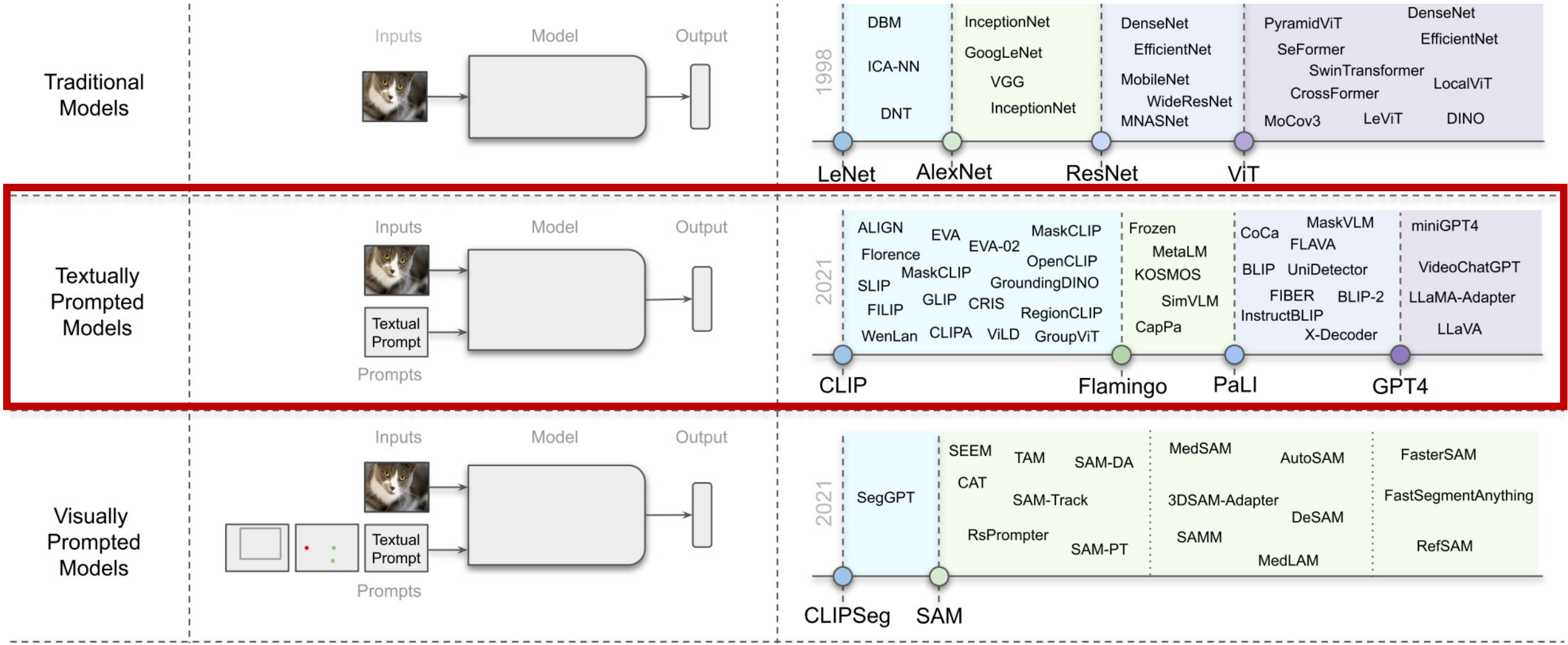
Large Language Models

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

M. Soleymani
Sharif University of Technology
Fall 2023

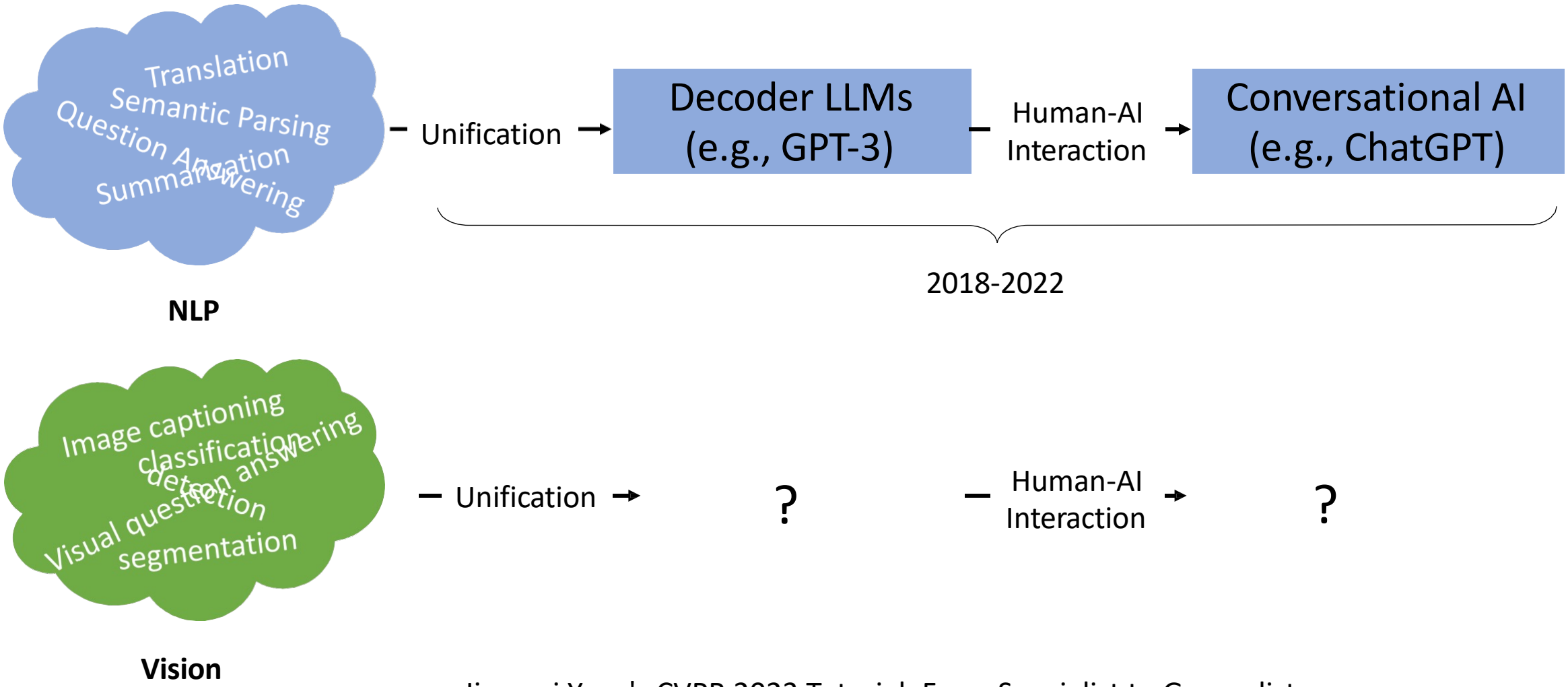
Multi-modal data

- 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)

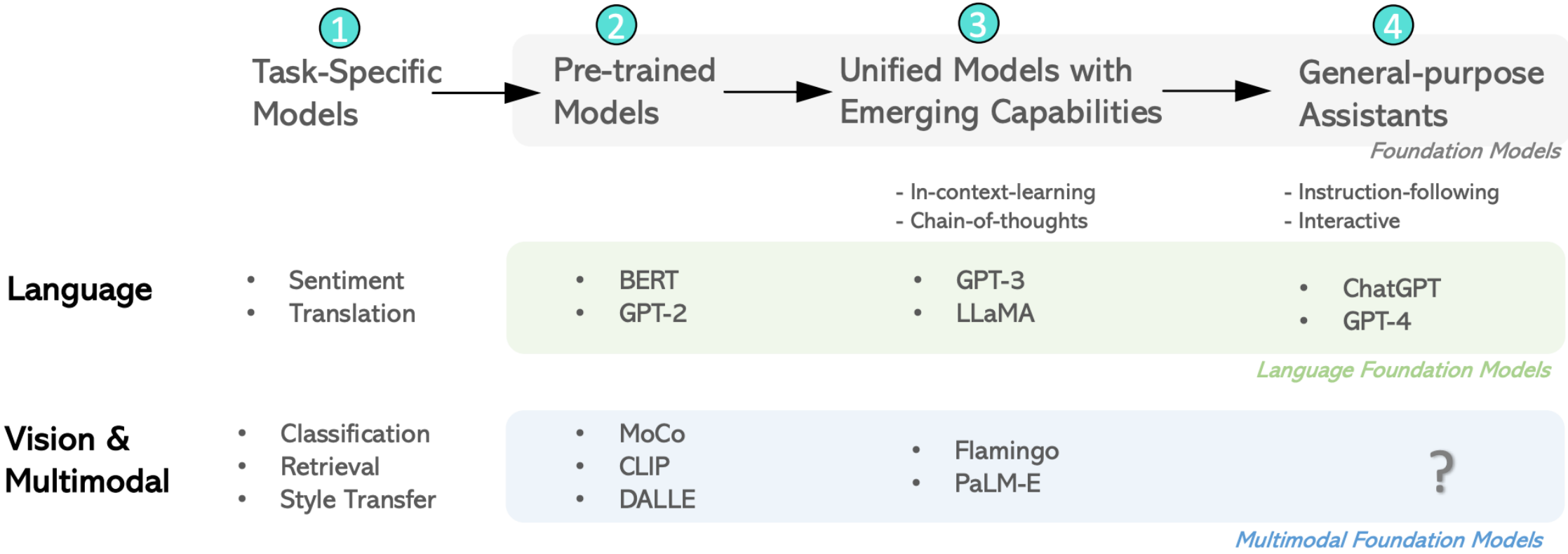


Awais et al., “Foundational Models Defining a New Era in Vision: A Survey and Outlook”, 2023

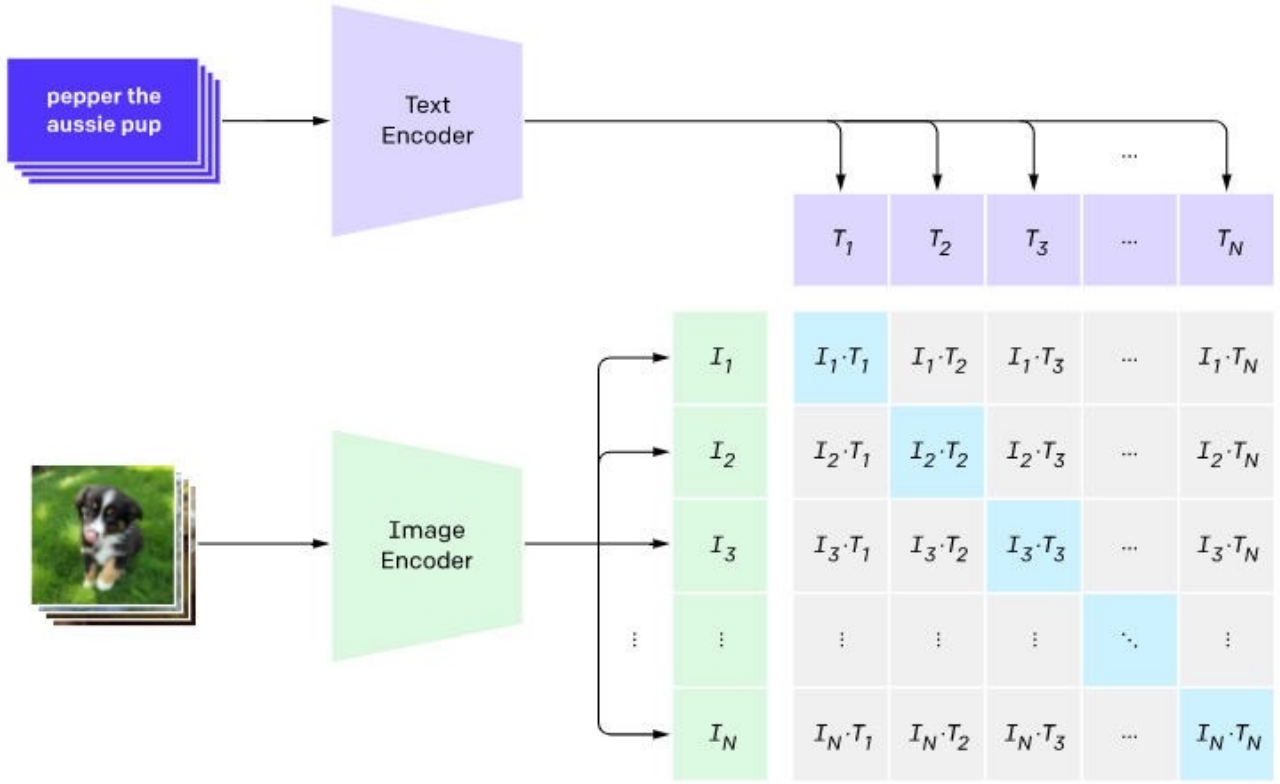
A Lesson from LLMs



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

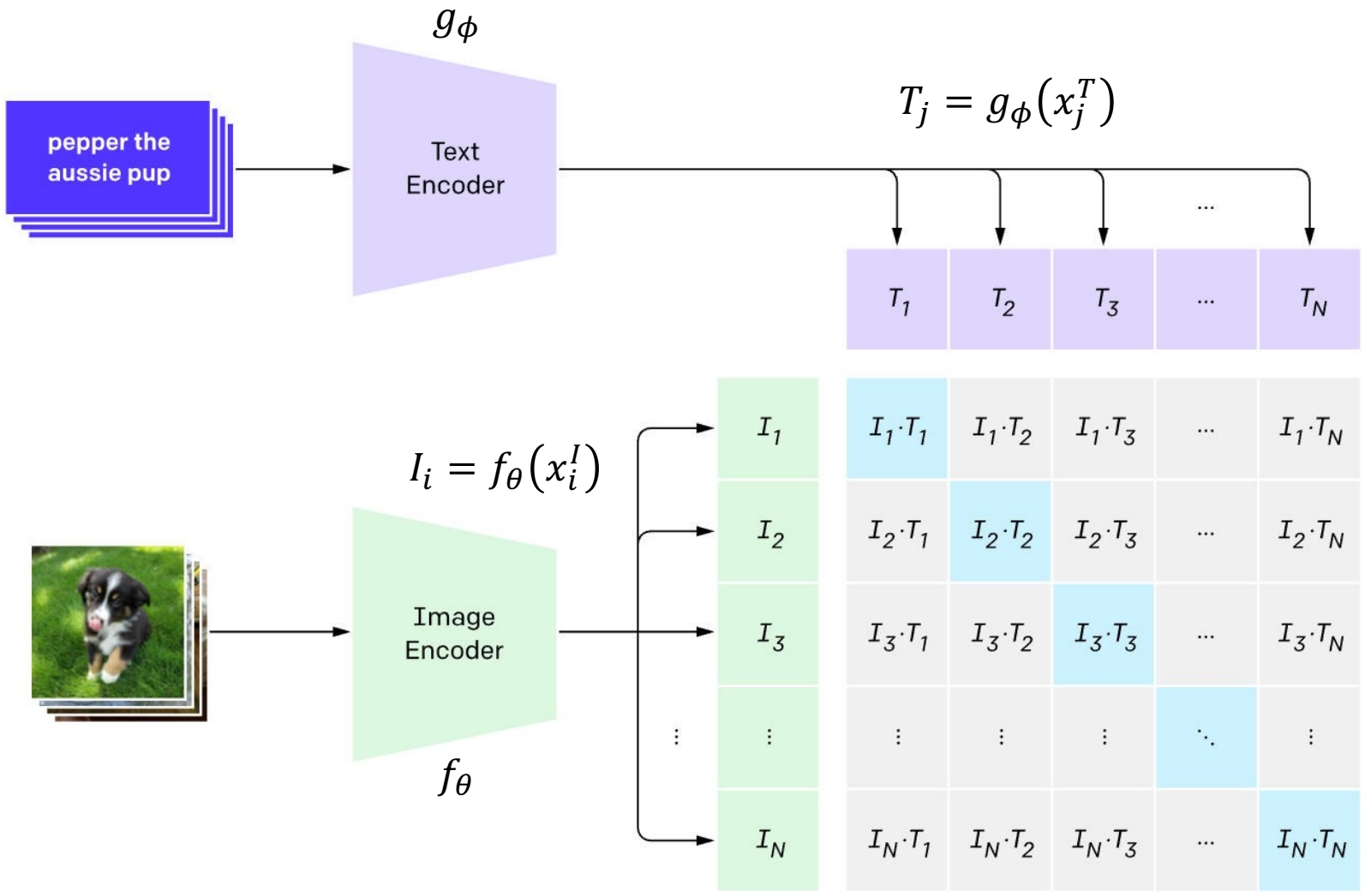
Radford et al. "Learning Transferable Visual Models From Natural Language Supervision", ICML 2021

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

1. Contrastive pre-training



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

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

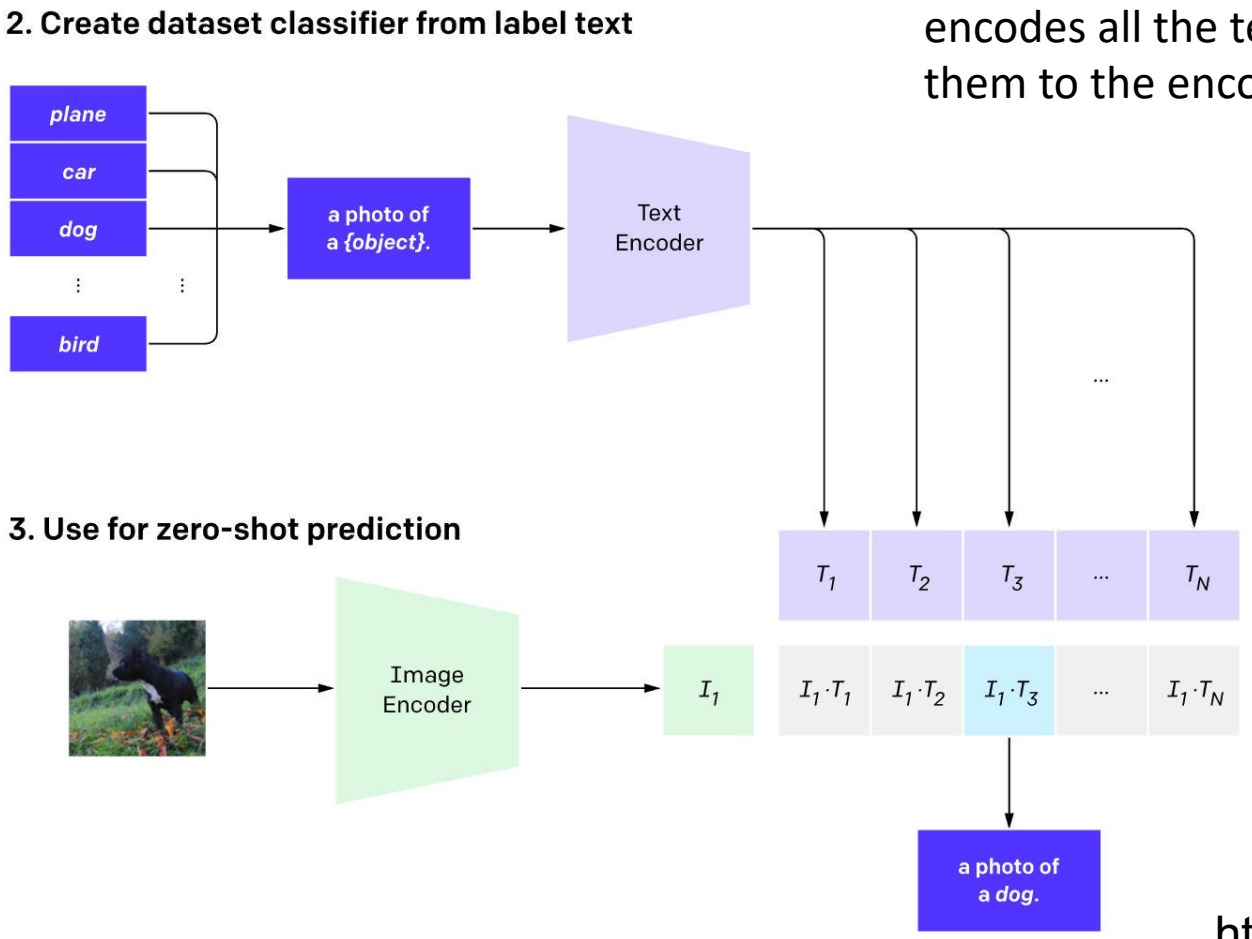
$$\mathcal{L}_j^T = -\log \frac{e^{s_{i,i}^T}}{\sum_{i=1}^N e^{s_{i,j}^T}}$$

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

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

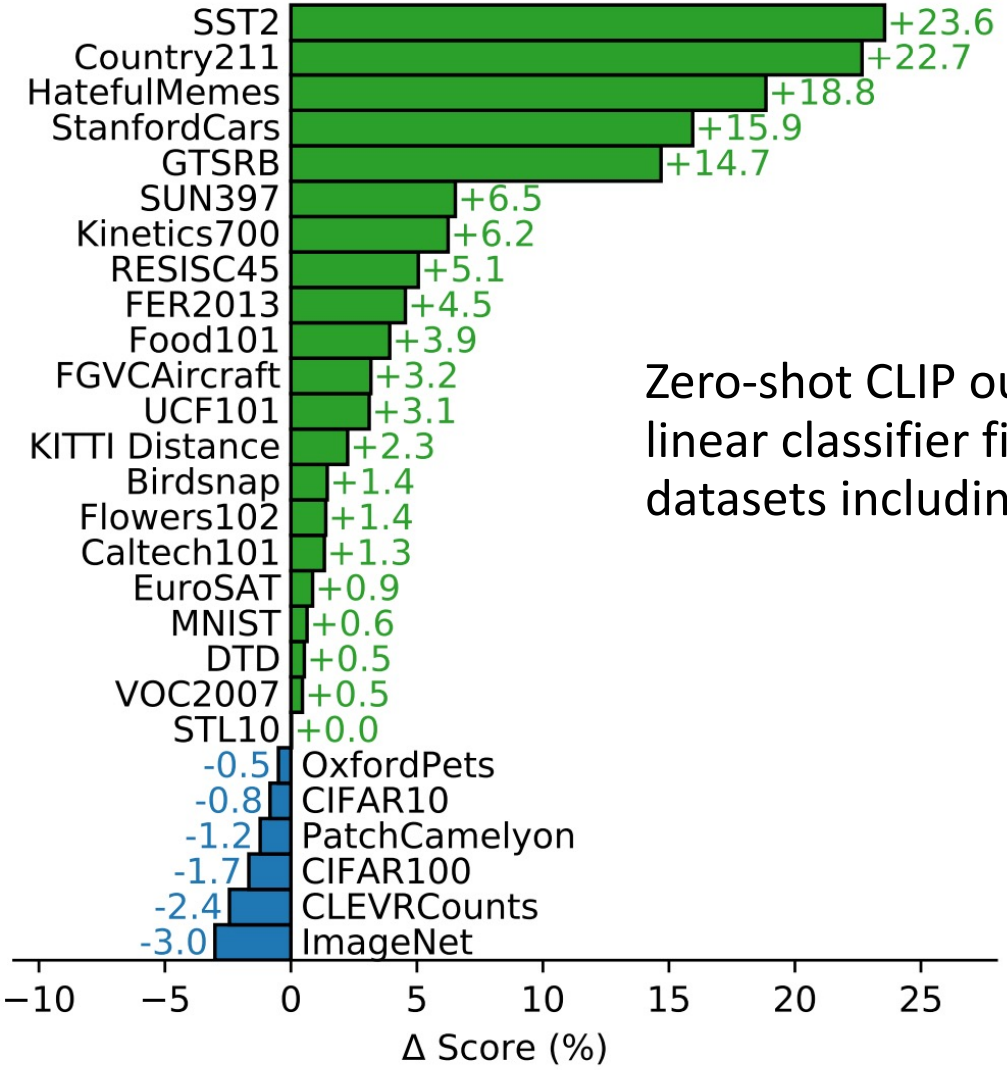
https://openai.com/blog/clip/

CLIP for zero-shot learning



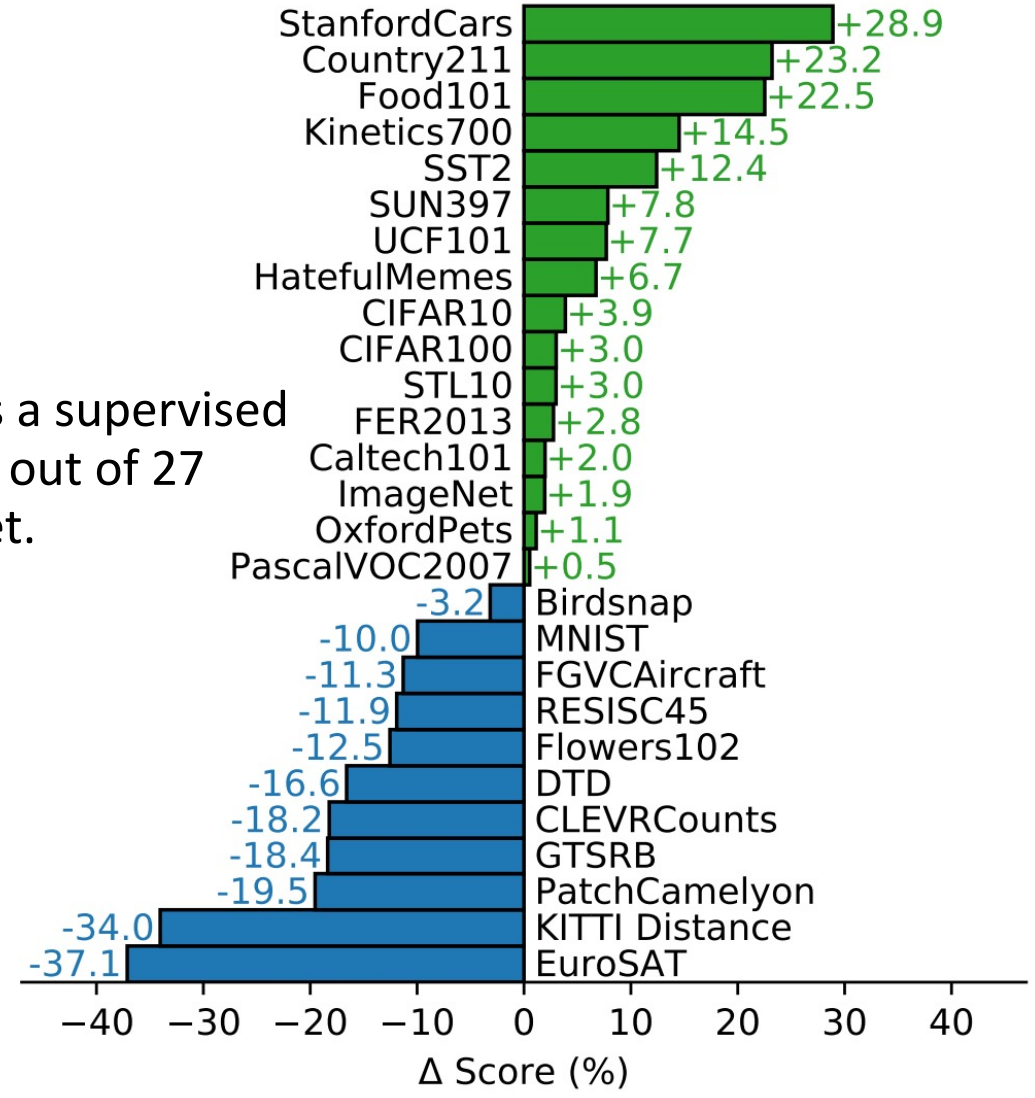
encodes all the text labels and compares them to the encoded image

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.

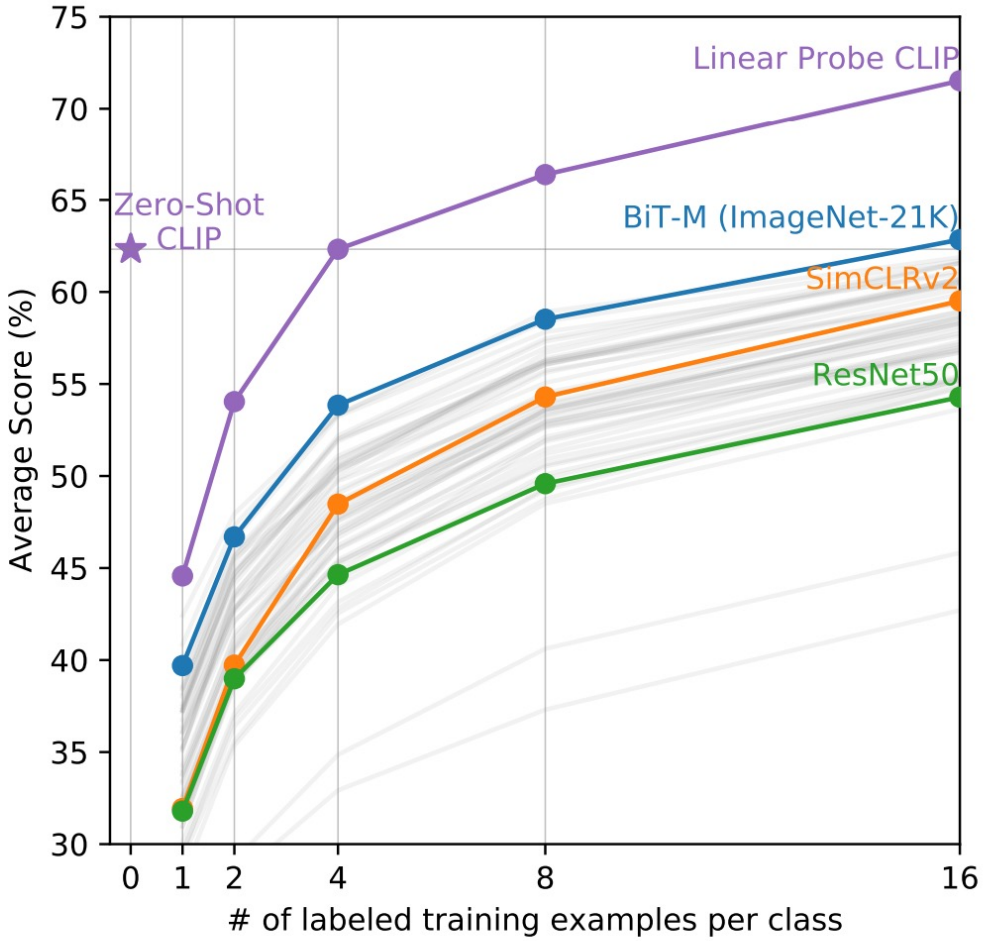


Zero-shot CLIP is competitive with a fully supervised linear-probing ResNet50

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

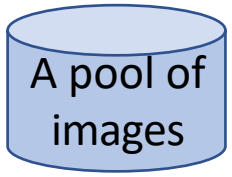
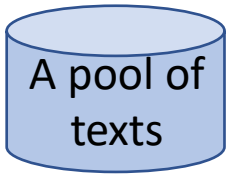
Zero-shot CLIP outperforms few-shot linear probes



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. 	Query:  	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

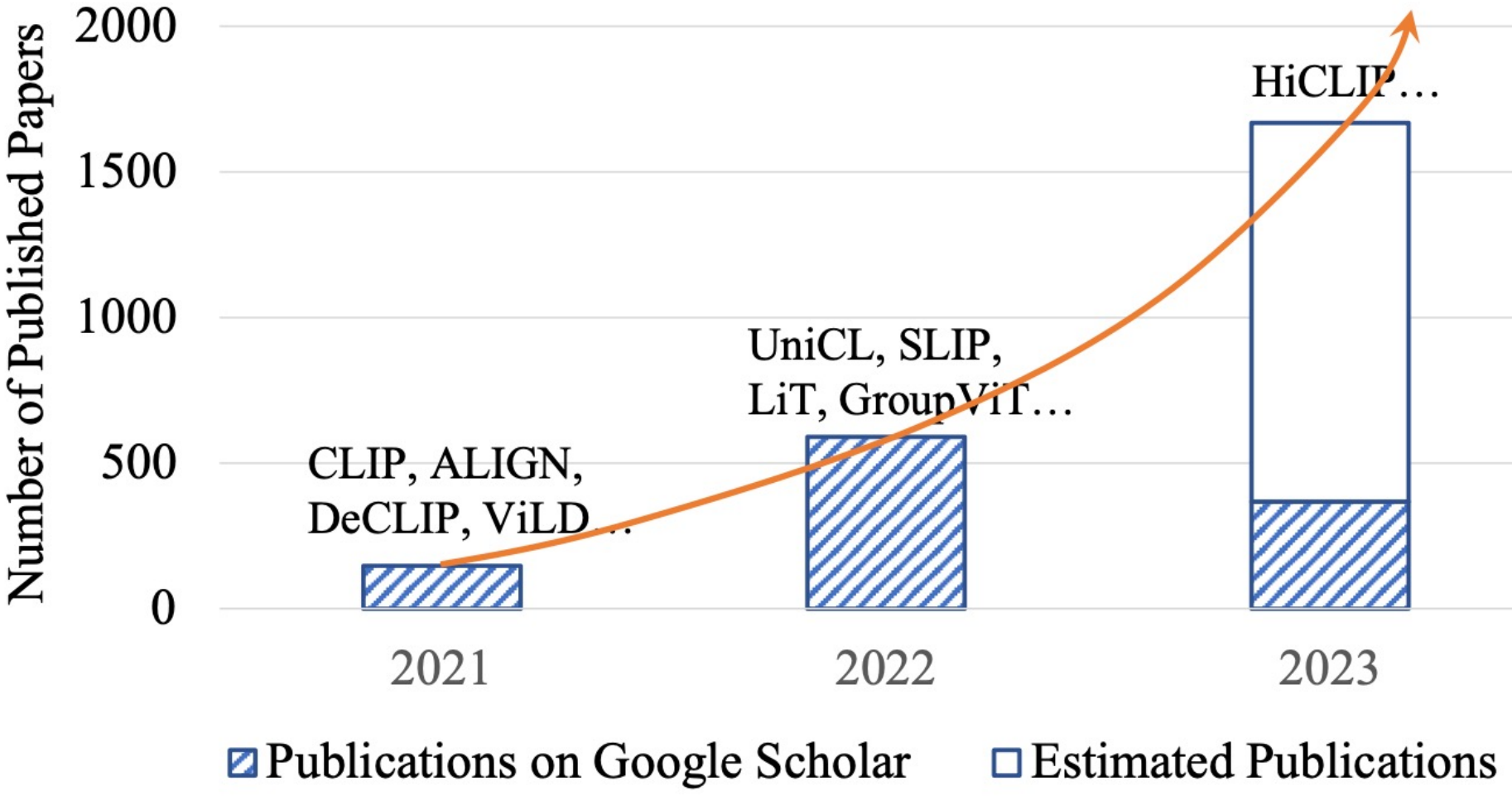
✓CLIP improved open-vocabulary visual recognition capabilities through learning from Internet-scale image-text pairs.

✗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

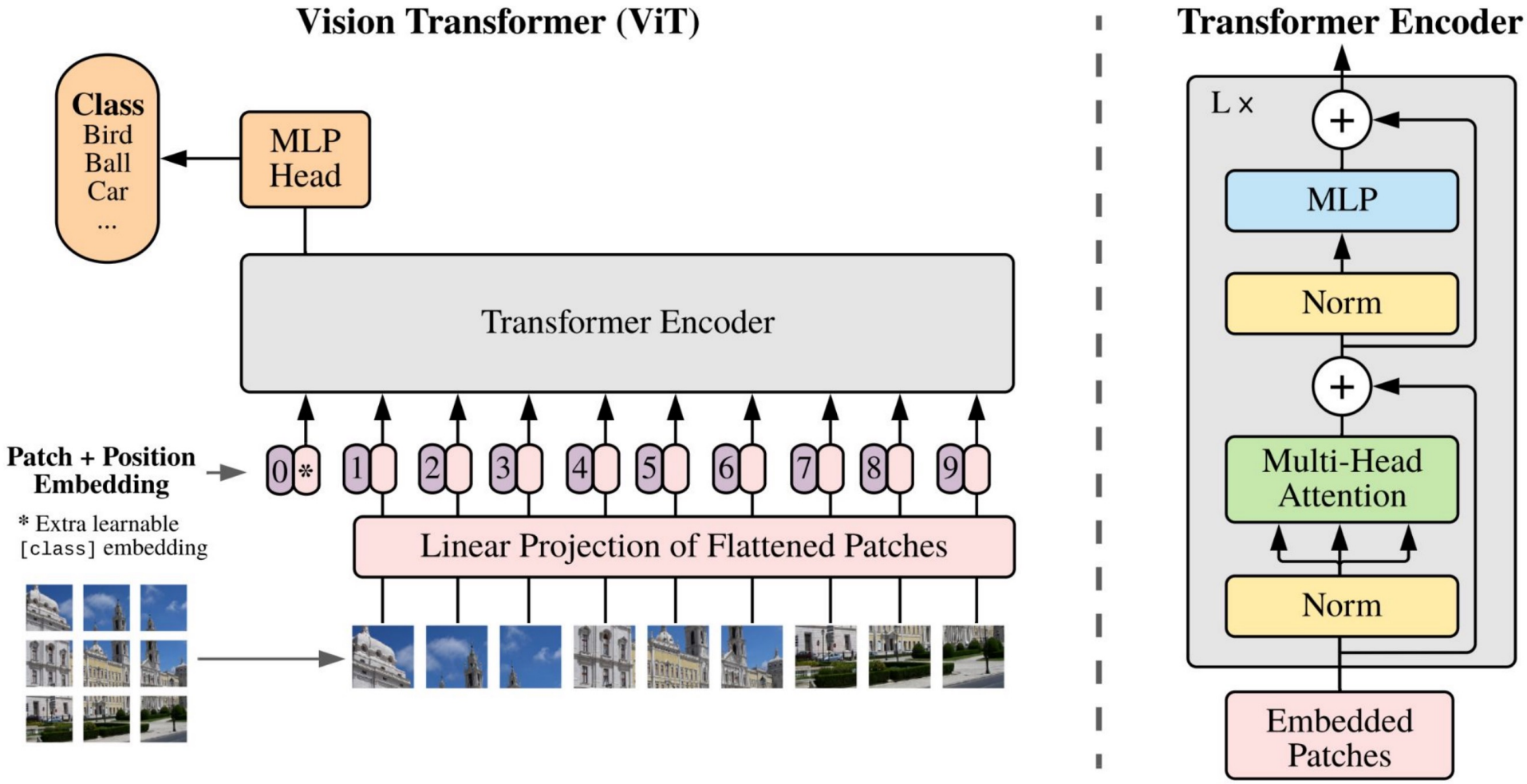


Zhang et al., “Vision-Language Models for Vision Tasks: A Survey”, 2023

CLIP Variants

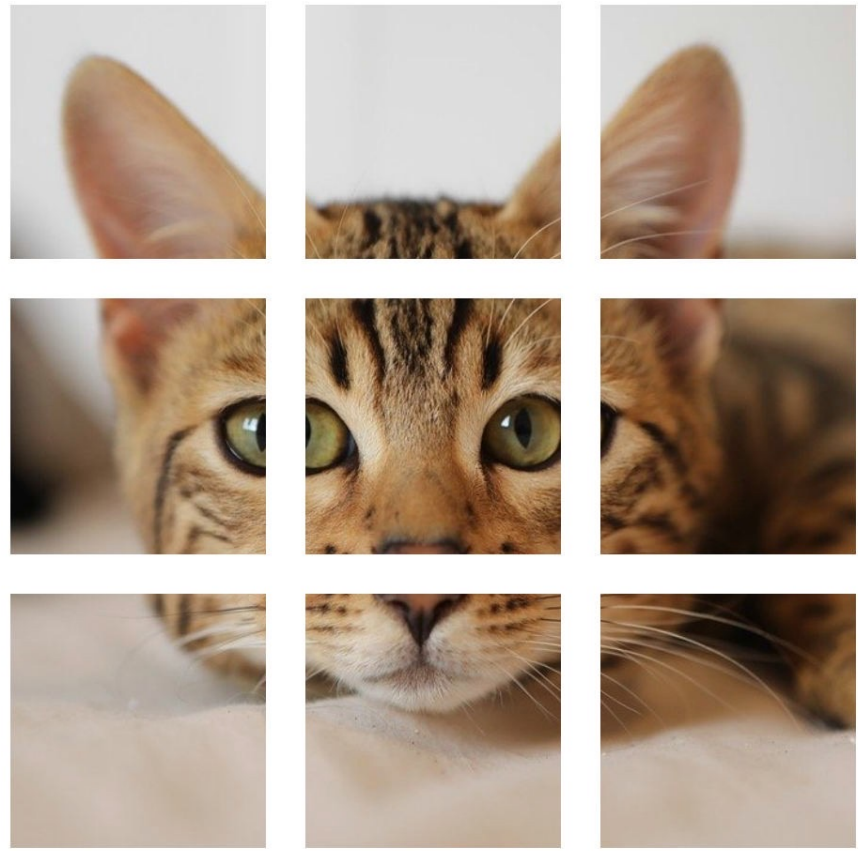
- Objective function or pretraining
  - Combining CLIP with label supervision (BASIC, UniCL, LiT, MOFI)
  - Contrastive + self-supervised image representation learning
    - Contrastive + Self-supervlised methods like SimCLR (SLIP, DeCLIP, nCLIP)
    - Contrastve + Masked Image Modeling (EVA, EVA-02,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

Vision Transfomer as Image Encoder Architecture

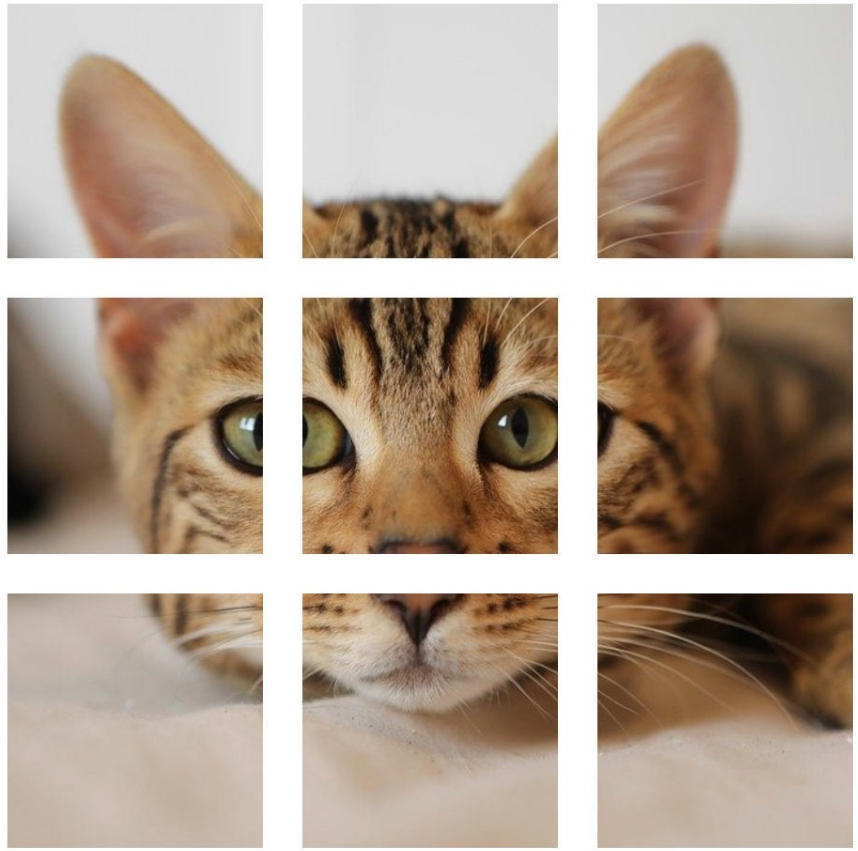


Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale", ICLR 2021

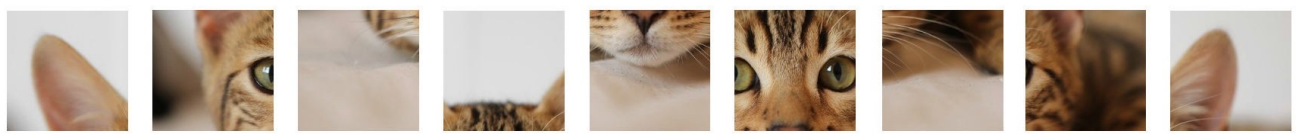
Slides from Justin Johnson



Slides from Justin Johnson



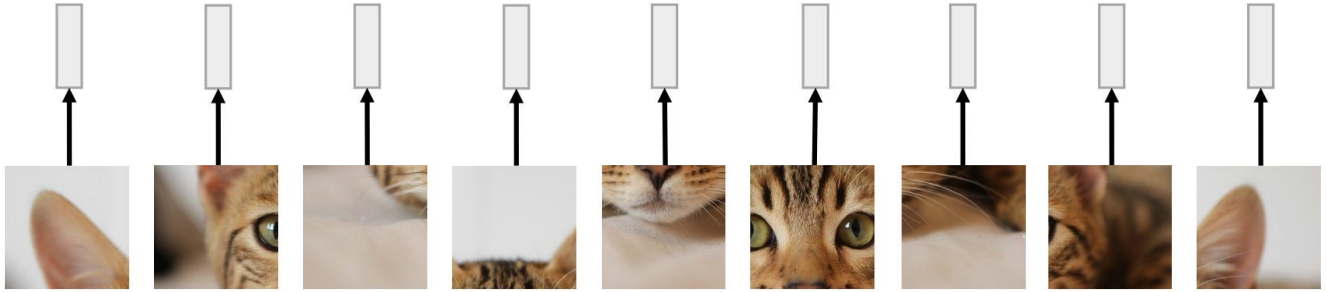
N input patches, each of shape 3x16x16



Slides from Justin Johnson

Linear projection to D-dimensional vector

N input patches, each of shape 3x16x16

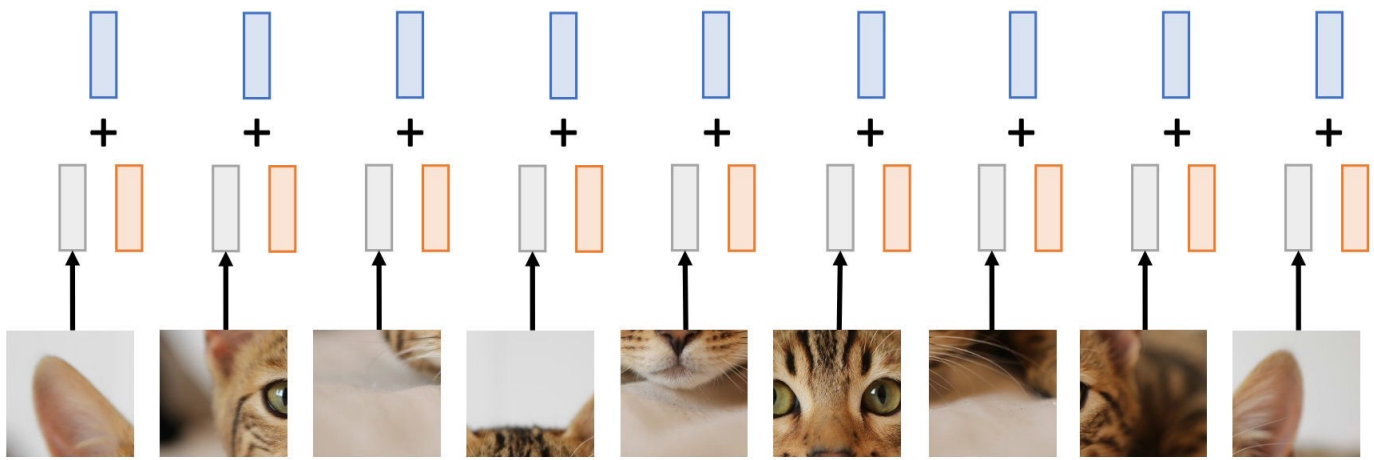


Slides from Justin Johnson

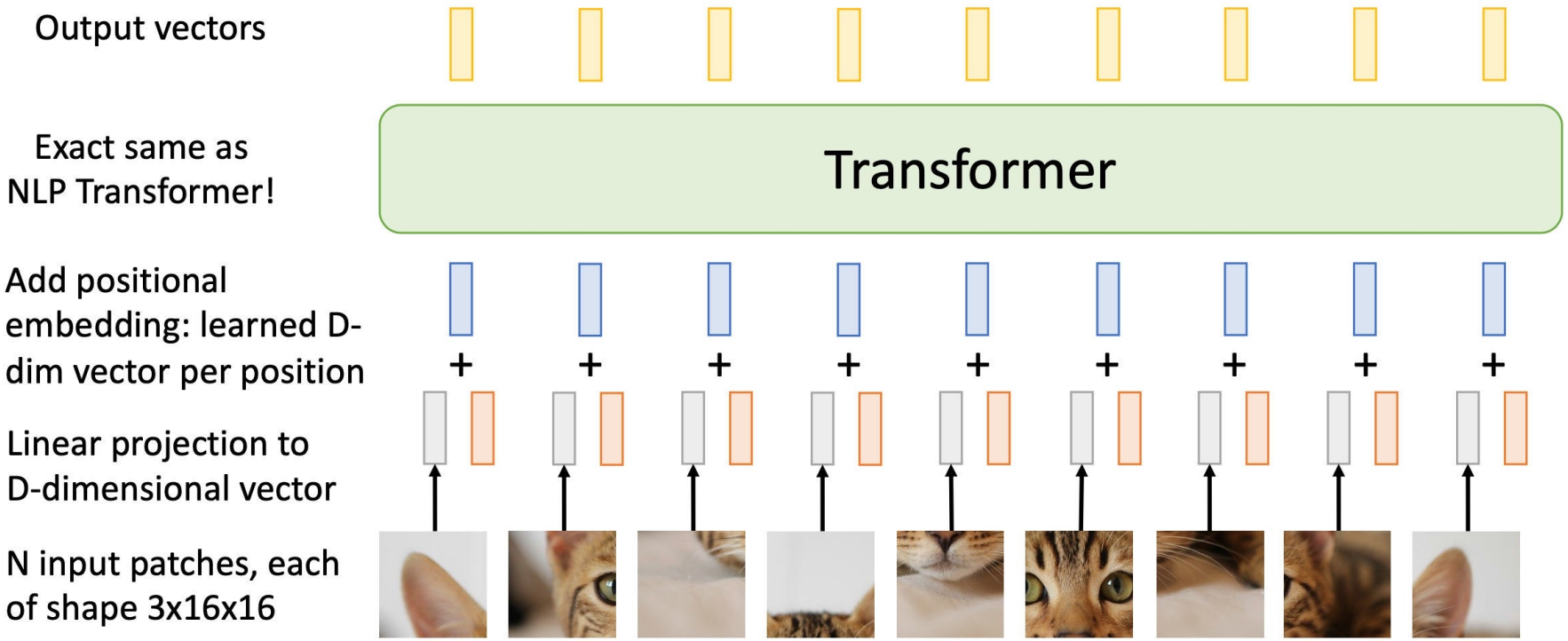
Add positional embedding: learned D-dim vector per position

Linear projection to D-dimensional vector

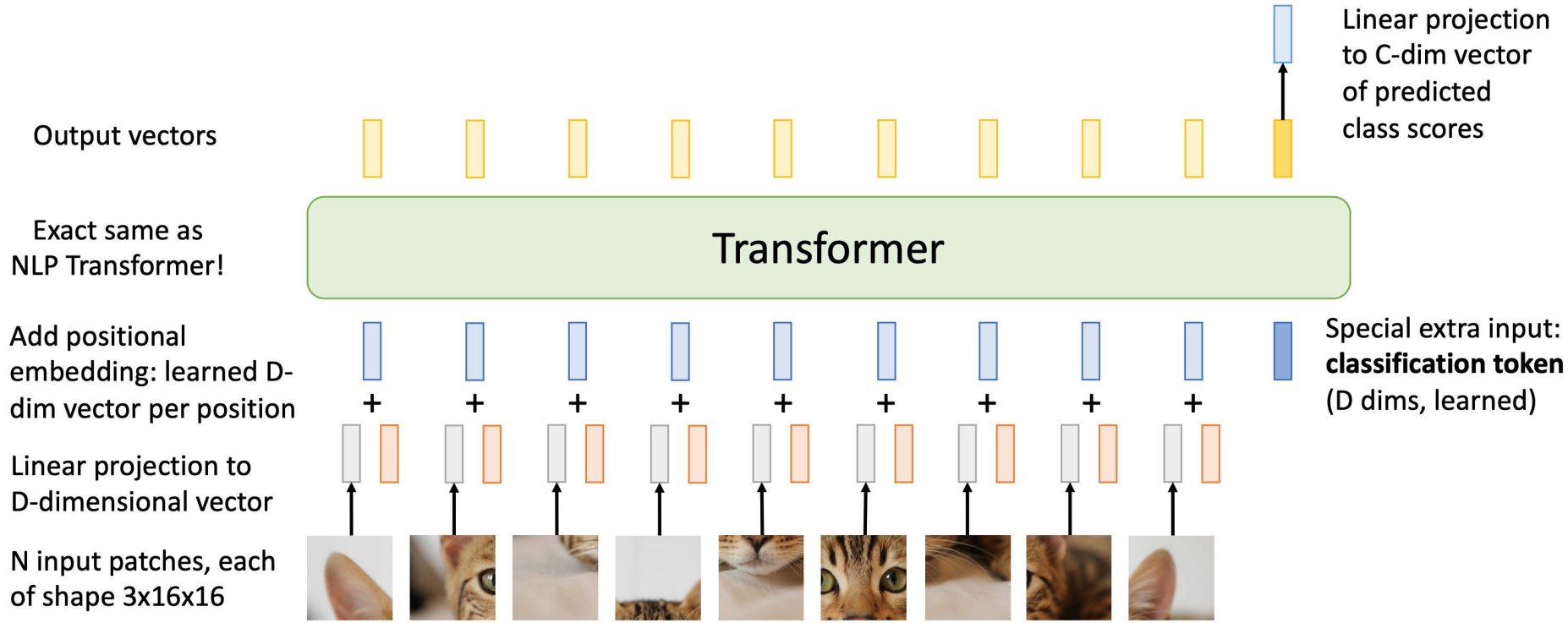
N input patches, each of shape 3x16x16



Slides from Justin Johnson



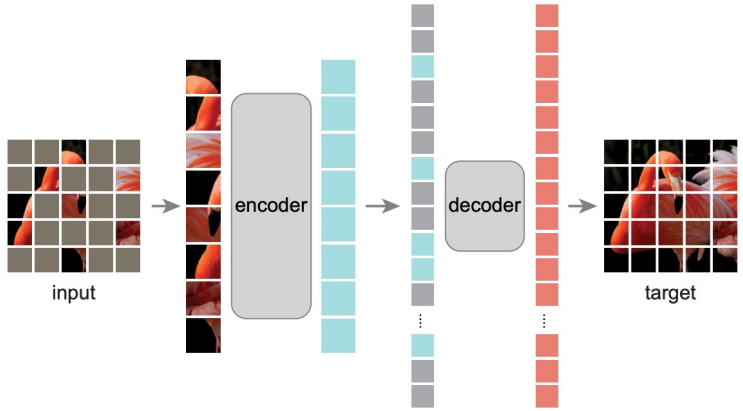
Slides from Justin Johnson



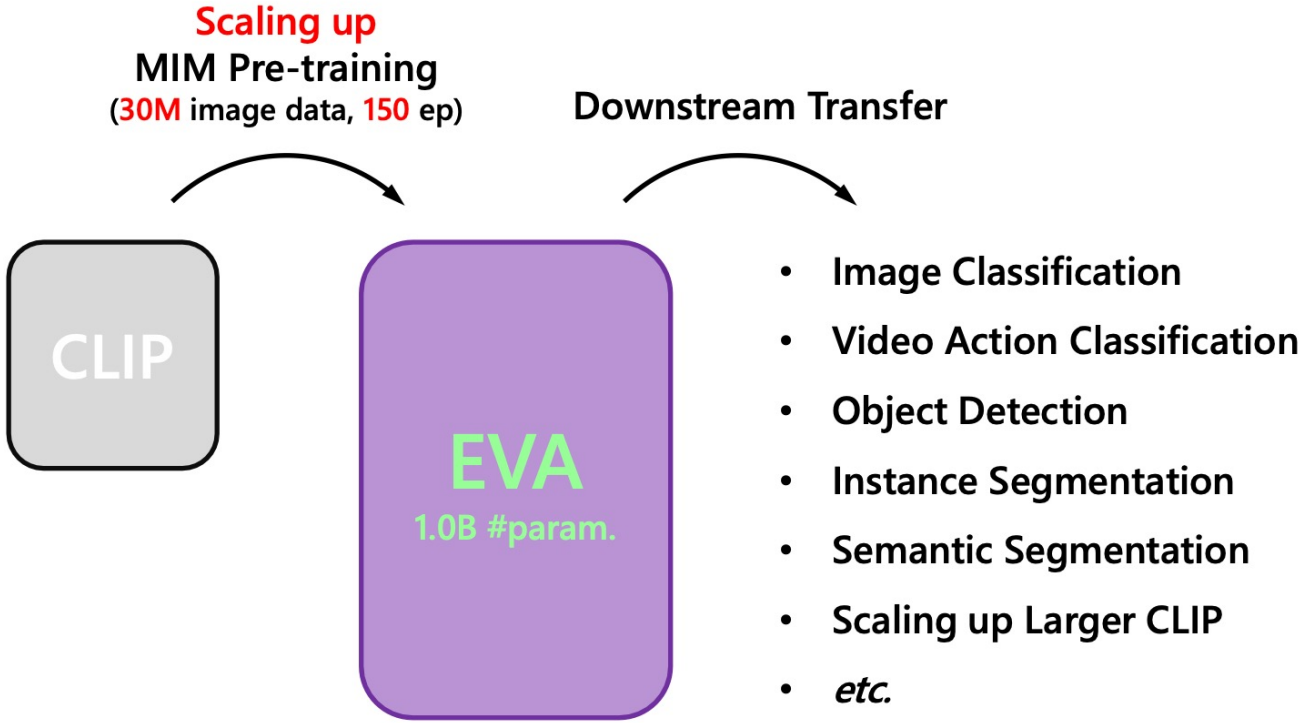
Slides from Justin Johnson

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.



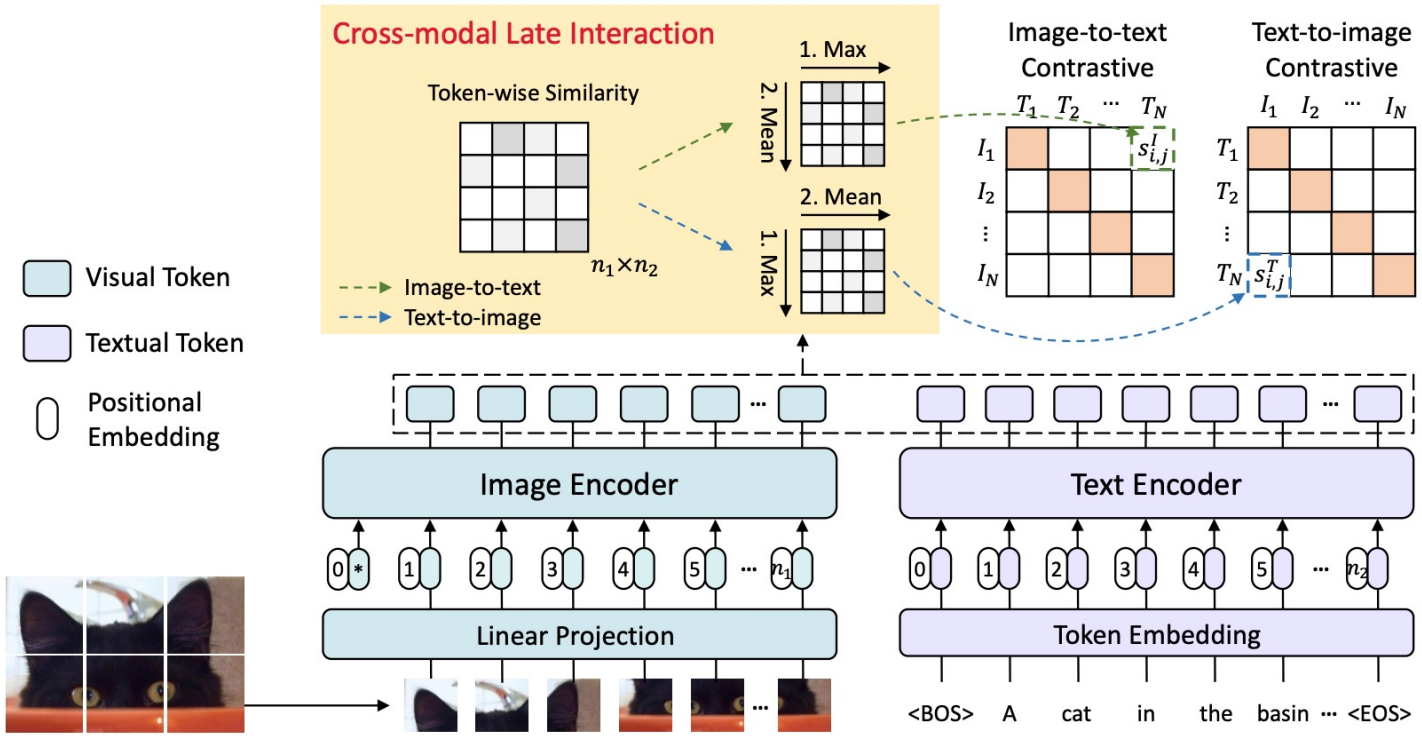
He et al., "Masked Autoencoders Are Scalable Vision Learners", 2021



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

Fine-grained alignment: FILIP

- First computing the token-wise similarity, and then aggregating the matrix by max pooling



$$s^I_{i,j} = \frac{1}{n_1}\sum_{k=1}^{n_1} \max_{l=1,\dots,n_2} I^T_{i,k} T_{j,l}$$

$$s^T_{i,j} = \frac{1}{n_2}\sum_{k=1}^{n_2} \max_{l=1,\dots,n_2} I^T_{i,l} T_{j,k}$$

$$\mathcal{L}^I_i = -\frac{1}{b}\log \frac{e^{s^I_{i,i}}}{\sum_{j=1}^{b} e^{s^I_{i,j}}}$$

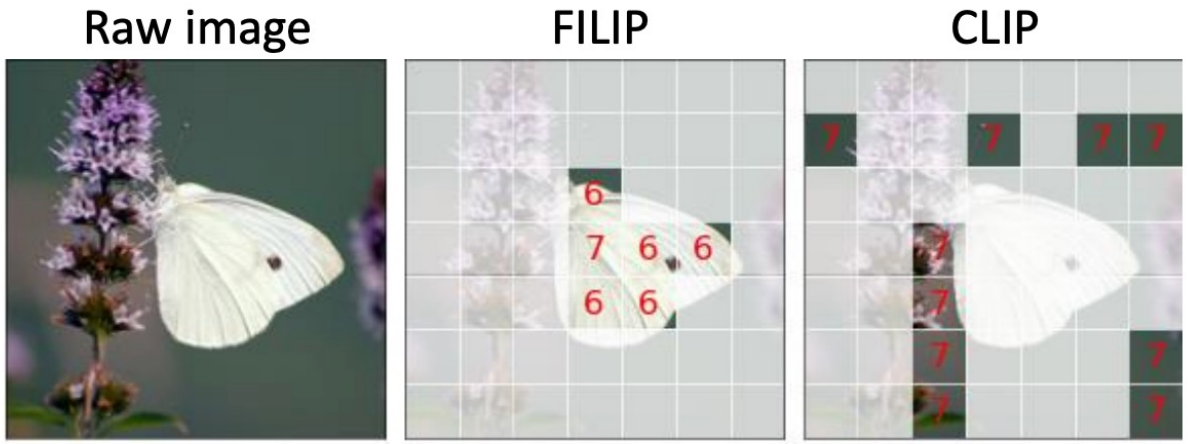
$$\mathcal{L}^T_j = -\frac{1}{b}\log \frac{e^{s^T_{i,i}}}{\sum_{i=1}^{b} e^{s^T_{i,j}}}$$

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{b} (\mathcal{L}^I_i + \mathcal{L}^T_j)$$

Yao et al., "FILIP: Fine-Grained Interactive Language Image Pre-training", ICLR 2022

FILIP vs. CLIP

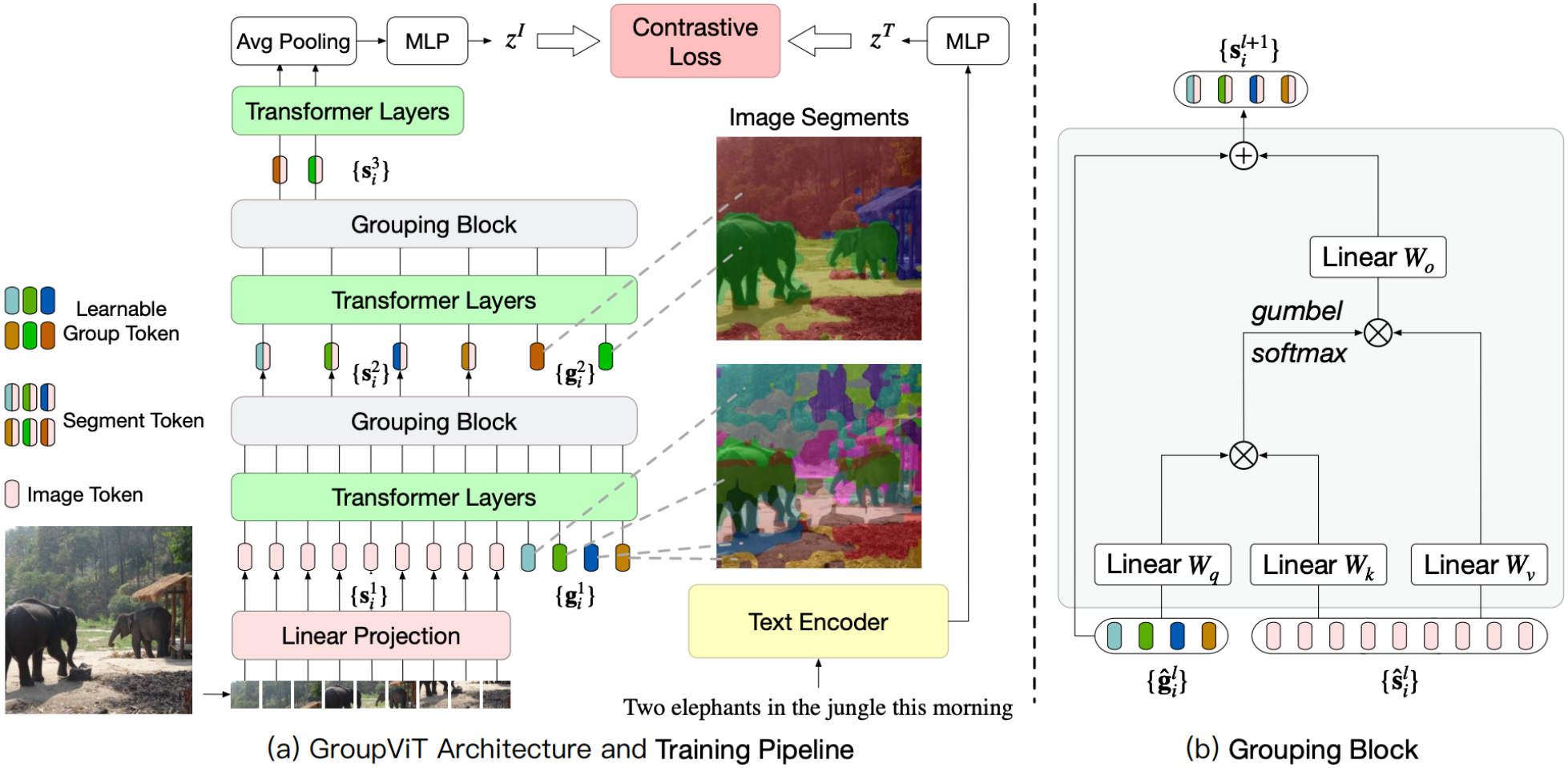
- FILIP learns word-patch alignment that is good for visualization



(c) Small white butterfly (5, 6, 7)

Yao et al., “FILIP: Fine-Grained Interactive Language Image Pre-training”, ICLR 2022

GroupVIT

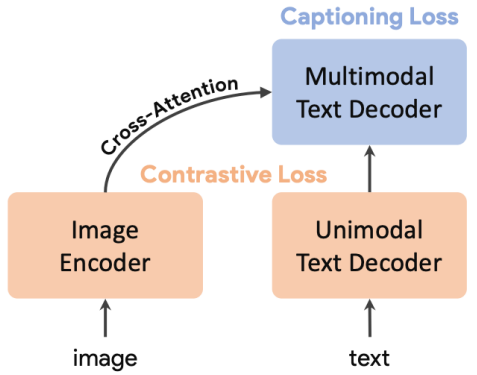


(a) GroupViT Architecture and Training Pipeline

(b) Grouping Block

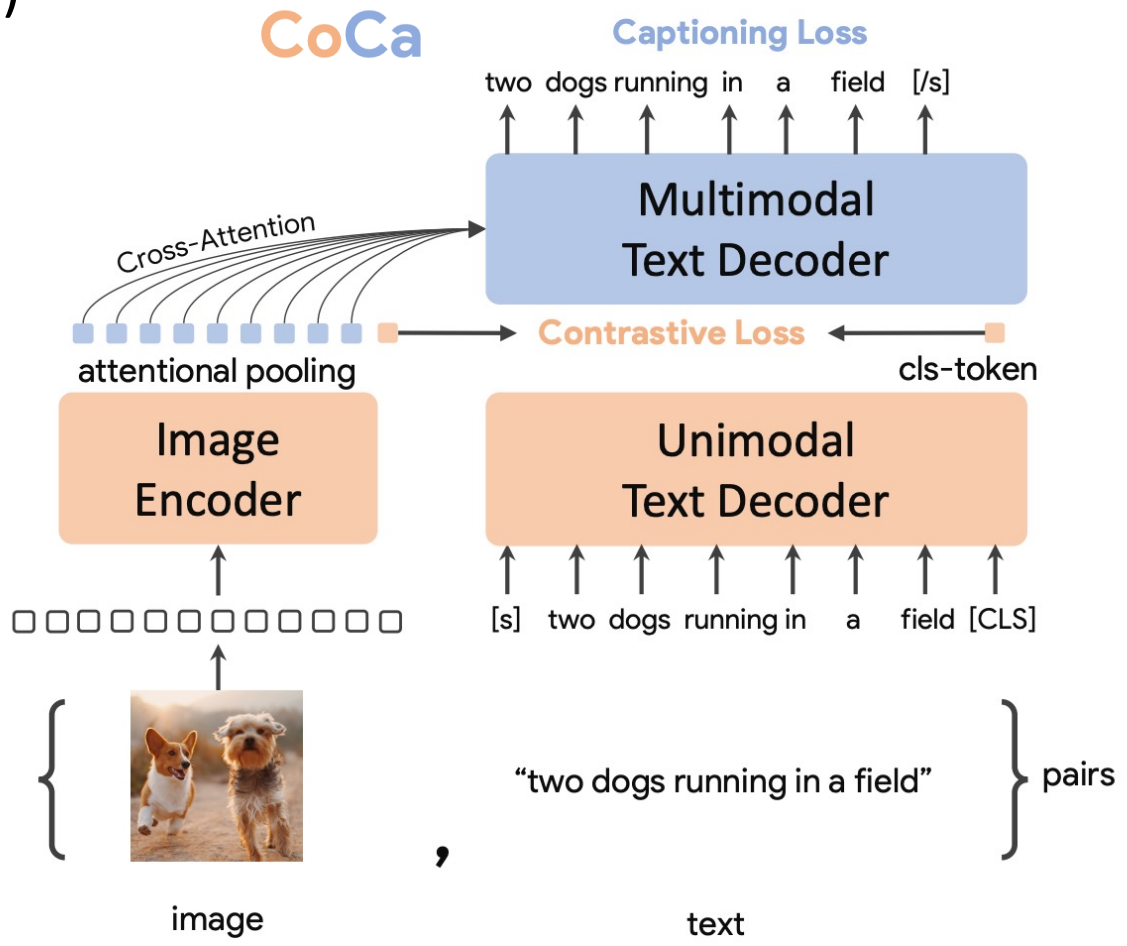
Xu et al., "GroupViT: Semantic Segmentation Emerges from Text Supervision", 2022

CoCa: Contrastive Captioner



Multi-modal fusion

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

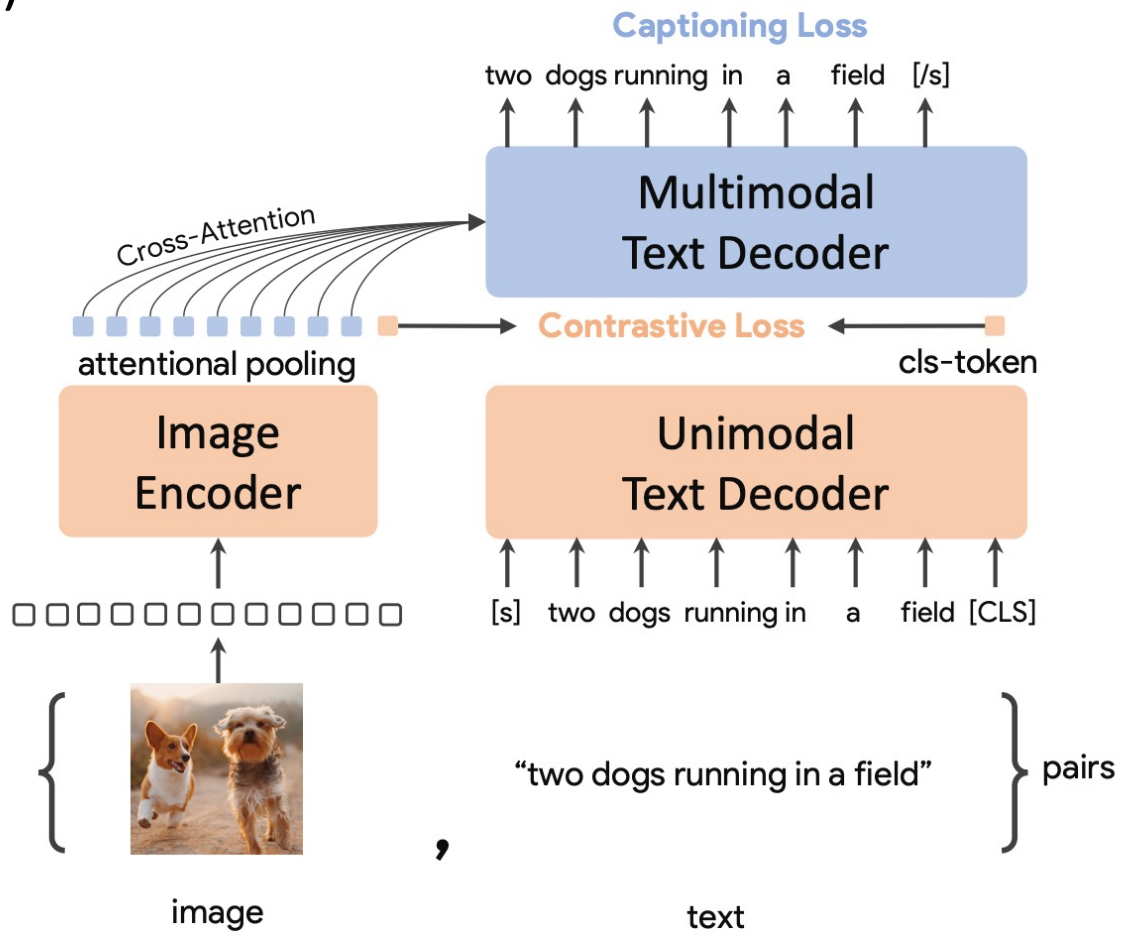


Yu et al., "Coca: Contrastive captioners are image-text foundation models", 2022

CoCa: Contrastive Captioner

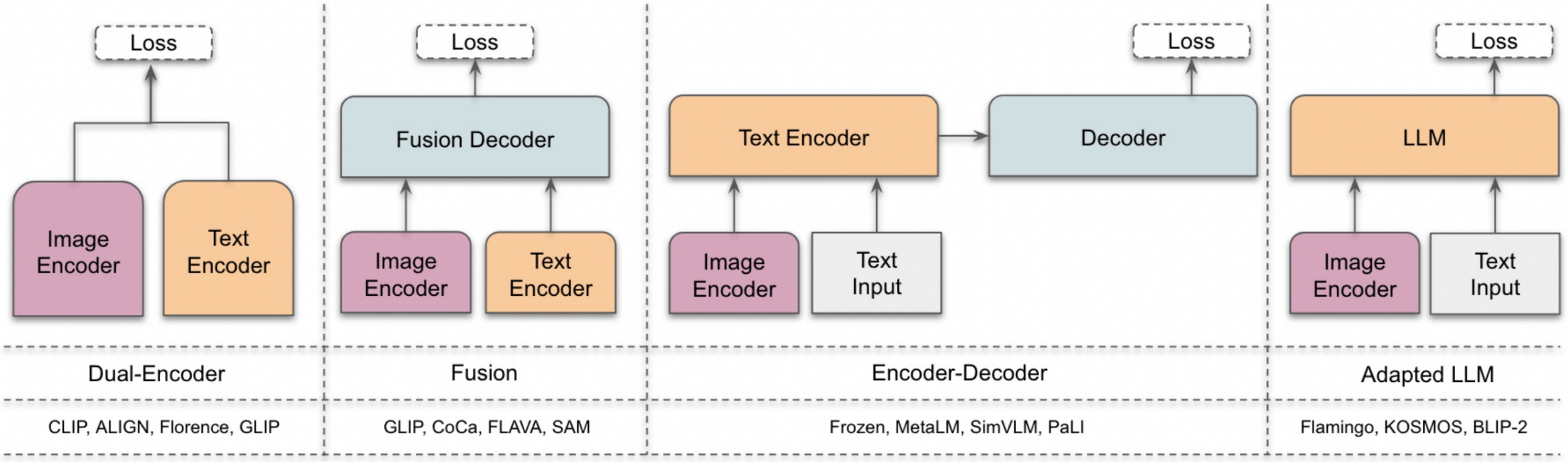
$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x).$$

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

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

Next Session: Multimodal LLMs
Large Multimodal Models (LMMs)

- A unified generalist framework that can integrate the strengths of LLMs with the specific requirements of vision-centric tasks.
  - More efficient multi-modal training
  - Multimodal In-Context Learning (M-ICL)
  - Multi-modal Instruction Tuning (M-IT)

Questions