Visual Question Answering Dataset for Bilingual Image Understanding: A Study of Cross-Lingual Transfer Using Attention Maps

Nobuyuki Shimizu[1], **Na Rong**[*2] , **Takashi Miyazaki**[1]
[1] Yahoo Japan Corporation, Tokyo, Japan
[2] Tokyo Institute of Technology, Tokyo, Japan
nobushim@yahoo-corp.jp, na@ks.cs.titech.ac.jp, takmiyaz@yahoo-corp.jp

Abstract

Visual question answering (VQA) is a challenging task that requires a computer system to understand both a question and an image. While there is much research on VQA in English, there is a lack of datasets for other languages, and English annotation is not directly applicable in those languages. To deal with this, we have created a Japanese VQA dataset by using crowdsourced annotation with images from the Visual Genome dataset. This is the first such dataset in Japanese. As another contribution, we propose a cross-lingual method for making use of English annotation to improve a Japanese VQA system. The proposed method is based on a popular VQA method that uses an attention mechanism. We use attention maps generated from English questions to help improve the Japanese VQA task. The proposed method experimentally performed better than simply using a monolingual corpus, which demonstrates the effectiveness of using attention maps to transfer cross-lingual information.

1 Introduction

Visual question answering (VQA) is the automated task of answering questions about a given image, which is a difficult task because the computer system must understand both the question (natural language processing, or NLP) and the image (computer vision). Recently there has been much research on VQA but the existing literature focuses only on English. Because each language differs in its grammar and resources, there is a pressing need to develop VQA systems for different languages. The effort to develop such systems is hindered by a lack of language resources. In this paper, we discuss two ways to address the lack of datasets for languages besides English. First, by using images from the Visual Genome dataset (Krishna et al., 2016) and annotation through crowdsourcing, we have created the first VQA dataset for Japanese. We call it the Japanese Visual Genome VQA dataset. It consists of 99,208 images with a total of 793,664 Japanese question answering (QA) pairs, for an average of eight QA pairs per image. The examples of obtained Japanese questions are shown along with the English version of Visual Genome questions in Figure 1. Second, we propose a cross-lingual method of exploiting the information in an existing English dataset to help improve the performance of our Japanese system in a VQA setting.

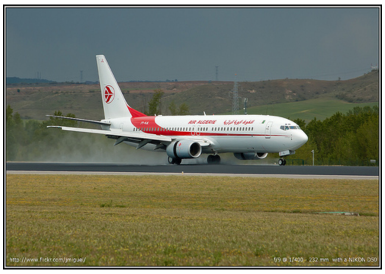
The difficulty of applying a cross-lingual method for VQA is the diversity of questions that one can ask about an image. For image captioning (Miyazaki and Shimizu, 2016), English and Japanese captions on the same image are generally considered as comparable corpora. Questions for VQA, however, can be quite diverse as compared to image captions, because they are not necessarily specific to the image. For example, questions about the time of day can be asked for almost any image available. As such questions might appear in one language but not in another, it is much tougher to exploit questions across languages.

On the other hand, striking features of an image are usually picked up by several questions. The motivation of our approach is based on experiences as a bilingual and a second language learner. Sometimes

[*] This study was conducted during an internship at Yahoo! JAPAN Research, Tokyo, Japan

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 27th International Conference on Computational Linguistics, pages 1918–1928
Santa Fe, New Mexico, USA, August 20-26, 2018.


http://www.flickr.com/jmiguel/
f/9 @ 1/400 - 232 mm with a NIKON D50

(a)
J1: kono norimono ha nandesuka
J2: hikouki no kitai ha naniiro desuka – E2
J3: kitai no moji ha naniiro desuka
J4: hikouki ha nani wo shiteimasuka
J5: hikouki ha doko ni imasuka – E14
J6: kokoha doko desuka – E1
J7: hikouki ha dochira ni hashitte imasuka
J8: enjin ha doko ni tsuite imasuka – E15

(b)
E1: Where was this picture taken?
E2: What color is the plane?
E3: When was this picture taken?
E4: How many planes are in the picture?
E5: What color is the sky?
E6: What color is the grass?
E7: Where is the body?
E8: What besides smoke could be seen behind plane?
E9: How could the sky be described?
E10: What is seen immediately behind runway?
E11: What is seen in the foreground of photo?
E12: How could the hills at skyline be described?
E13: What appears to be in the field on far right of photo?
E14: Where is the plane?
E15: Where is the turbine engine?
E16: What is red and white on the runway?
E17: What is in the distance?
E18: What website is on the bottom of the photo?
E19: What is under the plane?

Figure 1: Examples of (a) Japanese questions from our dataset and (b) English questions from the Visual Genome dataset. A Japanese question with an equivalent English question is matched, for example J2 – E2. The translations of J1, J3, J4, and J7 are as follows. J1: What form of transport is it? J3: What is the color of the letters on the body? J4: What is the airplane doing? J7: In which direction is the airplane going?

we can infer what a foreign language speaker is saying simply from the situation we are in. If we happen to be paying attention to the same object, we share the contexts and are likely to have similar perceptions.

We thus propose the first-ever method for cross-lingual VQA, enabling us to exploit questions in English to improve the performance of our VQA system in Japanese. In our method, we first generate attention maps for the English questions about an image. We then input these attention maps to the Japanese VQA system. We use the method given in Lu et al. (2016) both to generate the attention maps and to provide an experimental baseline. Our proposed cross-lingual method is simple yet effective, as we simply replace the input image feature from Lu et al. (2016) in our Japanese system with the attention maps generated from English questions. In our experiment, the proposed method using cross-lingual information produced better results than did the method using only a monolingual corpus. Because the attention maps generated from English questions tended to focus on salient objects in the images, our results also suggest the possibility of improving VQA through saliency modeling.

Our contribution is thus twofold. First, we have created a Japanese VQA dataset, which we believe is the first such dataset to be released[1]. Second, by using cross-lingual information to perform the VQA task, we obtain better results than by simply using a monolingual corpus, demonstrating the effectiveness of cross-lingual information.

The remainder of the paper is organized as follows. In section 2, we discuss English VQA datasets and related work using attention maps for the VQA task. In section 3 we introduce our dataset, including its construction and statistics. We then explain both the method used to generate attention maps from English questions and our proposed cross-lingual method in section 4. Experimental results are given in section 5 and we conclude the paper with a brief summary in section 6.

2 Related Work

In this section we discuss three related topics: first, four English datasets that have been generally used in VQA; second, some recent work on VQA with attention mechanisms, which have become a popular approach for VQA tasks; and last, cross-lingual models used in NLP. The attention-based approach involves generating image regions, called "attention maps," that are relevant to answering questions and then using the attention maps to generate answers.

[1] The dataset is to be released at `https://research-lab.yahoo.co.jp/en/software/`.

2.1 Datasets

A typical VQA dataset contains at least an image, a question for the image and the answer. Sometimes additional annotations, such as image regions relevant to the answers, or image captions, are provided as well. A number of datasets have been proposed for the VQA task in recent years; however, most of them are in English. In the following paragraphs we list the VQA datasets that have been generally used in this field and one multi-lingual VQA dataset with Chinese original VQA pairs and their English translations. The details are presented below.

The VQA dataset (Antol et al., 2015) contains 614,163 questions and 7,984,119 answers (including answers provided by humans either looking or not looking at the corresponding images) for 204,721 images from Microsofts COCO dataset (Lin et al., 2014), along with 150,000 questions and 1,950,000 answers for 50,000 abstract scenes. This dataset has two modalities for answering questions: open-ended and multiple-choice; in contrast, our dataset focuses on open-ended answers.

The Visual Genome (Krishna et al., 2016) dataset contains 1,773,258 QA pairs. On average, an image has 17 QA pairs. This dataset has two types of QA pairs (freeform QAs, which are based on the entire image, and region-based QAs, which are based on selected regions of the image) and six types of questions (what, where, when, who, why, and how). A subset of the freeform QA portion of Visual Genome is released as Visual 7W (Zhu et al., 2016). Because the release date of Visual 7W precedes that of Visual Genome, evaluations of VQA systems on freeform QA pairs are often conducted in Visual 7W instead of Visual Genome.

The COCO-QA (Ren et al., 2015) dataset was automatically generated from captions in the COCO dataset. It contains 78,736 training questions and 38,948 test questions, with the questions divided into four types: object questions, number questions, color questions, and location questions. Each answer consists of one word.

The DAQUAR (Malinowski and Fritz, 2014) dataset was built on top of the NYU-Depth V2 dataset (Silberman et al., 2012).The answers are mostly single-word answers. The complete dataset has 6,795 training QA pairs and 5,674 test pairs. On average, an image has nine QA pairs and the answers encompass 894 categories. The Reduced DAQUAR dataset is a subset of the complete DAQUAR dataset and contains 3,876 training samples, 297 test samples, and answers in 37 categories.

The Chinese Baidu dataset FM-IQA (Gao et al., 2015) uses 123,287 images sourced from the COCO dataset. The difference from COCO-QA is that the questions and answers were provided by human annotators through a crowdsourcing platform operated by Baidu. The annotators were free to add any type of question if it related to the content of the given image. This led to a much greater diversity of questions than in previously available datasets. Answering such questions typically requires both understanding the visual content of the image and incorporating prior "common sense" information. The dataset contains 120,360 images and 250,560 QA pairs, which were originally provided in Chinese and then converted to English by human translators.

Besides FM-IQA, the lack of datasets in languages other than English is striking. This situation hinders VQA research in other languages, such as Japanese. We intend our dataset to remedy this situation by providing resources in a morphologically rich language for the first time.

2.2 Attention-Based Methods for VQA

Previous studies have used information from whole images, but many questions and answers relate specifically to local regions in images. Many recent studies have thus focused on attention models, which select image regions relevant to answering questions, to deal with the VQA task (Xu and Saenko, 2016; Xiong et al., 2016; Shih et al., 2016; Yang et al., 2016; Lu et al., 2016). Shih et al. (2016) developed an approach for learning to answer visual questions by selecting image regions relevant to a text-based query. The approach maps textual queries and visual features from various regions into a shared space in which they are compared for relevance by applying an inner product. Yang et al. (2016) presented stacked attention networks (SANs), which account for the fact that VQA often requires multiple reasoning steps. The stacked attention model locates image regions relevant to the question for answer prediction via multi-step reasoning. While the above two approaches only use attention maps based on image regions, Lu et al. (2016) proposed a co-attention model for VQA, which jointly considers attention to both the image and the question. Attention to the question (which should be valued equally with attention to the image) represents the importance, in terms of probability, of each word in the question for answering the question. Because Lu et al. (2016) released work describing the implementation of this co-attention model and their results are easy to reproduce, we adopted their implementation as the foundation of our proposed method. We also used their work as a monolingual baseline.

2.3 Cross-Lingual Model

While resources in English have been quickly developed and are abundant by now, other languages have fallen behind in terms of the size and variety of their resources. To remedy this, some research leverages the existing knowledge in English to help process other languages. These multilingual resources capture valuable information that can be used in many fields and is especially useful for languages lacking resources. Many cross-lingual models have been proposed in NLP. These models are trained on a parallel corpus and find ways to connect between two languages, usually through supervised or semi-supervised learning. Cross-lingual information retrieval is another task that requires cross-lingual modeling (Jagarlamudi and Kumaran, 2007; Ballesteros and Croft, 1996), in which the query and results are written in different languages. In word embedding research, some studies have tried to transfer linguistic knowledge from one language to another, especially from English to low-resource languages, through distributed representations at the word level (Hermann and Blunsom, 2013; Haghighi et al., 2008; Klementiev et al., 2012).For image captions, Miyazaki and Shimizu (2016) proposed a caption generation model that transfers cross-lingual information from English to Japanese through pretraining of the model. Our proposed method shows that image attention maps can be used to transfer information cross-lingually. While the transfer in Miyazaki and Shimizu (2016) occurs at a fully connected layer after fc7 in the VGG16 model, our information transfer occurs at the attention-map level and is spatially interpretable.

Cross-lingual information is very useful, especially when dealing with sparse language resources. Although many cross-lingual learning models have been proposed, to our knowledge there has been no such prior research for the VQA task. This paper, we believe, reports the first work done on using cross-lingual information in the VQA field.

3 Statistics for Japanese Dataset

To create a Japanese version of a VQA corpus that would be comparable to the English version, we chose the Visual Genome dataset as a starting point. As noted previously, Visual Genome has two types of QA pairs: freeform and region-based. Our dataset is meant to be comparable to the freeform QA part of Visual Genome, which is similar to other VQA corpora, except for its focus on six types of questions (what, where, when, who, why, and how). Krishna et al. (2016) stated the benefits of focusing on those six question types as follows:

> First, they offer a considerable coverage of question types, ranging from basic perceptual tasks (e.g. recognizing objects and scenes) to complex common sense reasoning (e.g. inferring motivations of people and causality of events). Second, these categories present a natural and consistent stratification of task difficulty ... For instance, why questions that involve complex reasoning lead to the poorest performance ... of the six categories. This enables us to obtain a better understanding of the strengths and weaknesses of todays computer vision models, which sheds light on future directions in which to proceed.

3.1 Crowdsourcing Procedure

Visual Genome was created with Amazon Mechanical Turk (AMT), a well-known crowdsourcing platform for microtasks. To create a comparable Japanese dataset, we used a similar platform called Yahoo! Crowdsourcing, operated by Yahoo Japan Corporation. While AMT participants can be from anywhere in the world and have any mother language, participants in Yahoo! Crowdsourcing can be safely assumed to be proficient in Japanese, since such proficiency is required for signing up, navigating the user interface, and participating in the microtask market.

For collecting the freeform QA data in Visual Genome, consider a procedure with the following conditions: (1) Crowdsourcing participants are asked to view an image and write eight QA pairs related to it. (2) The participants are instructed that each question should start with one of the six question words - what, where, when, who, why, and how. (3) To encourage diversity, the participants are asked to write questions using at least three different question words out of the six.

While the above conditions are simple for English, we encountered several problems in creating a Japanese procedure with similar conditions. The second condition was especially troublesome. First, in the Japanese language, creating an interrogative sentence out of a declarative one does not require changing the word order. Since Japanese questions typically do not start with a question word, asking participants to write a question starting with such a word did not make sense. Second, just as English allows the phrase “what time” instead of “when,” Japanese allows “what place” for “where,” “what reason” for “why,” “what number” for “how many,” and so on. We thus attempted to list Japanese interrogative words similar to the six English question words used in Visual Genome: *nani* (what), *dare* (who), *doko* (where), *donna* (what kind), *dorekurai* (how much), *dou* (how), *itsu* (when), *ikutsu* (how many), and *naze* (why). We recognized, however, that the Japanese equivalent *nani* for “what” could become a catch-all category used for questions that would otherwise be asked with when/where/how/why and so on in English.

Once we came up with the Japanese equivalents of the six question words, we posted a pilot task that asked participants to view an image and write eight QA pairs. We found that some participants reused the same QA pairs over and over, on the basis of common knowledge. Consider, for example, the QA pair of “What color is the sky?” and “Blue.” Such a pair could be applicable to any outdoor image in the daytime. Given this experience with the pilot task, we modified the instructions to disallow such repeated QA pairs. All instructions and examples are provided in Japanese. The following instructions (translated into English) are the modified version:

> Please enter eight pairs consisting of a question and its answer in Japanese about the image linked by the URL.
>
> Please follow the following four rules in writing the question/answer pairs.
>
> Rule 1: Every question must use one of the following question words: *nani* (what), *dare* (who), *doko* (where), *donna* (what kind), *dorekurai* (how much), *dou* (how), *itsu* (when), *ikutsu* (how many), or *naze* (why). Example of a rule violation: “Question: Which is it, raining or sunny?” Reason for the violation: The word “which” is not in the list of question words.
>
> Rule 2: Every question must be distinct and ask something different. Furthermore, the eight questions must include at least three different question words from the list. Example of a rule violation: The eight questions use only the question words *nani* (what) and *dare* (who). Reason for the violation: The eight questions do not use at least three of the question words from the list.
>
> Rule 3: All eight questions must have a commonly agreed answer. Example of a rule violation: “Question: What is the man in the image thinking?” Reason for the violation: The answer is not commonly agreed.
>
> Rule 4: All eight questions must require the image content for the correct answer. Example of a rule violation: “Question: What is the color of the sunset? Answer: Red.” Reason for the violation: This question is answerable with common knowledge, so it does not require the image content.

After conducting the pilot task, we examined the results and selected promising participants (comprising a whitelist) for future task requests, so that only participants on the whitelist could perform the next task. We repeated this selection process until the final whitelist included about 1,500 participants. About 200-250 of them regularly participated in the actual VQA collection task. We posted tasks in small batches over the course of six months to prevent participants from working long hours. Despite

	Nine question types									other
	nani	*dare*	*doko*	*donna*	*dorekurai*	*dou*	*itsu*	*ikutsu*	*naze*	
	what	who	where	what kind	how much	how	when	how many	why	
Whole set	386,113	21,294	78,568	146,493	21,195	24,826	85,893	15,748	3,135	10,883
Test set	29,765	1,392	6,076	11,167	1,586	1,743	6,408	1,066	215	1,124

Table 1: Numbers of QA pairs in each category.

these measures, some participants eventually started to enter short, meaningless characters. To remove such noise, we removed sets of eight QA pairs whose question lengths averaged less than four characters.

3.2 Dataset Statistics and Analysis

The resulting Japanese VQA dataset consists of 99,208 images from Visual Genome, together with 793,664 QA pairs in Japanese, since every image has eight QA pairs.

Data quality. To ensure the consistency and integrity of the corpus, we randomly sampled and manually checked 800 QA pairs. Among those pairs, we found that six contained a minor typo, ten had a wrong answer, and two had an ambiguous question. In addition, four had questions answerable with common knowledge, while 29 had questions that were not based on one of the six question words and thus answerable with "yes" or "no." Thus, we found that 93.6% of the QA pairs correctly conformed to the specification. If we include the four pairs related to common knowledge, the figure increases to 94.1%, and if we also include the 29 pairs with a yes-no question and the six questions with a minor typo, it increases to 98.5%. Overall, we found that the quality of the QA pairs was very good.

Question type distribution. Unlike with English questions, which are easily classified according to the six question words (what, where, when, who, why, how), with Japanese questions grammar makes such classification more difficult. As mentioned in section 3.1, the questions in the Japanese dataset were classified into nine question types by nine words: *nani* (what), *dare* (who), *doko* (where), *donna* (what kind), *dorekurai* (how much), *dou* (how), *itsu* (when), *ikutsu* (how many), and *naze* (why). These nine words cover most Japanese questions. In addition to checking 800 QA pairs, we also examined 100 sets of eight QA pairs each and found twelve sets of questions that did not have at least three different question words. While the crowdsourcing participants sometimes did not follow the instructions exactly, we believe that this did not significantly decrease the diversity of question types. As shown by the statistics listed in Table 1, the question types in the corpus varied considerably.

Question and answer length distributions. We used MeCab to tokenize the Japanese questions and answers. Figure 2 shows the resulting distribution of average lengths for each category.

Comparison with English version. Manual examination of our corpus revealed that questions were very similar to the original English version of Visual Genome. An image with Japanese and English samples is shown in Figure 1. Generally, there are two to three times more QA pairs per image for the English version than for the Japanese version and questions often overlap. In Figure 1, J2 and E2, J5 and E14, J5 and E14, J6 and E1, and J8 and E15 are paired because they are essentially the same question.

4 Methodology

The method introduced in section 4.1 was used both to generate attention maps from English questions and to provide an experimental baseline. Section 4.2 describes how we used cross-lingual information (that is, the attention maps generated from the English data) to improve the performance of the Japanese VQA system.

4.1 Baseline

In this section, we briefly introduce material from Lu et al. (2016). That work proposed a VQA co-attention model, which jointly considers attention to both the image and the question. We used this approach for two purposes: first, to generate visual attention maps for English; and second, to provide a baseline for our experiment.

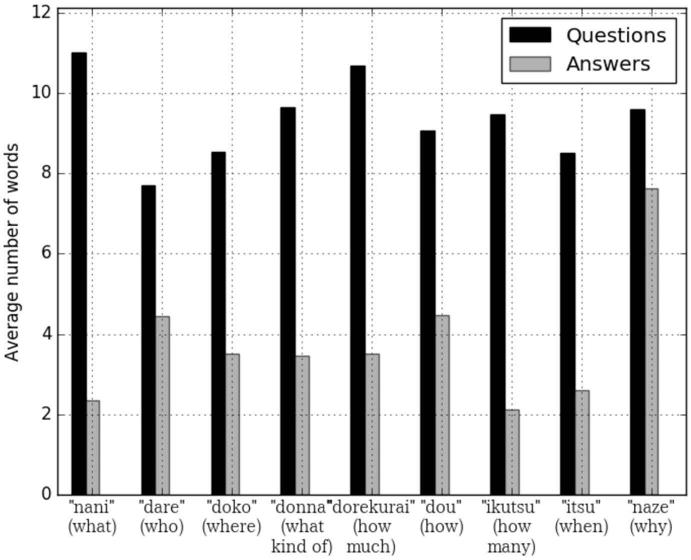


Figure 2: Average lengths of questions and answers for each question type.

A question consisting of T words is represented by $\mathbf{Q} = \{\mathbf{q}_1, ..., \mathbf{q}_T\}$, where $\mathbf{q}_t$ is the feature vector for the t-th word. Then, $\mathbf{q}_t^w$, $\mathbf{q}_t^p$, and $\mathbf{q}_t^s$ represent respectively the word embedding, phrase embedding, and question embedding at position t. The feature vectors are extracted the same way as in Lu et al. (2016) using LSTM (Hochreiter and Schmidhuber, 1997). The image feature is represented by

$$\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_i, ..., \mathbf{v}_N\}, \tag{1}$$

where $\mathbf{v}_i$ is the feature vector for the i-th spatial location. N is the number of grids in an image, which in our case is 96 (14 by 14). The co-attention features of the question and image at each level in the hierarchy are respectively denoted as $\hat{\mathbf{q}}^r$ and $\hat{\mathbf{v}}^r$, where $r \in \{w, p, s\}$ (i.e., the level of a word, phrase, or question).

As noted in section 2, there are two attention maps in this model, for attention to the image and to the question. Depending on the order in which the image and question attention maps are generated, there are two co-attention mechanisms in Lu et al. (2016): parallel co-attention, and alternating co-attention. Because Lu et al. showed that parallel co-attention outperforms alternating co-attention, we chose the former as a baseline model and building block of our proposed model. We forgo explaining the mechanism of the latter here.

Parallel Co-Attention. The parallel co-attention mechanism generates the image and question attention maps simultaneously. Given an image feature map $\mathbf{V} \in R^{d \times N}$ and a question representation $\mathbf{Q} \in R^{d \times T}$, the mechanism calculates the similarity between the image and question features for all pairs of image locations and question positions:

$$\mathbf{C} = \tanh(\mathbf{Q}^\top \mathbf{W}_b \mathbf{V}), \tag{2}$$

where $\mathbf{C} \in R^{T \times N}$ is the resulting affinity matrix, and $\mathbf{W}_b \in R^{d \times d}$ contains weights. Then, the mechanism uses the affinity matrix as a feature for learning to predict the image and question attention maps:

$$\begin{aligned}
\mathbf{H}^v &= \tanh(\mathbf{W}_v \mathbf{V} + (\mathbf{W}_q \mathbf{Q})\mathbf{C}) \\
\mathbf{H}^q &= \tanh(\mathbf{W}_q \mathbf{Q} + (\mathbf{W}_v \mathbf{V})\mathbf{C}^\top) \\
\mathbf{a}^v &= softmax(\mathbf{w}_{hv}^\top \mathbf{H}^v) \\
\mathbf{a}^q &= softmax(\mathbf{w}_{hq}^\top \mathbf{H}^q),
\end{aligned} \tag{3}$$

where $\mathbf{W}_v, \mathbf{W}_q \in R^{k \times d}$, $\mathbf{w}_{hv}$, and $\mathbf{w}_{hq} \in R^k$ are weight parameters. Then, the image attention map is $\mathbf{a}^v \in R^N$ and the question attention map is $\mathbf{a}^q \in R^T$. These represent respectively the attention probabilities for each image region $\mathbf{v_n}$ and word $\mathbf{q_t}$. Finally, the mechanism calculates the image and question attention vectors:

$$\hat{\mathbf{v}} = \sum_{n=1}^{N} a_n^v \mathbf{v}_n, \; \hat{\mathbf{q}} = \sum_{t=1}^{T} a_t^q \mathbf{q}_t. \tag{4}$$

The parallel co-attention mechanism is applied at each level in the hierarchy, leading to $\hat{\mathbf{v}}^r$ and $\hat{\mathbf{q}}^r$, where $r \in \{w, p, s\}$.

4.2 Proposed Method

The goal of our method is to use the information learned from the English dataset to help improve the performance of the Japanese dataset. First, we use the method introduced in the previous section to generate attention maps for each image by using English questions. Then, for the Japanese Visual Genome QA dataset, we define an image feature $\mathbf{V}_{new}$ created from the attention maps and the image feature defined by eq. (1). Finally, we replace the image feature $\mathbf{V}$ in eqs. (2) and (3) with $\mathbf{V}_{new}$, and the method then proceeds as in Lu et al. (2016). We discuss the details below.

First, our method learns information from the English dataset. For each English question and image pair in Visual Genome, we generate one visual attention map $\mathbf{a}^v = \{\mathbf{a}_1^v, ..., \mathbf{a}_i^v, ..., \mathbf{a}_N^v\}$, where $\mathbf{a}_i^v$ contains the attention probabilities for image region $\mathbf{v}_n$. An image with $\mathbf{M}$ English questions thus has $\mathbf{M}$ attention maps $\mathbf{a}^v$, which we average to obtain a final attention map:

$$\mathbf{a} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{a}_m^v, \; where\, \mathbf{a} \in R^N. \tag{5}$$

Then, the information learned from the English dataset (i.e., the image attention maps) is used for Japanese VQA prediction. The motivation for using these attention maps is that although the English and Japanese questions for the same image are usually different, the foci of attention most likely overlap. Letting $\mathbf{a} = (a_1, ..., a_N)$, we represent the image feature as $\mathbf{V}_{new} = \mathbf{Va} = (a_1\mathbf{v}_1, .., a_n\mathbf{v}_n)$. Finally, after replacing $\mathbf{V}$ with $\mathbf{V}_{new}$, our proposed method proceeds as in section 4.1 to learn the Japanese VQA model and predict Japanese answers. When the method is not considering information learned from the English dataset, it sets $\mathbf{a} = I$ (a vector with all elements set to 1).

5 Experiment

5.1 Setup

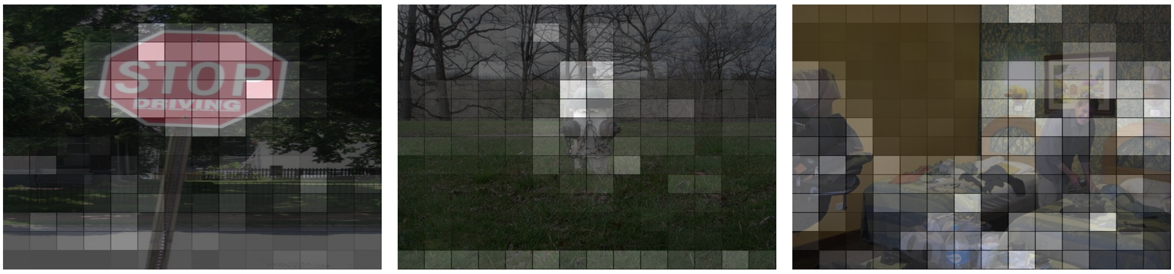
Our experimental baseline was the parallel co-attention model trained using the Japanese corpus. The proposed method uses the same model but with the attention maps initialized with the English corpus and then trained with the Japanese corpus. We divided our dataset into two parts for training and testing. The numbers of images in the training and test sets are respectively 91,609 and 7,599. The test set has 60,525 QA pairs. There are 135,740 unique answers in our dataset. We used the top 1,000 most frequent answers as the possible outputs, which covers 66.7% of the answers found in our dataset.

We used the RMSProp optimizer with a base learning rate of 4e-4, momentum 0.99, and weight decay 1e-8 and set the batch size to 20. We used VGG-19 (Simonyan and Zisserman, 2014), which is based on CNN (Fukushima, 1980; LeCun et al., 1989), to extract image features and MeCab (Kudo, 2005) to tokenize Japanese sentences. We performed 250,000 iterations.

5.2 Results and Analysis

For evaluation, we used the following definition of ***accuracy***:

$$\text{accuracy} = \frac{\text{No. of correctly classified questions}}{\text{No. of questions}} \tag{6}$$



(a) "mannaka ni utsutteiru hyoushiki ha naniiro desuka?" (b) "mannaka ni utsutteiru mono ha nan desuka?" (c) "ningen no mae ni donna kagu ga arimasu ka?"

Figure 3: Examples of attention maps for which the proposed method predicted the correct answer but the baseline method without cross-lingual attention maps did not. Questions: (a) What color is the road sign? (b) What is the object in the middle of the picture? (c) What is the furniture in front of the man?

	# images in training set		
Method	30,536	61,072	91,609
baseline	17.1%	17.7%	18.3%
proposed	**18.0%**	**18.8%**	**19.2%**

Table 2: Average accuracy with varying training set sizes. Each cell contains the average accuracy over four runs. All differences are statistically significant according to McNemar's test.

Figure 3 shows examples of attention maps that were generated from English questions and correctly predicted the answers for the Japanese dataset. For the same images, the baseline system without cross-lingual attention maps predicted wrong answers. We found that the attention maps usually focused on foreground objects and that accuracy tended to improve for images with clear foreground objects. Table 3 lists the accuracy by question type. For all question types except "why" questions, the cross-lingual attention maps improved the performance.

Because we trained and evaluated both our model and the baseline three times, Table 2 lists the average accuracies for each case. Our proposed model achieves 19.2% accuracy with 91,609 images in the training set. Note that our task is much tougher than in prior work (Zhu et al., 2016; Lu et al., 2016). In Zhu et al. (2016), the outputs are chosen from four multiple choices. While in Lu et al. (2016) the outputs are chosen from the same top 1,000 frequently occurring answers, the coverage of this set for their VQA corpus is 86.54%, unlike our 66.7%. The table also illustrates how the accuracy increased with the number of training images. Our method consistently performed around 1% better than the baseline method in all cases. As the performance increase was consistent across systems, we believe that using cross-lingual information should also improve performance in other situations. We can also see from Table 2 that the performance difference between the proposed method and the baseline did not decrease as the number of training images increased, which shows the value of our method for both larger and smaller datasets.

6 Conclusion

We have created a Japanese visual question answering (VQA) dataset comparable to the freeform question answering portion of the Visual Genome dataset (Krishna et al., 2016). This dataset is the first such dataset in Japanese. To show the utility of our corpus, we proposed a cross-lingual method for making use of English annotation to improve the Japanese VQA system. The proposed method experimentally

	Accuracy (%)								
Method	*nani* what	*dare* who	*doko* where	*donna* what kind	*dorekurai* how much	*dou* how	*itsu* when	*ikutsu* how many	*naze* why
baseline	19.9	26.1	14.4	20.3	15.5	24.4	53.1	18.9	**5.1**
proposed	**21.1**	**27.5**	**15.2**	**21.9**	**17.6**	**25.8**	**55.2**	**19.1**	**5.1**

Table 3: Accuracy for each of the nine Japanese question types.

performed better than simply using a monolingual corpus, which demonstrates the effectiveness of using attention maps to transfer cross-lingual information.

While VQA is mainly a testbed for monolingual image understanding, our data together with the original English Visual Genome allows modeling how a bilingual person understands images and two languages, which we call bilingual image understanding. We believe the release of our dataset will add significant resources to the research in this direction.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433.

Lisa Ballesteros and Bruce Croft. 1996. Dictionary methods for cross-lingual information retrieval. In *International Conference on Database and Expert Systems Applications*, pages 791–801. Springer.

Kunihiko Fukushima. 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, Apr.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*, pages 2296–2304.

Aria Haghighi, Percy Liang, Taylor Berg-Kirkpatrick, and Dan Klein. 2008. Learning bilingual lexicons from monolingual corpora. In *ACL*, volume 2008, pages 771–779.

Karl Moritz Hermann and Phil Blunsom. 2013. Multilingual distributed representations without word alignment. *arXiv preprint arXiv:1312.6173*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Jagadeesh Jagarlamudi and A Kumaran. 2007. Cross-lingual information retrieval system for indian languages. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 80–87. Springer.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*.

Taku Kudo. 2005. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. net/*.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural Comput.*, 1(4):541–551, December.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems*, pages 1682–1690.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1780–1790.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961.

Kevin J Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4613–4621.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2397–2406, New York, New York, USA, 20–22 Jun. PMLR.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 451–466, Cham. Springer International Publishing.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.