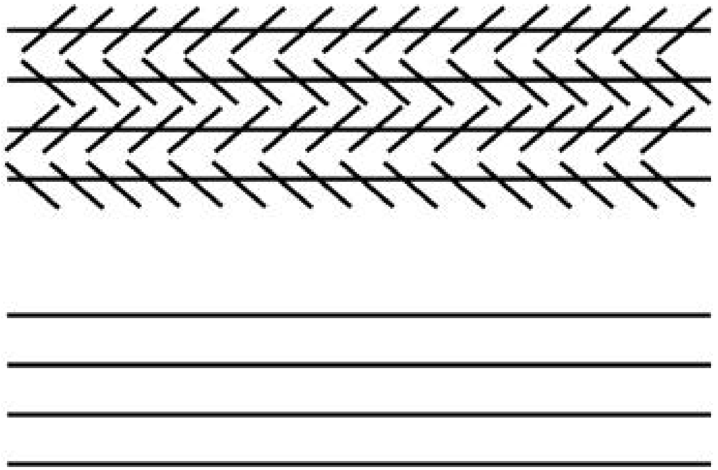
重磅！一文读懂
Attention注意力机制来龙去脉！

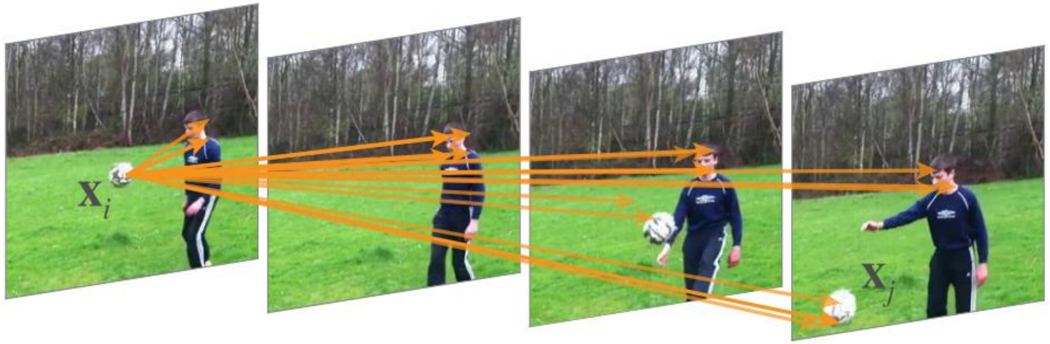
Presented by Li Ruiqi

Insights from HVS(Human Visual System)

- Bottom-up Influence
  - Factors that are low level, early, and normative
  - Light/dark contrast, edge detection, horizontality/verticality

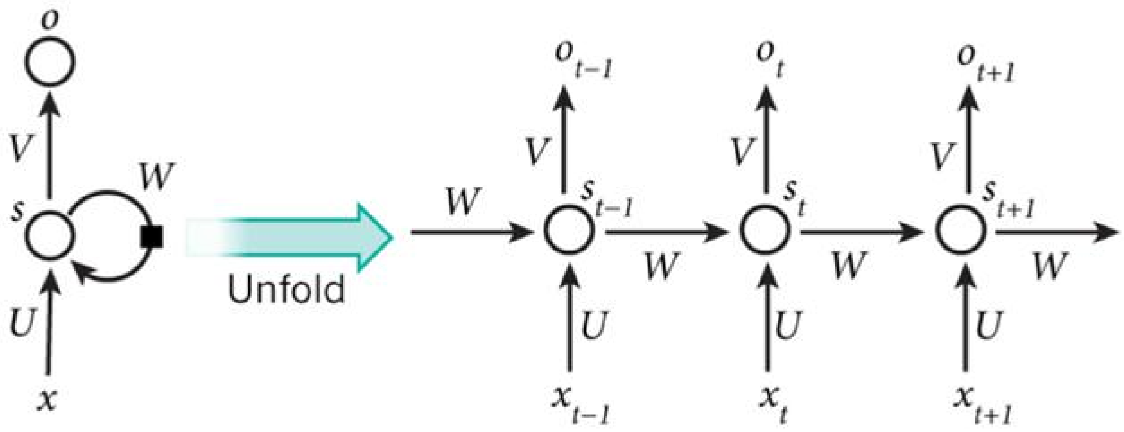


- Top Down Influence
  - High level, cognitive in nature, and individuating
  - Statement of the task, the test environment/use context, prior knowledge or experience level



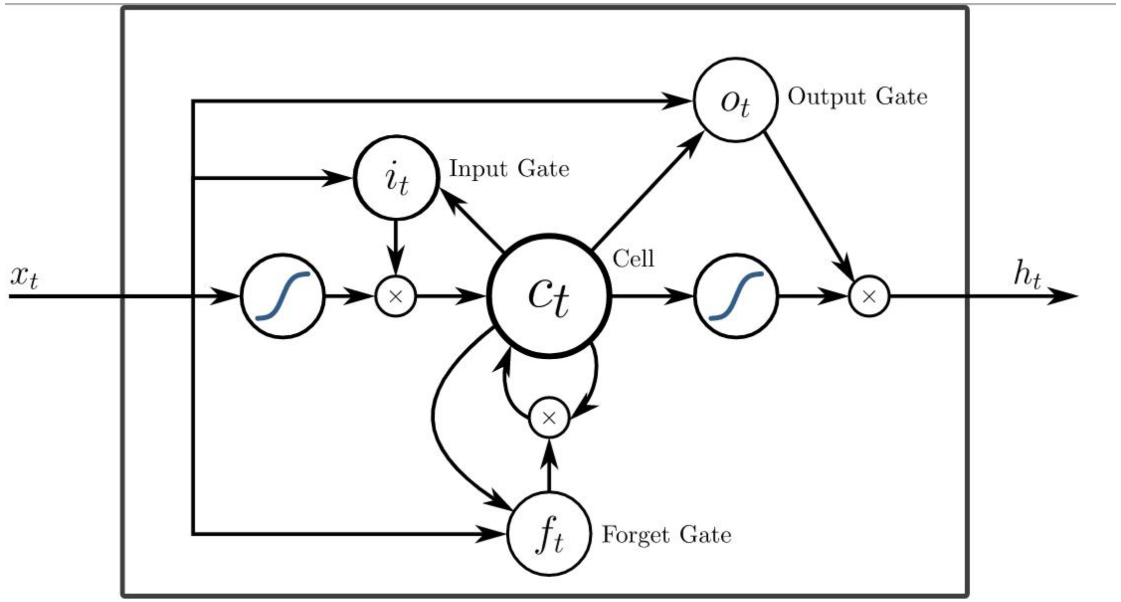
Start from Neural Machine Translation

- Recurrent Neural Network
  - Based on David Rumelhart's work in 1986
  - $$O_t = g\left(V \cdot S_t\right)$$
    $$S_t = f\left(U \cdot X_t + W \cdot S_{t-1}\right)$$



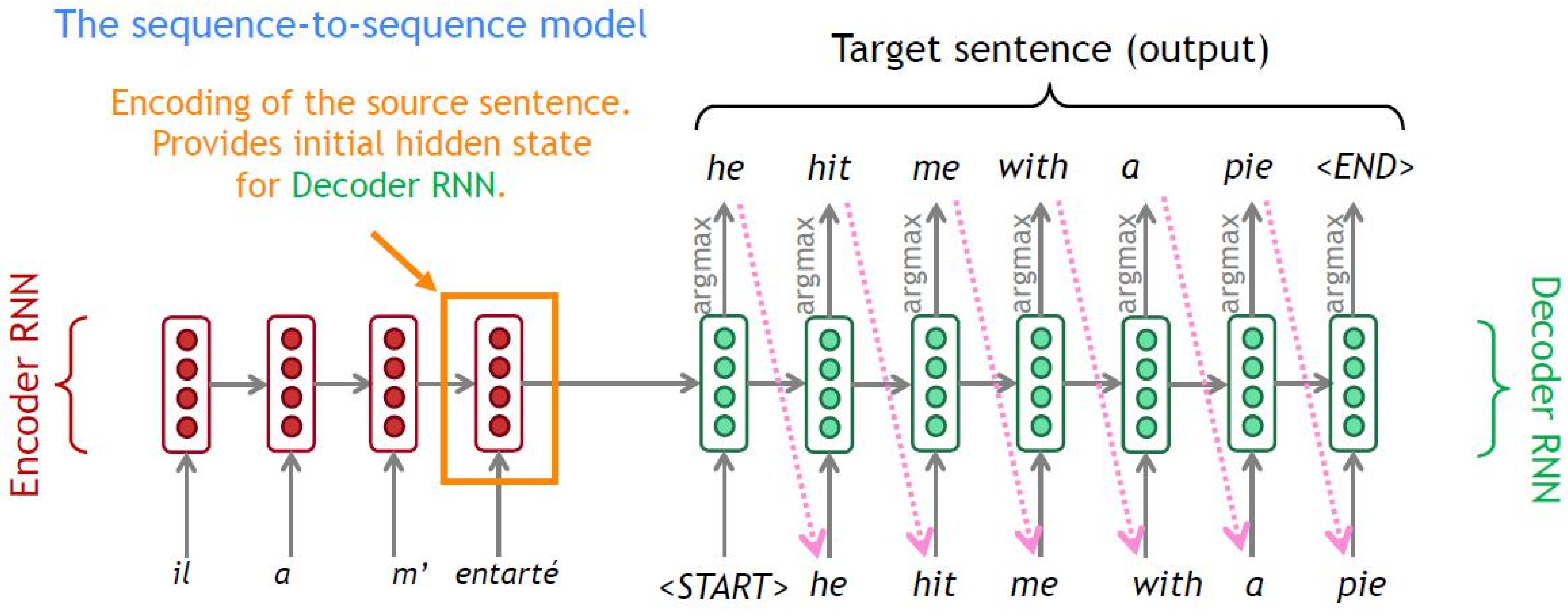
- LSTM
  - Proposed in 1997 by S. Hochreiter

$$
\begin{aligned}
f_t &= \sigma_g(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma_g(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma_g(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c}_t &= \sigma_h(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \\
h_t &= o_t \circ \sigma_h(c_t)
\end{aligned}
$$



Start from Neural Machine Translation

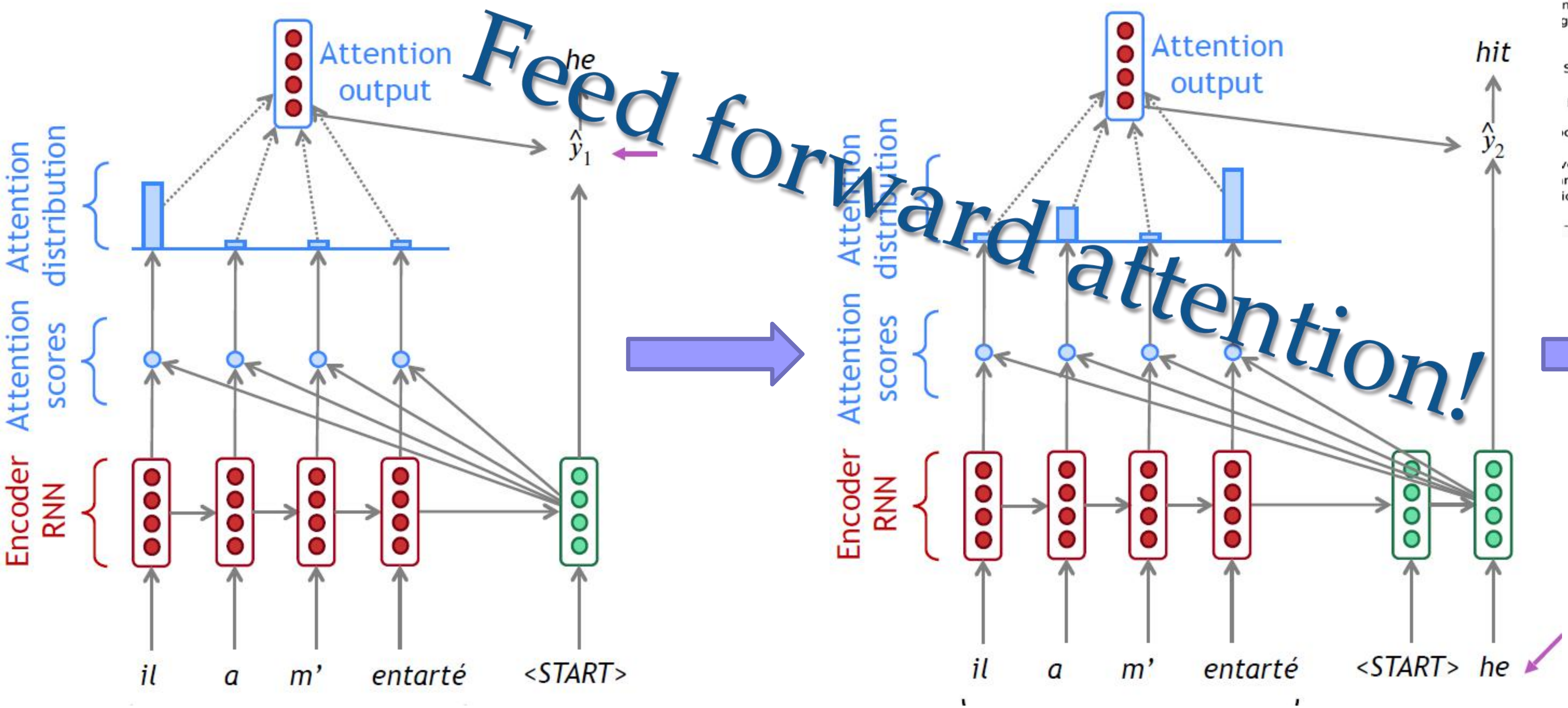
- Encoder-Decoder Structure: Seq2Seq



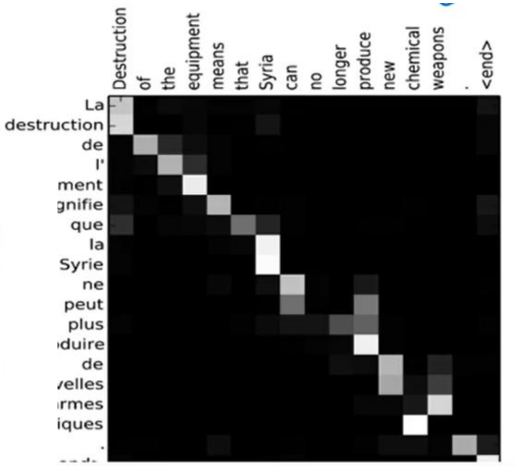
I. Sutskever et al. 2014. Sequence to Sequence Learning with Neural Networks. NIPS.

Attention in Seq2Seq

- NMT by Jointly Learning



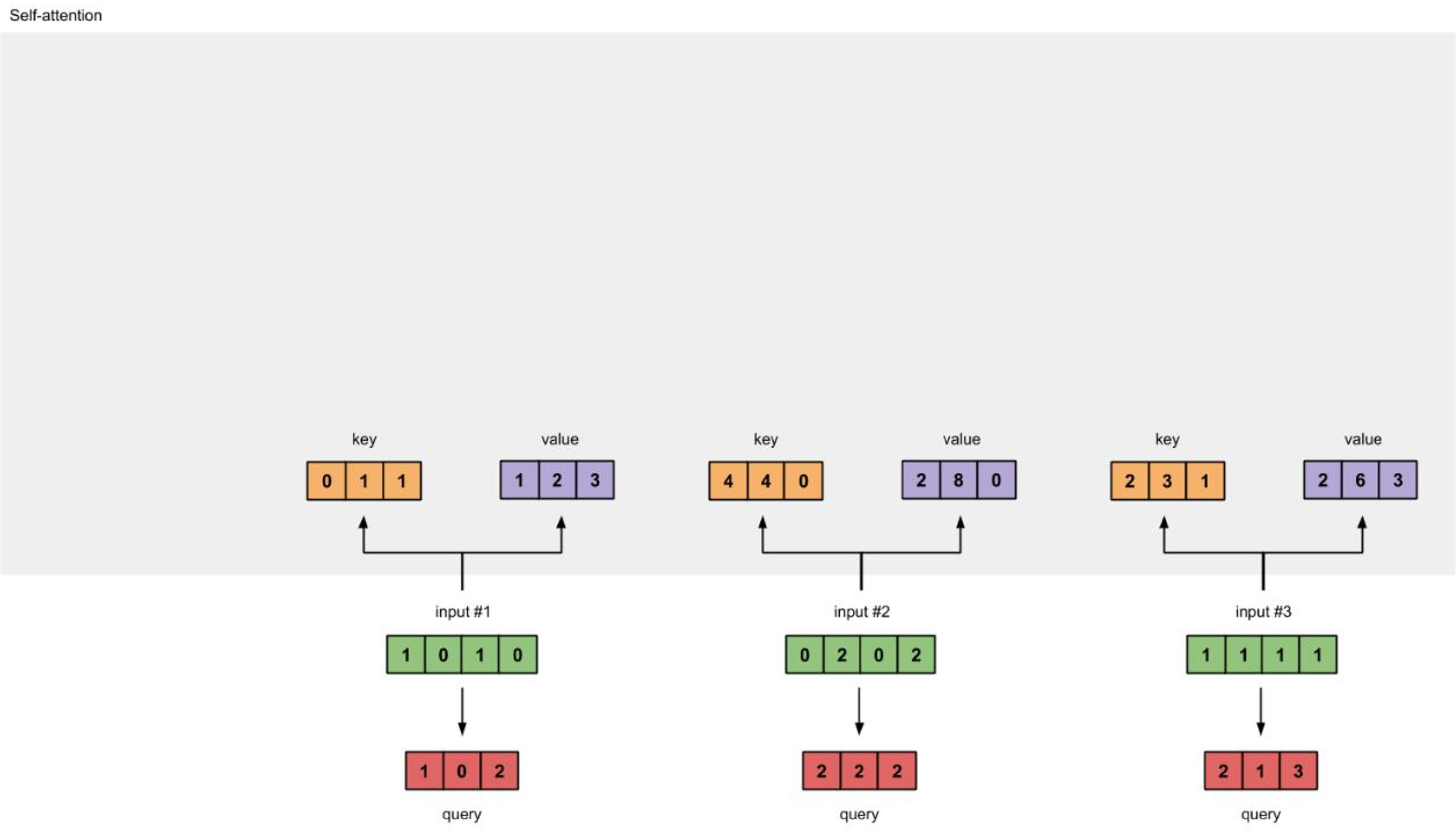
Feed forward attention!



……

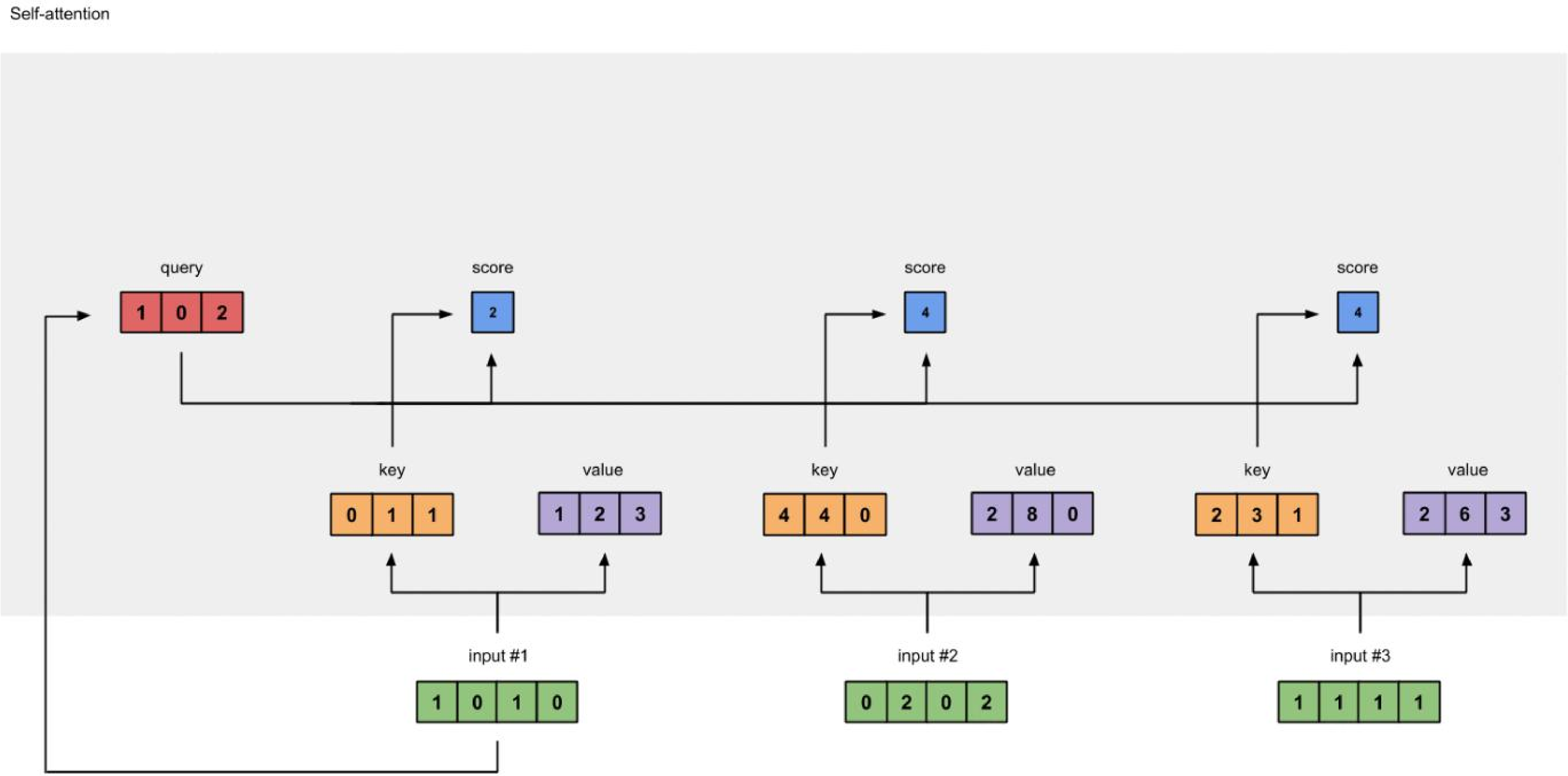
D. Bahdanau, 2015. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR.

Towards Transformer: Query, Key and Value



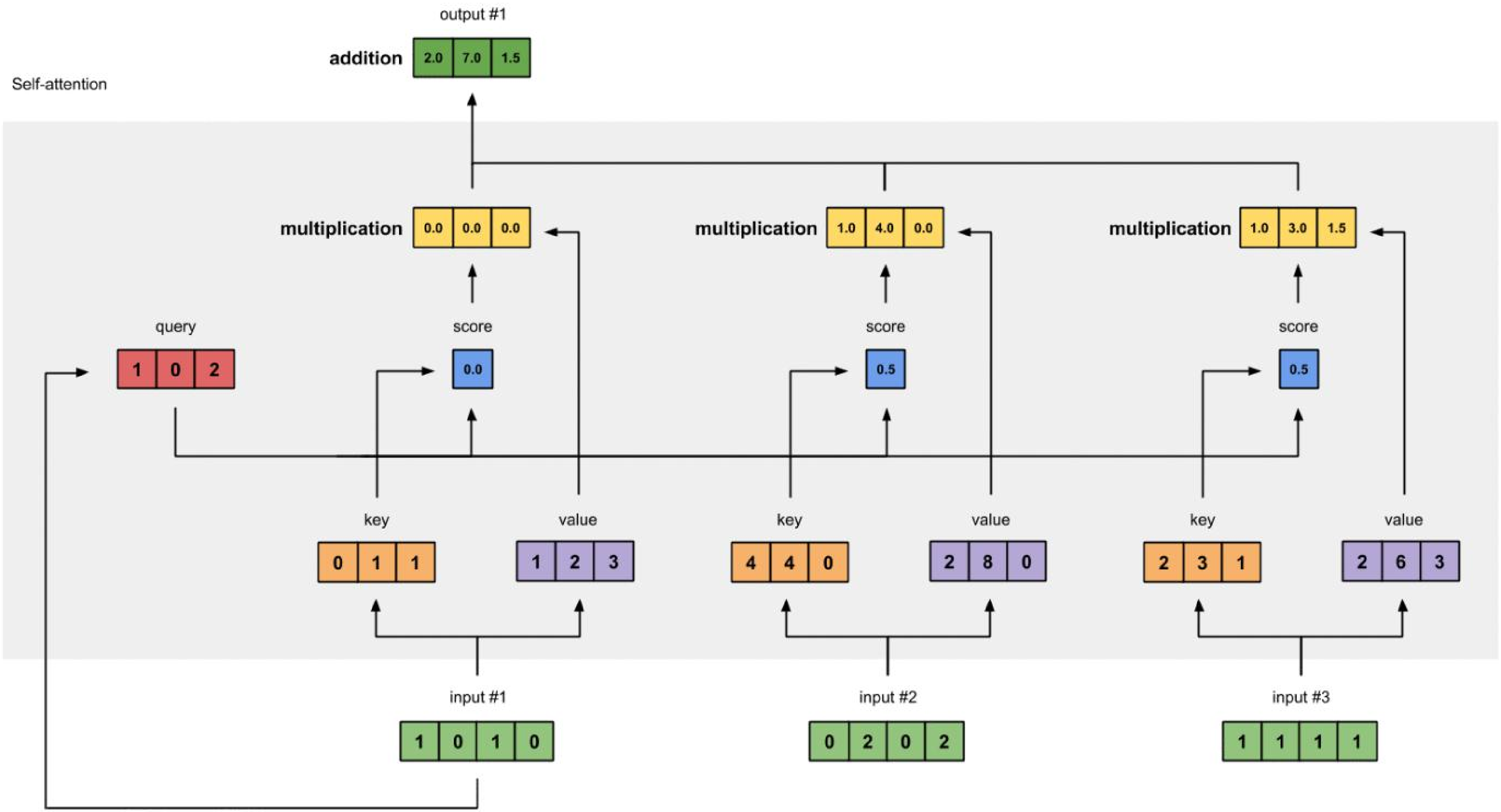
https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

Towards Transformer: Query, Key and Value



https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

Towards Transformer: Query, Key and Value



https://towardsdatascience.com/illustrated-self-attention-2d627e33b20a

Define Types of Attentions

- Feed Forward Attention
  - Query is learnable, donate as w
  - Key=Value!
  - $\alpha = Softmax(w^T v_1, w^T v_2, \cdots, w^T v_K)$
  - $Attn(\{v_i\}_{i=1}^K) = \sum_{i=1}^K \alpha_i v_i$
- Self Attention
  - Query=Key=Value
  - $Attn(V) = softmax_{row}(VV^T)V$

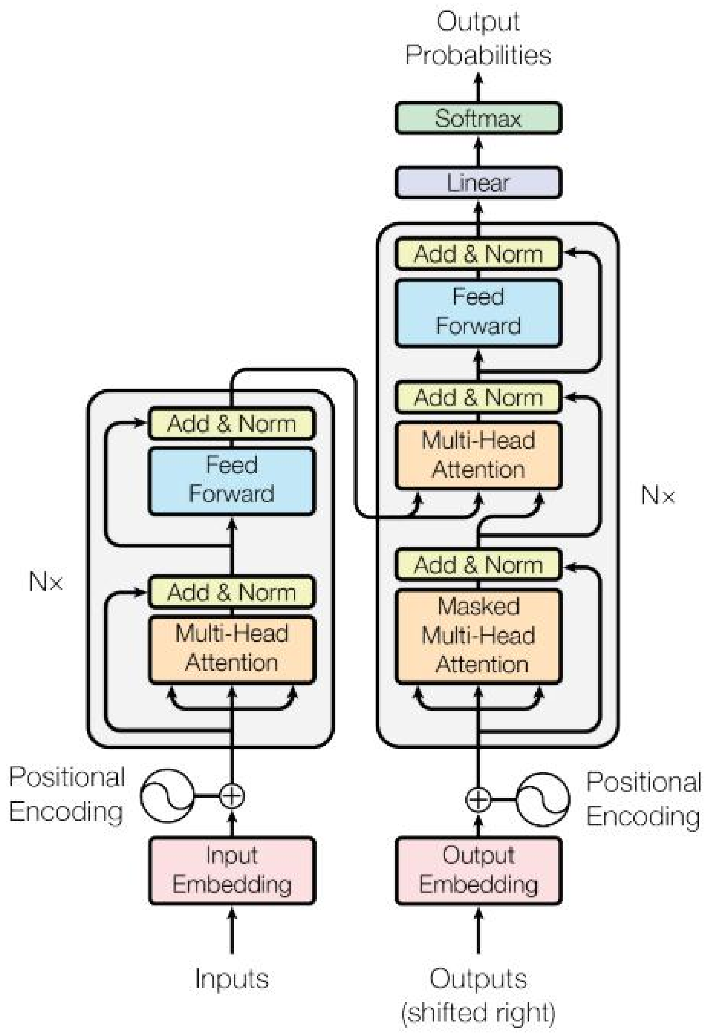
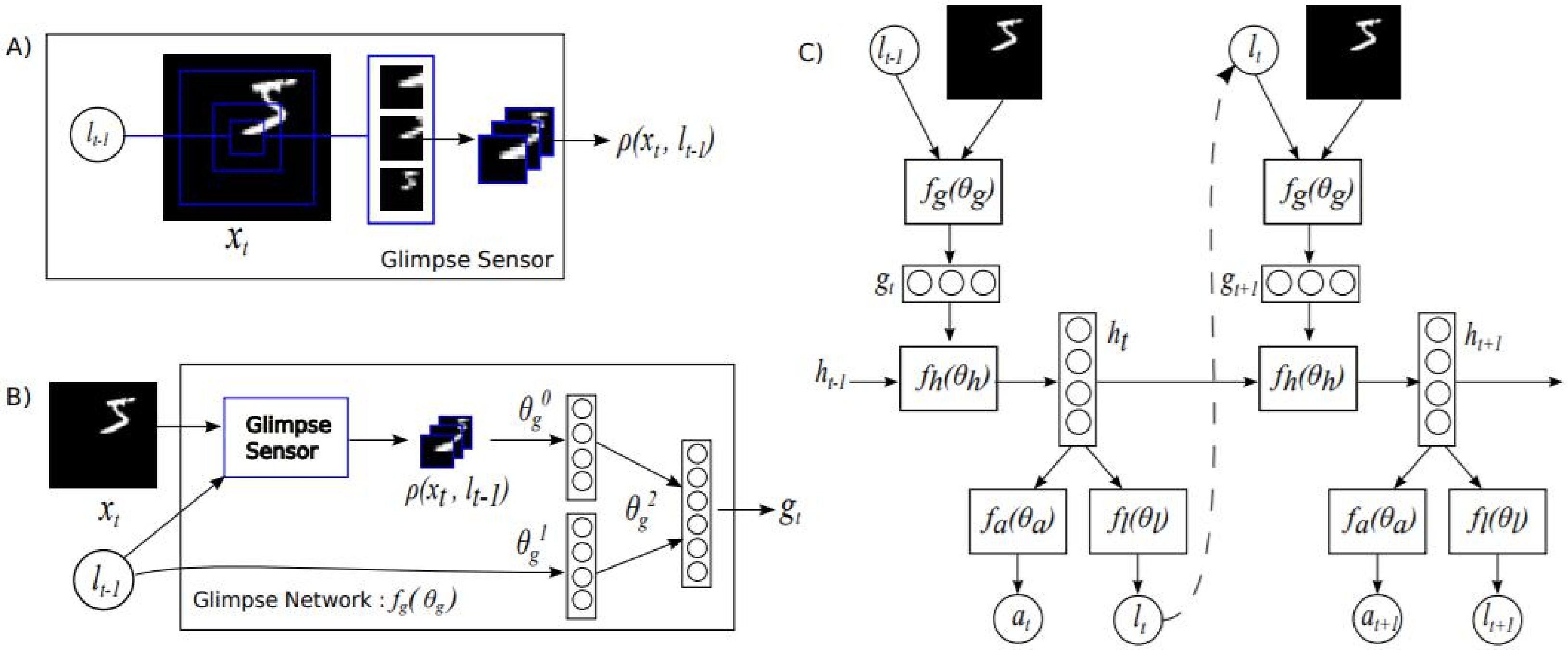


Figure 1: The Transformer - model architecture.

https://mp.weixin.qq.com/s/t6IboWbX5ztdscDqUjdxXg
https://www.bilibili.com/video/av48285039?p=92
C. Raffel et al. 2015. Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems. ICLR.

Attention in Vision

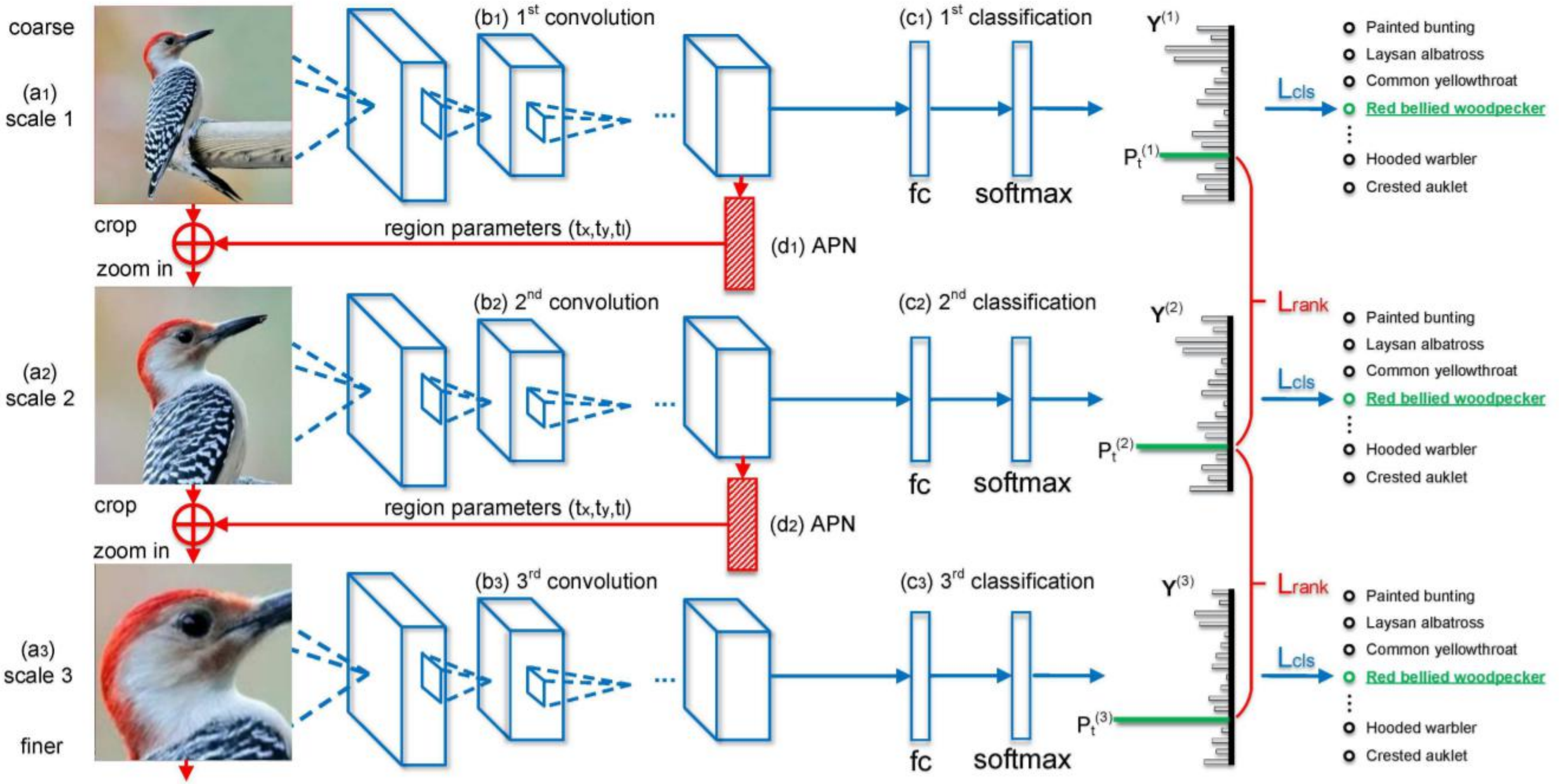
- Recurrent Models of Visual Attention
  - Based on pure RNN, image/video classification
  - l_t: attention location; a_t: classification



V. Mnih et al. 2014. Recurrent models of visual attention. NIPS.

Attention in Vision: Spatial

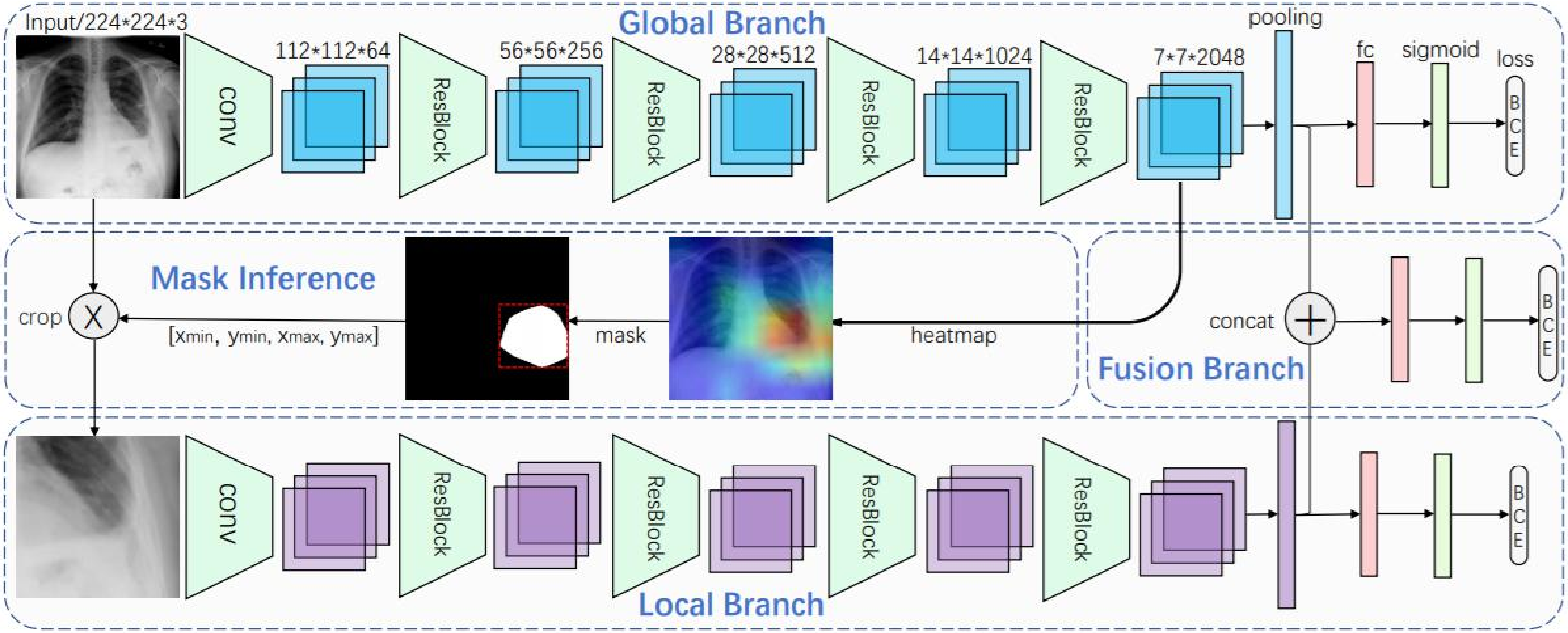
- Look Closer to See Better
  - APN inspired by RPN
  - Training:
    - keep APN, optimize L_{cls}
    - fix params, optimize L_{rank}
  - APN inspired by RPN



J. Fu et al. "Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition," 2017, CVPR.

Attention in Vision: Spatial

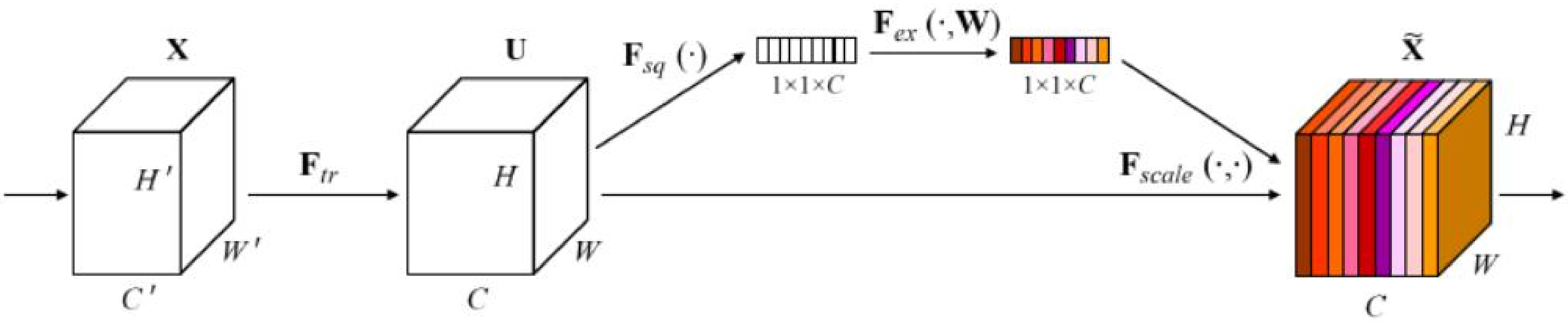
- Attention Guide CNN: Medical Image Analysis



Guan, Qingji, et al. "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification.", 2018

Attention in Vision: Channel

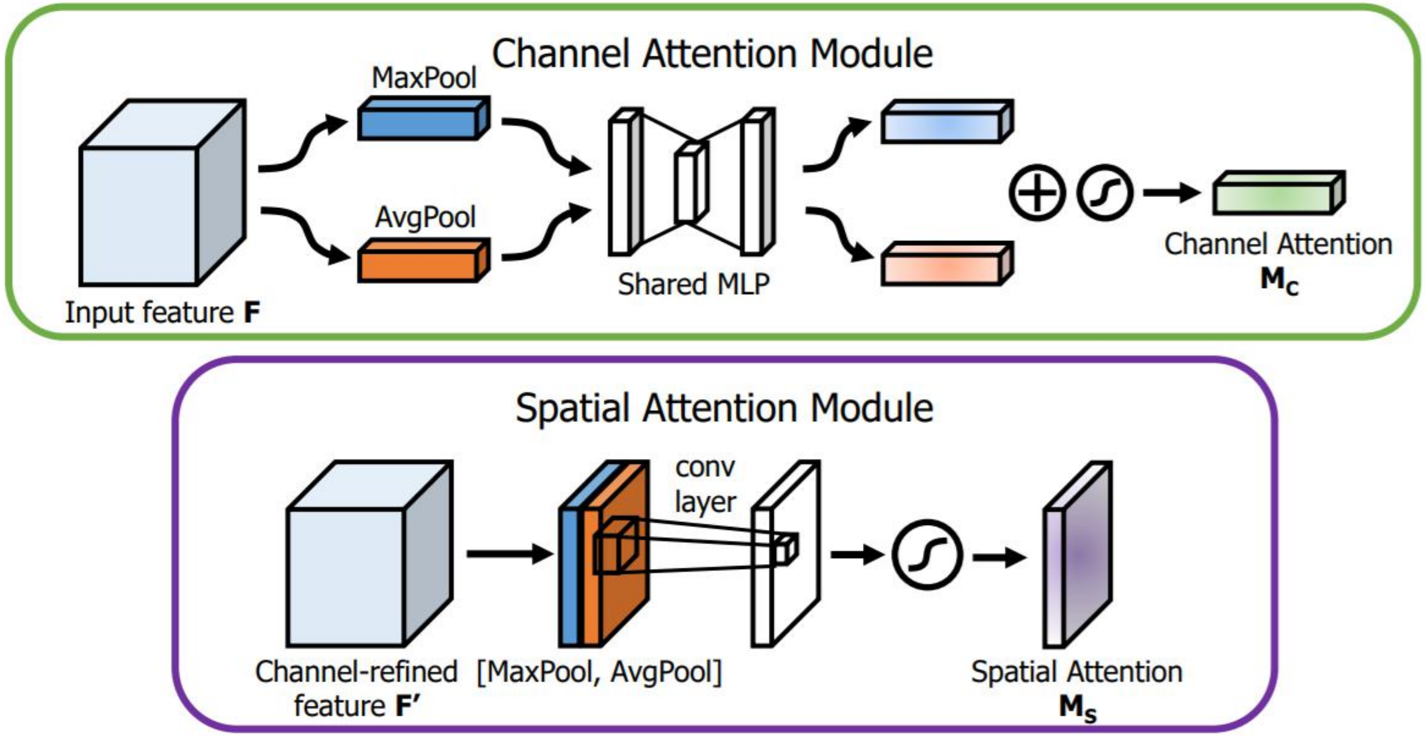
- Squeeze and Excitation Network
  - $$z_c = \mathbf{F}_{sq}(\mathbf{u}_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j). \quad \mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z}))$$
  - Main idea: include global information

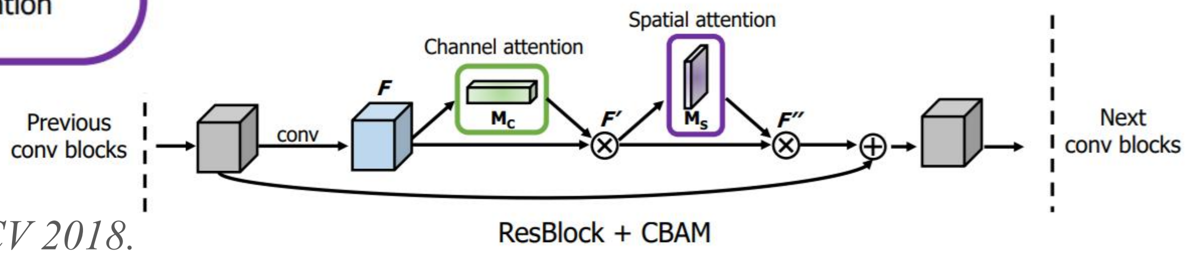


J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," CVPR, 2018.

Attention in Vision: Channel and Spatial

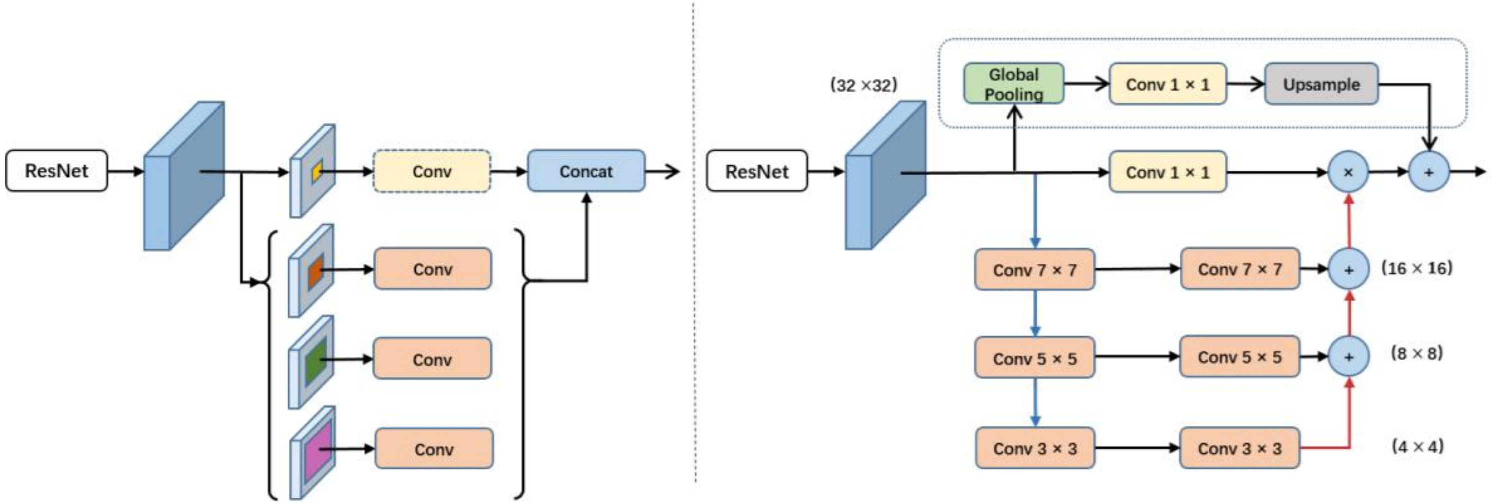
- Convolutional Block Attention Module





S Woo et al. (2018). CBAM: Convolutional Block Attention Module. ECCV 2018.

Attention in Visual: Pyramid Pooling in Segmentation



(a) Spatial Pyramid Pooling

(b) Feature Pyramid Attention

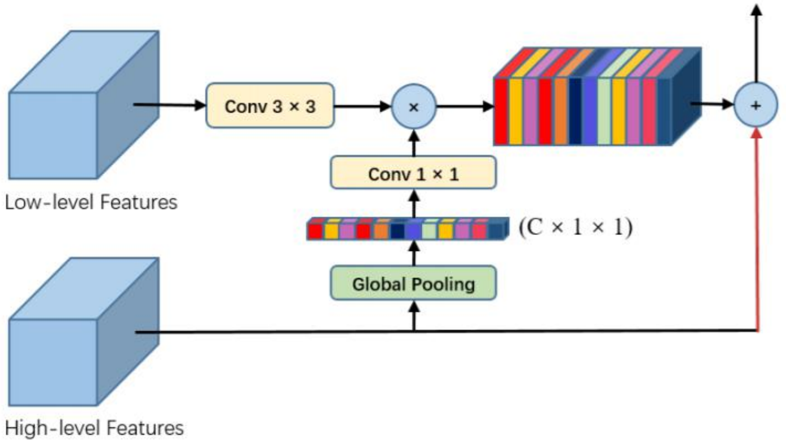
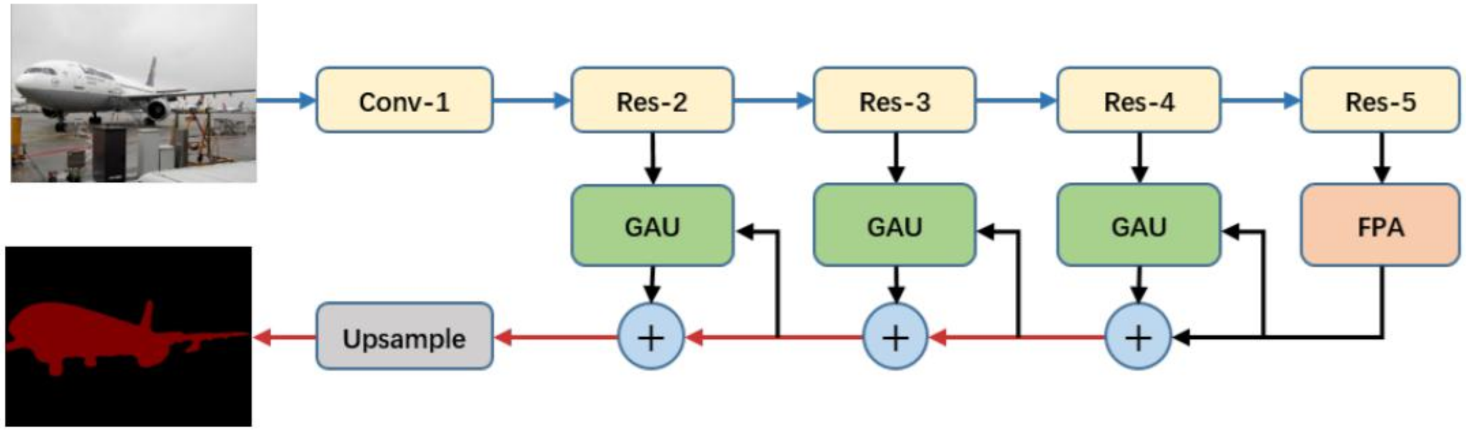


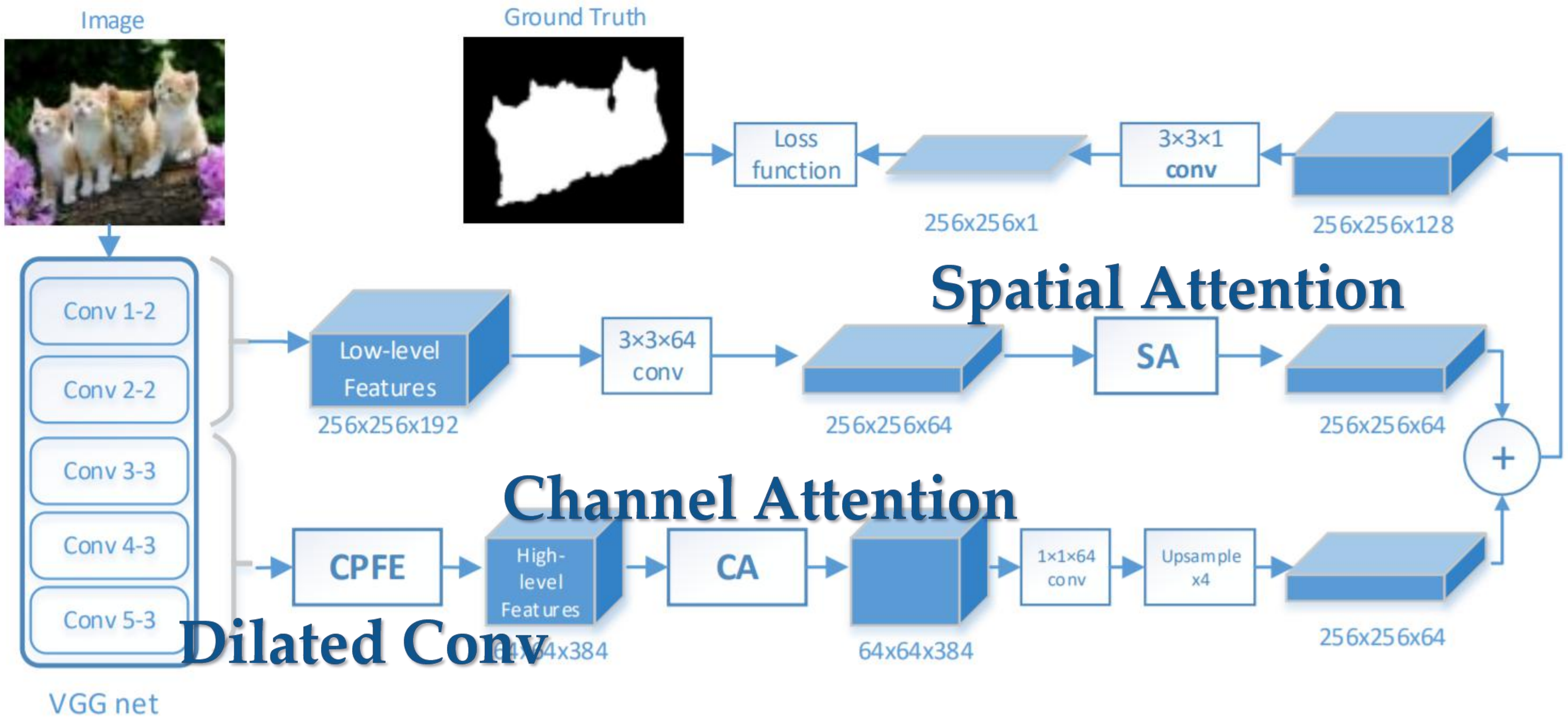
Figure 4: Global Attention Upsample module structure



Li, Hanchao et al. (2018). Pyramid Attention Network for Semantic Segmentation. BMVC, 2018.

Attention in Visual: Pyramid Feature

- Pyramid Feature Attention Network

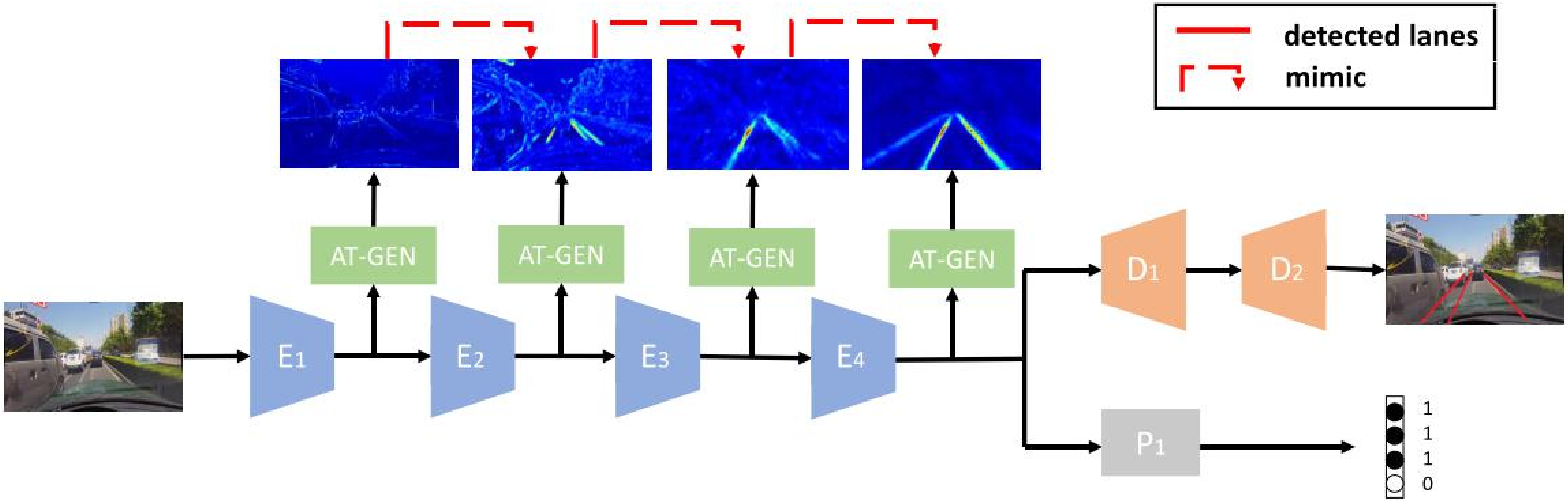


T. Zhao and X. Wu, "Pyramid Feature Attention Network for Saliency Detection," 2019 CVPR.

Attention in Lane Line Detection

Self Attention Distillation

- $\underbrace{\gamma \mathcal{L}_{\text{distill}}(A_m, A_{m+1})}_{\text{distillation loss}}$, without softmax+weighted sum up.



Y. Hou et al. 2019. Learning Lightweight Lane Detection CNNs by Self Attention Distillation.

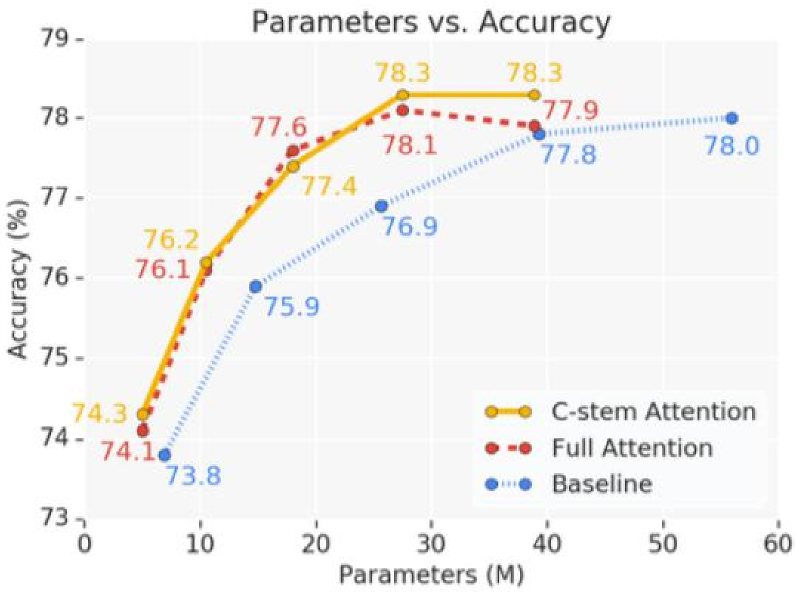
Attention is Really All You Need!

Stand-Alone Self-Attention in Visual Models

□ Recall that key, qeury, value form output y

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

□ Stand-Alone Self-Attention to replace convolution



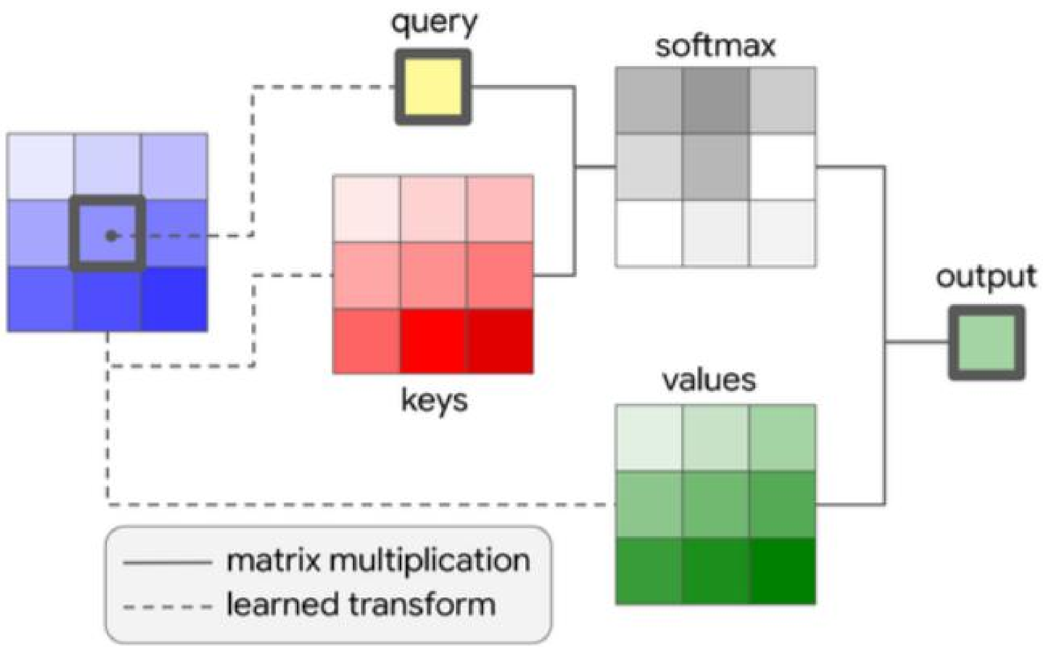


Figure 3: An example of a local attention layer over spatial extent of $k = 3$.

P. Ramachandran et al. *Stand-Alone Self-Attention in Visual Models*, 2019, NIPS

What the hell is attention?

WEIGHTED SUM!