Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)

Feature Enhancement in Attention for Visual Question Answering

Yuetan Lin, Zhangyang Pang, Donghui Wang*, Yueting Zhuang
College of Computer Science, Zhejiang University
Hangzhou, P. R. China
{linyuetan,pzy,dhwang,yzhuang}@zju.edu.cn

Abstract

Attention mechanism has been an indispensable part of Visual Question Answering (VQA) models, due to the importance of its selective ability on image regions and/or question words. However, attention mechanism in almost all the VQA models takes as input the image visual and question textual features, which stem from different sources and between which there exists essential semantic gap. In order to further improve the accuracy of correlation between region and question in attention, we focus on region representation and propose the idea of feature enhancement, which includes three aspects. (1) We propose to leverage region semantic representation which is more consistent with the question representation. (2) We enrich the region representation using features from multiple hierarchies and (3) we refine the semantic representation for richer information. With these three incremental feature enhancement mechanisms, we improve the region representation and achieve better attentive effect and VQA performance. We conduct extensive experiments on the largest VQA v2.0 benchmark dataset and achieve competitive results without additional training data, and prove the effectiveness of our proposed feature-enhanced attention by visual demonstrations.

1 Introduction

As a recent surge of interest, Visual Question Answering (VQA) comes as an interesting and promising task which produces meaningful results by processing and fusing the visual and textual data, involving the areas of computer vision, natural language processing and artificial intelligence. Concretely, it takes an image and a corresponding natural language question as input, and outputs the answer about it, of which the question can be as easy as asking the category of an object, or as hard as asking the reason of a phenomenon [Ren *et al.*, 2015a; Antol *et al.*, 2015; Goyal *et al.*, 2017]. In a nutshell, VQA requires deep analysis and understanding of image and question, including image recognition, object localization, attribute prediction, text tokenization, word representation, and even object counting, knowledge inference, etc [Malinowski *et al.*, 2015; Lu *et al.*, 2016; Fukui *et al.*, 2016; Xiong *et al.*, 2016; Andreas *et al.*, 2016; Wu *et al.*, 2016]. The solution of VQA problem will be a great progress in approaching the goal of Visual Turing Test, and will also be conductive to relevant tasks *e.g.* image captioning [Xu *et al.*, 2015; Lu *et al.*, 2017], image-sentence retrieval [Karpathy and Fei-Fei, 2015; Nam *et al.*, 2017], etc.

To capture fine-grained information for deeply understanding images, a vast array of VQA models explore the location-related visual representations, by dividing the image into multiple grids or region proposals. Attention mechanism [Xu *et al.*, 2015] provides a way of capturing question-relevant regions and representing desired visual information accurately. Concretely, it takes as input multiple region representations and a question representation, calculates the correlation values between them, where larger value denotes that the corresponding region is more relevant to the question. Due to the ability of capturing image regions, attention mechanism has become a standard configuration of VQA models [Shih *et al.*, 2016; Lu *et al.*, 2016; Kim *et al.*, 2017]. Seeing that the bottom-up region proposal (*e.g.* Faster R-CNN [Ren *et al.*, 2015b]) captures more meaningful image regions which are not limited to gridded regions or objects with closed curves, we resort to this mechanism used in bottom-up attention model [Anderson *et al.*, 2018], which has demonstrated great success on image captioning and VQA tasks, for finer-grained visual representation in our attention model.

However, the input features of attention models stem from different modal data. Specifically, the visual features usually come from CNNs trained on ImageNet [Deng *et al.*, 2009] image data, and textual features are from word embedding models [Pennington *et al.*, 2014] which are pre-trained on language corpora such as Wikipedia corpus. There is an essential semantic gap between them, which is often overlooked by attention-based methods.

To remedy this problem, we focus on the region representation in attention and propose the idea of feature enhancement, which includes three aspects as follows. (1) From an intuitively more plausible perspective, we take advantage of the semantic representations of image regions, *e.g.* the category or attribute labels. Since the semantic representation of a region is more abstract and has higher-level informa-

*Corresponding author

tion than the visual one, it could handle semantically more complicated questions. (2) Due to the monotony and information shortage of semantic representation, we supplement it with features from different hierarchies, *i.e.* the visual representation, which contains more detailed information. Thus it deals with questions about attributes or fine-grained types by providing whether the object is a red apple or a green one, while the semantic representation gives only the apple category. (3) We further refine the region representation using probabilistic semantic information, and achieve better performance. With these incremental feature enhancement mechanisms, we achieve state-of-the-art results of single model on VQA evaluation server without additional training data and rank second entry on the 2017 VQA test-standard leaderboard using ensemble models.

2 Related Work

2.1 Attention-based VQA Models

To tackle the problem of Visual Question Answering (VQA), deep understanding of the image is essential, and holistic image representation is insufficient to capture fine-grained information (*e.g.* spatial location, multiple objects and relationships) for answering the question. Attention mechanism (especially on the image) has been a standard practice for VQA, which is to select relevant image regions using the question representation. Attention-based models have demonstrated great success in VQA tasks [Shih *et al.*, 2016; Zhu *et al.*, 2016; Lu *et al.*, 2016]. [Yang *et al.*, 2016; Fukui *et al.*, 2016; Kim *et al.*, 2017] use multiple attention layers or generate multiple attention maps to achieve multiple steps of reasoning or multiple glimpses on the image. [Shih *et al.*, 2016] takes image region proposals and four binned semantic representations for question processed by Stanford Parser [De Marneffe *et al.*, 2006] to produce an attention weight for each region proposal, however it only aims at object proposals with closed curve, which limits the detectability of meaningful regions. Recently, [Anderson *et al.*, 2018] has proposed to compute attention at the level of objects and other salient image regions combining bottom-up and top-down attention. It extends the number of object categories from 200 to 1600 and is trained on the large Visual Genome dataset [Krishna *et al.*, 2017] which contains dense annotations of objects, attributes, region descriptions, etc. These semantics-richer region representations greatly enhance the performance on image captioning and VQA tasks [Anderson *et al.*, 2018; Teney *et al.*, 2017], which are adopted in our model.

2.2 Visual-Textual Embedding Methods

Due to the semantic gap existing between different sources (*e.g.* low-level image data and high-level text data), visual-textual embedding methods have been proposed to bridge or cross the gap [Frome *et al.*, 2013; Kiros *et al.*, 2014; Karpathy and Fei-Fei, 2015]. The deep visual-semantic embedding model [Frome *et al.*, 2013] maps the two embeddings to be similar using dot-product similarity. [Kiros *et al.*, 2014] maps image and text into a common multimodal joint embedding space to unify these two embeddings. However, these two methods use holistic image representation, which is not applicable for deeper image understanding. [Karpathy and Fei-Fei, 2015] learns the correspondences between language and image region data and shows good performance on retrieval and image captioning tasks.

3 Feature Enhancement for VQA

As in attention-based VQA models, attention mechanism takes as input the question and every image region feature, and computes correlation values between them. After normalization, the resulting values (*i.e.* the attention weights) indicate the probability of question selection on the image regions. The visual feature is the sum of all these region features weighted by their attention weights, and is fed to later fusion with question representation.

However, the question and image region features are not in the same modality, since they are trained using different types of data, and intuitively there is natural inconsistency. Through our experiments, we have found that the inconsistency between distinct sources and different hierarchical features affects the effect of attention. Therefore, we propose the idea of feature enhancement, which includes three aspects to make the question and image region features more consistent, the first two aspects are illustrated in Fig. 1 (b-d).

3.1 VQA Model Overview

Our VQA model is a basic attention model based on bottom-up attention model [Anderson *et al.*, 2018; Teney *et al.*, 2017], but different in two aspects. Firstly, we focus on bridging the gap between visual and textual features in attention mechanism. Secondly, we do not include the output classifier pre-trained using extra word embeddings and Google image data due to unavailability, and several tricks (*e.g.* sigmoid loss) are not adopted. Our model contains 3 elementary parts illustrated in Fig. 1, *i.e.* feature encoding, attention and feature fusion, and answer generation.

(1) In the feature encoding part, a question is first tokenized into a list of words and punctuations, then this list is trimmed or padded to have same length which is sufficient to express the meaning of question. The tokenized terms are represented by word embeddings (*i.e.* GloVe [Pennington *et al.*, 2014]), and are input into a recurrent model (*i.e.* gated recurrent units (GRU) [Cho *et al.*, 2014]) as if the recurrent model sees one term per time-step. The output at final time-step is regarded as the question representation $\mathbf{q} \in \mathbb{R}^{d_Q}$. The image is encoded via bottom-up attention model [Anderson *et al.*, 2018] implemented using Faster R-CNN [Ren *et al.*, 2015b], resulting in a certain number of CNN-encoded visual features as our image regions representation $\mathbf{V}_i \in \mathbb{R}^{d_I}$, where $i = 1, \cdots, k$ and k is the number of region proposals per image.

(2) In the attention part, we denote the region representation which is correlated with question representation to produce **a**ttention weights as $\mathbf{R}^{(a)}$ and the region representation to be **f**used by attention weights as $\mathbf{R}^{(f)}$, and models to be compared using different region representation pairs as $\mathbf{R}^{(a)}$/$\mathbf{R}^{(f)}$. We compute the correlation value by

$$\mathbf{x}_i = \mathbf{W}_C f(\mathbf{W}_V(\mathbf{q}; \mathbf{R}_i^{(a)})), \tag{1}$$

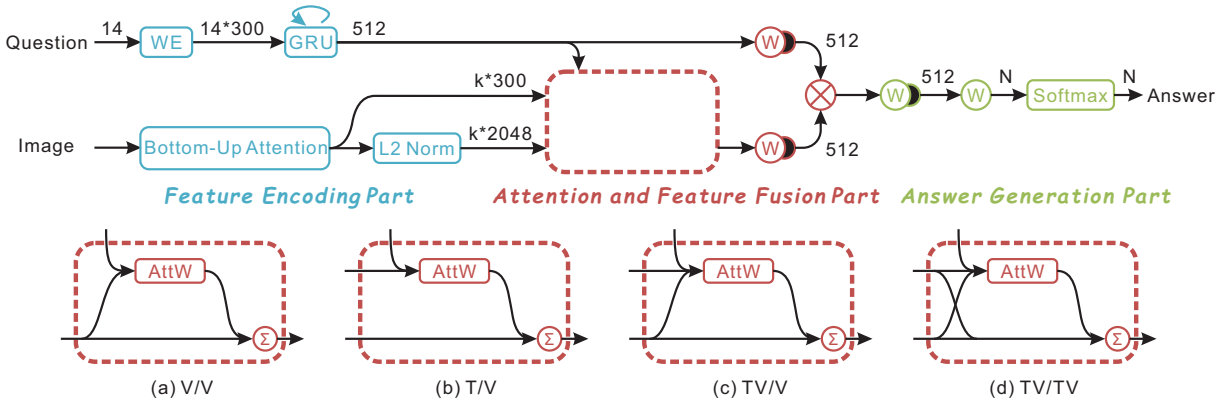

Figure 1: The model architecture with one of the mechanisms: (a) conventional attention and (b-d) proposed feature enhancement mechanisms. The sub-caption $\mathbf{R}^{(a)}$/$\mathbf{R}^{(f)}$ denotes region representation for generating attention weights ($\mathbf{R}^{(a)}$) and that for region feature fusion ($\mathbf{R}^{(f)}$), where $\mathbf{V}$ and $\mathbf{T}$ denote visual and textual features. The numbers above arrows are the feature sizes. N is the number of answers, and k is the number of region proposals per image. The abbreviations "WE", "GRU", "W", "AttW" denote word embedding, gated recurrent units, linear mapping, attention weights, respectively. The symbols "Σ", "⊗", solid circle denote weighted sum, Hadamard product, non-linear function, respectively.

where f is the non-linear activation function, $\mathbf{W}_C \in \mathbb{R}^{g \times d_C}$, $\mathbf{W}_V$ are trainable parameters and d_C is the common embedding size, g is the number of glimpses. We employ two-headed or two-glimpses ($g = 2$) attention as suggested in [Fukui *et al.*, 2016; Kim *et al.*, 2017] since intuitively it captures different explainable attentive selections. The attention weights are obtained via $\boldsymbol{\alpha} = \phi(\mathbf{x})$, where ϕ is the softmax function. The visual feature is obtained by summing the region representations weighted by attention weights,

$$\mathbf{v} = \Sigma_i^n \boldsymbol{\alpha}_i \mathbf{R}_i^{(f)}. \quad (2)$$

In the multi-modal feature fusion part, the visual feature $\mathbf{v}$ and the question representation $\mathbf{q}$ are mapped to a common space and fused via Hadamard product (element-wise multiplication) to get a fused feature $\mathbf{f} = (\mathbf{W}_q\mathbf{q}) \otimes (\mathbf{W}_v\mathbf{v})$.

(3) Then the fused feature is mapped through a linear layer and a softmax layer to produce the probabilities of the candidate answers $\mathbf{a} = \mathrm{softmax}(\mathbf{W}_f\mathbf{f})$, which are selected from the most commonly occurred answers in the training data.

3.2 V/V: Conventional Attention

In conventional attention mechanism, we use region visual features ($\mathbf{V}$) in both attention obtaining ($\mathbf{R}_i^{(a)} = \mathbf{V}_i$) and region feature fusion ($\mathbf{R}_i^{(f)} = \mathbf{V}_i$), therefore we denote this model as **V/V**.

In the following three sections, we will introduce mechanisms to enhance region features in attention and fusion.

3.3 T/V: Region Semantic Representation

Since the question words are encoded using word embeddings, it is easy and intuitive to represent each region as a semantic representation ($\mathbf{T}$) in the same word embedding space to obtain attention weights by correlating with the question representation. Specifically, we predict the object label of a region and further represent it by a word embedding model such as GloVe [Pennington *et al.*, 2014]. We denote the i-th region semantic representation as $\mathbf{T}_i \in \mathbb{R}^{d_T}$, where d_T is the dimension of the word embedding. For example, when asked "Where is the cat?", it is easier to correlate the question with a "cat" word embedding than a visual representation of a cat region of various cat breeds, postures, colors or illumination environments.

We keep the region feature to be fused unchanged ($\mathbf{R}_i^{(f)} = \mathbf{V}_i$), which itself preserves much information of the region. And we denote the model with this mechanism as **T/V**.

3.4 TV/V and TV/TV: Representation Enrichment using Multi-Level Features

Since the semantic representation of an image region comes from the final layer of a CNN (and further represented by a word embedding), it provides more abstract and higher-level information. However, the information of a textual label itself is inevitably monotonous. The visual features in relatively shallow layer of CNN provide richer visual information like color, shape, etc. For example, when asked "How many black cats are there?", the visual representation provides necessary determinative information.

Then the region feature used for producing attention weights is enhanced using the concatenation of semantic and visual representations (**TV**), *i.e.* $\mathbf{R}_i^{(a)} = \mathbf{T}_i \mid \mathbf{V}_i$, each of which provides complementary information to correlate with the question representation. The region feature to be fused by attention weights can be the visual representation ($\mathbf{R}_i^{(f)} = \mathbf{V}_i$) and can also be enriched ($\mathbf{R}_i^{(f)} = \mathbf{T}_i \mid \mathbf{V}_i$). Therefore the model is denoted as **TV/V** or **TV/TV** according to the region feature used in weighted sum.

3.5 rTV/V and rTV/rTV: Semantic Representation Refinement

A region can be described by different words, which means it can be expressed by more information. Imagine a region which contains a pony and the correct label "pony" is not

in the top-predicted entry, it is necessary to include the first few predicted object labels to increase the correctness of the semantic representation.

To improve the robustness of the semantic representation, we refine the semantic representation **rT** as the average of top-ranked m object label embeddings weighted by the output label probabilities β, *i.e.* $\mathbf{rT}_i = \sum_{j=1}^{m} \beta_j \mathbf{T}_j$. To this end, a region can be described by the semantic-richer representation with a visual one, denoted as **rTV/V** or **rTV/rTV**.

We take the two-headed attention with **rTV/rTV** as our final accurate model. Each attention head produces a set of attention weights for the region feature summation, and the question representation is also mapped twice. The two features after feature fusion are concatenated for later answer generation.

4 Experiments

In order to validate effectiveness of our proposed model, we conducted extensive experiments comparing our feature-enhancement mechanisms, existing reported results and some visualization results on the VQA v2.0 dataset.

4.1 Datasets

We trained our model and conducted comparative experiments on VQA v2.0 [Goyal *et al.*, 2017] dataset. VQA v2.0 is distinct from the first version of VQA dataset [Antol *et al.*, 2015] in that it balanced the question to have a pair of images which result in two different answers. This balanced version avoids the problem of predicting answer from question only and is almost twice larger than v1. VQA v2.0 provides question-answer pairs on MS COCO real images [Lin *et al.*, 2014] and abstract scenes, and we follow the common practice to evaluate on the real images only, which are closer to real world application and there is automatic assessment on the evaluation server. 10 human-labeled answer annotations per question are provided for training and validation splits. The evaluation metric is $\min(N/3, 1)$, where N is the number of human-labeled answers consistent with the prediction.

4.2 Training Details

In our experiment, we employ two model versions, *i.e.* the basic attention (*base-att*) model and the double attention (*double-att*) model. The base-att model contains less training parameters for fast computation and verification experiments, and the double-att model is slow but more accurate.

The common part of both two versions. Given a question image pair, we first tokenize the question into words and punctuations and trim them to have a maximum length of 14 as adopted in [Teney *et al.*, 2017], which covers 99.47% complete questions in VQA v2.0. We use a 300-dim GloVe [Pennington *et al.*, 2014] word embedding to denote a token item, as well as the semantic representation of a region. We take each word embedding as an input for each time-step of GRU, which has 1 hidden layer of size 512, and we take the 512-dim output of the final time-step as the question representation. We use the bottom-up attention model to encode the image and obtain 36 region visual features of size 2048. We employ *L2*-normalization on the visual features. The final layer for answer vector is of size about 3000.

We use hyperbolic tangent (tanh) as the non-linear activation function in our model. We use RMSprop optimizer with mini-batches of 512. The learning rate is initialized to be 3e-4 and kept fixed for the first 40 epochs, and is decayed every 10 epochs with a decay factor.

Difference. The main differences between two versions are the attention, dropout usage and learning rate changing manner. The attention weights are linear mapped from the combination of region and question representation. The base-att model takes the combination as the concatenation of both representations, while the double-att model first linear maps both representations to a common space with same dimension and takes the combination as the Hadamard product (element-wise multiplication) of the two representations. The base-att model produces one set of attention weights (one glimpse), and the double-att model uses two-headed attention (two glimpses). The base-att model uses dropout of 0.5 only after word embedding layer, before generating the attention weights and before generating answer, while the double-att uses dropout of 0.5 before every linear layer. The learning rate decay factor of two versions are 0.8 and 0.9, respectively.

4.3 Ablative Results

To verify the effectiveness of our proposed model, we compare our *base-att* model with respect to different feature enhancement mechanisms and report the accuracy on VQA v2.0 validation split in Table 1. The "Overall" accuracy denotes performance on all the testing instances, while "Yes/No", "Number" and "Other" accuracies are performances on testing instances of different answer types. We demonstrate the experimental results over **V/V**, **T/V**, **TV/V**, **TV/TV** and the semantic representation refined ones.

validation	Yes/No	Number	Other	Overall
V/V	79.51	40.53	52.48	61.07
T/V	79.82	41.93	52.18	61.22
TV/V	80.30	41.87	53.21	61.90
TV/TV	80.38	**42.38**	53.15	61.97
rT/V	79.97	42.03	52.84	61.61
rTV/V	**80.48**	41.90	**53.40**	**62.06**
rTV/rTV	**80.46**	**42.80**	**53.57**	**62.26**

Table 1: Ablative results (accuracy in %) of *base-att* model with respect to different feature enhancement mechanisms on VQA v2.0 validation split (no extra training data and no ensembles)

From Table 1 we see that using the region semantic representation in attention (*i.e.* **T/V**) instead of the visual one (*i.e.* **V/V**) promotes the overall accuracy, which proves that the semantic representation more consistently relevant to question performs better than the visual one in correlation with the question representation. And enriching the semantic representation using multiple hierarchical features improves performance consistently (*i.e.* **TV/V** and **TV/TV**),

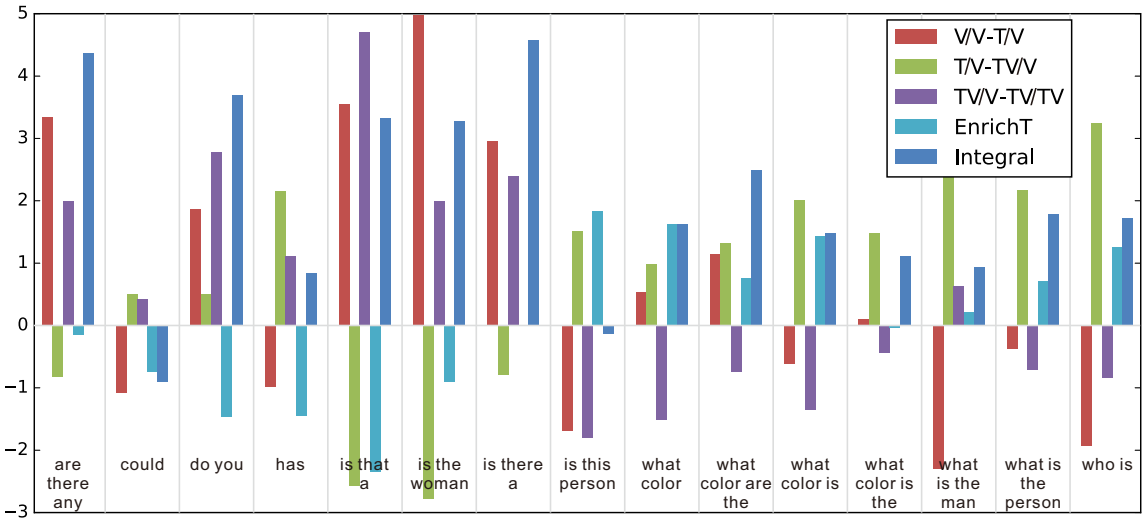

Figure 2: (Best view in color.) Per question-type accuracy improvement (in %, and decrease in negative value) of *base-att* model with respect to different feature enhancement mechanisms over several outstanding question types on VQA v2.0 validation split.

which demonstrates that enriched region representation contains richer information which handles better in attention and feature summation. Models using refined label consistently improve the performance over models using single label, showing that semantic representation refinement provides more precise representations. Note that we do not use extra Visual Genome data for training, and only use a single model without ensembles.

To demonstrate the effect of the cumulative ablations, we further illustrate the per question-type accuracy improvement (decrease in negative value) of *base-att* model with respect to different feature enhancement mechanisms over several outstanding question types on VQA v2.0 validation split in Fig. 2. The legend denotes from **A/B** to **X/Y** (A/B-X/Y), from **TV/TV** to **rTV/rTV** (EnrichT), and from **V/V** to **rTV/rTV** (Integral).

From Fig. 2, when we substitute the region visual representation for the semantic one (1.V/V-T/V) or enrich the visual representation using the semantic one (3.TV/V-TV/TV), we get performance improvement on questions asking about object(s), *e.g.* "Are there any", "Is that a", "Is there a", since the semantic representation gives directly the object label which is more consistent with the question representation. On the contrary, for questions asking about colors, *e.g.* "What color (is / are) (the)", the region visual representation plays a more important role in attention computation due to the consistent improvement when we enrich the region semantic representation using the visual one (2.T/V-TV/V). We lose the accuracy on these "color" questions when we enrich the visual representation using the semantic one (3.TV/V-TV/TV), probably because the single label brings interference information. But we gain some performance improvement when we refine the semantic label (4.EnrichT), which shows the robustness of semantic representation refinement. With all the feature enhancement mechanisms integrated (5.Integral), the vast majority of the question types get performance improvement, which demonstrates the effectiveness of our model.

test-dev	**Yes/No**	**Number**	**Other**	**Overall**
MFB	-	-	-	64.98
Bottom-Up Att	81.82	44.21	56.05	65.32
MFH	-	-	-	65.80
Our	**82.50**	**45.80**	**57.34**	**66.40**
test-std	**Yes/No**	**Number**	**Other**	**Overall**
MCB	78.82	38.28	53.36	62.27
Bottom-Up Att	82.20	43.90	56.26	65.67
Our	**82.44**	**44.93**	**57.60**	**66.52**

Table 2: Quantitative accuracy results of the *double-att* model without ensembles on VQA v2.0 **test-dev** and **test-std** splits.

4.4 Comparison with Existing Methods

We summarize our accuracy results of the *double-att* model on VQA v2.0 test-dev and test-std splits with existing methods in Table 2. Due to room shortage, we only compare with the reported accuracy results of three methods which are top ranked in 2017 VQA challenge, *i.e.* Bottom-Up Att [Teney *et al.*, 2017; Anderson *et al.*, 2018] the winner and MFB [Yu *et al.*, 2017], MFH [Yu *et al.*, 2018] the second place, and the winner of VQA Challenge 2016, *i.e.* MCB [Fukui *et al.*, 2016; Goyal *et al.*, 2017].

Unlike MFB, MFH and Bottom-Up Att, we do not use additional Visual Genome question-answer (QA) data for training and achieve state-of-the-art single model accuracy on test-dev and test-std splits. Our model with 8 ensembles utilizing the feature enhancement mechanism ranked second entry on the leaderboard [1], achieving **70.40%** overall accuracy.

[1] https://evalai.cloudcv.org/featured-challenges/1/leaderboard/3

4.5 Qualitative Results

To illustrate the effect of our feature enhancement mechanisms, we list several samples with their attention weights on the image and predicted answers below by our *base-att* model with respect to different feature enhancement mechanisms on VQA v2.0 validation split in Fig. 3. Since the attention weights are normalized to have sum of 1, we re-normalize the weights to have maximum of 1 and minimum of 0.1 for better illustration.

We can see from Fig. 3 that, the conventional attention mechanism (**V/V**) diverges on multiple regions (first example) or not focus on the correct regions (second example). While our models show focused attention regions (*i.e.* kite and kayak) and give correct answers. For the last sample, **T/V**, **TV/V** and **TV/TV** give a wrong answer, we speculate that this is because "frog" is an unseen label for our detector.

5 Conclusion

In this paper, we propose the idea of feature enhancement to tackle the issue in attention mechanism of directly correlating region visual feature and question textual representation, which stem from different sources and between which exists the essential semantic gap. The feature enhancement includes three aspects, *i.e.* region semantic representation, representation enrichment using multiple hierarchical features and semantic representation refinement. The general idea of combining multiple hierarchical features in attention based methods can be applied to many tasks in multi-modal community. We improve the region representation and achieve better attentive effect and VQA performance, demonstrated by empirical experiments and visual demonstration on the largest VQA v2.0 benchmark dataset. Note that our model is trained without extra QA data and show competitive results.

As we find in our experiments, enhanced region features provide semantically more consistent and complementary information. But the counting, direction and optical character recognition (OCR) questions are still hard to solve. Our future works include utilizing more trained region representations, *e.g.* activities, attributes, relationships and integrating specialized modules.

Figure 3: (Best view in color and zoom in.) Qualitative results of our *base-att* model with respect to different feature enhancement mechanisms on VQA v2.0 validation split. We show samples using different feature enhancement mechanisms, and "Q", "A", "P" denote question, ground-truth answer, predicted answer, respectively. The correctly predicted answer is in green and the false one in red.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61473256) and China Knowledge Center for Engineering Sciences and Technology (CKCEST).

References

[Anderson *et al.*, 2018] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. In *CVPR*. IEEE, 2018.

[Andreas *et al.*, 2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, pages 39–48, 2016.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zit-

nick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, pages 2425–2433, 2015.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014.

[De Marneffe *et al.*, 2006] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses. In *LREC*, volume 6, pages 449–454. Genoa Italy, 2006.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, pages 457–468. ACL, 2016.

[Goyal *et al.*, 2017] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, July 2017.

[Karpathy and Fei-Fei, 2015] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

[Kim *et al.*, 2017] Jin-Hwa Kim, Kyoung-Woon On, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *ICLR*, 2017.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, pages 289–297, 2016.

[Lu *et al.*, 2017] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, July 2017.

[Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, pages 1–9, 2015.

[Nam *et al.*, 2017] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, July 2017.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[Ren *et al.*, 2015a] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2953–2961, 2015.

[Ren *et al.*, 2015b] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[Shih *et al.*, 2016] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, June 2016.

[Teney *et al.*, 2017] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017.

[Wu *et al.*, 2016] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *CVPR*, June 2016.

[Xiong *et al.*, 2016] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 37, pages 2048–2057, 2015.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, June 2016.

[Yu *et al.*, 2017] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *ICCV*, Oct 2017.

[Yu *et al.*, 2018] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *T-NNLS*, 2018.

[Zhu *et al.*, 2016] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *CVPR*, June 2016.