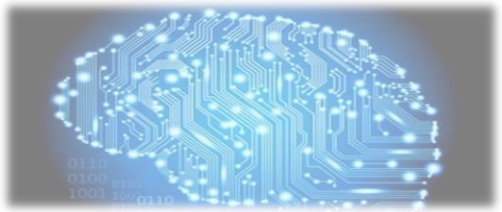
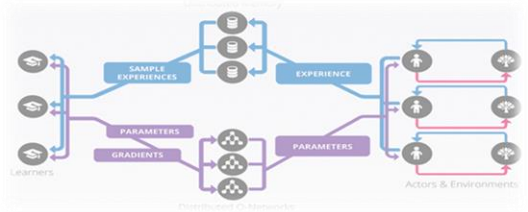



Visual Question Answering Models

AutoML 연구실
석사과정 문아성
21.09.07

Visual Question Answering Models

- 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.

- In Defense of Grid Features for Visual Question Answering

Jiang, Huaizu, et al. "In defense of grid features for visual question answering."

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

The Winning Entry to the VQA Challenge 2020., Facebook AI Research

- Deep Multimodal Neural Architecture Search

Yu, Zhou, et al. "Deep Multimodal Neural Architecture Search."

Proceedings of the 28th ACM International Conference on Multimedia. 2020.

Deep Modular Co-Attention Networks for Visual Question Answering

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

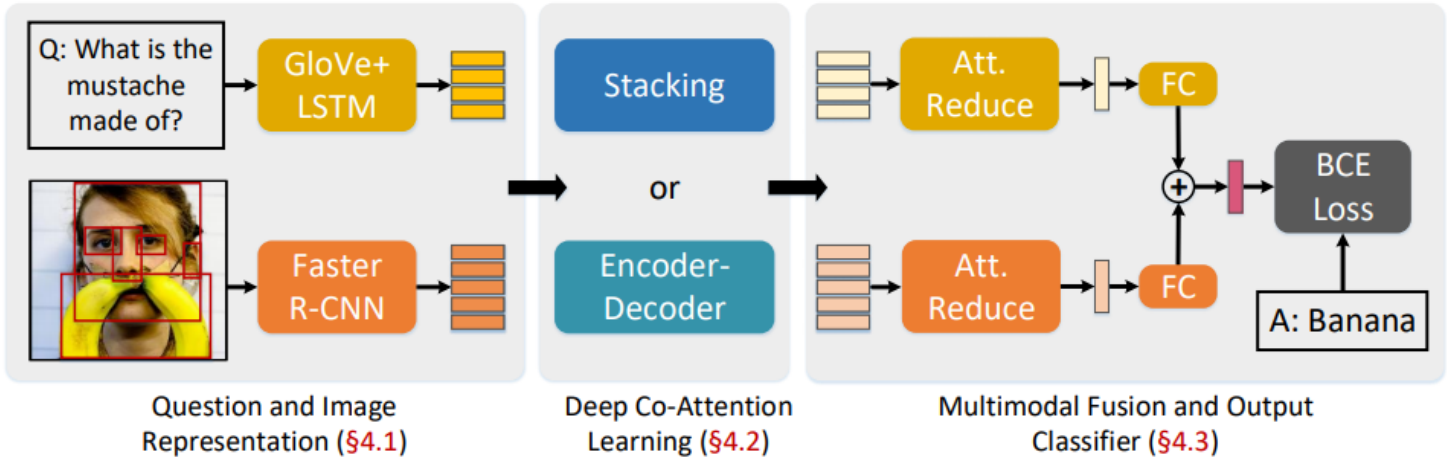


Figure 4: Overall flowchart of the deep Modular Co-Attention Networks (MCAN). In the Deep Co-attention Learning stage, we have two alternative strategies for deep co-attention learning, namely *stacking* and *encoder-decoder*.

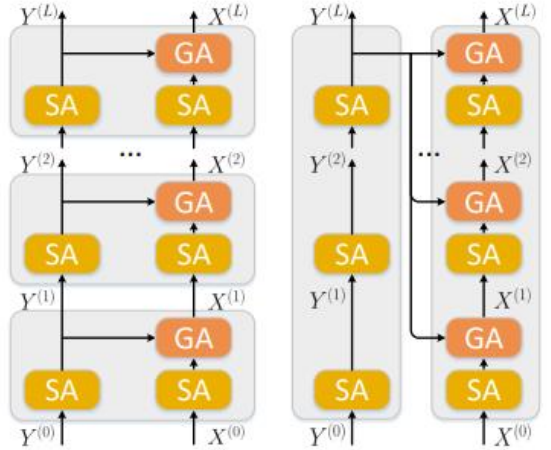


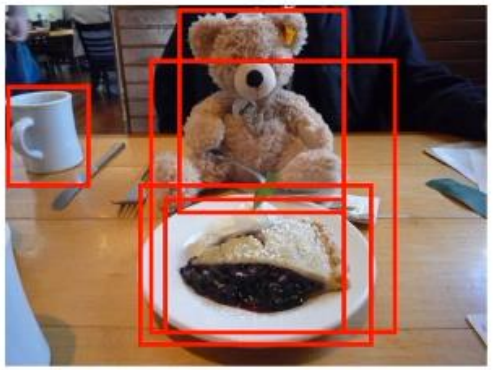
Figure 5: Two deep co-attention models based on a cascade of MCA layers (*e.g.*, SA(Y)-SGA(X,Y)).

In Defense of Grid Features for Visual Question Answering

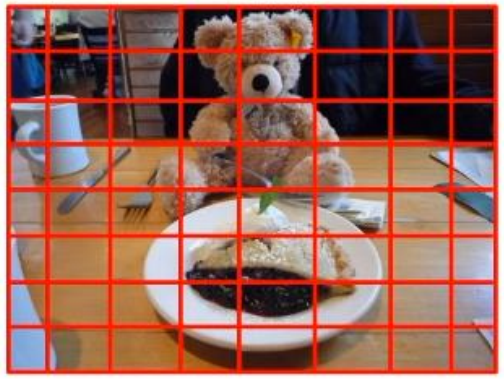
- Bottom-Up Attention with Regions

  MCAN: Region features ➡ Grid features

- Running Time: improvement compared to Regions
  (0.89s -> 0.02s)



Region features [Anderson et al]



Grid feature (ours)

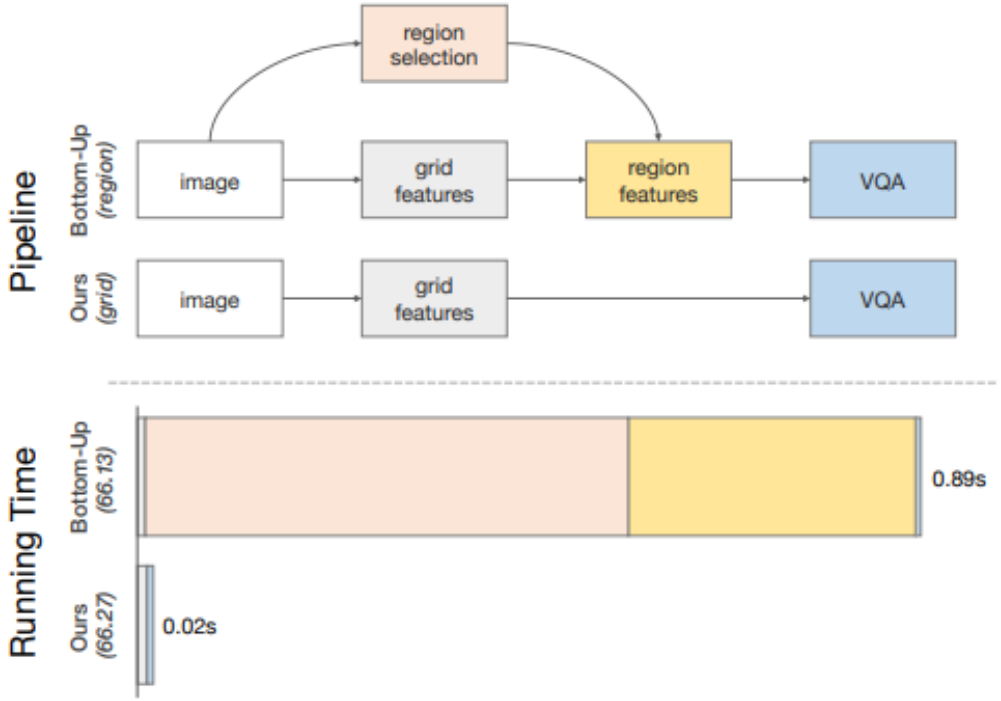
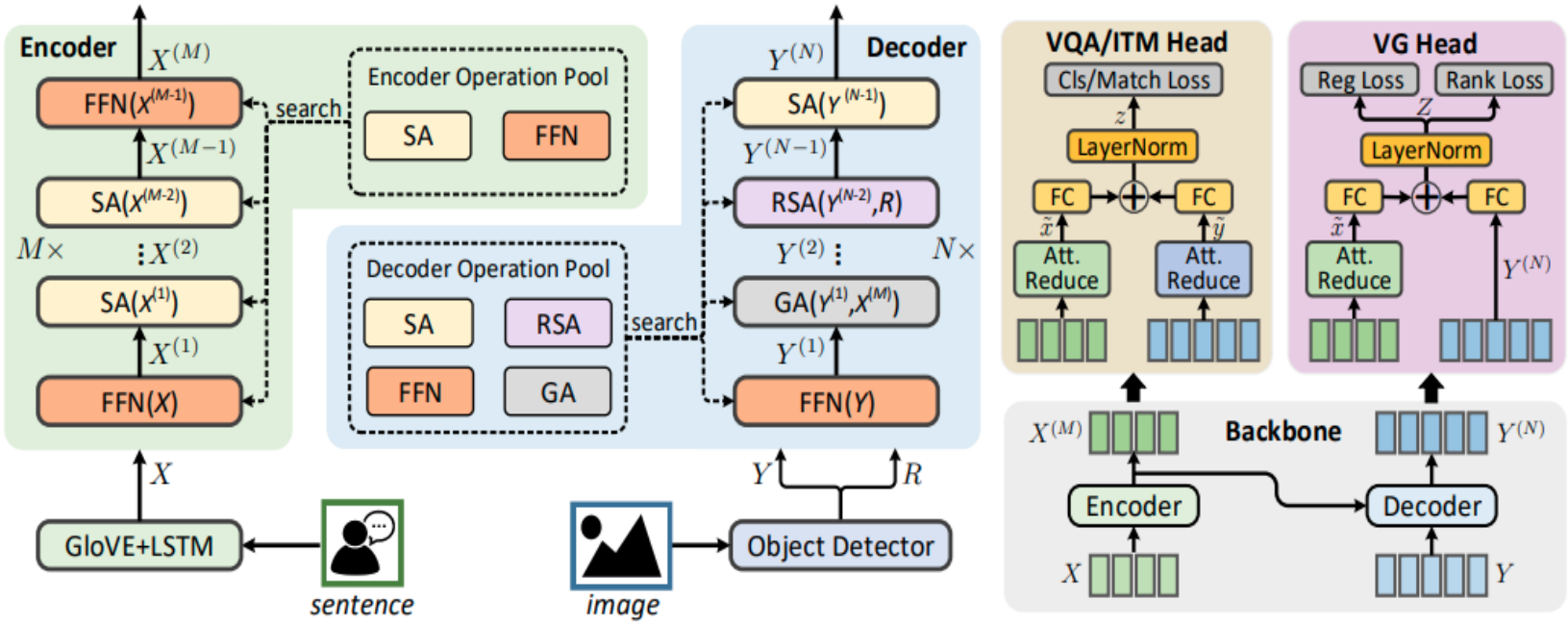


Figure 1: We revisit **grid**-based convolutional features for VQA, and find they can *match* the accuracy of the dominant **region**-based features from bottom-up attention [2], provided that one closely follow the pre-training process on Visual Genome [21]. As computing grid features skips the expensive region-related steps (shown in colors), it leads to significant speed-ups (all modules run on GPU; timed in the same environment).

Deep Multimodal Neural Architecture Search

- Unified Encoder-Decoder Backbone
  - *Self-attention operation*
  - *Guided-attention operation*
  - *Feed-forward network operation*
  - *Relation self-attention operation*



(a) Unified Encoder-Decoder Backbone

(b) Task-specific Heads

Figure 2: The flowchart of the MMnas framework, which consists of (a) unified encoder-decoder backbone and (b) task-specific heads on top the backbone for visual question answer (VQA), image-text matching (ITM), and visual grounding (VG).

Visual Question Answering Models

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

  Text(Question): 300-D GloVe word Embedding

Experiment Results

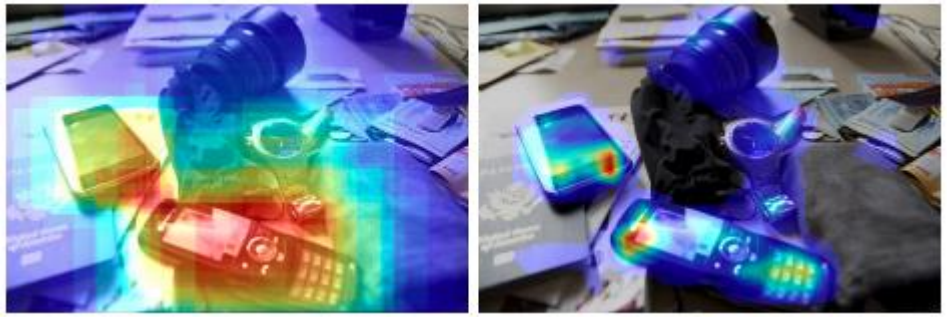
Method	All	Yes/No	Num	Other
MCAN	**70.30**	**86.91**	**53.42**	**60.24**
GridFeature	**71.20**	-	-	-
MMnasNet	**67.53**	**84.93**	**51.74**	**58.47**

Attention Map *(MCAN vs GridFeat)*

Q: Which devices do you see?

GT-A: Phones

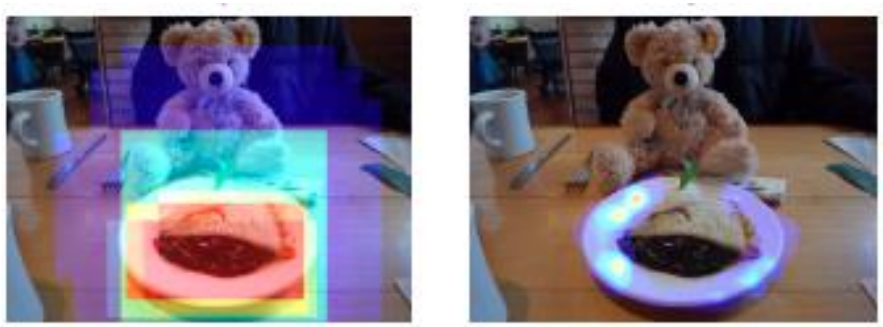
A(MCAN): *Phones* ***A(GridFeat)***: *Phones*



Q: Is the plate white?

GT-A: Yes

A(MCAN): *Yes* ***A(GridFeat)***: *Yes*



Q: Questions *GT-A: ground-truth answers*

Attention Map *(MCAN vs GridFeat)*

Q: Has the pizza been eaten?

GT-A: No

A(MCAN): *No* ***A(GridFeat)***: *Yes*

Q: What breed of dog is this?

GT-A: pug

A(MCAN): *pug* ***A(GridFeat)***: *bulldog*

Q: What color are the curtains?

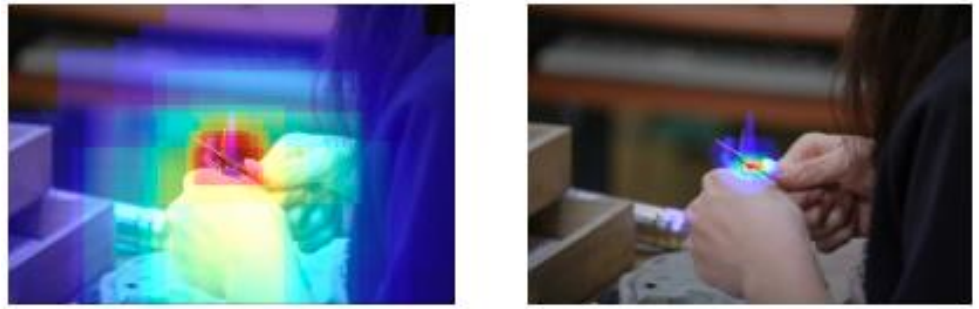
GT-A: red and white

A(MCAN): *red* ***A(GridFeat)***: *red and white*

Q: What is the person doing?

GT-A: cutting

A(MCAN): *texting* ***A(GridFeat)***: *cutting*

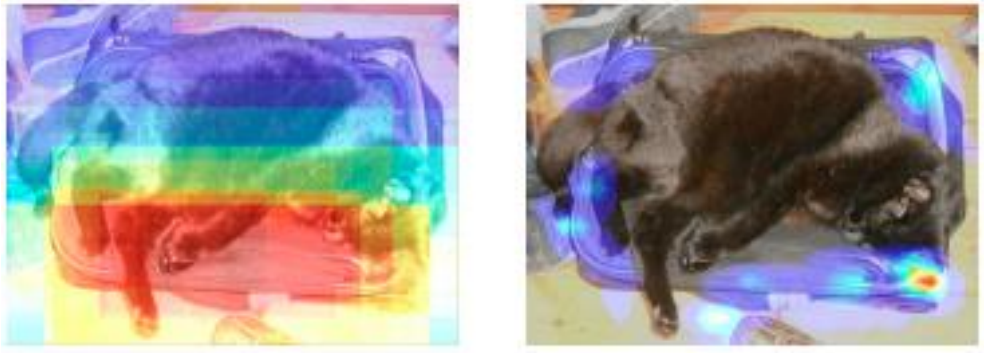


Attention Map *(MCAN vs GridFeat)*

Q: What is the cat laying on?

GT-A: suitcase

A(MCAN): *shoes* ***A(GridFeat)***: *shoe*



Q: How many boats do you see?

GT-A: 7

A(MCAN): *5* ***A(GridFeat)***: *4*

Q: Questions *GT-A: ground-truth answers*