Visual Question Generation for Class Acquisition of Unknown Objects

ECCV2018

2019.05.09

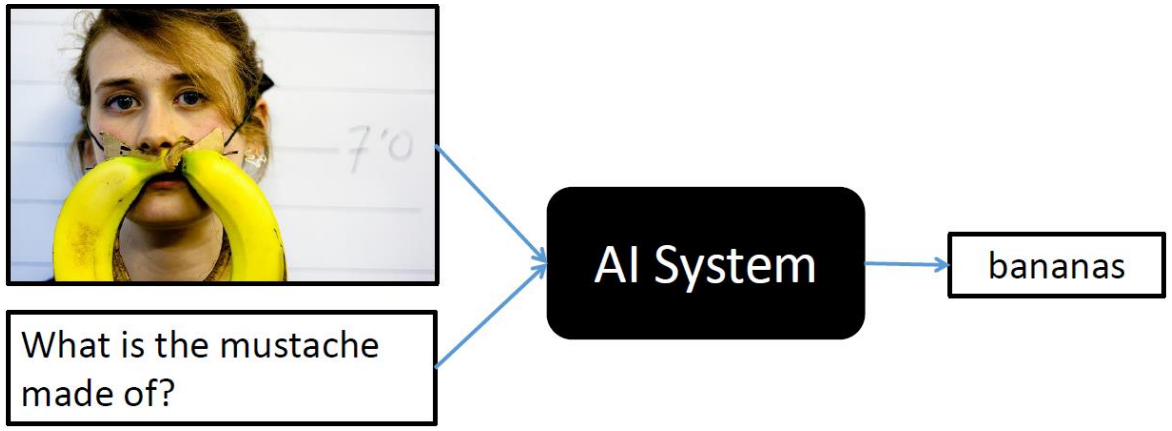
발표자 박성현

DAVIAN
Data and Visual Analytics Lab

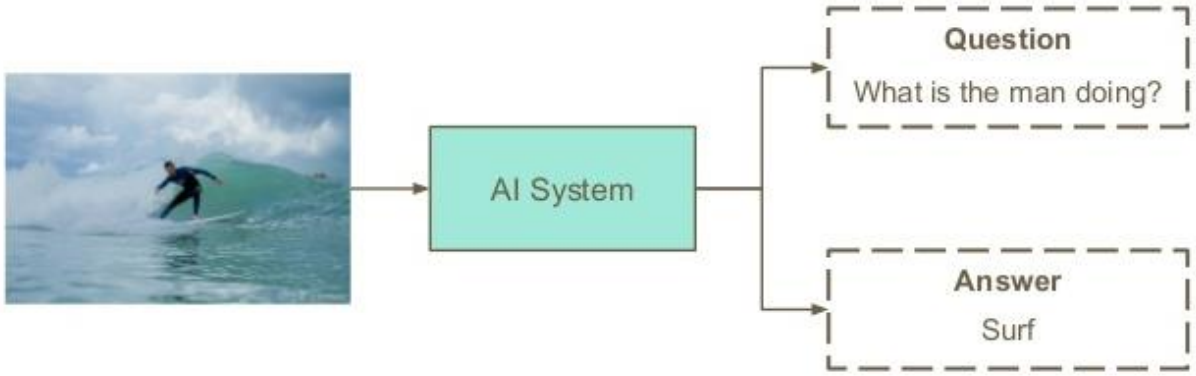
KOREA UNIVERSITY

1 Introduction

Visual Question Answering & Generation



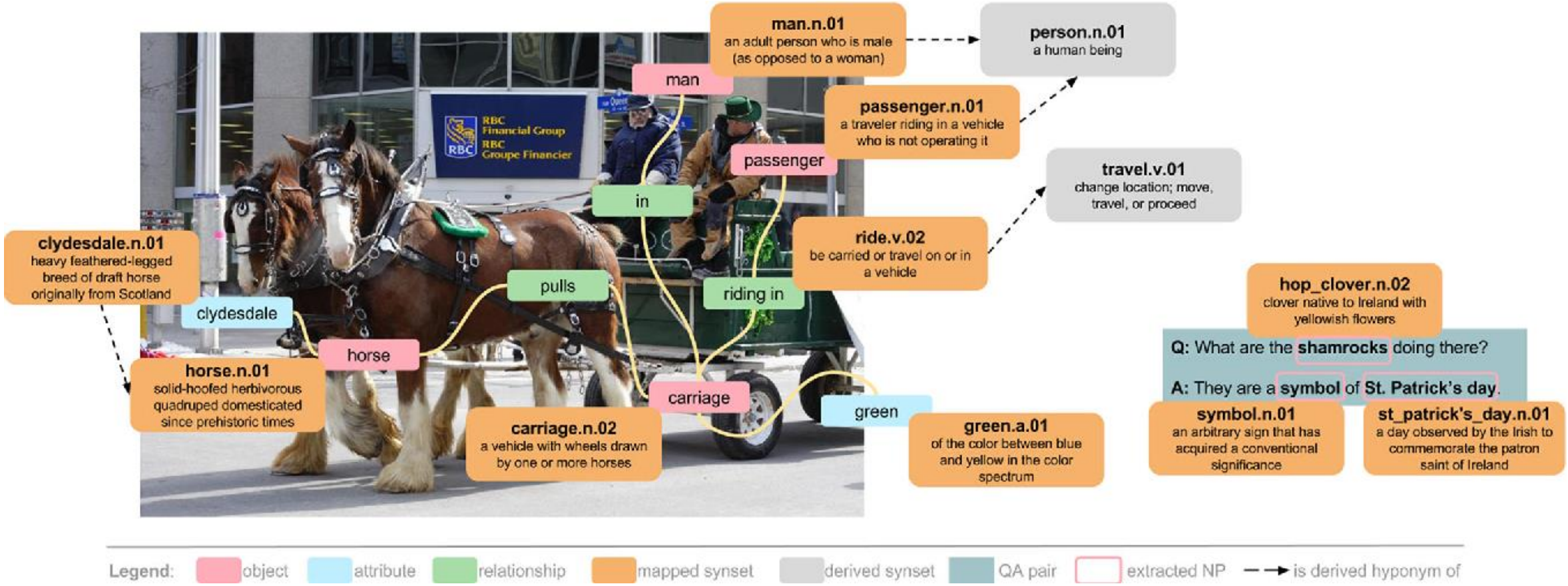
[Visual Question Answering]



[Visual Question Generation]

1 Introduction

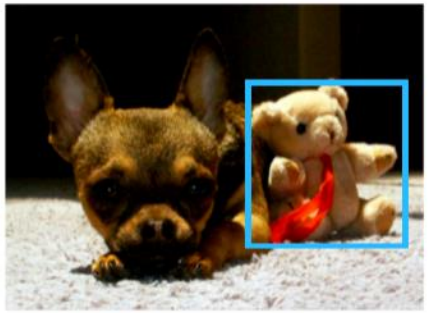
Visual Genome Dataset



Visual Genome : https://visualgenome.org/

1 Introduction

Motivation



✓ (a) What is the stuffed toy sitting next to the dog?

✗ (b) What is this?

✗ (c) Where is this picture taken?

Fig. 1. Examples of suitable/unsuitable questions for unknown objects. A suitable question should specify the target object (*stuffed toy*), so the answer is the class of the unknown object (*teddy bear*). Therefore, questions such as (a) are suitable. On the other hand, simple questions such as (b) and questions about location such as (c) are unsuitable.

→ **Unknown Object로 Question Generation하는 게 목표**

2 Model

Overview of the proposed model

1. **Object Region Proposal (Selective Search)**

2. **Unknown Object Classification and Target Selection**
   - Unknown object classification with CNN + Uncertainty Sampling
   - Select the most salient unknown object (Target selection)

3. **Visual Question Generation**
   - Encoding of Image Features
   - Question Target

2 Model

Overview of the proposed model

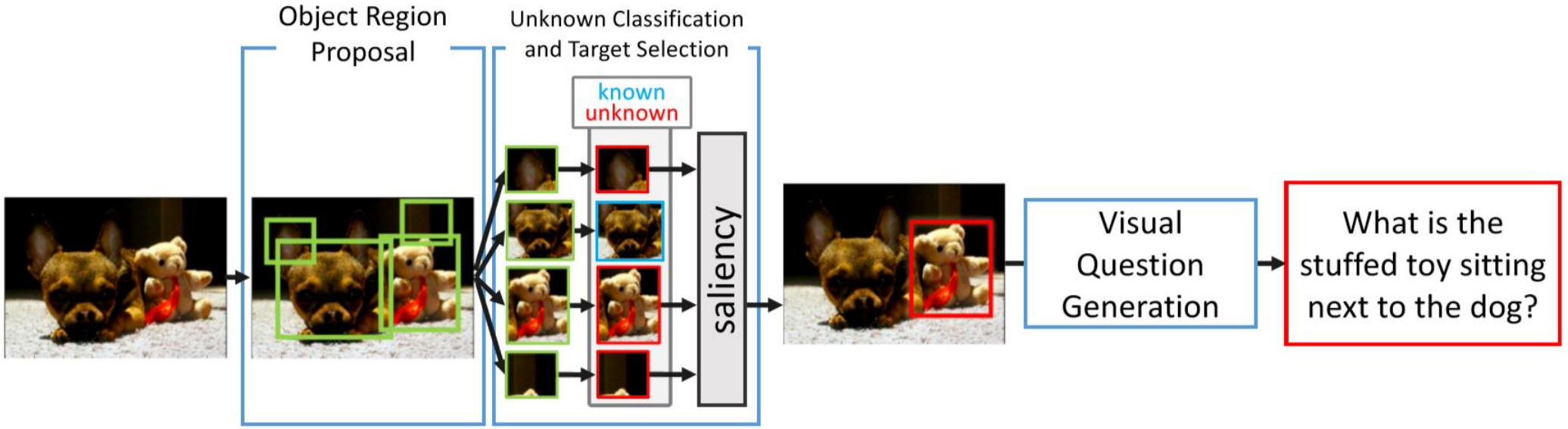
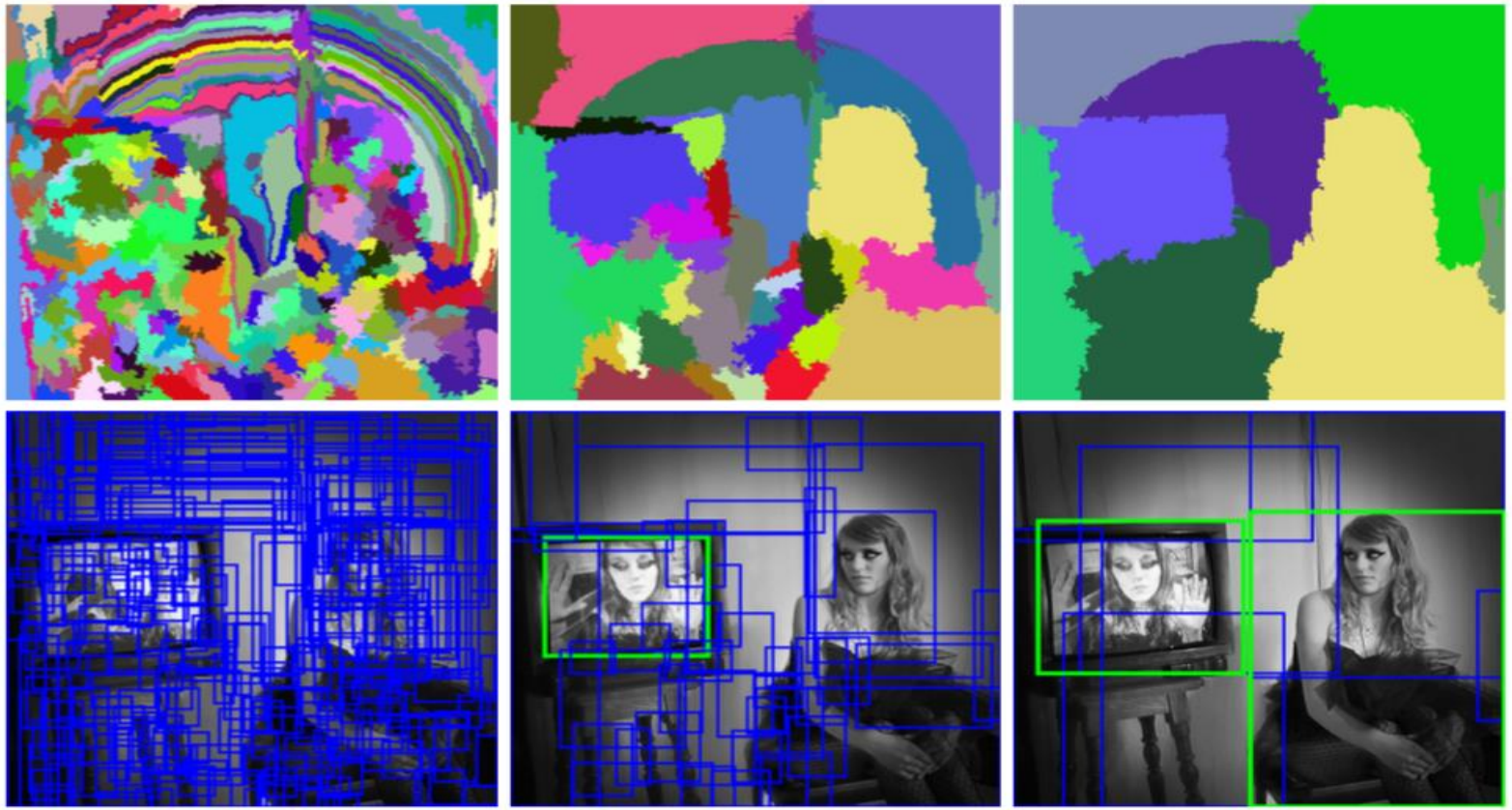


Fig. 2. Overview of the proposed method. First, regions from objects in the image (including unknown objects) are detected. Then, unknown objects are classified and the target region is selected. Finally, the target region along with the whole image is coded into a feature vector, and a question for the unknown object is generated

2 Model

Object Region Proposal

Unknown Object Detection Task는 Unsupervised → Selective Search를 사용



[Selective Search]

2 Model

Unknown Object Classification and Target Selection

- **Unknown object classification & Uncertainty Sampling**
  Perform unknown object classification by estimating the dispersion of the probability distribution using an entropy measure.

$$E = -\sum_{j=1}^{K} p_j \log_2 p_j$$

- **Select the most salient unknown object (Target selection)**
  Background regions are likely to be classified as unknown, but they don't contain an object to ask about.
  Calculate the saliency map for selecting the target region.

$$I_{region} = \sum_{I(p) \geq \theta} I(p) \times \frac{S_{salient}}{S_{region}}$$

2 Model

Visual Question Generation

- **Encoding of Image Features**
  Pretrained CNN model to extract the features f_I of the entire image and the features f_R of the target region.

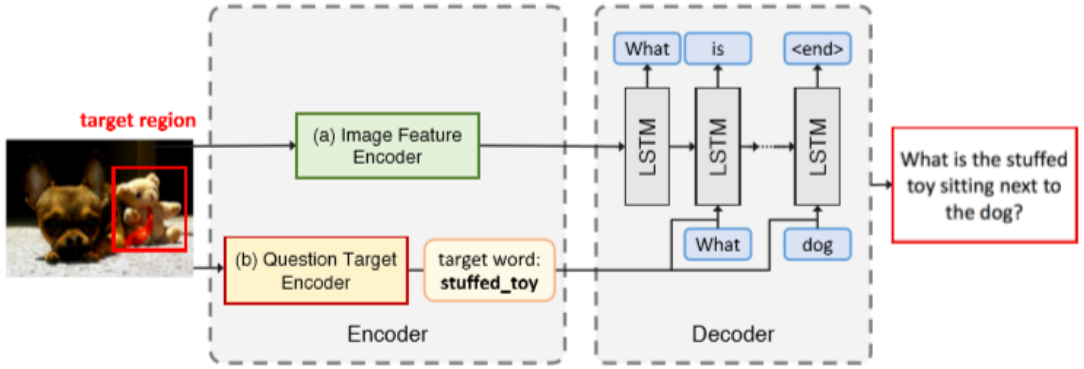
$$l_R = \left[\frac{x_{tl}}{W}, \frac{y_{tl}}{H}, \frac{x_{br}}{W}, \frac{y_{br}}{H}, \frac{S_R}{S_I}\right] \qquad f = [f_R, f_I, l_R]$$

  Concatenate f_R, f_I, I_R and let the 2005 dimensional vector f be the image feature encoding

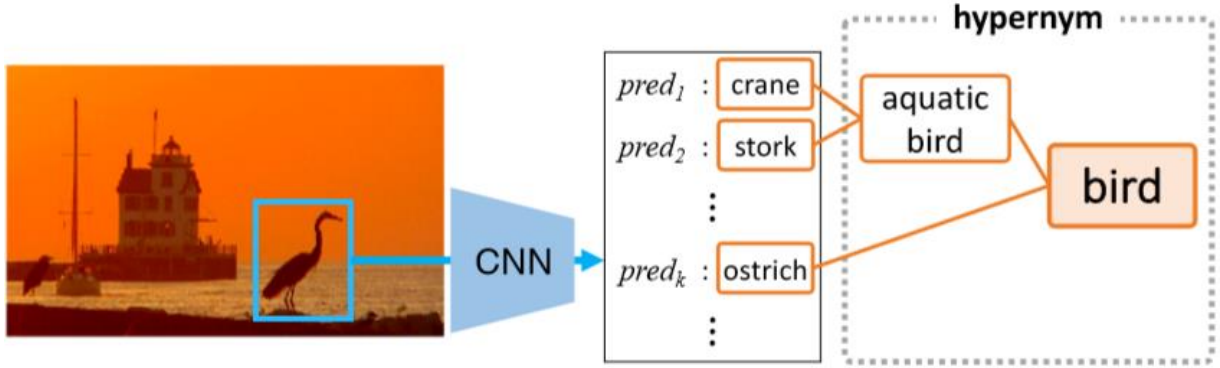
- **Question Target**
  Use WordNet to get the hierarchical relationship of words.
  Use the k predicted classes with the highest confidence of the classification result and select the word with the lowest level among the common hypernyms of the class labels.

2 Model

Visual Question Generation



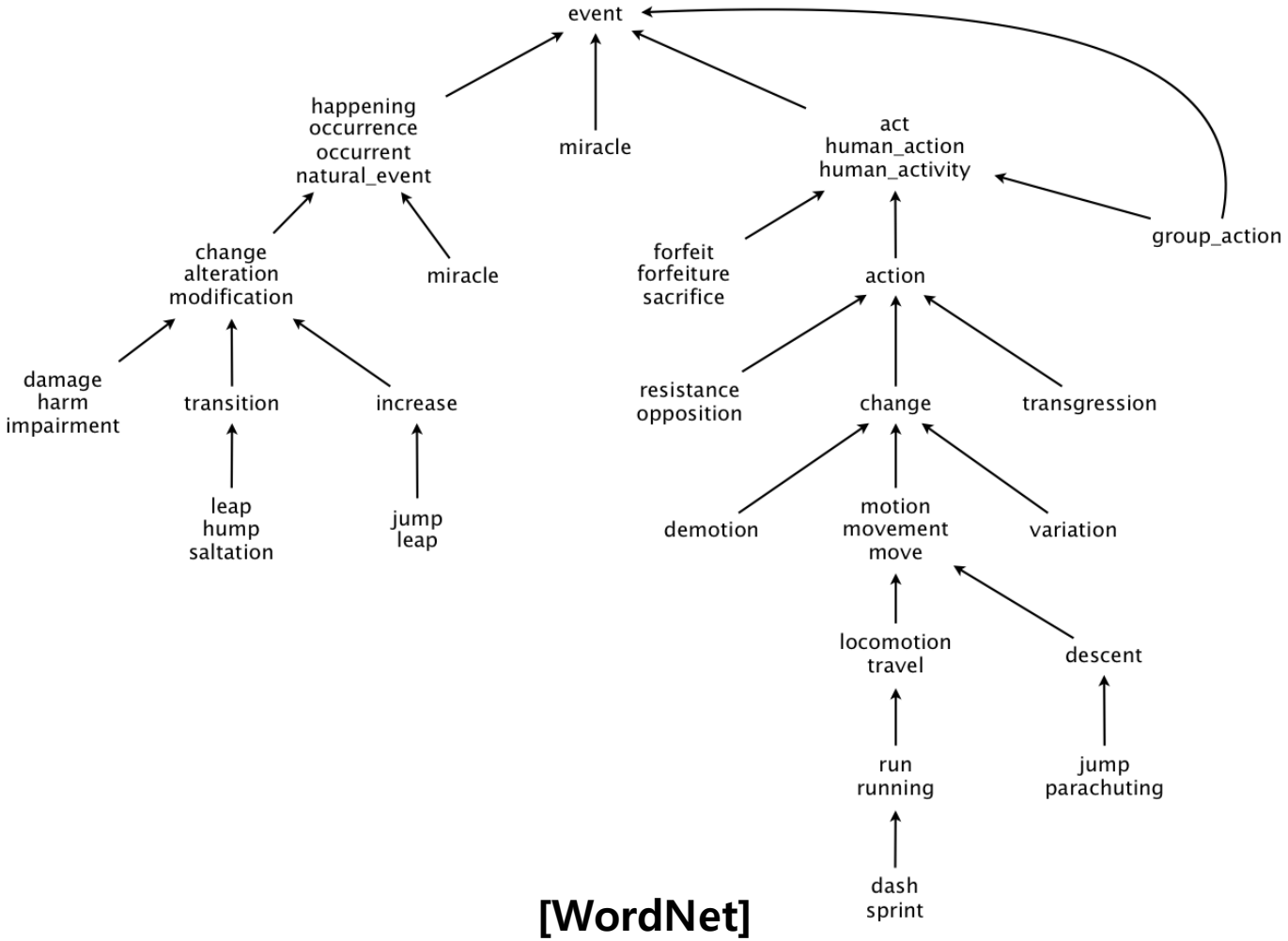
[Visual Question Generation]



[Question Target Module]

2 Model

WordNet



[WordNet]

3 Experiments

Evaluation of the Unknown Object Classification

Table 1. Comparison of the proposed unknown object classification method in terms of F measure results ± standard error. We performed experiments on CaffeNet, VGGNet, and ResNet. In all three cases, the proposed method outperformed the other methods

	F measure		
	CaffeNet	VGGNet	ResNet
Ours	$\mathbf{0.526 \pm 1.1 \cdot 10^{-3}}$	$\mathbf{0.602 \pm 0.2 \cdot 10^{-3}}$	$\mathbf{0.654 \pm 0.9 \cdot 10^{-3}}$
Least Confident	$0.522 \pm 1.1 \cdot 10^{-3}$	$0.590 \pm 1.5 \cdot 10^{-3}$	$0.635 \pm 1.2 \cdot 10^{-3}$
Bendale et al. [19]	$0.524 \pm 0.9 \cdot 10^{-3}$	$0.553 \pm 0.6 \cdot 10^{-3}$	$0.624 \pm 1.7 \cdot 10^{-3}$

Table 2. Comparison of the proposed unknown object classification method in terms of execution time, with CaffeNet as a classifier. We performed classification for 100 images and showed the average time per image ± standard error

	time (sec/image)
Ours	0.0400 ± 0.0017
Least Confident	$\mathbf{0.0365 \pm 0.0019}$
Bendale et al. [19]	15.6 ± 0.7

3 Experiments

Evaluation of the Visual Question Generation

Table 3. Comparison between our method and the baseline in terms of automatic evaluation metrics. The proposed method outperformed baseline methods

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Ours	**0.518**	**0.359**	**0.244**	**0.175**	**0.197**
CNN + LSTM	0.456	0.296	0.175	0.110	0.163
Retrieval	0.438	0.275	0.157	0.094	0.151

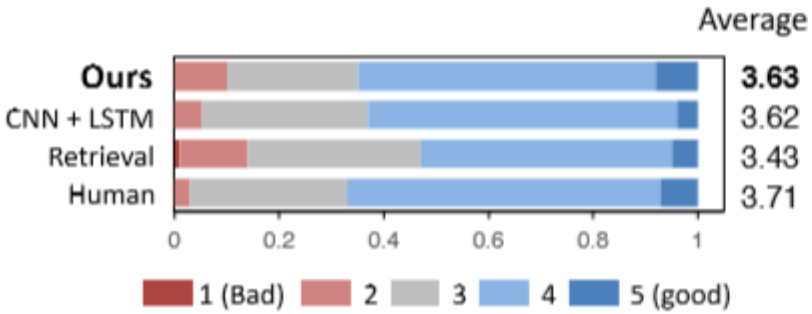


Fig. 5. (1) Human evaluation results on the naturalness of questions

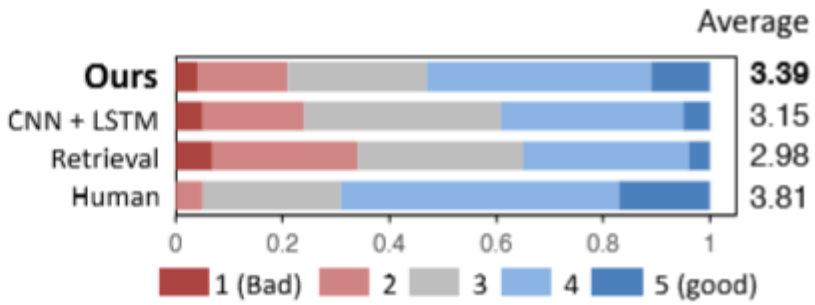
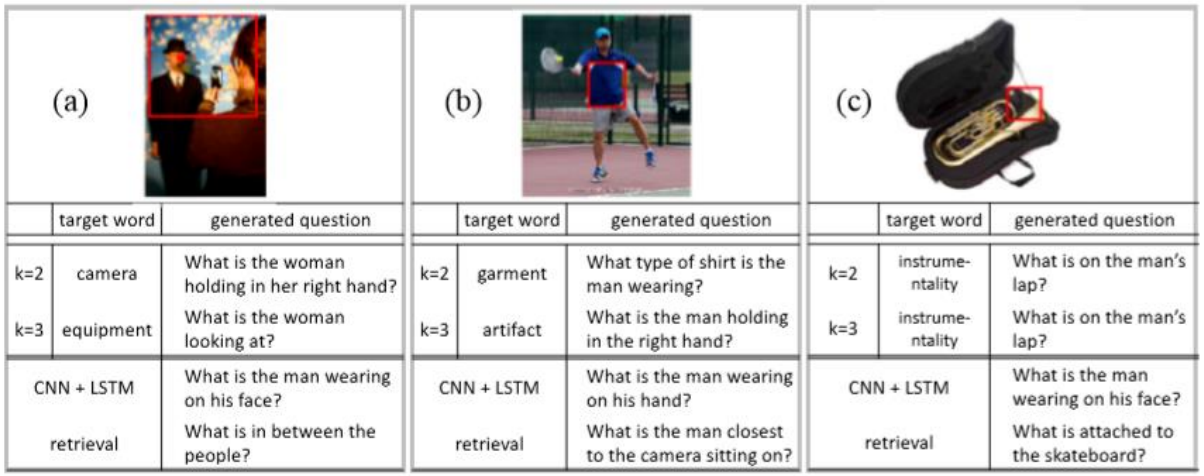


Fig. 6. (2) Human evaluation results on the relevance of questions to their region

3 Experiments

Evaluation of VQG for Unknown Objects



	target word	generated question
k=2	camera	What is the woman holding in her right hand?
k=3	equipment	What is the woman looking at?
CNN + LSTM		What is the man wearing on his face?
retrieval		What is in between the people?

(b)

	target word	generated question
k=2	garment	What type of shirt is the man wearing?
k=3	artifact	What is the man holding in the right hand?
CNN + LSTM		What is the man wearing on his hand?
retrieval		What is the man closest to the camera sitting on?

(c)

	target word	generated question
k=2	instrume-ntality	What is on the man's lap?
k=3	instrume-ntality	What is on the man's lap?
CNN + LSTM		What is the man wearing on his face?
retrieval		What is attached to the skateboard?

Fig. 7. Examples of input images (upper), the target words and generated questions by our proposed VQG method for unknown objects (middle), and the generated questions by the *CNN + LSTM* and *retrieval* baselines (lower).

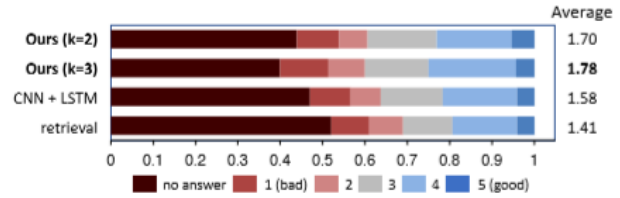


Fig. 8. Comparison of our method with the baseline in terms of the human evaluation in task (2). Task (2) evaluates whether or not the generated question, the image region, and the obtained answer are related. The greater the score, the higher the relevance.

Table 4. The number of generated questions that successfully allowed acquiring information on unknown objects (out of 300). We counted only the questions whose answers (task (1)) are not included in the known classes of the classifier, and the relevance of the question and target region in the image (task (2)) is four or more.

Ours($k = 2$)	**61**
Ours($k = 3$)	49
CNN + LSTM	46
Retrieval	45