



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 6.2: Alignment and Representation

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Spring 2021 edition taught by Yonatan Bisk*

Administrative Stuff

Second Project Assignment (Due Sunday 10/10)

Main goals:

- Get familiar with unimodal representations
 - Learn about tools based on CNNs, word2vec, BERT, ...
- Understand the structure in your unimodal data
 - Perform some visualization of the unimodal data
- Explore qualitatively the unimodal data
 - How does it relate to your labels? Look at specific examples

Examples of unimodal 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



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

Lecture 6.2: Alignment and Representation

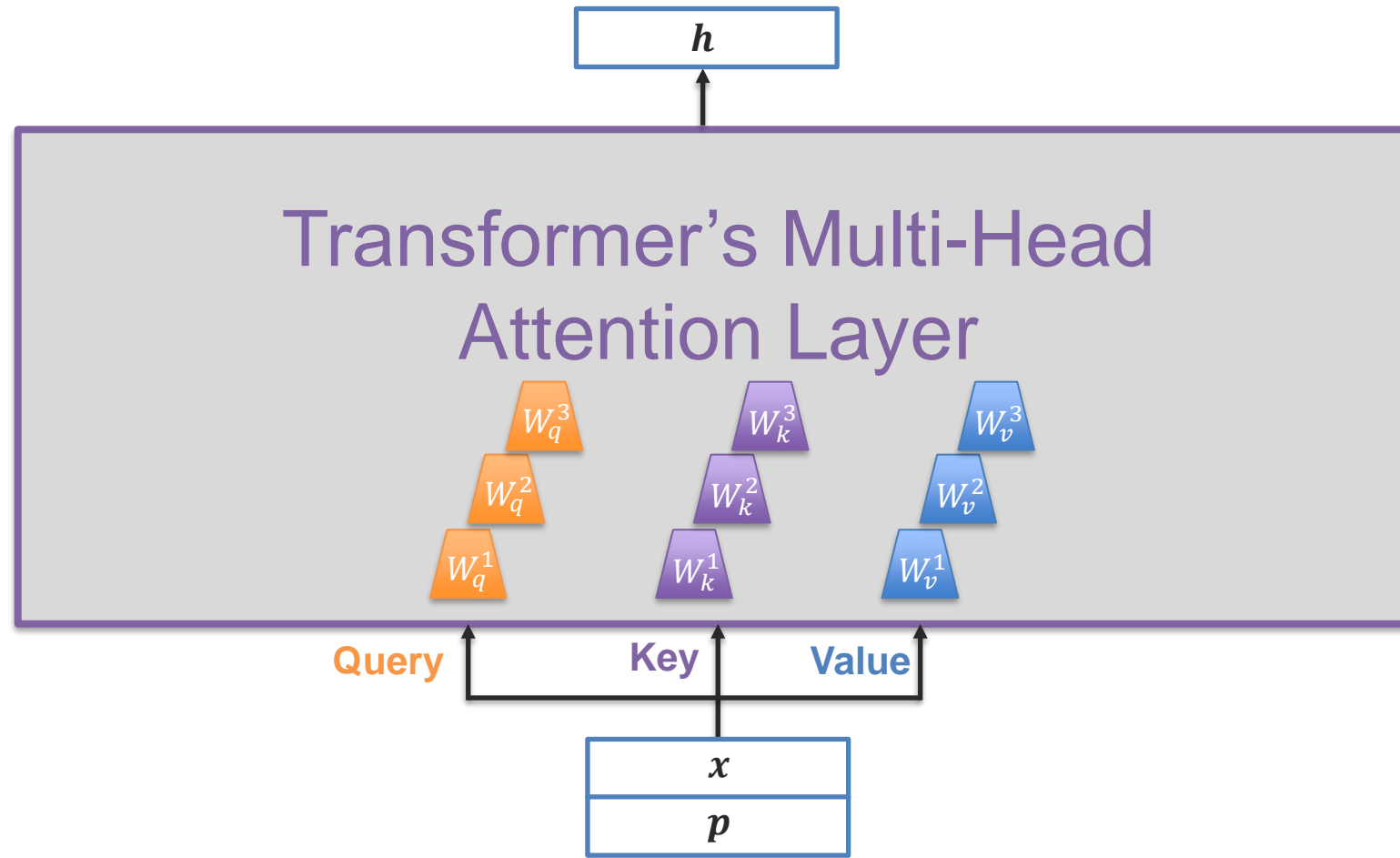
Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.
Spring 2021 edition taught by Yonatan Bisk*

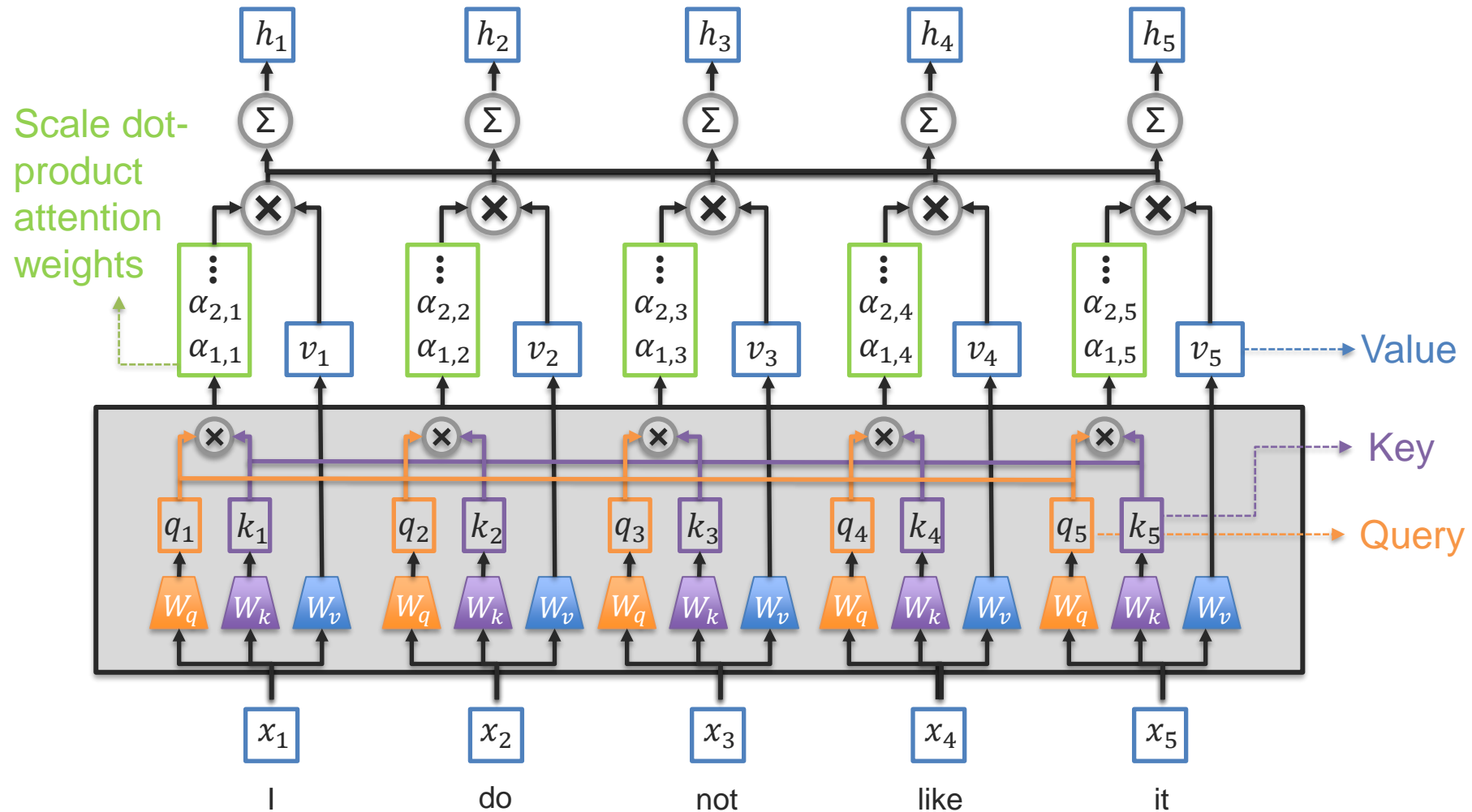
Objectives of today's class

- Transformer pre-training
 - BERT: Bidirectional Encoder Representations from Transformers
- Multimodal transformers (Image and language)
 - Concatenated transformers (VisualBERT, Uniter)
 - Crossmodal transformers (ViLBERT, LXMERT)
- Video and language transformers
 - VideoBERT, ActBERT
- Visual transformers
 - Vision transformer, CLIP
- Graph representations
 - Graph neural networks

Transformer Multi-Head Attention Model



Transformer Self-Attention

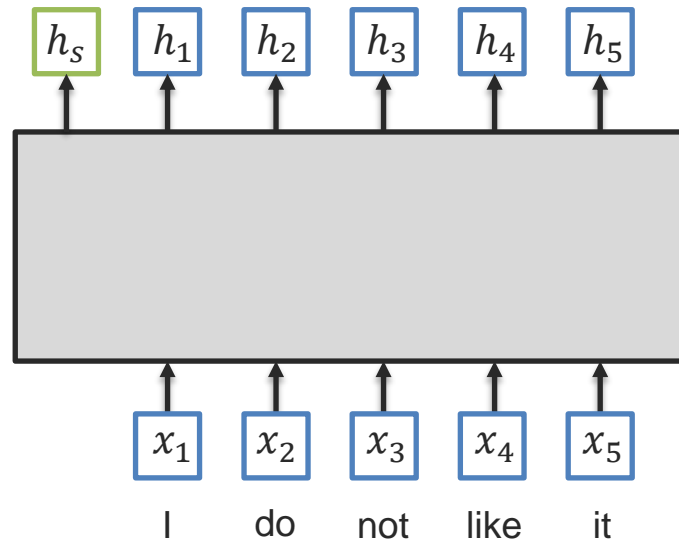


Pre-training Language Models

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

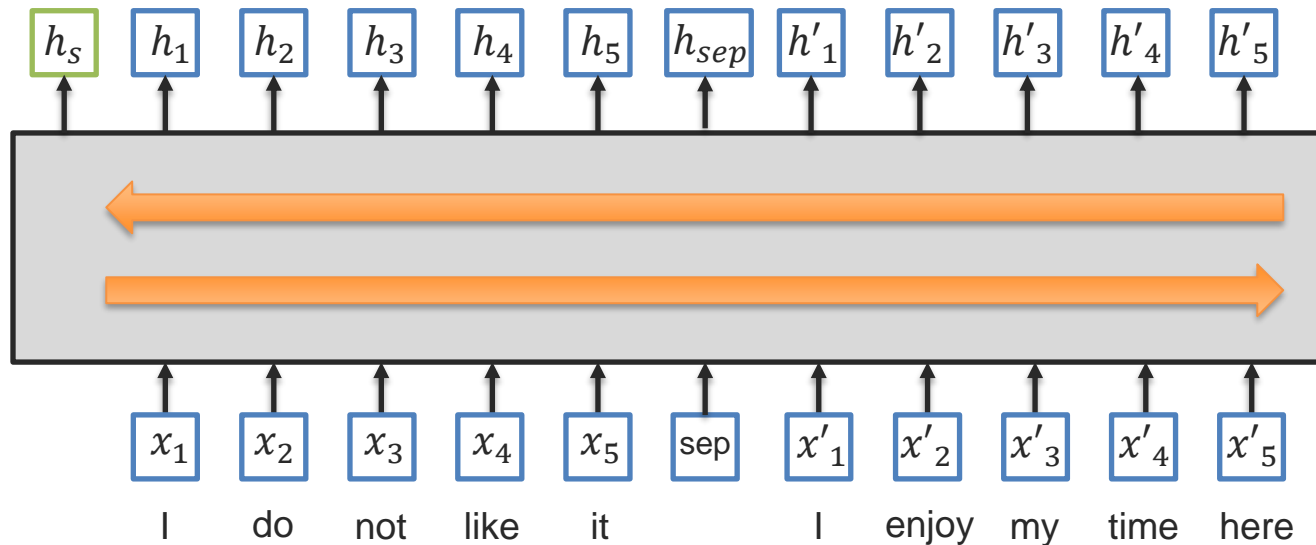


BERT: Bidirectional Encoder Representations from Transformers

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

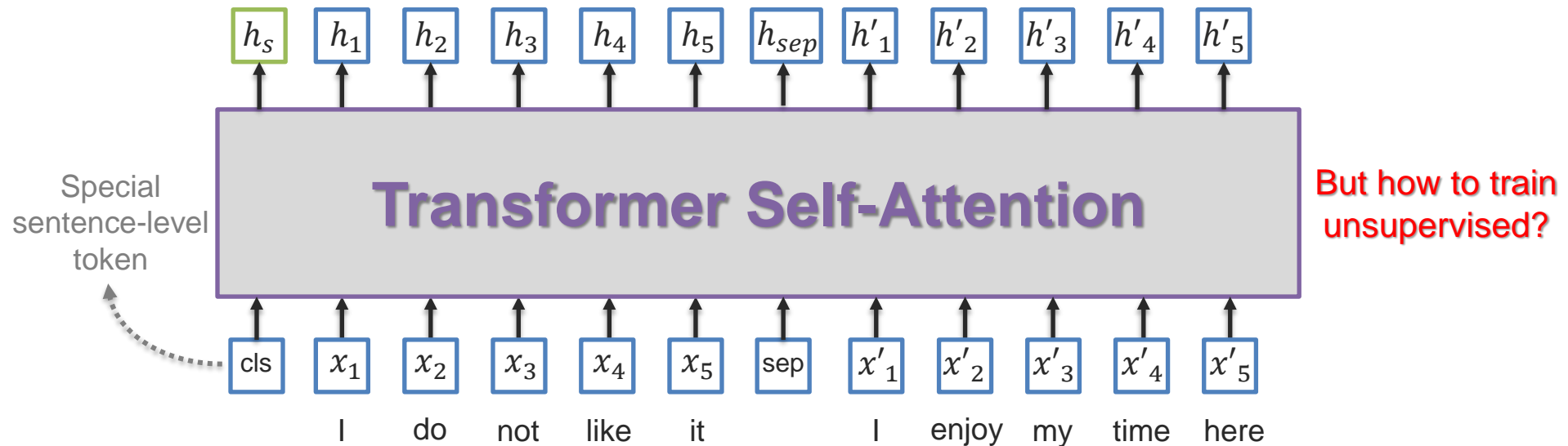
How can
we do all
this?



BERT: Bidirectional Encoder Representations from Transformers

Advantages:

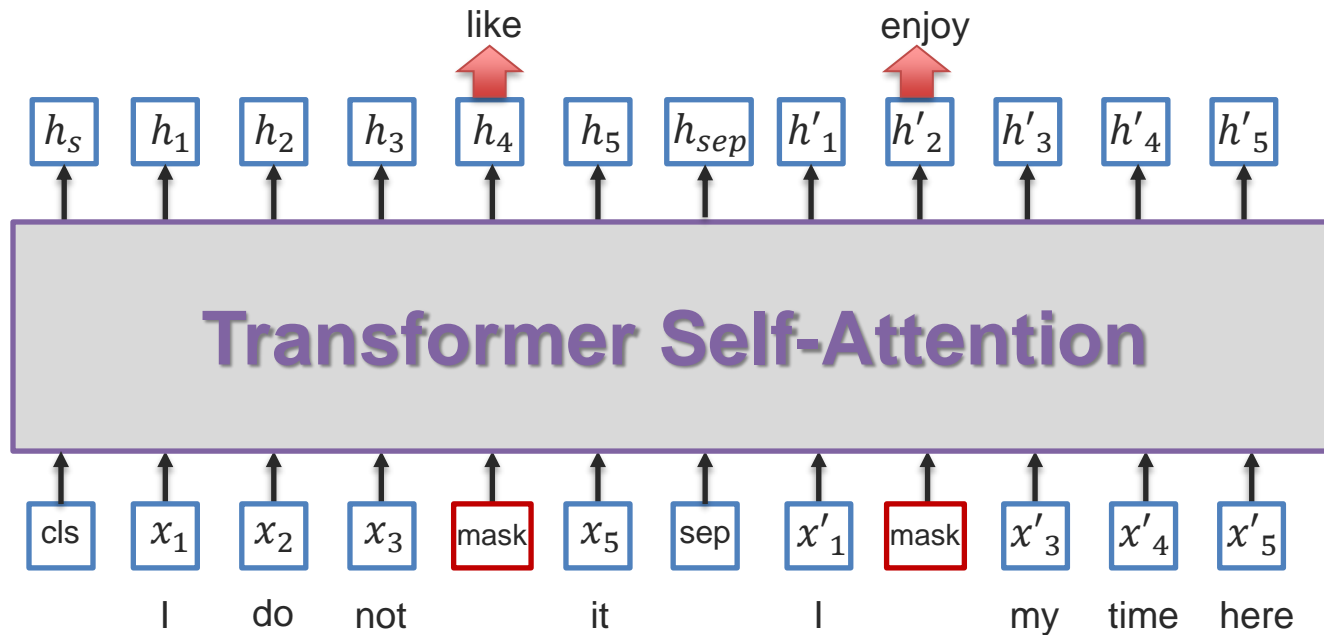
- 1 Jointly learn representation for token-level and sentence level
- 2 Same network architecture for pre-training and fine-tuning
- 3 Can be used learn relationship between sentences
- 4 Models bidirectional interactions between tokens



Pre-training BERT Model

A Masked Language Model

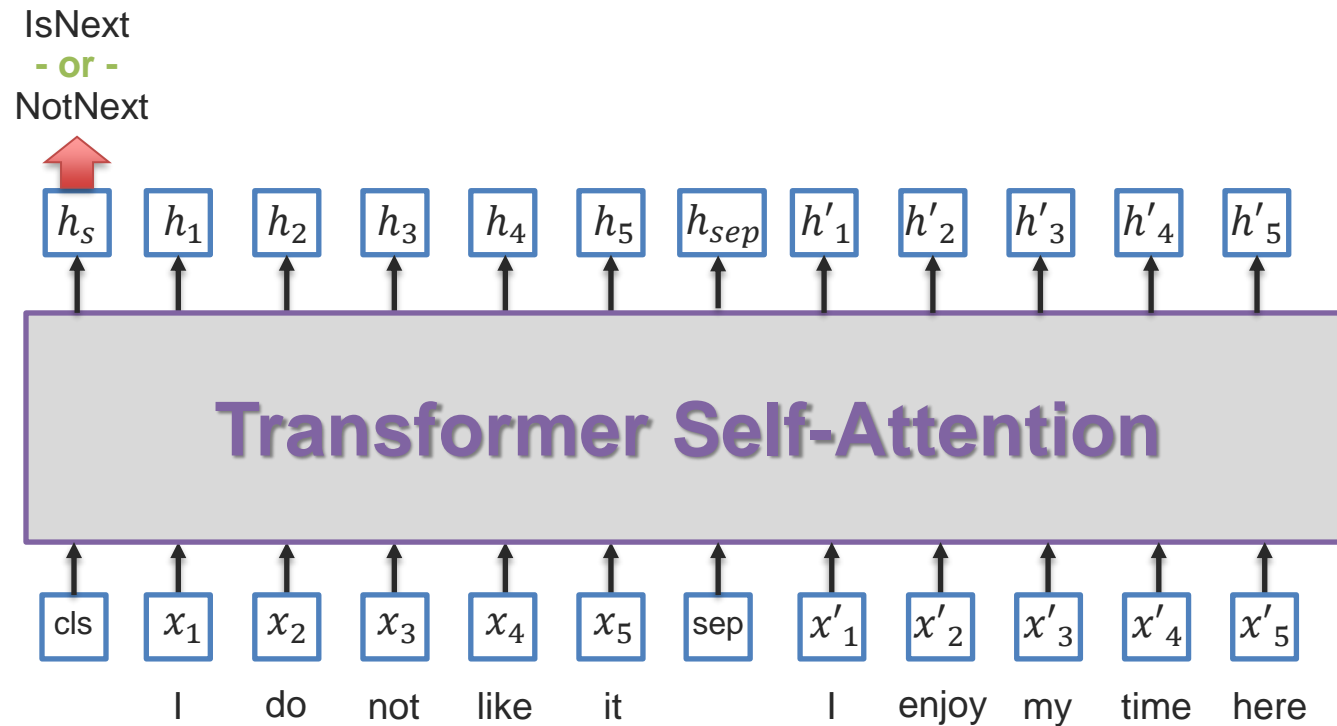
Randomly mask input tokens and then try to predict them



Pre-training BERT Model

B Next Sentence Prediction

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

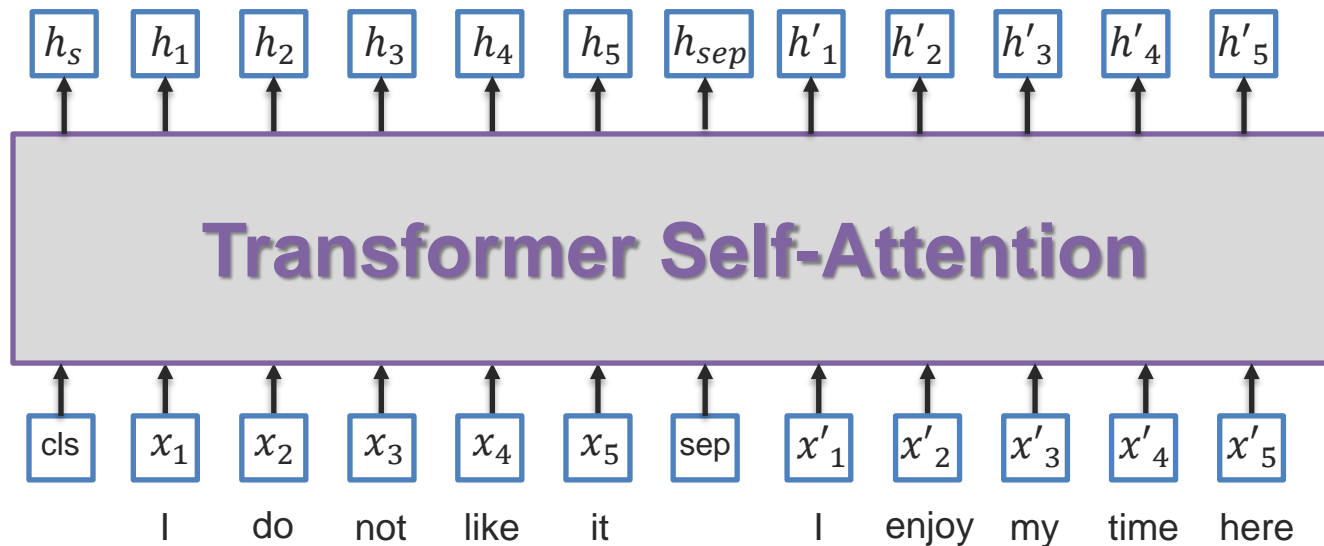


Fine-Tuning BERT

- 1 Sentence-level classification for only one sentence

Examples: sentiment analysis, document classification

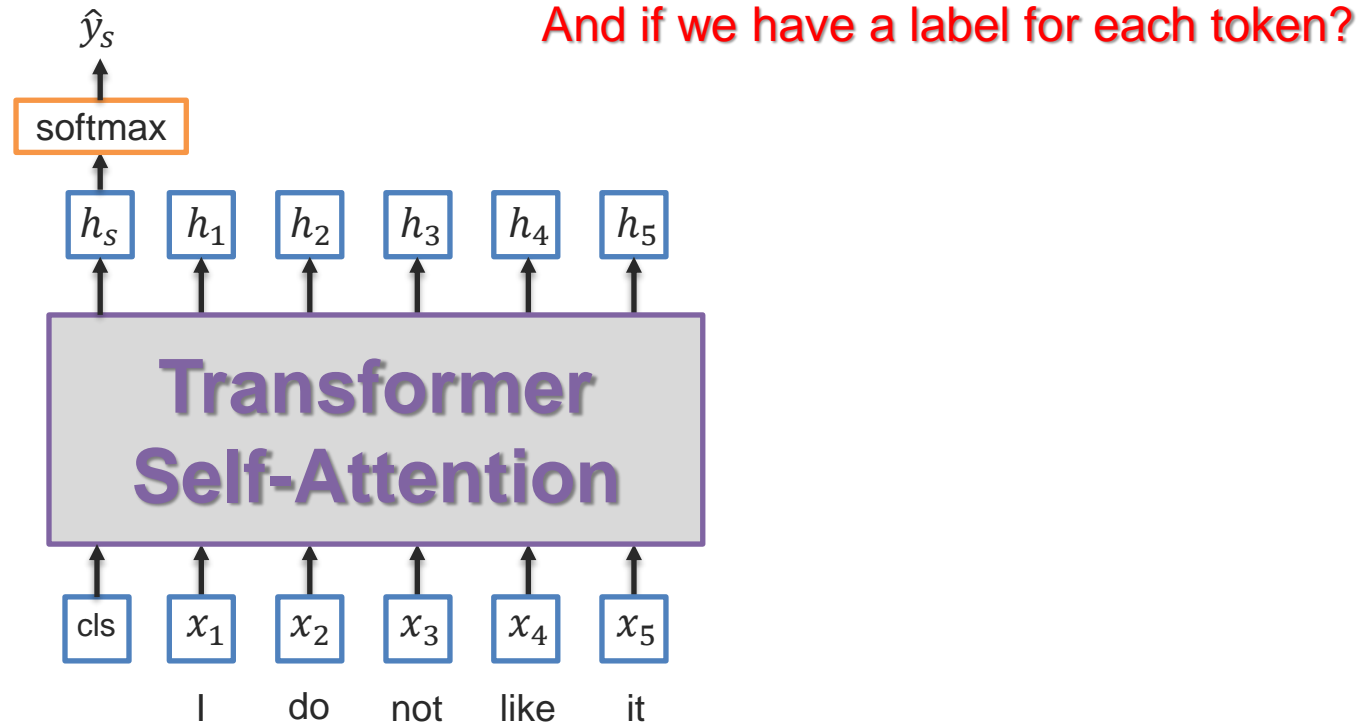
How?



Fine-Tuning BERT

- 1 Sentence-level classification for only one sentence

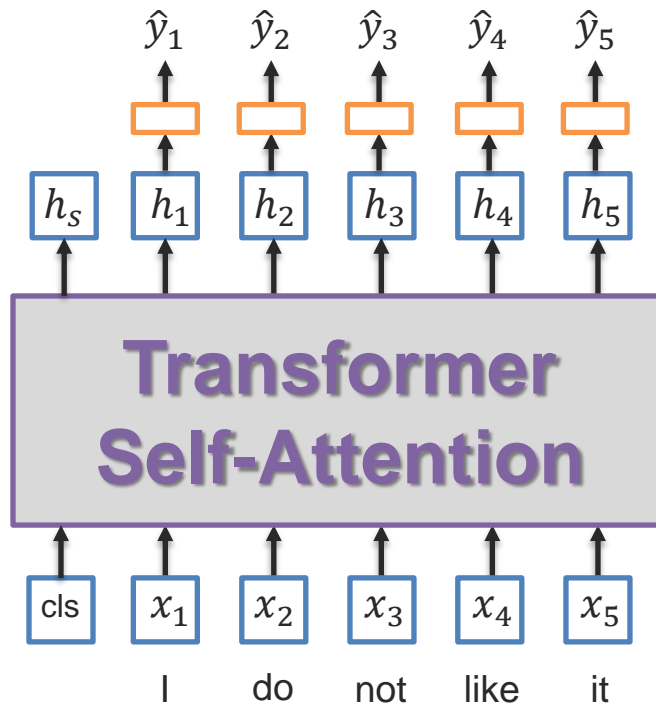
Examples: sentiment analysis, document classification



Fine-Tuning BERT

2 Token-level classification for only one sentence

Examples: part-of-speech tagging, slot filling

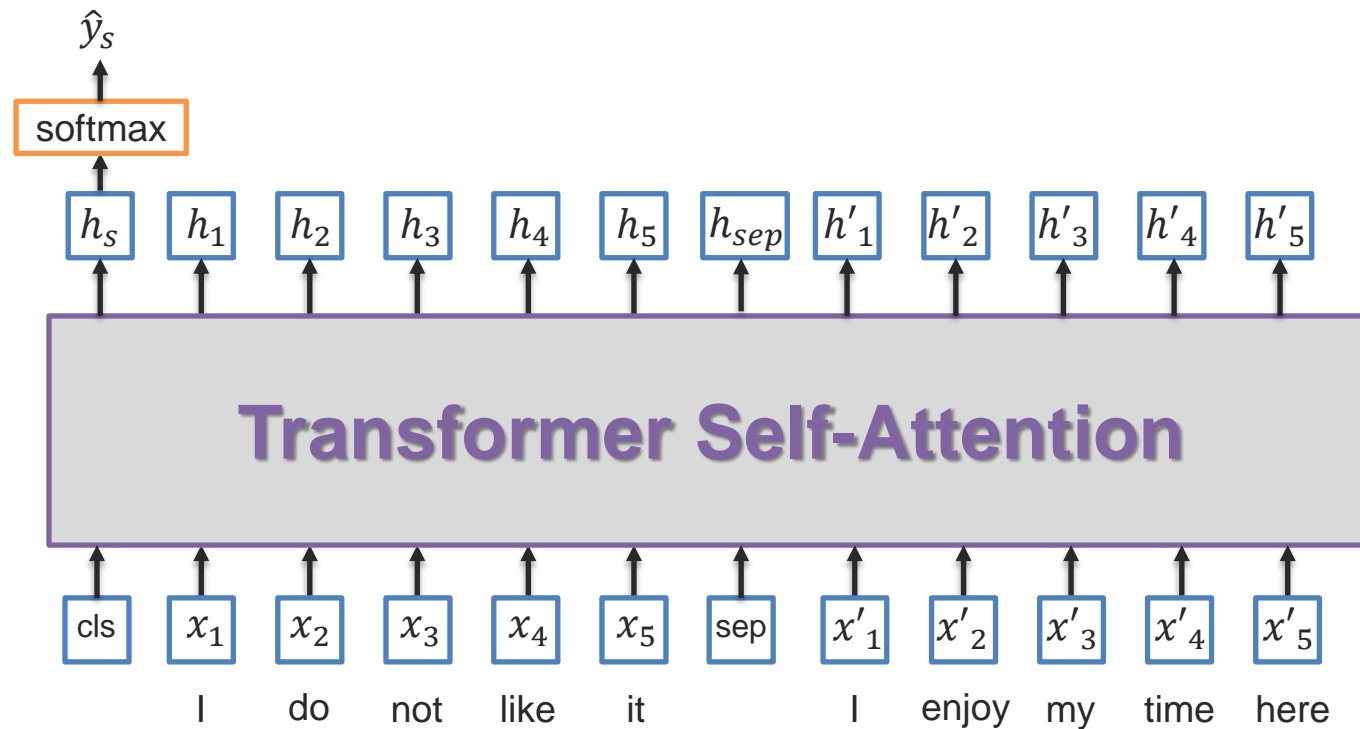


How to compare two sentences?

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

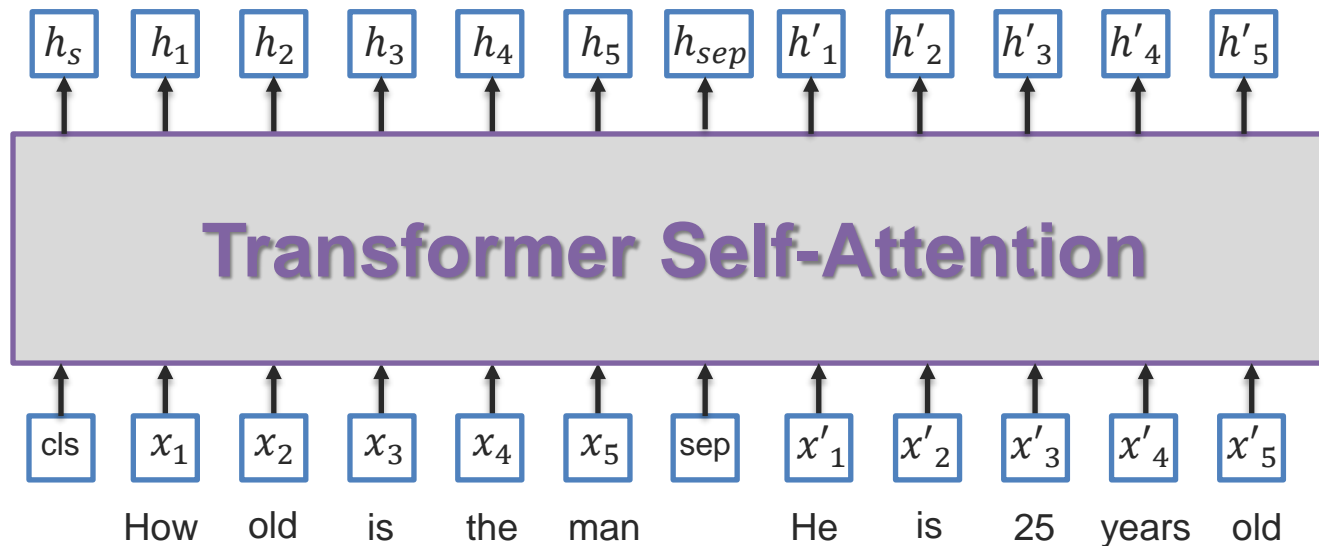
Paragraph: “... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition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

Question 1: “Which laws faced significant opposition?”

Plausible Answer: later laws

Question 2: “What was the name of the 1937 treaty?”

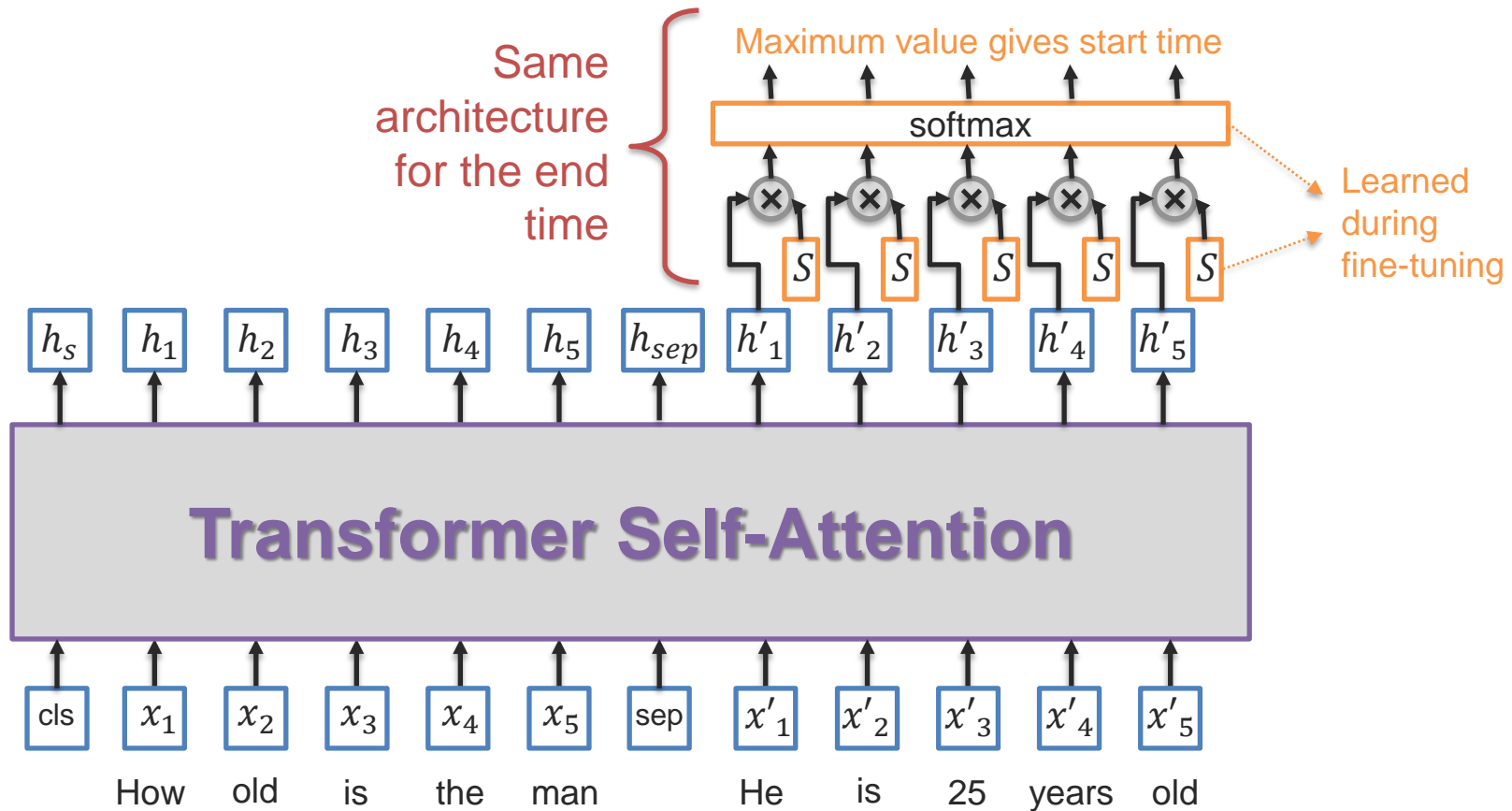
Plausible Answer: Bald Eagle Protection Act



How?

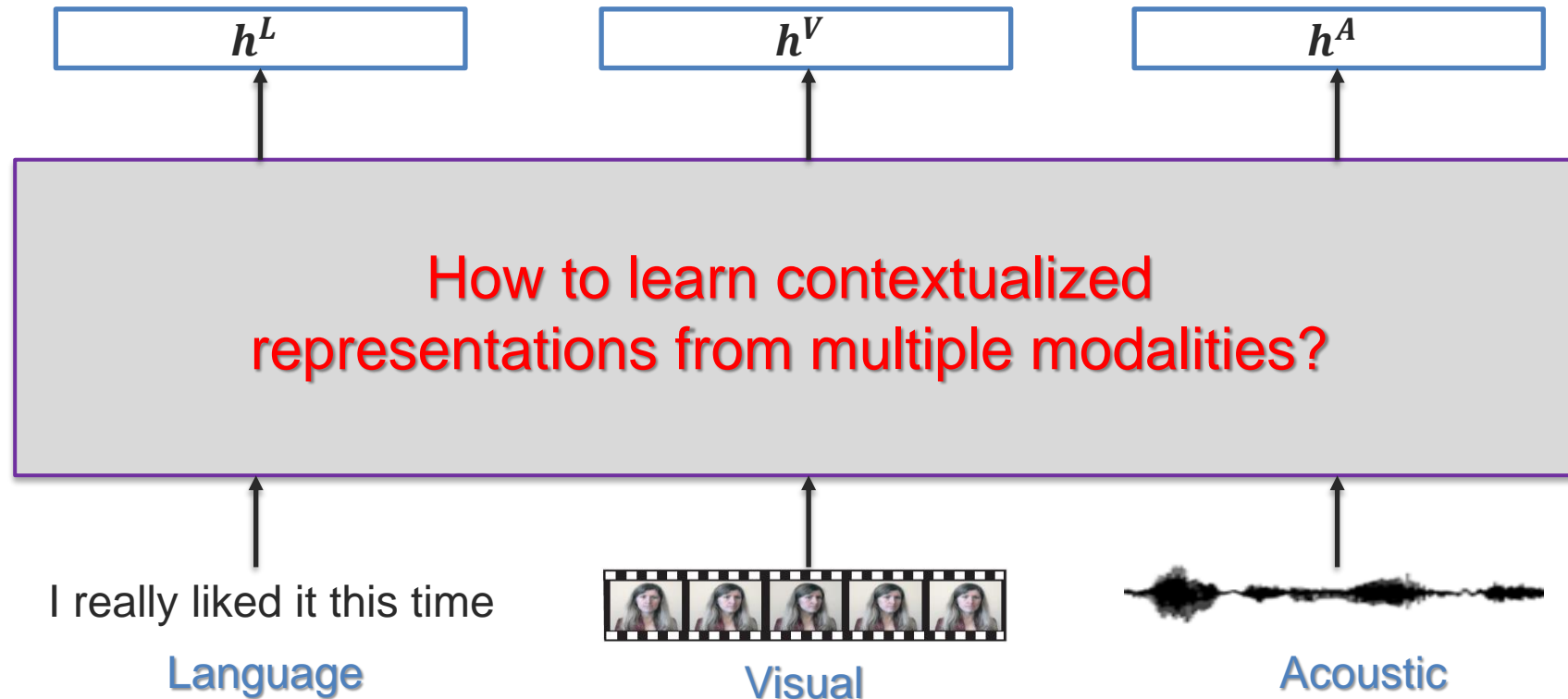
Fine-Tuning BERT

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

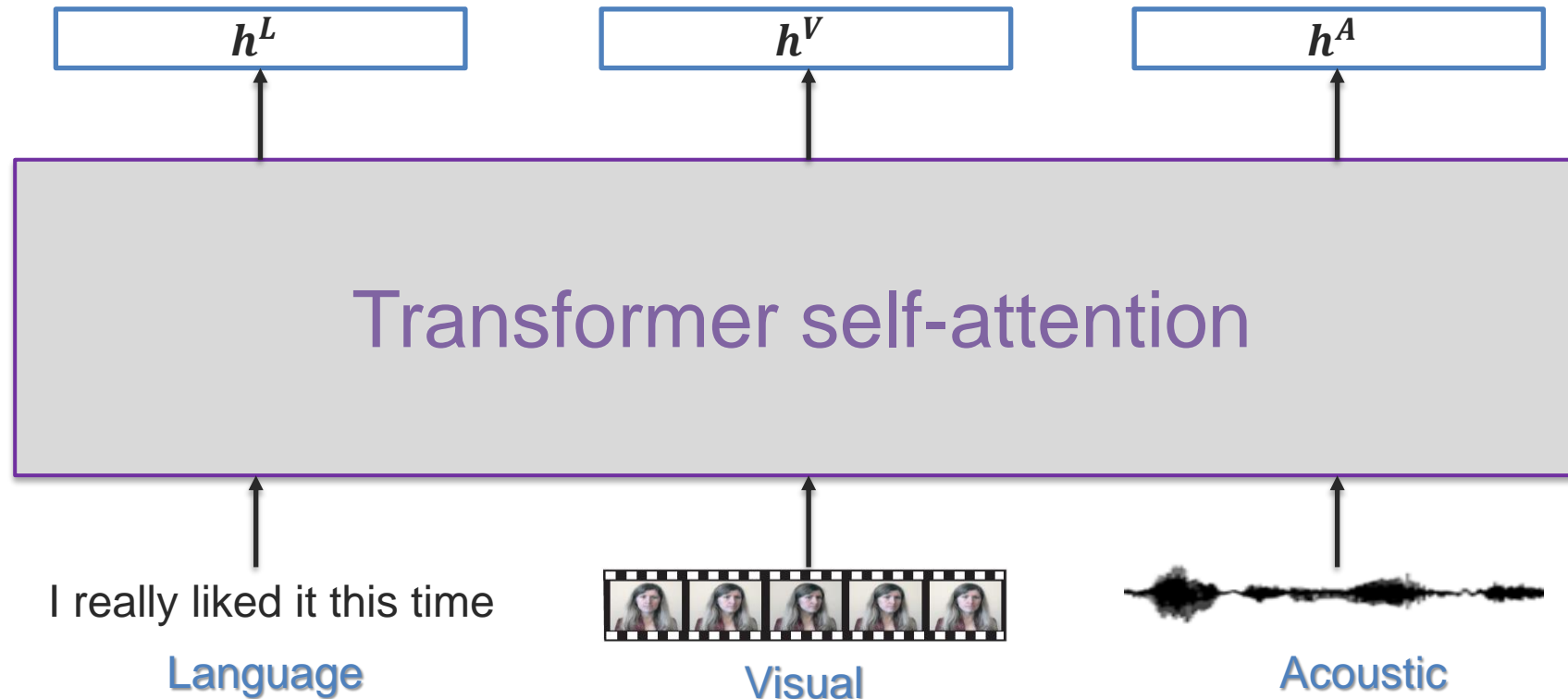


Multimodal Transformers

Multimodal Embeddings



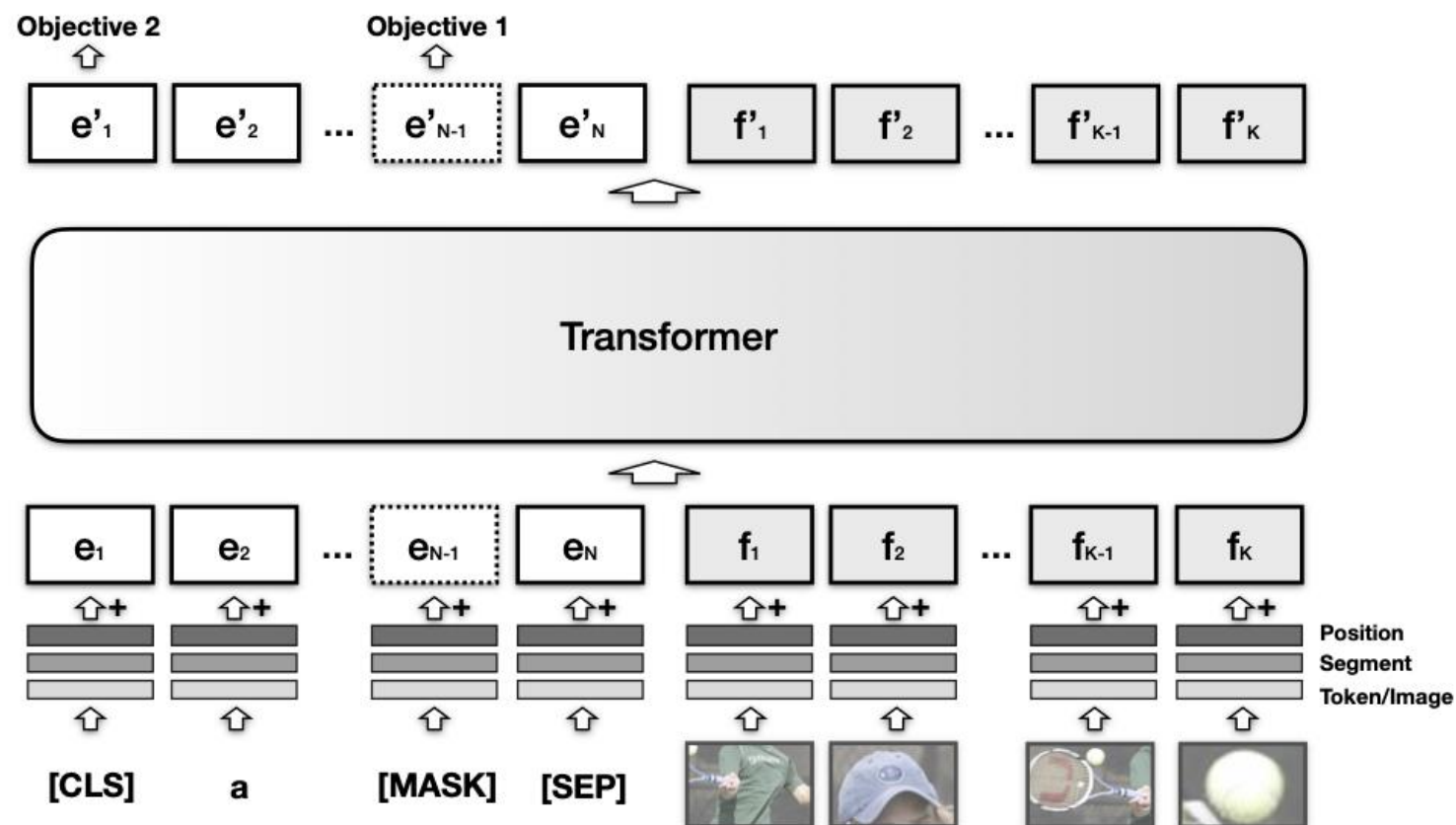
Simple Solution: Contextualized Multimodal Embeddings



VisualBERT



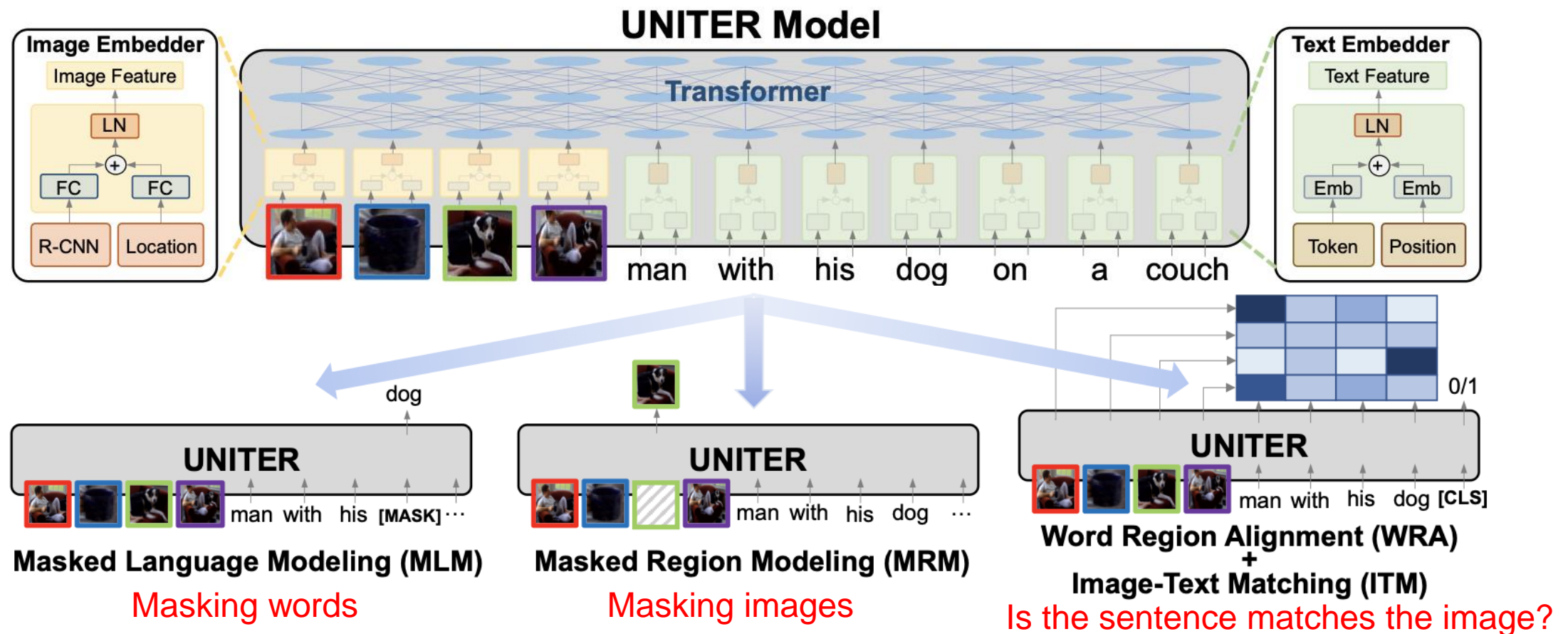
A person hits a ball with a tennis racket



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

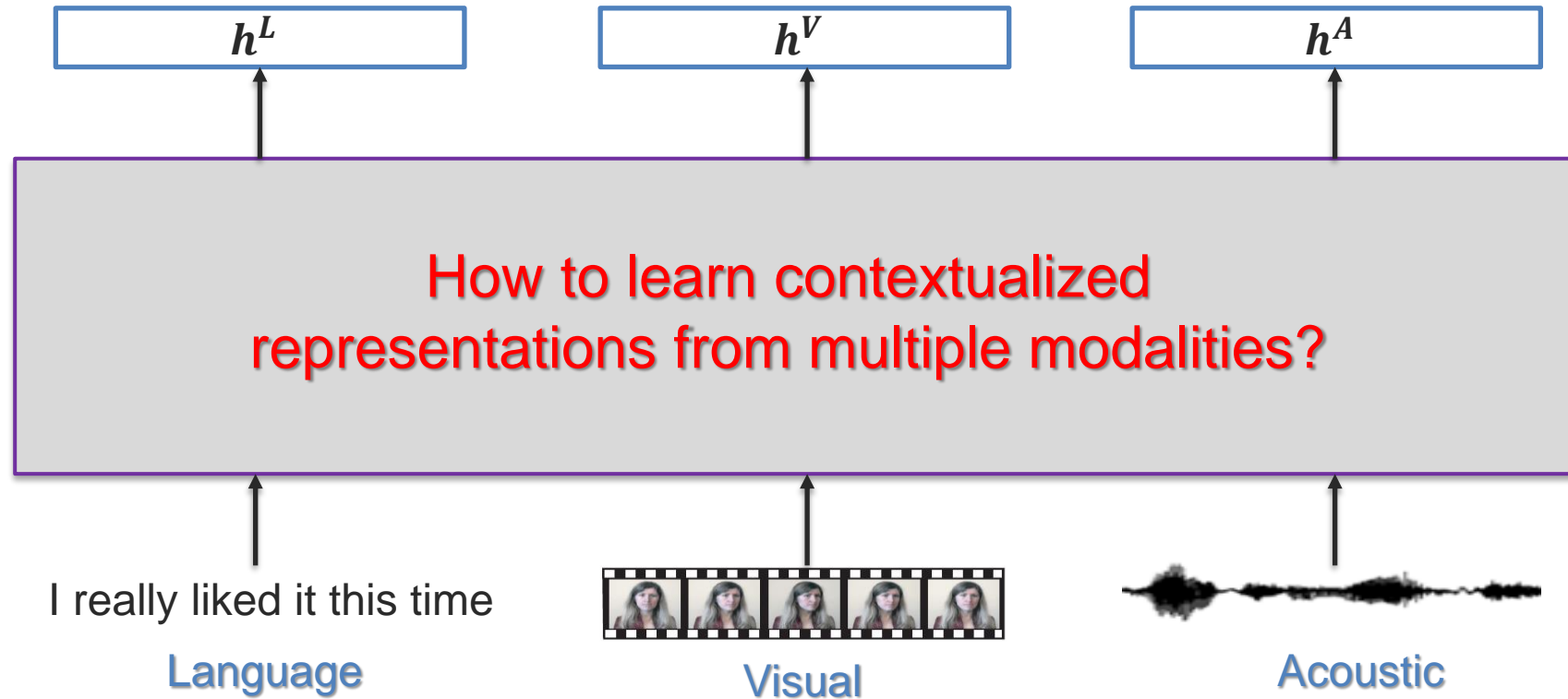
UNITER

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



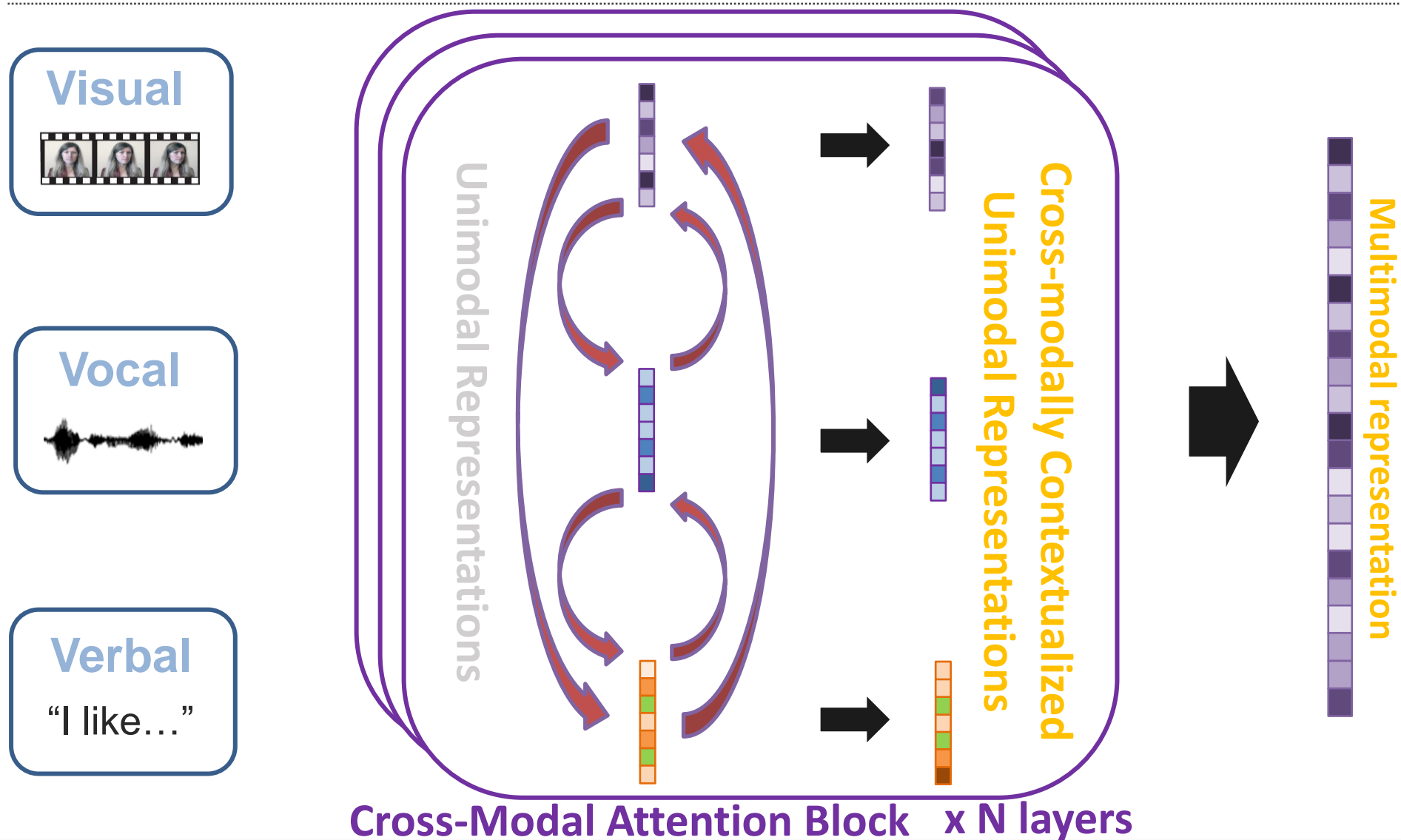
Chen, Yen-Chun, et al. "Uniter: Universal image-text representation learning." *European conference on computer vision*. 2020.

Multimodal Embeddings

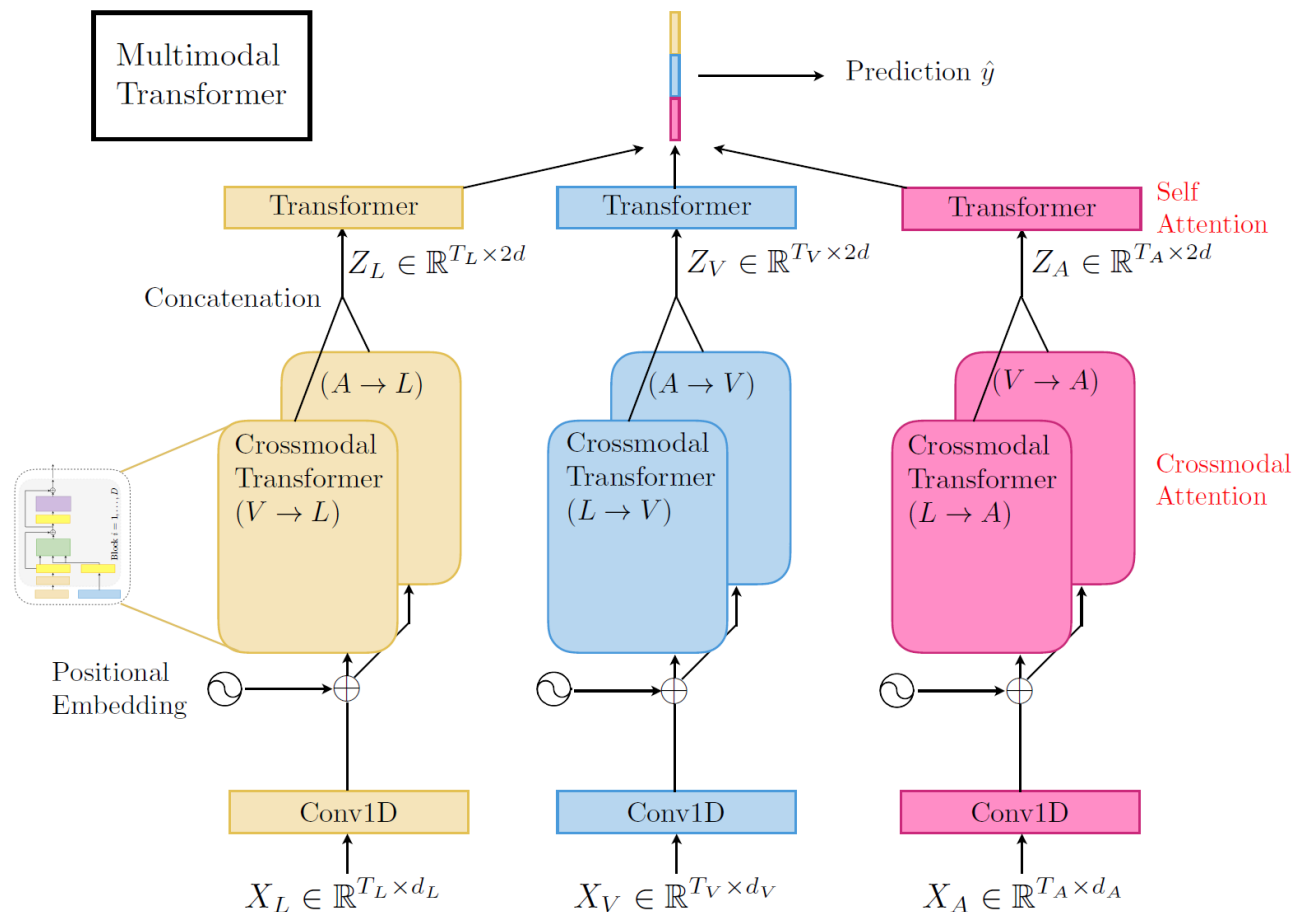


Look at pairwise interactions between modalities

Multimodal Transformer – Pairwise Cross-Modal



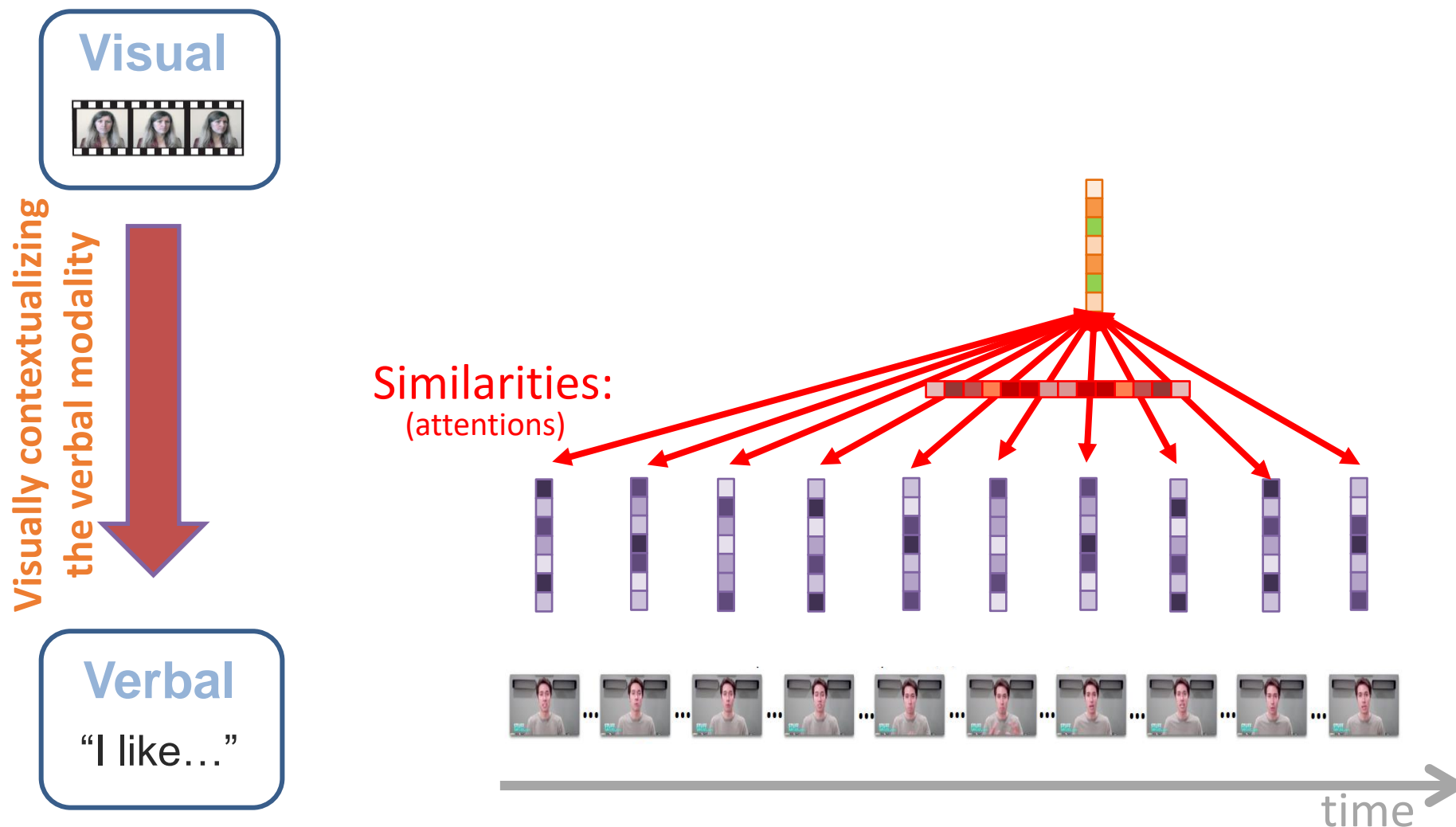
Multimodal Transformer – Pairwise Cross-Modal



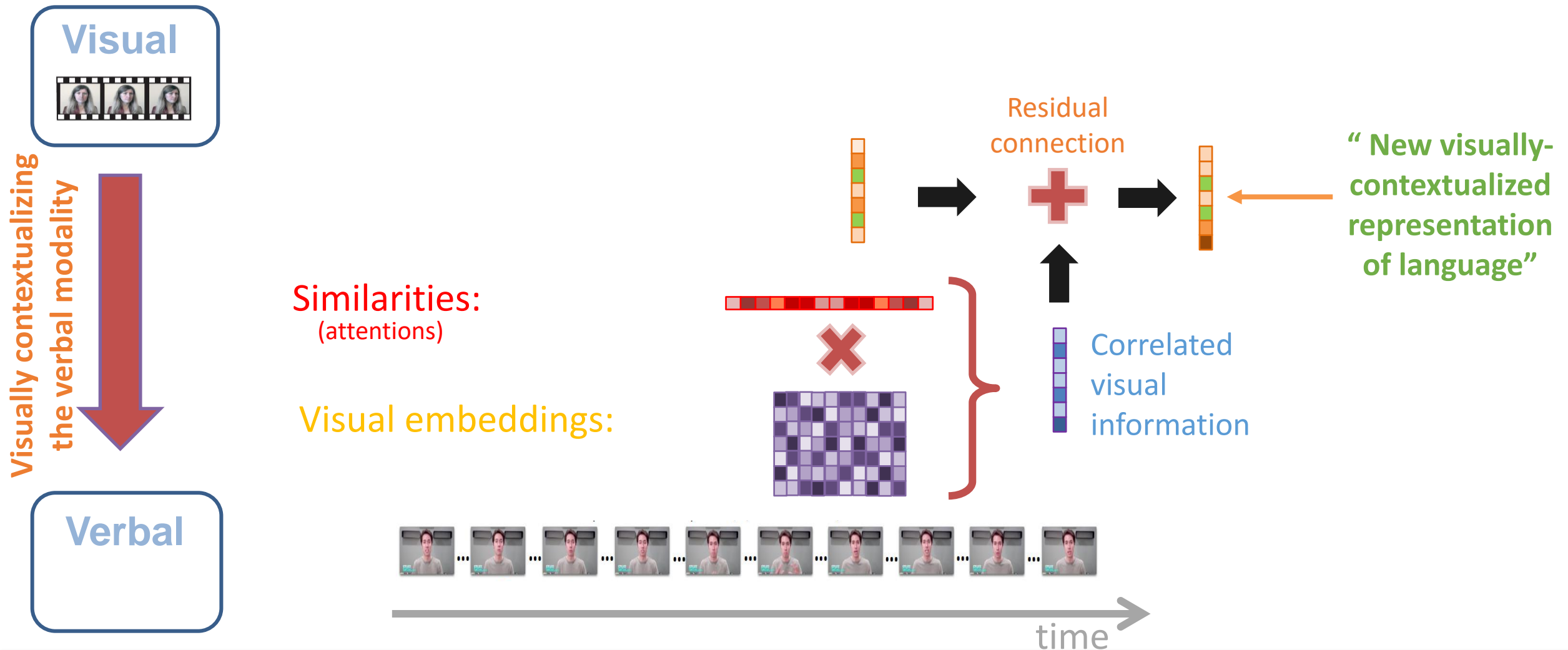
How should we connect Query, Key and Value in the crossmodal transformer?

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

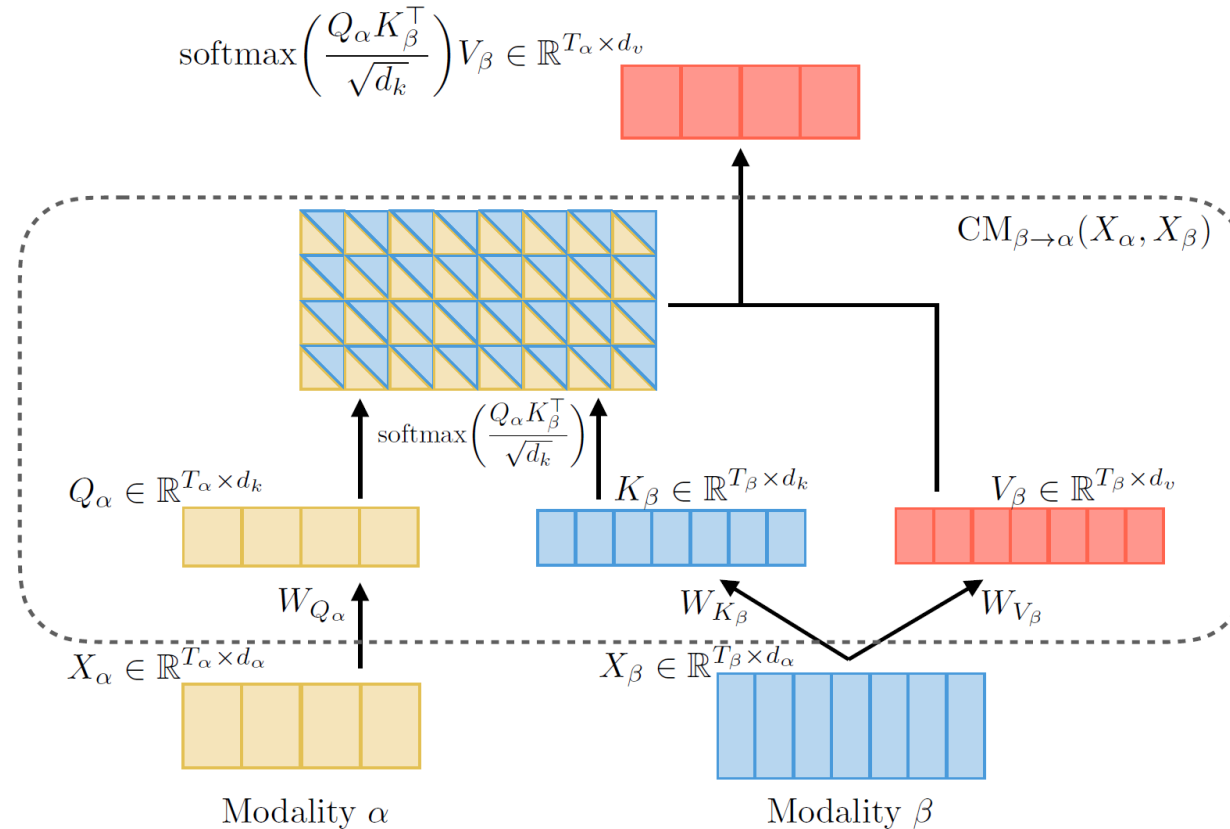
Cross-Modal Transformer Module ($V \rightarrow L$)



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

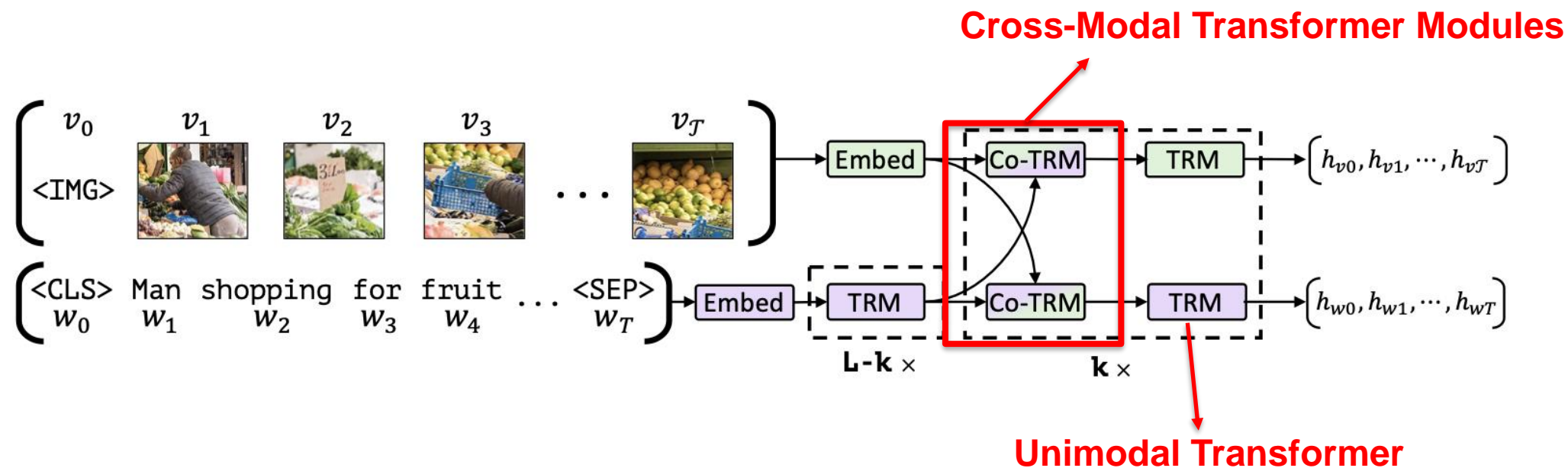


Cross-Modal Transformer Module ($\beta \rightarrow \alpha$)



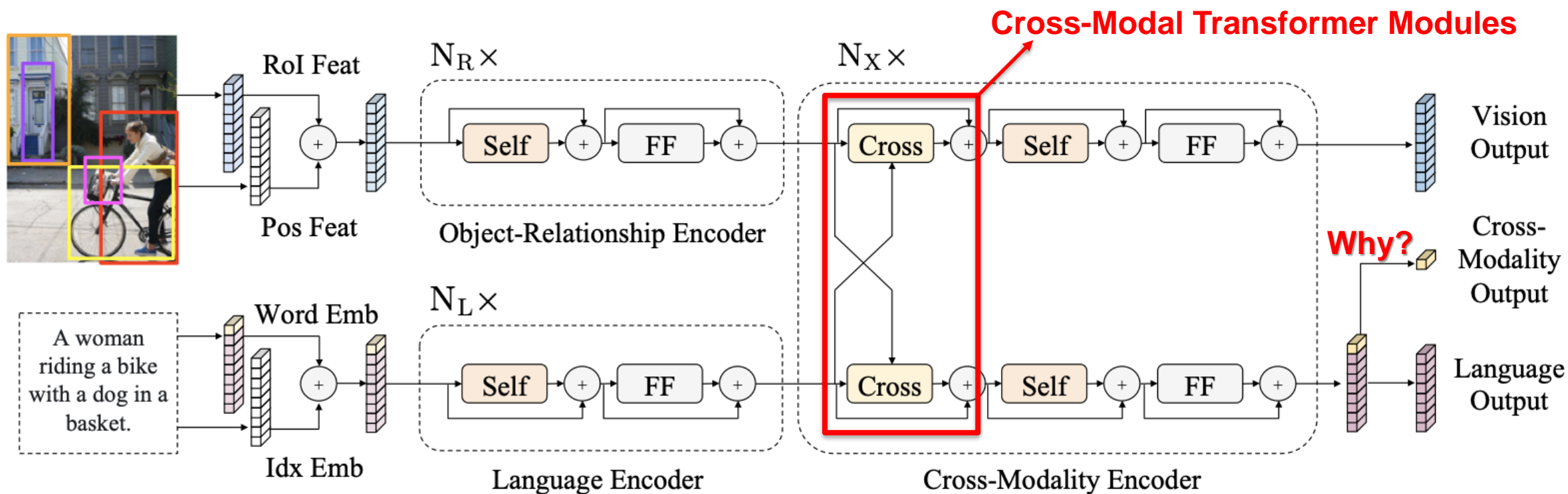
Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

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).

Video and Language Transformers

HowTo100M

🔍 Iron cloth



🔍 Cut paper



🔍 Cut wood



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

Table 2: Number of tasks, videos and clips within each category.

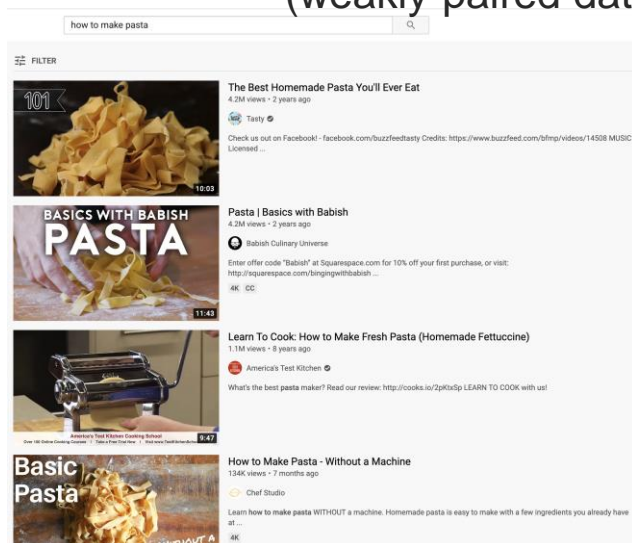
Visual Representations from Uncurated Instructional Videos

Goal: Learn better visual representations...

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

Co-Learning!

Instructional videos
(weakly-paired data)



it's turning into a much thicker mixture



The biggest mistake is not kneading it enough



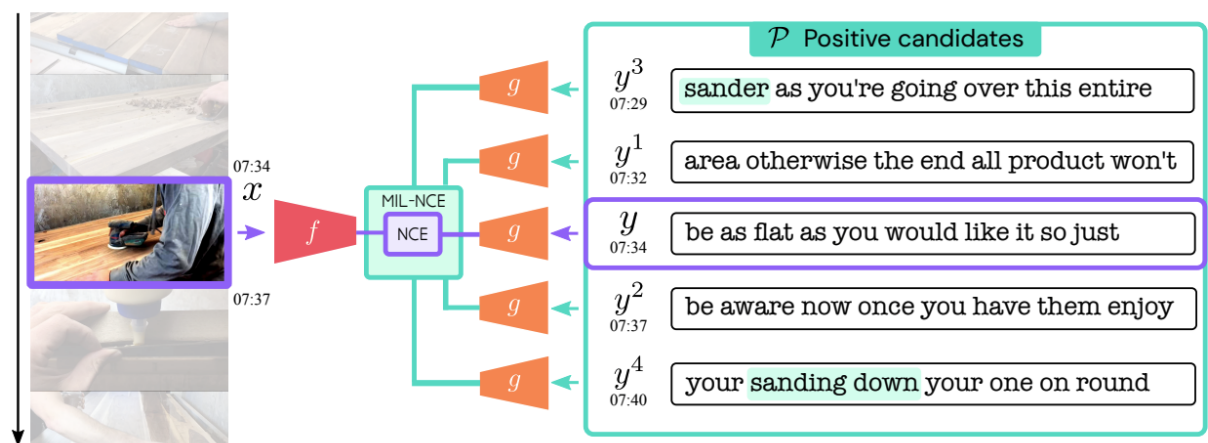
...

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

Multiple Instance Learning Noise Contrastive Estimation

Objective

Given video x and text y from a positive set P_i and a negative set N_i , maximize the positive / total score ratio




$$\max_{f,g} \sum_{i=1}^n \log \left(\frac{\sum_{(x,y) \in P_i} e^{f(x)^\top g(y)}}{\sum_{(x,y) \in P_i} e^{f(x)^\top g(y)} + \sum_{(x',y') \sim \mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

Note: Doing so requires maximizing $f(x)^\top g(y)$ for only positive examples

1. Using sets of positive and negative examples to ~wash out the misaligned text
2. Ideally, we would maximize all positives over all possible negatives (intractable)

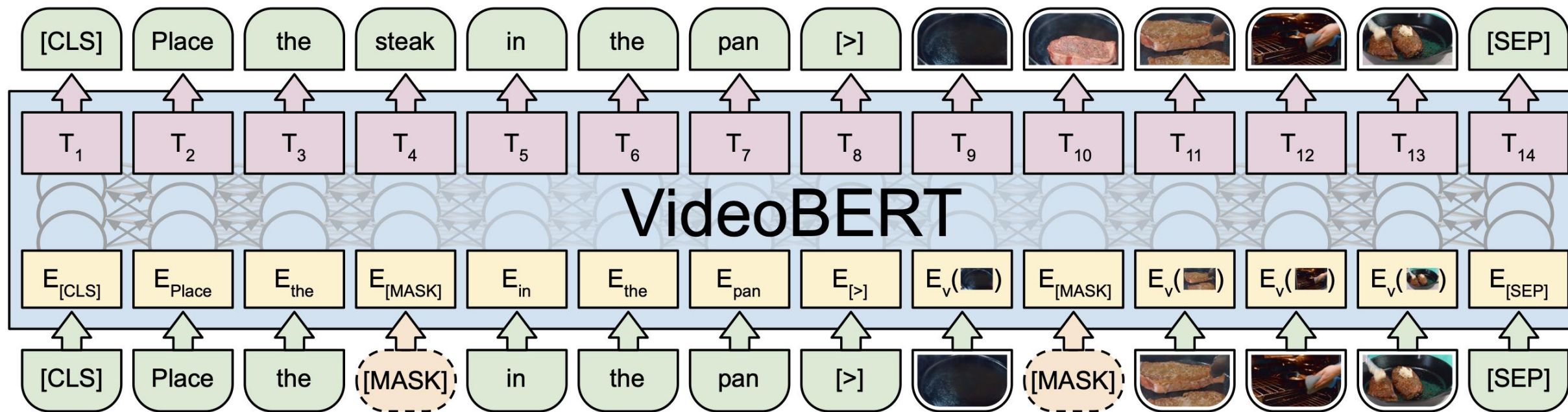
End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Experiments – HowTo100M Dataset

	<p>\mathcal{P} Positive candidates</p> <ul style="list-style-type: none">.60 it's quite a simple technique for.53 beginners to learn and basically all I.63 do is squeeze out three little circles.49 then with the back of a teaspoon.47 simply press the teaspoon into the
	<p>\mathcal{P} Positive candidates</p> <ul style="list-style-type: none">.50 main body of the laptop cover the.63 duct tape with aluminum cover all.61 remaining gaps edges with aluminum.56 tape use the leftover poster board to.50 create the keyboard keys I made my
	<p>\mathcal{P} Positive candidates</p> <ul style="list-style-type: none">.67 spinach what's the name.57 keep it simple you just want to add.58 fresh herbs maybe some oregano.59 you can add cilantro basil they give.50 it a couple more copies and when you

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

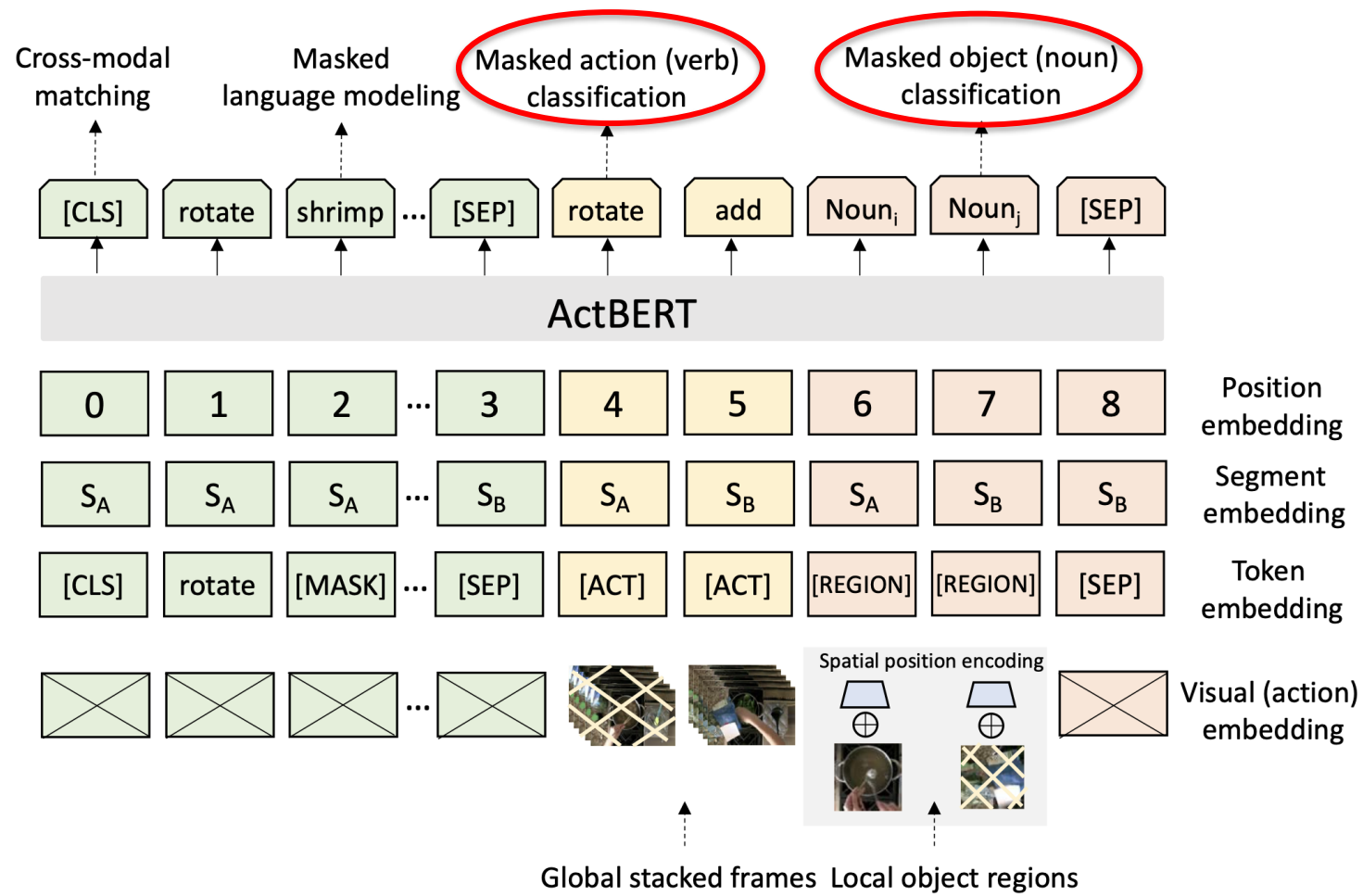
VideoBERT



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

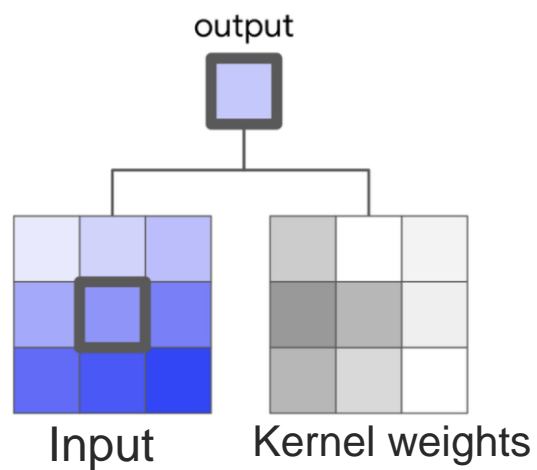
ActBERT



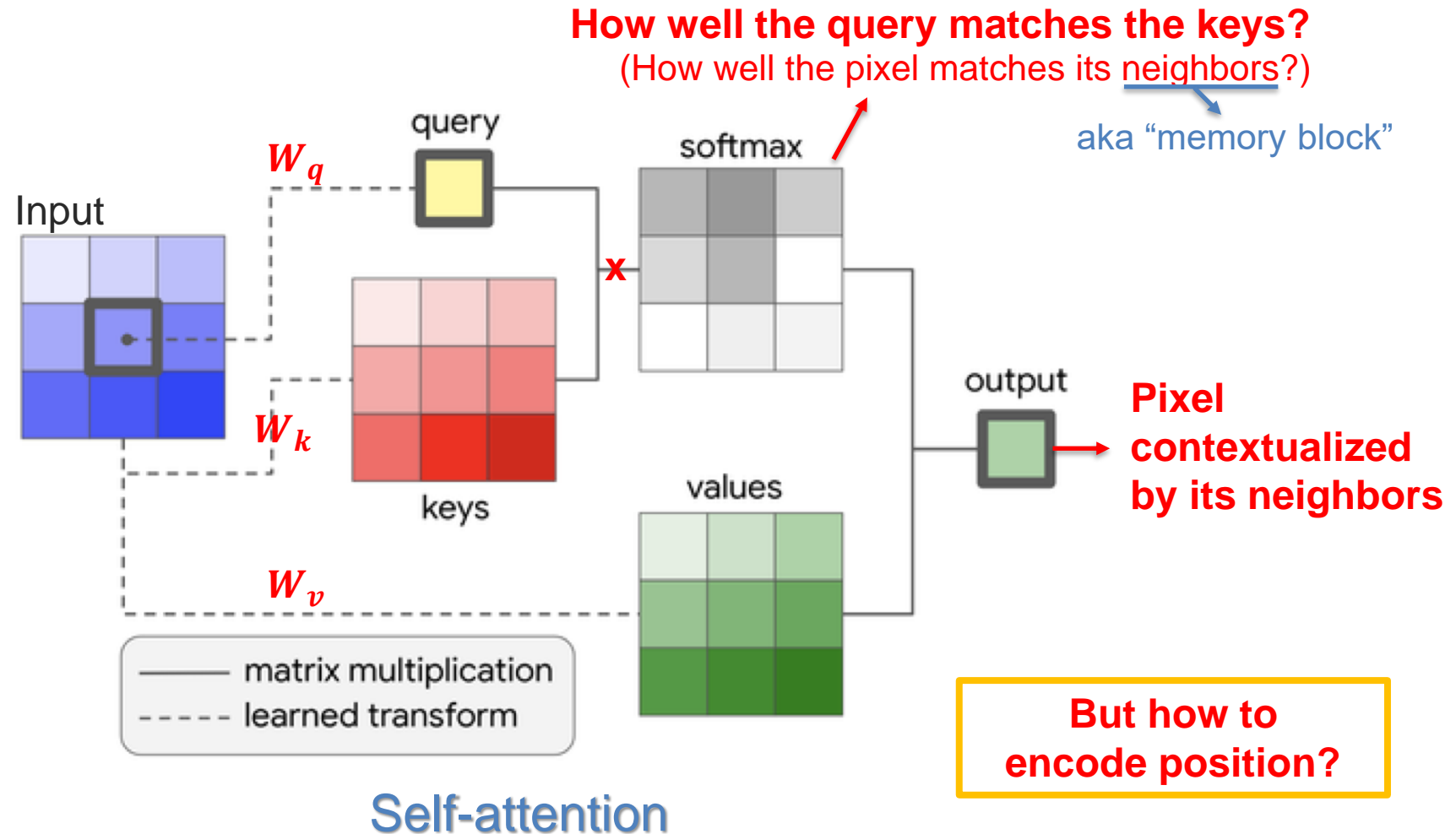
Going Beyond CNNs...

(Vision Transformer)

Replacing a CNN w/ Self-Attention



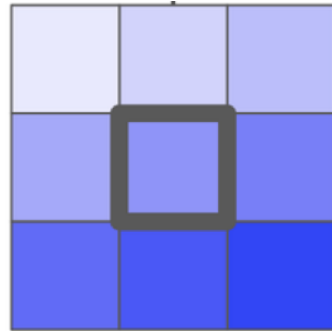
Convolution



<https://arxiv.org/abs/1906.05909>

Replacing a CNN w/ Self-Attention

Image patch



2D relative
position
embedding

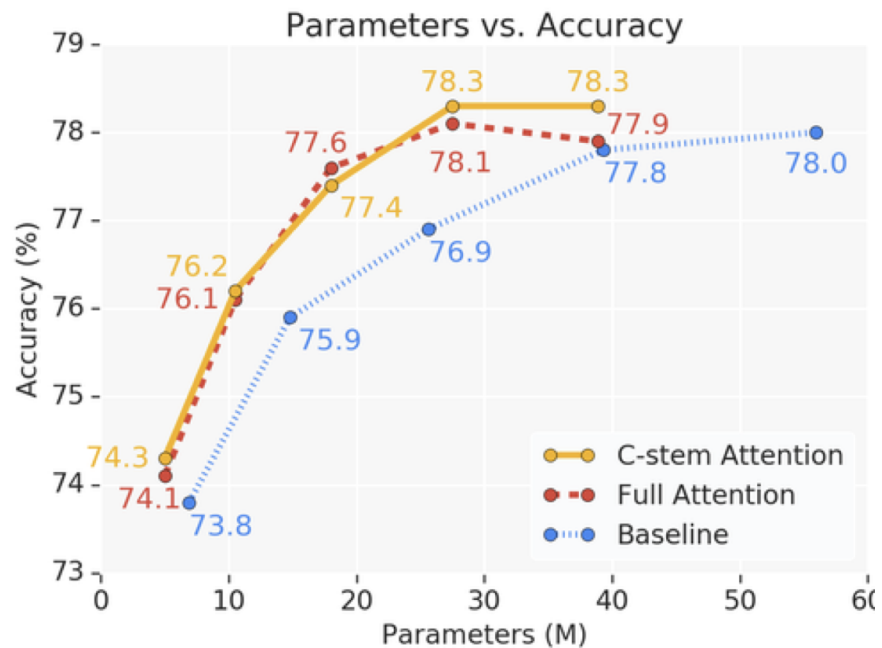
-1, -1	-1, 0	-1, 1	-1, 2
0, -1	0, 0	0, 1	0, 2
1, -1	1, 0	1, 1	1, 2
2, -1	2, 0	2, 1	2, 2

Position embedding is added to the key:

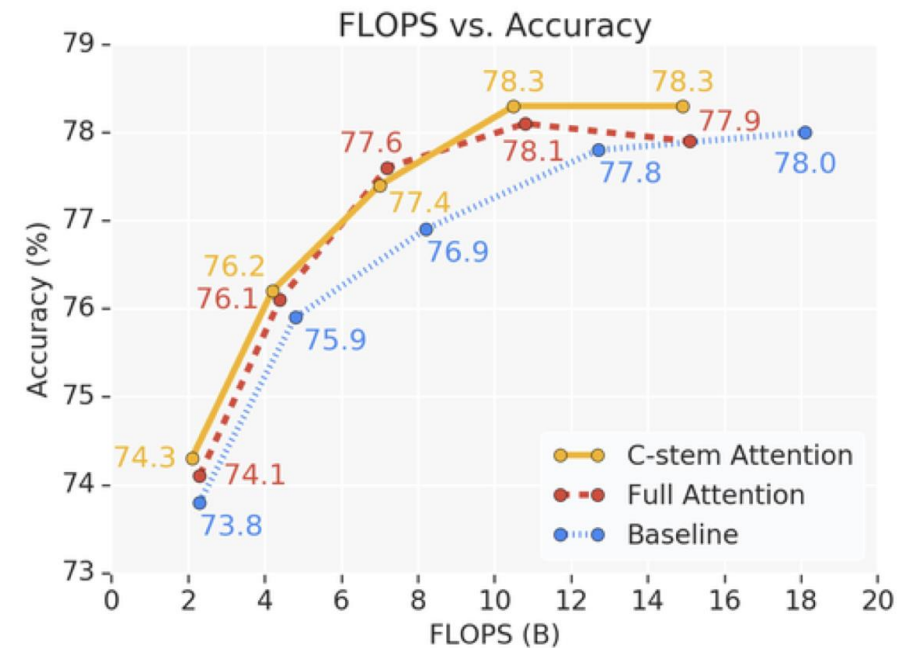
$$y_{ij} = \sum_{a,b \in \mathcal{N}_k(i,j)} \text{softmax}_{ab} \left(q_{ij}^\top k_{ab} + q_{ij}^\top r_{a-i,b-j} \right) v_{ab}$$

Replacing a CNN w/ Self-Attention

It reduces number of parameters:



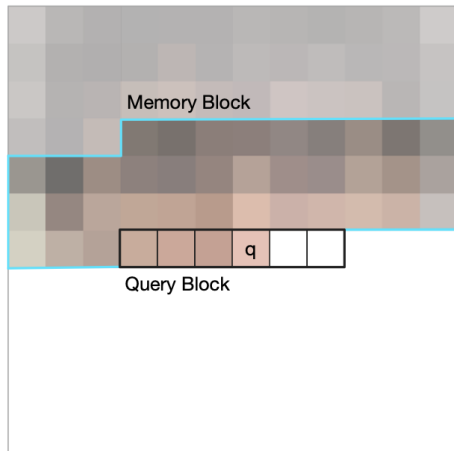
It improves computation time:



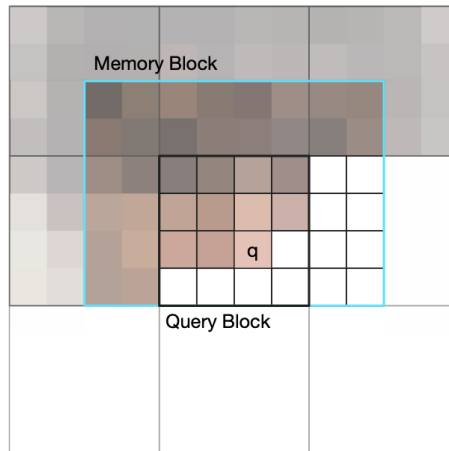
<https://arxiv.org/abs/1906.05909>

Image Transformer

Local 1D Attention



Local 2D Attention



<https://arxiv.org/abs/1802.05751>

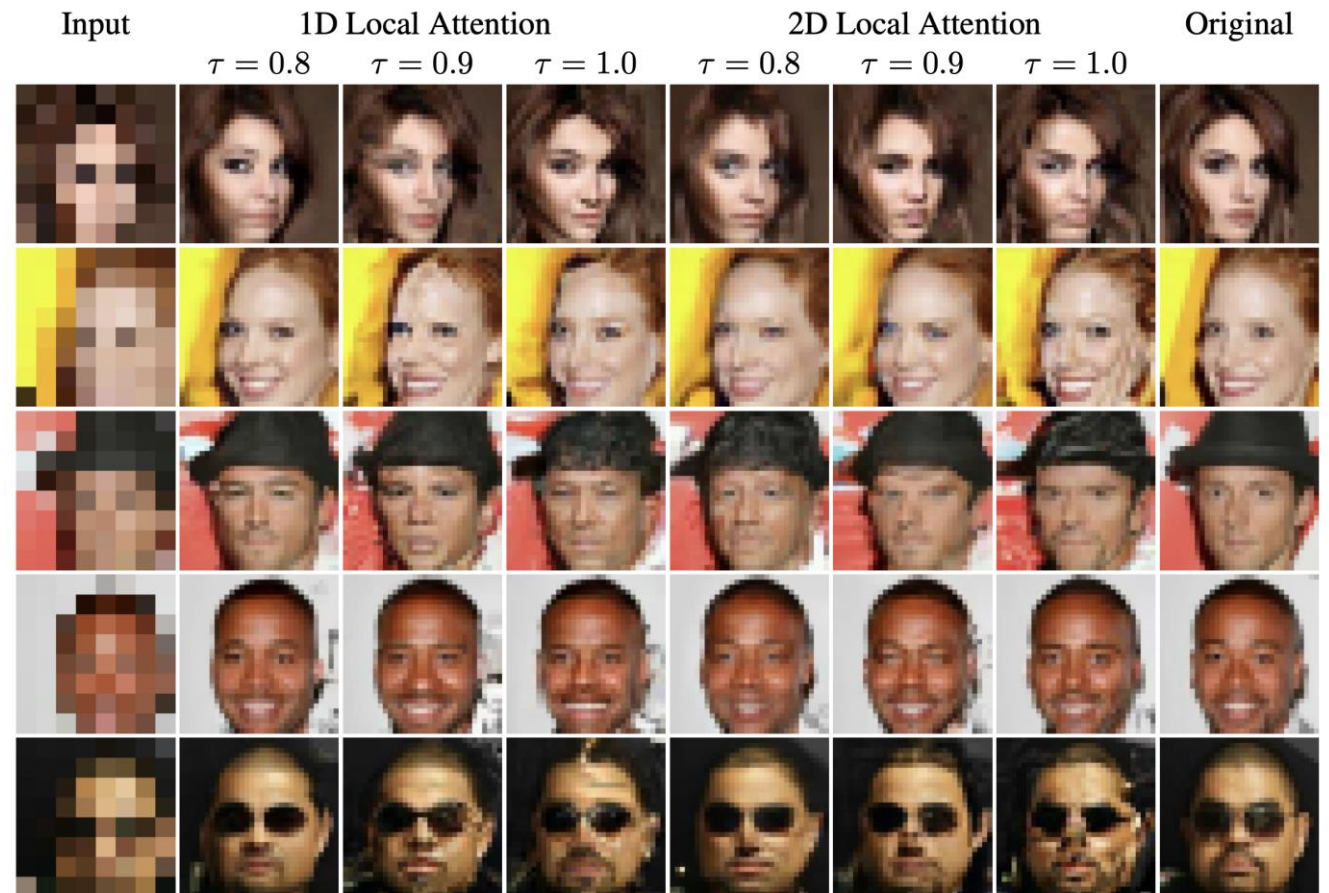


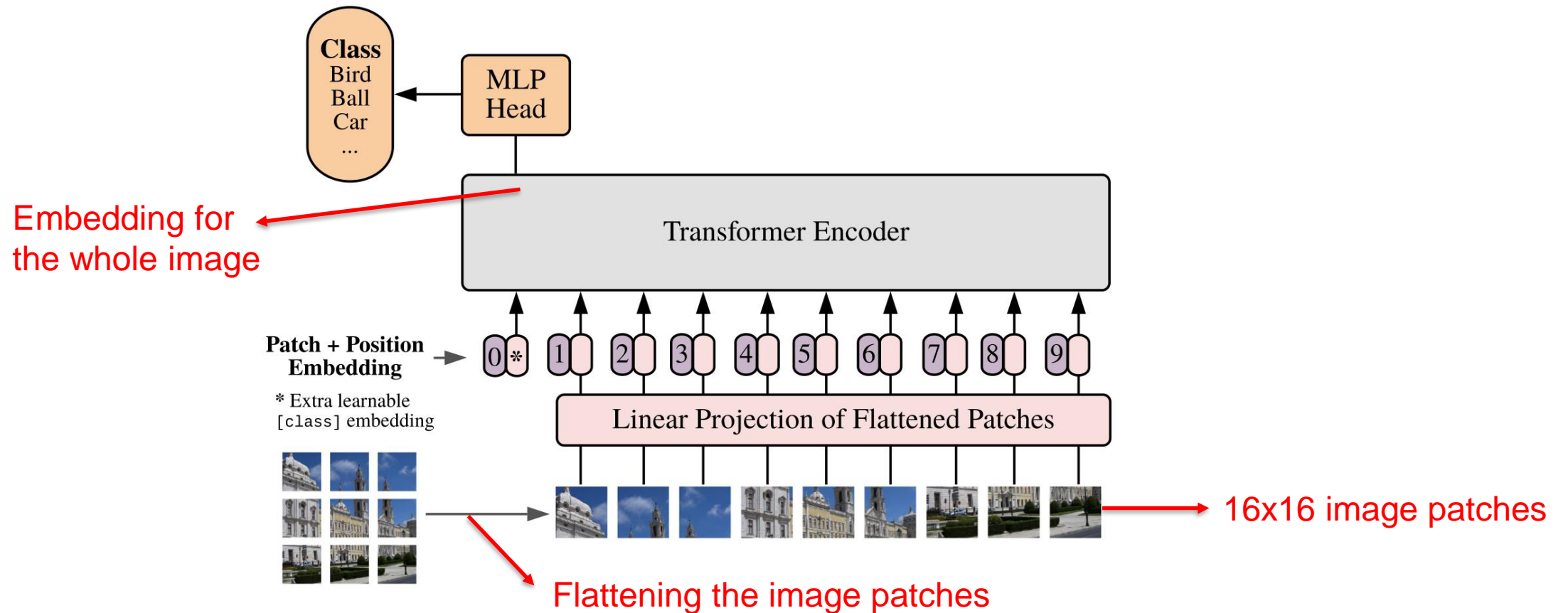
Image super-resolution

Vision Transformer (ViT)



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

Vision Transformer (ViT)

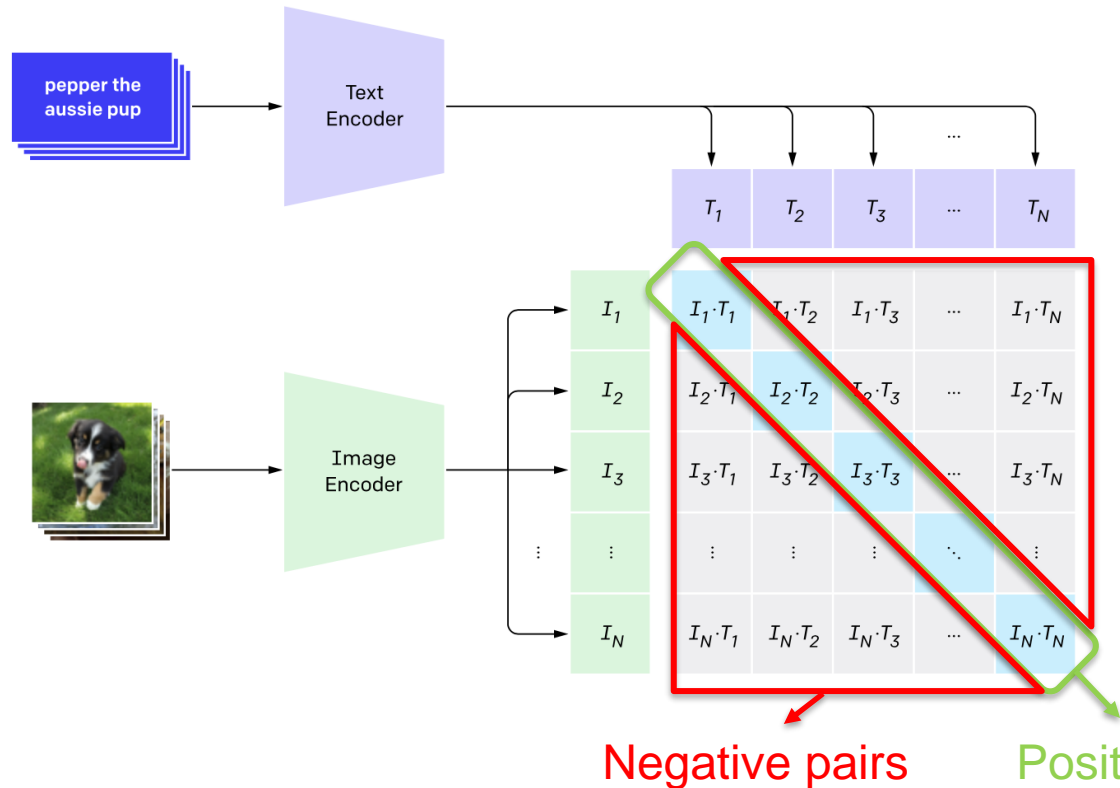


Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

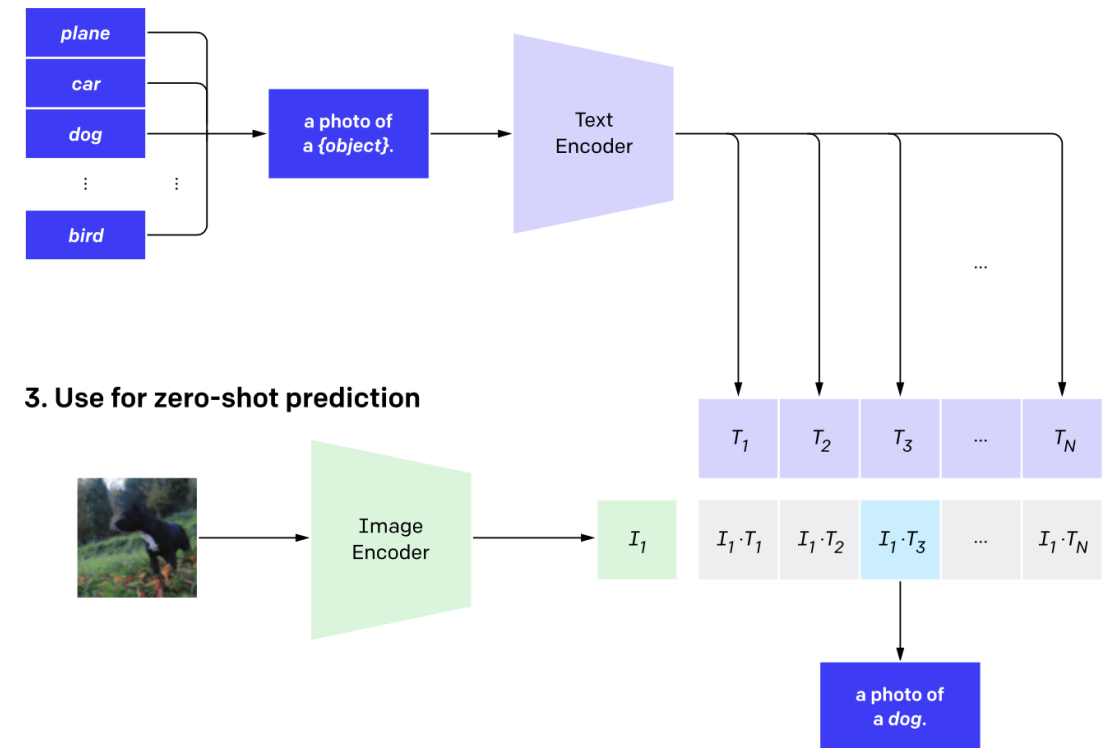
CLIP — 400 million (image, text) pairs

Brings **positive pairs** close to each others and
negative pairs farther from each others

1. Contrastive pre-training



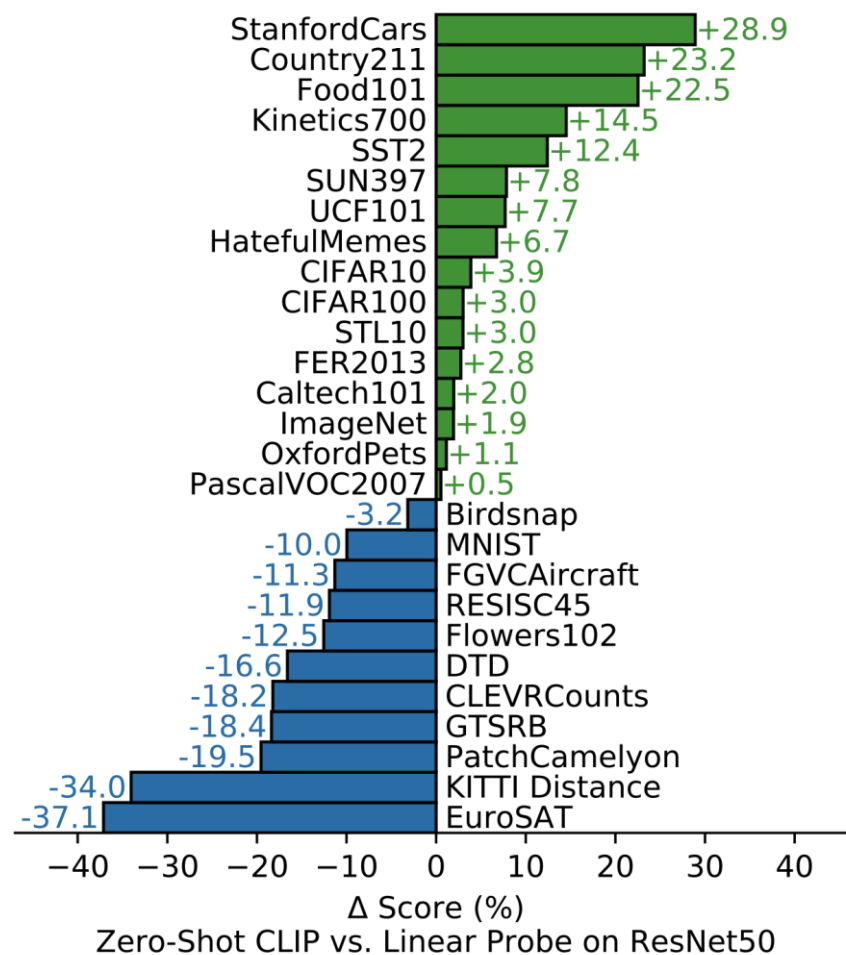
2. Create dataset classifier from label text



3. Use for zero-shot prediction

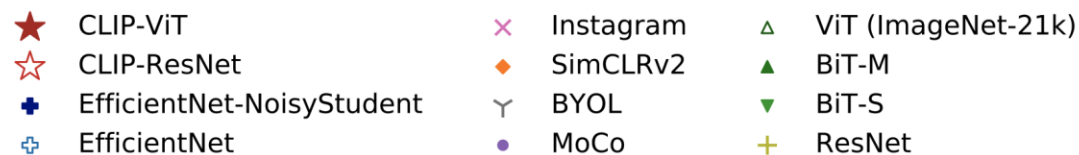
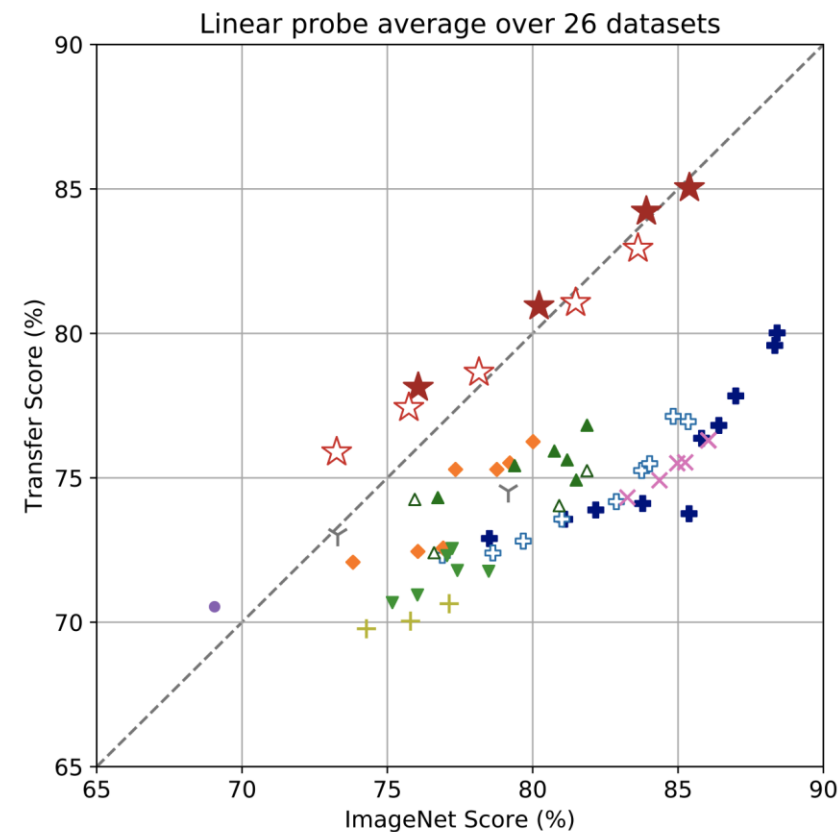
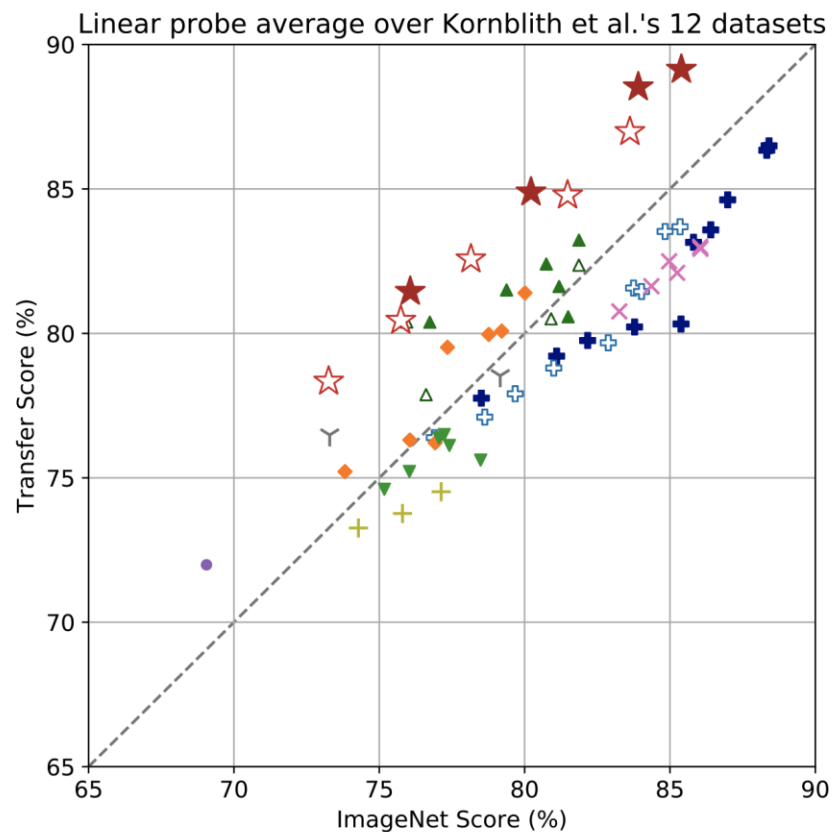
<https://openai.com/blog/clip/>

CLIP - Zero-Shot Testing



<https://openai.com/blog/clip/>

CLIP Using Either ViT or ResNet

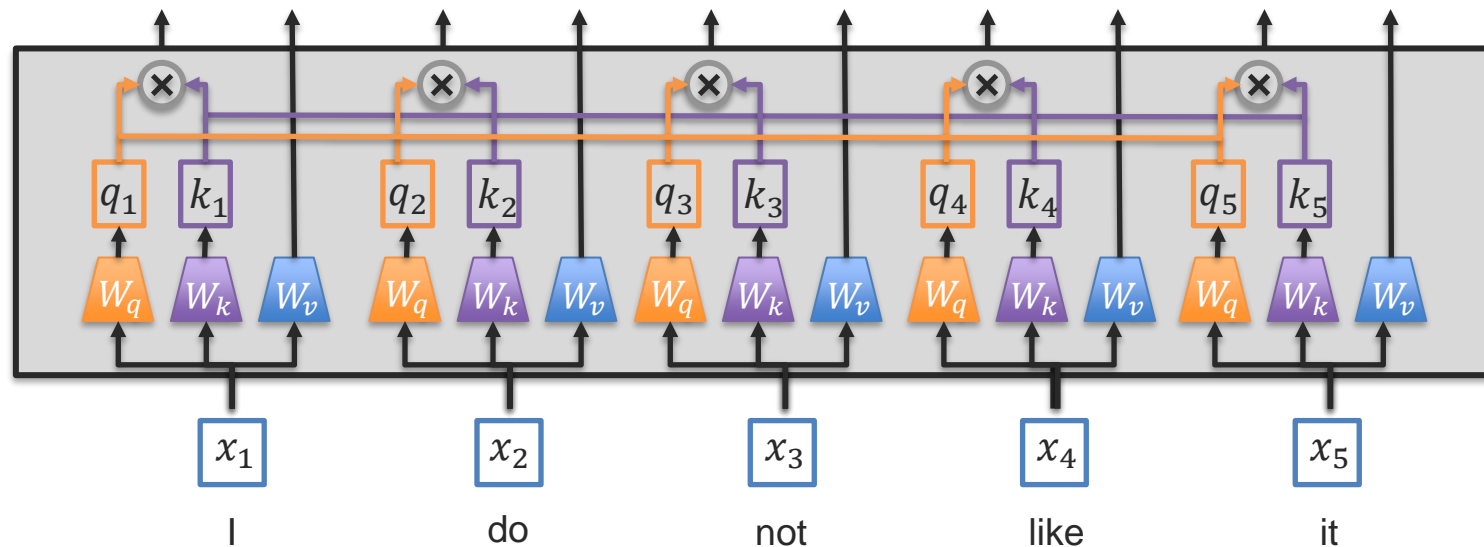


<https://openai.com/blog/clip/>

Going Beyond Sequences: Graph Representations

*slides adapted from Leskovec, Representation Learning on Networks. WWW 2018

Transformers – Fully-Connected Sequences



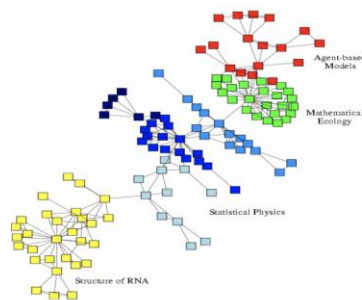
Should everything be connected to everything?

What if we have domain knowledge about connections?

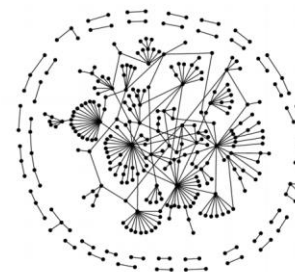
Graphs (aka “Networks”)



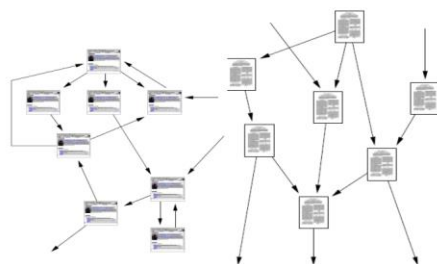
Social networks



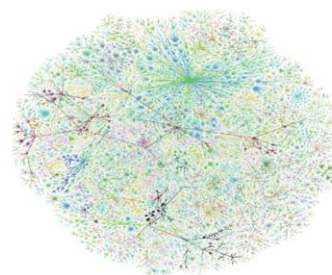
Economic networks



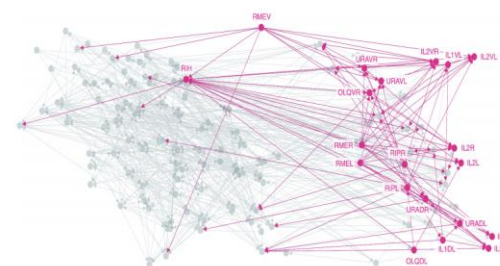
Biomedical networks



Information networks: Web & citations



Internet

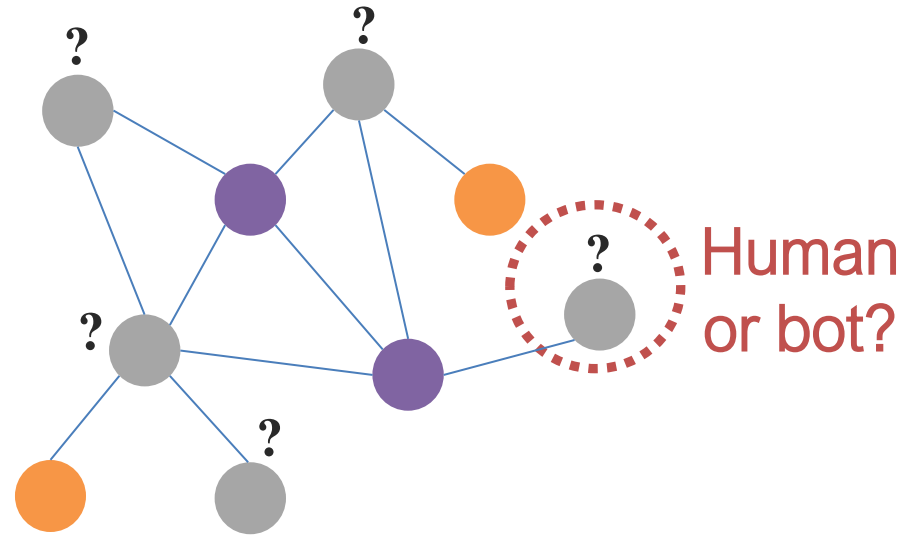


Networks of neurons

Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019

Graphs – Supervised Task

Goal: Learn from labels associated with a subset of nodes (or with all nodes)



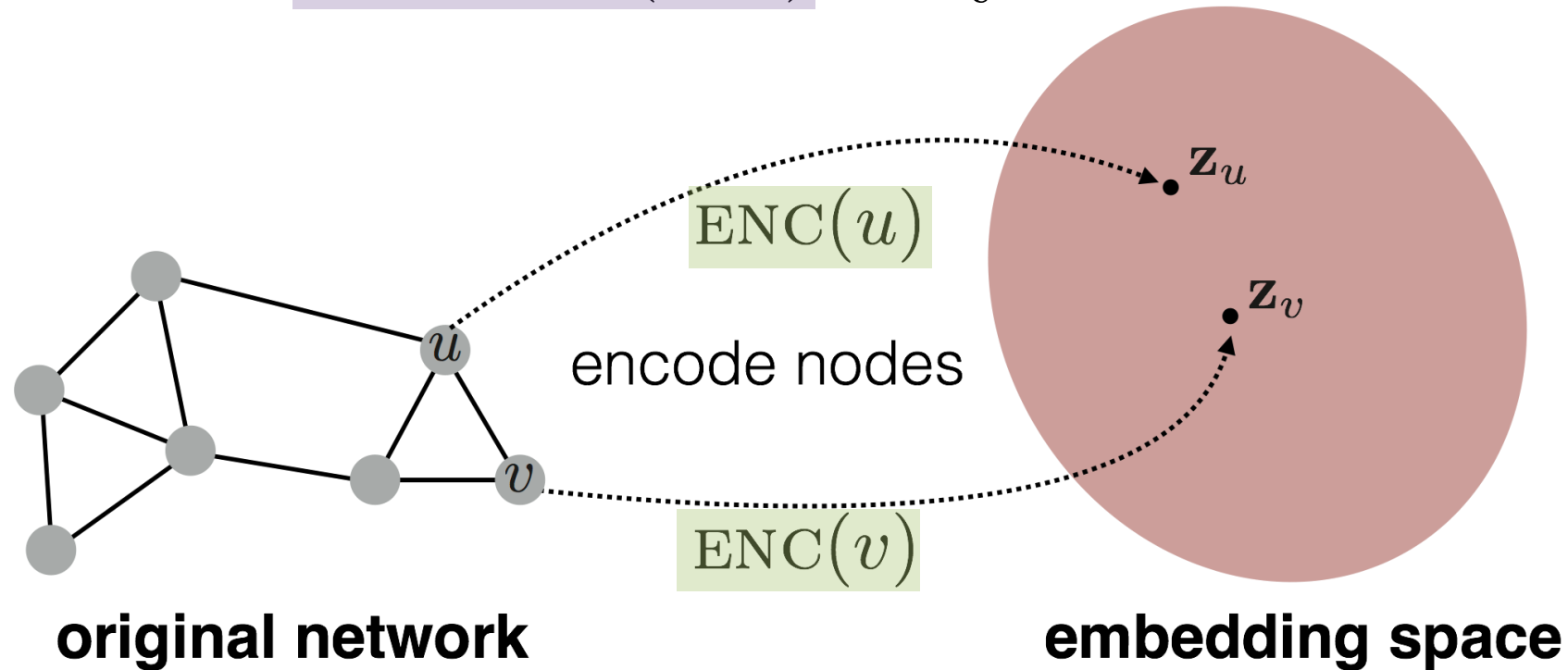
e.g., an online social network

55

Graphs – Unsupervised Task

Goal: Learn an embedding space where

$$\text{similarity}(u, v) \approx \mathbf{z}_v^\top \mathbf{z}_u$$



56

Graph Neural Nets

Assume we have a graph \mathbf{G} :

\mathbf{V} is the set of vertices

\mathbf{A} is the binary adjacency matrix

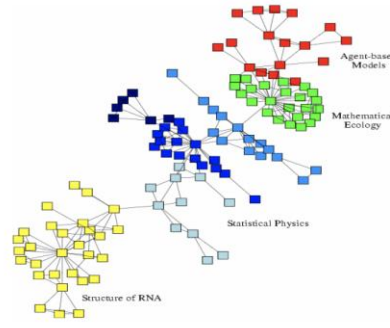
\mathbf{X} is a matrix of node features:

- Categorical attributes, text, image data
e.g. profile information in a social network
- ...

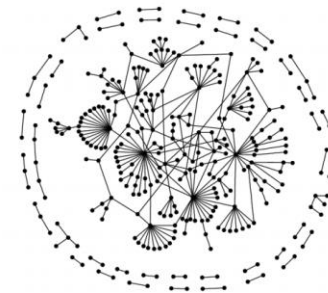
\mathbf{Y} is a vector of node labels (optional)



Social networks



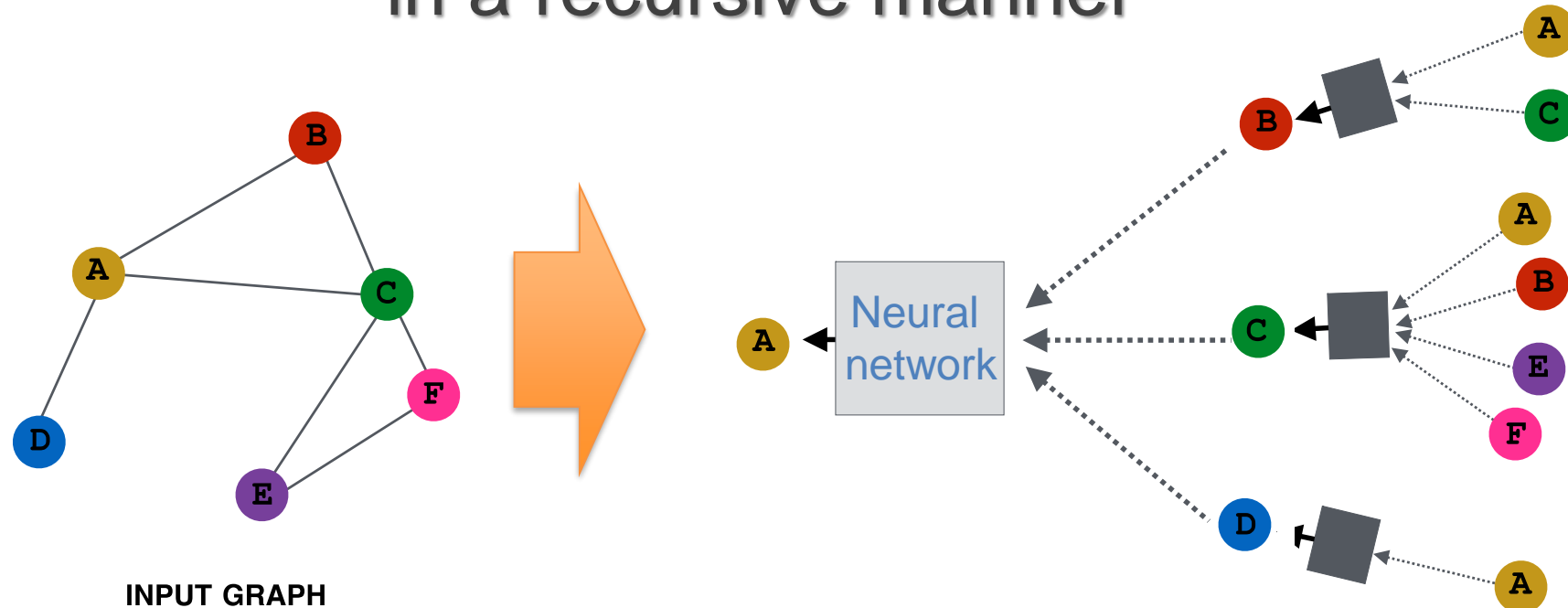
Economic networks



Biomedical networks

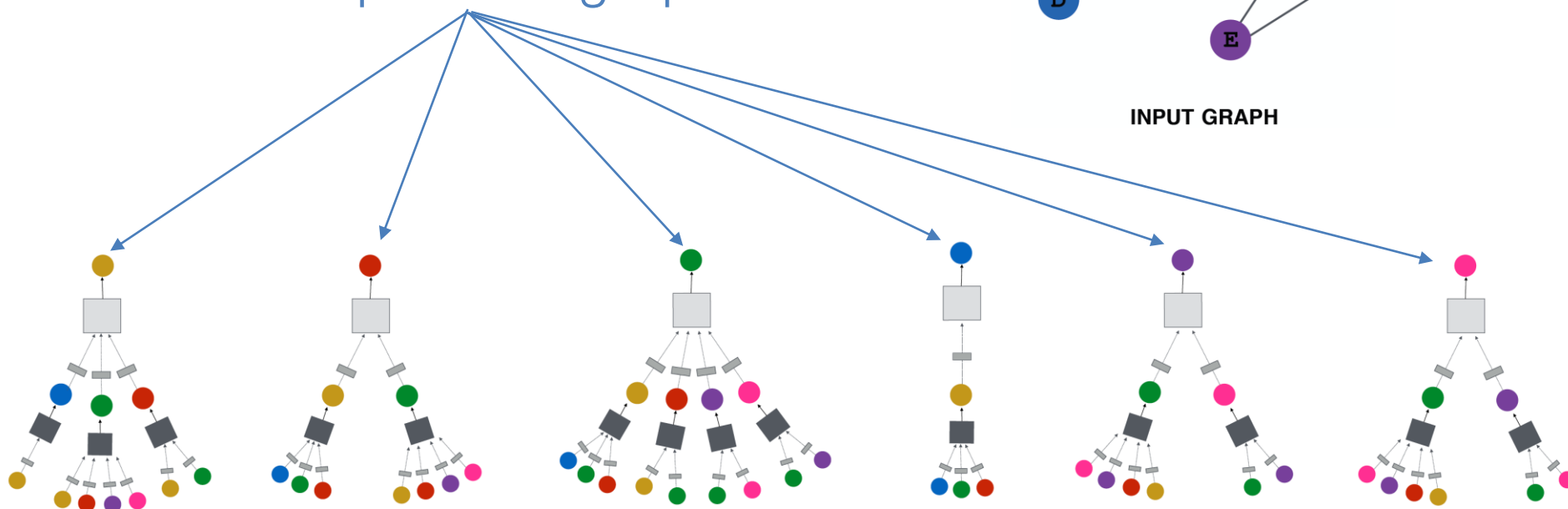
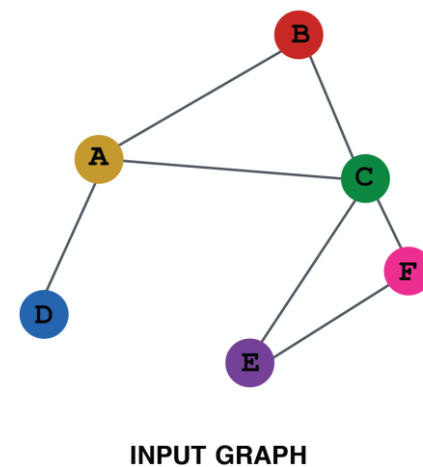
Graph Neural Nets

Key idea: Generate node embeddings based on local neighborhoods in a recursive manner



Graph Neural Nets

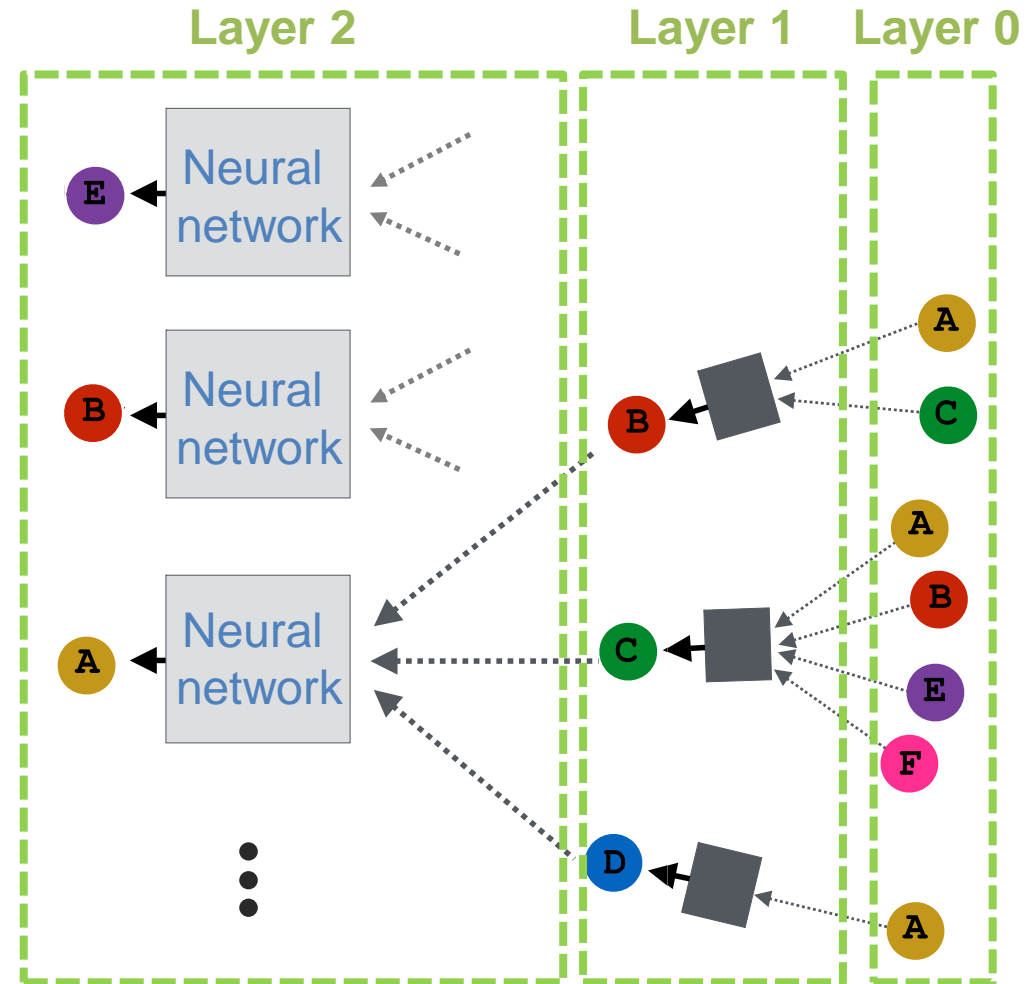
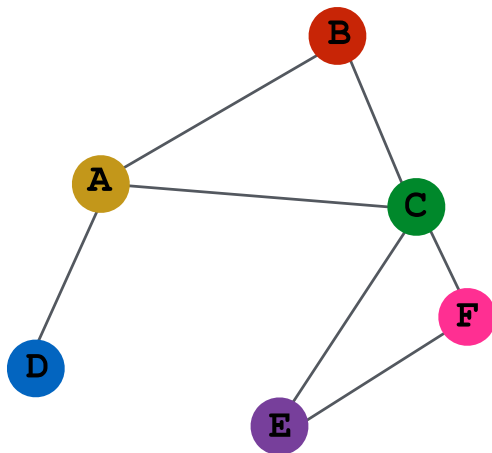
Every node defines a unique
computation graph!



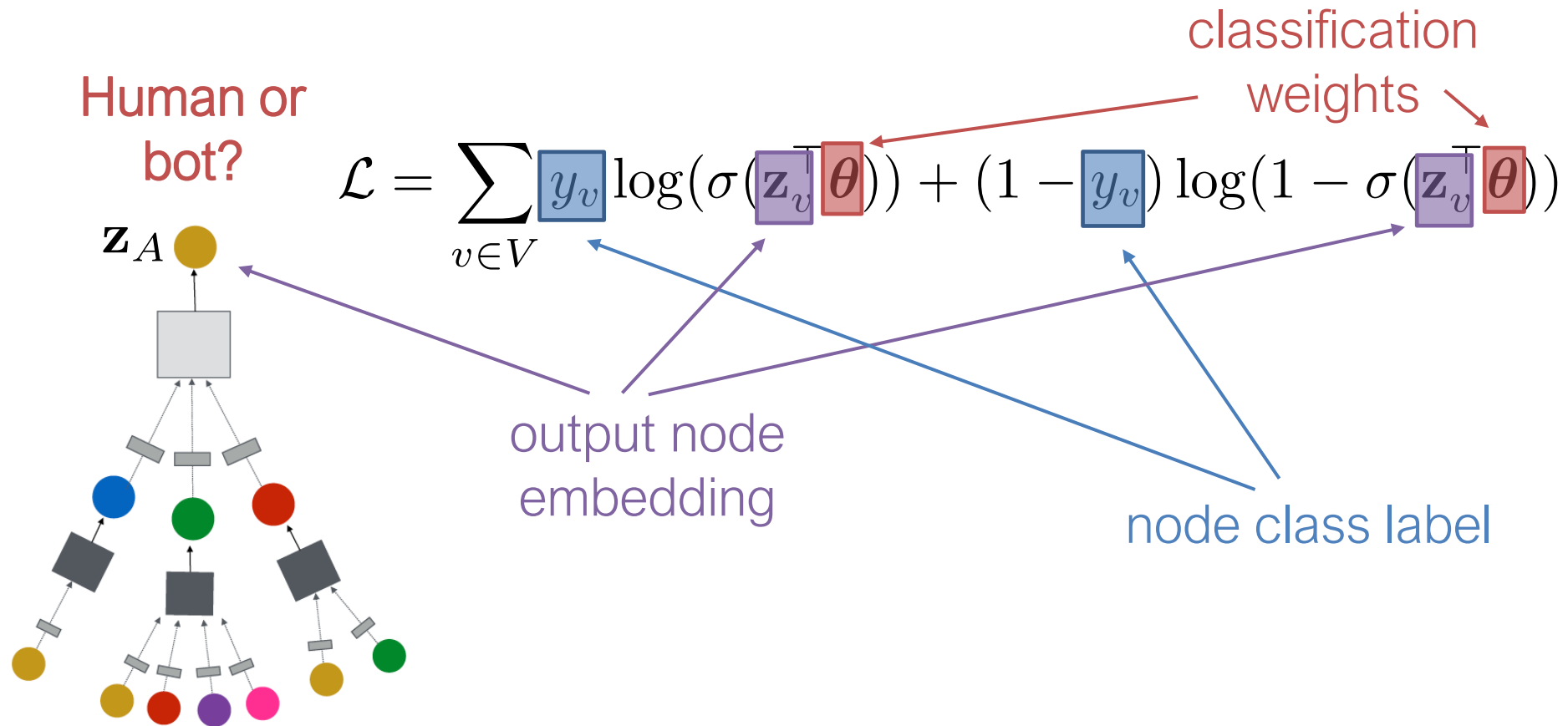
Graph Neural Nets

And multiple layers!

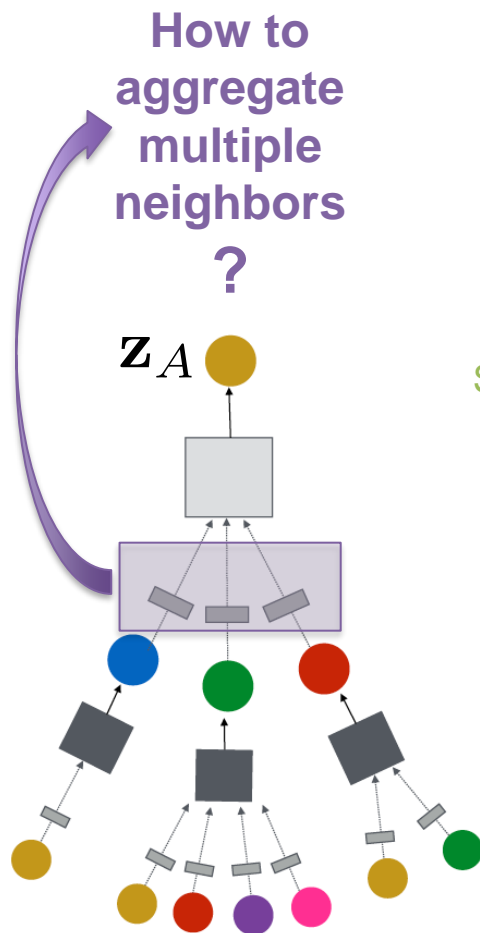
- ➡ Shared parameters within a specific layer
- ➡ “layer-0” is the input feature x_u



Graph Neural Nets – Supervised Training



Graph Neural Nets – Neighborhood Aggregation



Average pooling (Scarselli et al., 2005)

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1} \right)$$

Different weights for neighbors and self

Graph Convolution Network (Kipf et al., 2017)

Same weights

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_u^{k-1}}{\sqrt{|N(u)||N(v)|}} \right)$$

Different normalization

➡ It can be efficiently implemented

Graph Attention Network (Velickovic et al., 2018)

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\alpha_{uv} \mathbf{h}_u^{k-1}}{\sqrt{|N(u)||N(v)|}} \right)$$

Attention weights

➡ Very similar to a self-attention transformer