A Gaze-grounded Visual Question Answering Dataset for Clarifying Ambiguous Japanese Questions

Shun Inadumi[1,2], Seiya Kawano[2,1], Akishige Yuguchi[3,2]
Yasutomo Kawanishi[2,1], Koichiro Yoshino[2,1]
1. Nara Institute of Science and Technology
2. Guardian Robot Project, RIKEN, 3. Tokyo University of Science

RIKEN Guardian Robot Project

Dataset link

Motivation

Backgrounds:

- Situated conversations often contain ambiguities [Taniguchi+, AR 2019]
  Caused by **directives** and **ellipsis subjective** or **objective**
- **Referring to user gaze** is a key idea to resolve this problem
  Deal with this problem in VQA setting [Antol+, ICCV 2015]

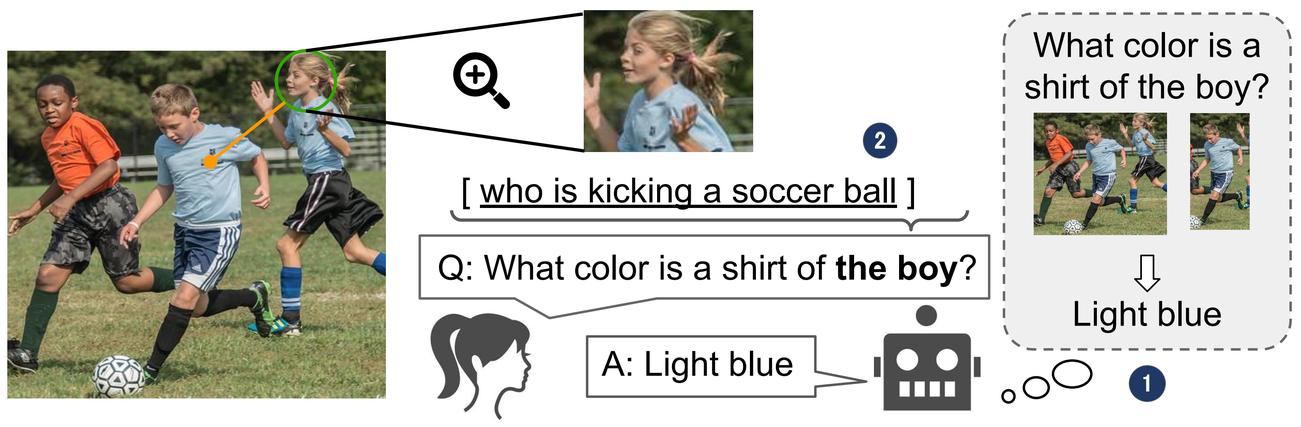
Contributions:

- Construct **a Gaze-grounded VQA Dataset** (**GazeVQA**)
  A VQA task to incorporate with gaze context
- Propose **a VQA model for integrating gaze targets** by extending existing VQA models

Research Questions:

Can gaze context;

1. Improve the accuracy of VQA?
2. Clarify an ambiguous question?

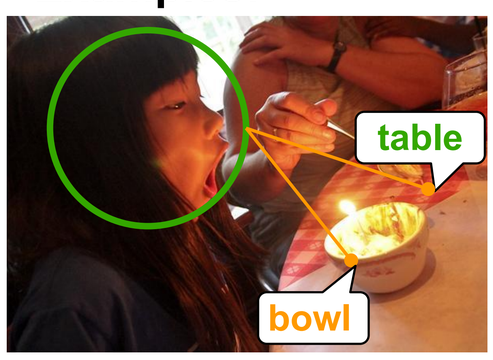


GazeVQA

Overview:

- **10,760 images** :
  An image shows **a first-person view** from the **system side**.
- **17,276 QA pairs** :
  A user asks an ambiguous questions (AQs) about **user gaze targets**.
- **User gaze annotation** :
  Source and target of gaze in GazeFollow [Recasens+, NIPS 2015]
- **Object annotation** :
  Object label in COCO [Lin+, ECCV 2014]

Examples:



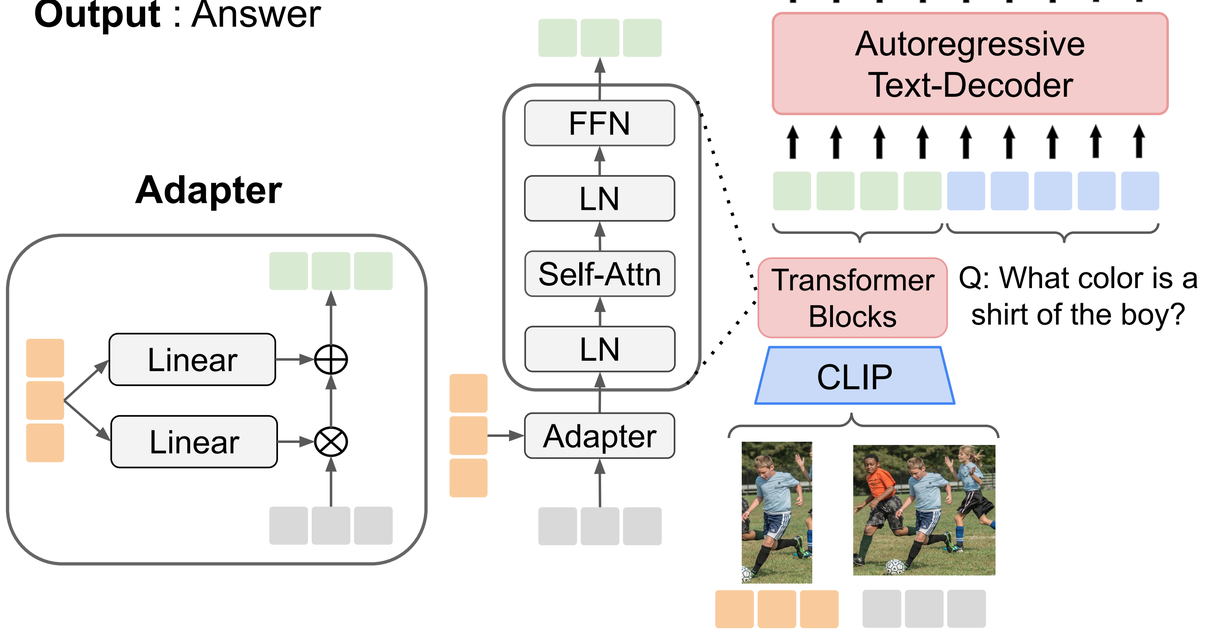
GazeVQA test-set has **clarified questions** (CQs).

Q_1 : What is lit up [in the bowl]?
Q_2 : What is placed on top of it [table]?

A_1 : candle
A_2 : bowl

Model

Inputs : Image, Question, **Gaze RoI**
Output : Answer



- Predict **region of interest** (**RoI**) from source of gaze and image
- **Integrate of image and gaze target features** using **element-wise affine transformation** [Dumoulin+, Distill 2018]

Results

Settings:

Datasets:

- Japanese Captions [Yoshihara+, ACL 2017]
- Japanese VQA [Shimizu+, COLING 2018]
- GazeVQA
  train : valid : test = 13,785 : 1,811 : 1,680

Evaluation Metrics:

- Acc : VQA score
- Bs : BERT score

Baseline: ClipCap [Mokady+, 2021]
CLIP RN×4 [Radford+, ICML 2021] + GPT-2 [Radford+, 2021]

Does our model outperform the baseline?

Setup:

- Report 5-trial results
- |θ| : Trainable parameters [M]
- Baseline : ClipCap
- Our model : ClipCap + Adapter

Models		|θ|	Acc	Bs
Fine-tuned Decoder & Trans. Blocks	ClipCap	410	**36.80**	**81.75**
	ClipCap + Adapter	426	34.15	81.28
Fine-tuned Trans. Blocks	ClipCap	74	35.83	81.21
	ClipCap + Adapter	90	**38.11**	**81.71**
FT. Adapter Only	**ClipCap + Adapter**	**16**	**39.03**	**81.92**

- Our model, **when the adapter are well-trained**, outperforms a baseline

What factors contribute to improve GazeVQA task?

Qualitative Results :

User: **green circle**, Pred. RoI: **orange box**, Ground truth : **blue box**

AQ : What is this man wearing?
A : A (Gray) Suit
Pred. : A Coat; A Suit; A Shirt
✔ GT : **A Suit**

- Our model, which inputs **GT to adapter provides consistent** answer to AQs about **attributes of gaze targets**.

Evaluation with Clarified Questions (CQs):

- Discuss on modern V&L models [Cho+, ICML2021; OpenAI, 2023]
- Without fine-tuned GazeVQA

Models	Q	Acc
ClipCap [Mokady+, 2021]	AQ	21.55
	CQ	**25.11**
VL-T5 [Cho+, ICML2021]	AQ	32.33
	CQ	**34.11**
GPT-4V [OpenAI, 2023]	AQ	34.11
	CQ	**39.33**

- Acc improved 1-5 points
- **An approach to rewrite AQ to CQ** is also effective for GazeVQA task.

Future Directions

- Apply our model to an actual system, such as a robot
- Develop a method for clarifying questions (RQ 2)