Visual Explanations from Hadamard Product in Multimodal Deep Networks

Jin-Hwa Kim
Seoul National University
Seoul, Republic of Korea
jhkim@bi.snu.ac.kr

Byoung-Tak Zhang
Seoul National University
Seoul, Republic of Korea
btzhang@bi.snu.ac.kr

Abstract

The visual explanation of learned representation of models helps to understand the fundamentals of learning. The attentional models of previous works used to visualize the attended regions over an image or text using their learned weights to confirm their intended mechanism. Kim et al. (2016) show that the Hadamard product in multimodal deep networks, which is well-known for the joint function of visual question answering tasks, implicitly performs an attentional mechanism for visual inputs. In this work, we extend their work to show that the Hadamard product in multimodal deep networks performs not only for visual inputs but also for textual inputs simultaneously using the proposed gradient-based visualization technique. The attentional effect of Hadamard product is visualized for both visual and textual inputs by analyzing the two inputs and an output of the Hadamard product with the proposed method and compared with learned attentional weights of a visual question answering model.

1 Introduction

As a multimodal joint function, Hadamard product is widely used in the multimodal learning tasks. Many state-of-the-art models used the Hadamard product as a joint function to achieve competitive performance [5, 6, 9, 12] for the visual question answering (VQA) [1], and one of them won the recent VQA challenge [12].

The characteristic of Hadamard product is studied in deep neural networks. MI-RNN [13] uses this to integrate different information flows within RNN. They show that hidden activations are not saturated toward ± 1.0 comparing to the addition, which implies that the gradient of *tanh* function is not vanished. For the MLB [6], they show that Hadamard product performs low-rank bilinear pooling in deep neural networks.

In this paper, we show that the analysis of the input and output of Hadamard product in multimodal deep networks is sufficient to visualize the cross-grounding between two modalities. Unlike the previous works, this method does not need an annotated label nor attentional weights. These results suggest that Hadamard product as a multimodal joint function gives not only excellent performance in the task but also visual explanations for the vision-language, input modalities.

2 Previous Works

Class Activation Mapping (CAM) is proposed to identify discriminative regions in image classification tasks [14]. CAM utilizes global average pooling layers to get the representations to localize the regions. However, it has the limitation of the CNN architecture and the modification of architecture requires re-training. Grad-CAM [10] generalizes this to various CNN models. The Grad-CAM can

NIPS 2017 Workshop on Visually-Grounded Interaction and Language (ViGIL), Long Beach, CA, USA.

also visualize in the other tasks, *e.g.* image captioning and visual question answering, with the similar approach. Unlike the previous methods, our method is an unsupervised visualization, and a direct visualization of vision-language cross-groundings occurred in the joint function.

3 Visual Explanations from Hadamard Product

In their work [5], they visualize the difference between the intermediate visual input $\mathcal{V}_i$ and the output of Hadamard product $\mathcal{F}_i$ between the intermediate visual input $\mathcal{V}_i$ and the intermediate textual input $\mathcal{Q}_i$, in the space of image for three layers using standard back-propagation.

$$\frac{\partial \mathcal{L}_i}{\partial \mathcal{I}} = \frac{\partial\big(\frac{1}{2}||\mathcal{V}_i - \mathcal{F}_i||_2^2\big)}{\partial \mathcal{I}} = \frac{\partial \mathcal{V}_i}{\partial \mathcal{I}}(\mathcal{V}_i - \mathcal{F}_i) \tag{1}$$

where $\mathcal{I}$ is an input image to ResNet-152, and $\mathcal{V}_i$ is a function of $\mathcal{I}$; however, though $\mathcal{F}_i$ is also a function of $\mathcal{V}_i$, the $\mathcal{F}_i$ is treated as a constant for the visualization purpose. We speculate that the constant $\mathcal{F}_i$ set the virtual target of joint representation and the mean squared error between $\mathcal{F}_i$ and $\mathcal{V}_i$ indicates the amount of deviation of $\mathcal{V}_i$ from $\mathcal{F}_i$. Empirically, this way of visualization is more effective than not fixing the $\mathcal{F}_i$.

We define the visual explanation of vision and language using the inputs to Hadamard product and the output of the product:

$$\nabla_v := (\mathcal{V} - \mathcal{F})\partial\mathcal{V}/\partial\mathcal{I} \tag{2}$$

$$\nabla_q := (\mathcal{Q} - \mathcal{F})\partial\mathcal{Q}/\partial q \tag{3}$$

Unlike the their work, we use *guided back-propagation* [11] for ReLU activations, which uses the only positive gradient of output for back-propagation, for a better visualization. This imputed version of gradient is calculated using the ResNet-152, the feature extractor. Equation 3 is newly introduced with the same gist. Here, the q is the embedded word vectors for a given question which are looked up by the indices of tokens. Notice that the dimensions of $\mathcal{V}$, $\mathcal{Q}$, and $\mathcal{F}$ are the same since $\mathcal{F}$ is the output of element-wise multiplication of $\mathcal{V}$ and $\mathcal{Q}$. Moreover, these definitions can be generalized to the other models which uses Hadamard product as multimodal joint function.

4 Experiments

We use the multimodal low-rank bilinear attention networks (MLB) [6] as the VQA model for visualization, to compare the visual explanation for visual input with attentional weights, and to show the visual explanation for textual input. The MLB provides an efficient attention mechanism for visual question-answering tasks, based on the interpretation of Hadamard product as a key operator for low-rank bilinear pooling.

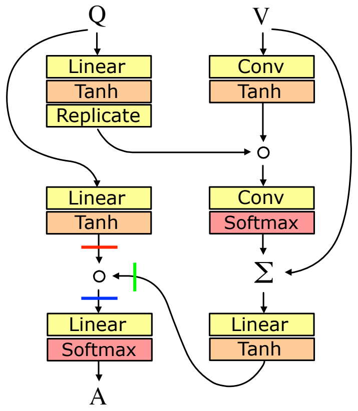

Figure 1: A diagram of MLB indicating the analyzing points; **red**: textual input $\mathcal{Q}$, **green**: visual input $\mathcal{V}$, and **blue**: the output of Hadamard product $\mathcal{F}$.

4.1 Multimodal Low-rank Bilinear Attention Networks

The inputs are a question embedding vector $\mathbf{q}$, which is the output of a learnable Skip-thought Vectors model [7], and a set of visual feature vectors $\mathbf{F}$ over $S \times S$ lattice space, which is the output of the fixed ResNet-152 model [3]. In this section, we briefly describe the structure of MLB.

Attention mechanism uses an attention probability distribution α over $S \times S$ lattice space. Here, using low-rank bilinear pooling, α is defined as:

$$\alpha = \text{softmax}\Big(\mathbf{P}_\alpha^T\big(\sigma(\mathbf{U}_\mathbf{q}^T\mathbf{q}\cdot \mathbb{1}^T)\circ\sigma(\mathbf{V}_\mathbf{F}^T\mathbf{F}^T)\big)\Big) \tag{4}$$

where $\alpha \in \mathbb{R}^{G\times S^2}$, $\mathbf{P}_\alpha \in \mathbb{R}^{d\times G}$, σ is a hyperbolic tangent function, $\mathbf{U}_\mathbf{q} \in \mathbb{R}^{N\times d}$, $\mathbf{q} \in \mathbb{R}^N$, $\mathbb{1} \in \mathbb{R}^{S^2}$, $\mathbf{V}_\mathbf{F} \in \mathbb{R}^{M\times d}$, and $\mathbf{F} \in \mathbb{R}^{S^2\times M}$. If $G > 1$, multiple glimpses are explicitly expressed as in Fukui et al. [2], conceptually similar to Jaderberg et al. [4]. And, the softmax function applies to each row vector of α. The bias terms are omitted for simplicity.

Attended visual feature $\hat{\mathbf{v}}$ is a linear combination of $\mathbf{F}_i$ with the corresponding coefficients $\alpha_{g,i}$. Each attention probability distribution α_g is for a glimpse g. For $G > 1$, $\hat{\mathbf{v}}$ is the concatenation of resulting vectors $\hat{\mathbf{v}}_g$ as:

$$\hat{\mathbf{v}} = \mathop{\Big|\Big|}_{g=1}^{G} \sum_{s=1}^{S^2} \alpha_{g,s} \mathbf{F}_s \tag{5}$$

where || denotes concatenation of vectors. The posterior probability distribution is an output of a softmax function, whose input is the result of another low-rank bilinear pooling of $\mathbf{q}$ and $\hat{\mathbf{v}}$ as:

$$\mathcal{Q} := \sigma(\mathbf{W}_{\mathbf{q}}^T \mathbf{q}), \quad \mathcal{V} := \sigma(\mathbf{V}_{\hat{\mathbf{v}}}^T \hat{\mathbf{v}}), \quad \mathcal{F} := \mathcal{Q} \circ \mathcal{V} \tag{6}$$

$$p(a|\mathbf{q}, \mathbf{F}; \Theta) = \text{softmax}\big(\mathbf{P}_o^T \mathcal{F}\big) \tag{7}$$

$$\hat{a} = \arg\max_{a \in \Omega} p(a|\mathbf{q}, \mathbf{F}; \Theta) \tag{8}$$

where $\hat{a}$ denotes a predicted answer, Ω is a set of candidate answers and Θ is an aggregation of entire model parameters. Our method uses the intermediate representations, $\mathcal{Q}$, $\mathcal{V}$, and $\mathcal{F}$.

In Figure 1 indicates the analyzing points, $\mathcal{Q}$, $\mathcal{V}$, and $\mathcal{F}$. The *Replicate* module copies an question embedding vector to match with S^2 visual feature vectors. *Conv* modules indicate 1×1 convolution to project channel dimension, which is computationally equivalent to linear projection for the channel.

4.2 Post-processing

Using Equation 2 and 3, we get the gradients of $\nabla_v \in \mathbb{R}^{C \times H \times W}$ and $\nabla_q \in \mathbb{R}^{\rho \times D}$. The ∇_v has the same size of a RGB image and the ρ is the number of tokens in a question, and D is the dimension of word embedding vector. Then, each pixel $\nabla_{v(i,j)}$ is normalized (channel-wise) by:

$$\hat{\nabla}_{v(i,j)} = (\nabla_{v(i,j)} - \mu(\nabla_v))/\sigma(\nabla_v) \tag{9}$$

where μ denotes mean and σ denotes standard deviation.

For the question, we take the absolute values of ∇_q, followed by the summation over D for the i-th token:

$$\hat{\nabla}_{q(i)} = \sum_{d=1}^{D} |\nabla_{q(i,d)}| \tag{10}$$

Then, the standard score z_i is calculated to get relative importance of each token as follows:

$$z_i = \big(\hat{\nabla}_{q(i)} - \mu(\hat{\nabla}_q)\big)/\sigma(\hat{\nabla}_q). \tag{11}$$

5 Results and Discussions

Figure 2 shows an example of the visual explanations. The first column shows an input image, and the second and third columns show the first and second attention maps representing α_1 and α_2 (MLB uses the G of 2). Notice that $\hat{\mathbf{v}}$ represents the concatenation of g attended visual features. Through this, the model learns to generate appropriate attention probability distributions α_g in parallel. The first and second rows show that α_1 and α_2 has similar distributions; however, the third row shows some difference.

The fourth column shows the visualization of our proposed method. Although this visualization comes from the analysis of Hadamard product, it shows a similar result to the attention maps. The first and second attend to *a donut on a plate* and *a scarf in the neck*, respectively. In the third row, the proposed visualization seems to represent both of the first and second attention maps in an additive way.

The visual explanation of textual input shows a plausible result that the nouns are significantly attended (red), whereas *wh* words, verbs, adjectives, and articles are less attended (blue) in the bottom

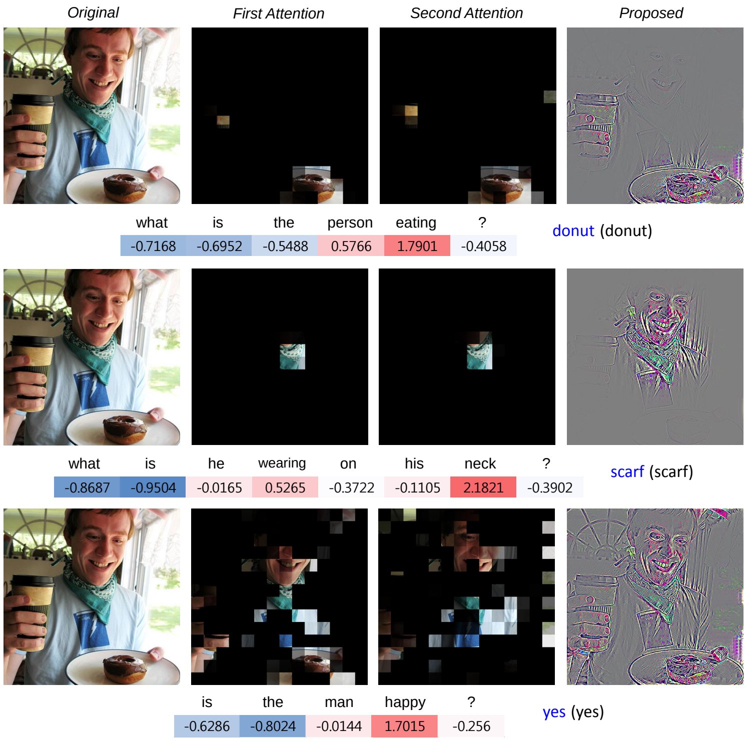

Figure 2: The visualization of attentional weights α_g (second and third columns), visual explanations for visual (forth column), and textual (bottom of each row) inputs. The two attention maps are represented on the 14×14 lattice space, while the proposed method represents an attended region on the image pixels. The plate of donut is visualized in the proposed method (first row), and the scarf (second row).

of each row. Take-home message is two-fold: **1)** these results are competitive with the explicit textual attention models [9, 8]. The explicit textual attention of theirs are not giving unprecedented attention to the text (notice that Hadamard product is a well-known joint function in the VQA tasks), rather it might be working as the regularization using selective weights. **2)** there is no sufficient evidence that the visual attention is based on the comprehension of a question. Because the textual attention is not expanding to the verbs and propositions connected to the nouns (*e.g.* *'on'* of *'wearing on his neck'*), which phenomena seem to be consistent with the co-attention models [9, 8], although the work [8] tried to mitigate the problem using word, phrase, and question-level features.

6 Conclusions

In this work, we show that the Hadamard product in multimodal deep networks implicitly performs an attentional mechanism not only for visual inputs but also for textual inputs simultaneously using the proposed gradient-based visualization technique in a visual question answering model. Though this technique is based on the analysis of Hadamard product in multimodal deep networks, it shows competitive results with the explicit visualization of learned attentional weights. Our results suggest that the explicit textual attention is not providing a unique attentional mechanism to the textual input. Instead, it might be the regularization using selective weights which are learned by training. Moreover, we cautiously argue that textual attention is biased toward the noun words appeared in a given text which limits the inferential capability of the model.

Acknowledgments. JK is supported by 2017 Google PhD Fellowship, and partly by NAVER Corp. and the Korea government (IITP-2017-0-01772-VTT, IITP-R0126-16-1072-SW.StarLab, KEIT-10060086-RISF).

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.

[2] 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 arXiv:1606.01847*, 2016.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Advances in Neural Information Processing Systems 28*, pages 2008–2016, 2015.

[5] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal Residual Learning for Visual QA. In *Advances in Neural Information Processing Systems 29*, pages 361–369, 2016.

[6] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017.

[7] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-Thought Vectors. In *Advances in Neural Information Processing Systems 28*, pages 3294–3302, 2015.

[8] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical Question-Image Co-Attention for Visual Question Answering. *Advances in Neural Information Processing Systems 29*, pages 289–297, 2016.

[9] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[10] Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *International Conference on Computer Vision*, 2017.

[11] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for Simplicity: The All Convolutional Net. In *ICLR 2015 Workshop Track*, 2015.

[12] 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.

[13] Yuhuai Wu, Saizheng Zhang, Ying Zhang, Yoshua Bengio, and Ruslan Salakhutdinov. On Multiplicative Integration with Recurrent Neural Networks. In *Advances in Neural Information Processing Systems 29*, pages 2856–2864, 2016.

[14] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.

A Appendix

The supplementary examples of our method are shown in Figure 3, 4, and 5. This example emphasizes the importance of visual explanation. In the first row, the question is ‘*what color is the toddler’s hair?*’ and the corresponding answer is ‘*blonde*’. Without the visualization, we do not know whether the model is *purely* biased from the data distribution or not. Although there is the possibility that the model is biased toward ‘*blonde hair*’ *and* attends the hair in the given image, this visual explanation helps to assess the model.

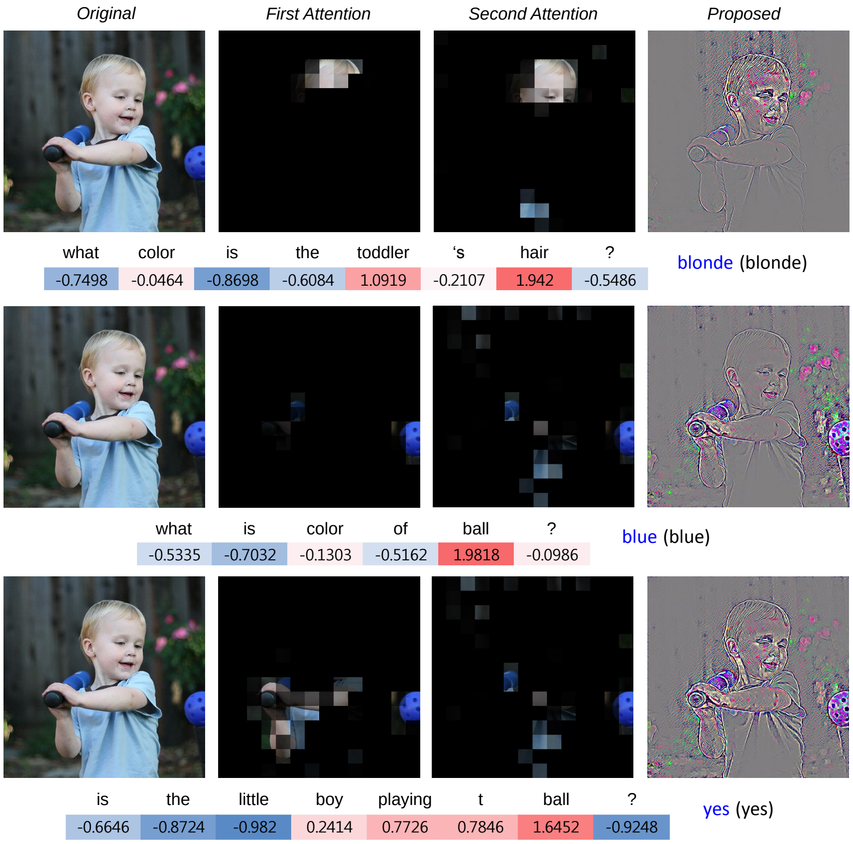

Figure 3: Another examples of the visualization.

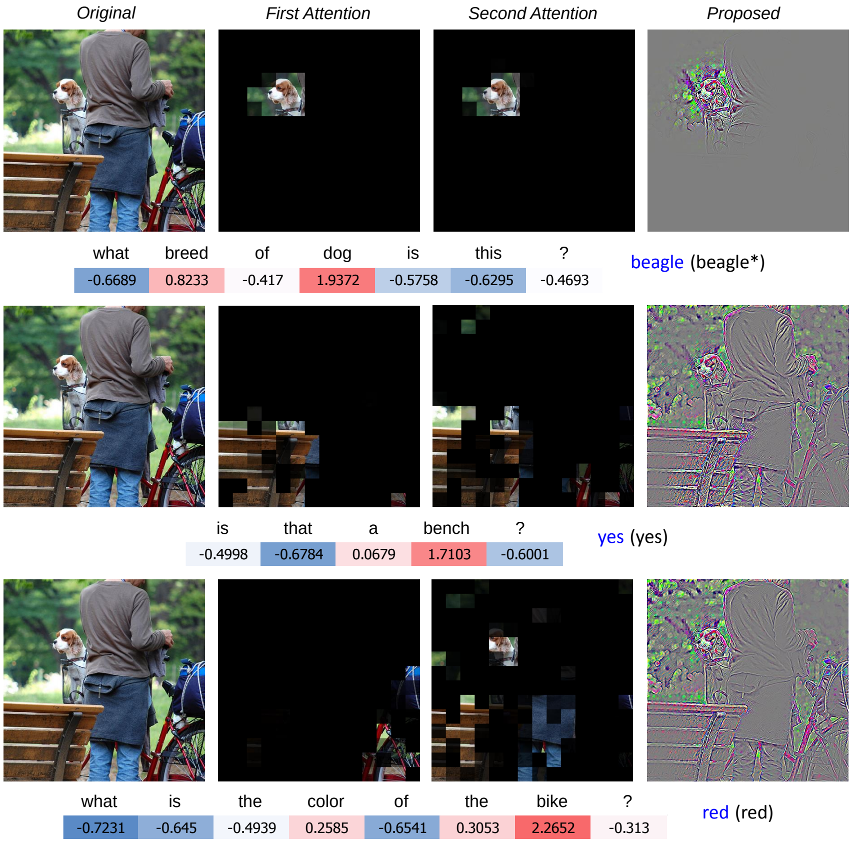

Figure 4: Another examples of the visualization. In the second and third rows of the forth column, the attended area of bench and bike is slightly different. Interestingly, the attention maps of the third row are different from each other, the first attention shows the part of bike, whereas the second attention shows the other salient objects. *Which is unclear to be Beagle or Charles Spaniel.

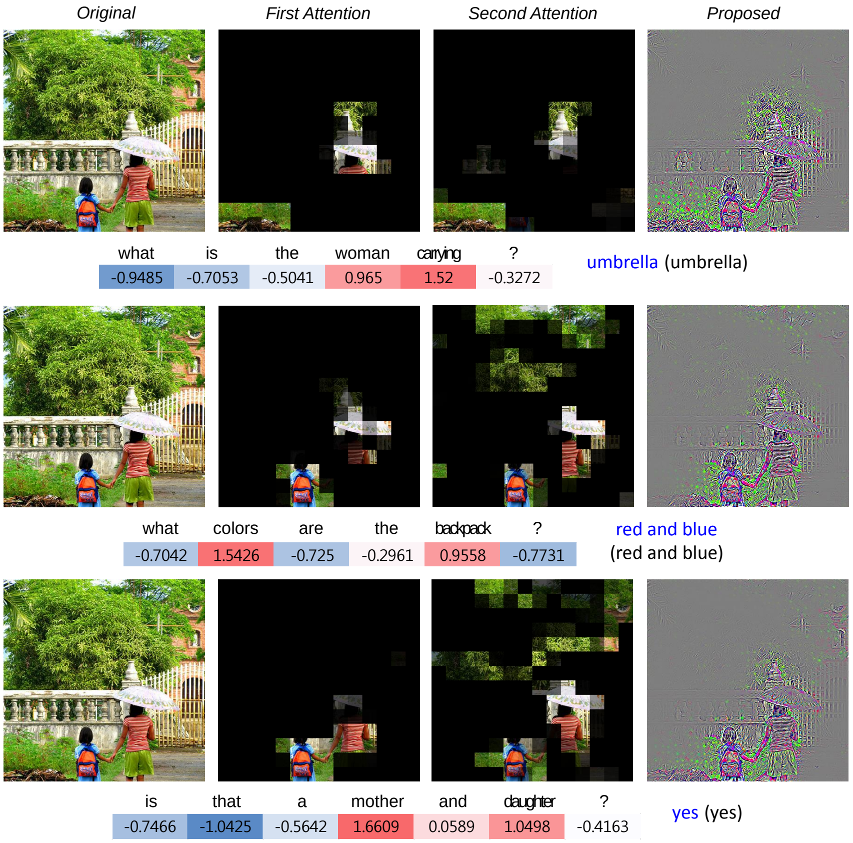

Figure 5: Another examples of the visualization. In the first and second rows of the fourth column, there is a very subtle difference between umbrella and backpack (distinguishable with two standard deviation threshold). The backpack has blue color in its side area (second row). We do not know the relationship between the two but only can infer from the pose of hand in hand.