Towards visual explanations that ground multiple modalities

Madhura Keshava Ummettuguli
Georgia Institute of Technology
mummettuguli3@gatech.edu

Abstract

We try to produce more accurate visual explanations for decisions from multi-modal networks such as Visual Question Answering (VQA) models. With the existing technique, i.e. Grad-CAM [3], the maps generated for two questions about an image having the same answer, are very similar. Grad-CAM uses the answer to visualize where the model must look to give the answer. In this computation, the question has very little influence. The motivation for this work comes from the fact that the question plays an equally important role as the answer in determining where the model did look to give the answer. Especially for binary questions such as 'Is there a cat?', the answer 'yes' or 'no' by itself does not make sense without the question. We see this manifesting in the visualizations being similar for questions that have the same answer. We propose a technique to improve these visualizations by allowing only a part of the gradients to flow backward during back propagation from the combine layer (where the question and image features combine) to the last convolutional layer where the Grad-CAM map is computed. Only the gradients corresponding to the combine layer neurons which are more sensitive to the question, calculated from the top n indices of the $\frac{\partial A}{\partial q}$ vector are allowed to flow back. Our visualizations show that the model does look at different regions in the image depending on the question, unlike the previous visualizations where it showed the model looking at the same region to produce the same answer despite the question being different. We identify the future direction to arrive at a generic solution.

1. Introduction

Grad-CAM maps generated for two questions about an image having the same answer, are very similar. Given the utility of visual explanations in lending insights into failure modes of models, making models interpretable and building trust in them, having more accurate explanations is of importance.

For the Deeper LSTM + normalized CNN model, given the Figure 1 and the question 'What is the color of the sink?', Figure 2 shows the Grad-CAM heat map for the answer 'white'. Given the Figure 1 and the question 'What color are the towels?, Figure 3 shows the Grad-CAM heat map for the answer 'white'. We see that the heat maps generated are very similar when the answer is the same.

Typical VQA [1] pipelines consist of a CNN to process images and an RNN language model for questions. The image and the question representations are fused to predict the answer, typically with a 1000-way classification (1000 being the size of the answer space). Since this is a classification problem, in the Grad-CAM approach we pick an answer and use its score to compute Grad-CAM visualizations over the image to explain the answer. The question has very little influence in this computation. The gradients flowing back for the computation are very similar for the same answer, which is why the Grad-CAM maps are very similar.

We try different approaches to enhance the influence of the question in the Grad-CAM computation.With the Deeper LSTM + normalized CNN model we find that with a mechanism to identify important neurons at the combine layer, allowing gradients to flow backward only via these important neurons produces Grad-CAM maps which are better at localizing the regions under consideration. An example of the sink and towel regions being localized is shown in Figures 5 and 6.

2. Related Work

VQA. A VQA system takes as input an image and a question about the image and produces a natural language answer as the output. Various models have been developed for this task, improving over previous ones. For our work, we are experimenting with Grad-CAM for the Pythia VQA model [2] and the Deeper LSTM + normalized CNN model [1].

Grad-CAM. In order to obtain the localization map Grad-CAM for a VQA model $L^c_{Grad-CAM} \in R^{u \times v}$ of width u and height v, the gradient of the score for answer c, y^c is calculated with respect to feature map activations A^k of a convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k}$.These gradients flowing back are then global-average-pooled over the width and

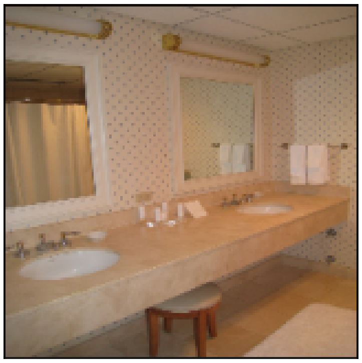

Figure 1. Original Image

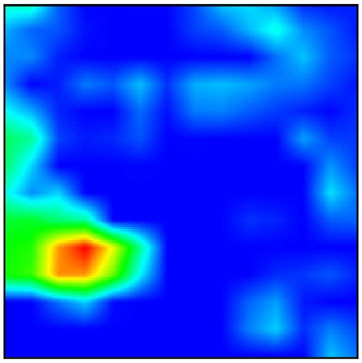

Figure 2. Grad-CAM heat map for Q: What is the color of the sink? A: white (Deeper LSTM+normalized CNN model)

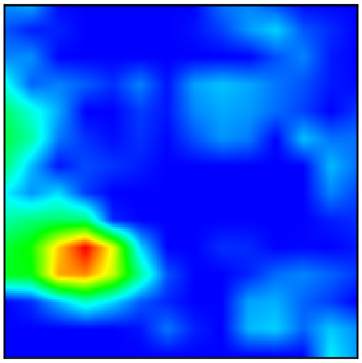

Figure 3. Grad-CAM heat map for Q: What color are the towels? A: white (Deeper LSTM+normalized CNN model)

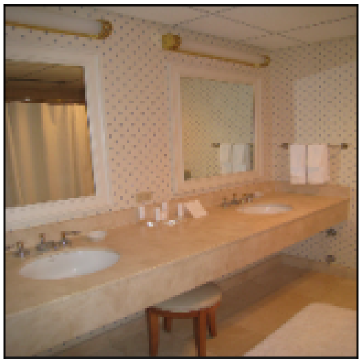

Figure 4. Original Image

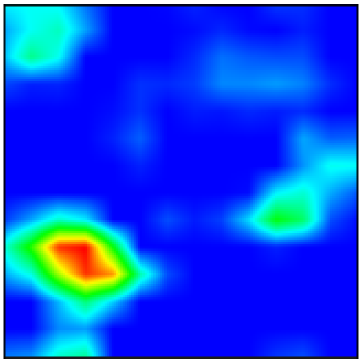

Figure 5. Grad-CAM heat map using our technique for Q: What is the color of the sink? A: white
(Deeper LSTM+normalized CNN model)

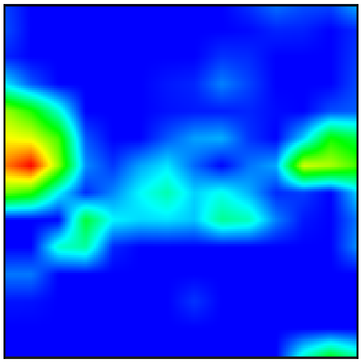

Figure 6. Grad-CAM heat map using our technique for Q: What color are the towels? A: white
(Deeper LSTM+normalized CNN model)

height dimensions (indexed by i and j respectively) to obtain the neuron importance weights α_k^c

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

A weighted combination of forward activation maps is followed by a ReLU to obtain,

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{2}$$

3. Approach

We generated Grad-CAM visualizations of the ground truth answers for the Pythia VQA model from the VQA Validation dataset for question type 'What color' and answer type 'yes/no'. Examples shown in Figures 7-9, 13-19. Thus, we validated the problem with the Grad-CAM visualizations generated for the Pythia model as well.

3.1. Visualizing Grad-CAM maps for different questions and answers for the same image

We see how the activations change for a given image with (a) different questions having the same answer: we saw that the activations are different, and their signs also differ quite a bit (b) different answers for the same question: we saw that the activations are same.

Figures 17 and 18 show Grad-CAM maps for the questions 'What is the man doing?' and 'What is the woman doing?', with both having the answer as 'sitting'. While these

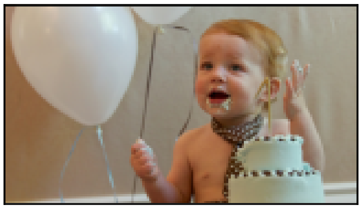

Figure 7. Original Image

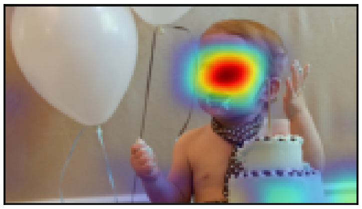

Figure 8. Grad-CAM output for Pythia model generated for Q: What color is the cake? A: white

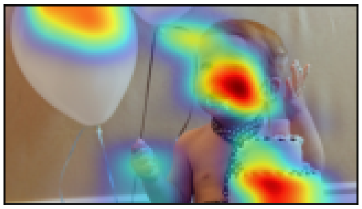

Figure 9. Grad-CAM output for Pythia model generated for Q: What color are the balloons? A: white

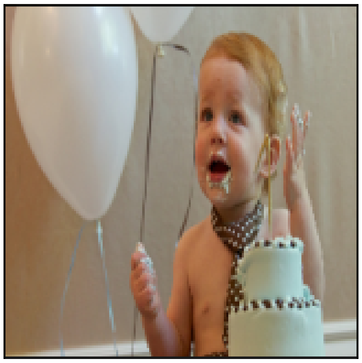

Figure 10. Original Image

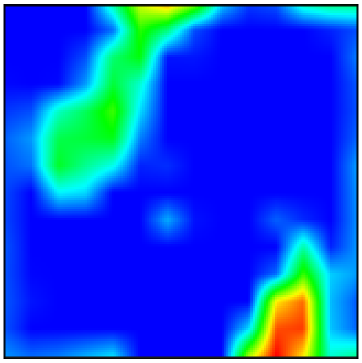

Figure 11. Grad-CAM output for Deeper LSTM+normalized CNN model generated for Q: What color is the cake? A: white

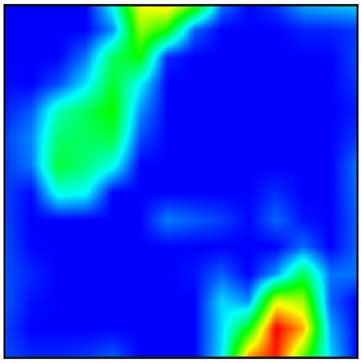

Figure 12. Grad-CAM output for Pythia generated for Q: What color are the balloons? A: white

maps are similar, we see that when the answer is different, as in Figure 19, the Grad-CAM map differs.

When there is ReLU, the gradients which are backpropagated depends on the sign of the activations. Only the neurons which have a positive activation will allow the gradient to flow back [4]. Hence the idea was to see where the signs of the Activations are different for different questions for the layers after the question interaction in the model.

But during this experiment we observed that Grad-CAM when applied on the Pythia VQA model does not seem to be as good at localizing regions as Grad-CAM when applied on the Deeper LSTM + normalized CNN model. If we look at the visualization in Figure 8 and compare the same with the visualization in Figure 11, we see that Grad-CAM for Pythia is not showing the cake at all while the Grad-CAM for Deeper LSTM + normalized CNN model shows the cake. Similarly, As seen in Figure 21, Grad-CAM for the Deeper LSTM + normalized CNN model localizes the man sitting clearly whereas the corresponding visualization of the Pythia model in Figure 17 is not as good. Hence, we decide to first experiment with Grad-CAM on the Deeper LSTM + normalized CNN model before moving to the Pythia model.

3.2. Calculating partial derivative of the Activations with respect to the question

We calculate the partial derivative of the Activations at the combine layer of Deeper LSTM + normalized CNN with respect to the question embedding ($\frac{\partial A}{\partial q}$). We compare the top 10 indices of $\frac{\partial A}{\partial q}$ for different questions and images.

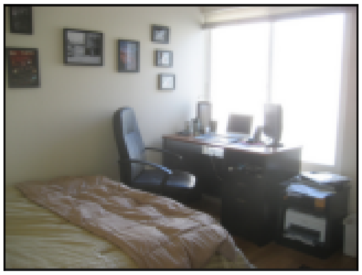

Figure 13. Original Image

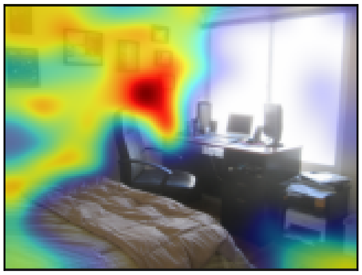

Figure 14. Grad-CAM output for Pythia generated for Q: Is that a folding chair? A: no

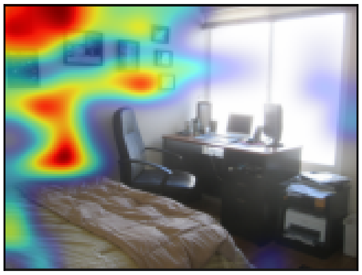

Figure 15. Grad-CAM output for Pythia generated for Q: Are these twin mattresses? A: no

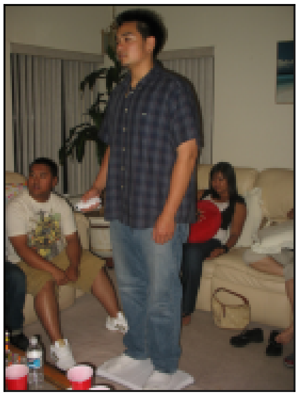

Figure 16. Original Image

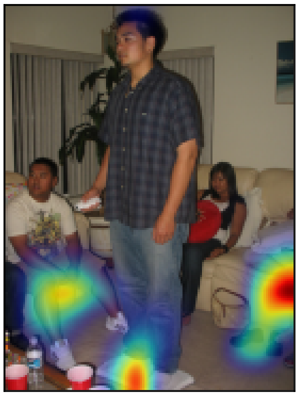

Figure 17. Grad-CAM output for Pythia generated for Q: What is the man doing? A: sitting

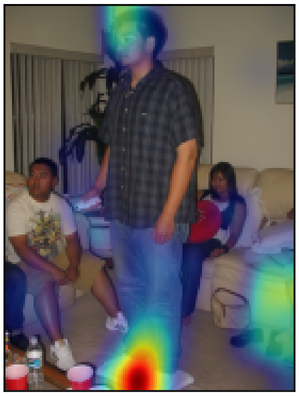

Figure 18. Grad-CAM output for Pythia generated for Q: What is the woman doing? A: sitting

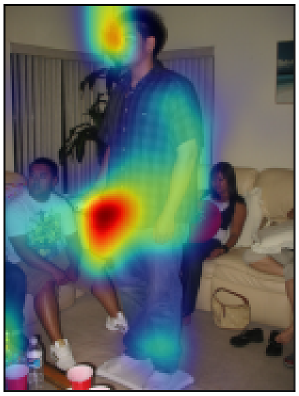

Figure 19. Grad-CAM output for Pythia generated for Q: What is the man doing? A: standing

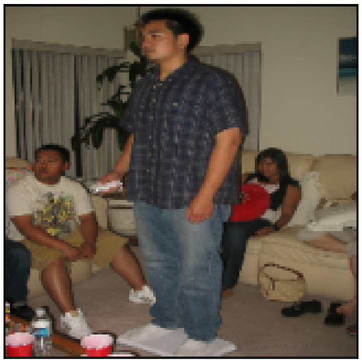

Figure 20. Original Image

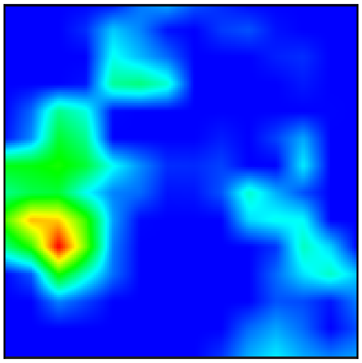

Figure 21. Grad-CAM output for Deeper LSTM+normalized CNN model generated for Q: What is the man doing? A: sitting

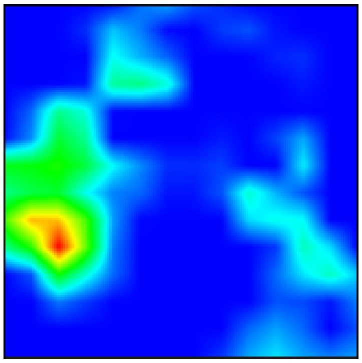

Figure 22. Grad-CAM output for Deeper LSTM+normalized CNN model generated for Q: What is the woman doing? A: sitting

Top 10 values of $\frac{\partial A}{\partial q}$ for input Figure 20	Top 10 indices of $\frac{\partial A}{\partial q}$ for input Figure 20	Top 10 values of $\frac{\partial A}{\partial q}$ for input Figure 25	Top 10 indices of $\frac{\partial A}{\partial q}$ for input Figure 25
0.6974	304	0.6975	304
0.6597	281	0.6597	281
0.4048	457	0.3944	457
0.2965	362	0.2792	362
0.2205	72	0.2253	72
0.1558	627	0.1452	627
0.1309	673	0.1296	673
0.1000	184	0.0927	184
0.0865	450	0.0869	450
0.0809	51	0.0809	51

Table 1. Top 10 values and indices of $\frac{\partial A}{\partial q}$ for the question ‘What is the man doing?’ in Figures 20 and 25. Indices are similar for the same question across images

Top 10 values of $\frac{\partial A}{\partial q}$ for input Figure 23	Top 10 indices of $\frac{\partial A}{\partial q}$ for input Figure 23	Top 10 values of $\frac{\partial A}{\partial q}$ for input Figure 24	Top 10 indices of $\frac{\partial A}{\partial q}$ for input Figure 24
0.6967	304	0.6971	304
0.6597	281	0.6597	281
0.2954	362	0.3373	457
0.2466	457	0.2972	362
0.1300	673	0.2579	72
0.0857	450	0.1615	814
0.0809	51	0.1302	673
0.0788	72	0.1102	627
0.0605	701	0.0869	450
0.0600	184	0.0809	51

Table 2. Top 10 values and indices of $\frac{\partial A}{\partial q}$ for the question ‘What is the man doing?’ in Figures 23 and 24. Indices are similar for the same question across images

Top 10 values of $\frac{\partial A}{\partial q}$ for ‘Where are they looking?’	Top 10 indices of $\frac{\partial A}{\partial q}$ for ‘Where are they looking?’	Top 10 values of $\frac{\partial A}{\partial q}$ for ‘What is the woman doing?’	Top 10 indices of $\frac{\partial A}{\partial q}$ for for ‘What is the woman doing?’
5.1008	713	2.4855	457
4.5618	304	1.6732	51
3.6857	504	1.2460	104
3.0891	296	1.2099	450
2.4732	374	1.0379	362
2.2428	234	0.8052	506
2.1852	61	0.7399	829
1.9189	362	0.5800	304
1.1856	397	0.2729	224
1.0953	118	0.2533	882

Table 3. Top 10 values and indices of $\frac{\partial A}{\partial q}$ for the questions ‘Where are they looking?’ and ‘What is the woman doing?’ in Figure 20. Indices are different for different questions

Top 10 values of $\frac{\partial A}{\partial q}$ for ‘What color is the cake?’	Top 10 indices of $\frac{\partial A}{\partial q}$ for ‘What color is the cake?’	Top 10 values of $\frac{\partial A}{\partial q}$ for ’What color are the balloons?’	Top 10 indices of $\frac{\partial A}{\partial q}$ for ’What color are the balloons?’
3.9076	929	4.0458	995
3.4301	379	3.4439	245
2.8445	214	2.9227	628
2.8362	783	2.8059	337
2.6715	831	2.5965	902
2.1125	612	2.4618	981
1.6200	903	2.3719	792
1.5807	337	2.1470	836
1.4561	154	1.8327	612
1.4277	6	1.6557	874

Table 4. Top 10 values and indices of $\frac{\partial A}{\partial q}$ for the questions ‘What color is the cake?’ and ’What color are the balloons?’ in Figure 10. Indices are different for different questions

Figure 23. Image of a man skiing

Figure 24. Image of a man skiing

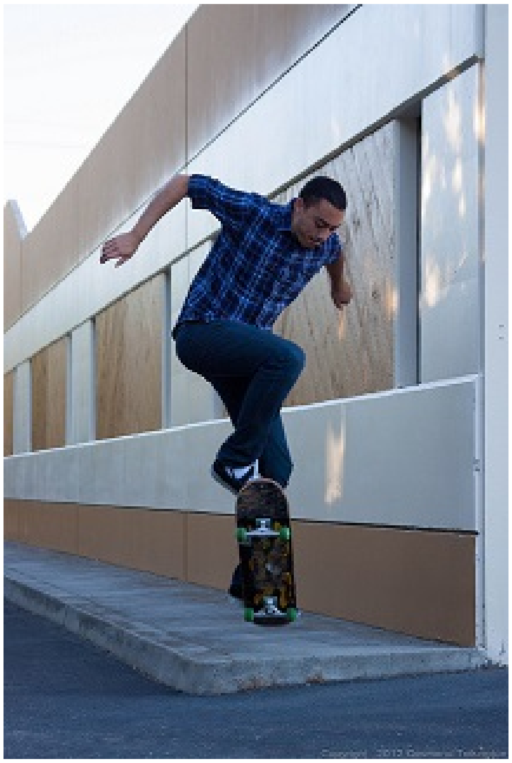

Figure 25. Image of a man skateboarding

We see that top 10 indices of $\frac{\partial A}{\partial q}$ for different questions (taking the same image and different images) is different and top 10 indices for the same question with different images is similar. Top 10 values and indices of $\frac{\partial A}{\partial q}$ for the question ‘What is the man doing?’ in Figures 20, 23, 24 and 25 are shown in Tables 1 and 2.The indices are very similar. This shows that the top 10 neurons which are sensitive to ‘What is the man doing?’ across images are similar.

3.3. Backprop only via top n neurons at the combine layer

We try to use this information about the top neurons from $\frac{\partial A}{\partial q}$ in our Grad-CAM computation. As shown in Figure 26, at the combine layer (where the question and image features combine), during back propagation we allow only a part of the gradients to flow backwards to the last convolutional layer where the Grad-CAM map is computed. Only the gradients corresponding to the combine layer neurons which are more sensitive to the question, calculated from the top n indices of the $\frac{\partial A}{\partial q}$ vector are allowed to flow back. Gradients are zeroed at indices other than the top n.

In the original Grad-CAM computation, all the gradients $\frac{\partial y^c}{\partial A^k}$ in Equation (1) are allowed to flow back, whereas in our approach we allow a subset of $\frac{\partial y^c}{\partial A^k}$ to flow back. We experiment with the value of n ranging from 10 to 1024, 1024 being the size of the combine layer (Taking Top 1024 is equivalent to the original method of computing Grad-CAM). We also calculate the rank correlation between the values of $\frac{\partial A}{\partial q}$ vectors corresponding to different questions about the same image.

Figures 27-77 show the Grad-CAM visualizations computed using gradients flowing from the top n neurons at the combine layer. For this example, we see n=25 shows the best visualization, where the sink and the towels are localized. We also note that the variations in the visualizations are quite interesting, especially how the visualization changes so drastically at top 200 for the question ‘What is the color of the sink?’, the sink loses the attention which is back again at top 300.

For other examples we see the visualizations having better localization at different values of n such as n=200 for the trees and grass example (Figures 81-83) and n=400 for the wall and toilets example (Figures 87-89). We also note that the spearman rank correlation between $\frac{\partial A}{\partial q}$ vectors are quite high as shown in Tables 5 and 6.

4. Conclusion

With these preliminary results we see that allowing only a part of the gradients to flow backwards from the combine layer to the last convolutional layer where the Grad-CAM

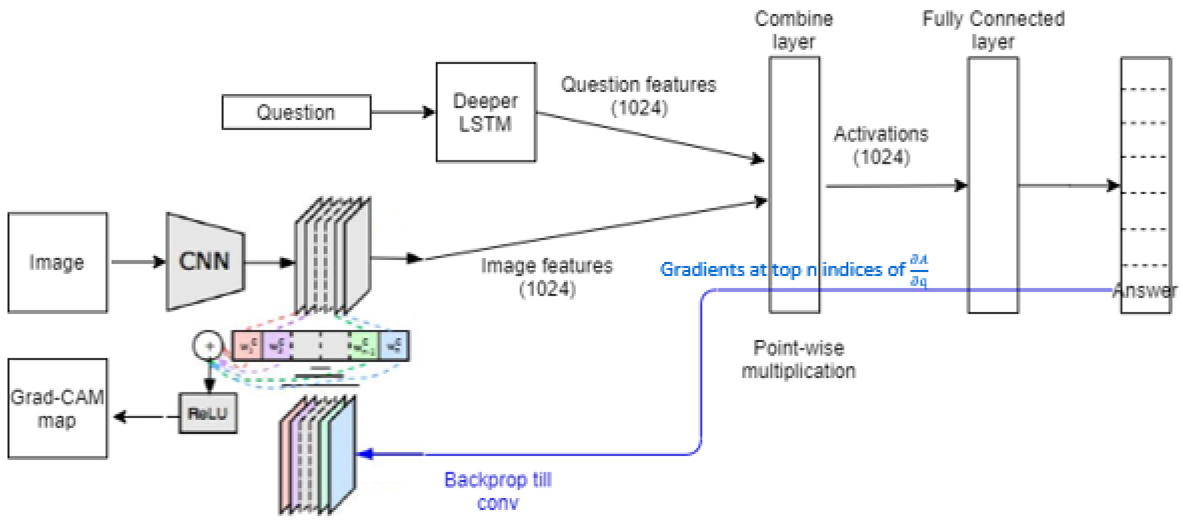

Figure 26. Deeper LSTM+normalized CNN model Architecture showing the proposed approach of backprop only via top n neurons at the combine layer for Grad-CAM computation

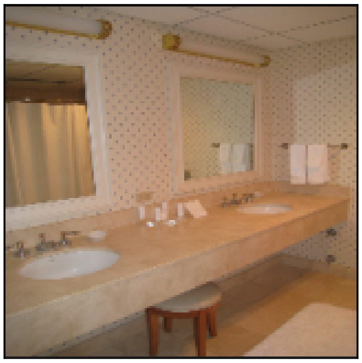

Figure 27. Original Image

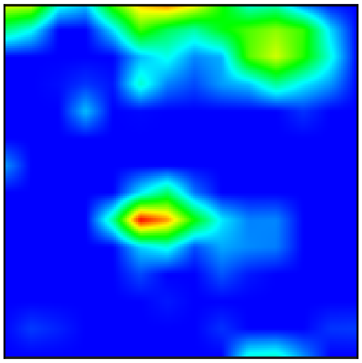

Figure 28. Top 10 for Q: What is the color of the sink? A: white

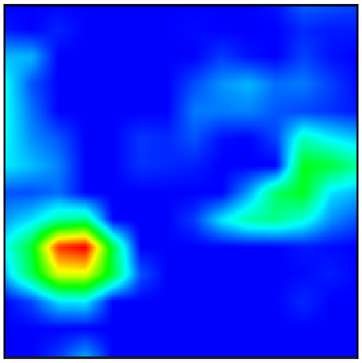

Figure 29. Top 10 for Q: What color are the towels? A: white

map is computed looks promising. Our visualizations show that the model actually does look at different regions in the image depending on the question, unlike the original Grad-CAM visualizations where it showed the model looking at the same region to produce the same answer despite the question being different. Although we have not yet arrived at a solution which is able to generalize across examples, we have identified the direction to get there.

5. Future Work

Since the rank correlation is very high, taking top n indices of $\frac{\partial A}{\partial q}$ may not be the best approach. In some cases we see the localization around top 25 is good (Figures 36-38, 90-92, 96-98). For others, at smaller values of n, 'color' seems take more importance than the object under consideration as seen in Figure 101, the color white has more importance than the object wall. Hence next we will calculate the partial derivative of the Activations at the combine layer of Deeper LSTM + normalized CNN with respect to a word embedding, such as the word embedding for 'wall' when the question is 'What color is the wall?'. We will then take top n indices of this vector $\frac{\partial A}{\partial w}$ and allow the combine layer gradients at these indices to flow back to compute Grad-CAM maps. Another approach would be to take the top n indices which are different , i.e leave out the indices which are common in the two vectors.

6. Appendix

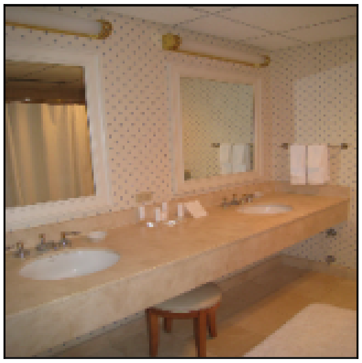

Figure 30. Original Image

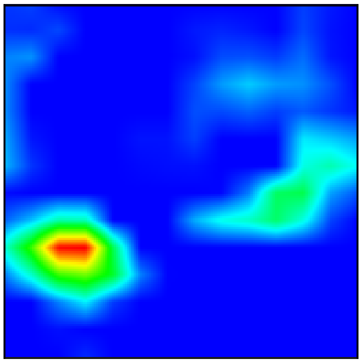

Figure 31. Top 15 for Q: What is the color of the sink? A: white

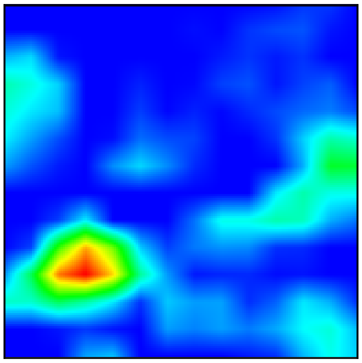

Figure 32. Top 15 for Q: What color are the towels? A: white

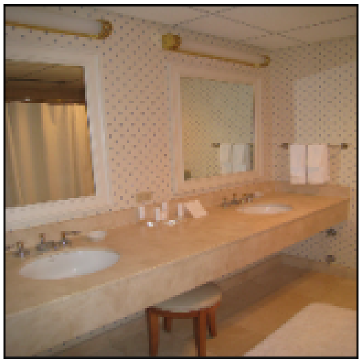

Figure 33. Original Image

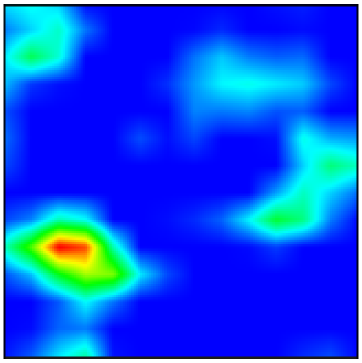

Figure 34. Top 20 for Q: What is the color of the sink? A: white

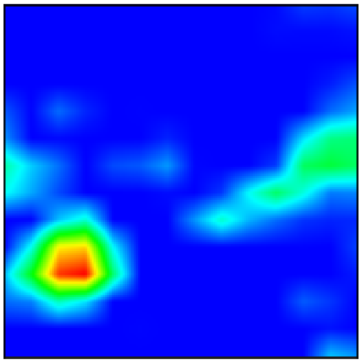

Figure 35. Top 20 for Q: What color are the towels? A: white

Spearman rank correlation between $\frac{\partial A}{\partial q}$ for the questions ‘What is the color of the sink?’ and ’What color are the towels?’ in Figure 27	0.860921
Spearman rank correlation between $\frac{\partial A}{\partial q}$ for the questions ‘What color are the trees?’ and ’What color is the grass?’ in Figure 78	0.932396
Spearman rank correlation between $\frac{\partial A}{\partial q}$ for the questions ‘What color are the toilets?’ and ’What color is the wall?’ in Figure 84	0.812344

Table 5. Spearman rank correlation between $\frac{\partial A}{\partial q}$ for different image + question combinations

Spearman rank correlation between $\frac{\partial A}{\partial q}$ for the questions ‘What color is the cake?’ and ’What color are the balloons?’ in Figure 10	0.896294
Spearman rank correlation between $\frac{\partial A}{\partial q}$ for the questions ‘What color is the phone?’ and ’What color is her top?’ in Figure 93	0.872120

Table 6. Spearman rank correlation between $\frac{\partial A}{\partial q}$ for different image + question combinations

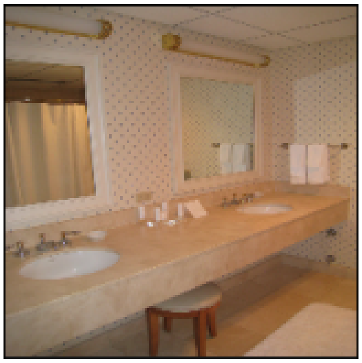

Figure 36. Original Image

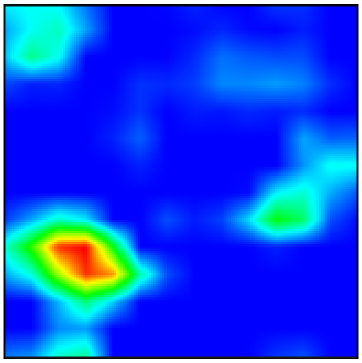

Figure 37. Top 25 for Q: What is the color of the sink? A: white

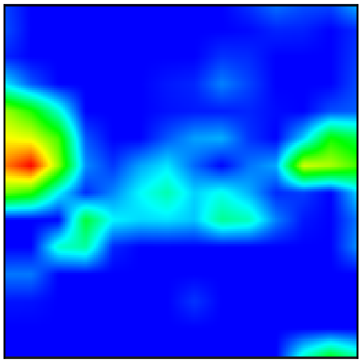

Figure 38. Top 25 for Q: What color are the towels? A: white

Figure 39. Original Image

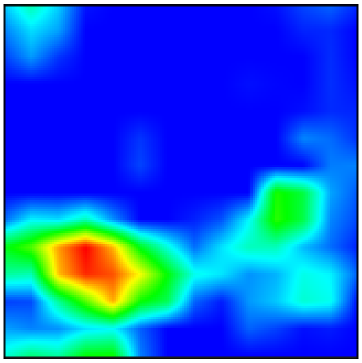

Figure 40. Top 30 for Q: What is the color of the sink? A: white

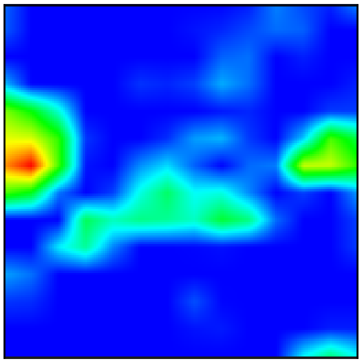

Figure 41. Top 30 for Q: What color are the towels? A: white

Figure 42. Original Image

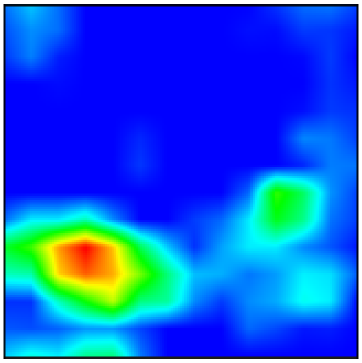

Figure 43. Top 35 for Q: What is the color of the sink? A: white

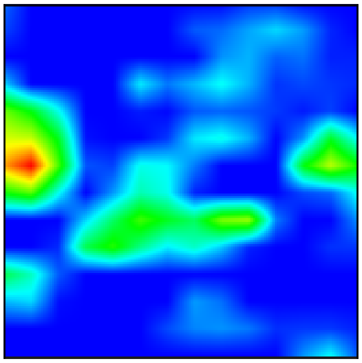

Figure 44. Top 35 for Q: What color are the towels? A: white

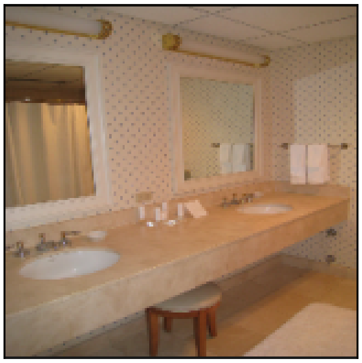

Figure 45. Original Image

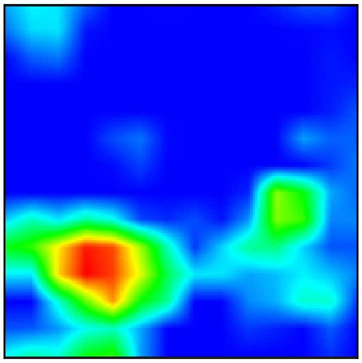

Figure 46. Top 40 for Q: What is the color of the sink? A: white

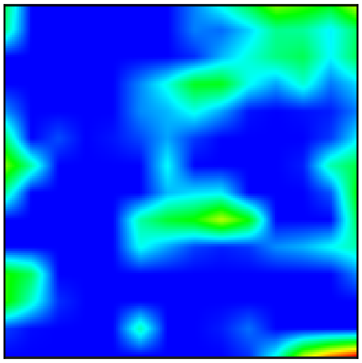

Figure 47. Top 40 for Q: What color are the towels? A: white

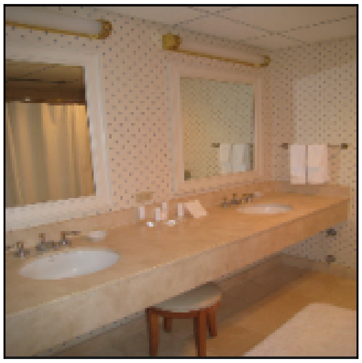

Figure 48. Original Image

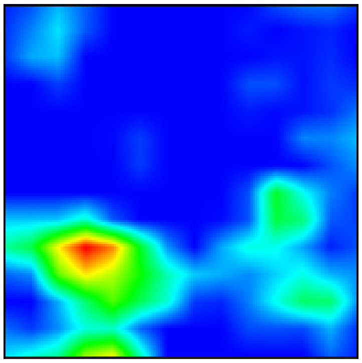

Figure 49. Top 45 for Q: What is the color of the sink? A: white

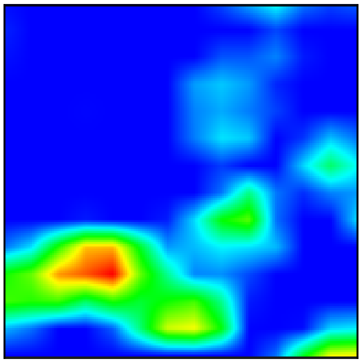

Figure 50. Top 45 for Q: What color are the towels? A: white

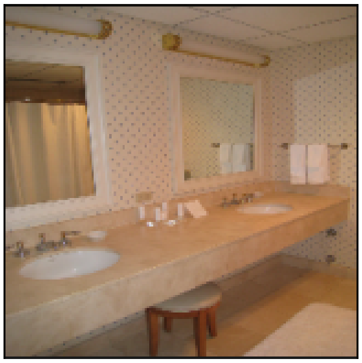

Figure 51. Original Image

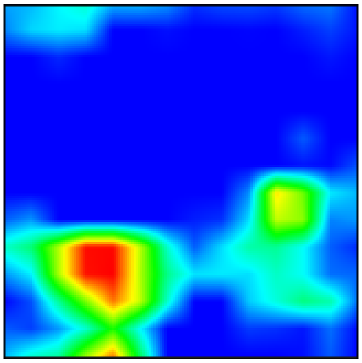

Figure 52. Top 50 for Q: What is the color of the sink? A: white

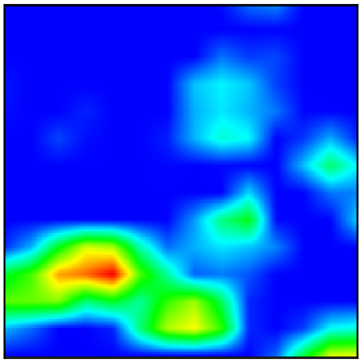

Figure 53. Top 50 for Q: What color are the towels? A: white

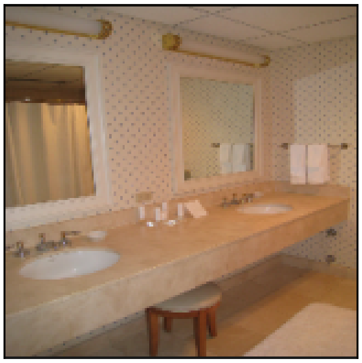

Figure 54. Original Image

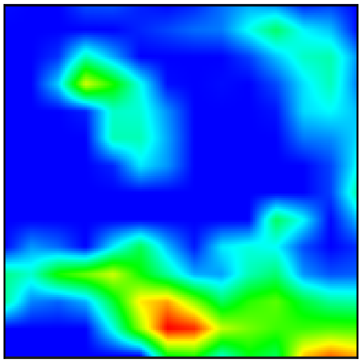

Figure 55. Top 100 for Q: What is the color of the sink? A: white

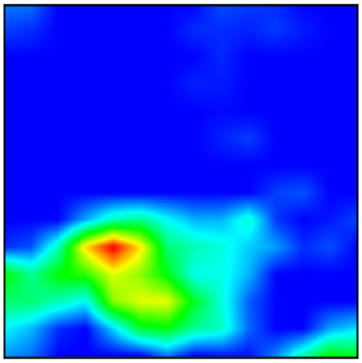

Figure 56. Top 100 for Q: What color are the towels? A: white

Figure 57. Original Image

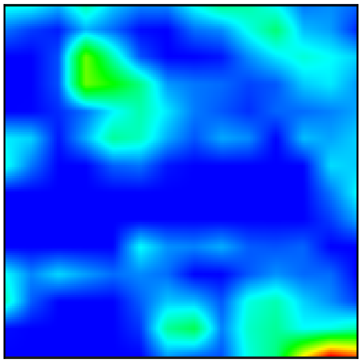

Figure 58. Top 200 for Q: What is the color of the sink? A: white

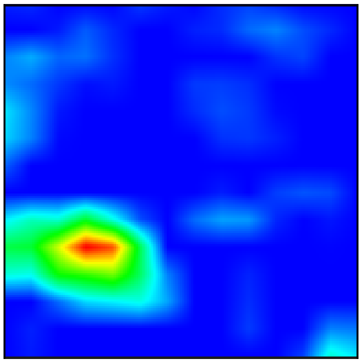

Figure 59. Top 200 for Q: What color are the towels? A: white

Figure 60. Original Image

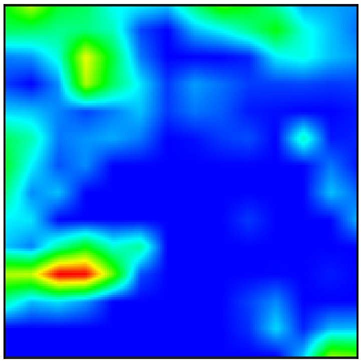

Figure 61. Top 300 for Q: What is the color of the sink? A: white

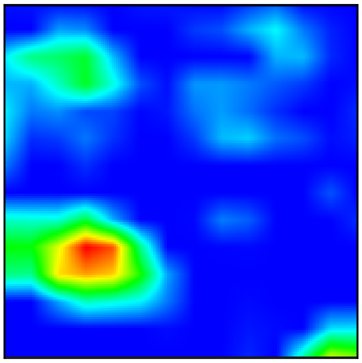

Figure 62. Top 300 for Q: What color are the towels? A: white

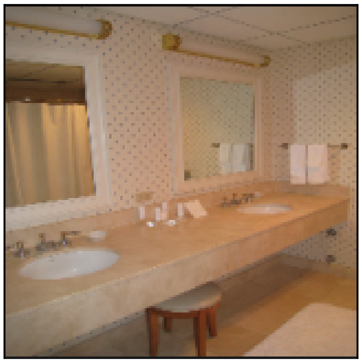

Figure 63. Original Image

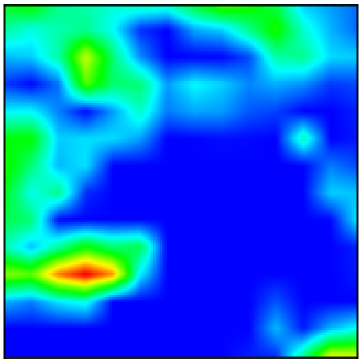

Figure 64. Top 400 for Q: What is the color of the sink? A: white

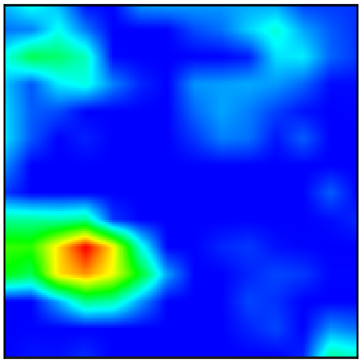

Figure 65. Top 400 for Q: What color are the towels? A: white

Figure 66. Original Image

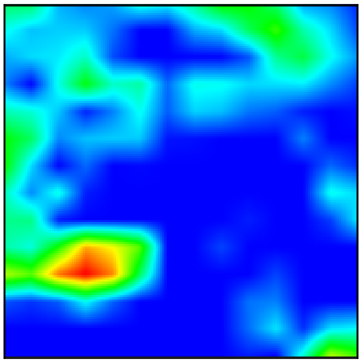

Figure 67. Top 500 for Q: What is the color of the sink? A: white

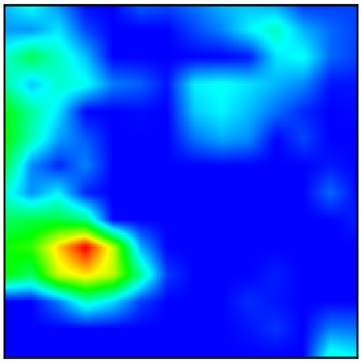

Figure 68. Top 500 for Q: What color are the towels? A: white

Figure 69. Original Image

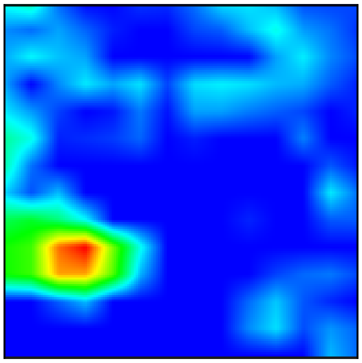

Figure 70. Top 750 for Q: What is the color of the sink? A: white

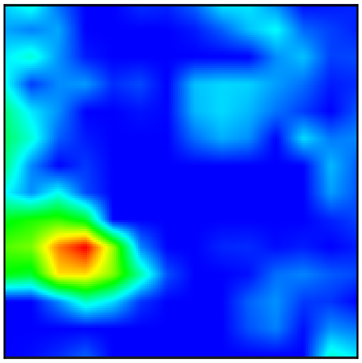

Figure 71. Top 750 for Q: What color are the towels? A: white

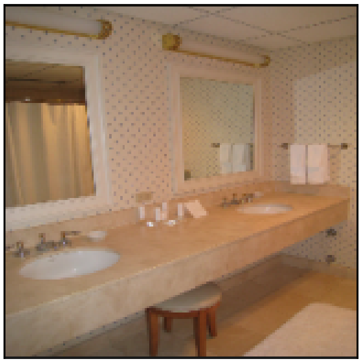

Figure 72. Original Image

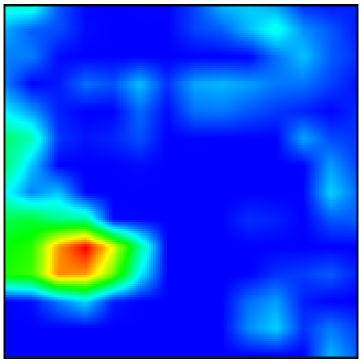

Figure 73. Top 1000 for Q: What is the color of the sink? A: white

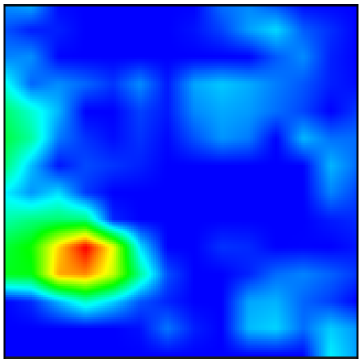

Figure 74. Top 1000 for Q: What color are the towels? A: white

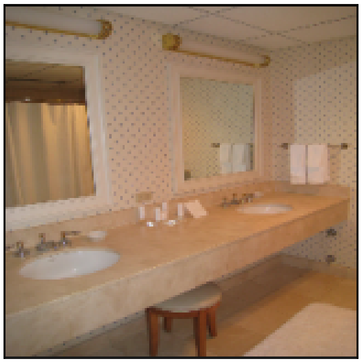

Figure 75. Original Image

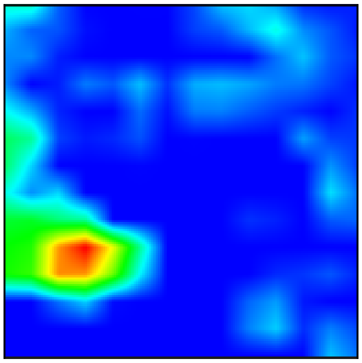

Figure 76. Top 1024 (Original Grad-CAM) for Q: What is the color of the sink? A: white

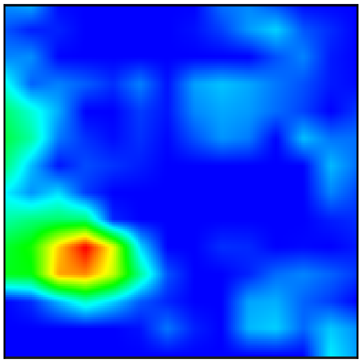

Figure 77. Top 1024 (Original Grad-CAM) for Q: What color are the towels? A: white

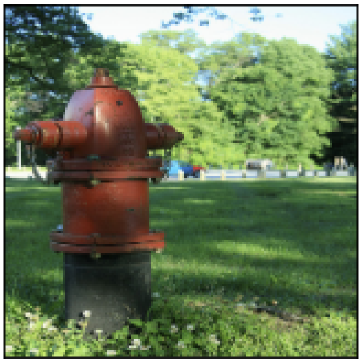

Figure 78. Original Image

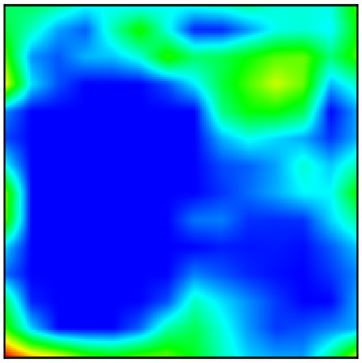

Figure 79. Top 1024 (Original Grad-CAM) for Q: What color are the trees? A: green

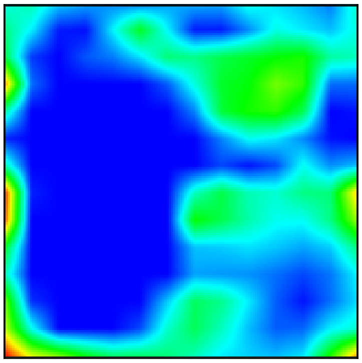

Figure 80. Top 1024 (Original Grad-CAM) for Q: What color is the grass? A: green

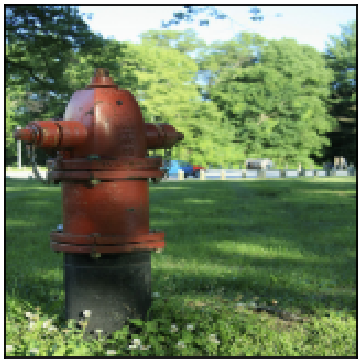

Figure 81. Original Image

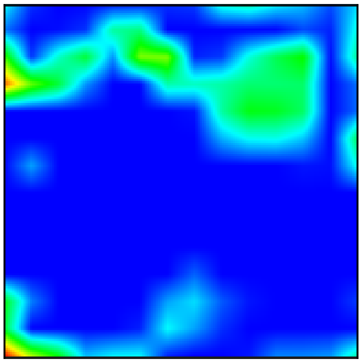

Figure 82. Top 200 for Q: What color are the trees? A: green

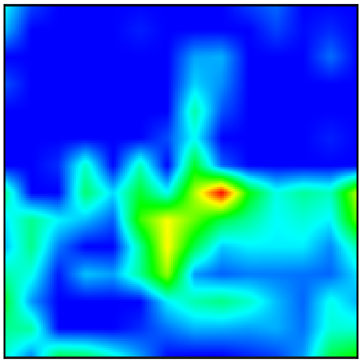

Figure 83. Top 200 for Q: What color is the grass? A: green

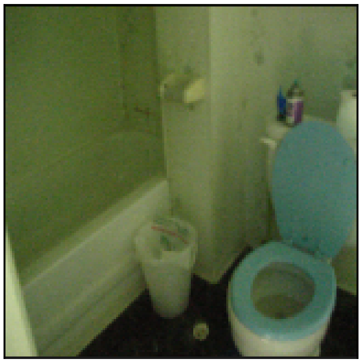

Figure 84. Original Image

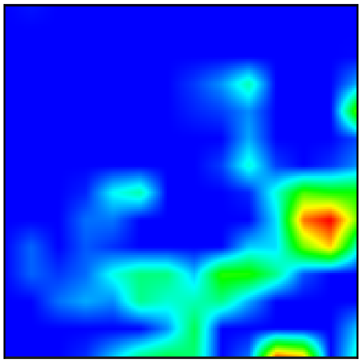

Figure 85. Top 1024 (Original Grad-CAM) for Q:What color are the toilets? A: white

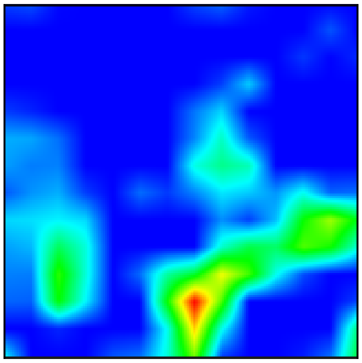

Figure 86. Top 1024 (Original Grad-CAM) for Q: What color is the wall? A: white

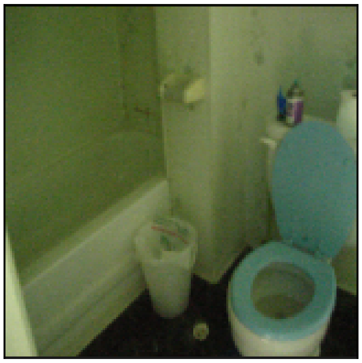

Figure 87. Original Image

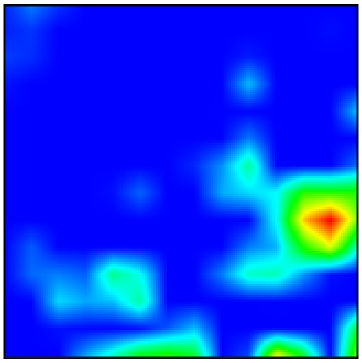

Figure 88. Top 400 for Q: What color are the toilets? A: white

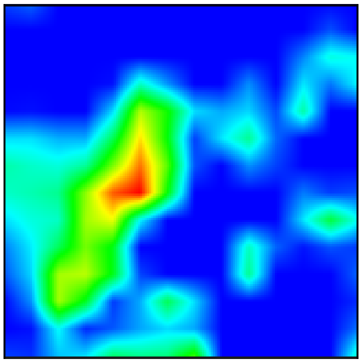

Figure 89. Top 400 for Q: What color is the wall? A: white

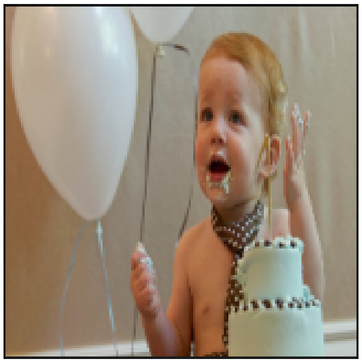

Figure 90. Original Image

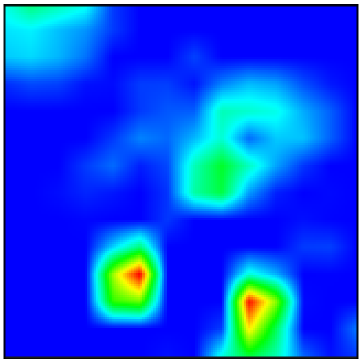

Figure 91. Top 25 for Q: What color is the cake? A: white

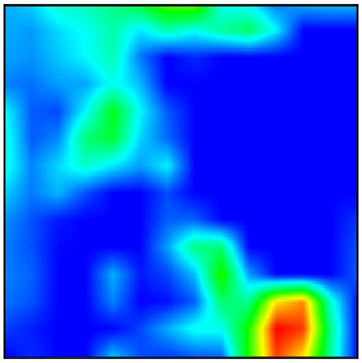

Figure 92. Top 25 for Q: What color are the balloons? A: white

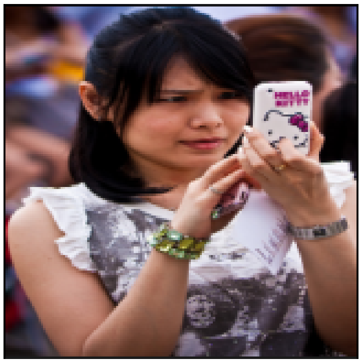

Figure 93. Original Image

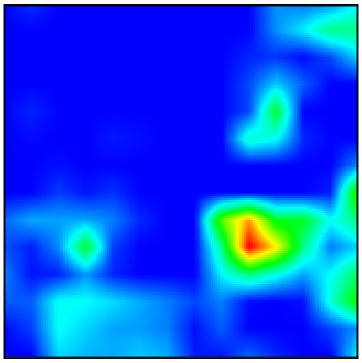

Figure 94. Top 1024 (Original Grad-CAM) for Q:What color is the phone? A: white

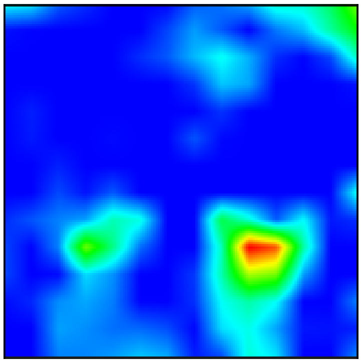

Figure 95. Top 1024 (Original Grad-CAM) for Q: What color is her top? A: white

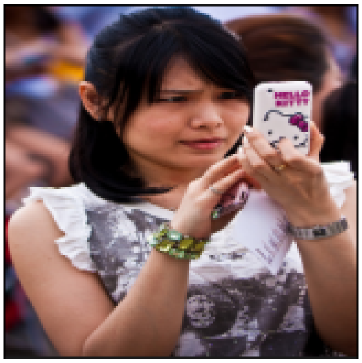

Figure 96. Original Image

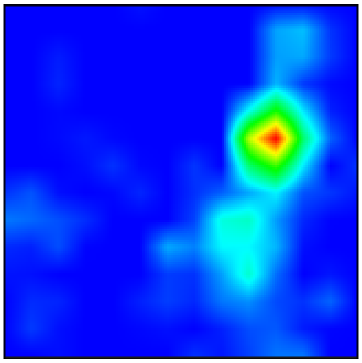

Figure 97. Top 25 for Q: What color is the phone? A: white

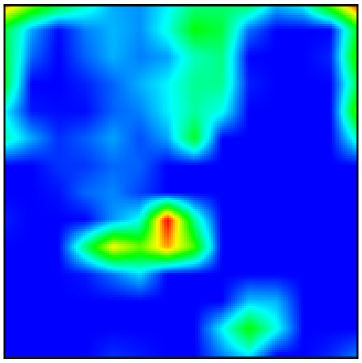

Figure 98. Top 25 for Q: What color is her top? A: white

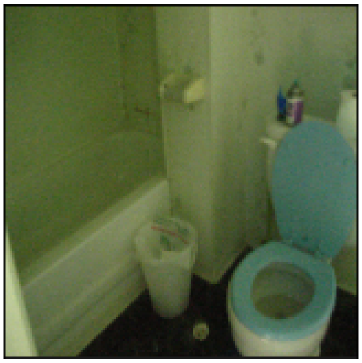

Figure 99. Original Image

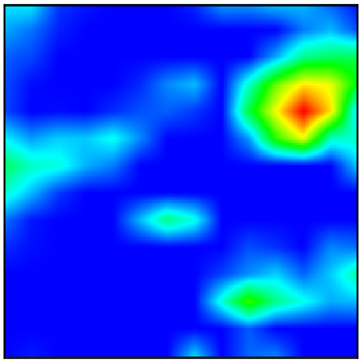

Figure 100. Top 25 for Q: What color are the toilets? A: white

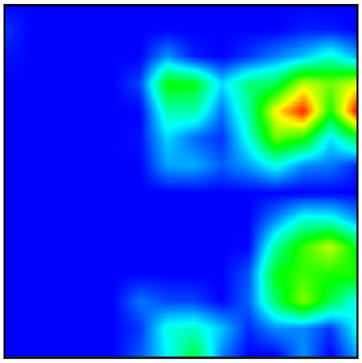

Figure 101. Top 25 for Q: What color is the wall? A: white

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 1

[2] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the winning entry to the vqa challenge 2018, 2018. 1

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. 1

[4] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015. 3