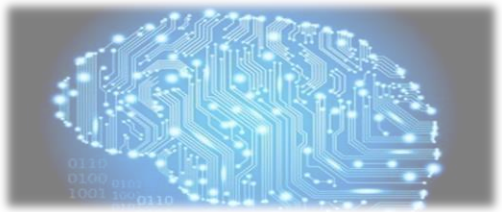
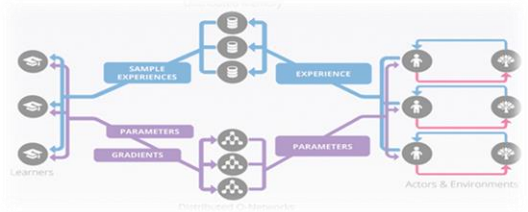



# Deep Modular Co-Attention Networks for Visual Question Answering

AutoML 연구실
석사과정 문아성

21.08.24

# Deep Modular Co-Attention Networks for Visual Question Answering

- Deep Modular Co-Attention Networks for Visual Question Answering

*Yu, Zhou, et al. "Deep modular co-attention networks for visual question answering."*

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.*

*The Winning Entry to the VQA Challenge 2019.*

- Self-Attention & Guided-Attention Units
- Modular Co-Attention (MCA)
- Modular Co-Attention Network

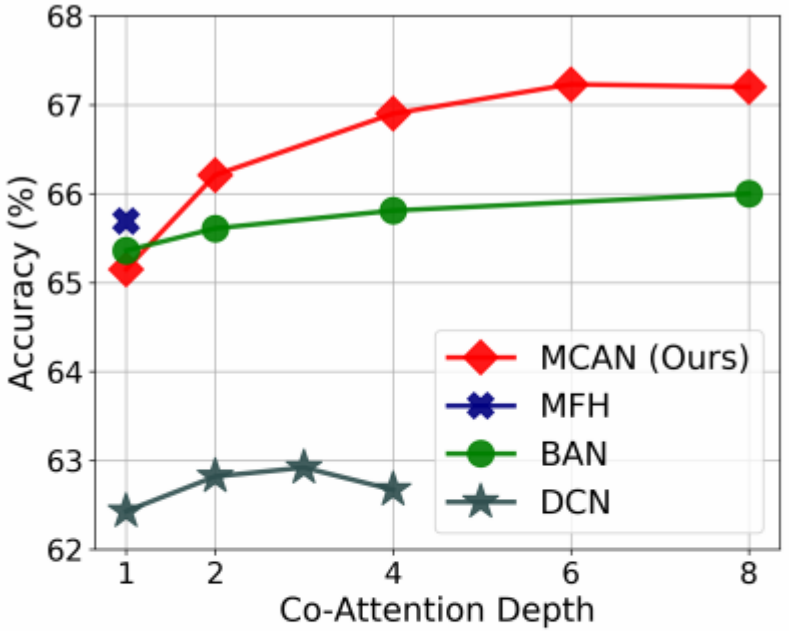



Figure 1: Accuracies *vs.* co-attention depth on VQA-v2 *val* split. We list most of the state-of-the-art approaches with (deep) co-attention models. Except for DCN [24] which uses the convolutional visual features and thus leads to inferior performance, all the compared methods (*i.e.*, MCAN, BAN [14] and MFH [33]) use the same bottom-up attention visual features to represent images [1].

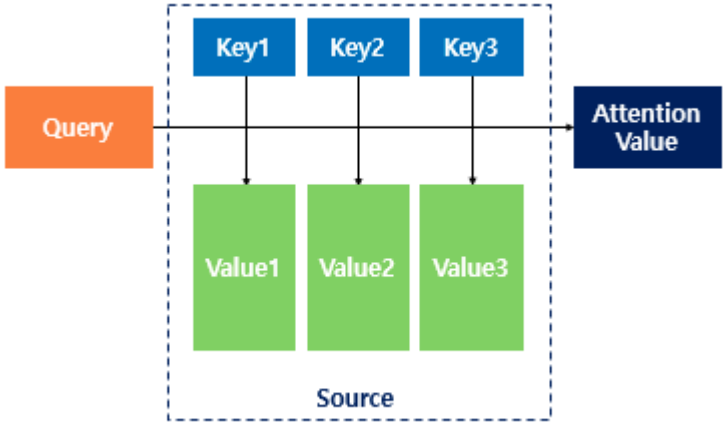
## Self-Attention & Guided-Attention Units

- Multi-head Attention

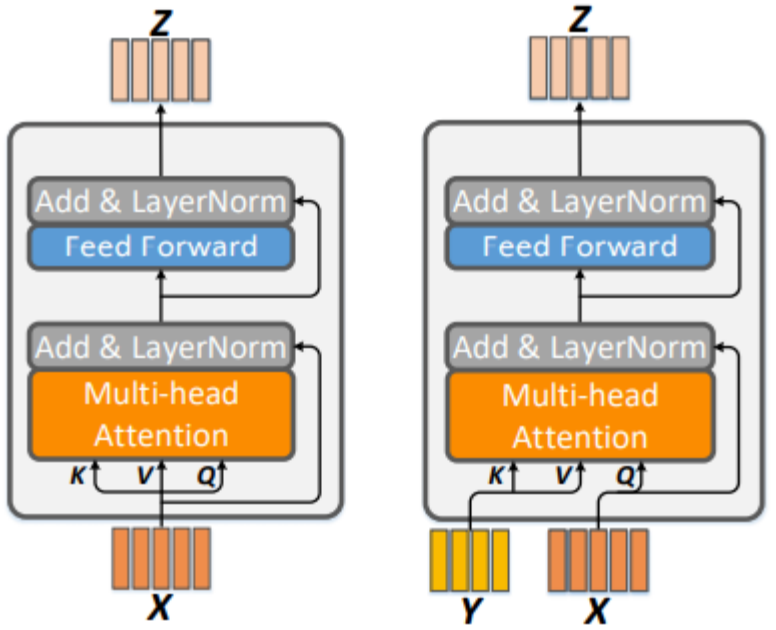
$$f = A(q, K, V) = \text{softmax}(\frac{qK}{\sqrt{d}})V$$

$$f = MA(q, K, V) = [\text{head}_1, \text{head}_2, ..., \text{head}_h]W^o$$

$$\text{head}_j = A(qW_j^Q, KW_j^K, VW_j^V)$$




- q = Query
- K = Key
- V = Value




(a) Self-Attention (SA) (b) Guided-Attention (GA)

Figure 2: Two basic attention units with multi-head attention for different types of inputs. SA takes one group of input features $X$ and output the attended features $Z$ for $X$; GA takes two groups of input features $X$ and $Y$ and output the attended features $Z$ for $X$ guided by $Y$.

## Modular Co-Attention (MCA)

- Identity – Guided-Attention
- Self-Attention – Guided-Attention
- Self-Attention – Self & Guided-Attention

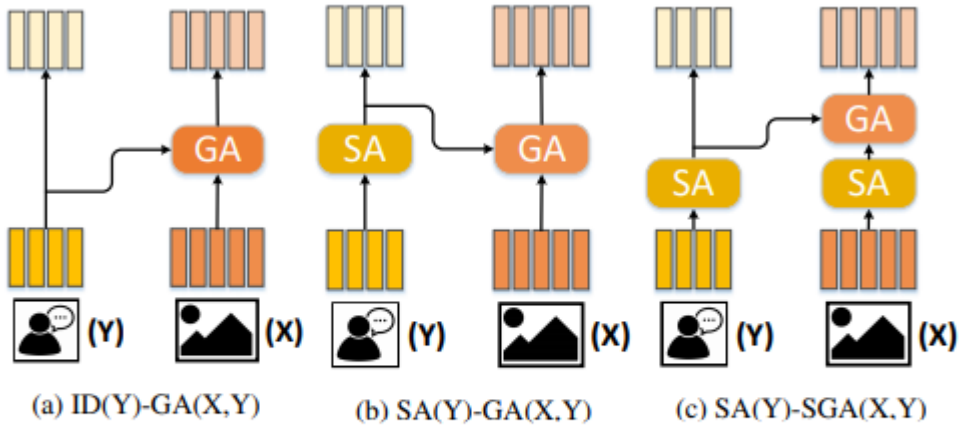



Figure 3: Flowcharts of three MCA variants for VQA. (Y) and (X) denote the question and image features respectively.

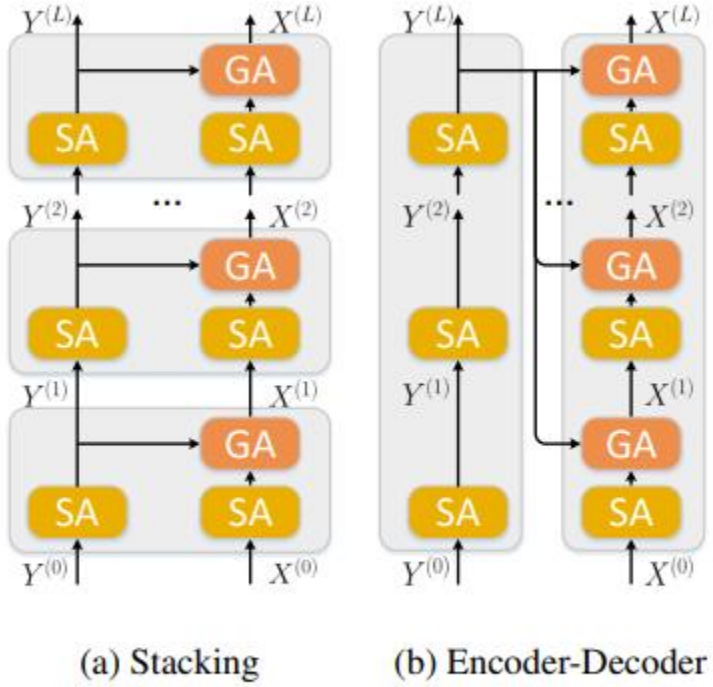



Figure 5: Two deep co-attention models based on a cascade of MCA layers (*e.g.*, SA(Y)-SGA(X,Y)).

## Modular Co-Attention Network

- Multimodal Fusion and Output Classifier

$$\alpha = \text{softmax}(\text{MLP}(X^{(L)}))$$

$$\tilde{x} = \sum_{i=1}^{m} \alpha_i x_i^{(L)} \qquad z = \text{LayerNorm}(W_x^T \tilde{x} + W_y^T \tilde{y})$$

- $MLP: FC(d) - ReLU - Dropout(0.1) - FC(1)$

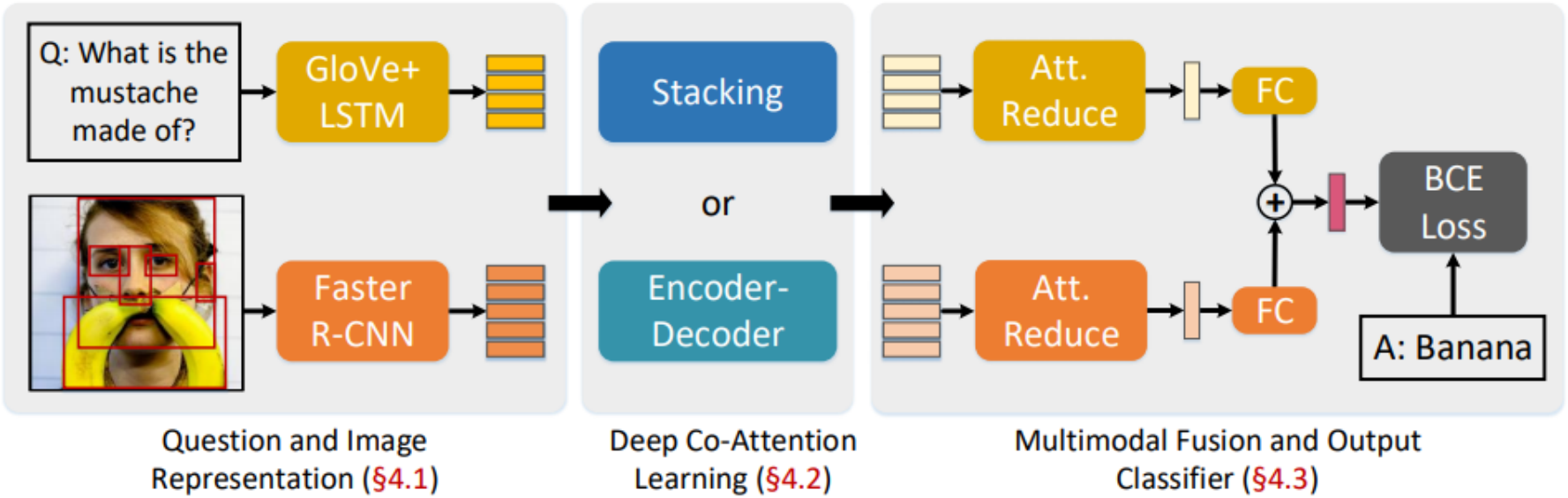



Figure 4: Overall flowchart of the deep Modular Co-Attention Networks (MCAN). In the Deep Co-attention Learning stage, we have two alternative strategies for deep co-attention learning, namely *stacking* and *encoder-decoder*.

## Experiment Settings

- Dataset: VQA-v2

  Yes/No, Number, Other

  - Train: 80k images,  444k Q/A pairs

  - Test: 80k images, 448k Q/A pairs

  - Validation: 40k images, 214 Q/A pairs

- Input Preprocessing

  Image: Faster R-CNN with ResNet-101

  Question:  300-D GloVe word Embedding

## Experiment Results

**Table 2: Accuracies of single-model on the test-dev and test-standard splits to compare with the state-of-the-art methods. All the methods use the same bottom-up attention visual features [1], and are trained on the train+val+vg sets (vg denotes the augmented VQA samples from Visual Genome).**

| *Model* | *Test-dev* | | | | *Test-std* |
|---|---|---|---|---|---|
| | *All* | *Y/N* | *Num* | *Other* | *All* |
| Bottom-Up | 65.32 | 81.82 | 44.21 | 56.05 | 65.67 |
| MFH | 68.76 | 84.27 | 49.56 | 59.89 | - |
| BAN | 69.52 | 85.31 | 50.93 | 60.26 | - |
| BAN+Counter | 70.04 | 85.42 | **54.04** | 60.52 | 70.35 |
| $MCAN_{ed} - 6$ (paper) | **70.63** | **86.82** | 53.26 | **60.72** | **70.90** |
| $MCAN_{ed} - 6$ (our) | 70.30 | 86.91 | 53.42 | 60.24 | 70.61 |

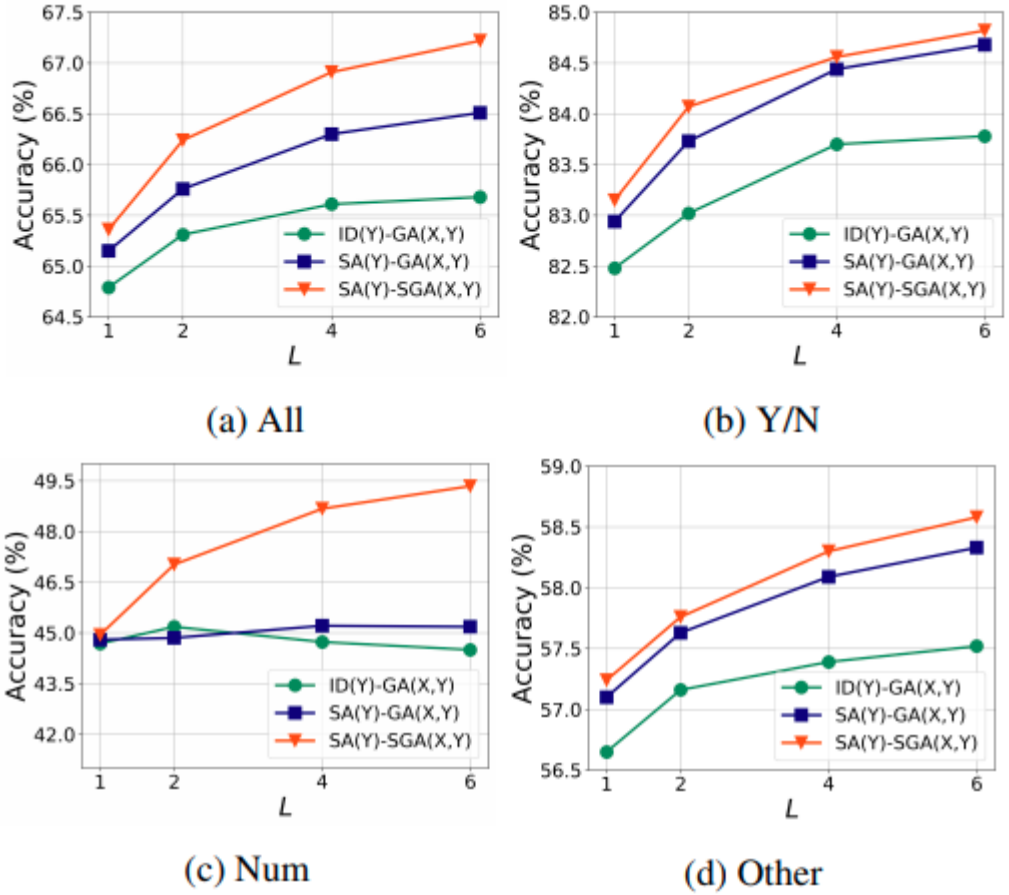



**Figure 6: The overall and per-type accuracies of the $MCAN_{ed} - L$ models equipped with different MCA variants, where the number of layers $L \in \{1, 2, 4, 6\}$. All the reported results are evaluated on the val split.**