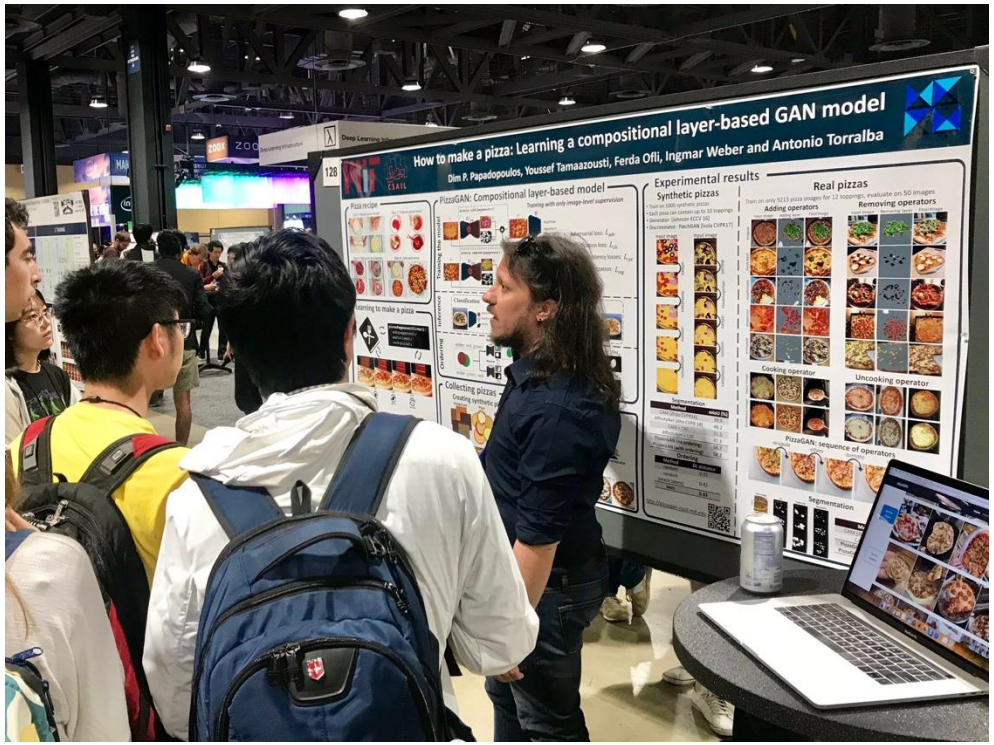
Multimodal Learning

Dimitrios Papadopoulos
Associate Professor, DTU Compute

Dimitrios (Dim) Papadopoulos



- **Associate Professor** at the section of Visual Computing DTU Compute
- **Research interests:** Computer Vision and Deep Learning
- **Prior to DTU:**
  - Postdoc, MIT, MA, USA (2018 - 2021)
  - Visiting student, ETH Zurich, Switzerland (2016 - 2017)
  - PhD, University of Edinburgh, UK (2013 - 2017)
  - MSc and Meng at Democritus University of Thrace (2006-2012)

What is Multimodal Learning?

Multimodal

Dictonary definition…

Multimodal: having or involving several modes or modalities

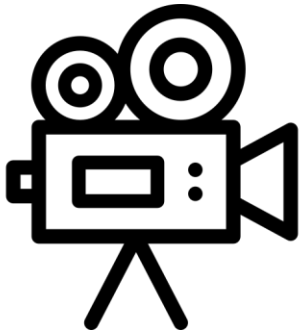
Research-oriented definition…

Multimodal is the scientific study of heterogenous and interconnected data

Connected + Interacting

What is a modality?

Modality *refers to the way in which something expressed or perceived.*



sensor

Raw Modalities (closest to sensor)

Abstract Modalities (farthest from sensor)

Examples:

Speech signal	Language	Sentiment intensity
Image	Detected objects	Object classes
Video	Tracks	3D meshes

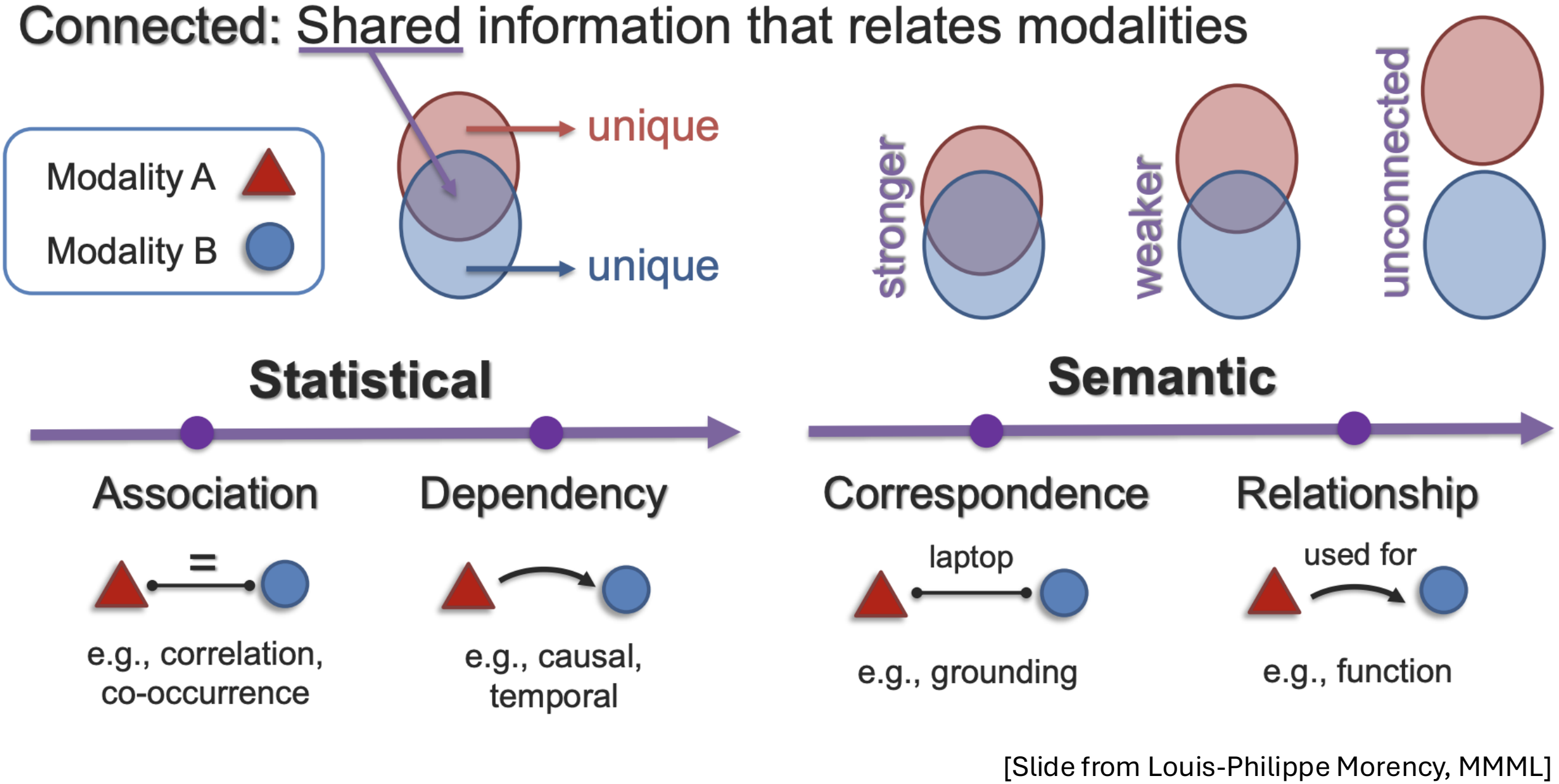
Heterogeneous Modalities

Information present in different modalities will often show diverse qualities, structures and representations

Examples:

Images from 2 cameras	Text from 2 different languages	Vision and language	??? ??? ???

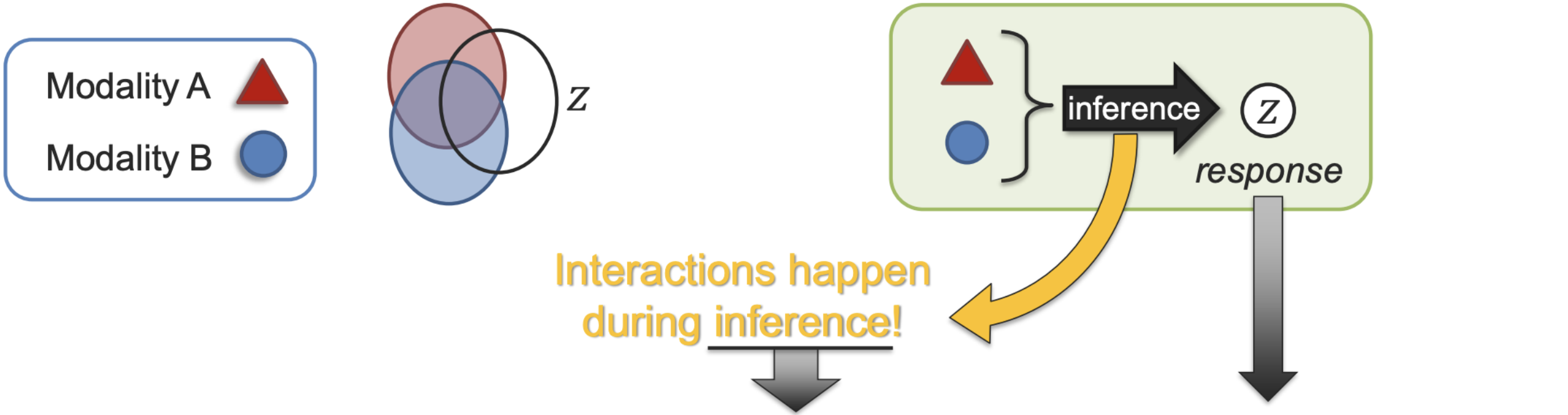
Connected



[Slide from Louis-Philippe Morency, MMML]

Interacting

Interacting: process affecting each modality, creating new response



"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation

representation

$\hat{y}$ *prediction*

modality C

[Slide from Louis-Philippe Morency, MMML]

Multimodal

Multimodal is the scientific study of heterogenous and interconnected data

Connected + Interacting

Multimodal Learning

Multimodal (machine) Learning *is the study of computer algorithms that integrate and process data from multiple modalities, such as images, text, audio, or videos.*

Applications

Visual Question Answering

What is the mustache made of?

Text-to-Image Generation

“a cute cat in Copenhagen”

Cross-modal retrieval

Ingredients

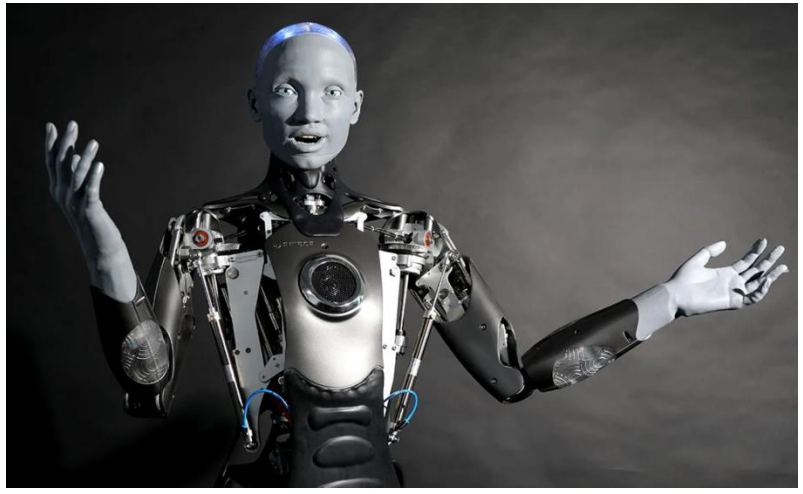
- *(8 ounce) package linguini pasta*
- *½ pound sweet Italian sausage*
- *2 red bell peppers, chopped*
- *1 onion, chopped*
- *1 clove garlic, minced*
- *1 cup white wine*
- *¼ cup grated Parmesan cheese*

Instructions

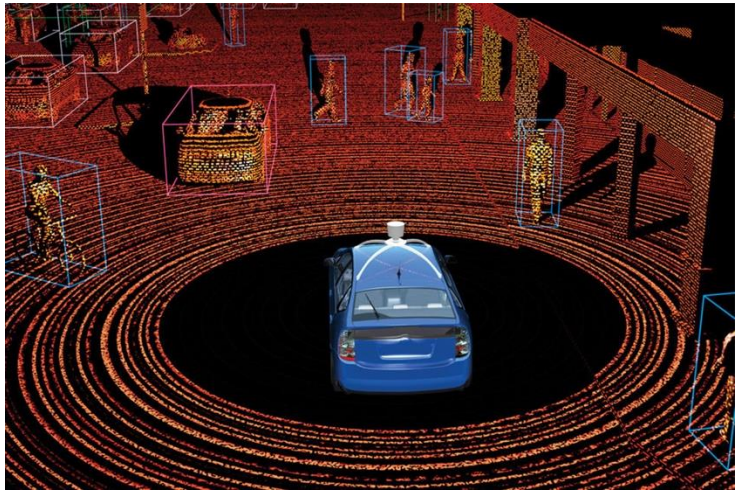
- *Cook pasta in a large pot of boiling salted water until al dente.*
- *While the pasta is cooking, prepare the sauce.*
- *Sauté sausages in a heavy skillet over medium high heat until light brown, breaking up clumps with back of spoon.*
- *Add peppers, onion, and garlic; saute until tender.*
- *Add wine and simmer until liquid is slightly reduced, about 6 minutes.*
- *Drain pasta, and add to the skillet.*
- *Toss to combine.*
- *Serve.*

Healthcare

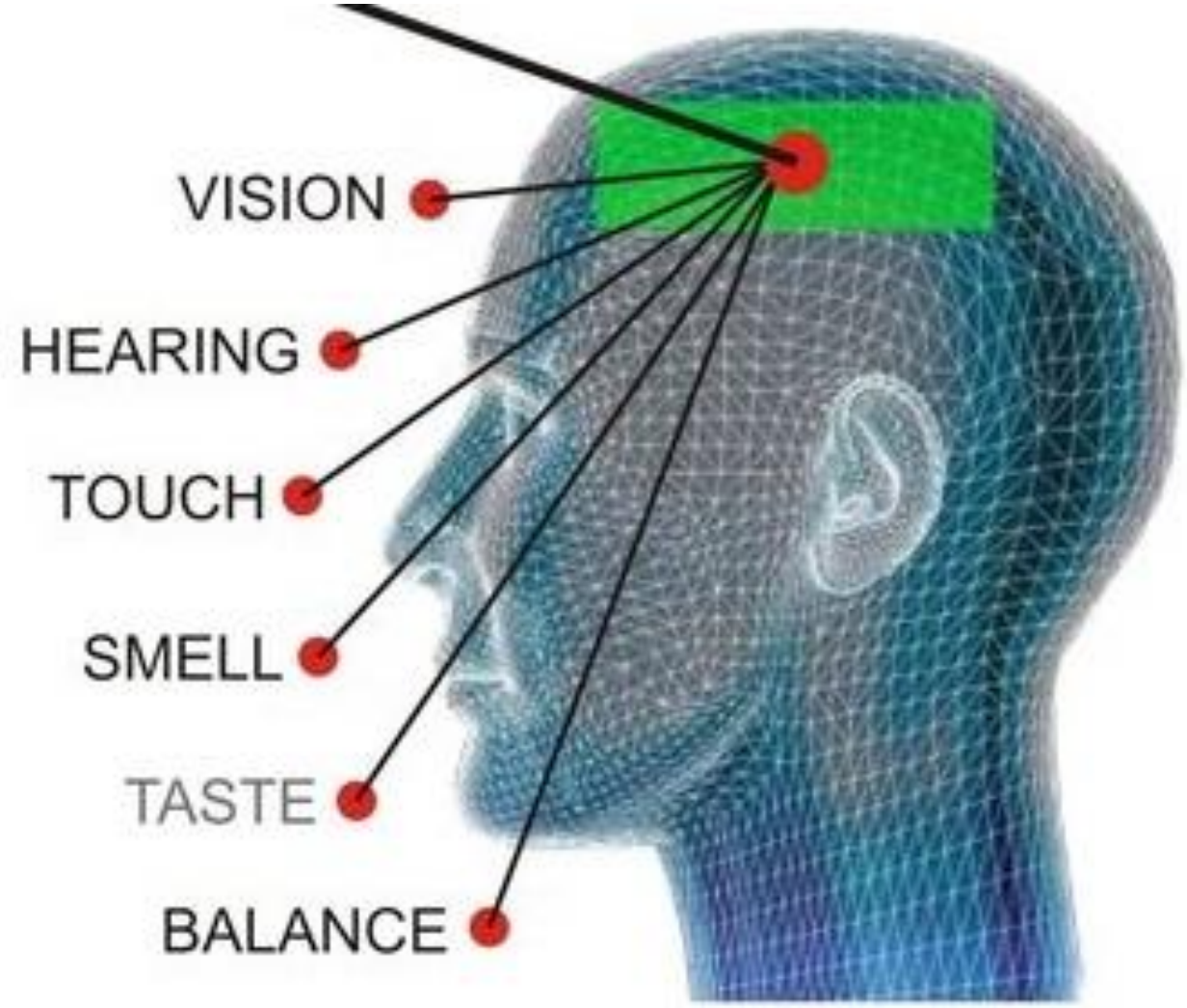
Robotics



Autonomous Driving



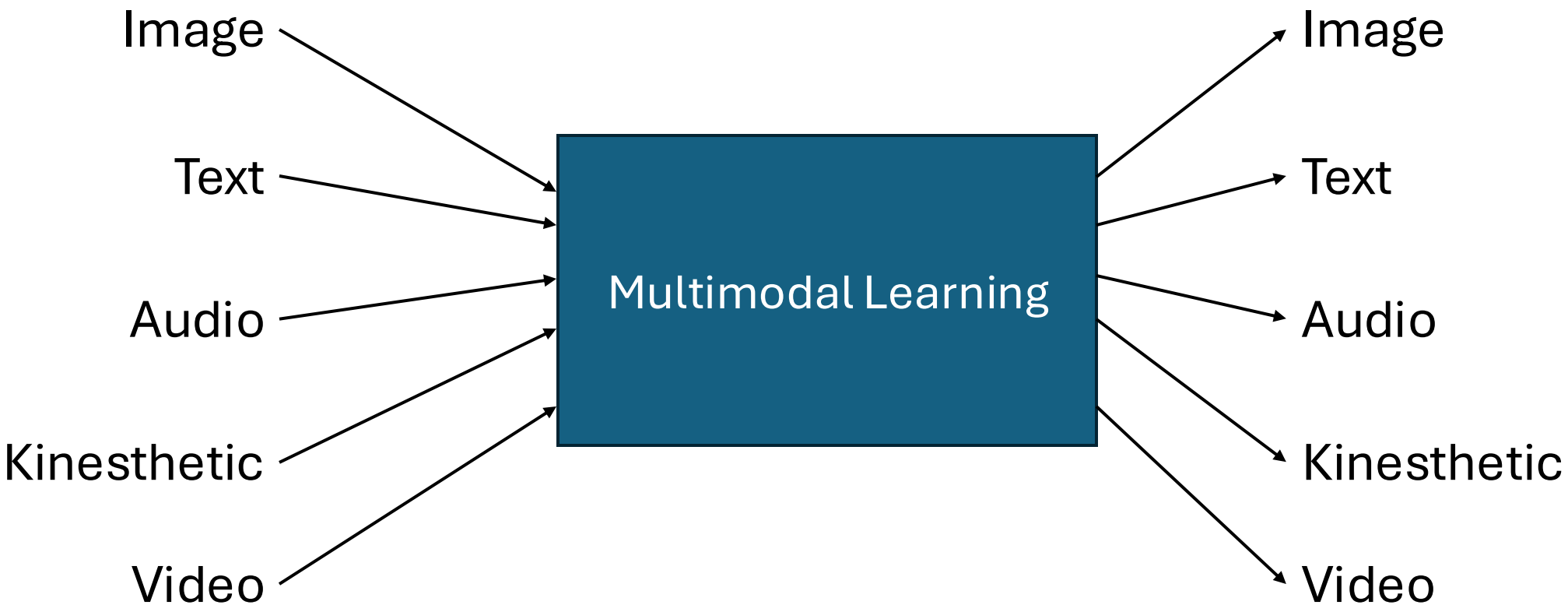
Why Multimodal Learning?



***Human perception** is inherently multimodal*

Humans explore the world through diverse senses: sight, sound, touch, and scent.

Multimodal Learning framework



ChatGPT

What do you see in the image?

The image shows a large group of people gathered outdoors on a grassy area. They are standing and sitting close together, smiling, and raising their hands in the air, giving a celebratory or group photo vibe. The setting appears casual and friendly, possibly for an event like a workshop, summer school, team-building day, or community gathering. In the background,

Generate a bounding box around the tree of the image

Estimate the number of people in the image

2/2

By counting roughly row by row, it looks like there are about 80–90 people in the image. My closer estimate is around **85 people**.

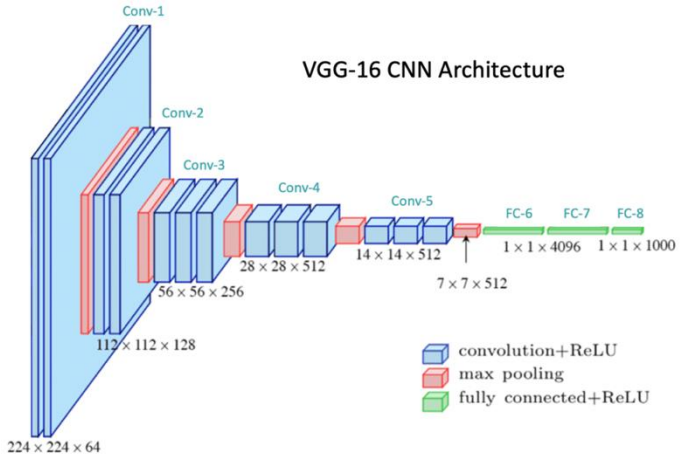
Generate a similar image with double the amount of people

Background

(disclaimer)

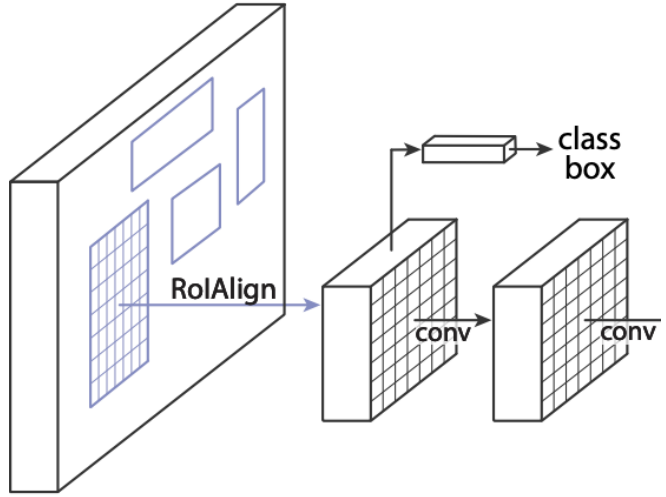
Image Recognition

IMAGE CLASSIFICATION



cat

OBJECT DETECTION



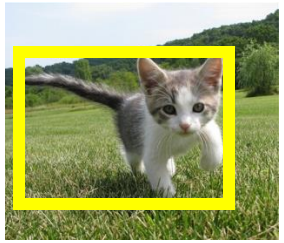
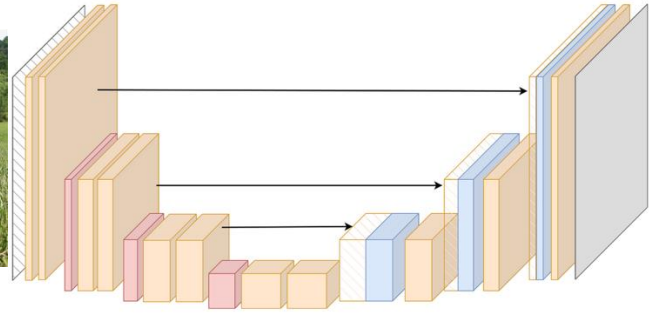
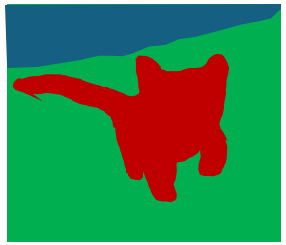
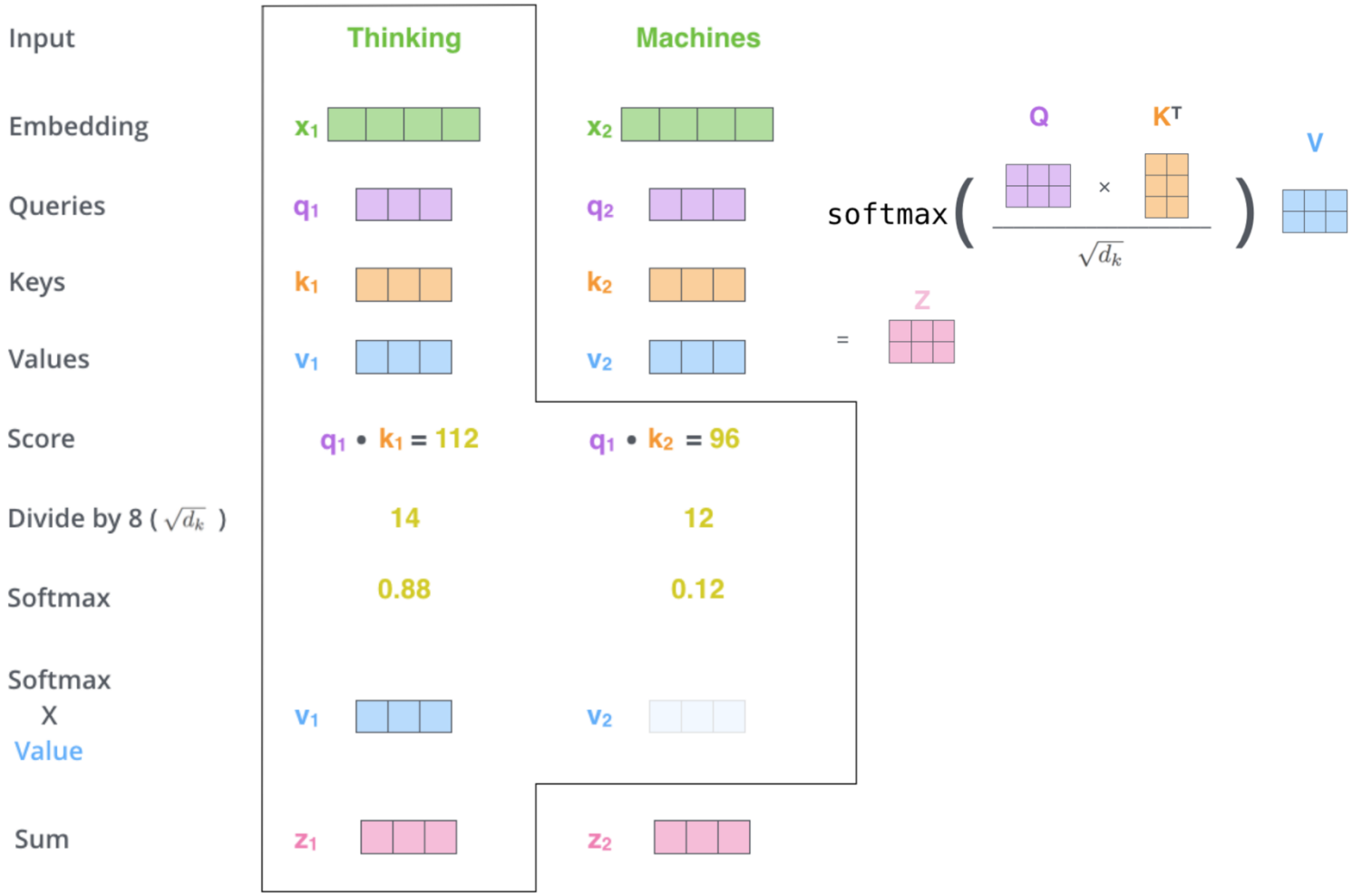


IMAGE SEGMENTATION

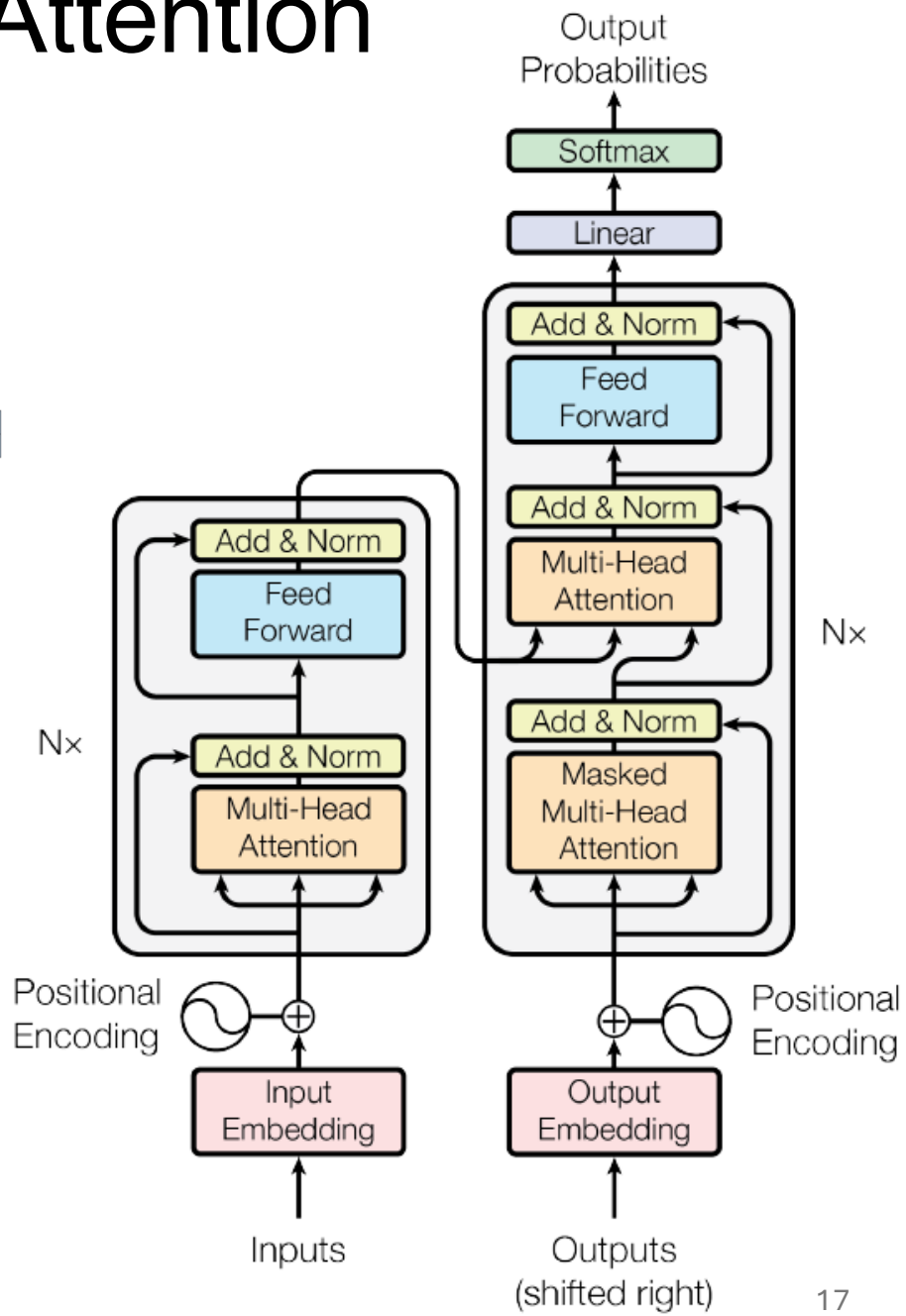




Transformers and Attention



$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V = Z$$



[Slides from Illustrated Transformer]

VIT (Vision Transformer)

An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

Alexey Dosovitskiy[*,†], **Lucas Beyer**[*], **Alexander Kolesnikov**[*], **Dirk Weissenborn**[*], **Xiaohua Zhai**[*], **Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby**[*,†]

[*]equal technical contribution, [†]equal advising

Google Research, Brain Team

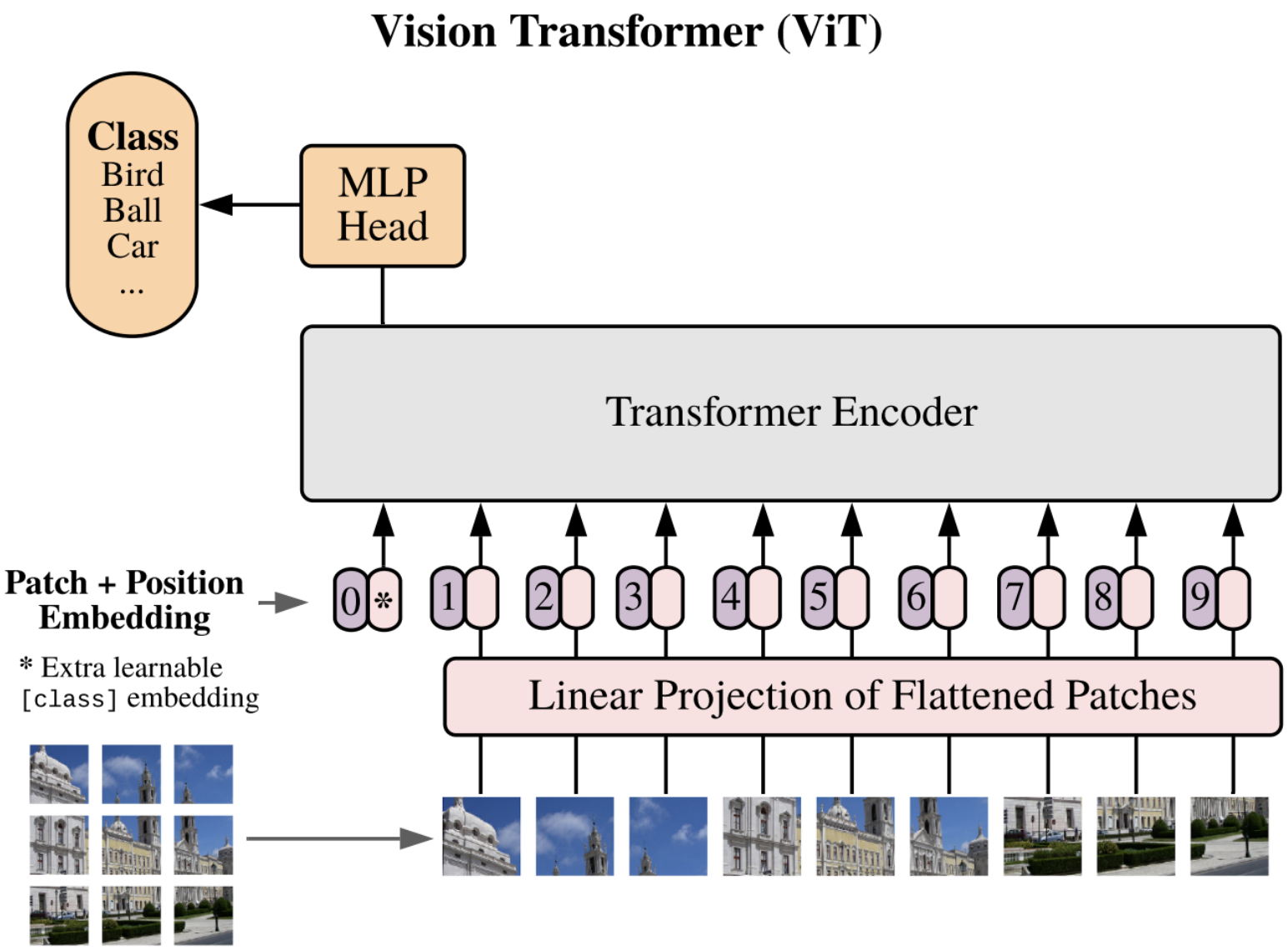
{adosovitskiy, neilhoulsby}@google.com

Abstract

While the Transformer architecture has become the de-facto standard for natural language processing tasks, its applications to computer vision remain limited. In vision, attention is either applied in conjunction with convolutional networks, or used to replace certain components of convolutional networks while keeping their overall structure in place. We show that this reliance on CNNs is not necessary and a pure transformer applied directly to sequences of image patches can perform very well on image classification tasks. When pre-trained on large amounts of data and transferred to multiple mid-sized or small image recognition benchmarks (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) attains excellent results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.[1]

1 Introduction

Self-attention-based architectures, in particular Transformers (Vaswani et al., 2017), have become the model of choice in natural language processing (NLP). The dominant approach is to pre-train on a large text corpus and then fine-tune on a smaller task-specific dataset (Devlin et al., 2019). Thanks to Transformers' computational efficiency and scalability, it has become possible to train models of



Overview

A. Tutorial: Core Multimodal Learning Paradigms

- Representation Learning (fusion-based, joint learning, cross-modal retrieval, etc)
- Alignment (semantic and temporal, visual grounding)
- Generation (Image captioning, text-to-image, text-to-speech, VQA)
- Leveraging Large Language Models

B. Past and Ongoing Research on Multimodal Learning

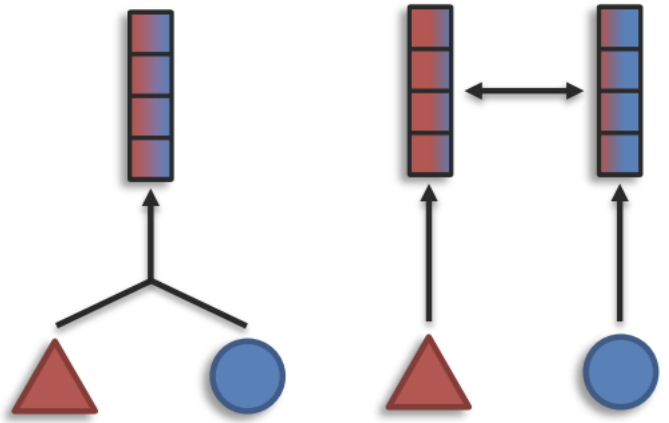
- Cross-modal retrieval [CVPR 2022, CVPR 2019, Submitted 2025]
- Image generation and editing [CVPR-W 2025, CVPR-W 2024, SCIA 2025]
- Test-time scaling and augmentation [WIP… 2025]

Modality A

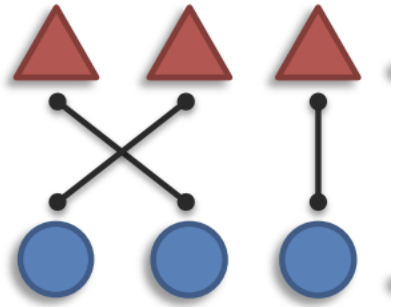
Modality B

Multimodal learning paradigms

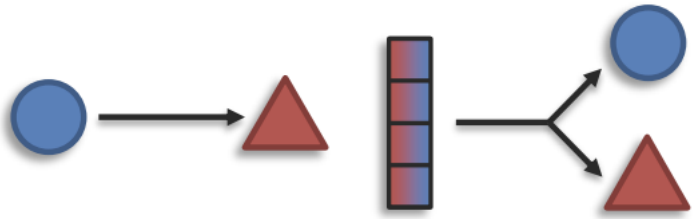
REPRESENTATION: Learning representations that reflect cross-modal interactions between individual elements, across different modalities



ALIGNMENT: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure



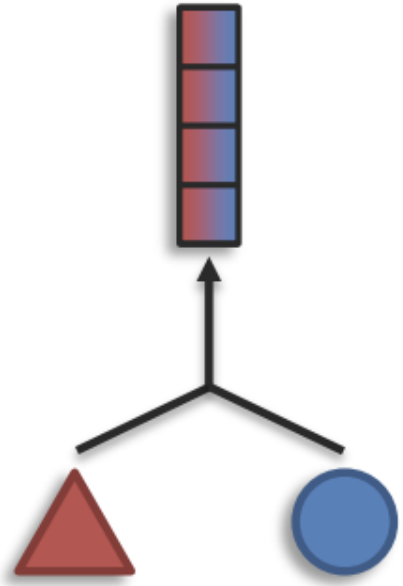
GENERATION: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence



Multimodal Representation Learning

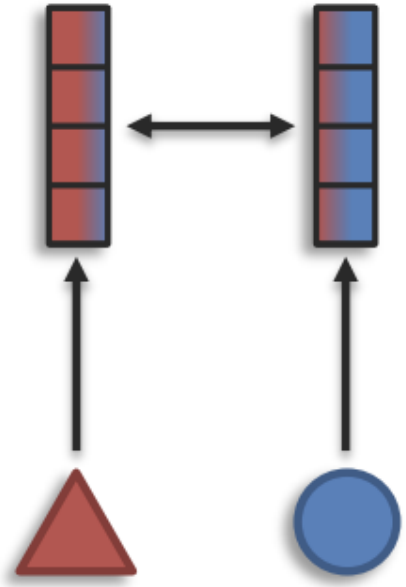
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities.

Fusion



- modalities **>** representations
- joint representation
- Multimodal classification and prediction

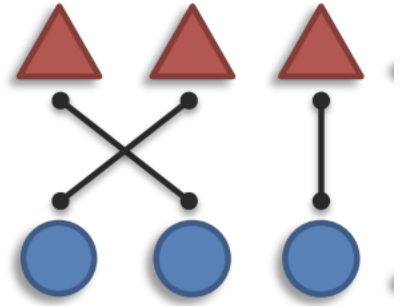
Coordination



- modalities **=** representations
- multimodally-contextualized representations
- Cross-modal retrieval, zero-shot capabilities

Multimodal Alignment

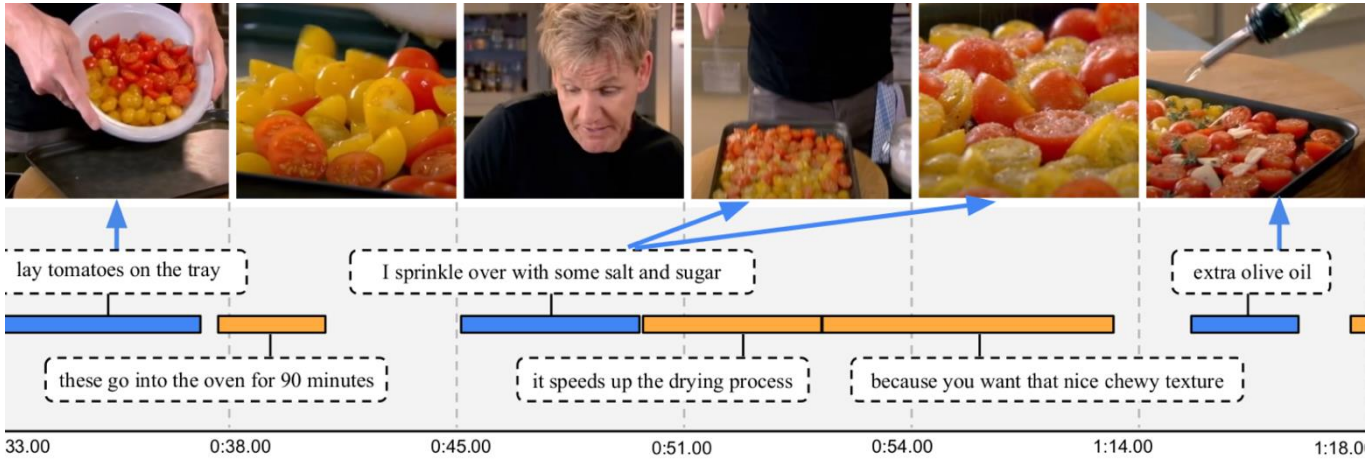
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure



Semantic alignment

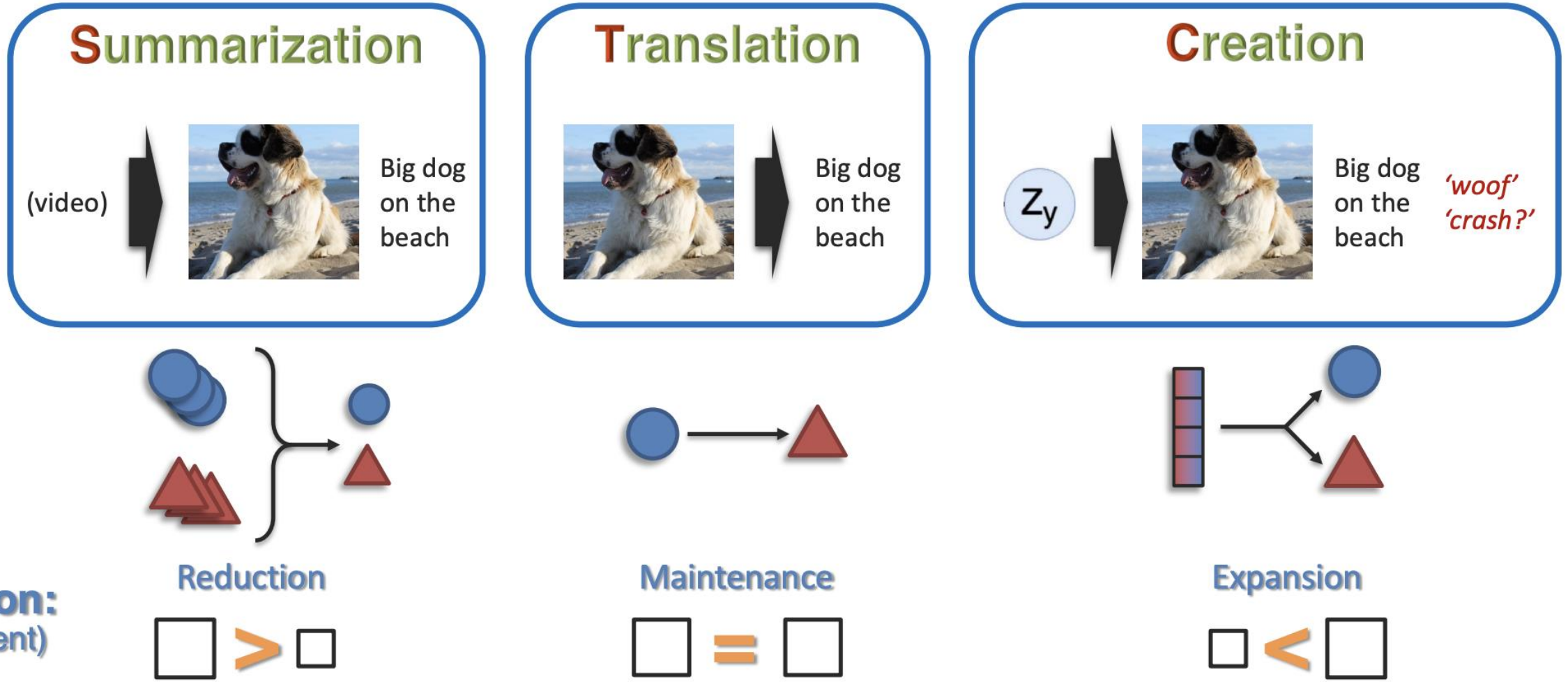
A dog is lying on the grass next to a frisbee.

Temporal alignment



Generation

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence



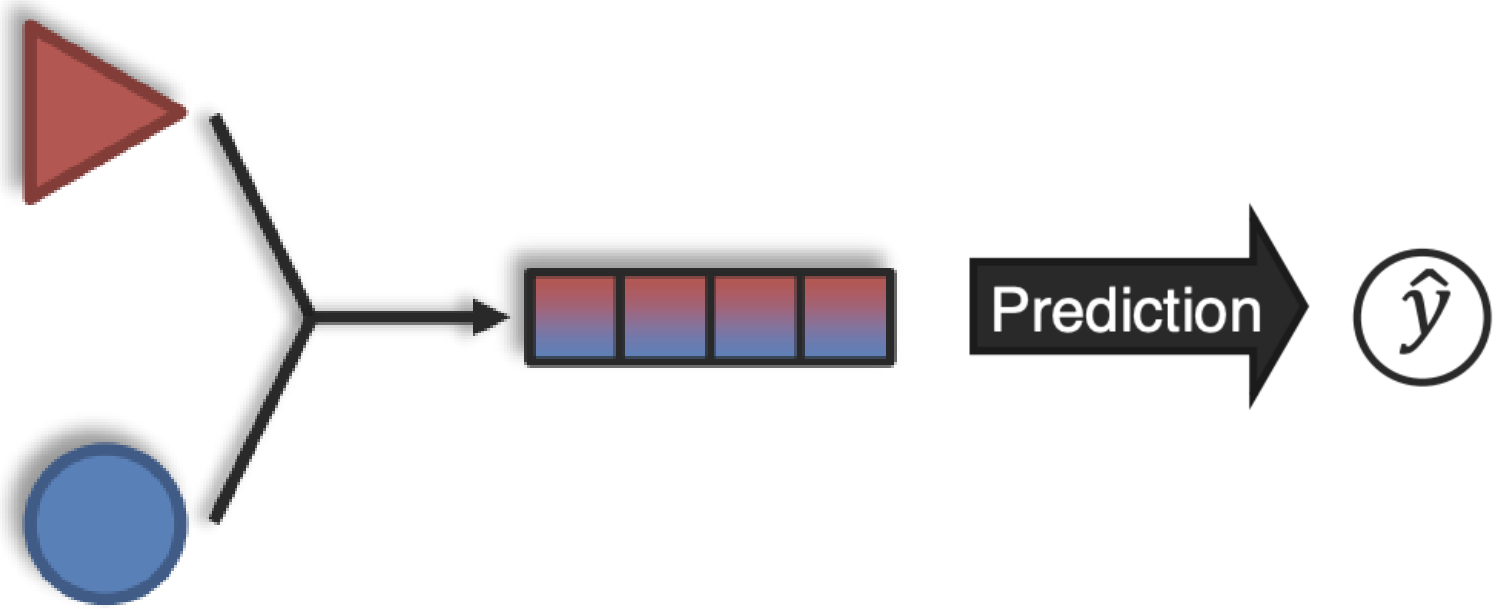
Multimodal tasks (=Vision and Language Tasks)

- Multimodal Classification
- Image-Text Retrieval
- Visual Grounding
- Visual Question Answering and Visual Reasoning
- Image Captioning
- Text-to-image Generation

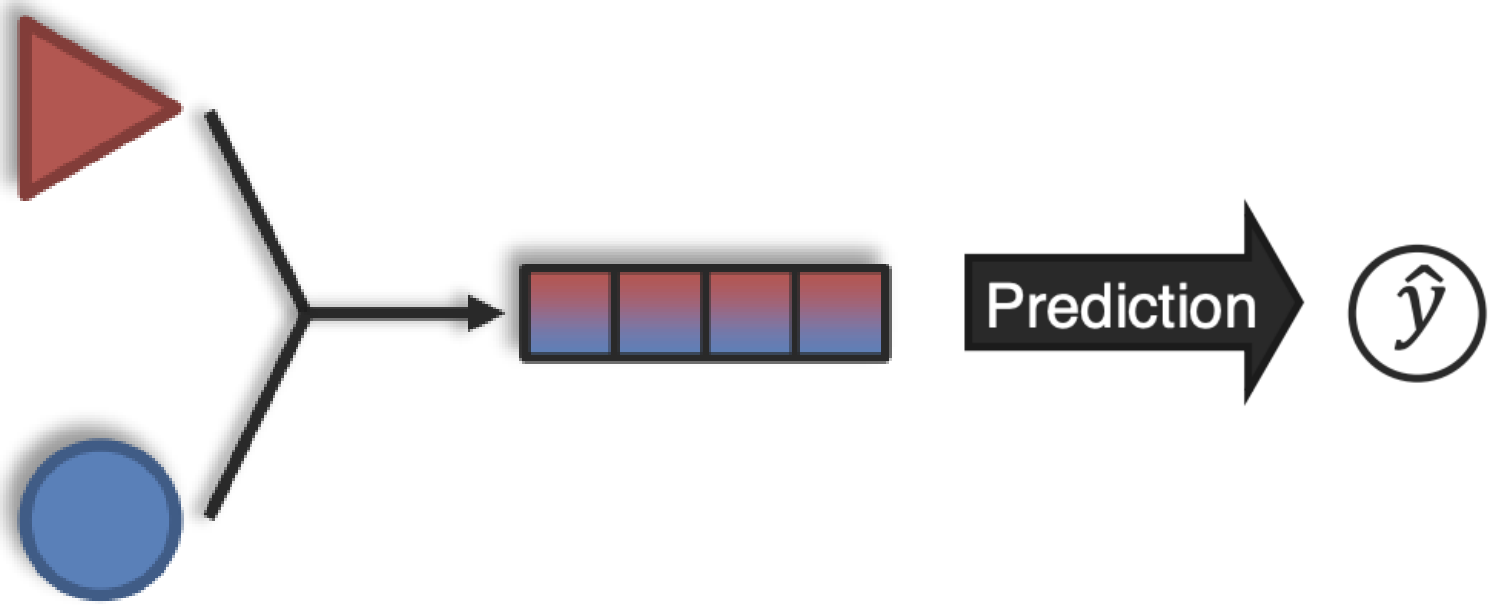
Multimodal tasks (=Vision and Language Tasks)

- **Multimodal Classification**
- Image-Text Retrieval
- Visual Grounding
- Visual Question Answering and Visual Reasoning
- Image Captioning
- Text-to-image Generation

Multimodal Classification



Multimodal Classification

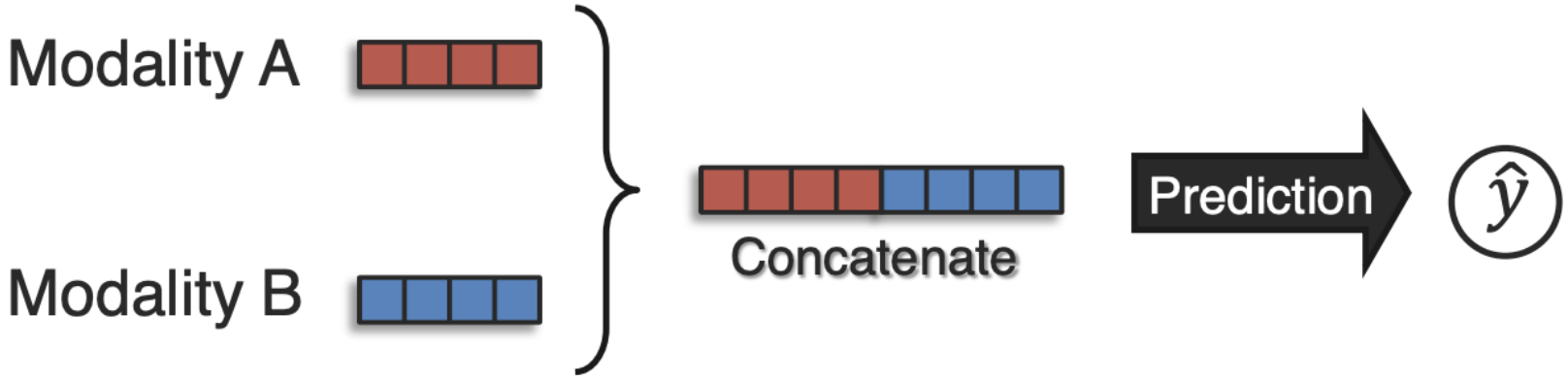


1) Where to fuse?

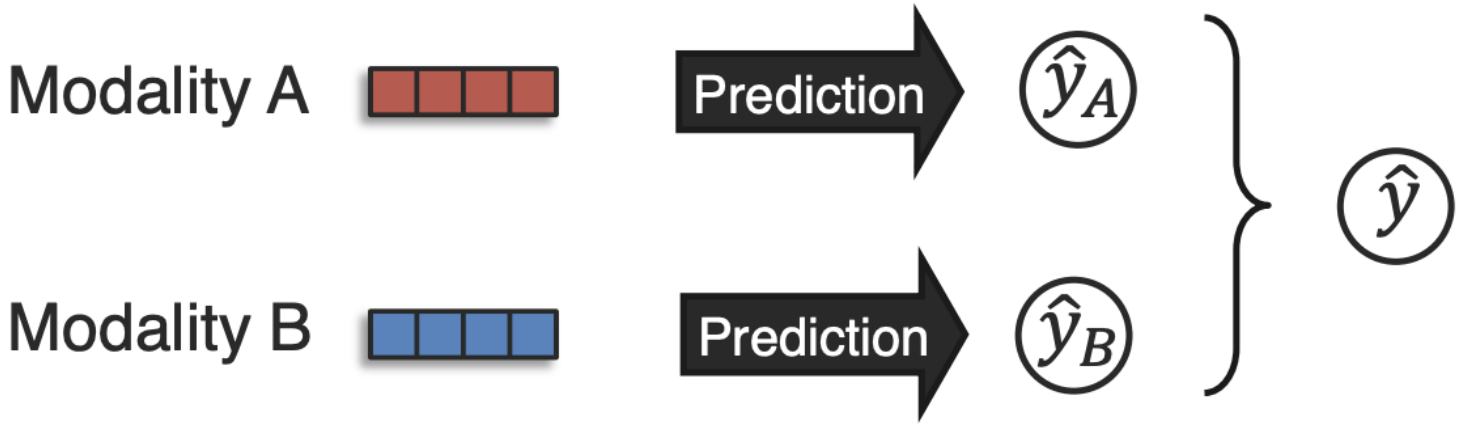
2) How to fuse?

Multimodal Classification

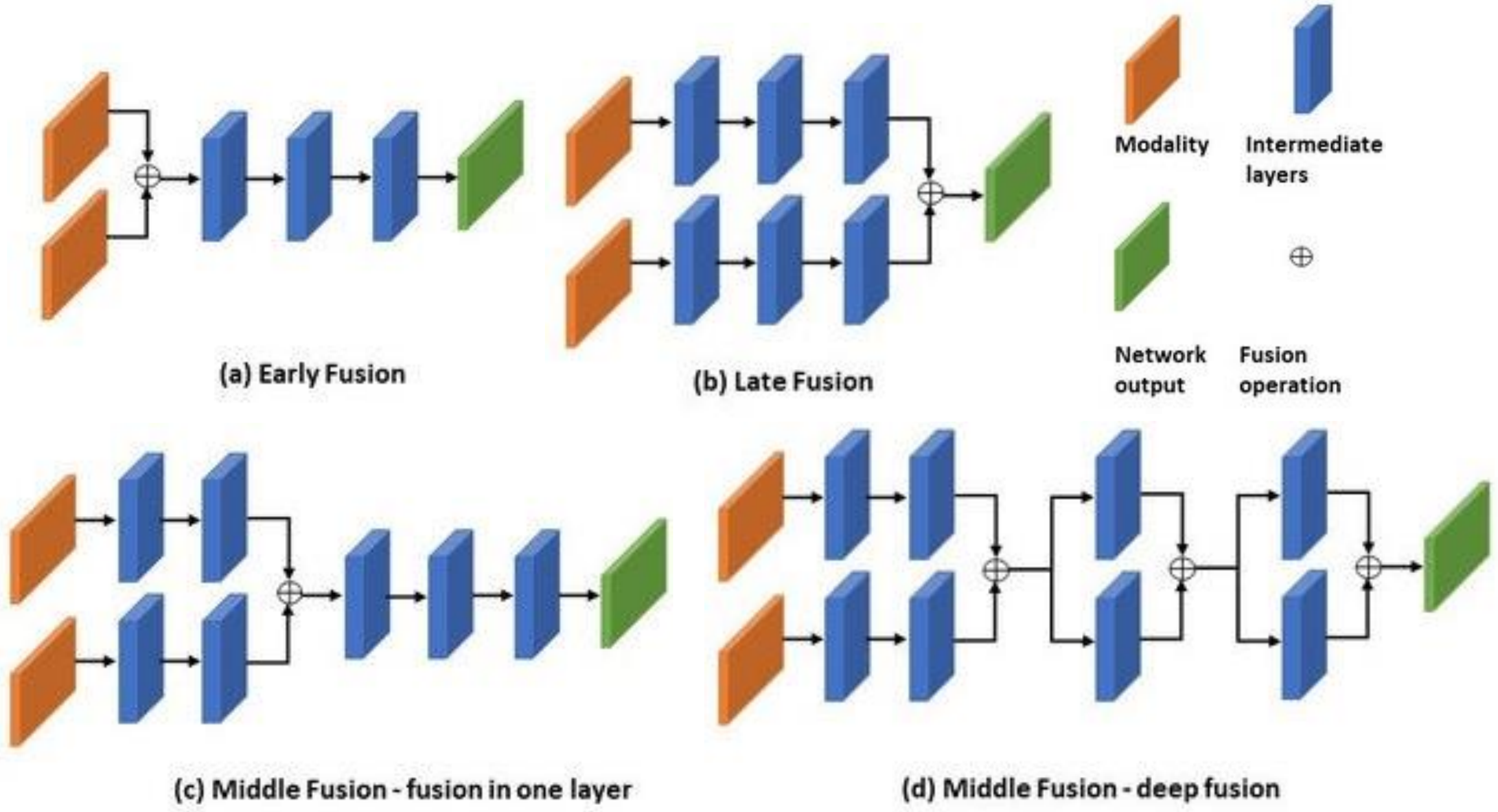
Early fusion:



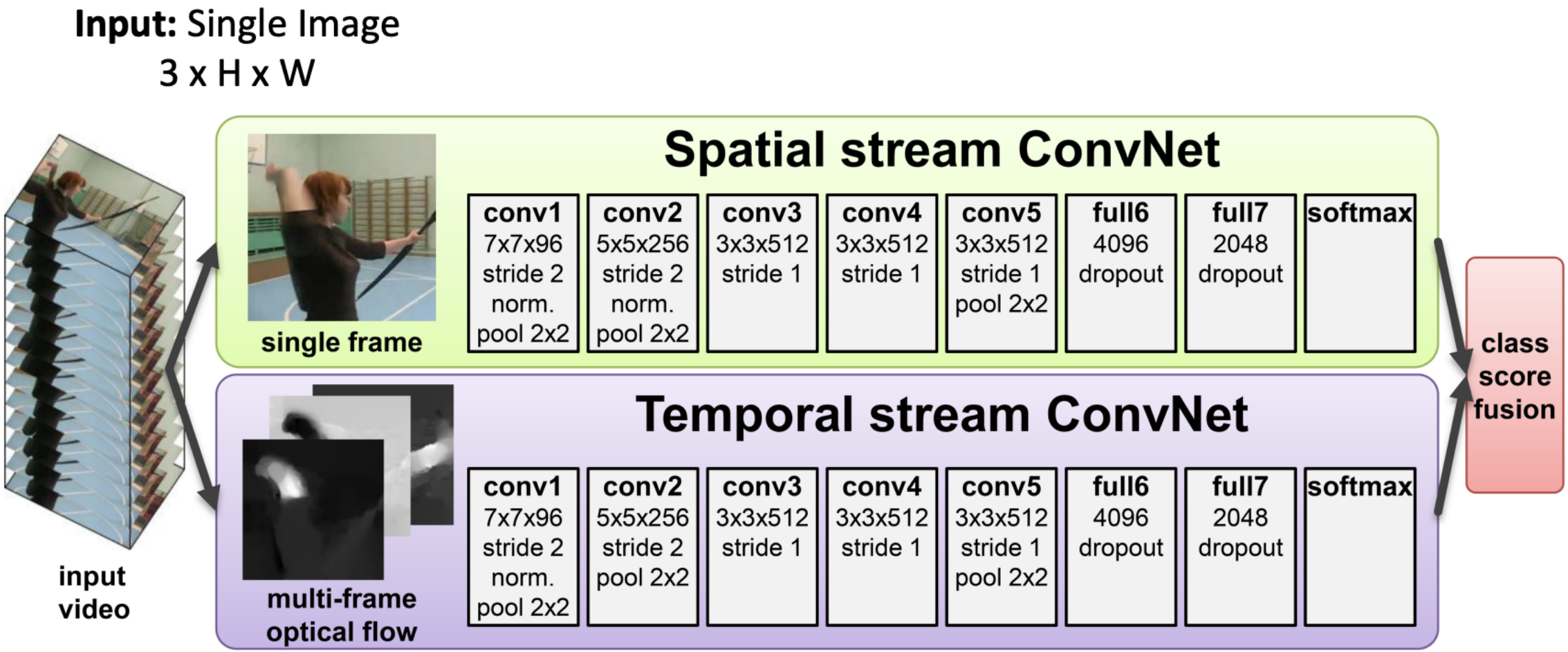
Late fusion:



Multimodal Classification



Video Understanding



[Simonyan and Zisserman, "Two-stream convolutional networks for action recognition in videos", NeurIPS 2014]

Look, Listen and Learn (= AudioVisual Classification)

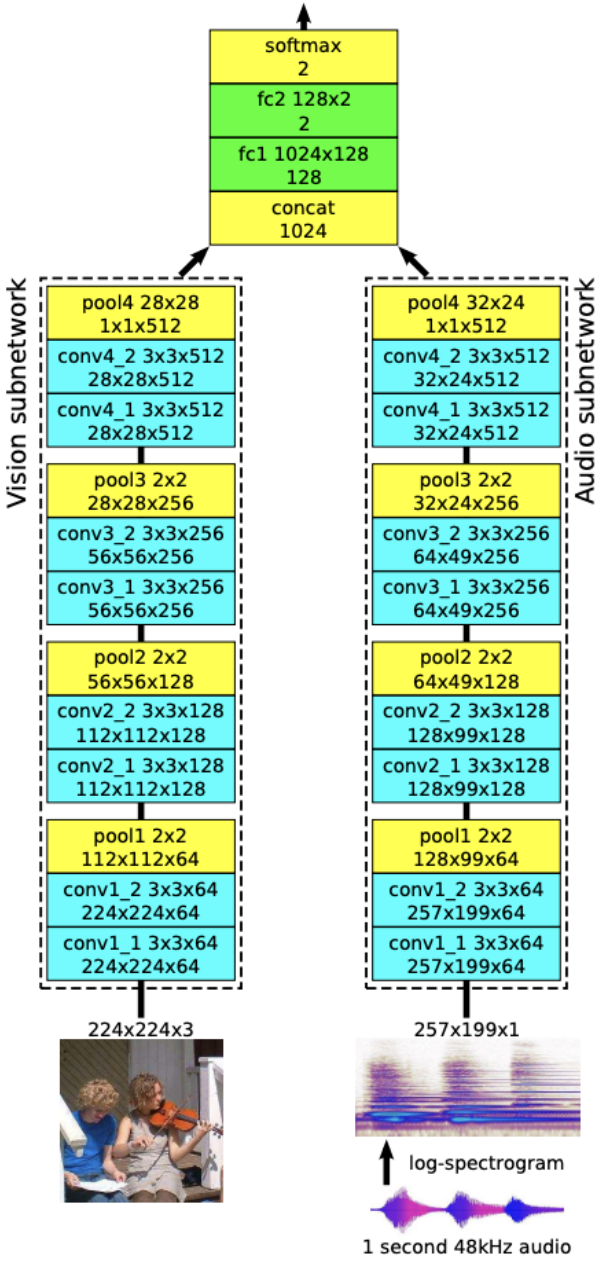
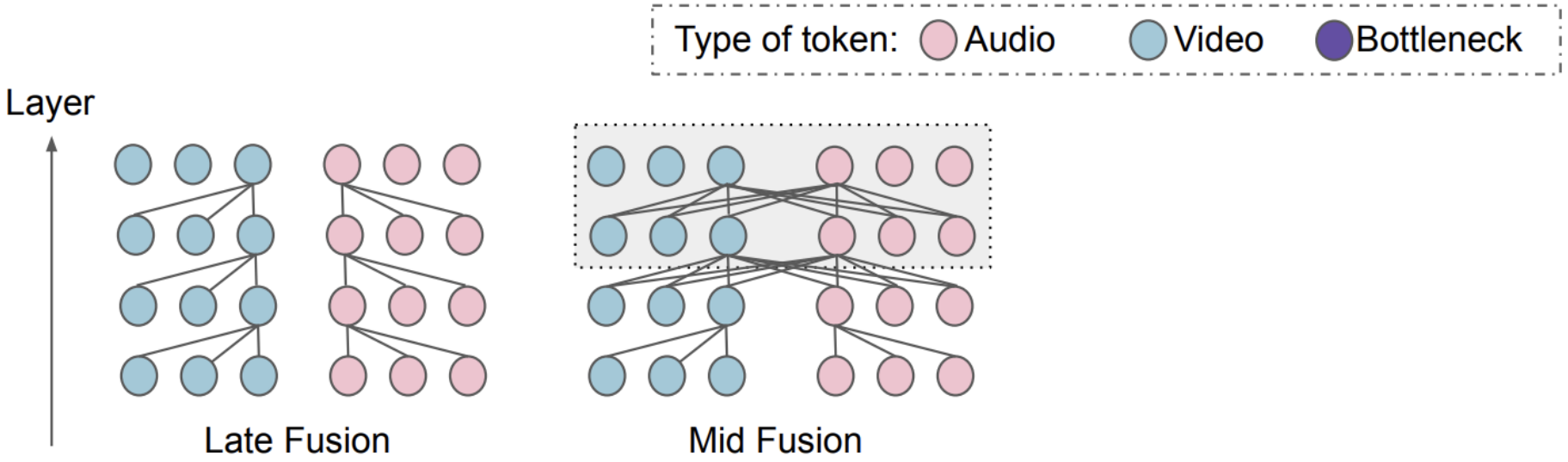


Figure 2. L^3**-Net architecture**. Each blocks represents a single layer with text providing more information – first row: layer name and parameters, second row: output feature map size. Layers with a name prefix `conv`, `pool`, `fc`, `concat`, `softmax` are convolutional, max-pooling, fully connected, concatenation and softmax layers, respectively. The listed parameters are: conv – kernel size and number of channels, pooling – kernel size, fc – size of the weight matrix. The stride of pool layers is equal to the kernel size and there is no padding. Each convolutional layer is followed by batch normalization [13] and a ReLU nonlinearity, and the first fully connected layer (`fc1`) is followed by ReLU.

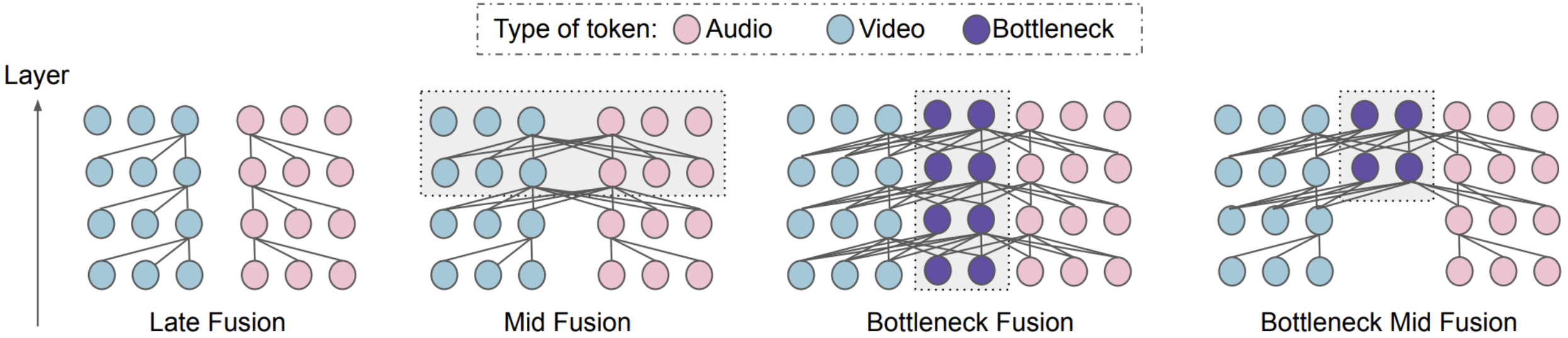
[Arandjelovic and Zisserman, ICCV 2017]

Multimodal Bottleneck Transformer



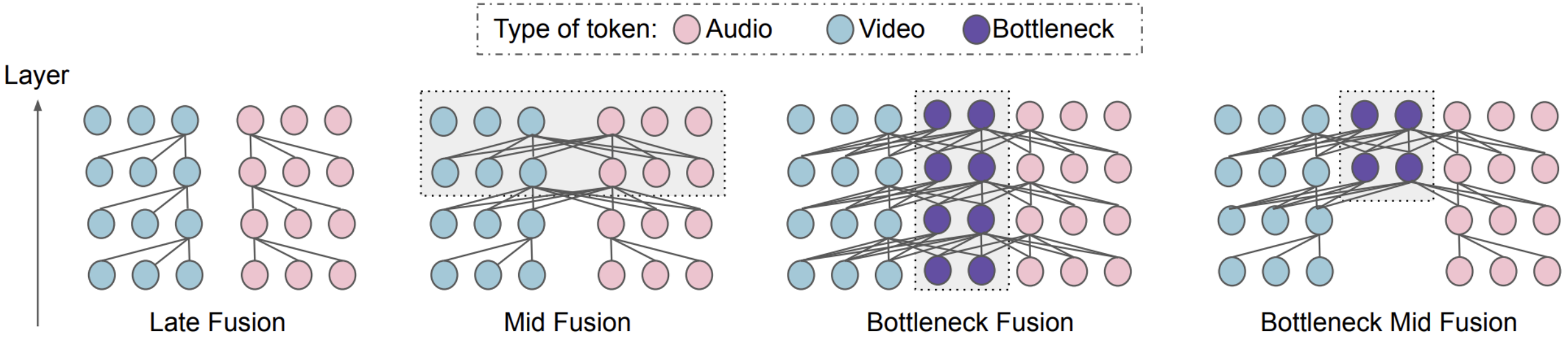
[Nagrani et al, NeurIPS 2022]

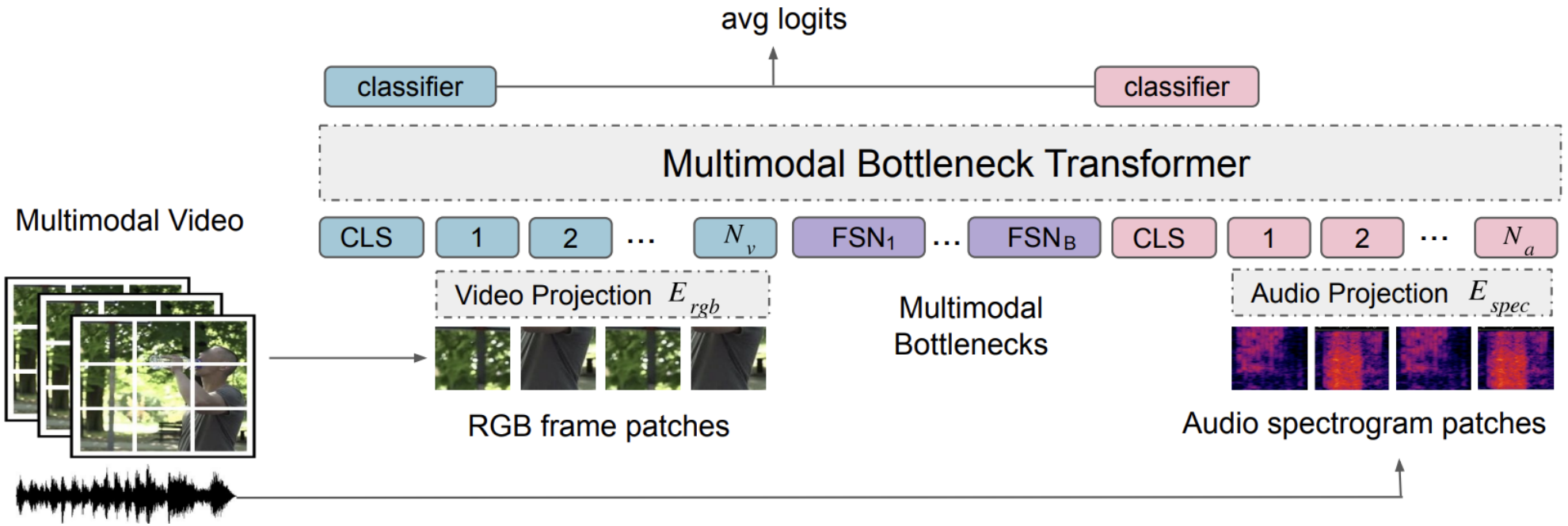
Multimodal Bottleneck Transformer



[Nagrani et al, NeurIPS 2022]

Multimodal Bottleneck Transformer



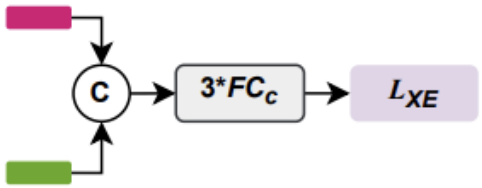


[Nagrani et al, NeurIPS 2022]

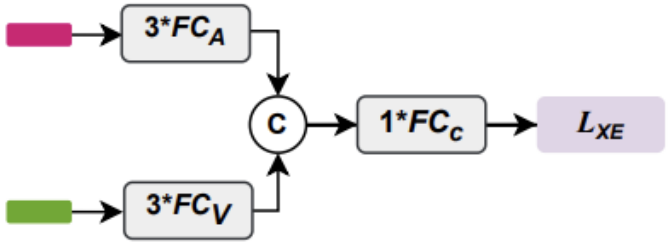
How to fuse? Concat and FC

(a) FC Audio

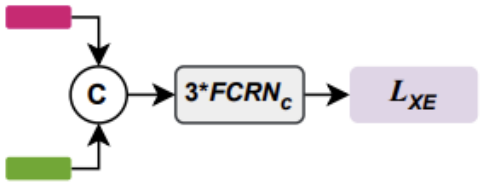
(b) FC Visual



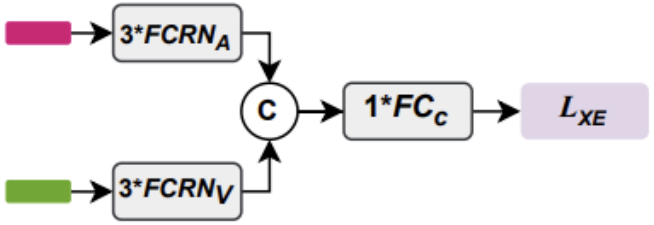
(c) FC Early Fusion



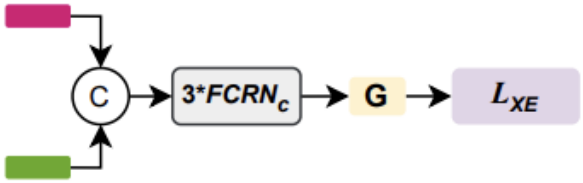
(d) FC Late Fusion



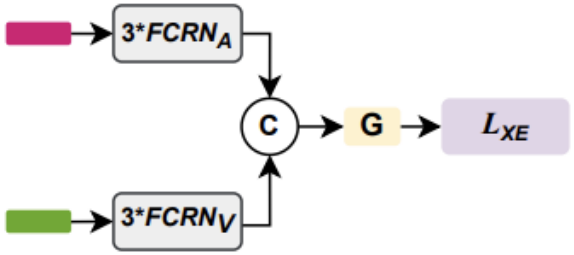
(e) FC Residual Early Fusion



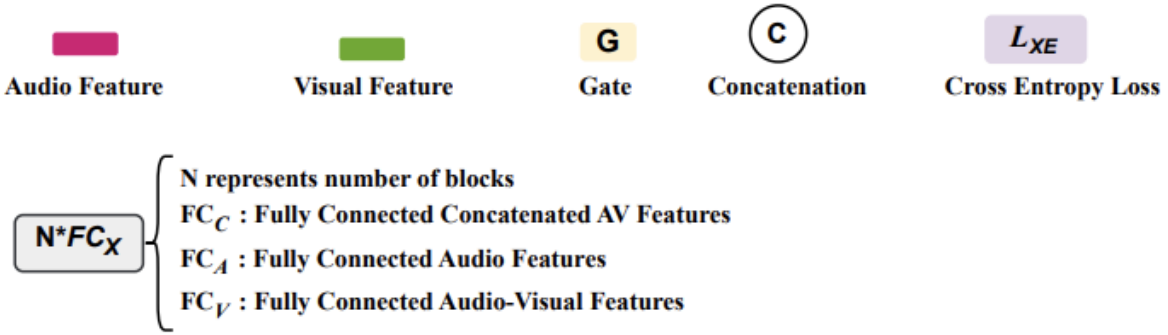
(f) FC Residual Late Fusion



(g) FC Residual Gated Early Fusion

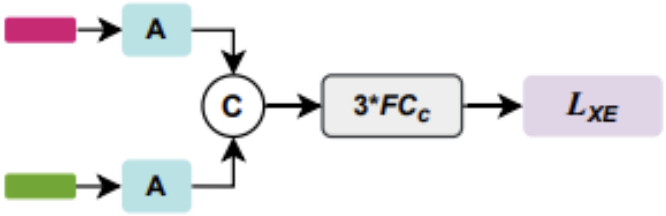


(h) FC Residual Gated Late Fusion

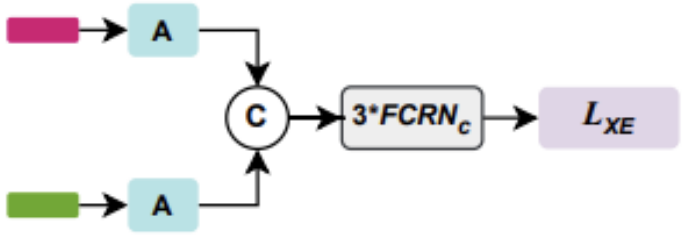


[Figures from Awan ECCV 2024]

How to fuse? Concat and FC

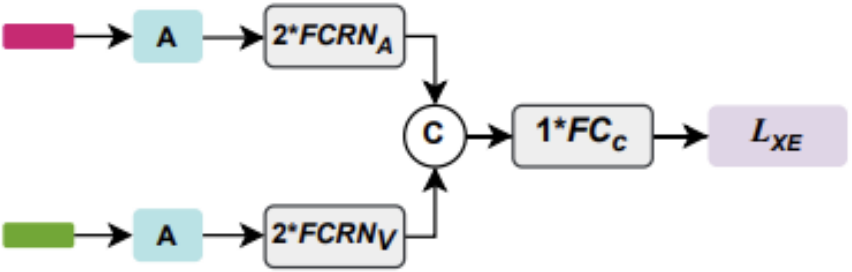


(a) FC Attention

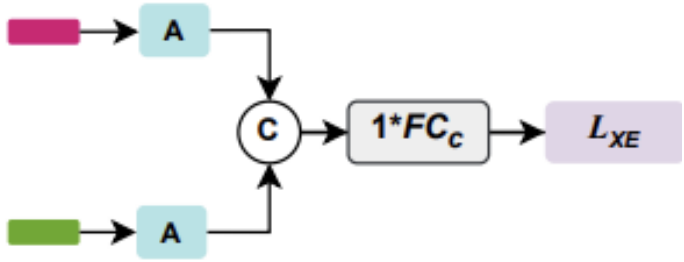


(b) Residual Attention Early Fusion

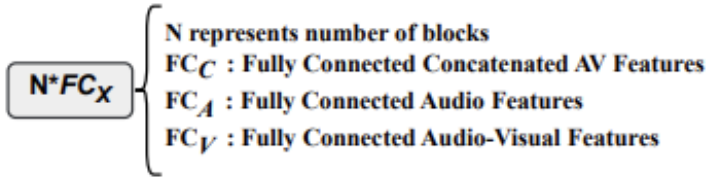
(g) Self-Attended Cross-Modal FCRN Network

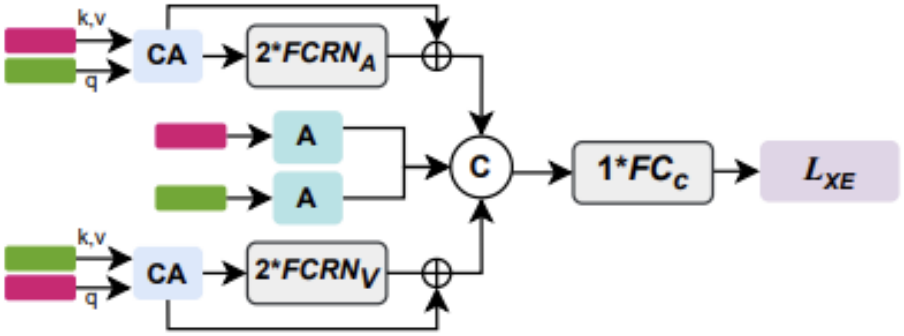


(c) Residual Attention Late Fusion

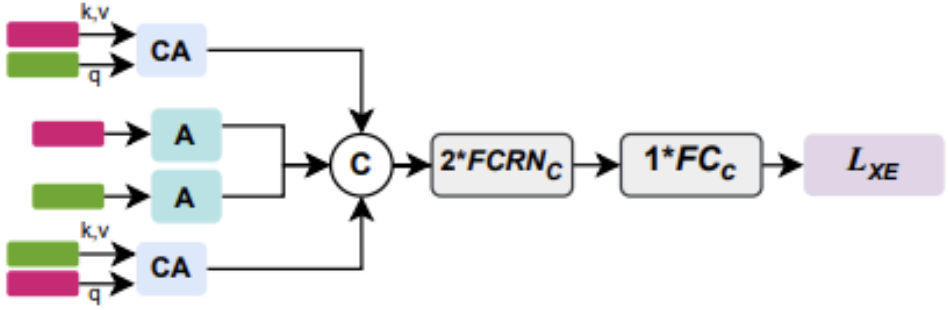


(d) Attend-Fusion





(e) Audio-Visual Attention Network



(f) Self and Cross Modal Attention Network

[Figures from Awan ECCV 2024]

Multimodal tasks (=Vision and Language Tasks)

- Multimodal Classification
- **Image-Text Retrieval**
- Visual Grounding
- Visual Question Answering and Visual Reasoning
- Image Captioning
- Text-to-image Generation

Image-text retrieval

Image-text Retrieval (Text-to-Image Retrieval)

Text Query: A dog lying on the grass next to a frisbee

Match

Not Match

Image-text retrieval

Image-text Retrieval (Text-to-Image Retrieval)
Text Query: A dog lying on the grass next to a frisbee

Match

Not Match

Coordination

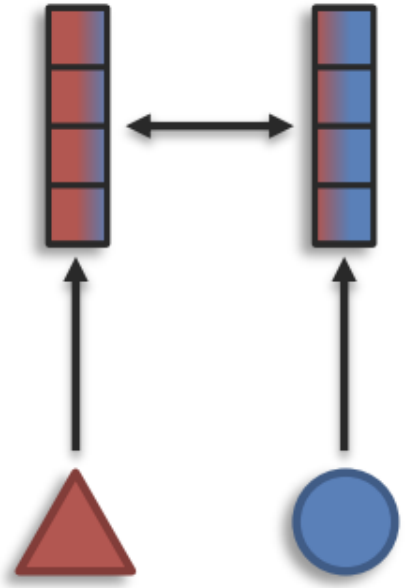
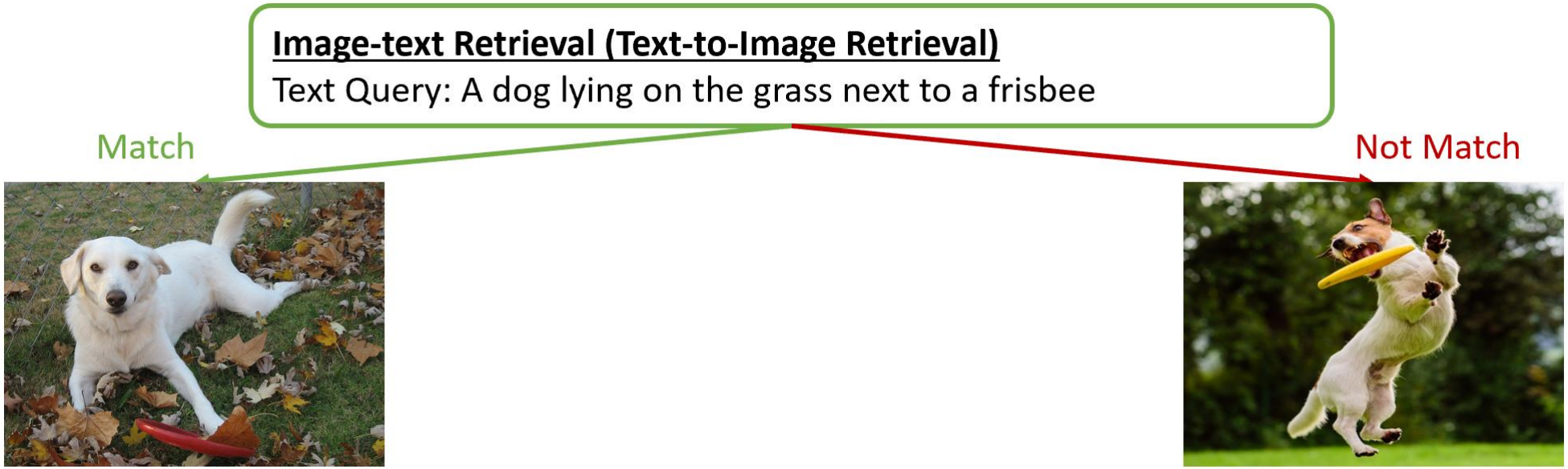
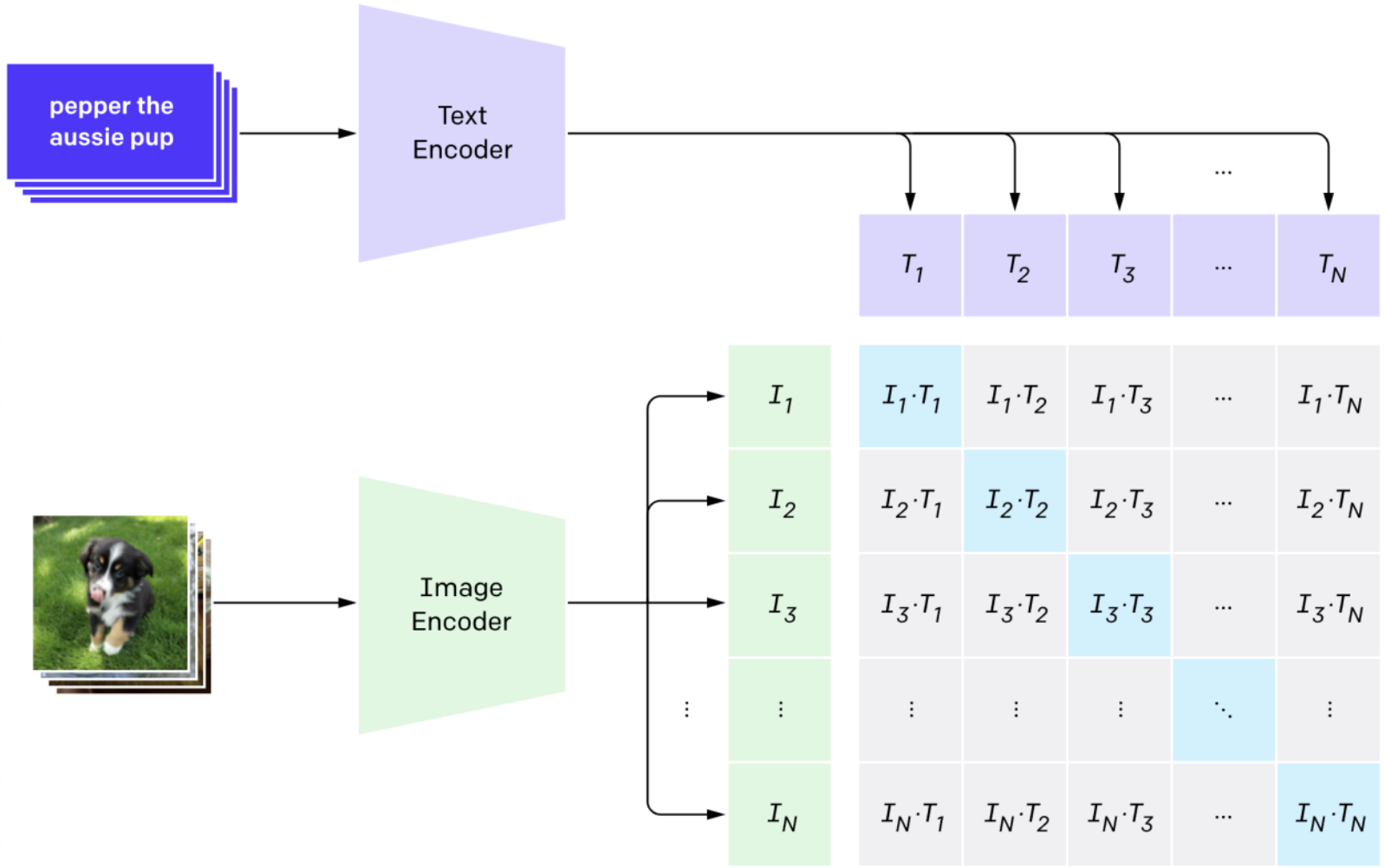


Image-text retrieval



- **Inputs:** Images and Text
- **Outputs:**
  - Relevant images: When a text query is given, the system returns a ranked list of images most relevant to the text.
  - Relevant text: When an image query is given, the system returns a ranked list of text descriptions or captions that best describe the image.
- **Tasks:**
  - Image-to-text retrieval: Given an image as input, retrieve text descriptions or captions that accurately describe its content.
  - Text-to-image retrieval: Given a text query, retrieve images that visually match the concepts and entities mentioned in the text.

CLIP



Contrastive loss: Each image predicts which caption matches

Self-Supervised Learning

Build methods that learn from "raw" data

no annotations required

Self-Supervised Learning

Supervised Learning

Data: (x, y)

x is data, y is label

Goal: Learn a *function* to map x -> y

Examples: Classification, regression, object detection, semantic segmentation, image captioning, etc.

Unsupervised Learning

Data: x

Just data, no labels!

Goal: Learn some underlying hidden *structure* of the data

Examples: Clustering, dimensionality reduction, feature learning, density estimation, etc.

Self-Supervised Learning

Let's build methods that learn from "raw" data: ***no annotations required***

- **Unsupervised Learning:** Model isn't told what to predict. Older terminology, not used as much today.
- **Self-Supervised Learning:** Model is trained to predict some naturally occurring signal in the raw data rather than human annotations.

Self-Supervised Learning

Let's build methods that learn from "raw" data: ***no annotations required***

- **Unsupervised Learning:** Model isn't told what to predict. Older terminology, not used as much today.
- **Self-Supervised Learning:** Model is trained to predict some naturally occurring signal in the raw data rather than human annotations.

Self-Supervised Learning

2 step process: First **Pretext** task, Then **downstream** task

Pretext tasks

Generative: Predict part of the input signal

- Autoencoders (sparse, denoising, masked)
- Autoregressive
- GANs
- Colorization
- Inpainting

Discriminative: Predict something about the input signal

- Context prediction
- Rotation
- Clustering
- Contrastive

Multimodal: Use some additional signal in addition to RGB images

- Video
- 3D
- Sound
- Language

Contrastive Learning

Hadsell et al, “Dimensionality Reduction by Learning and Invariant Mapping”, CVPR 2006
Wu et al, “Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination”, CVPR 2018
Van den Oord et al, “Representation Learning with Contrastive Predictive Coding”, NeurIPS 2018

Hjelm et al, “Learning deep representations by mutual information estimation and maximization”, ICLR 2019
Bachman et al, “Learning Representations by Maximizing Mutual Information Across Views”, NeurIPS 2019
Henaff et al, “Data-Efficient Image Recognition with Contrastive Predictive Coding”, ICML 2020

Tian et al, “Contrastive Multiview Coding”, ECCV 2020
He et al, “Momentum Contrast for Unsupervised Visual Representation Learning”, CVPR 2020
Chen et al, “A Simple Framework for Contrastive Learning of Visual Representations”, ICML 2020

Contrastive Learning

Batch of
N images

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

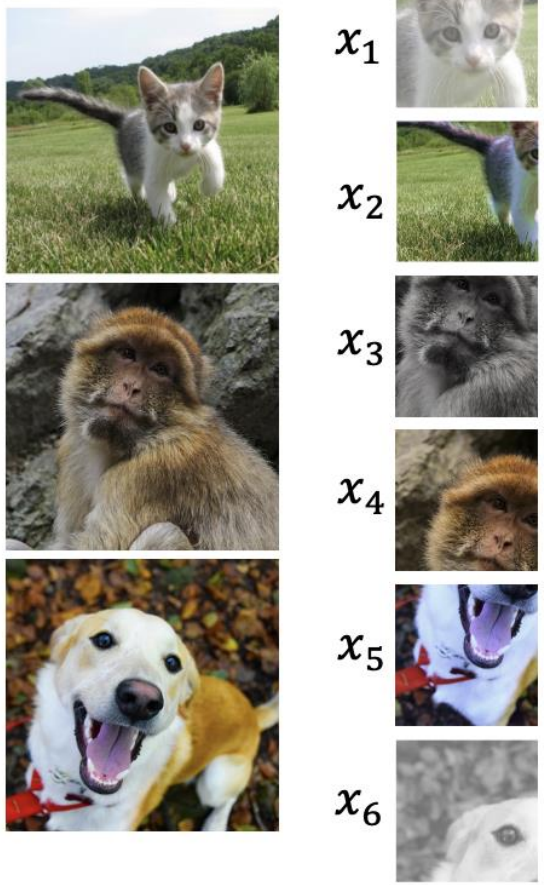
Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020

Contrastive Learning

Batch of N images

Two augmentations for each image



x_1 x_2 x_3 x_4 x_5 x_6

Hadsell et al, "Dimensionality Reduction by Learning and Invariant Mapping", CVPR 2006
Wu et al, "Unsupervised Feature Learning by Non-Parametric Instance-Level Discrimination", CVPR 2018
Van den Oord et al, "Representation Learning with Contrastive Predictive Coding", NeurIPS 2018

Hjelm et al, "Learning deep representations by mutual information estimation and maximization", ICLR 2019
Bachman et al, "Learning Representations by Maximizing Mutual Information Across Views", NeurIPS 2019
Henaff et al, "Data-Efficient Image Recognition with Contrastive Predictive Coding", ICML 2020

Tian et al, "Contrastive Multiview Coding", ECCV 2020
He et al, "Momentum Contrast for Unsupervised Visual Representation Learning", CVPR 2020
Chen et al, "A Simple Framework for Contrastive Learning of Visual Representations", ICML 2020