Self Supervision for Attention Networks-Supplementary

Badri N Patro *†
IIT Kanpur
badri@iitk.ac.in

Kasturi G S *
NSUT‡
gs.it@nsit.net.in

Ansh Jain *
NSUT
ansh.it@nsit.net.in

Vinay P Namboodiri
University of Bath
vpn22@bath.ac.uk

1. Analysis of mask generation for text and image

In this section, we conduct experiments to generate both text based supervision and visual supervision for the attention map.

Table 1. Experiment for obtaining Visual (Image) Supervision map for attention

Experiment	# Masks	# Ques	# Images	# Sample
Experiment 1	100	6	2	8
Experiment 2	500	30	10	8
Experiment 3	1000	30	10	12

1.1. Visual (Image) Supervision Experiments

The parameters for different experiments is shown in table 1

- Experiment 1: We conduct an experiment for obtaining Visual (Image) attention map for the baseline model using 100 masks. The result of this experiment is shown in figure- 1. The result shows experiment results with a best answer confidence score and a second-best answer confidence score for given question and image. This experiment shows that the baseline does not attend very well to images, so we increase the number of masks.
- Experiment 2: We conduct another experiment for obtaining Visual (Image) Supervision map for attention with different set of parameters mentioned in table-1. The result of this experiment is shown in figure- 2. The result shows experiment results with a best answer confidence score and a second-best answer confidence score for given question and image. We obtained better results than those obtained from 100 masks. We still obtained vague answers for some question-image pairs, so we decided to increase the number of masks further.
- Experiment 3: We conduct another experiment for obtaining Visual (Image) Supervision map for attention with different set of parameters mentioned in table-1. The result of this experiment is shown in figure-3. The result shows experiment results with a best answer confidence score and a second-best answer confidence score for a given question and image. With 1000 masks, we got better results, but still, the supervision mask is not suitable for many images.

We hypothesized that this might be due to the reason that the model was learning more from questions rather than the image. So to further check our hypothesis, we applied masks for questions.

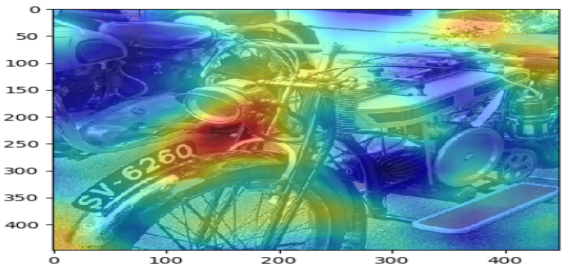
1.2. Text (Question) Supervision Experiment

For this task, we find the importance map of a question and ask each masked question from an image and check the output. This task would help us find the most important words in a question for an answer. In this section, we provide results on question supervision map generation. The analysis results are provide in the figure-4.

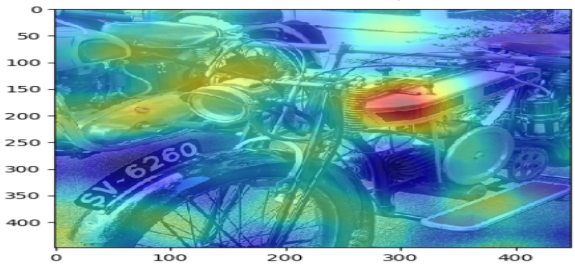
*Equal contribution
†Currently working at Google
‡Netaji Subhas University of Technology

Example 1: Q-: Is this a bike or a motorcycle?

(a) BEST ANSWER : bike (71.6%)

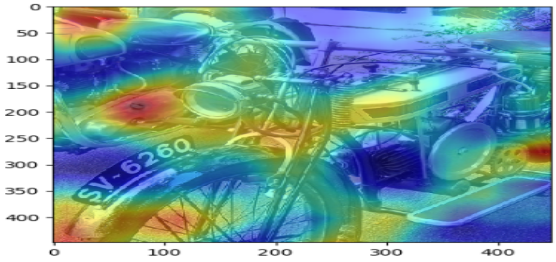


(b) 2nd BEST ANSWER : motorcycle (18.4%)

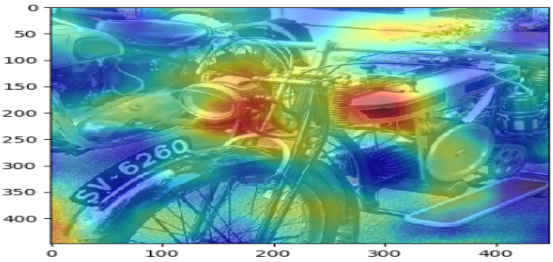


Example 2: Q-: What color is the bike?

(a) BEST ANSWER : black (24.8%)

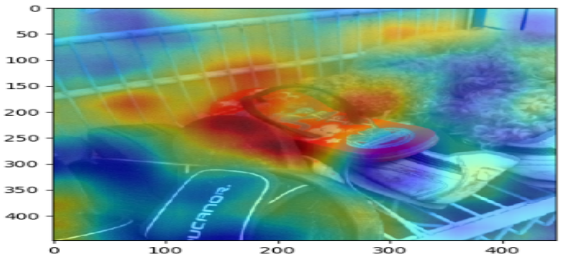


(b) 2nd BEST ANSWER : silver (22.5%)

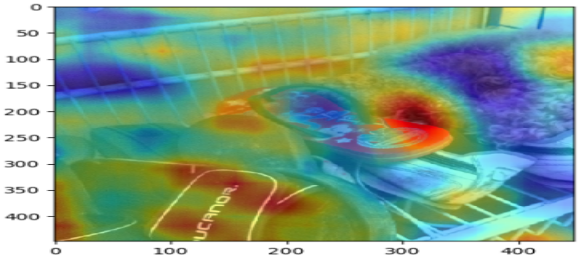


Example 3: Q-: What color are the gym shoes?

(a) BEST ANSWER : red (15.52%)

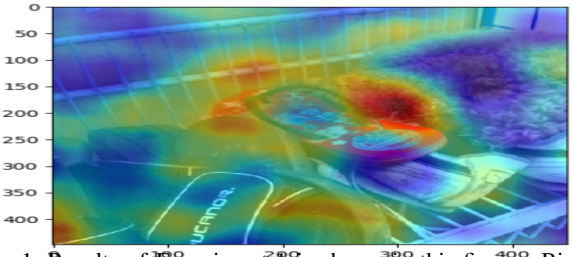


(b) 2nd BEST ANSWER : red and blue (10.18%)



Example 4: Q-: Is the red sandal here?

(a) BEST ANSWER : yes (80.6%)



(b) 2nd BEST ANSWER : no (19.19%)

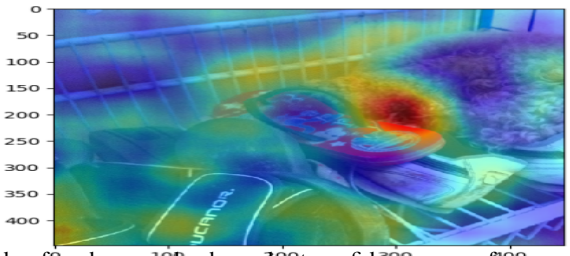
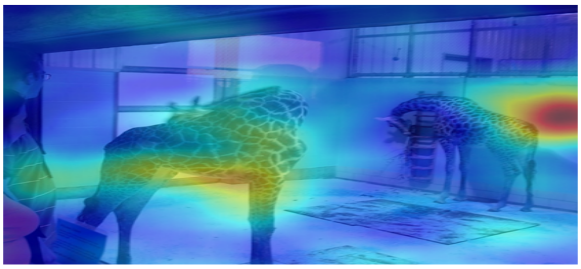
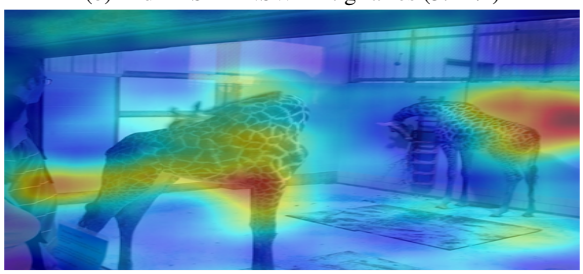


Figure 1. Results of Experiment-1 is shown in this figure. Right side of each example shows best confidence score for answer class probability and left side shows second best confidence score for answer class probability.

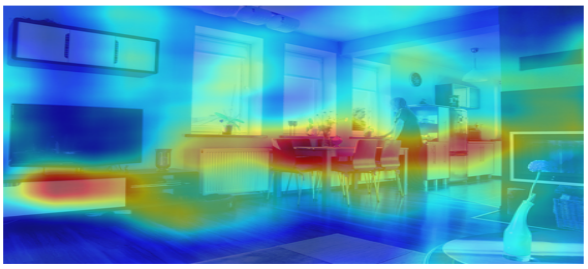
Example 1: Q-: What kind of animal is shown?
(a) BEST ANSWER : giraffe (84.5%)



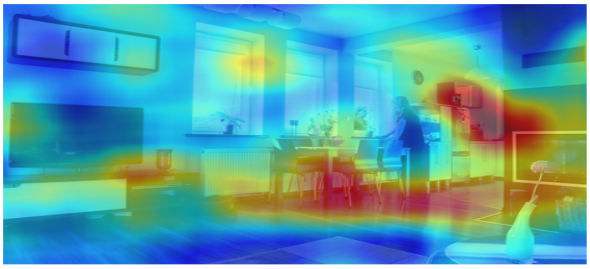
(b) 2nd BEST ANSWER : giraffes (5.21%)



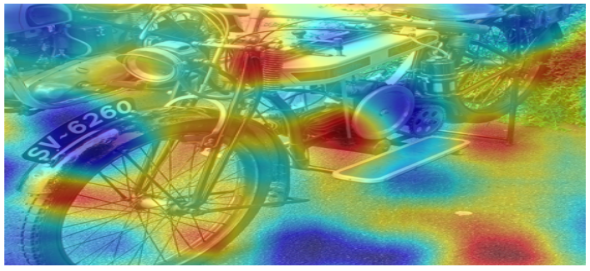
Example 2: Q-: what is the woman in room doing?
(a) BEST ANSWER : sitting (45.4%)



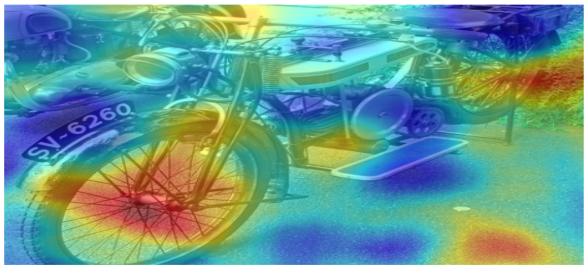
(b) 2nd BEST ANSWER : standing (8.4%)



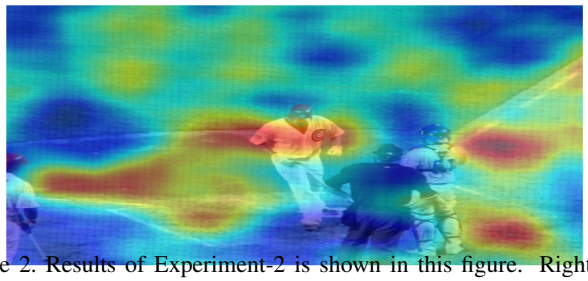
Example 3: Q-: What color is the bike?
(a) BEST ANSWER : red (45.52%)



(b) 2nd BEST ANSWER : red and blue (8.18%)



Example 4: Q-: what sport is this?
(a) BEST ANSWER : baseball (78.2%)



(b) 2nd BEST ANSWER : baseball game (2.02%)

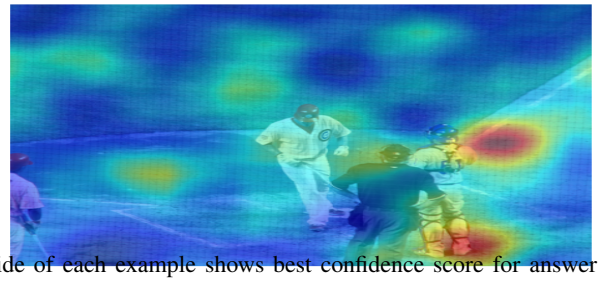
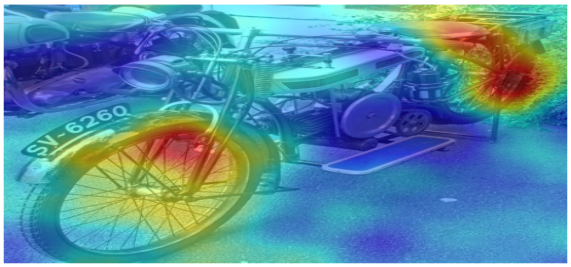
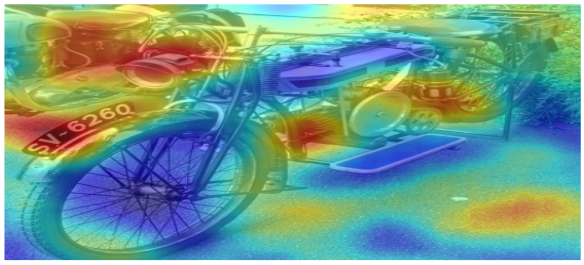


Figure 2. Results of Experiment-2 is shown in this figure. Right side of each example shows best confidence score for answer class probability and left side shows second best confidence score for answer class probability.

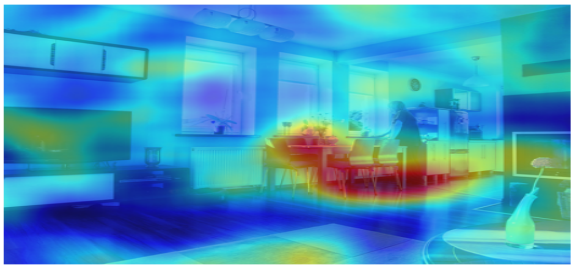
Example 1: Q-: Is this a bike or motorcycle?
(a) BEST ANSWER : bike (71.65%)



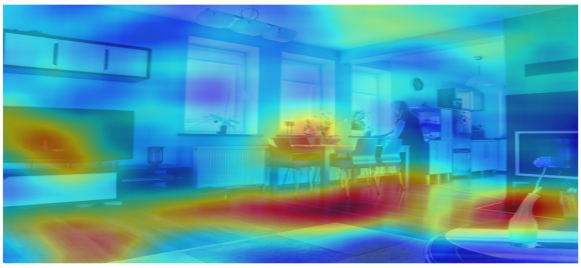
(b) 2nd BEST ANSWER : motorcycle (18.4%)



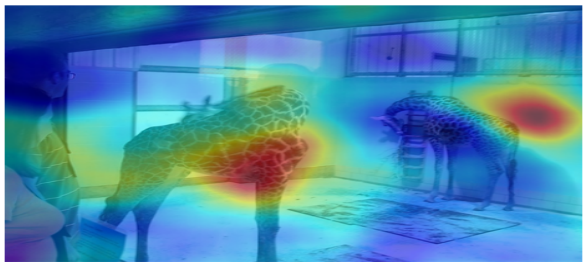
Example 2: Q-: what is the woman in room doing?
(a) BEST ANSWER : sitting (63.14%)



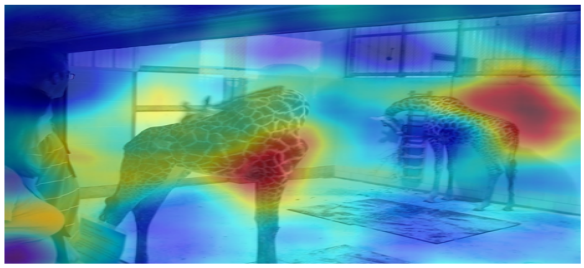
(b) 2nd BEST ANSWER : standing (6.27%)



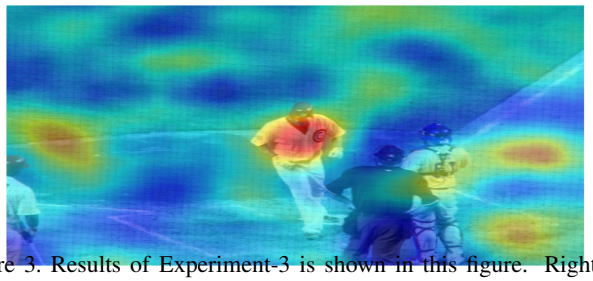
Example 3: Q-: What kind of animal is shown?
(a) BEST ANSWER : giraffe (93.6%)



(b) 2nd BEST ANSWER : giraffes (6.13%)



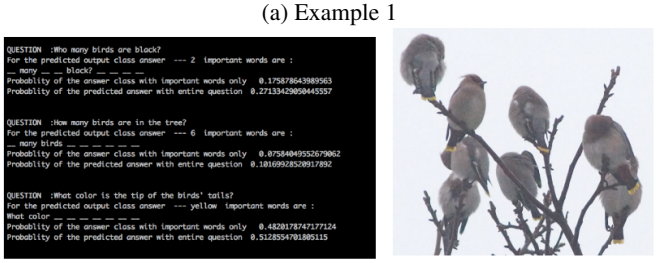
Example 4: Q-: what sport is this?
(a) BEST ANSWER : baseball (99.2%)



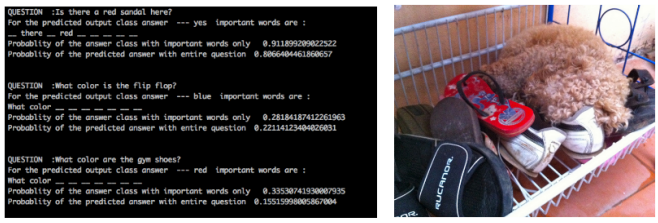
(b) 2nd BEST ANSWER : baseball game (0.02%)

Figure 3. Results of Experiment-3 is shown in this figure. Right side of each example shows best confidence score for answer class probability and left side shows second best confidence score for answer class probability.

(a) Example 1



(b) Example 2



(a) Example 3

(b) Example 4

(a) Example 5

(b) Example 6

Figure 4. Right side of each example shows original image and left side shows importance of each words and corresponding answer prediction probability if only unmasked words are sent as questions