Supplementary Material for "Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering"

Duy-Kien Nguyen[1] and Takayuki Okatani[1,2]

[1]Graduate School of Information Sciences, Tohoku University
[2]RIKEN Center for Advanced Intelligence Project
{kien, okatani}@vision.is.tohoku.ac.jp

This document contains: more details of our experimental setups (Sec.A), the evaluation of effects in the employment of Contextualized Word Vectors (Sec.B), more visualization of attention maps generated in the answer prediction layer including failure cases (Sec.C), and an analysis of the attention mechanism employed in the image feature extraction (Sec.D).

A. More Details of the Experimental Setups

In our experiments, images and questions are preprocessed as follows. All the images were resized to 448×448 before feeding into the CNN. All the questions were tokenized using Python Natural Language Toolkit (nltk) [2]. We used the vocabulary provided by the CommonCrawl-840B Glove model for English word vectors [11], and set out-of-vocabulary words to *unk*. As mentioned in the main paper, we chose the correct answer appearing more than 5 times (= 3,014 answers) for VQA 1.0, and 8 times (= 3,113 answers) for VQA 2.0 as in [12]. We capped the maximum length of questions at 14 words and then performed dynamic unrolling for each question to allow for questions of different lengths.

Throughout the experiments, we used three-layer DCNs, that is, DCNs with three dense co-attention layers (L = 3). This number of layers were chosen based on our preliminary experiments. The Bi-LSTM was initialized following the recommendation in [5] and all the other parameters were initialized as suggested by Glorot *et al*. [4]. In the training procedure, the ADAM [8] optimizer was used to train our model for 16 and 21 epochs on VQA and VQA 2.0 with batch size of 160 and 320, respectively; weight decay with rate of 0.0001 was added. We used exponential decay to gradually decrease the learning rate as

$$\alpha_{step} = 0.5^{\frac{\text{epochs}}{\text{decay epochs}}}\alpha,$$

where the initial learning rate α was set to $\alpha = 0.001$, and the decay epochs was set to 4 and 7 epochs for VQA and VQA 2.0 in turn; we set $\beta_1 = 0.9$, and $\beta_2 = 0.99$.

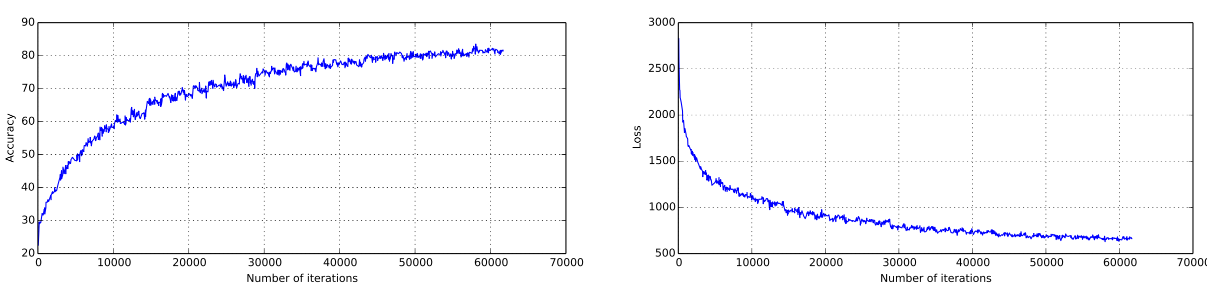


Figure 5: Learning curves for DCN.

B. Effects of the Employment of Contextualized Word Vectors

To extract word features from input questions, some of the previous studies [6, 7, 3, 1] employed pretrained RNNs (specifically, GRU networks pre-trained with Skip-thought) [9]. In this study, we initially pursued a similar approach; we perform fine-tuning of a pretrained LSTM, specifically a two-layer Bi-LSTM trained as a CoVe (Context Vector) encoder [10]. Conducting comparative experiments, we eventually employ a single-layer Bi-LSTM with random initialization, as explained in the main paper. We report here the results of the experiments.

Table 5 shows the performances of DCNs with the CoVe-pretrained Bi-LSTM and with the randomly initialized Bi-LSTM. Note that the former is a two-layer model and the later has only one layer. Here, the VQA 2.0 test-dev dataset was used. It is observed that for DCNs with the answer prediction layer of (16), the one with the CoVe-pretrained model performs slightly better than the one with the randomly initialized model, but their differences are small. For DCNs with the answer prediction layers of (17) and (18), the one with the randomly initialized model performs better with a less number of parameters.

It should be noted, however, that the employment of CoVe-pretrained models, together with the answer prediction layer of (16), enables to compute meaningful answer representation (s_A) for answers that have not been seen before, i.e., those that are not included in training data. Table 6 shows the results of DCN (16) with the CoVe-pretrained model for *Multiple Choice* answers, which include a lot of unseen answers. This is not the case with DCNs (17) and (18) that compute scores of a fixed set of predetermined answers— the common approach of most of the recent studies.

Table 5: Performances of DCNs with the CoVe-pretrained LSTM and with the randomly initialized LSTM on the VQA 2.0 test-dev set.

Model	Overall	Other	Number	Yes/No	No. params
DCN (16) + CoVe	67.06	57.44	46.91	83.69	31M
DCN (16)	66.87	57.26	46.61	83.51	28M
DCN (17) + CoVe	66.21	56.71	46.01	82.72	34M
DCN (17)	66.72	56.77	46.65	83.70	31M
DCN (18) + CoVe	66.31	56.62	45.78	83.14	35M
DCN (18)	66.60	56.72	46.60	83.50	32M

Table 6: Effectiveness of DCN (16) + CoVe-pretrained LSTM on *Multiple Choice* answers.

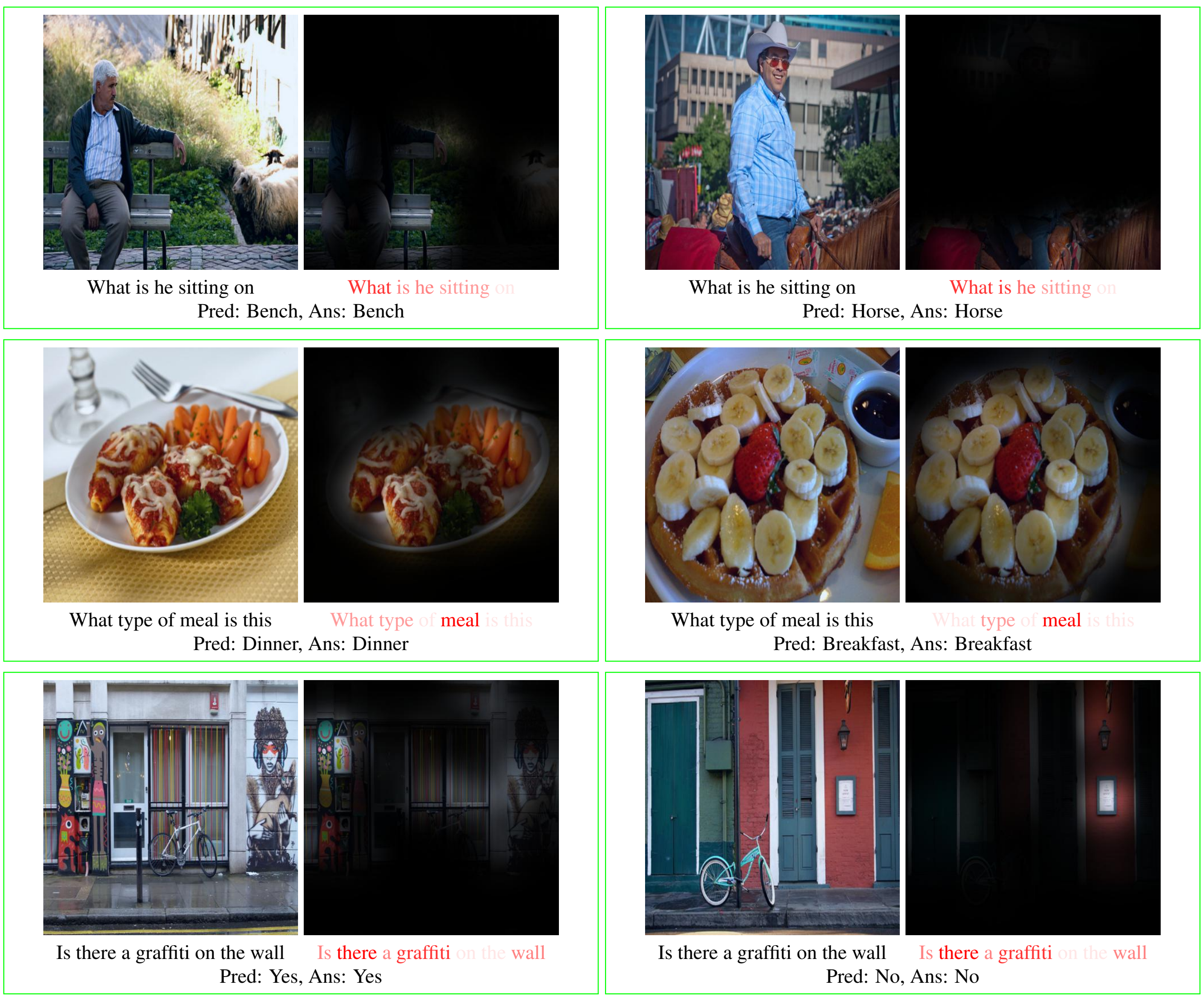
Model	Test-dev				Test-std			
	Overall	Other	Number	Yes/No	Overall	Other	Number	Yes/No
DCN (16) + CoVe	71.37	66.10	45.48	84.39	71.20	65.93	44.13	84.23

C. Visualization of Attention Maps in the Answer Prediction Layer

We have shown a few examples of attention maps generated in the answer prediction layer of our DCNs in Fig.4 of the main paper. We show here more examples for success cases (Sec.C.1) and also for failure cases (Sec.C.2).

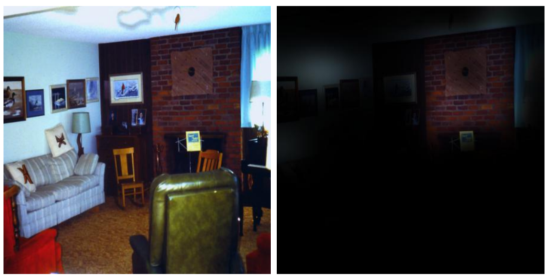
C.1. Success Cases

We consider the visualization of complementary pairs to analyze the behaviour of our DCNs. Each row shows a complementary pair having the same question and different images. It can be seen from the examples shown below that the image and question attention maps are generated appropriately for most of success cases.

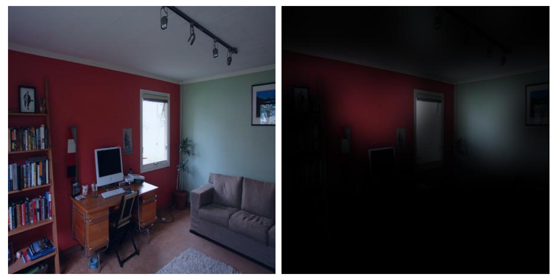


How many vases are in the photo How many vases are in the photo
Pred: 2, Ans: 2

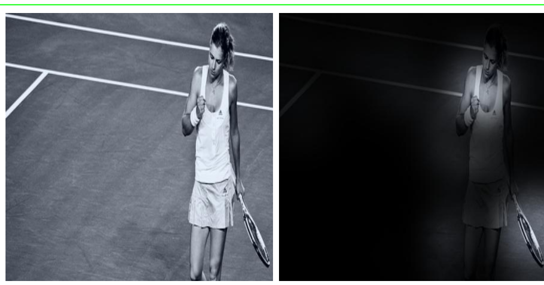
How many vases are in the photo How many vases are in the photo
Pred: 1, Ans: 1



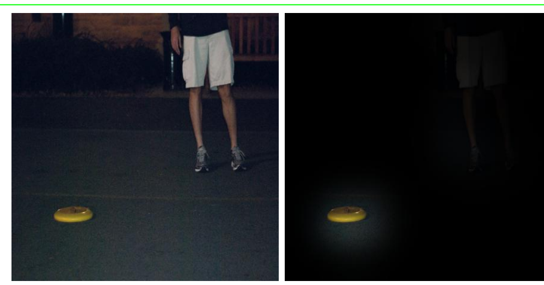
What is the darker wall made of What is the darker wall made of
Pred: Brick, Ans: Brick



What is the darker wall made of What is the darker wall made of
Pred: Drywall, Ans: Drywall

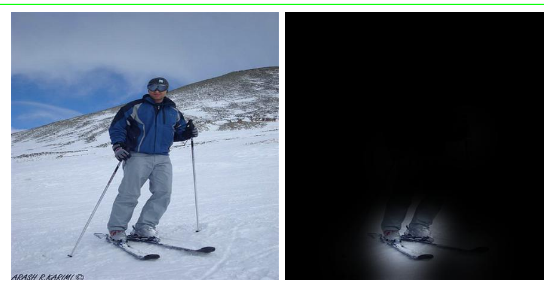


What sport is this woman playing What sport is this woman playing
Pred: Tennis, Ans: Tennis

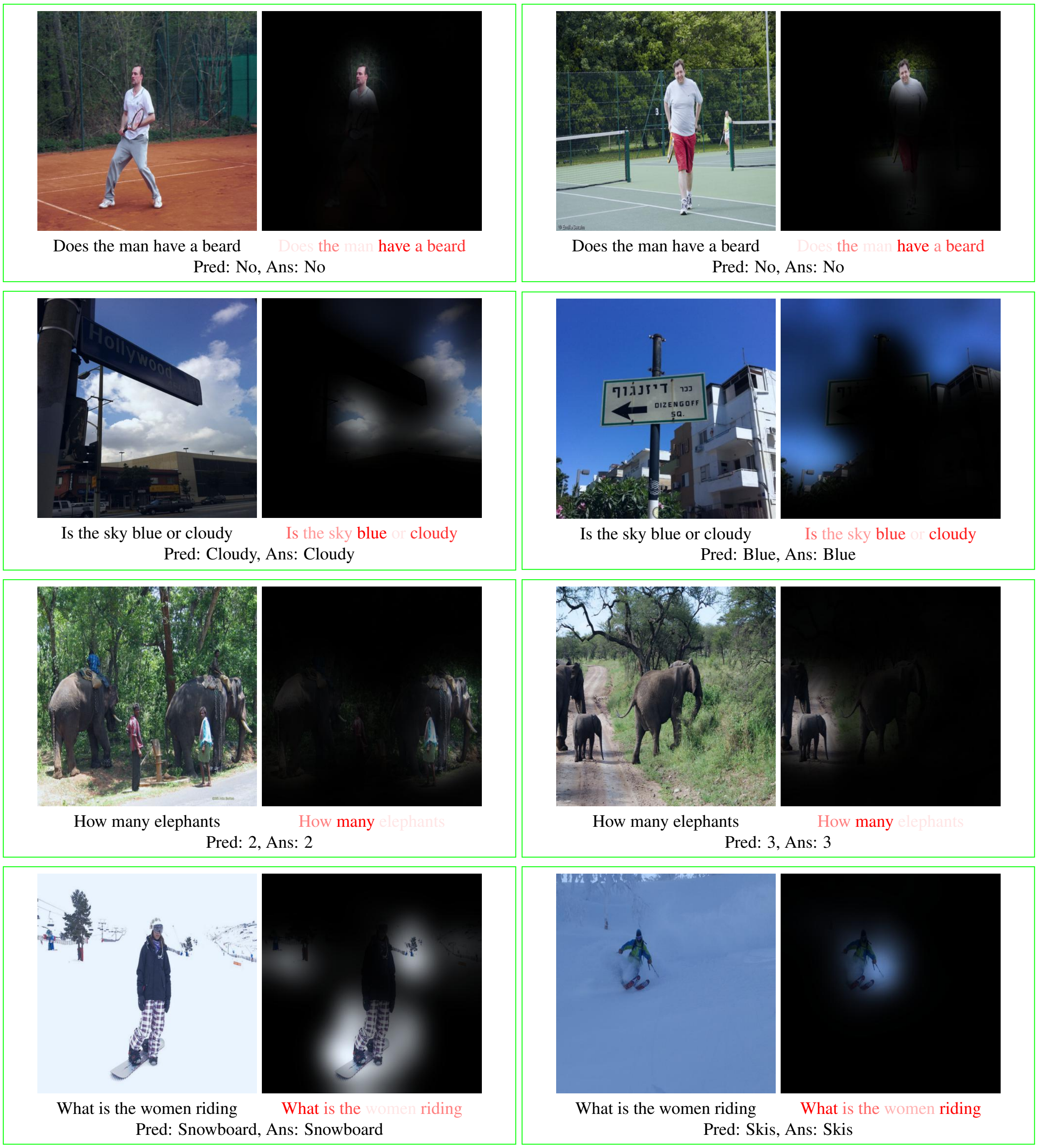


What sport is this woman playing What sport is this woman playing
Pred: Frisbee, Ans: Frisbee

What color are the skiers shoes What color are the skiers shoes
Pred: Yellow, Ans: Yellow



What color are the skiers shoes What color are the skiers shoes
Pred: White, Ans: White


Does the man have a beard
Does the man have a beard
Pred: No, Ans: No
Does the man have a beard
Does the man have a beard
Pred: No, Ans: No
Is the sky blue or cloudy
Is the sky blue or cloudy
Pred: Cloudy, Ans: Cloudy
Is the sky blue or cloudy
Is the sky blue or cloudy
Pred: Blue, Ans: Blue
How many elephants
How many elephants
Pred: 2, Ans: 2
How many elephants
How many elephants
Pred: 3, Ans: 3
What is the women riding
What is the women riding
Pred: Snowboard, Ans: Snowboard
What is the women riding
What is the women riding
Pred: Skis, Ans: Skis

C.2. Failure Cases

According to our analysis, failure cases can be categorized into the following four types:

- *Type-1* Although the DCN is able to locate appropriate image regions and words, it fails to distinguish two different objects or concepts that have similar appearance. This may be attributable to that the extracted image features are not rich enough to distinguish them (e.g. *mutt* and *lab*; and *spoon* and *fork*).

- *Type-2* Although the DCN is able to locate appropriate image regions and words, it fails to yield correct answers due to the bias of the dataset or missing instances of some objects/concepts in the dataset. For example, there are many samples of an *american flag* but no sample of a *dragon flag* in the training set.

- *Type-3* The DCN fails to locate appropriate image regions. This tends to occur when some image regions have similar appearance to the region that the DCN should attend, or the region that it should attend is too small.

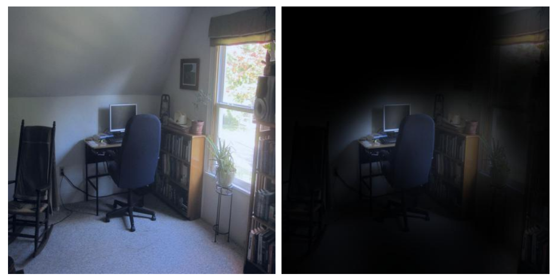
- *Type-4* Although the DCN does yield conceptually correct answers, they are not listed in the given set of answers in the dataset and thus judged incorrect. For instance, while the given correct answer is *water*, the DCN outputs *beach*, which should also be correct, as in one of the examples below.

As in the above success cases, each row shows a complementary pair having the same question and different images. In each row, at least either one of the two has an erroneous prediction. The red bounding boxes indicate erroneous answers and the green ones indicate correct answers. The numbers in the failure examples indicate the error types we categorize above.

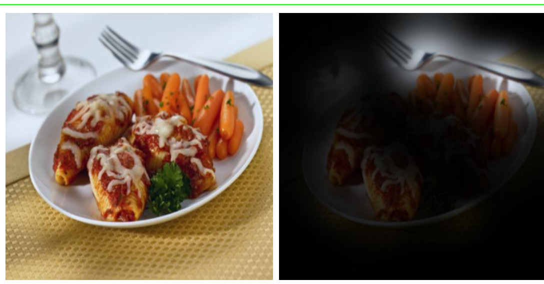
What breed of dog is this What breed of dog is this
Pred: Mutt, Ans: Lab *(Error type: 1)*

What breed of dog is this What breed of dog is this
Pred: Terrier, Ans: Terrier

What room is this What room is this
Pred: Bedroom, Ans: Bedroom



What room is this What room is this
Pred: Living room, Ans: Office *(Error type: 1)*

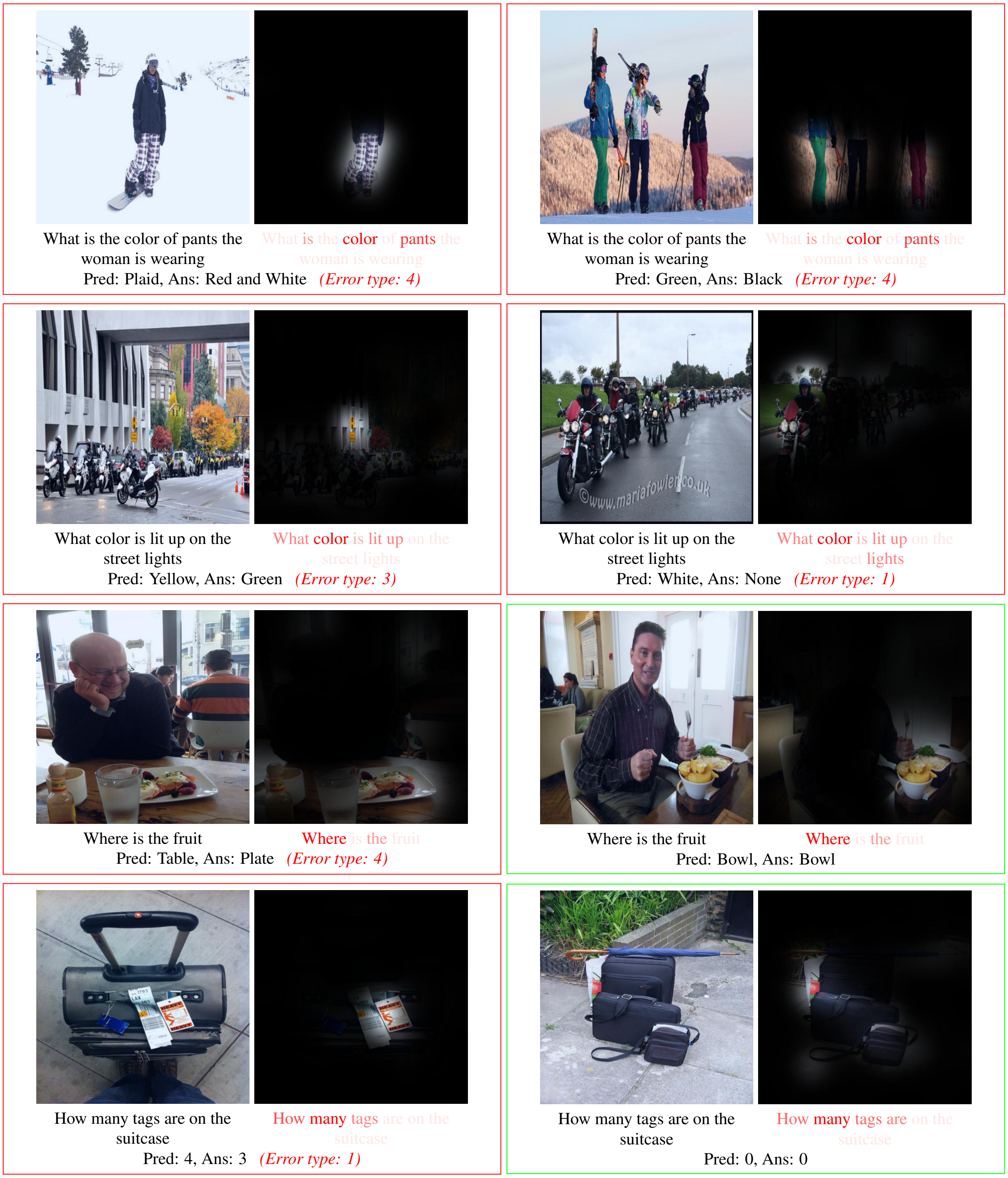


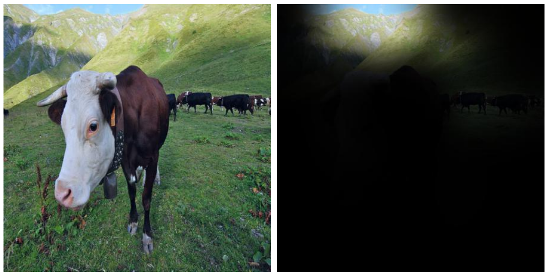
What is the name of the utensil What is the name of the utensil
Pred: Fork, Ans: Fork

What is the name of the utensil What is the name of the utensil
Pred: Fork, Ans: Spoon *(Error type: 1)*

How tall is he How tall is he
Pred: 5 feet, Ans: Tall *(Error type: 1)*

How tall is he How tall is he
Pred: 5 feet, Ans: 6 feet *(Error type: 2)*


What is the color of pants the woman is wearing
What is the color of pants the woman is wearing
Pred: Plaid, Ans: Red and White (Error type: 4)
What is the color of pants the woman is wearing
What is the color of pants the woman is wearing
Pred: Green, Ans: Black (Error type: 4)
What color is lit up on the street lights
What color is lit up on the street lights
Pred: Yellow, Ans: Green (Error type: 3)
©www.mariafowler.co.uk
What color is lit up on the street lights
What color is lit up on the street lights
Pred: White, Ans: None (Error type: 1)
Where is the fruit
Where is the fruit
Pred: Table, Ans: Plate (Error type: 4)
Where is the fruit
Where is the fruit
Pred: Bowl, Ans: Bowl
How many tags are on the suitcase
How many tags are on the suitcase
Pred: 4, Ans: 3 (Error type: 1)
How many tags are on the suitcase
How many tags are on the suitcase
Pred: 0, Ans: 0



What landforms are behind the cows | What landforms are behind the cows

Pred: Mountains, Ans: Mountains

What landforms are behind the cows | What landforms are behind the cows

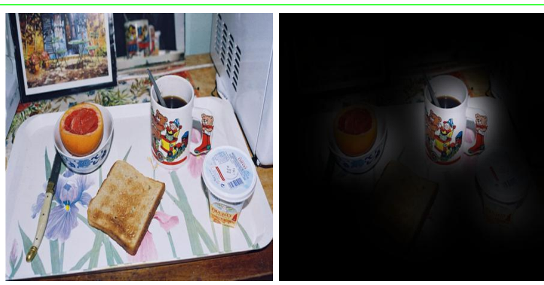
Pred: Beach, Ans: Water *(Error type: 4)*

What flag is that | What flag is that

Pred: American, Ans: American

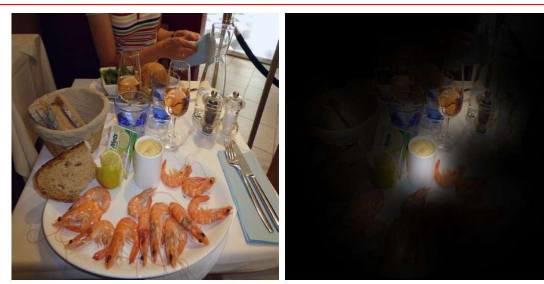
What flag is that | What flag is that

Pred: American, Ans: Dragon *(Error type: 2)*



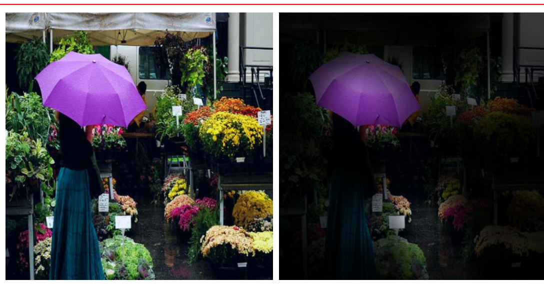
What is in the mug | What is in the mug

Pred: Coffee, Ans: Coffee



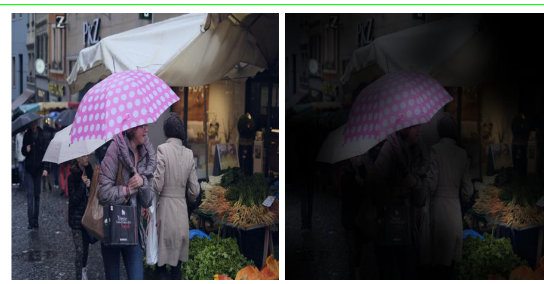
What is in the mug | What is in the mug

Pred: Wine, Ans: Butter *(Error type: 1)*



Where is this woman at | Where is this woman at

Pred: Outside, Ans: Farmers market *(Error type: 4)*



Where is this woman at | Where is this woman at

Pred: Market, Ans: Market

D. Layer Attention in the Image Feature Extraction Step

As explained in the main paper (Sec.3.1), our DCN extracts visual features from an input image using a pre-trained ResNet at the initial step. The features are obtained by computing the weighted sum of the activations (i.e., outputs) of the four convolutional layers of the ResNet, where the attention weights generated conditioned on the input question are used. We examine here how this attention mechanism works for different types of questions. Specifically, utilizing the fifty five question types provided in the VQA-2.0, we compute the mean and standard deviation of the four attention weights for the questions belonging to each question type. We used all the questions in the validation set and our DCN trained only on train set for this computation.

Figure 6 shows the results. The bars in four colors represent the means of the four layer weights for each question type, and the thin black bars attached to the color bars represent their standard deviations. The fifty five question types are ordered by their similarity in the horizontal axis. From the plot, we can make the following observations:

- Layer 1 (the lowest one) has a certain level of weights only for *Yes/No* questions (shown on about the left half of the plots) and no weight for other types of questions (on the right half);
- Layer 2 has a small weight only for *Yes/No* questions and no weight for other types of questions;
- Layer 3 tends to have large weights for questions about colors (e.g., *"what color"*) and questions about presence of a given object(s) (e.g., *"are there"* and *"how many"*);
- Layer 4 (the highest one) has the largest attention weights in most of the question types, indicating its importance in answering them.
- Specific questions, such as *"what color "* and *"what sport is"*, tend to have smaller standard deviations than nonspecific questions, such as *"is the woman"* and *"do you"*.

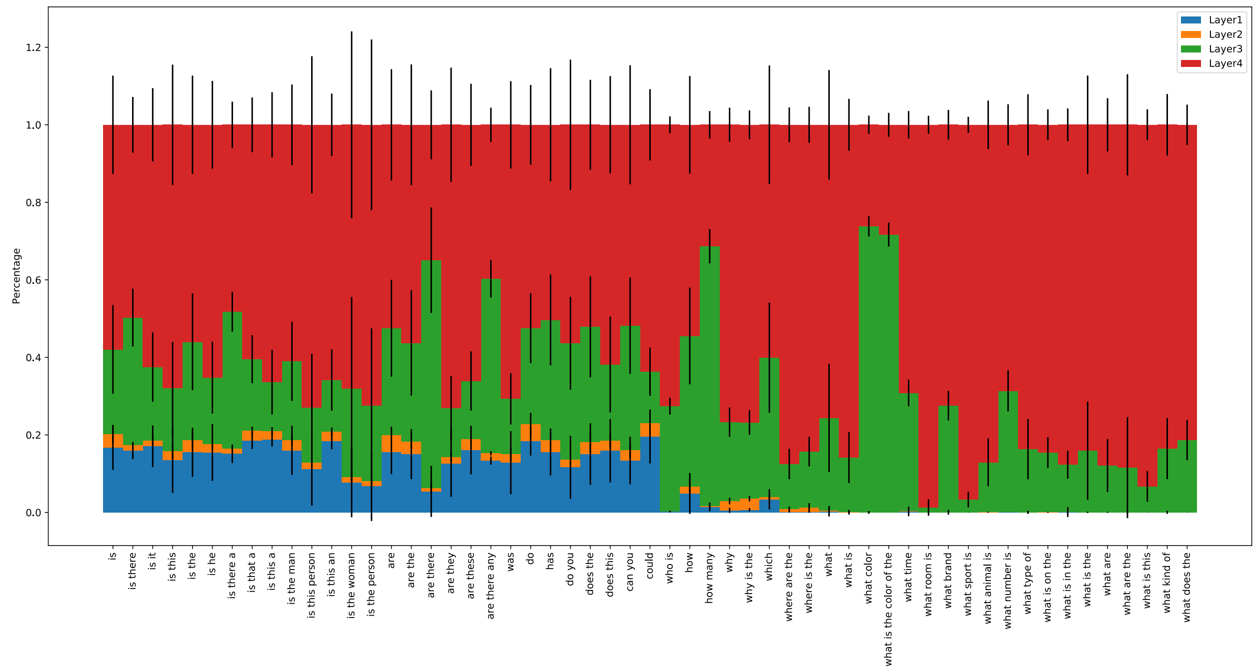


Figure 6: Statistics (means and standard deviations) of the attention weights on the four convolutional layers generated in the image feature extraction step for different types of questions.

References

[1] H. Ben-younes, R. Cadène, M. Cord, and N. Thome. MUTAN: multimodal tucker fusion for visual question answering. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[2] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009. 1

[3] Z. Chen, Z. Yanpeng, H. Shuaiyi, T. Kewei, and M. Yi. Structured attentions for visual question answering. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[4] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010. 1

[5] R. Jozefowicz, W. Zaremba, and I. Sutskever. An empirical exploration of recurrent network architectures. *Journal of Machine Learning Research*, 2015. 1

[6] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal Residual Learning for Visual QA. In *International Conference on Neural Information Processing Systems (NIPS)*, 2016. 2

[7] J.-H. Kim, K.-W. On, W. Lim, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 1

[9] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. In *International Conference on Neural Information Processing Systems (NIPS)*, 2015. 2

[10] B. McCann, J. Bradbury, C. Xiong, and R. Socher. Learned in translation: Contextualized word vectors. *arXiv preprint arXiv:1708.00107*, 2017. 2

[11] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 1

[12] D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. *arXiv preprint arXiv:1708.02711*, 2017. 1