KLE Technological University
Creating Value
Leveraging Knowledge

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Weakly Supervised Visual Question Answer Generation

Charani Alampalle, Shamanthak Hegde, Soumya Jahagirdar, Shankar Gangisetty

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

PROBLEM AND MOTIVATION

- Conversational agents which promote **two-way human-computer communications** unlike one-way chatbots and systems which help **child education by asking and answering visual questions** have become an active area of research in AI.
- Many communities like robotics and VQA have started contributing in this field but they end up generating generic questions. Good QA pair is the one that has a tightly focused purpose and must be relevant to the image content.
- Works in this domain are dependent on large datasets to generate question answer pairs for images.
- **In this work, we address this problem by introducing a method of visual question answer generation in a weakly supervised manner by utilizing the visual(image) and text(caption) information.**

CONTRIBUTIONS

- We propose a method for addressing the problem of weakly supervised QA pair generation by creating **cloze question**, for a given image and its caption.
- We generate a new vocabulary on our generated QA pair and then **fine-tune the ViLBERT model** to get better QA pairs.
- We experimentally evaluated our QA pairs on standard **VQA dataset** and compared our results with state of the art models.

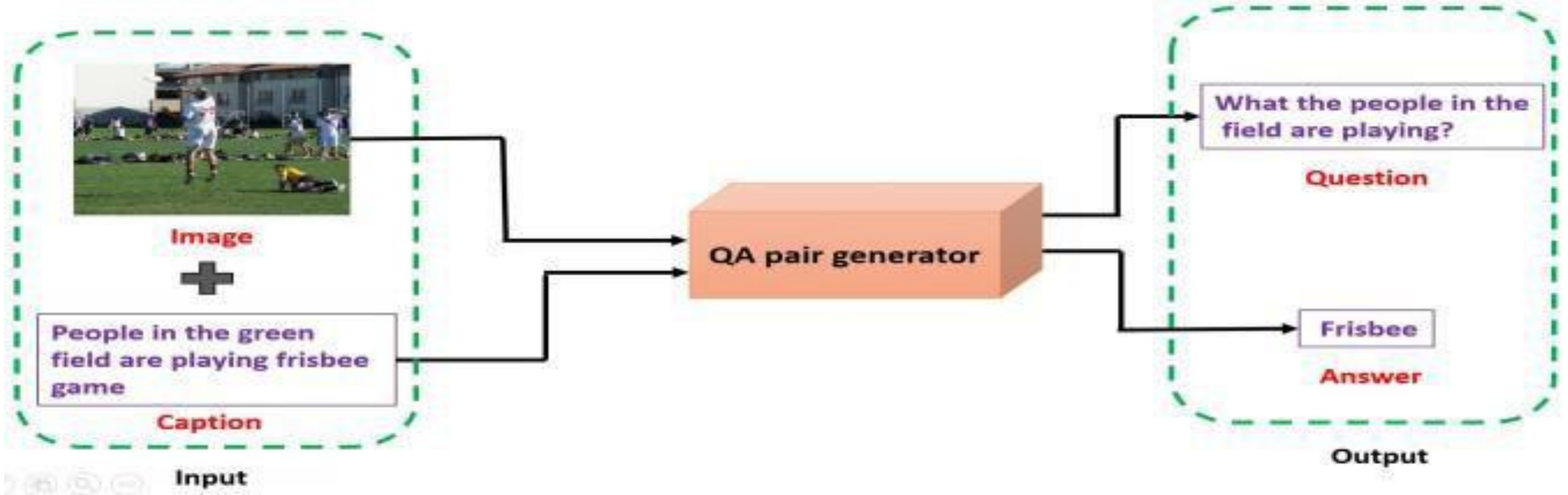


Image
People in the green field are playing frisbee game
Caption
Input
QA pair generator
What the people in the field are playing?
Question
Frisbee
Answer
Output

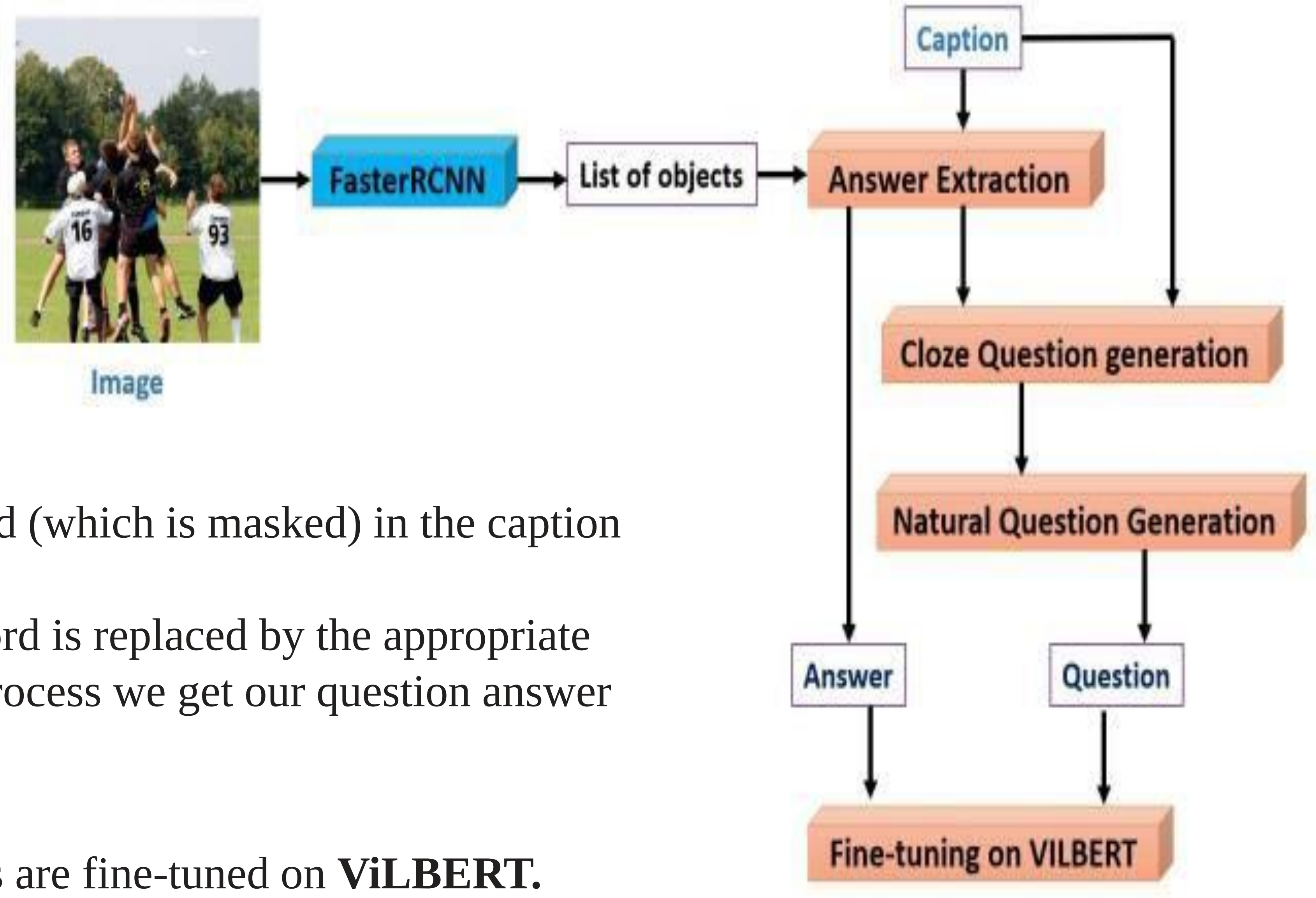
VISUAL QUESTION ANSWER GENERATION

(i) Answer Extraction Module: Given an image and its captions, a list of objects ($O_1, . . . , O_n$) are identified from the image and an answer word is extracted from captions ($C_1, . . . , C_n$) that are part of the list of objects or words identified by **NER and noun chunkers** (W_1 OR W_2,. . . ,W_n).

If O_i in < $C_1, . . . , C_n$ > :
ans = O_i
If O_i not in < $C_1, . . . , C_1$ > :
ans = <W_1 OR W_2 . . . W_n>

(ii) Question Generation Module: We use the method of **Cloze Question generation**: Answer word (which is masked) in the caption is replaced by one of **category word** such as THING, PERSON, LOCATION etc.
Natural Question generation: Using a dependency tree reconstruction method, the category word is replaced by the appropriate **question word.** (THING by "what", PERSON by "who", LOCATION by "where"). After this process we get our question answer pairs that can be used to train a VQA model for question answering task.

(iii) Fine-tuning Module: Based on QA pair generated we create new vocabulary and the QA pairs are fine-tuned on **ViLBERT.**



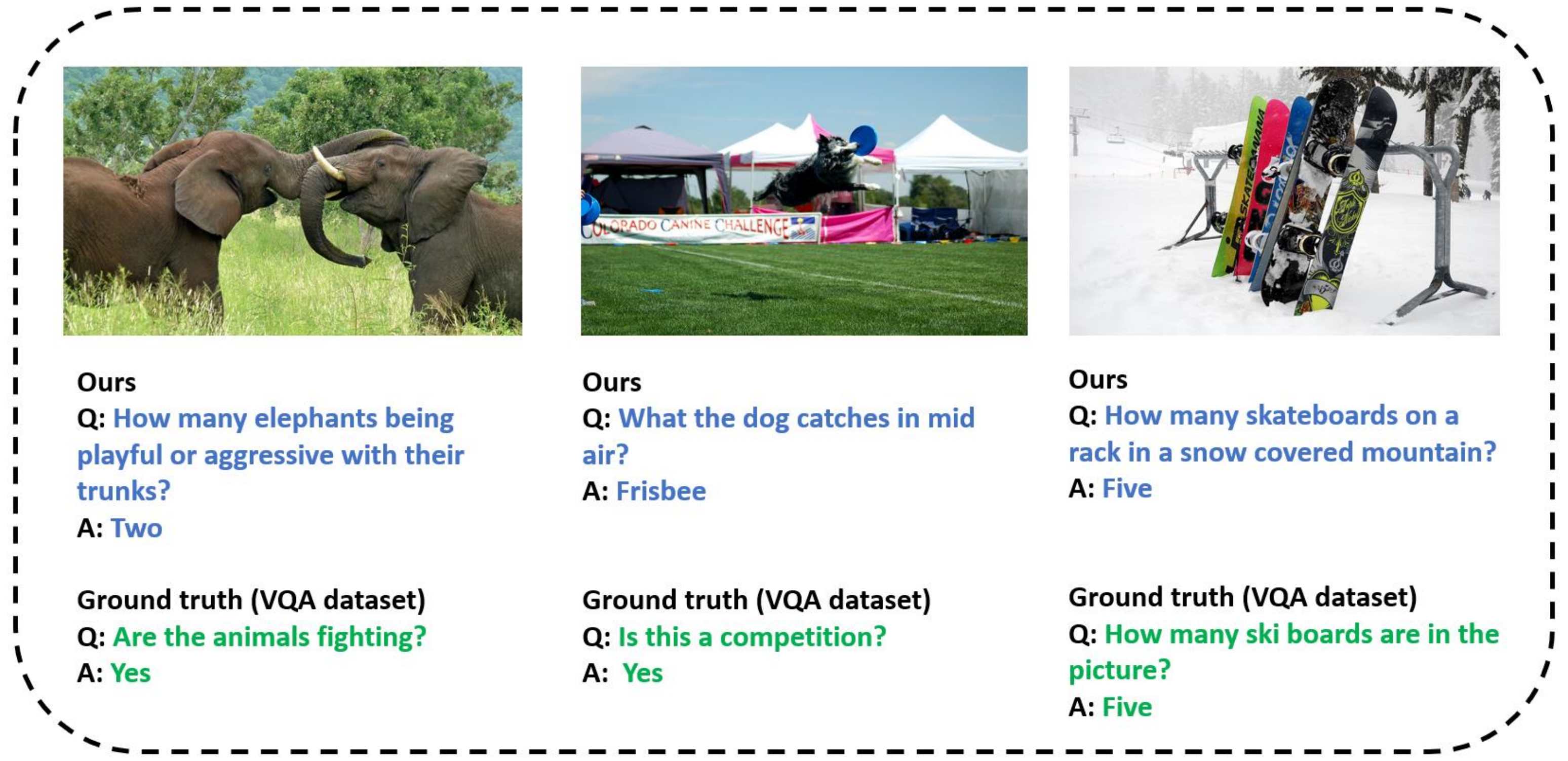
RESULTS

Fig - 1 Comparison of our results with SOTA

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
IA2Q[3]	30.42	13.55	6.23	4.44	9.42	-
V-IA2Q[3]	35.40	25.55	14.94	10.78	13.35	-
IMVQG[2]	31.2	16.20	11.18	6.24	12.11	40.27
C3VQG[1]	41.87	22.11	14.96	10.04	13.60	42.34
Ours	**47.78**	**8.08**	**1.79**	**0.35**	**27.61**	**18.89**

- Our method gave **BLEU score of 47.78** which is more than the SOTA works by value 6 due to the better utilization of both visual and textual information to generate QA pairs.

Fig - 2 Results: QA pairs generated by our method vs VQA



Ours
Q: How many elephants being playful or aggressive with their trunks?
A: Two

Ground truth (VQA dataset)
Q: Are the animals fighting?
A: Yes

Ours
Q: What the dog catches in mid air?
A: Frisbee

Ground truth (VQA dataset)
Q: Is this a competition?
A: Yes

Ours
Q: How many skateboards on a rack in a snow covered mountain?
A: Five

Ground truth (VQA dataset)
Q: How many ski boards are in the picture?
A: Five

- The questions are more detailed and relevant to the image than ground truth.
- The **test accuracy on VQA** after finetuning on ViLBERT is **49.367**.

CONCLUSIONS & FUTURE WORK

We proposed a method of image-based QA pair generation in a weakly supervised manner because of proper utilization of visual information. Our method generates more detailed and relevant QA pair for a given image. **BLEU score value got increased by 6** in our method compared to SOTA. As qualitative and quantitative scores are better by this method, it can be used **to generate large datasets with minimum human efforts.**

REFERENCES

[1] Uppal et al Category Consistent Cyclic Visual Question Generation, *ACM MM Asia 20*

[2] Krishna et. al. Information Maximizing Visual Question Generation, *CVPR 19*

[3] Banerjee et. Al. WeaQA: Weak Supervision via Captions for Visual Question Answering. *IJCNLP 21*