Multimodal Deep Learning

Till Aczél

Mentor: Zhao Meng

26.04.2022

ETH Zürich - Seminar in Deep Neural Networks

Vision Transformer (ViT)

Vision Transformer (ViT)

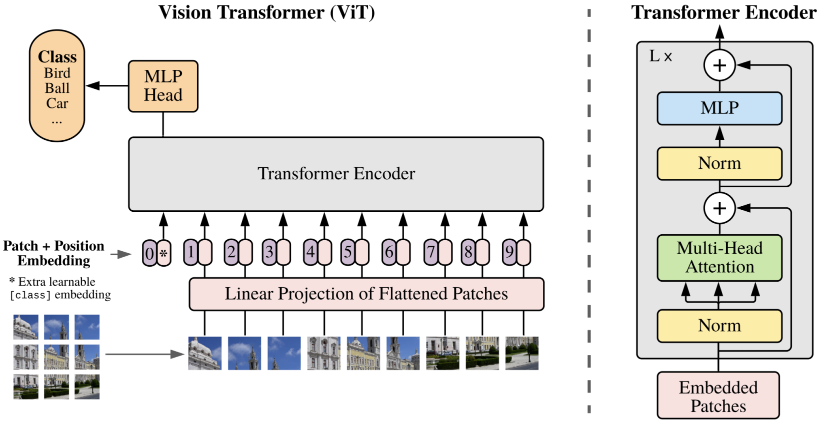


Figure taken from Dosovitskiy et al. 2020

Example

Video taken from Caron et al. 2021

Related work

Comparison:

- Do Vision Transformers See Like Convolutional Neural Networks?, Raghu et al. 2021
- Transformers in vision: A survey, Khan et al. 2021

Improving ViTs:

- Training data-efficient image transformers & distillation through attention, Touvron et al. 2021
- Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, Ze Liu et al. 2021
- Cvt: Introducing convolutions to vision transformers, Wu et al. 2021

Attention for CNNs:

- A ConvNet for the 2020s, Zhuang Liu et al. 2022

Motivation and Tasks

Motivation

- Direct user interaction
- Easier dataset collection
- Discrete categories are to strict

Tasks

- Image Captioning
- Visual Question Answering
- Natural Language for Visual Reasoning
- Image Text Retrieval

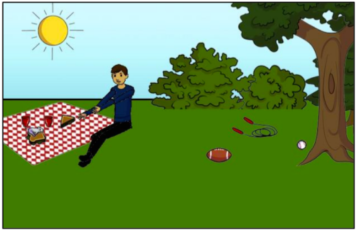
Image Captioning

Figure taken from Hodosh, Young, and Hockenmaier 2013

Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Figure taken from Antol et al. 2015

Natural Language for Visual Reasoning

The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

One image shows exactly two brown acorns in back-to-back caps on green foliage.

Figure taken from Suhr et al. 2018

Image Text Retrieval

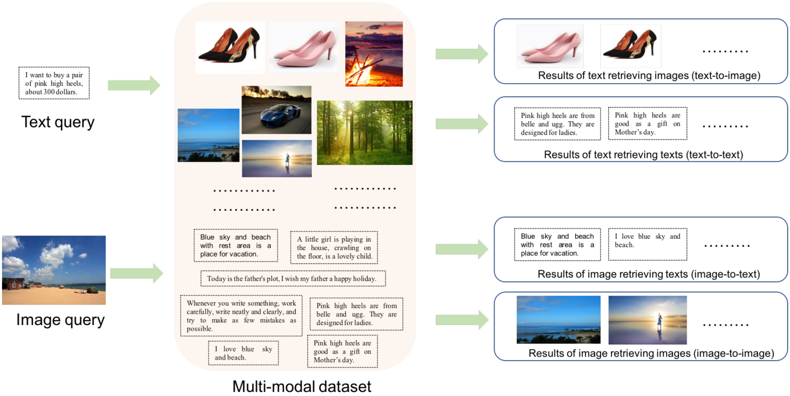


Figure taken from Hua, Yang, and Du 2020

How do multi-modal models work?

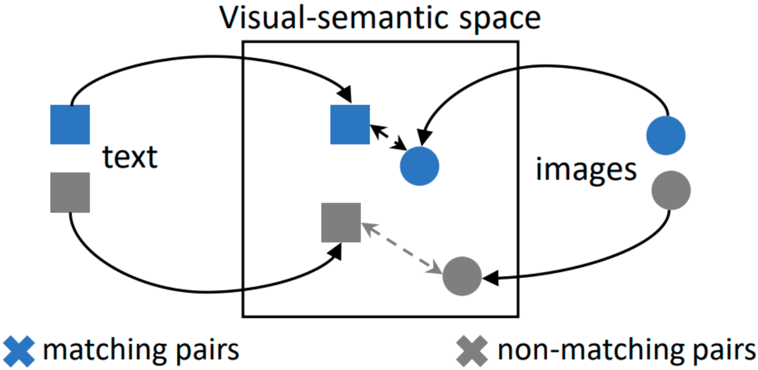


Figure taken from Cornia et al. 2018

CLIP

CLIP training

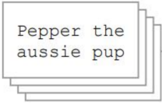


Figure taken from Radford et al. 2021

CLIP training

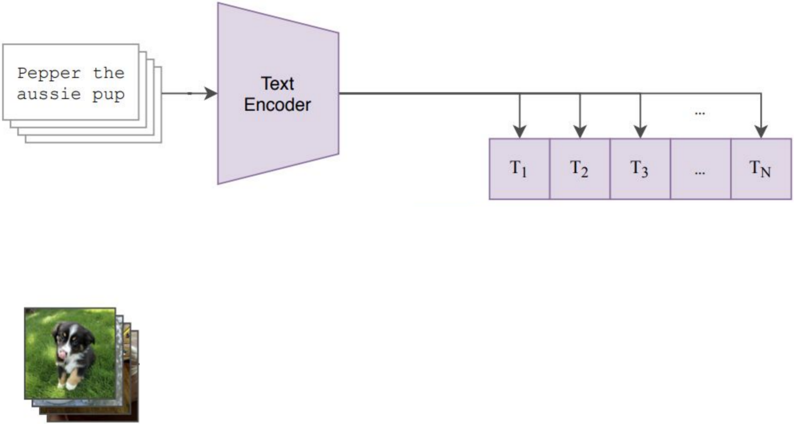


Figure taken from Radford et al. 2021

CLIP training

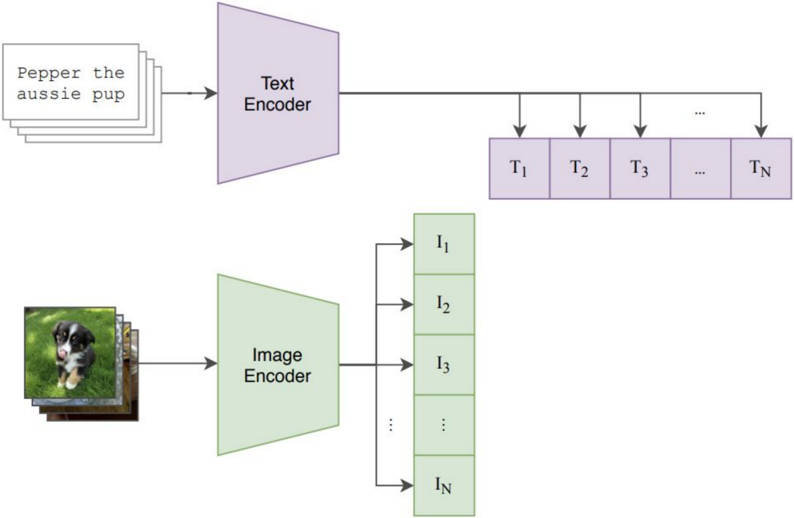


Figure taken from Radford et al. 2021

CLIP training

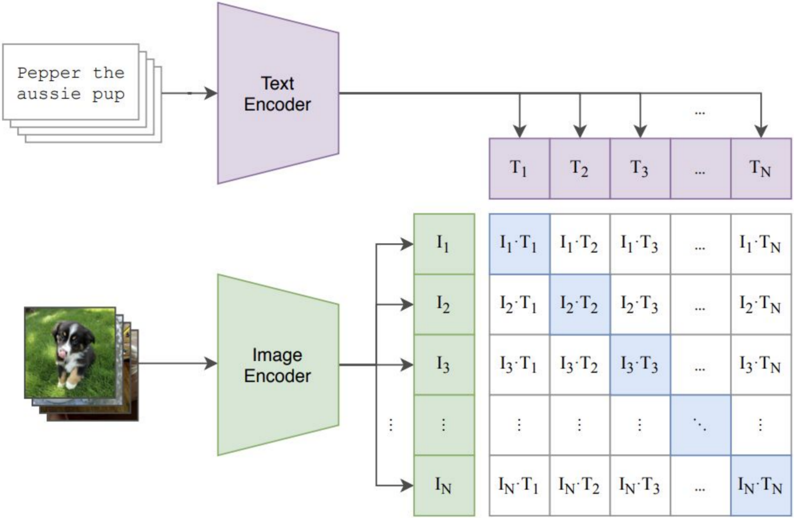


Figure taken from Radford et al. 2021

CLIP training

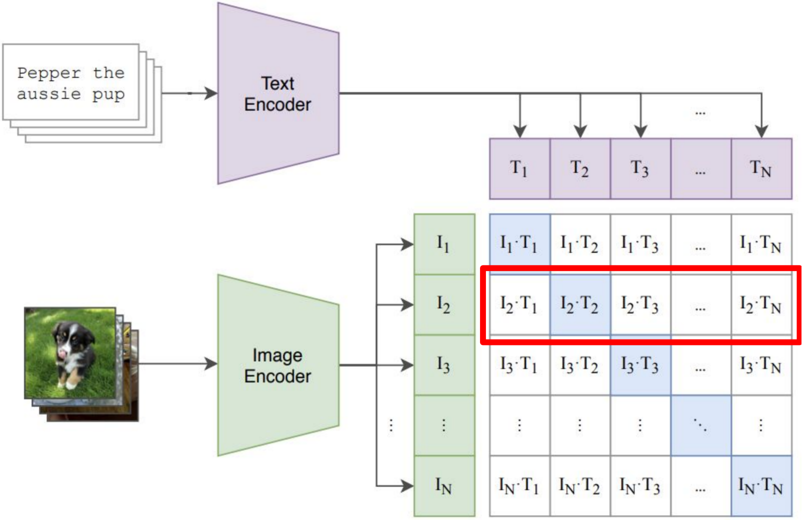


Figure taken from Radford et al. 2021

CLIP training

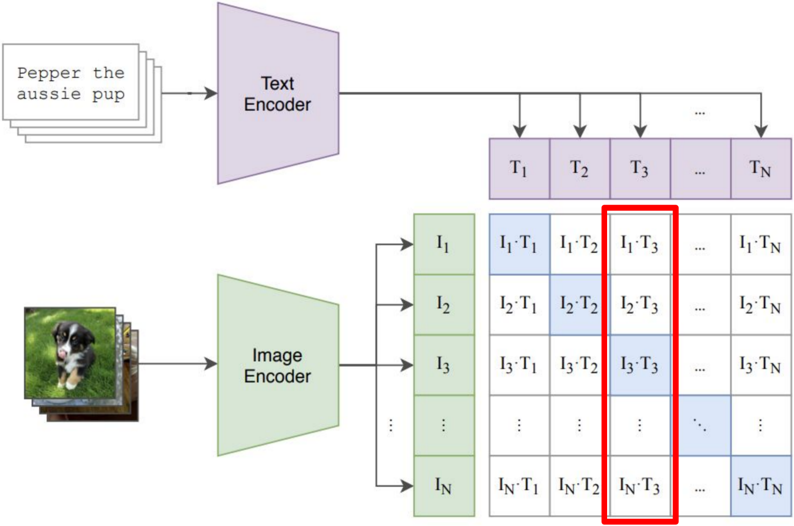


Figure taken from Radford et al. 2021

CLIP training

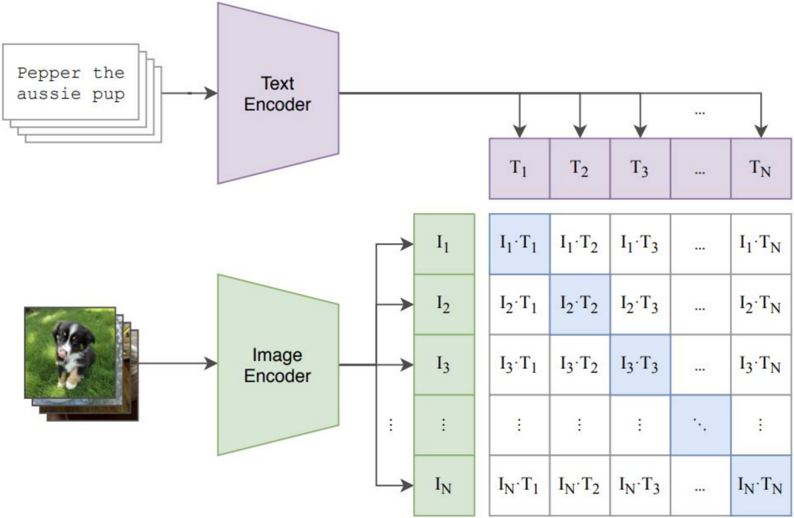


Figure taken from Radford et al. 2021

CLIP inference

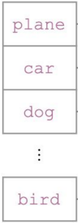


Figure taken from Radford et al. 2021

CLIP inference

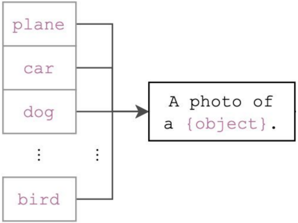


Figure taken from Radford et al. 2021

CLIP inference

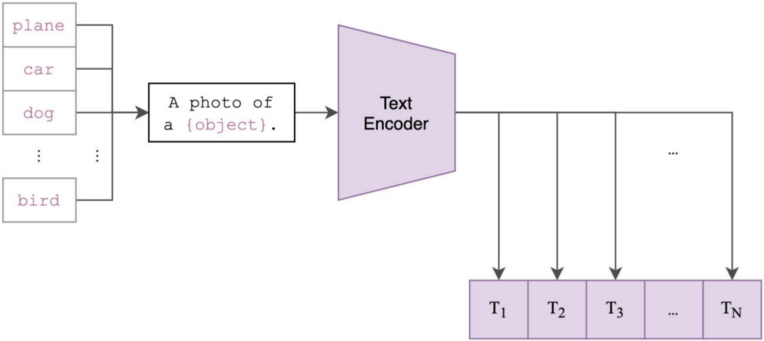


Figure taken from Radford et al. 2021

CLIP inference

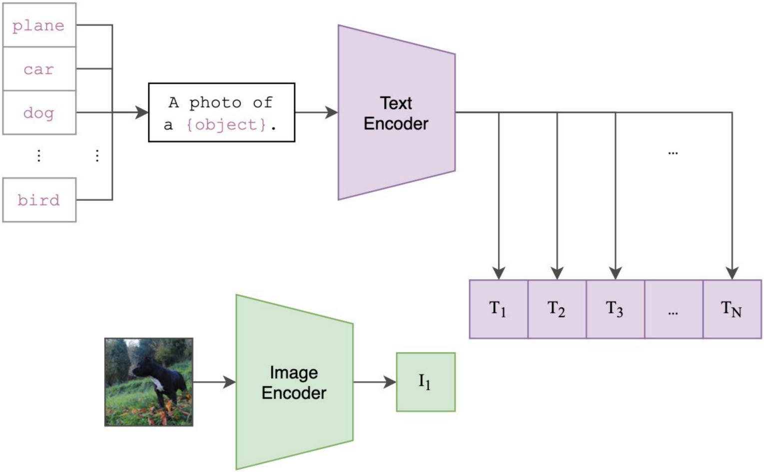


Figure taken from Radford et al. 2021

CLIP inference

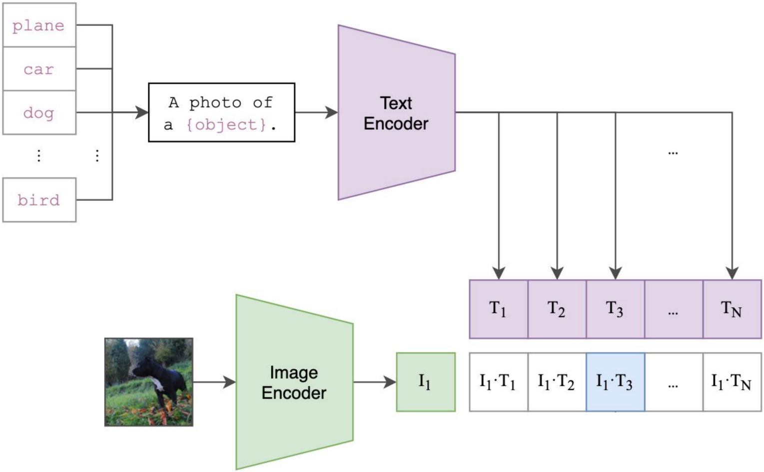


Figure taken from Radford et al. 2021

CLIP inference

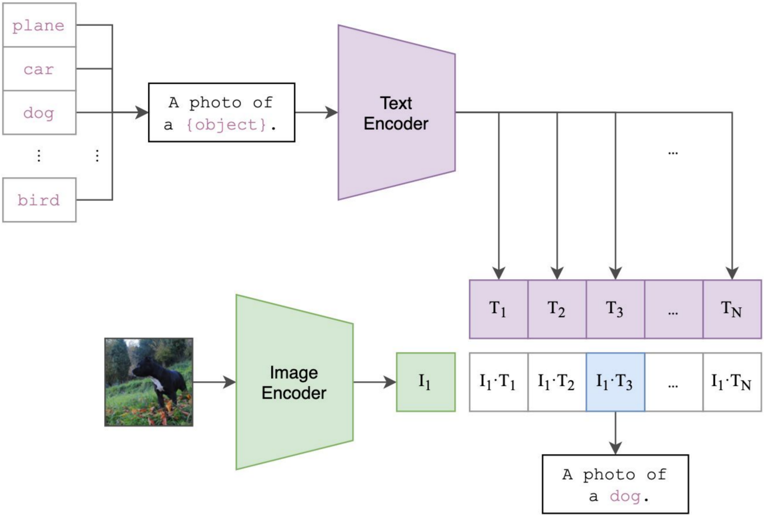


Figure taken from Radford et al. 2021

CLIP

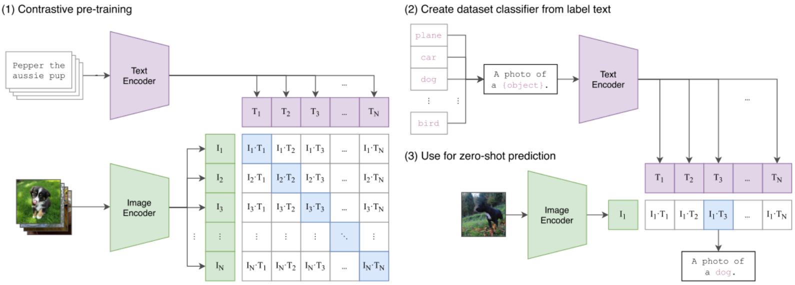


Figure taken from Radford et al. 2021

CLIP robustness

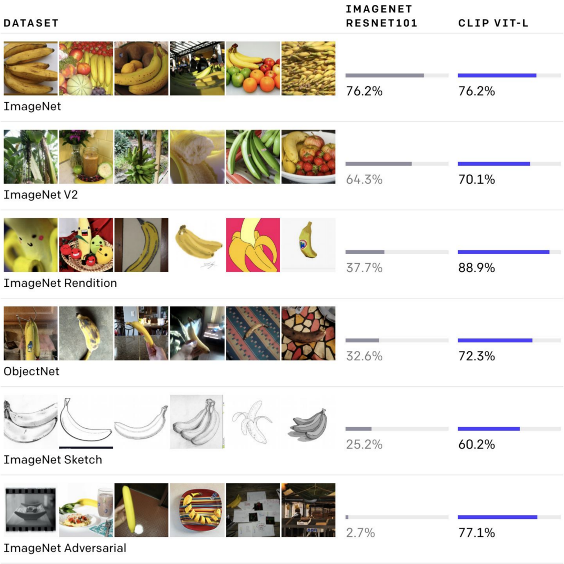
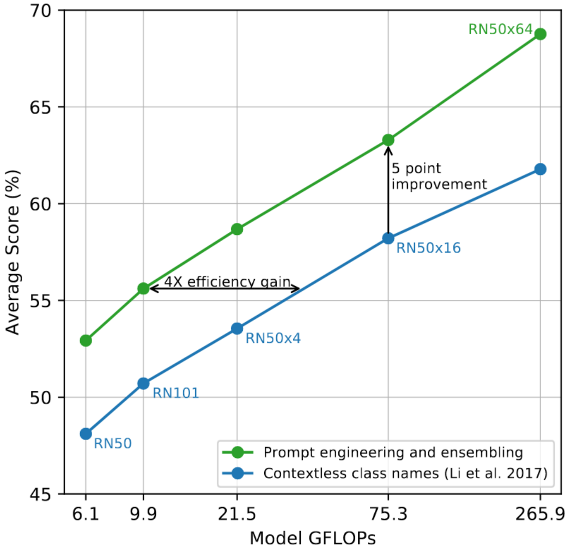


Figure taken from Radford et al. 2021

CLIP[1] prompt engineering



[1] Radford et al. 2021.

CLIP limitations

- Zero-shot performance is well below the SOTA
- Especially weak on abstract tasks such as counting
- Poor on out-of-distribution data such as MNIST
- Susceptible to adversarial attacks

CLIP typographic attacks

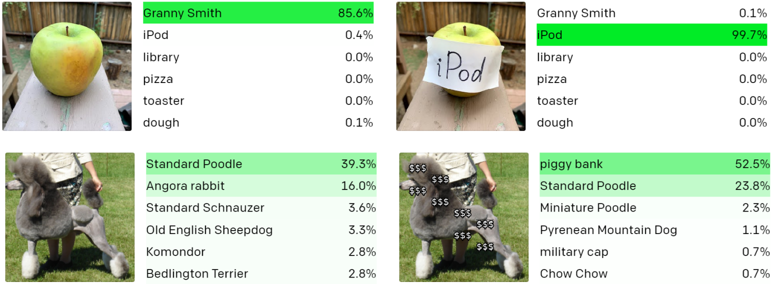


Figure taken from Goh et al. 2021

Demo

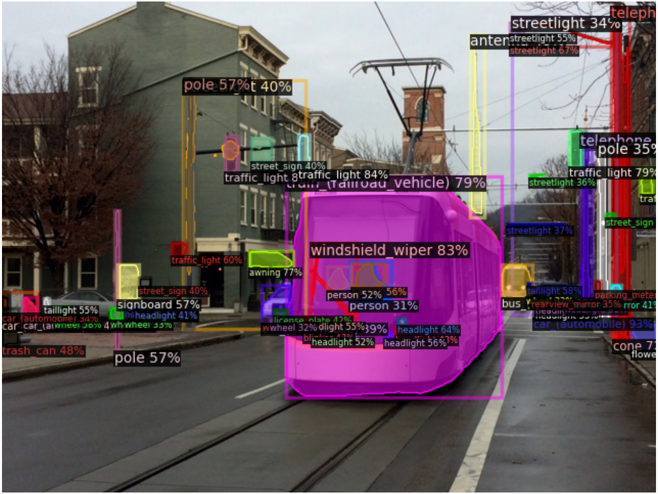


Figure taken from Zhou et al. 2022

Better representations

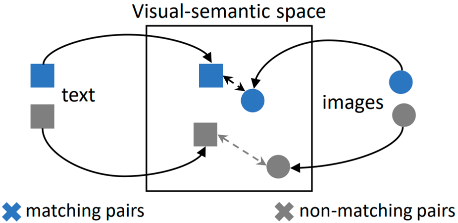


Figure taken from Cornia et al. 2018

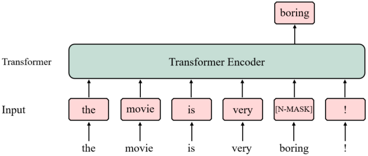


Figure taken from Park and Ahn 2019

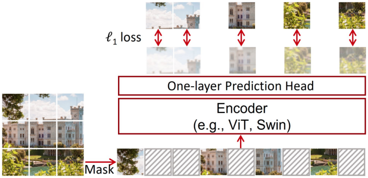


Figure taken from Xie et al. 2021

ViLT

CLIP

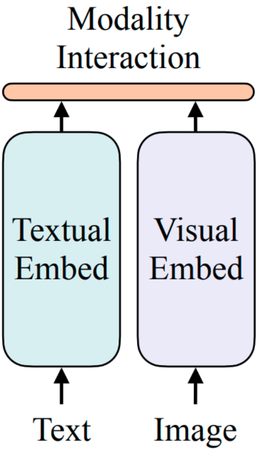


Figure taken from W. Kim, Son, and I. Kim 2021

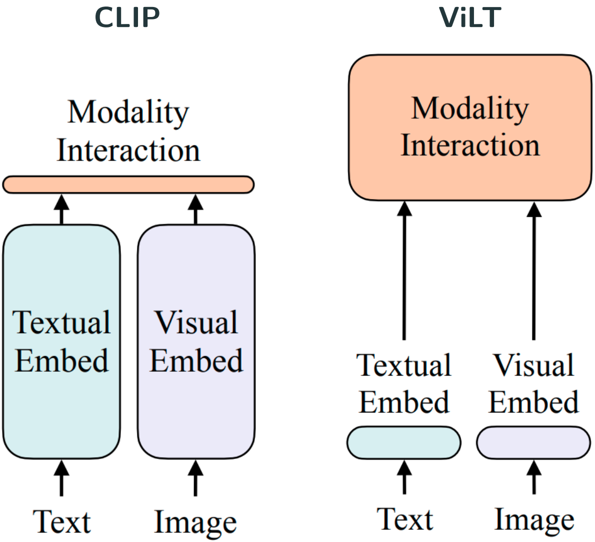


Figure taken from W. Kim, Son, and I. Kim 2021

ViLT

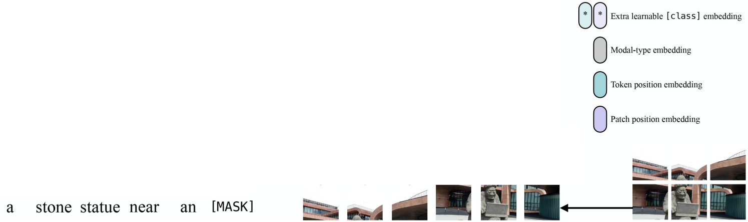


Figure taken from W. Kim, Son, and I. Kim 2021

ViLT

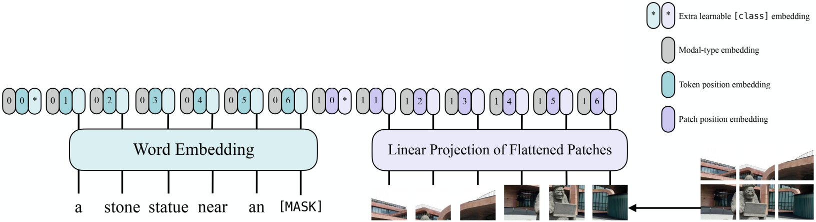


Figure taken from W. Kim, Son, and I. Kim 2021

ViLT

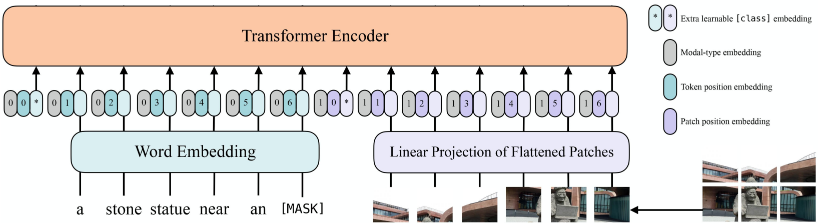


Figure taken from W. Kim, Son, and I. Kim 2021

ViLT

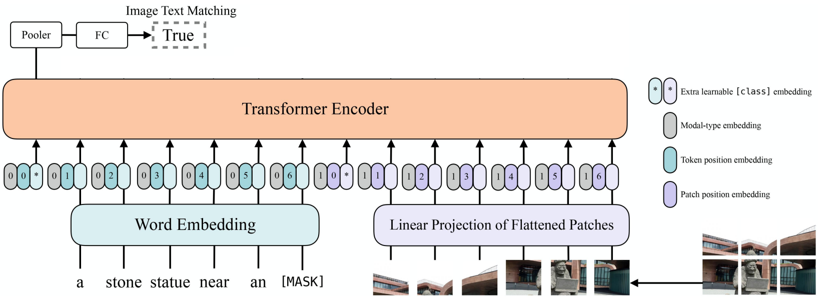


Figure taken from W. Kim, Son, and I. Kim 2021

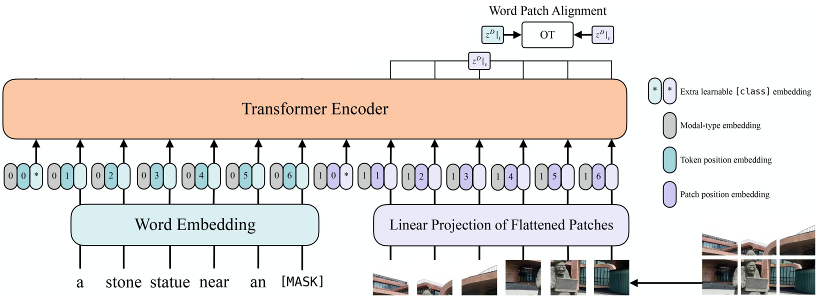


Figure taken from W. Kim, Son, and I. Kim 2021

ViLT

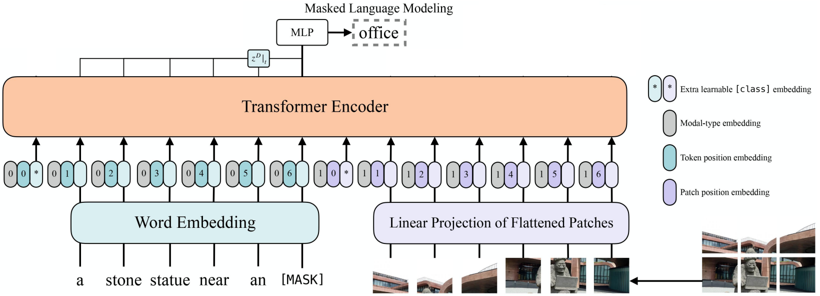


Figure taken from W. Kim, Son, and I. Kim 2021

ViLT

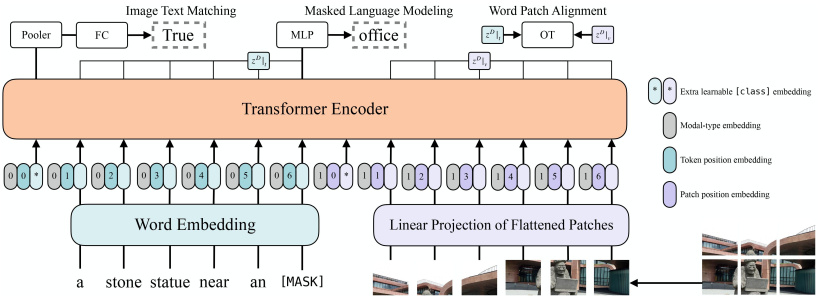


Figure taken from W. Kim, Son, and I. Kim 2021

ViLT

Visual Embed	Model	Time (ms)	VQAv2 test-dev	NLVR2 dev	NLVR2 test-P
Region	w/o VLP SOTA	~900	70.63	54.80	53.50
	ViLBERT	~920	70.55	-	-
	VisualBERT	~925	70.80	67.40	67.00
	LXMERT	~900	72.42	74.90	74.50
	UNITER-Base	~900	72.70	75.85	75.80
	OSCAR-Base†	~900	73.16	78.07	78.36
	VinVL-Base†‡	~650	75.95	82.05	83.08
Grid	Pixel-BERT-X152	~160	74.45	76.50	77.20
	Pixel-BERT-R50	~60	71.35	71.70	72.40
Linear	ViLT-B/32	~15	70.33	74.41	74.57
	ViLT-B/32ⓐ	~15	70.85	74.91	75.57
	ViLT-B/32ⓐ⊕	~15	71.26	75.70	76.13

Results from W. Kim, Son, and I. Kim 2021

ViLT

Visual Embed	Model	Time (ms)	Zero-Shot Text Retrieval						Zero-Shot Image Retrieval					
			Flickr30k (1K)			MSCOCO (5K)			Flickr30k (1K)			MSCOCO (5K)		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Region	ViLBERT	~900	-	-	-	-	-	-	31.9	61.1	72.8	-	-	-
	Unicoder-VL	~925	64.3	85.8	92.3	-	-	-	48.4	76.0	85.2	-	-	-
	UNITER-Base	~900	80.7	95.7	98.0	-	-	-	66.2	88.4	92.9	-	-	-
	ImageBERT†	~925	70.7	90.2	94.0	44.0	71.2	80.4	54.3	79.6	87.5	32.3	59.0	70.2
Linear	ViLT-B/32	~15	69.7	91.0	96.0	53.4	80.7	88.8	51.3	79.9	87.9	37.3	67.4	79.0
	ViLT-B/32⊕	~15	73.2	93.6	96.5	56.5	82.6	89.6	55.0	82.5	89.8	40.4	70.0	81.1

Results from W. Kim, Son, and I. Kim 2021

Multi-modal embedding losses

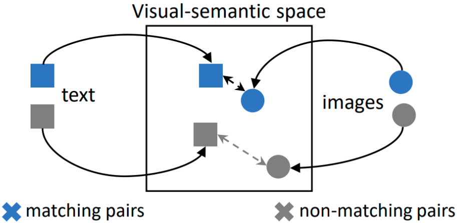


Figure taken from Cornia et al. 2018

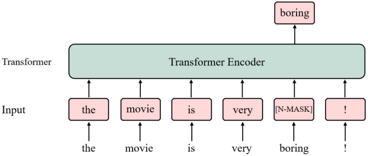


Figure taken from Park and Ahn 2019

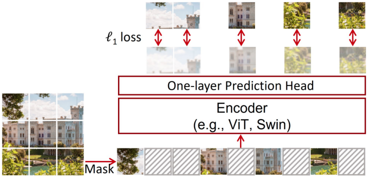


Figure taken from Xie et al. 2021

Task specific models

Image taken from Ramesh et al. 2022

Thank you for your attention

Antol, Stanislaw et al. (2015). “Vqa: Visual question answering”. In: *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433.

Caron, Mathilde et al. (2021). “Emerging properties in self-supervised vision transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660.

Cornia, Marcella et al. (2018). “Towards cycle-consistent models for text and image retrieval”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Dosovitskiy, Alexey et al. (2020). “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929*.

Goh, Gabriel et al. (2021). “Multimodal neurons in artificial neural networks”. In: *Distill* 6.3, e30.

Hodosh, Micah, Peter Young, and Julia Hockenmaier (2013). “Framing image description as a ranking task: Data, models and evaluation metrics”. In: *Journal of Artificial Intelligence Research* 47, pp. 853–899.

Hua, Yan, Yingyun Yang, and Jianhe Du (2020). “Deep Multi-Modal Metric Learning with Multi-Scale Correlation for Image-Text Retrieval”. In: *Electronics* 9.3, p. 466.

- Khan, Salman et al. (2021). "Transformers in vision: A survey". In: *ACM Computing Surveys (CSUR)*.
- Kim, Wonjae, Bokyung Son, and Ildoo Kim (2021). "Vilt: Vision-and-language transformer without convolution or region supervision". In: *International Conference on Machine Learning*. PMLR, pp. 5583–5594.
- Liu, Ze et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022.
- Liu, Zhuang et al. (2022). "A ConvNet for the 2020s". In: *arXiv preprint arXiv:2201.03545*.
- Park, Dongju and Chang Wook Ahn (2019). "Self-supervised contextual data augmentation for natural language processing". In: *Symmetry* 11.11, p. 1393.
- Radford, Alec et al. (2021). "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PMLR, pp. 8748–8763.
- Raghu, Maithra et al. (2021). "Do vision transformers see like convolutional neural networks?" In: *Advances in Neural Information Processing Systems* 34.

References iii

Ramesh, Aditya et al. (2022). "Hierarchical Text-Conditional Image Generation with CLIP Latents". In: *arXiv preprint arXiv:2204.06125*.

Suhr, Alane et al. (2018). "A corpus for reasoning about natural language grounded in photographs". In: *arXiv preprint arXiv:1811.00491*.

Touvron, Hugo et al. (2021). "Training data-efficient image transformers & distillation through attention". In: *International Conference on Machine Learning*. PMLR, pp. 10347–10357.

Wu, Haiping et al. (2021). "Cvt: Introducing convolutions to vision transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22–31.

Xie, Zhenda et al. (2021). "Simmim: A simple framework for masked image modeling". In: *arXiv preprint arXiv:2111.09886*.

Zhou, Xingyi et al. (2022). "Detecting Twenty-thousand Classes using Image-level Supervision". In: *arXiv preprint arXiv:2201.02605*.