Bridging Vision and Language: Overview of Multimodal Large Language Models

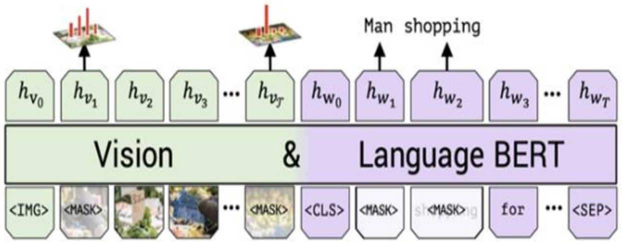
Adel Ahmadyan

Meta

Vision Language Model Overview

Earlier ...

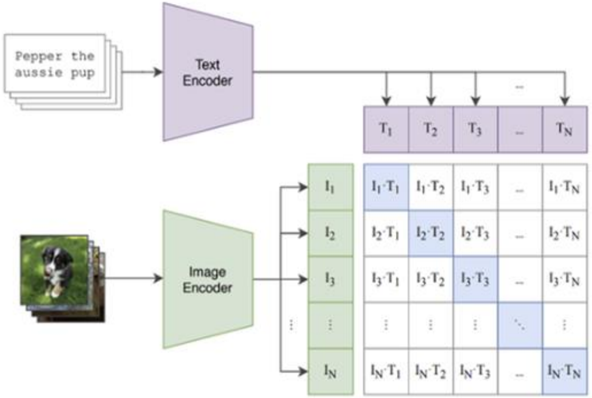
Bert-based Models



VilBert, VisualBert, VL-BERT, UNITER, ALBERF,HERO, ...

Text models are usually very small
Limited language understanding

Dual-encoder Contrastive Models

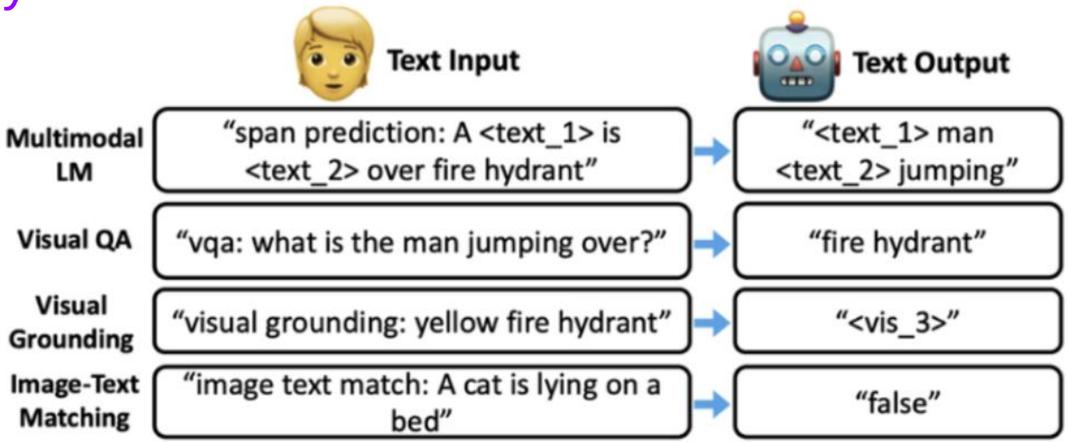


CLIP, ALIGN, CoCa, Florence, MIL-NCE, BASIC, LiT, FILIP, MMV, ...

paperswithcode.com/methods/category/vision-and-language-pre-trained-models

Vision Language Model Overview

Today ...



Vision Language Model

- Capable of performing different tasks, including VQA.
- Additional knowledge and intelligence comes from adding a large language model to the VLM

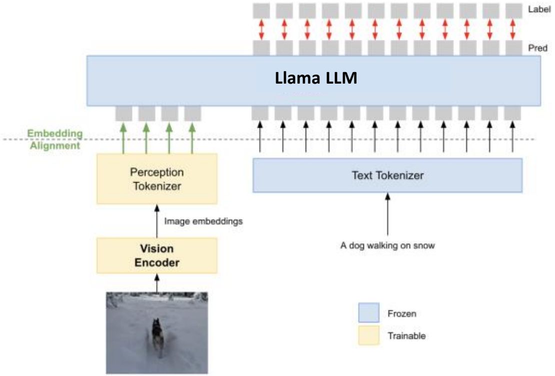
How LLMs Understand Images?

Fusing vision tokens with Large Language Models

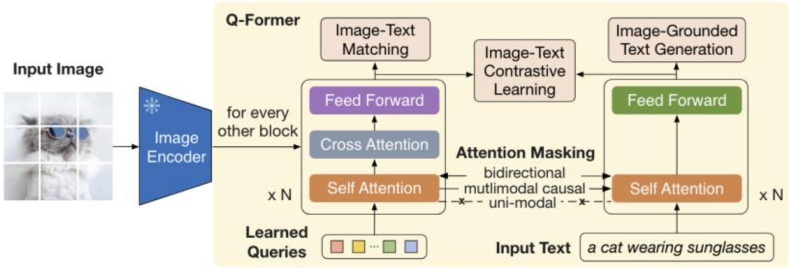
Why is the vision encoder important for MM-LLM?

- Allows LLMs to input/ understand images

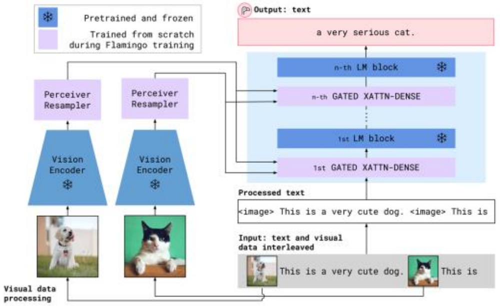
How to fuse and align modality tokens with the large language model?



AnyMAL | LLaVa | Fuyu



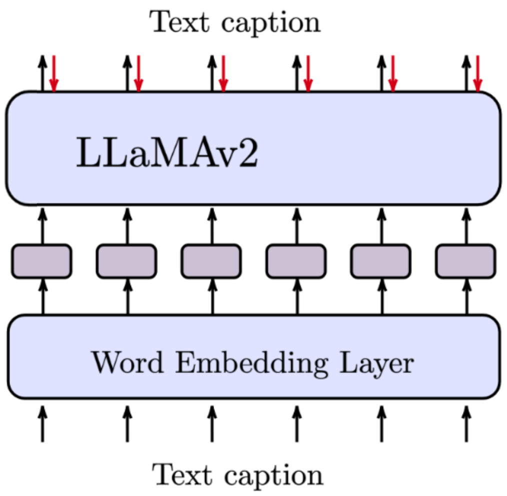
BLIP-v2



Flamingo

Li et. al, BLIP-2: Bootstrapping Language-Image Pre-training.
Moon et. al, AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model
Alayrac et. al, Flamingo: a Visual Language Model for Few-Shot Learning

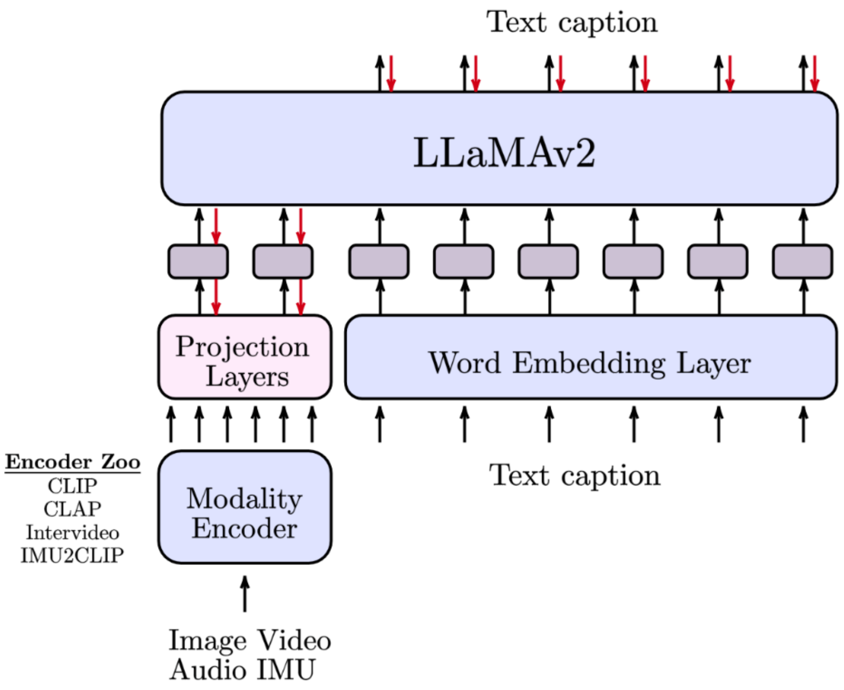
AnyMAL Overview (1)



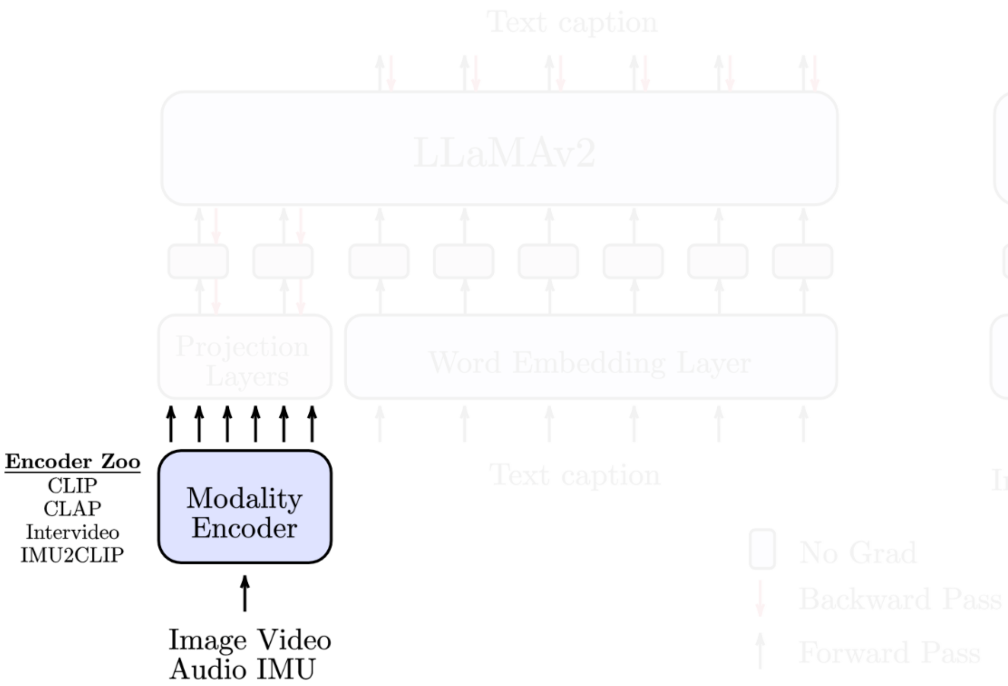
Base Language Model

- Preserves the strong language-based capabilities
- Variations
  - OPT
  - FlanT5
  - Llama
  - Llama-2-chat

AnyMAL Overview (2)

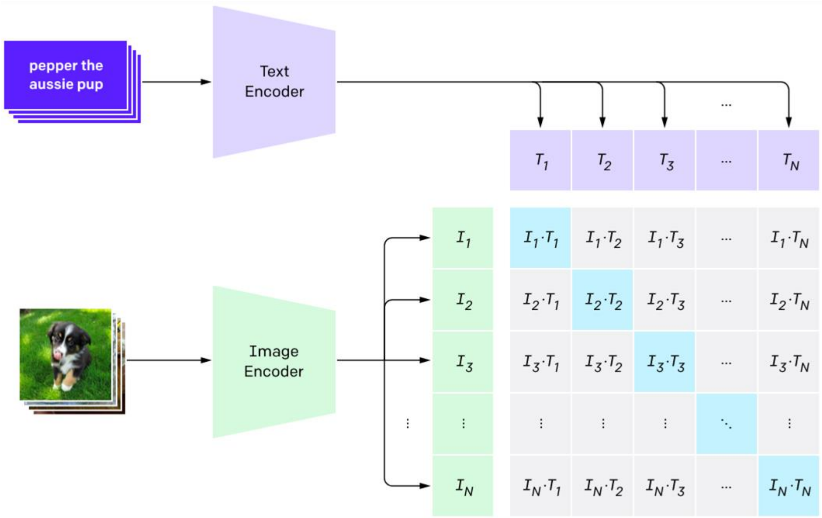


AnyMAL Overview (3)



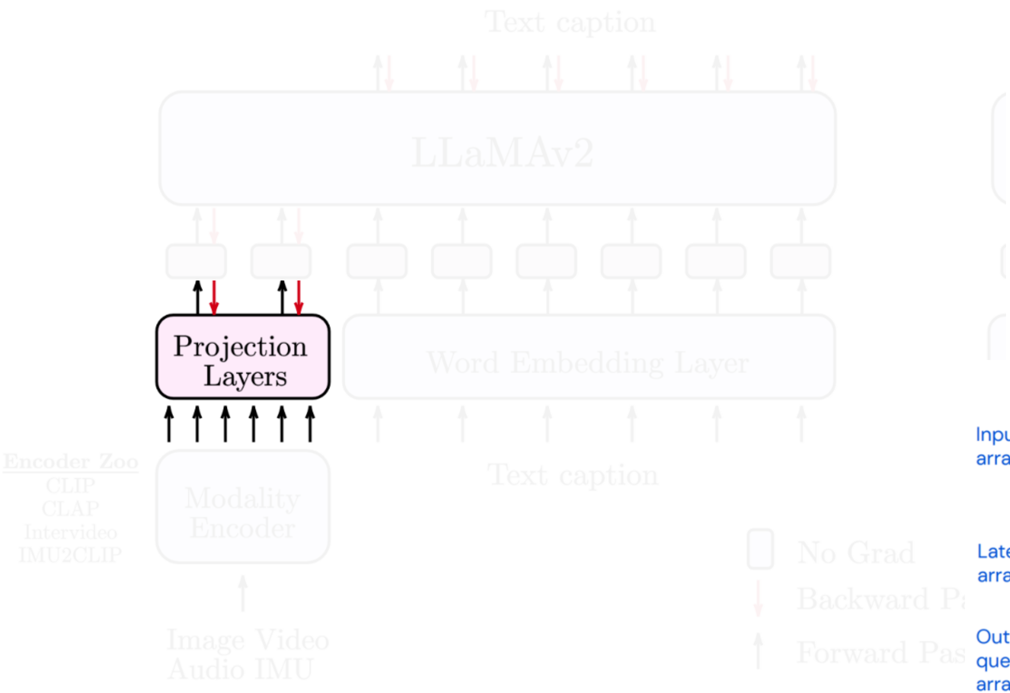
Modality Encoder

- Trained with contrastive loss (text & other modality) for the best alignment in the text space



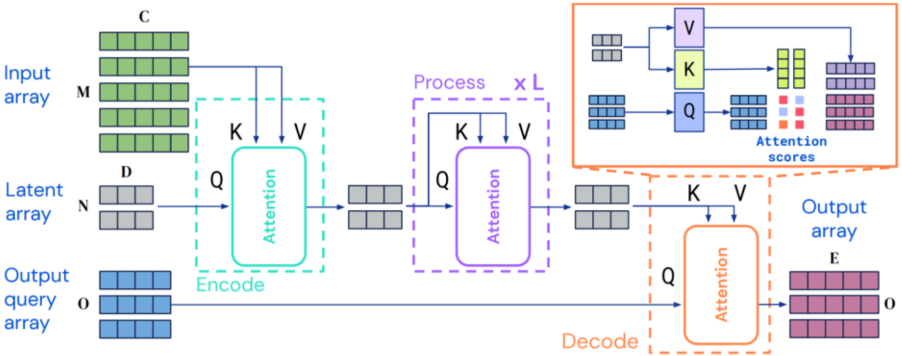
OpenAI, "CLIP: Connecting Text and Images", 2021

AnyMAL Overview (4)

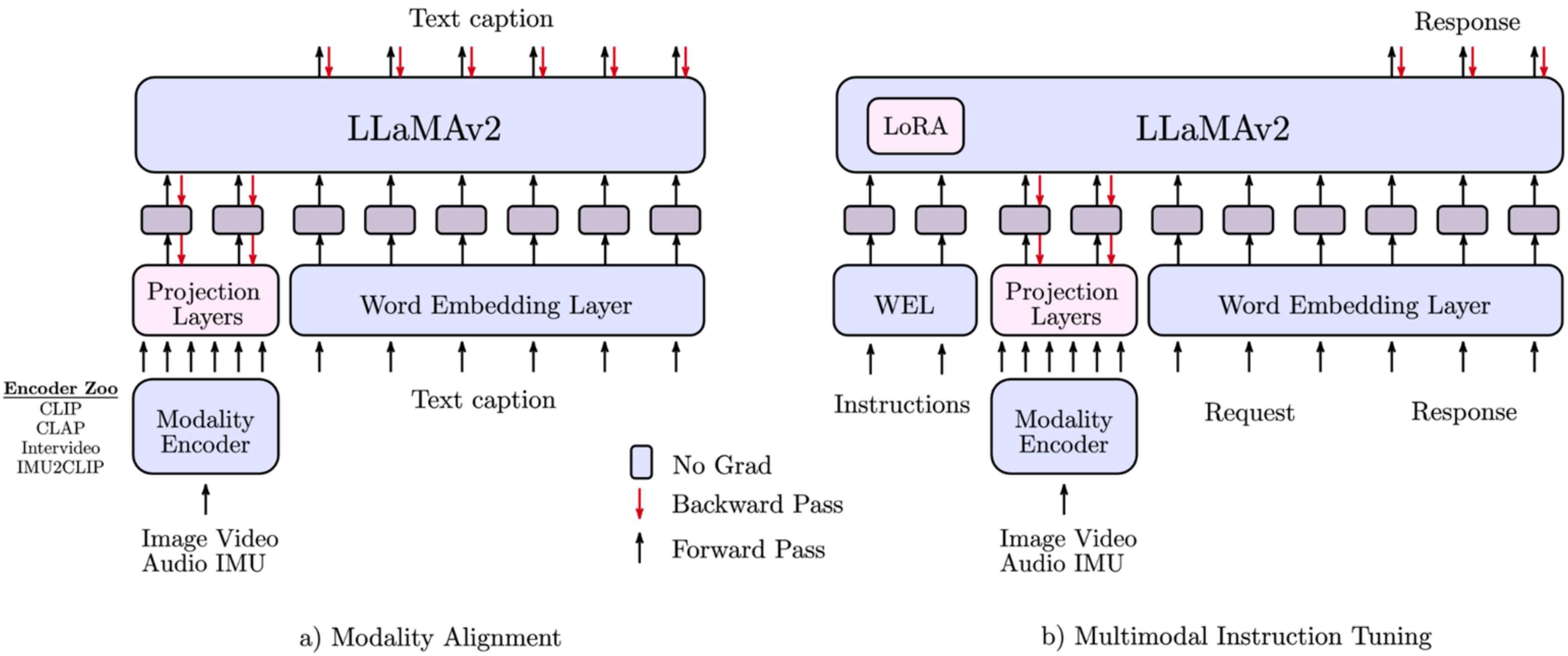


Projection Layers

- Perceiver Resampler to resample patch embeddings into a sequence of Llama-compatible tokens



Training: Pre-training & Fine-tuning



Moon et. al, AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model

Vision language model training
=
Vision encoder with language alignment

What Does the Vision Encoder See?

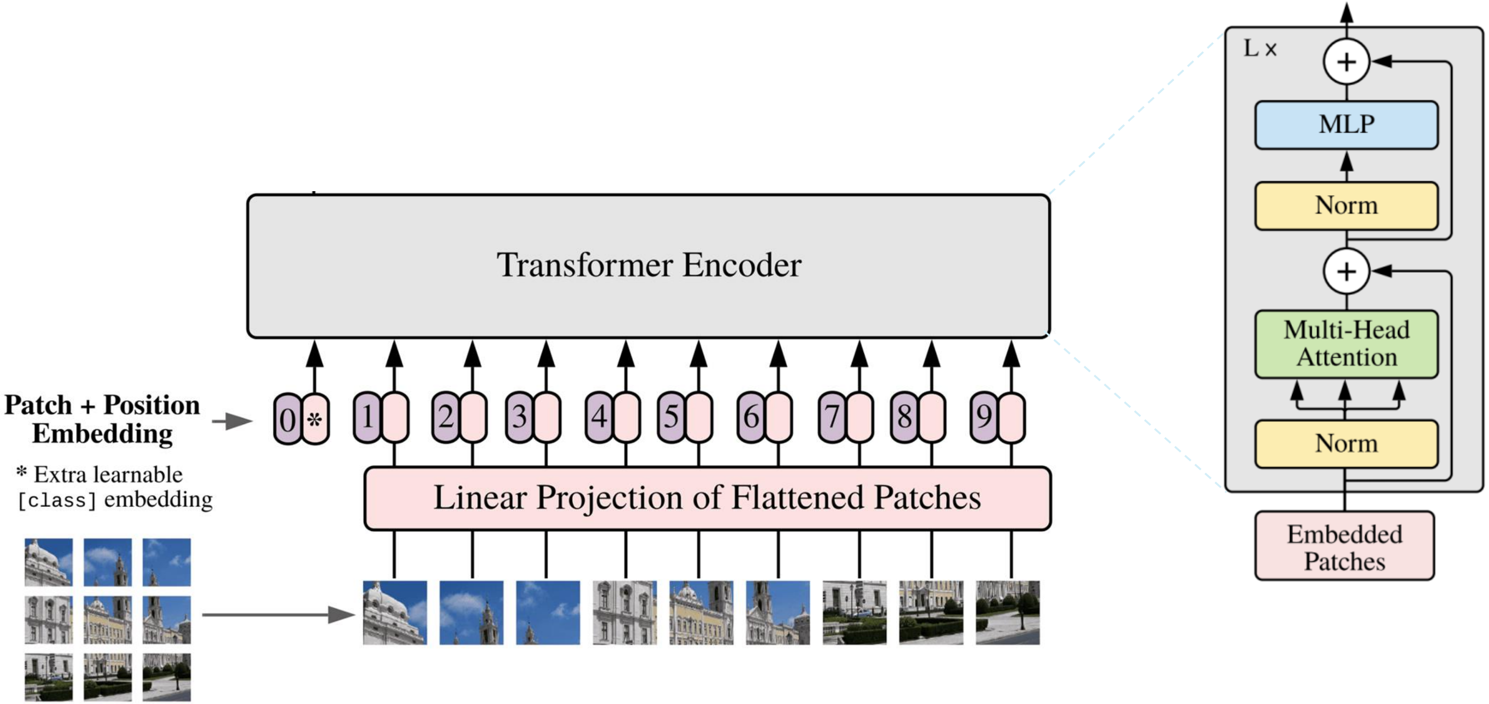
Look at that person

What is the name of that hotel?

What is that?

Li et. al. CLIP Surgery for Better Explainability with Enhancement in Open-Vocabulary Tasks

Vision Transformer Backbone



Dosovitskiy et. al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

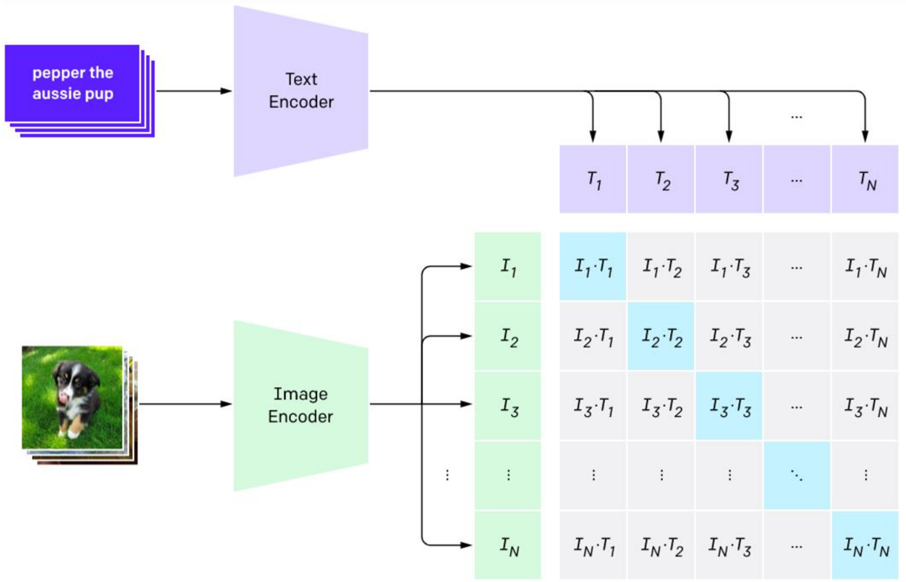
Training a High Quality CLIP Vision Encoder (1)

Key ingredients:

- Data volume and quality
- Model scale (number of parameters)
- Efficient pre-training recipes
- Domain fine-tuning: higher resolutions, diverse data, better caption quality

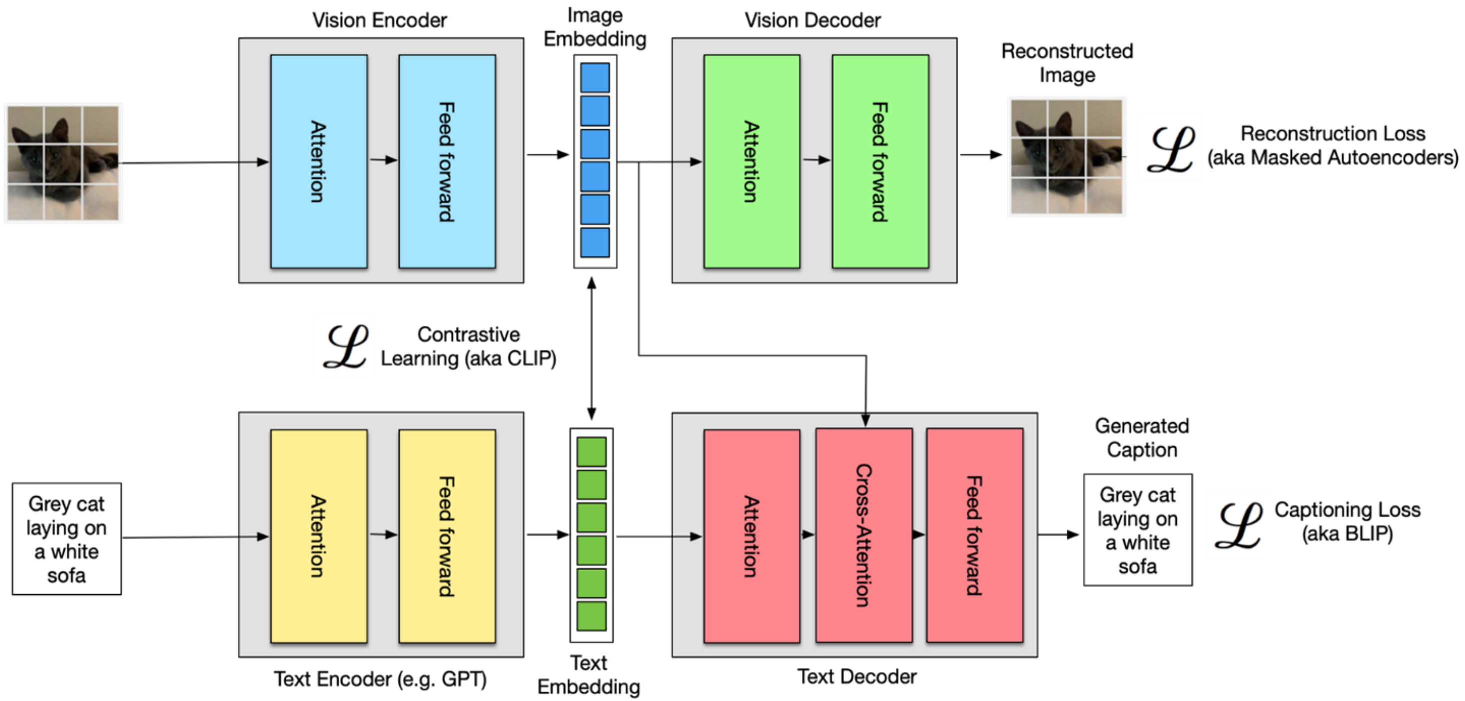
Measuring quality:

- Zero-shot ImageNet (IN1k-0 shot)
- Post-alignment metrics
  - VQA accuracy
  - Captioning accuracy



Radford et. al. Learning Transferable Visual Models From Natural Language Supervision
Li et. al, An Inverse Scaling Law for CLIP Training

Training a High Quality CLIP Vision Encoder (2)



Auxiliary loss functions

Improving alignment performance with LLM

Yu et. al. CoCa: Contrastive Captioners are Image-Text Foundation Models

Resources

We are hiring!

www.metacareers.com

AI at Meta
ai.meta.com

For more information:

AnyMAL
Any-Modality Augmented Language Model

Seungwhan Moon*, Andrea Madotto*, Zhaojiang Lin*, Tushar Nagarajan*, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, Kavya Srinet, Babak Damavandi, Anuj Kumar

Meta Reality Labs & FAIR