

Semantic Equivalent Adversarial Data Augmentation for Visual Question Answering Supplementary Material

Ruixue Tang¹, Chao Ma^{1*}, Wei Emma Zhang², Qi Wu², and Xiaokang Yang¹

¹ MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{alicetang, chaoma, xkyang}@sjtu.edu.cn

² University of Adelaide
{wei.e.zhang, qi.wu01}@adelaide.edu.au

Additional Results of Augmented Questions

We provide more qualitative examples of augmented questions generated by our method with edit distance threshold $e = 4$ and top two semantic score. Questions that cause the answers to flip are presented in Fig. 2 and questions that preserve the answers are presented in Fig. 1.











	Are they all planning to skate? (no) Ground truth: no (0.034) Are they all going to skate? (no) (0.013) Are they all planning skating? (no)		Are all the dogs looking in the same direction? (no) Ground truth: yes (0.032) Are all the dogs looking at the same direction? (no) (0.017) Do all the dogs look at the same direction? (no)
	Are there bed headboards present in the photo? (no) Ground truth: no (0.066) Are there bed headboards in the picture? (no) (0.058) Are there some headboards in the picture? (no)		What size mattress would you need for this bed? (twin) Ground truth: twin (0.046) Which size would you need for this bed? (twin) (0.015) How big a mattress would you need for this bed? (twin)
	Are there lights on in the two buildings? (no) Ground truth: yes (0.032) Are there any lights on the two buildings? (no) (0.009) Are there lights on the two buildings? (no)		What is the name for a room like this? (kitchen) Ground truth: kitchen (0.092) What is the name of a room like this? (kitchen) (0.049) What is the name for a room like that? (kitchen)
	Are these people dining together? (yes) Ground truth: yes (0.119) Are these people eating together? (yes) (0.087) Are those people eating together? (yes)		What is moving on the street? (bus) Ground truth: bus (0.083) What is moving in the street? (bus) (0.074) What is moving across the street? (bus)
	Is anyone crossing the bridge? (no) Ground truth: yes (0.088) Is anybody crossing the bridge? (no) (0.057) Is anyone walking across the bridge? (no)		Is the seagull in danger of getting entangled in these boat sails? (no) Ground truth: no (0.017) Is the seagull in danger of getting stuck in these boat sails? (no) (0.009) Is the seagull in danger of getting stuck in this boat? (no)

Fig. 1. Examples of our generated questions that preserve the answers. The first question in bold in each block is the original question. The words in brackets are model predictions of the corresponding question with positive predictions in green and negative predictions in red. The numbers in brackets are the semantic scores of generated questions.

* Corresponding author.





















	<p>What flag is being held up on the boat? (british) Ground truth: american (0.022) What flag is held on the boat? (american) (0.020) Which flag is being held on the boat? (american)</p>		<p>Is the woman wearing a backpack or shoulder bag? (bag) Ground truth: shoulder bag (0.013) Is women wearing a backpack or a shoulder bag ? (woman) (0.007) Does the woman wear a backpack or a shoulder bag ? (both)</p>
	<p>How is this keyboard unlike a typical keyboard? (wireless) Ground truth: wireless (0.070) How does this keyboard differ from a typical keyboard? (yes) (0.061) How can this keyboard be unlike a typical keyboard? (yes)</p>		<p>How many dishes of food are in the picture? (6) Ground truth: 19 (0.088) How many food dishes are there in the picture? (7) (0.033) How many food plates are in the picture? (7)</p>
	<p>Is the giraffe a baby? (yes) Ground truth: yes (0.049) Is the giraffe an infant? (no) (0.029) Is this giraffe a baby? (no)</p>		<p>Could this be a hotel room? (yes) Ground truth: no (0.025) Can this be a hotel room? (no) (0.017) Can it be a hotel room? (no)</p>
	<p>How many blue benches are visible in this photo? (3) Ground truth: 2 (0.065) How many blue benches are visible in that picture? (2) (0.045) How many blue benches are seen in this picture? (2)</p>		<p>Is the power connected to the laptop? (no) Ground truth: yes (0.055) Is the power connected to a laptop? (yes) (0.015) Is the power connected to laptop? (yes)</p>
	<p>What direction is the sun? (west) Ground truth: right (0.069) Which direction is the sun? (left) (0.013) What direction is sunlight? (right)</p>		<p>Are people in the crosswalk? (yes) Ground truth: yes (0.017) Are the people in the crosswalk? (no) (0.009) Are people in a crosswalk? (no)</p>
	<p>Are the people getting on the bus or off the bus? (off) Ground truth: on (0.061) Are people getting on the bus or off the bus? (on) (0.042) Are people getting on the bus or out of the bus? (no)</p>		<p>What is the woman doing with the cell phone? (talking on phone) Ground truth: talking (0.076) What is the woman doing with the mobile phone? (talking) (0.041) What is the woman doing with the cellphone? (talking)</p>
	<p>How many engines are visible on the plane? (4) Ground truth: 2 (0.018) How many motors are visible on the plane? (2) (0.006) How many engines are visible in the plane? (2)</p>		<p>What does the word on the front of the bus mean? (fire) Ground truth: bus (0.066) What does the word on the front of the bus means? (stop) (0.051) What does the word in front of the bus mean? (stop)</p>
	<p>Why is the equipment on the side of the road? (parked) Ground truth: parked (0.074) Why is the equipment on the road side? (yes) (0.059) Why are the equipment on the side of the road? (yes)</p>		<p>Are these kids having fun on the pier? (no) Ground truth: yes (0.089) Do these children have fun on the pier? (yes) (0.062) Do these kids have fun on the pier? (yes)</p>
	<p>What does the man's facial expression suggest? (happy) Ground truth: angry (0.066) What does the facial expression of man suggest? (smiling) (0.051) What does man's facial expression suggest? (smiling)</p>		<p>Has the ball left the pitcher's hand? (yes) Ground truth: yes (0.077) Has the ball left the hand of the pitcher? (no) (0.043) Did the ball leave the pitcher's hand? (no)</p>
	<p>Has anyone eaten a slice of this pizza yet? (yes) Ground truth: no (0.078) Has anybody eaten a slice of this pizza yet? (no) (0.034) Did someone eat a slice of this pizza yet? (no)</p>		<p>What activities can be held in the building? (nothing) Ground truth: classes (0.057) What activities can be done in the building? (none) (0.023) What activities can be carried out in the building? (none)</p>

Fig. 2. Examples of our generated questions that cause the answers to flip. The first question in bold in each block is the original question. The words in brackets are model predictions of the corresponding question with positive predictions in green and negative predictions in red. The numbers in brackets are the semantic scores of generated questions.