Supplementary Material: Analyzing the Behavior of Visual Question Answering Models

Aishwarya Agrawal[*], **Dhruv Batra**[†,*], **Devi Parikh**[†,*]
[*]Virginia Tech [†]Georgia Institute of Technology
`{aish, dbatra, parikh}@vt.edu`

1 Overview

In this supplementary material, we provide:

1. Behavioral analysis for question-only and image-only VQA models (Section 2).
2. Scatter plot of average distance of test instances from nearest neighbor training instances w.r.t. VQA accuracy (Section 3).
3. Additional qualitative examples for "generalization to novel test instances" (Section 4).
4. The analyses on "complete question understanding" for different question types (Section 5).
5. Additional qualitative examples for "complete question understanding" (Section 6).
6. The analyses on "complete image understanding" for different question types (Section 7).
7. Additional qualitative examples for "complete image understanding" (Section 8).

2 Behavioral analysis for question-only and image-only VQA models

We evaluated the performance of both CNN+LSTM and ATT models by just feeding in the question (and mean image embedding) and by just feeding in the image (and mean question embedding). We computed the percentage of responses that change on feeding the question as well, compared to only feeding in the image and the percentage of responses that change on feeding the image as well, compared to only feeding in the question. We found that that the responses changed much more (about 40% more) on addition of the question than they did on addition of the image. So this suggests that the VQA models are heavily driven by question rather than the image.

3 Scatter plot of average distance of test instances from nearest neighbor training instances w.r.t. VQA accuracy

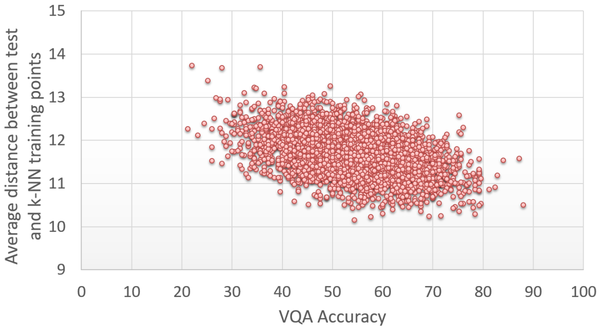


Figure 1: Test accuracy vs. average distance of the test points from k-NN training points for the CNN+LSTM model.

Fig. 1 shows the variation of accuracy of test point w.r.t their average distance from k-NN training points for the CNN+LSTM model. Each point in the plot represents average statistics (accuracy and average distance) for a random subset of 25 test points. We can see that for the test points with low accuracy, the average distance is higher compared to test points with high accuracy. The correlation between accuracy and average distance is significant (-0.41 at $k = 50$.[1])

---

[1] $k = 50$ leads to highest correlation

4 Additional qualitative examples for “generalization to novel test instances”

Fig. 2 shows test QI pairs for which the CNN+LSTM model produces the correct response and their nearest neighbor QI pairs from training set. It can be seen that the nearest neighbor QI pairs from the training set are similar to the test QI pair. In addition, the GT labels in the training set are similar to the test GT label.

Fig. 3 shows test QI pairs for which the CNN+LSTM model produces incorrect response and their nearest neighbor QI pairs from training set. Some of the mistakes are probably because the test QI pair does not have similar QI pairs in the training set (rows 2, 4 and 5) while other mistakes are probably because the GT labels in the training set are not similar to the GT test label (rows 1 and 3).

5 Analyses on “complete question understanding” for different question types

We show the breakdown of our analyses from the main paper – (i) whether the model ‘listens’ to the entire question; and (ii) which POS tags matter the most – over the three major categories of questions – “yes/no”, “number” and “other” as categorized in (Antol et al., 2015). “yes/no” are questions whose answers are either “yes” or “no”, “number” are questions whose answers are numbers (e.g., “Q: How many zebras are there?”, “A: 2”), “other” are rest of the questions.

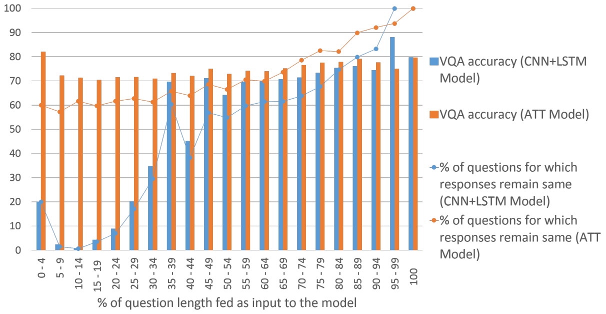


Figure 4: X-axis shows length of partial “yes/no” question (in %) fed as input. Y-axis shows percentage of “yes/no” questions for which responses of these partial “yes/no” questions are the same as full “yes/no” questions and VQA accuracy of partial “yes/no” questions.

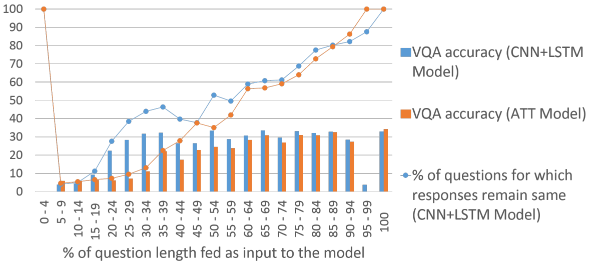


Figure 5: X-axis shows length of partial “number” question (in %) fed as input. Y-axis shows percentage of “number” questions for which responses of these partial “number” questions are the same as full “number” questions and VQA accuracy of partial “number” questions.

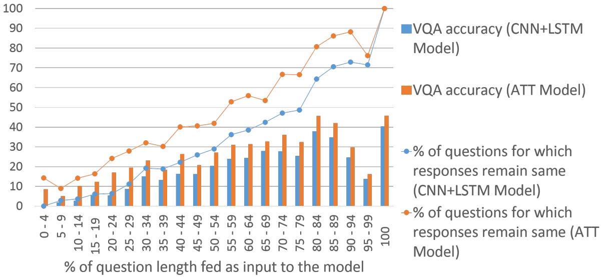


Figure 6: X-axis shows length of partial “other” question (in %) fed as input. Y-axis shows percentage of “other” questions for which responses of these partial “other” questions are the same as full “other” questions and VQA accuracy of partial “other” questions.

For “yes/no” questions, the ATT model seems particularly ‘jumpy’ – converging on a predicted answer listening to only the first few words of the question (see Fig. 4). Surprisingly, the accuracy is also as much as the final accuracy (after listening to entire question) when making predictions based on first few words of the question. In contrast, the CNN+LSTM model converges on a predicted answer later, after listening to atleast 35% of the question, achieving as much as the final accuracy after convergence. For “number” and “other” questions, both ATT and CNN+LSTM model show similar trends (see Fig. 5 for “number” and Fig. 6 for “other”).

It is interesting to note that VQA models are most sensitive to adjectives for “yes/no” questions (compared to wh-words for all questions) (see Fig. 7). This is probably because often the “yes/no” questions are about attributes of objects (e.g., “Is the cup empty?”). For “number” questions, the CNN+LSTM model is most sensitive to adjectives whereas the ATT model is most sensitive to wh-words (see Fig. 8). For “other” questions, both the

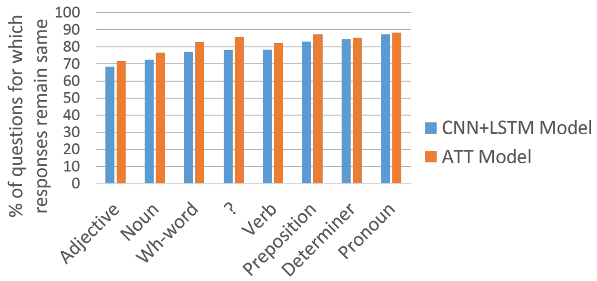


Figure 7: Percentage of "yes/no" questions for which responses remain same (compared to entire "yes/no' question) as a function of POS tags dropped from the "yes/no' question.

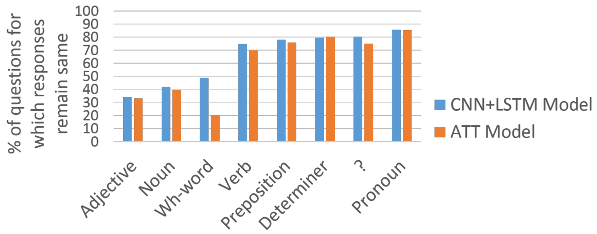


Figure 8: Percentage of "number" questions for which responses remain same (compared to entire "number" question) as a function of POS tags dropped from the "number" question.

models are most sensitive to "nouns" (see Fig. 9).

6 Additional qualitative examples for "complete question understanding"

Fig. 10 shows examples where the CNN+LSTM model converges on a predicted answer without listening to the entire question. On doing so, the model gets the answer correct for some QI pairs (first three rows) and incorrect for others (last two rows).

7 Analyses on "complete image understanding" for different question types

Fig. 11, Fig. 12 and Fig. 13 show the breakdown of percentage of questions for which the model produces same answer across images for "yes/no", "number" and "other" respectively. The ATT model seems to be more "stubborn" (does not change its answers across images) for "yes/no" questions compared to the CNN+LSTM model, and less "stubborn" for "number" questions compared to the CNN+LSTM model.

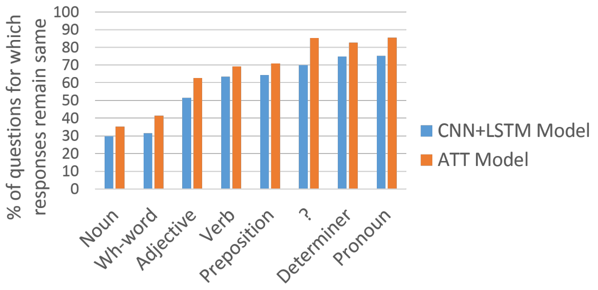


Figure 9: Percentage of "other" questions for which responses remain same (compared to entire "other" question) as a function of POS tags dropped from the "other" question.

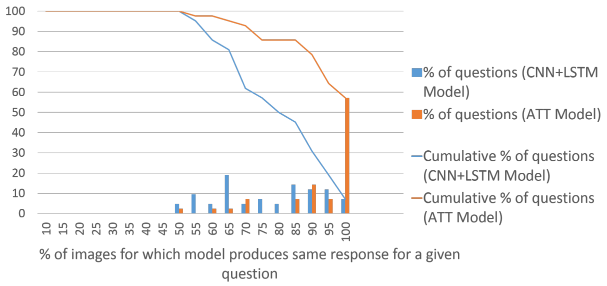


Figure 11: Histogram of percentage of images for which model produces same answer for a given "yes/no" question. The cumulative plot shows the % of "yes/no" questions for which model produces same answer for *atleast* x % of images.

8 Additional qualitative examples for "complete image understanding"

Fig. 14 shows examples where the CNN+LSTM model produces the same answer for atleast half the images for a given question and the accuracy achieved by the model for such QI pairs.

References

[Antol et al.2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*. 2

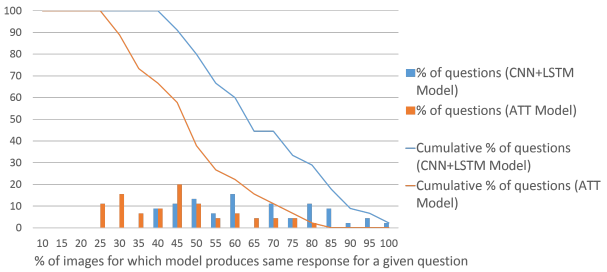


Figure 12: Histogram of percentage of images for which model produces same answer for a given "number" question. The cumulative plot shows the % of "number" questions for which model produces same answer for *atleast* x % of images.

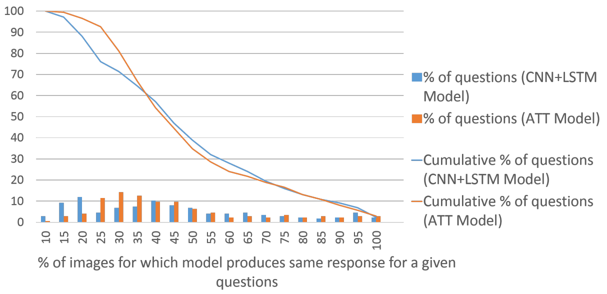
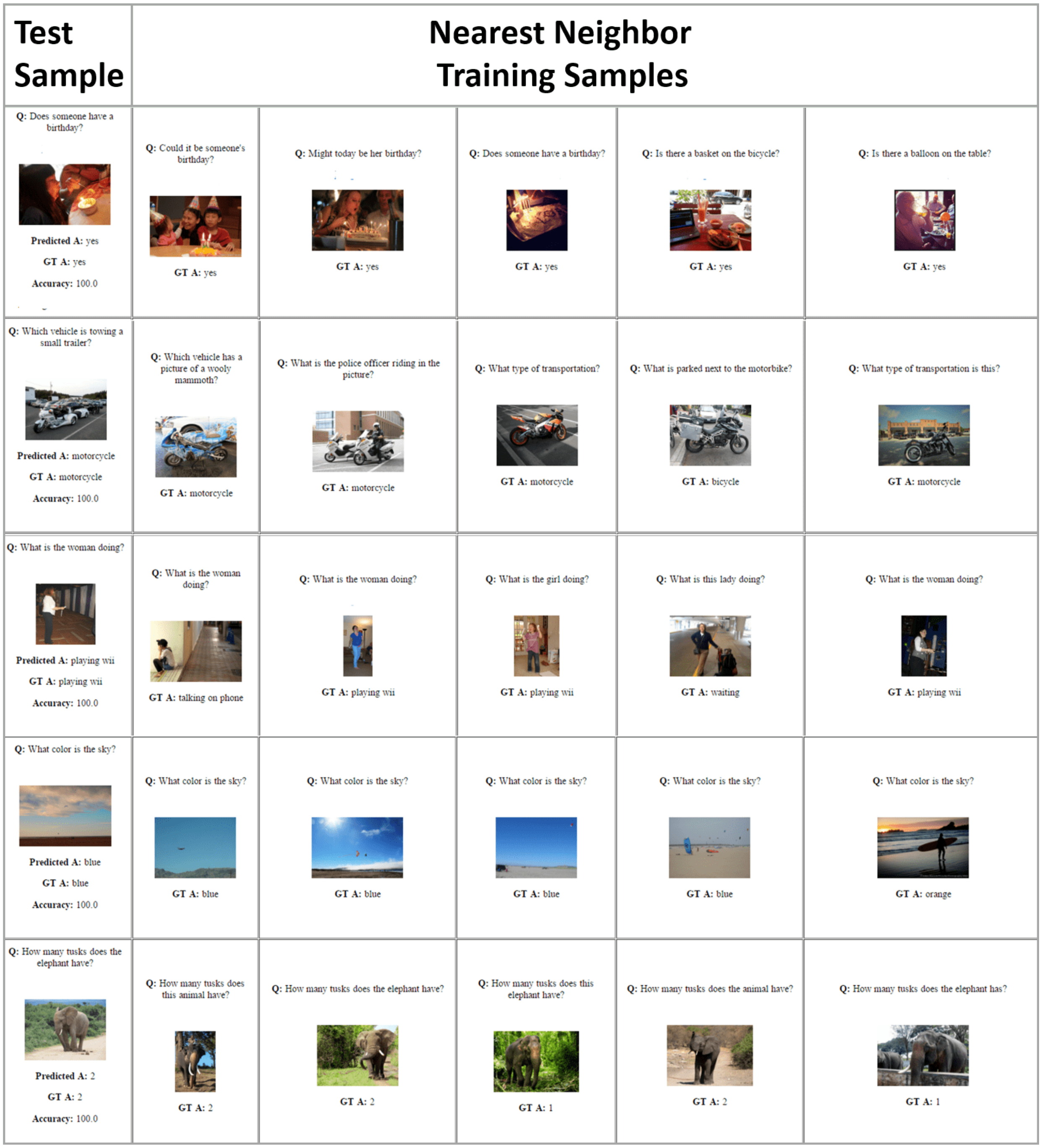


Figure 13: Histogram of percentage of images for which model produces same answer for a given "other" question. The cumulative plot shows the % of "other" questions for which model produces same answer for *atleast* x % of images.



Test Sample	Nearest Neighbor Training Samples				
Q: Does someone have a birthday? **Predicted A:** yes **GT A:** yes **Accuracy:** 100.0	**Q:** Could it be someone's birthday? **GT A:** yes	**Q:** Might today be her birthday? **GT A:** yes	**Q:** Does someone have a birthday? **GT A:** yes	**Q:** Is there a basket on the bicycle? **GT A:** yes	**Q:** Is there a balloon on the table? **GT A:** yes
Q: Which vehicle is towing a small trailer? **Predicted A:** motorcycle **GT A:** motorcycle **Accuracy:** 100.0	**Q:** Which vehicle has a picture of a wooly mammoth? **GT A:** motorcycle	**Q:** What is the police officer riding in the picture? **GT A:** motorcycle	**Q:** What type of transportation? **GT A:** motorcycle	**Q:** What is parked next to the motorbike? **GT A:** bicycle	**Q:** What type of transportation is this? **GT A:** motorcycle
Q: What is the woman doing? **Predicted A:** playing wii **GT A:** playing wii **Accuracy:** 100.0	**Q:** What is the woman doing? **GT A:** talking on phone	**Q:** What is the woman doing? **GT A:** playing wii	**Q:** What is the girl doing? **GT A:** playing wii	**Q:** What is this lady doing? **GT A:** waiting	**Q:** What is the woman doing? **GT A:** playing wii
Q: What color is the sky? **Predicted A:** blue **GT A:** blue **Accuracy:** 100.0	**Q:** What color is the sky? **GT A:** blue	**Q:** What color is the sky? **GT A:** blue	**Q:** What color is the sky? **GT A:** blue	**Q:** What color is the sky? **GT A:** blue	**Q:** What color is the sky? **GT A:** orange
Q: How many tusks does the elephant have? **Predicted A:** 2 **GT A:** 2 **Accuracy:** 100.0	**Q:** How many tusks does this animal have? **GT A:** 2	**Q:** How many tusks does the elephant have? **GT A:** 2	**Q:** How many tusks does this elephant have? **GT A:** 1	**Q:** How many tusks does the animal have? **GT A:** 2	**Q:** How many tusks does the elephant has? **GT A:** 1

Figure 2: Test QI pairs for which the CNN+LSTM model produces the correct response and their nearest neighbor QI pairs from training set.

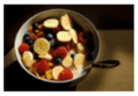
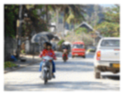
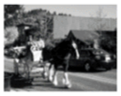
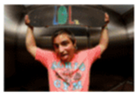
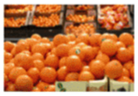
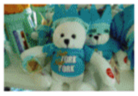
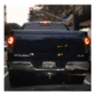
Test Sample	**Nearest Neighbor Training Samples**				
Q: What kind of food is this?  **Predicted A:** dessert **GT A:** cereal with fruit **Accuracy:** 0.0	**Q:** What kind of food is this? **GT A:** dessert	**Q:** What kind of food is this? **GT A:** pizza	**Q:** What kind of food is this? **GT A:** lunch	**Q:** What type of food is this? **GT A:** pizza	**Q:** What kind of food is this? **GT A:** salad
Q: What is red and driving down the road?  **Predicted A:** car **GT A:** bus **Accuracy:** 0.0	**Q:** What is the back of the motorbike? **GT A:** box	**Q:** What is on the pole behind the bike? **GT A:** sign	**Q:** What is around the corner to the right? **GT A:** store	**Q:** What is the bike locked up to? **GT A:** tree	**Q:** What is green and behind the people? **GT A:** trees
Q: What breed of horse is this?  **Predicted A:** black and white **GT A:** clydesdale **Accuracy:** 0.0	**Q:** What breed of horse is this? **GT A:** brown	**Q:** What kind of horse is this? **GT A:** brown	**Q:** What kind of horse is this? **GT A:** brown	**Q:** What type of horse is this? **GT A:** brown	**Q:** Which kind of horse is this? **GT A:** brown
Q: Is this Miley Cyrus?  **Predicted A:** yes **GT A:** no **Accuracy:** 0.0	**Q:** Is the train blue? **GT A:** yes	**Q:** Does this look right? **GT A:** no	**Q:** Is the skateboard flying? **GT A:** yes	**Q:** Is the bus driver visible? **GT A:** no	**Q:** Is the person female? **GT A:** yes
Q: What is the name a state that grows these fruits?  **Predicted A:** new york **GT A:** florida **Accuracy:** 0.0	**Q:** What state is the can from? **GT A:** new york	**Q:** What state is the mug from? **GT A:** new york	**Q:** What face does the topmost fruit have? **GT A:** happy	**Q:** What state is the bear representing?  **GT A:** new york	**Q:** What state is the truck from?  **GT A:** california

Figure 3: Test QI pairs for which the CNN+LSTM model produces incorrect response and their nearest neighbor QI pairs from training set.

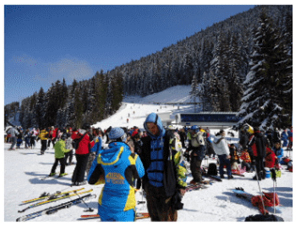
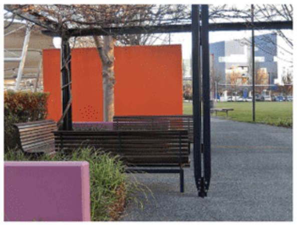
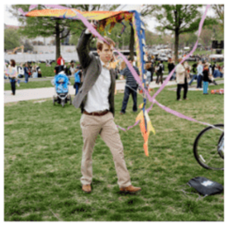
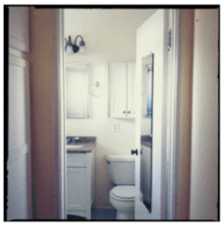
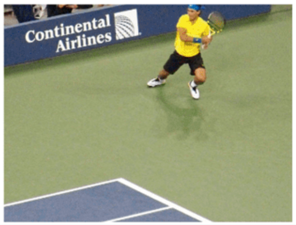
	**GT A:** no **Accuracy of predicted answer for full question: 100.0**	**Q:** Is there a tram to the west of where the people are? **A:** yes **Q:** Is **A:** outside **Q:** Is there **A:** beach **Q:** Is there a **A:** yes **Q:** Is there a tram **A:** yes **Q:** Is there a tram to **A:** yes **Q:** Is there a tram to the **A:** yes **Q:** Is there a tram to the west **A:** yes **Q:** Is there a tram to the west of **A:** yes **Q:** Is there a tram to the west of where **A:** yes **Q:** Is there a tram to the west of where the **A:** yes **Q:** Is there a tram to the west of where the people **A:** yes **Q:** Is there a tram to the west of where the people are? **A:** yes **Q:** Is there a tram to the west of where the people are? **A:** yes
	**GT A:** 3 **Accuracy of predicted answer for full question: 90.0**	**Q:** How many different directions are the benches facing? **A:** 2 **Q:** How **A:** yes **Q:** How many **A:** 2 **Q:** How many different **A:** 2 **Q:** How many different directions **A:** 2 **Q:** How many different directions are **A:** 2 **Q:** How many different directions are the **A:** 2 **Q:** How many different directions are the benches **A:** 2 **Q:** How many different directions are the benches facing? **A:** 2 **Q:** How many different directions are the benches facing? **A:** 2
	**GT A:** grass **Accuracy of predicted answer for full question: 100.0**	**Q:** What type of surface is the man standing on? **A:** grass **Q:** What **A:** umbrellas **Q:** What type **A:** shadow **Q:** What type of **A:** kite **Q:** What type of surface **A:** grass **Q:** What type of surface is **A:** grass **Q:** What type of surface is the **A:** grass **Q:** What type of surface is the man **A:** grass **Q:** What type of surface is the man standing **A:** grass **Q:** What type of surface is the man standing on? **A:** grass **Q:** What type of surface is the man standing on? **A:** grass
	**GT A:** bathroom **Accuracy of predicted answer for full question: 0.0**	**Q:** Where is the light fixture in the photo? **A:** window **Q:** Where **A:** bathroom **Q:** Where is **A:** outside **Q:** Where is the **A:** bathroom **Q:** Where is the light **A:** counter **Q:** Where is the light fixture **A:** on left **Q:** Where is the light fixture in **A:** window **Q:** Where is the light fixture in the **A:** window **Q:** Where is the light fixture in the photo? **A:** window **Q:** Where is the light fixture in the photo? **A:** window
	**GT A:** continental airlines **Accuracy of predicted answer for full question: 0.0**	**Q:** What company is a sponsor of this match? **A:** polo **Q:** What **A:** shadow **Q:** What company **A:** nike **Q:** What company is **A:** polo **Q:** What company is a **A:** polo **Q:** What company is a sponsor **A:** polo **Q:** What company is a sponsor of **A:** polo **Q:** What company is a sponsor of this **A:** polo **Q:** What company is a sponsor of this match? **A:** polo **Q:** What company is a sponsor of this match? **A:** polo

Figure 10: Examples where the CNN+LSTM model converges on a predicted answer without listening to the entire question.

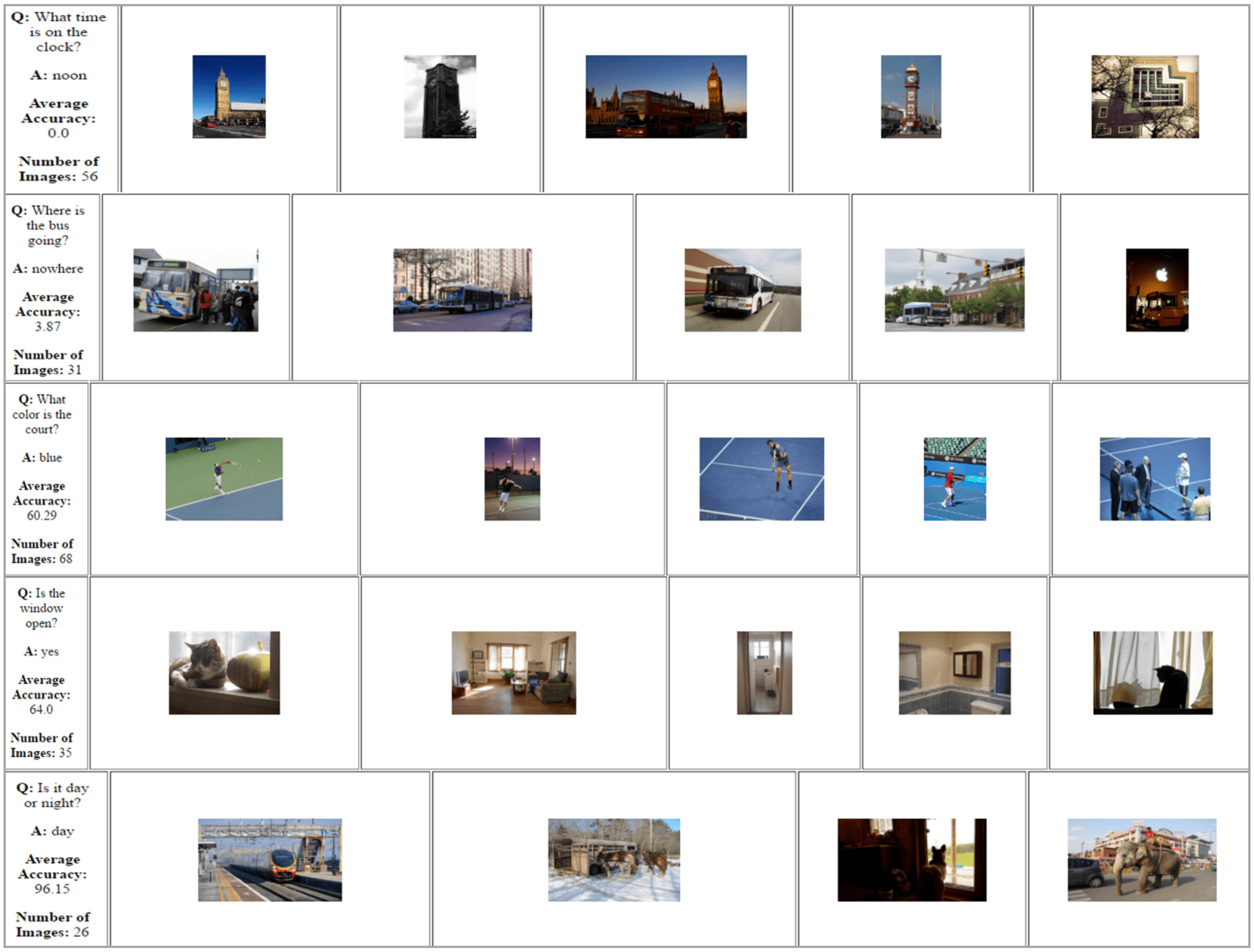


Figure 14: Examples where the CNN+LSTM model produces the same answer for atleast half the images for each of the questions shown above. “Q” denotes the question for which model produces same response for atleast half the images, “A” denotes the answer predicted by the model (which is same for atleast half the images), “Number of Images” denotes the number of images for which the question is repeated in the VQA validation set and “Average Accuracy” is the VQA accuracy for these QI pairs (with same question but different images).