Thursday Learning Hour

Getting Started with Multi-Modal Machine Learning

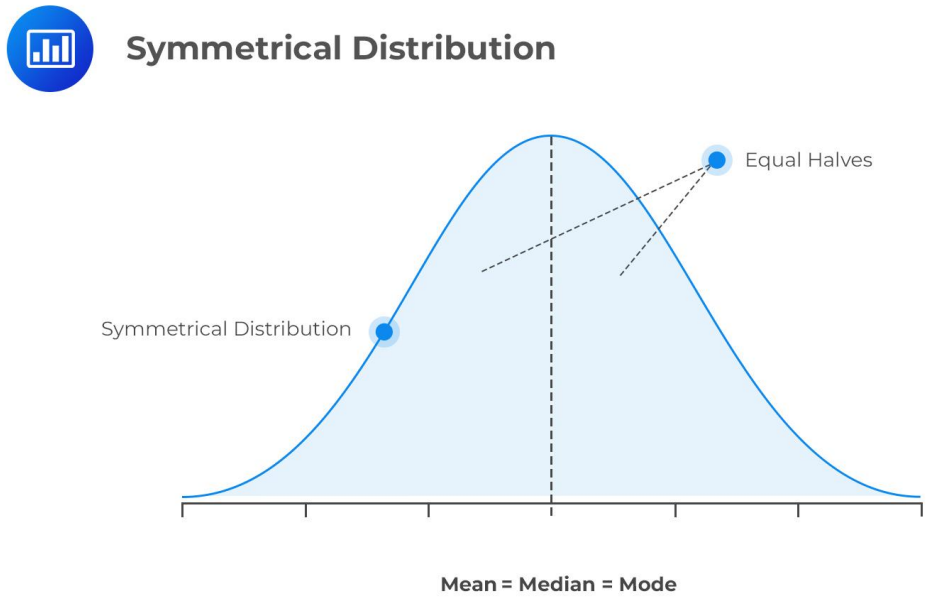
Prabakaran Chandran

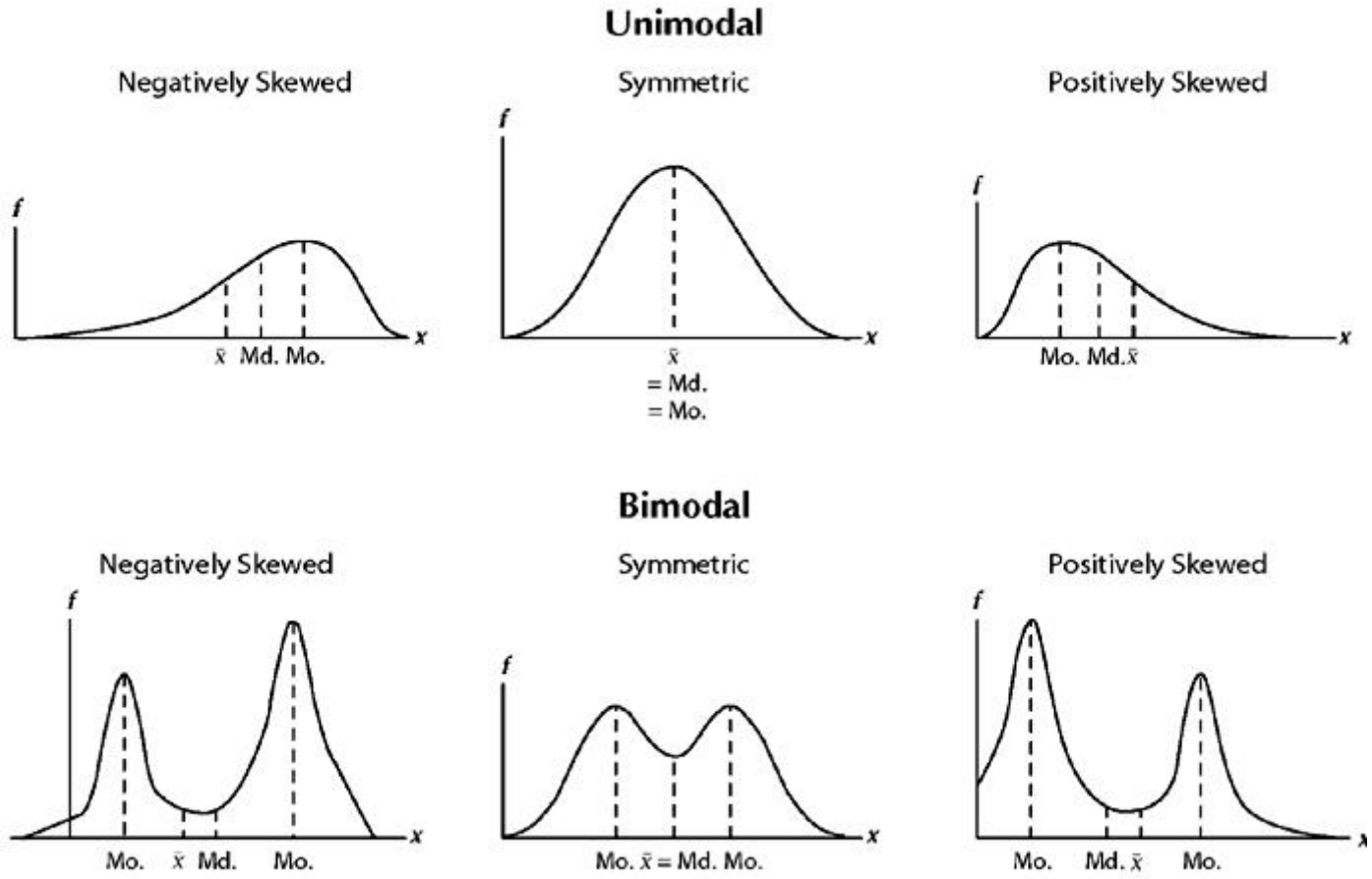
22nd Sept 2022

Agenda for the day

- What is Modality ?
- Unimodal Machine Learning and Pretraining
- Multi Modality – Introduction
- Where do we need Multi Modality based Intelligence
- Multi Modal Machine learning – Conceptual Introduction
- Challenges in Multi Modal Learning
- MMML Architectures
- MMML Deep learning Networks
- Example : Vision + Language Modeling

Modality ?



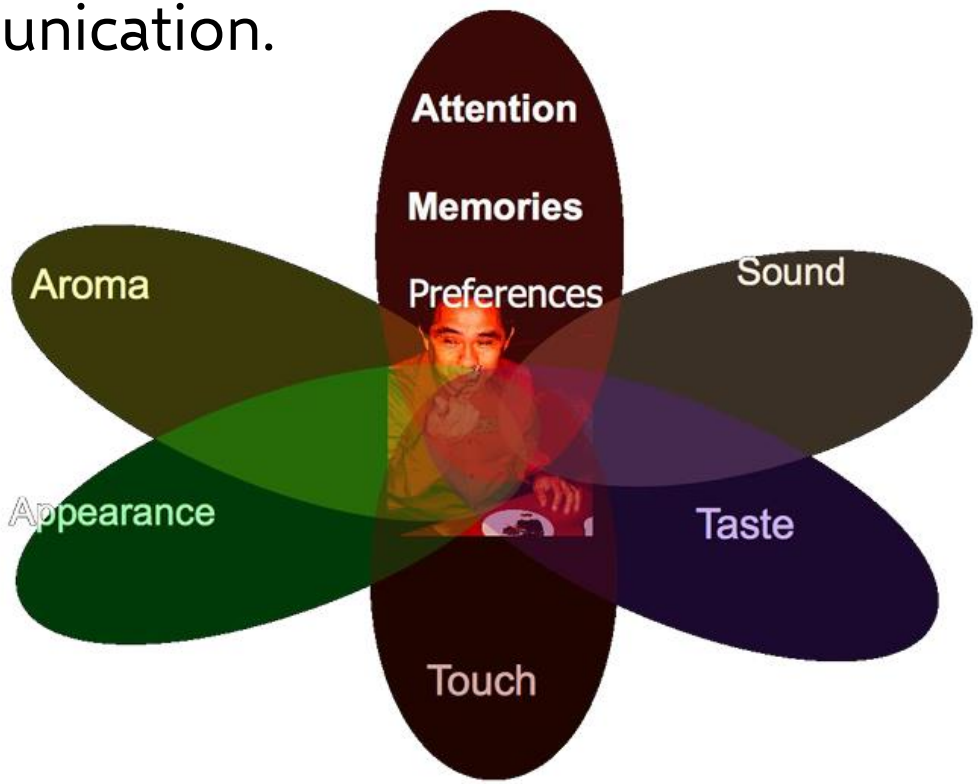


Multiple modes, i.e., distinct “peaks” (local maxima) in the probability density function

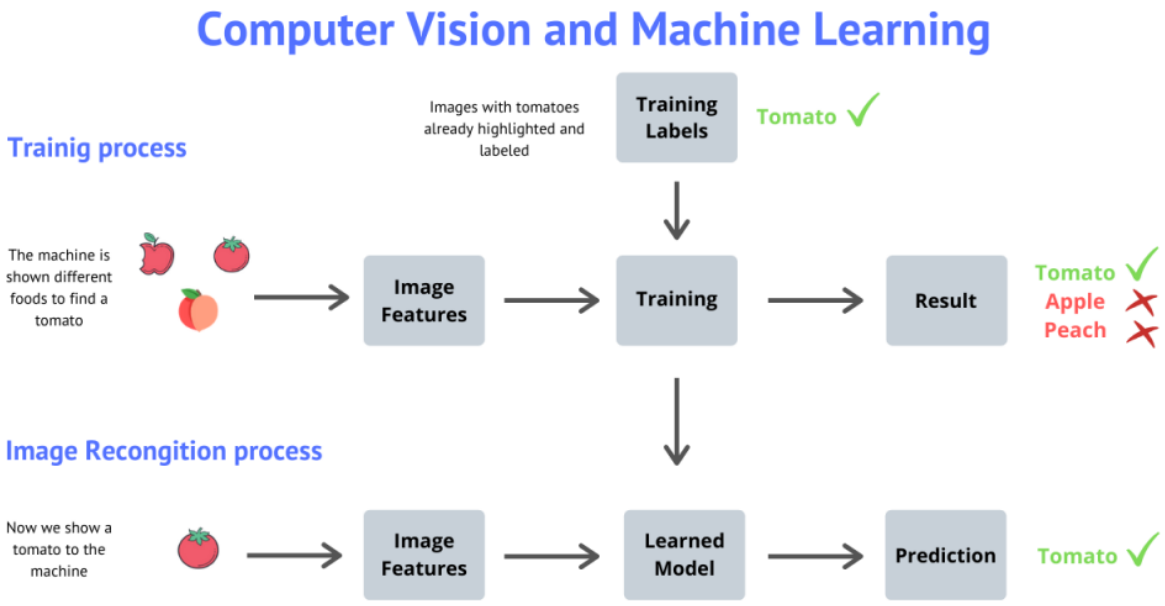
Modality ? – Another Explanation

Modality The way in which something happens or is experienced.

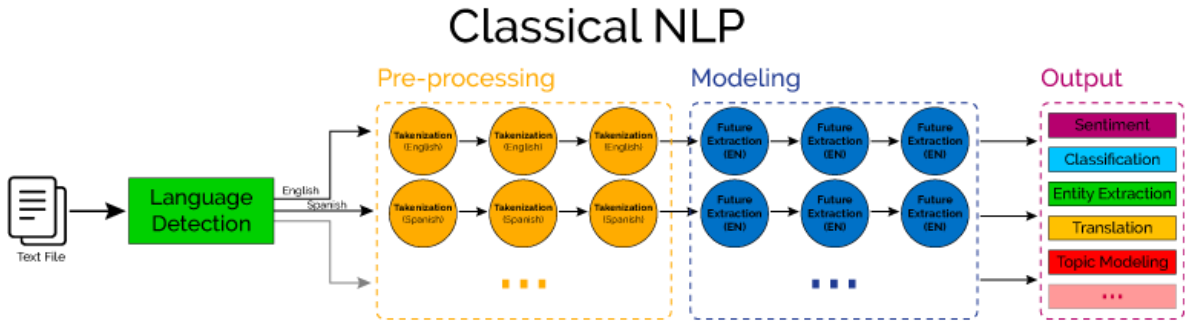
- Modality refers to a certain type of information and/or the representation format in which information is stored.
- Sensory modality: one of the primary forms of sensation, as vision or touch; channel of communication.



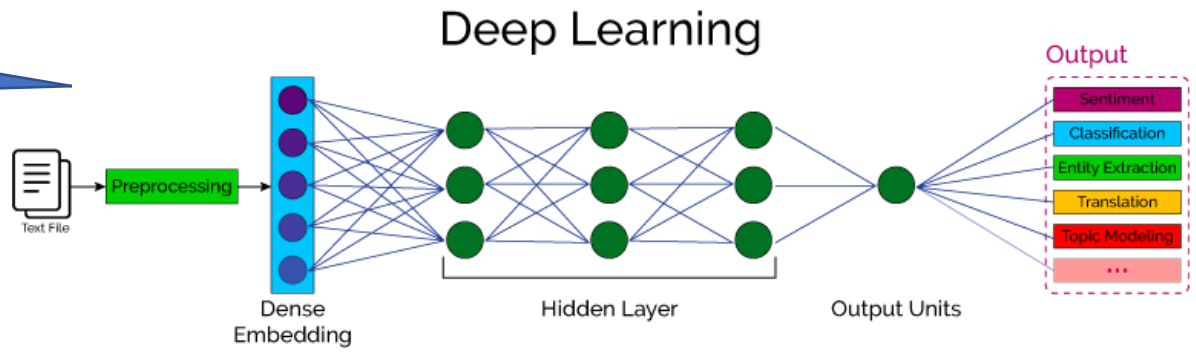
Uni-Modal Machine learning



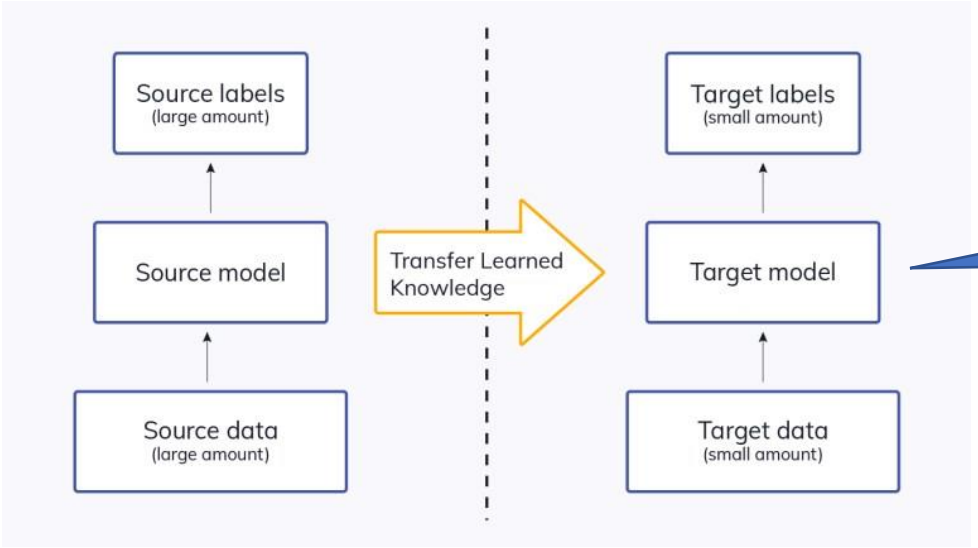
Computer vision - Image Classification



Natural Language Tasks

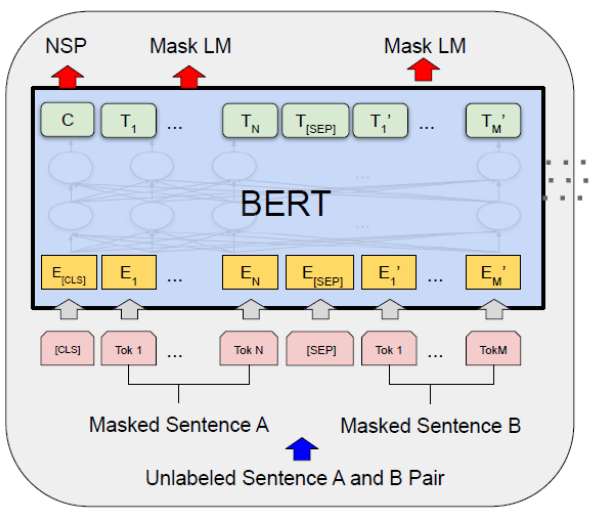


Uni-Modal Pretraining and Transfer learning

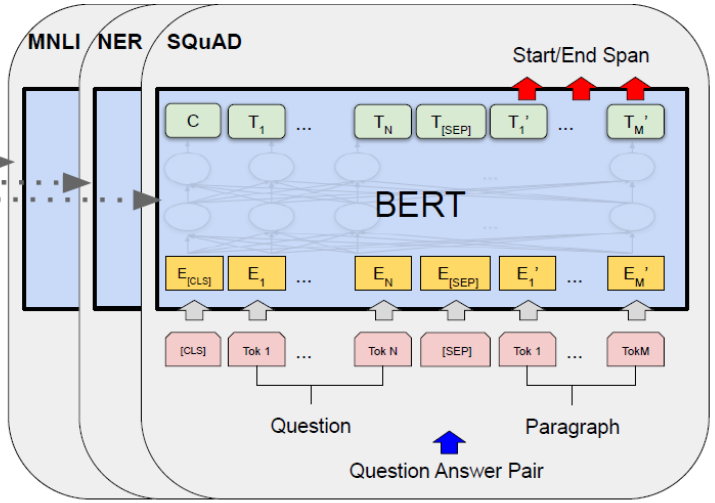


Transfer learning

BERT Pretraining



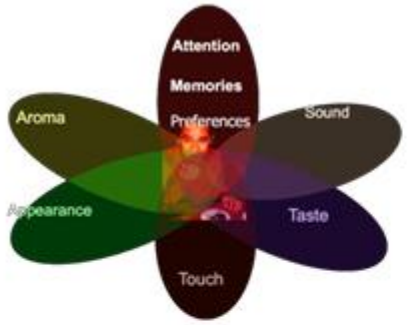
Pre-training



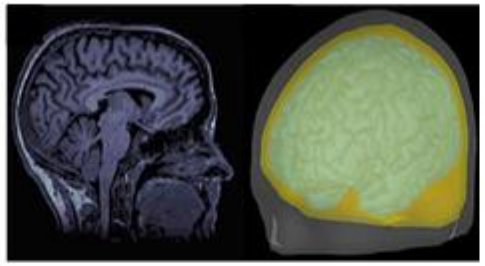
Fine-Tuning

Multi-Modality Introduction

Multimodal data refers to data that spans different types and contexts (e.g., imaging, text, or genetics).



Psychology



Medical

Speech

Vision

Language

Multimedia

Robotics

Learning

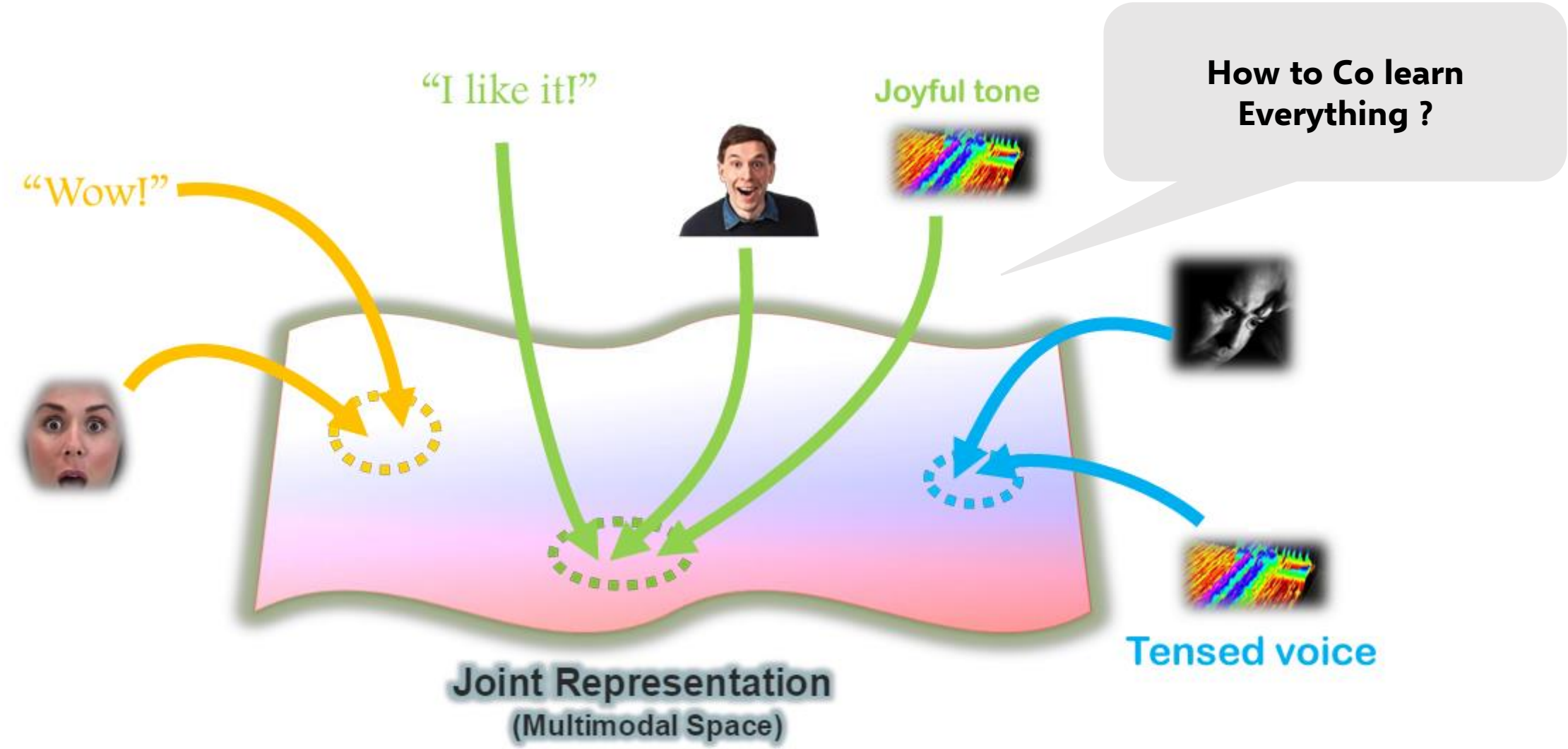
Multi-Modality Introduction

Multimodal data refers to data that spans different types and contexts (e.g., imaging, text, or genetics).

- Natural language
- Visual (both spoken or written) (from images or videos) Auditory (including voice, sounds and music)
- Haptics / touch
- Smell, taste and self motion
- Physiological signals
  - Electrocardiogram (ECG)
  - skin conductance
- Other modalities
  - Infrared images, depth images, fMRI

There are tons of Machine Learning tasks we can do on individual modalities

Multi-Modality Feature Extraction



Where do we need Multi Modality based Intelligence

- Affect recognition
  - Emotion
  - Persuasion
  - Personality traits
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

Where do we need Multi Modality based Intelligence – Vision + Language

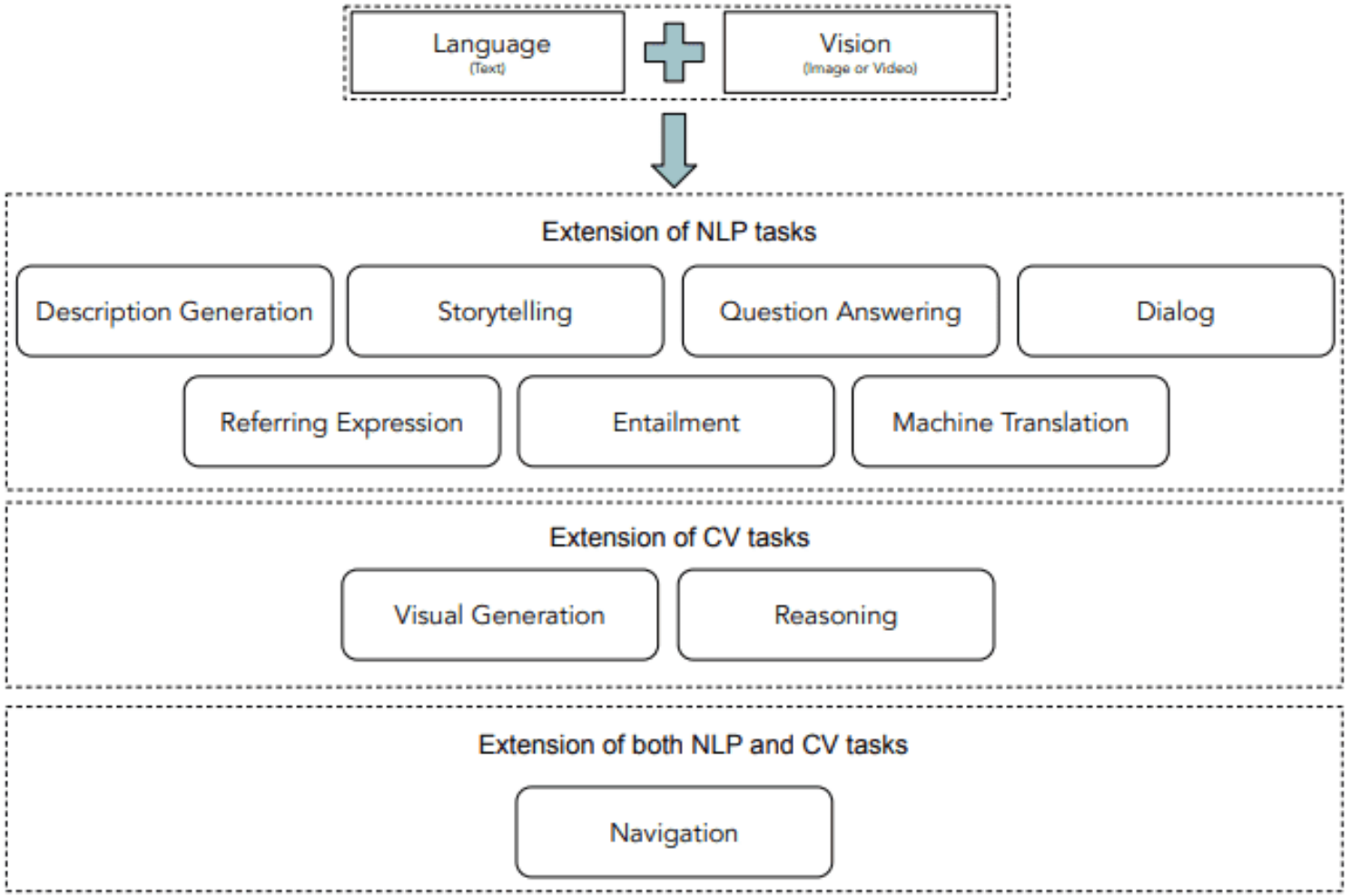
Vision-language tasks
Vision-language models have gained a lot of popularity in recent years due to the number of potential applications. We can roughly categorize them into 3 different areas. Let's explore them along with their subcategories.

Generation tasks
•**Visual Question Answering (VQA)** refers to the process of providing an answer to a question given a visual input (image or video).
•**Visual Captioning (VC)** generates descriptions for a given visual input.
•**Visual Commonsense Reasoning (VCR)** infers common-sense information and cognitive understanding given a visual input.
•**Visual Generation (VG)** generates visual output from a textual input, as shown in the image.

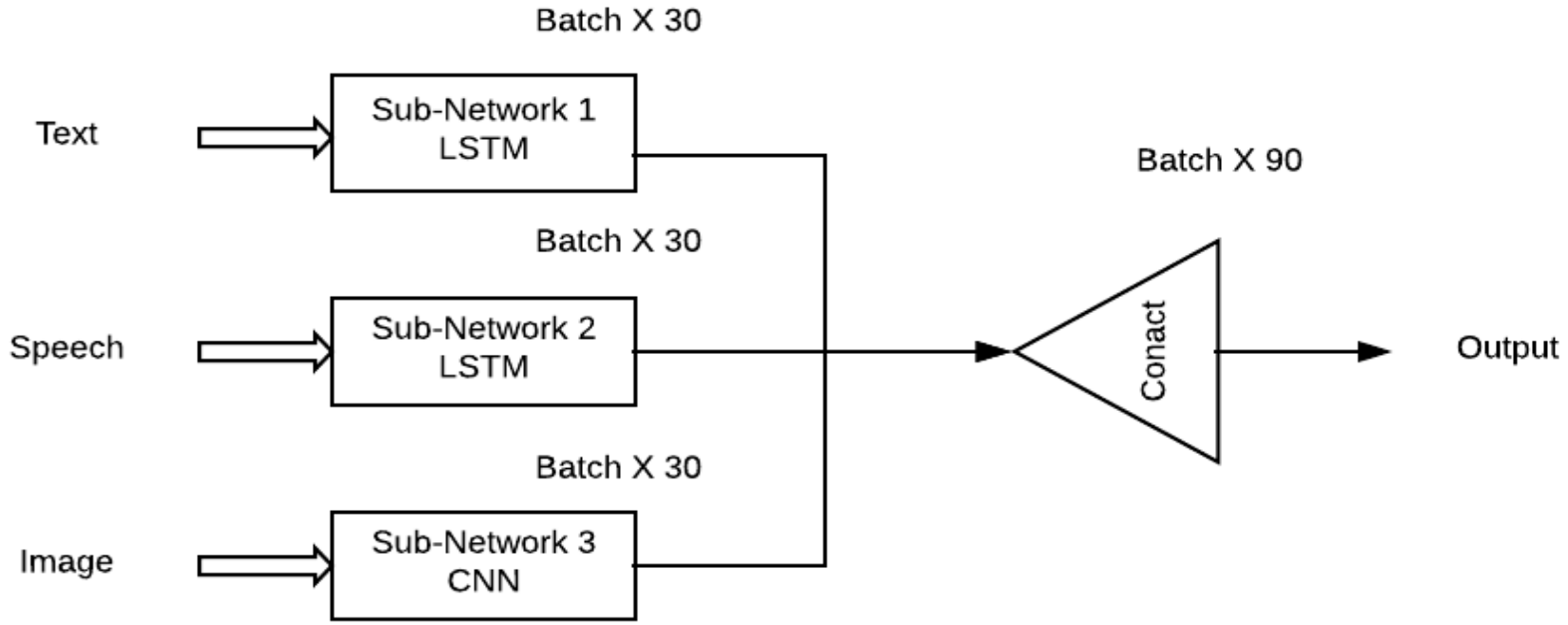
Classification tasks
•**Multimodal Affective Computing (MAC)** interprets visual affective activity from visual and textual input. In a way, it can be seen as multimodal sentiment analysis.
•**Natural Language for Visual Reasoning (NLVR)** determines if a statement regarding a visual input is correct or not.

Retrieval tasks
•**Visual Retrieval (VR)** retrieves images based only on a textual description.
•**Vision-Language Navigation (VLN)** is the task of an agent navigating through a space based on textual instructions.
•**Multimodal Machine Translation (MMT)** involves translating a description from one language to another with additional visual information.

Where do we need Multi Modality based Intelligence – Vision + Language

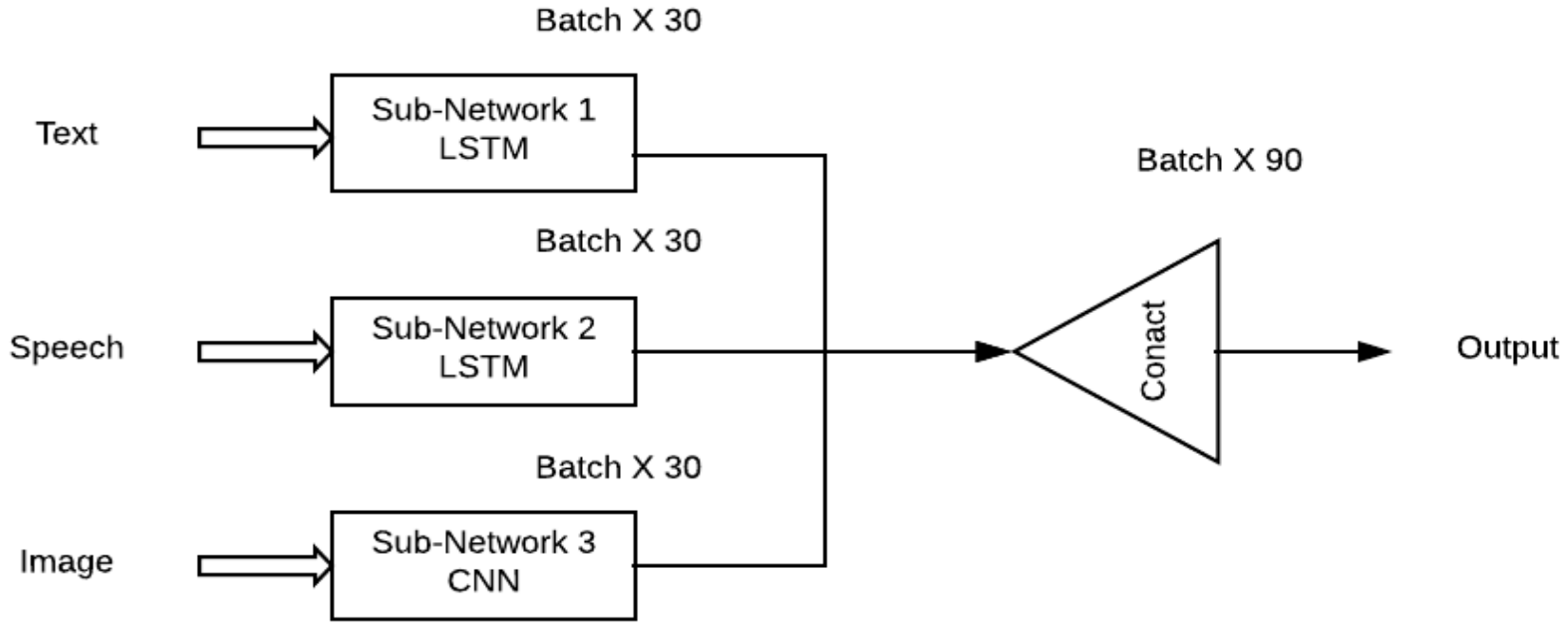


Multi Modal Machine learning – Conceptual Introduction (In the Perspective of V+L)



Though combining different modalities or types of information for improving performance seems intuitively appealing task, but in practice, it is challenging to combine the varying level of noise and conflicts between modalities. Moreover, modalities have different quantitative influence over the prediction output. The most common method in practice is to combine high-level embeddings from the different inputs by concatenating them and then applying a softmax.

Multi Modal Machine learning – Conceptual Introduction (In the Perspective of V+L)



Though combining different modalities or types of information for improving performance seems intuitively appealing task, but in practice, it is challenging to combine the varying level of noise and conflicts between modalities. Moreover, modalities have different quantitative influence over the prediction output. The most common method in practice is to combine high-level embeddings from the different inputs by concatenating them and then applying a softmax.

Multi Modal Machine learning – Conceptual Introduction (In the Perspective of V+L)

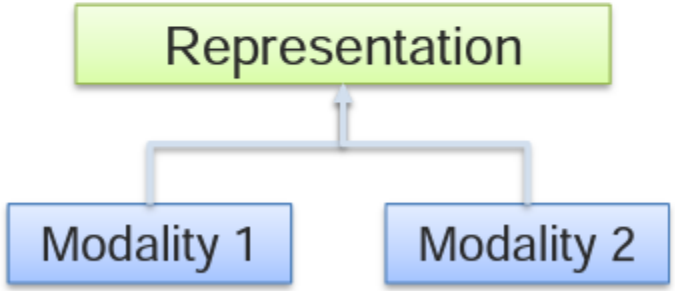
The core challenges are multiple:

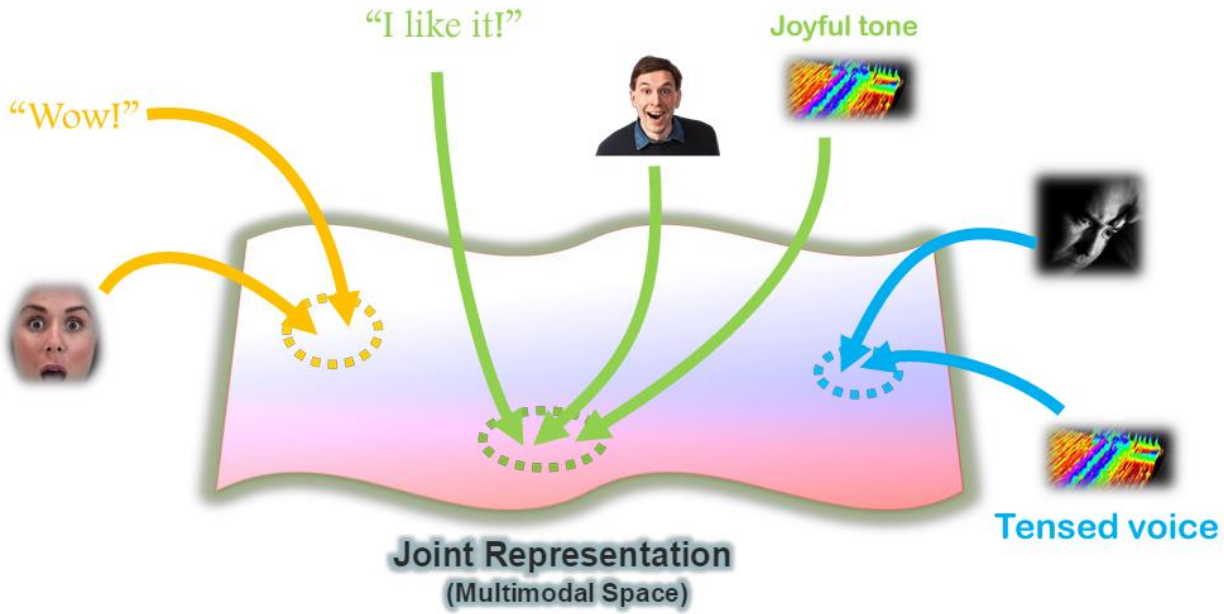
- **representation** with the goal to learn computer interpretable descriptions of heterogenous data from multiple modalities
- **translation** which represents the process of changing data from one modality to another
- **alignment** where we want to identify relations between elements from two or more different modalities
- **fusion** which represents the process of joining information from two or more modalities to perform a prediction task, and finally
- **co-learning** with the goal of transferring knowledge between modalities and their representations.

Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

Ⓐ Joint representations:

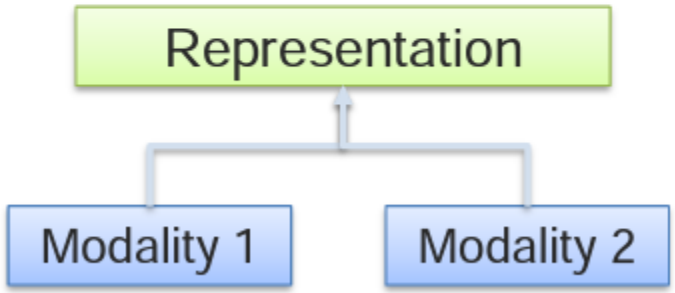


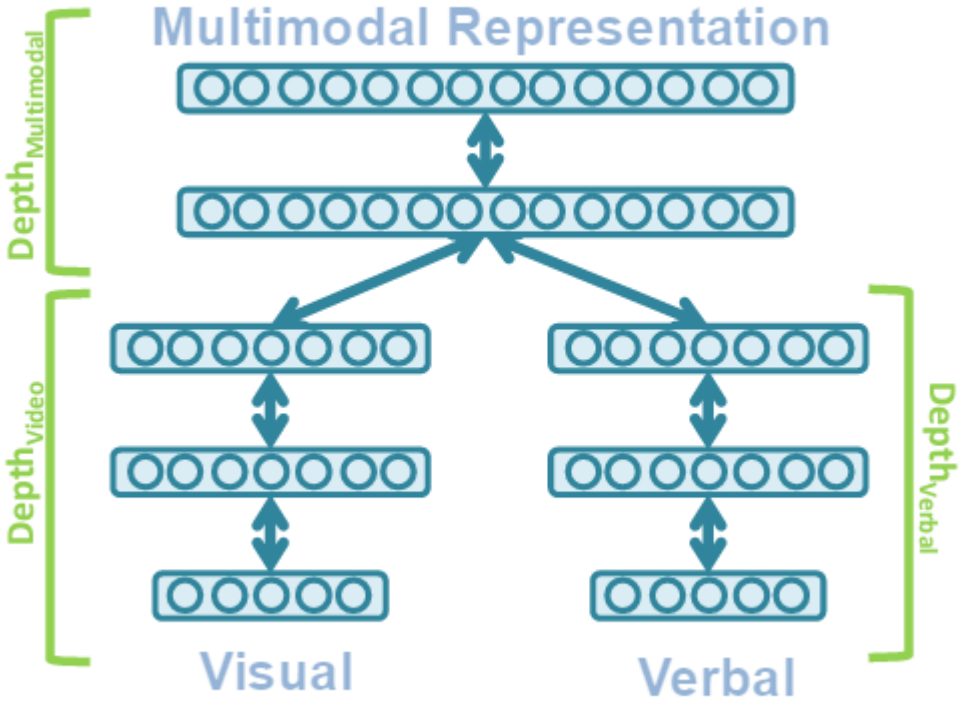


Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

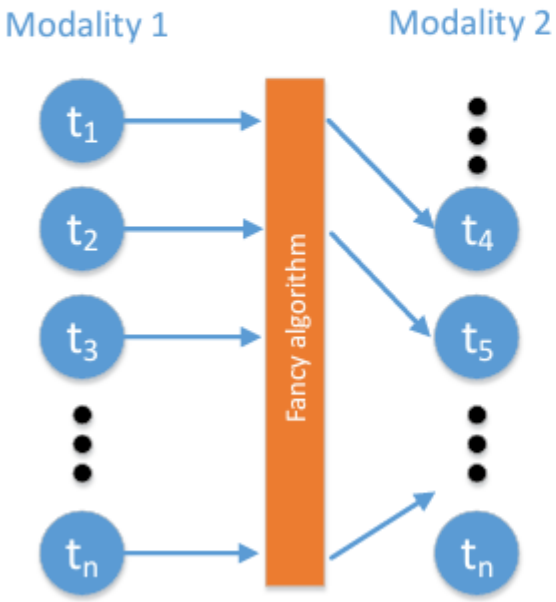
Ⓐ Joint representations:





Alignment

Definition: Identify the direct relations between (sub)elements from two or more different modalities.



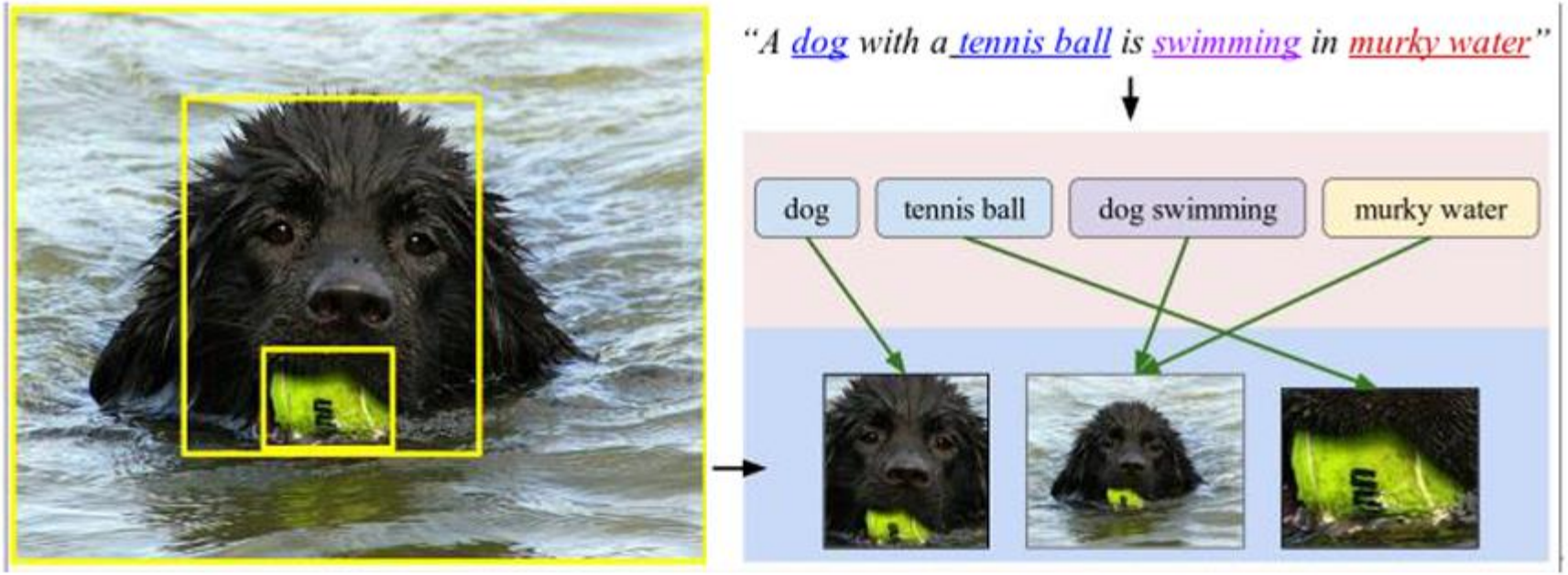
A Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

Uses internally latent alignment of modalities in order to better solve a different problem

Alignment

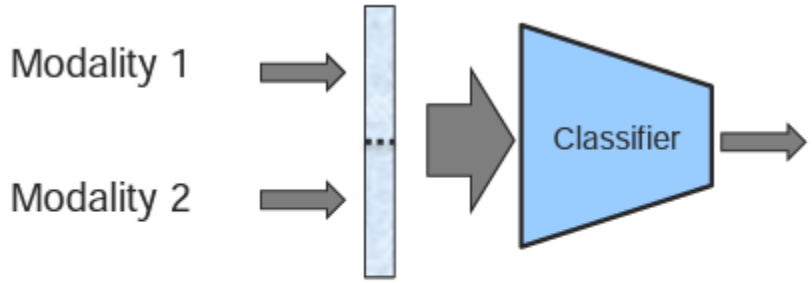

"A dog with a tennis ball is swimming in murky water"
dog
tennis ball
dog swimming
murky water

Fusion

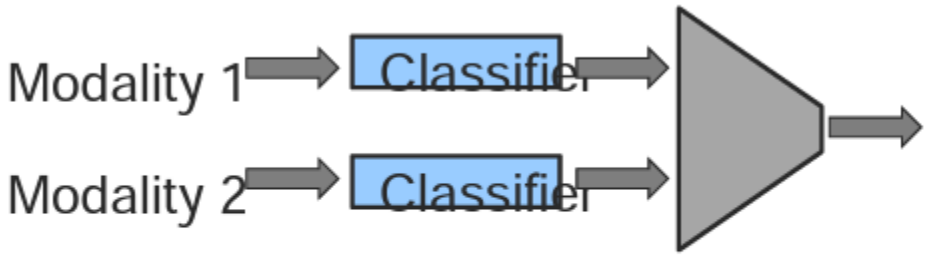
Definition: To join information from two or more modalities to perform a prediction task.

A Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion



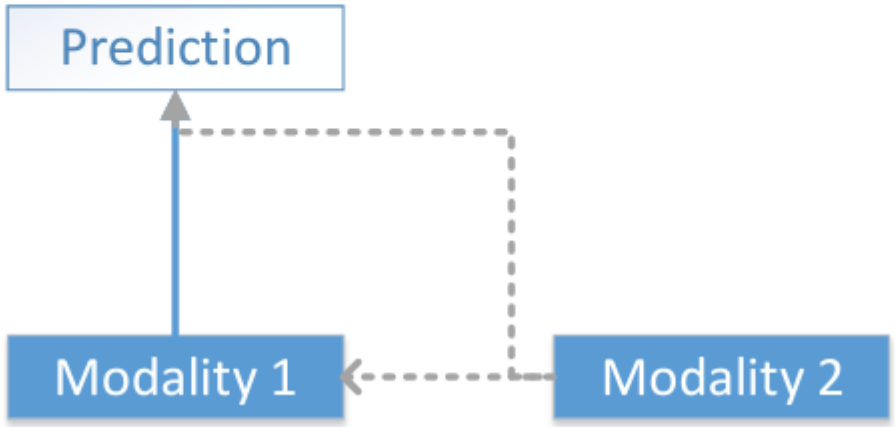
Translation

Visual gestures
(both speaker and listener gestures)

Transcriptions
+
Audio streams

Co Learning

Definition: Transfer knowledge between modalities, including their representations and predictive models.



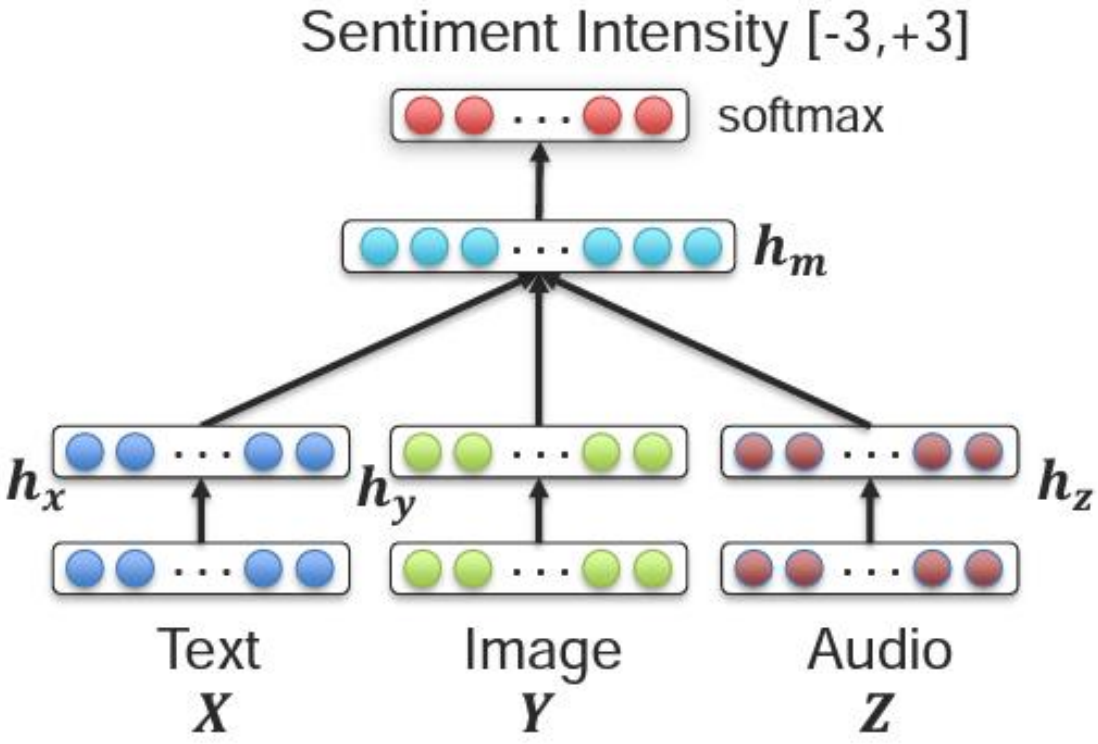
Multi Modal Sentiment Analysis

MOSI dataset (Zadeh et al, 2016)

- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

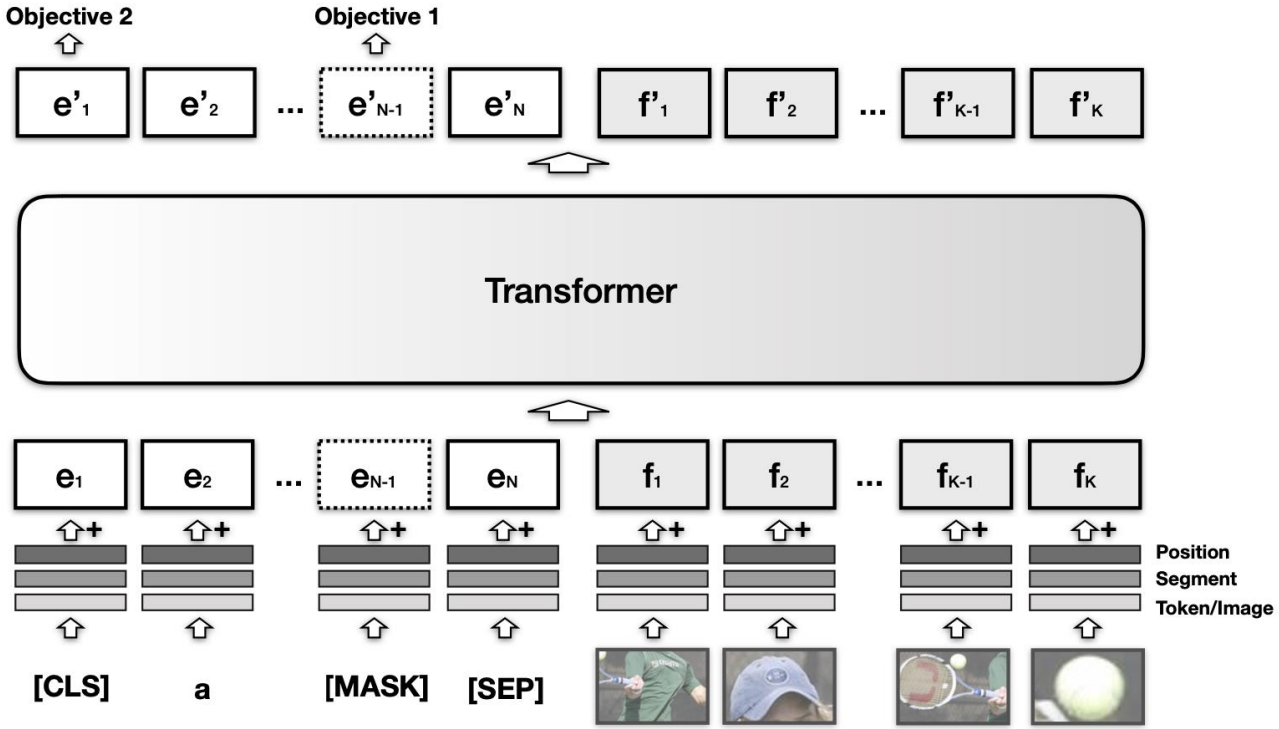
Multimodal joint representation:

$$h_m = f\left(W \cdot [h_x, h_y, h_z]\right)$$



V+L Deep learning Networks - Visual Bert

A person hits a ball with a tennis racket



V+L Deep learning Networks - Visual Bert

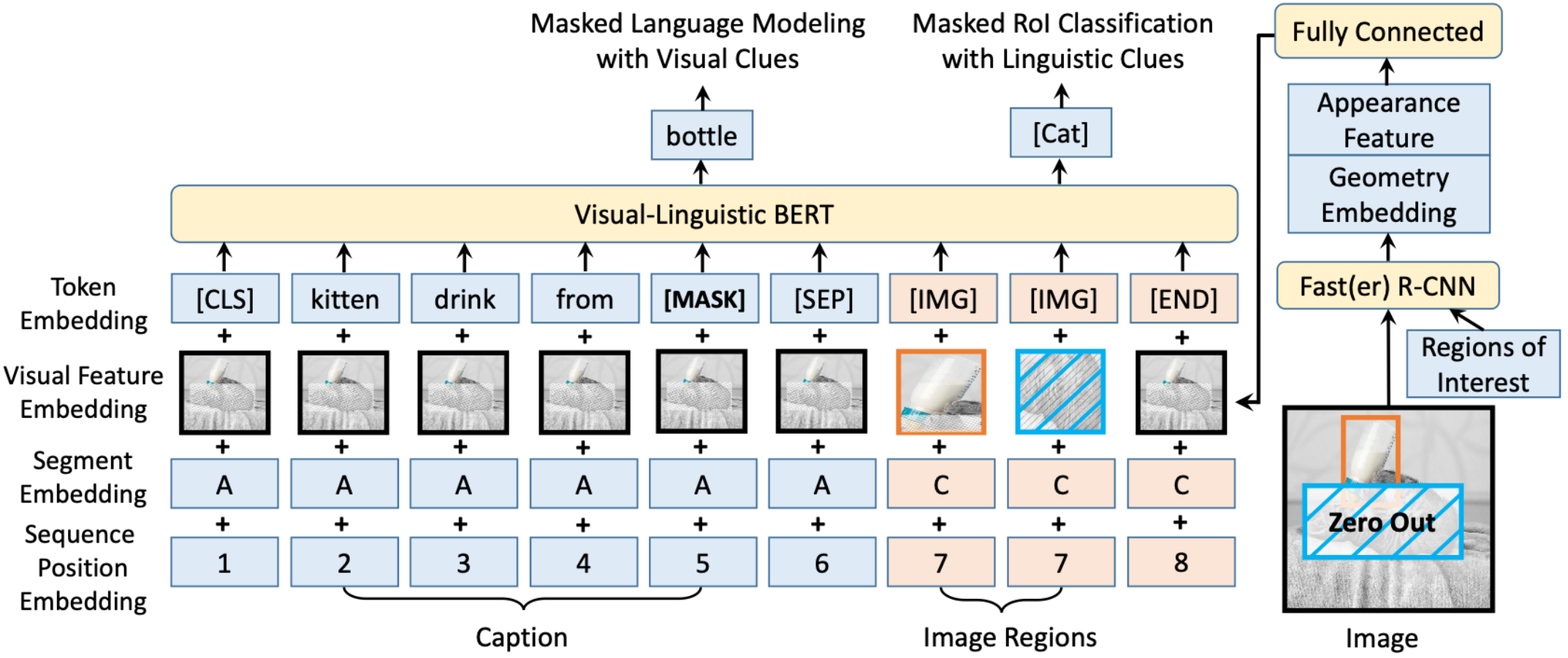


Figure 1. Architecture for Pre-training VL-BERT

V+L Deep learning Networks - simple VLM