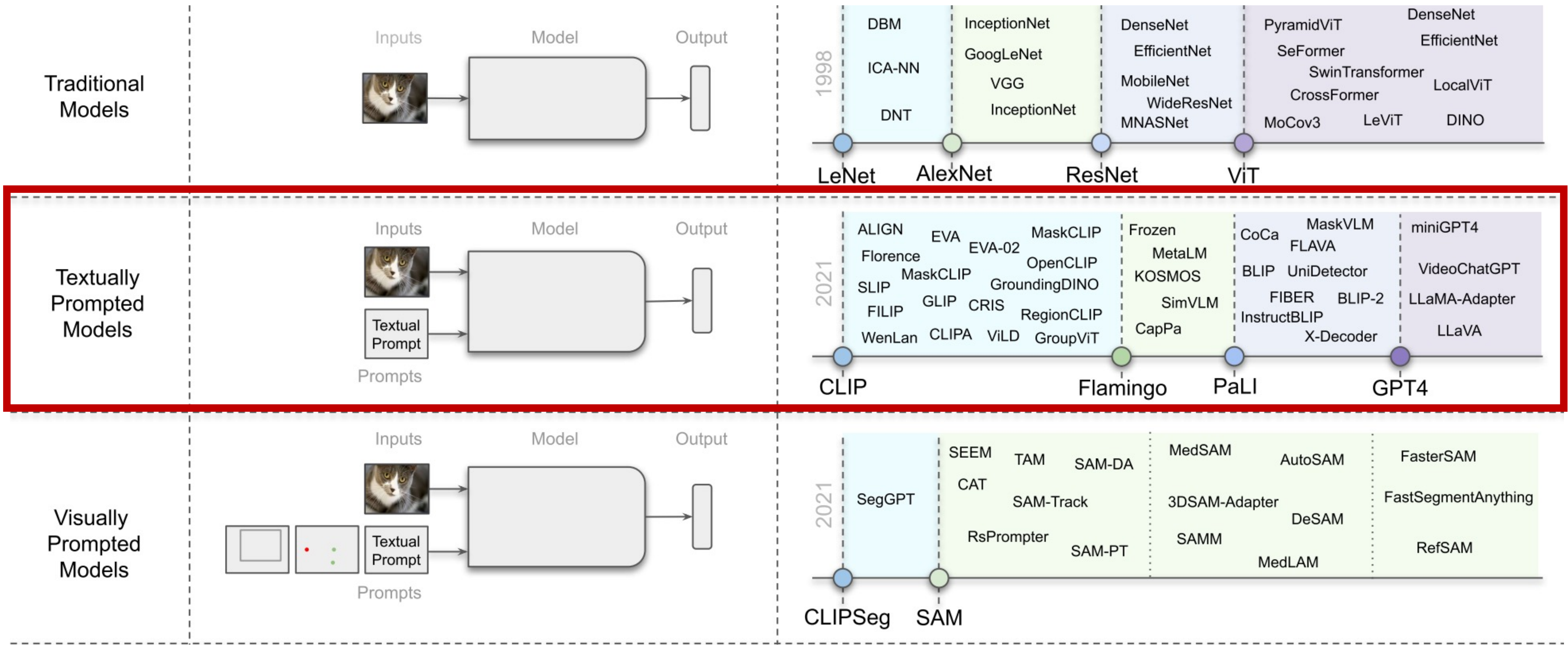
Large Language Models

Multi-modal Foundation Models: Vision-Language Models (VLMs)

M. Soleymani
Sharif University of Technology
Fall 2023

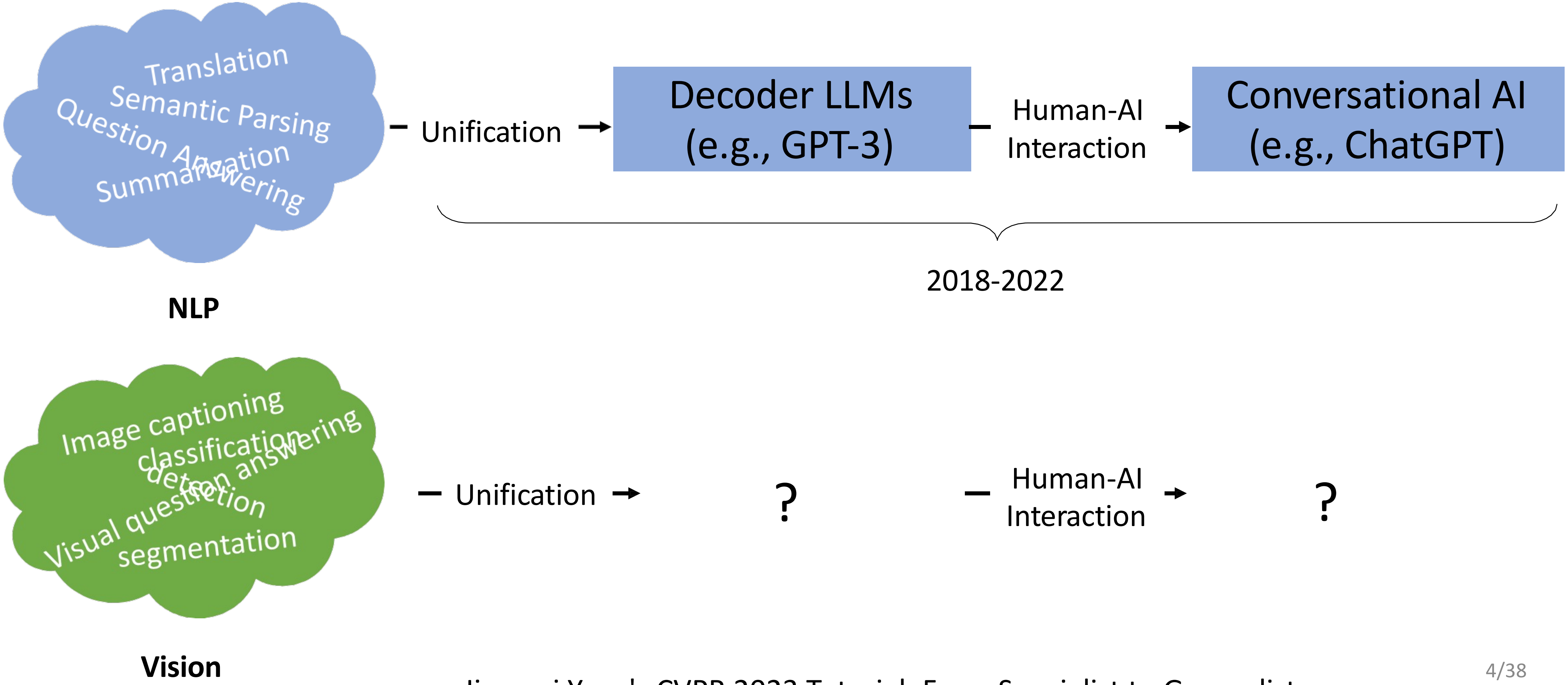
Multi-modal data

- Multimodal data:
  - Input and output from different modalities (e.g. text-to-image, image-to-text)
  - Inputs are multimodal (e.g. a system that can process both text and images)
  - Outputs are multimodal (e.g. a system that can generate both text and images)

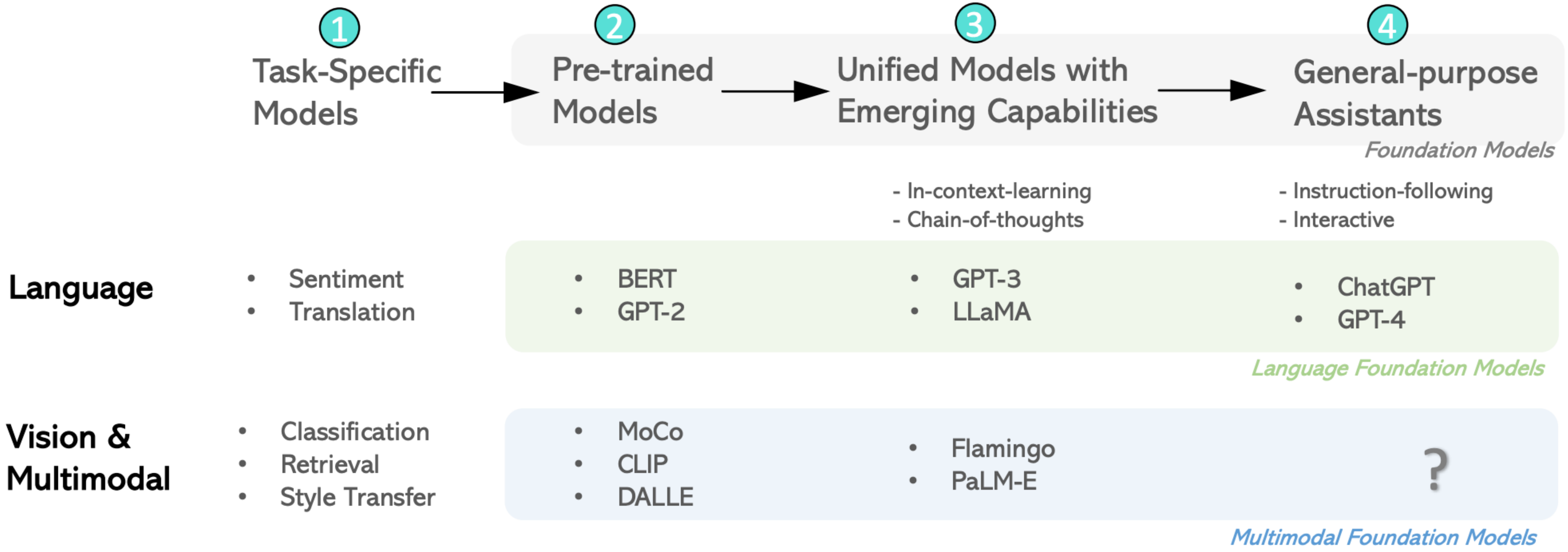


Awais et al., “Foundational Models Defining a New Era in Vision: A Survey and Outlook”, 2023

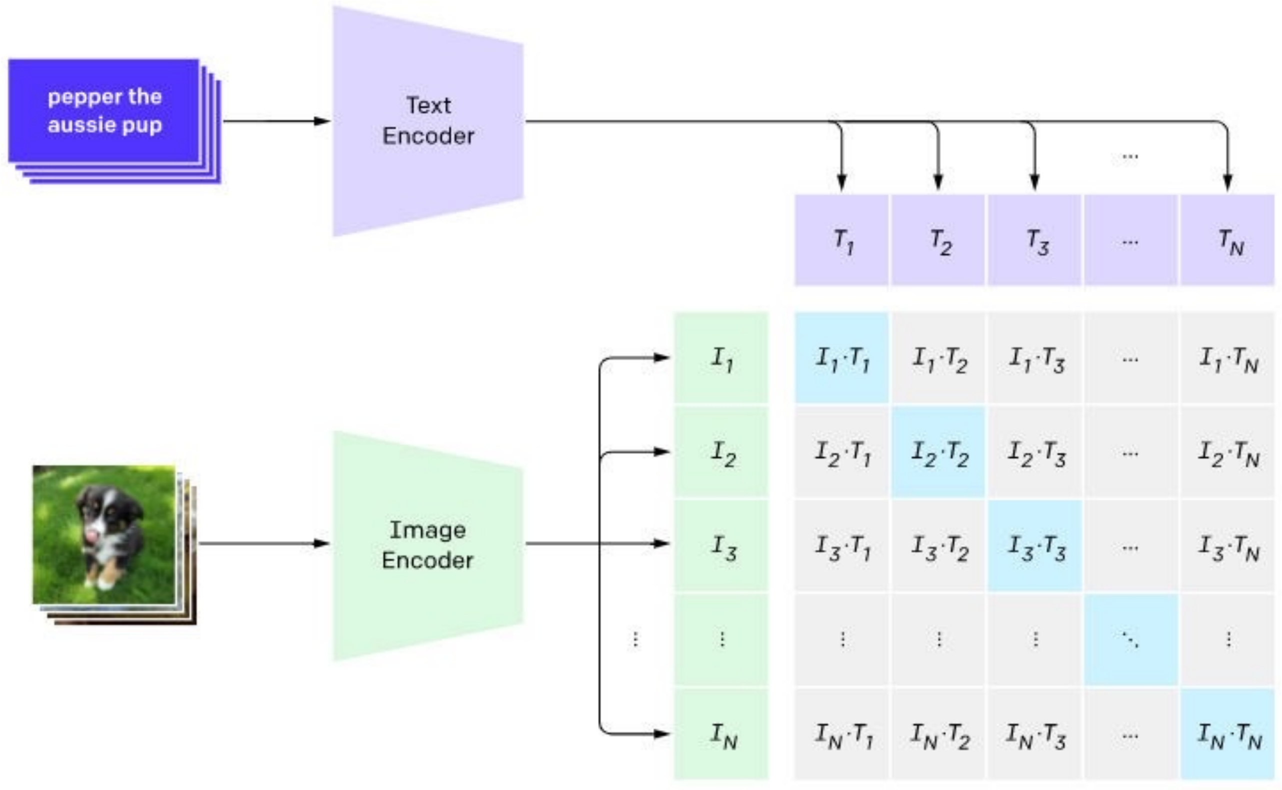
A Lesson from LLMs



A Lesson from LLMs



CLIP: Models and Training Complexity



- Text encoder:
  - 12-layer Transformer with causal mask
- Image encoder:
  - ResNet families: RN50, RN101, RN50x4, RN50x16, RN50x64
  - ViT families: ViT-B/32, ViT-B/16, ViT-L/14

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision", ICML 2021

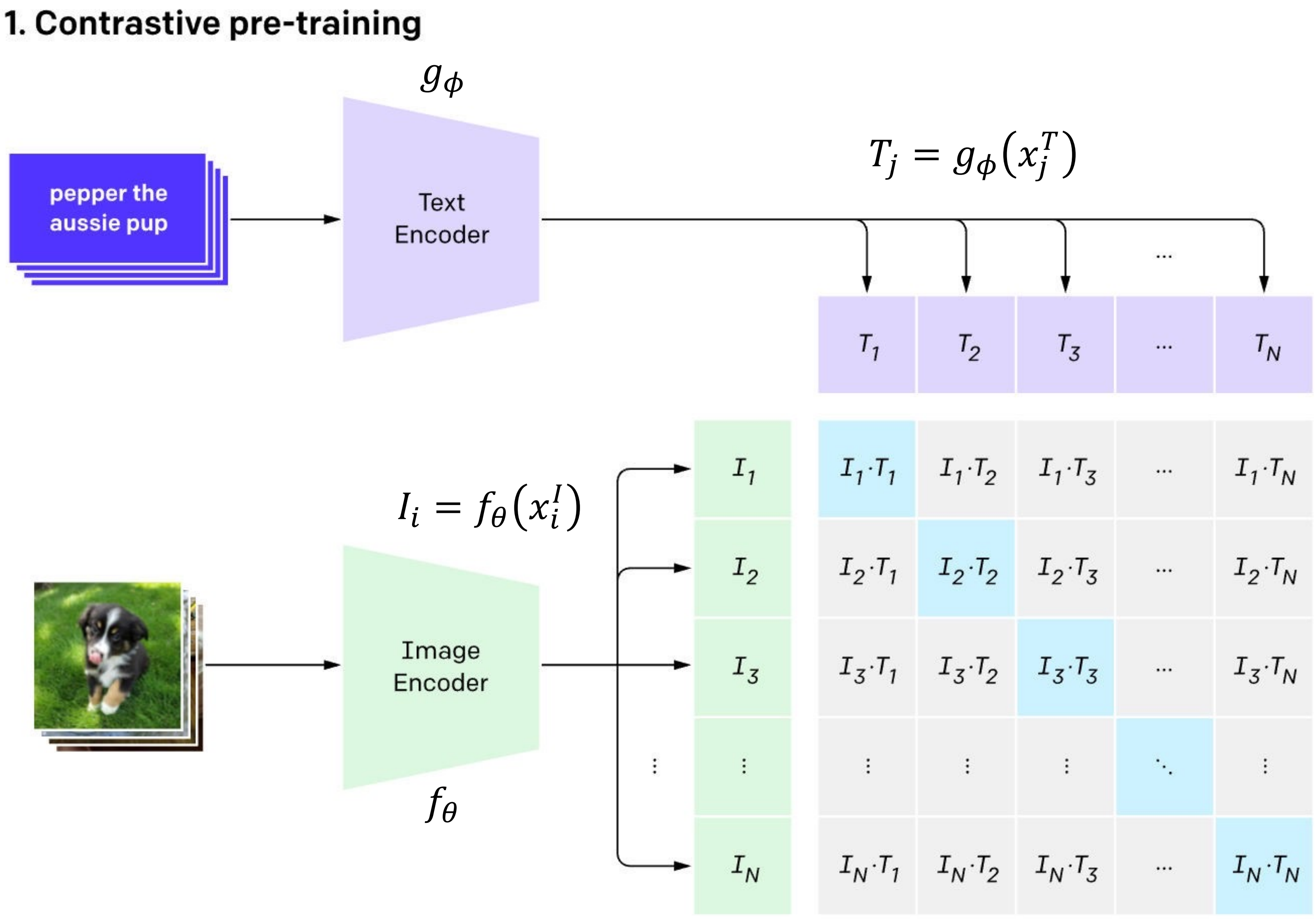
Vision-language models: Contrastive learning

- Contrastive training to bridge the image and text embedding spaces

- Making embedding of (image, text) pairs similar and that of non-pairs dissimilar

- This embedding space is super helpful for performing searches across modalities
  - Can return the best caption given an image
  - Has impressive capabilities for zero-shot adaptation to unseen tasks, without the need for fine-tuning

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision”, ICML 2021



$$s_{i,j}^T = s_{i,j}^I = I_i^T T_j$$

$$\mathcal{L}_i^I = -\log \frac{e^{s_{i,i}^I}}{\sum_{j=1}^N e^{s_{i,j}^I}}$$

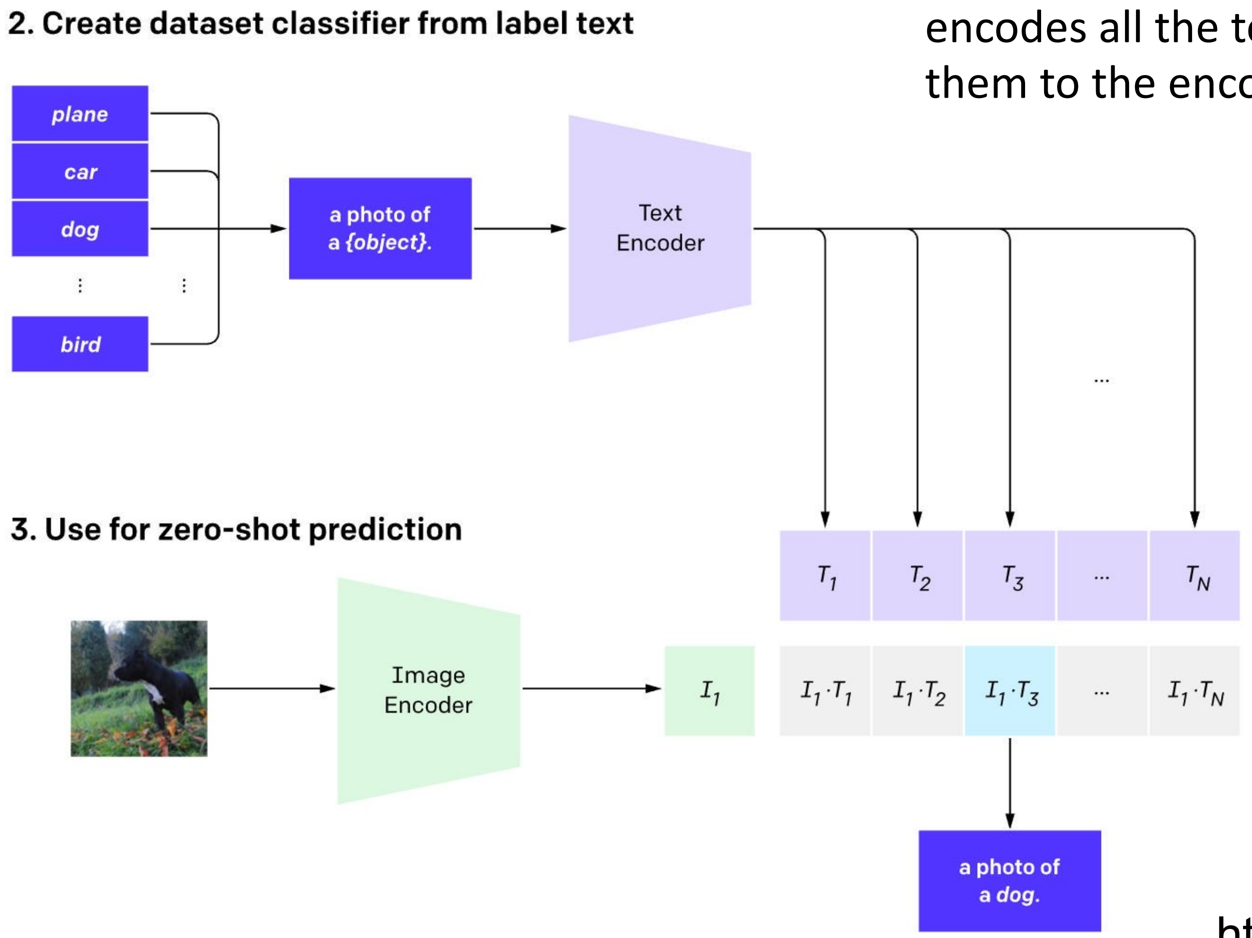
$$\mathcal{L}_j^T = -\log \frac{e^{s_{i,i}^T}}{\sum_{i=1}^N e^{s_{i,j}^T}}$$

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^N (\mathcal{L}_i^I + \mathcal{L}_j^T)$$

- Training batchsize: 32,768
- Training time:
  - RN50x64: 18 days on 592 V100 GPUs
  - ViT-L/14: 12 days on 256 V100 GPUs

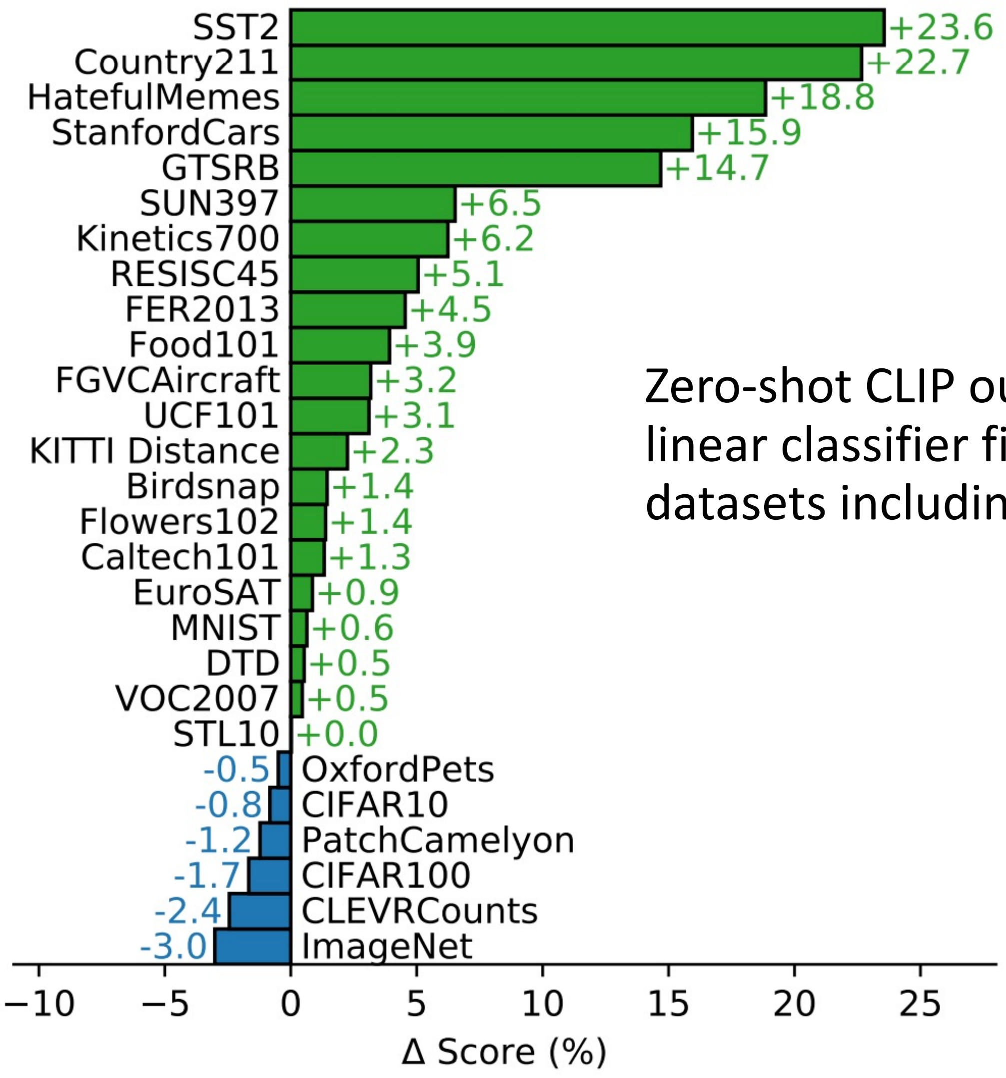
https://openai.com/blog/clip/

CLIP for zero-shot learning



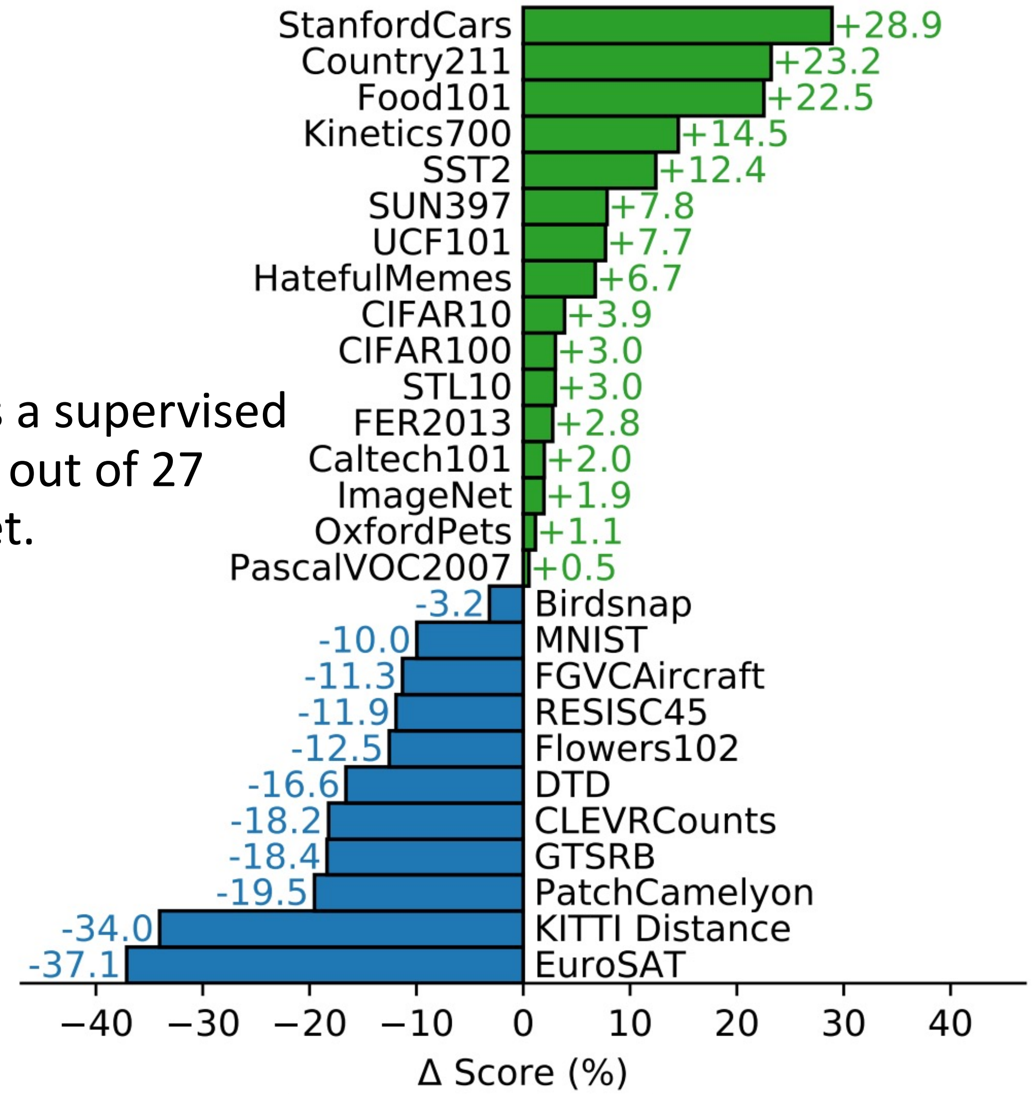
encodes all the text labels and compares them to the encoded image

https://openai.com/blog/clip/



Linear-probing CLIP outperforms the linear probing Noisy Student EfficientNet-L2

Zero-shot CLIP outperforms a supervised linear classifier fitted on 16 out of 27 datasets including ImageNet.

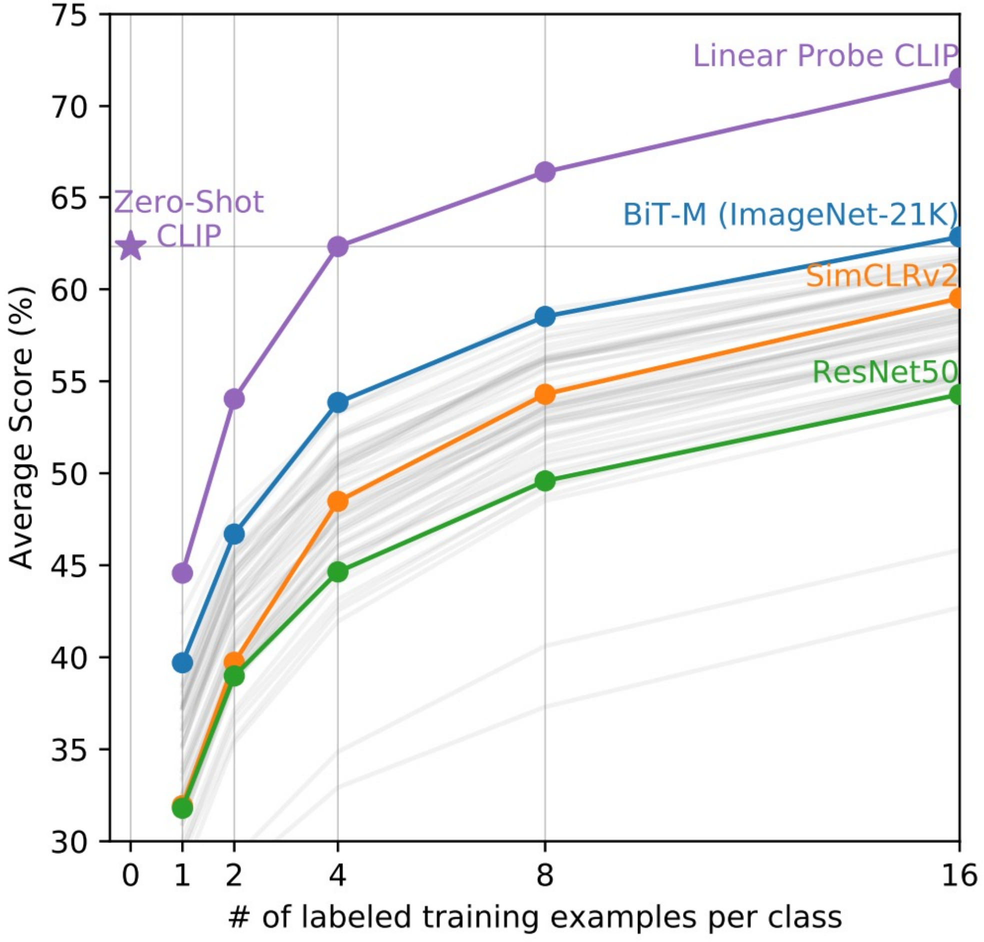


Zero-shot CLIP is competitive with a fully supervised linear-probing ResNet50

CLIP's features outperform the features of the best ImageNet model on a wide variety of datasets.

Radford et al. "Learning Transferable Visual Models From Natural Language Supervision", ICML 2021

Zero-shot CLIP outperforms few-shot linear probes



Radford et al. "Learning Transferable Visual Models From Natural Language Supervision", ICML 2021

Vision Language Tasks

Large Multi-modal Models (LMMs) in their current form is primarily generates a text sequence.

	Image Captioning	Text-to Image Retrieval	Image-to-Text Retrieval	VQA	Text-to-Image Generation
Input	Image: 	Query: A couple of zebra walking across a dirt road. A pool of images	Query:  A pool of texts	Image:  Q: why did the zebra cross the road?	Text: A couple of zebra walking across a dirt road.
Output	A couple of zebra walking across a dirt road.		A couple of zebra walking across a dirt road.	A: to get to the other side (Selected from a pool of 3,129 answers in VQAv2 or generate answer)	
	Generation	**Understanding**	**Understanding**	**Understanding/Generation**	**Generation**

Source: P. Zhang and L. Zhang, A Tutorial On Vision Language Intelligence

CLIP: Summary

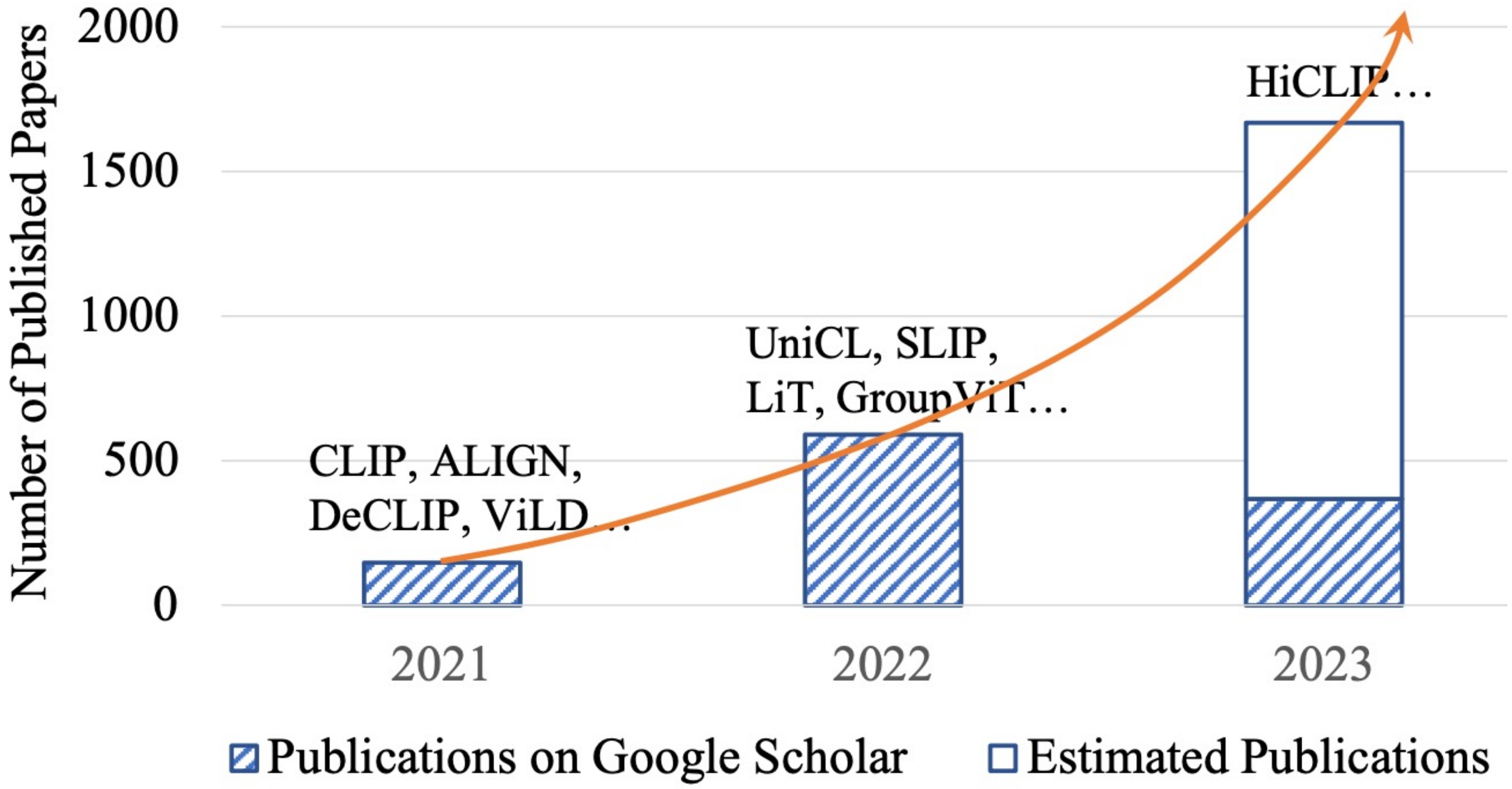
✓ CLIP improved open-vocabulary visual recognition capabilities through learning from Internet-scale image-text pairs.

✗ CLIP doesn't go directly from image to text or vice versa. It just connecting the image and text embedding spaces

- CLIP can only address limited use cases such as classification
- It crucially lack the ability to generate language which makes them less suitable to more open-ended tasks such as captioning or visual question answering

Survey of VLMs

Publication on VLMs

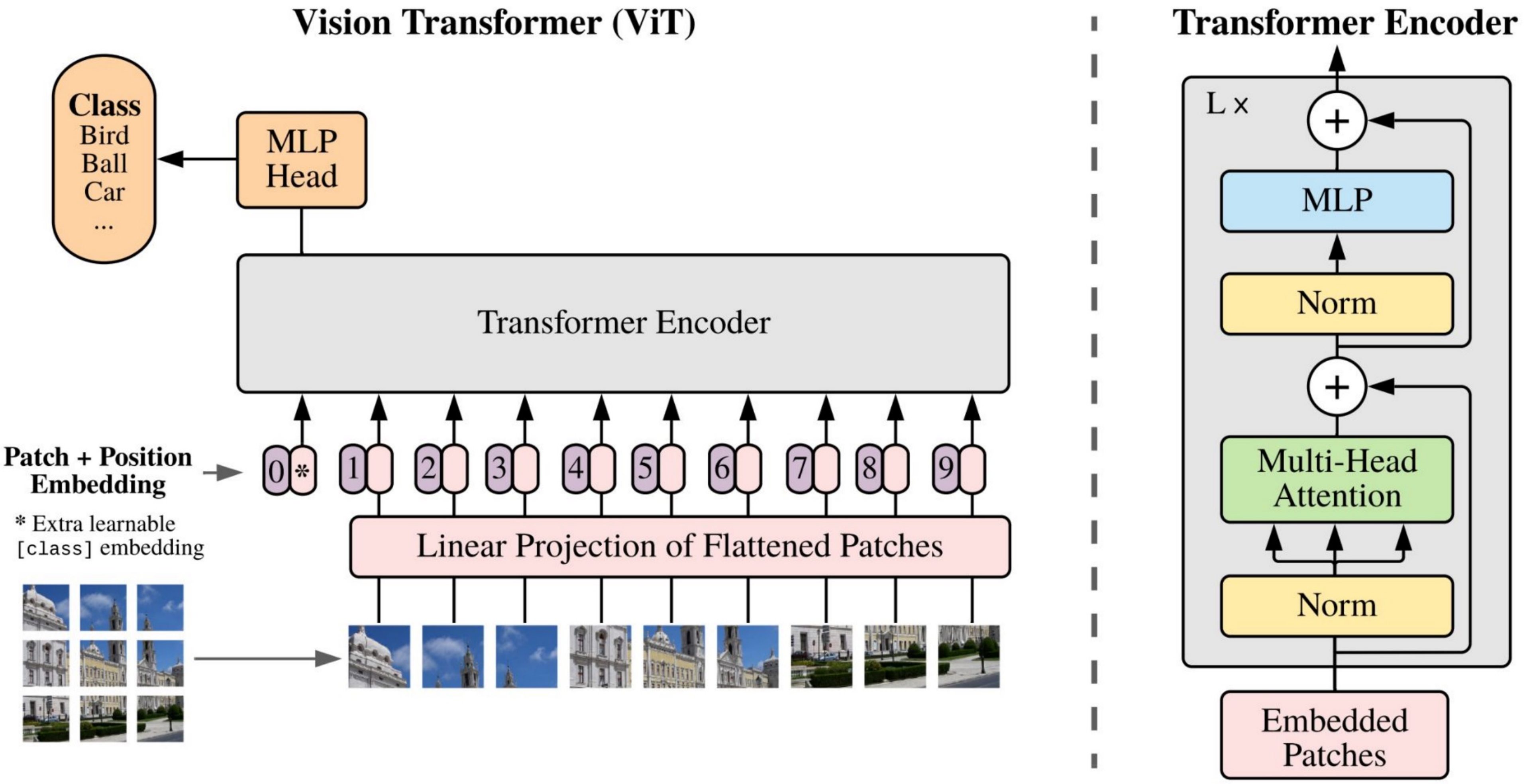


Zhang et al., “Vision-Language Models for Vision Tasks: A Survey”, 2023

CLIP Variants

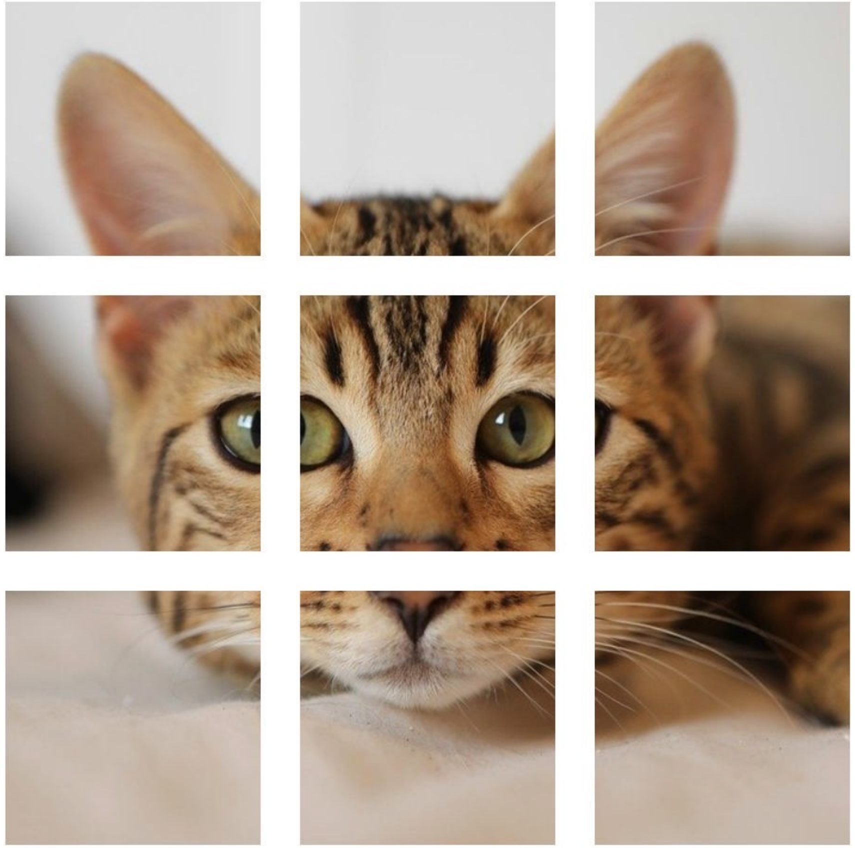
- Objective function or pretraining
  - Combining CLIP with label supervision (BASIC, UniCL, LiT, MOFI)
  - Contrastive + self-supervised image representation learning
    - Contrastive + Self-supervlised methods like SimCLR (SLIP, DeCLIP, nCLIP)
    - Contrastve + Masked Image Modeling (EVA, EVA-02, MVP)
  - Fine-grained matching loss (FILIP)
  - Region-level pretraining (RegionCLIP, GLIP)
  - Sigmoid loss for language-image pre-training (SigCLIP)

Vision Transfomer as Image Encoder Architecture

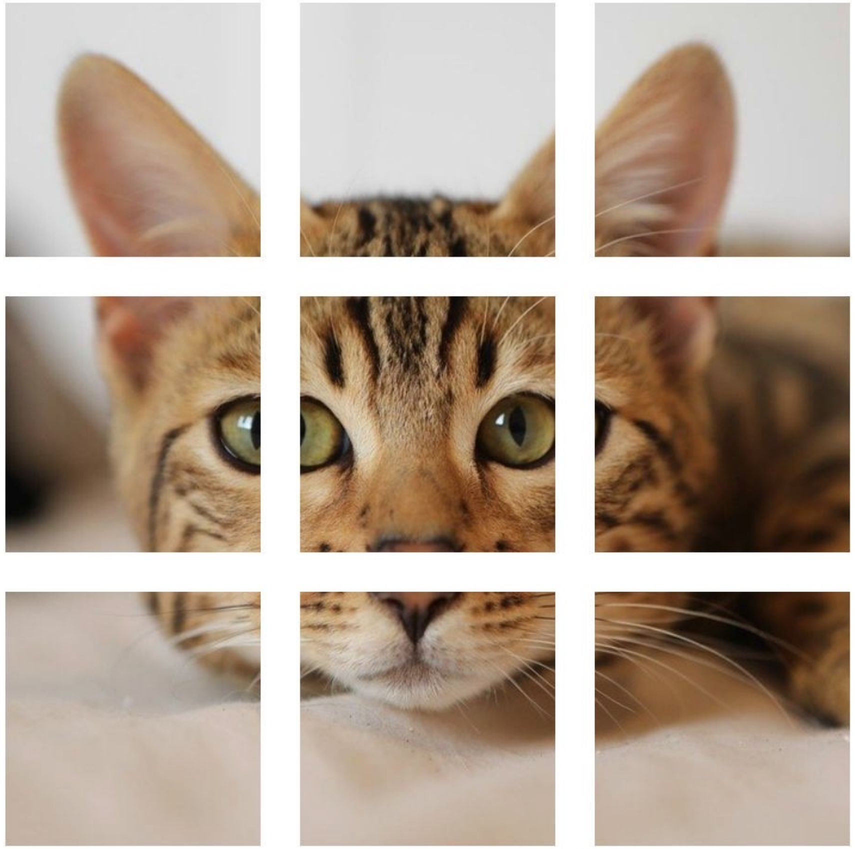


Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale", ICLR 2021

Slides from Justin Johnson



Slides from Justin Johnson

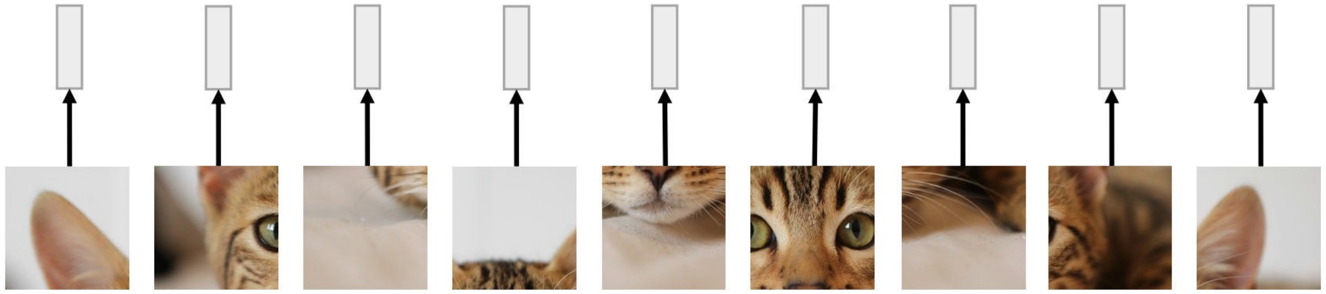


N input patches, each
of shape 3x16x16

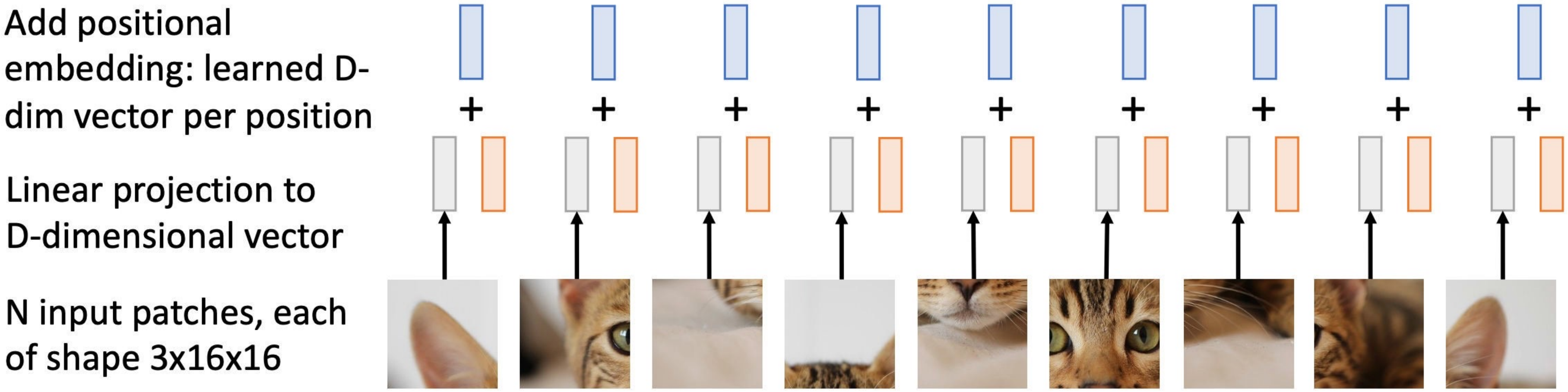
Slides from Justin Johnson

Linear projection to
D-dimensional vector

N input patches, each
of shape 3x16x16



Slides from Justin Johnson



Slides from Justin Johnson

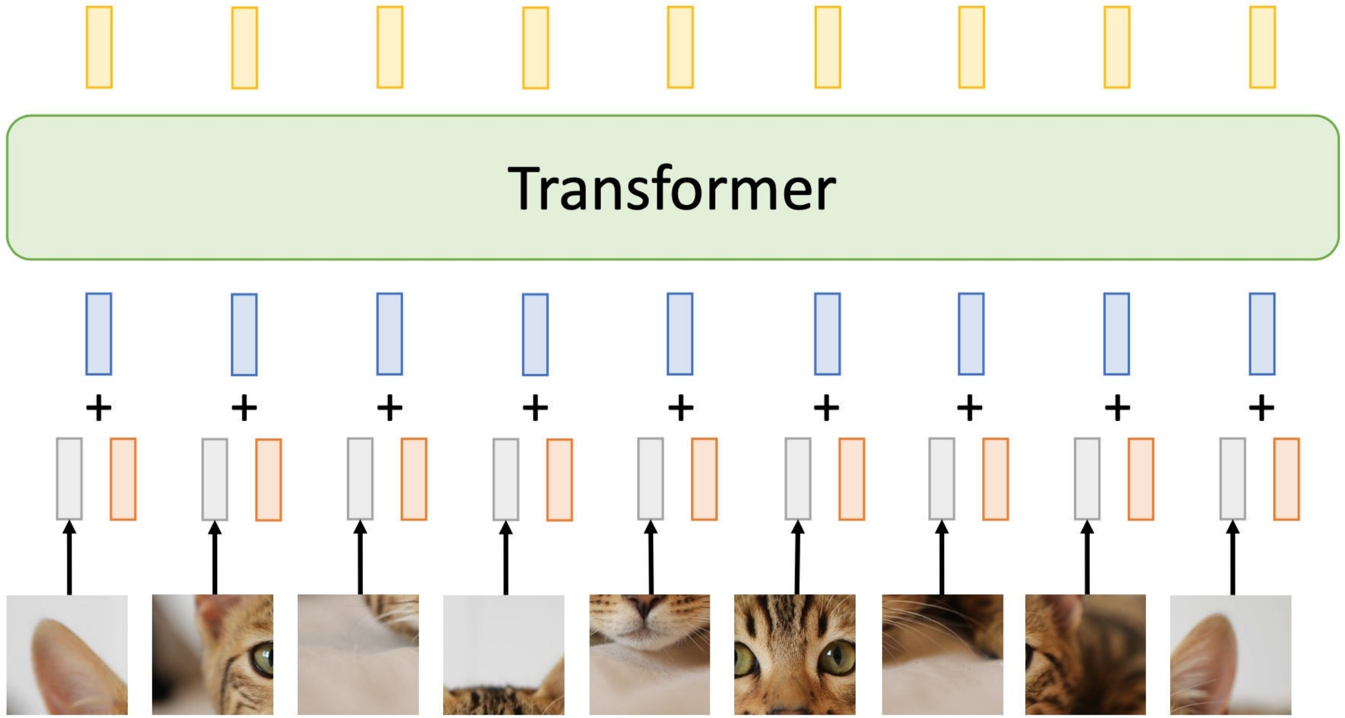
Output vectors

Exact same as
NLP Transformer!

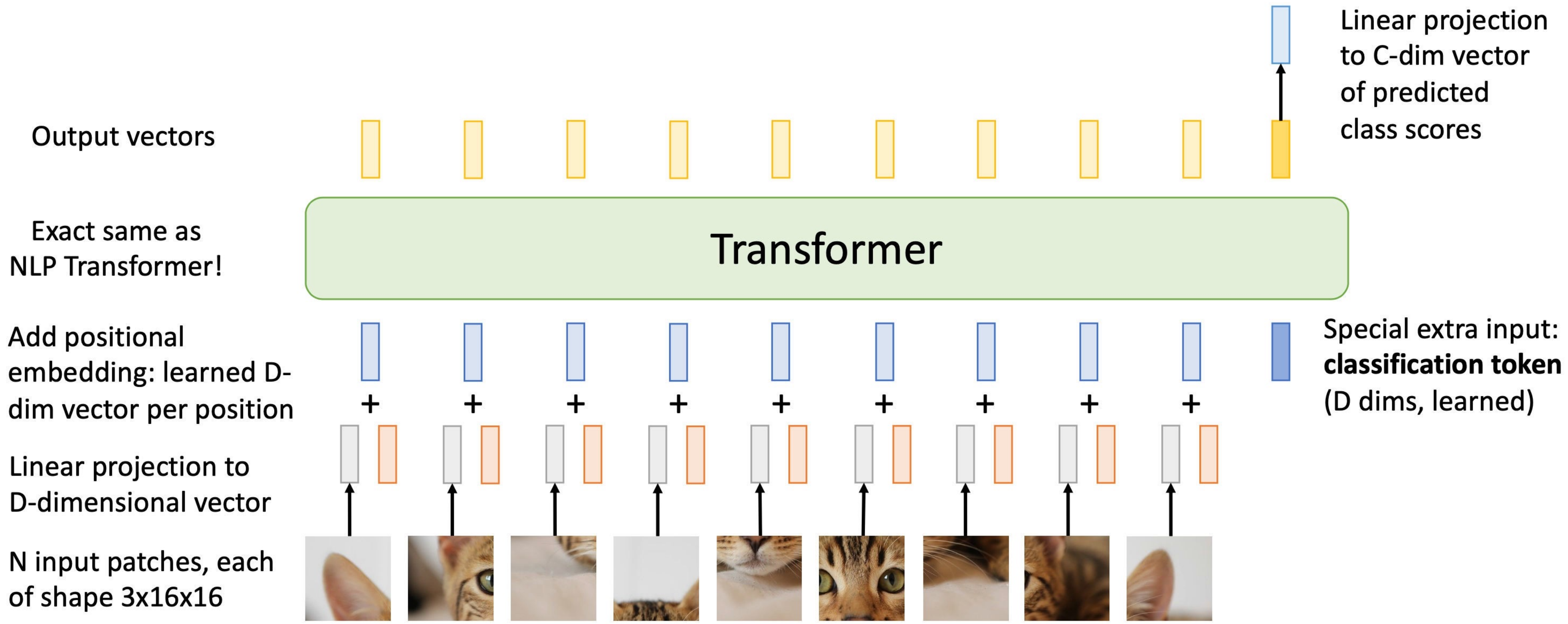
Add positional
embedding: learned D-
dim vector per position

Linear projection to
D-dimensional vector

N input patches, each
of shape 3x16x16



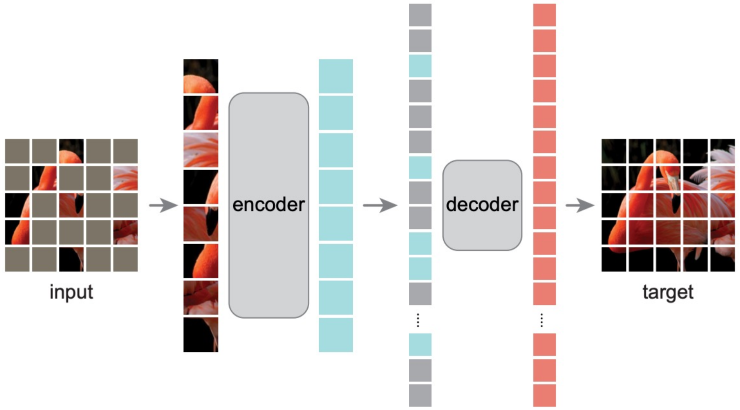
Slides from Justin Johnson



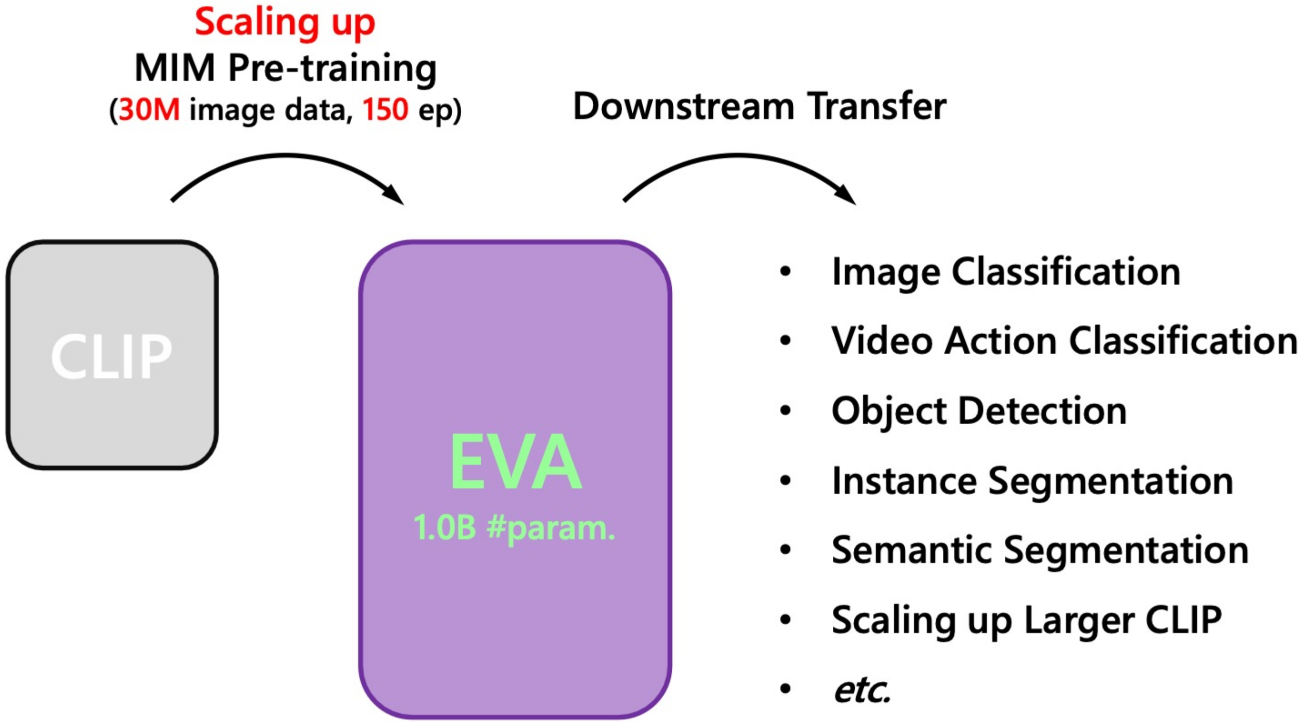
Slides from Justin Johnson

EVA

- Simply regressing the masked out image-text aligned vision features (*i.e.*, CLIP features) scales up well (to 1.0B parameters) and transfers well to various downstream tasks.

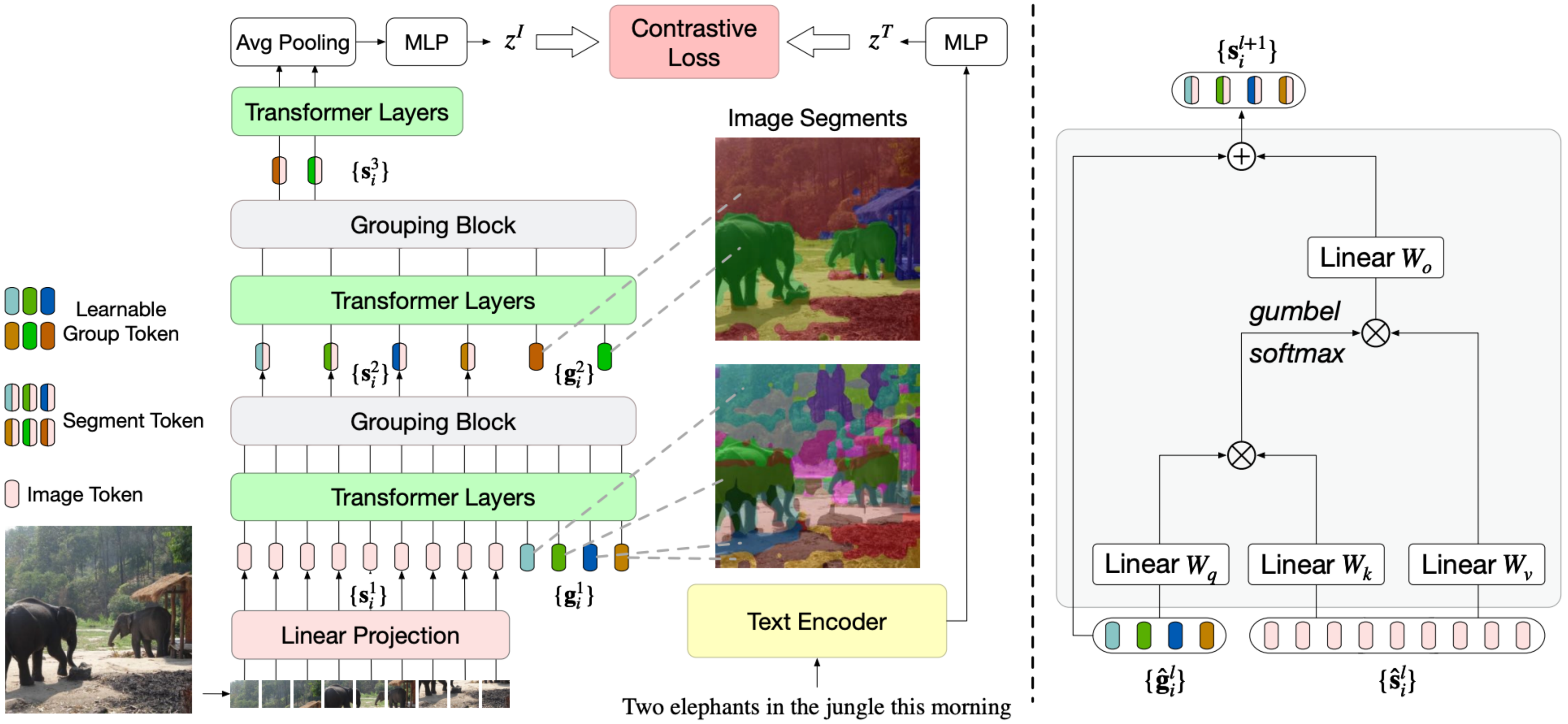


He et al., “Masked Autoencoders Are Scalable Vision Learners”, 2021



Fang et al., “EVA: Exploring the Limits of Masked Visual Representation Learning at Scale”, 2022.

GroupVIT



(a) GroupViT Architecture and Training Pipeline

(b) Grouping Block

Xu et al., “GroupViT: Semantic Segmentation Emerges from Text Supervision”, 2022

Learning to Prompt for VLMs

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
[V]$_1$ [V]$_2$ … [V]$_M$ [CLASS].	**91.83**

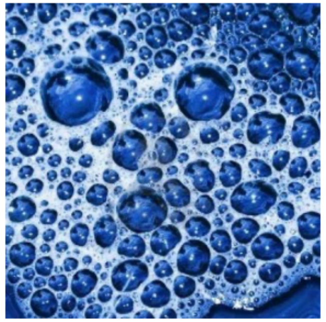
(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
[V]$_1$ [V]$_2$ … [V]$_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)



Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
[V]$_1$ [V]$_2$ … [V]$_M$ [CLASS].	**63.58**

(c)

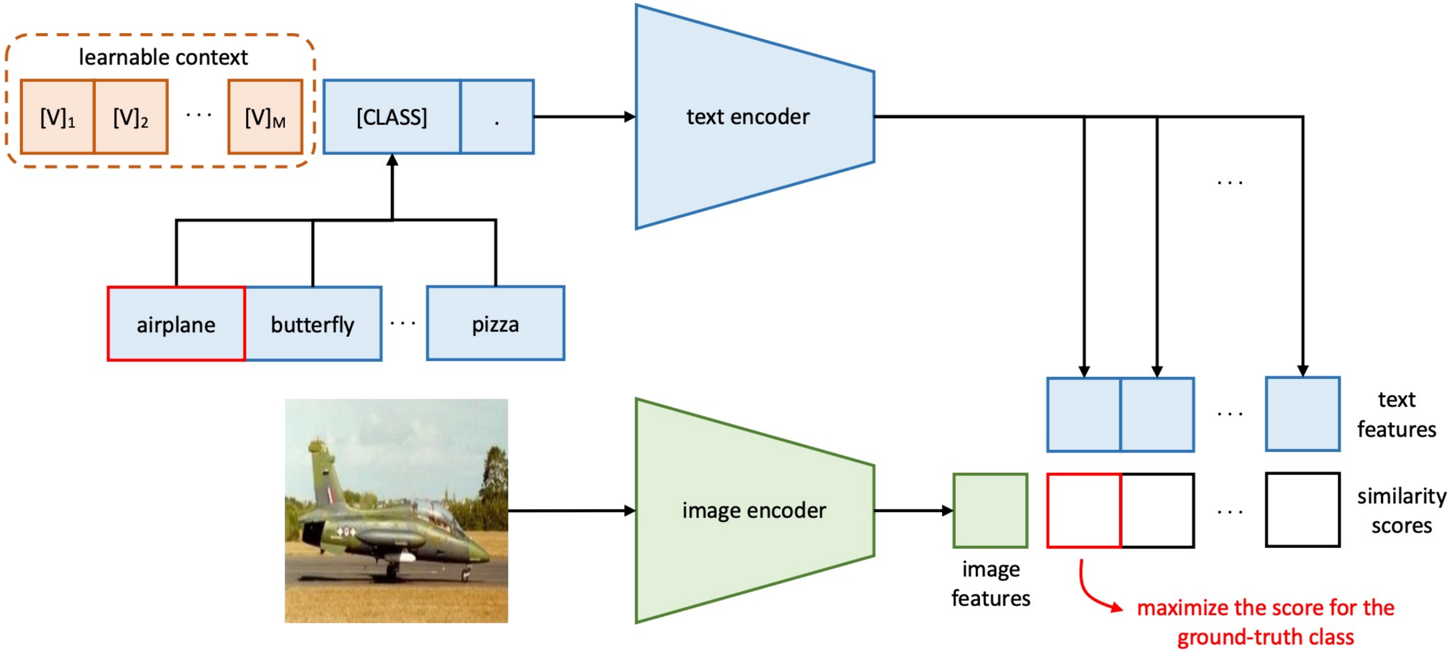
EuroSAT

Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
[V]$_1$ [V]$_2$ … [V]$_M$ [CLASS].	**83.53**

(d)

Zhu et al., "Learning to Prompt for Vision-Language Models", 2021

Learning to Prompt for VLMs

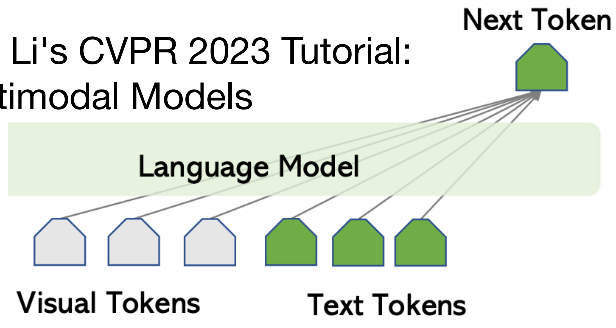


Method	Source ImageNet	Target -V2	-Sketch	-A	-R
ResNet-50					
Zero-Shot CLIP	58.18	51.34	33.32	21.65	56.00
Linear Probe CLIP	55.87	45.97	19.07	12.74	34.86
CLIP + CoOp ($M = 16$)	62.95	55.11	32.74	22.12	54.96
CLIP + CoOp ($M = 4$)	**63.33**	**55.40**	**34.67**	**23.06**	**56.60**
ResNet-101					
Zero-Shot CLIP	61.62	54.81	38.71	28.05	64.38
Linear Probe CLIP	59.75	50.05	26.80	19.44	47.19
CLIP + CoOp ($M = 16$)	**66.60**	**58.66**	39.08	28.89	63.00
CLIP + CoOp ($M = 4$)	65.98	58.60	**40.40**	**29.60**	**64.98**
ViT-B/32					
Zero-Shot CLIP	62.05	54.79	40.82	29.57	**65.99**
Linear Probe CLIP	59.58	49.73	28.06	19.67	47.20
CLIP + CoOp ($M = 16$)	**66.85**	58.08	40.44	30.62	64.45
CLIP + CoOp ($M = 4$)	66.34	**58.24**	**41.48**	**31.34**	65.78
ViT-B/16					
Zero-Shot CLIP	66.73	60.83	46.15	47.77	73.96
Linear Probe CLIP	65.85	56.26	34.77	35.68	58.43
CLIP + CoOp ($M = 16$)	**71.92**	64.18	46.71	48.41	74.32
CLIP + CoOp ($M = 4$)	71.73	**64.56**	**47.89**	**49.93**	**75.14**

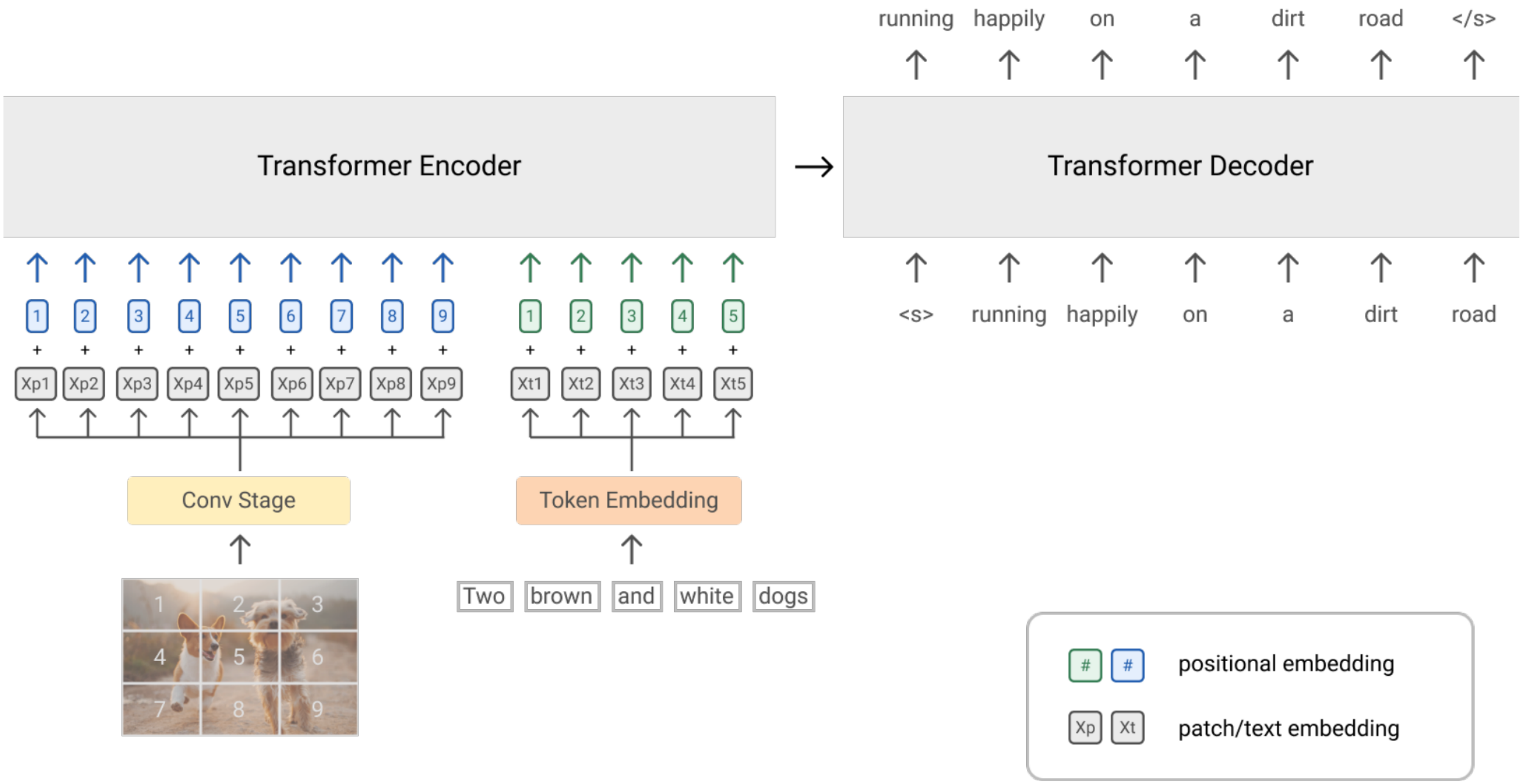
Zhu et al., "Learning to Prompt for Vision-Language Models", 2021

Vision-Language Models: Toward generative models

- Architecture
  - Dual encoders → CLIP & its mentioned variants
  - Encoder-decoder
  - Fusion decoder

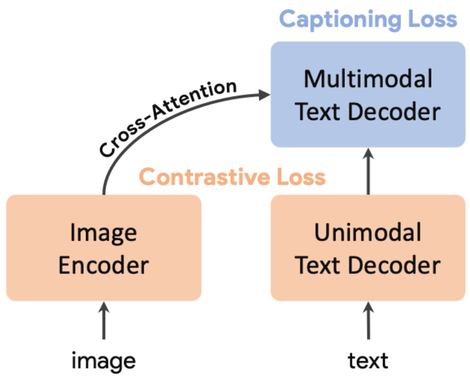


SimVLM



Wang et al., "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision", ICLR 2022

CoCa: Contrastive Captioner

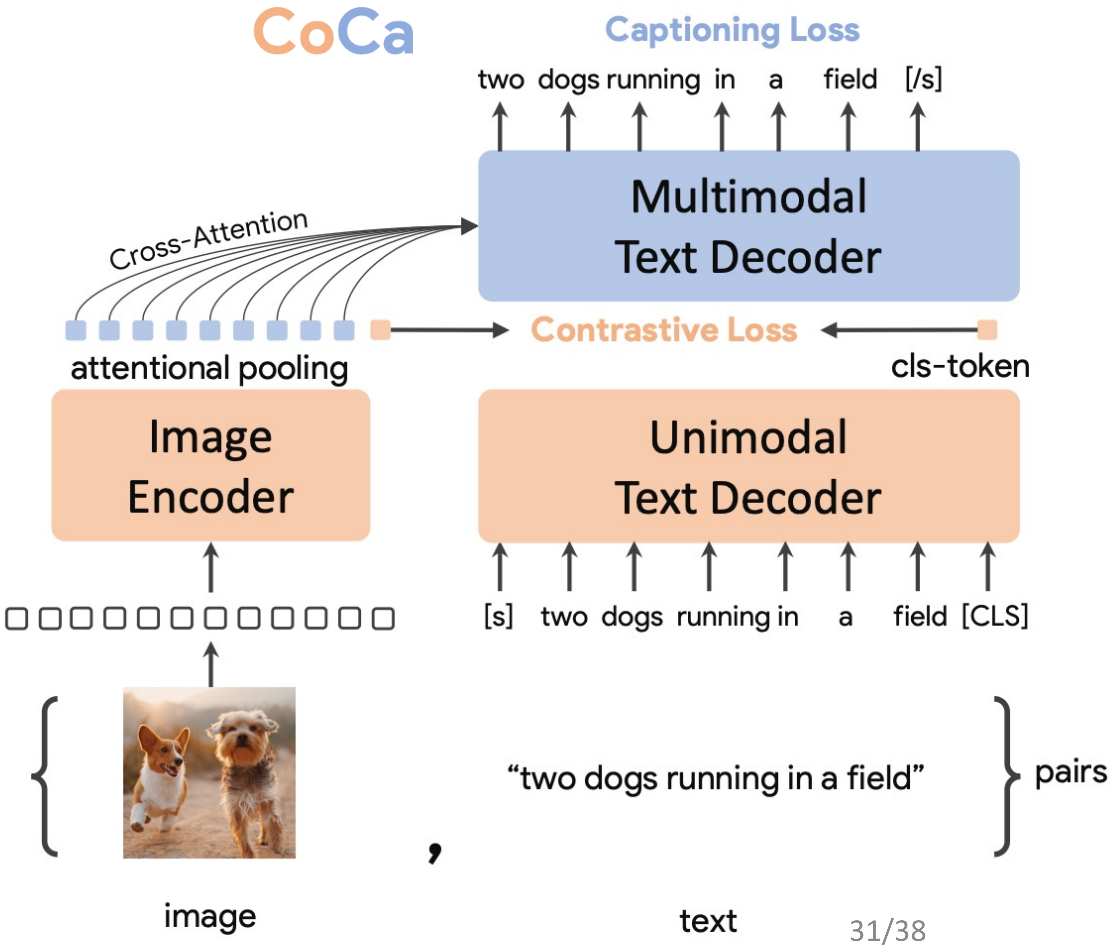


Multi-modal fusion

- Use mixed image-text and image-label (JFT-3B) data for pre-training

- A generative branch for enhanced performance and enabling new capabilities (image captioning and VQA)

- CoCa aims to learn a better image encoder from scratch

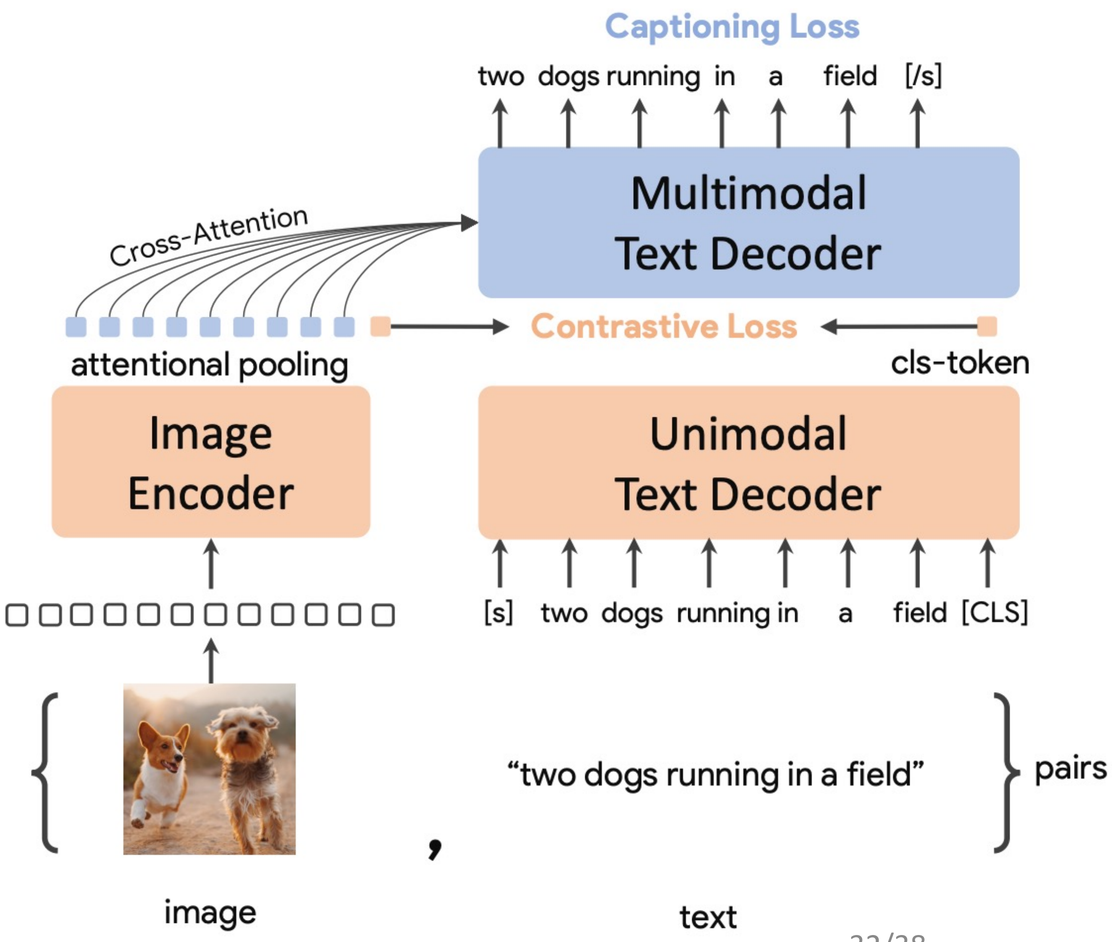


Yu et al., "Coca: Contrastive captioners are image-text foundation models", 2022

CoCa: Contrastive Captioner

- Use mixed image-text and image-label (JFT-3B) data for pre-training
- A generative branch for enhanced performance and enabling new capabilities (image captioning and VQA)
- CoCa aims to learn a better image encoder from scratch

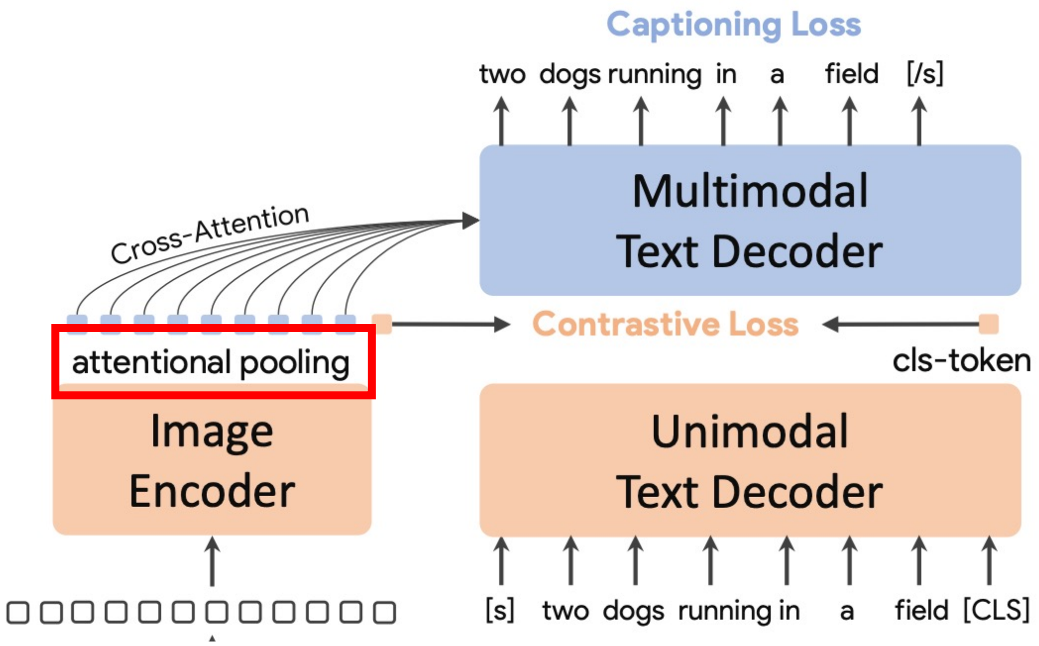
$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x).$$



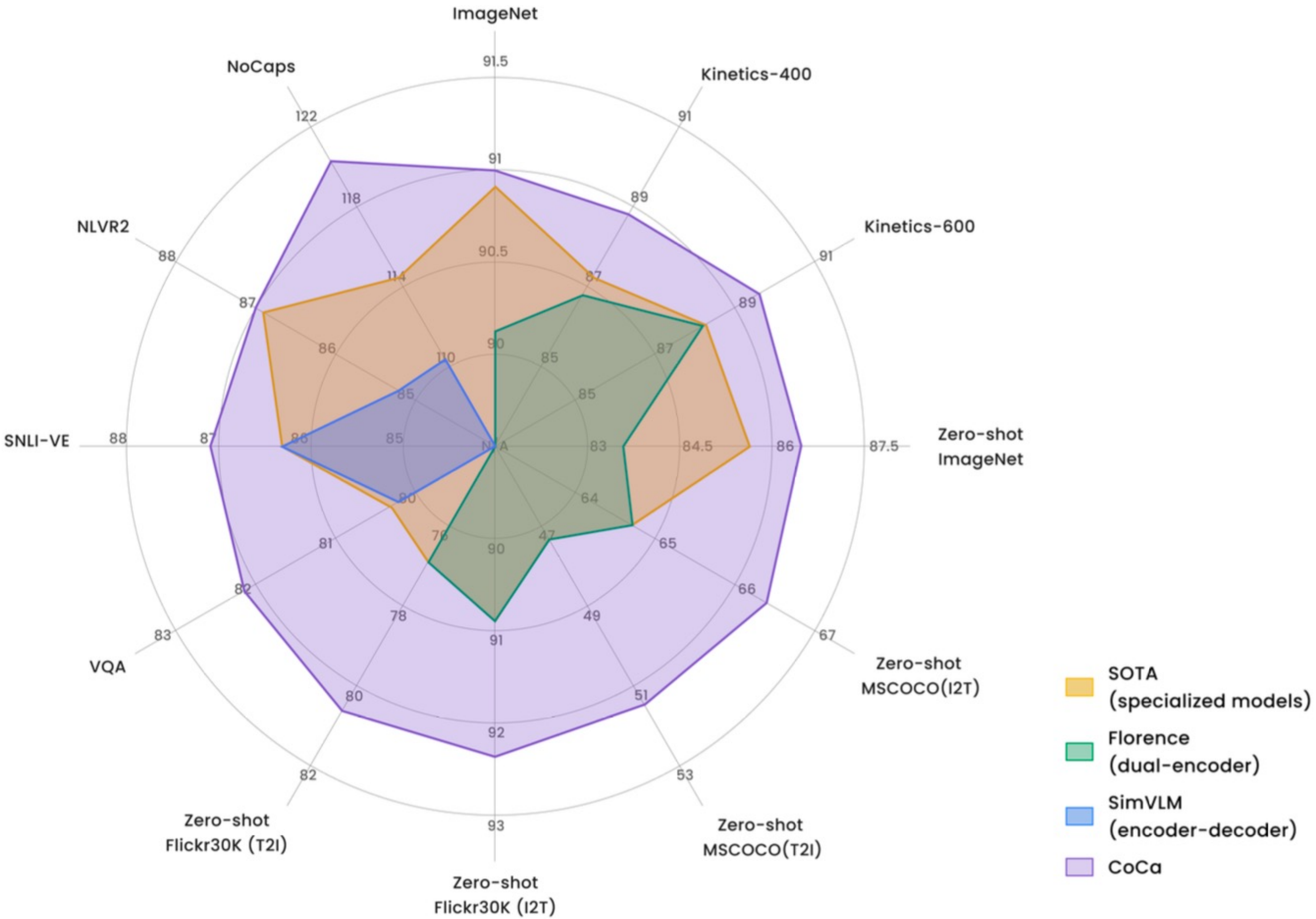
Yu et al., "Coca: Contrastive captioners are image-text foundation models", 2022

CoCa Architecture

- Unified single-encoder, dual-encoder, and encoder-decoder paradigms
  - one image-text foundation model with the capabilities of all three approaches
- Cross-attention is omitted in unimodal decoder layers to encode text-only representations
- Multimodal decoder cross-attending to image encoder outputs to learn multimodal representations.

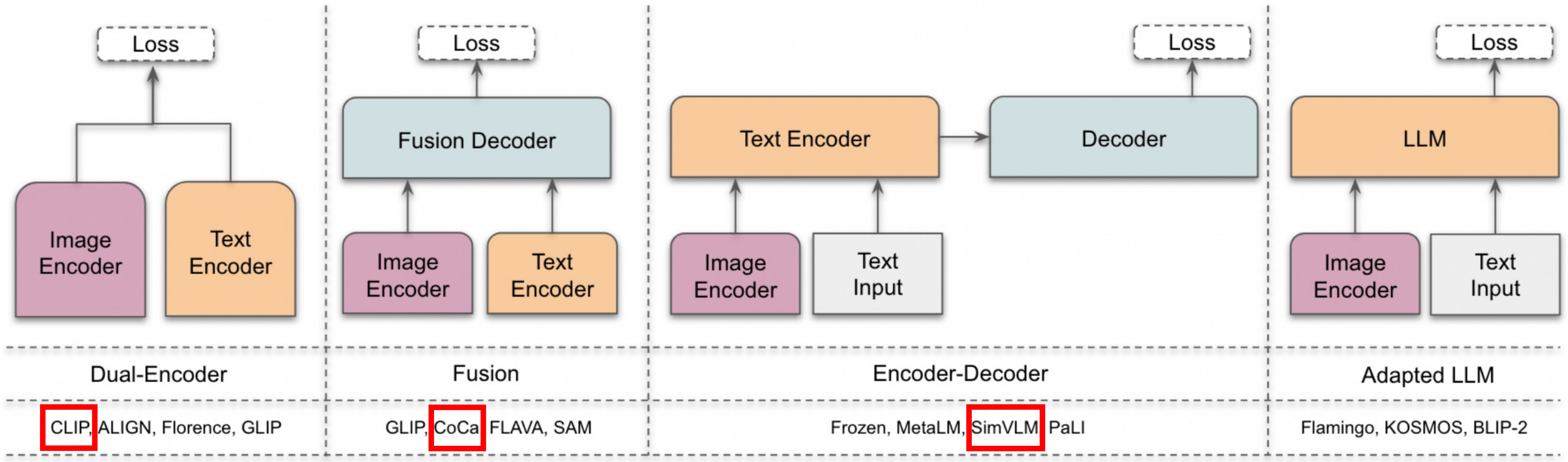


Yu et al., "Coca: Contrastive captioners are image-text foundation models", 2022



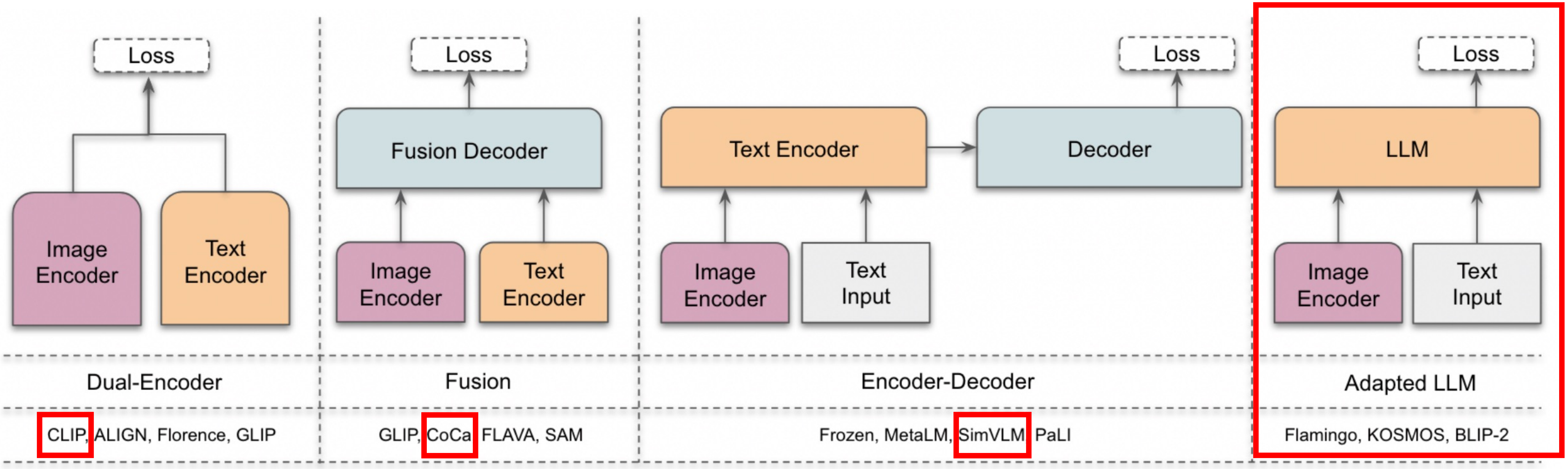
Yu et al., “Coca: Contrastive captioners are image-text foundation models”, 2022

Architecture of Multimodal Models



Li et al., "Multimodal Foundation Models: From Specialists to General-Purpose Assistants", 2023

Architecture of Multimodal Models



Li et al., “Multimodal Foundation Models: From Specialists to General-Purpose Assistants”, 2023

Conclusion

- VLMs bridge the vision and language spaces
- VLMs showcase impressive capabilities for zero-shot adaptation to unseen tasks
- However, they are still restricted to tasks in a pre-defined form, struggling to match the open-ended task capabilities of LLMs
- A unified generalist framework is required that will be discussed in the next session

Questions