ICASSP2020

MULTI-LAYER CONTENT INTERACTION THROUGH QUATERNION PRODUCT FOR VISUAL QUESTION ANSWERING

Lei Shi[†], *ShijieGeng*[□], *KaiShuang*[†], *ChioriHori*[*], *SongxiangLiu*[○], *Peng Gao*[○], *SenSu*[†]

†*Beijing University of Posts and Telecommunications*

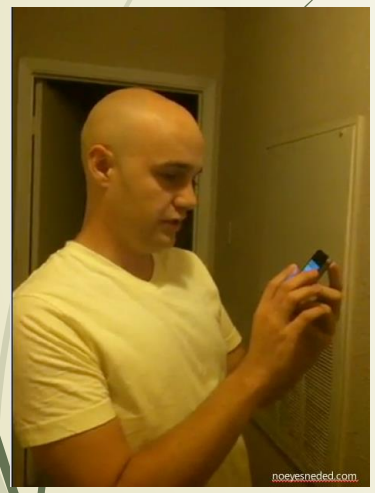
□*Rutgers University*

○*CUHK*

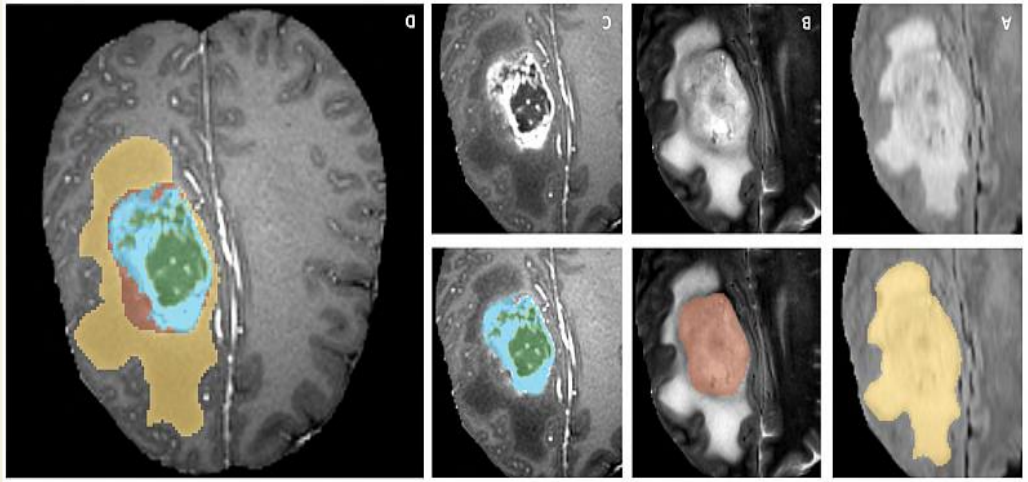
∗*Mitsubishi Electric Research Laboratories* (*MERL*)

MULTI-LAYER CONTENT INTERACTION THROUGH QUATERNION PRODUCT FOR VISUAL QUESTION ANSWERING

- Why do people need a visual question answering system?
  - Help blind people live a better life.
  - Smart medical
  - Satellite image analysis



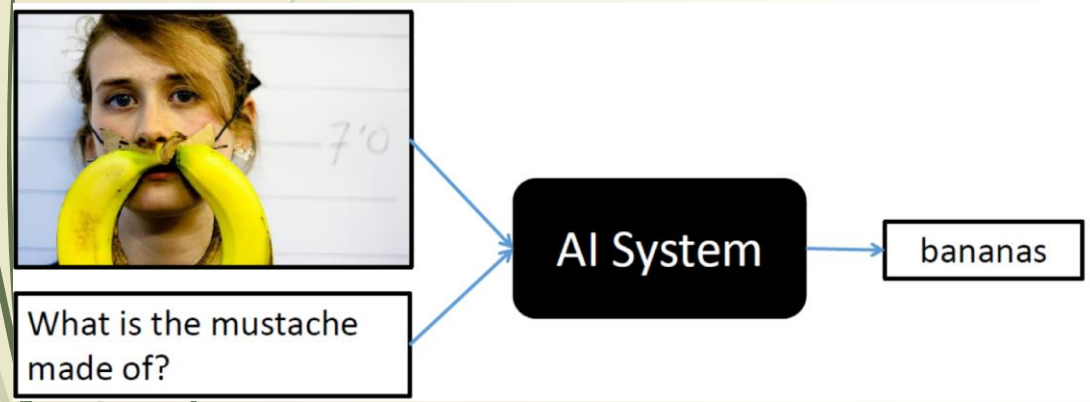
What is the temperature now?

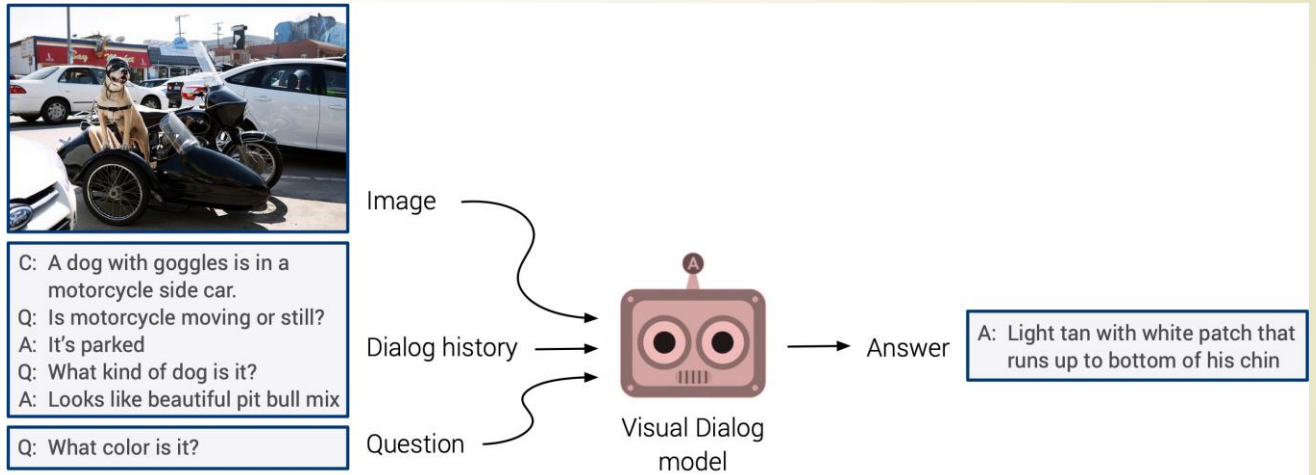


Is the tumor malignant?

Is there an airport in this picture?

Single-round and Multi-round visual question and answer



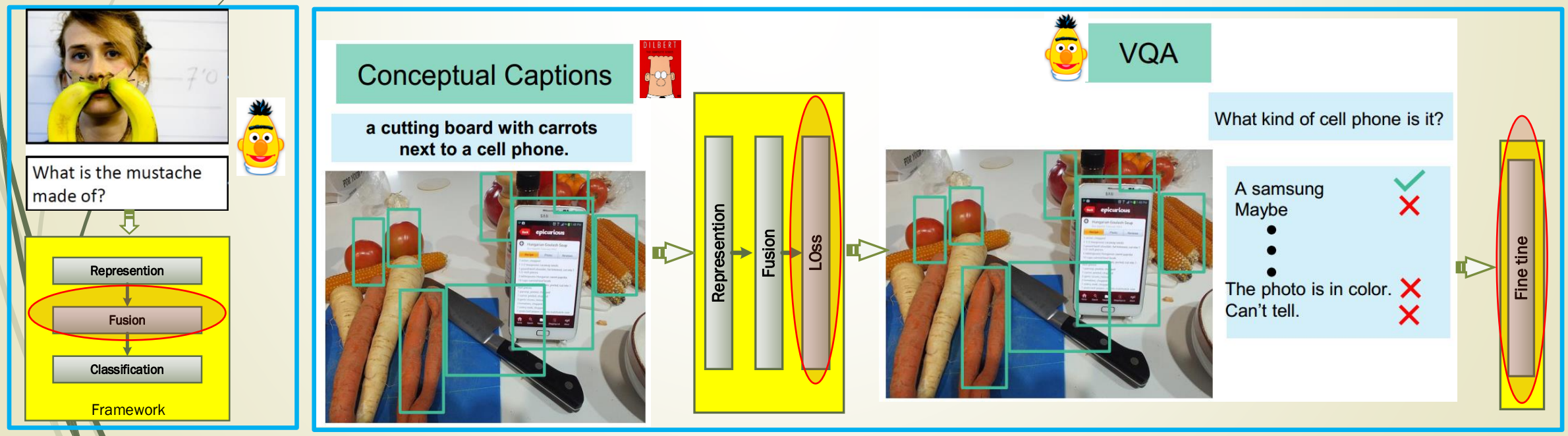


- Single-round visual question answer： VQA
- Input： One image, one question
- Output:The answer to the question

- Multi-round visual question answer： Visual Dialog
- Input： one image， Dialogue history， one question
- Output： The answer to the question

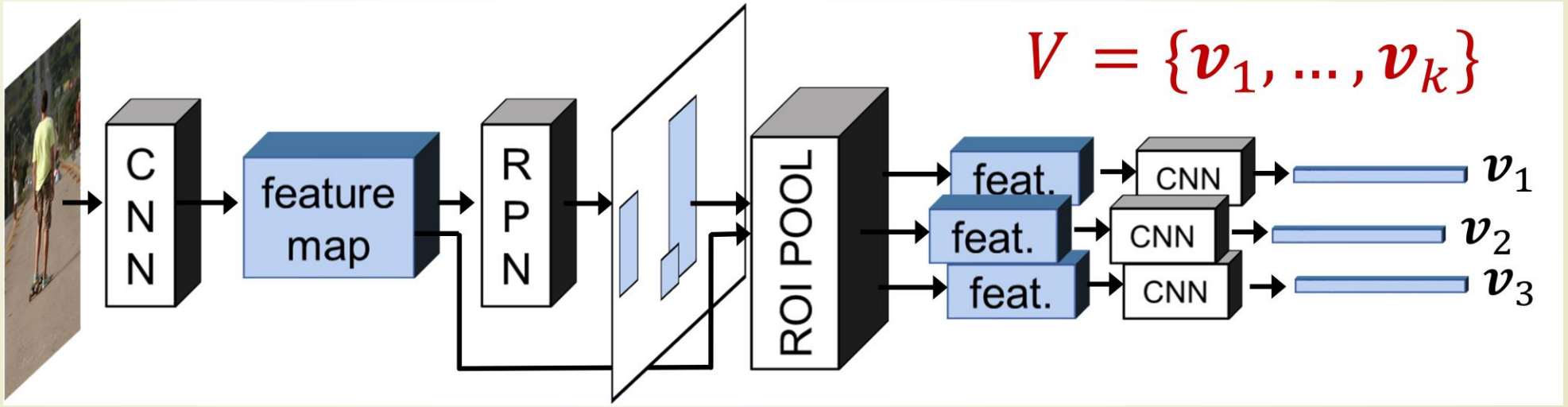
Visual question answering under no-pre-training and pre-training conditions

- no-pre-training：
- Directly answer the question through the visual question and answer task data set, establish the connection between the image and the question feature Directly answer the question through the visual question and answer task data set, establish the connection between the image and the question feature
- pre-training conditions：
- First, through proxy tasks, learn the more generalized representation of images and text features, and then train the visual question and answer task to answer questions.



MULTI-LAYER CONTENT INTERACTION THROUGH QUATERNION PRODUCT FOR VISUAL QUESTION ANSWERING

- motivation
  - Layer-based information interaction cannot fully integrate multimodal information, because the abstraction of image and text features is not equivalent;
  - The existing methods are not conducive to exploring cross-layer information interaction between shallow and deep Impact on model performance.



MULTI-LAYER CONTENT INTERACTION THROUGH QUATERNION PRODUCT FOR VISUAL QUESTION ANSWERING

- Related work
  - Hierarchical Question-Image Co-Attention for Visual Question Answering, NIPS 2017
  - Bilinear Attention Networks，ECCV 2018

Fusion Match

  - Relation Networks for Object Detection CVPR 2018

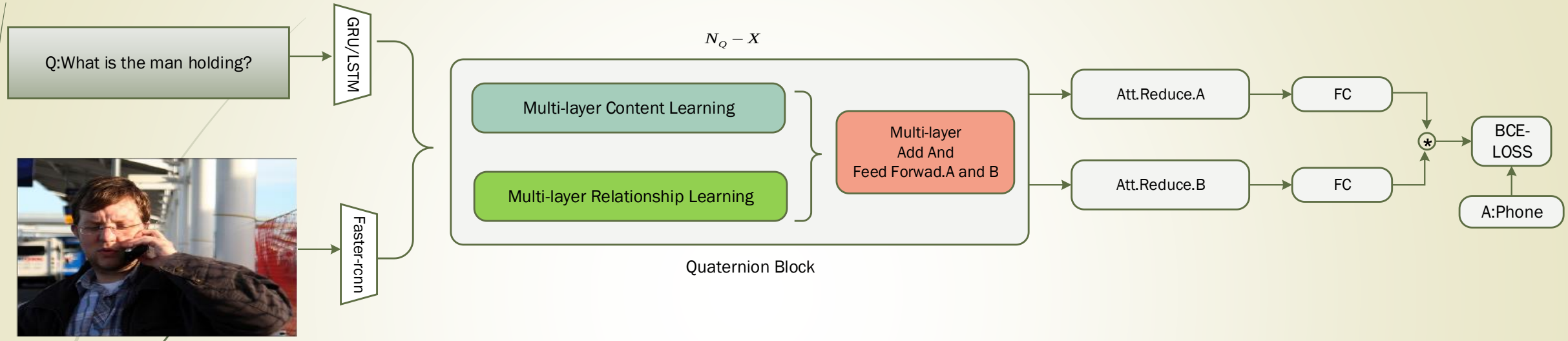
Representation

  - Dynamic Fusion with Intra- and Inter-modality Attention Flow for Visual Question Answering，cvpr2019
  - Deep Modular Co-Attention Networks for Visual Question Answering，cvpr2019
- Challenge

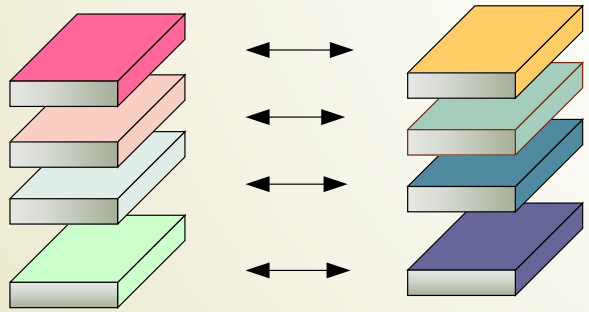
Represent and fusion

  - How to reduce the matching deviation of image-text features due to different levels of abstraction
  - How to realize the cross-layer flow of information
  - How to preprocess image features to remove redundant information that is not relevant to the task

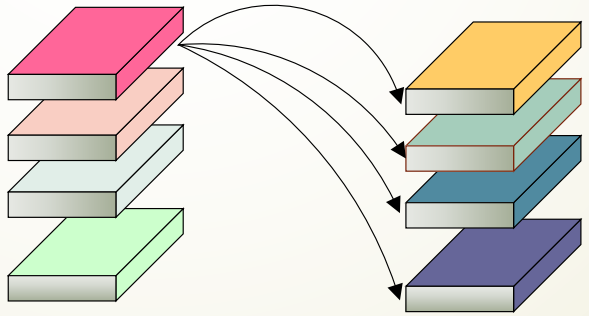
Overall framework



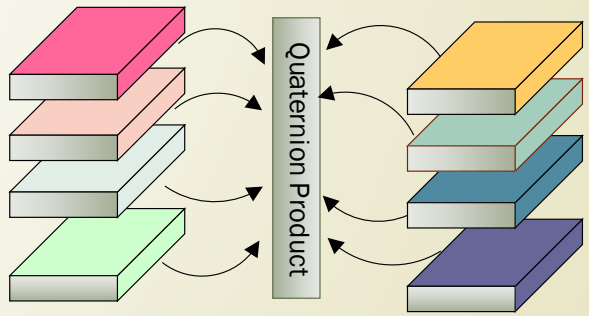
Comparison with existing models



Val:66.6 DFAF interaction



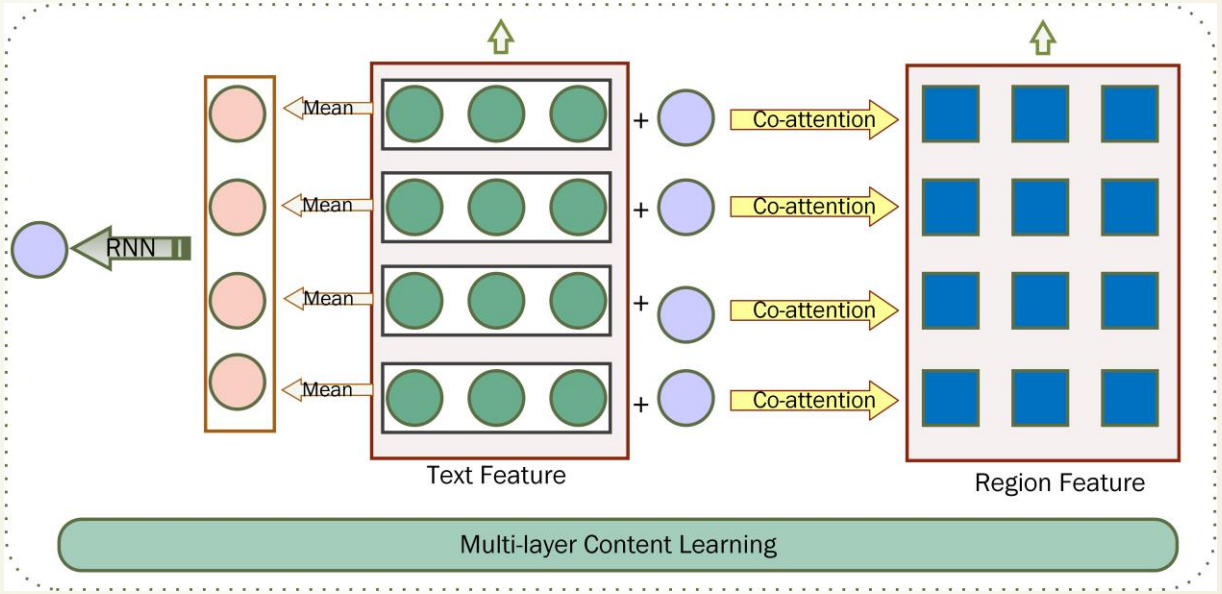
Val:67.1 MCAN Interaction



Val:67.45 Quaternion Product Interaction

Multi-layer content learning

- In the visual question answering task, because the text layer and the visual layer of the same layer have different levels of abstraction, the single layer of information interaction will cause matching deviation.



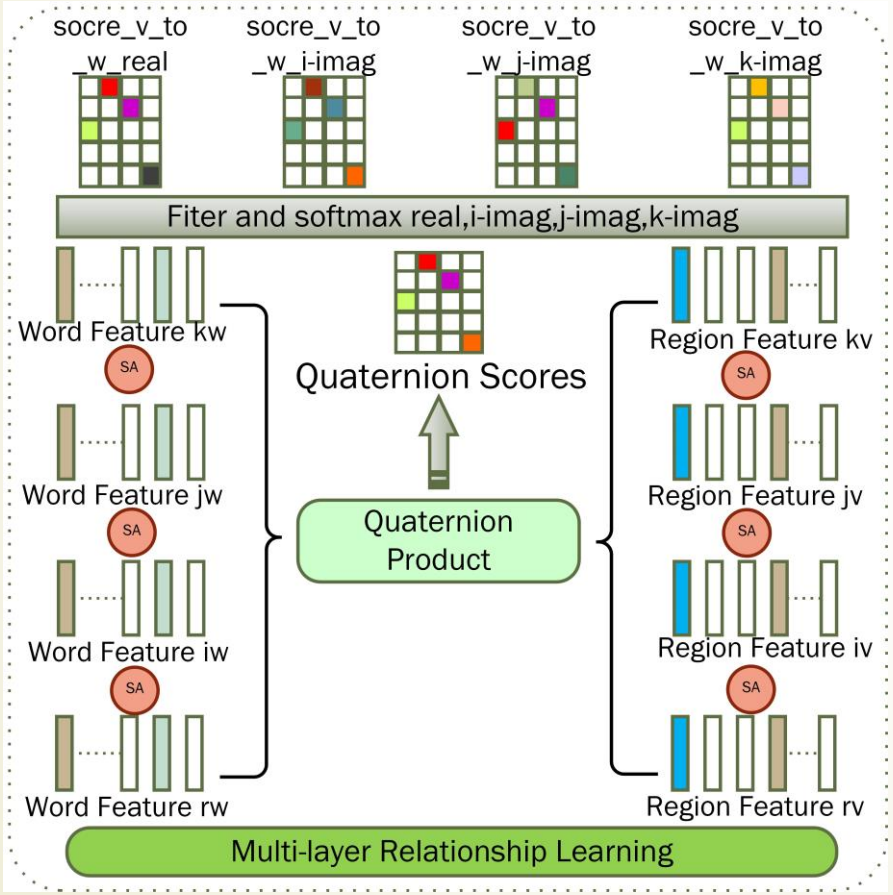
$$rv = v \;; iv = \mathrm{Sa}(rv) \;; jv = \mathrm{Sa}(iv) \;; kv = \mathrm{Sa}(jv)$$
$$rw = w \;; iw = Sa(rw) \;; jw = Sa(iw) \;; kw = Sa(jw)$$

$$xw_{\mathrm{mean}} = \mathrm{Mean}(xw) \quad x \in \{r, i, j, k\}$$

$$\mathrm{Multi}_{\mathrm{context}}, \mathrm{Multi}_{q_n} = \mathrm{RNN}\left(rw_{\mathrm{mean}}, iw_{\mathrm{mean}}, jw_{\mathrm{mean}}, kw_{\mathrm{mean}}\right)$$

Multi-layer relationship

The single-layer information interaction method can only connect the visual and text features of the same layer. In order to establish a durable relationship between multi-modal information between multiple layers, We use the inner product of quaternary complex numbers to model multi-layer relationships

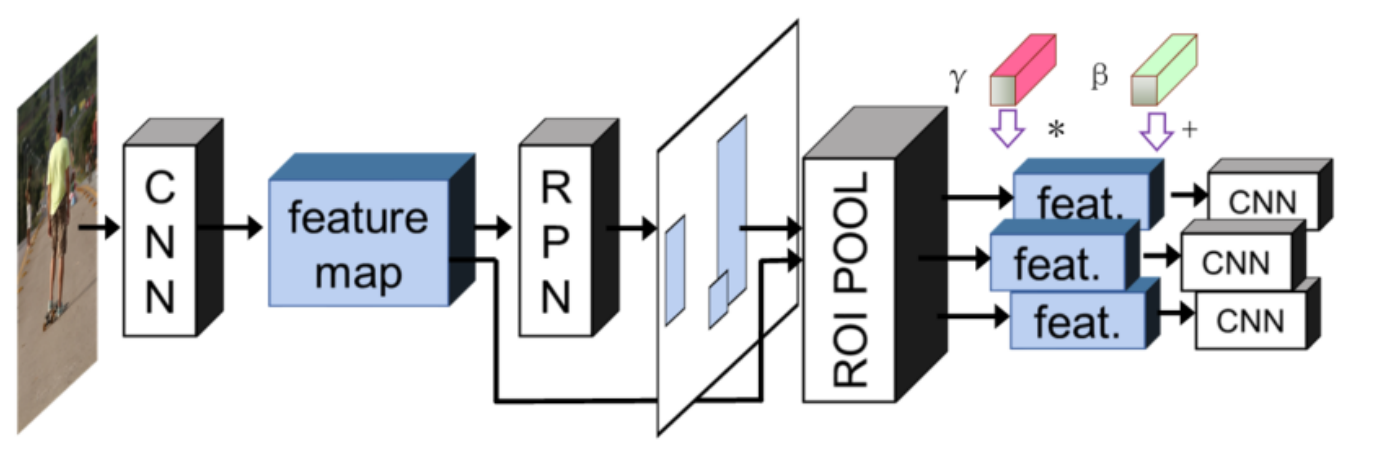


$$W \otimes V = \left(rv*rw - iv*iw - jv* jw - kv*kw\right) + \left(iv*rw + rv*iw - kv* jw + jv*kw\right)\vec{I} + \left(jv*rw + kv*iw + rv* jw - iv*kw\right)\vec{J} + \left(kv*rw - jv*iw + iv* jw + rv*kw\right)\vec{K}$$

$$Attention(Q,K,V) = Soft\max\left(\frac{Q*K}{\sqrt{\dim}}\right)*G_{quaternion-score-soft\max}*V$$

Preprocessing of image features

- For the same picture, the problem changes dynamically, but the initial characteristics of the image are fixed. Need to dynamically preprocess the visual signs, remove redundant information, and retain information related to the problem.



$$\gamma_{i,c} = f_c(q_t)$$

$$\beta_{i,c} = h_c(q_t)$$

$$v = Averagepooling\left(\gamma_{i,c} * R_{i,c} + \beta_{i,c}\right)$$

Visualization of the updating process of attention weights learned by our Quaternion Block Networks. If the weight of an object ishigher, the color of that bounding box will be more “gray”. From the two examples, we can see that after three iterations, our model attendsto the right objects and generate the right answers

Thankyou!!!