Multimodal LLMs

CS 5624: Natural Language Processing

Spring 2025

https://tuvllms.github.io/nlp-spring-2025

Tu Vu

VIRGINIA TECH.

Logistics

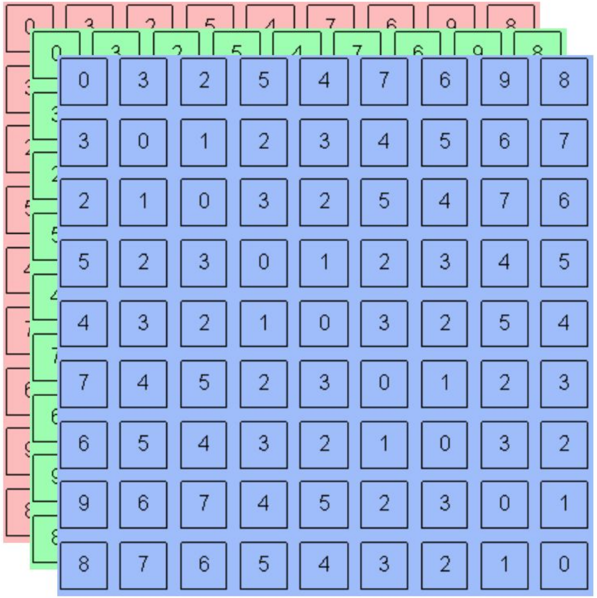
- Homework 2 due **5/5**
- Final project presentations **5/6**
- Final project report due **5/9**
- Final grades due **5/16**

Grayscale images

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

Slide from Mohit Iyyer

Color images

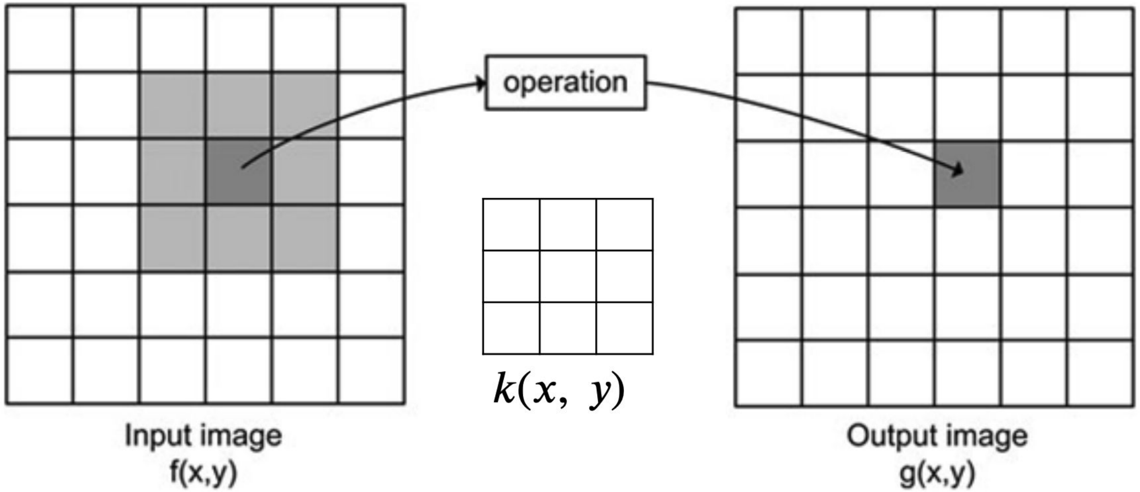


channel x height x width

Channels are usually RGB: Red, Green, and Blue

Slide from Mohit Iyyer

Convolution operator



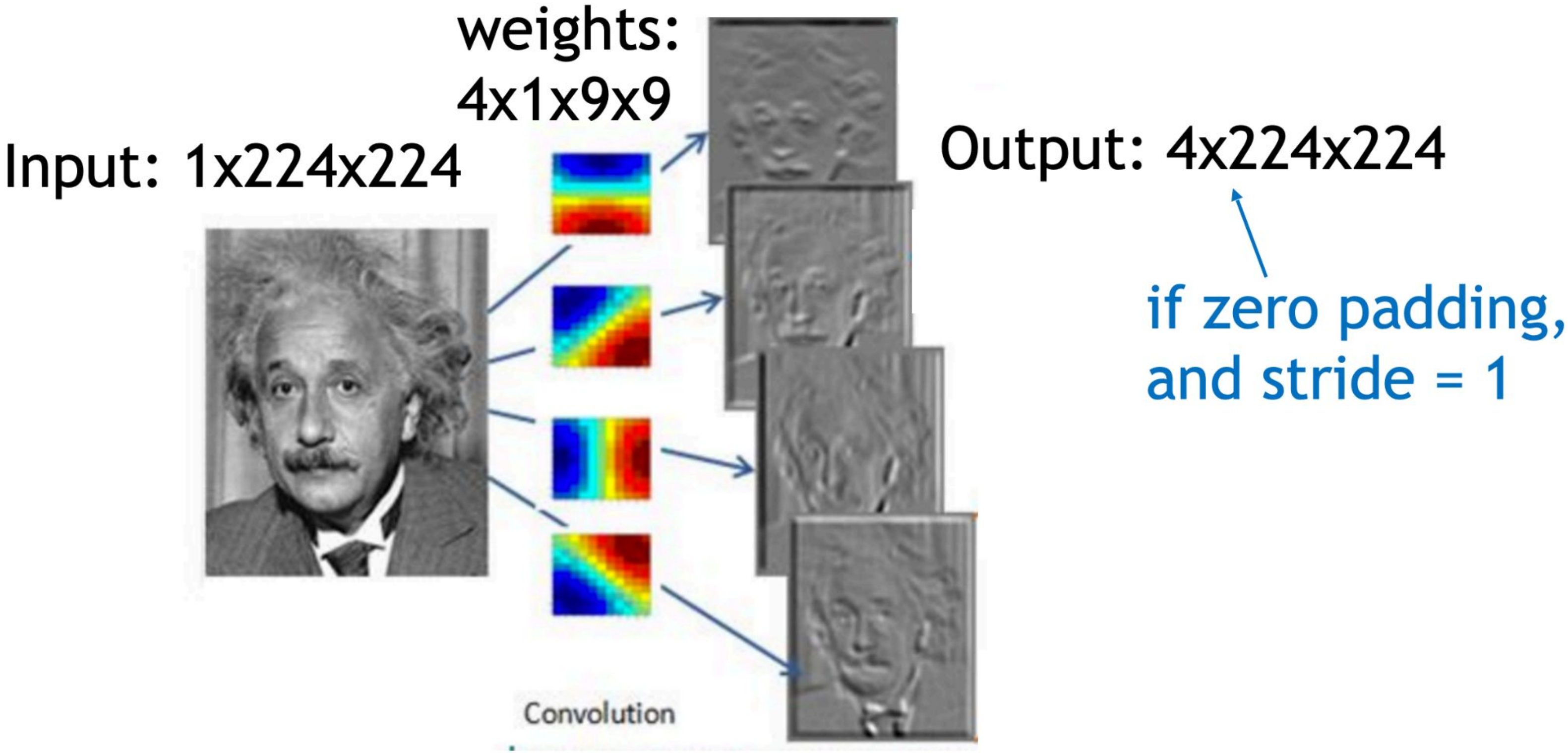
$$g(x, y) = \sum_{v} \sum_{u} k(u, v) f(x - u, y - v)$$

http://what-when-how.com/introduction-to-video-and-image-processing/neighborhood-processing-introduction-to-video-and-image-processing-part-1/

Demo

- https://setosa.io/ev/image-kernels/

Convolutional layer (with 4 filters)



Slide from Mohit Iyyer

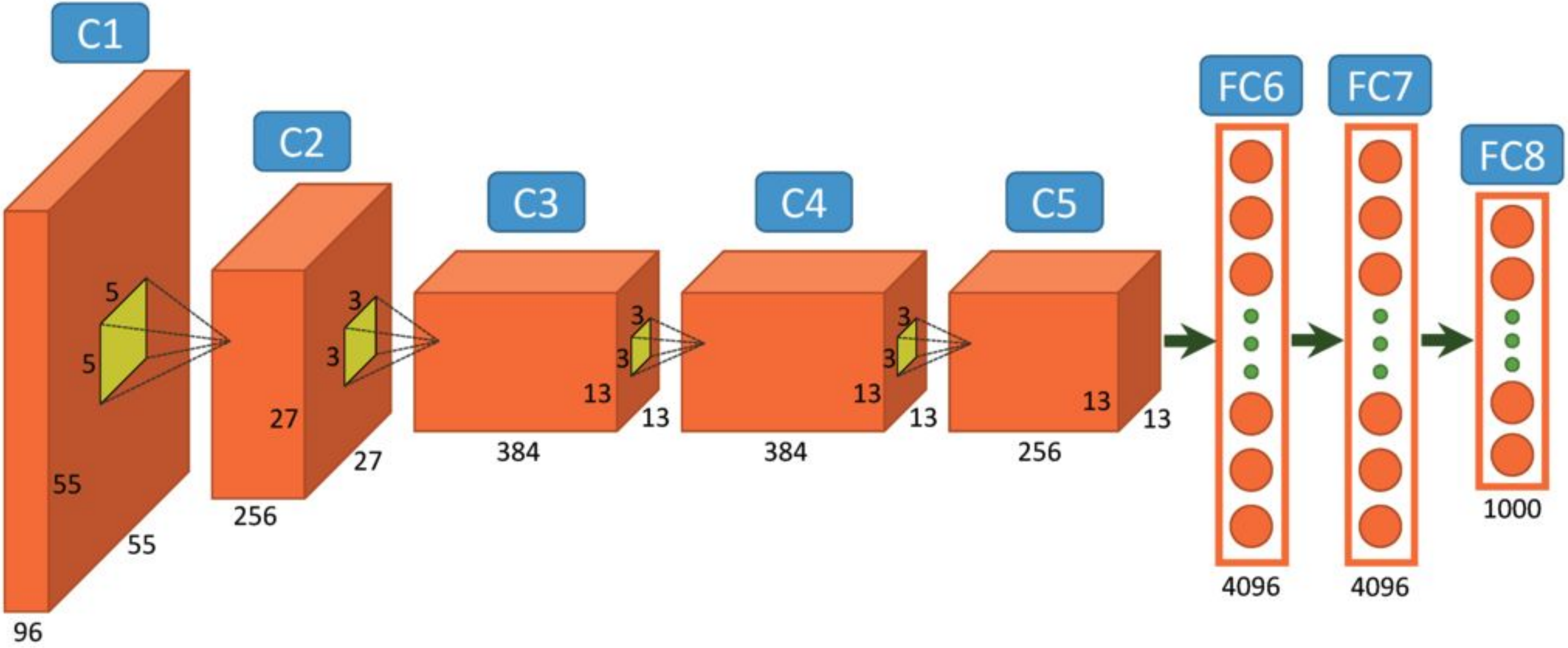
ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

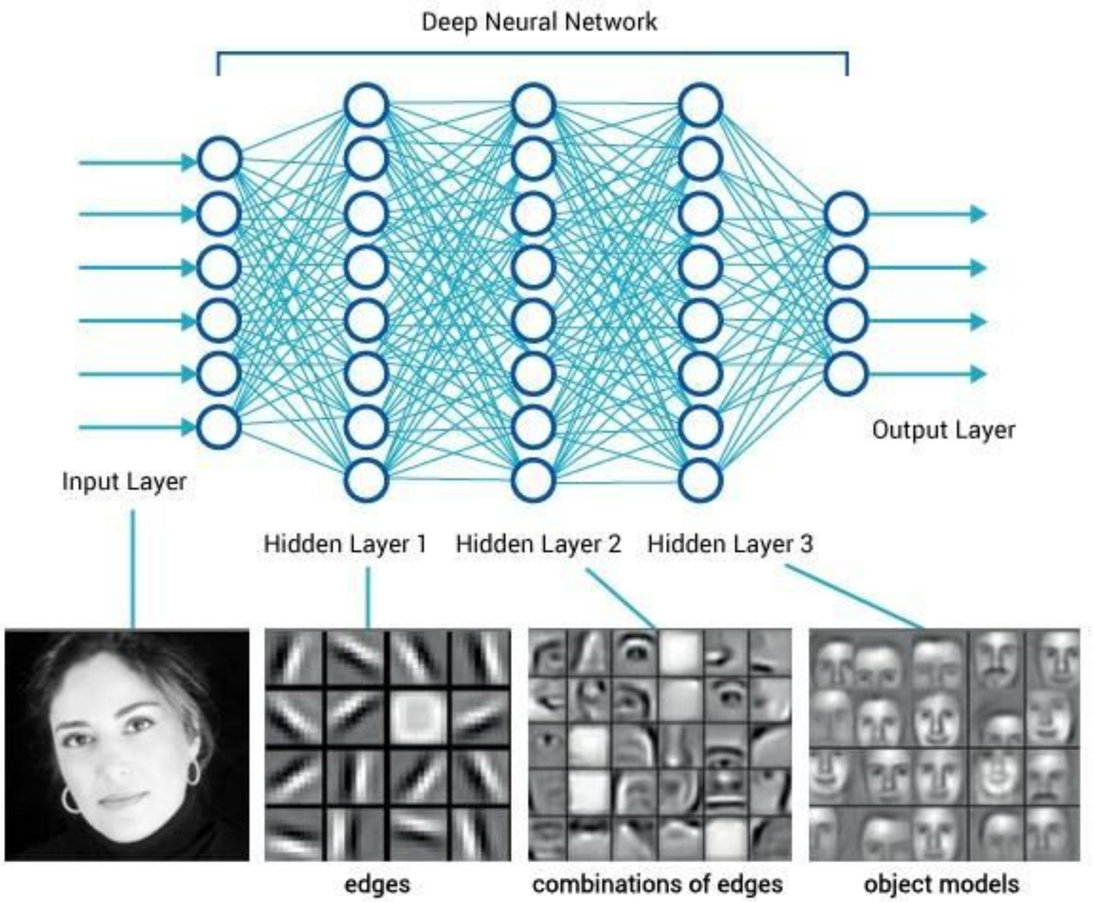
Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

AlexNet



https://www.saagie.com/en/blog/object-detection-part1/

AlexNet (cont'd)



https://www.saagie.com/en/blog/object-detection-part1/

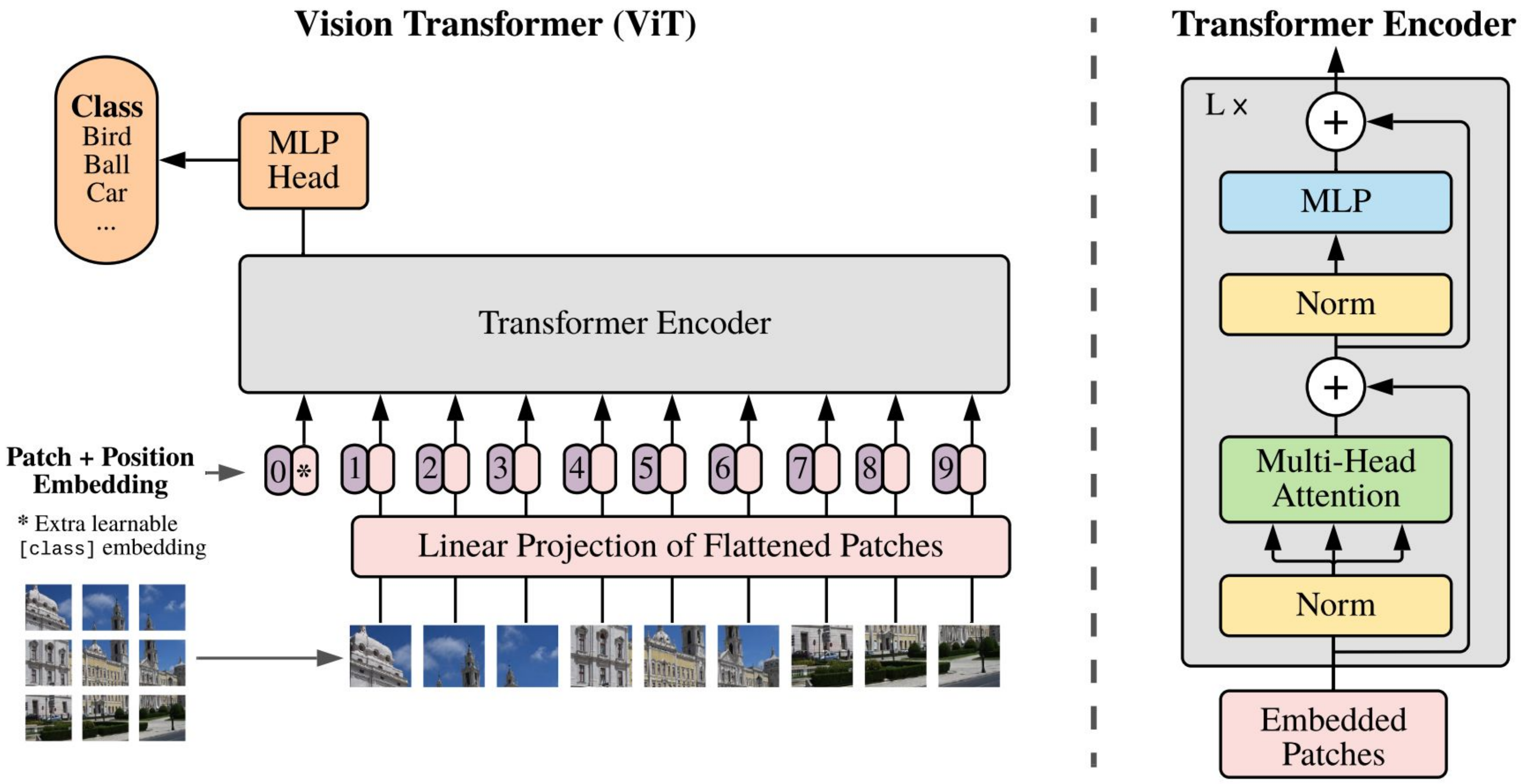
AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE

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


Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0 * 1 2 3 4 5 6 7 8 9
Linear Projection of Flattened Patches
Transformer Encoder
L ×
MLP
Norm
Multi-Head Attention
Norm
Embedded Patches

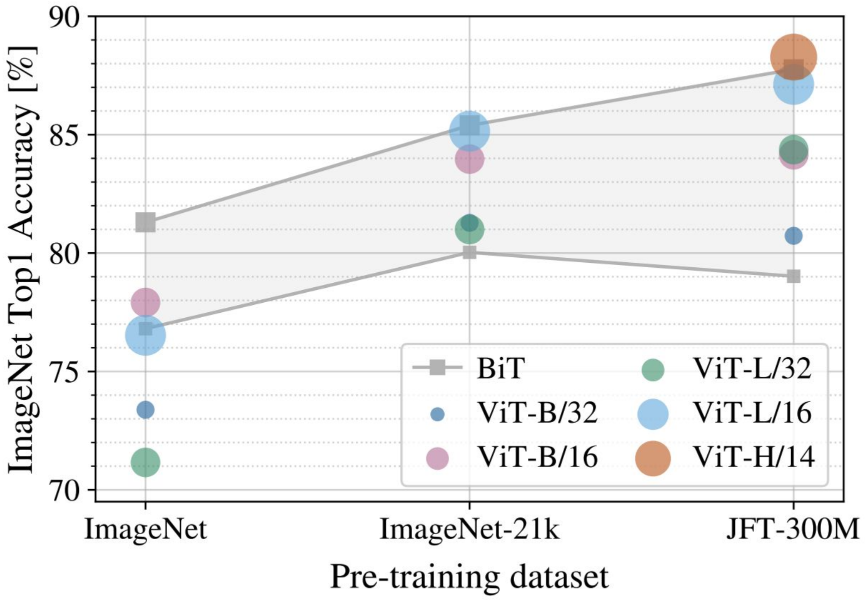
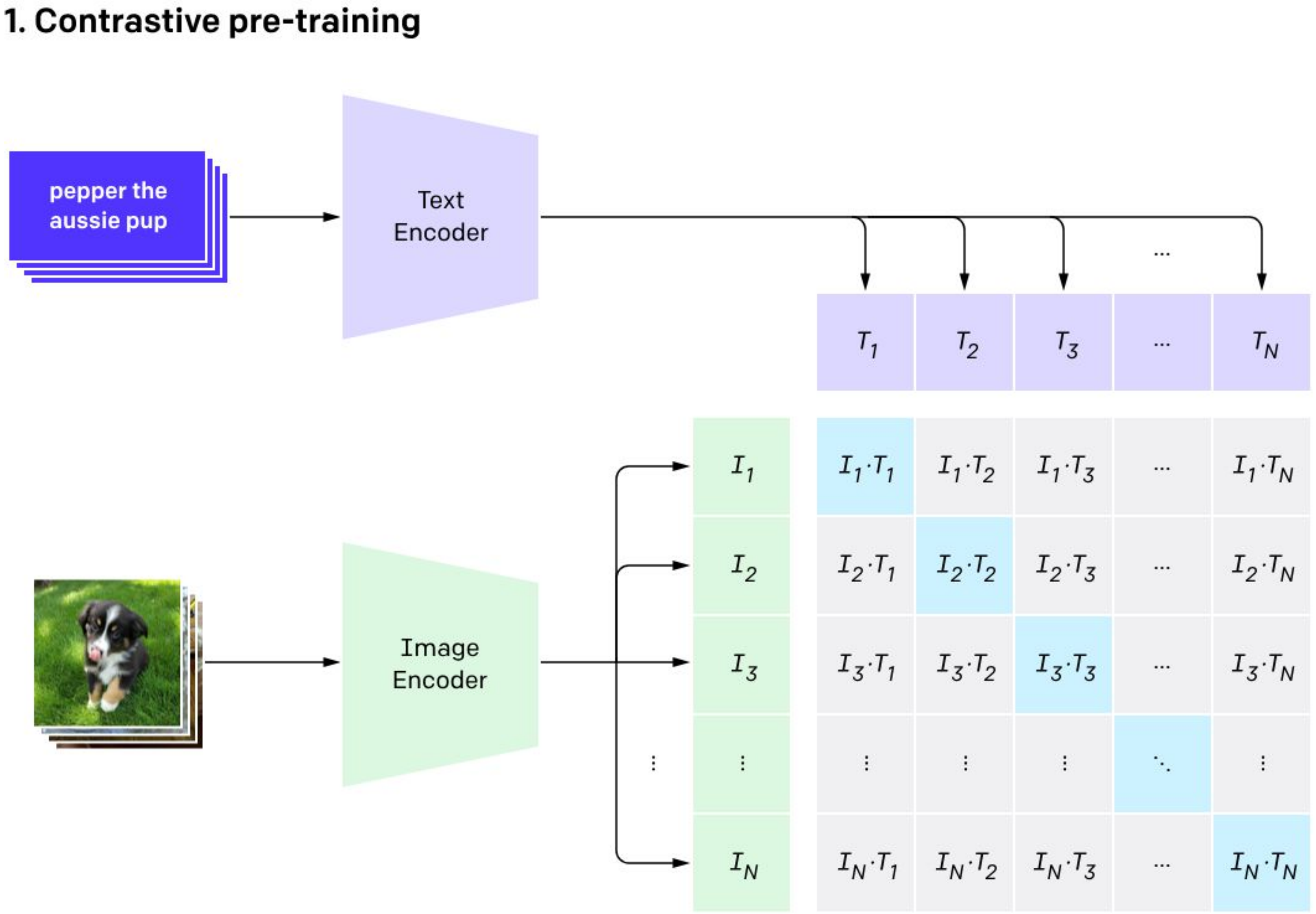


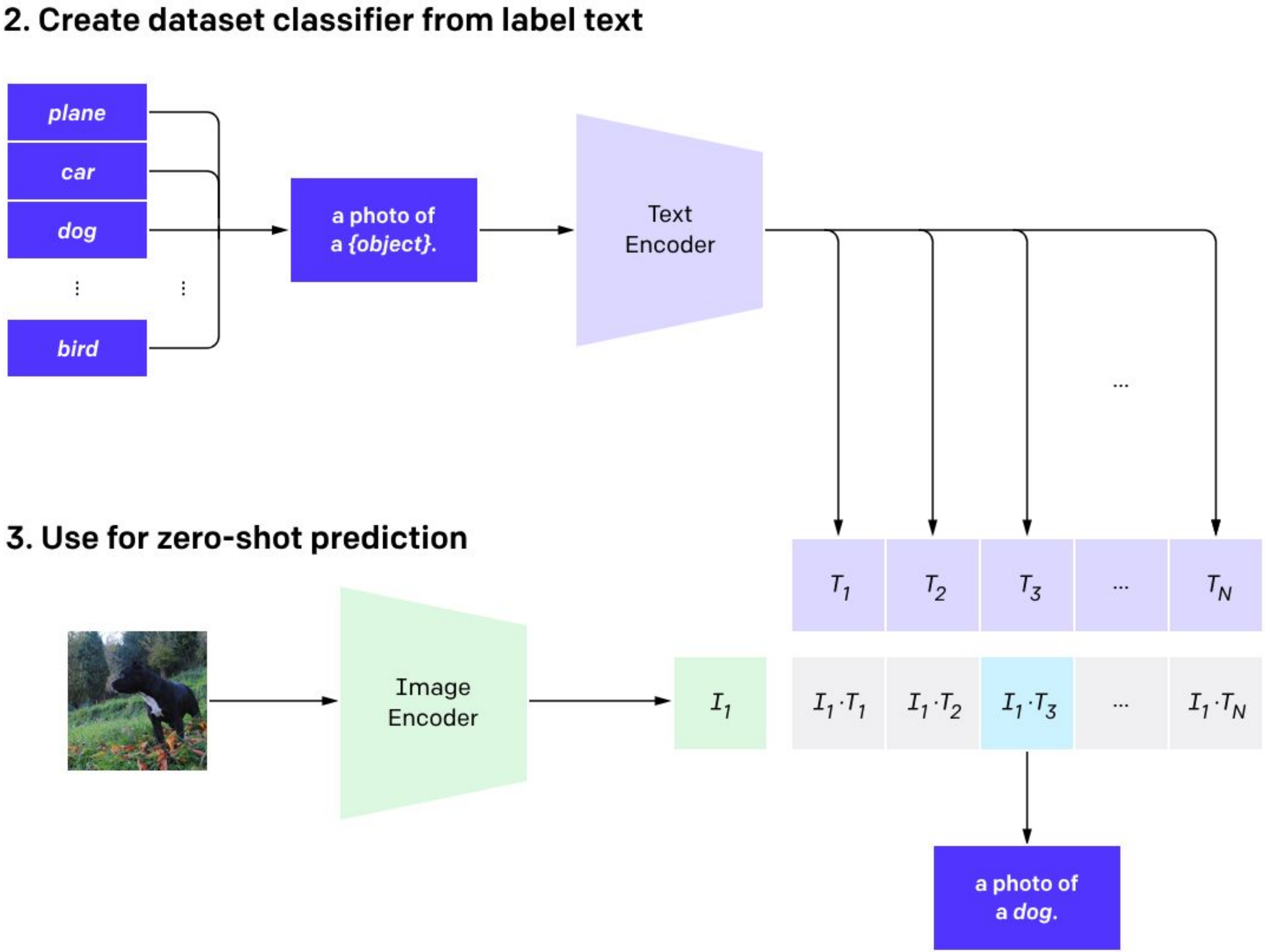
Figure 3: Transfer to ImageNet. While large ViT models perform worse than BiT ResNets (shaded area) when pre-trained on small datasets, they shine when pre-trained on larger datasets. Similarly, larger ViT variants overtake smaller ones as the dataset grows.

OpenAI's CLIP



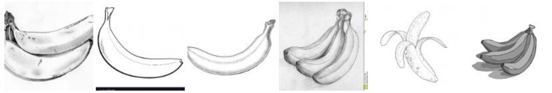
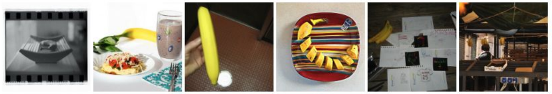
https://openai.com/index/clip/

OpenAI's CLIP (cont'd)



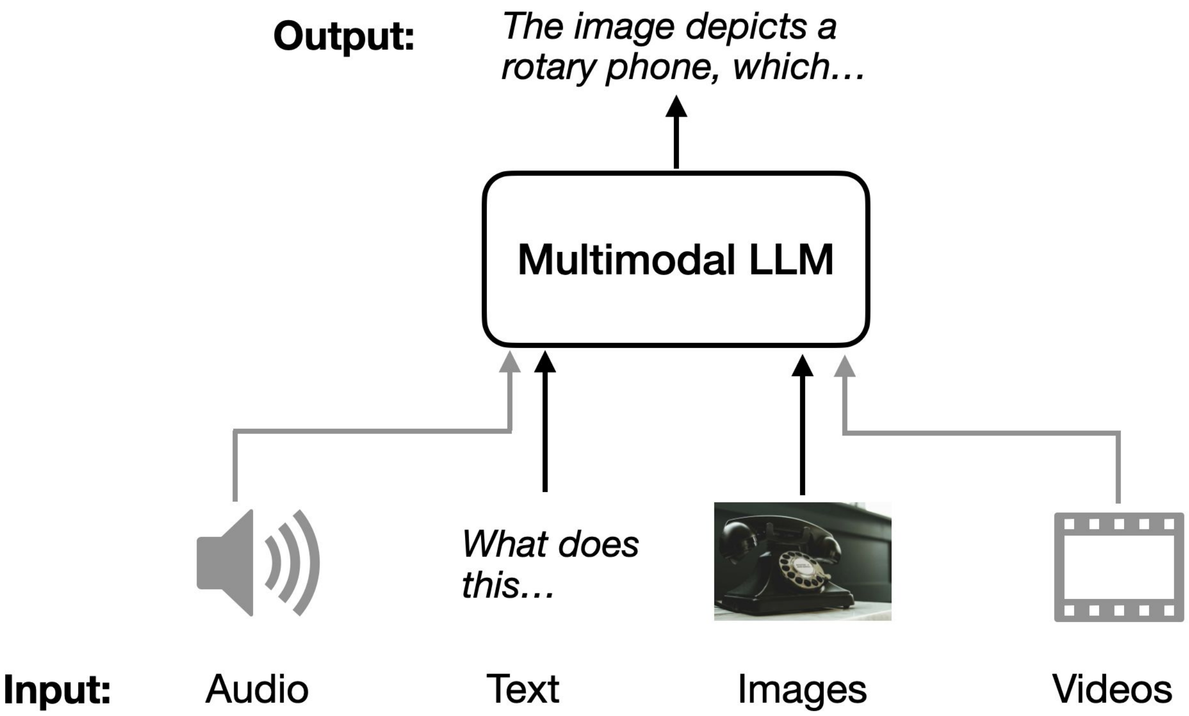
https://openai.com/index/clip/

OpenAI's CLIP (cont'd)

Dataset	ImageNet ResNet101	CLIP ViT-L
 ImageNet	76.2%	76.2%
 ImageNet V2	64.3%	70.1%
 ImageNet Rendition	37.7%	88.9%
 ObjectNet	32.6%	72.3%
 ImageNet Sketch	25.2%	60.2%
 ImageNet Adversarial	2.7%	77.1%

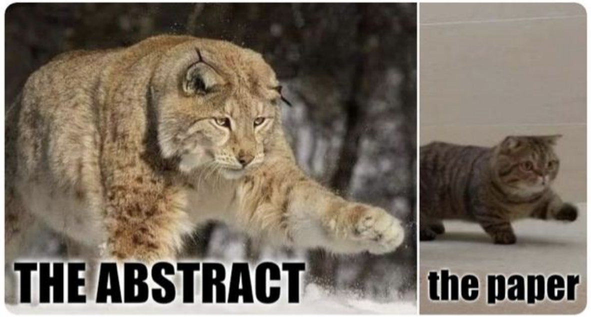
https://openai.com/index/clip/

A multimodal LLM can accept inputs in multiple modalities and emit text outputs



[Understanding Multimodal LLMs by Sebastian Raschka]

Use cases of multimodal LLMs



Explain the meme

This meme humorously captures the common experience many researchers and readers encounter when reading academic papers. It shows a powerful, fierce lynx labeled "THE ABSTRACT" on the left and a smaller, cautious-looking domestic cat labeled "the paper" on the right.

The joke here is that an abstract (the brief summary at the beginning of a paper) often makes the research sound exciting, groundbreaking, or impactful, like a "fierce lynx." However, the actual content of the paper may feel less impressive, or overly complex, dense, or even disappointing by comparison, represented by the "scaredy-cat" on the right. This is a playful take on how expectations can differ from reality in academic writing.

[Understanding Multimodal LLMs by Sebastian Raschka]

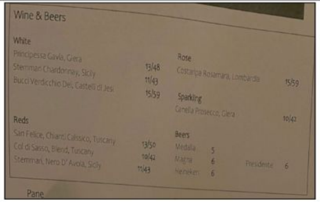
Use cases of multimodal LLMs (cont'd)

Fig. 1: MM1 can perform in-context predictions thanks to its large-scale multimodal pre-training. This allows MM1 to (a) count objects and follow custom formatting, (b) refer to parts of the images and perform OCR, (c) demonstrate common-sense and word knowledge about everyday objects, and (d) perform basic math functions. Images are from the COCO 2014 validation set [72].

[MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training]

Use cases of multimodal LLMs (cont'd)

User:
Photo: <img1> Menu: <img2> How much should I pay for all the beer on the table according to the price on the menu?



MM1-30B-Chat (Ours): 12

Emu-Chat-37B: According to the menu, the price for all the beer on the table is 15.99.

LLaVA-NeXT-34B: The image shows a menu with beer prices, but it's not entirely clear which specific beers are on the table. However, I can provide you with an estimate based on the visible prices. From the menu, it appears that the prices for beer are as follows: Magna: 5, Medalla: 5, Heineken: 6 Presidente: 6
Assuming that the beers on the table are Magna, Medalla, Heineken, and Presidente, and there are two of each, you would pay:
2 x Magna: 2 x 5 = 10, 2 x Medalla: 2 x 5 = 10, 2 x Heineken: 2 x 6 = 12, 2 x Presidente: 2 x 6 = 12 Total: 10 + 10 + 12 + 12 = 44
Please note that this is an estimate based on the visible prices and the assumption that there are two of each type of beer. The actual cost may vary if the beers on the table are not the ones listed on the menu or if there are different quantities.

User: Explain why.

MM1-30B-Chat (Ours): There are 2 beers on the table, and each beer costs 6 according to the price on the menu. So, 2*6=12.

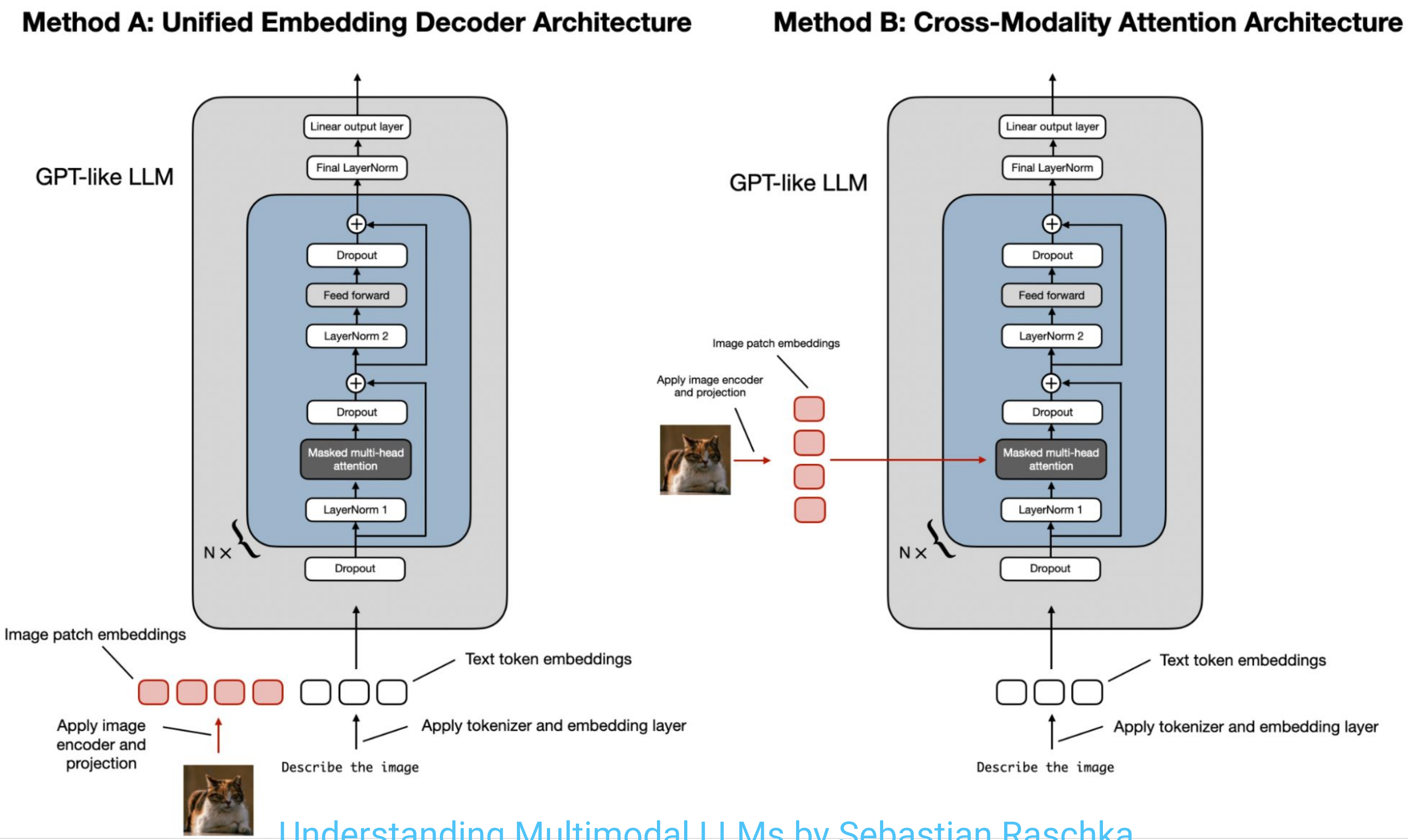
Emu-Chat-37B: The price for all the beer on the table is 15.99 because each beer has a price of 10.99, and there are a total of five beers on the table. When you multiply the number of beers (5) by the price per beer (10.99), you get a total cost of 54.95 (rounded up to 15.99).

LLaVA-NeXT-34B: The image shows a menu with beer prices, but it's not entirely clear which specific beers are on the table. However, I can provide you with an estimate based on the visible prices. From the menu, it appears that the prices for beer are as follows: (...)

Fig. 2: MM1 can follow instructions and reason across images. Example and images from VILA [71]; VILA answers correctly when prompted with chain-of-thought.

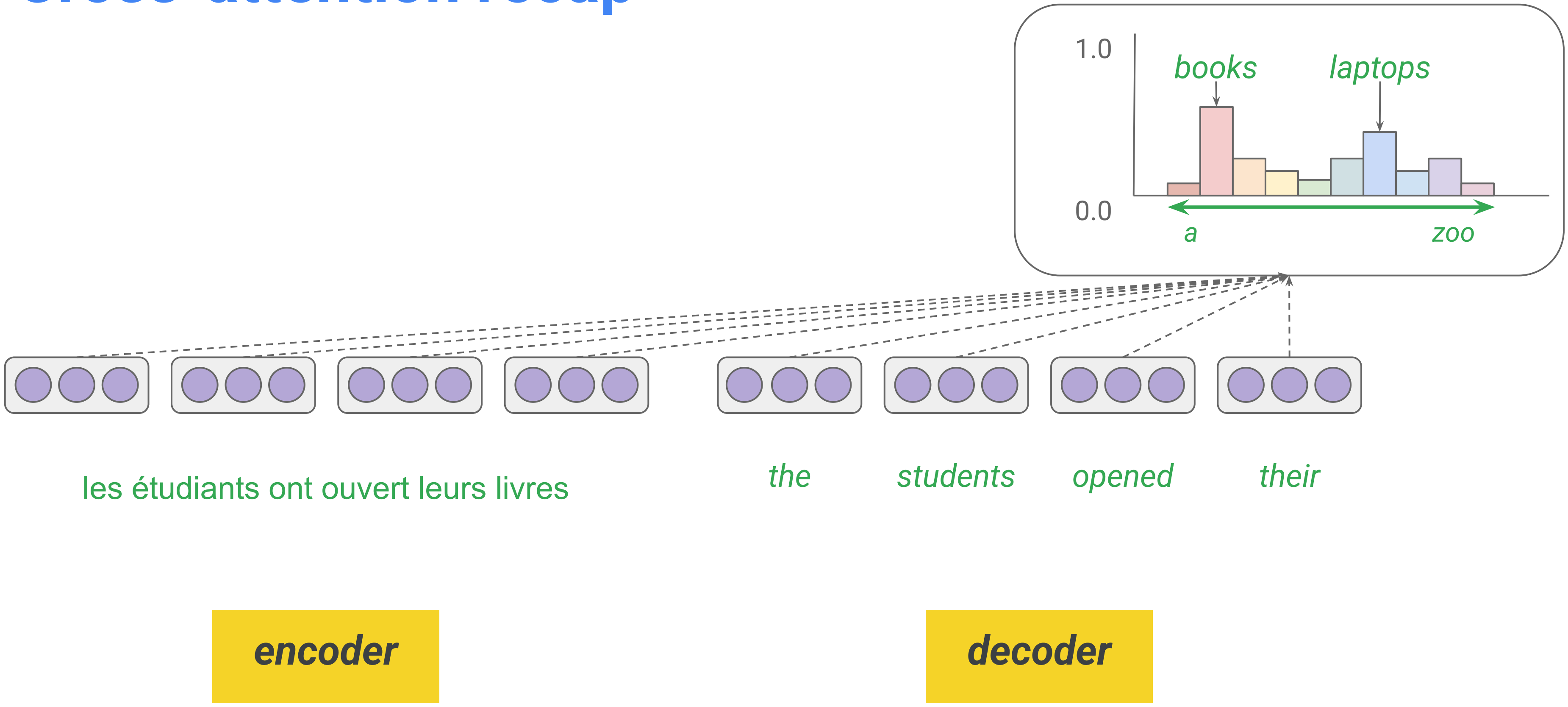
MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

Common approaches to building multimodal LLMs

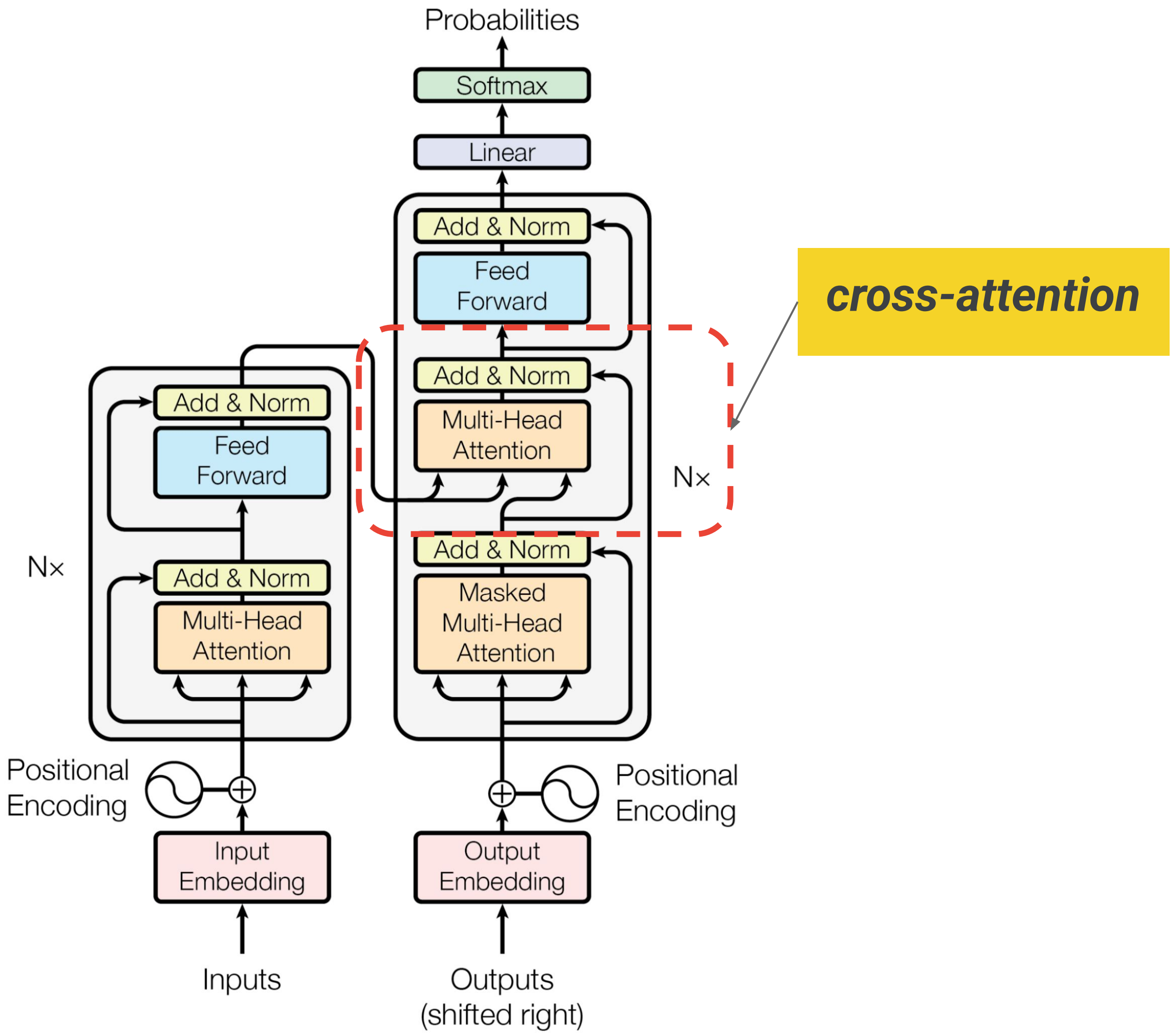


[Understanding Multimodal LLMs by Sebastian Raschka]

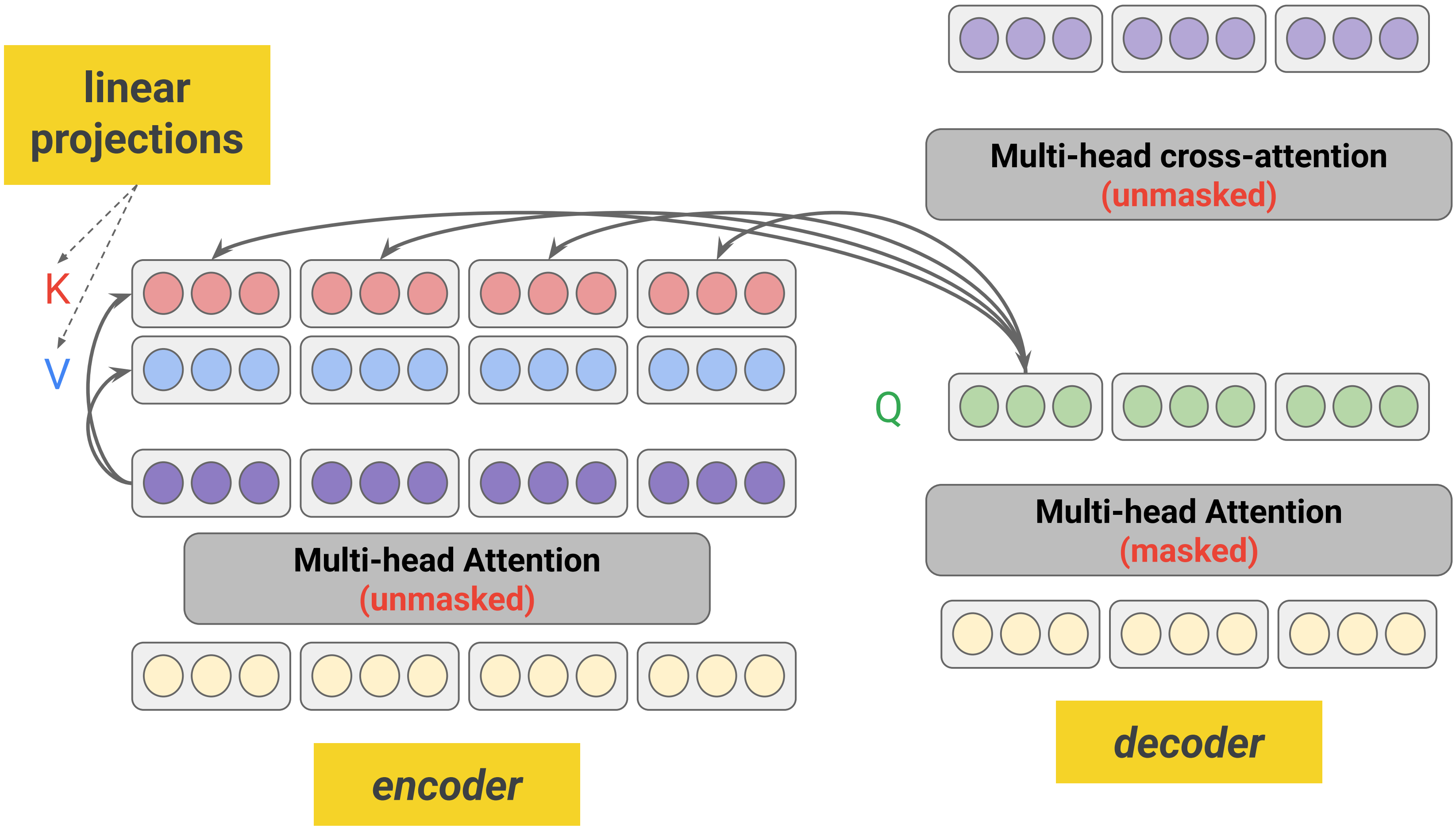
Cross-attention recap



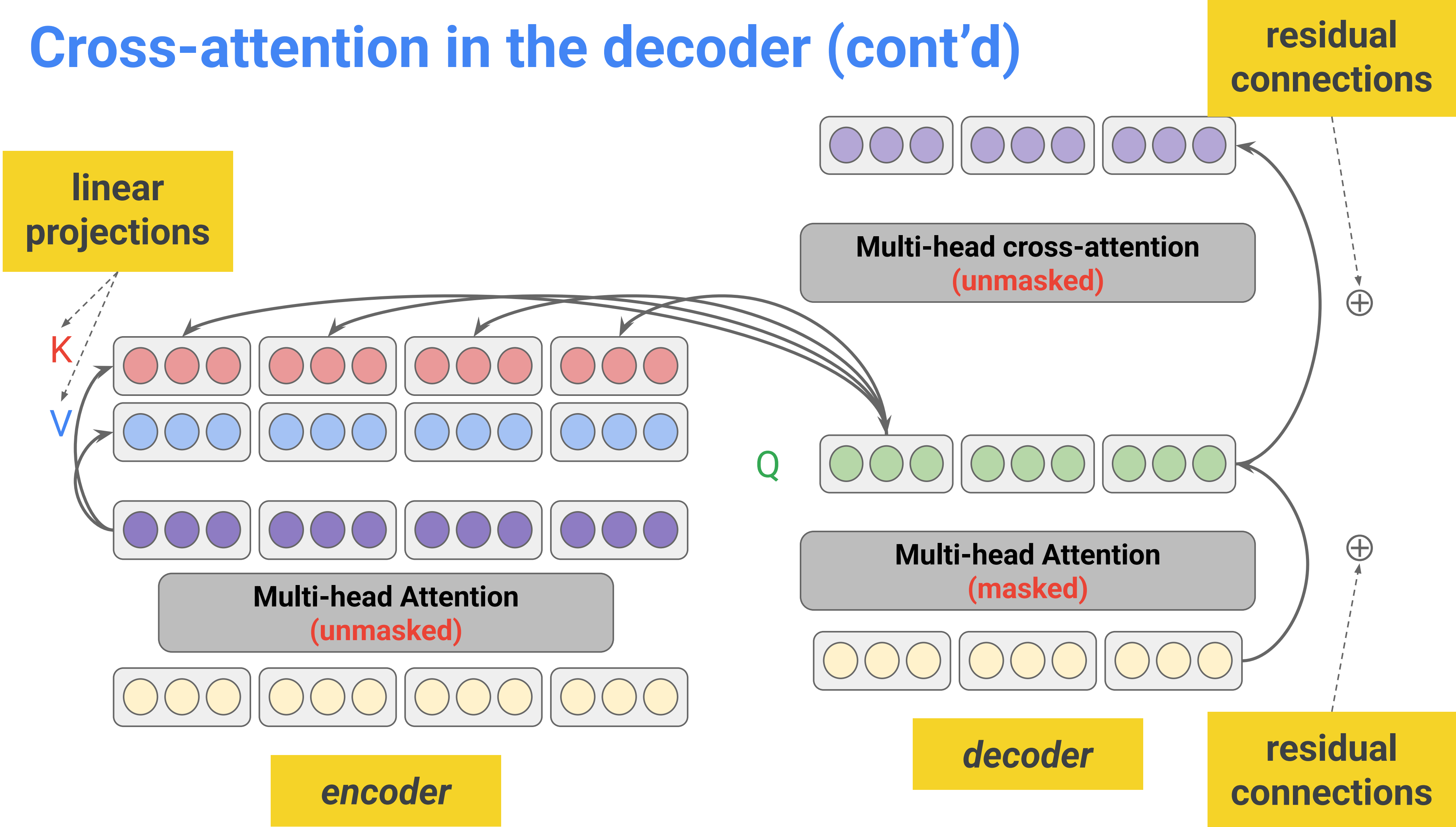
Cross-attention in the decoder



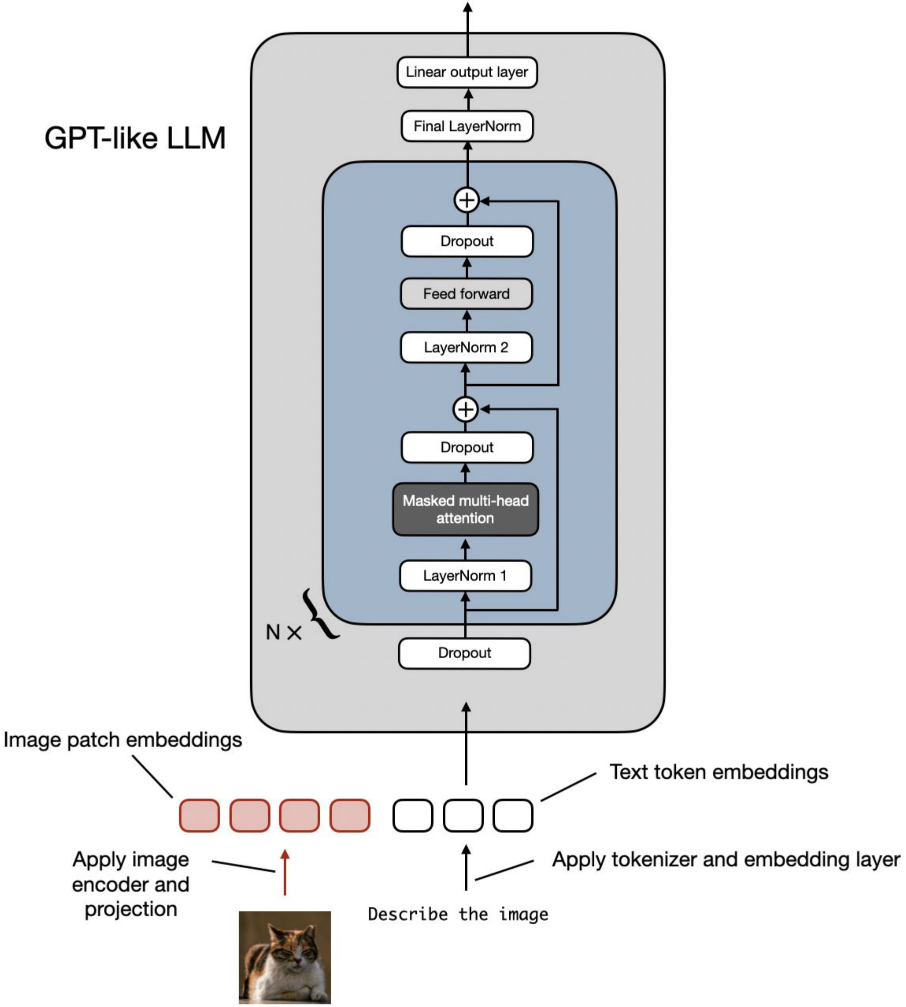
Cross-attention in the decoder



Cross-attention in the decoder (cont'd)

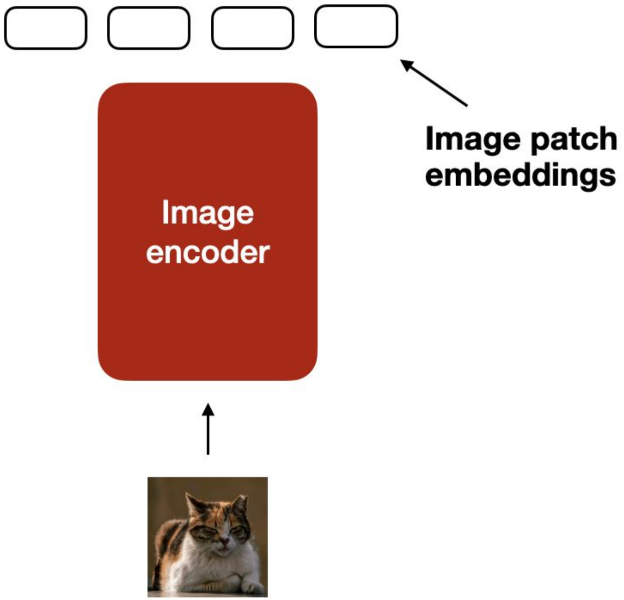


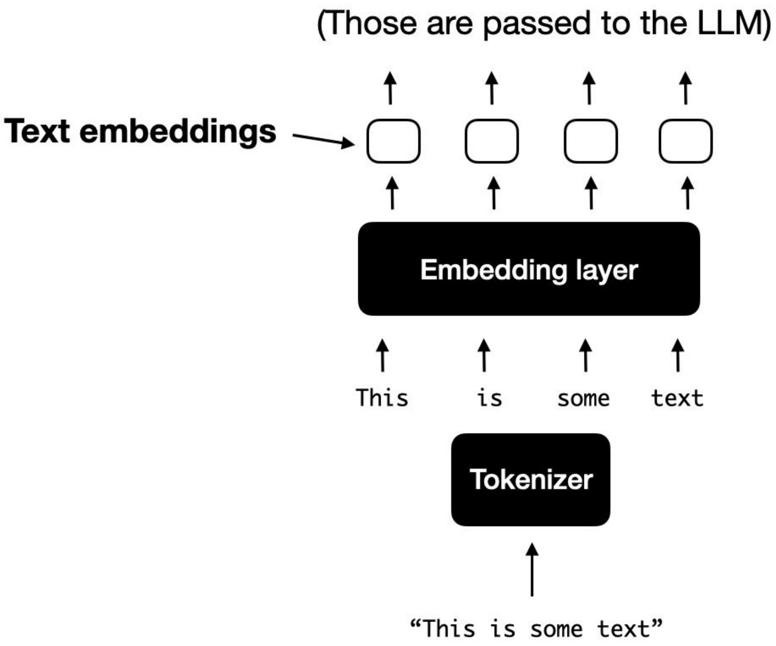
Method A: Unified embedding decoder architecture



[Understanding Multimodal LLMs by Sebastian Raschka]

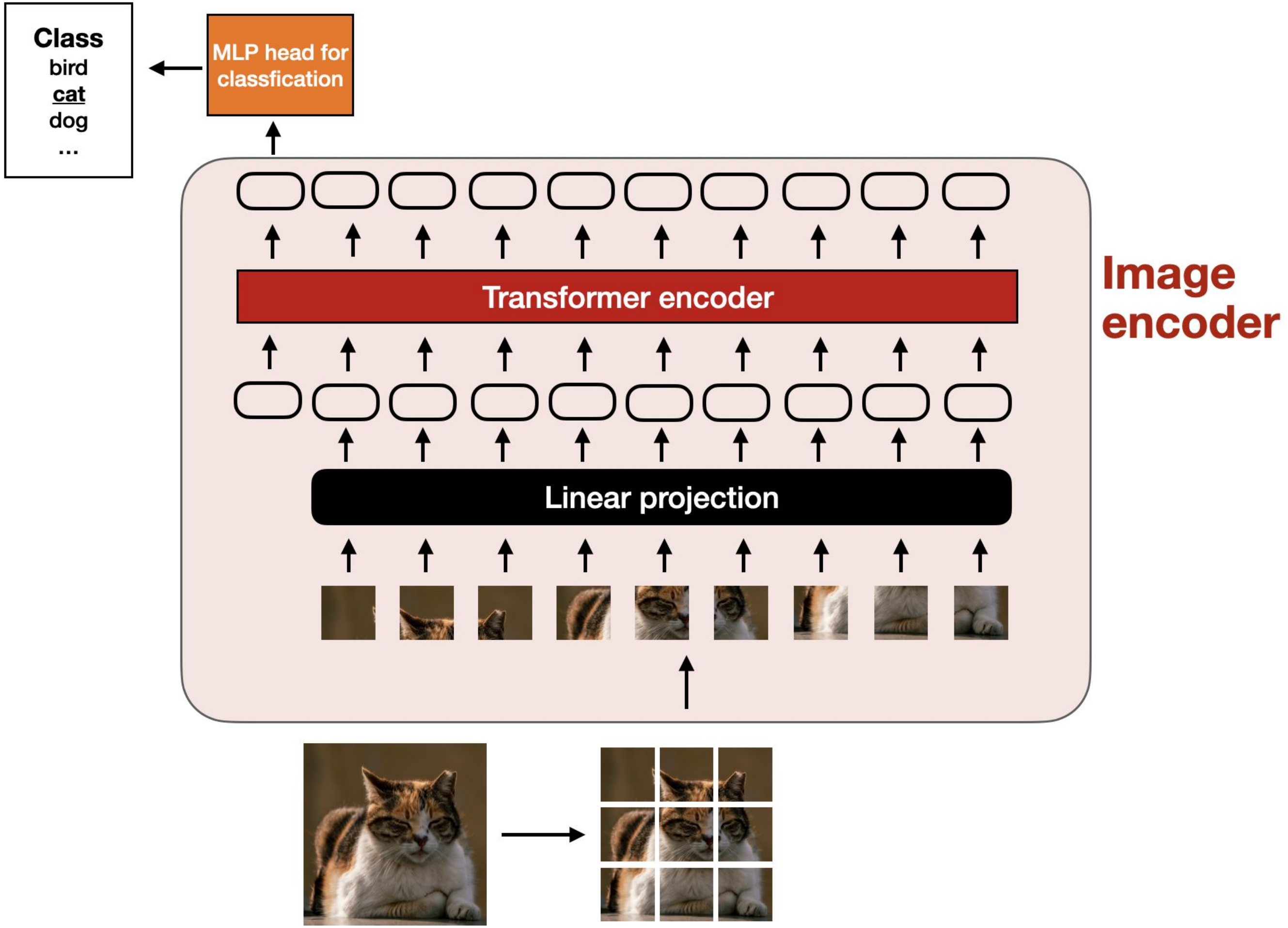
Understanding image encoders





[Understanding Multimodal LLMs by Sebastian Raschka]

Understanding image encoder (cont'd)



Understanding Multimodal LLMs by Sebastian Raschka

BERT recap

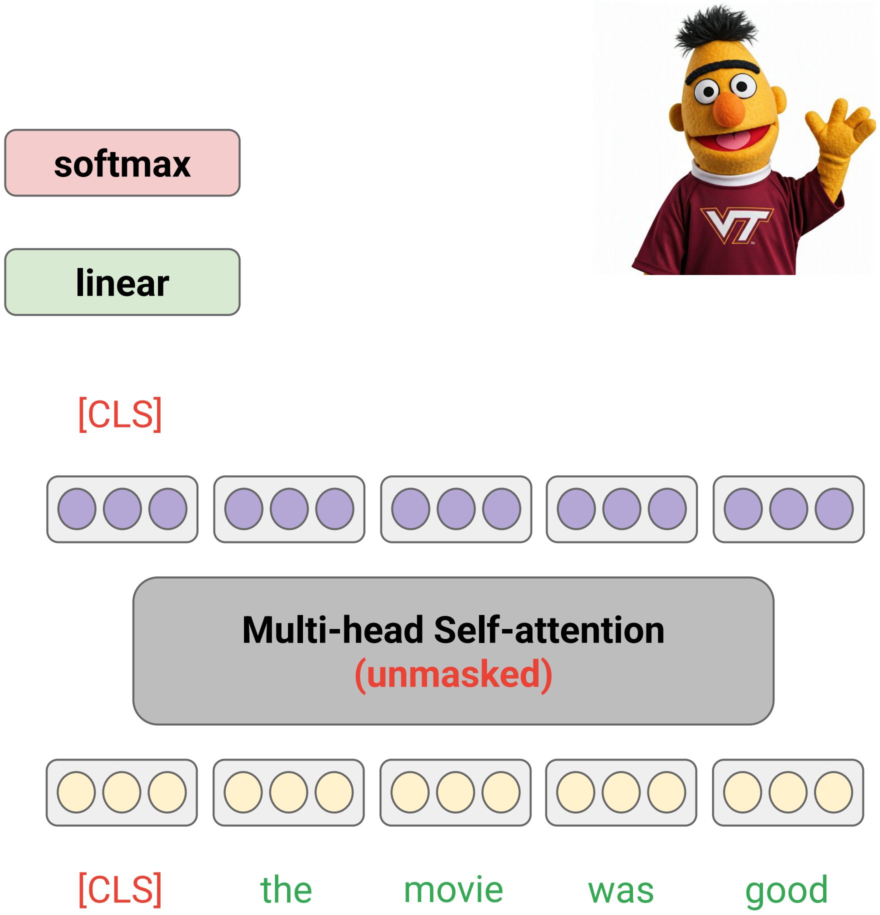
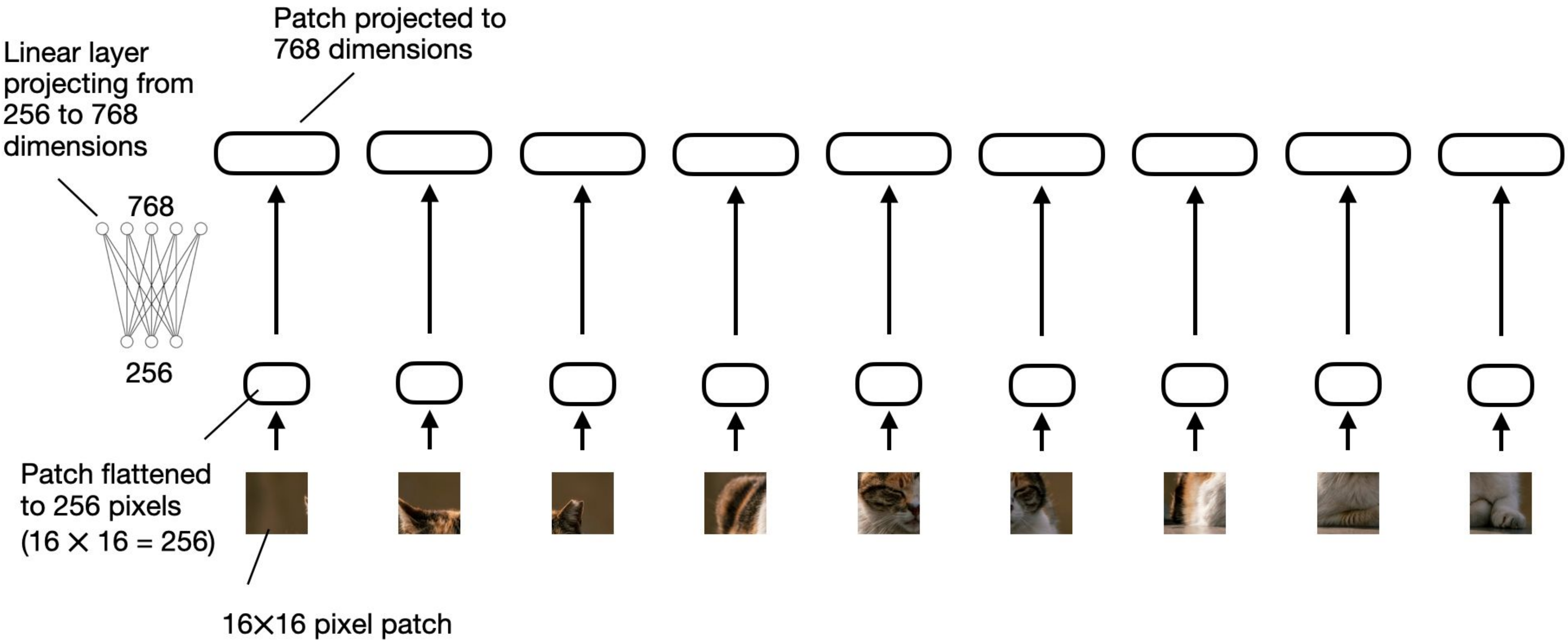


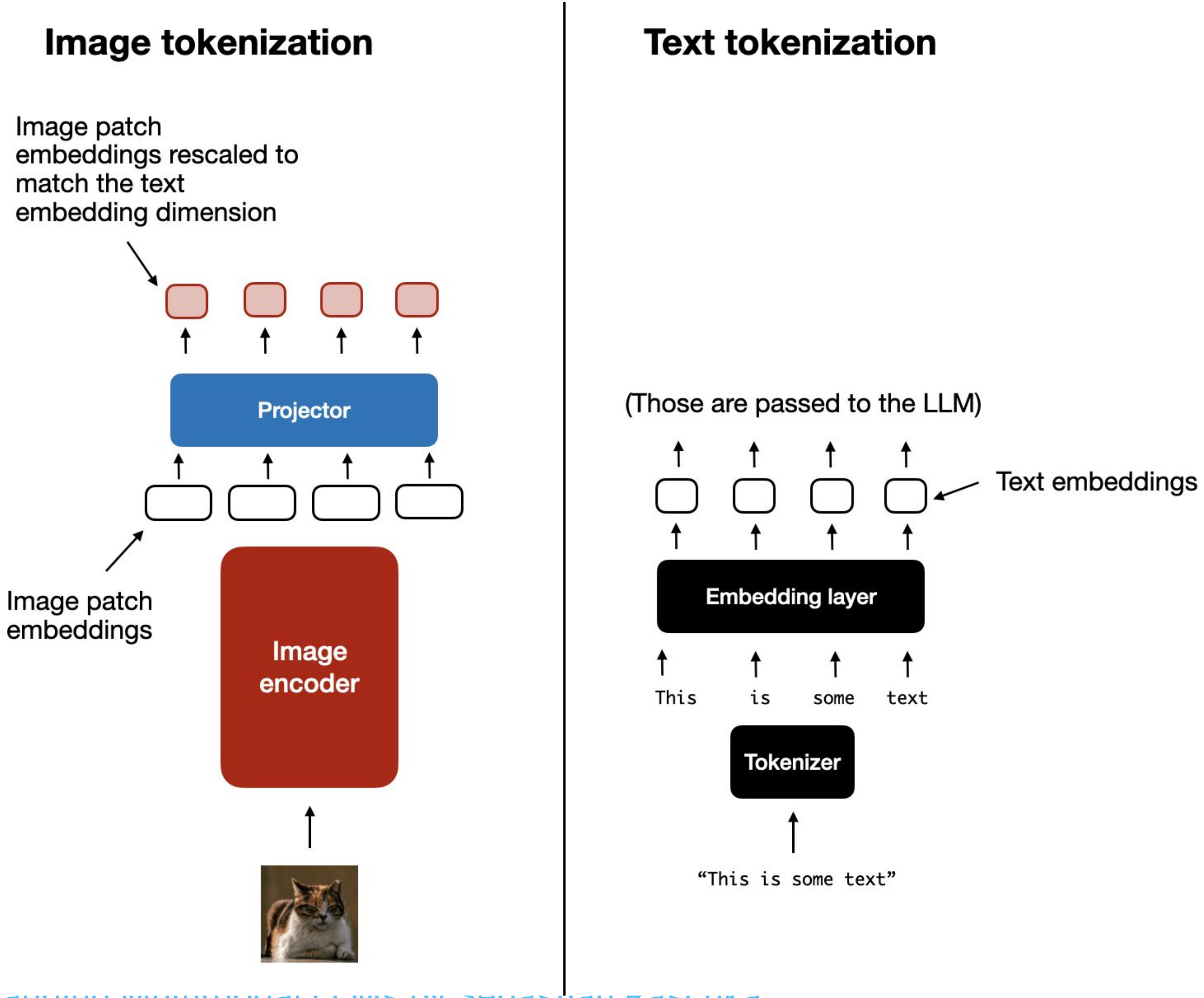
Image created by Gemini

The role of the linear projection module



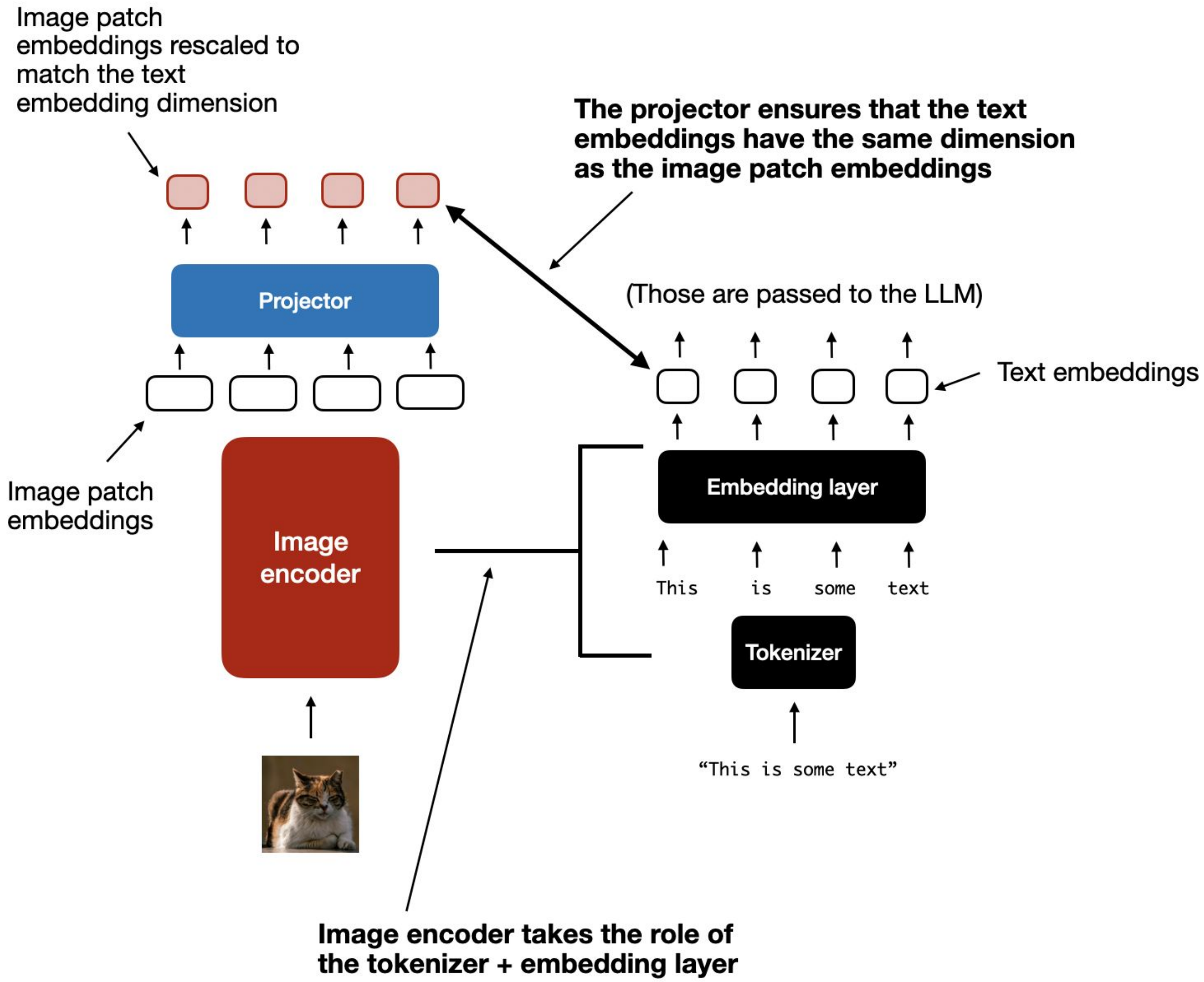
Understanding Multimodal LLMs by Sebastian Raschka

Text and image tokenization and embedding



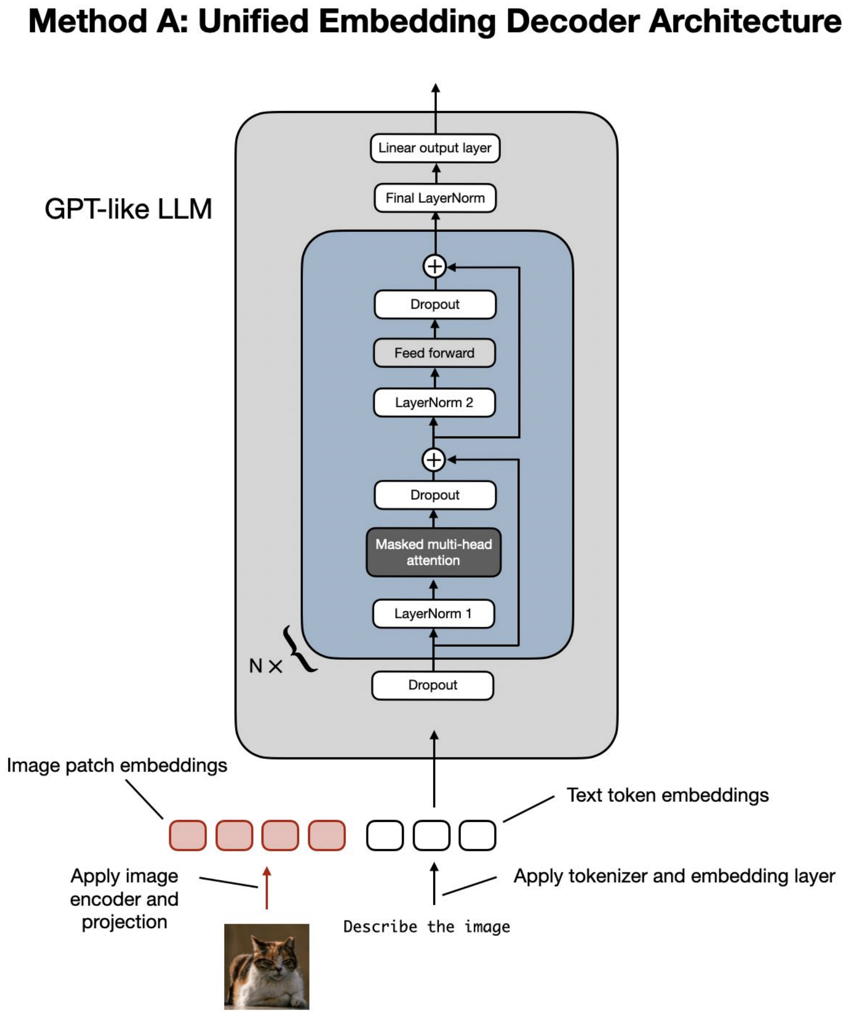
[Understanding Multimodal LLMs by Sebastian Raschka]

The role of the projector is to match the text token embedding dimensions



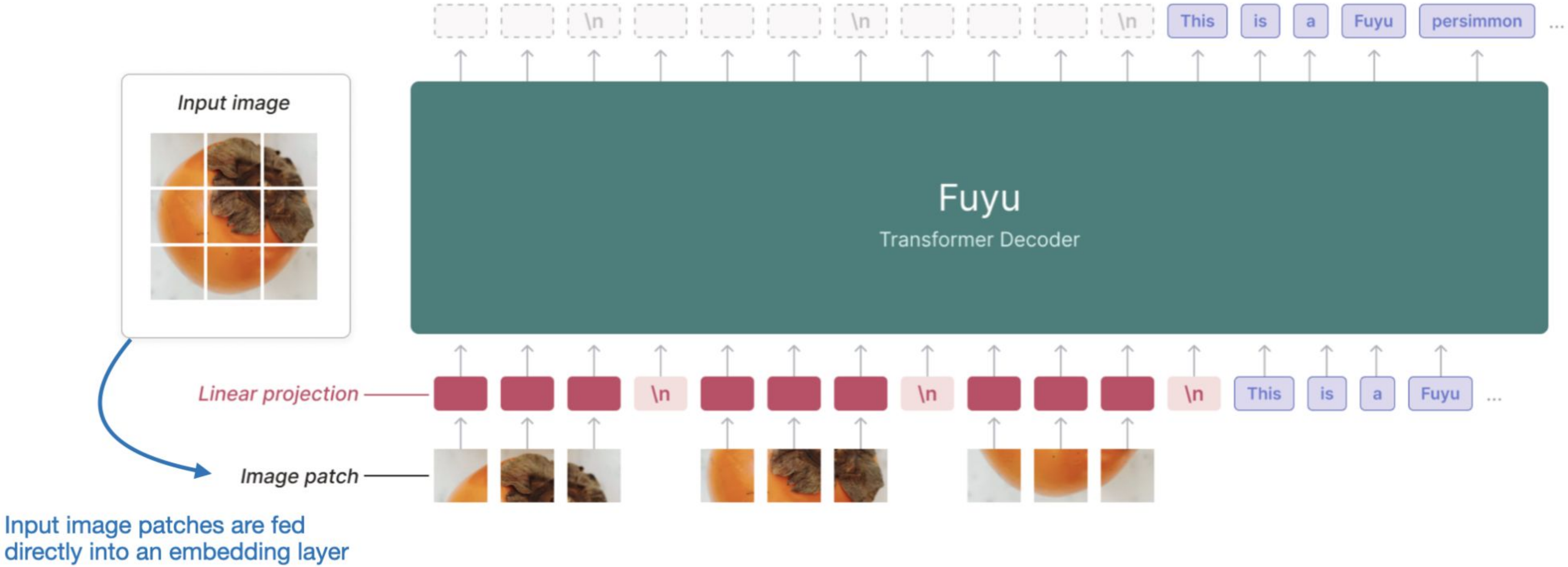
[Understanding Multimodal LLMs by Sebastian Raschka]

We can simply concatenate image and text embeddings



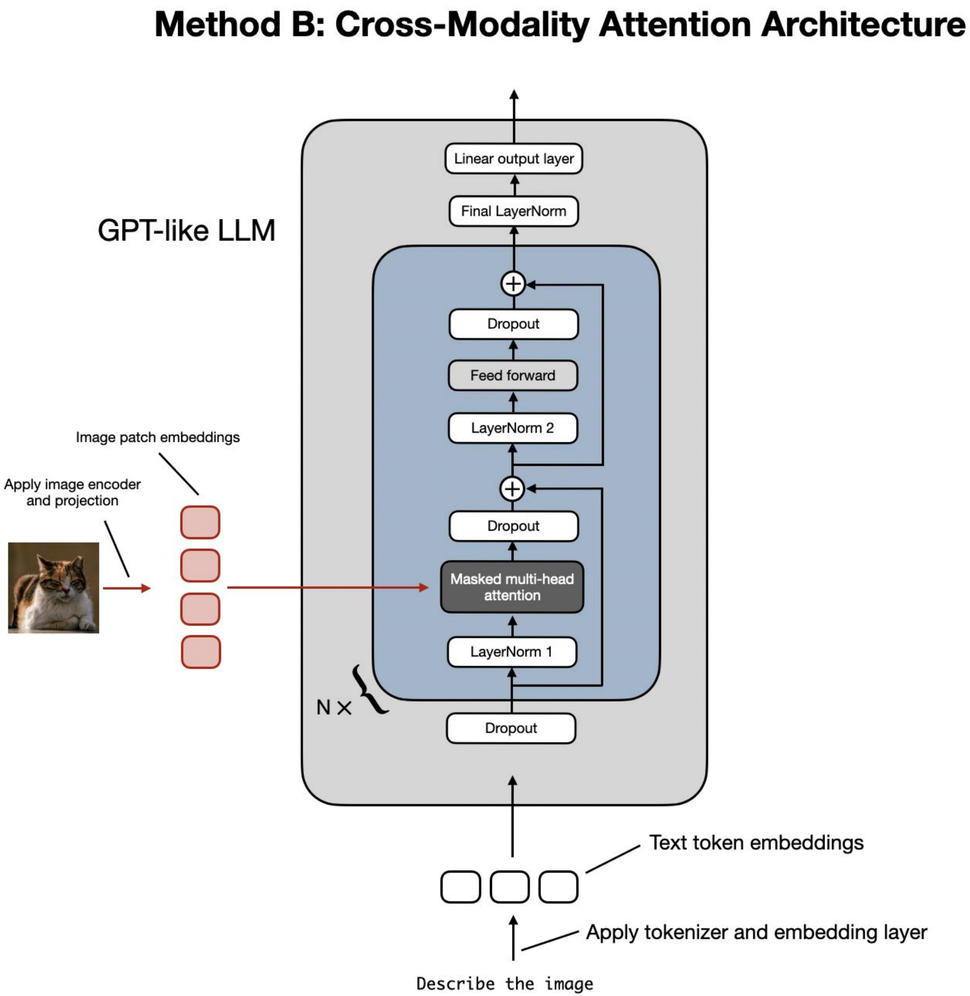
[Understanding Multimodal LLMs by Sebastian Raschka]

Versions of Method A that operate directly on patches



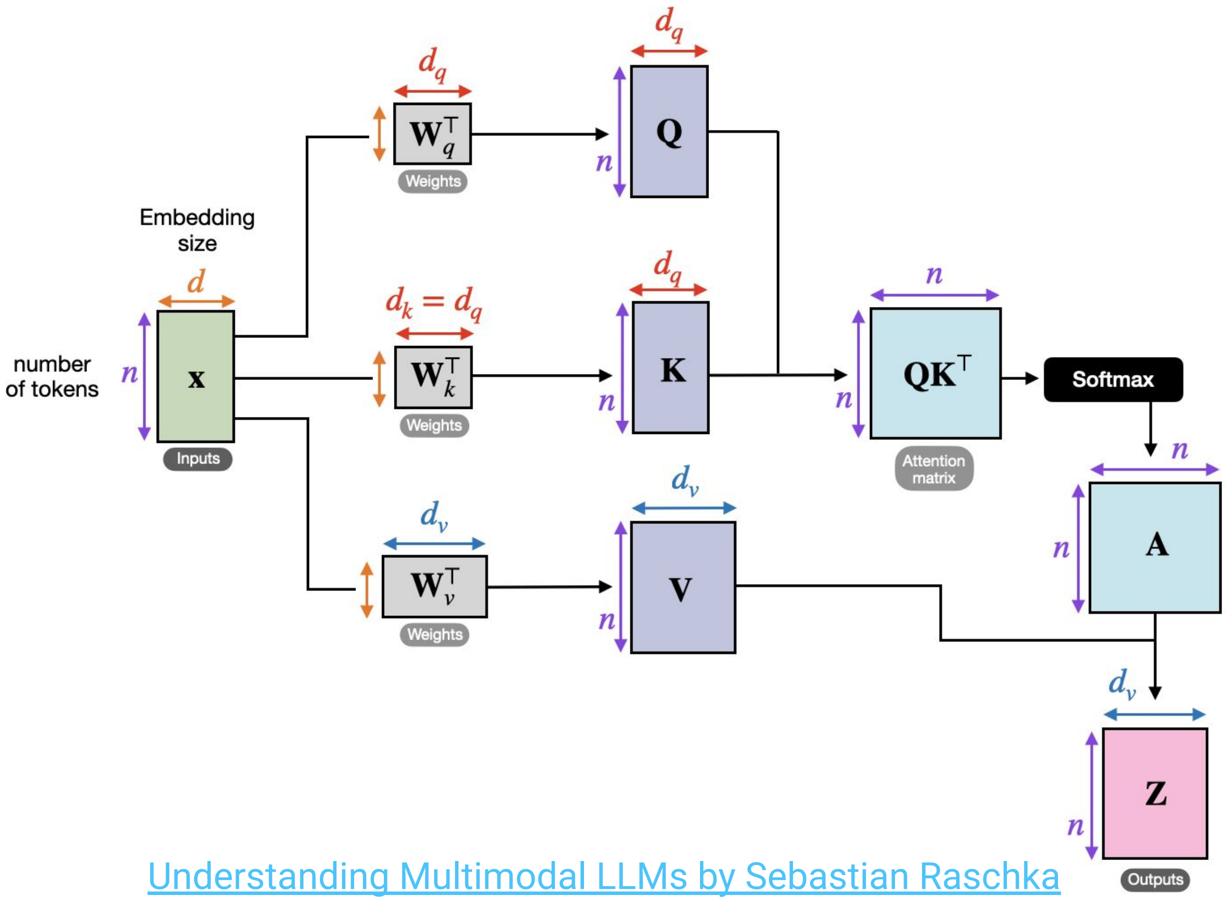
https://www.adept.ai/blog/fuyu-8b
[Understanding Multimodal LLMs by Sebastian Raschka]()

Method B: Cross-modality attention architecture



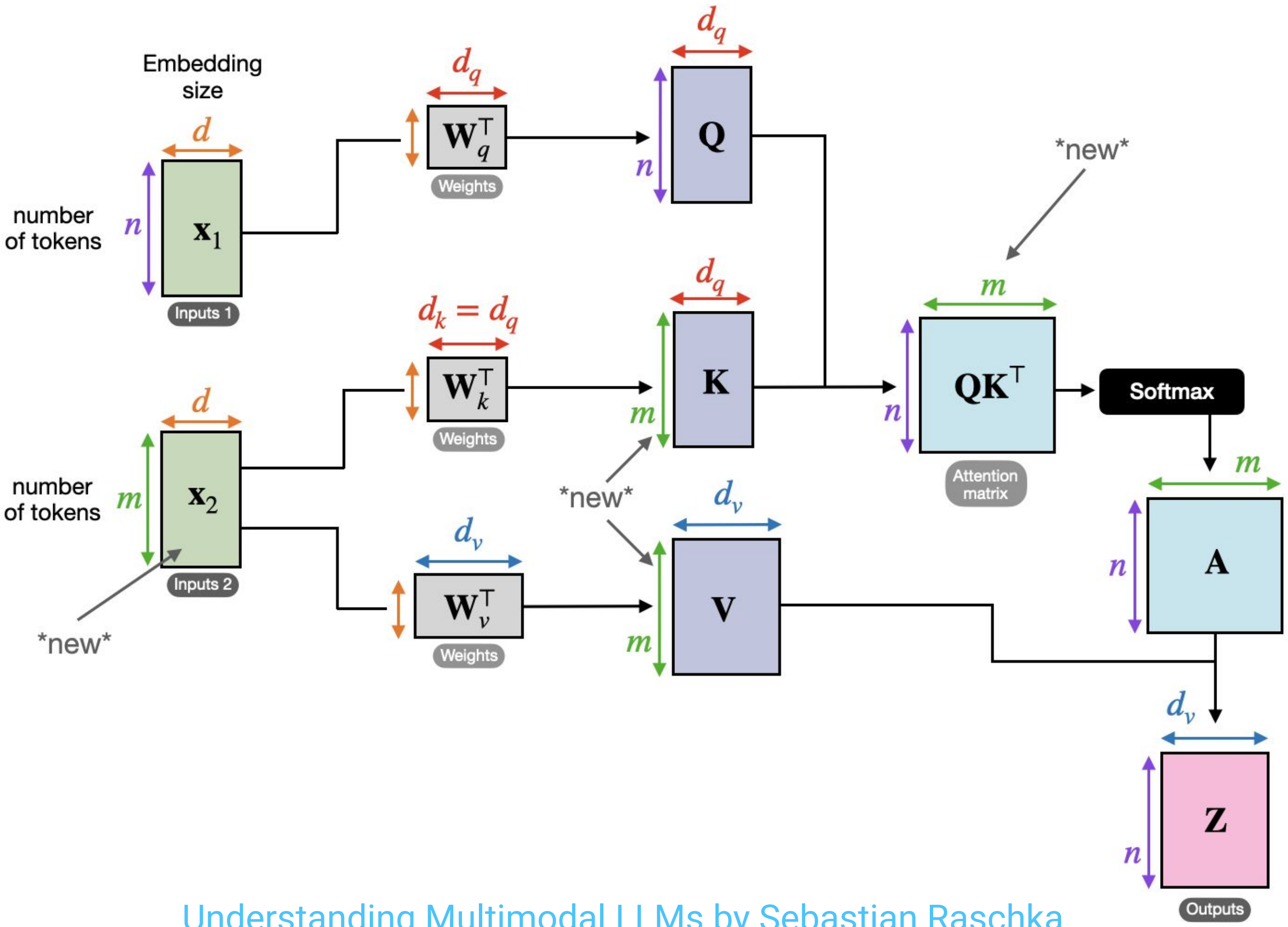
[Understanding Multimodal LLMs by Sebastian Raschka]

Regular self-attention



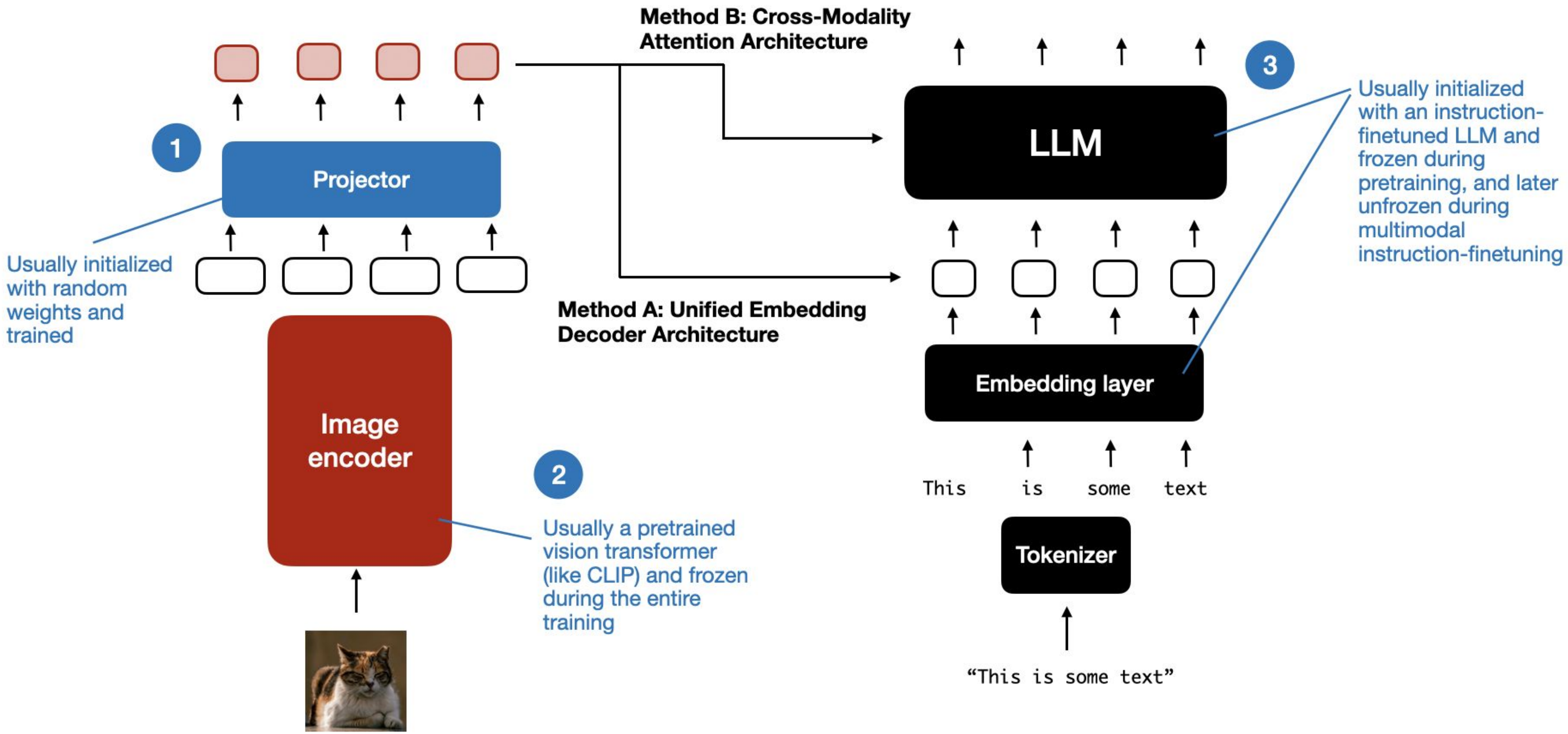
[Understanding Multimodal LLMs by Sebastian Raschka]

Cross-attention



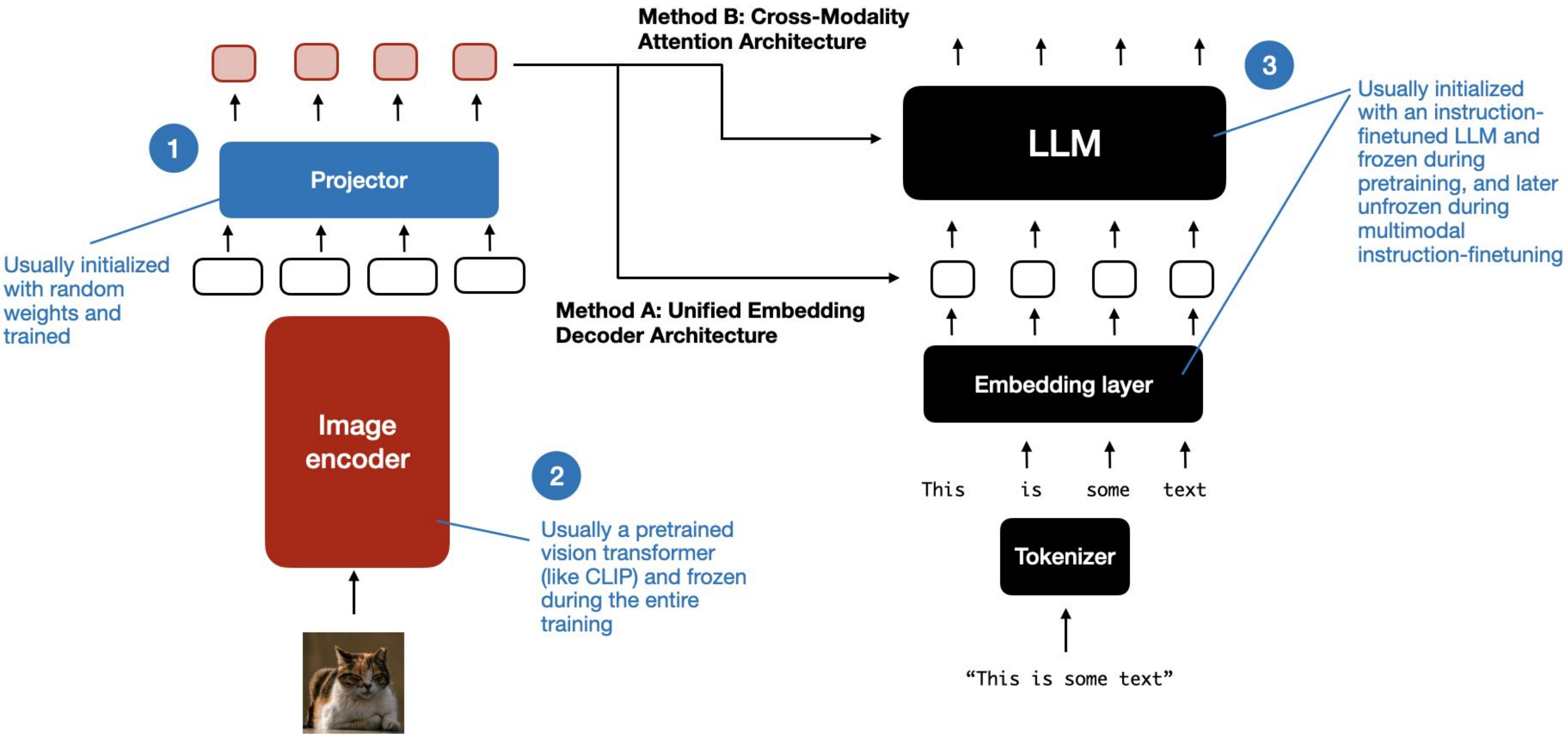
Understanding Multimodal LLMs by Sebastian Raschka

Unified decoder and cross-attention model training



Understanding Multimodal LLMs by Sebastian Raschka

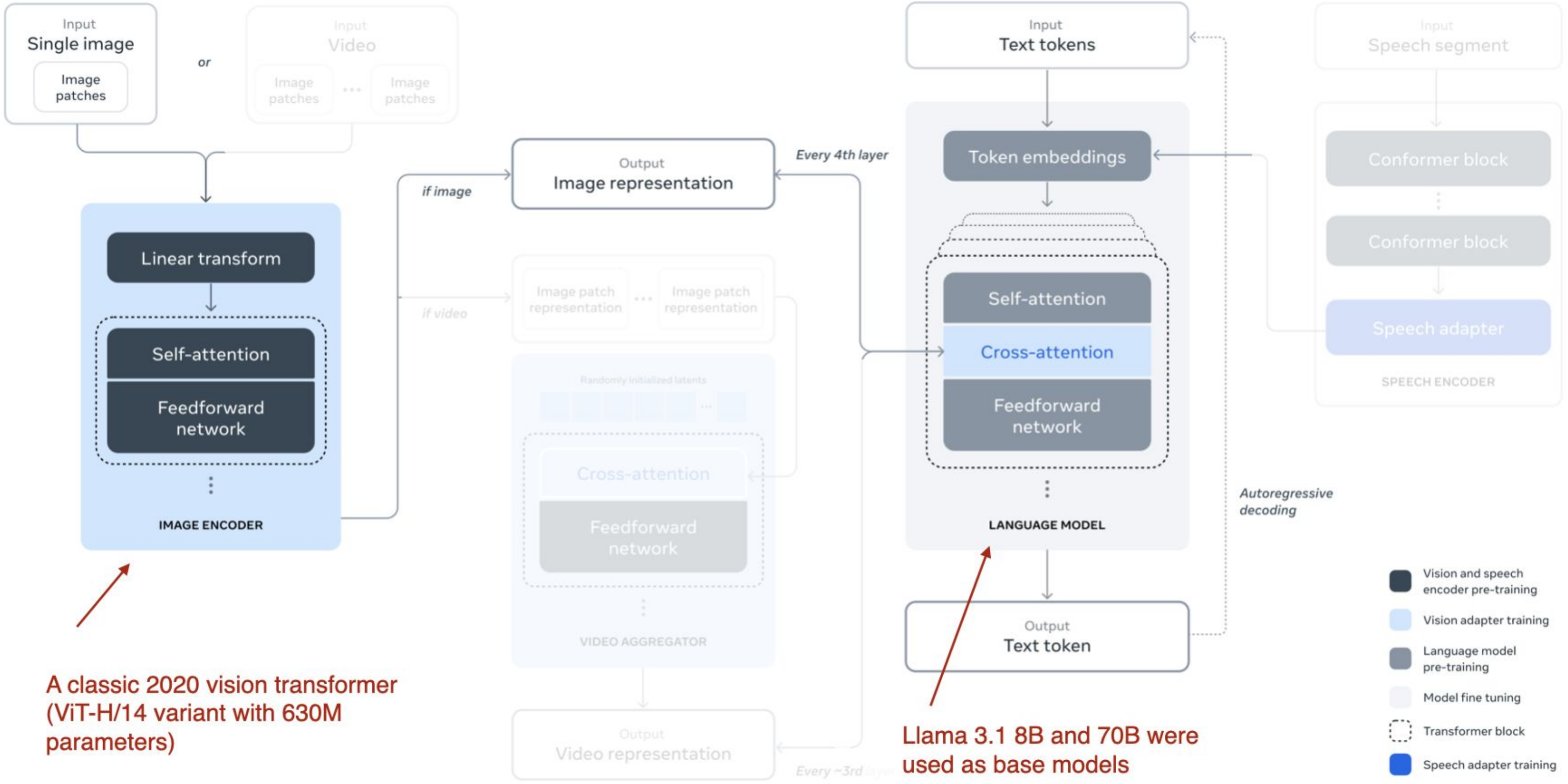
Unified decoder and cross-attention model training



Understanding Multimodal LLMs by Sebastian Raschka

Recent multimodal models and methods

The Llama 3 herd of models

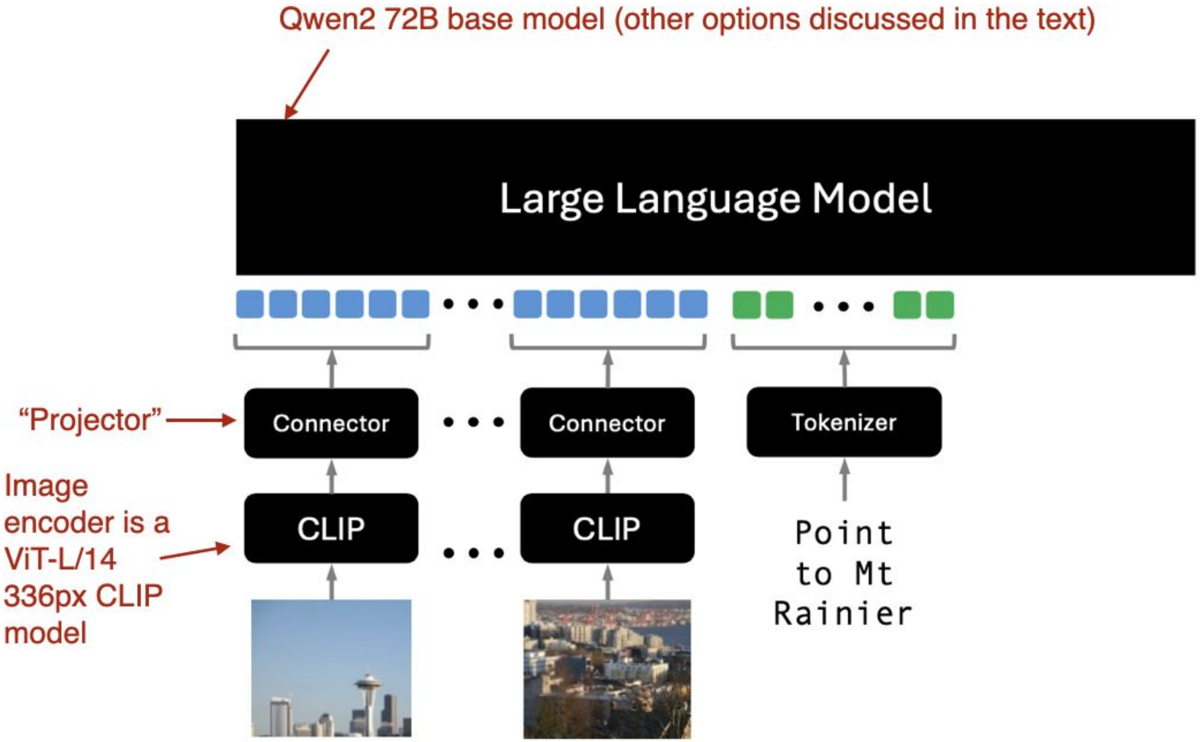


A classic 2020 vision transformer (ViT-H/14 variant with 630M parameters)

Llama 3.1 8B and 70B were used as base models

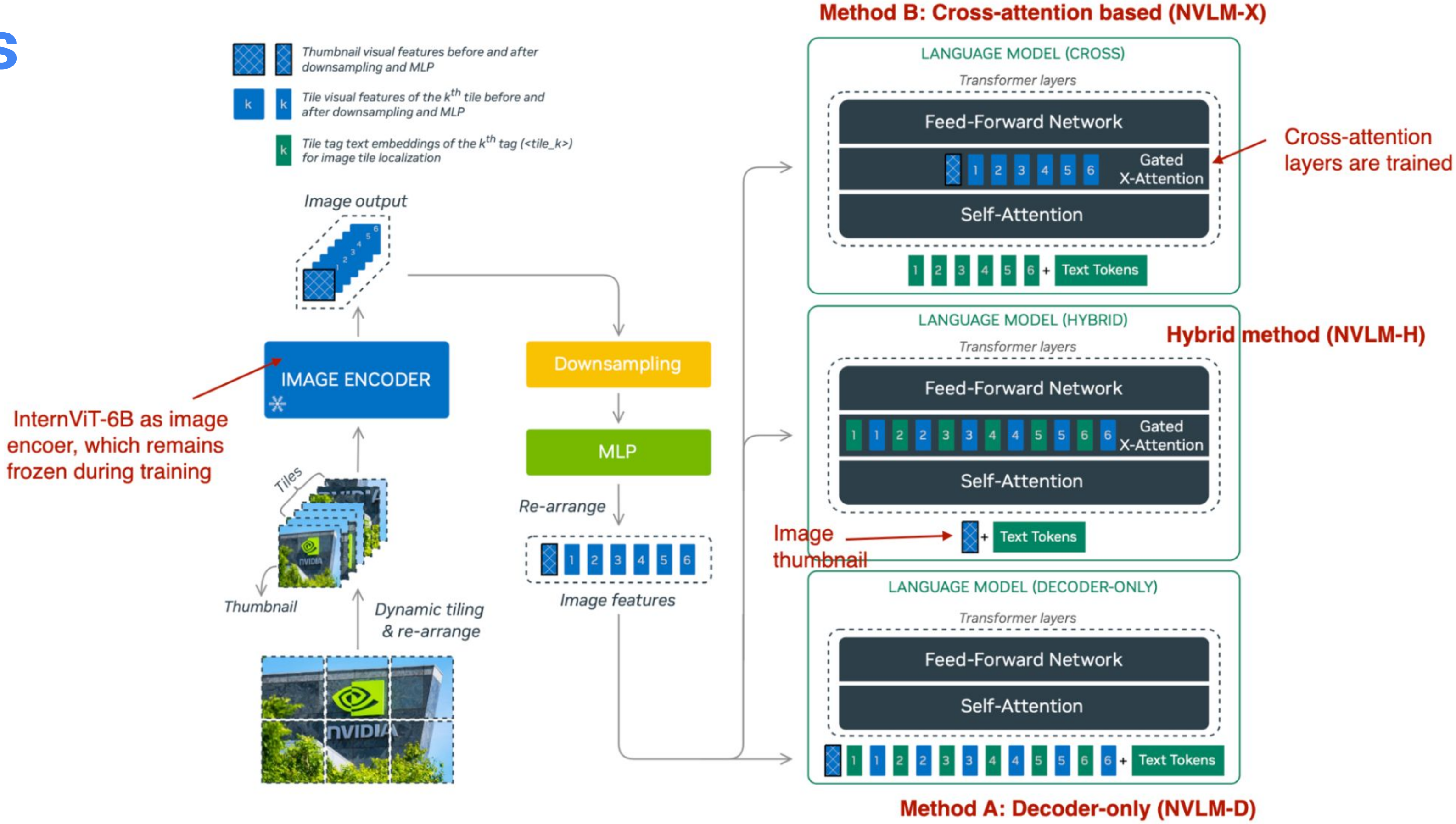
https://arxiv.org/abs/2407.21783
Understanding Multimodal LLMs by Sebastian Raschka

AI2's Molmo and PixMo



https://www.arxiv.org/abs/2409.17146
Understanding Multimodal LLMs by Sebastian Raschka

NVIDIA's NVLM



https://arxiv.org/abs/2409.11402

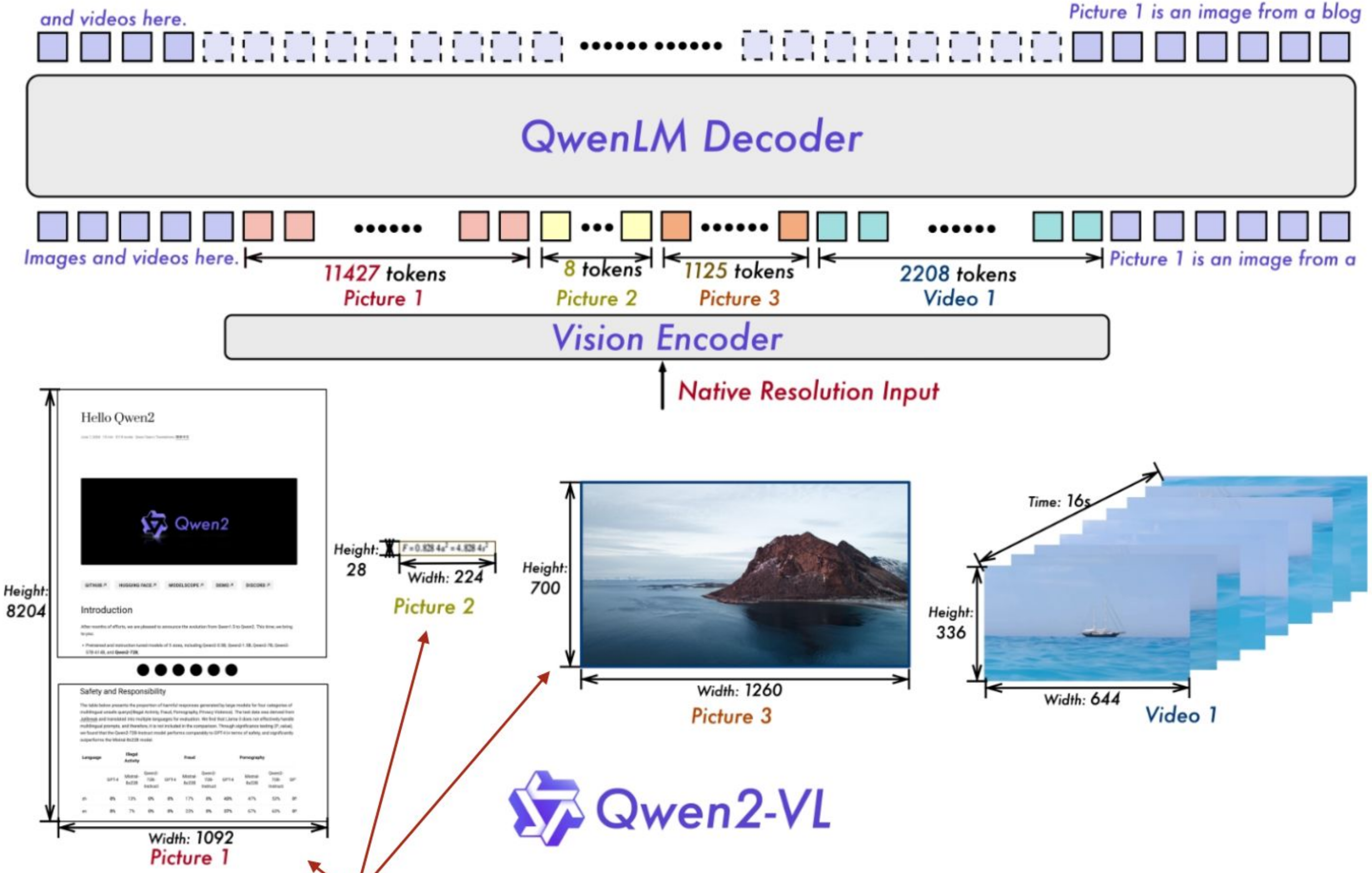
[Understanding Multimodal LLMs by Sebastian Raschka]()

NVIDIA's NVLM

- **NVLM-X** (cross-attention)demonstrates superior computational efficiency for high-resolution images.
- **NVLM-D** (unified embedding) achieves higher accuracy in OCR-related tasks.
- **NVLM-H** combines the advantages of both methods.

https://arxiv.org/abs/2409.11402
Understanding Multimodal LLMs by Sebastian Raschka

Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution

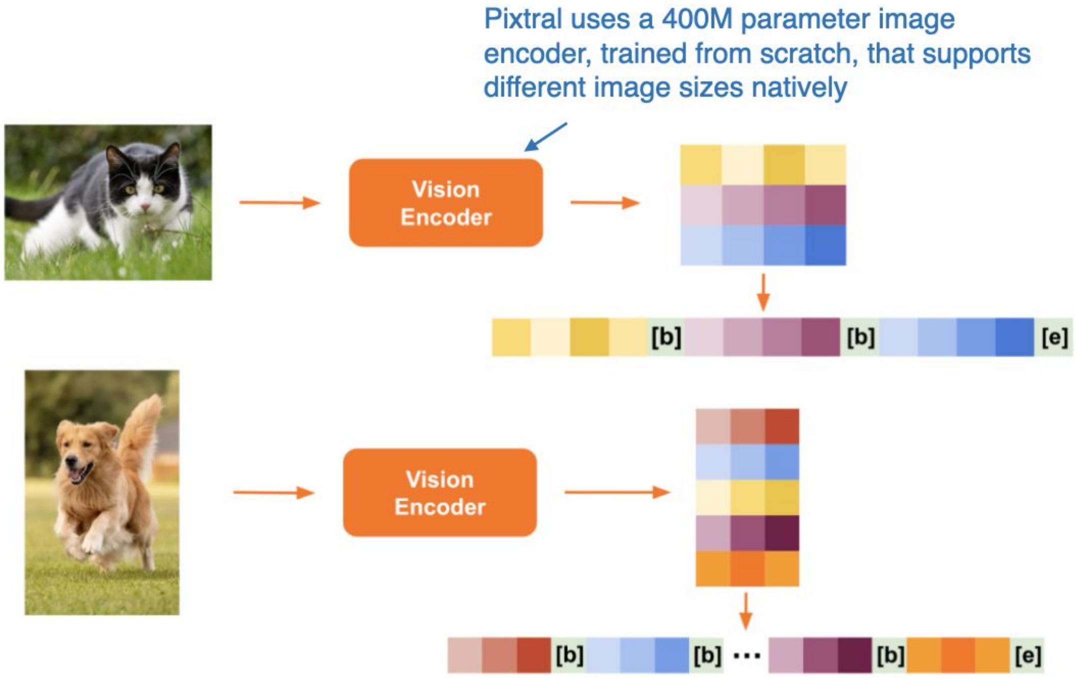


Input images are not resized, so they enter the LLM with a different number of tokens

https://arxiv.org/abs/2409.12191

[Understanding Multimodal LLMs by Sebastian Raschka]()

Pixtral 12B



https://mistral.ai/news/pixtral-12b
Understanding Multimodal LLMs by Sebastian Raschka

MM1: Methods, analysis & insights

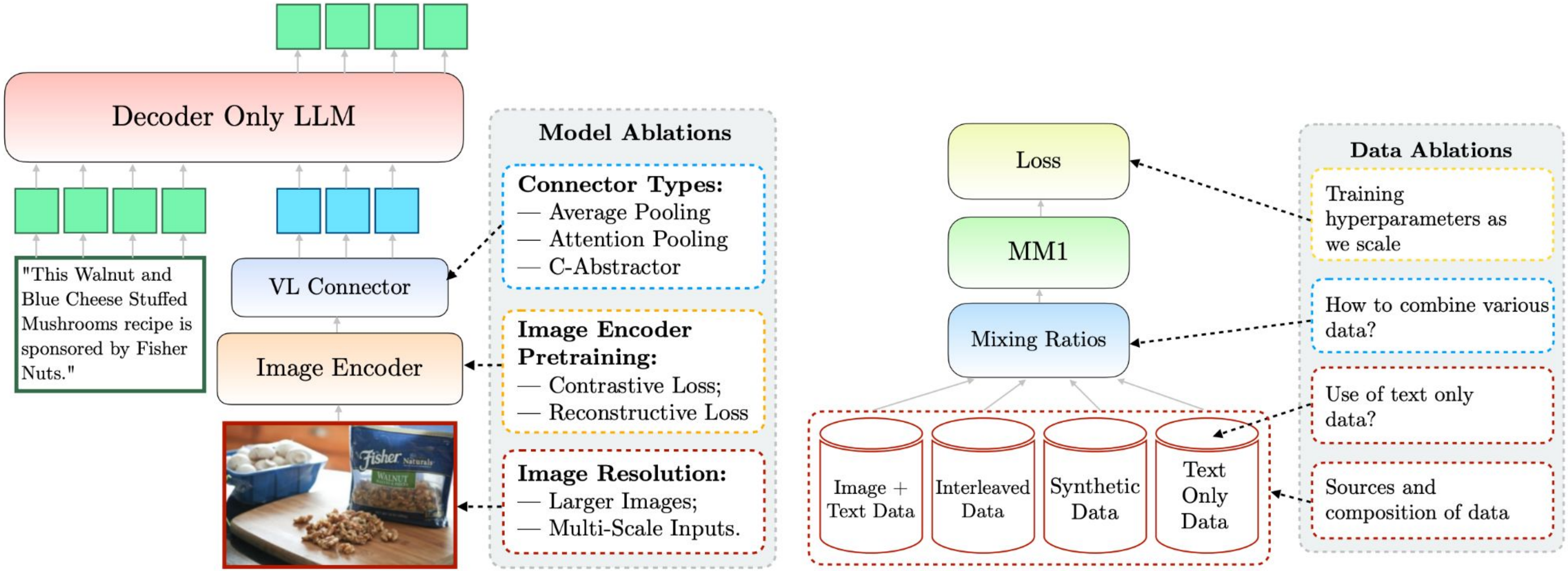
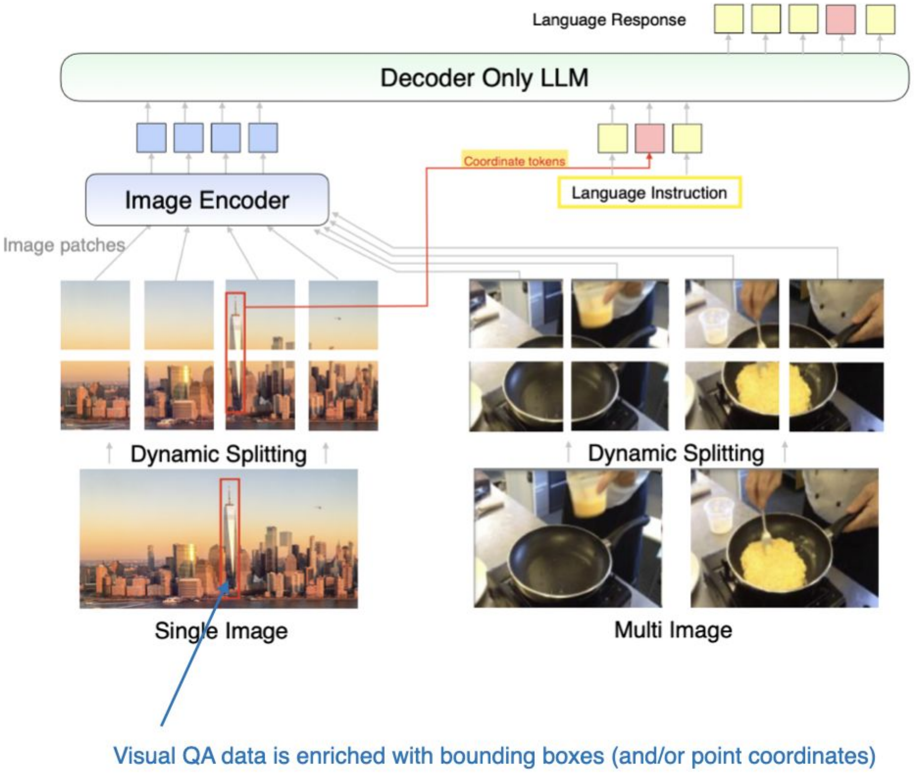


Fig. 3: *Left:* Model ablations: what visual encoder to use, how to feed rich visual data, and how to connect the visual representation to the LLM. *Right:* Data ablations: type of data, and their mixture.

MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training

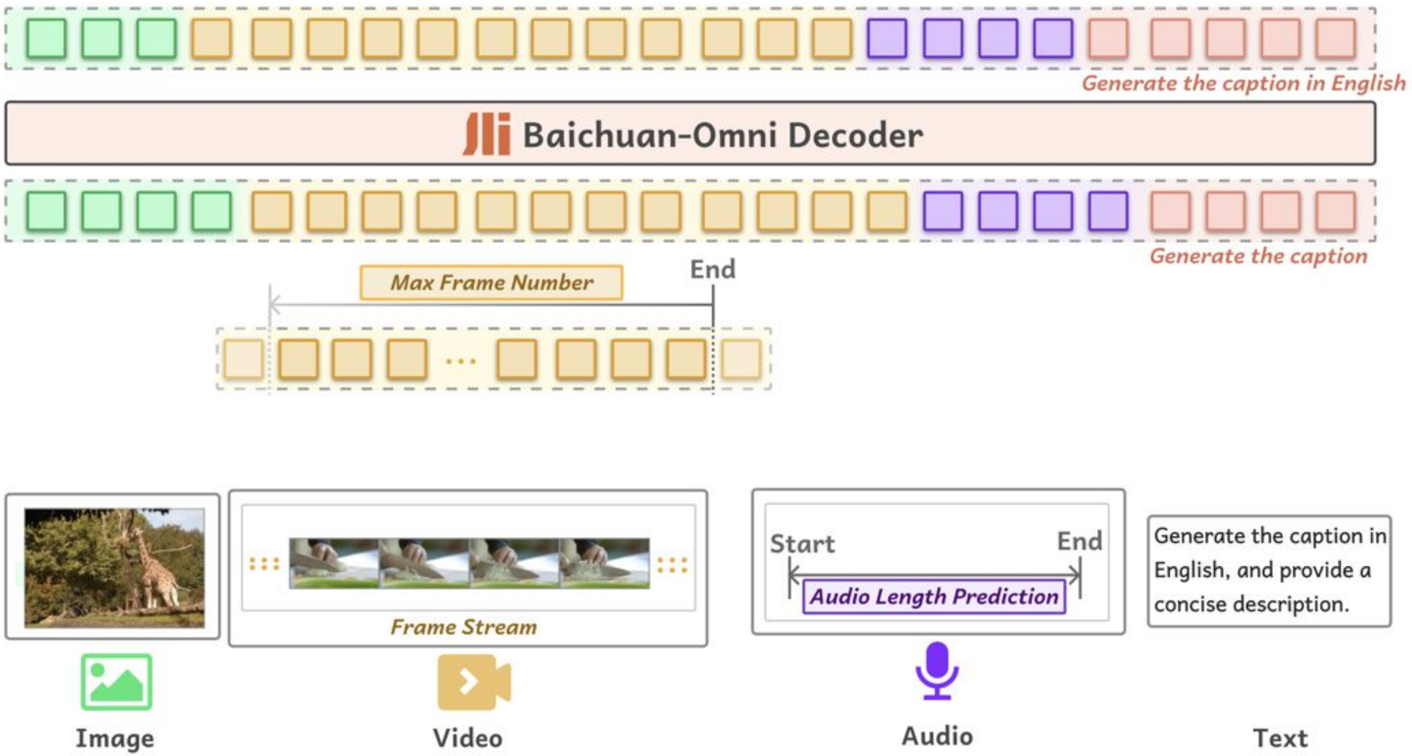
MM1.5: Methods, analysis & insights



https://arxiv.org/abs/2409.20566
[Understanding Multimodal LLMs by Sebastian Raschka]()

Baichuan-Omni

Baichuan-Omni uses the Unified Embedding Decoder Architecture setup



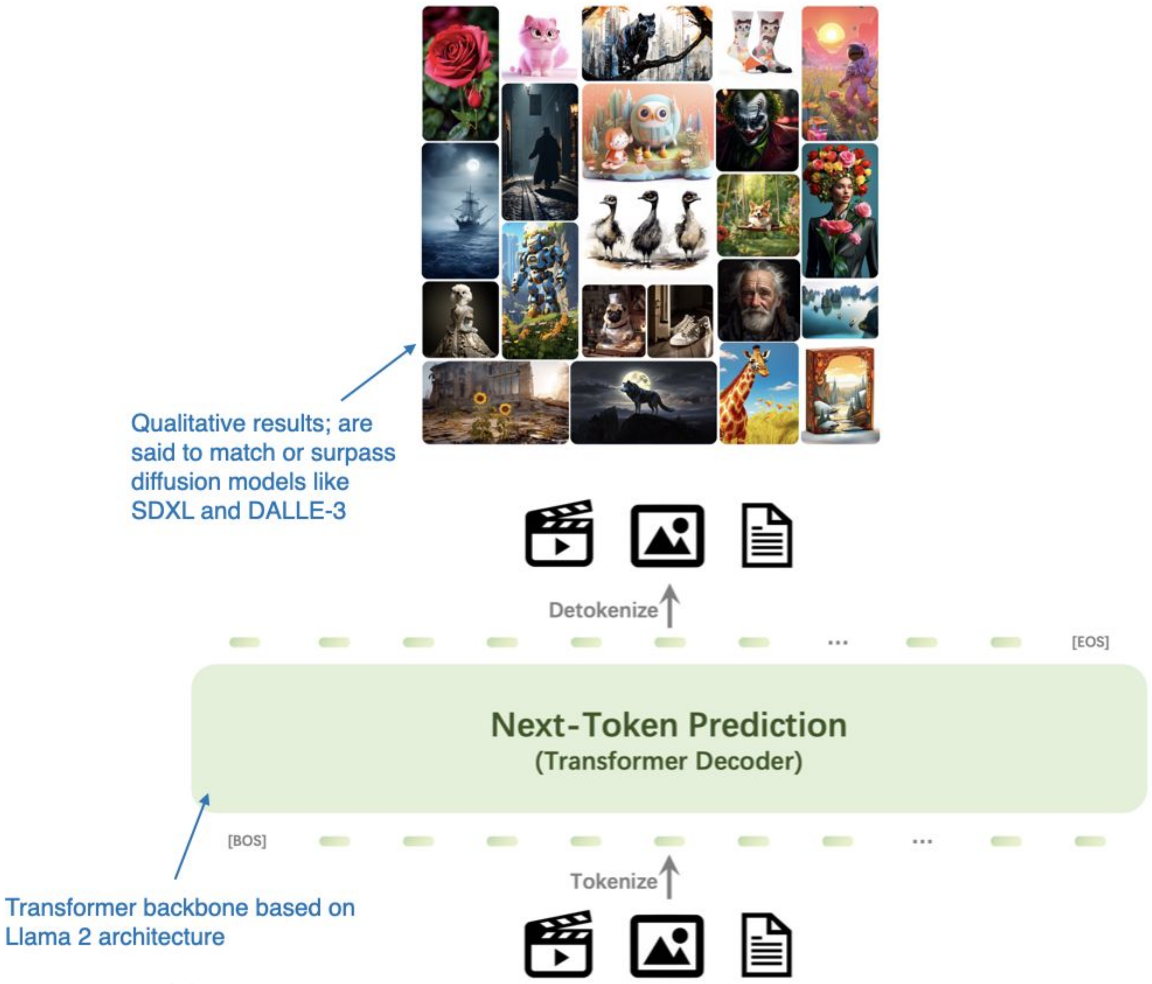
https://arxiv.org/abs/2410.08565
Understanding Multimodal LLMs by Sebastian Raschka

Baichuan-Omni (cont'd)

- **Projector training:** Initially, only the projector is trained, while both the vision encoder and the language model (LLM) remain frozen.
- **Vision encoder training:** Next, the vision encoder is unfrozen and trained, with the LLM still frozen.
- **Full model training:** Finally, the LLM is unfrozen, allowing the entire model to be trained end-to-end.

https://arxiv.org/abs/2410.08565
Understanding Multimodal LLMs by Sebastian Raschka

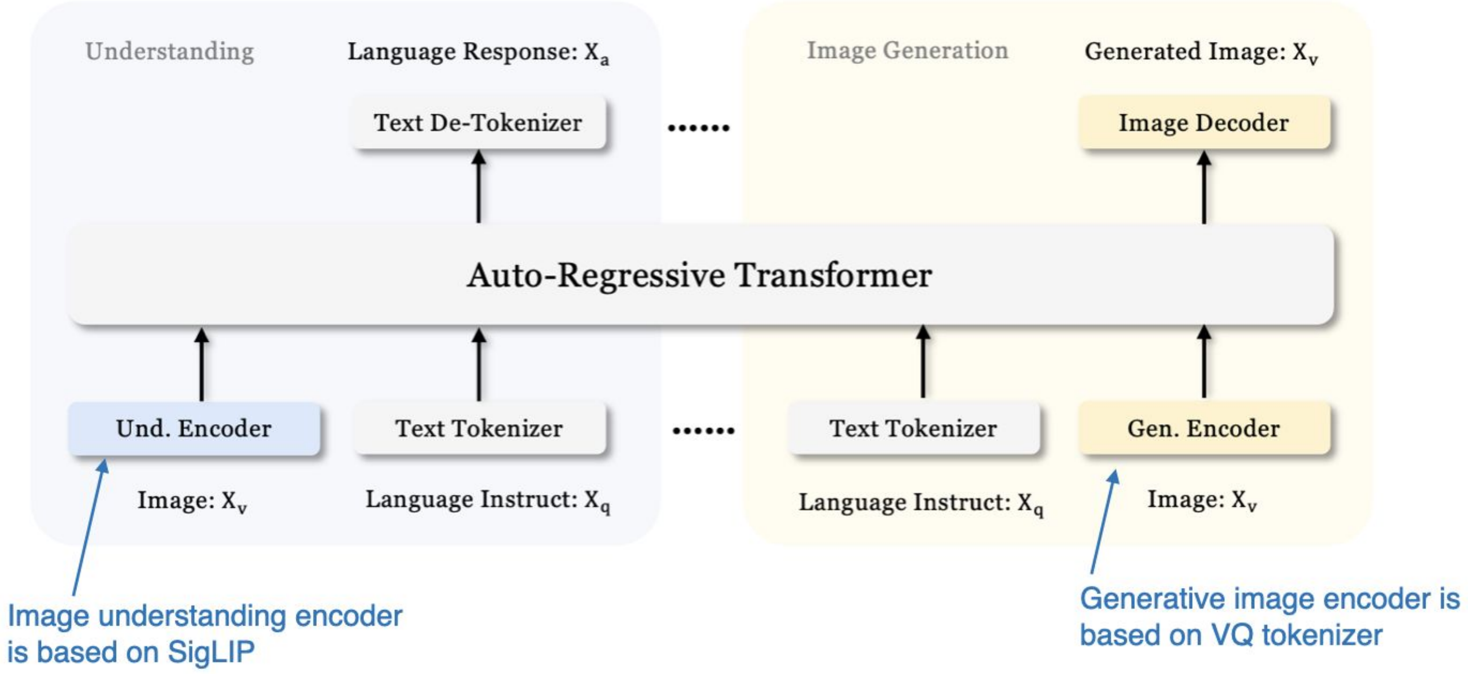
Emu3: Next-token prediction is all you need



https://arxiv.org/abs/2409.18869

Understanding Multimodal LLMs by Sebastian Raschka

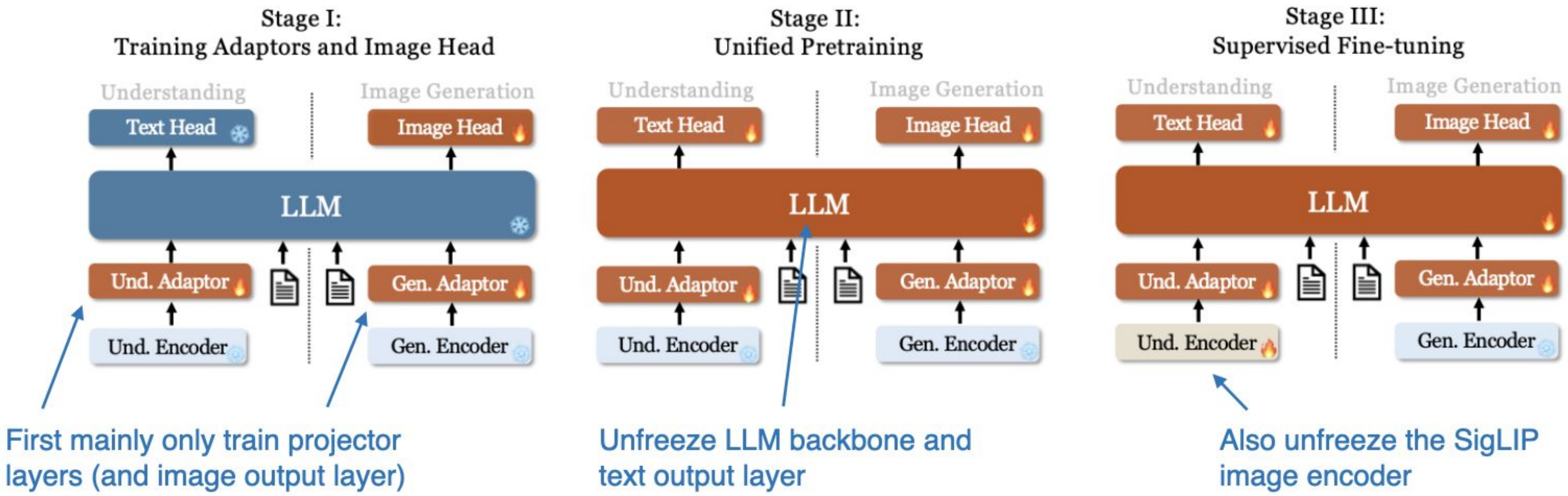
Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation



https://arxiv.org/abs/2410.13848

Understanding Multimodal LLMs by Sebastian Raschka

Janus: Decoupling Visual Encoding for Unified Multimodal Understanding and Generation



https://arxiv.org/abs/2410.13848
Understanding Multimodal LLMs by Sebastian Raschka

Thank you!