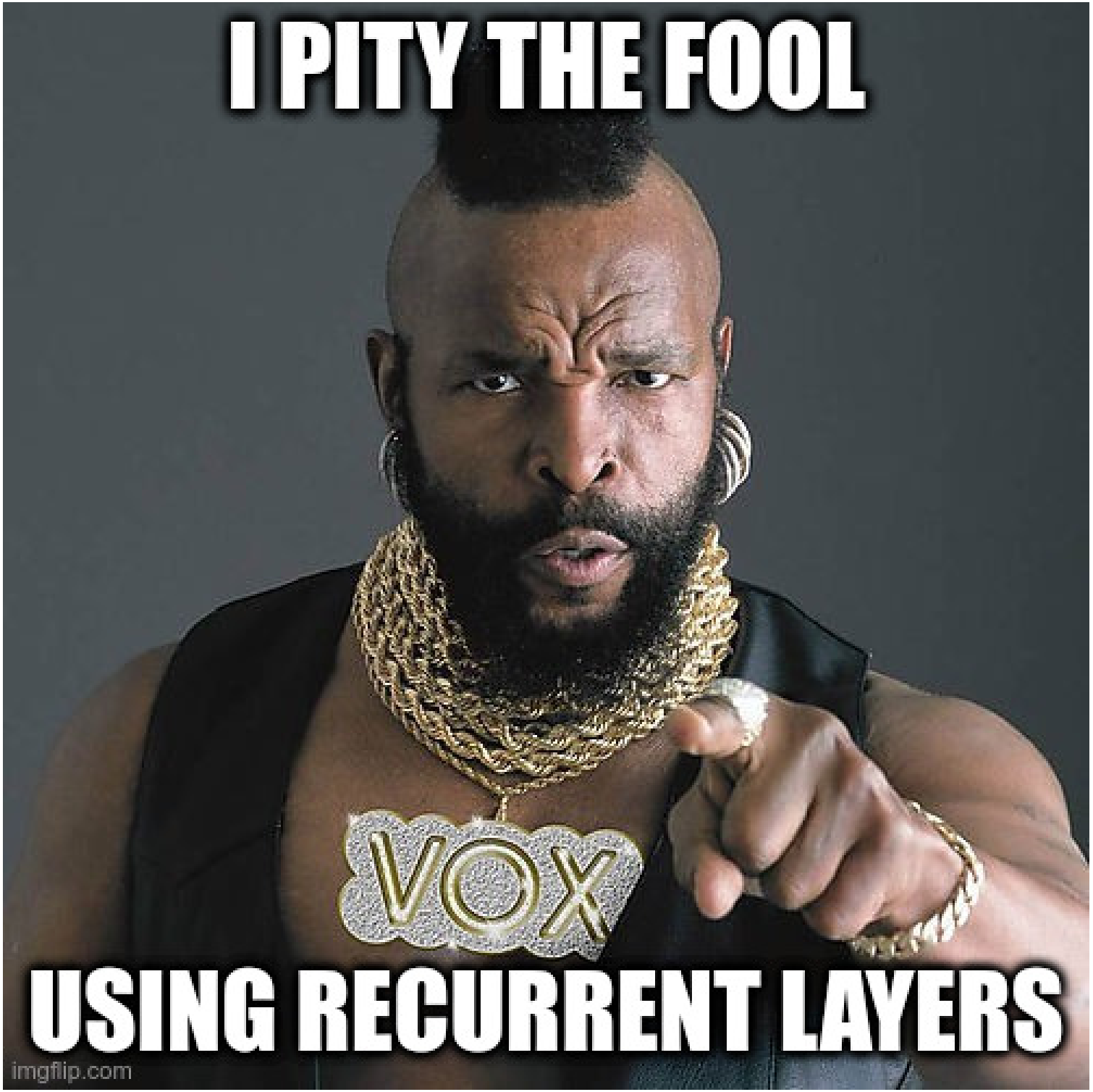
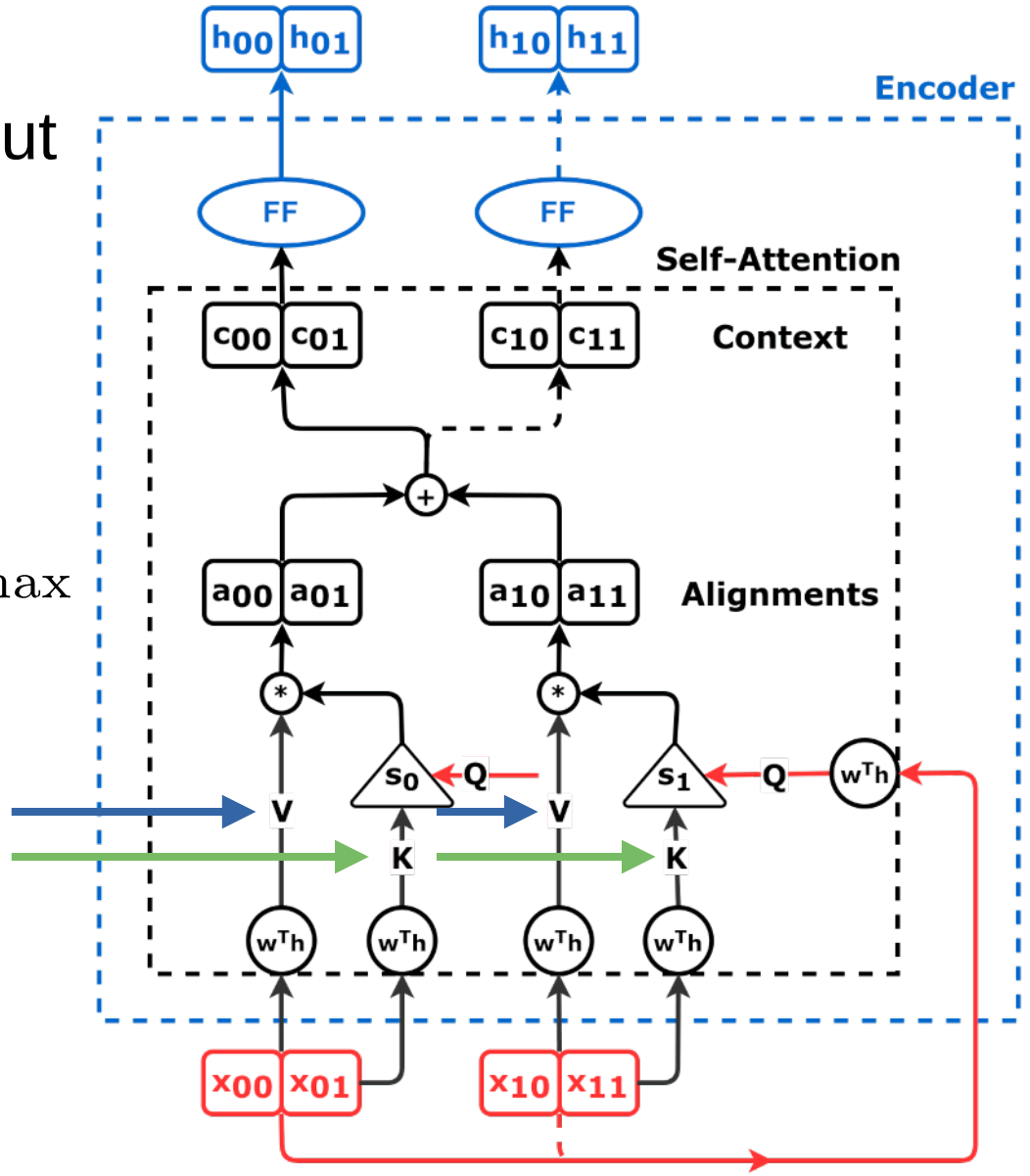
Self-Attention (Transformer)


I PITY THE FOOL
VOX
USING RECURRENT LAYERS
imgflip.com

Self-Attention

Compute V, K, Q from the input sequence $X = [x_1, \ldots, x_L]$

$$C = [W_v X] \left([W_k X]^T [W_q X] \right)_{\text{ssmax}}$$



Vaswani et al., Attention is all you need, 2017

Convolution vs Self-Attention

$$C = [W_v X] \left([W_k X]^T [W_q X] \right)_{\text{ssmax}}$$

$$c_i = \sum_j \langle q_i, k_j \rangle v_j$$

$$W_v = W_k = W_q = I$$

$$C = [X] \left([X]^T [X] \right)_{\text{ssmax}}$$

Convolution:

$$c_i = \sum_j h_{i-j} x_j$$

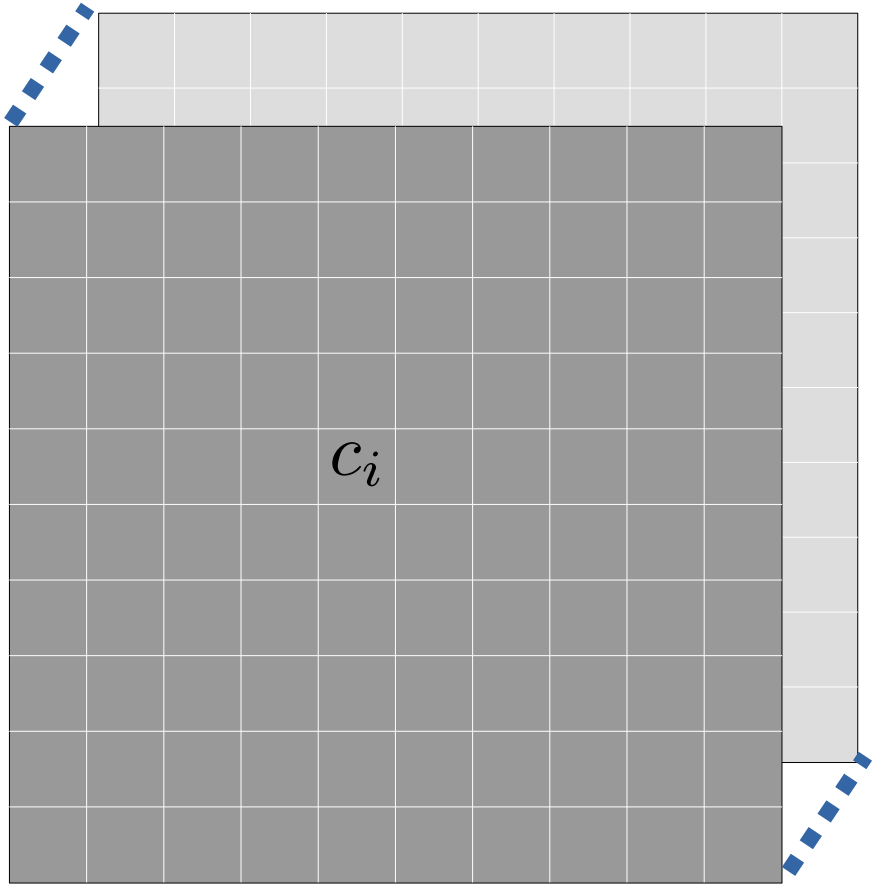
$$c_i = \sum_j \langle x_i, x_j \rangle x_j$$

With ssmax:

$$c_i = \sum_j \frac{\exp(\langle x_i, x_j \rangle)}{\sum_k \exp(\langle x_i, x_k \rangle)} x_j$$

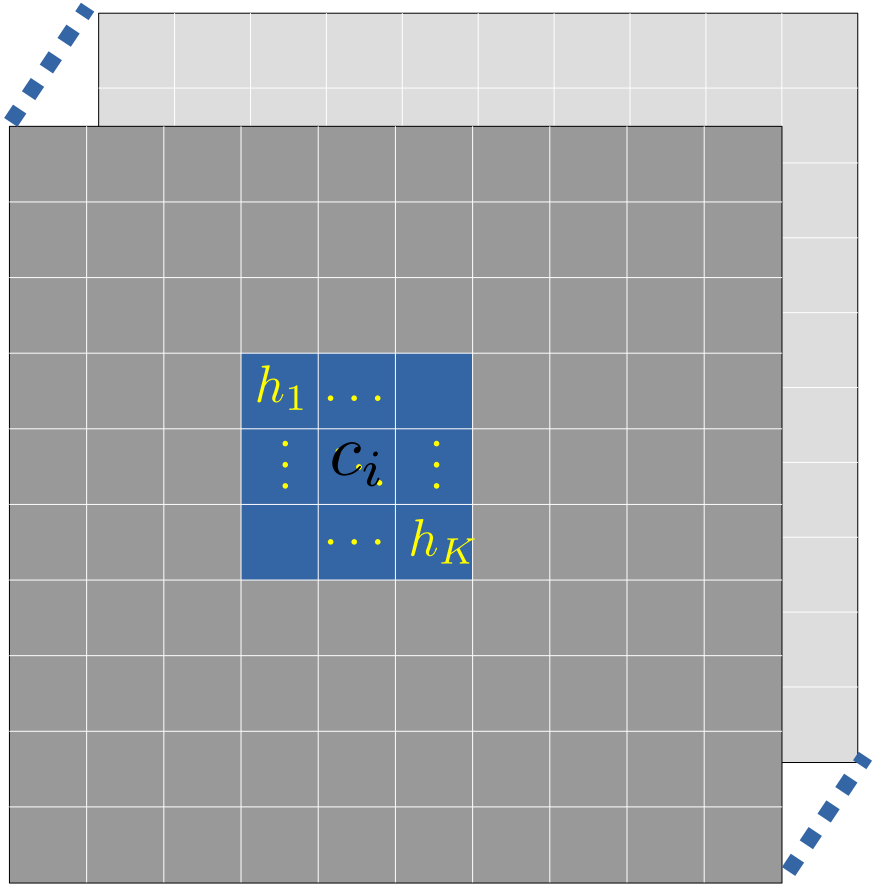
Convolution vs Self-Attention

$$c_i = \sum_j h_{i-j} x_j$$



Convolution vs Self-Attention

$$c_i = \sum_j h_{i-j} x_j$$

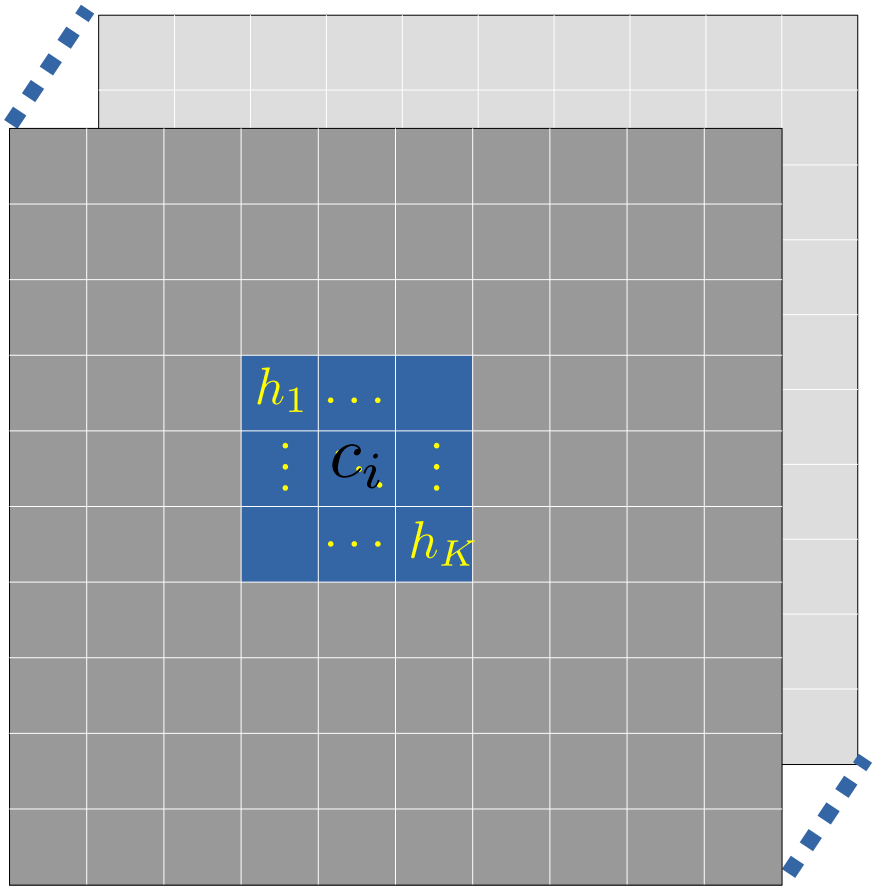


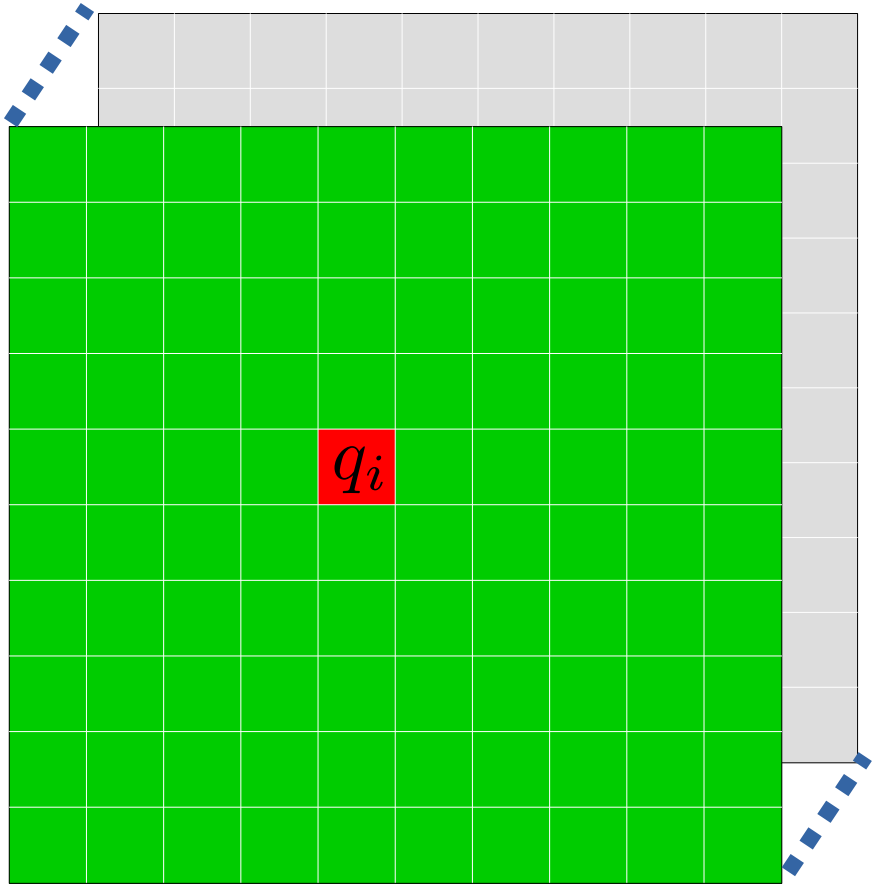
Local and fixed
during inference

Convolution vs Self-Attention

$$c_i = \sum_j h_{i-j} x_j$$

$$c_i = \sum_j \alpha_{ij} \langle q_i, k_j \rangle v_j$$



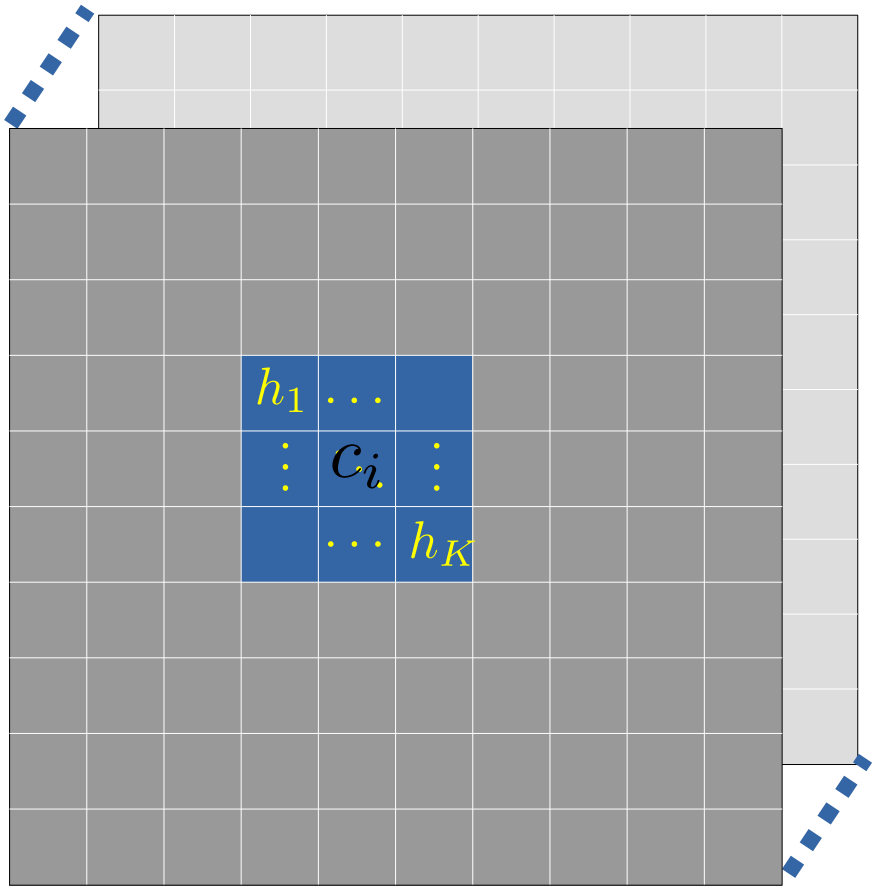


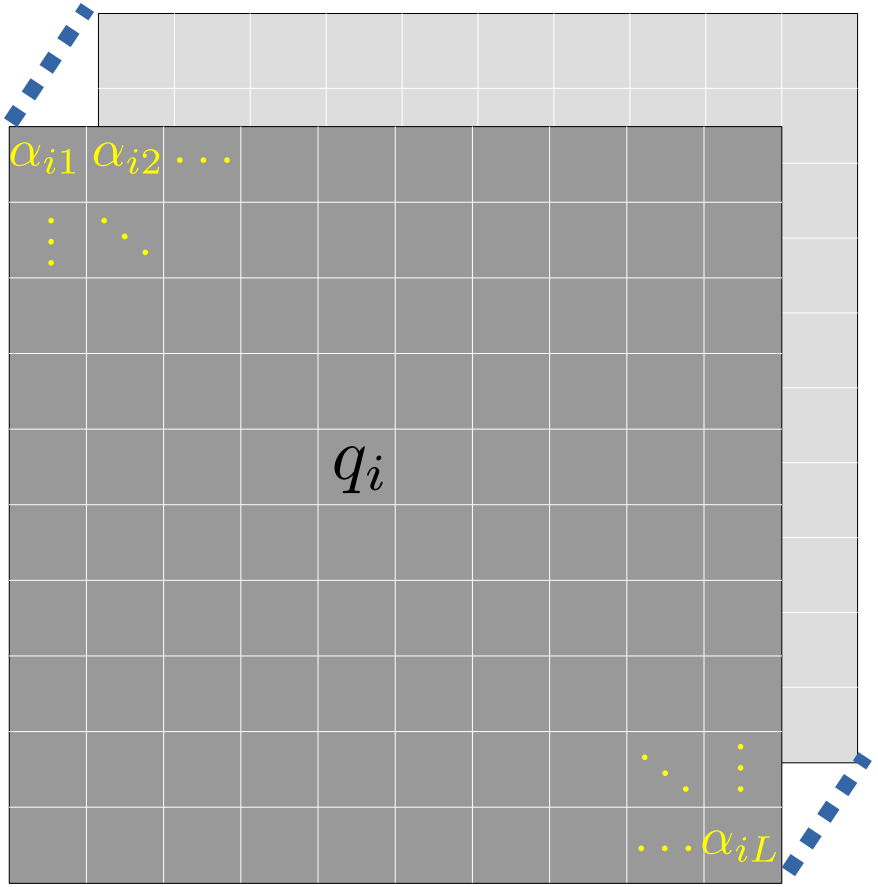
Local and fixed
during inference

Convolution vs Self-Attention

$$c_i = \sum_j h_{i-j} x_j$$

$$c_i = \sum_j \alpha_{ij} \langle q_i, k_j \rangle v_j$$



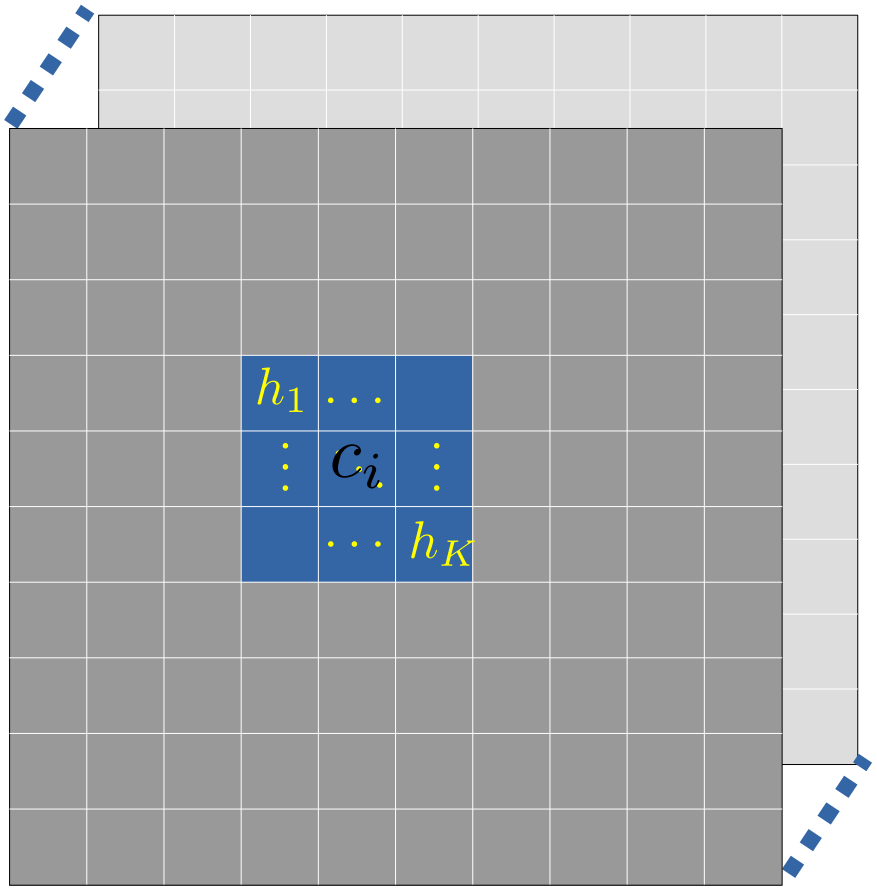


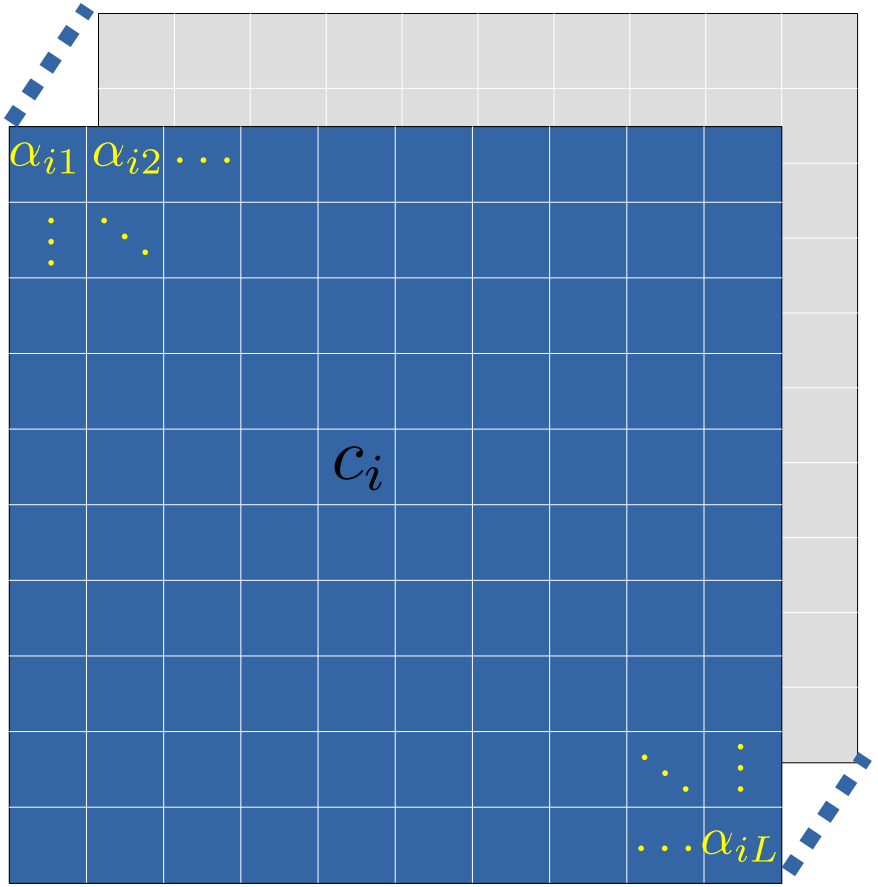
Local and fixed
during inference

Convolution vs Self-Attention

$$c_i = \sum_j h_{i-j} x_j$$

$$c_i = \sum_j \alpha_{ij} \langle q_i, k_j \rangle v_j$$





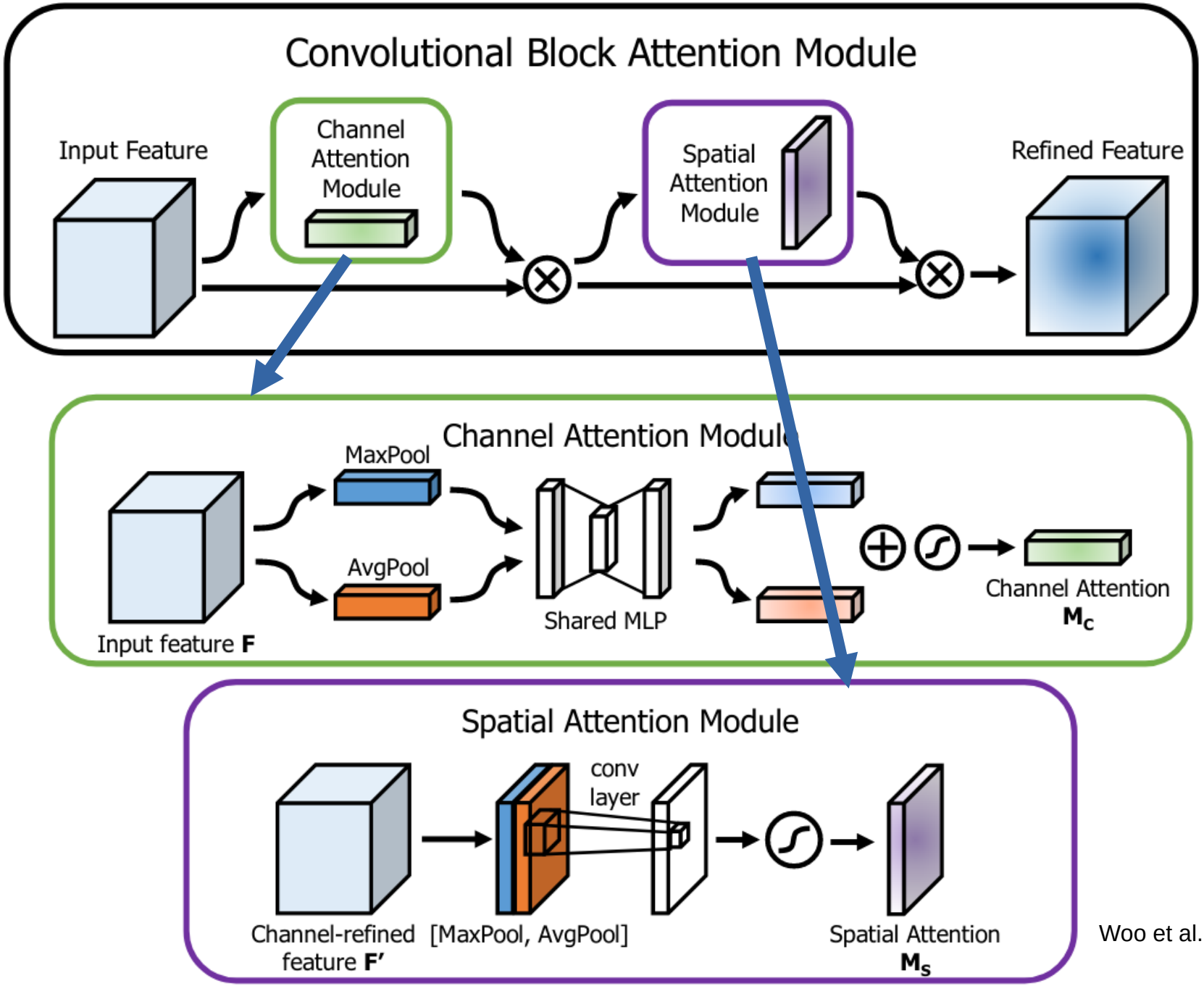
Local and fixed during inference

Global and variable during inference

Attention in Image Processing

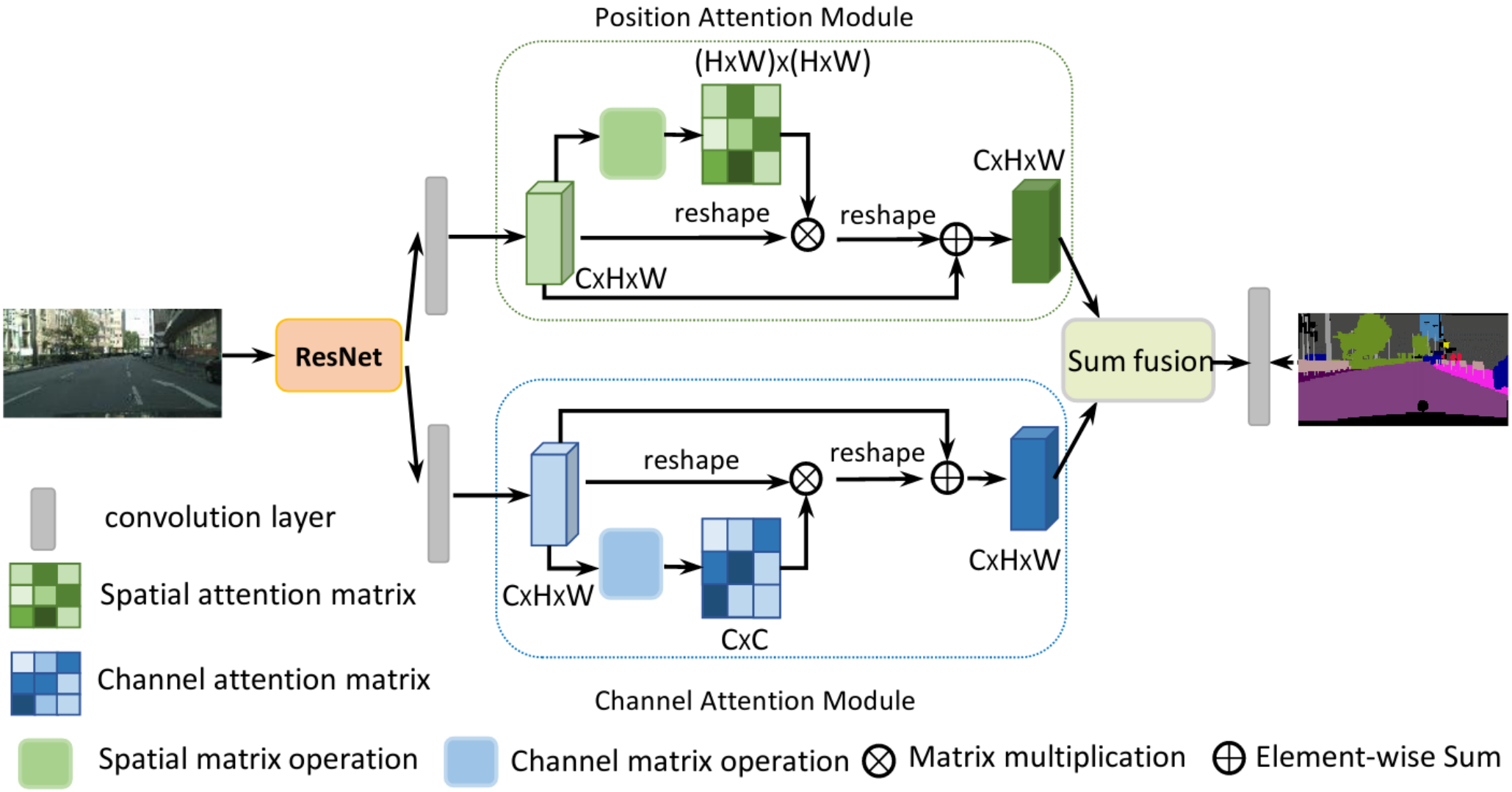
- Image Classification & Detection:
  - CBAM (Convolutional Block Attention Module)
  - Dual Attention (Spatial and Channel)
  - ViT (Vision Transformer)
  - CoAtNet (Convolution with Self-Attention)

Convolutional Block Attention Module



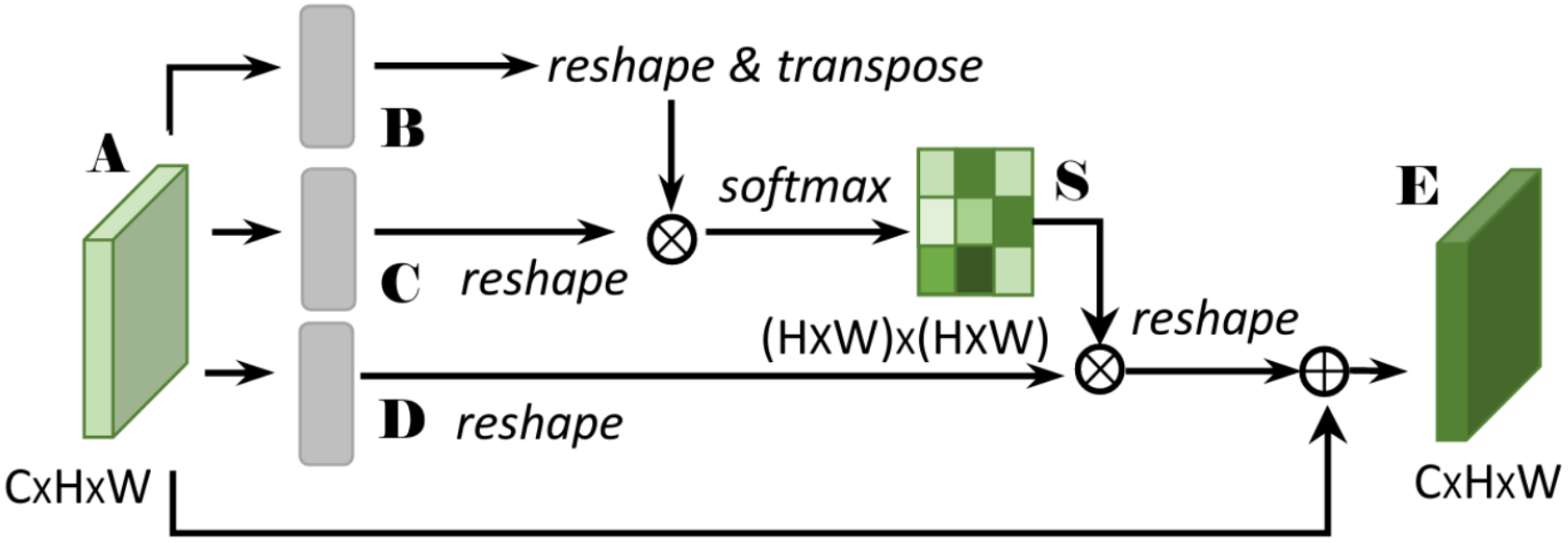
Woo et al., CBAM, 2018

Dual Attention

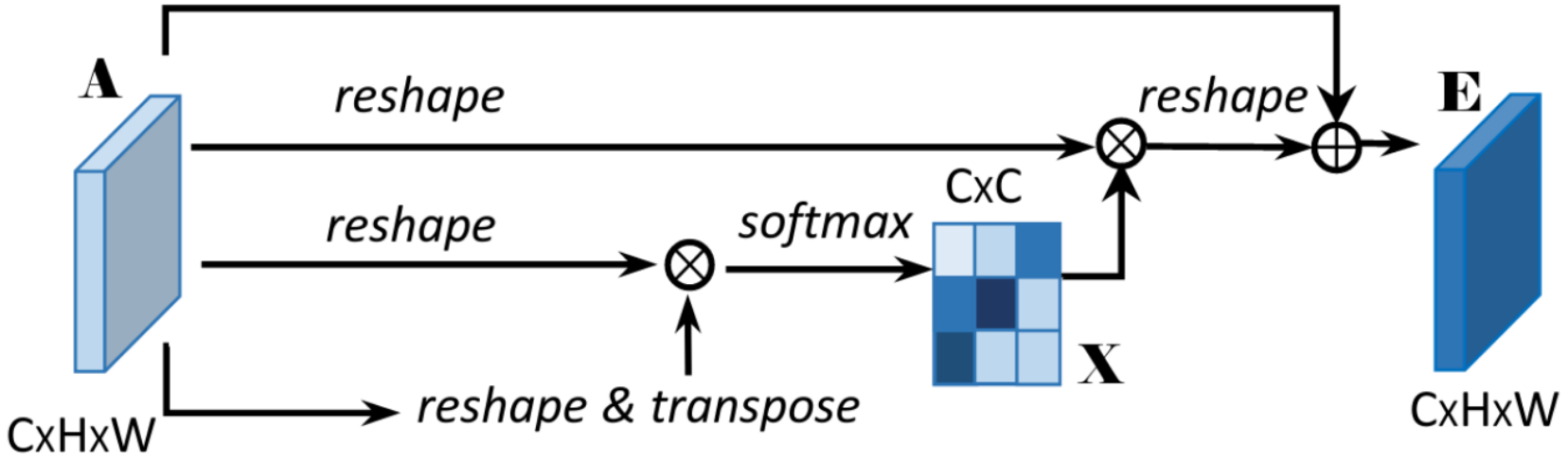


Fu et al., DANet, 2019

Dual Attention



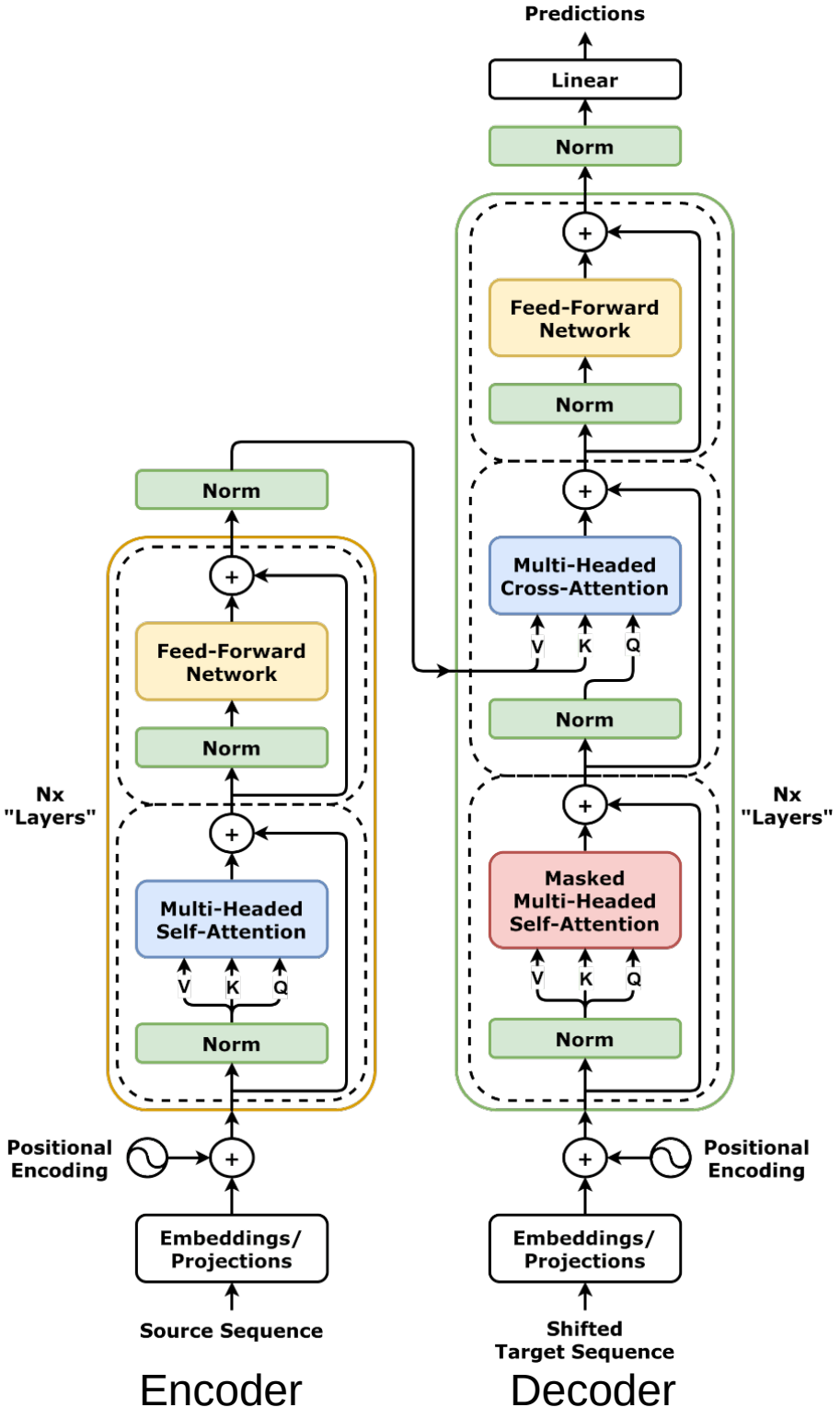
A. Position attention module



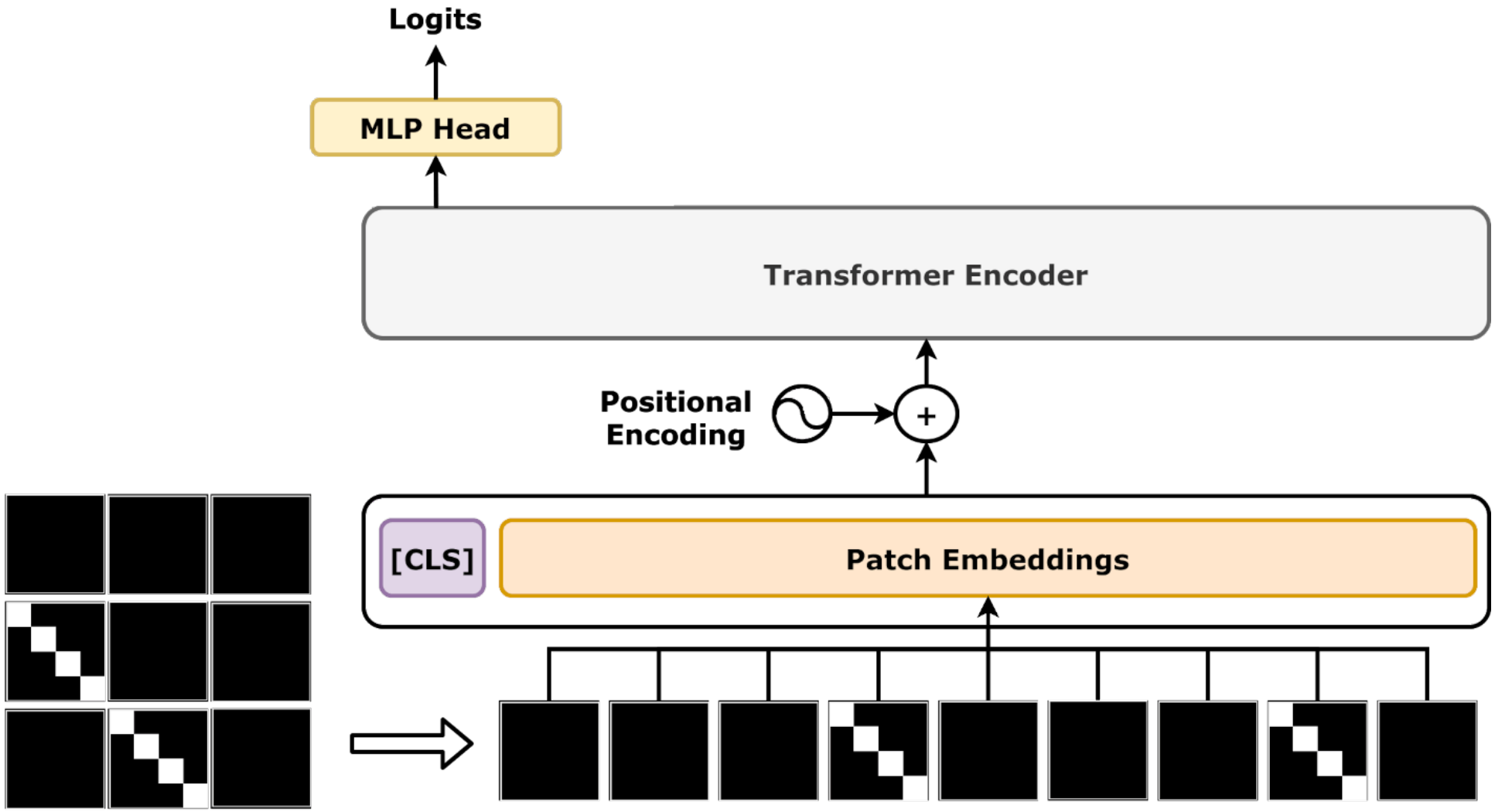
B. Channel attention module

The Transformer

- Multi-Headed Self- and Cross-Attention
- Masked Multi-Headed Self-Attention
- Layer Normalization
- Linear + ReLU
- Positional Encoding
- Residual Connection
- Dropouts



Vision Transformer (ViT)



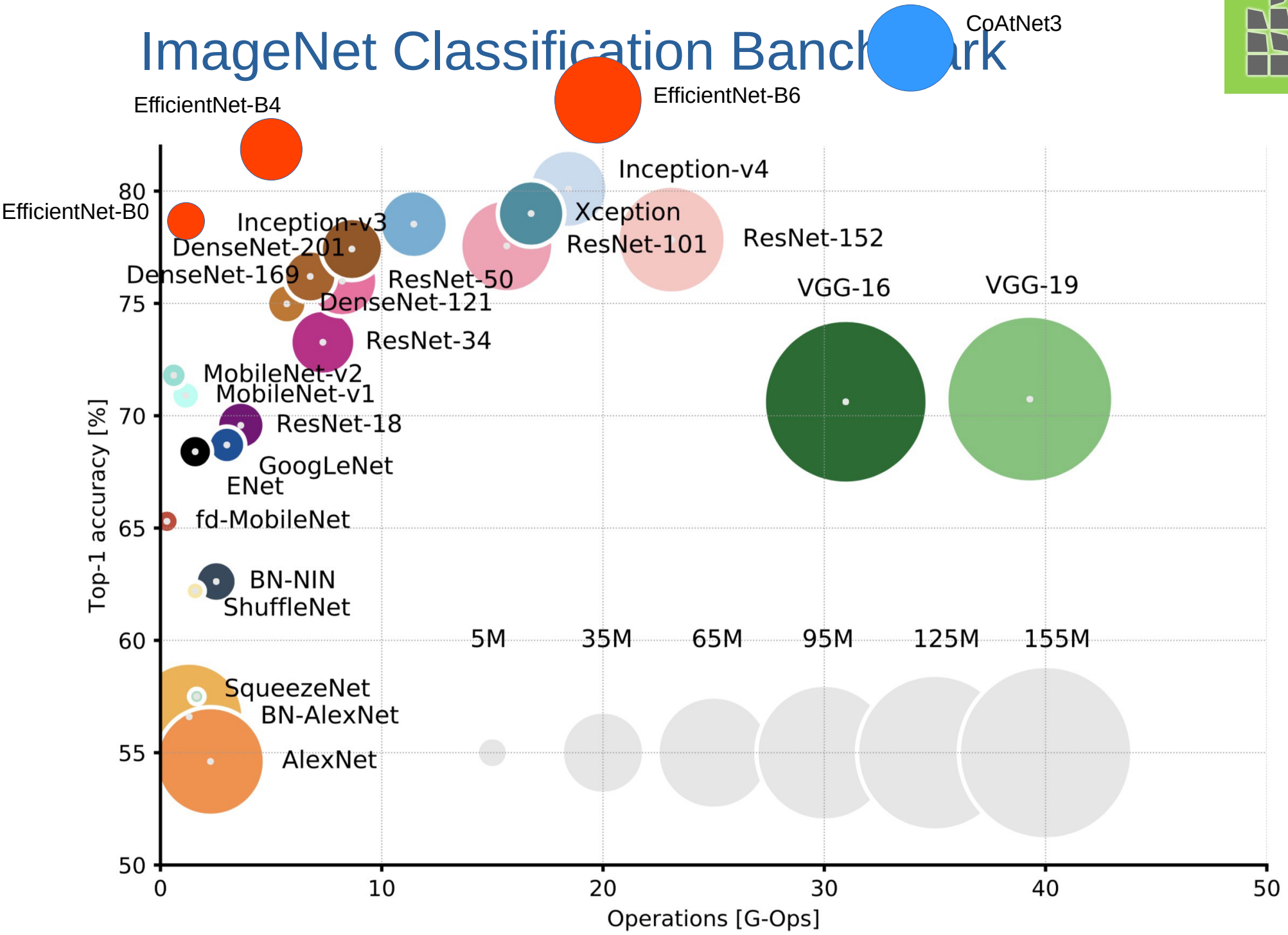
CoAtNet

- Marry convolution and attention

$$c_i = \sum_j \frac{\exp(\langle x_i, x_j \rangle)}{\sum_k \exp(\langle x_i, x_k \rangle)} x_j$$

$$c_i = \sum_j \frac{\exp(\langle x_i, x_j \rangle + w_{i-j})}{\sum_k \exp(\langle x_i, x_k \rangle + w_{i-k})} x_j$$

ImageNet Classification Benchmark



Bianco et al., IEEE Access, 2018

https://bit.ly/zoi-sem01