

 APPLICABLE ML

Understanding Transformer From Computer Vision Perspective

Sangbum Daniel Choi

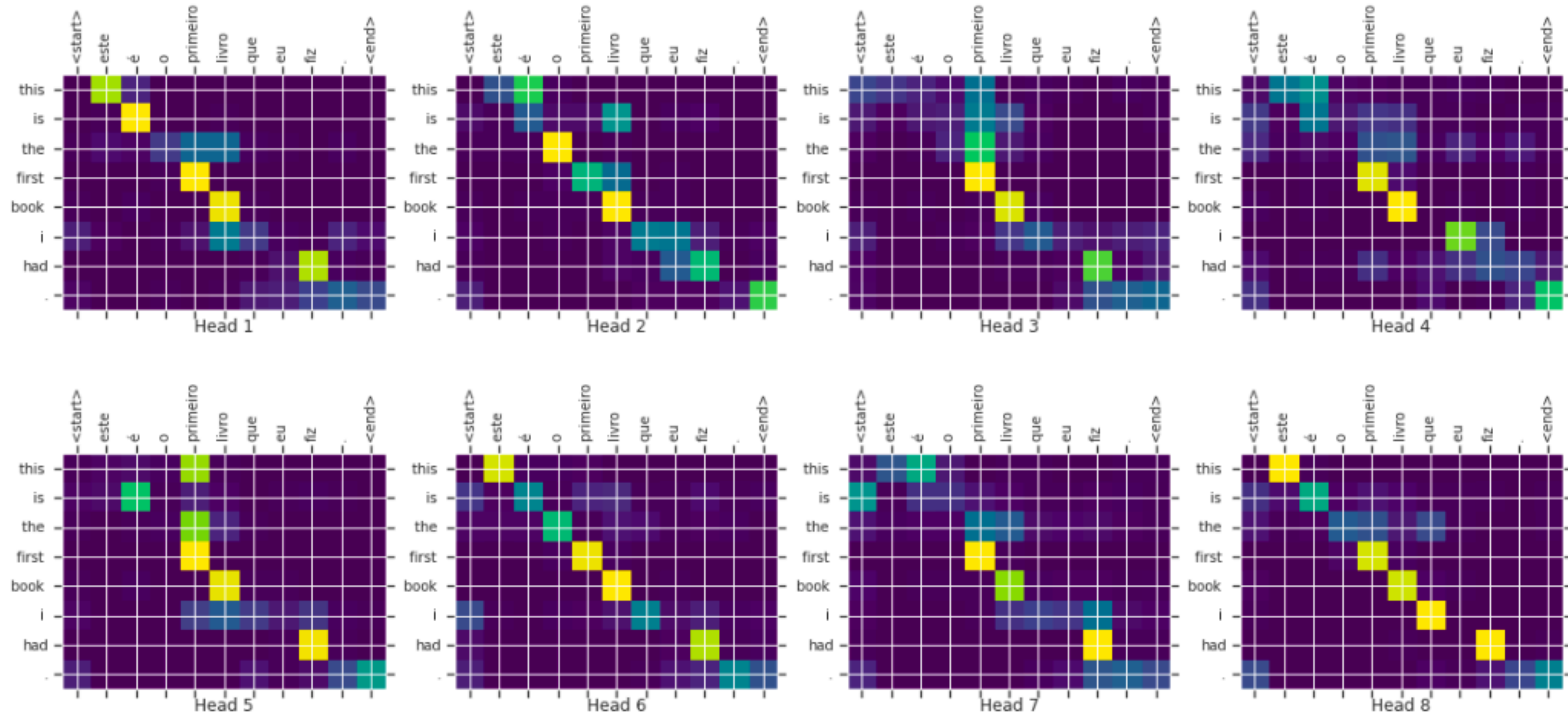
Transformer?

The word “**Transformer**” is originated from the paper ‘Attention Is All You Need’ NIPS 2017 paper.

As the quotation in paper ‘We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely’.

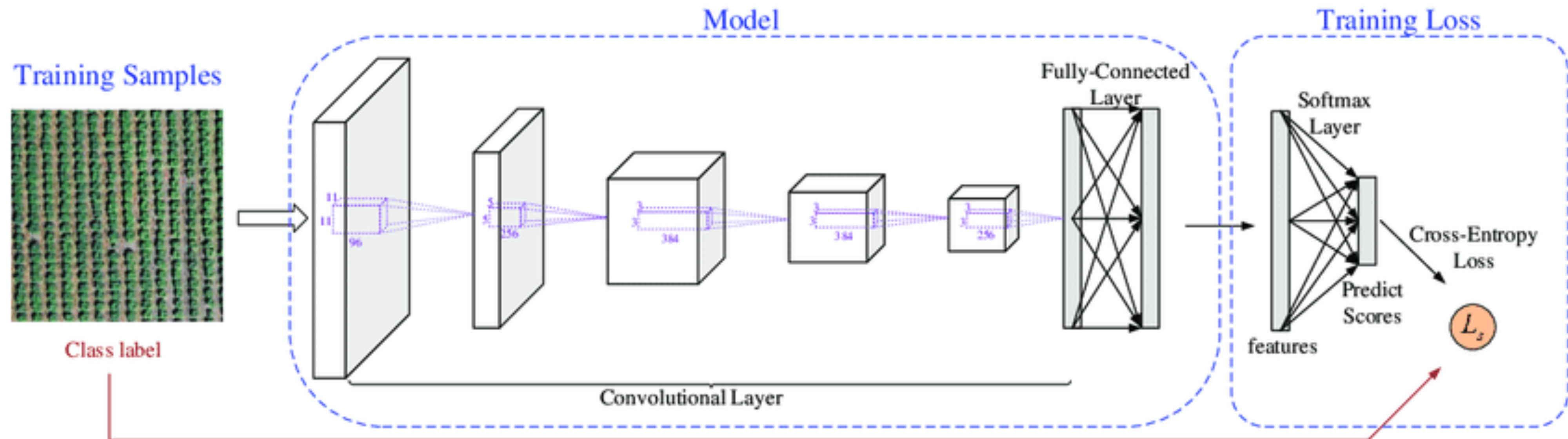
Transformers are usually used in Natural Language Processing (NLP) domain due to concept of attention.

Attention



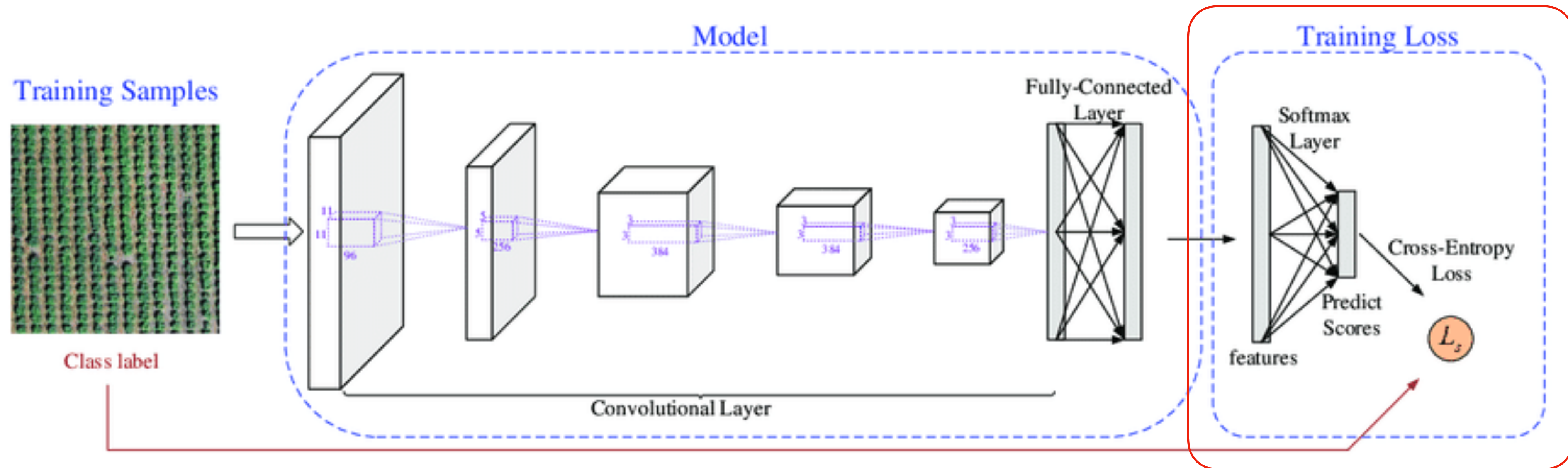
In translation problem, attention is used to look between 'words' to make more sparse feature to look at.

Attention

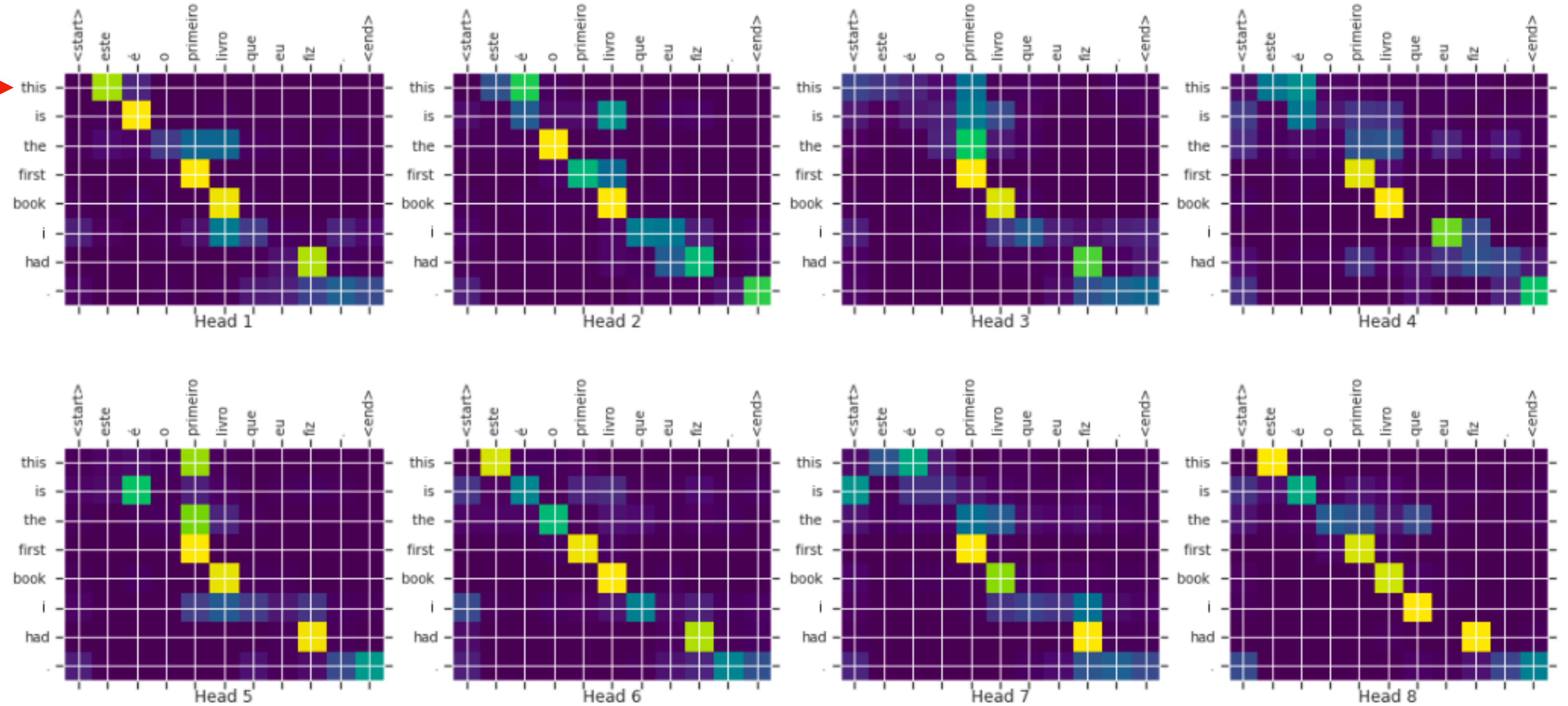
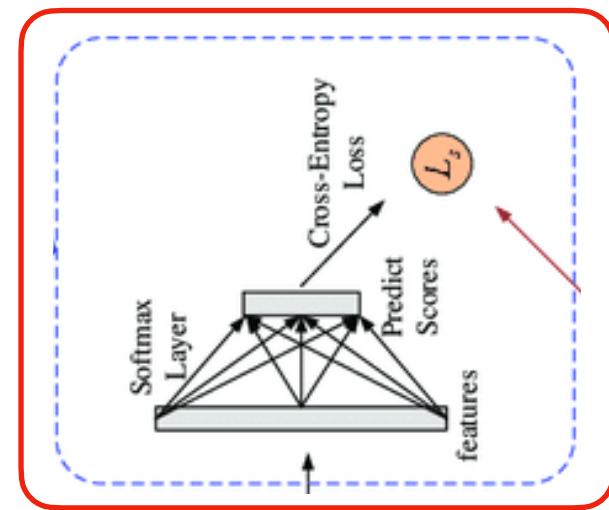


Normal procedure of classification problem using CNNs.

Attention

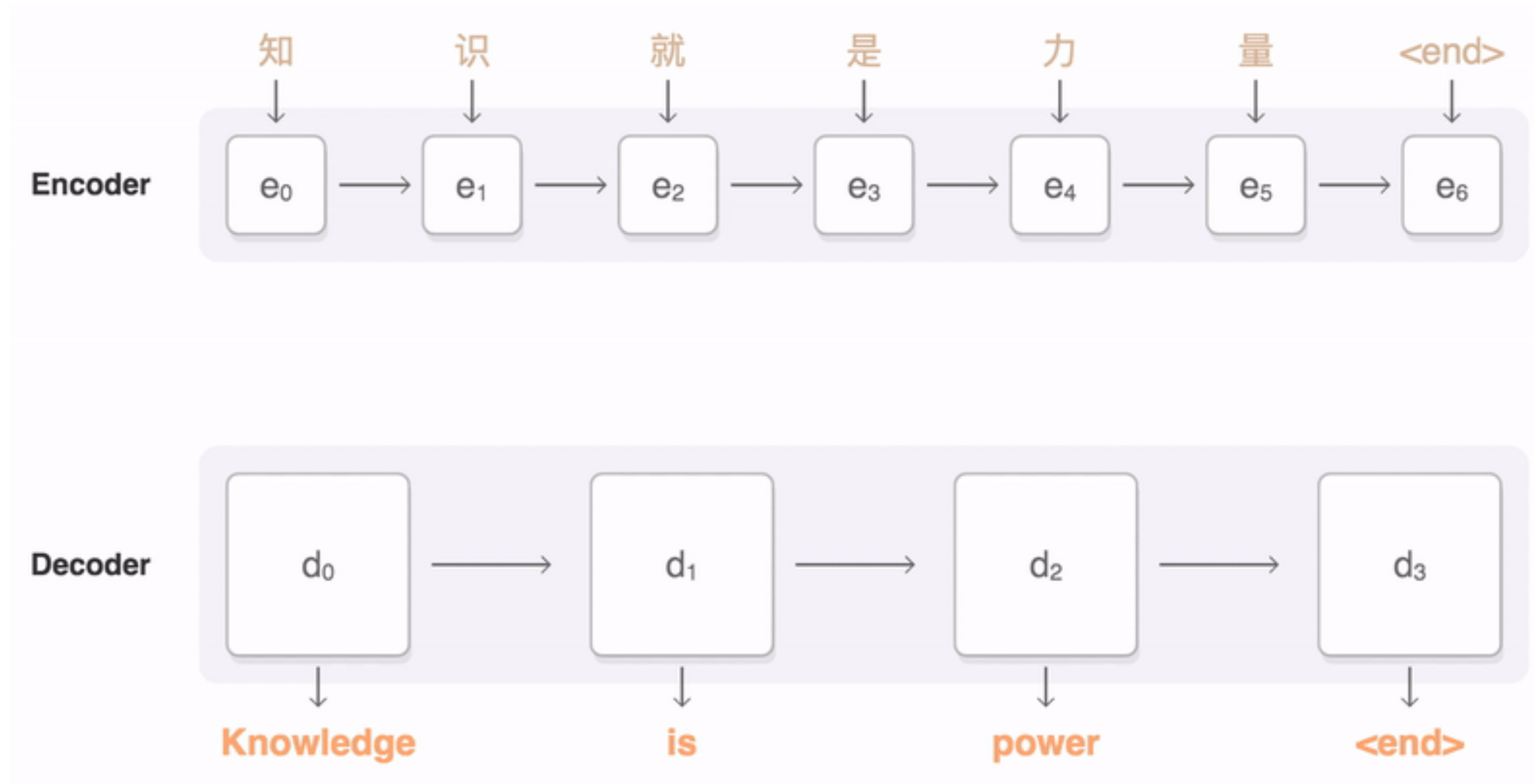


Attention



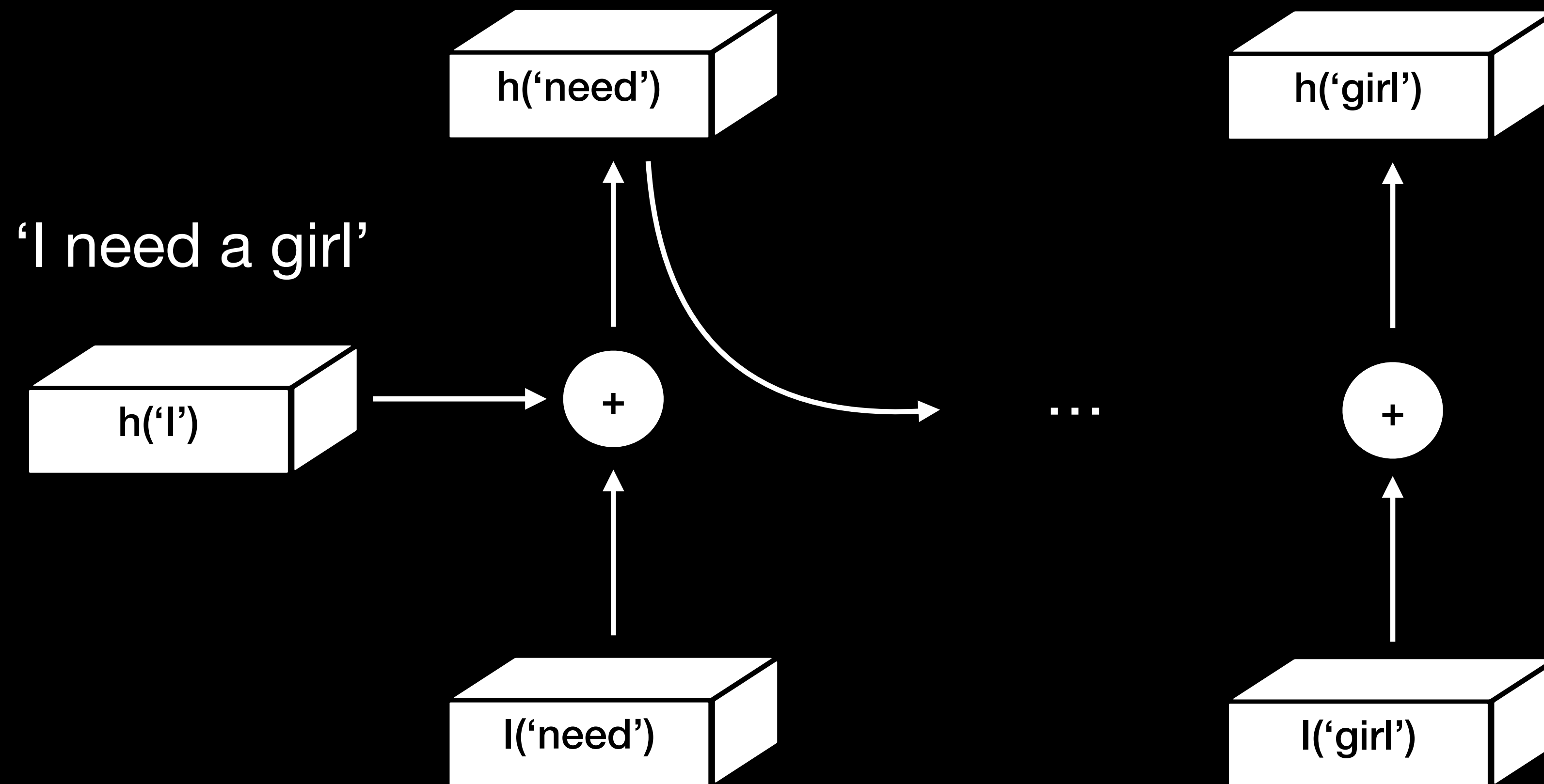
End of the classification training procedure in computer vision is very similar to one attention layer in Transformer.

Seq2Seq



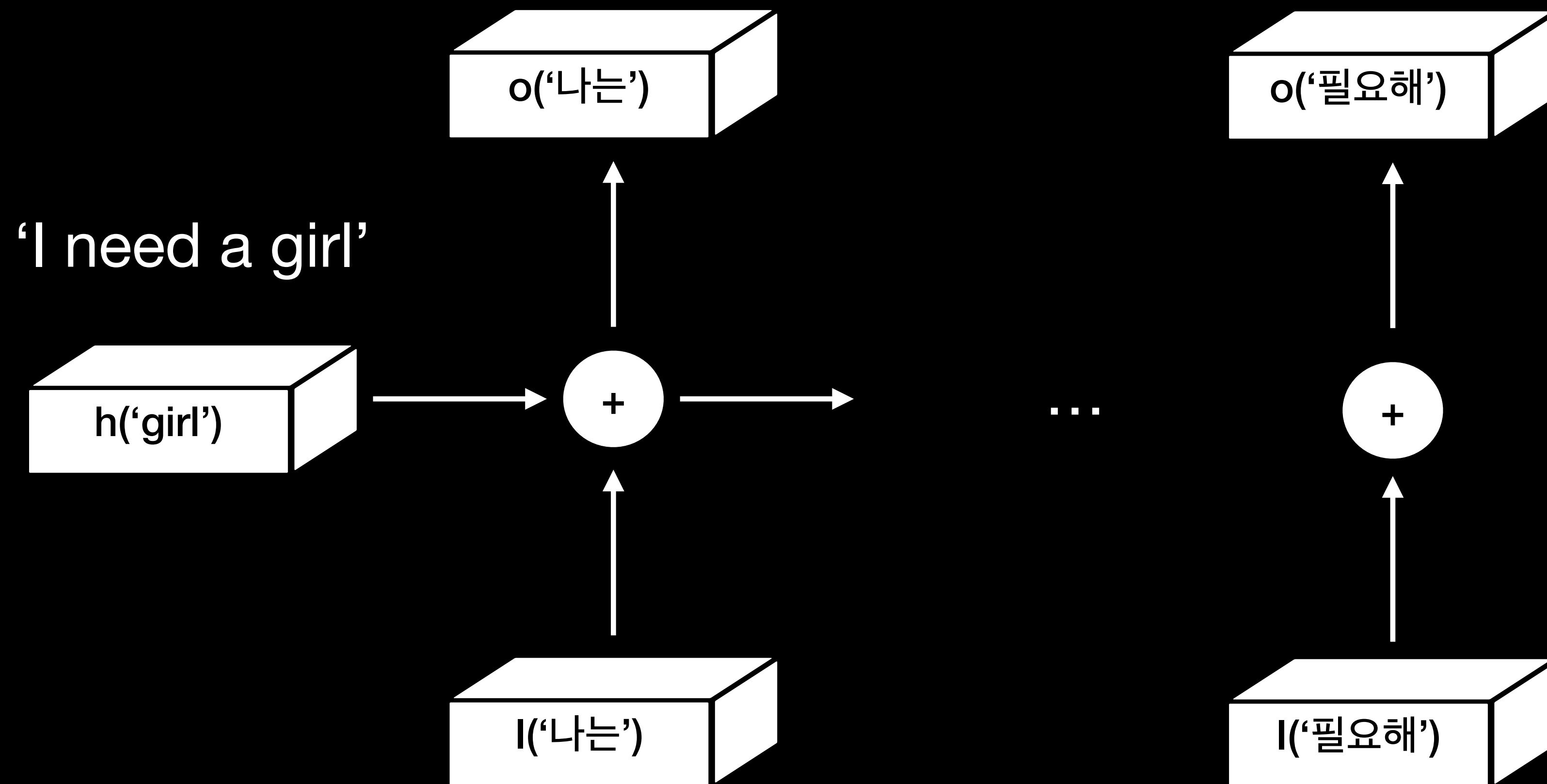
Cons of Seq2Seq is getting worse as the input sequence gets larger.

Seq2Seq - encoder



Can't use hidden space vector at 't-2' in progress of time step 't'

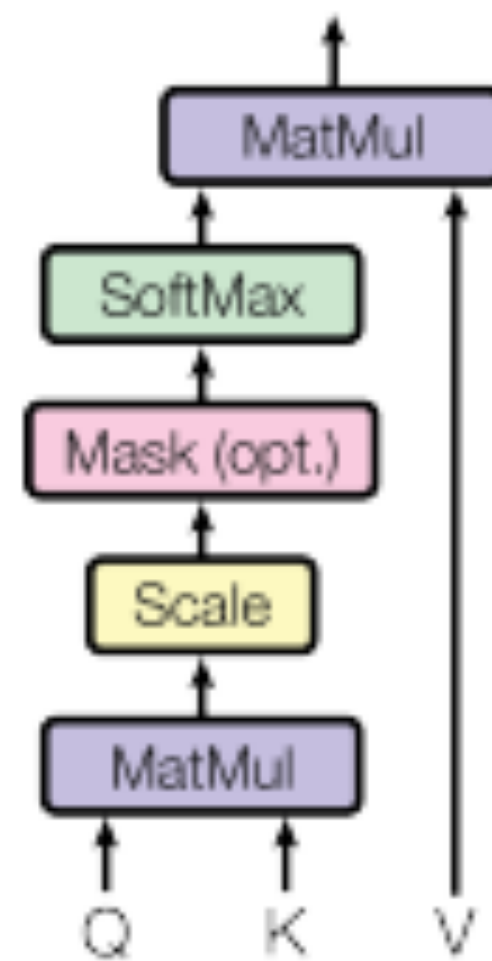
Seq2Seq - decoder



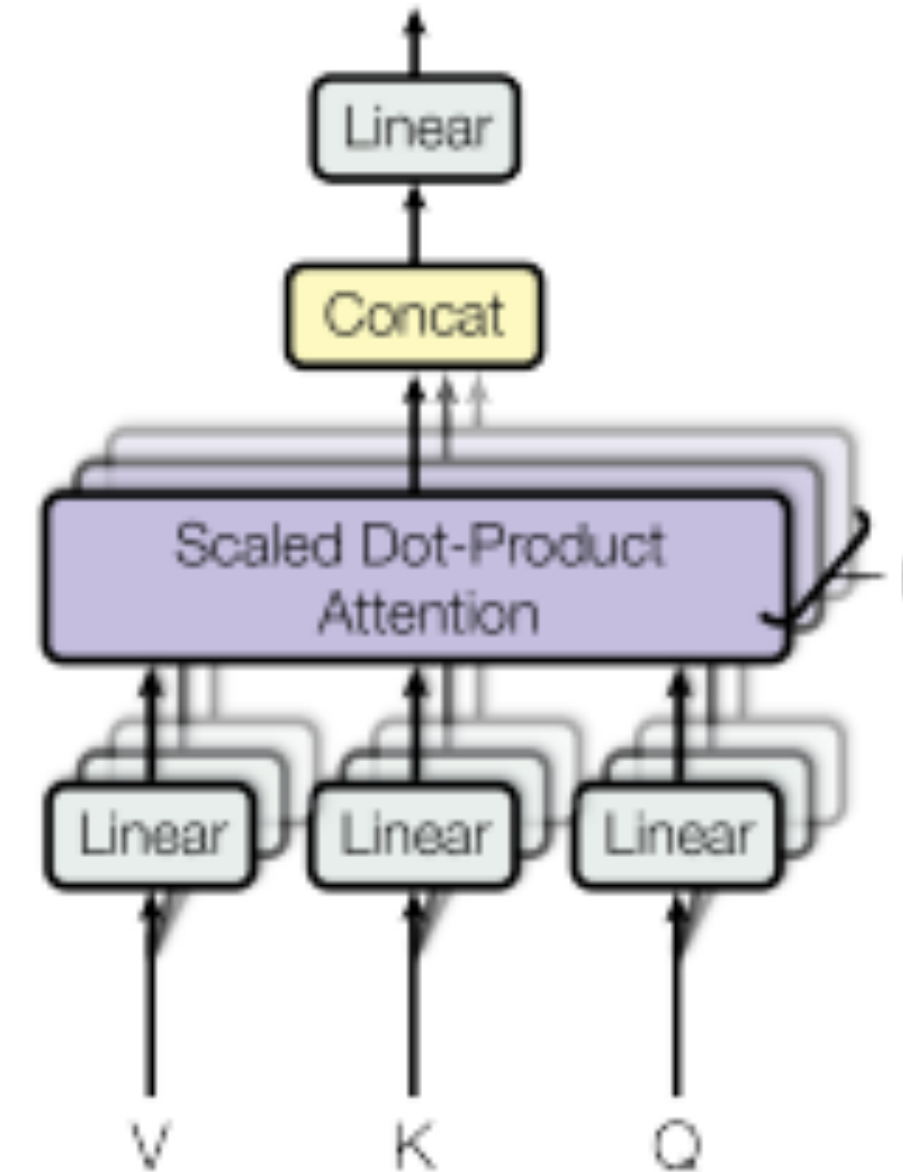
Can't use hidden space vector at 't-2' in progress of time step 't'

Attention

Scaled Dot-Product Attention

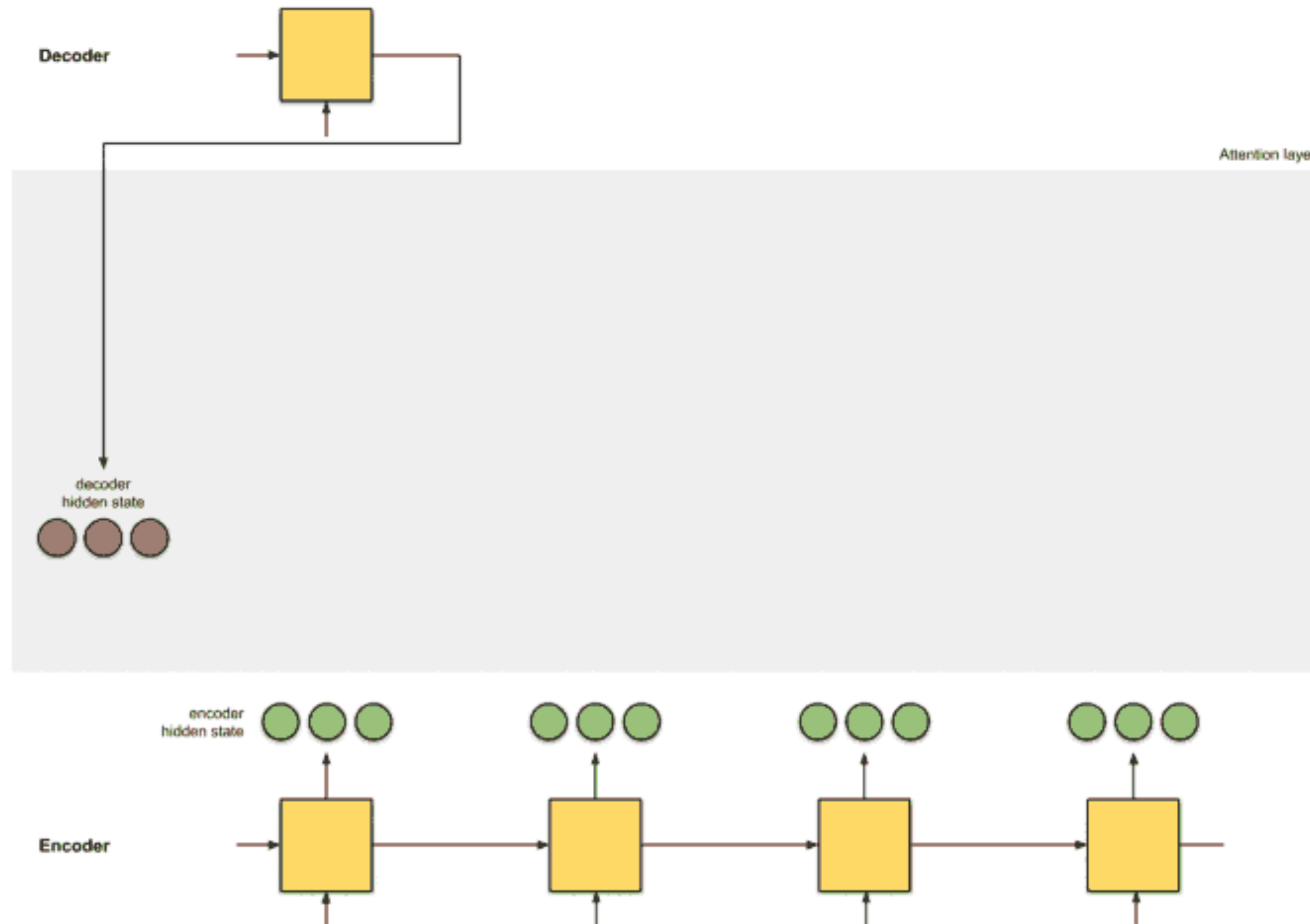


Multi-Head Attention

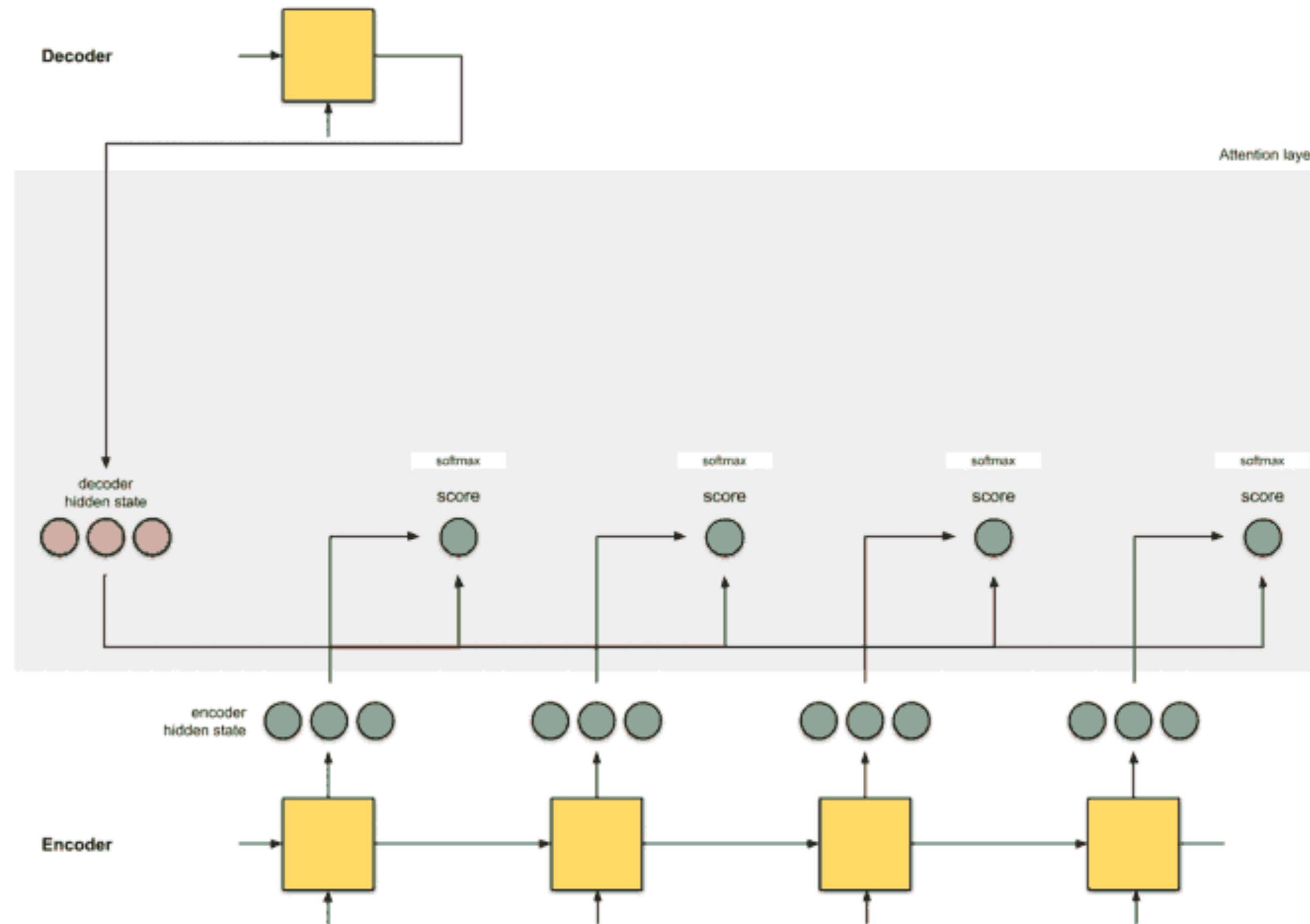


Key, Query, Value is from the concept of retrieval system.

Seq2Seq + Attention \approx Transformer



Seq2Seq + Attention \approx Transformer











Key, Query, Value is a form of vector like hidden space output and diverse by different modeling (like GPT3, BERT, etc...)

Seq2Seq + Attention \neq Transformer

There are so many additional factors that describes transformer

- Positional Embedding
- Position-wise Feed-Forward Networks
- Self Attention

Vision Transformer

Rank	Model	Top 1 Accuracy 	Top 5 Accuracy	Number of params	Extra Training Data	Paper	Code	Result	Year	Tags 
1	CoAtNet-7	90.88%		2440M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes			2021	Conv+Transformer JFT-3B
2	ViT-G/14	90.45%		1843M	✓	Scaling Vision Transformers			2021	Transformer JFT-3B
3	CoAtNet-6	90.45%		1470M	✓	CoAtNet: Marrying Convolution and Attention for All Data Sizes			2021	Conv+Transformer JFT-3B
4	ViT-MoE-15B (Every-2)	90.35%		14700M	✓	Scaling Vision with Sparse Mixture of Experts			2021	Transformer JFT-3B

Vision Transformer

Can't explain the phenomena of '**inductive bias**' that happens in CNN.

Look out for more detail

http://cs231n.stanford.edu/slides/2021/lecture_11.pdf

Vision Transformer

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (Vision in Transformer) (ICLR 2021)

BEiT: BERT Pre-Training of Image Transformers (Arxiv)

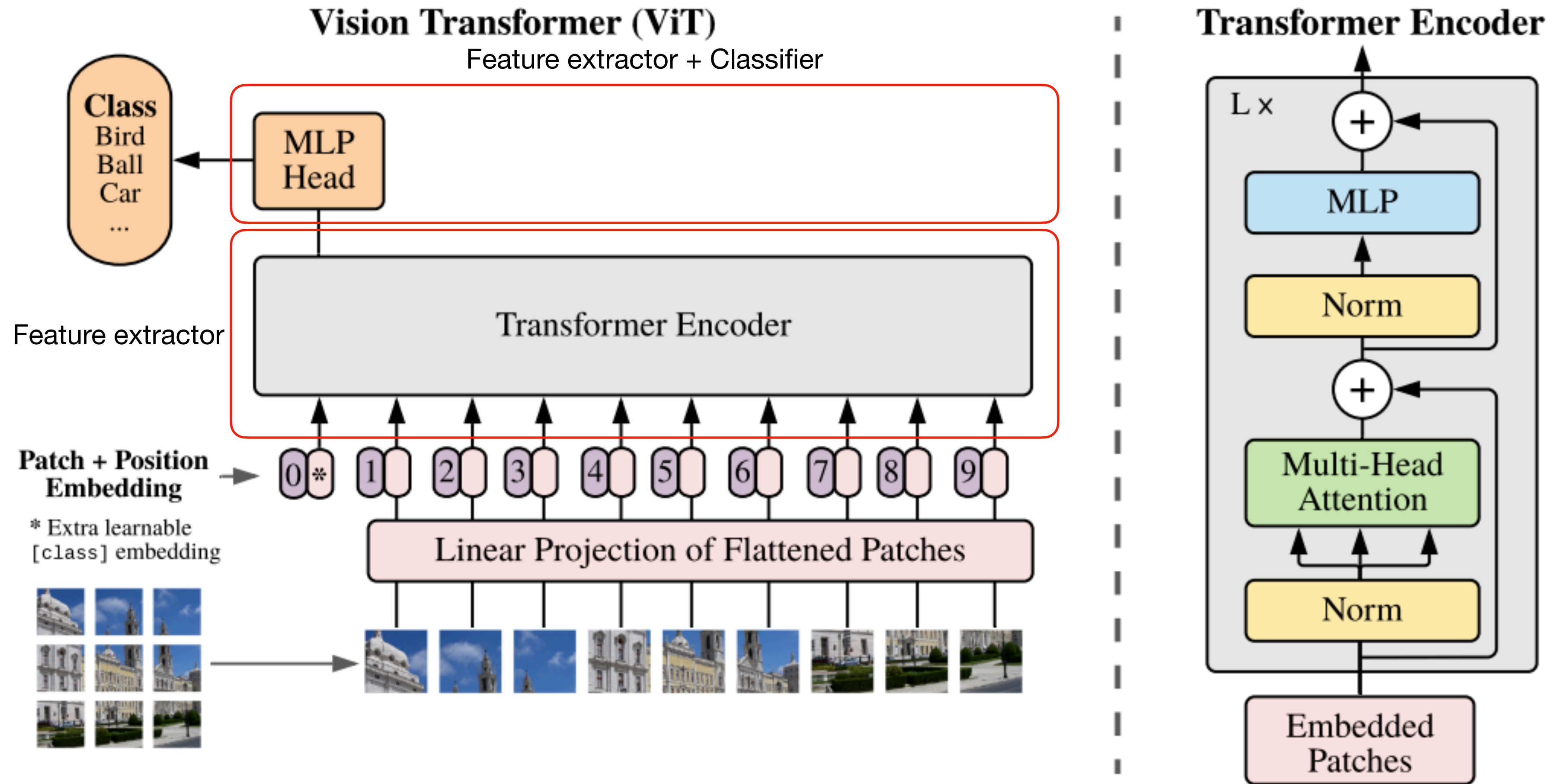
Swin Transformer : Hierarchical Vision Transformer using Shifted Windows (ICCV 2021 Best paper)

MobileViT: Light-weight, General-purpose, and Mobile friendly Vision Transformer (Arxiv)

AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ViT)

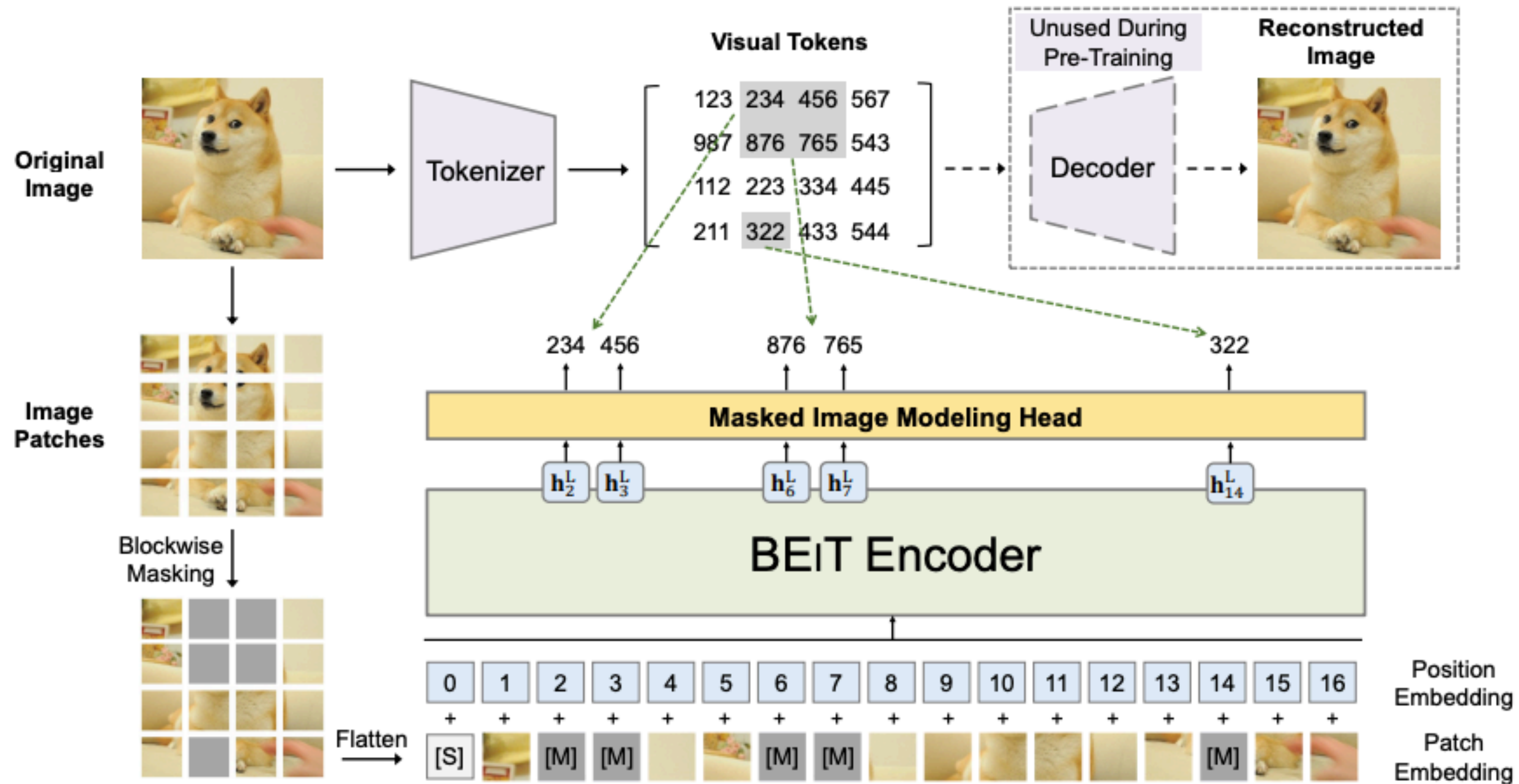


AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE (ViT)



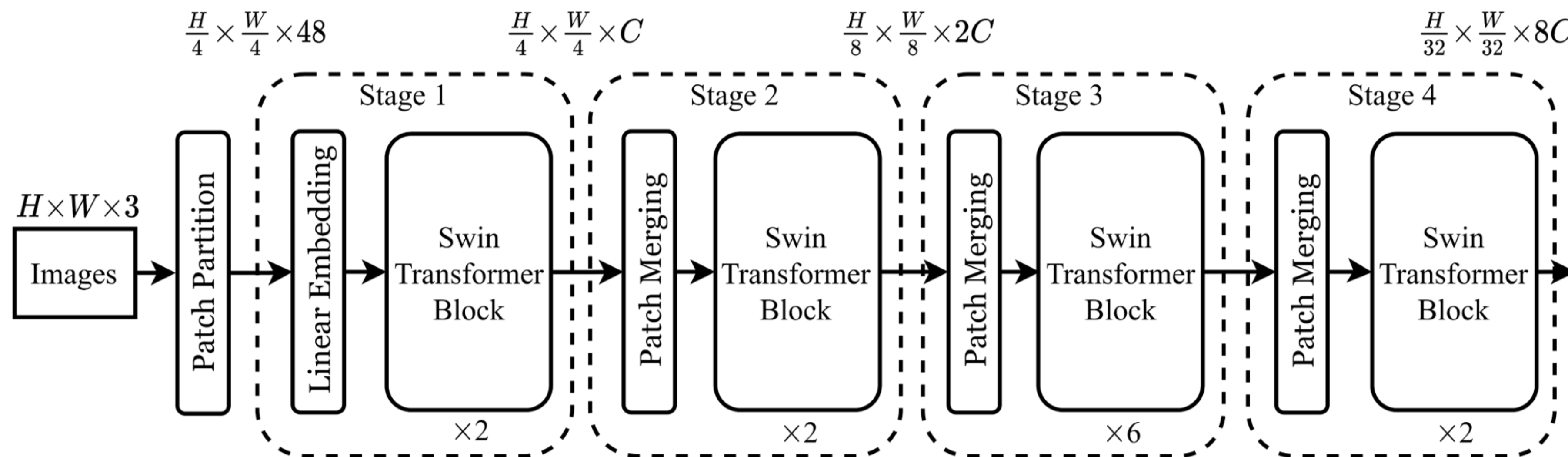
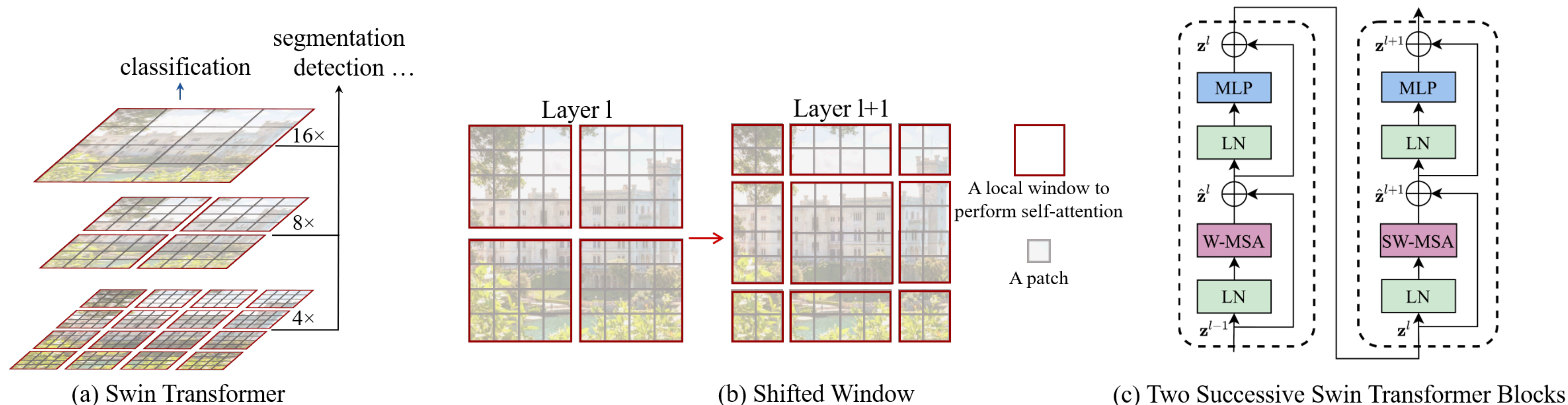
They used in house JFT-300M dataset which are way more larger than ImageNet-21k dataset.

BEIT: BERT Pre-Training of Image Transformers



Self-supervised vision representation model. However, decoder to use in various tasks and Visual Tokens need to be trained on supervised manner (COCO or DALL-E).

Swin Transformer : Hierarchical Vision Transformer using Shifted Windows



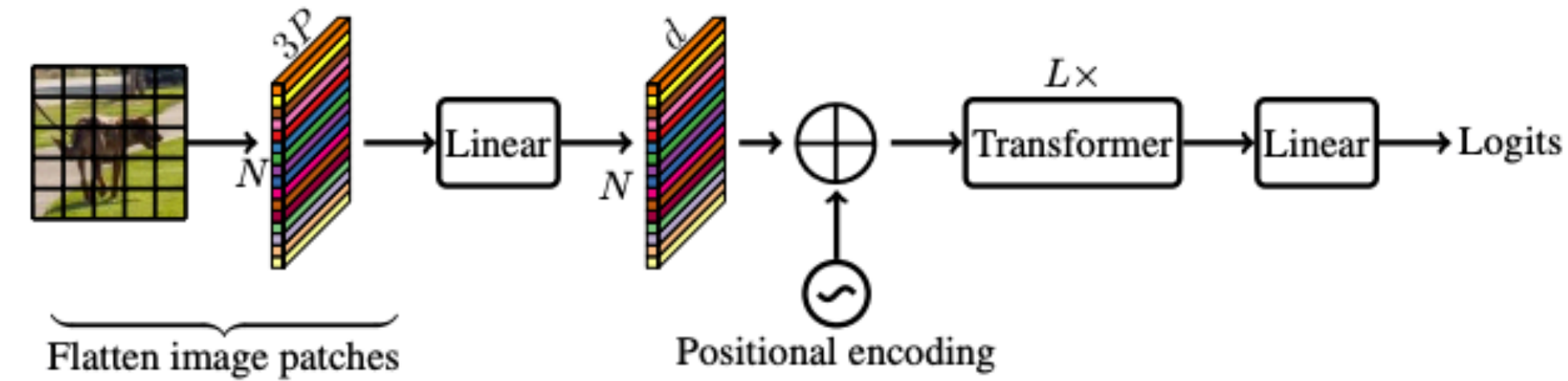
(d) Architecture

Swin Transformer : Hierarchical Vision Transformer using Shifted Windows

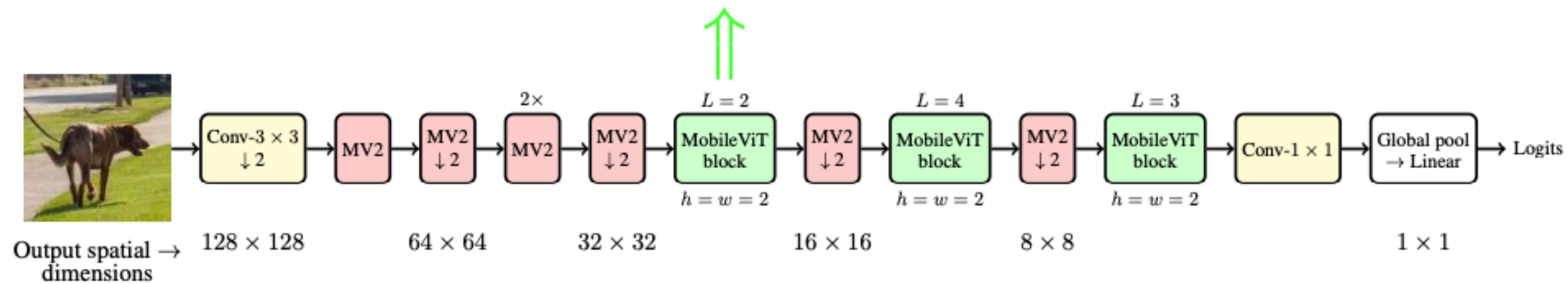
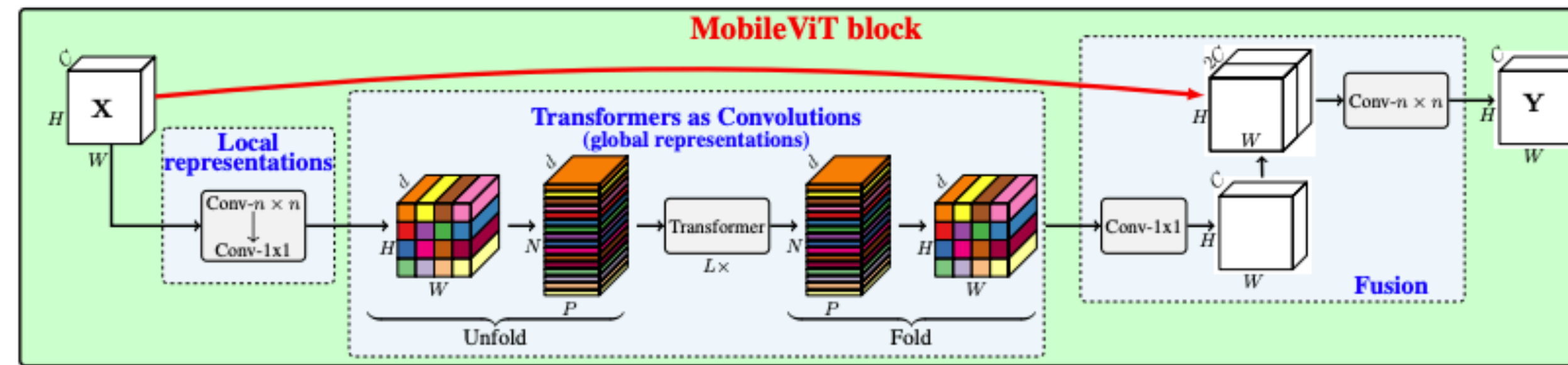
Method	mini-val		test-dev		#param. FLOPs	
	AP ^{box}	AP ^{mask}	AP ^{box}	AP ^{mask}		
RepPointsV2* [12]	-	-	52.1	-	-	-
GCNet* [7]	51.8	44.7	52.3	45.4	-	1041G
RelationNet++* [13]	-	-	52.7	-	-	-
SpineNet-190 [21]	52.6	-	52.8	-	164M	1885G
ResNeSt-200* [78]	52.5	-	53.3	47.1	-	-
EfficientDet-D7 [59]	54.4	-	55.1	-	77M	410G
DetectoRS* [46]	-	-	55.7	48.5	-	-
YOLOv4 P7* [4]	-	-	55.8	-	-	-
Copy-paste [26]	55.9	47.2	56.0	47.4	185M	1440G
X101-64 (HTC++)	52.3	46.0	-	-	155M	1033G
Swin-B (HTC++)	56.4	49.1	-	-	160M	1043G
Swin-L (HTC++)	57.1	49.5	57.7	50.2	284M	1470G
Swin-L (HTC++)*	58.0	50.4	58.7	51.1	284M	-

Table 2. Results on COCO object detection and instance segmentation. † denotes that additional decovolution layers are used to produce hierarchical feature maps. * indicates multi-scale testing.

MOBILEViT: LIGHT-WEIGHT, GENERAL-PURPOSE, AND MOBILE-FRIENDLY VISION TRANSFORMER



(a) Standard visual transformer (ViT)



(b) MobileViT. Here, **Conv- $n \times n$** in the MobileViT block represents a standard $n \times n$ convolution and **MV2** refers to MobileNetv2 block. Blocks that perform down-sampling are marked with $\downarrow 2$.