Supplementary Material: High-Order Attention Models for Visual Question Answering

1 Qualitative results

In Fig. 1 we show additional attention heat maps. For the picture provided in the first column we demonstrate the difference in the obtained attention for two different questions in columns 2-4 and columns 5-7 respectively. Intermediate attention results for the unary and pairwise modalities are provided in columns 2 & 3, and columns 5 & 6 for the two different questions. Column 4 and column 7 depict the final attention result of our mechanism. The unary attention usually attends to strong features of the image. The pairwise potential is created from the correlation of image features with the textual feature. While in many cases the pairwise potential is accurate, it is clear that the combined attention is more accurate in most cases.

In Fig. 4 we compare results of different approaches that use attention for the VQA task. These additional images provide insights regarding the importance of attention to generate the correct answer. In the following we discuss the results illustrated in the left column more carefully, while noting that the results demonstrated in the right column are similar. Left column, first row, the approach of [2] answer "red," while the answer is obviously black. The reason for the answers could be explained by the attention modules. Left column, second row, we observe that attention not focused on the skiers results in a wrong answer for one of the approaches, while ours predicts the correct answer.

We want to also further explain on the process of generating these attention maps. For comparable reasons we used two modalities architecture, which described in Fig. 2. Lu *et al*. [2] generates attention maps for each textual representation, word, phrase, and question. In this analysis, the attention maps are taken from the question (last) representation, because they indicate the most meaningful attention in this model. Fukui *et al*. [1] uses two attention maps, also known as glimpses. As in the demo of [1][1], we are showing the first attention map for their model, because it is visually more plausible.

2 Ternary Potentials Implementation

We further describe our implementations of C_3 tensor. In a sense C_3 is constructed by multiplication of three metrics. Support for high order operations is very limited in deep learning architectures. Our solution is based on existing building blocks, using a dynamic neural networks that depends on the spatial size of the input metrics. First, we split the question tensor over the spatial dimension, Assuming $n_q < n_a < n_v$. For each spatial location $q_i \in \mathbb{R}^d$ we perform a element wise multiplication with each spatial location of A, we then perform a matrix multiplication operation with V the result equals to $(C_3)_{1,:} \in \mathbb{R}^{1\times n_v \times n_a}$. Last, we concatenate the n_q outputs on the first dimension to achieve $C_3 \in \mathbb{R}^{n_q \times n_v \times n_a}$. This process is illustrated in Fig. 3.

References

[1] 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. *arXiv preprint*

---

[1] http://demo.berkeleyvision.org

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

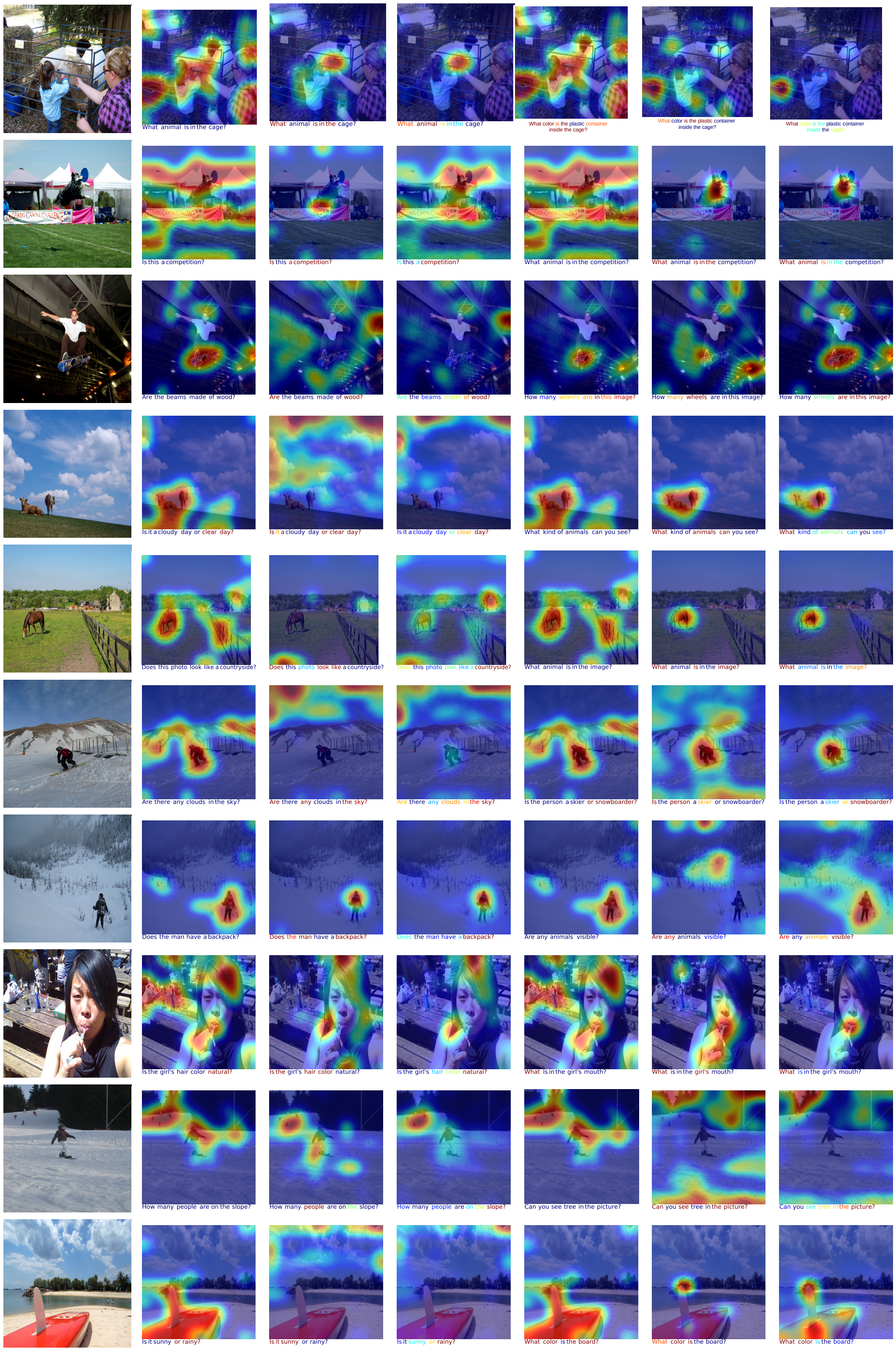

What animal is in the cage?
What animal is in the cage?
What animal is in the cage?
What color is the plastic container inside the cage?
What color is the plastic container inside the cage?
What color is the plastic container inside the cage?
Is this a competition?
Is this a competition?
Is this a competition?
What animal is in the competition?
What animal is in the competition?
What animal is in the competition?
Are the beams made of wood?
Are the beams made of wood?
Are the beams made of wood?
How many wheels are in this image?
How many wheels are in this image?
How many wheels are in this image?
Is it a cloudy day or clear day?
Is it a cloudy day or clear day?
Is it a cloudy day or clear day?
What kind of animals can you see?
What kind of animals can you see?
What kind of animals can you see?
Does this photo look like a countryside?
Does this photo look like a countryside?
Does this photo look like a countryside?
What animal is in the image?
What animal is in the image?
What animal is in the image?
Are there any clouds in the sky?
Are there any clouds in the sky?
Are there any clouds in the sky?
Is the person a skier or snowboarder?
Is the person a skier or snowboarder?
Is the person a skier or snowboarder?
Does the man have a backpack?
Does the man have a backpack?
Does the man have a backpack?
Are any animals visible?
Are any animals visible?
Are any animals visible?
Is the girl's hair color natural?
Is the girl's hair color natural?
Is the girl's hair color natural?
What is in the girl's mouth?
What is in the girl's mouth?
What is in the girl's mouth?
How many people are on the slope?
How many people are on the slope?
How many people are on the slope?
Can you see tree in the picture?
Can you see tree in the picture?
Can you see tree in the picture?
Is it sunny or rainy?
Is it sunny or rainy?
Is it sunny or rainy?
What color is the board?
What color is the board?
What color is the board?

Figure 1: For each image (1^{st} column) we show the attention generated for two different questions in columns 2-4 and columns 5-7 respectively. The attentions are ordered as unary attention, pairwise attention and combined attention for both the image and the question. We observe the combined attention to significantly depend on the question.

arXiv:1606.01847, 2016.

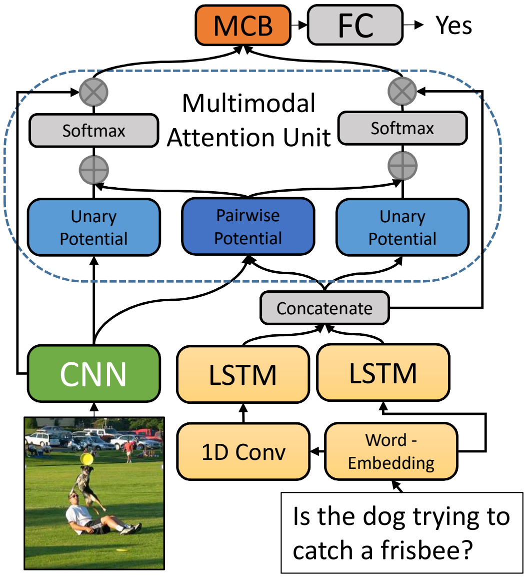


Figure 2: Overview of our attention based mechanism for two modalities.

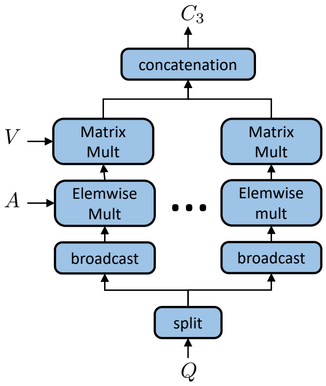
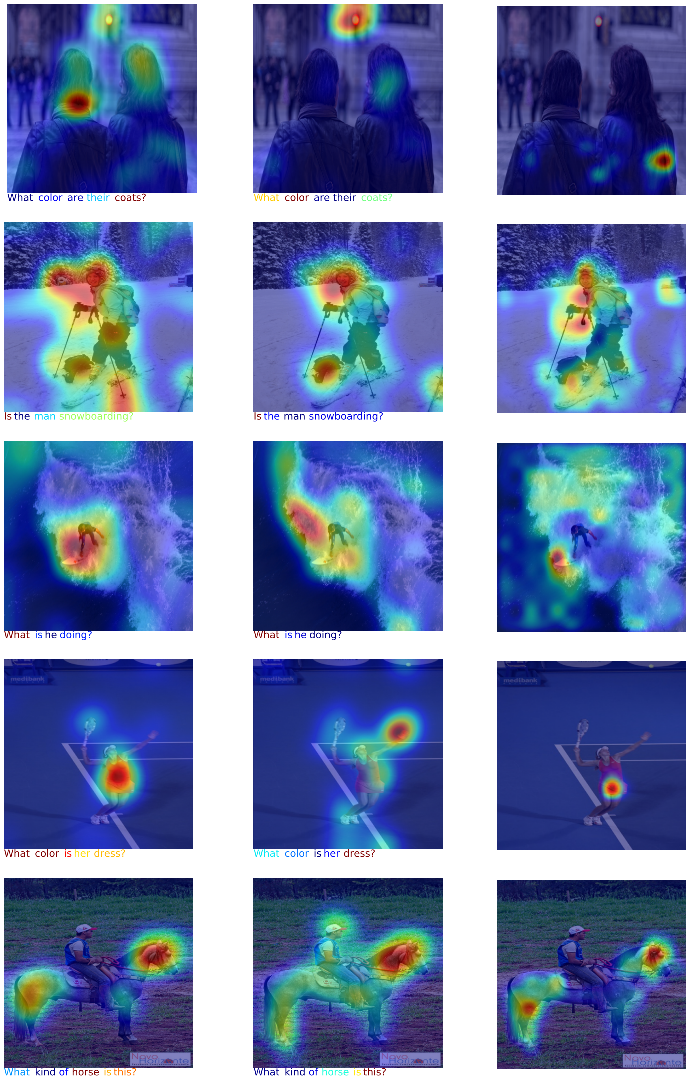


Figure 3: Illustration of our architecture to produce C_3

[2] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.



What color are their coats?
Ours: black
[2]: red
[1]: black
GT: black

Is the man snowboarding?
Ours: no
[2]: yes
[1]: no
GT: no

What is he doing?
Ours: surfing
[2]: yes
[1]: surfing
GT: surfing

What color is her dress?
Ours: red
[2]: white
[1]: red
GT: red

What kind of horse is this?
Ours: gray
[2]: gray
[1]: clydesdale
GT: gray

Figure 4: Comparison of our attention results (1st column) with attention provided by [2] (2nd column) and [1] (3rd column). The fourth column provides the question and the answer of the different techniques.