Attention Mechanism

Jun Gao

Peking University

8.11

Outline

- History of Attention
  - End-to-End Attention
  - Attention with Deep Learning
- Attention with Image Caption
- Attention with Computer Vision
  - Image classification
  - Fine-grained classification

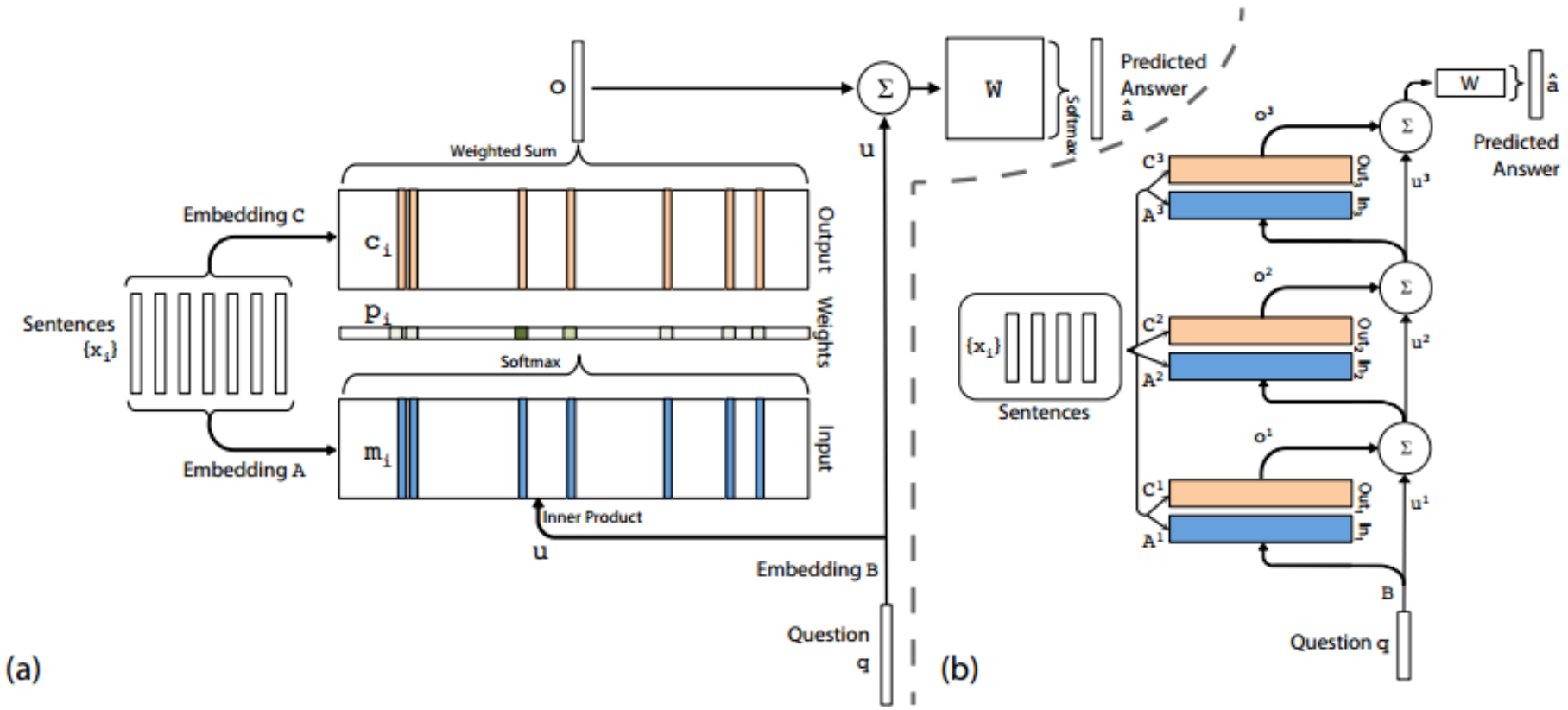
End-To-End Memory Networks

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus
FAIR
In NIPS 2015, cited by 322

Question Answering

Sam walks into the kitchen.	Brian is a lion.	Mary journeyed to the den.
Sam picks up an apple.	Julius is a lion.	Mary went back to the kitchen.
Sam walks into the bedroom.	Julius is white.	John journeyed to the bedroom.
Sam drops the apple.	Bernhard is green.	Mary discarded the milk.
Q: Where is the apple?	Q: What color is Brian?	Q: Where was the milk before the den?
A. Bedroom	A. White	A. Hallway

End-to-End Model



$$p_i = \text{Softmax}(u^T m_i).$$

$$o = \sum_i p_i c_i.$$

$$\hat{a} = \text{Softmax}(W(o+u))$$

Extensions

- Weight tying
- Sentence representation: position encoding

$$x_i = \{x_{i1}, x_{i2}, ..., x_{in}\}$$

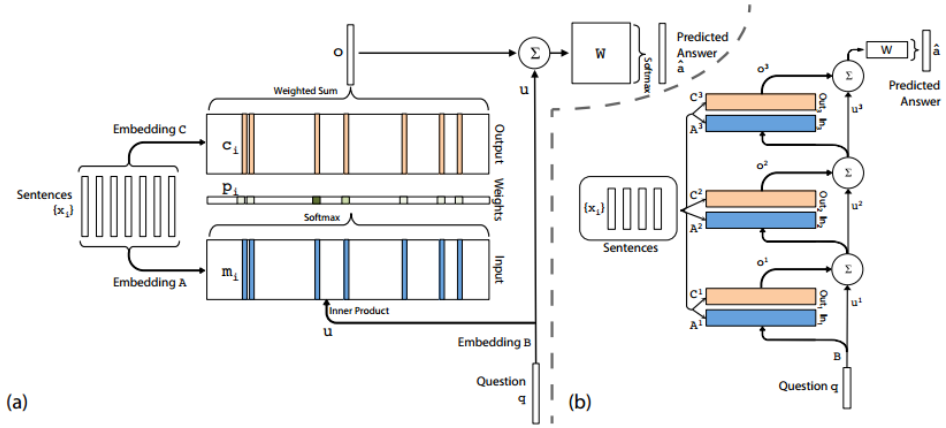
$$m_i = \sum_j l_j \cdot A x_{ij}$$

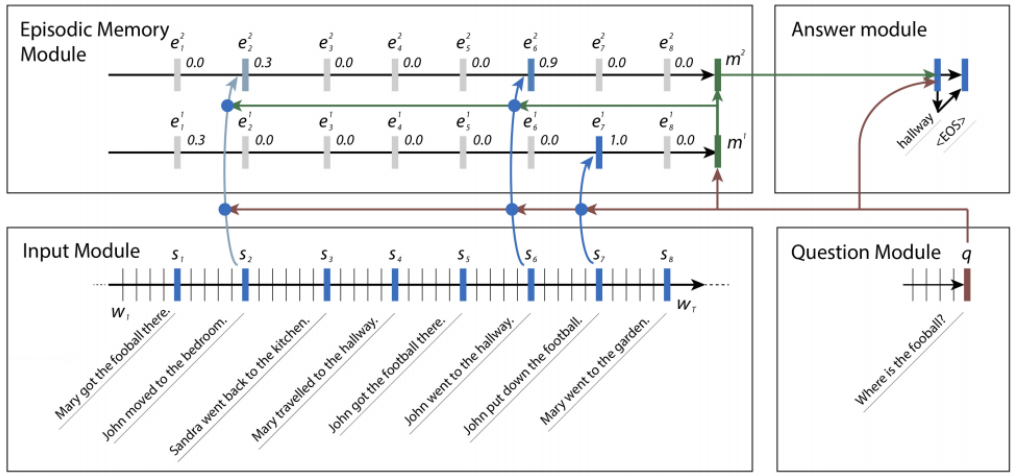
- Linear start

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing

Ankit Kumar, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus, Richard Socher
Metamind
In ICML 2016, cited by 133

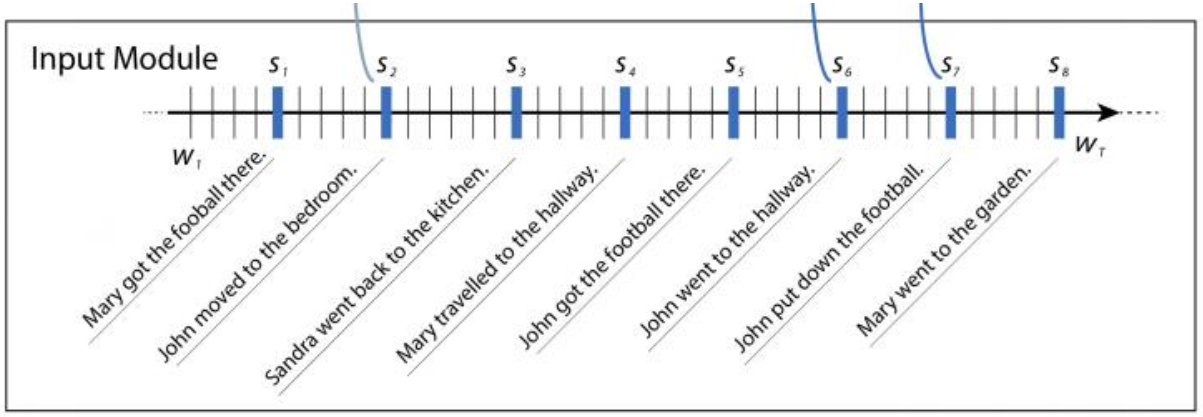
A deep version of end-to-end memory network





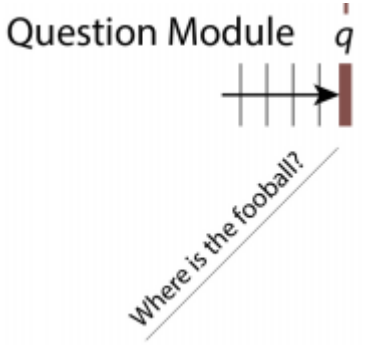
Modules

- Input Module:



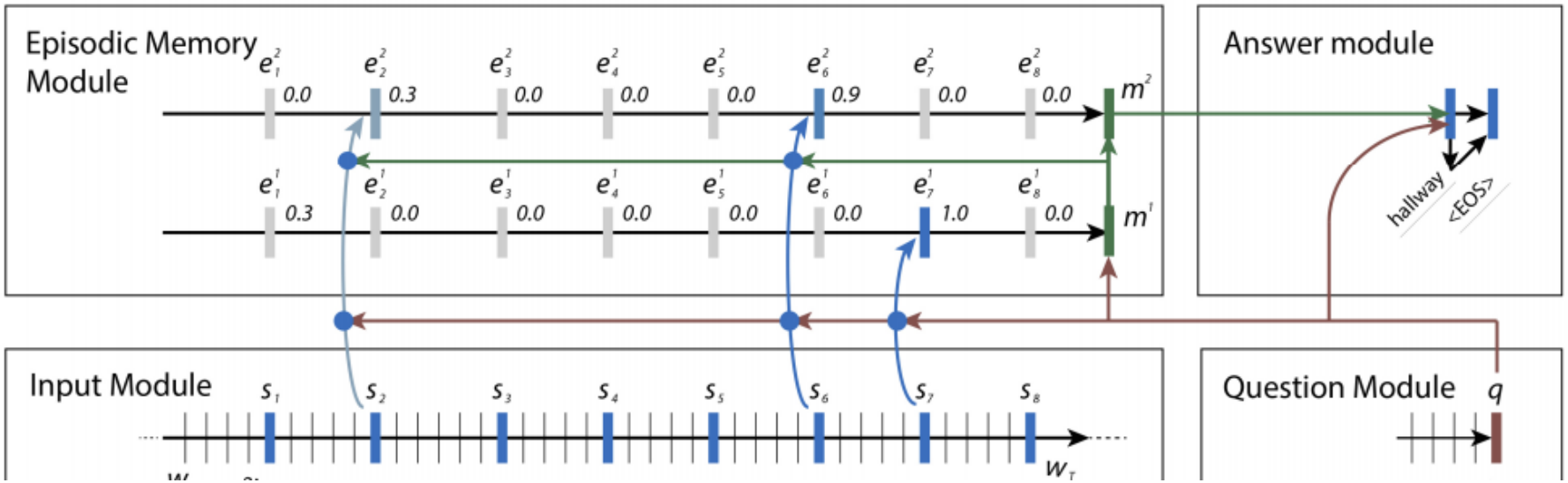
$$h_t = RNN(L[w_t], h_{t-1})$$

- Question Module:



$$q_t = GRU(L[w_t^Q], q_{t-1})$$

Episodic Memory



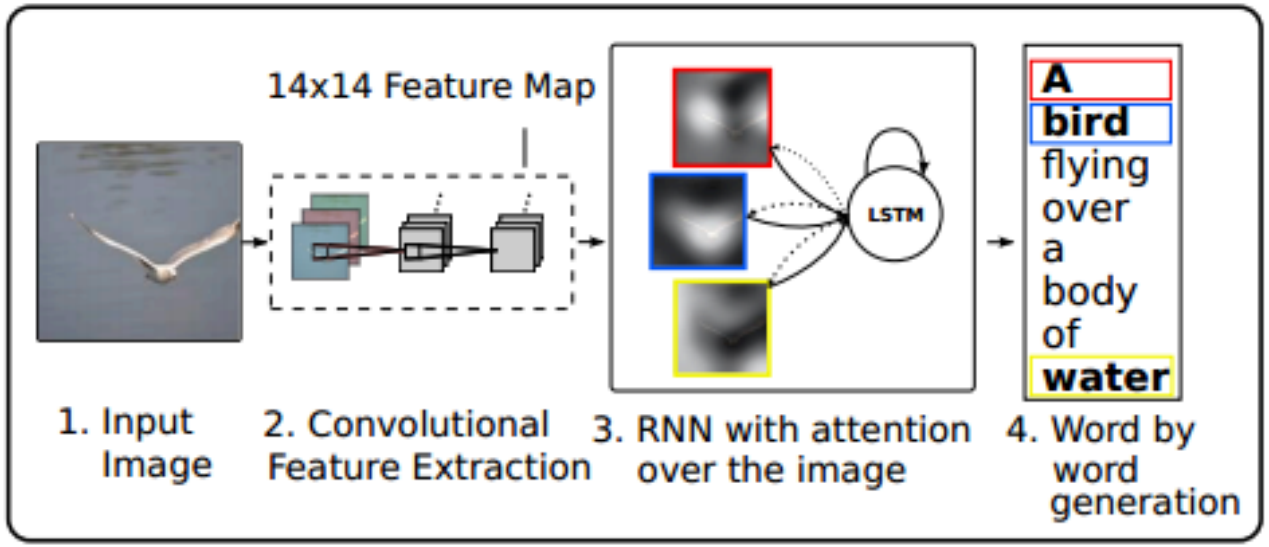
- Two layer feedforward neural network $g_t^i = G(c_t, m^{i-1}, q)$
- Input: $\left[c, m, q, c \circ q, c \circ m, |c - q|, |c - m|, c^T W^{(b)} q, c^T W^{(b)} m\right]$
- Memory update:

$$h_t^i = g_t^i GRU(c_t, h_{t-1}^i) + (1 - g_t^i) h_{t-1}^i$$
$$e^i = h_{T_C}^i$$

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio
UoT, UMontreal
In ICML 2015, cited by 902

Image Caption



Model

- Encoder: CNN

$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \ \mathbf{a}_i \in \mathbb{R}^D$$

- Decoder: LSTM
  - Input:

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$$

  - Output: word probability

Attention Mechanism

- Hard Attention

$$p(s_{t,i} = 1 \mid s_{j<t}, \mathbf{a}) = \alpha_{t,i}$$
$$\hat{\mathbf{z}}_t = \sum_i s_{t,i} \mathbf{a}_i.$$

- Soft Attention

$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i} \mathbf{a}_i$$

- Doubly Stochastic Attention

$$\phi(\{\mathbf{a}_i\}, \{\alpha_i\}) = \beta \sum_i^L \alpha_i \mathbf{a}_i \qquad \sum_t \alpha_{ti} \approx 1$$

Results

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

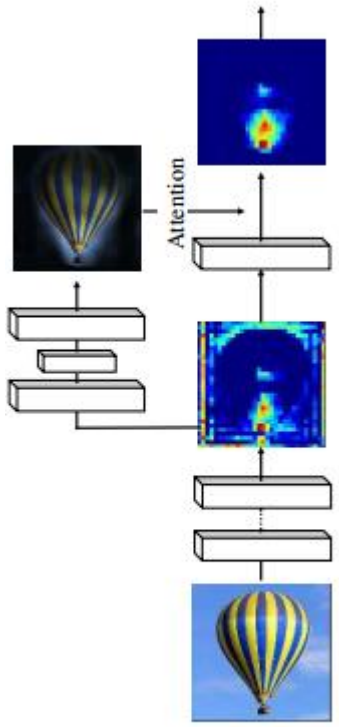
A giraffe standing in a forest with trees in the background.

Residual Attention Network for Image Classification

Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li,Honggang Zhang, Xiaogang Wang, Xiaoou Tang
SenseTime
In CVPR 2017, SpotLight, cited by 1

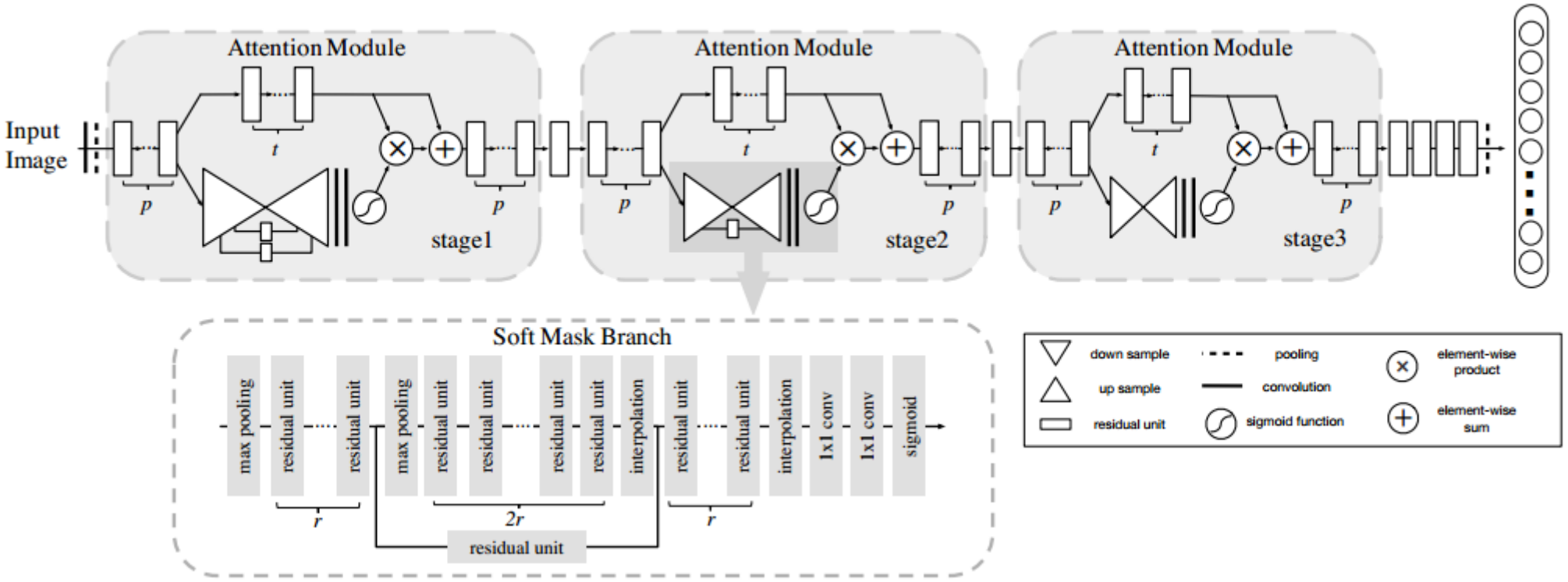
Intuition

- Attention will help for classification
- Direct attention will diminish information



Attention mechanism

Model


Attention Module
Attention Module
Attention Module
Input Image
t
p
stage1
stage2
stage3
Soft Mask Branch
max pooling
residual unit
interpolation
1x1 conv
sigmoid
r
2r
down sample
up sample
residual unit
pooling
convolution
sigmoid function
element-wise product
element-wise sum

Attention

- Mixed attention
- Channel attention
- Spatial attention

$$f_1(x_{i,c}) = \frac{1}{1 + exp(-x_{i,c})}$$

$$f_2(x_{i,c}) = \frac{x_{i,c}}{\|x_i\|}$$

$$f_3(x_{i,c}) = \frac{1}{1 + exp(-(x_{i,c} - \text{mean}_c)/\text{std}_c)}$$

Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition

Jianlong Fu, Heliang Zheng, Tao Mei
MSRA
In CVPR Oral