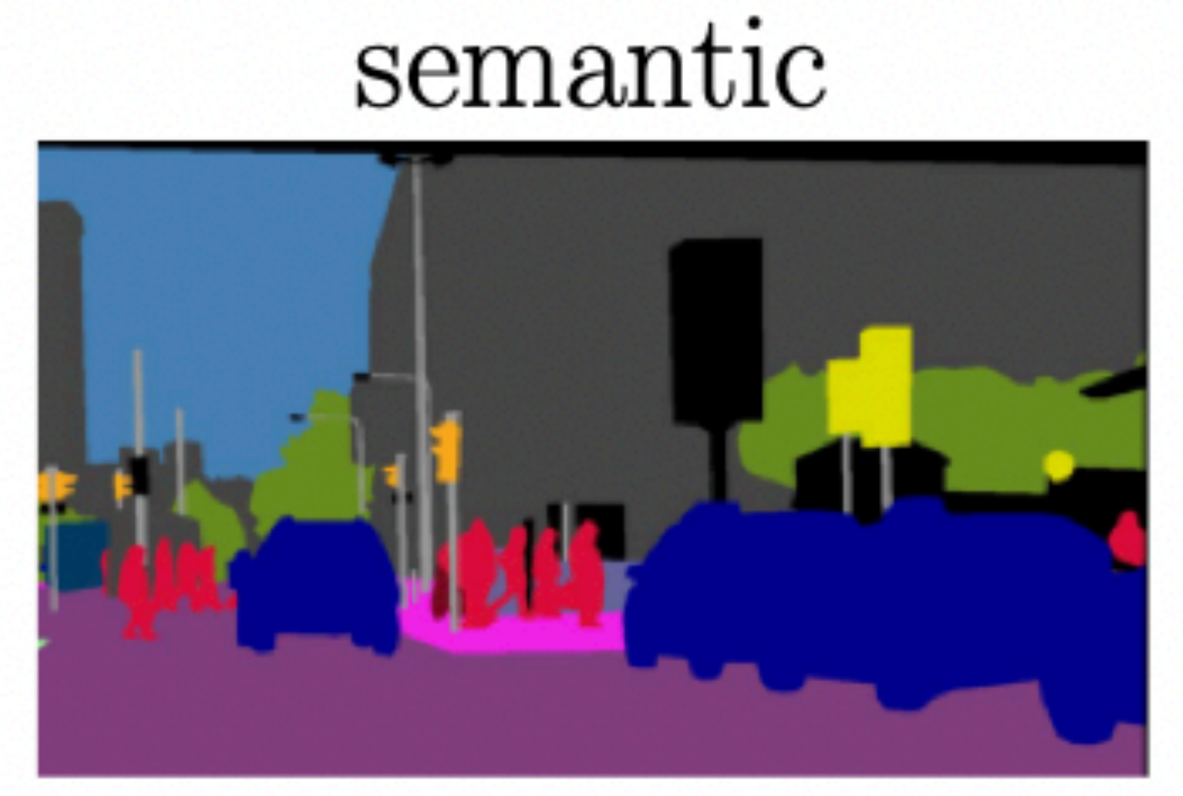
Panoptic Segmentation with Transformers Tutorial

Mennatullah Siam

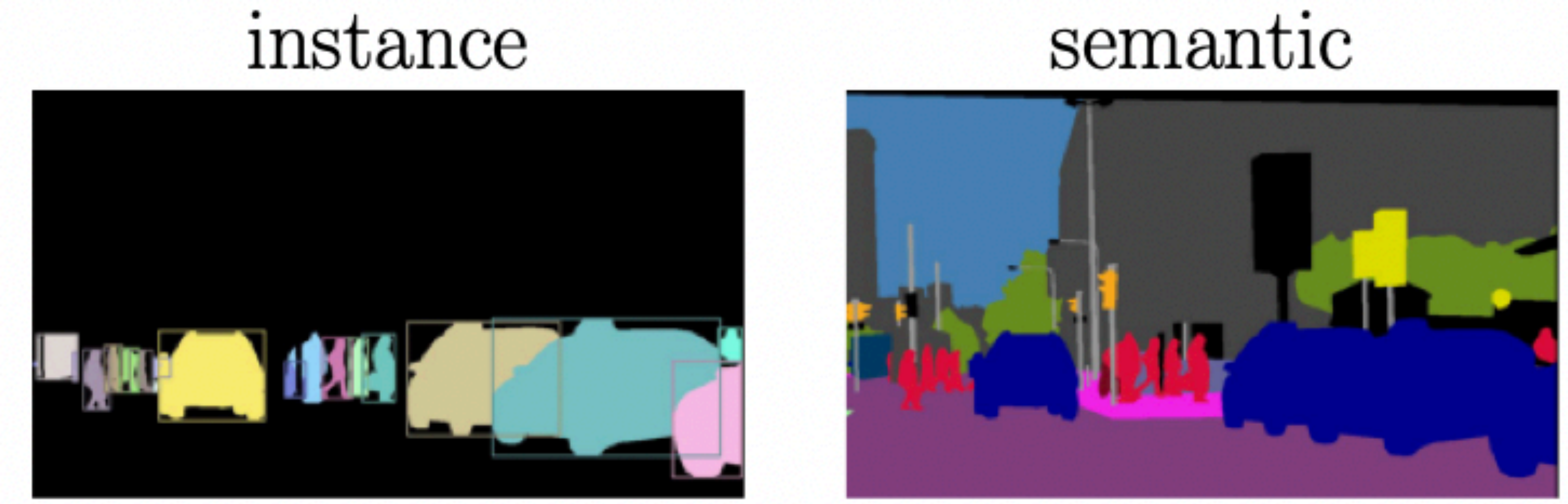
Accra, Ghana

9 September, 2023

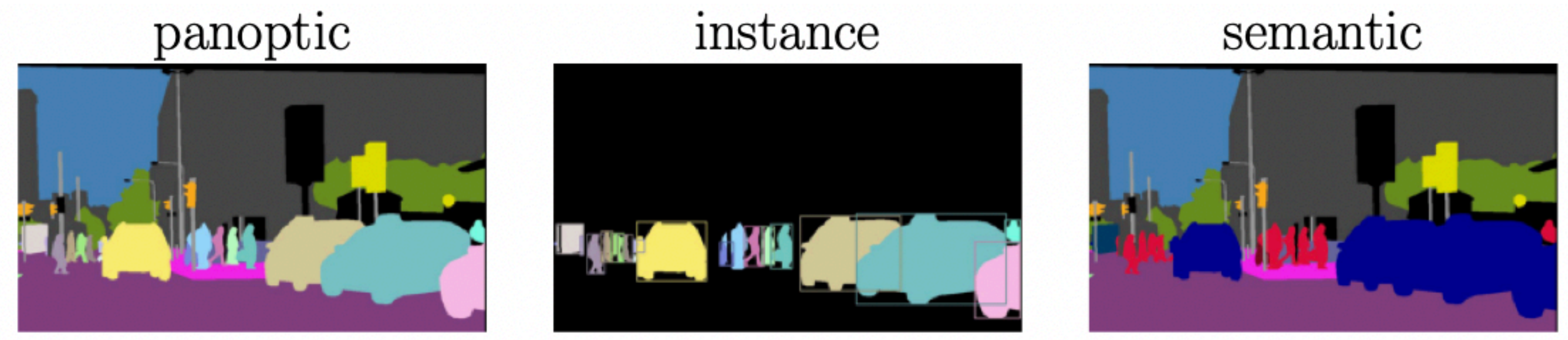
Semantic/Instance/Panoptic Segmentation



Semantic/Instance/Panoptic Segmentation



Semantic/Instance/Panoptic Segmentation



Holistic Scene Understanding

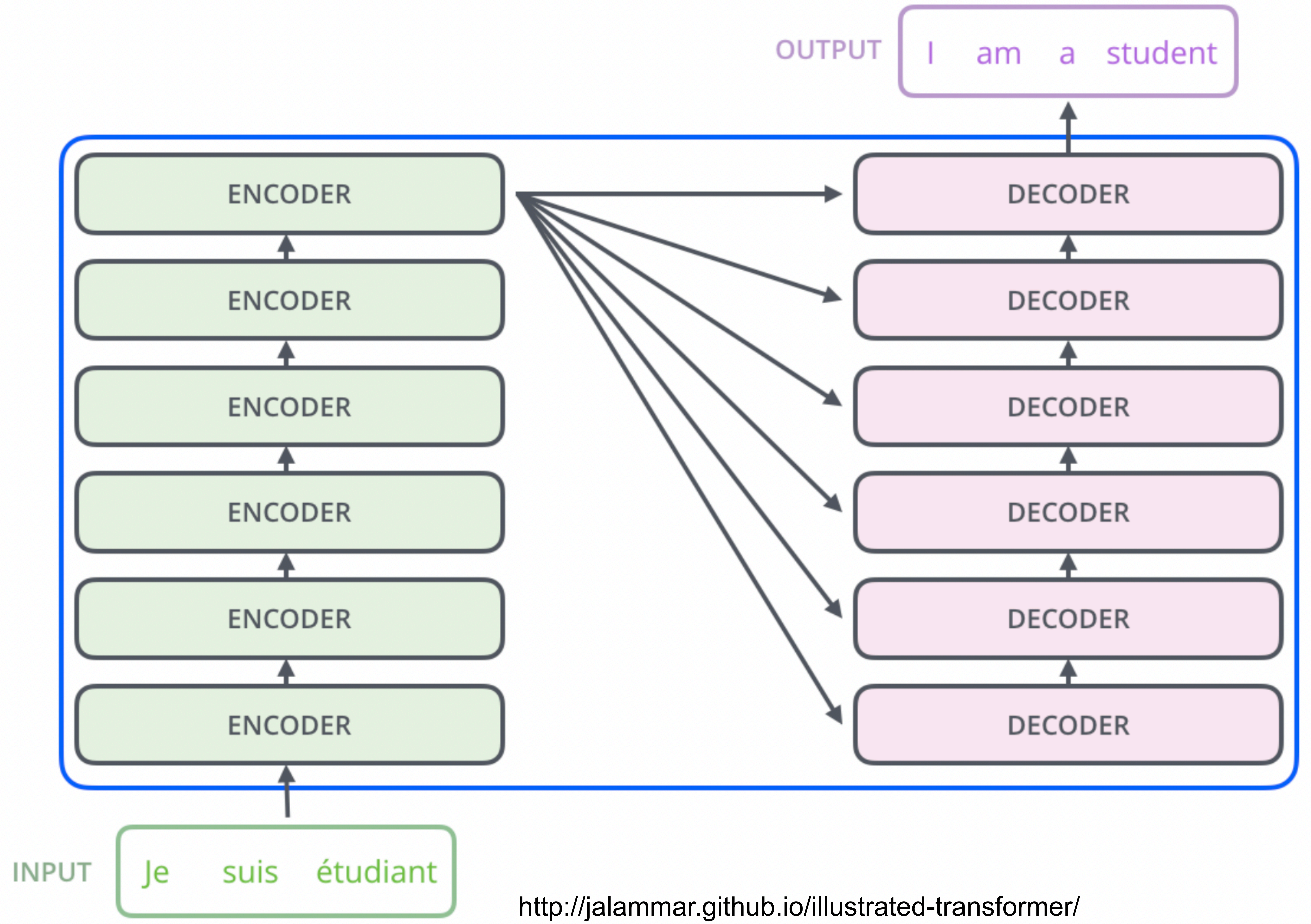
Transformers

Example

http://jalammar.github.io/illustrated-transformer/

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

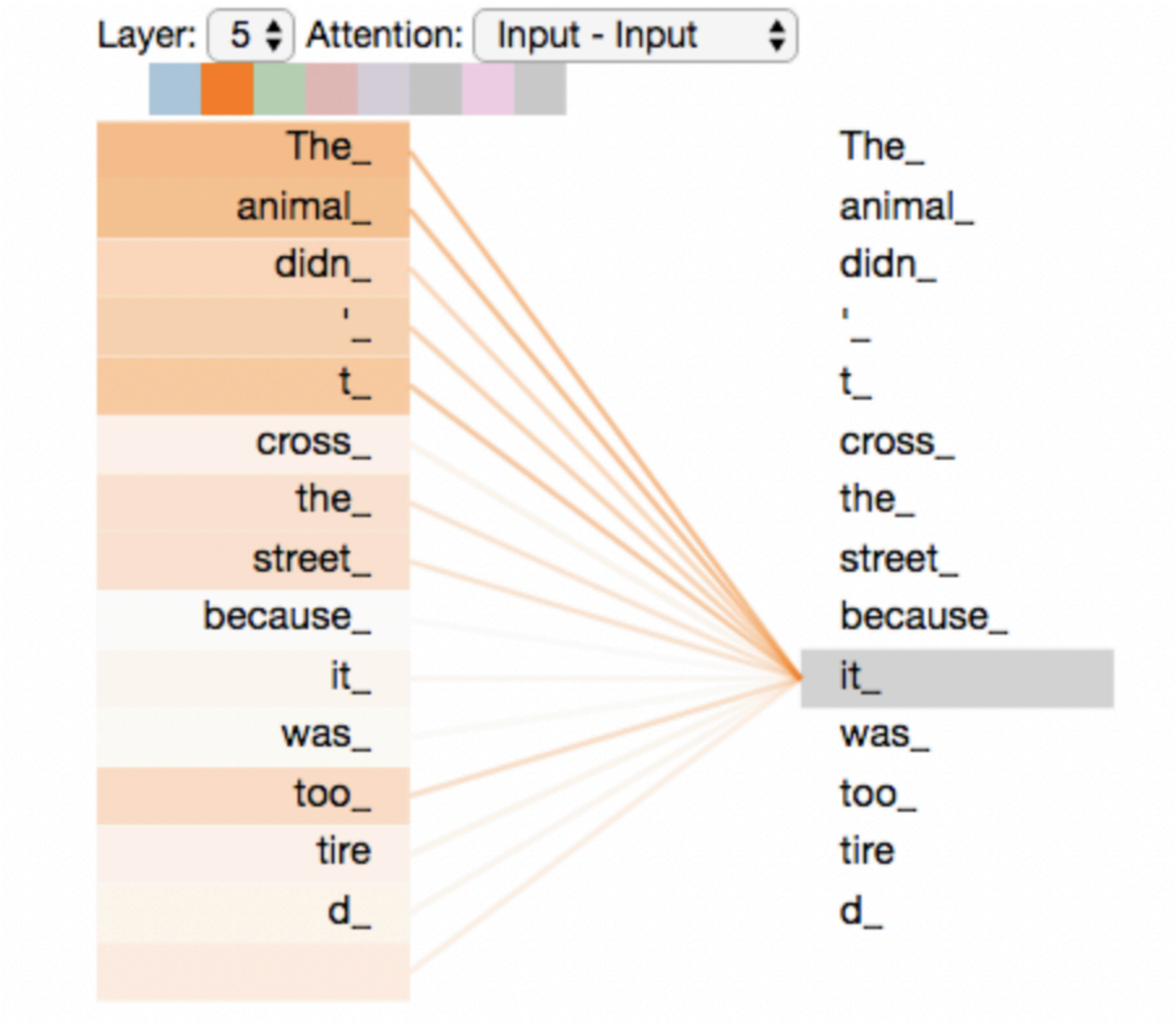
Transformers



http://jalammar.github.io/illustrated-transformer/

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

Self Attention

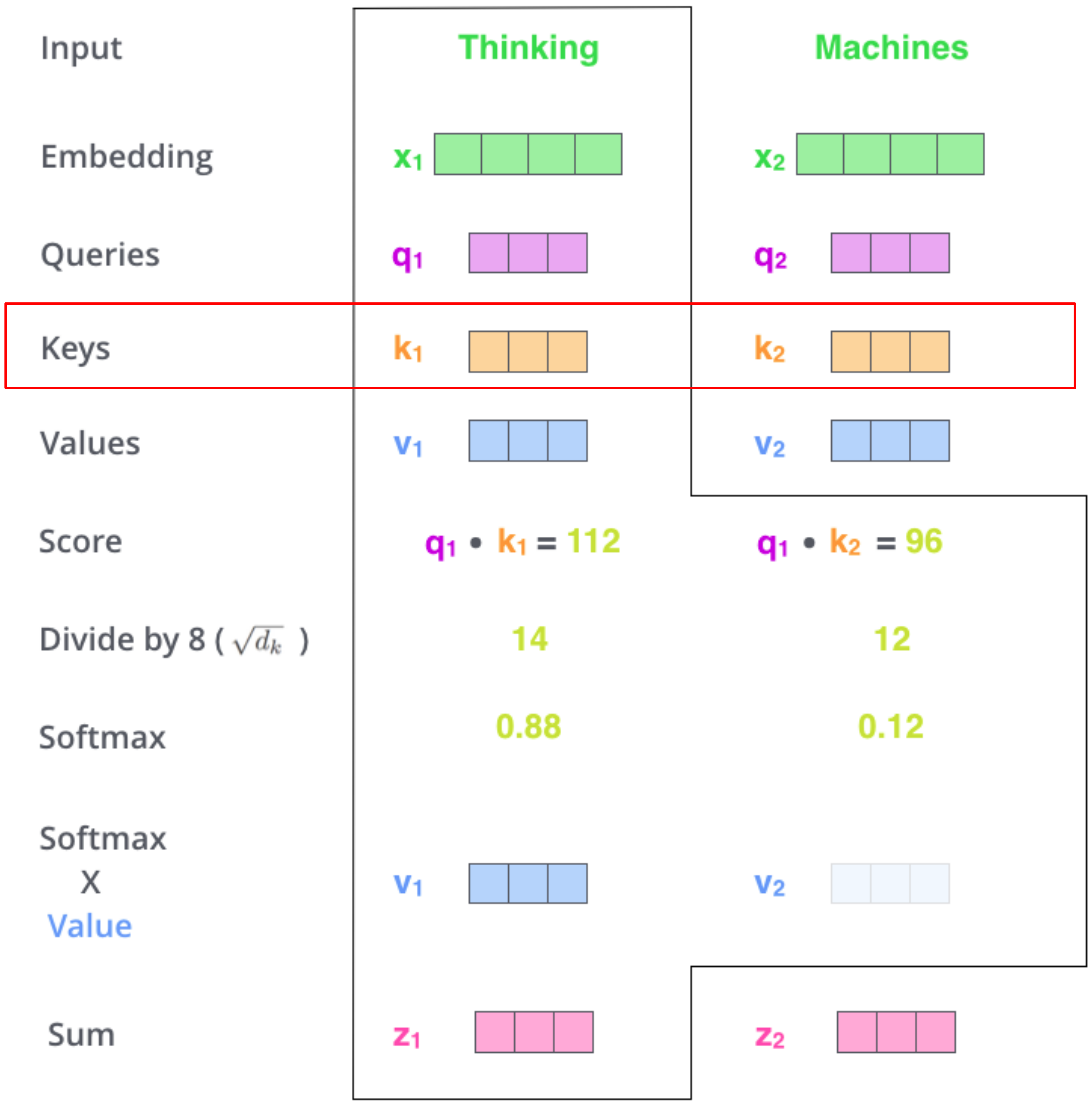


http://jalammar.github.io/illustrated-transformer/

Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

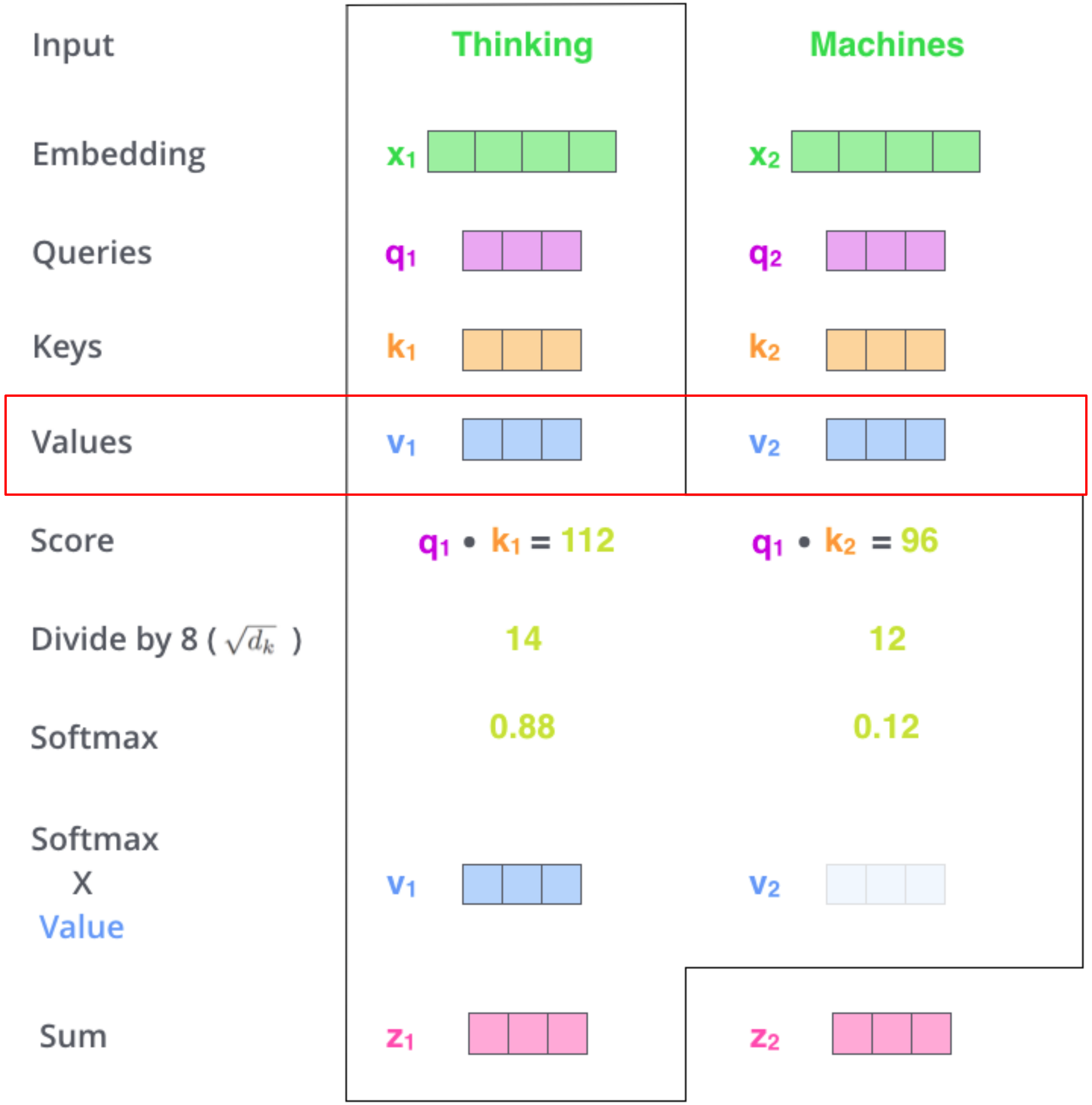
Think of this as your Dictionary keys used for addressing

Single Head Attention



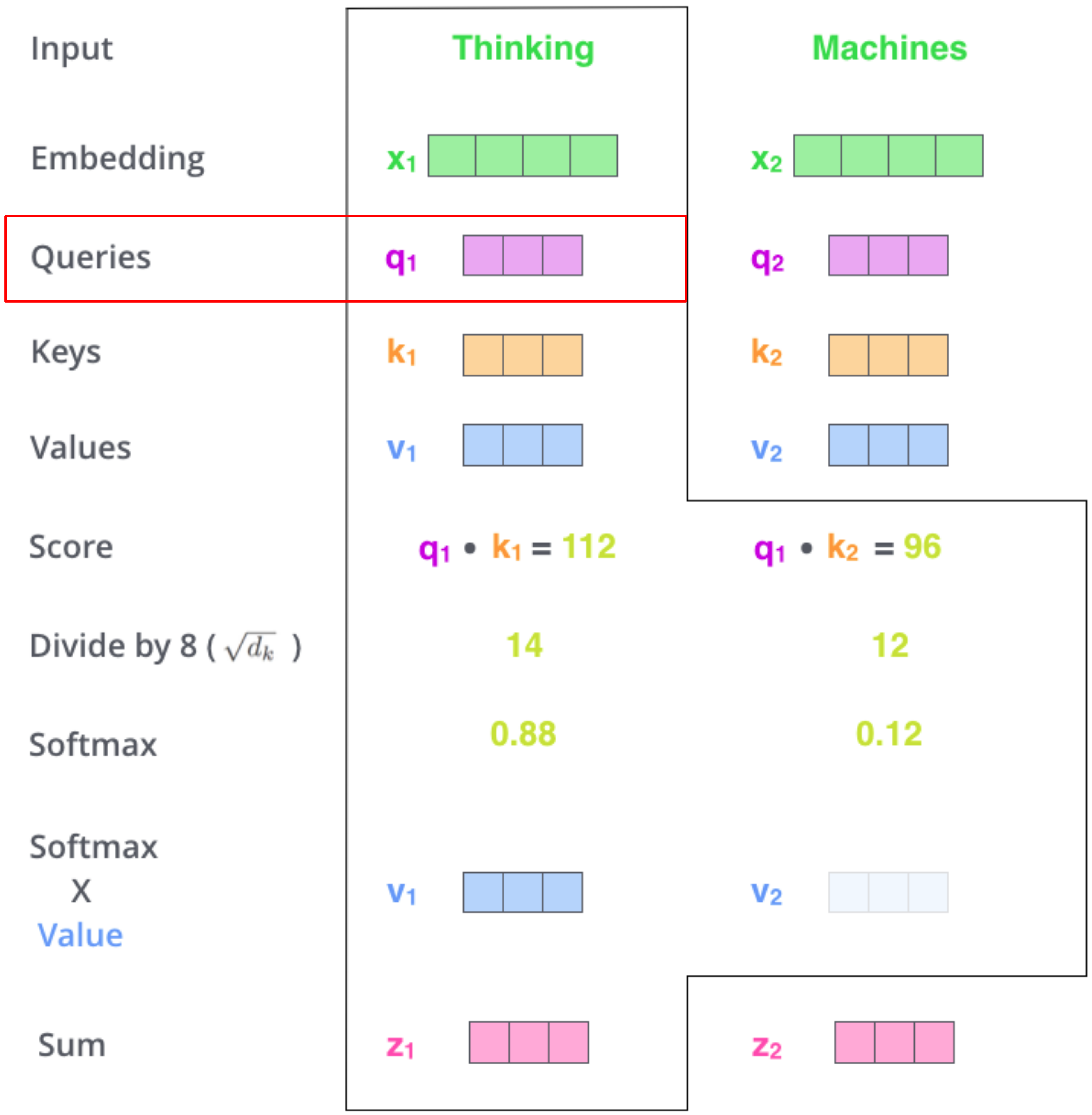
http://jalammar.github.io/illustrated-transformer/

Detailed information - what I want to read out from memory



http://jalammar.github.io/illustrated-transformer/

My current word embedding

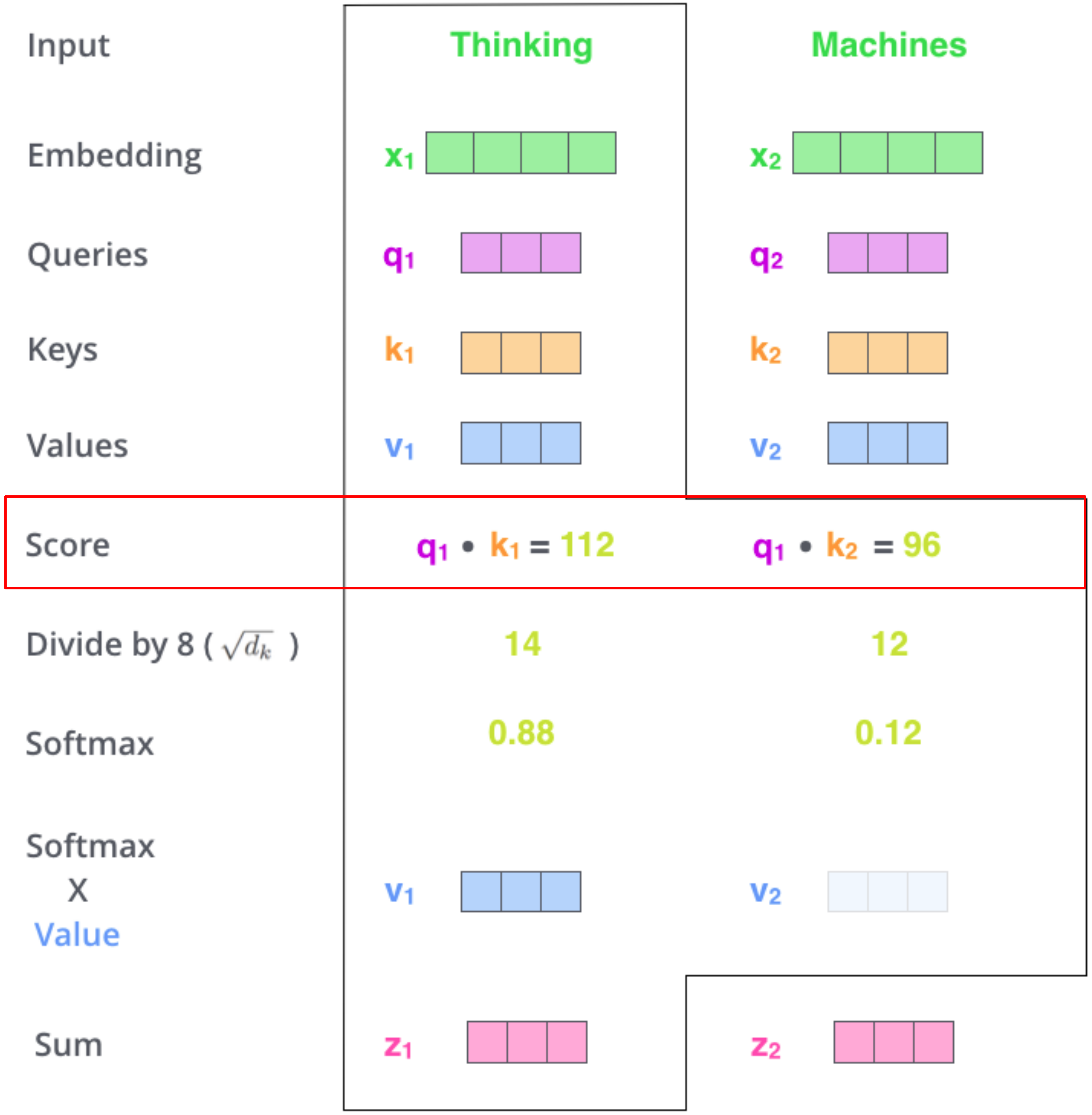


http://jalammar.github.io/illustrated-transformer/

1] Relate

Compatibility bet. each embedding in the dictionary to myself

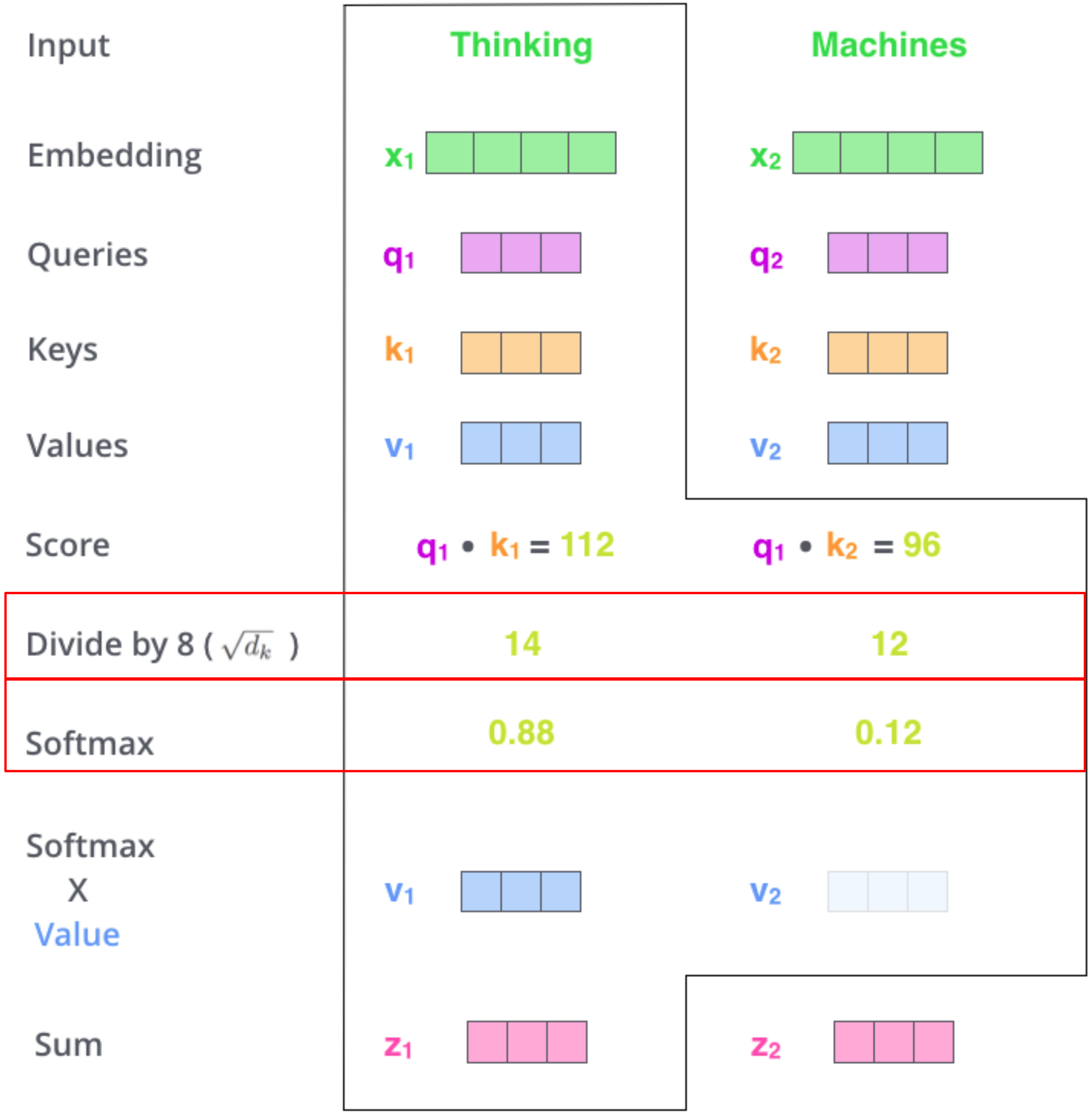
Scalar, Pairwise



http://jalammar.github.io/illustrated-transformer/

1] Relate

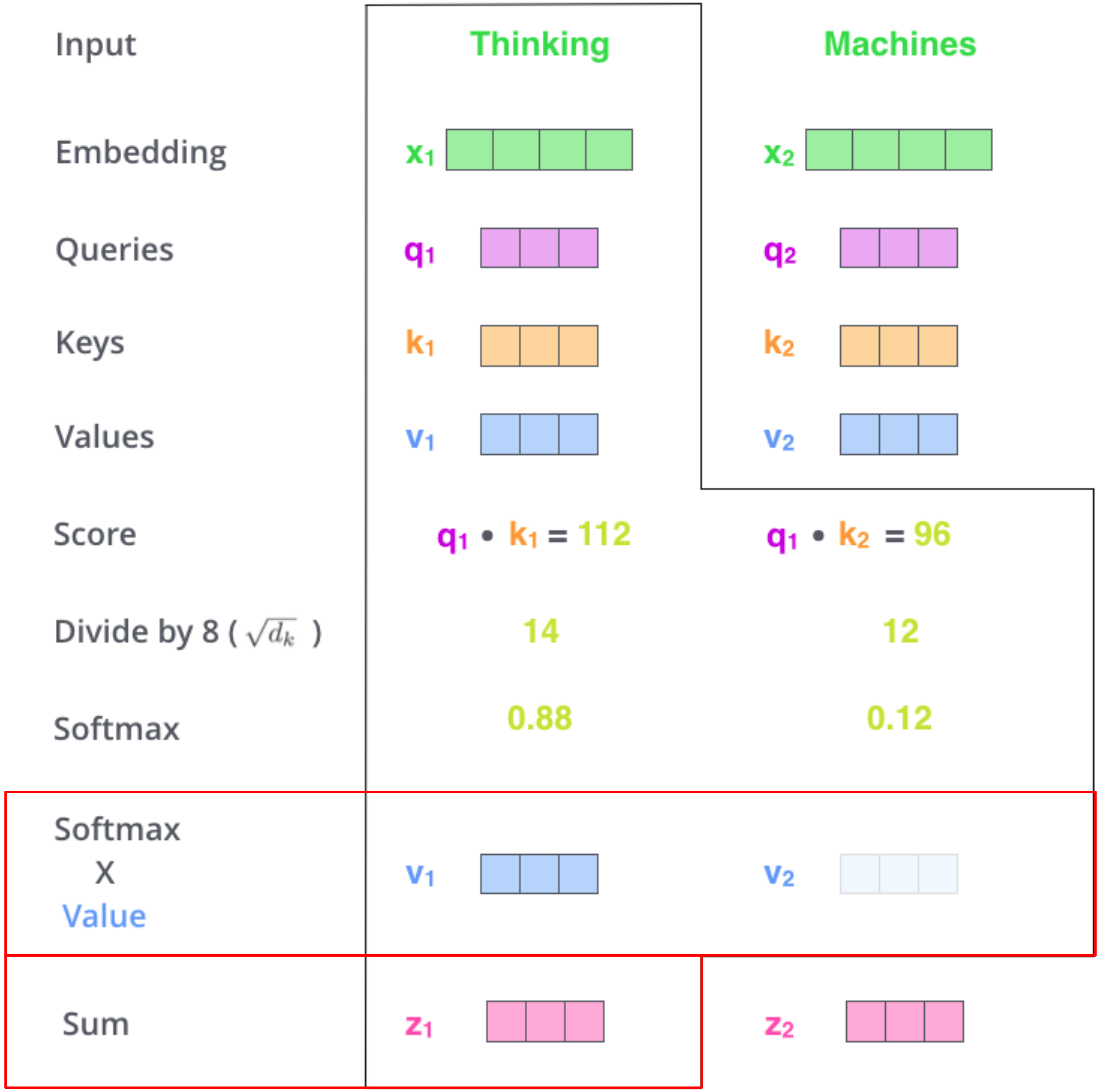
Scaled: because for large values of d_k
→large values of dot product
→ pushes the softmax to have small gradients.



http://jalammar.github.io/illustrated-transformer/

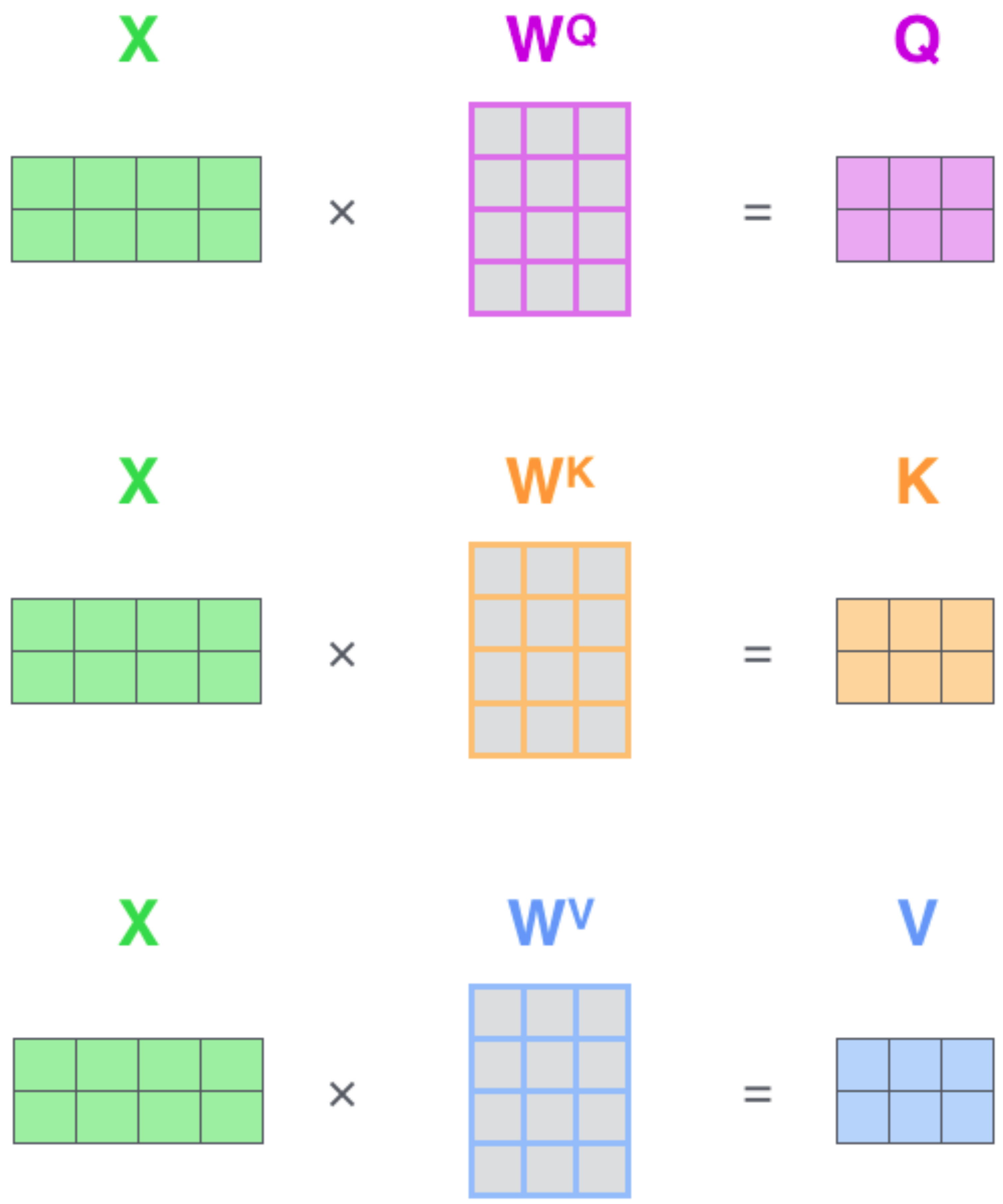
2] Aggregate

Aggregate information from all tokens



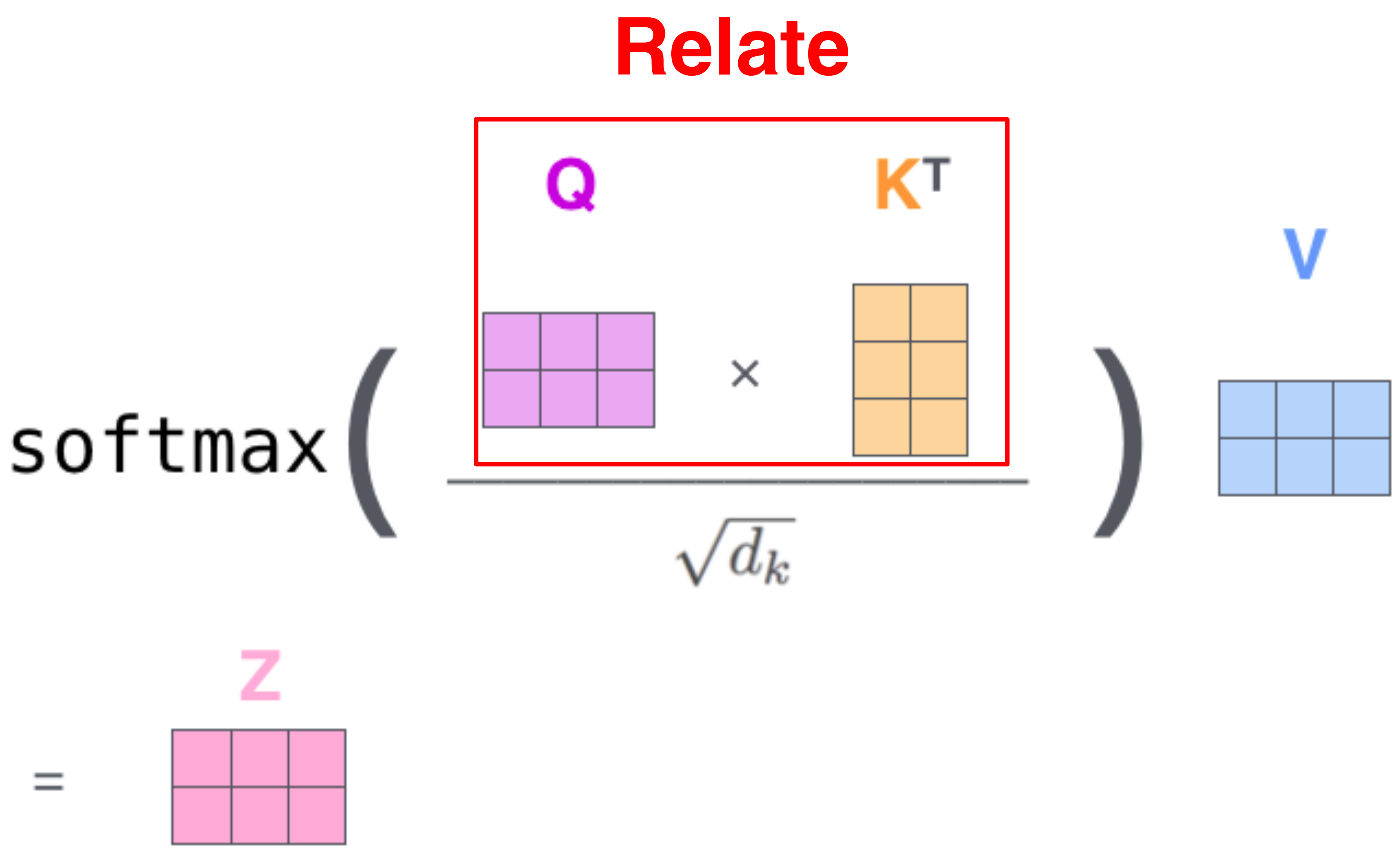
http://jalammar.github.io/illustrated-transformer/

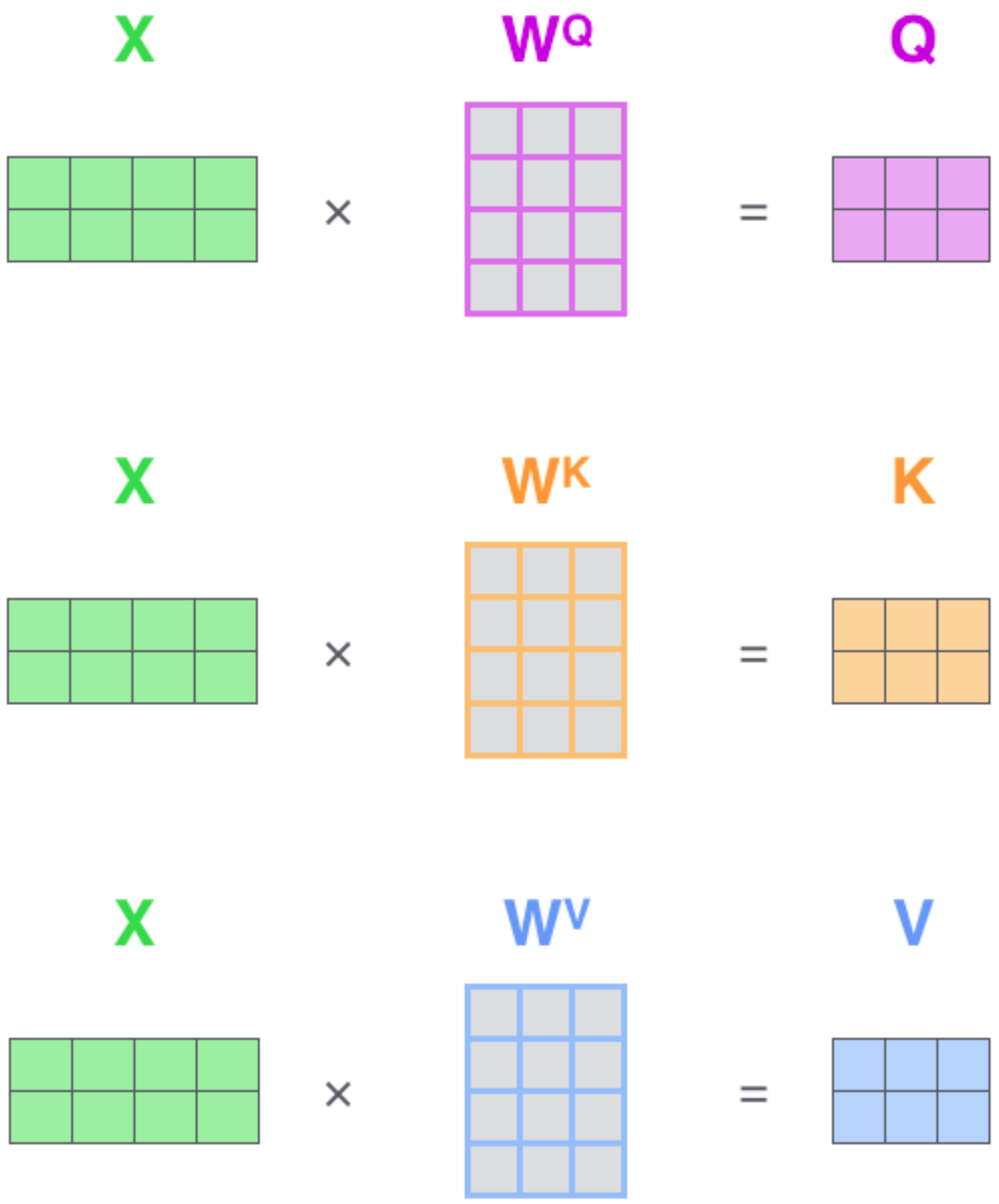
Single-Headed Attention



http://jalammar.github.io/illustrated-transformer/

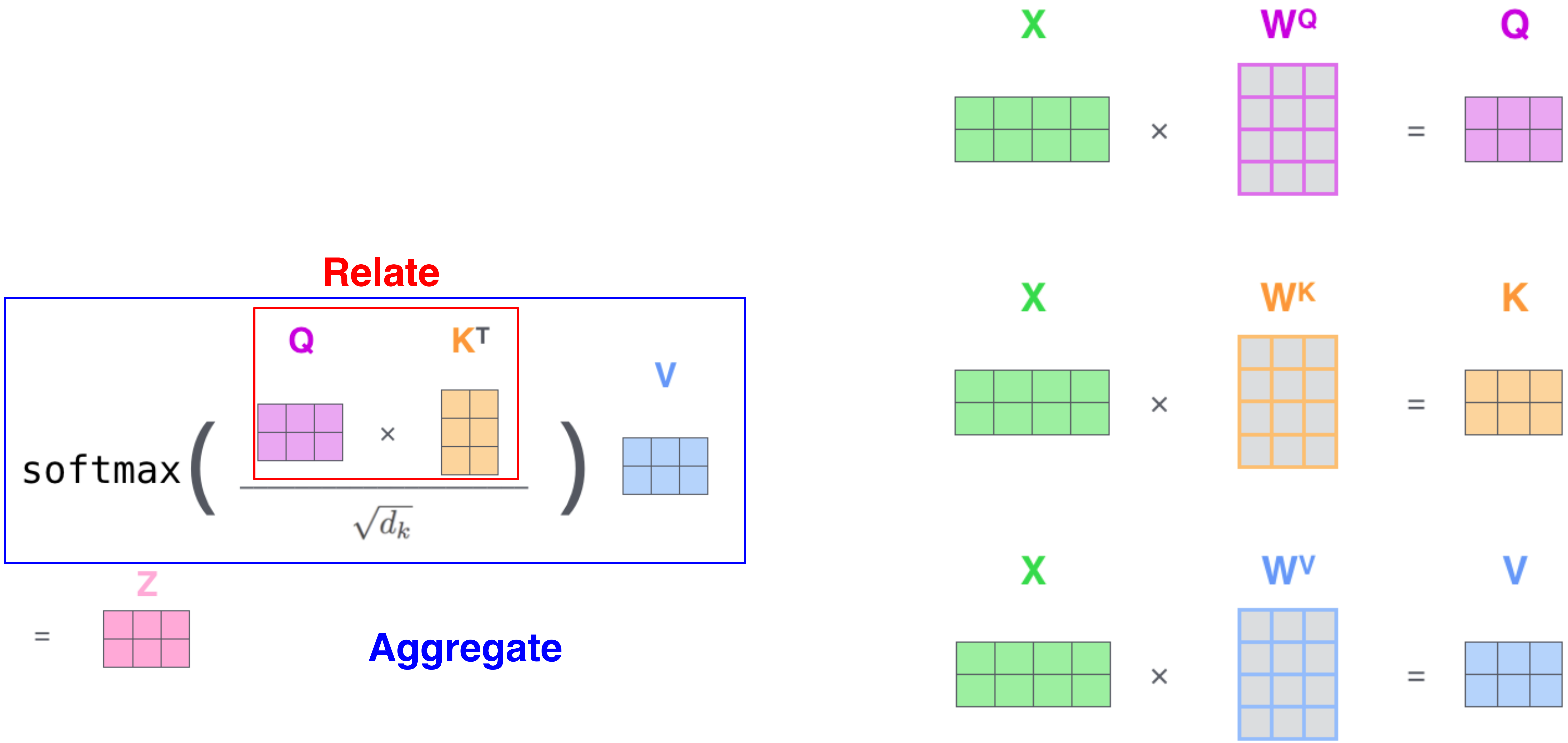
Single-Headed Attention





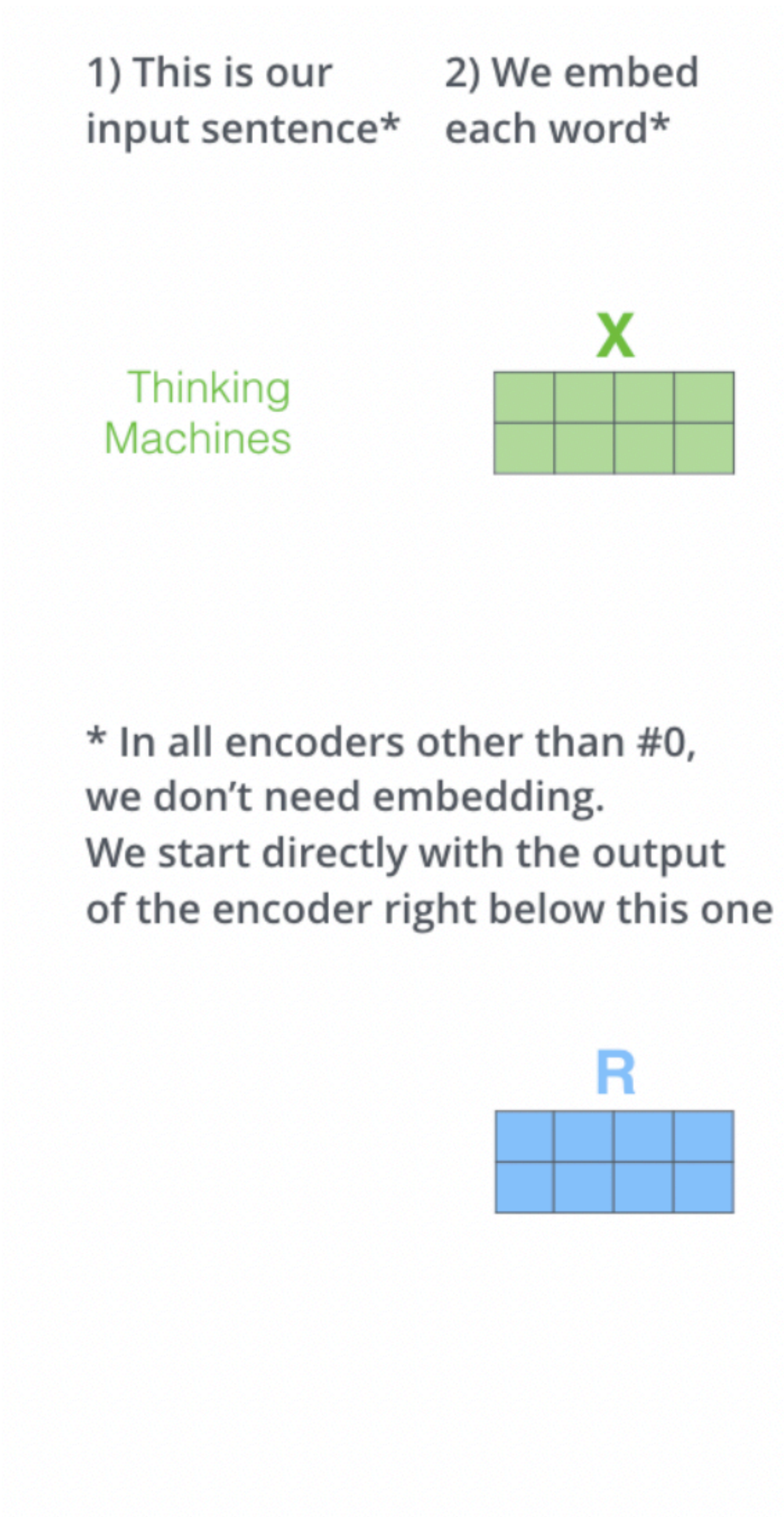
http://jalammar.github.io/illustrated-transformer/

Single-Headed Attention

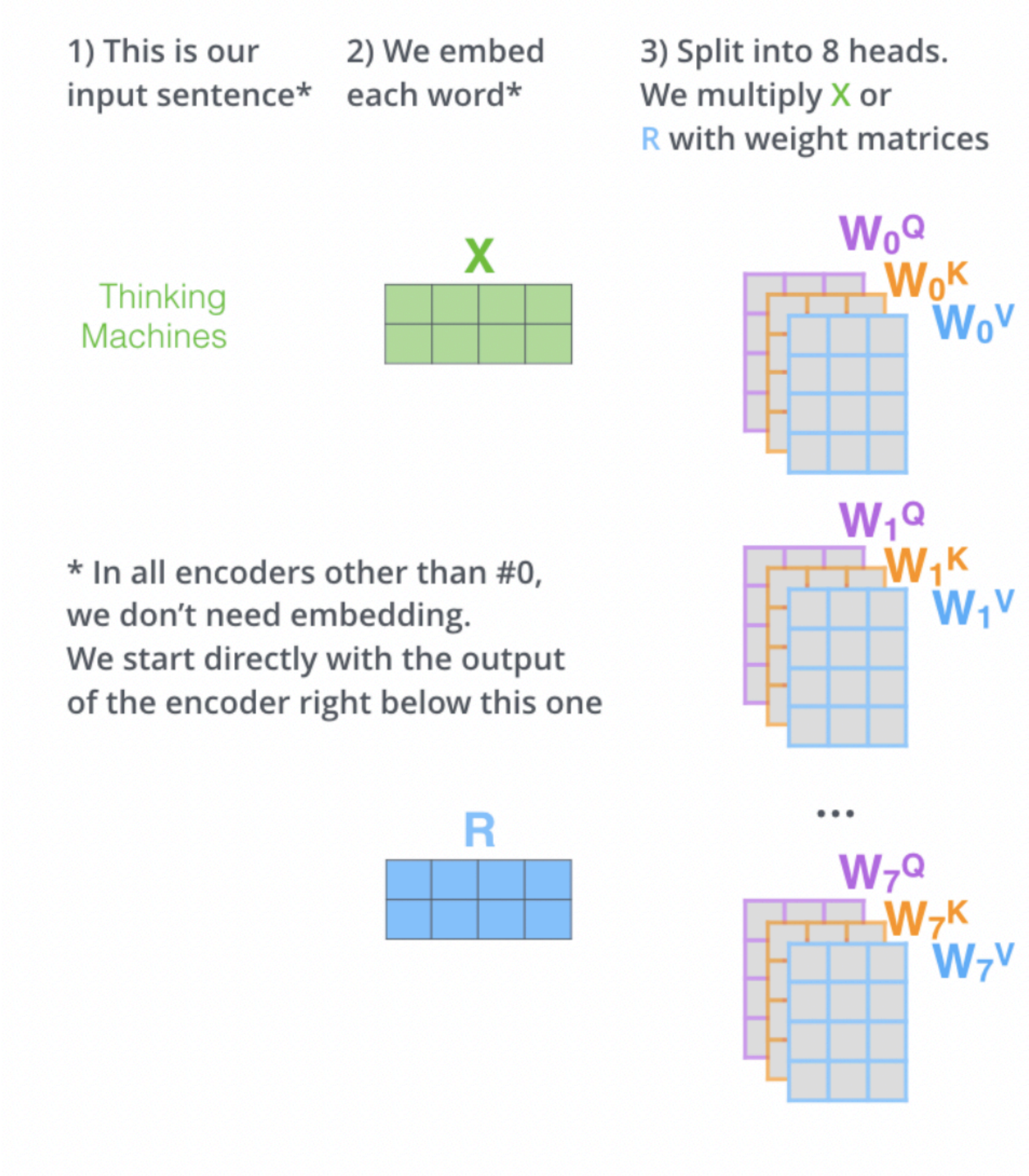


http://jalammar.github.io/illustrated-transformer/

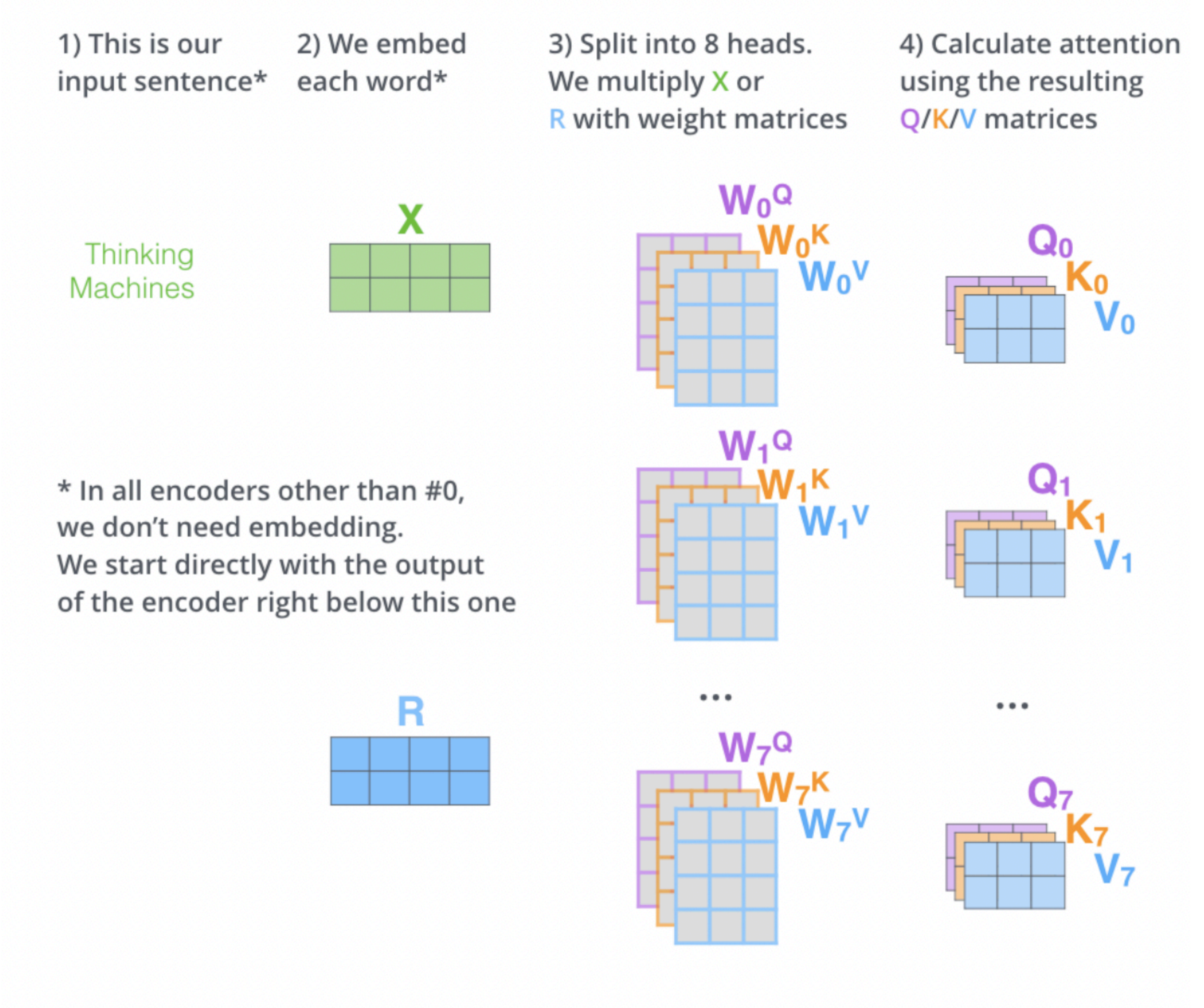
Multihead Attention



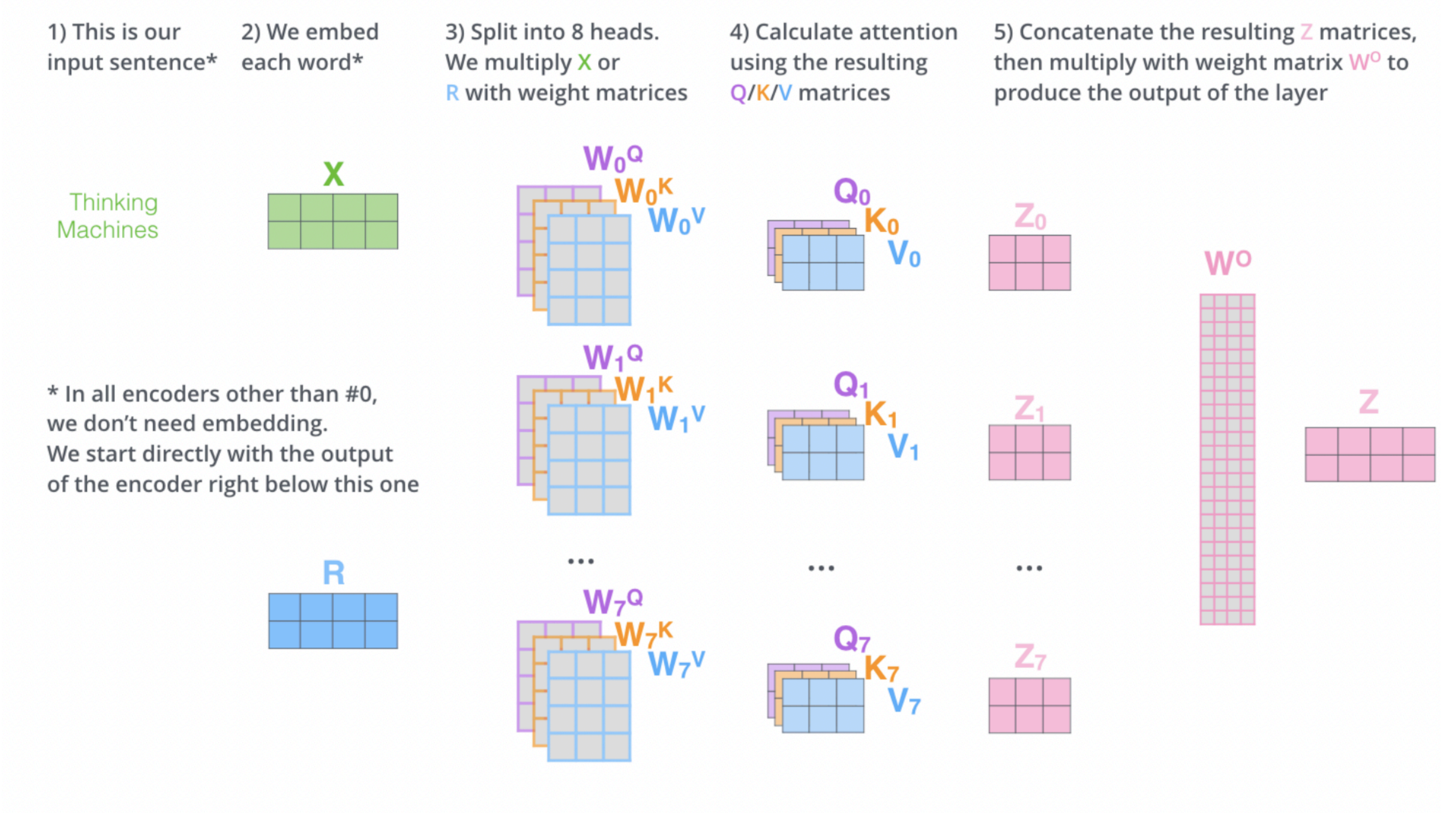
Multihead Attention



Multihead Attention



Multihead Attention

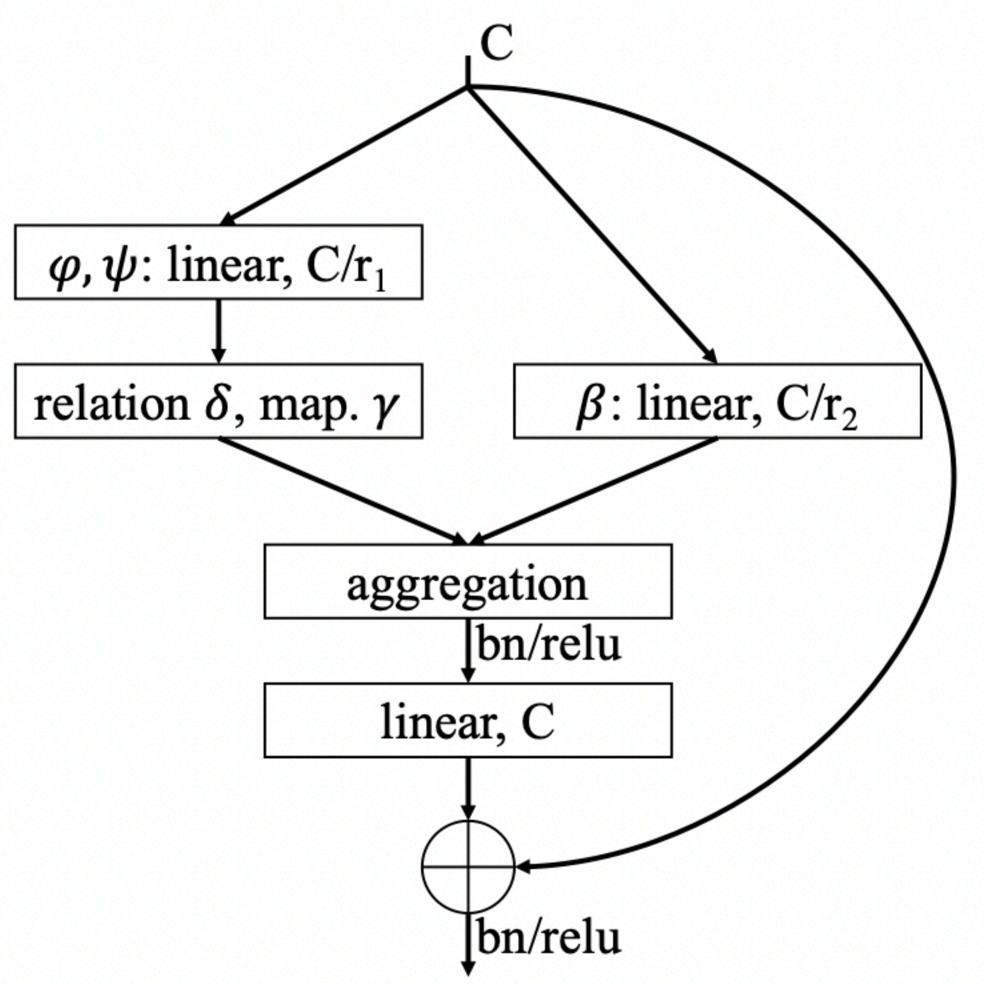


Transformers in Computer Vision

April 2020 — **Exploring Self Attention for Image Recognition**

May 2020 — **DetR**

October 2020 — **Vision Transformers**

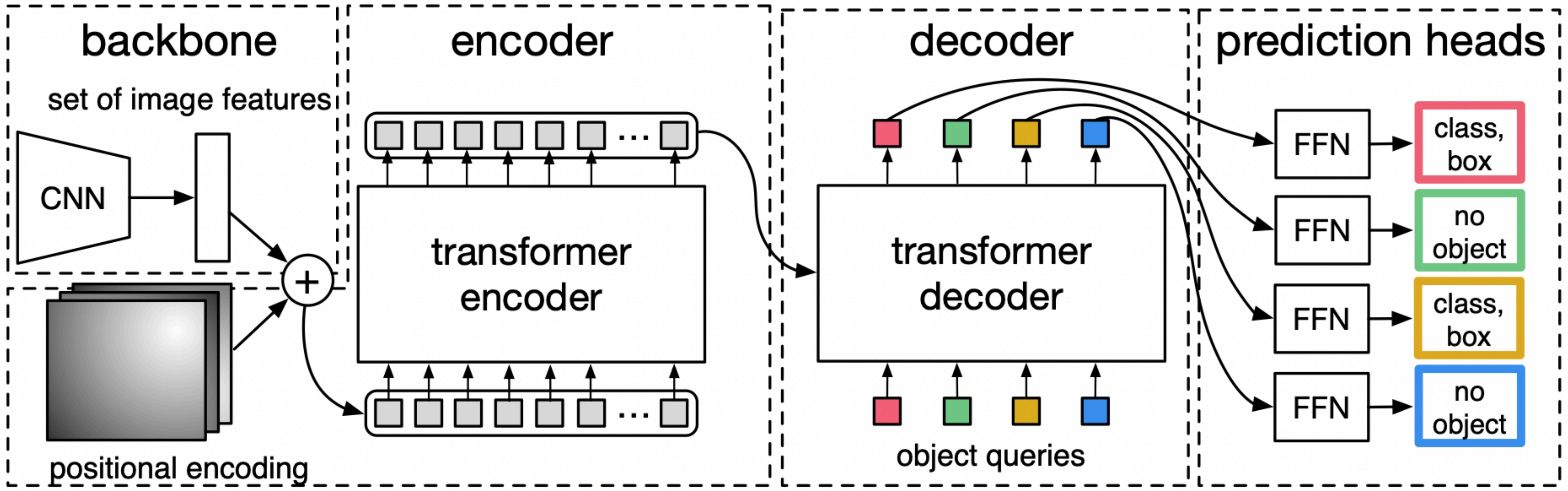


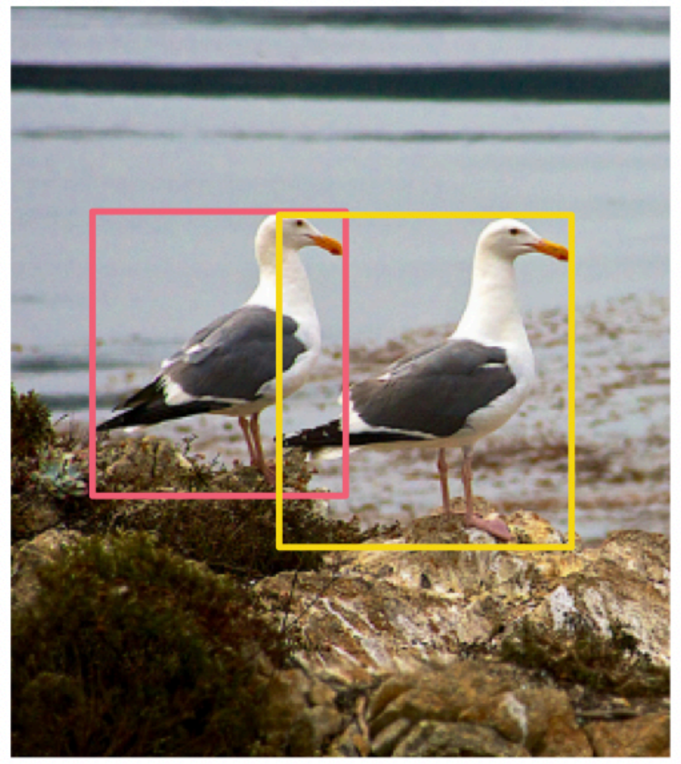
Transformers in Computer Vision

April 2020 — **Exploring Self Attention for Image Recognition**

May 2020 — **DetR**

June 2021 — **Vision Transformers**



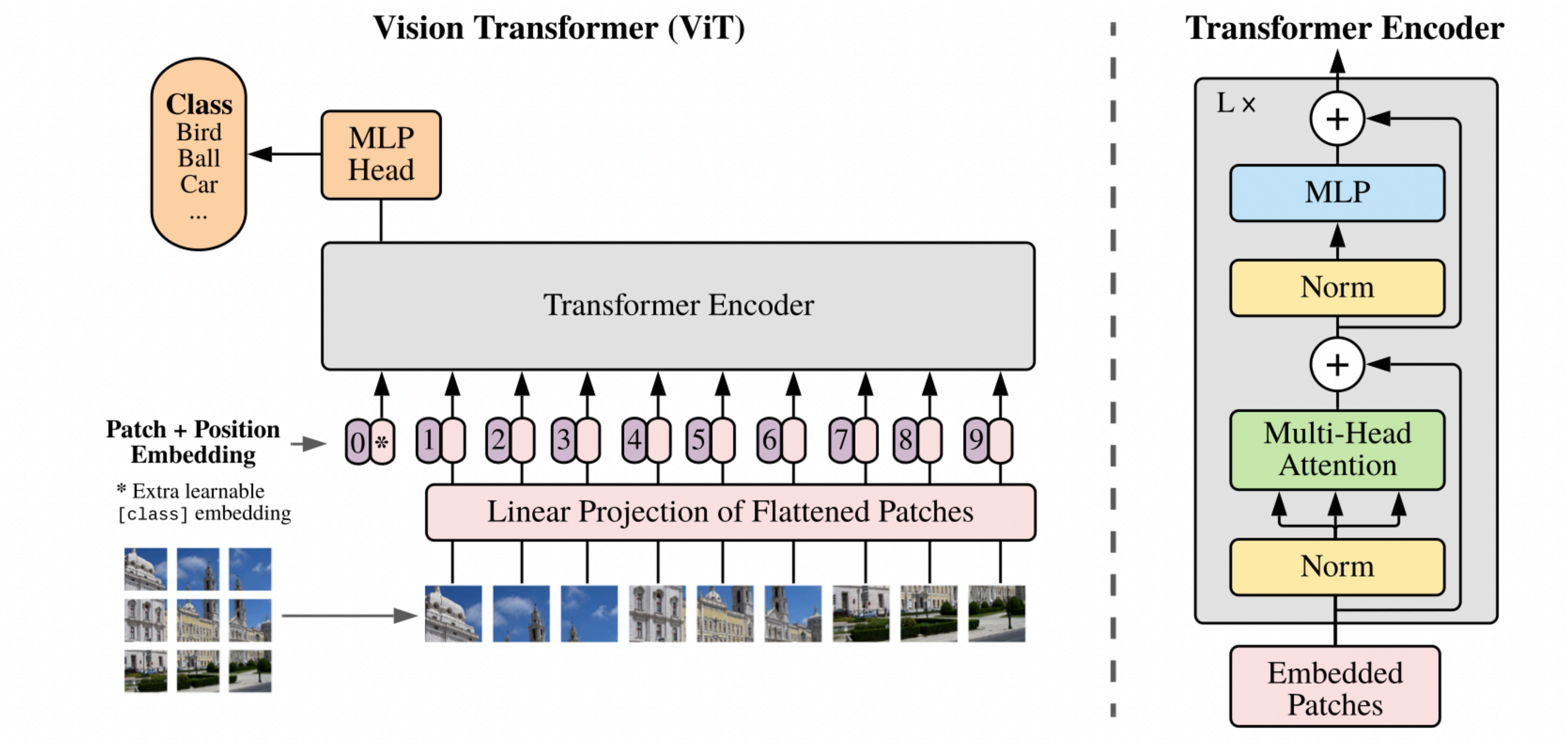


Transformers in Computer Vision

April 2020 — **Exploring Self Attention for Image Recognition**

May 2020 — **DetR**

June 2021 — **Vision Transformers**

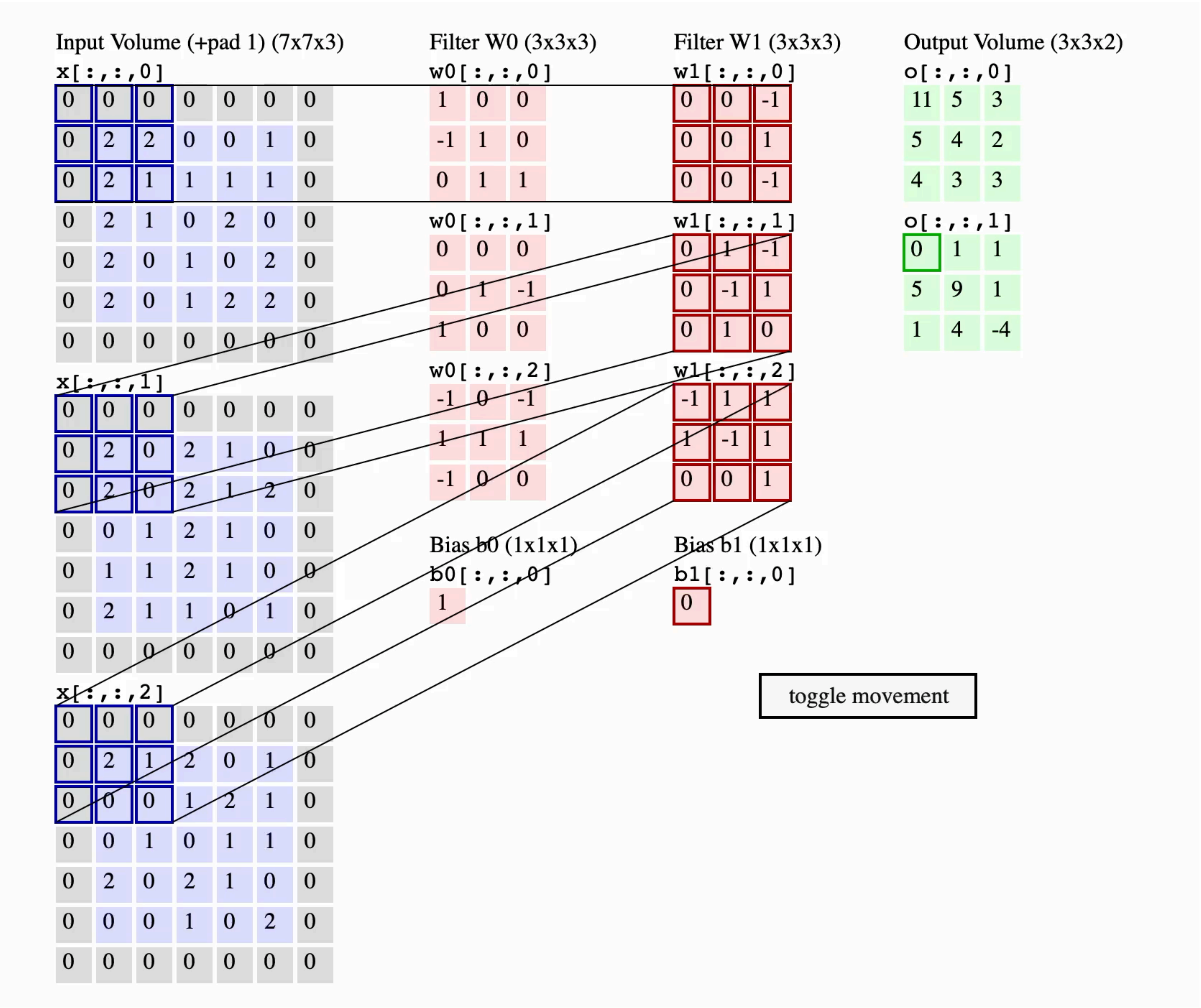


Self Attention: A New Perspective

Convolutional Networks

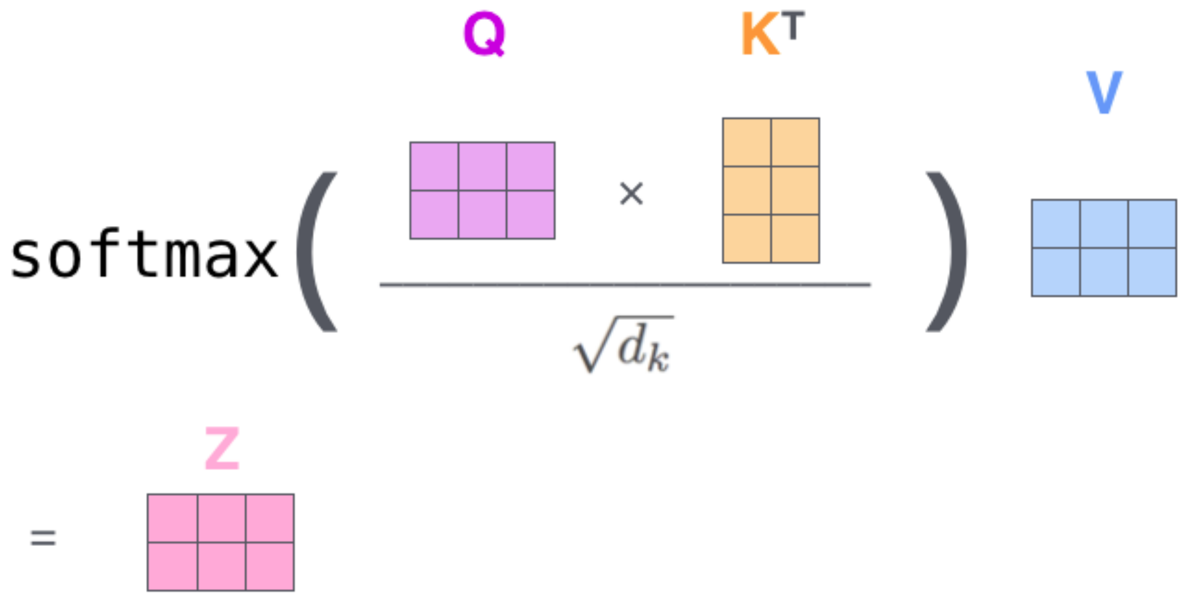
Aggregate Function

Fixed Weights



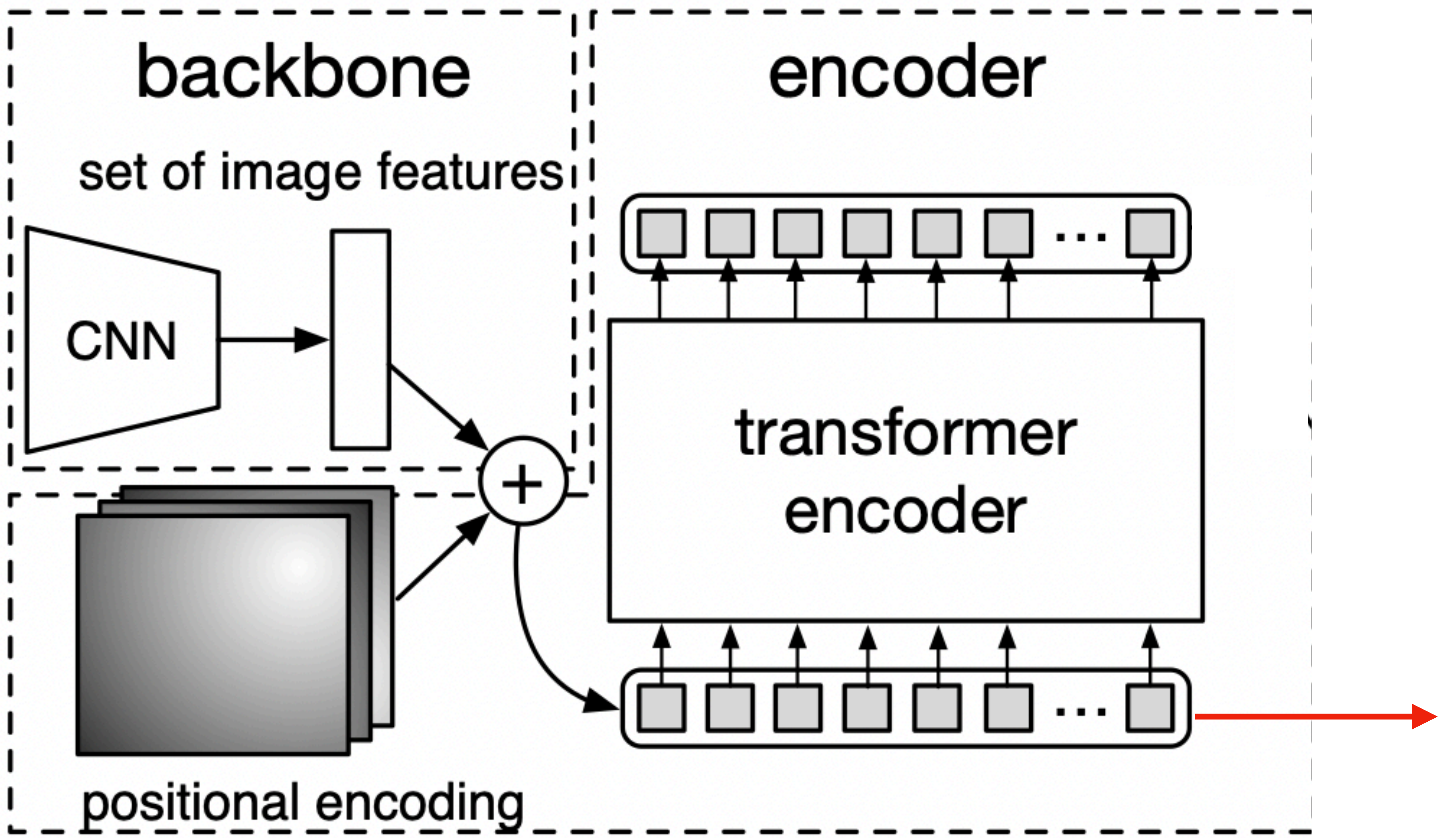
Self Attention: A New Perspective

Transformers



Aggregate Function

Content Adaptive Weights



Think of these as Visual Words

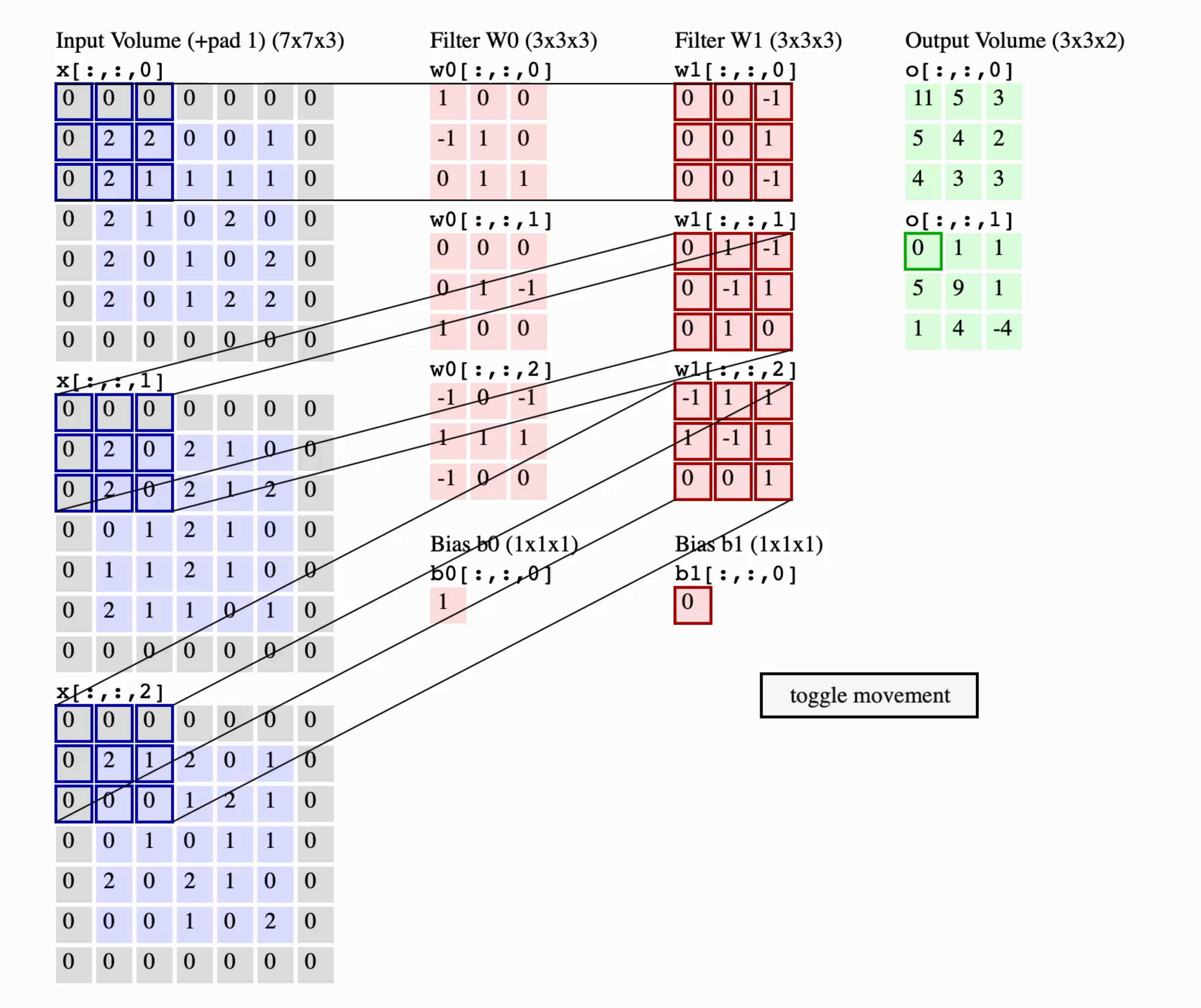
Tokens: Flattened CNN features

Self Attention: A New Perspective

Convolutional Networks

Aggregate Function

Local

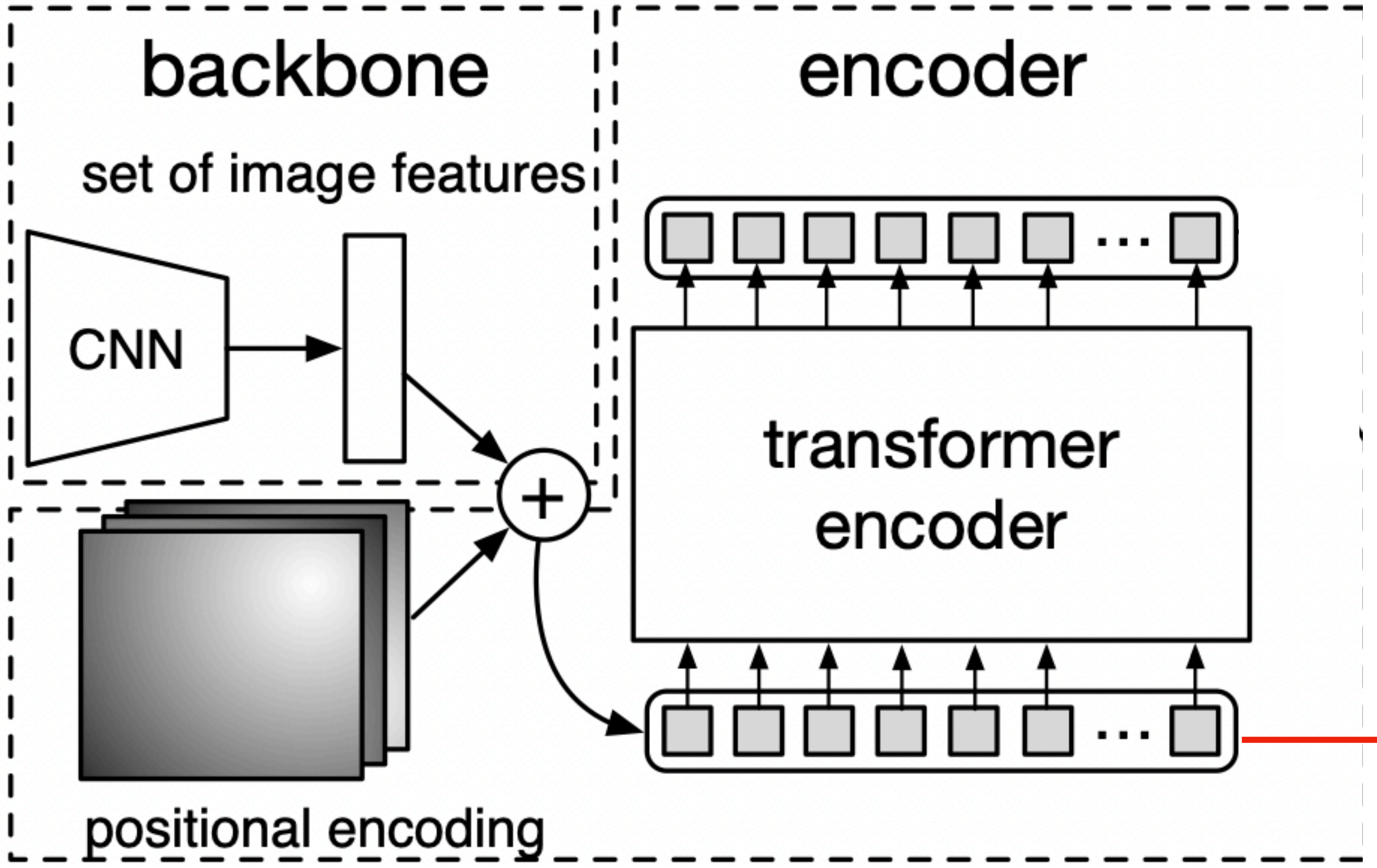


Self Attention: A New Perspective

Transformers

Aggregate Function

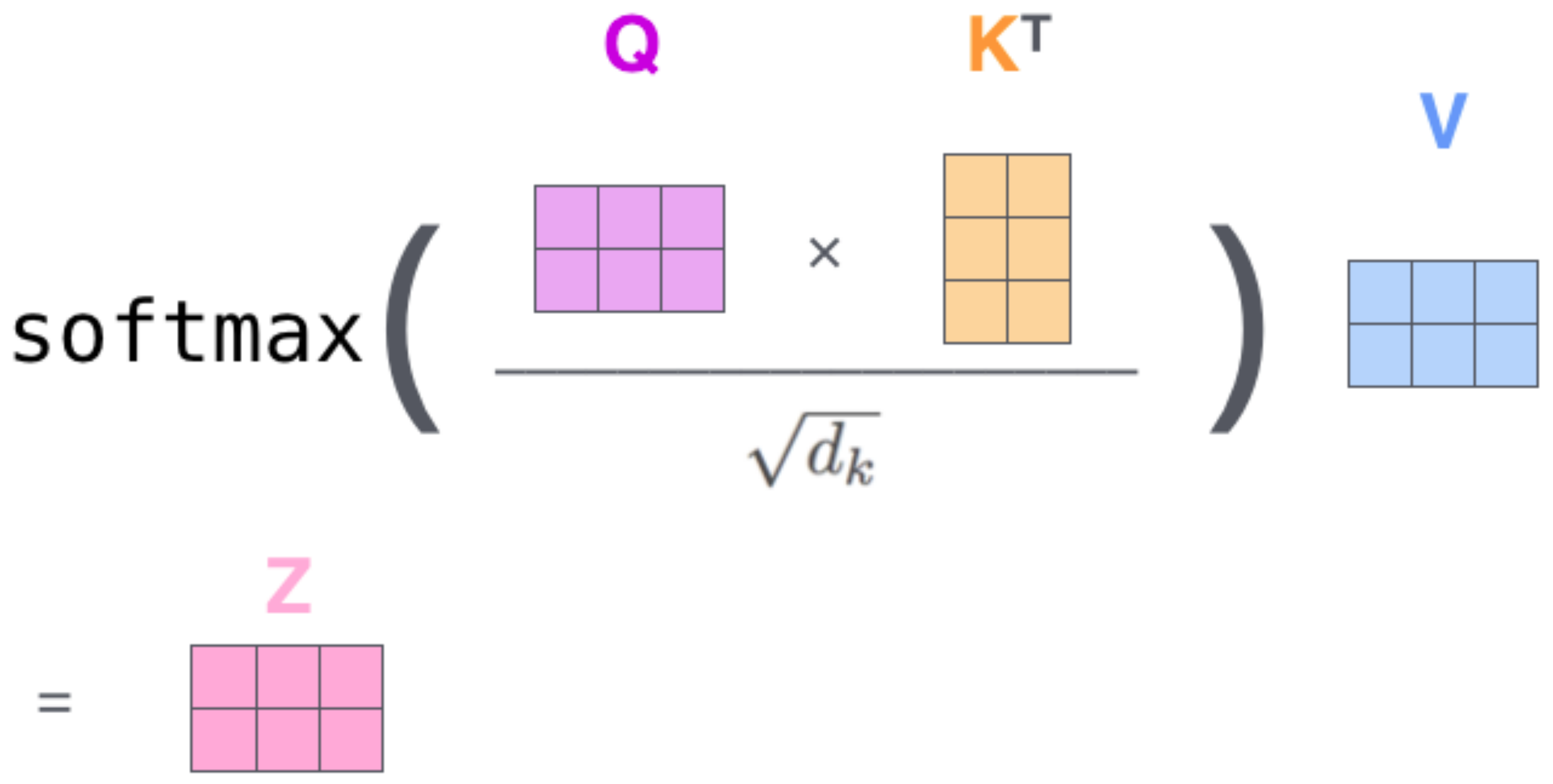
Global Context



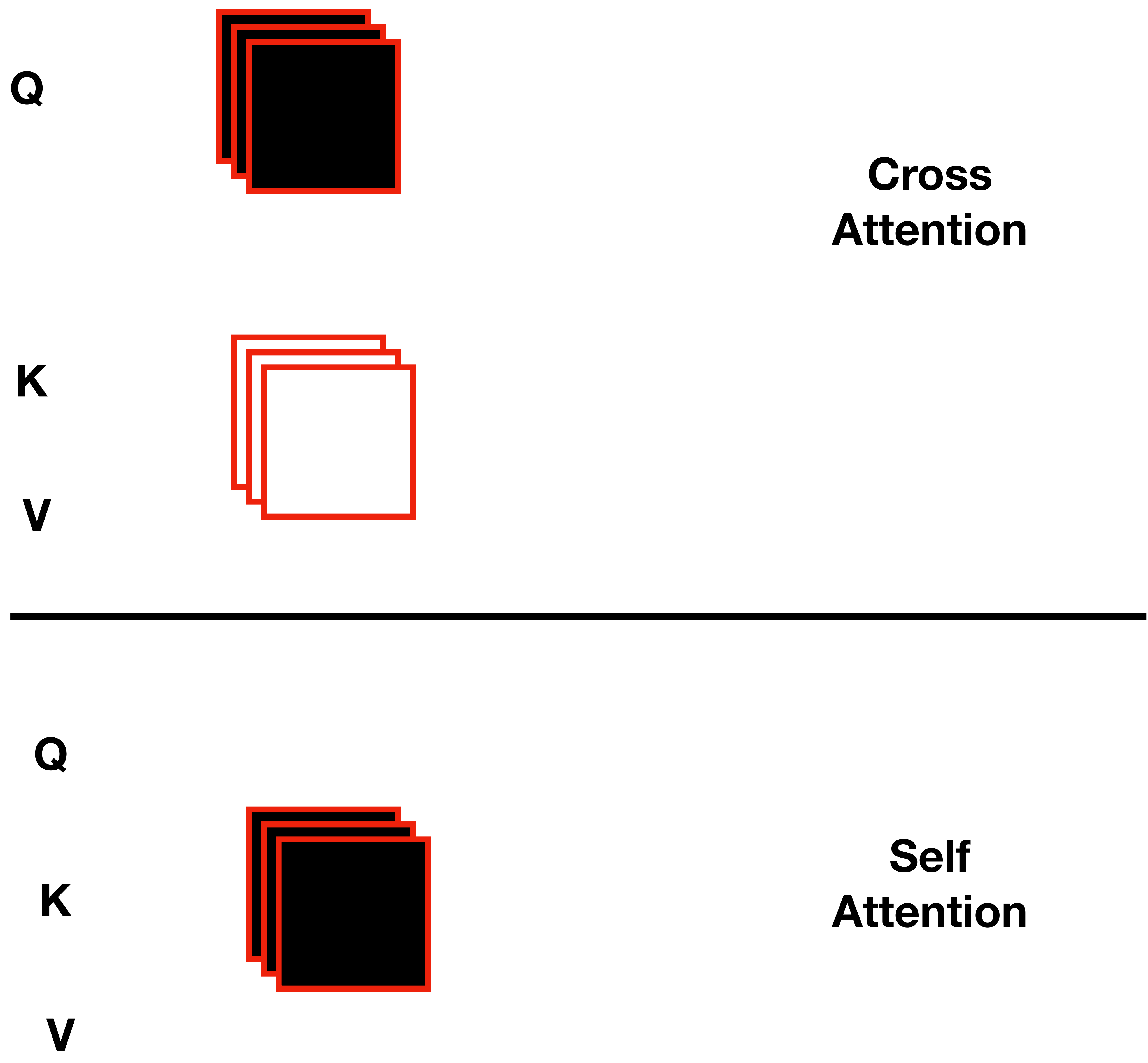
Self Attention across all the features not locally

Tokens: Flattened CNN features

Self Attention vs Cross Attention



MultiHeaded Attention in both

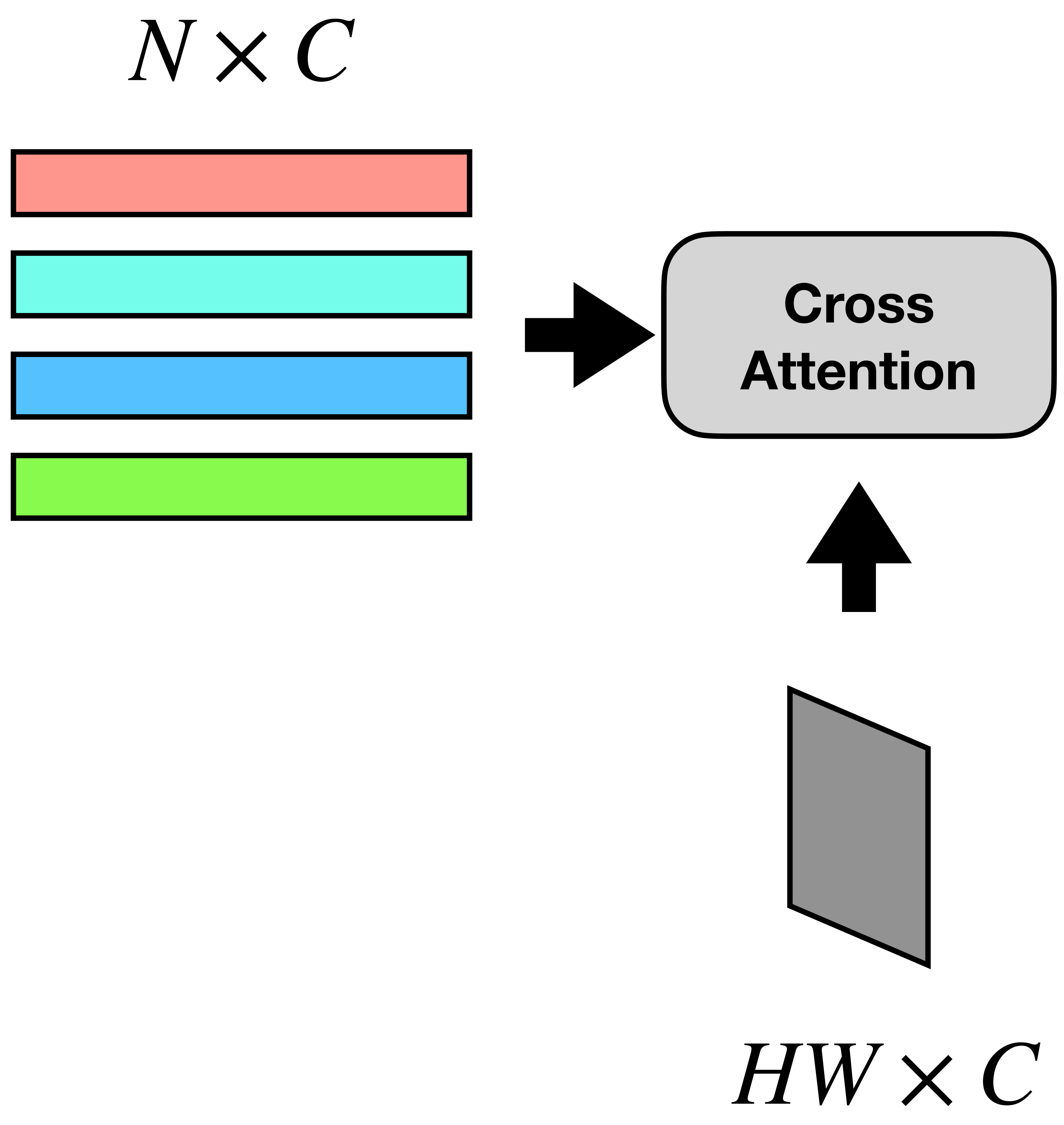


Cross Attention

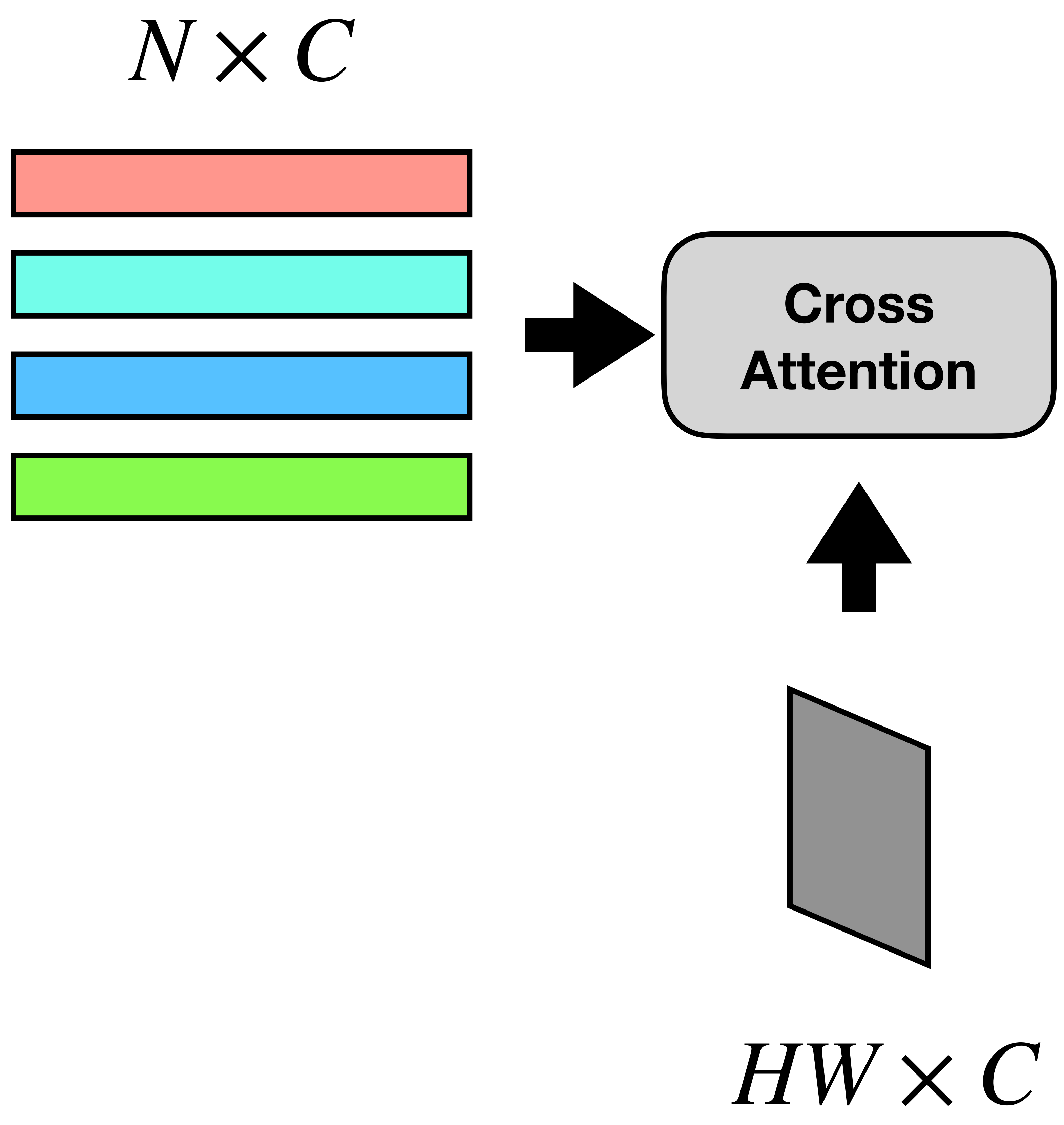
$N \times C$

Learnable Queries

Cross Attention

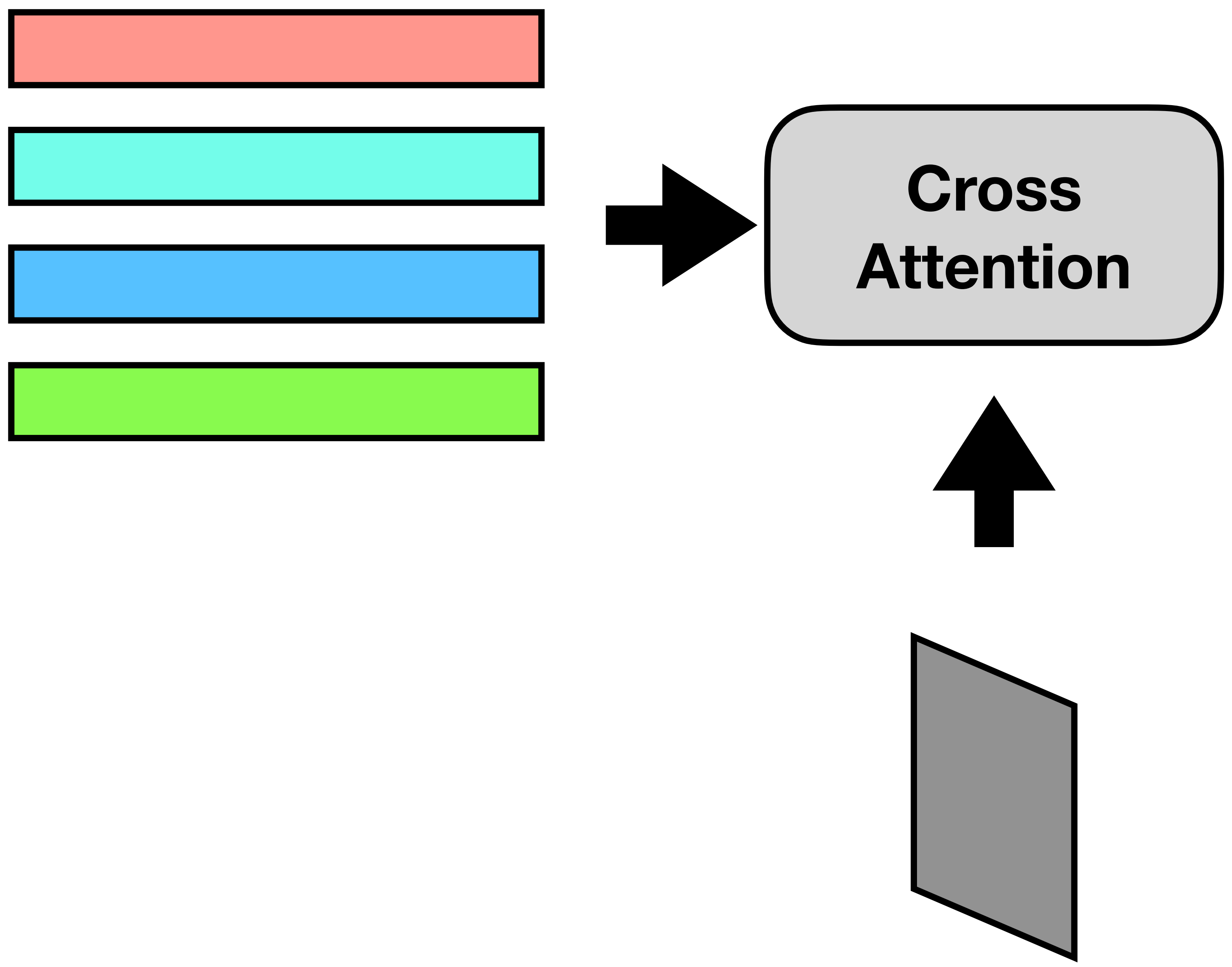


Cross Attention



$$\mathbf{X}_l = \mathrm{softmax}(\mathbf{Q}_l \mathbf{K}_l^{\mathrm{T}})\mathbf{V}_l + \mathbf{X}_{l-1}.$$

Cross Attention



$$\mathbf{X}_l = \mathrm{softmax}(\mathbf{Q}_l\mathbf{K}_l^{\mathrm{T}})\mathbf{V}_l + \mathbf{X}_{l-1}.$$

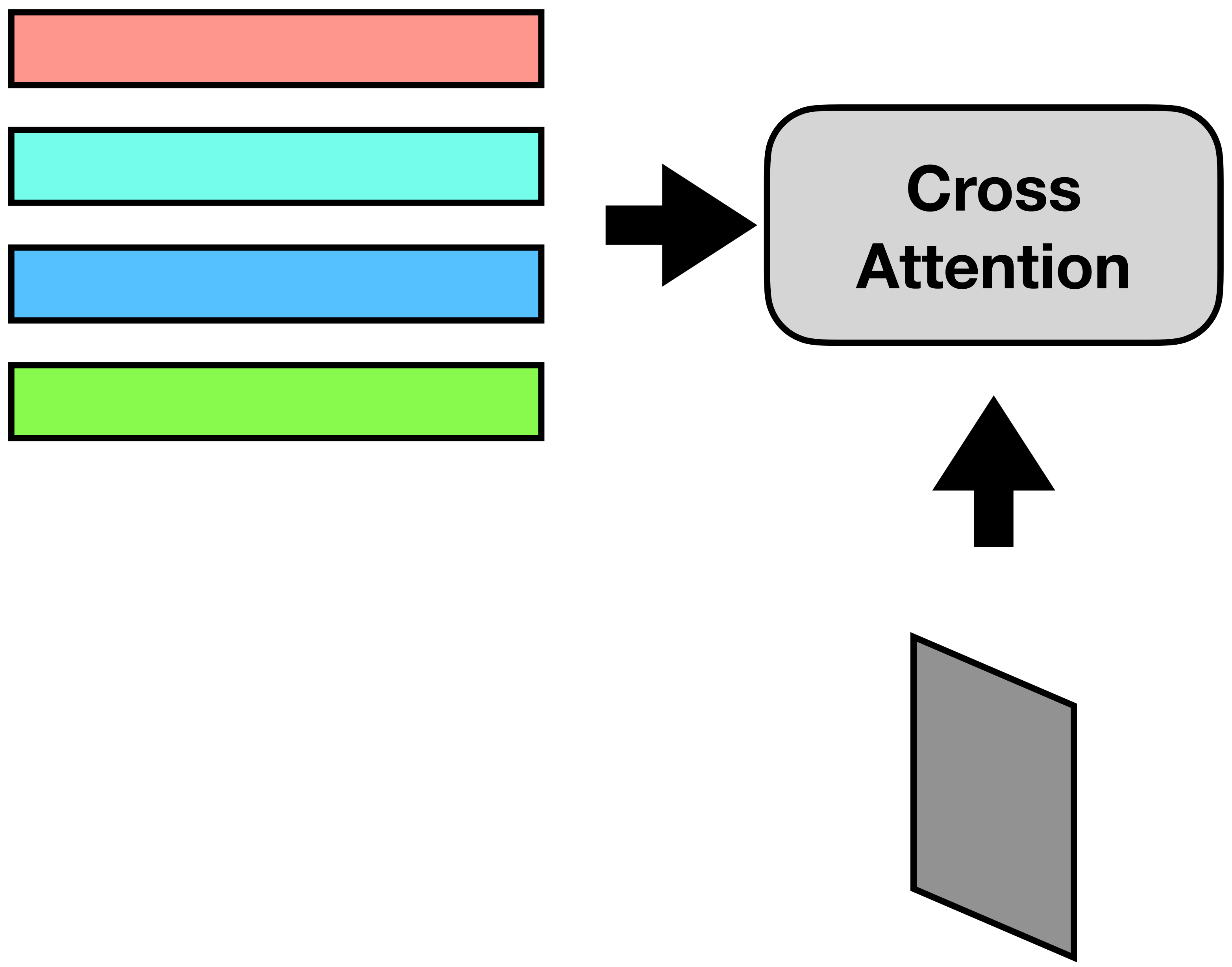
$N \times C \qquad C \times HW$

$N \times HW$

Attention Maps

Relate

Cross Attention



$$\mathbf{X}_l = \mathrm{softmax}(\mathbf{Q}_l \mathbf{K}_l^{\mathrm{T}}) \mathbf{V}_l + \mathbf{X}_{l-1}.$$

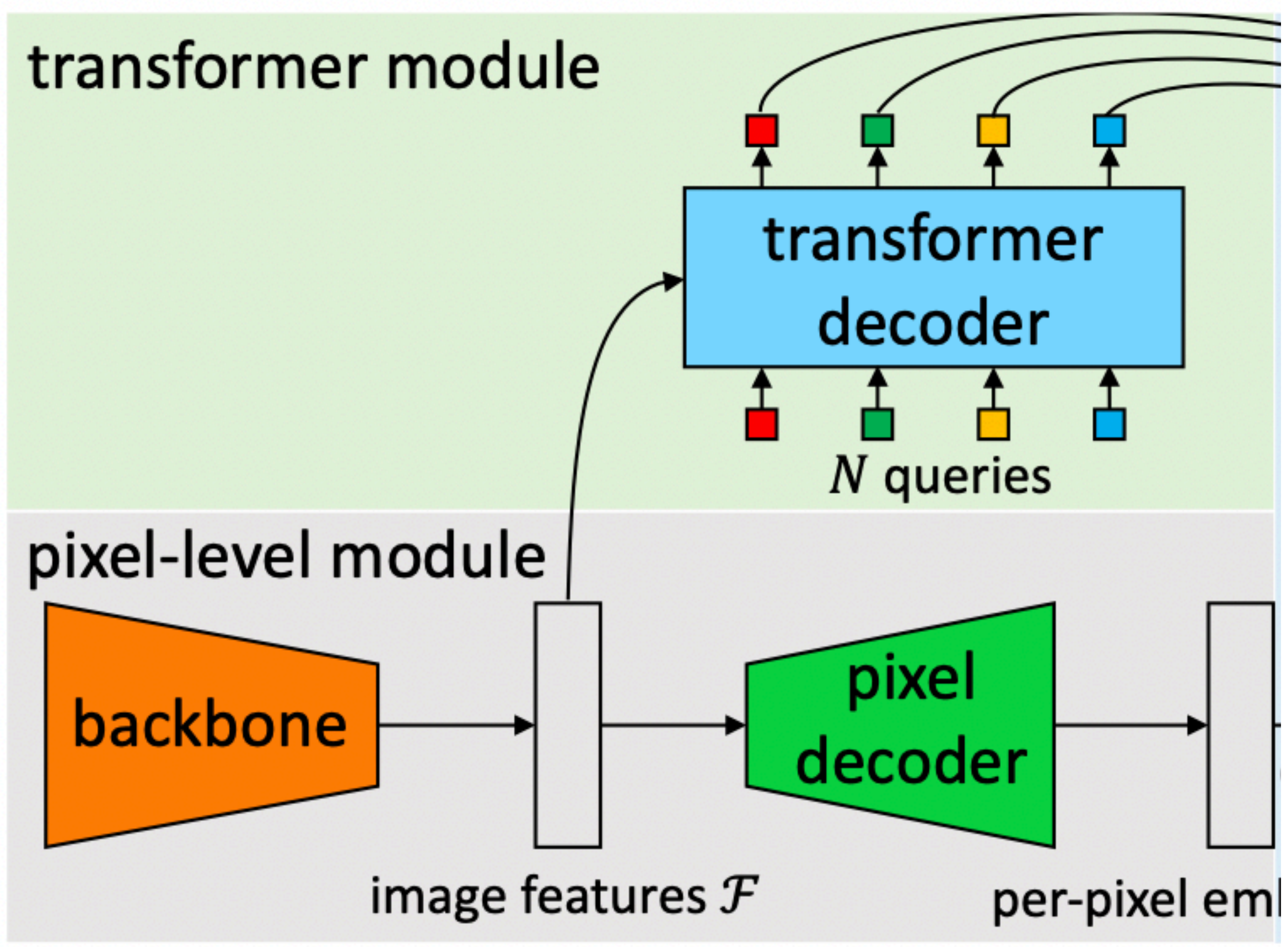
$N \times HW$ $HW \times C$

Attention Maps

$N \times C$

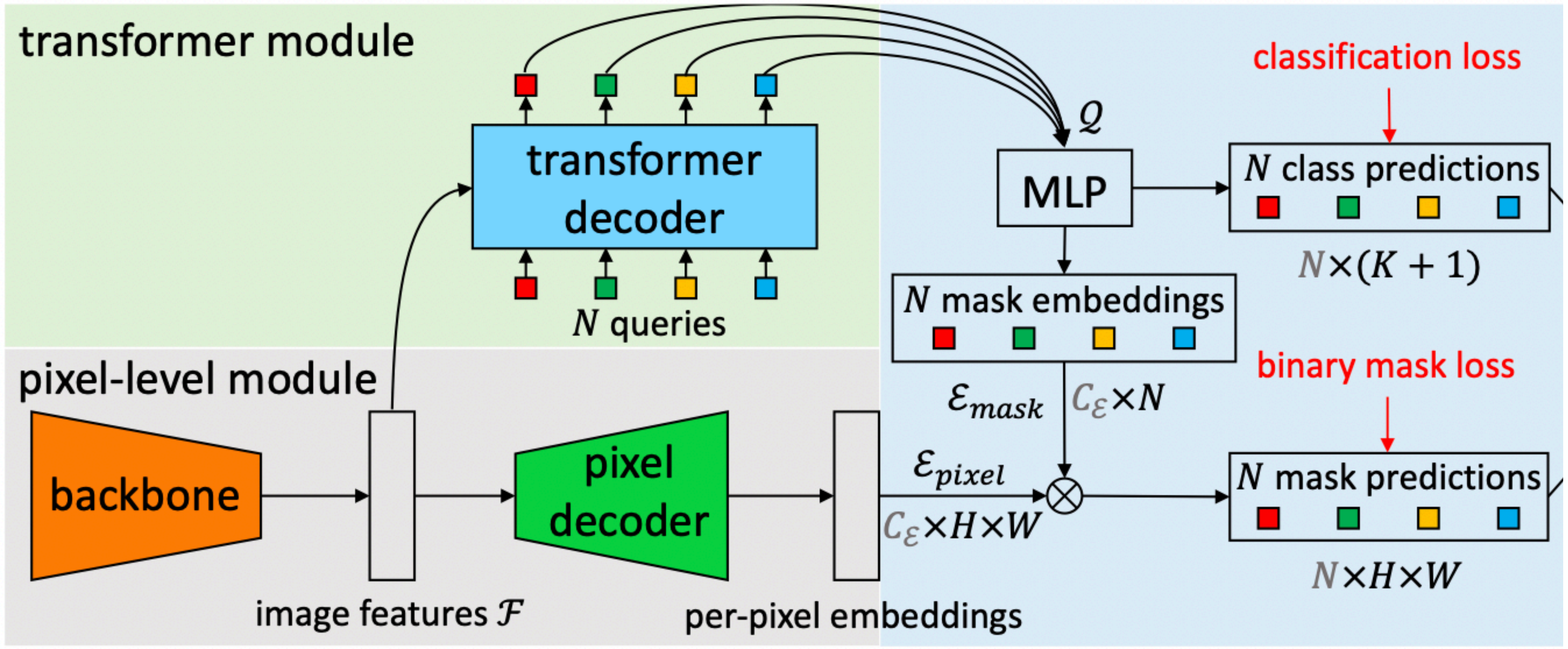
Aggregate

MaskFormer



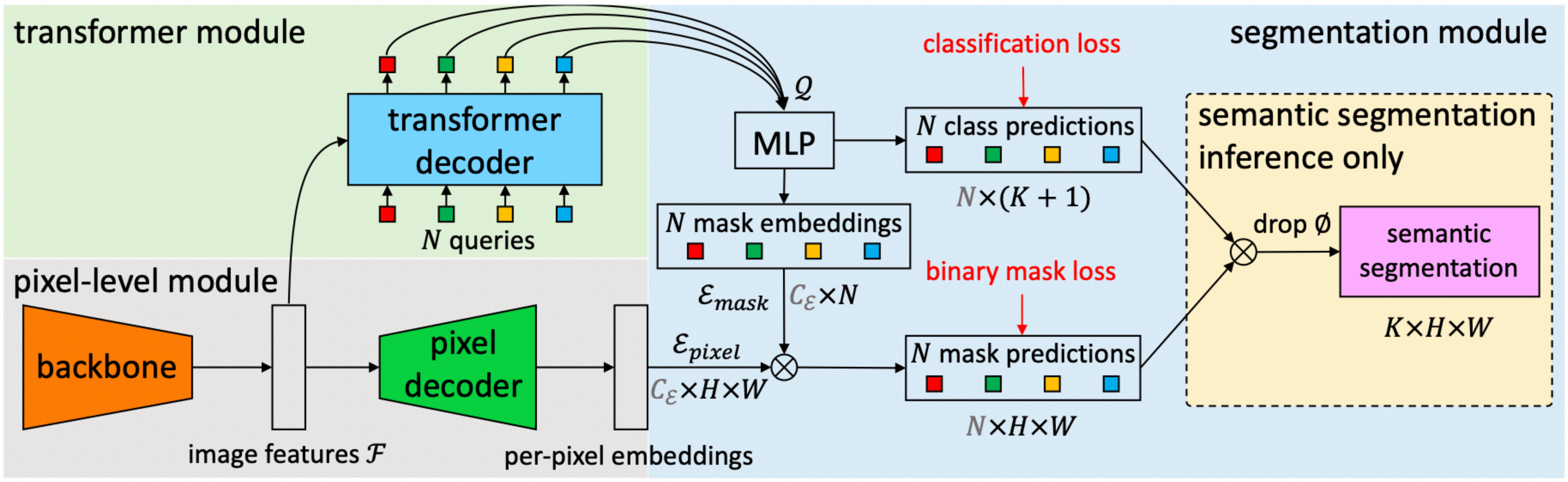
Cheng, Bowen, Alex Schwing, and Alexander Kirillov. "Per-pixel classification is not all you need for semantic segmentation." *Advances in Neural Information Processing Systems* 34 (2021): 17864-17875.

MaskFormer



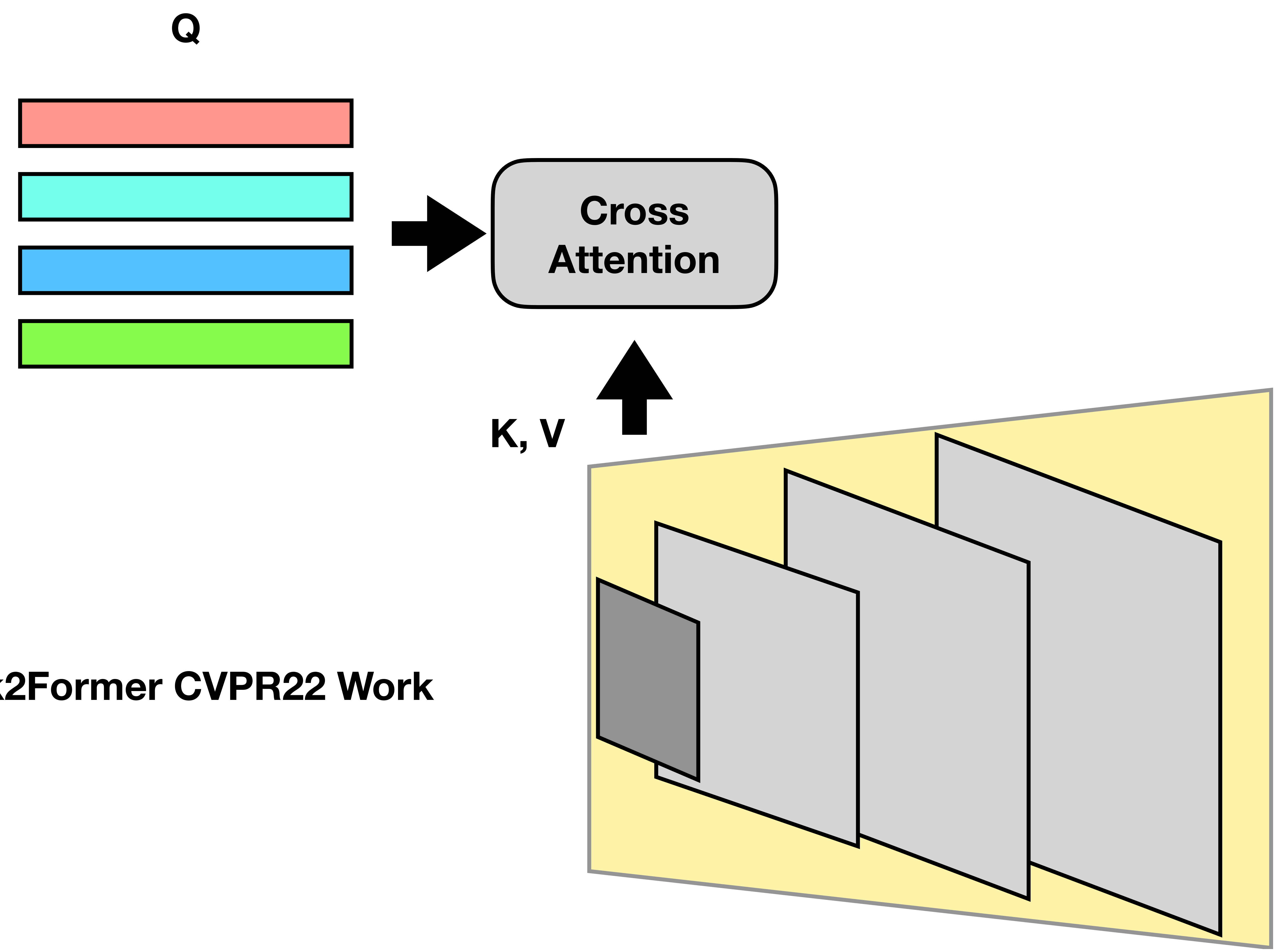
Cheng, Bowen, Alex Schwing, and Alexander Kirillov. "Per-pixel classification is not all you need for semantic segmentation." *Advances in Neural Information Processing Systems* 34 (2021): 17864-17875.

MaskFormer



Cheng, Bowen, Alex Schwing, and Alexander Kirillov. "Per-pixel classification is not all you need for semantic segmentation." *Advances in Neural Information Processing Systems* 34 (2021): 17864-17875.

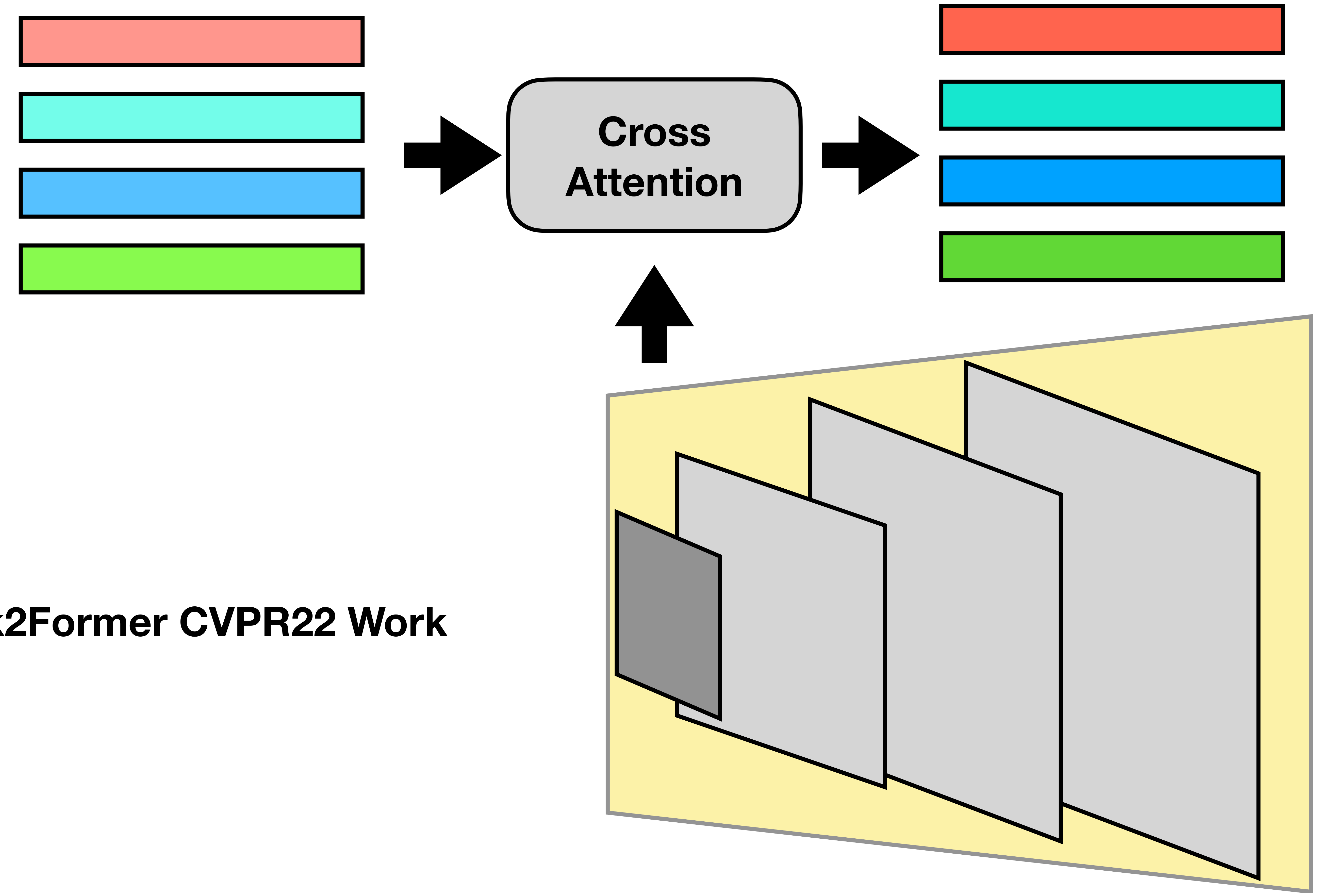
Mask2Former
Multiscale Transformer Decoder



Mask2Former CVPR22 Work

Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

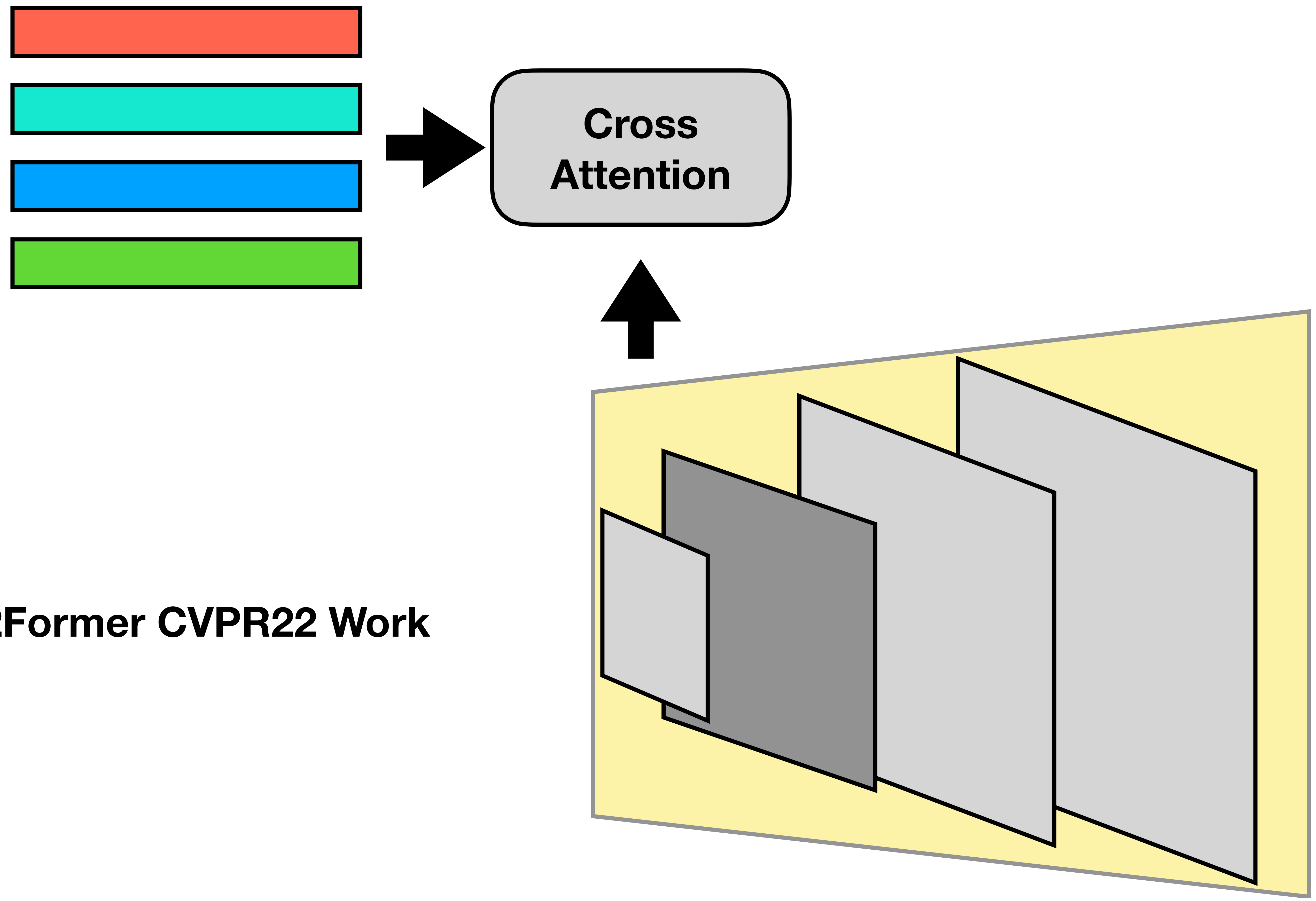
Mask2Former
Multiscale Transformer Decoder



Mask2Former CVPR22 Work

Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

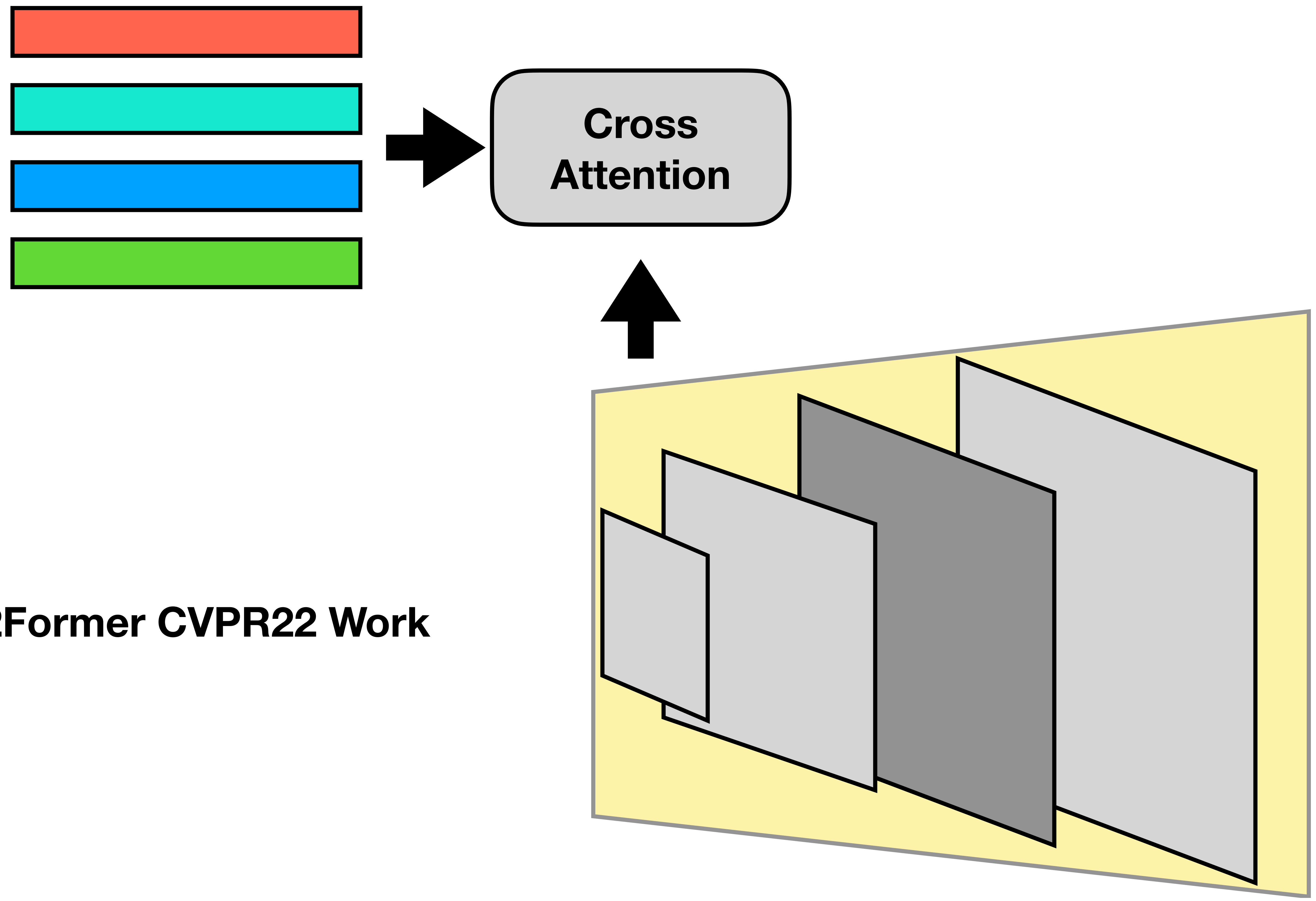
Mask2Former
Multiscale Transformer Decoder



Mask2Former CVPR22 Work

Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

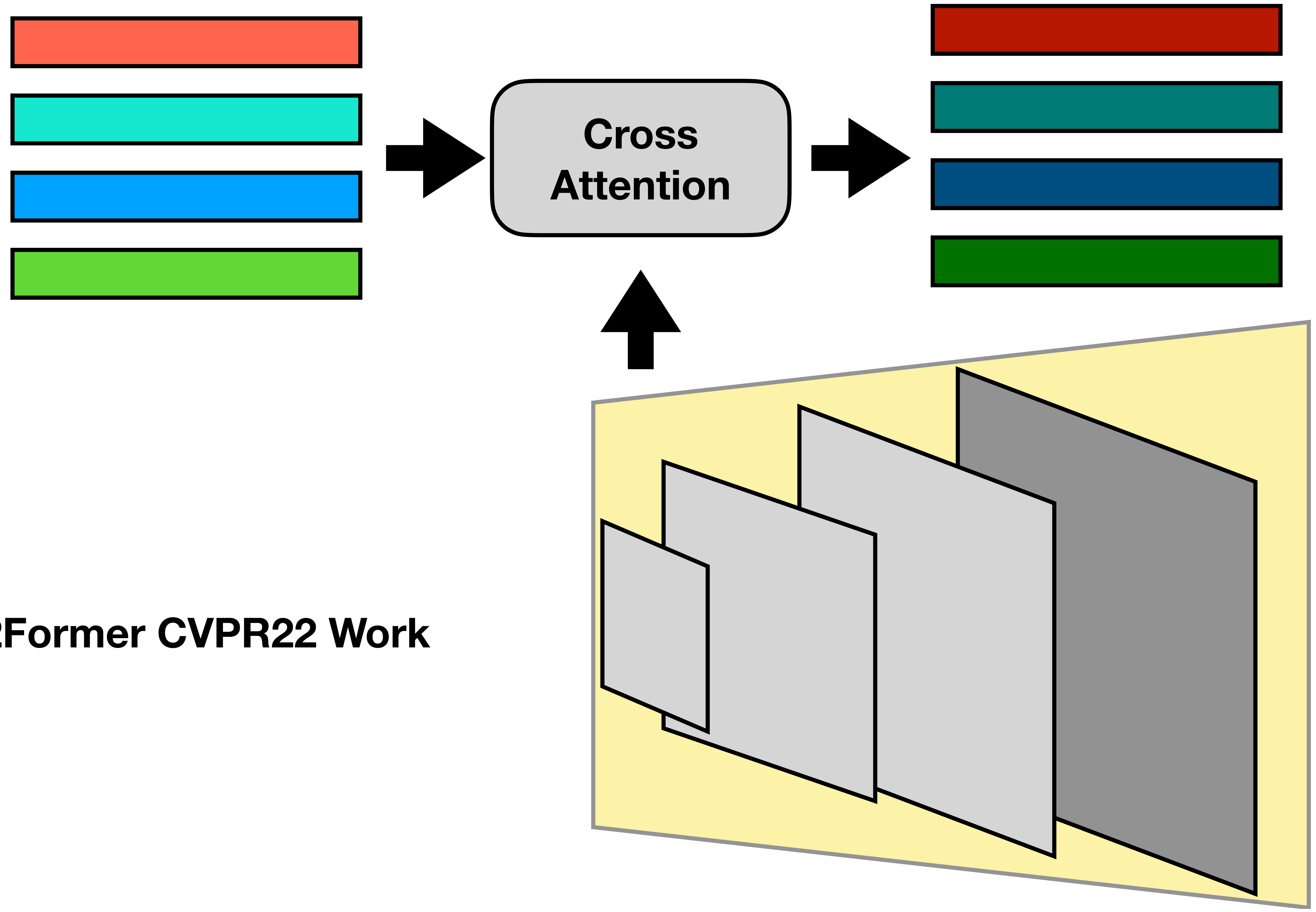
Mask2Former
Multiscale Transformer Decoder



Mask2Former CVPR22 Work

Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

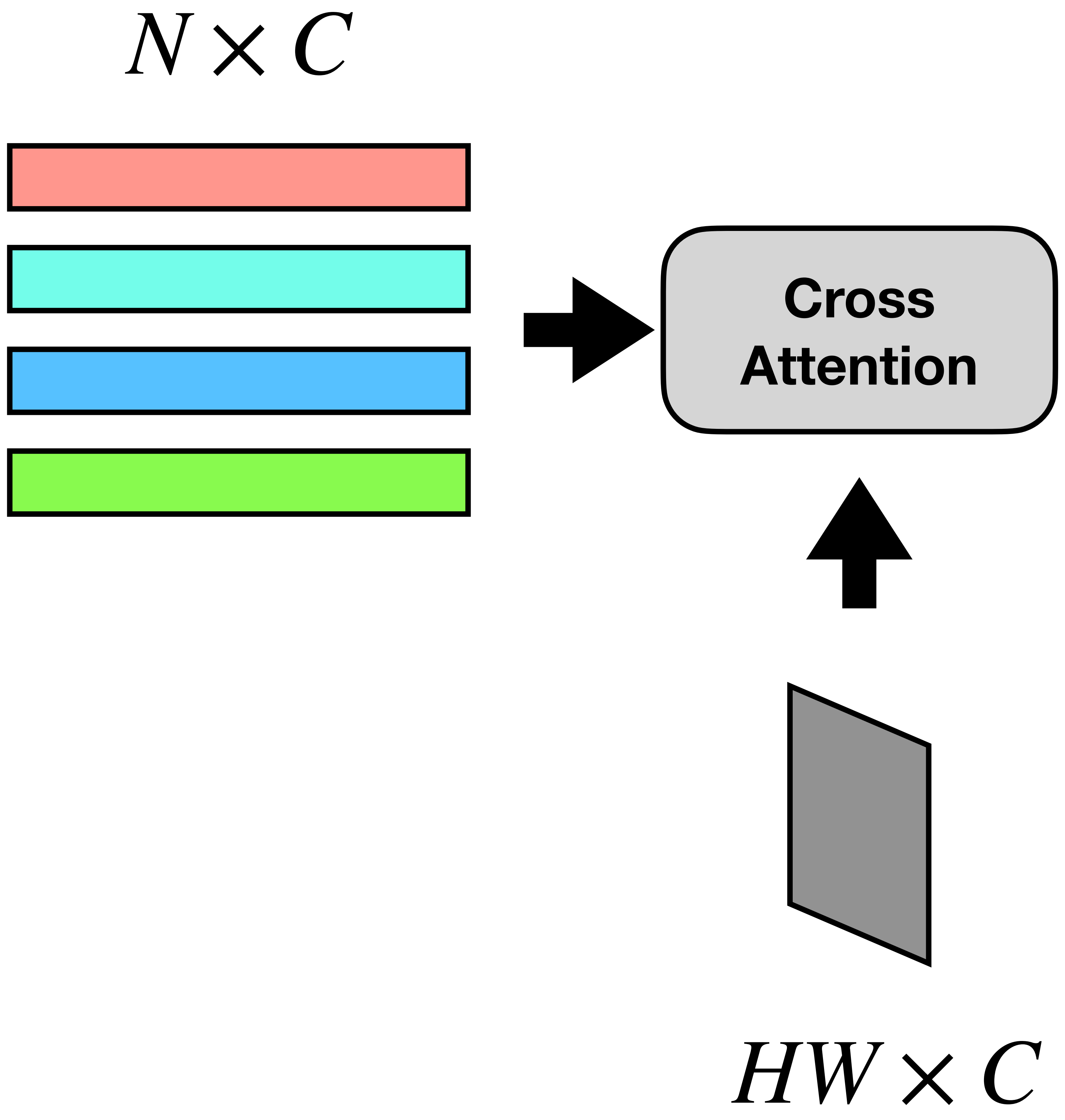
Mask2Former
Multiscale Transformer Decoder



Mask2Former CVPR22 Work

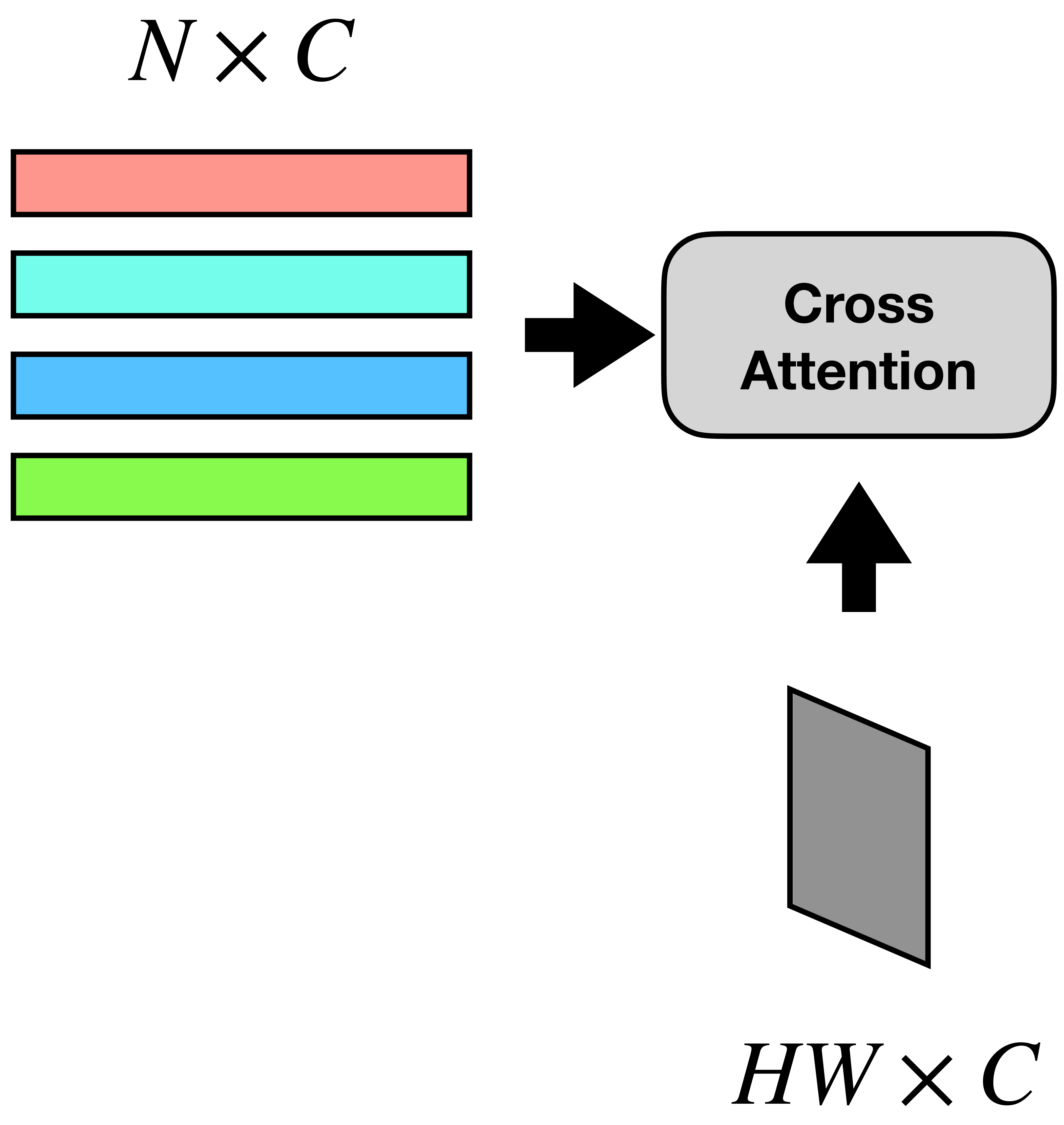
Cheng, Bowen, et al. "Masked-attention mask transformer for universal image segmentation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.

Mask2Former
Masked Attention



$$\mathbf{X}_l = \mathrm{softmax}(\mathcal{M}_{l-1} + \mathbf{Q}_l \mathbf{K}_l^{\mathrm{T}})\mathbf{V}_l + \mathbf{X}_{l-1}.$$

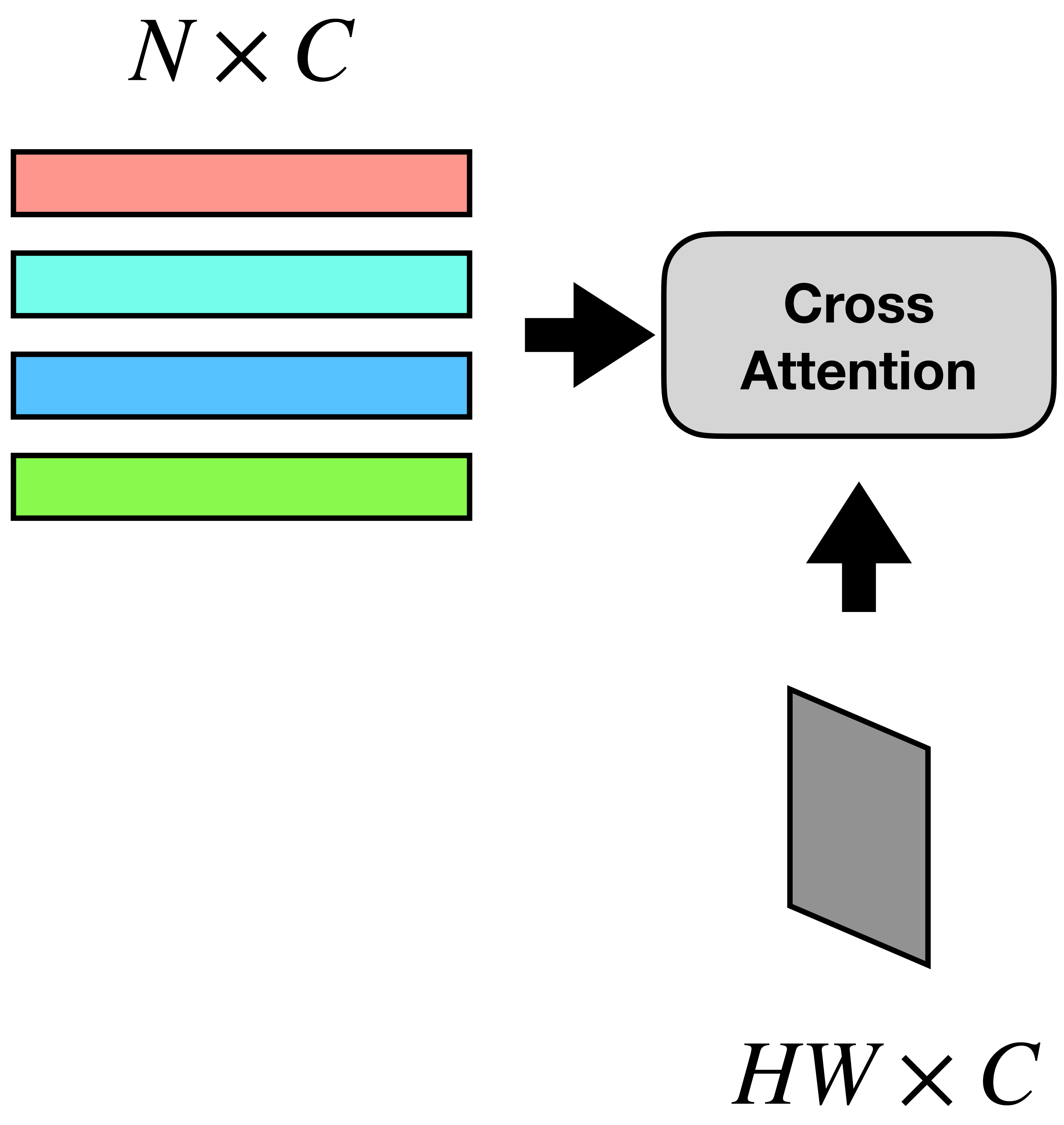
Mask2Former
Masked Attention



$$\mathbf{X}_l = \mathrm{softmax}(\boldsymbol{\mathcal{M}}_{l-1} + \mathbf{Q}_l\mathbf{K}_l^{\mathrm{T}})\mathbf{V}_l + \mathbf{X}_{l-1}.$$

$$\boldsymbol{\mathcal{M}}_{l-1}(x, y) = \begin{cases} 0 & \text{if } \mathbf{M}_{l-1}(x, y) = 1 \\ -\infty & \text{otherwise} \end{cases}$$

Mask2Former
Masked Attention



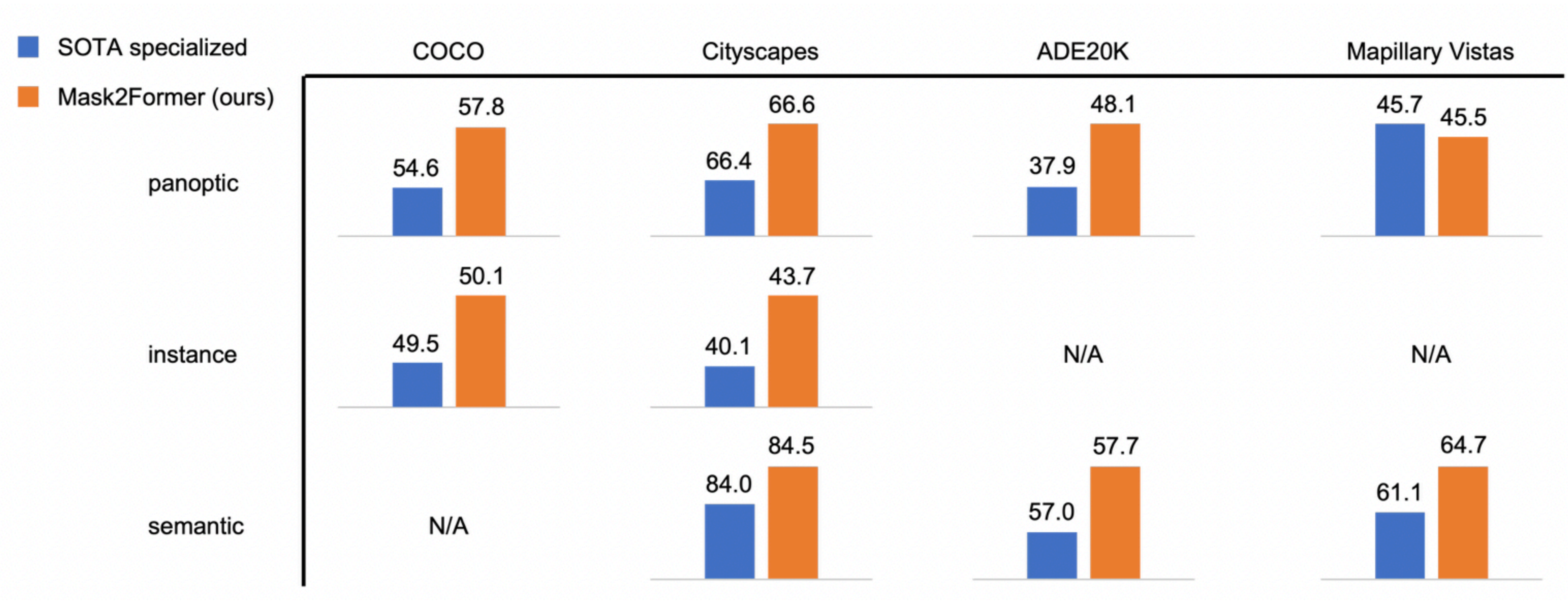
M_{l-1} $N \times HW$

Binary Mask Predictions {0, 1}

$$\mathbf{X}_l = \mathrm{softmax}(\boldsymbol{\mathcal{M}}_{l-1} + \mathbf{Q}_l \mathbf{K}_l^{\mathrm{T}})\mathbf{V}_l + \mathbf{X}_{l-1}.$$

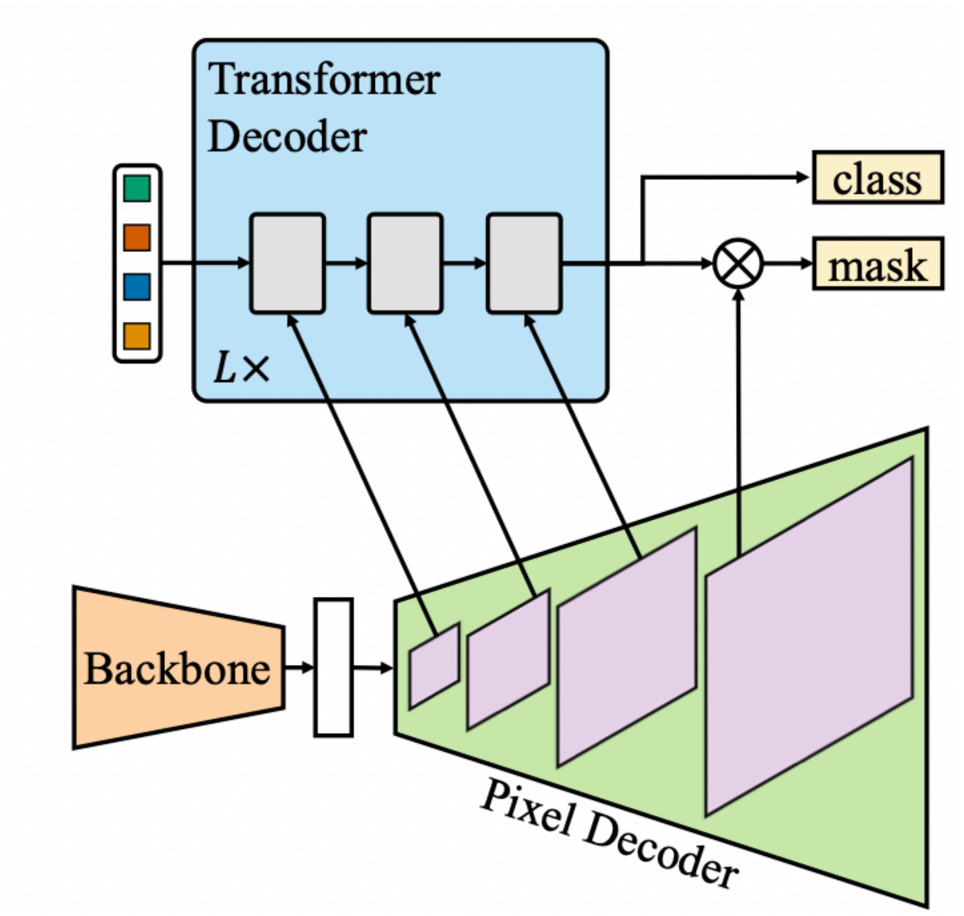
$$\boldsymbol{\mathcal{M}}_{l-1}(x, y) = \begin{cases} 0 & \text{if } \mathbf{M}_{l-1}(x, y) = 1 \\ -\infty & \text{otherwise} \end{cases}$$

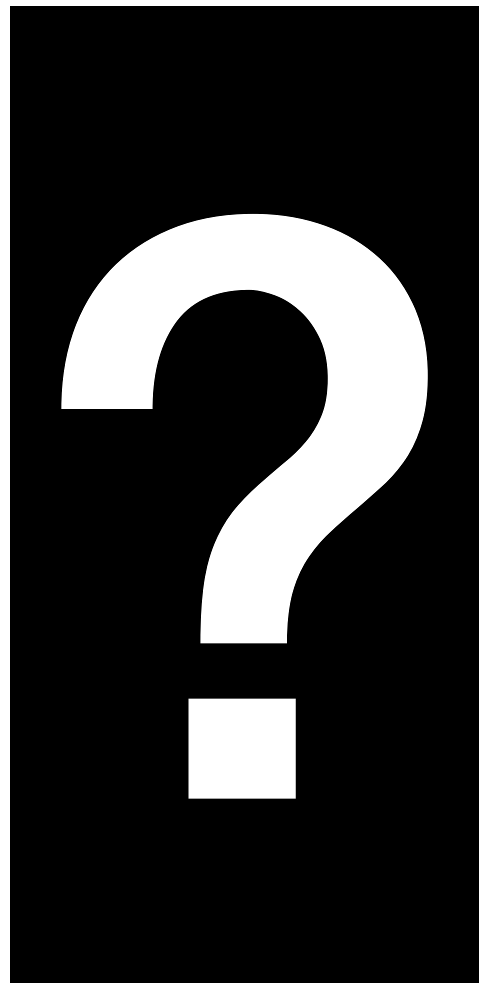
Mask2Former
Masked Cross Attention



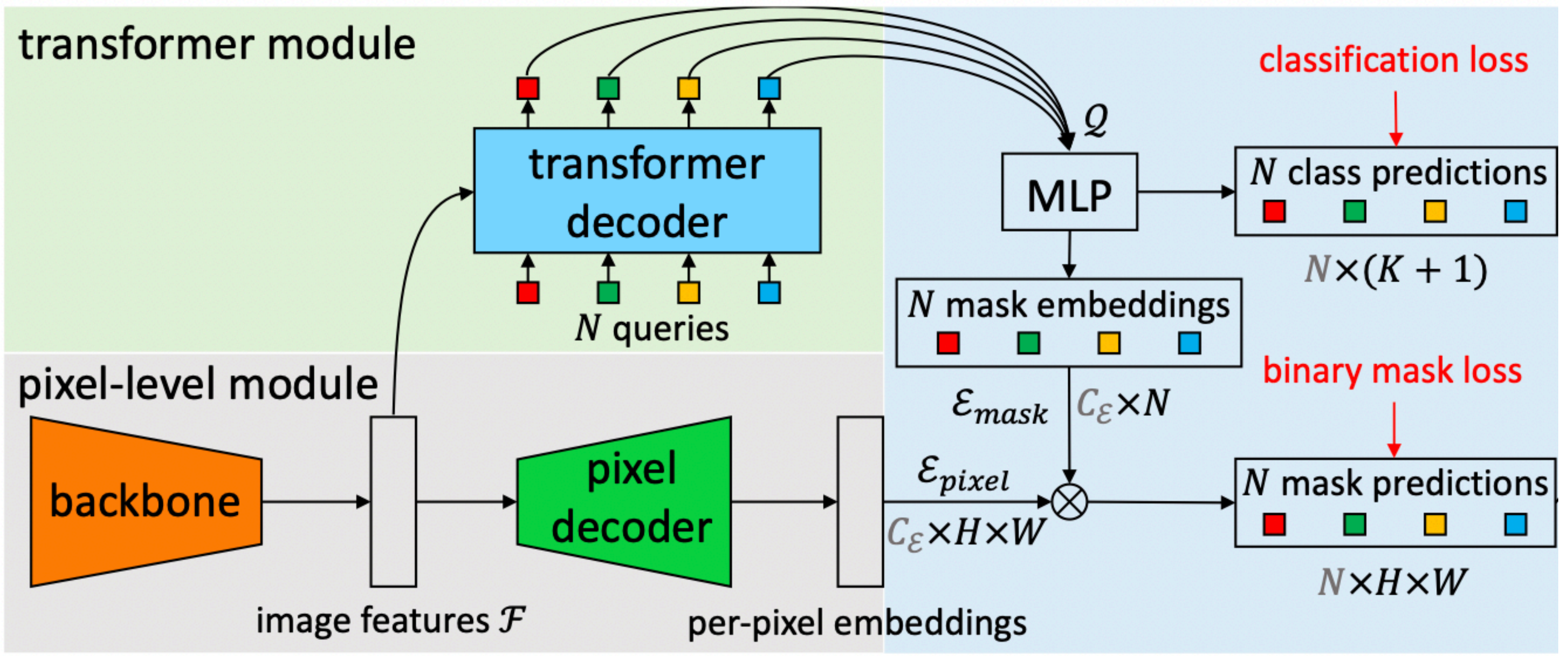
Questions

Mask2Former, CVPR'22





MaskFormer, NeurIPS'21



MED-VT, CVPR'23