

Customized Image Narrative Generation via Interactive Visual Question Generation and Answering

Andrew Shin¹ Yoshitaka Ushiku¹ Tatsuya Harada^{1,2}

¹The University of Tokyo, ²RIKEN

{andrew, ushiku, harada}@mi.t.u-tokyo.ac.jp

Abstract

Image description task has been invariably examined in a static manner with qualitative presumptions held to be universally applicable, regardless of the scope or target of the description. In practice, however, different viewers may pay attention to different aspects of the image, and yield different descriptions or interpretations under various contexts. Such diversity in perspectives is difficult to derive with conventional image description techniques. In this paper, we propose a customized image narrative generation task, in which the users are interactively engaged in the generation process by providing answers to the questions. We further attempt to learn the user's interest via repeating such interactive stages, and to automatically reflect the interest in descriptions for new images. Experimental results demonstrate that our model can generate a variety of descriptions from single image that cover a wider range of topics than conventional models, while being customizable to the target user of interaction.

1. Introduction

Recent advances in visual language field enabled by deep learning techniques have succeeded in bridging the gap between vision and language in a variety of tasks, ranging from describing the image [14, 7, 26, 27] to answering questions about the image [2, 5]. Such achievements were possible under the premise that there exists a set of ground truth references that are universally applicable regardless of the target, scope, or context. In real-world setting, however, image descriptions are prone to an infinitely wide range of variabilities, as different viewers may pay attention to different aspects of the image in different contexts, resulting in a variety of descriptions or interpretations. Due to its subjective nature, such diversity is difficult to obtain with conventional image description techniques.

In this paper, we propose a customized image narrative generation task, in which we attempt to actively engage the

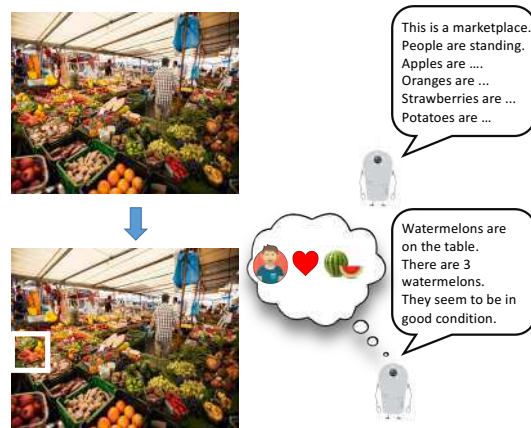


Figure 1: Example of conventional image description (top) and customized image narrative (bottom).

users in the description generation process by asking questions and directly obtaining their answers, thus learning and reflecting their interest in the description. We use the term *image narrative* to differentiate our image description from conventional one, in which the objective is fixed as depicting factual aspects of global elements. In contrast, *image narratives* in our model cover a much wider range of topics, including subjective, local, or inferential elements.

We first describe a model for automatic image narrative generation from single image without user interaction. We develop a self Q&A model to take advantage of wide array of contents available in visual question answering (VQA) task, and demonstrate that our model can generate image descriptions that are richer in contents than previous models. We then apply the model to interactive environment by directly obtaining the answers to the questions from the users. Through a wide range of experiments, we demonstrate that such interaction enables us not only to customize the image description by reflecting the user's choice in the current image of interest, but also to automatically apply the learned preference to new images (Figure 1).

2. Related Works

Visual Language: The workflow of extracting image features with convolutional neural network (CNN) and generating captions with long short-term memory (LSTM) [10] has been consolidated as a standard for image captioning task. [14] generated region-level descriptions by implementing alignment model of region-level CNN and bidirectional recurrent neural network (RNN). [12] proposed DenseCap that generates multiple captions from an image at region-level. [11] built SIND dataset whose image descriptions display a more casual and natural tone, involving aspects that are not factual and visually apparent. While this work resembles the motivation of our research, it requires a sequence of images to fully construct a narrative.

Visual question answering (VQA) has escalated the interaction of language and vision to a new stage, by enabling a machine to answer a variety of questions about the image, not just describe certain aspects of the image. A number of different approaches have been proposed to tackle VQA task, but classification approach has been shown to outperform generative approach [1, 13]. [8] proposed multimodal compact bilinear pooling to compactly combine the visual and textual features. [23] proposed an attention-based model to select a region from the image based on text query. [18] introduced co-attention model, which not only employs visual attention, but also question attention.

User Interaction: Incorporating interaction with users into the system has rapidly become a research interest. Visual Dialog [5] actively involves user interaction, which in turn affects the responses generated by the system. Its core mechanism, however, functions in an inverse direction from our model, as the users ask the questions about the image, and the system answers them. Thus, the focus is on extending the VQA system to a more context-dependent, and interactive direction. On the other hand, our model’s focus is on generating customized image descriptions, and user interaction is employed to learn the user’s interest, whereas Visual Dialog is not concerned about the users themselves.

[6] introduces an interactive game, in which the system attempts to localize the object that the user is paying attention to by asking relevant questions that narrow down the potential candidates, and obtaining answers from the users. This work is highly relevant to our work in that user’s answers directly influence the performance of the task, but our focus is on contents generation instead of object localization or gaming. Also, our model not only utilizes user’s answer for current image, but further attempts to apply it to new images. Recent works in reinforcement learning (RL) have also employed interactive environment by allowing the agents to be taught by non-expert humans [4]. However, its main purpose is to assist the training of RL agents, while our goal is to learn the user’s interest specifically.

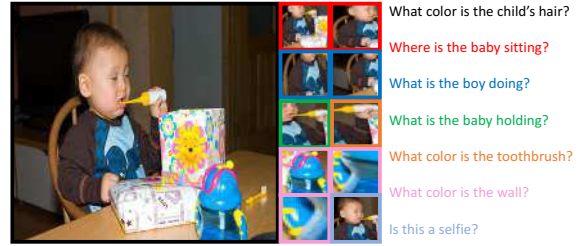


Figure 2: Example of regions extracted from the image, and the questions generated from each region.

3. Automatic Image Narrative Generation

We first describe a model to generate image narrative that covers a wide range of topics without user interaction. We propose a self Q&A model where questions are generated from multiple regions, and VQA is applied to answer the questions, thereby generating image-relevant contents.

Region Extraction: Following [9], we first extract region candidates from the feature map of an image, by applying linear SVM trained on annotated bounding boxes at multiple scales, and applying non-maximal suppression. The region candidates then go through inverse cascade from upper, fine layer to lower, coarser layers of CNN, in order to better-localize the detected objects. This results in region proposals that are more contents-oriented than selective search [25] or Edge Boxes [16]. We first extracted top 10 regions per image. Figure 2 shows an example of the regions extracted in this way. In the experiments to follow, we set the number of region proposals K as 5, since the region proposals beyond top 5 tended to be less congruent, thus generating less relevant questions.

Visual Question Generation: In image captioning task, it is conventional to train an LSTM with human-written captions as ground truth annotations. On the other hand, in VQA task, questions are frequently inserted to LSTM in series with fixed image features, and the answers to the questions become the ground truth labels to be classified. Instead, we replace the human-written captions with human-written questions, so that LSTM is trained to predict the question, rather than caption.

Given an image I and a question $Q = (q_0, \dots, q_N)$, the training proceeds as in [26]:

$$x_{-1} = CNN(I), x_t = W_e q_t, p_{t+1} = LSTM(x_t) \quad (1)$$



where W_e is a word embedding, x_t is the input features to LSTM at t , and p_{t+1} is the resulting probability distribution for the entire dictionary at t . In the actual generation of questions, it will be performed over all region proposals $r_0, \dots, r_N \in I$:

$$x_{-1} = CNN(r_i), x_t = W_e q_{t-1} \quad (2)$$

$$q_t = \max_{q \in p} p_{t+1} = \operatorname{argmax} LSTM(x_t)$$

for $q_0, \dots, q_N \in Q_{r_i}$. Figure 2 shows examples of questions generated from each region including the entire image. As

Table 1: Examples of questions generated using non-visual questions in VQG dataset.

Image	Generated Questions
	<ul style="list-style-type: none"> • What is the player’s name? • What is he speaking about? • What is the score? • Is this costume for a race? • Has he worked there?
	<ul style="list-style-type: none"> • Can the boy win the prize? • Was this a charity event? • What is she looking at? • What are they waiting for? • Who is that guy? • What is he looking at?

shown in the figure, by focusing on different regions and extracting different image features, we can generate multiple image-relevant questions from single image.

So far, we were concerned with generating “visual” questions. We also seek to generate “non-visual” questions. [20] generated questions that a human may naturally ask and require common-sense and inference. We examined whether we can train a network to ask multiple questions of such type by visual cues. We replicated the image captioning process described above, with 10,000 images of MS COCO and Flickr segments of VQG dataset, with 5 questions per image as the annotations. Examples of questions generated by training the network solely with non-visual questions are shown in Table 1.

Visual Question Answering: We now seek to answer the questions generated. We train the question answering system with VQA dataset [2]. Question words are sequentially encoded by LSTM as one-hot vector. Hyperbolic tangent non-linearity activation was employed, and element-wise multiplication was used to fuse the image and word features, from which softmax classifies the final label as the answer for visual question. We set the number of possible answers as 1,250.

As we augmented the training data with “non-visual” questions, we also need to train the network to “answer” those non-visual answers. Since [20] provides the questions only, we collected the answers to these questions on Amazon Mechanical Turk. Since many of these questions cannot be answered without specific knowledge beyond what is seen in the image (e.g. “*what is the name of the dog?*”), we encouraged the workers to use their imagination, but required them to come up with answers that an average person might also think of. For example, people frequently answered the question “*what is the name of the man?*” with “*John*” or “*Tom*.” Such non-visual elements add vividness and story-like characteristics to the narrative as long as they are compatible with the image, even if not entirely verifiable.

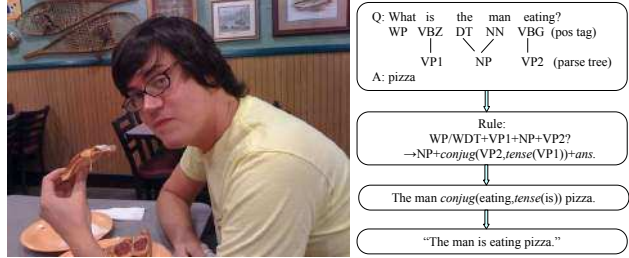


Figure 3: Example of question and answer converted to a declarative sentence by conversion rule.

Natural Language Processing: We are now given multiple pairs of questions and answers about the image. By design of the VQA dataset, which mostly comprises simple questions regarding only one aspect with the answers mostly being single words, the grammatical structure of most questions and answers can be reduced to a manageable pool of patterns. Exploiting these design characteristics, we combine the obtained pairs of questions and answers to a declarative sentence by application of rule-based transformations, as in [22, 24].

We first rephrase the question to a declarative sentence by switching word positions, and then insert the answers to its appropriate position, mostly replacing *wh*-words. For example, a question “*What is the man holding?*” is first converted to a declarative statement “*The man is holding what*” and the corresponding answer “*frisbee*” replaces “*what*” to make “*The man is holding frisbee.*” Part-of-speech tags with limited usage of parse tree were used to guide the process, particularly conjugation according to tense and plurality. Figure 3 illustrates the workflow of converting question and answer to a declarative sentence. See Supplemental Material for specific conversion rules. Part-of-speech tag notation is as used in PennTree I Tags [19].

4. Interactive Image Narrative Generation

We now extend the automatic image narrative generation model described in Section 3 to interactive environment, in which users participate in the process by answering questions about the image, so that generated narrative varies depending on the user input provided.

4.1. Applying Interaction within the Same Images

4.1.1 Question with Multiple Possible Answers

As discussed earlier, we attempt to reflect user’s interest by asking questions that provide visual context. The foremost prerequisite for the interactive questions to perform that function is the possibility of various answers or interpretations. In other words, a question whose answer is so obvious that it can be answered in an identical way would not be valid as an interactive question. In order to make sure that each generated question allows for multiple possible

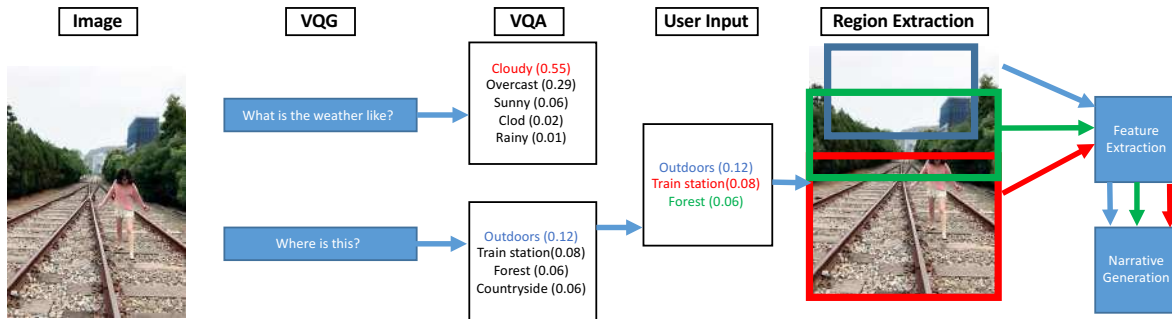


Figure 4: Questions that allow for multiple responses are generated to reflect user’s interest and corresponding regions proceed to image narrative generation process.

answers, we internally utilize the VQA module. The question generated by the VQG module is passed on to VQA module, where the probability distribution p_{ans} for all candidate answers C is determined. If the most likely candidate $c_i = \max p_{ans}$, where $c_i \in C$, has a probability of being answer over a certain threshold α , then the question is considered to have a single obvious answer, and is thus considered ineligible. The next question generated by VQG is passed on to VQA to repeat the same process until the the following requirement is met:

$$c_i < \alpha, c_i = \max p_{ans} \quad (3)$$

In our experiments, we set α as 0.33. We also excluded the yes/no type of questions. Figure 4 illustrates an example of a question where the most likely answer had a probability distribution over the threshold (and is thus ineligible), and another question whose probability distribution over the candidate answers was more evenly distributed (and thus proceeds to narrative generation stage).

4.1.2 Region Extraction

Once the visual question that allows for multiple responses is generated, a user inputs his answer to the question, which is assumed to reflect his interest. We then need to extract a region within the image that corresponds to the user’s response. We slightly modify the attention networks introduced in [28] in order to obtain the coordinates of the region that correspond to the user response. In [28], the question itself was fed into the network, so that the region necessary to answer that question is “attended to.” On the other hand, we are already given the answer to the question by the user. We take advantage of this by making simple yet efficient modification, in which we replace the *wh*-question terms with the response provided by the user. For example, a question “what is on the table?” with a user response “pizza” will be converted to a phrase “pizza is on the table,” which is fed into attention network. This is similar to the rule-based NLP conversion in Section 3. We obtain the coordinates of the region from the second attention layer, by obtaining minimum and maximum values for x -axis and y -

axis in which the attention layer reacts to the input phrase. Since the regions are likely to contain the objects of interest at very tight scale, we extracted the regions at slightly larger sizes than coordinates. A region r_i of size (w_{r_i}, h_{r_i}) with coordinates $x_{0_i}, y_{0_i}, x_{max_i}, y_{max_i}$ for image I of size (W, H) is extracted with a magnifying factor α (set as 0.25):

$$r'_i = (\max(0, x_{0_i} - w_{r_i}\alpha), \max(0, y_{0_i} - h_{r_i}\alpha), \min(W, x_{max_i} + w_{r_i}\alpha), \min(H, y_{max_i} + h_{r_i}\alpha)) \quad (4)$$

Given the region and its features, we can now apply the image narrative generation process described in Section 3 with minor modifications in setting. Regions are further extracted, visual questions are generated and answered, and rule-based natural language processing techniques are applied to organize them. Figure 4 shows an overall workflow of our model.

4.2. Applying Interaction to New Images

We represent each instance of image, question, and user choice as a triplet consisting of image feature, question feature, and the label vector for the user’s answer. In addition, collecting multiple choices from identical users enables us to represent any two instances by the same user as a pair of triplets, assuming source-target relation. With these pairs of triplets, we can train the system to predict a user’s choice on a new image and a new question, given the same user’s choice on the previous image and its associated question. User’s choice x_{ans_i} is represented as one-hot vector where the size of the vector is equal to the number of possible choices. We refer to the fused feature representation of this triplet consisting of image, question, and the user’s choice as **choice vector**.

We now project the image feature x_{img_j} and question feature x_{q_j} for the second triplet onto the same embedding space as the choice vector. We can now train a softmax classification task in which the feature from the common embedding space predicts the user’s choice x_{ans_j} on new question. In short, we postulate that the answer with index u , which maximizes the probability calculated by LSTM, is to be chosen as x_{ans_i} by the user who chose x_{ans_k} , upon

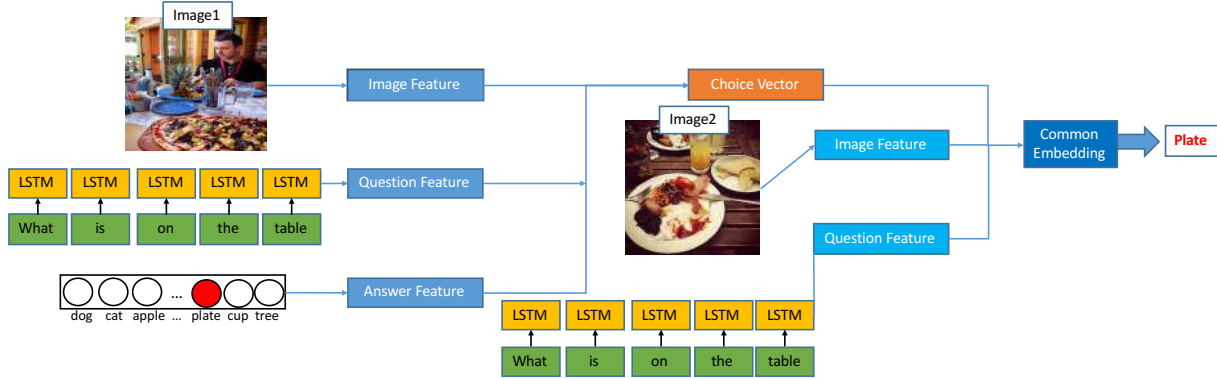


Figure 5: Training with pair of choices made by the same user. Given the choice vector for image 1 and new image feature and question feature for image 2, it is trained to predict the answer for the question on image 2.

seeing a tuple (x_{img_l}, x_{q_l}) of new image and new question:

$$u = \arg \max_v P(v; c_k, x_{img_l}, x_{q_l}) \quad (5)$$

where P is a probability distribution determined by softmax over the space of possible choices, and c_k is the choice vector corresponding to $(x_{img_k}, x_{q_k}, x_{ans_k})$. This overall procedure and structure are essentially identical as in VQA task, except we augment the feature space to include choice vector. Figure 5 shows the overall workflow for training.

5. Experiments

5.1. Automatic Image Narrative Generation

5.1.1 Setting

We applied the model described in Section 3 to 40,775 images in test 2014 split of MS COCO [17]. We compare our proposed model to three baselines as following:

Baseline 1 (COCO): general captioning trained on MS COCO applied to both images in their entirety and the region proposals

Baseline 2 (SIND): captions with model trained on MS SIND dataset [11], applied to both images in their entirety and the region proposals

Baseline 3 (DenseCap): captions generated by DenseCap [12] at both the whole images and regions with top 5 scores using their own region extraction implementation.

5.1.2 Evaluation

Automatic Evaluation: It is naturally of our interest how humans would actually write image narratives. Not only can we perform automatic evaluation for reference, but we can also have a comprehension of what characteristics would be shown in actual human-written image narratives. We collected image narratives for a subset of MS COCO dataset ¹. We asked the workers to write a 5-sentence narrative about the image in a story-like way. We made it clear that the

¹<http://www.mi.t.u-tokyo.ac.jp/projects/narrative>

Table 2: Examples of human-written image narratives collected on Amazon Mechanical Turk.

Image	Human-written Narrative
	This cat is having fun. She is very confused about the change in carpet. It is funny that this has interested her so much. Cats are very picky and they do not like changes. She is probably mad about this.
	The pizza cook makes the pizza. The couple looks forward to pizza. The oven is very hot. He is a master at making pizza. He was born in Italy.

Table 3: Performances of the generated image narratives with human-written image narratives as ground truth.

Model	BLEU1	BLEU2	BLEU3	BLEU4
COCO	13.97	6.13	2.85	1.39
SIND	13.39	2.99	0.82	0.18
DenseCap	20.77	9.26	4.15	1.90
Ours	20.87	8.71	3.58	1.41

description can involve not only factual description of the main event, but also local elements, sentiments, inference, imagination, etc., provided that it can relate to the visual elements shown in the image. Table 2 shows examples of actual human-written image narratives collected and they display a number of intriguing remarks. On top of the elements and styles we asked for, the participants actively employed many other elements encompassing humor, question, suggestion, etc. in a highly creative way. It is also clear that conventional captioning alone will not be able to capture or mimic the semantic diversity present in them.

We performed automatic evaluation with BLEU [21] with collected image narratives as ground truth annotations. Table 3 shows the results. While resemblance to human-

Table 4: Each model’s performance on DIANE.

Metric	COCO	SIND	DenseCap	Ours
Diversity	2.972	2.060	3.102	3.580
Interesting	2.875	2.100	3.336	3.489
Accuracy	2.812	2.105	3.188	3.132
Naturalness	2.754	2.059	3.146	3.374
Expressivity	2.819	2.141	3.257	3.381
Overall	2.846	2.093	3.201	3.391
% of Win.	.300	.195	.357	.400

Table 5: Against each model on χ^2 with 2 degrees of freedom, and one-sided p -value from binomial probability.

vs. Model	>	=	<	χ^2	p -value
COCO	2,208	1,222	1,570	133.37	1.4e-25
SIND	2,970	538	1,492	812.93	1.1e-11
DenseCap	1,890	1,454	1,656	271.33	4.5e-05

written image narratives may not necessarily guarantee better qualities, our model, along with DenseCap, showed highest resemblance to human-written image narratives. As we will see in human evaluation, such tendency turns out to be consistent, suggesting that resemblance to human-written image narratives may indeed provide a meaningful reference.

Human Evaluation: We asked the workers to rate each model’s narrative with 5 metrics that we find essential in evaluating narratives; *Diversity*, *Interestingness*, *Accuracy*, *Naturalness*, and *Expressivity* (DIANE). Evaluation was performed for 5,000 images with 2 workers per image, and all metrics were rated in the scale of 1 to 5 with 5 being the best performance in each metric. We asked each worker to rate all 4 models for the image on all metrics.

Table 6 shows example narratives from each model. Table 4 shows the performance of each model on the evaluation metrics, along with the percentage of each model receiving the highest score for a given image, including par with other models. Our model obtained the highest score on *Diversity*, *Interestingness* and *Expressivity*, along with the highest overall score and the highest percentage of receiving best scores. In all other metrics, our model was the second highest, closely trailing the models with highest scores. Table 5 shows our model’s performance against each baseline model, in terms of the counts of wins, losses, and pars. χ^2 values on 2 degrees of freedom are evaluated against the null hypothesis that all models are equally preferred. The rightmost column in Table 5 corresponds to the one-sided p -values obtained from binomial probability against the same null hypothesis. Both significance tests provide an evidence that our model is clearly preferred over others.

Discussion: General image captioning trained on MS COCO shows weaknesses in accuracy and expressivity. Lower score in accuracy is presumably due to quick diversion from the image contents as it generates captions

directly from regions. Since it is restricted by an objective of describing the entire image, it frequently generates irrelevant description on images whose characteristics differ from typical COCO images, such as regions within an image as in our case. Story-like captioning trained on MS SIND obtained the lowest scores in all metrics. In fact, examples in Table 6 display that the narratives from this model are almost completely irrelevant to the corresponding images, since the correlation between single particular image and assigned caption is very low. DenseCap turns out to be the most competitive among the baseline models. It demonstrates the highest accuracy among all models, but shows weaknesses in interestingness and expressivity, due to their invariant tone and design objective of factual description. Our model, highly ranked in all metrics, demonstrates superiority in many indispensable aspects of narrative, while not sacrificing the descriptive accuracy.

5.2. Interactive Image Narrative Generation

5.2.1 Setting

We first need to obtain data that reflect personal tendencies of different users. Thus, we not only need to collect data from multiple users so that individual differences exist, but also to collect multiple responses from each user so that individual tendency of each user can be learned.

We generated 10,000 questions that allow for multiple responses following the procedure described in Section 4. We grouped every 10 questions into one task, and allowed 3 workers per task so that up to 3,000 workers can participate. Since multiple people are participating for the same group of images, we end up obtaining different sets of responses that reflect each individual’s tendency.

We have permutation of 10 choose 2, $P(10, 2) = 90$ pairs of triplets for each user, adding up to 270,000 pairs of training data. Note that we are assuming a source-to-target relation within the pair, so the order within the pair does matter. We randomly split these data into 250,000 and 20,000 for training and validation splits, and performed 5-fold validation with training procedure described in Section 4. With 705 labels as possible choices, we had an average of 68.72 accuracy in predicting the choice on new image, given the previous choice by the same user. Randomly matching the pairs with choices from different users seemingly drops the average score down to 45.17, confirming that the consistency in user choices is a key point in learning preference.

5.2.2 Evaluation

Question Generation: For question generation, our interest is whether our model can generate questions that allow for various responses, rather than single fixed response. We asked the workers on Amazon Mechanical Turk to decide

Table 6: Examples of image narratives. See Supplemental Material for many more examples.





Image	COCO	SIND	DenseCap	Ours
	An elephant standing in a field of grass. A large elephant standing in a field of grass. A bathroom with a sink and a mirror. A large building with a clock on it.	The dog was very happy to see the animals. We had a great time. I went to the museum today. We went to the city to see the sights. We saw a lot of old buildings.	An elephant standing in a field of grass. A gray elephant. Elephant trunk is curled. Elephant in the photo. Trunk of an elephant. Elephants walking on the road.	This is a baby elephant. The elephants are standing on grass. They are bored. The elephant is sitting. This is not a zoo.
	A teddy bear sitting on top of a wooden table. A teddy bear sitting on top of a bed. A stuffed bear is sitting on a bed.	The cake was delicious. I had a great time. The food was delicious.	A teddy bear sitting on top of a wooden table. Teddy bear on a table. A brown teddy bear. A teddy bear. A teddy bear on a table. The head of a teddy bear.	Bear is on the cake. That stuffed animal has a funny face. It is mine. The bear is wearing hat. The cake is white.

Table 7: Evaluation results on whether the generated questions allow for multiple responses.

Model	# Overall	# Yes	# No
Ours	1,000	664	336
VQG	1,000	217	783
Overall	2,000	881	1119

Table 8: Examples of generated questions using our proposed model and VQG respectively.

Image	Generated Questions
	Ours
	What is the color of the shirt?
	VQG
	How many children are there?
	Ours
	What is on the table?
	VQG
	What is the table made of?

whether the question can be answered in various ways or has multiple answers, given an image. 1,000 questions were generated with our proposed model using both VQG and VQA, and another 1,000 questions were generated using VQG only.

Table 7 shows the number of votes for each model. It is very clear that the questions generated from our proposed model of parallel VQG and VQA outperformed by far the questions generated from VQG only. This is inevitable in a sense that VQG module was trained with human-written questions that were intended to train the VQA module, i.e. with questions that mostly have clear answers. On the other hand, our model deliberately chose the questions from VQG that have evenly distributed probabilities for answer labels, thus permitting multiple possible responses. Table 8 shows examples of visual questions generated from our model and VQG only respectively. In questions generated from our

model, different responses are possible, whereas the questions generated from VQG only are restricted to single obvious answer.

Reflection of User’s Choice on the Same Image: Our next experiment is on the user-dependent image narrative generation. We presented the workers with 3,000 images and associated questions, with 3 possible choices as a response to each question. Each worker freely chooses one of the choices, and is asked to rate the image narrative that corresponds to the answer they chose, considering how well it reflects their answer choices. As a baseline model, we examined a model where the question is absent in the learning and representation, so that only the image and the user input are provided. Rating was performed over scale of 1 to 5, with 5 indicating highly reflective of their choice. Table 11 shows the result. Agreement score among the workers was calculated based on [3]. Agreement score for our model falls into the range of ‘moderate’ agreement, whereas, for baseline model, it is at the lower range of ‘fair’ agreement, as defined by [15], demonstrating that the users more frequently agreed upon the reliability of the image narratives for our model. Our model clearly has an advantage over using image features only with a margin considerably over standard deviation. Table 9 shows examples of images, generated question, and image narratives generated depending on the choice made for the question respectively.

Reflection of User’s Choice on New Images: Finally, we experiment with applying user’s interest to new images. As in the previous experiment, each worker is presented with an image and a question, with 3 possible choices as an answer to the question. After they choose an answer, they are presented with a new image and a new image narrative. Their task is to determine whether the newly presented image narrative reflects their choice and interest. As a baseline, we again examined a model where the question is ab-

Table 9: Examples of image narratives generated depending on the user choices.





Image	Answers, Regions and Narratives		
	Skateboard	Motorcycle	Car
			
Generated Question What is the man riding?	The man is riding skateboard. The man is skateboarding. The color of the jacket is red.	The man is riding motorcycle. It is white. The motorcycle is honda.	The man is riding car. This is a modern car. It is a black and white photo.

Table 10: Examples of image narratives generated on new images, depending on the choices made.





Image & Question	Choice	New Image	Image Narrative
 What animal is this?	giraffe		The giraffe is standing. The weather is sunny.
	zebra		Zebra is thinking. It is not in a zoo.
	rhino		2 animals are in the picture. The sky is blue.
 What kind of animal is that?	dog		The horse is running. The car is white.
	sheep		The boy is wearing red shirt. Tree is in the background.
	person		The man is riding horse. The man is wearing hat.

Table 11: Evaluation results on how well the generated image narrative reflects the choices they made.

Model	Avg. Score	Agreement
Ours	3.851±1.12	.601
image only	2.636±1.01	.432

Table 12: Evaluation results on how well the generated image narrative for new images reflects their interest.

Model	Avg. Score	Agreement
Ours	3.455±0.93	.527
random match	2.772±0.79	.489
image only	2.238±1.24	.428

sent in the learning and representation stages. In addition, we performed an experiment in which we trained preference learning module with randomly matched choices. This allows us to examine whether there exists a consistency in user choices that enables us to apply the learned preferences to new image narratives.

Table 12 shows the result. As in previous experiment, our model clearly has an advantage over using image features only. Inter-rater agreement score is also more stable for our model. Training preference learning module with randomly matched pairs of choices resulted in a score below our proposed model, but above using the image features only. This may imply that, even with randomly matched pairs, it is better to train with actual choices made by the

users with regards to specific questions, rather than with conspicuous objects only. Overall, the result confirms that it is highly important to provide a context, in our case by generating visual questions, for the system to learn and reflect the user’s specific preferences. It also shows that it is important to train with consistent choices made by identical users. Table 10 shows examples of image narratives generated for new images, depending on the choice the users made for the original image, given the respective questions.

6. Conclusion

We proposed a customized image narrative generation task, where we proposed a model to engage the users in image description generation task, by directly asking questions to the users, and collecting answers. Experimental results demonstrate that our model can successfully diversify the image description by reflecting the user’s choice, and that user’s interest learned can be further applied to new images.

Acknowledgments

This work was partially funded by the ImPACT Program of the Council for Science, Technology, and Innovation (Cabinet Office, Government of Japan), and was partially supported by CREST, JST.

References

- [1] A. Agrawal, D. Batra, and D. Parikh. Analyzing the behavior of visual question answering models. In *EMNLP*, 2016.
- [2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.
- [3] E. Bennett, R. Alpert, and A. Goldstien. Communications through limited-response questioning. *Public Opinion Quarterly*, 18:303–308, 1954.
- [4] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. <https://arxiv.org/abs/1706.03741>, 2017.
- [5] A. Das, S. Kottur, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *CVPR*, 2017.
- [6] H. de Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. C. Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*, 2017.
- [7] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *CVPR*, 2015.
- [8] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. *Emnlp*. 2016.
- [9] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. J. V. Gool. Deepproposal: Hunting objects by cascading deep convolutional layers. In *ICCV*, 2015.
- [10] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [11] T.-H. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell. Visual storytelling. In *NAACL*, 2016.
- [12] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *CVPR*, 2016.
- [13] K. Kafle and C. Kanan. Visual question answering: Datasets, algorithms, and future challenges. <https://arxiv.org/abs/1610.1465>, 2016.
- [14] A. Karpathy and F.-F. Li. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015.
- [15] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 1977.
- [16] P. D. Larry Zitnick. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.
- [17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014.
- [18] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016.
- [19] M. Marcus, G. Kim, M. A. Marcinkiewicz, R. MacIntyre, A. Bies, M. Ferguson, K. Katz, and B. Schasberger. The penn treebank: annotating predicate argument structure. In *HLT*, 1994.
- [20] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *ACL*, 2016.
- [21] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002.
- [22] M. Ren, R. Kiros, and R. Zemel. Exploring Models and Data for Image Question Answering. In *NIPS*, 2015.
- [23] K. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, 2016.
- [24] A. Shin, Y. Ushiku, and T. Harada. The color of the cat is gray: 1 million full-sentences visual question answering (fsvqa). *arXiv:1609.6657*, 2016.
- [25] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *IJCV*, 104:154–171, 2013.
- [26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, 2015.
- [27] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*, 2015.
- [28] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.