Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 6.1: Multimodal Transformers (Part 2)

Mehul Agarwal, Louis-Philippe Morency

** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Administrative Stuff

Second Project Assignment (Due Sunday 10/8)

Main goals:

1. Help clarify and expand your research ideas
   - Build qualitative intuitions by directly studying the original data
   - Perform analyses on your dataset, relevant to your research ideas
2. Understand the structure in your data and modalities
   - Perform analyses and visualizations to understand each modality
   - Study representations from language and visual modalities

Two types of analyses:

- Idea-oriented analyses
- Modality-oriented analyses

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 8/29 & 8/31	**Course introduction** • Multimodal core challenges • Course syllabus	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/5 & 9/7 Read due: 9/9	**Unimodal representations** • Dimensions of heterogeneity • Visual representations	**Unimodal representations** • Language representations • Signals, graphs and other modalities
Week 3 9/12 & 9/14 Read due: 9/16 Proj. Due: 9/13	**Multimodal representations** • Cross-modal interactions • Multimodal fusion	**Multimodal representations** • Coordinated representations • Multimodal fission
Week 4 9/19 & 9/21 *Proj. due: 9/24*	**Multimodal alignment and grounding** • Explicit alignment • Multimodal grounding	**Alignment and representations** • Self-attention transformer models • Masking and self-supervised learning
Week 5 9/26 & 9/28 Read due: 9/30	**Multimodal transformers – Part 1** • Language pretraining • Multimodal transformers	**Multimodal Reasoning** • Structured and hierarchical models • Memory models
Week 6 10/3 & 10/5 *Proj. due: 10/8*	**Multimodal transformers – Part 2** • Image and video transformers • Vision-language transformers	***Multimodal language grounding*** • Guest lecturer: Jack Hessel • Vision, language and grounding

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 6.1: Multimodal Transformers (Part 2)

Mehul Agarwal, Louis-Philippe Morency

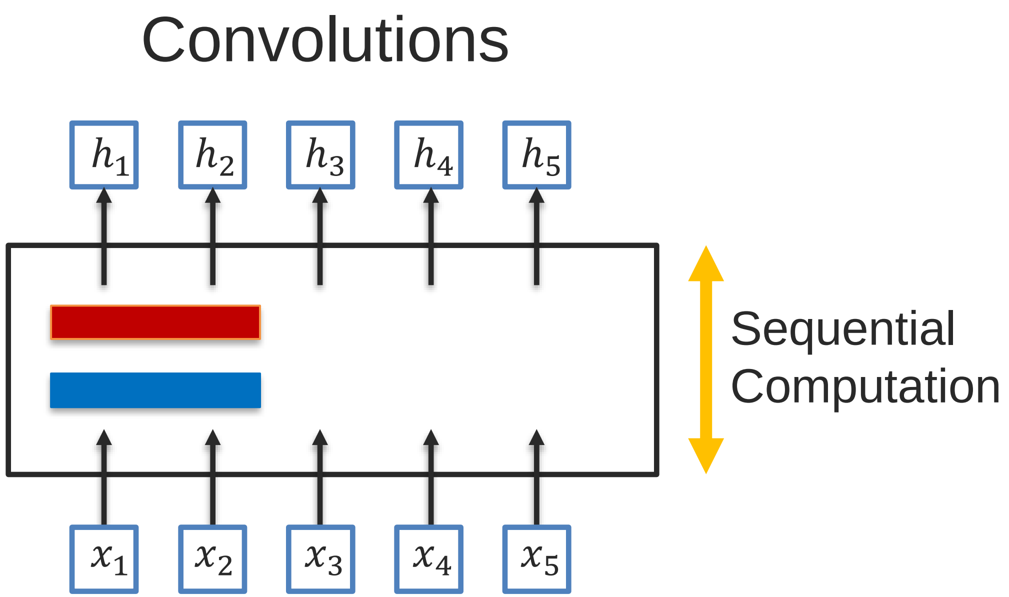
** Co-lecturer: Paul Liang. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

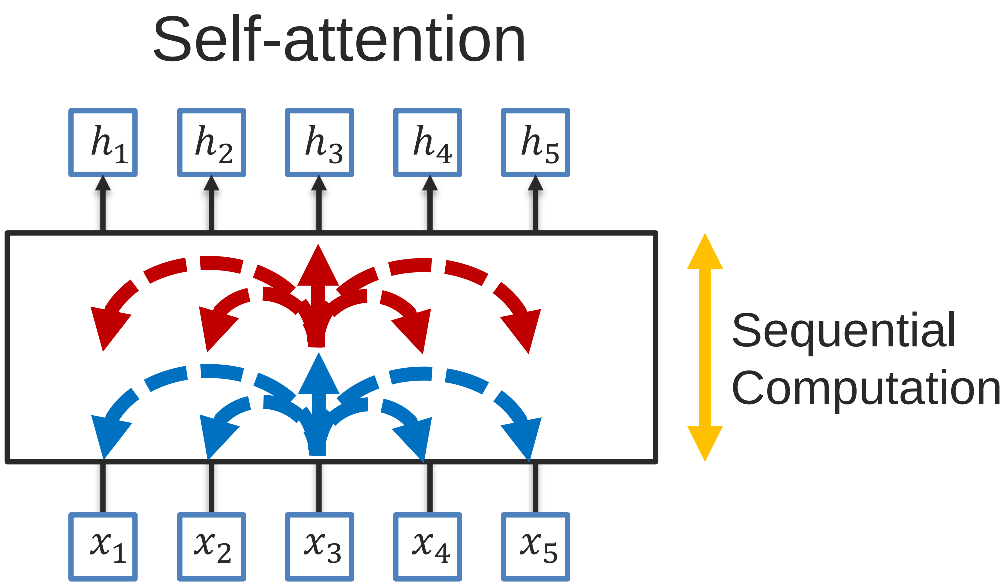
Objectives of today's class

- Visual transformers
  - Vision transformer (VIT)
  - Masked Auto-Encoder (MAE)
- Visual-language transformers:
  - ViLT = VIT+BERT
  - Vision-Language Caption (MAE+BERT)
- Video transformers

Vision Transformers

Recap: CNNs vs Transformers





Can be parallelized!

But modeling long-range dependencies requires many layers.

And convolutional kernels are static.

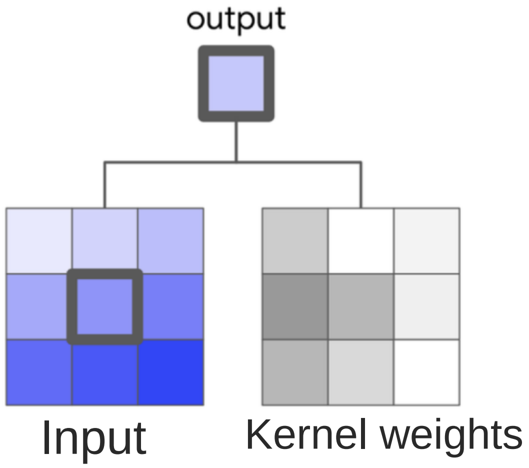
Can be parallelized!

Long-range dependencies

Dynamic attention weights

No inductive bias toward locality

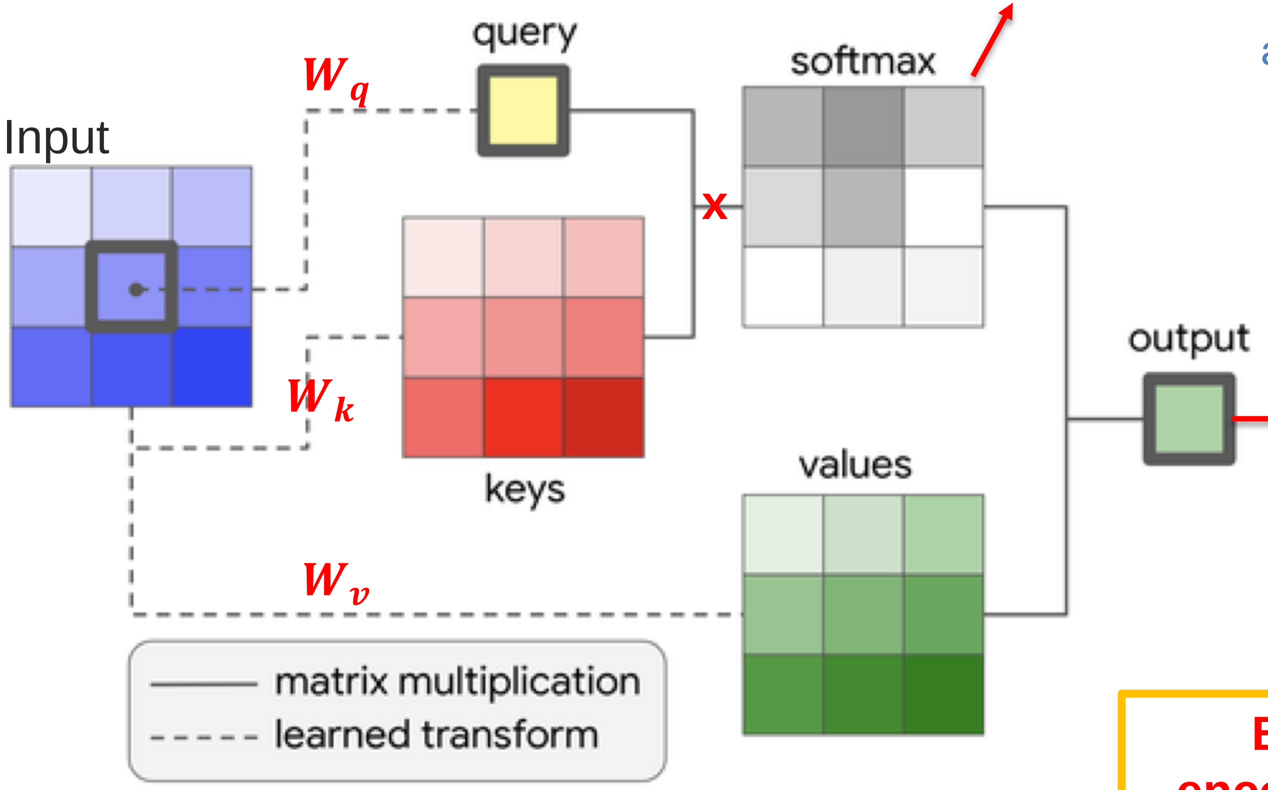
Replacing a CNN w/ Self-Attention



Convolution

How well the query matches the keys?
(How well the pixel matches its neighbors?)

aka “memory block”



Pixel contextualized by its neighbors

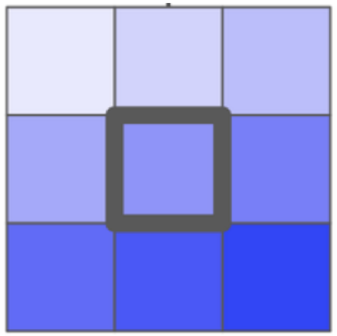
Self-attention

But how to encode position?

https://arxiv.org/abs/1906.05909

Replacing a CNN w/ Self-Attention

Image patch



2D relative position embedding

-1, -1	-1, 0	-1, 1	-1, 2
0, -1	0, 0	0, 1	0, 2
1, -1	1, 0	1, 1	1, 2
2, -1	2, 0	2, 1	2, 2

Position embedding is added to the key:

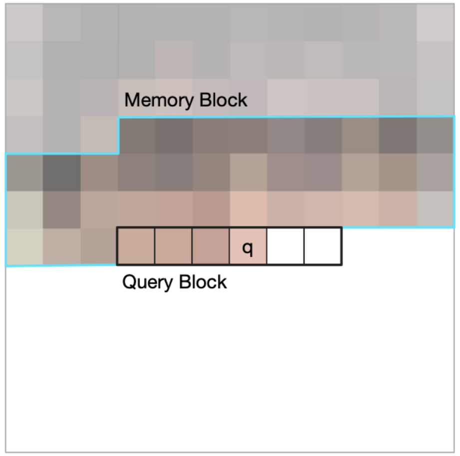
$$y_{ij} = \sum_{a,b \in \mathcal{N}_k(i,j)} \texttt{softmax}_{ab}\left(q_{ij}^\top k_{ab} + q_{ij}^\top r_{a-i,b-j}\right) v_{ab}$$

https://arxiv.org/abs/1906.05909

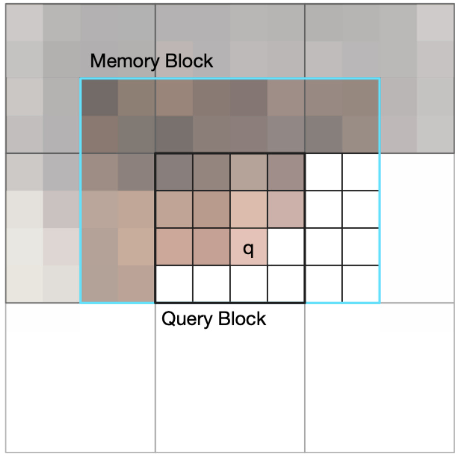
Pixel-Based Image Generation via Transformers

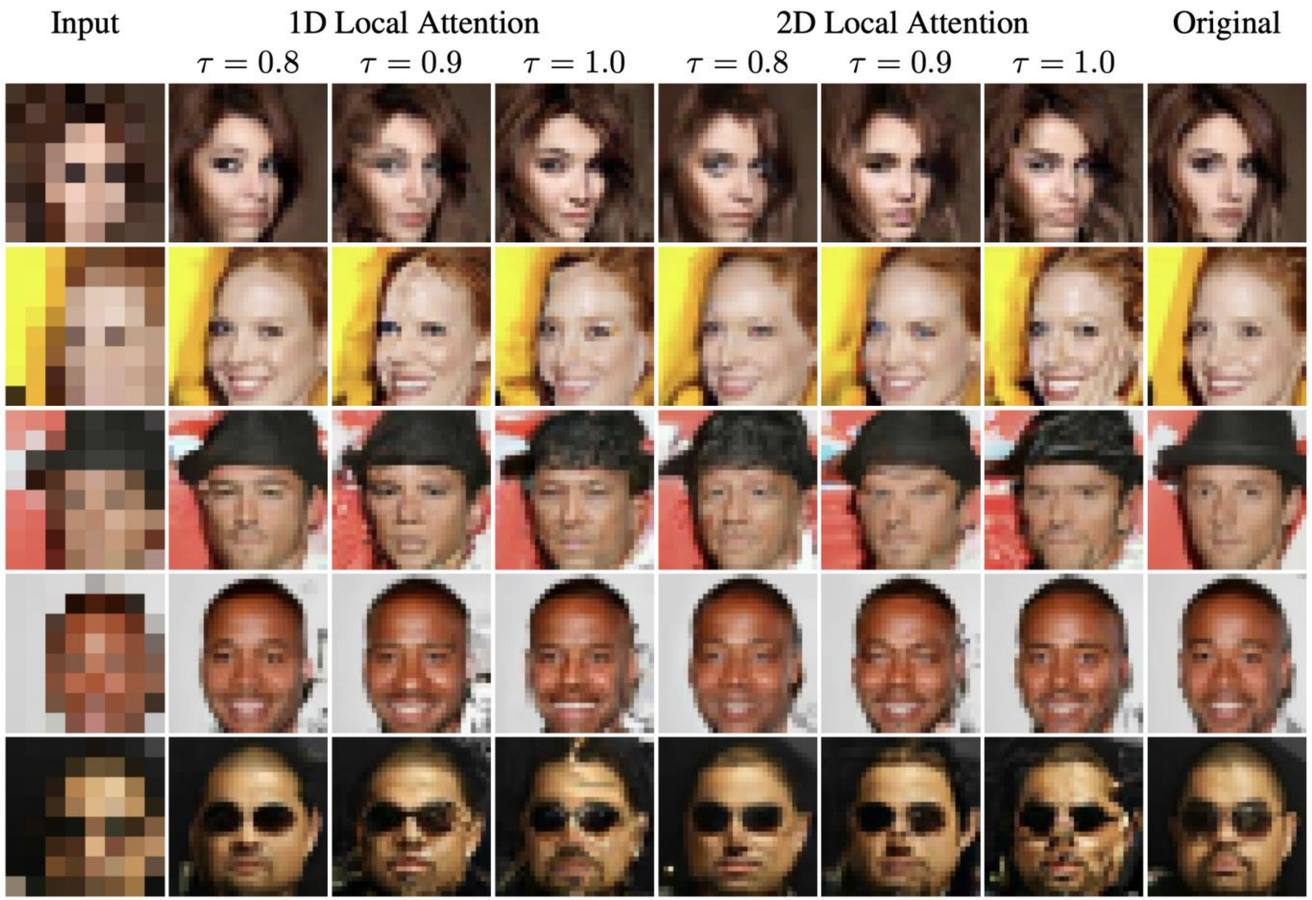
Produce 32x32 images, one channel of each pixel at a time. 3 x 32 x 32 = 3072 positions

Local 1D Attention



Local 2D Attention

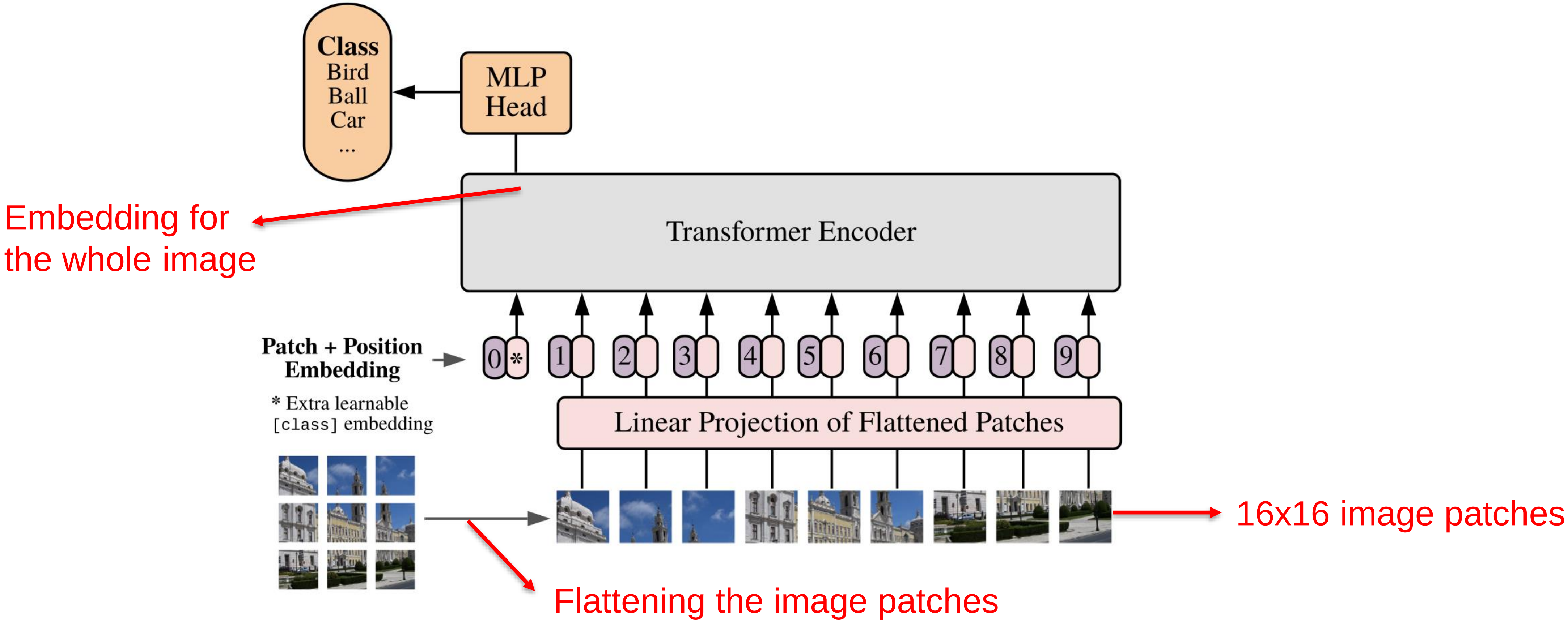




Vision Transformer (ViT)

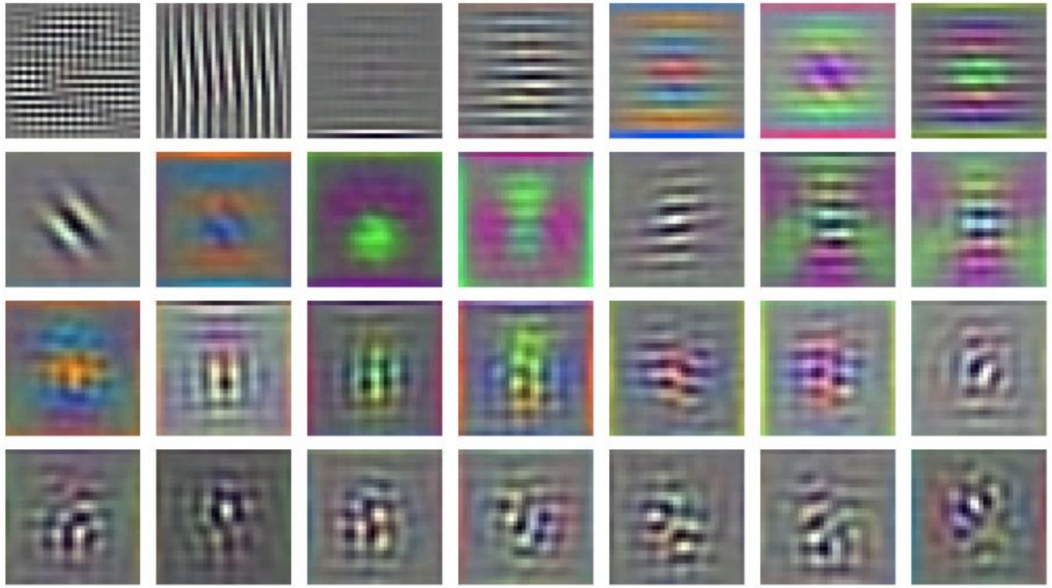
Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

Vision Transformer (ViT)

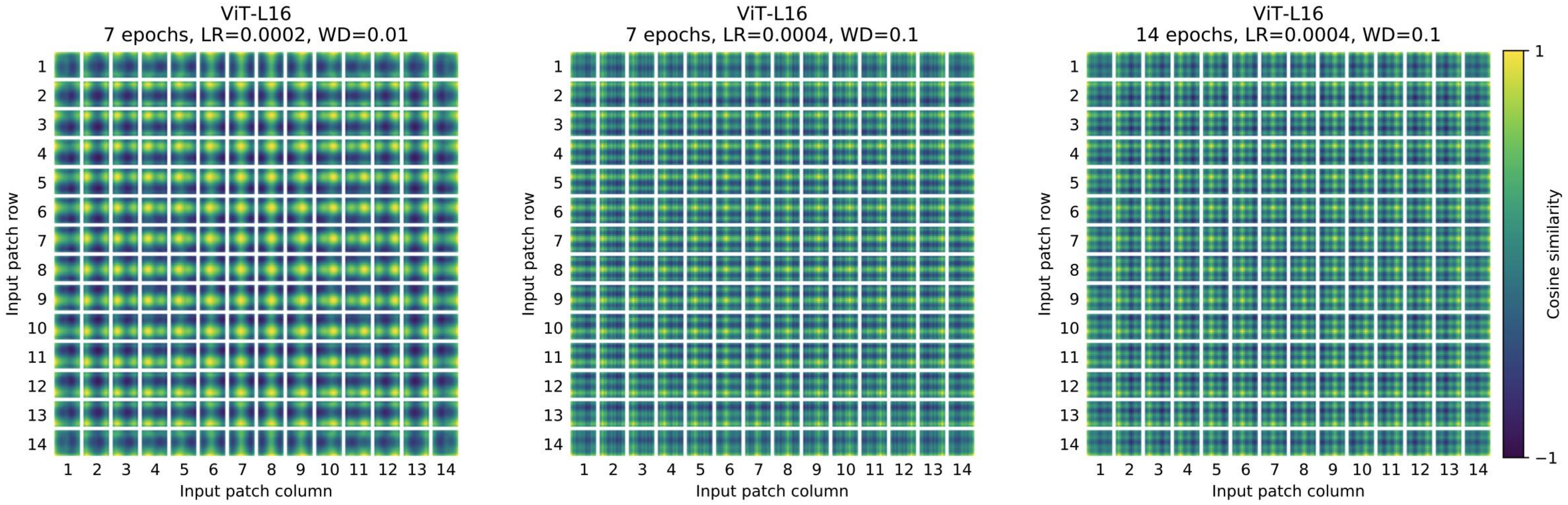


Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

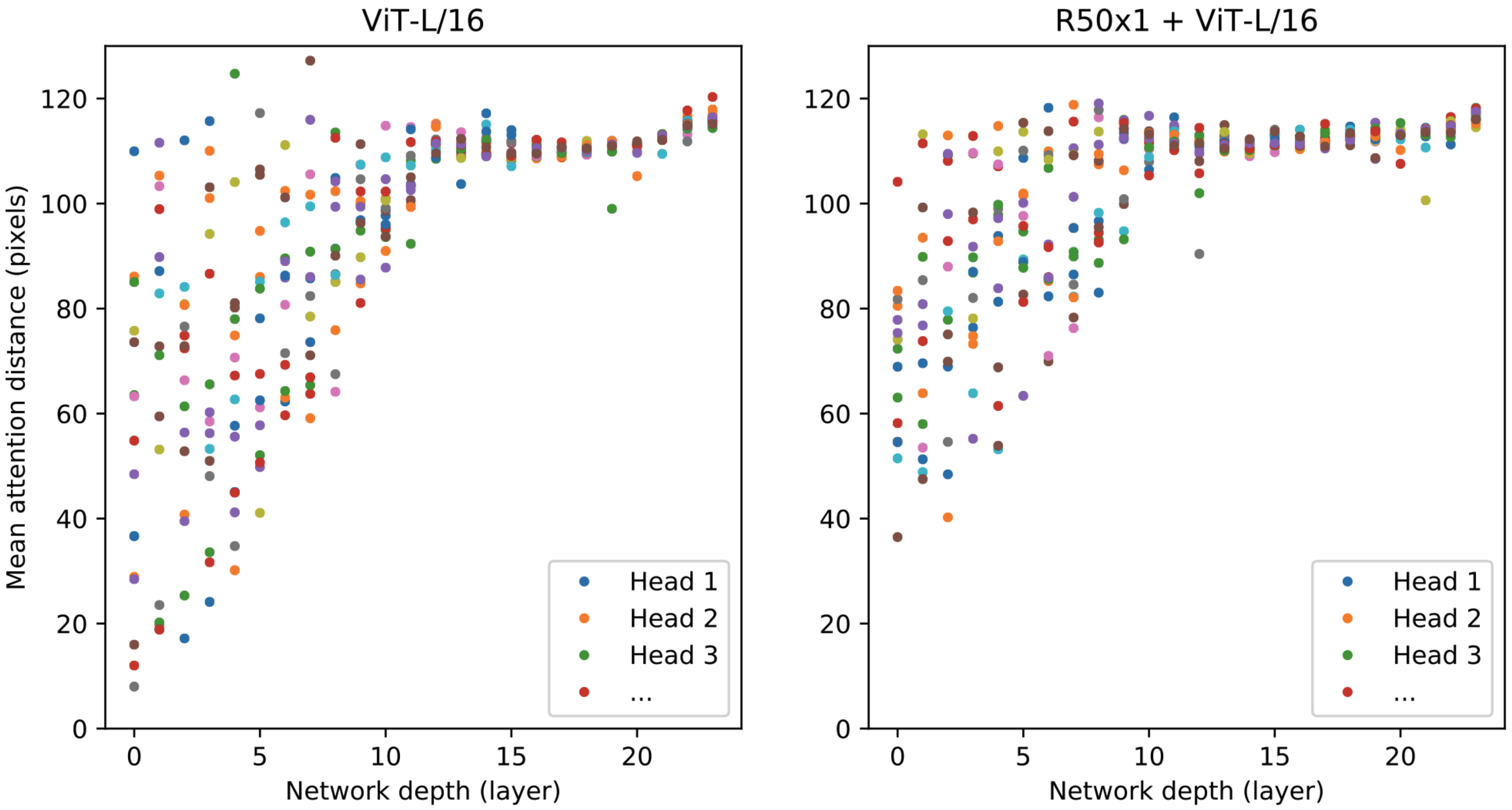
Filters



Learning Location



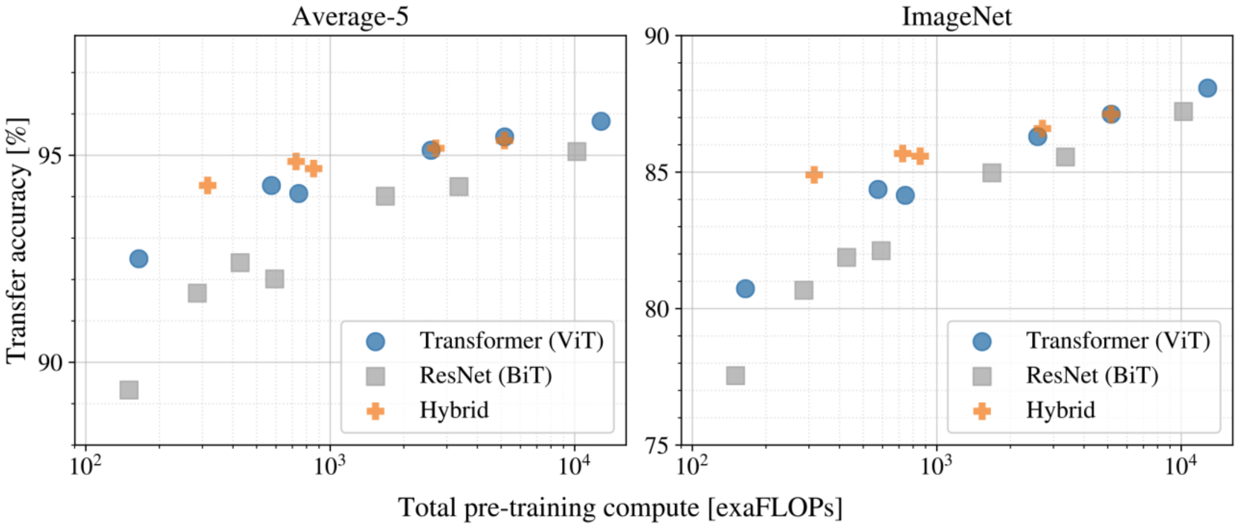
Learning Location



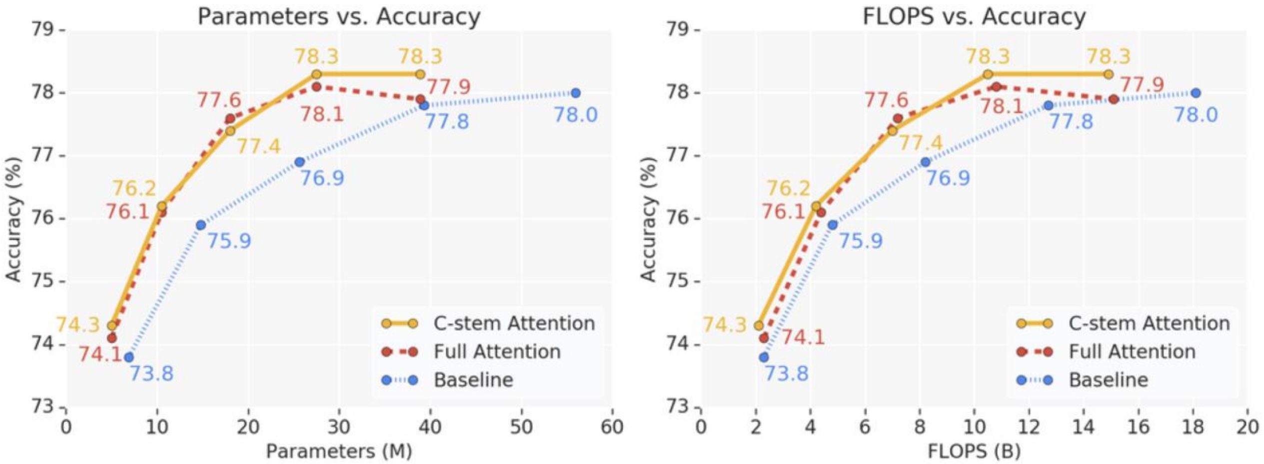
Which is the best?

ImageNet	ImageNet ReaL	CIFAR-10	CIFAR-100	Pets	Flowers	exaFLOPs
80.73	86.27	98.61	90.49	93.40	99.27	164
84.15	88.85	99.00	91.87	95.80	99.56	743
84.37	88.28	99.19	92.52	95.83	99.45	574
86.30	89.43	99.38	93.46	96.81	99.66	2586
87.12	89.99	99.38	94.04	97.11	99.56	5172
88.08	90.36	99.50	94.71	97.11	99.71	12826
77.54	84.56	97.67	86.07	91.11	94.26	150
82.12	87.94	98.29	89.20	93.43	97.02	592
80.67	87.07	98.48	89.17	94.08	95.95	285
81.88	87.96	98.82	90.22	94.17	96.94	427
84.97	89.69	99.06	92.05	95.37	98.62	1681
85.56	89.89	99.24	91.92	95.75	98.75	3362
87.22	90.15	99.34	93.53	96.32	99.04	10212
84.90	89.15	99.01	92.24	95.75	99.46	315
85.58	89.65	99.14	92.63	96.65	99.40	855
85.68	89.04	99.24	92.93	96.97	99.43	725
86.60	89.72	99.18	93.64	97.03	99.40	2704
87.12	89.76	99.31	93.89	97.36	99.11	5165

Curves



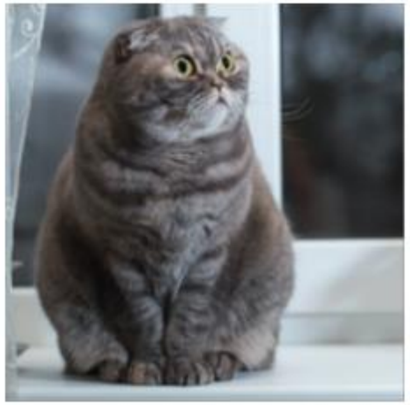
Average-5
ImageNet
Transfer accuracy [%]
Total pre-training compute [exaFLOPs]
Transformer (ViT)
ResNet (BiT)
Hybrid

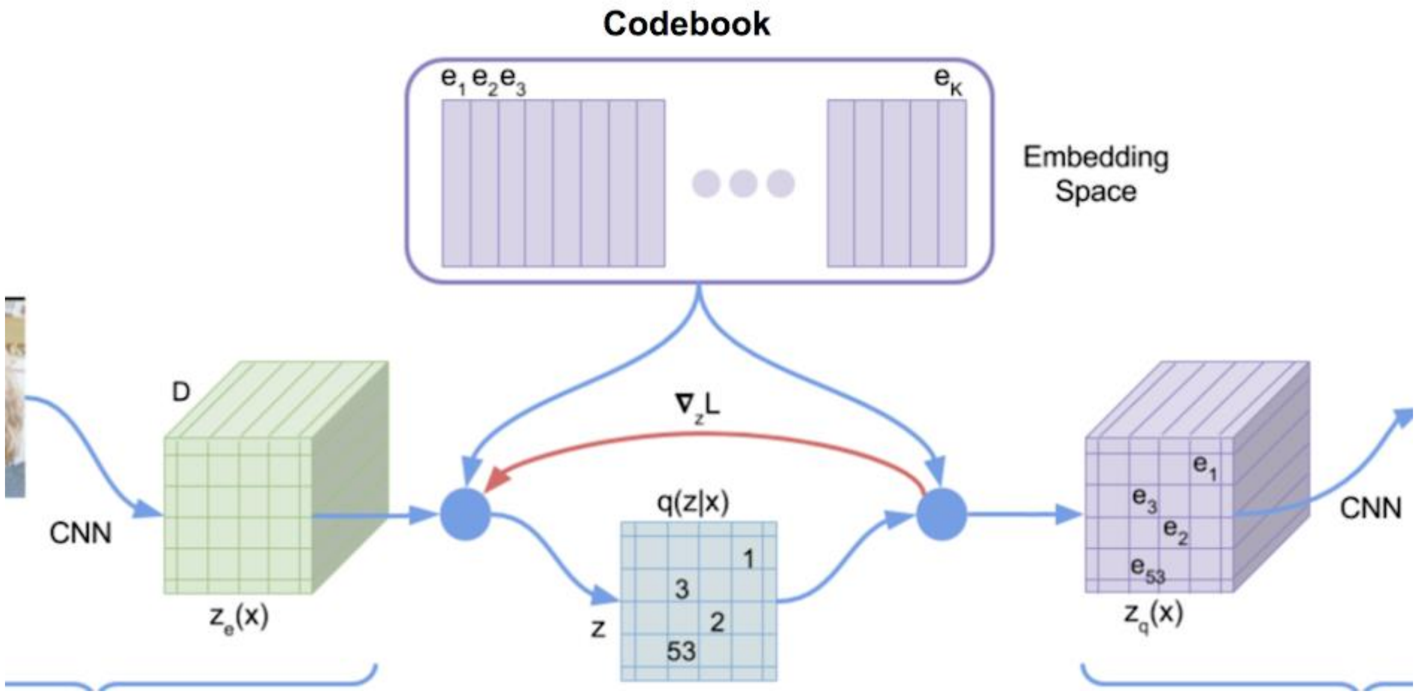


Parameters vs. Accuracy
FLOPS vs. Accuracy
Accuracy (%)
Parameters (M)
FLOPS (B)
C-stem Attention
Full Attention
Baseline

Visual Tokens

DALL-E’s Discrete Variational Autoencoder



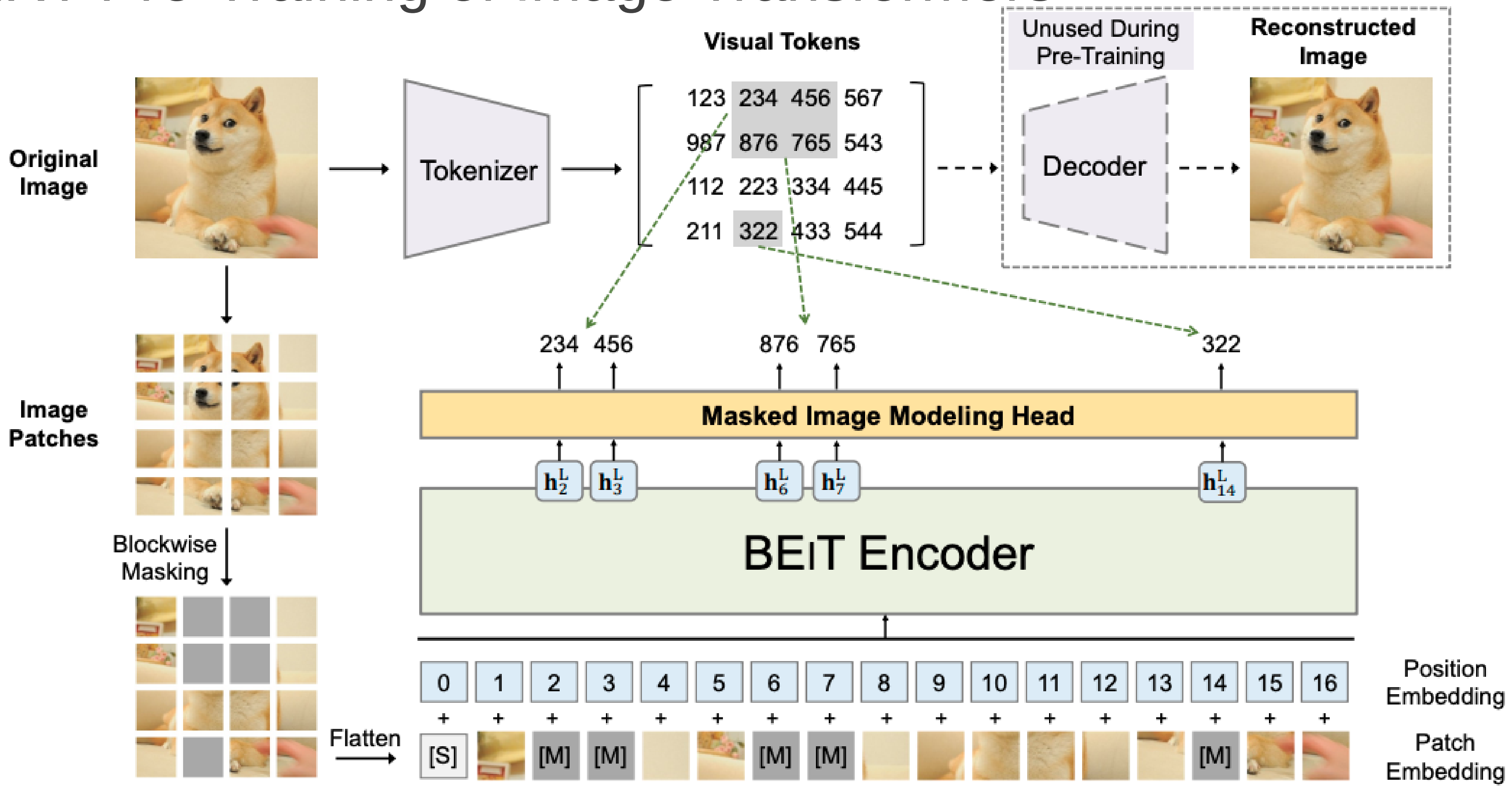


32 x 32 grid of digits, [0… 8192]
Each digit is a “visual token”

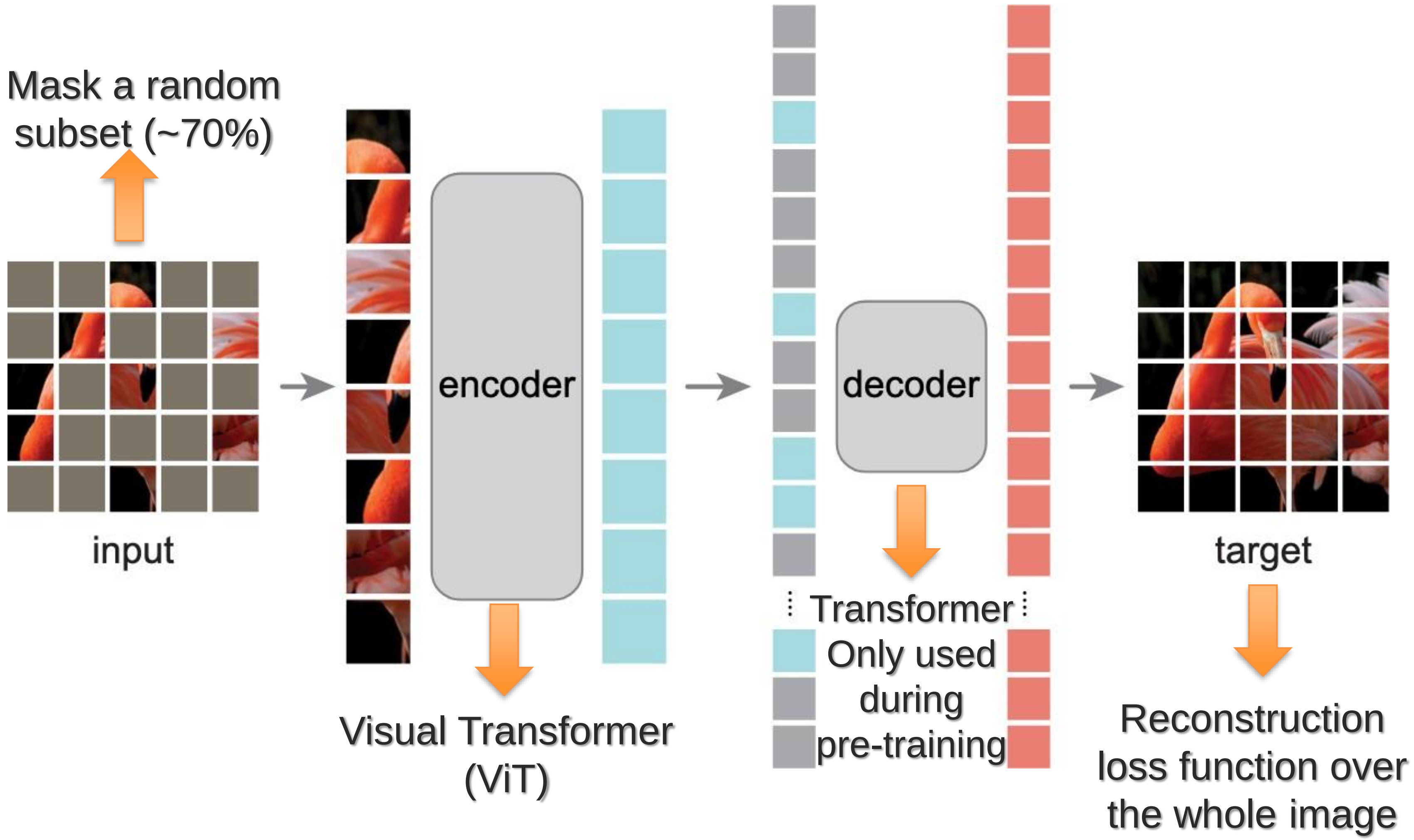
https://arxiv.org/abs/2102.12092,

Visual Tokens

BeIT: BERT Pre-Training of Image Transformers

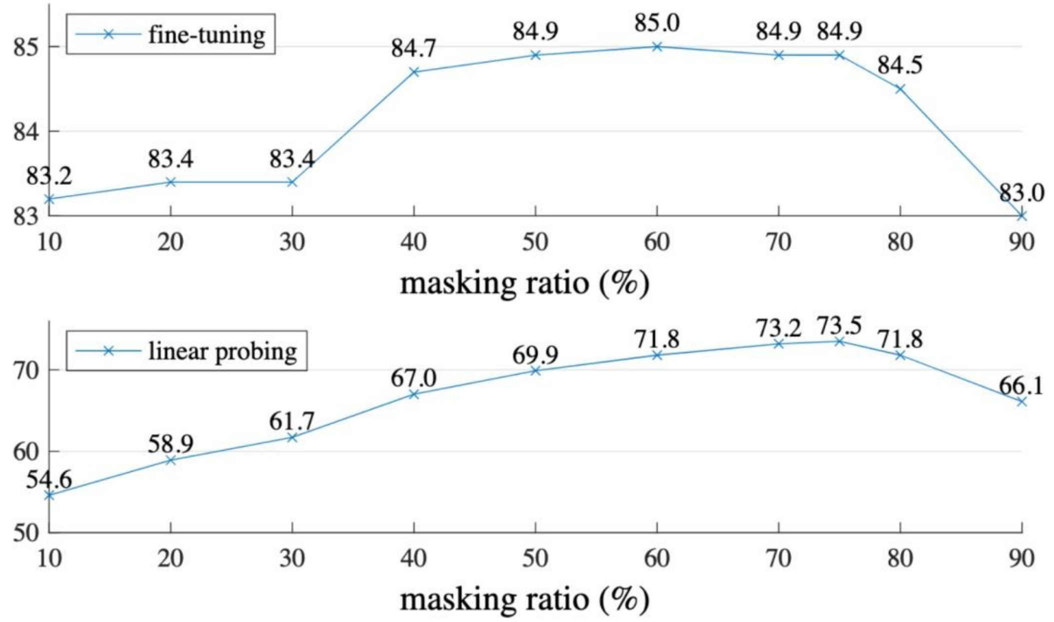


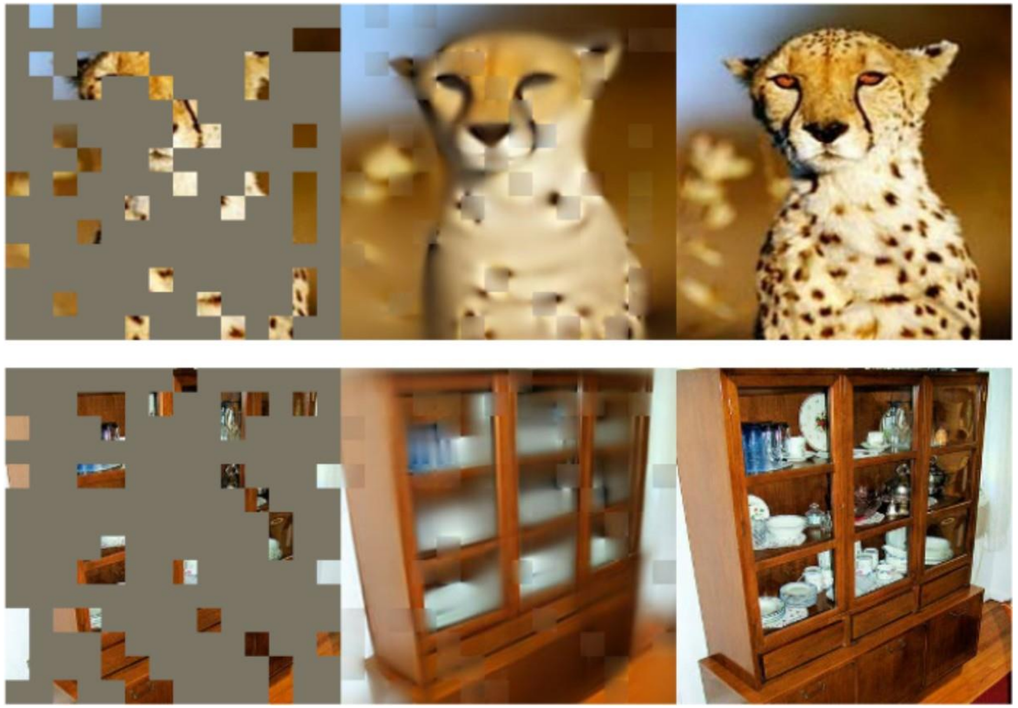
Masked Auto-Encoder (MAE)



He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

Masked Auto-Encoder (MAE)

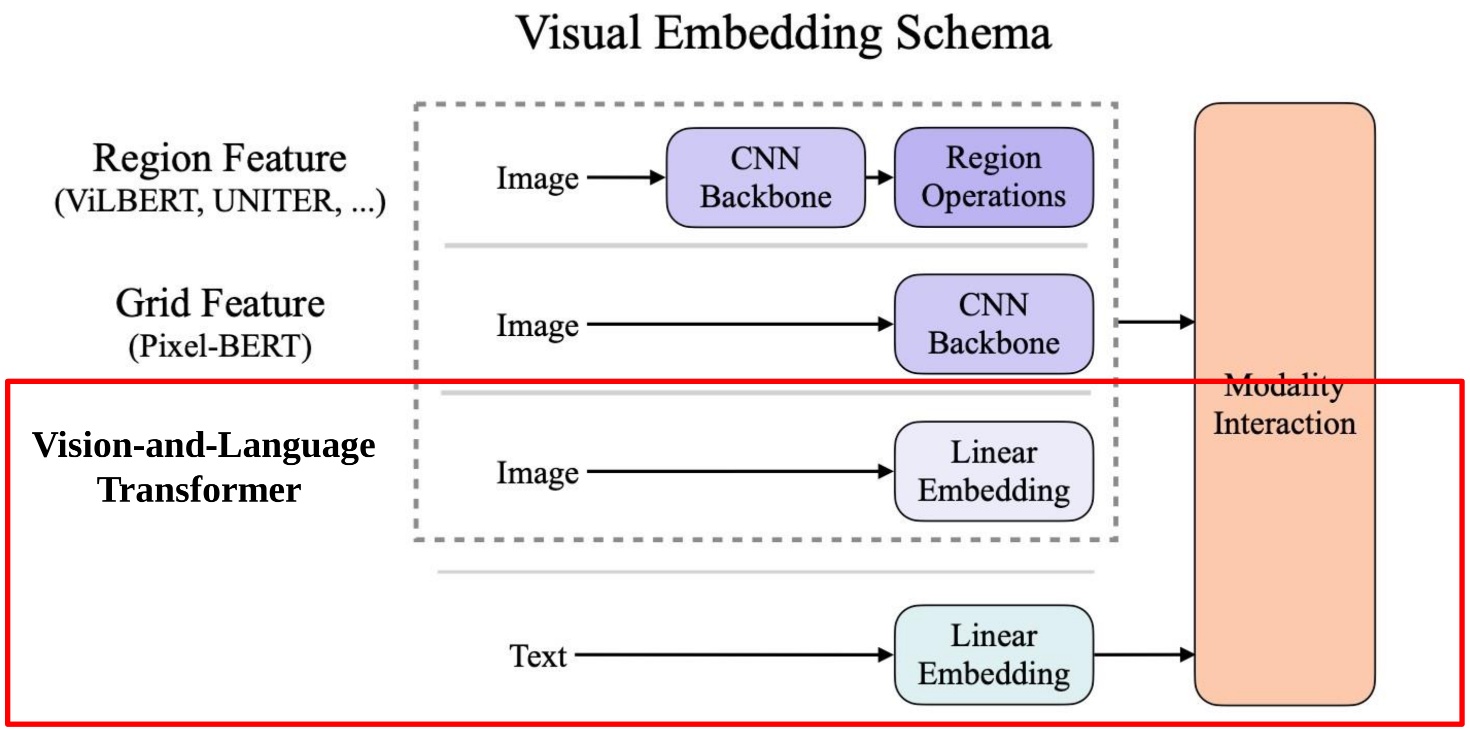




He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

Vision-Language Transformers

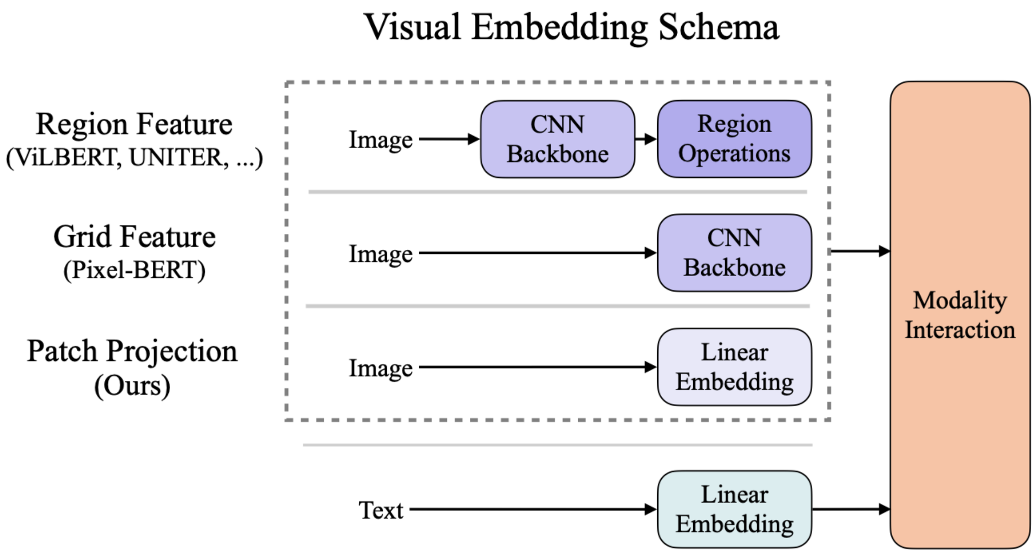
Visual Transformers for Multimodal Learning



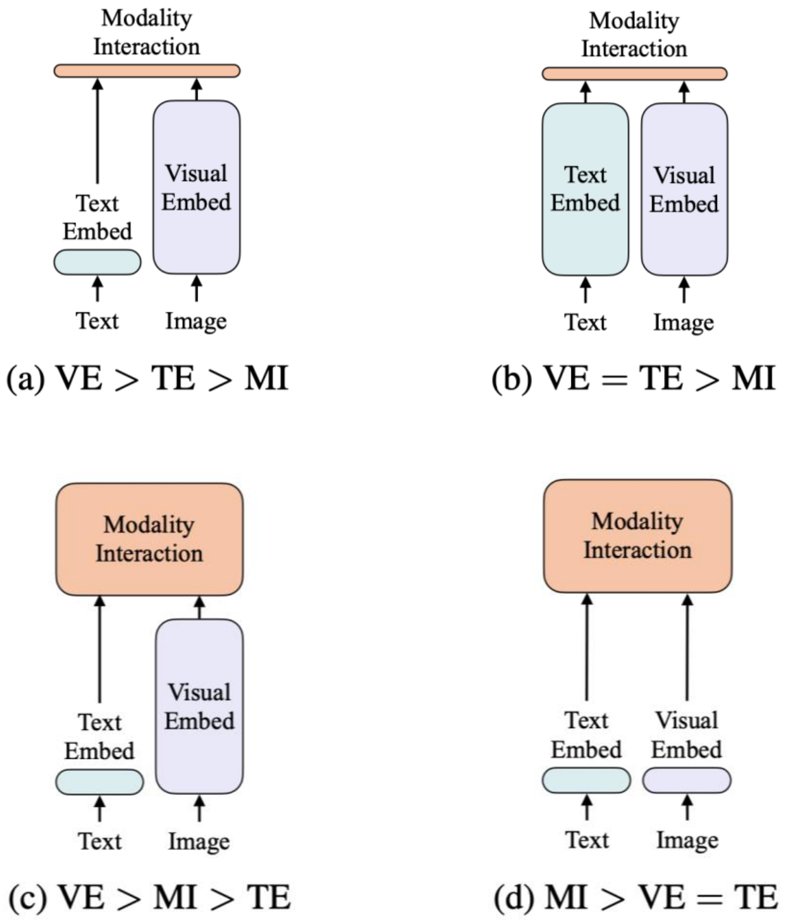
(≈ BERT + ViT)

https://arxiv.org/abs/2102.03334

Visual Transformers for Multimodal Learning



e.g. CLIP

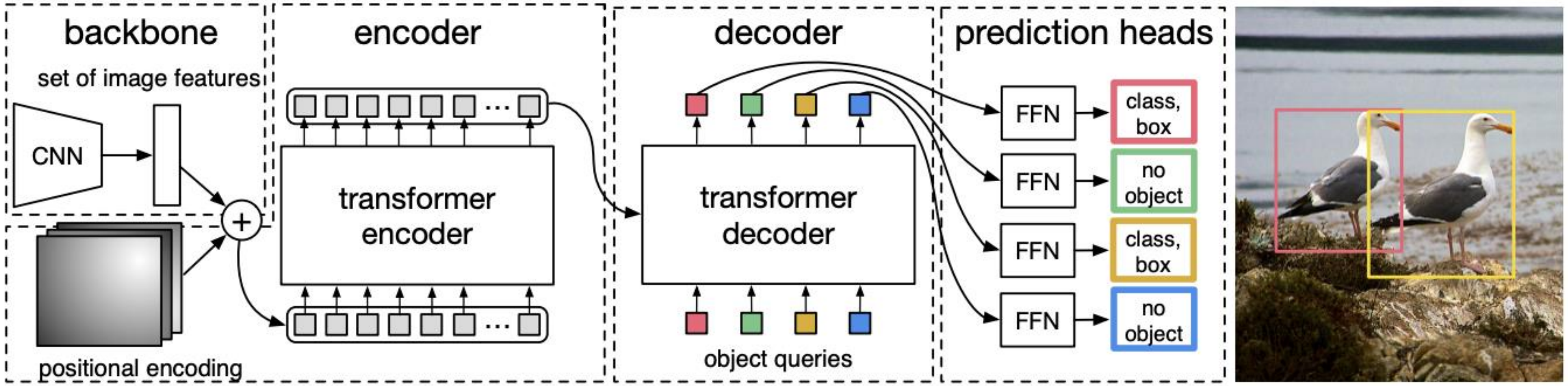


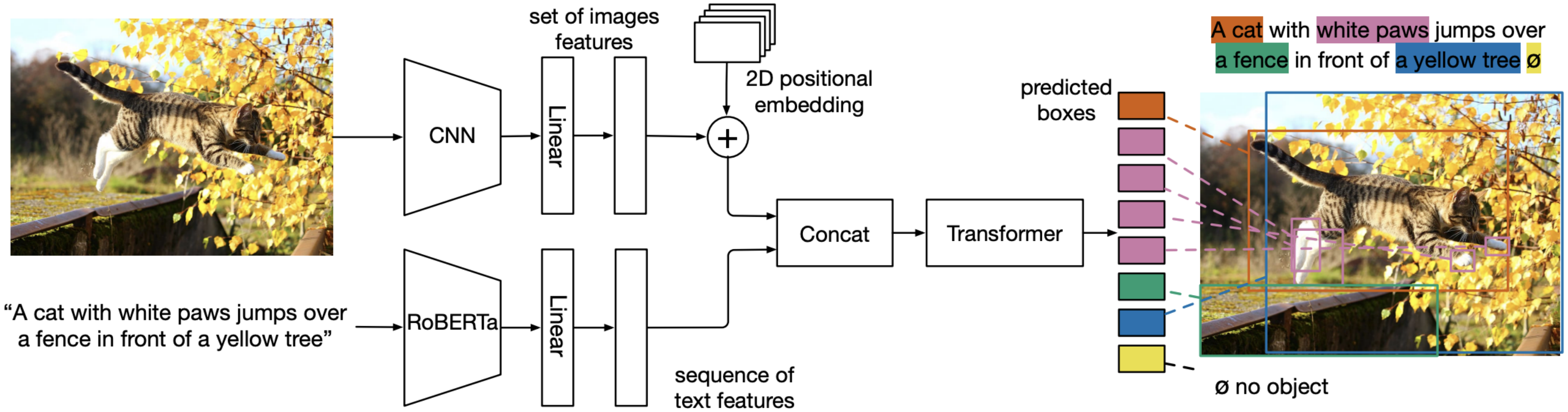
e.g. LXMERT

ViLT

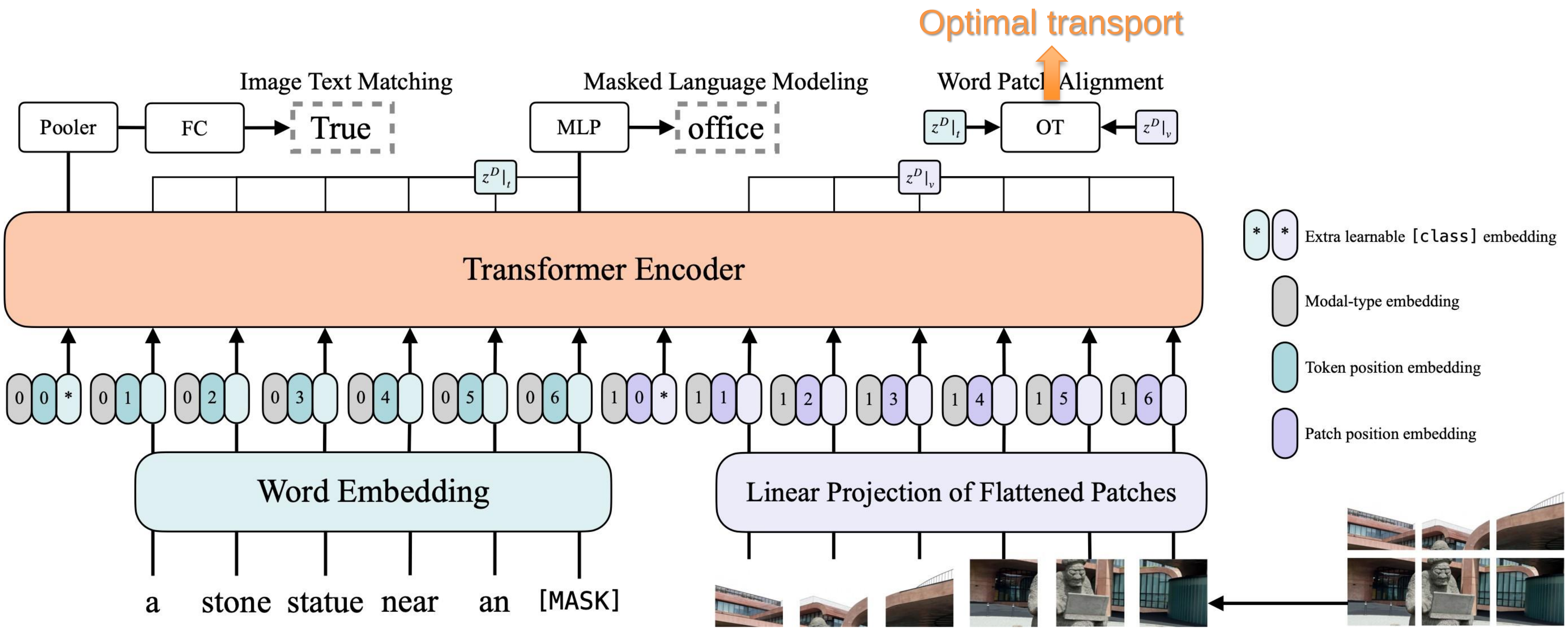
DETR / MDETR (CNN+BERT)

Predicting bounding boxes from images (and text)





Visual-and-Language Transformer (ViLT) (≈ BERT + ViT)



https://arxiv.org/abs/2102.03334

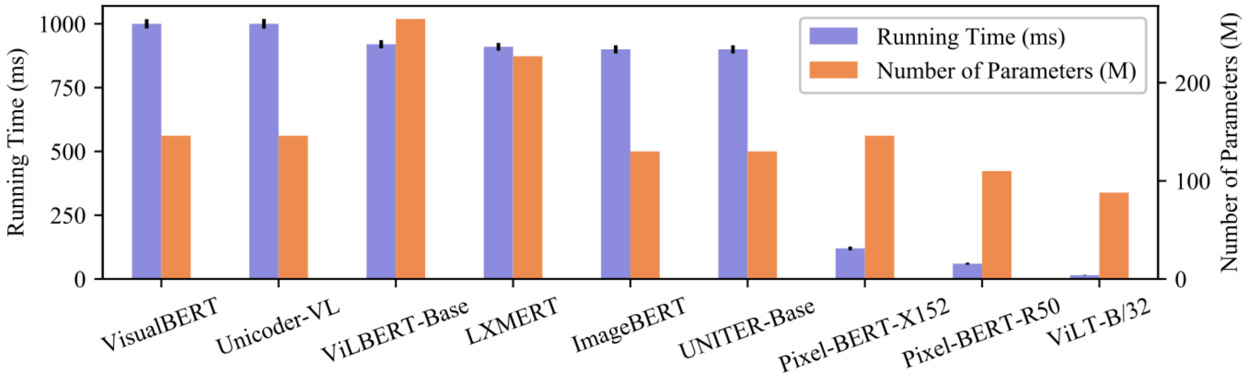
Visual-and-Language Transformer (ViLT)

Example of alignment between modalities:

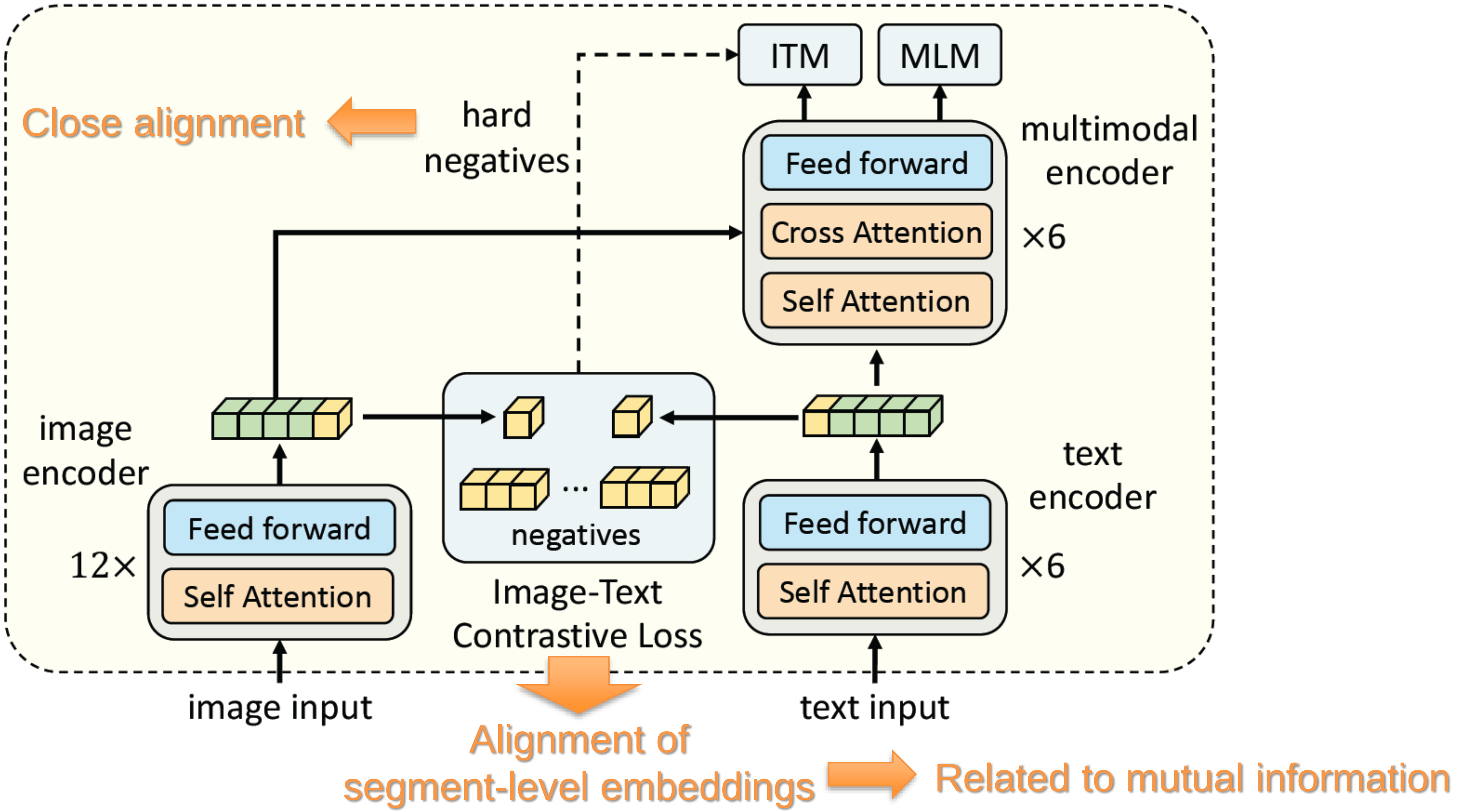
https://arxiv.org/abs/2102.03334

ViLT: Faster Inference?

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-Base	~1000	70.80	67.40	67.00
	LXMERT	~910	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†‡	~1000	75.95	82.05	83.08
Grid	Pixel-BERT-X152	~120	74.45	76.50	77.20
	Pixel-BERT-R50	~60	71.35	71.70	72.40
Linear	ViLT-B/32	~15	70.34	74.56	74.66
	ViLT-B/32ⓐ	~15	70.94	75.24	76.21



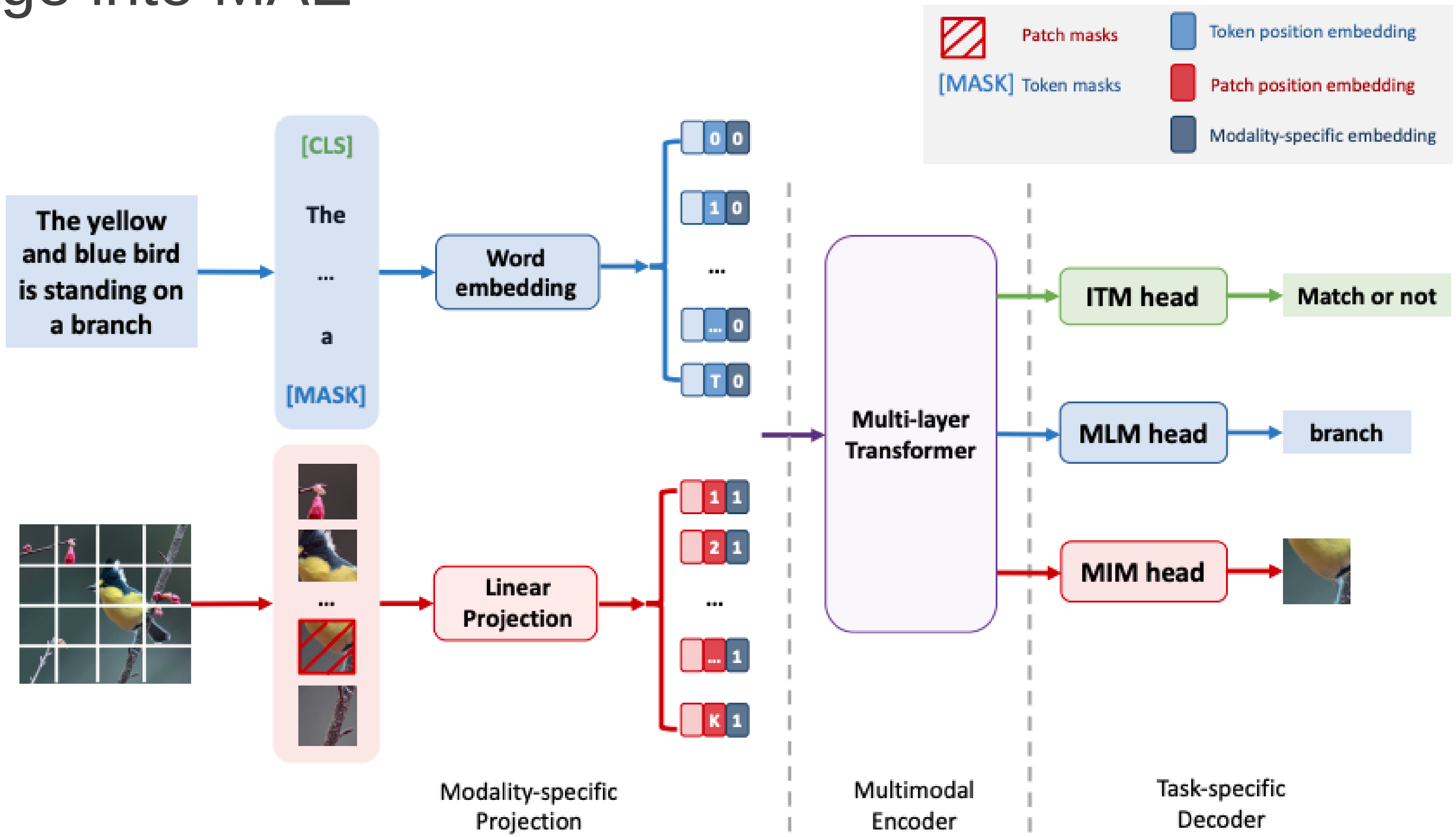
ALBEF: Align Before Fusion (≈ BERT + ViT + CLIP-ish)



Li et al., Align before Fuse: Vision and Language Representation Learning with Momentum Distillation, Neurips 2021

Vision-Language from Captions (VLC)

Add language into MAE

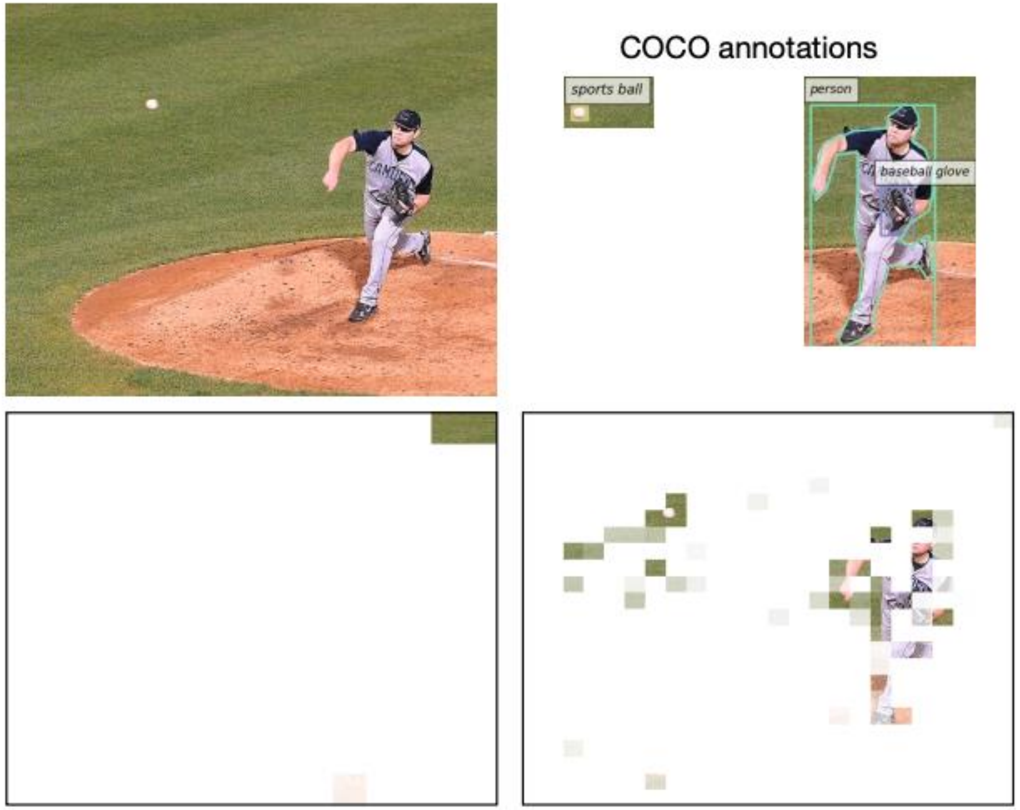


https://arxiv.org/abs/2205.09256

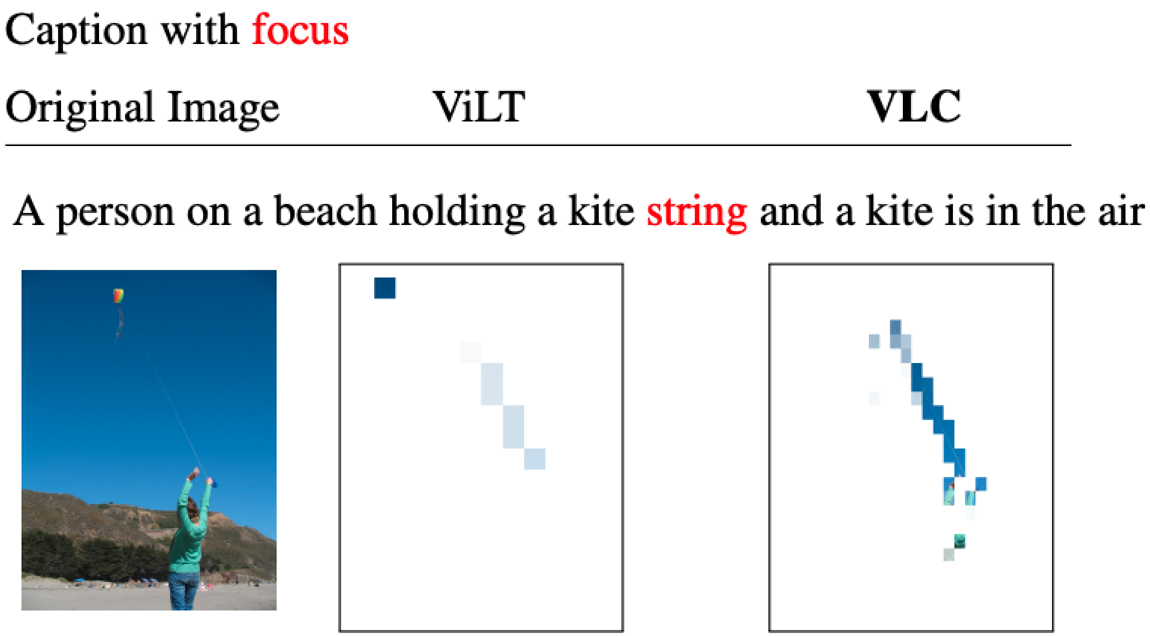
Vision-Language from Captions (VLC)

What are we learning?

*A pitcher at a baseball game who has just **thrown** the ball.*



ViLT (supervised with ImageNet) Ours (no BBox/class supervision)



Vision-Language from Captions (VLC)

What are we learning?

Model	Params	Image Retrieval Flickr30K (1K) @1	@5	@10	MSCOCO (5K) @1	@5	@10
ViLT [23]	86M	64.4	88.7	93.8	42.7	72.9	83.1
VLC-Base (ours – 5.6M)	86M	72.4	93.4	96.5	50.7	78.9	88.0

Model	Params	VQAv2 test-dev	test-std	NLVR^2 dev	test
ViLT [23]	86M	71.26	-	75.70	76.13
No supervised classes or bounding boxes					
VLC-Base (ours – 4M)	86M	72.98	73.03	77.04	78.51

Video Transformers

Video-based Representation and Alignment

HowTo100M benchmark dataset

Category	Tasks	Videos	Clips
Food and Entertaining	11504	497k	54.4M
Home and Garden	5068	270k	29.5M
Hobbies and Crafts	4273	251k	29.8M
Cars & Other Vehicles	810	68k	7.8M
Pets and Animals	552	31k	3.5M
Holidays and Traditions	411	27k	3.0M
Personal Care and Style	181	16k	1.6M
Sports and Fitness	205	16k	2.0M
Health	172	15k	1.7M
Education and Communications	239	15k	1.6M
Arts and Entertainment	138	10k	1.2M
Computers and Electronics	58	5k	0.6M
Total	23.6k	1.22M	136.6M

https://www.di.ens.fr/willow/research/howto100m/

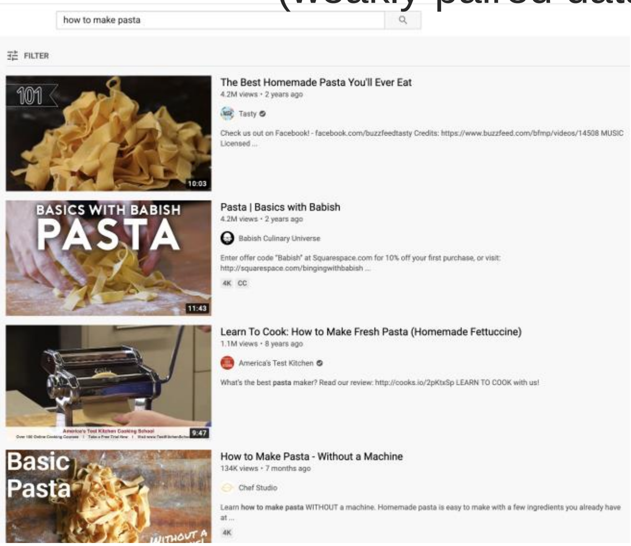
Visual Representations from Uncurated Instructional Videos

Goal: Learn better visual representations…

… by taking advantage of large-scale video+language resources

Instructional videos

(weakly-paired data)



it's turning into a much thicker mixture

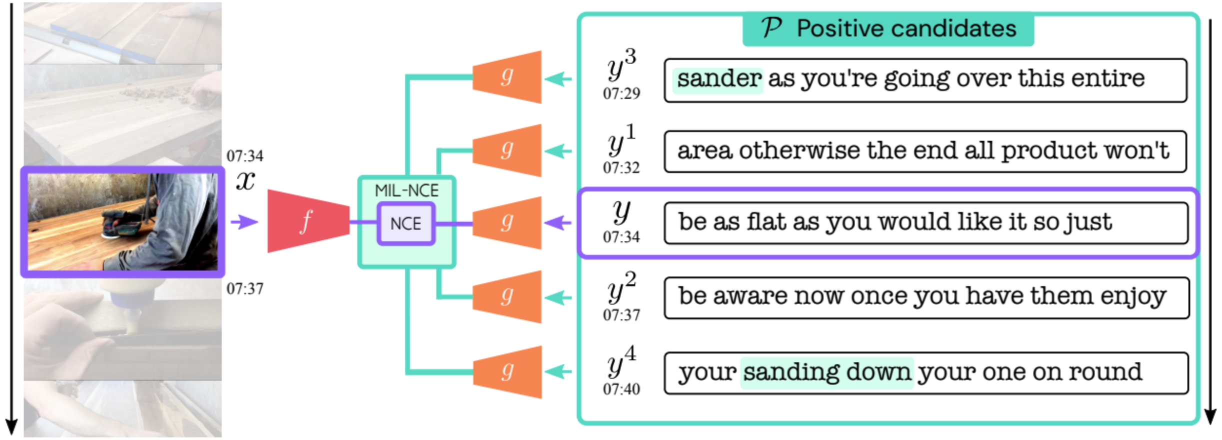
The biggest mistake is not kneading it enough

…

End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Weakly Paired Data

Data point: "a short 3.2 seconds video clip (32 frames at 10 FPS) together with a small number of words (not exceeding 16)"

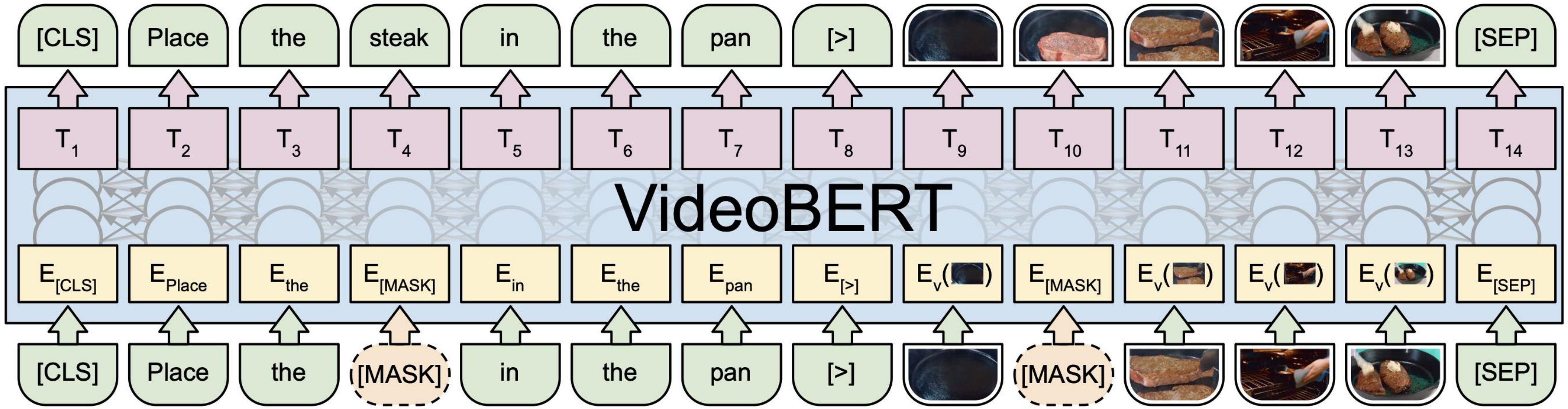


How to handle this misalignment? Multi-instance learning!

How to do it self-supervised? Contrastive learning!

End-to-End Learning of Visual Representations from Uncurated Instructional Videos
Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman – CVPR 2020

Another Approach for Weakly-Paired Video Data

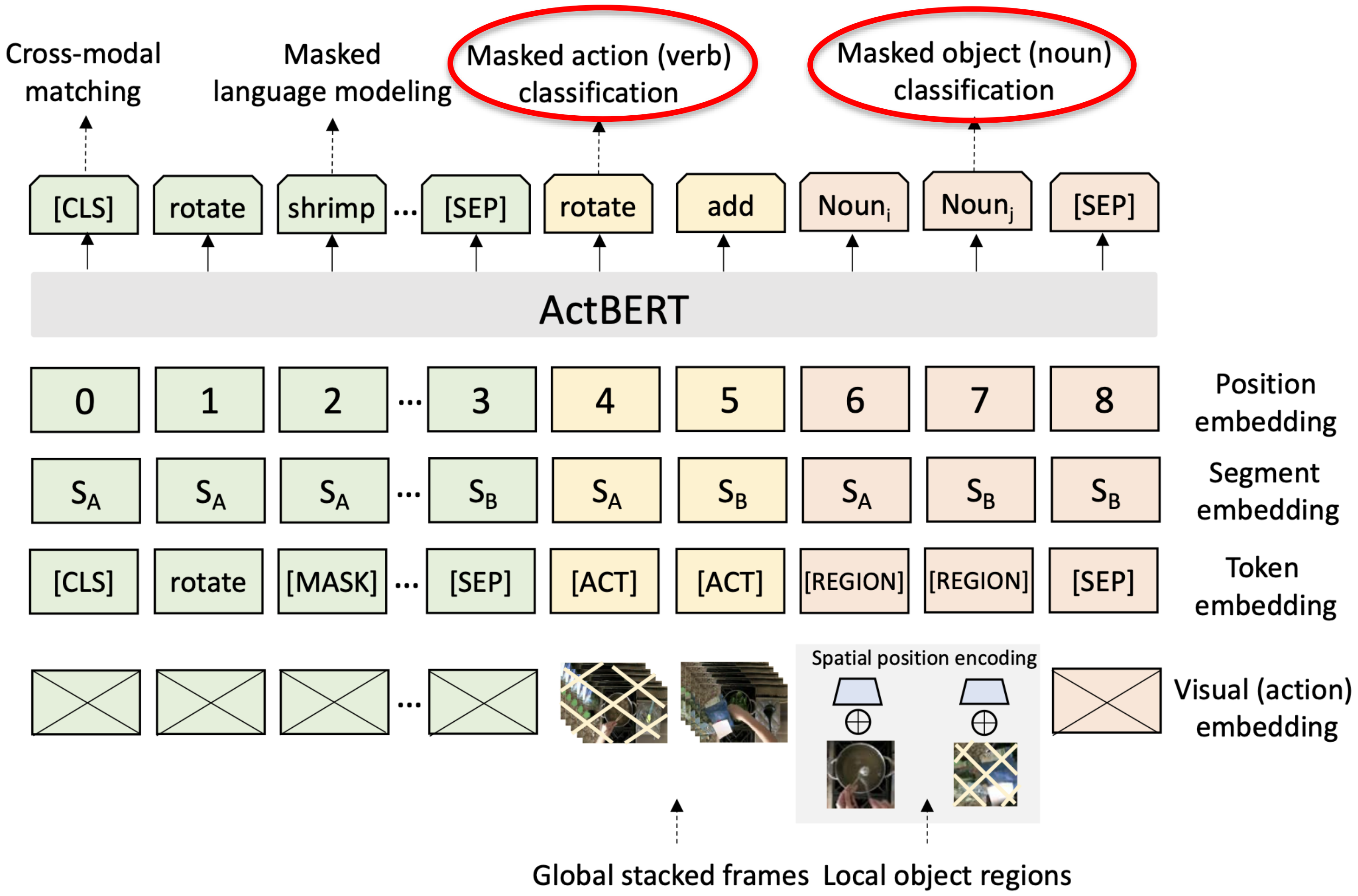


How do we get visual words now?

K-mean clustering + centroid

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, Cordelia Schmid; VideoBERT: A Joint Model for Video and Language Representation Learning ICCV, 2019

ActBERT



Zhu and Yang, ActBERT: Learning Global-Local Video-Text Representations, CVPR 2020


"That's all Folks!"