
An Interpretable (Conversational) VQA model using Attention based Weighted Contextual Features

Yu Wang

Samsung Research America
yu.wang1@samsung.com

Abhishek Patel

Samsung Research America
abhishek.p@samsung.com

Yilin Shen

Samsung Research America
yilin.shen@samsung.com

Hongxia Jin

Samsung Research America
hongxia.jin@samsung.com

Larry Heck

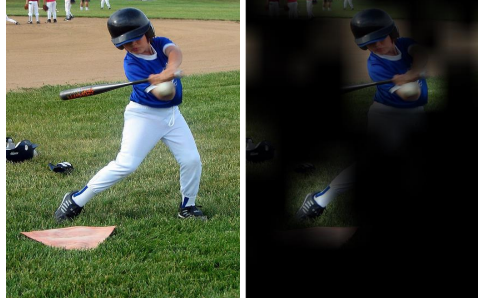
Samsung Research America
larry.h@samsung.com

Abstract

Visual question answering (VQA) is a challenging task which is required a deep understanding of language and images. Currently, most of the algorithms focus on finding the correlation between the basic question embeddings and image features by using element-wise product or bilinear pooling between these two vectors. In this paper, a deeper analysis of these features by further extracting their contextual information (without using external features) is given. A novel interpretable attention-based weighted contextual feature model is proposed for VQA task by assigning adaptive weights to their contextual features and themselves based on the importance. Due to the model's interpretability, we also show in our paper that how this model can be applied in tasks like Conversational VQA by asking user interpretable questions to explain ambiguity.

Introduction

Visual question answering(VQA) has become a popular research problem which is being studied from the perspectives of multiple disciplines, like natural language understanding and computer vision. The target of a VQA task is to generate an answer based on an image and a related given question. The answer can be a number, yes/no or just word phrases. The task itself is not trivial as it needs to understand the image and its corresponding question and further find the correlation between the question-image pairs based on their own features and even some auxiliary external features (Frome et al. (2013); Gong et al. (2014); Malinowski et al. (2015); Yang et al. (2016); Rajani & Mooney (2016); Lu et al. (2016); Teney et al. (2017); Yu et al. (2018); Zhang et al. (2018)). Currently, most of the approaches take advantage of the concept of multiple modalities by representing images and queries separately using two embedding or feature vectors Fukui et al. (2016). The vision modal extracts the image features using a convolutional neural network (CNN)Krizhevsky et al. (2012). Comparatively, the question understanding modal generates a question embedding vector representing its semantic meanings, by using either a recurrent neural network (RNN) or Bag-of-Words approach Harris (1954). In order to locate the important information from a question-image pair, most of current approaches use the concept of attention/co-attention in their algorithms. By assigning different weights to image features, a good attention algorithm is able to select the most important features in an image that is related to a question asked. There are many ways to generate attention weights, for instance like, element-wise sum or product, multimodal compact bi-linear



Original Image Attention Map
 Q: Is the boy wearing protective headgear
 playing basketball?
 Ans: Yes (Incorrect)

Figure 1: An example giving wrong answer using a VQA model

pooling (MCB)Fukui et al. (2016) and etc., most of them have decent performance on VQA task. Despite the decent results obtained by using different attention mechanisms for VQA, the overall performance is still not comparable to human beings Lu et al. (2016). One of the possible reason is that human can use more context-level information in a question to imply the answer based on an image and also the image’s contextual information. By observing the attention maps and test results generated by one of the current state-of-the-art VQA model, we notice that most of the wrong answers are generated due to two reasons: 1. The question and its context-level information is not fully understood by the system, hence a correct answer cannot be generated despite the right regions of an image can be located.

2. Due to the attention mechanism applied, part of the important context-level information in an image is missing, hence may adversely affect the system performance.

As an example shown in Figure 1: the orange color text is a contextual information should be de-emphasized by the model since the text in purple color is the real question need to be addressed. Though the attention map locates the correct regions in the image, but the model still cannot find the right answers correspondingly.

In order to overcome these two potential difficulties, in this paper, an interpretable new structure for VQA is designed by considering the context-level information from questions and images. Weighted contextual features structure is also proposed to balance the information from the question/image and the their contextual information by assigning appropriate ratios through learning.

Our main discussions and contributions in the paper are from the following three aspects:

1. We propose a framework using contextual features extracted from multiple sources (image and query) to improve VQA performance, by further considering their cross-impact with different types of data in VQA.
2. We introduce interpretable weighted contextual feature (WCF) modules to balance the wights of the different contextual features based on their importance.
3. A brief explanation is given to explain that how to expand the model to more complicated conversational VQA tasks.

The paper structure is organized as following: A detailed system design based on the context-level information and the attention-based weighted contextual feature model will be explained in section 2. In section 3, experiments are performed to demonstrate the performance of our new system on standard VQA datasets, and examples are given to show how the weighted context information can help on Conversational VQA by asking user interpretable questions to explain ambiguity.

An Attention-based Weighted Contextual Features (WCF) Model for Visual Question Answering (VQA)

As described earlier, it is possible to extract contextual features in semantics using an RNN based encoder-decoder structure, or contextual features in images using MDLSTM based encoder-structure. Specifically, the RNN structure in our model is chosen as a bidirectional LSTM (BLSTM)

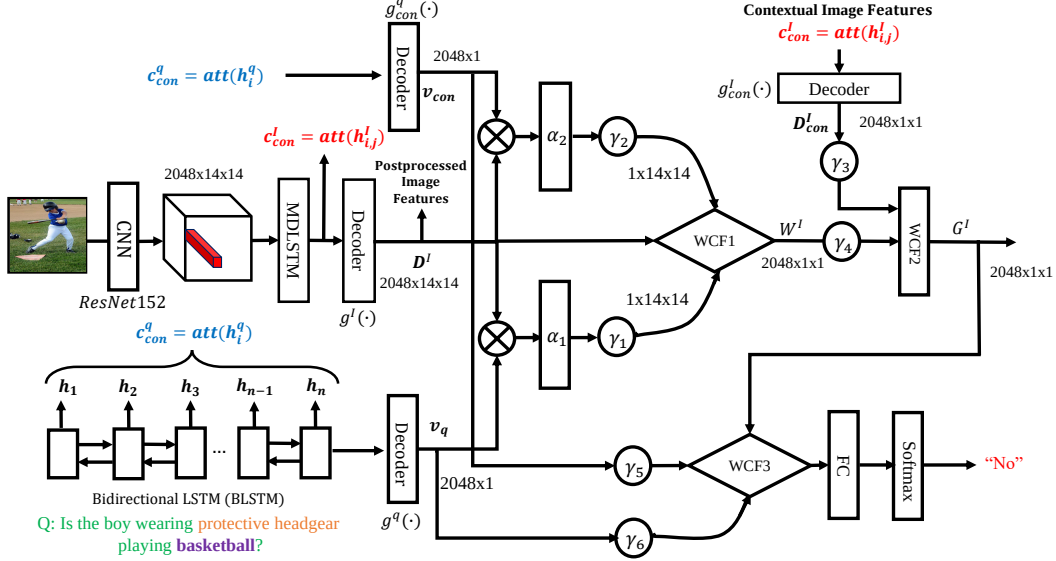


Figure 2: A general structure for the attention based weighted contextual features (WCF) model

structure Graves & Schmidhuber (2005). Also as the VQA example given in Figure 1, it is noticed that many misinterpreted VQA task is due to the misunderstanding of the context level information from question (which is major) or images. Inspired by these observations and model features, we proposed an attention-based models using contextual features from both questions and images for the VQA task. The basic structure is as shown in figure 2.

As shown in the figure, the attention based model structure contains several sub-modules:

1. The image contextual feature extraction (I-CFE) module
 2. The weighted contextual features (WCF1,2) based question-image understanding module and
 3. The weighted contextual feature (WCF3) based answer generation module
- Their detailed descriptions will be given in this section.

The Image Contextual Feature Extraction (I-CFE) Module

Following the Multidimensional LSTM (MDLSTM) definition as shown in Graves & Schmidhuber (2009), a detailed explanation on building the image contextual feature extraction (I-CFE) module based on MDLSTM is given in this section.

The image feature tensor containing k image features is represented by $I \in \mathbb{R}^{k \times n \times n}$, which is normally extracted by passing the raw image data into an convolutional neural network (CNN) based image classifier, then extracting its embedding features before the last softmax layer. Then the MDLSTM encoder reads in each image feature $I_{i,j} \in I$ together with hidden state vectors $h_{i,j-1}$ and $h_{i-1,j}$, which are generated from the image features $I_{i,j-1}$ and $I_{i-1,j}$. These hidden states will be used as the input to another MDLSTM based decoder, such that decoded image features can be further generated. By using the similar attention mechanism as illustrated in section ??, an image contextual feature is generated as:

$$c_{con}^I = att(h_{i,j}^I) = \sum_{i=1}^n \sum_{j=1}^n \alpha_{i,j} h_{i,j}^I \quad (1)$$

where $n \times n$ is the number of image features generated by the CNN from the raw image inputs, and $\alpha_{i,j}$ is calculated as:

$$\alpha_{i,j} = \frac{\exp(\tau_{i,j})}{\sum_{i=1}^n \sum_{j=1}^n \exp(\tau_{i,j})} \quad (2)$$

$$\tau_{i,j} = \phi(s_I, h_{i,j}^I)$$

where $\phi(\cdot)$ is a feed-forward neural network and $\mathbf{s}_I$ is the last hidden state generated by the MDLSTM decoder. Since there is only one feature vector generated by the decoder for each image, only the last one is applied, *i.e.* $\mathbf{s}_I$.

The post-processed image features $\mathbf{D}^I \in \mathbb{R}^{k \times n \times n}$ is generated by taking the hidden states from the encoder as the decoder's input:

$$\mathbf{D}_{i,j}^I = g^I(\mathbf{s}_{i-1,j}^I, \mathbf{s}_{i,j-1}^I, \mathbf{h}_{i,j}^I) \quad (3)$$

where $g^I(\cdot)$ is an MDLSTM decoder. Similarly, the contextual feature's decoder output $\mathbf{D}_{con}^I \in \mathbb{R}^{k \times 1 \times 1}$ is calculated as following:

$$\mathbf{D}_{con}^I = g_{con}^I(\mathbf{s}_{n-1,n}^I, \mathbf{s}_{n,n-1}^I, \mathbf{c}_{con}^I) \quad (4)$$

where $\mathbf{s}_{n-1,n}^I$ and $\mathbf{s}_{n,n-1}^I$ are the previous decoder hidden states of $\mathbf{s}_{n,n}^I$, and $\mathbf{c}_{con}^I$ is the contextual image feature generated as shown in (1).

Remarks:

It is worth noticing that $\mathbf{D}_{i,j}^I$ and $\mathbf{D}_{con}^I$ are two different image-level features, one is generated by the image features decoder $g^I(\cdot)$ and the other is generated by the image contextual features decoder $g_{con}^I(\cdot)$. Also their dimensions are different as $\mathbf{D}^I \in \mathbb{R}^{k \times n \times n}$ and $\mathbf{D}_{con}^I \in \mathbb{R}^{k \times 1 \times 1}$, since $\mathbf{D}^I$ corresponds to all image features and $\mathbf{D}_{con}^I$ is only from one contextual feature $\mathbf{c}_{con}^I$.

The Weighted Contextual Features (WCF) based Question-Image Understanding Module

After extracting the contextual feature from questions and images separately, the next important question is that how to use these features effectively based on their importance. In this paper, we proposed a weighted contextual feature (WCF) based approach to take advantage of the contextual vectors generated by our model. The first application of this technique in our problem is to understand the questions on our given images, *i.e.* find the features of an image that is most related to the image-question pair.

One of the most effective ways to deal with this problem is the (co-)attention mechanism by assigning an attention weight to each image feature vector then sum them together. There are several popular approaches to generate the attention weights including projection Shih et al. (2016), HieCoAtt Lu et al. (2016), MCB Fukui et al. (2016) and etc. In our model we will use the projection approach to find the attention weights as in (5):

$$\begin{aligned} \alpha_1 &= \sigma((\mathbf{D}^I)^T \mathbf{v}_q) \\ \sigma(z_j) &= \frac{e^{z_j}}{\sum_{k=1}^n e^{z_k}} \quad \text{for } j = \{1, \dots, n\} \end{aligned} \quad (5)$$

where α_1 is the normalized softmax weighting ($\sigma(\cdot)$) of the inner products of the post-processed image feature vector $\mathbf{D}^I$ and the decoded question feature vector $\mathbf{v}_q$. Since $\mathbf{D}^I \in \mathbb{R}^{k \times n \times n}$ and $\mathbf{v}_{con} \in \mathbb{R}^{k \times 1 \times 1}$, the attention weights generated by (5) has a dimension of $\alpha_1^T \in \mathbb{R}^{1 \times n \times n}$.

So α_1 is an attention weights of image features from the perspective of the question embedding. Similarly, we can generate another attention weight vector α_2 by projecting the question contextual features $\mathbf{v}_{con}^q$ on the post-processed image feature vector $\mathbf{D}^I$, such that the useful image features based on the question contextual information can be highlighted by assigning larger attention weights:

$$\alpha_2 = \sigma((\mathbf{D}^I)^T \mathbf{v}_{con}) \quad (6)$$

where $\alpha_2^T \in \mathbb{R}^{1 \times n \times n}$.

Despite that the attention weights generated by using question embedding $\mathbf{v}_q$ and the question contextual feature $\mathbf{v}_{con}$ has the same dimension, the naive sum of the two attention weights won't give a decent performance in our model since the importance of a sentence embedding and that of its contextual feature may not be weighted equally. This inspires us to assign another weight parameter γ_i to each of the attention weight α_i , such that the importance of attention from different sources can be further learned.

$$\begin{aligned} \alpha_{1,2} &= \gamma_1 \alpha_1 + \gamma_2 \alpha_2 \\ \mathbf{W}^I &= \alpha_{1,2}^T \mathbf{D}^I \end{aligned} \quad (7)$$

where $\gamma_2 = 1 - \gamma_1$ and $0 \leq \gamma_i \leq 1$ ($i = \{1, 2\}$), and $\mathbf{W}^I$ is the attention generated image feature by using the weighted attention vector $\alpha_{1,2}$. γ_1 is updated during training through back-propagation.

The generated image feature $\mathbf{W}^I$ using attention still doesn't contain the image contextual features $\mathbf{c}_{con}^I$ generated earlier. In order to further contain this part of information, we applied our weighted mechanism again to these two vectors as

$$\mathbf{G}^I = \gamma_3 \mathbf{D}_{con}^I + \gamma_4 \mathbf{W}^I \quad (8)$$

where $\gamma_4 = 1 - \gamma_3$ and $0 \leq \gamma_i \leq 1$ ($i = \{3, 4\}$). γ_3 is updated during training through back-propagation. The generated vector $\mathbf{G}^I$ can be further interpreted as a projected image vector based on its understanding of the question and their contextual information.

The next step is to generate the answers based on the the question and our projected image vector $\mathbf{G}^I$.

The Weighted Contextual Feature (WCF) based Answer Generation Module

In order to generate the answer based on the given question related vectors ($\mathbf{v}_q$ and $\mathbf{v}_{con}$) and our generated projected image vector $\mathbf{G}^I$, a weighted contextual feature based convolution algorithm is designed.

$$\begin{aligned} \mathbf{v}_{q,con} &= \gamma_5 \mathbf{v}_{con} + \gamma_6 \mathbf{v}_q \\ \mathbf{A}^{q,I} &= \mathbf{G}^I * \mathbf{v}_{q,con} \end{aligned} \quad (9)$$

where $\mathbf{A}^{q,I}$ the vector generated by convolution $\mathbf{G}^I * \mathbf{v}_{q,con}$. $\gamma_5 = 1 - \gamma_6$ and $0 \leq \gamma_i \leq 1$ ($i = \{5, 6\}$), and γ_5 is updated during training through back-propagation.

The convolution operation can be further rewritten as

$$\mathbf{G}^I * \mathbf{v}_{q,con} = FFT^{-1}(FFT(\mathbf{v}_{q,con}) \cdot FFT(\mathbf{G}^I)) \quad (10)$$

$FFT(\cdot)$ stands for the Fourier transform, and FFT^{-1} is the inverse Fourier transform.

The final answer is generated by passing $\mathbf{A}^{q,I}$ through one fully connected layer and one more softmax layer, such that a one-hot vector is generated for the answer.

Remarks:

It is worth to noticing that the weighted multiple models (multi-modal) structure has been widely used in many research domains, including reinforcement learning (Narendra et al. (2015a, 2016)), system identification (Wang (2017); Narendra et al. (2014, 2015b)), image processing (Bowyer et al. (2006); Collignon et al. (1995)), NLU (Wang et al. (2018a,b)), Quesiton-Answering (Wang & Jin (2019)) and adaptive control (Murray-Smith & Johansen (1997)) etc. Some of them uses two models to compensate each other, and the others use more than two model to boost the performances using collective information. It has shown extremely good results on these applications using these various multiple models structures.

Experiment

In this section, we evaluate our new model's performance on the several visual question understanding datasets

Datasets

MSCOCO VQA Datasets: This dataset contains a total of more than 200K images, in which 82,783 images are for training, 40,504 images are for validation, and another 81,434 images for testing. There are 3 questions per image and 10 answers per question. There is another 25% of the test dataset are used as the test-dev data. Currently there are two versions of VQA datasets (v1 Antol et al. (2015) and v2 Goyal et al. (2017)), with same number of images but different number of questions, and we use VQA v2 on which to perform our experiment.

We report our evaluation results using both the test-standard dataset and the test-dev dataset. Also, the results are evaluated on the open-ended (VQA v2) tasks. The model is trained on the train and validation set, and compared the results with the current state-of-the art models in each category and the whole dataset.



Figure 3: Comparison examples between A-WCF and one of the state-of-the-art models

Experiment Setup

Different Models

We test our attention based WCF (A-WCF) model with several different setups:

A-WCF model without question contextual features: In this model setup, we remove the question contextual features v_{con} in Figure 2, hence the weight parameters γ_1 , γ_2 , γ_3 and γ_4 are removed from the system during training. The entire model only contains one set of weighted contextual features, *i.e.* the image contextual features D_{con}^I .

A-WCF model without image contextual features: The second model setup is by removing the image contextual features from the A-WCF model which eliminate two weight parameters γ_3 and γ_4 .

A-WCF model with both question and image contextual features: This is the original model as shown in Figure 2.

Architecture Parameters

The image features are extracted from a 152-layer Residual Network He et al. (2016), which is pretrained on the ImageNet dataset Deng et al. (2009). The features extracted before the last classifier layer ("pool5") are used as the image feature inputs to the MDLSTM classifier. Following the model setup in Fukui et al. (2016), the image feature's dimension generated is $I \in \mathbb{R}^{2048 \times 14 \times 14}$. The dimension of MDLSTM encoder-decoder is set as 2, and the number of units is also chosen as 2048 to keep the dimension consistent. On the question side, a given question is firstly tokenized into words, and 100-D word vectors are generated from GloVe word2vec representations Pennington et al. (2014). Then these word vectors are fed into the BLSTM structure with 2048 number of units. The decoder RNN has the same number of units such that $v_q \in \mathbb{R}^{2048 \times 1}$. The other vectors' dimension information can be found directly in Figure 2.

The Adam stochastic optimizer is used with a learning rate 0.002 and also early-stopping is also applied if there is no improvement for the last 30 epochs.

Experiments

A-WCF: An Interpretable Structure

In this subsection, we would like to give some visual examples showing how A-WCF structure is interpreted by taking advantage of the image and contextual features. Figure 3 gives two examples using three different A-WCF setups. In the first example, there is a post-position "kicking the ball" for "the person", where a question level contextual feature should be used to understand the question correctly. Otherwise, a model may not be able to focus on the correct sub-regions on image since it doesn't fully understand the question. In the third setup, by removing the image level contextual feature, the model now can locate the correct sub-regions on image, however, it still cannot generate the correct answers since contextual correlations between the masked sub-regions cannot be fully understood, so wrong answers are generated in the case when there is not image context features.

Similarly, in the second example, the question-level semantic context feature helps locate the correct sub-region(s) on the image, where the "type of material" should be identified. Then the image context features will help to generate the correct answer by filtering out some noisy image information (like 5 baseball bats) by giving the a lower weight, hence generate the correct answer.

From the above two examples, we can see that the question-level semantic context features help the system locate the correct question-related sub-regions (like masking) on the image, and the image level context feature mainly help the system to generate correct answers by assigning appropriate weights to the extracted sub-regions.

In the next section, more quantitative results will be given using our model, and compared with some base models on VQA datasets.

Experiments on VQA datasets

Our model with three different set-ups are trained on the train+validation MSCOCO VQA v2 dataset, and further compared with the state-of-the-art models. The results are shown in Table 1.

One observation for Table 1 is that the model without question contextual features has far inferior performance than the one without image contextual features and the model with both contextual features. It demonstrates the importance of question contextual features to our system. The model A-WCF with both image and question contextual features outperforms the previous state-of-the-art results on most of the categories in VQA v2. It is also excited to show that our model even gives a better performance than the ensemble/stacking based models.

Table 1: Results on Open-ended Answers on Test-dev and Test-standard in VQA v2

VQA v2	Test-dev (%)				Test-standard (%)			
	Y/N	Num.	Other	All Categories	Y/N	Num.	Other	All Categories
MFH Ensemble Yu et al. (2018)	NA	NA	NA	NA	86.65	51.13	61.75	70.92
Yan Model Zhang et al. (2018)	83.14	51.62	58.97	68.09	83.56	51.39	59.11	68.41
FAIR A-STAR	87.82	51.54	63.41	72.12	87.82	51.59	63.43	72.25
A-WCF w/o Question Context	83.26	50.32	59.84	69.63	83.38	50.12	59.82	69.56
A-WCF w/o Image Context	85.24	51.25	61.22	70.98	85.63	51.26	61.25	71.02
A-WCF	88.93	51.48	61.98	72.86	88.98	51.65	62.62	72.89

Application to Conversational VQA tasks

Due to the interpretability of the model, it is possible for us to take advantage of its properties to get even better VQA results based on a conversation between user and the system. For example, when the VQA system is not sure about the correctness of its answer, it can ask the user to clarify those parts in his question where our system generates higher attention weights, or ask the user whether specific regions in an image with high attention weights should be highlighted. One conversational VQA example using our model is illustrated in Figure 4.

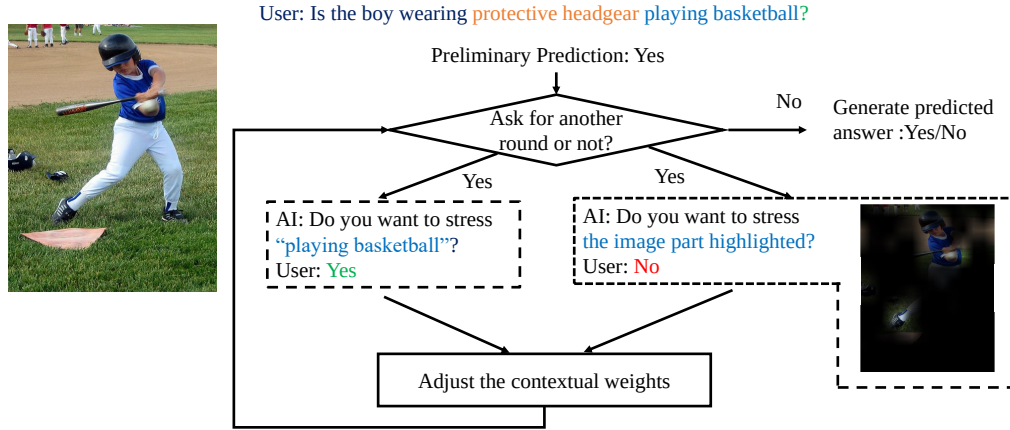


Figure 4: An example of conversational QA using an A-WCF model

Following is an analysis of the example in Figure 4:

1. As shown in the example, in the first round, the result image attention map highlights the protective headgear, but the query attention gives a higher attention weights on phrase *playingbasketball*. There is a conflict between the stress points of the image and query.
2. Assume that the difference between the weight of the query feature (γ_5) and that of the image feature (γ_6) is larger than a threshold T_r , then trigger the next step to ask for user another round to clarify. There are two ways for asking questions: one is to ask question on query and the other is to ask question on image. We will choose the one has a larger weight to ask.
3. If the user says “yes”, then raise the value of contextual image/query weight, *i.e.* $\gamma_5 \setminus \gamma_6$, and the other weight is equal to $1 - (\gamma_5 \setminus \gamma_6)$.
4. Do steps 1-3 until there is no conflict between between the stress points of the image and query, and also the difference between the weights of query and image features are within our threshold T_r , then use the answer at that step as our final answer.

This is one of the examples showing that how to effectively use our attention weights to build conversations between our system and users, such that a better VQA results can be obtained. There are many other approaches to use these weights on conversational VQA tasks, which are currently under investigation.

Conclusion

In this paper, we proposed a novel interpretable attention based weighted contextual feature model to address the visual question answering problem. By using the contextual features with adaptive weights from both questions and images, it gives us the advantage to pinpoint the most important part of questions and images, and de-emphasize the less important features. We generate the new state-of-the-art results on the MSCOCO VQA v2 dataset for the open-ended answers task. As a relative general technique, the weighted contextual feature can be further expanded to other text or image related tasks, like question-answering, summerization or visual grounding. We also show a conversational VQA example by leveraging the interpretability of our structure, which demonstrates that it may have even greater potentials on applying to more complex conversational VQA tasks.

References

- Antol, Stanislaw, Agrawal, Aishwarya, Lu, Jiasen, Mitchell, Margaret, Batra, Dhruv, Zitnick, C. Lawrence, and Parikh, Devi. Vqa: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.
- Bowyer, Kevin W, Chang, Kyong, and Flynn, Patrick. A survey of approaches and challenges in 3d and multi-modal 3d+ 2d face recognition. *Computer vision and image understanding*, 101(1):1–15, 2006.

- Collignon, André, Maes, Frederik, Delaere, Dominique, Vandermeulen, Dirk, Suetens, Paul, and Marchal, Guy. Automated multi-modality image registration based on information theory. In *Information processing in medical imaging*, volume 3, pp. 263–274, 1995.
- Deng, Jia, Dong, Wei, Socher, Richard, Li, Li-Jia, Li, Kai, and Fei-Fei, Li. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. IEEE, 2009.
- Frome, Andrea, Corrado, Greg S, Shlens, Jon, Bengio, Samy, Dean, Jeff, Mikolov, Tomas, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pp. 2121–2129, 2013.
- Fukui, Akira, Park, Dong Huk, Yang, Daylen, Rohrbach, Anna, Darrell, Trevor, and Rohrbach, Marcus. Multi-modal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 457–468, 2016.
- Gong, Yunchao, Wang, Liwei, Hodosh, Micah, Hockenmaier, Julia, and Lazebnik, Svetlana. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*, pp. 529–545. Springer, 2014.
- Goyal, Yash, Khot, Tejas, Summers-Stay, Douglas, Batra, Dhruv, and Parikh, Devi. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Graves, Alex and Schmidhuber, Jürgen. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.
- Graves, Alex and Schmidhuber, Jürgen. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in neural information processing systems*, pp. 545–552, 2009.
- Harris, Zellig S. Distributional structure. *Word*, 10(2-3):146–162, 1954.
- He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- Lu, Jiasen, Yang, Jianwei, Batra, Dhruv, and Parikh, Devi. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pp. 289–297, 2016.
- Malinowski, Mateusz, Rohrbach, Marcus, and Fritz, Mario. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9. IEEE Computer Society, 2015.
- Murray-Smith, Roderick and Johansen, T. *Multiple model approaches to nonlinear modelling and control*. CRC press, 1997.
- Narendra, Kumpati S, Wang, Yu, and Chen, Wei. Stability, robustness, and performance issues in second level adaptation. In *American Control Conference (ACC)*, 2014, pp. 2377–2382. IEEE, 2014.
- Narendra, Kumpati S, Mukhopadhyay, Snehasis, and Wang, Yu. Improving the speed of response of learning algorithms using multiple models. *arXiv preprint arXiv:1510.05034*, 2015a.
- Narendra, Kumpati S, Wang, Yu, and Chen, Wei. Extension of second level adaptation using multiple models to siso systems. In *American Control Conference (ACC)*, 2015, pp. 171–176. IEEE, 2015b.
- Narendra, Kumpati S, Wang, Yu, and Mukhopadhyay, Snehasis. Fast reinforcement learning using multiple models. In *Decision and Control (CDC), 2016 IEEE 55th Conference on*, pp. 7183–7188. IEEE, 2016.
- Pennington, Jeffrey, Socher, Richard, and Manning, Christopher. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.
- Rajani, Nazneen Fatema and Mooney, Raymond J. Stacking with auxiliary features. *arXiv preprint arXiv:1605.08764*, 2016.
- Shih, Kevin J, Singh, Saurabh, and Hoiem, Derek. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4613–4621, 2016.
- Teney, Damien, Anderson, Peter, He, Xiaodong, and Hengel, Anton van den. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017.
- Wang, Yu. A new concept using lstm neural networks for dynamic system identification. In *American Control Conference (ACC)*, 2017, pp. 5324–5329. IEEE, 2017.
- Wang, Yu and Jin, Hongxia. A deep reinforcement learning based multi-step coarse to fine question answering (mscqa) system. In *AAAI, 2019*. AAAI, 2019.
- Wang, Yu, Patel, Abhishek, and Jin, Hongxia. A new concept of deep reinforcement learning based augmented general tagging system. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1683–1693. Association for Computational Linguistics, 2018a. URL <http://aclweb.org/anthology/C18-1143>.
- Wang, Yu, Shen, Yilin, and Jin, Hongxia. A bi-model based rnn semantic frame parsing model for intent detection and slot filling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2*, volume 2, pp. 309–

314, 2018b.

Yang, Zichao, He, Xiaodong, Gao, Jianfeng, Deng, Li, and Smola, Alex. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 21–29, 2016.

Yu, Zhou, Yu, Jun, Xiang, Chenchao, Fan, Jianping, and Tao, Dacheng. Beyond bilinear: Generalized multi-modal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

Zhang, Yan, Hare, Jonathon, and Prügel-Bennett, Adam. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018.