SAMSUNG

NOVA EXECUTIVE EDUCATION
NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Samsung Innovation Campus

Chapter 9. Deep Learning Module

Chapter 9.

Vision and Language

(Plus the CLIP Model)

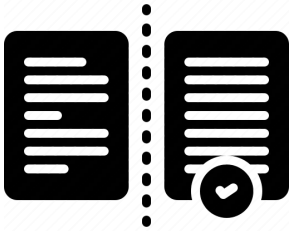
(and SigLiP as a bonus)

In the latest episode of the Deep Learning Module…

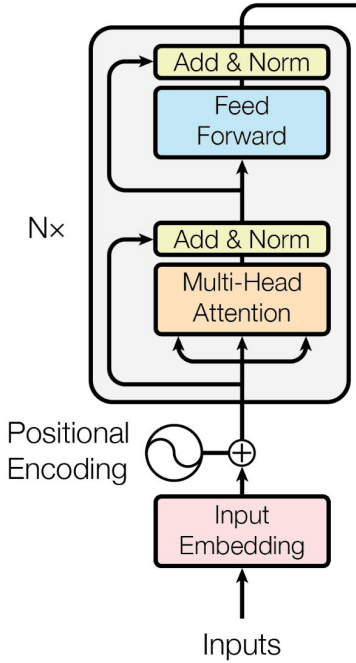
You saw the encoder architecture and how it can be used for many tasks in Natural Language Processing (NLP).

Retrieval

Sentiment



Similarity

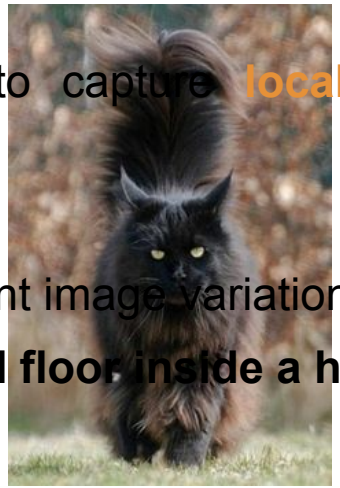


You even got to meet BERT!

Episode 2: What about Vision?

Vision tasks have **different requirements:**

1. **Complex** visual information: We need to understand individual objects, **scenes**, and **compositions**



2. **Global Context:** We need to capture **local** and **long-range dependencies**

3. **Generalization:** Across different image variations and different tasks

“A black cat on a hardwood floor inside a house.”

Episode 3: What about Vision and Languages?

Vision and Language (V&L) tasks are crucial for models to understand the world in a more “human-like” way.

If humans naturally combine both modalities, why shouldn’t models do the same?

Visual Question-Answering

Image Captioning

Cross-modal Retrieval

V&L Tasks: Visual Question-Answering

Input: Image and a question

What is this museum?

Output: Answer

This can either be similar to a **multiple question** test - where the model knows all the possible answers *a priori* - or **generated!**

V&L Tasks: Image Captioning

Input: Image

Besides simply describing the image, this can be used for **search:** where the query *is* the image description and the retrieved documents may be either images or text

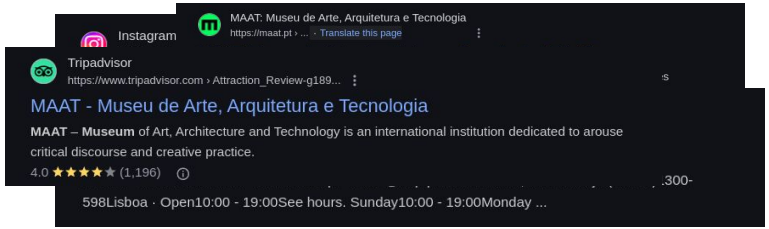
Output: Caption

A picture of the MAAT museum in Lisbon, taken at night

V&L Tasks: Cross-modal Retrieval

Input: Image

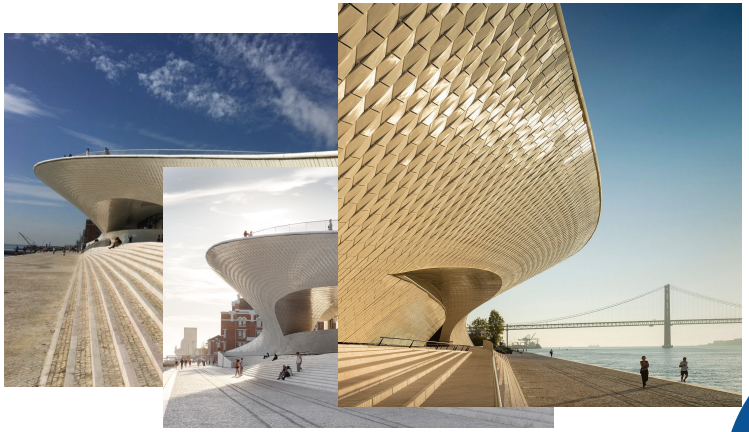
Output: Text



Input: Text

maat museum

Output: Image



V&L Models: How?

Visual and Language Models are a collection of models that can learn both from **textual** and **visual** information.

Although these models are often **generative** - capable of generating captions or an image, we first need to focus on how can these models…

… understand **text**?

Text Encoder

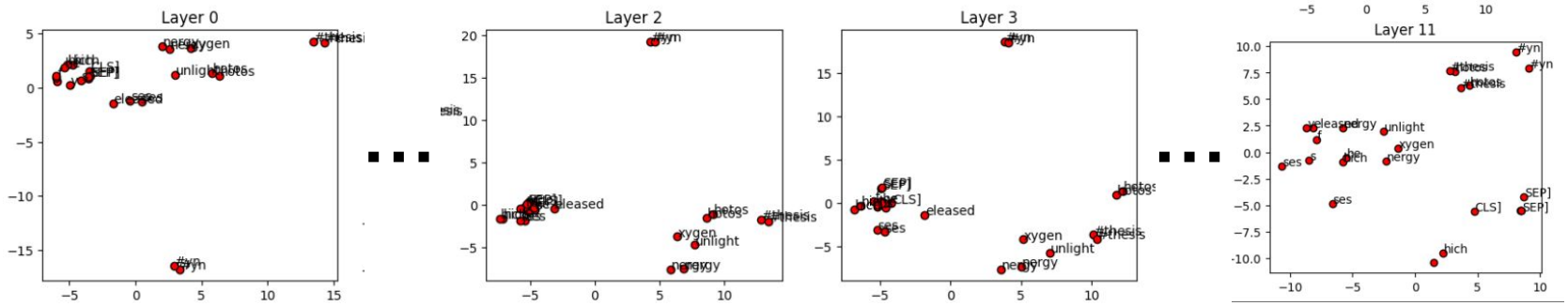
… understand **images**?

Vision Encoder

Model Input: Text

Text is encoded as word embeddings plus their positional embeddings.

Its is then processed with the usual pipeline in order to obtain the **contextual embeddings**.



1 - Based on WiSDoM slides by Professor João Magalhães.

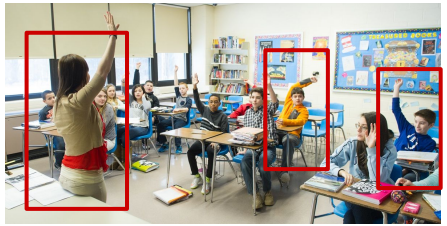
Model Input: Image

Are there visual embeddings?

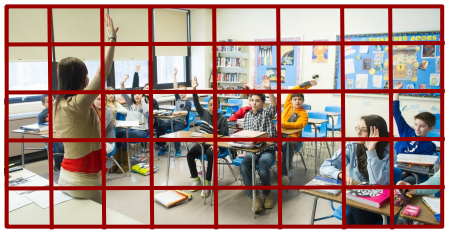
Yes! We need vector representations of images in order to represent them on the **same** higher dimensional space as text.

How can we get them?

Region-based Visual Embeddings



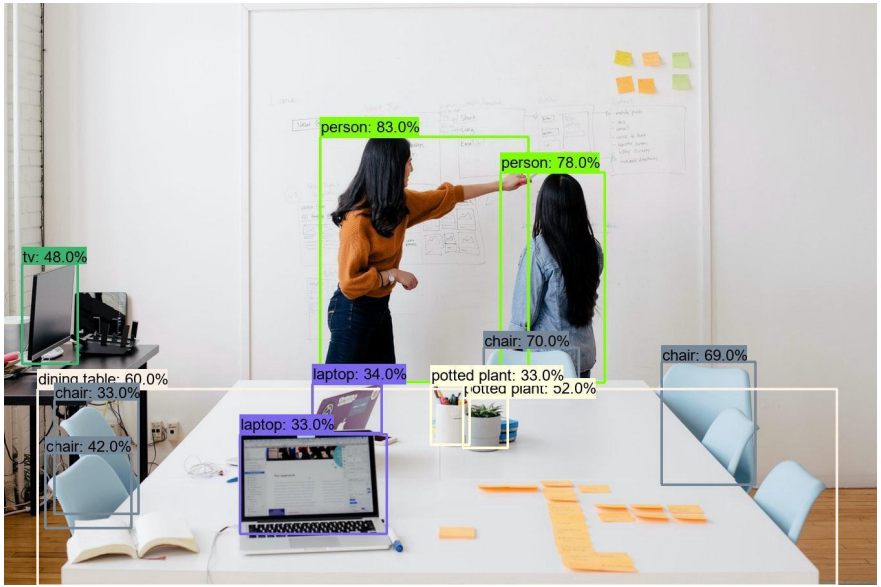
Grid-based Visual Embeddings



Region-based Embeddings

Image region features:

- Bounding boxes
- Visual features extracted from a pre-trained object detection model



This can be achieved with CNNs!

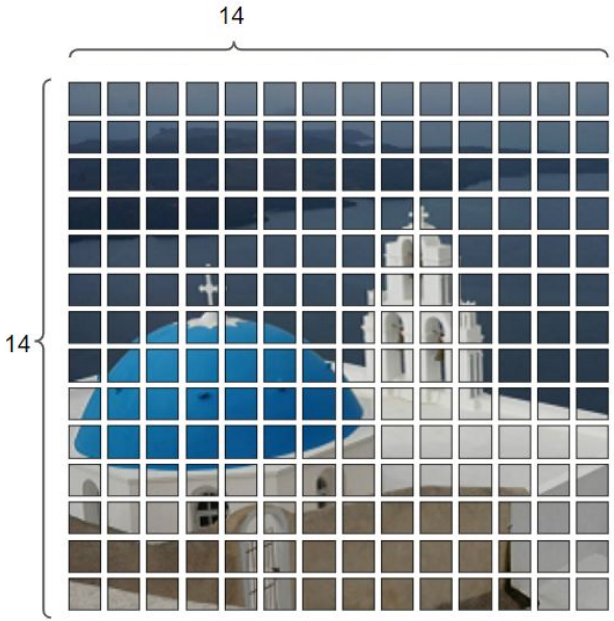
Grid-based Embeddings

Goal: Divide an image into a uniform grid and extract individual embeddings for each square **(patch)**.

1. The image is **center cropped** to a **224x224** pixel image

Enough detail is retained

Training is efficient



2. Create a **14x14** grid

1 - Based on WiSDoM slides by Professor João Magalhães

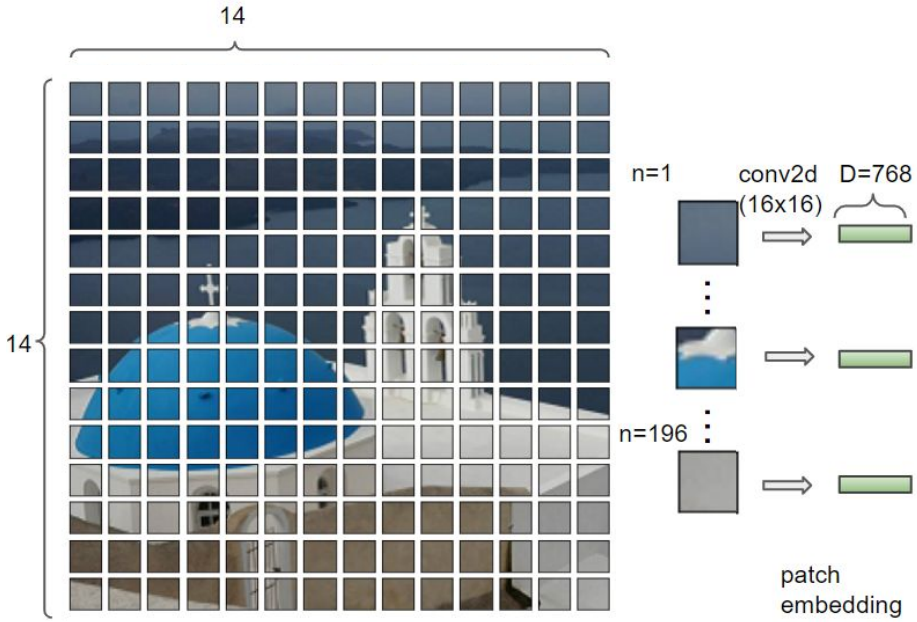
Grid-based Embeddings

Assuming a 14x14 grid, each patch will be **16x16** pixels.

Total nº of patches: 196

Total of pixels per patch: 768 (256 per channel in an RGB image)

3. A **transformation** is applied to each patch to obtain a single 768 dimensional vector

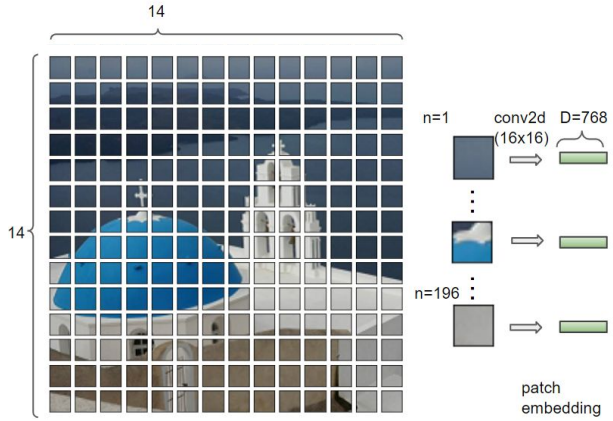


1 - Based on WiSDoM slides by Professor João Magalhães

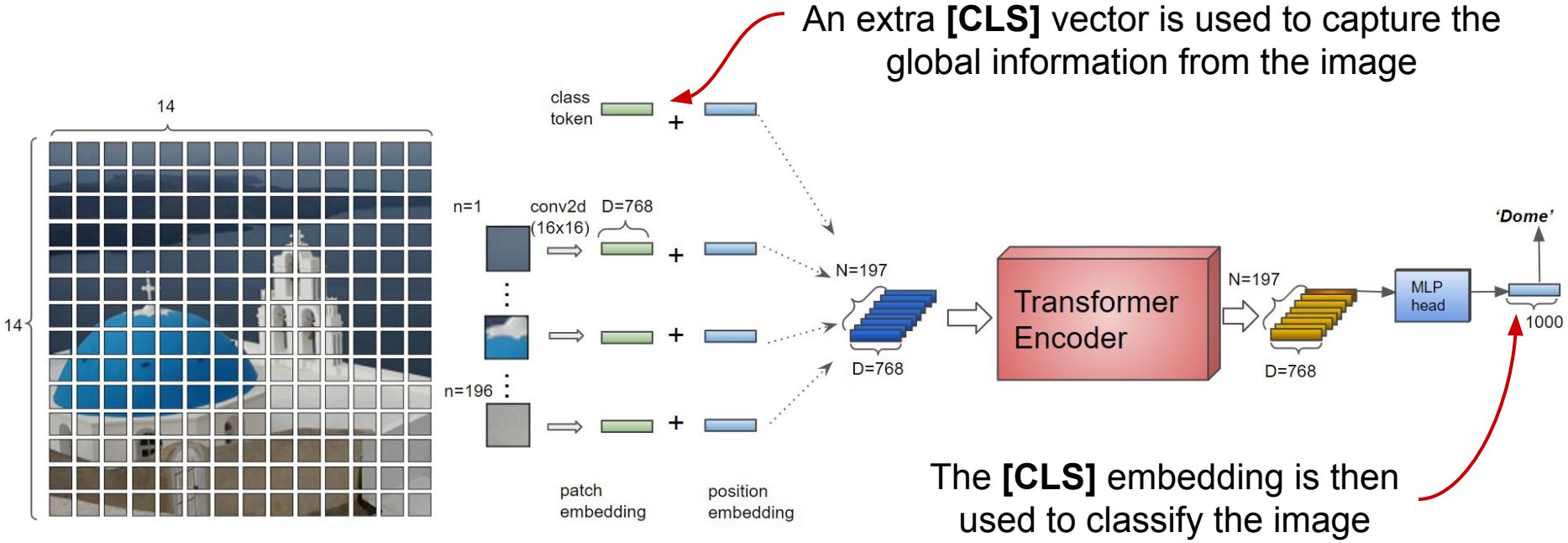
Grid-based Embeddings

This transformation can be:

- **Convolutional Transformation:** The patches are processed to obtain information about the edges, textures, and patterns
- **Linear Transformation:** Flattens the image embeddings into a vector

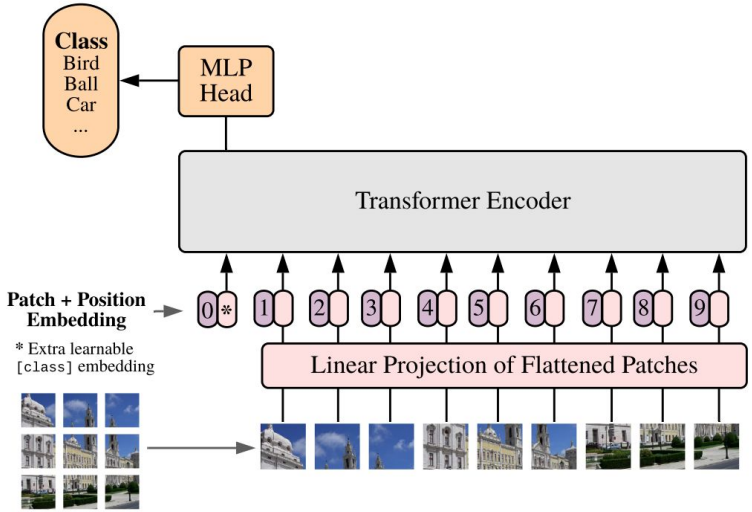


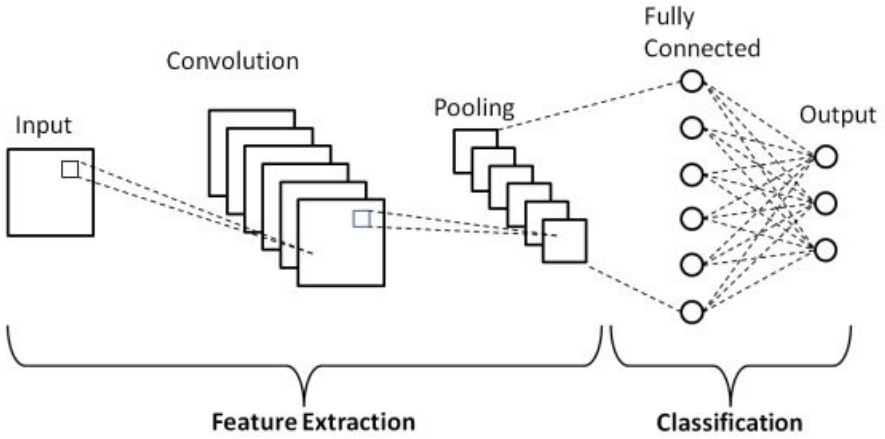
Vision Transformer (ViT)



1 - Based on WiSDoM slides by Professor João Magalhães

Did ViT overthrow CNNs? Not quite..





Both of these architectures co-exist!

Both with their respective advantages and disadvantages

1 - An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale; 2- CNN Overview

ViT vs. CNNs: The Showdown

The choice depends on **dataset size**, **computational resources**, and **task complexity**.

ViT

Advantages

Captures **global-context**

Few inductive biases

Disadvantages

Computationally **expensive**

Data hungry

CNN

Advantages

Computationally **efficient**

Smaller datasets

Disadvantages

Global context is **limited**

Limited flexibility

Vision

Text

ViT & Multimodal Tasks

It is very simple to leverage ViT for multimodal tasks since both vision and text quickly **align within the same embedding space** due to similarities in architecture and data representation.

How do we combine both modalities?

Fusion Techniques

Single Modality
(no fusion)

Single-stream Transformer
(early fusion)

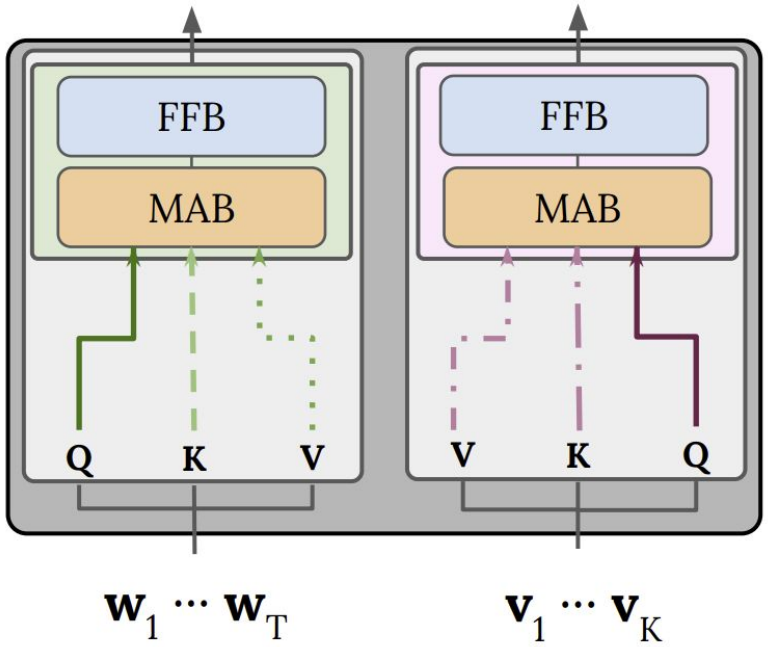
Dual-stream Transformer
(late fusion)

Dual Encoders

One **encoder** per **modality!**

Each modality is processed **independently** leading to **text** and **vision** embeddings being computed **separately**

The embeddings are only aligned in a **shared embedding space**!



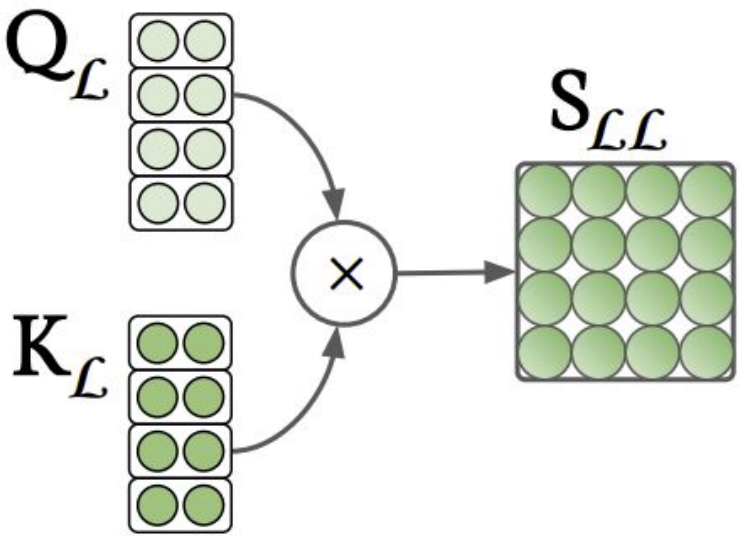
MAB = Multi-Head Attention Block
FFB = Feed-Forward Block

1 - Based on WiSDoM slides by Professor João Magalhães

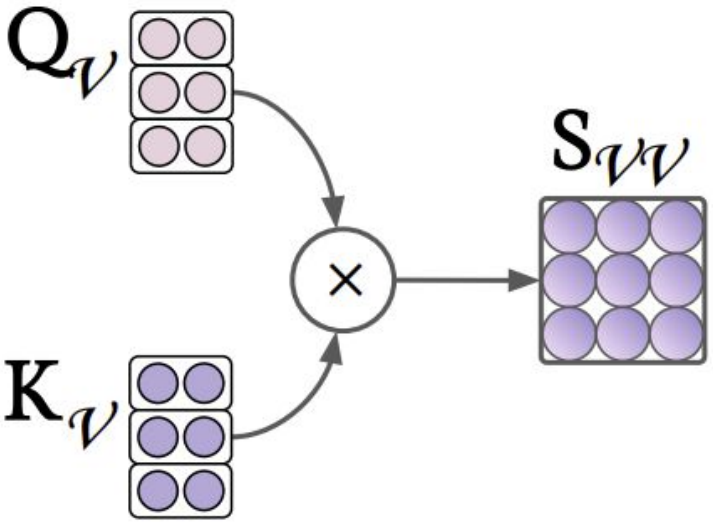
Single Modality
(no fusion)

Dual Encoders: Self-Attention

Text Self-Attention



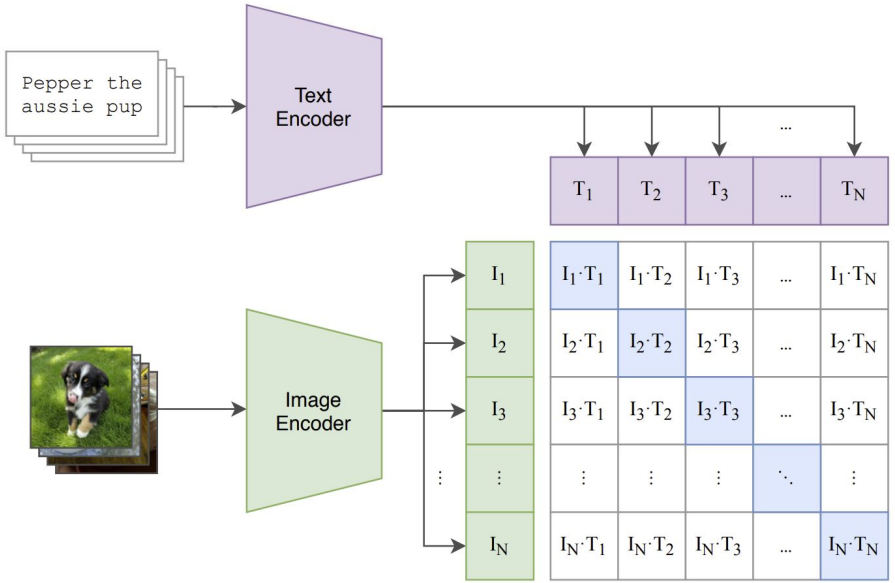
Vision Self-Attention



1 - Based on WiSDoM slides by Professor João Magalhães

CLIP: Contrastive Language-Image Pre-training

Multimodal model that aligns visual and textual embeddings in a shared semantic space using a **contrastive learning** objective.

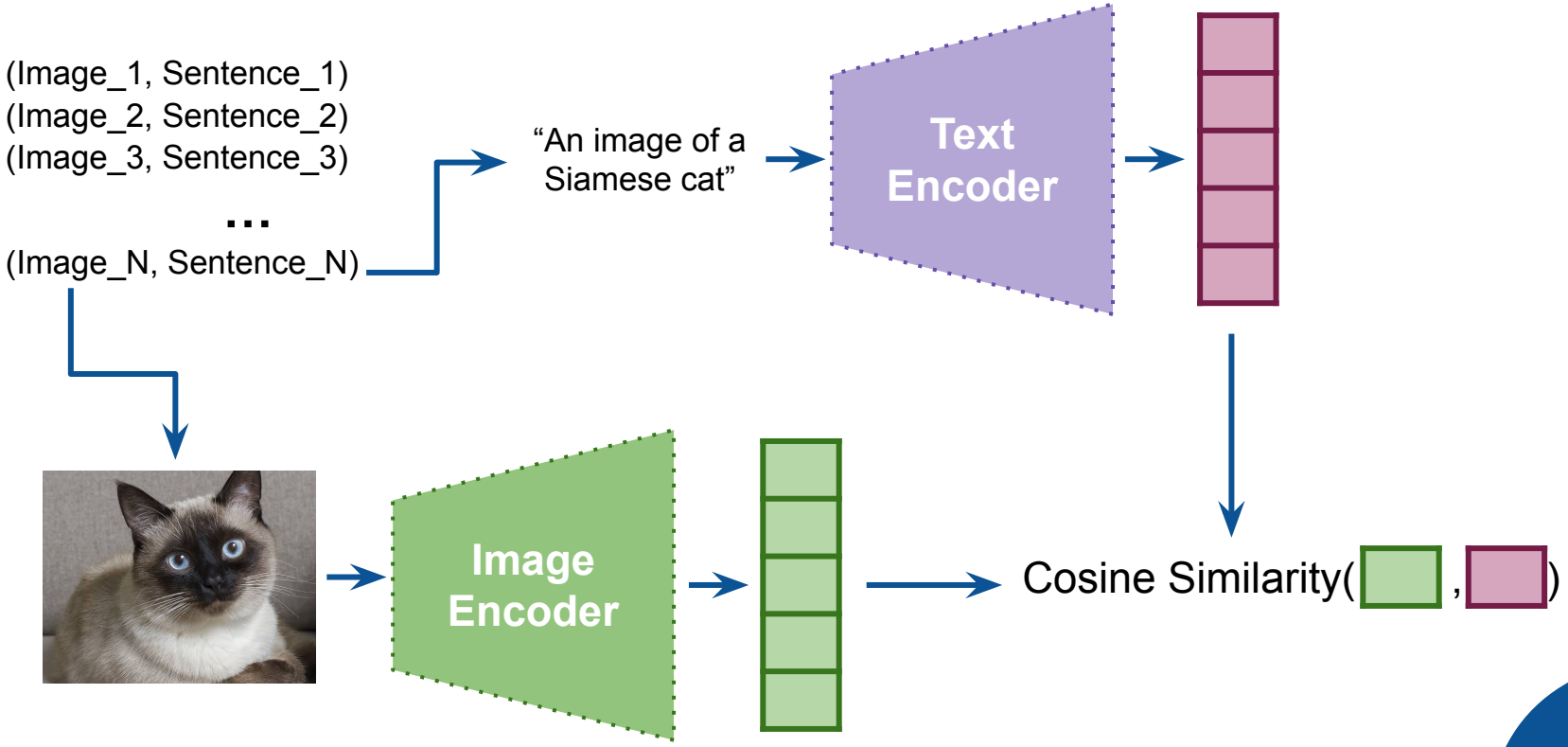


Trained on **400 million** image/sentence pairs: 1 image-sentence pair/second for 12 days!

1. **Two encoders**, one per modality
2. **Independent** embedding computation
3. **Shared** embedding space

1 - CLIP paper; 2 - Based on WiSDoM slides by Professor João Magalhães

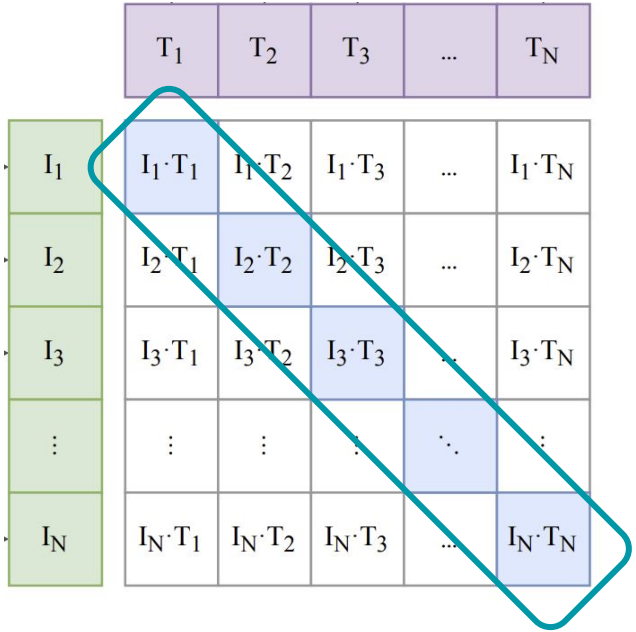
CLIP: Two independent encoders



1 - CLIP paper

CLIP: Contrastive Learning Objective

This model learns by **maximizing** the similarity of **correct** image-text pairs while **minimizing** the similarity of **incorrect** pairs.



We use the similarity scores to obtain two probability distributions: one for the **text-to-image (T2I)** and one for **image-to-text (I2T)**

Softmax

1. Compute Cross-entropy loss **T2I**
2. Compute Cross-entropy loss **I2T**
3. **Average** them!

1 - CLIP paper;

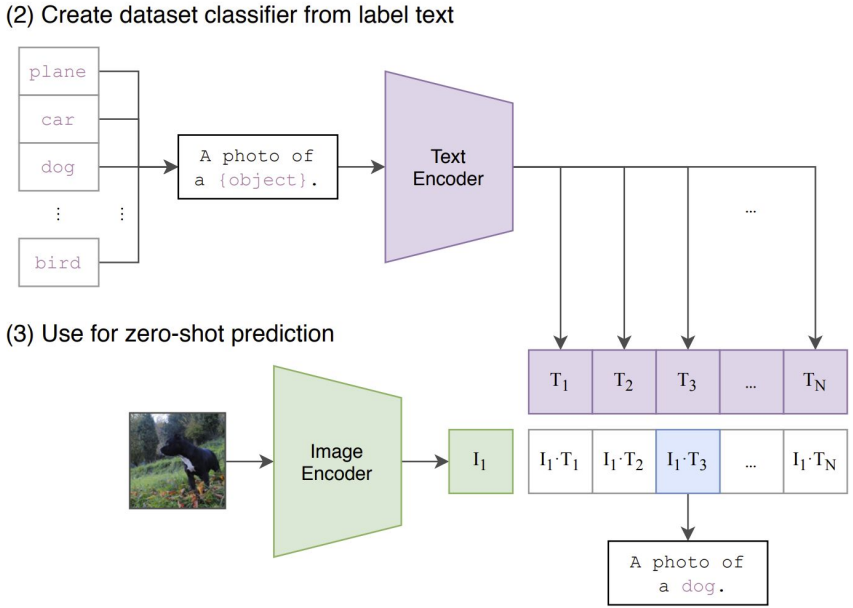
CLIP: Zero-Shot Capabilities

This model **generalizes** extremely well in tasks for which it was never trained for (**zero-shot prediction**).

Image annotation can be achieved with prompts like:

A photo of a _____.

It can also be used for **image-to-text** and **text-to-image** retrieval.



CLIP: Pre-training and Data

A major motivation for natural language supervision is the large quantities of data of this form available publicly on the internet. Since existing datasets do not adequately reflect this possibility, considering results only on them would underestimate the potential of this line of research. To address this, we constructed a new dataset of 400 million (image, text) pairs collected form a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries.[1] We approximately class

[1]The base query list is all words occurring at least 100 times in the English version of Wikipedia. This is augmented with bi-grams

their local batch of embeddings. The largest ResNet model, RN50x64, took 18 days to train on 592 V100 GPUs while the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336 pixel resolution for one additional epoch to boost performance similar to FixRes (Touvron et al., 2019). We denote this model as ViT-L/14@336px. Unless otherwise specified, all results reported in this paper as "CLIP" use this model which we found to perform best.

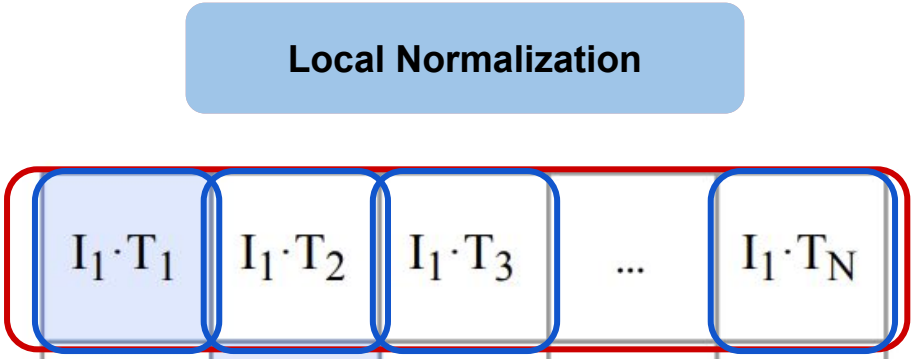
1 - CLIP paper; 2 - Based on WiSDoM slides by Professor João Magalhães

Single Modality
(no fusion)

SigLIP: The New Age

SigLIP (Sigmoid Loss for Language-Image Pre-training) replaces the softmax loss with a **pairwise sigmoid loss**.

Operates independently for each image-pair!



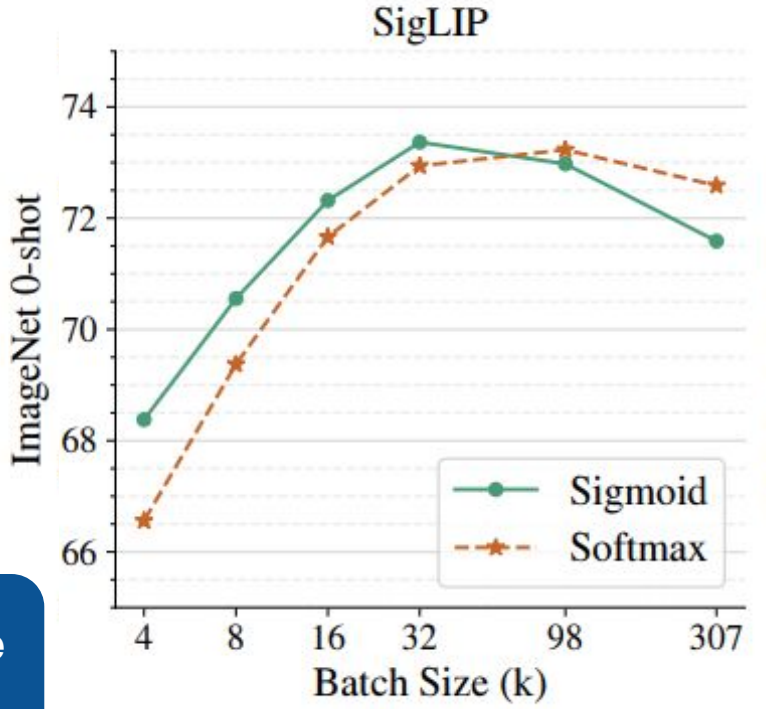
1 - SigLIP Paper

SigLIP: Why?

This is a more efficient alternative, reducing the need for such intense computational resources, as the loss is now computed over the **entire similarity matrix**

Performs **comparably or better** than softmax, especially at smaller batch sizes

Scales more efficiently as batch size increases



1 - SigLIP Paper

SigLIP: The proof is in the pudding

SigLIP (2023)

	Image	Text	BS	#TPUv4	Days
SigLIP	B/16	B	16 k	16	3
SigLIP	B/16	B	32 k	32	2
SigLIP	B/16	B	32 k	32	5

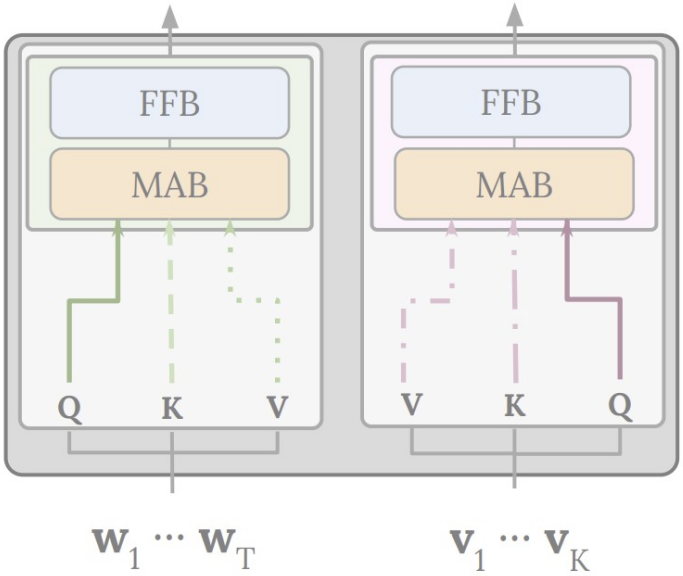
CLIP (2021)

Batch size: 32

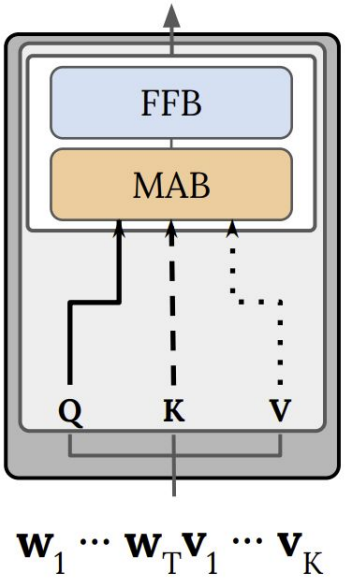
the largest Vision Transformer took 12 days on 256 V100 GPUs. For the ViT-L/14 we also pre-train at a higher 336

1 - SigLIP Paper; 2 - CLIP paper

Single Modality
(no fusion)



Single-stream Transformer
(early fusion)



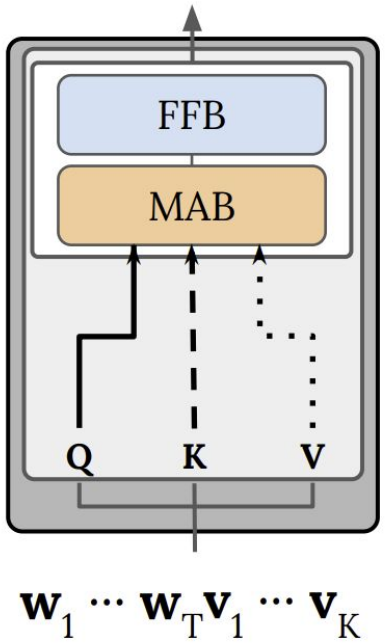
1 - Based on WiSDoM slides by Professor João Magalhães

Early Fusion: One Single Encoder

Both modalities are combined at **input level** by **concatenating** both text and image features.

Allows the model to learn **cross-modal interactions** from the beginning by using **one single encoder**

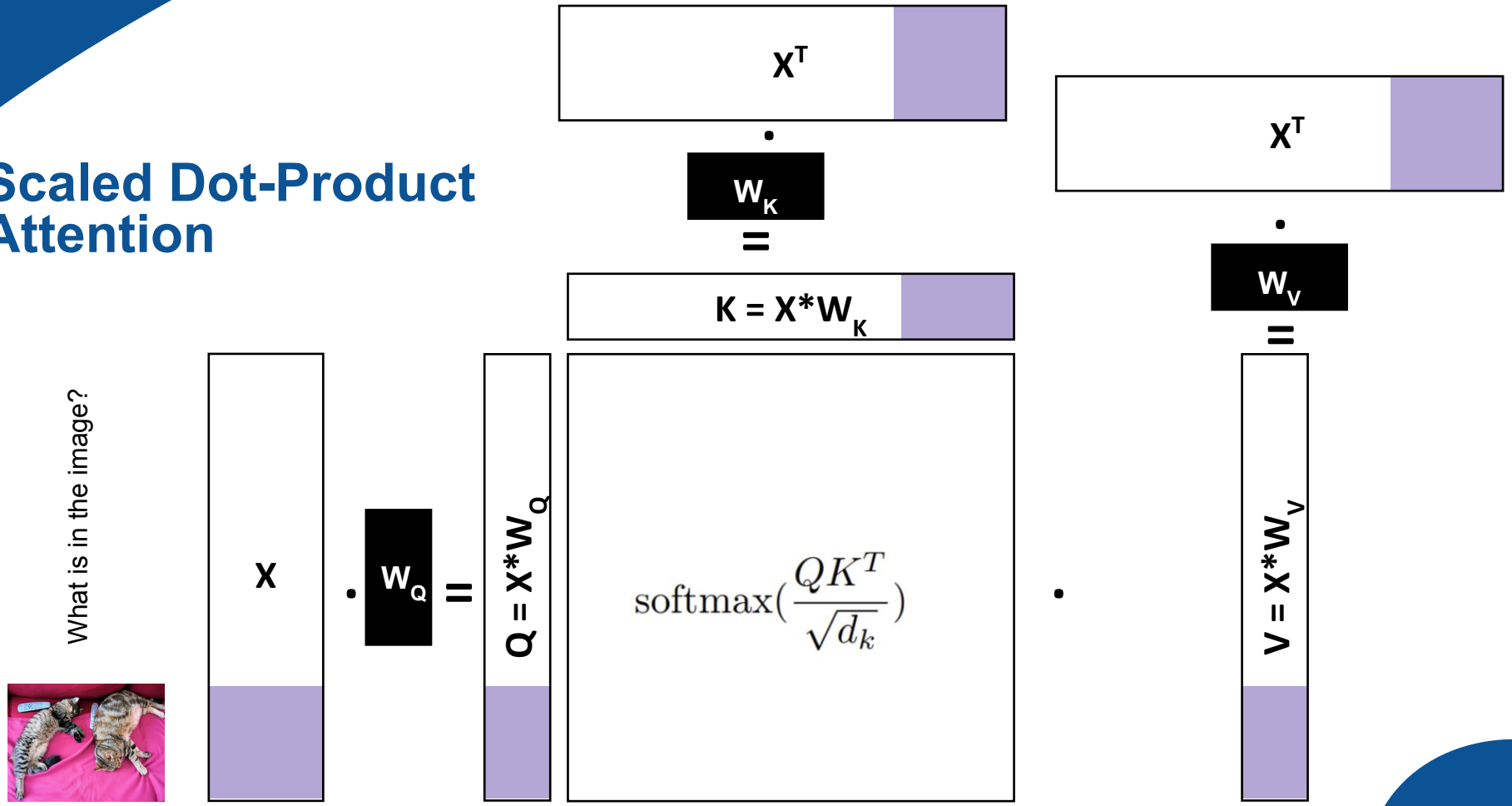
Attention can be computed over both modalities!



MAB = Multi-Head Attention Block
FFB = Feed-Forward Block

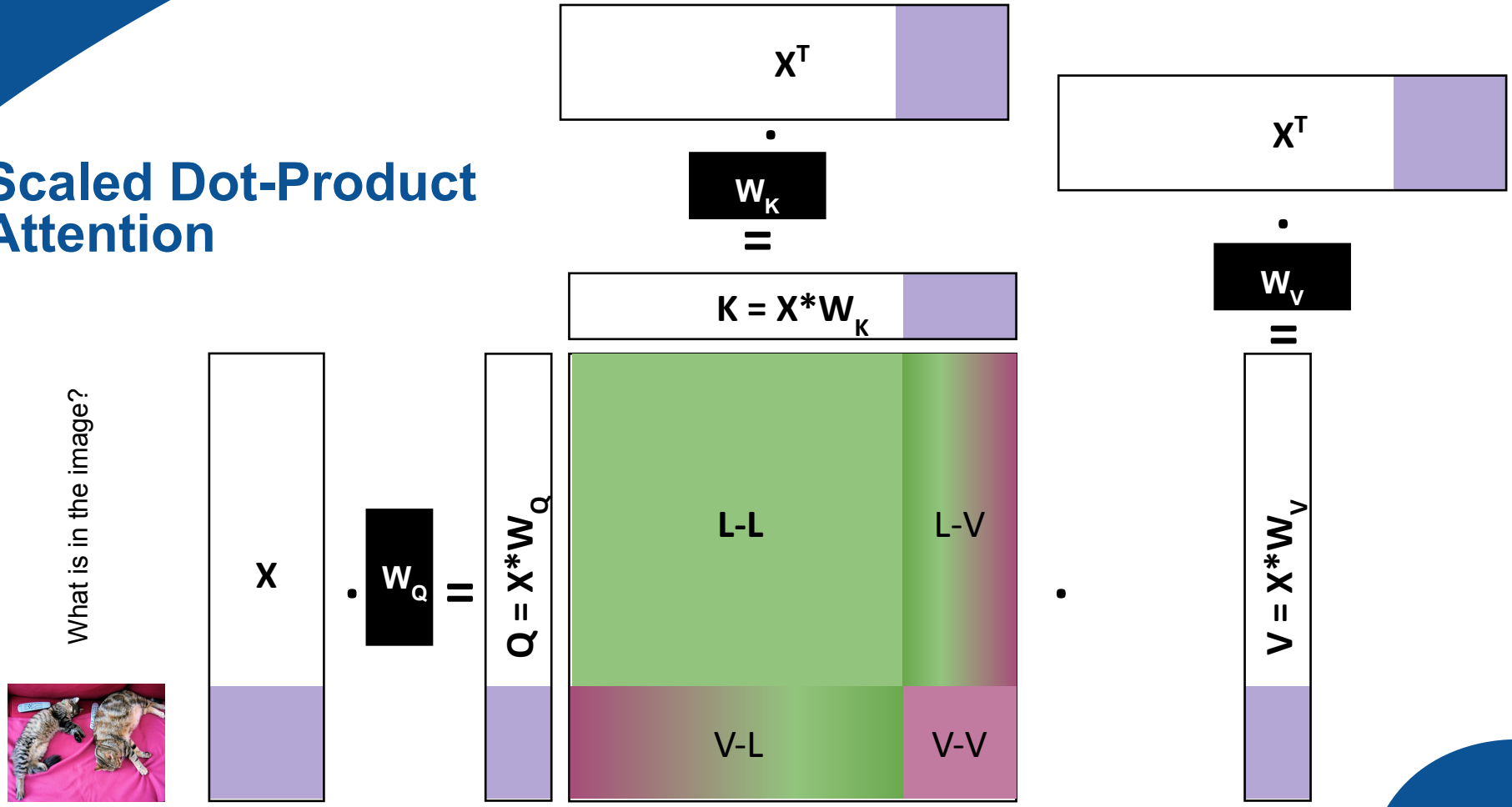
1 - Based on WiSDoM slides by Professor João Magalhães

Scaled Dot-Product Attention



1 - Based on WiSDoM slides by Professor João Magalhães

Scaled Dot-Product Attention



1 - Based on WiSDoM slides by Professor João Magalhães

Early Fusion: VL BERT

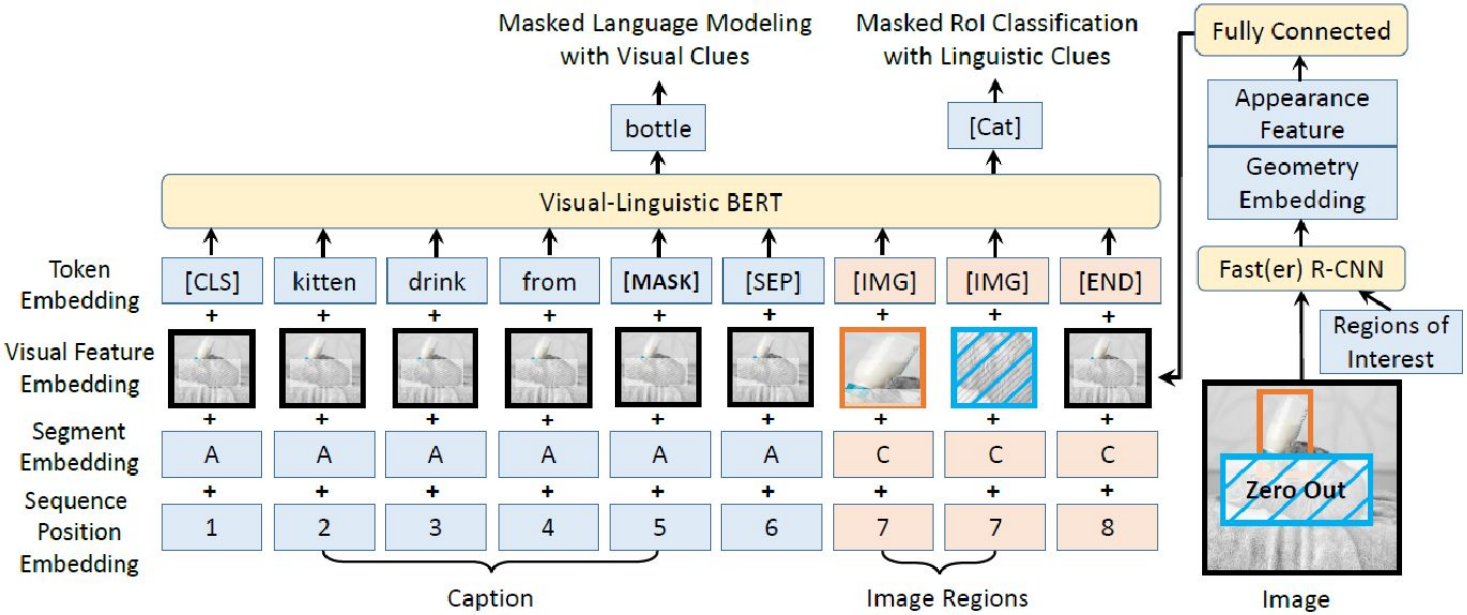
An extension of the original BERT model designed to **jointly** process visual and language data

Pre-training tasks:

- **Masked Language Modeling** - It uses the **multimodal context** to predict the masked word
- **Masked Object Classification** - Similar to MLM, a region of the image is masked (i.e., A bounding box of a dog)

1 - VL BERT Paper; 2 - Based on WiSDoM slides by Professor João Magalhães

Early Fusion: VL BERT



1 - VL BERT Paper

Early Fusion: Is it better?

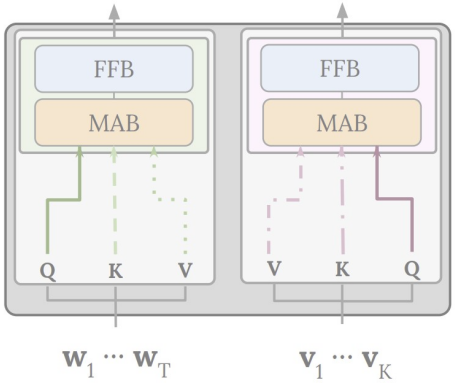
Single Modality
(no fusion)

- ✅ Efficient Training
- ✅ Preserves independent modality representations
- ❌ Limited cross-modal interaction

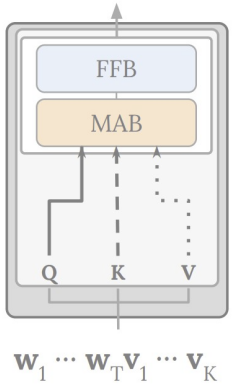
Single-stream Transformer
(early fusion)

- ✅ Enables early cross-modal interactions (joint attention)
- ✅ Ideal for tasks requiring deep vision-language reasoning
- ❌ Computationally expensive

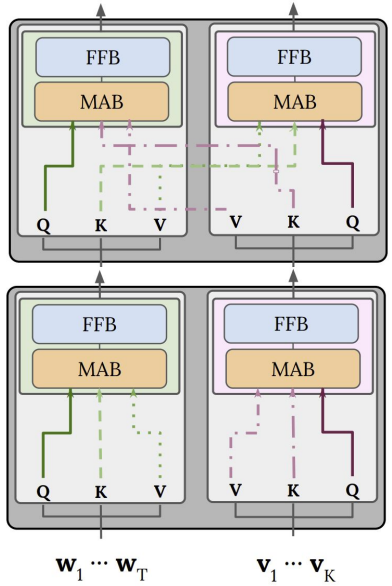
Single Modality
(no fusion)



Single-stream Transformer
(early fusion)



Dual-stream Transformer
(late fusion)



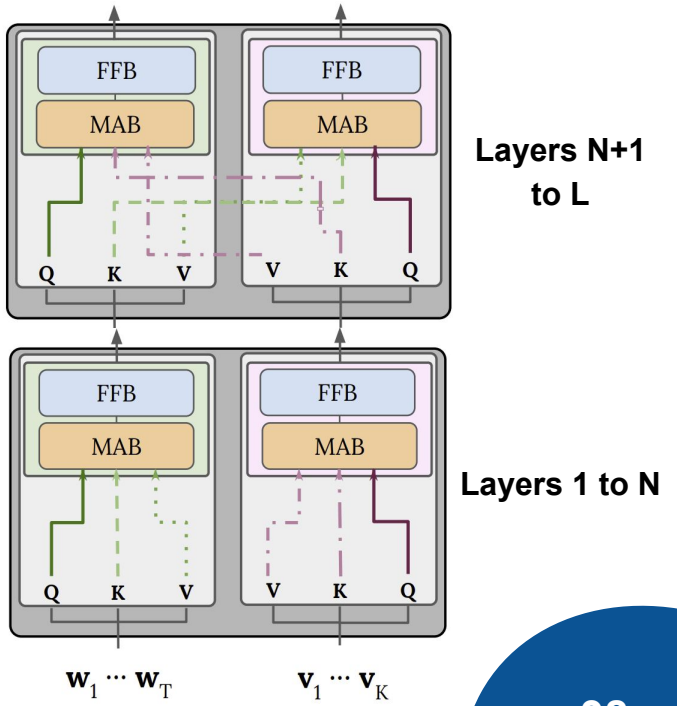
1 - Based on WiSDoM slides by Professor João Magalhães

Late Fusion: Two types of Layers

The **dual stream** approaches involves two different types of transformer layers:

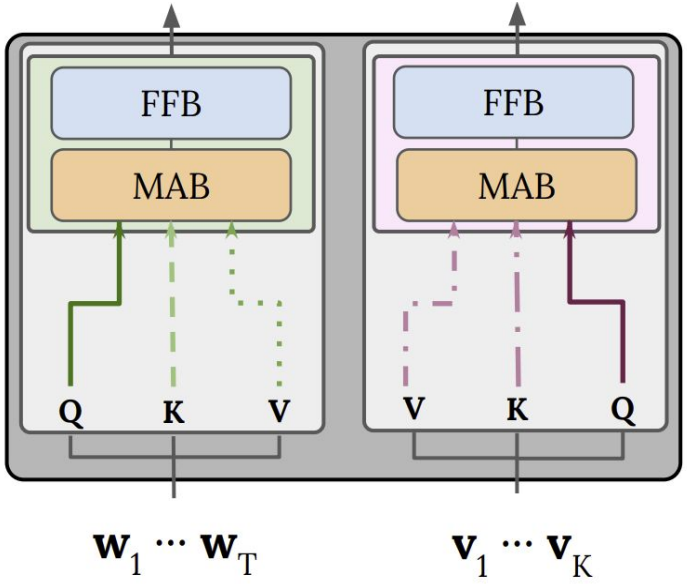
Inter-model Layers: Embeddings from different modalities are exchanged across streams

Intra-model Layers: Separate streams for language and vision



1 - Based on WiSDoM slides by Professor João Magalhães

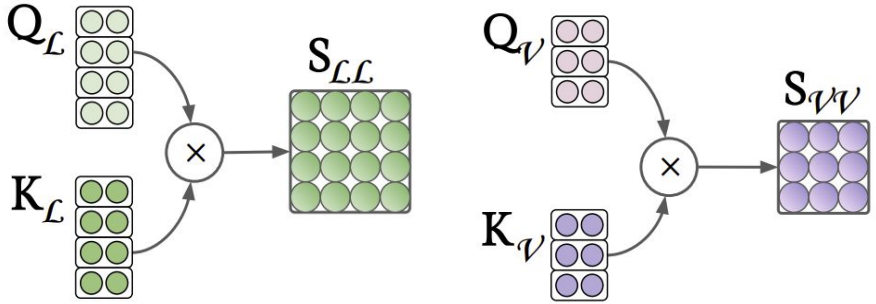
Late Fusion: Intra-modal



Layers 1 to N

Does this look familiar?

Single Modality
(no fusion)

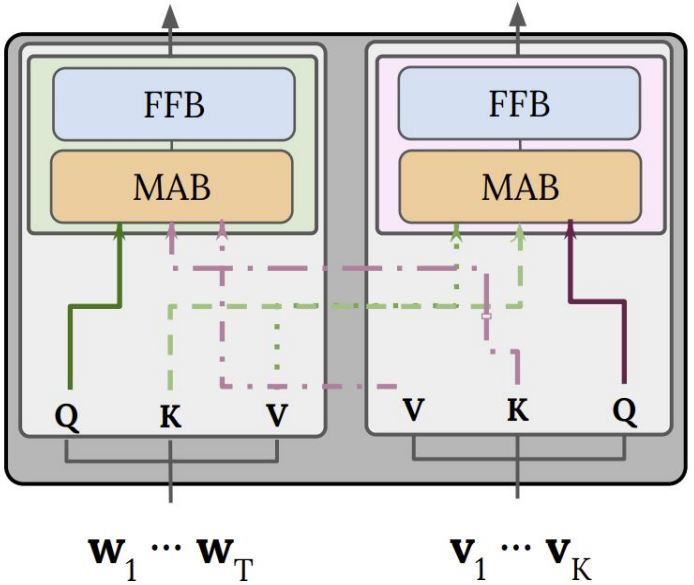


1 - Based on WiSDoM slides by Professor João Magalhães

Late Fusion: Inter-modal

These layers **explicitly** model cross-modal interaction via a **cross-modal attention module**

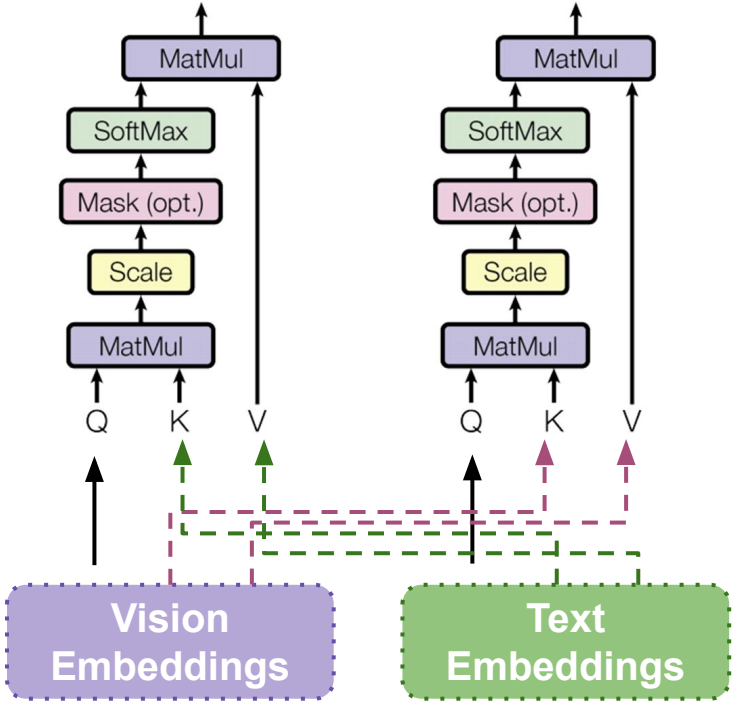
Co-attention mechanism



MAB = Multi-Head Attention Block
FFB = Feed-Forward Block

1 - Based on WiSDoM slides by Professor João Magalhães

Co-attention Mechanism: Inter-modal



The **keys** and **values** from each modality are used as input to the other modality's multi-headed attention block.

$$\text{Attention}(Q_v, K_w, V_w) = \text{softmax}\left(\frac{Q_V \cdot K_w}{\sqrt{d_k}}\right) \cdot V_w$$

$$\text{Attention}(Q_w, K_v, V_v) = \text{softmax}\left(\frac{Q_w \cdot K_v}{\sqrt{d_k}}\right) \cdot V_v$$

1 - Based on WiSDoM slides by Professor João Magalhães

Co-attention Mechanism: Inter-modal

The attention block produces **attention-pooled features** for each modality conditioned on the other:

1. Language attention is conditioned in the visual stream

   Keys and **Values** come from **Image embeddings**

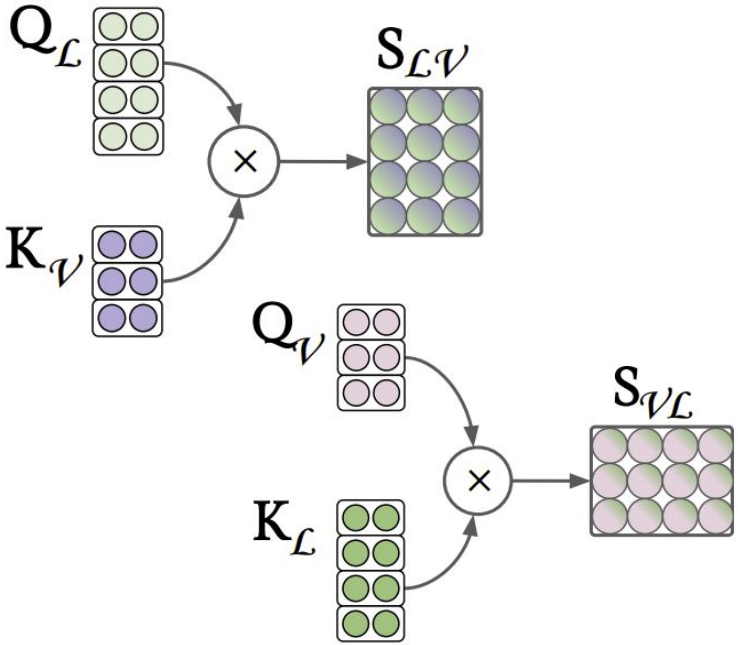
2. Image attention is conditioned in the linguistic stream

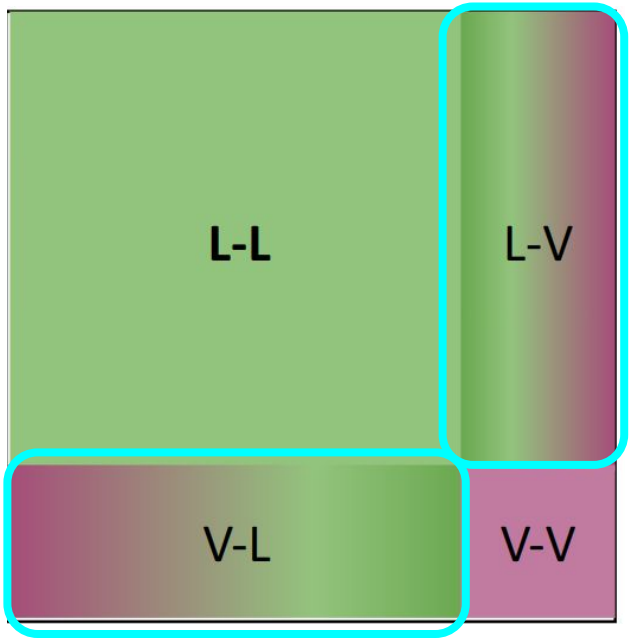
   Keys and **Values** come from **Text embeddings**

This mechanism ensures that each modality learns context-aware representations by leveraging information from the other modality!

1 - Based on WiSDoM slides by Professor João Magalhães

Multimodal Co-Attention



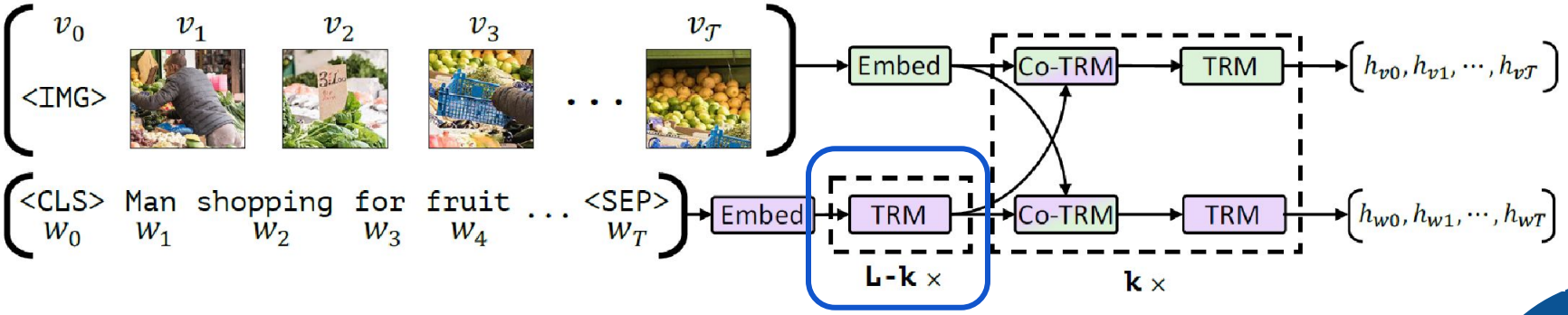


1 - Based on WiSDoM slides by Professor João Magalhães

Late Fusion: ViLBERT

This model leverages **two separate streams** with the modalities interacting through **co-attention layers**.

There's also a difference in **processing** depth between both modalities (vision involves fairly more high-level features)



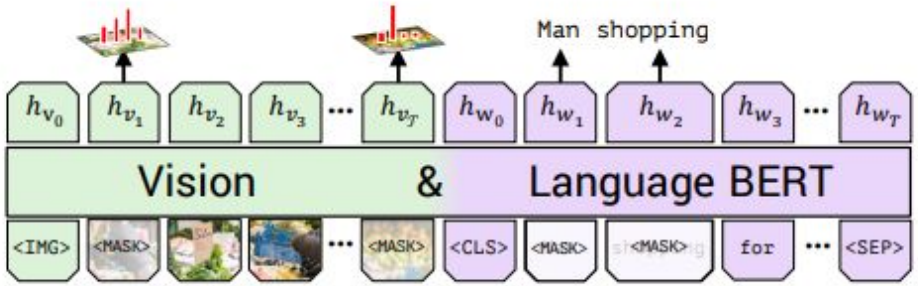
TRM = Transformer Blocks

1 - ViL BERT Paper; 2- Based on WiSDoM slides by Professor João Magalhães

ViLBERT: Pre-training

Masked Multimodal Learning: Similar to the standard MLM but incorporates both vision and language.

- **Masking text tokens** in a sentence and predicting them using visual context and surrounding text.
- **Masking image regions** and predicting their categories using text context and visual information

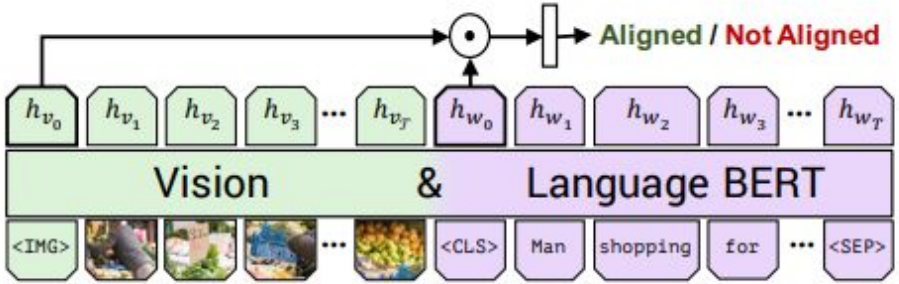


(a) Masked multi-modal learning

1 - ViL BERT Paper; 2- Based on WiSDoM slides by Professor João Magalhães

ViLBERT: Pre-training

Multimodal-Alignment Prediction: The model predicts whether or not the caption describes the image content (i.e., whether the text and the image are **aligned** or not)



(b) Multi-modal alignment prediction

1 - ViL BERT Paper; 2- Based on WiSDoM slides by Professor João Magalhães

Single Modality (no fusion)	**Single-stream Transformer** (early fusion)	**Dual-stream Transformer** (late fusion)
✅ Efficient Training	✅ Enables early cross-modal interactions (joint attention)	✅ Explicit cross-modal alignment
✅ Preserves independent modality representations	✅ Ideal for tasks requiring deep vision-language reasoning	✅ Different processing needs
❌ Limited cross-modal interaction	❌ Computationally expensive	❌ Higher complexity

Exercise Time