Aligning Visual Regions and Textual Concepts for Semantic-Grounded Image Representations

Fenglin Liu[1]*, Yuanxin Liu[3,4]*, Xuancheng Ren[2]*, Xiaodong He[5], Xu Sun[2]

*Equal contribution

[1] ADSPLAB, School of ECE, Peking University, [2] MOE Key Laboratory of Computational Linguistics, School of EECS, Peking University, [3] Institute of Information Engineering, Chinese Academy of Sciences, [4] School of Cyber Security, University of Chinese Academy of Sciences, [5] JD AI Research

北京大學 PEKING UNIVERSITY

Introduction

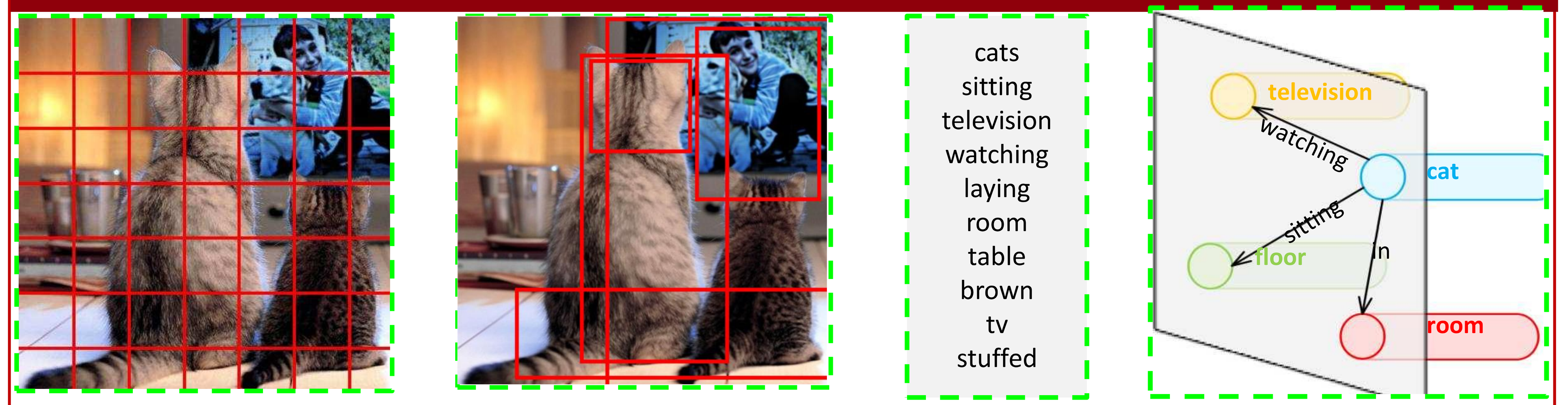

Figure 1. Illustrations of commonly-used image representations (from left to right): CNN-based grid visual features, RCNN-based region visual features, textual concepts, and scene-graphs.

An image in vision-and-language tasks, e.g., image captioning and visual question answering, is typically represented in two fundamental forms: visual features and textual concepts (see Figure 1).

Limitation & Challenge:

- Most existing downstream systems integrate visual features and the textual concepts in the decoding process, mostly ignoring the innate alignment between the two modalities.
- The systems have to learn the alignment between each individual visual feature and textual concept.
- These representations only contain individual features, lacking the meaningful combinations and structural relationships among them.

Those problems hinder the system from understanding images efficiently.

Solution:

- We propose the Mutual Iterative Attention (MIA) module to align the visual features and textual concepts in the encoding process. Using textual concepts to query and integrate visual features with attention, we could get image representations centered upon each concept forming meaning visual feature groups, and vice versa. The representations are refined by applying MIA iteratively.

Approach

Our approach based on the Multi-Head Attention (MHA) and Feed-Forward Network (FCN) from Transformer [1].

Mutual Attention

Given visual features I and textual concepts T, the mutual attention is conducted as:

$$I' = \mathrm{FCN}(\mathrm{MHA}(T, I)), \quad T' = \mathrm{FCN}(\mathrm{MHA}(I', T)) \qquad (1)$$

i.e., visual features are first integrated according to textual concepts, and then textual concepts are integrated according to integrated visual features.

Mutual Iterative Attention (MIA)

We perform mutual attention iteratively to refine both visual features and textual concepts:

$$I_N = \mathrm{FCN}(\mathrm{MHA}(T_{N-1}, I_{N-1})), \quad T_N = \mathrm{FCN}(\mathrm{MHA}(I_N, T_{N-1})) \qquad (2)$$

Semantic-Grounded Image Representations

Since the visual features and the textual concepts are already aligned, we can add them up to get the semantic-grounded image representations:

$$\mathrm{MIA}(I, T) = \mathrm{LayerNorm}(I_N + T_N) \qquad (3)$$

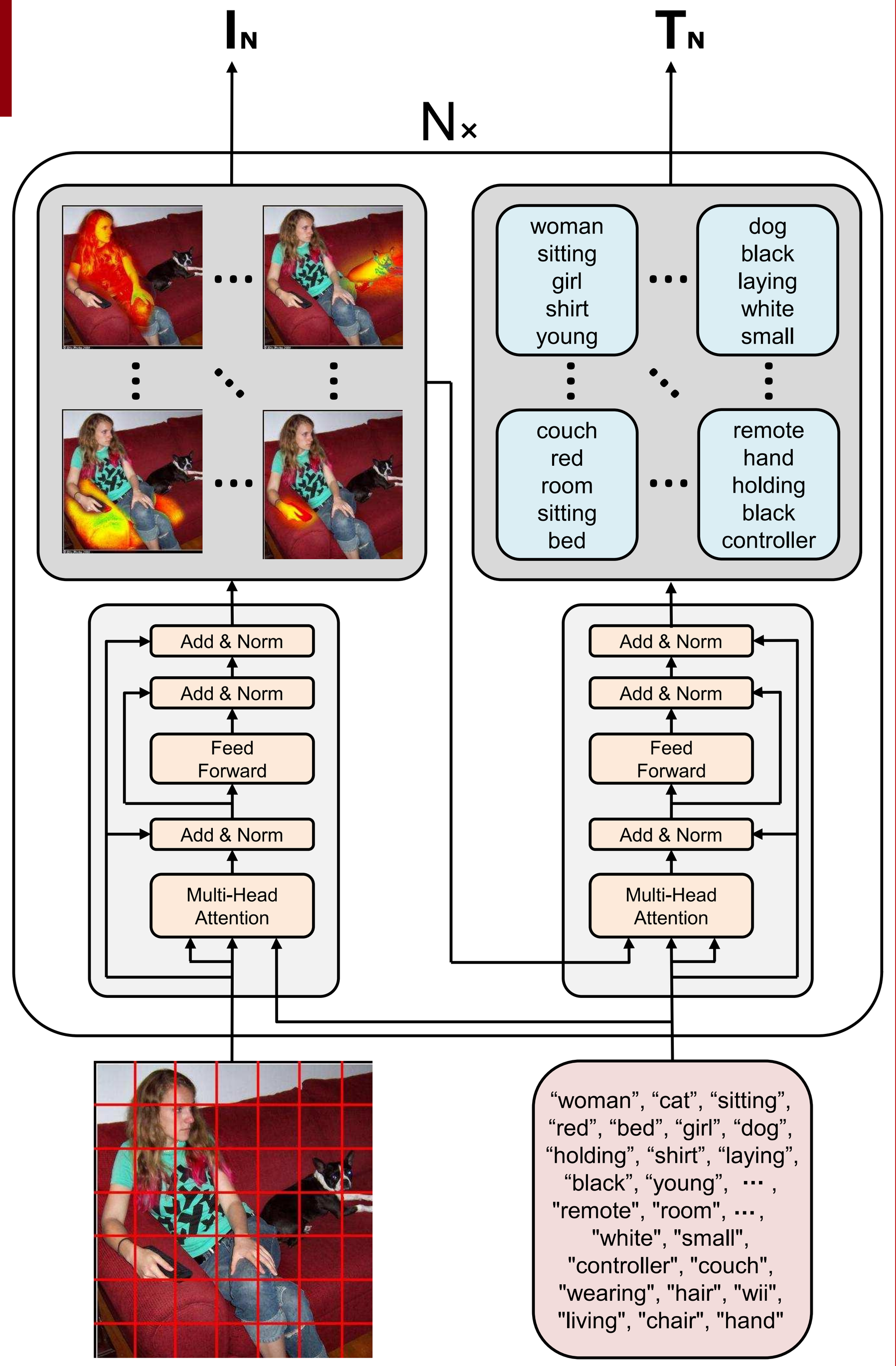

Figure 2. MIA combines the individual features (the lower) from each domain, resulting in integrated image representations (the upper) reflecting certain semantics of the image.

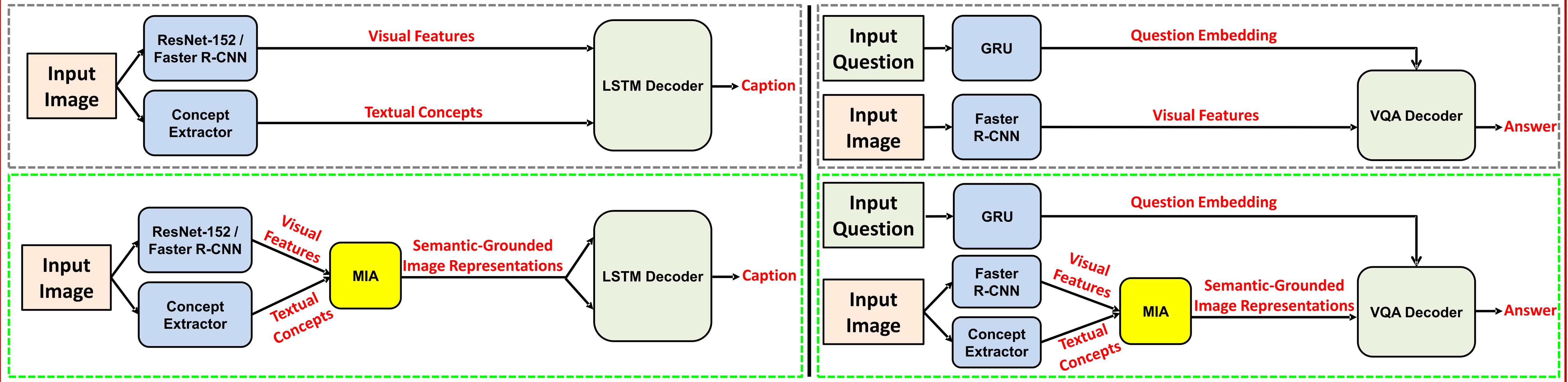

Figure 3. Illustration of how to equip the baseline models with our MIA. MIA aligns and integrates the original image representations from two modalities. Left: For image captioning, the semantic-grounded image representations are used to replace both kinds of original image features. Right: For VQA, MIA only substitutes the image representations, and the question representations are preserved.

Experiments

- We evaluate the proposed MIA on two multi-modal tasks (image captioning and visual question answering (VQA)).

Methods	B-1	B-2	B-3	B-4	M	R	C	S
Visual Attention	72.6	56.0	42.2	31.7	26.5	54.6	103.0	19.3
w/ MIA	**74.5**	**58.4**	**44.4**	**33.6**	**26.8**	**55.8**	**106.7**	**20.1**
Concept Attention	72.6	55.9	42.5	32.5	26.5	54.4	103.2	19.4
w/ MIA	**73.8**	**57.4**	**43.8**	**33.6**	**27.1**	**55.3**	**107.9**	**20.3**
Visual Condition	73.3	56.9	43.4	33.0	26.8	54.8	105.2	19.5
w/ MIA	**73.9**	**57.3**	**43.9**	**33.7**	**26.9**	**55.1**	**107.2**	**19.8**
Concept Condition	72.9	56.2	42.8	32.7	26.4	54.4	104.4	19.3
w/ MIA	**73.9**	**57.3**	**43.9**	**33.7**	**26.9**	**55.1**	**107.2**	**19.8**
Visual Regional Attention	75.2	58.9	45.2	34.7	27.6	56.0	111.2	20.6
w/ MIA	**75.6**	**59.4**	**45.7**	**35.4**	**28.0**	**56.4**	**114.1**	**21.1**

Table 1. Results of representative systems on the MSCOCO image captioning dataset. B-n, M, R, C and S are short for BLEU-n, METEOR, ROUGE-L, CIDEr and SPICE, respectively.

Methods	B	M	R	C	S
Up-Down[2]	36.5	28.0	57.0	120.9	21.5
w/ MIA	**37.0**	**28.2**	**57.4**	**122.2**	**21.7**
Transformer	39.0	28.4	58.6	126.3	21.7
w/ MIA	**39.5**	**29.0**	**58.7**	**129.6**	**22.7**

Table 2. Results of systems under the reinforcement learning setting..

Methods	Test-dev	Test-std
Up-Down[2]	67.3	67.5
w/ MIA	**68.8**	**69.1**
BAN[3]	69.6	69.8
w/ MIA	**70.2**	**70.3**

Table 3. The overall accuracy on the VQA task.

- As we can see, the proposed MIA exhibits compelling effectiveness in boosting the baseline systems.

References

[1] Attention is all you need. *In NIPS, 2017.*

[2] Bottom-up and top-down attention for image captioning and VQA. *In CVPR, 2018.*

[3] Bilinear attention networks. *In NeurIPS, 2018.*

Contact Us

- fenglinliu98@pku.edu.cn
- liuyuanxin@iie.ac.cn
- renxc@pku.edu.cn,
- xusun@pku.edu.cn
- xiaodong.he@jd.com

➢ arxiv

➢ code