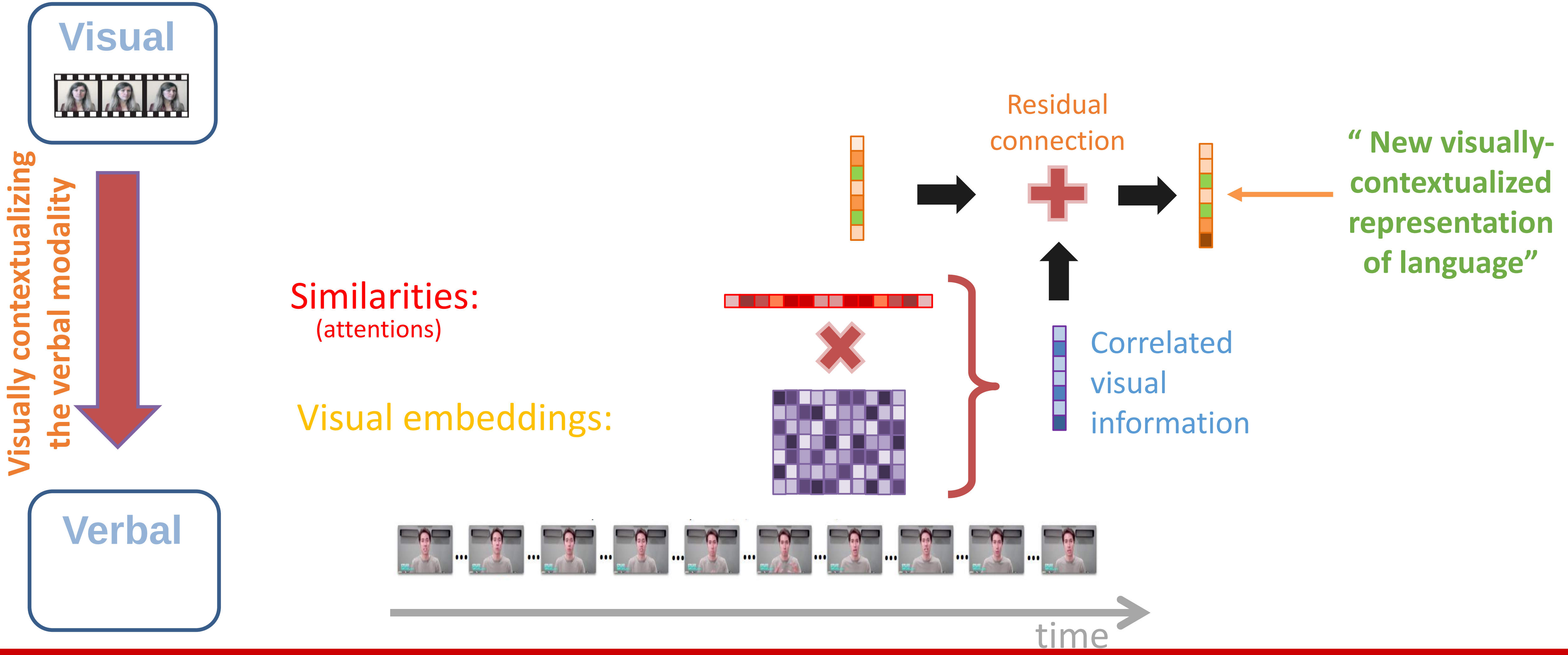
Multimodal Transformer – Pairwise Cross-Modal



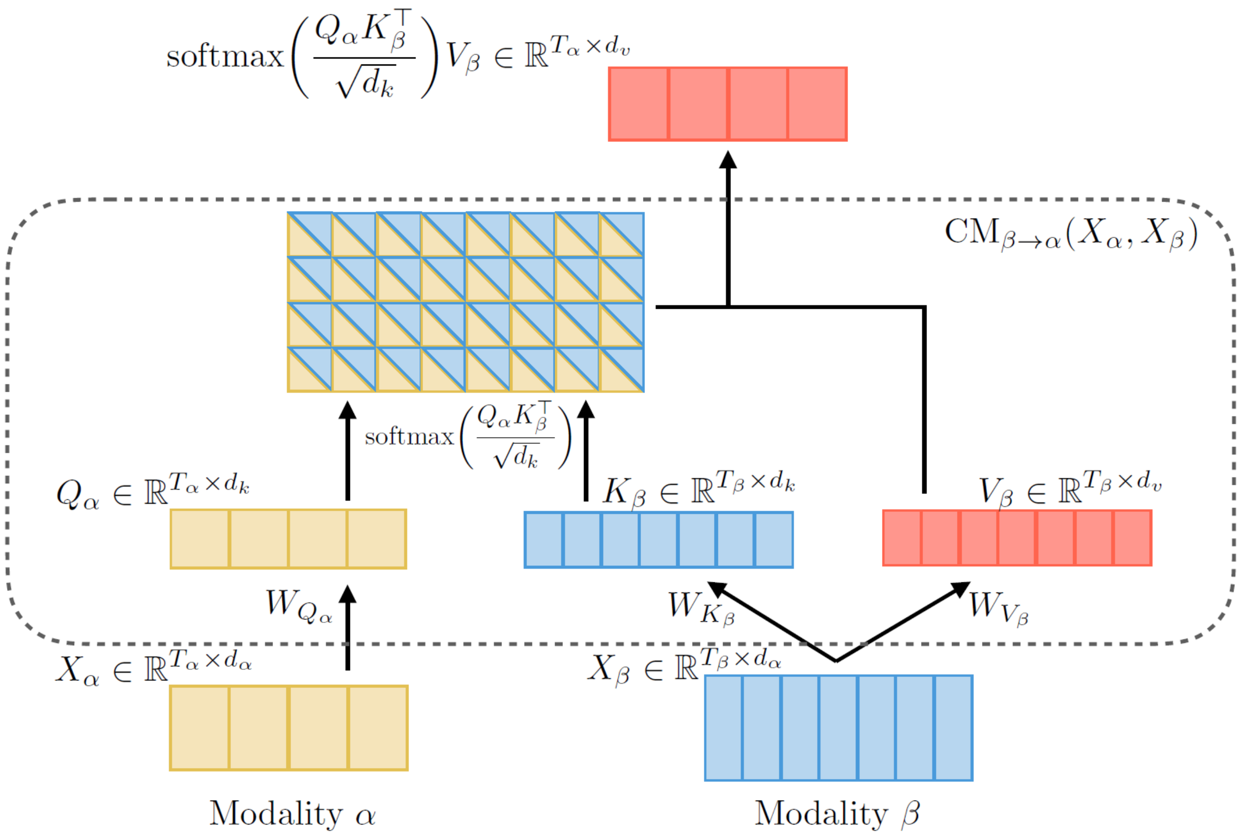
Cross-Modal Transformer Module ($V \rightarrow L$)



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

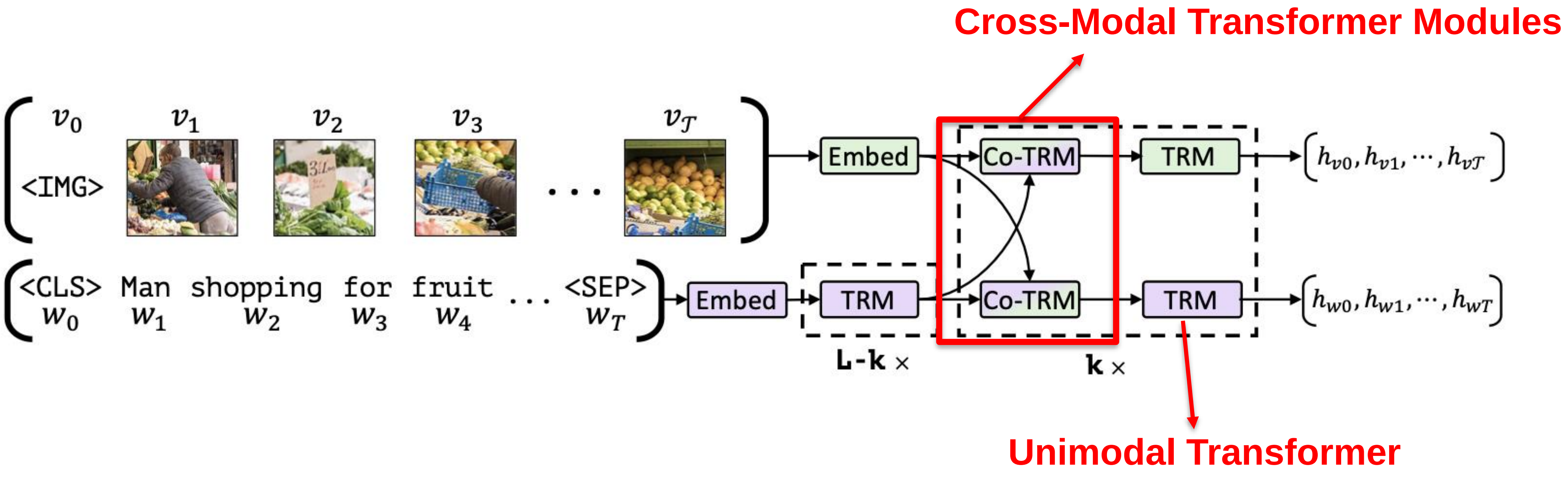


Cross-Modal Transformer Module ($\beta \rightarrow \alpha$)



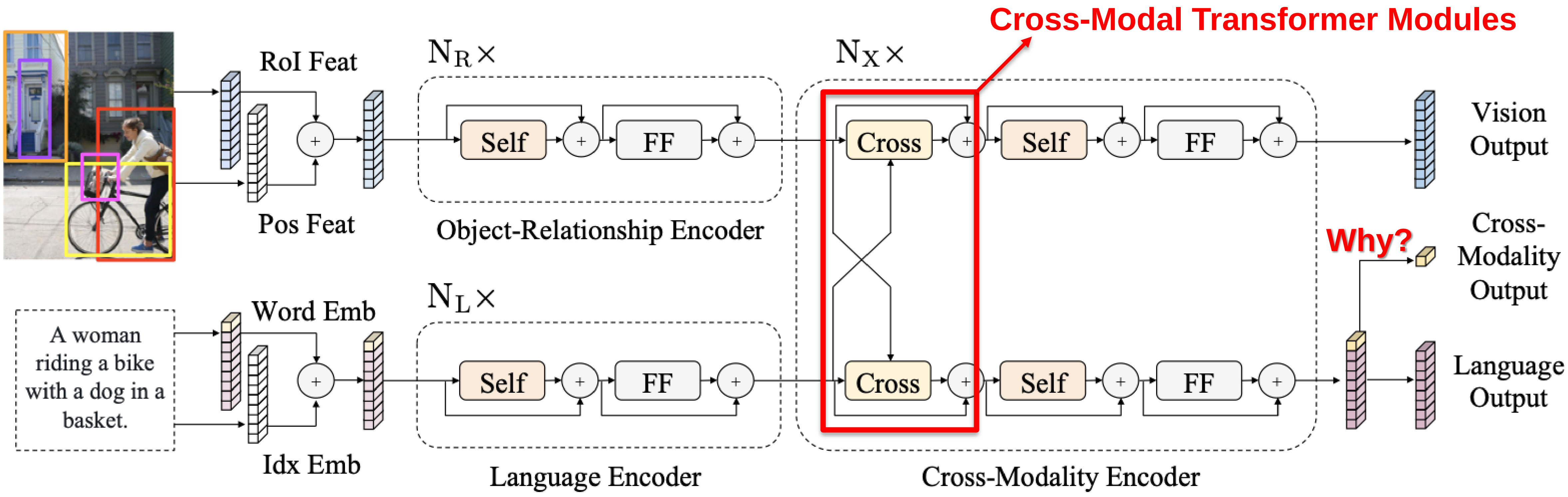
Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

ViLBERT



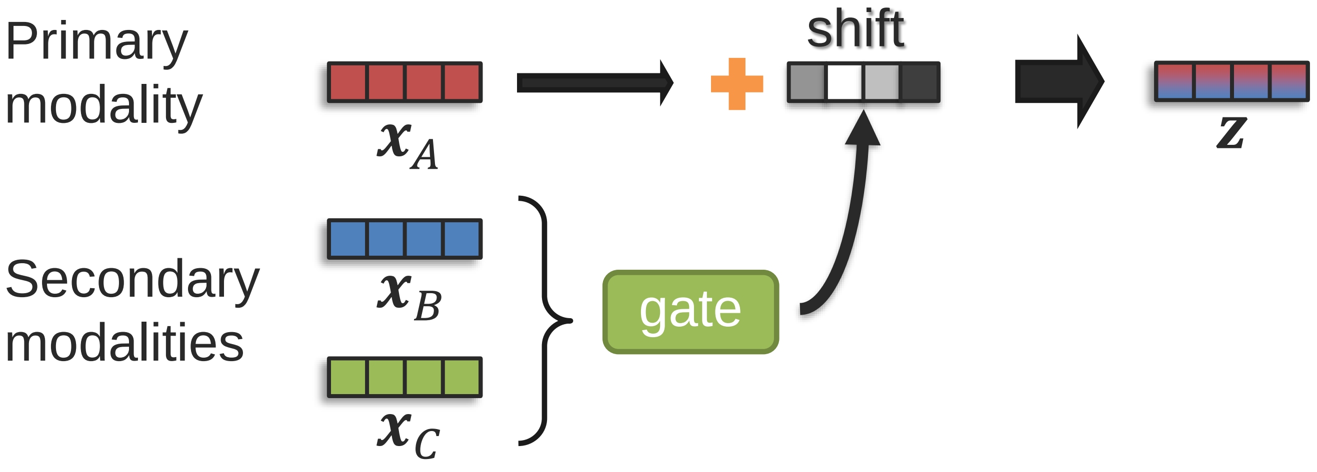
Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *arXiv* (August 6, 2019).

LXMERT



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

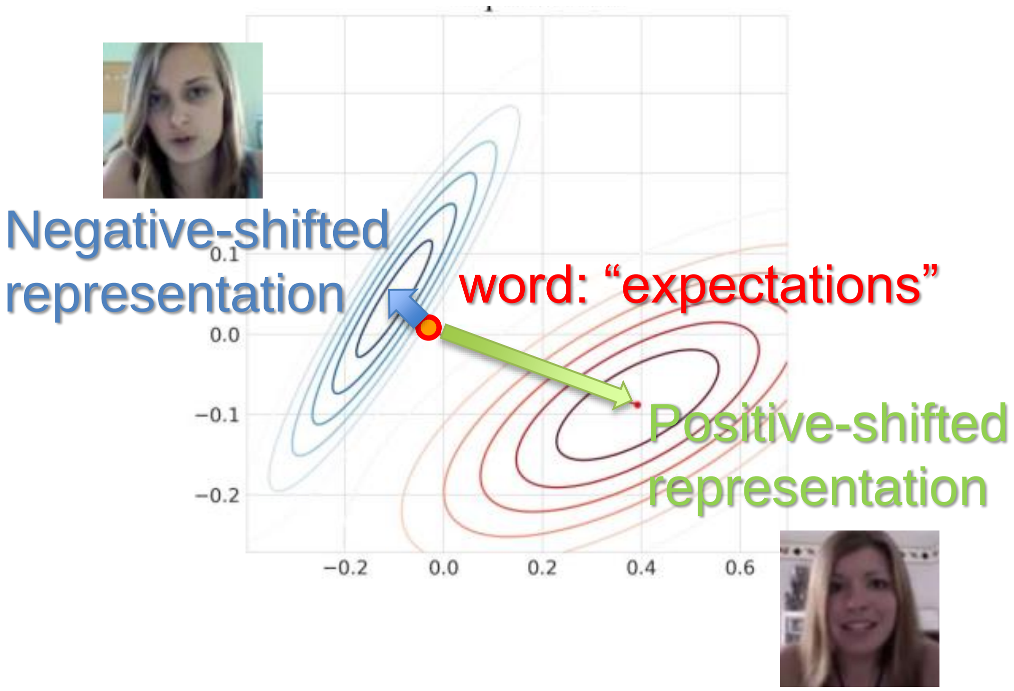
Reminder: Modality-Shifting Fusion



Example with language modality:

Primary modality: language

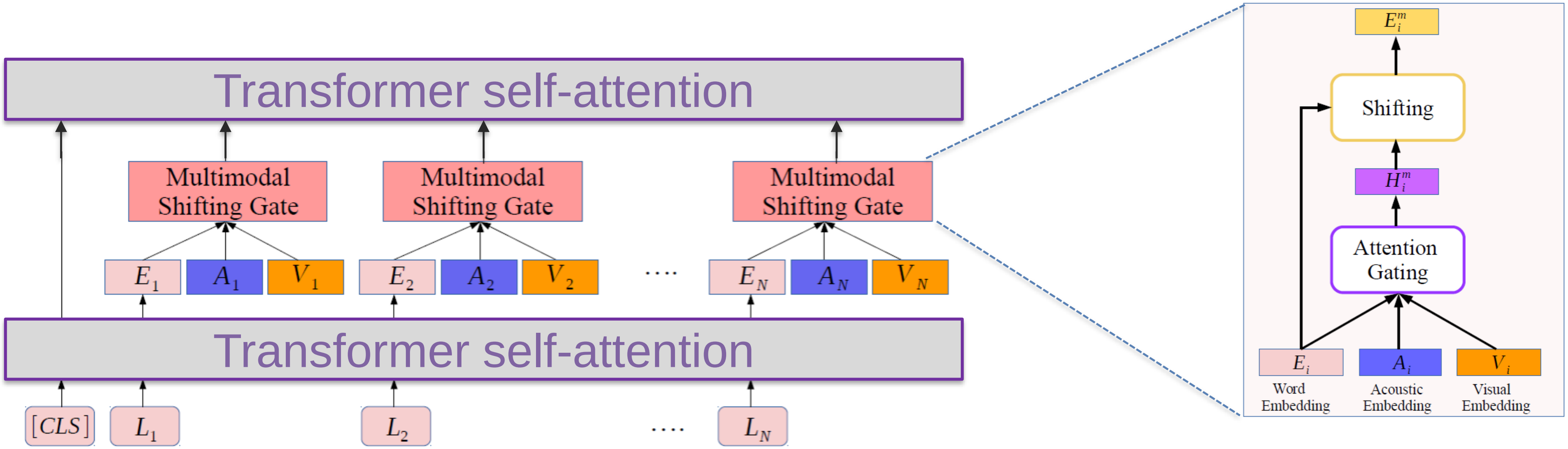
Secondary modalities: acoustic and visual



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

Video-based Representation and Alignment

HowTo100M benchmark dataset

Category	Tasks	Videos	Clips
Food and Entertaining	11504	497k	54.4M
Home and Garden	5068	270k	29.5M
Hobbies and Crafts	4273	251k	29.8M
Cars & Other Vehicles	810	68k	7.8M
Pets and Animals	552	31k	3.5M
Holidays and Traditions	411	27k	3.0M
Personal Care and Style	181	16k	1.6M
Sports and Fitness	205	16k	2.0M
Health	172	15k	1.7M
Education and Communications	239	15k	1.6M
Arts and Entertainment	138	10k	1.2M
Computers and Electronics	58	5k	0.6M
Total	23.6k	1.22M	136.6M

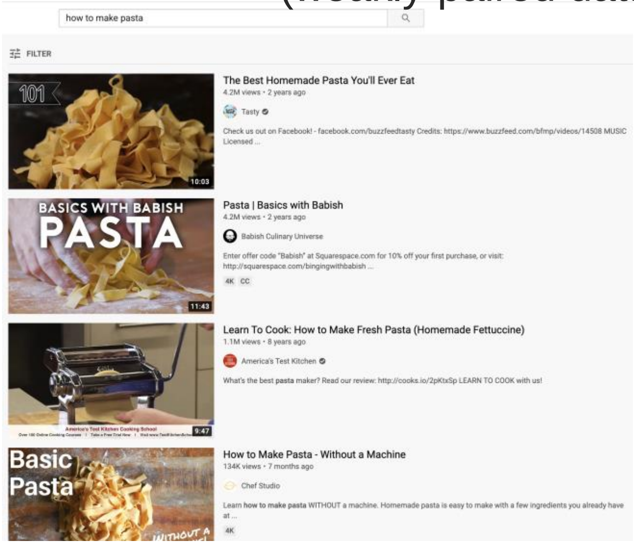
https://www.di.ens.fr/willow/research/howto100m/

Visual Representations from Uncurated Instructional Videos

Goal: Learn better visual representations…

… by taking advantage of large-scale video+language resources

Instructional videos
(weakly-paired data)



it's turning into a much thicker mixture

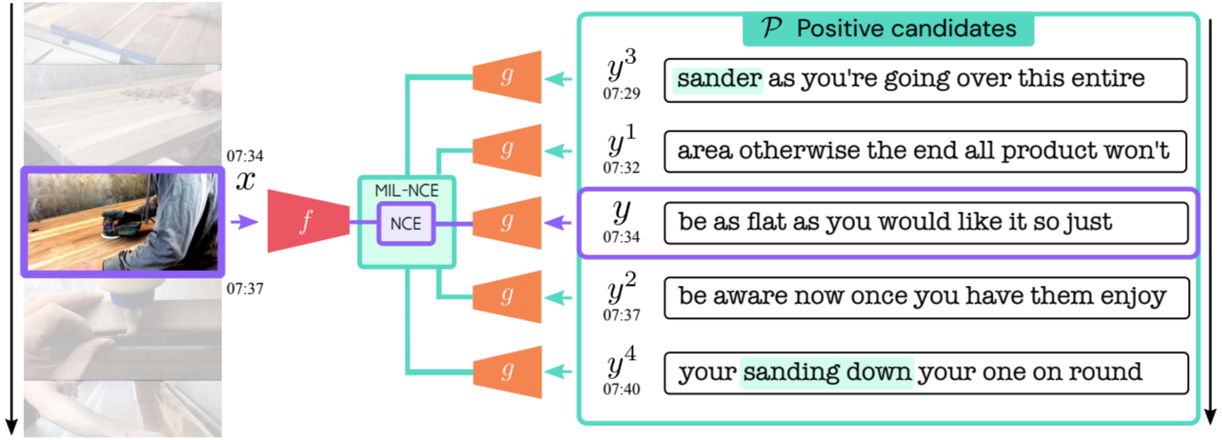
The biggest mistake is not kneading it enough

…

End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Weakly Paired Data

Data point: "a short 3.2 seconds video clip (32 frames at 10 FPS) together with a small number of words (not exceeding 16)"

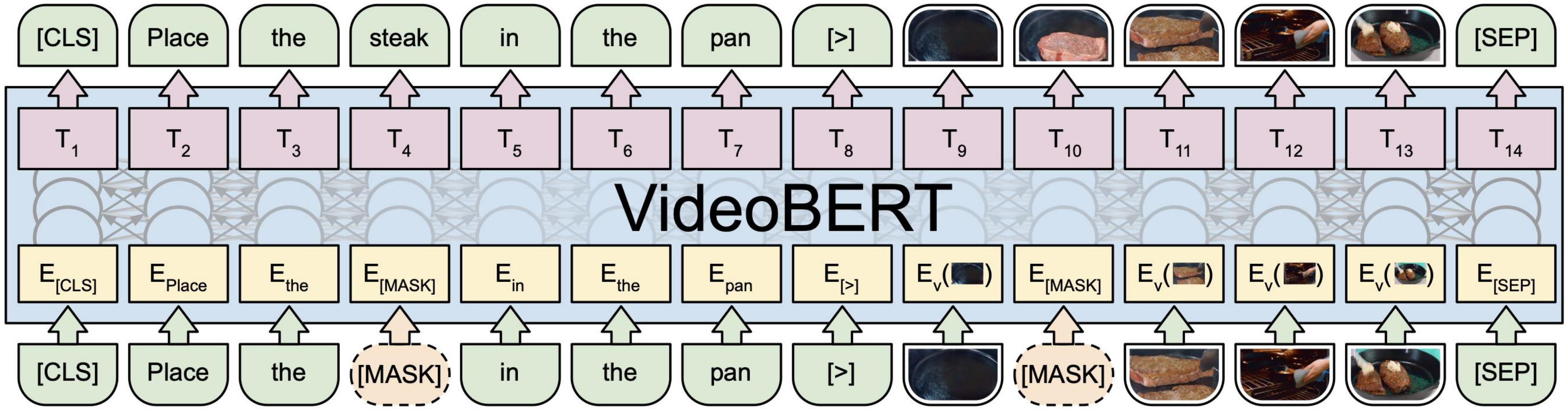


How to handle this misalignment? — Multi-instance learning!

How to do it self-supervised? — Contrastive learning!

End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Another Approach for Weakly-Paired Video Data

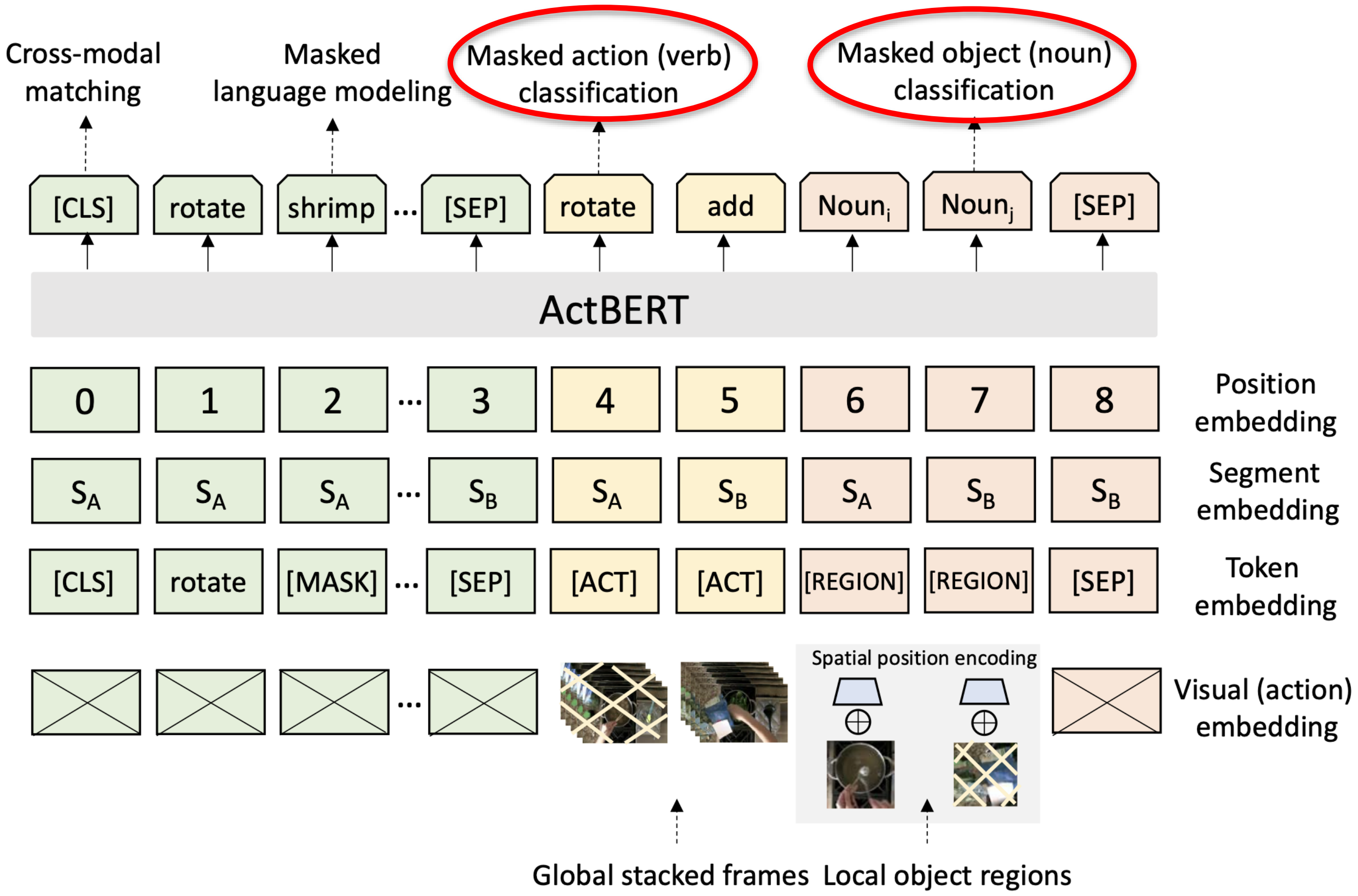


How do we get visual words now?

K-mean clustering + centroid

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid; VideoBERT: A Joint Model for Video and Language Representation Learning ICCV, 2019

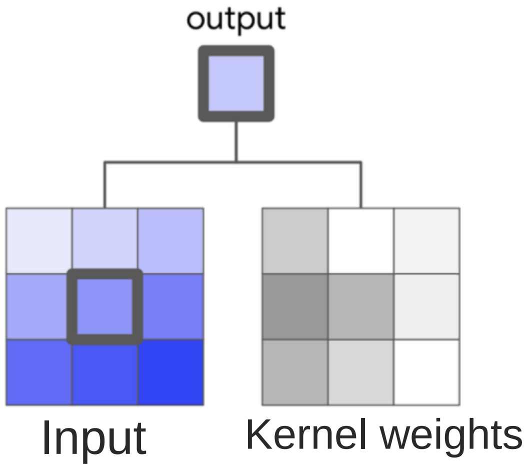
ActBERT



Zhu and Yang, ActBERT: Learning Global-Local Video-Text Representations, CVPR 2020

Going Beyond CNNs…
Vision Transformers (and more!)

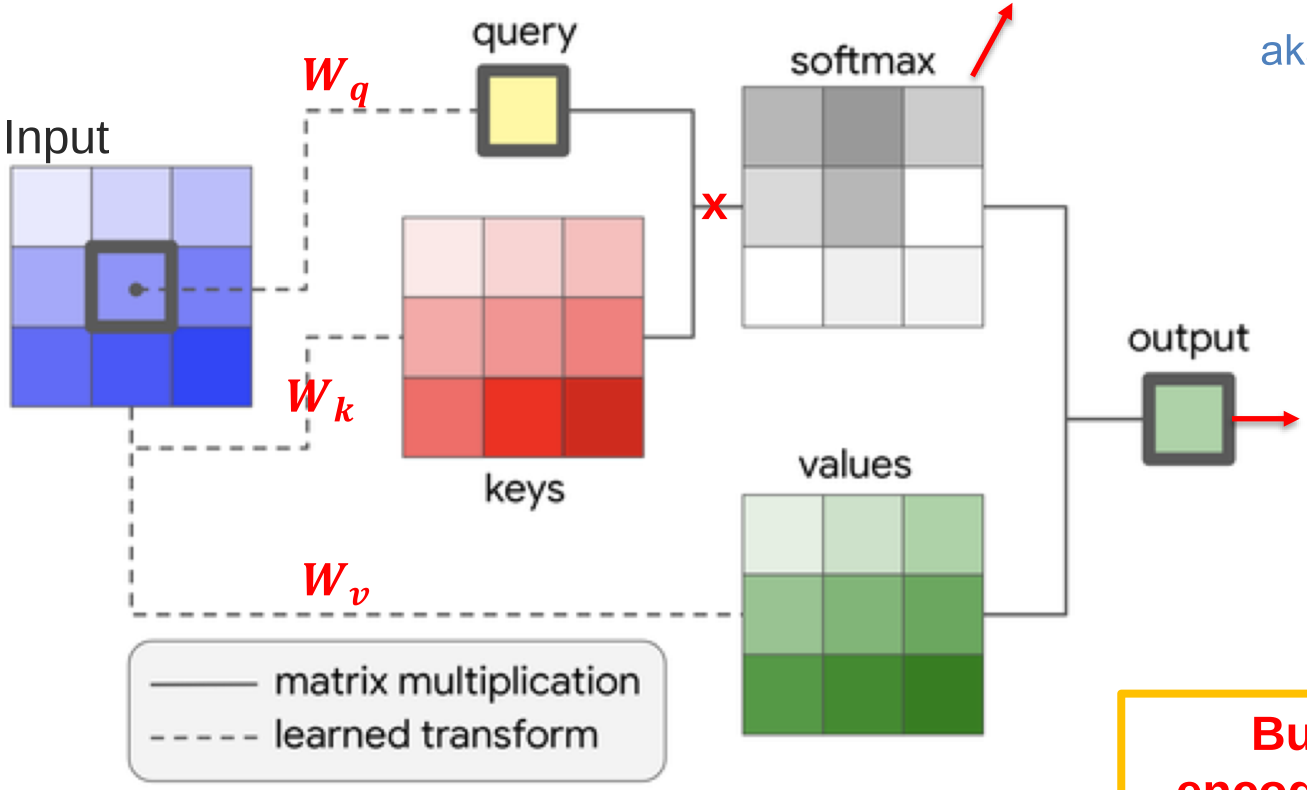
Replacing a CNN w/ Self-Attention



Convolution

How well the query matches the keys?
(How well the pixel matches its neighbors?)

aka “memory block”



Pixel contextualized by its neighbors

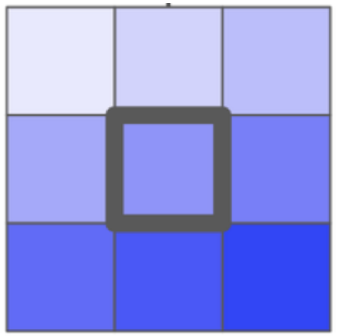
But how to encode position?

Self-attention

https://arxiv.org/abs/1906.05909

Replacing a CNN w/ Self-Attention

Image patch



2D relative position embedding

-1, -1	-1, 0	-1, 1	-1, 2
0, -1	0, 0	0, 1	0, 2
1, -1	1, 0	1, 1	1, 2
2, -1	2, 0	2, 1	2, 2

Position embedding is added to the key:

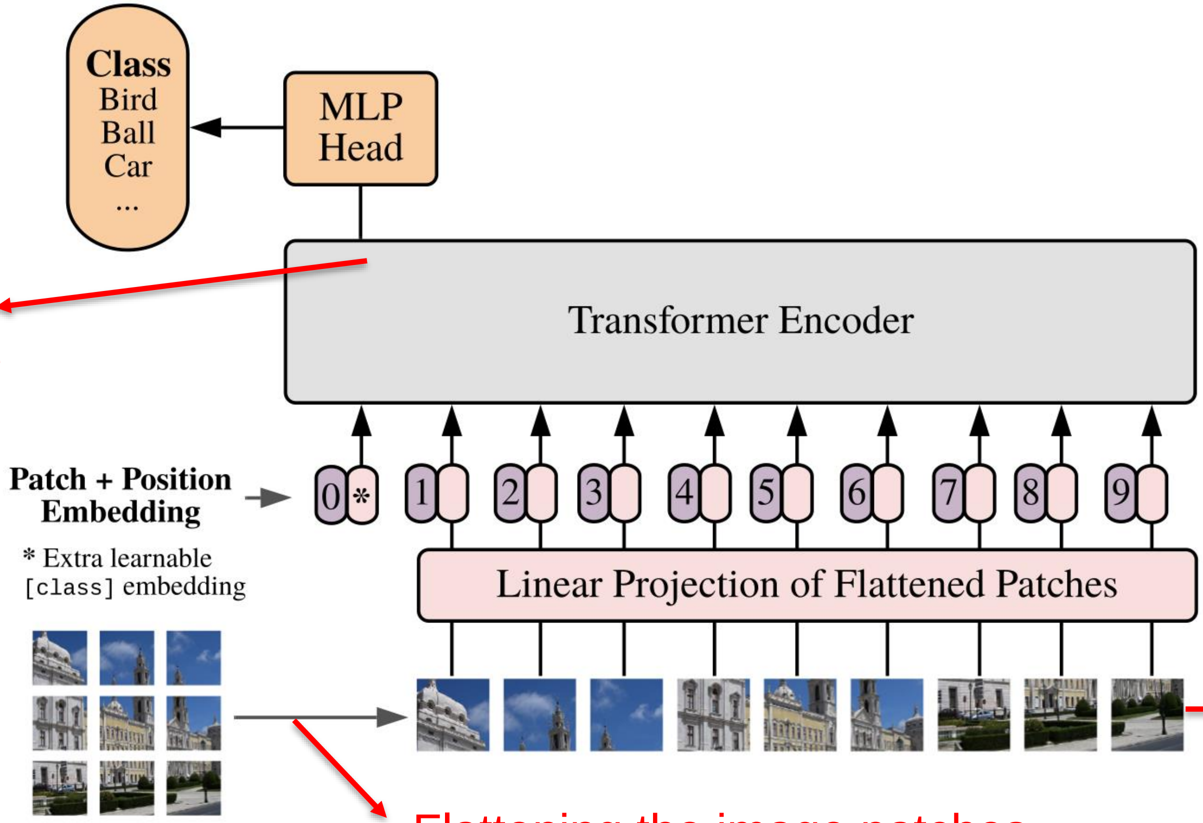
$$y_{ij} = \sum_{a,b \in \mathcal{N}_k(i,j)} \texttt{softmax}_{ab} \left(q_{ij}^\top k_{ab} + q_{ij}^\top r_{a-i,b-j} \right) v_{ab}$$

https://arxiv.org/abs/1906.05909

Vision Transformer (ViT)

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

Vision Transformer (ViT)



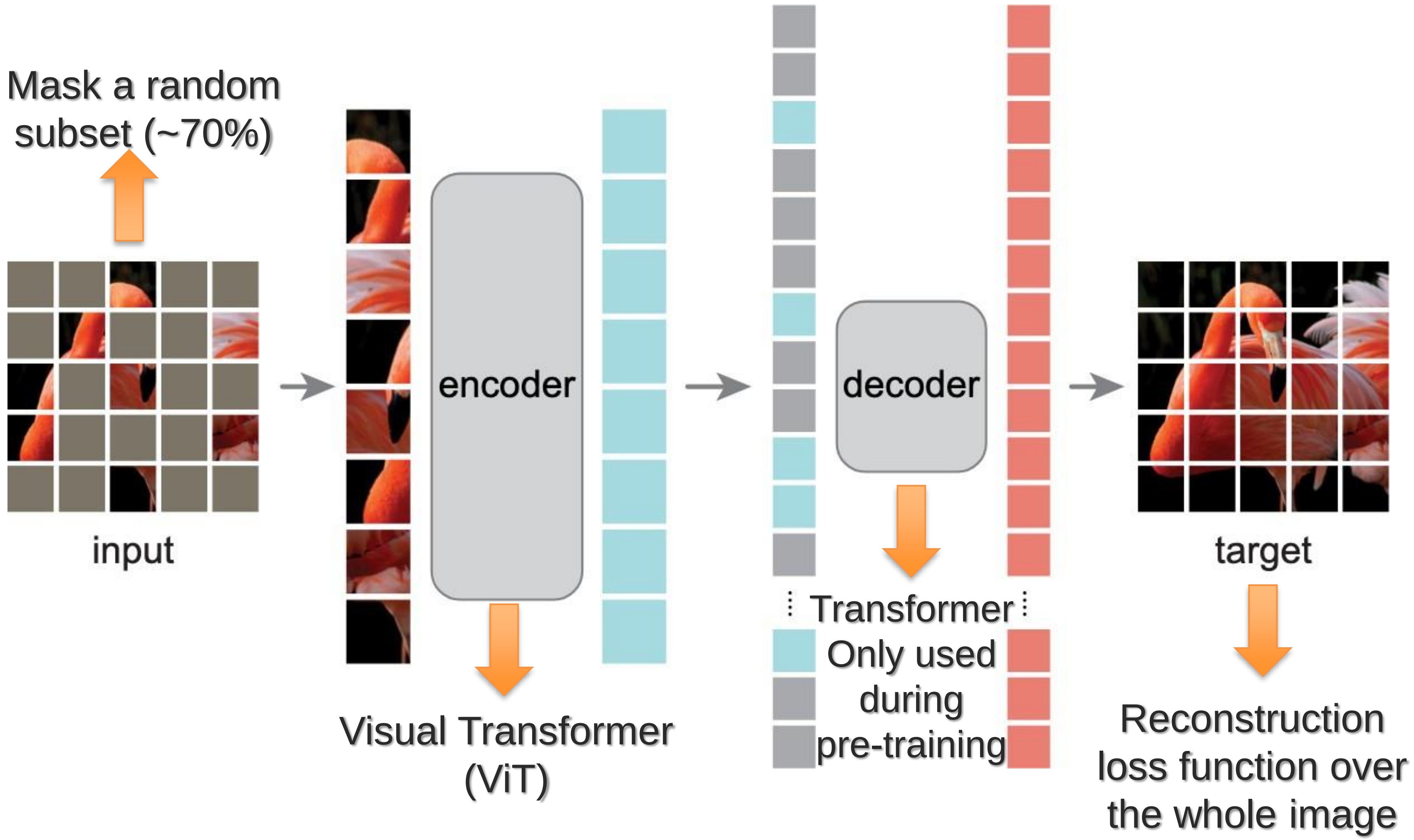
Embedding for the whole image

16x16 image patches

Flattening the image patches

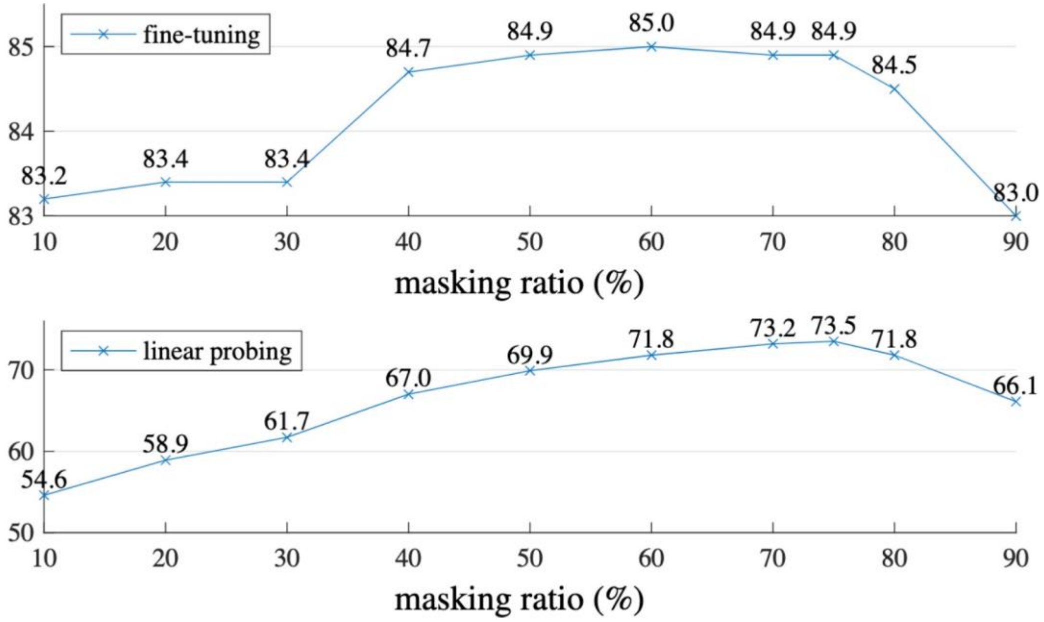
Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

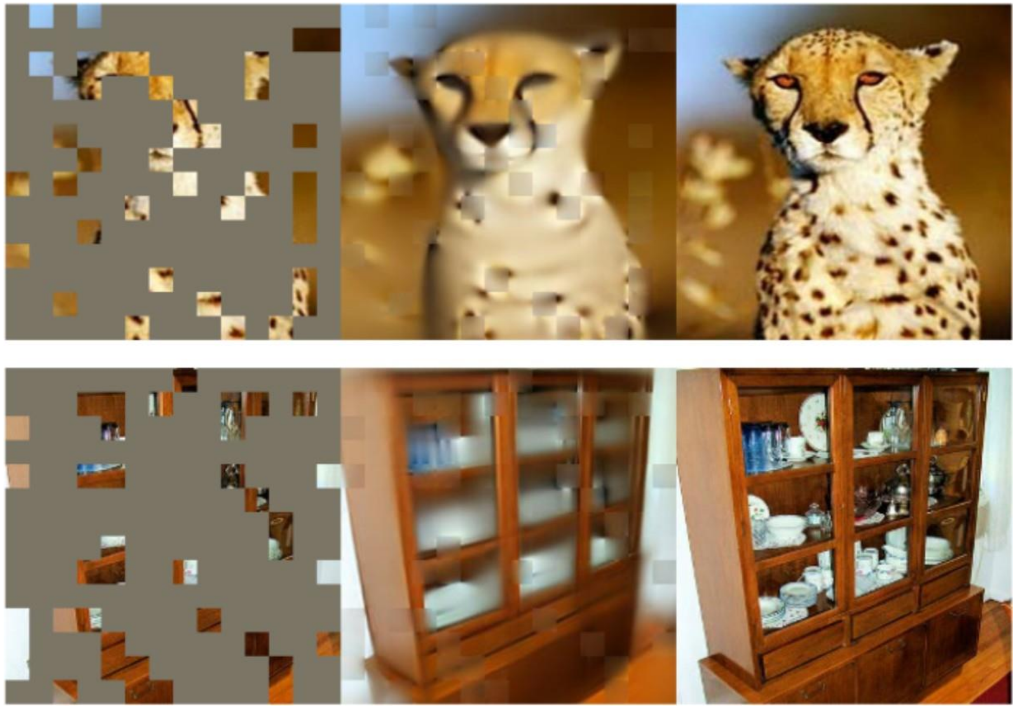
Masked Auto-Encoder (MAE)



He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

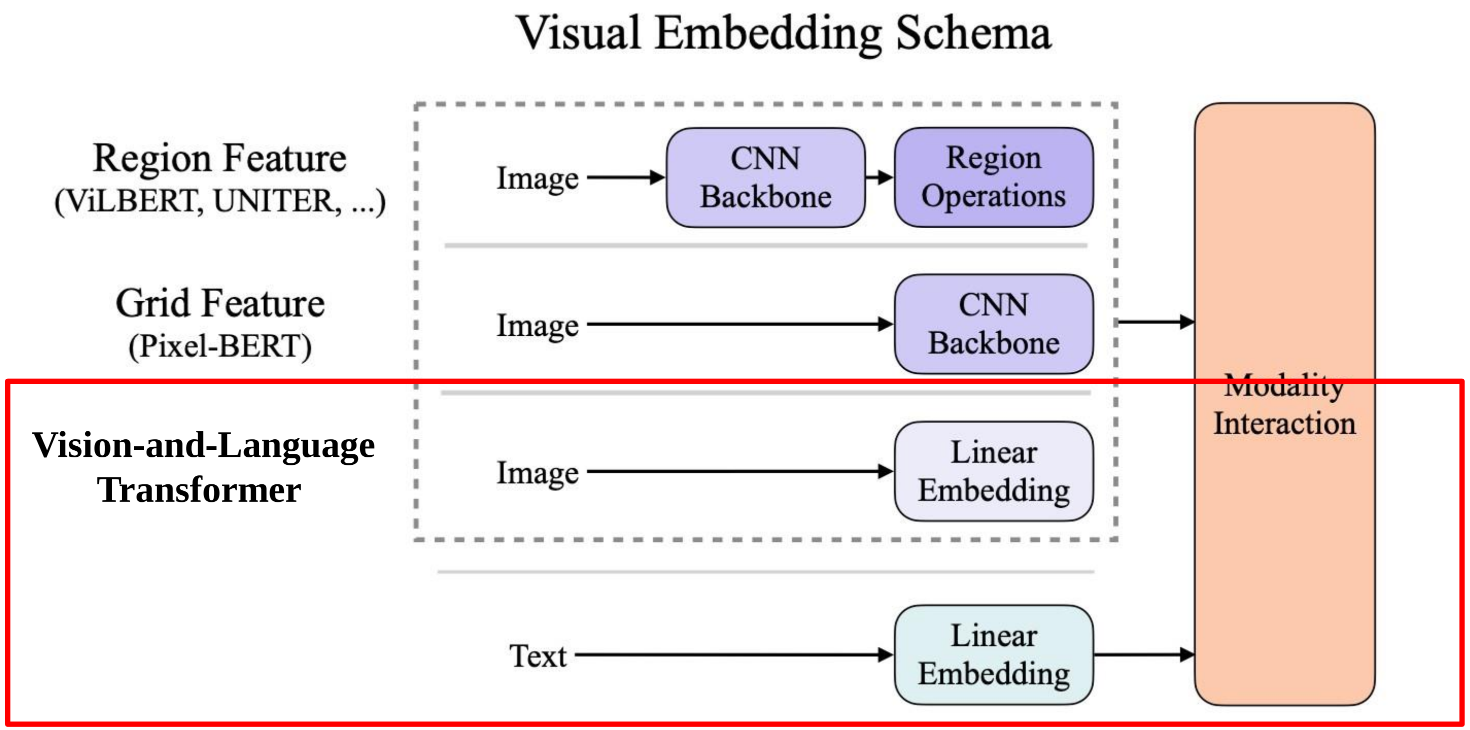
Masked Auto-Encoder (MAE)





He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

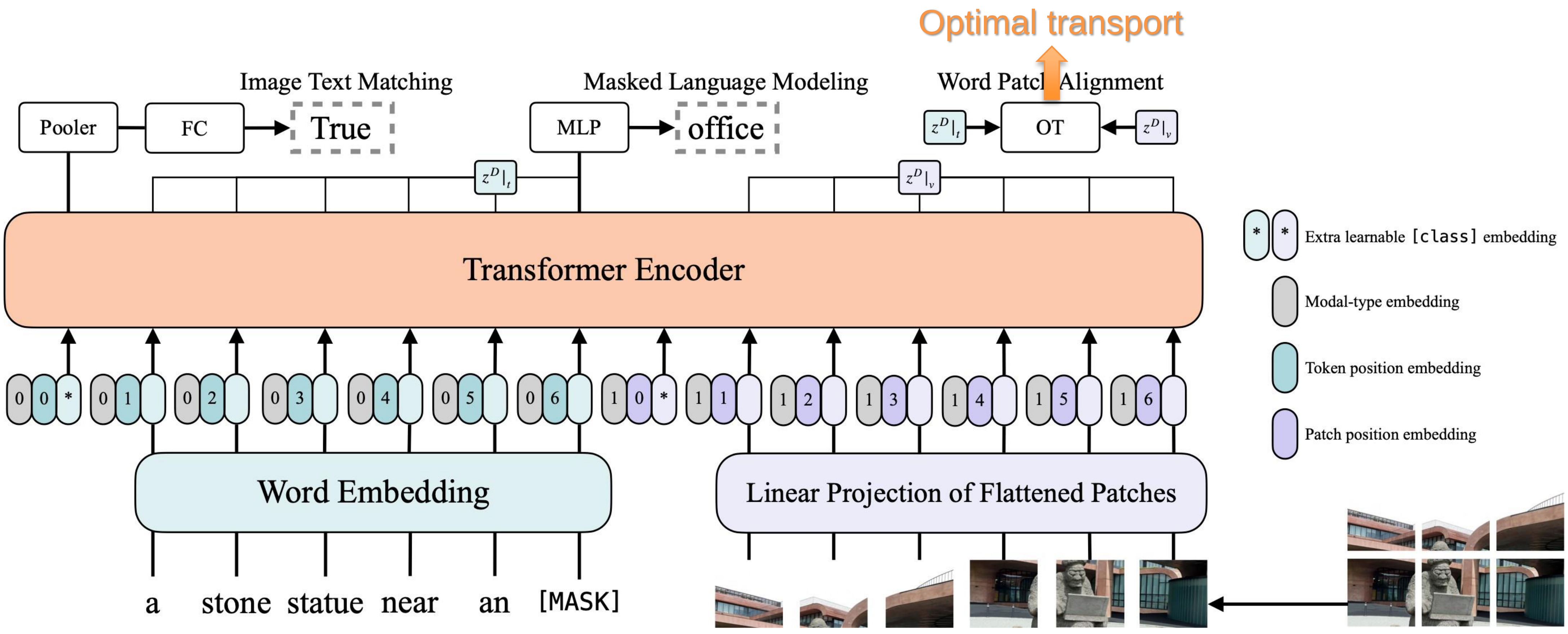
Visual Transformers for Multimodal Learning



(≈ BERT + ViT)

https://arxiv.org/abs/2102.03334

Visual-and-Language Transformer (ViLT) (≈ BERT + ViT)



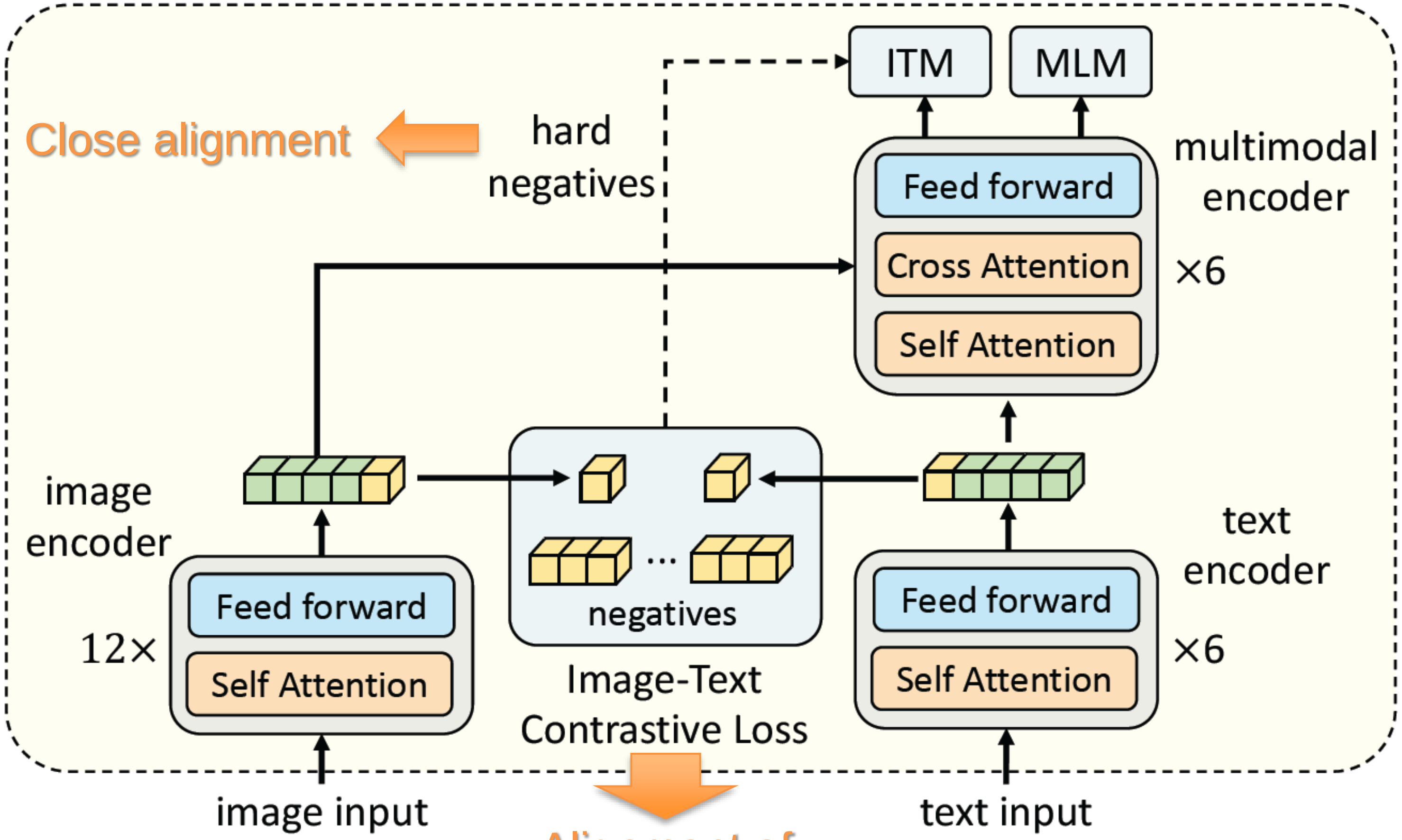
https://arxiv.org/abs/2102.03334

Visual-and-Language Transformer (ViLT)

Example of alignment between modalities:

https://arxiv.org/abs/2102.03334

ALBEF: Align Before Fusion (≈ BERT + ViT + CLIP-ish)



Alignment of segment-level embeddings → Related to mutual information

Li et al., Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, Neurips 2021