Let's Talk! Striking Up Conversations via Conversational Visual Question Generation

Shih-Han Chan[1] **Tsai-Lun Yang**[1] **Yun-Wei Chu**[1] **Chi-Yang Hsu**[1,2]
Ting-Hao (Kenneth) Huang[2] **Yu-Shian Chiu**[3] **Lun-Wei Ku**[1]
[1]Academia Sinica, [2]The Pennsylvania State University, [3]Institute for Information Industry
[1]{hank08tw,a7532ariel,yunweichu,lwku}@iis.sinica.edu.tw, [2]{cxh5437,txh710}@psu.edu,
[3]samuelchiu@iii.org.tw

Abstract

An engaging and provocative question can open up a great conversation. In this work, we explore a novel scenario: a conversation agent views a set of the user's photos (for example, from social media platforms) and asks an engaging question to initiate a conversation with the user. The existing vision-to-question models mostly generate tedious and obvious questions, which might not be ideals conversation starters. This paper introduces a two-phase framework that first generates a visual story for the photo set and then uses the story to produce an interesting question. The human evaluation shows that our framework generates more response-provoking questions for starting conversations than other vision-to-question baselines.

Introduction

Question-asking play an essential role in human conversations. Studies have shown that people who ask more questions in interpersonal conversations are better liked by their conversational partners (Huang et al. 2017). For automated social bots, prompting the user with a question is known to be an effective way to initiate conversations. For example, Wang et al. (2018) generates questions in diverse yet relevant topics to enhance the interactiveness and persistence of conversations. Pan et al. (2019) also uses a Reinforced Dynamic Reasoning network to produce meaningful questions to engage users in conversations. In this work, we explore a novel scenario: an automated conversation agent "views" a user's photos — for example, from social media platforms, or shared by the user proactively — and asks an engaging question to initiate a conversation with the user. This scenario uses images shared by the user to start a conversation, enriching the plausible topics of the human-agent conversations. For example, when a user posts a set of wedding photos with only a little or even no descriptions, the proposed conversational agent can still ask questions about the wedding. Such machine-generated questions can be used to engage user for social purposes; to allow robots to elicit situational information from passersby (Krishna, Bernstein, and Fei-Fei 2019); or to support memory therapies (*e.g.*, reminiscence therapy) or activities that use the patients personal images as memory prompts (Bhar 2014).

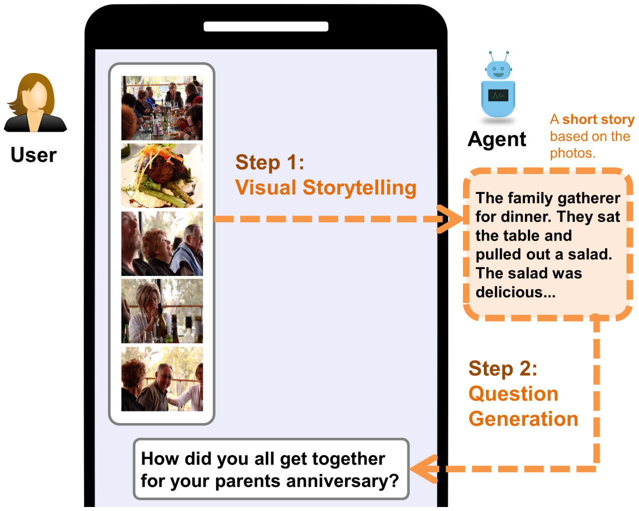

User
Agent
A short story based on the photos.
Step 1: Visual Storytelling
The family gatherer for dinner. They sat the table and pulled out a salad. The salad was delicious...
Step 2: Question Generation
How did you all get together for your parents anniversary?

Figure 1: The agent receives five images from the user and comes up with an short story based on the images. Then the agent asks a response-provoking question based on the story.

However, the existing vision-to-question models, such as visual question answering (VQA) (Antol et al. 2015), do not generate questions with the purpose of engaging users or provoking users' responses (Krishna, Bernstein, and Fei-Fei 2019). For example, many VQA questions are about the factual property, such as color, size, and shape, of the objects in the image, rather than human activities or broader contexts of the image. In this paper, instead of producing generic questions using one image, our proposed model takes a sequence of images as input and generates an engaging and provocative question based on these images to start a conversation with the user.

This paper introduces a two-phase framework that first generates a **visual story** for the image sequence and then uses the story to produce an interesting question. This approach takes advantage of the existing "visual storytelling" (VIST) technologies, where the model generates an engaging short story based on a sequence of images (Huang et al. 2016). We use one of the state-of-the-art visual storytelling model, KG-Story (Hsu et al. 2020), to produce a visual story, which is then fed into a Transformer-based model to create engaging questions. Figure 1 illustrates the scenario of the proposed framework. The human evaluation shows that our framework generates more response-provoking questions than other vision-to-question baselines.

The contribution of this work is three-fold:

- We are the first to introduce a unique scenario: inputting a sequence of images, the system then asks the user an engaging and response-provoking question to start a conversation.
- We propose a two-stage framework to perform visual question generation: the system first generates stories based on the image sequence. It then generates an engaging question based on the story.
- We conduct a human evaluation using the VIST dataset to show that the questions generated by the proposed approach are better at invoking human desire to chat than traditional vision-to-question methods.

Related Work

Question Generation Question generation usually takes the datasets designed for question answering and generates questions based on given textual context (*e.g.*, paragraph or historical conversation). Most traditional approaches design an end-to-end structure by implementing the recurrent neural network (Duan et al. 2017). Some implement attention mechanism (Bahdanau, Cho, and Bengio 2014) to enhance embedding features (Zhou et al. 2017). With the benefit brought by Transformer (Devlin et al. 2018) for language modeling, Kettip Kriangchaivech (2019) starts developing question generation model based on Transformer. However, existing text-based question generation is hard to start a conversation since the lack of informative material, and most works still use historical conversation to generate a question.

Visual Question Generation With the help of external sources, Visual Question Generation (VQG) aims to generate a question based on a given image. The idea of VQG came from Visual Question Answering (VQA) (Antol et al. 2015). However, the questions in the VQA task are designed limited to objects, colors, numbers, or locations. Mostafazadeh et al. (2016b) introduced the VQG dataset, where the system is asked to generate a question for people to answer. Most of approaches focus on leveraging seq2seq model to generate questions (Patro et al. 2018). However, generating questions from the image feature lacks a comprehensive understanding of visual input as the model still briefly asks questions about the objects in the image.

Image Captioning We conducted a survey on image captioning, a task that the model should use a sentence to describe one image, which can help comprehensively understand the visual inputs. The model should learn representations of the interdependence between the objects/concepts in the image and use them to describe the image factually. Kiros, Salakhutdinov, and Zemel (2014) first introduced a deep learning based method on this topic by using CNN to extract image features and a language model to generate captions. Some work proposed architecture or applied attention mechanism based on recurrent neural network (Vinyals et al. 2014; Xu et al. 2015). However, describing the image is usually not considered attractive by humans. If we want our system to communicate with humans, it must capture their interests and avoid just stating the obvious.

Visual Storytelling Visual storytelling (VIST) was proposed by (Huang et al. 2016). Unlike image captioning which generates a sentence describing the image, VIST asks the model to generate a story based on 5 images. Visual stories should be descriptive, relevant to the images, and appealing to human readers. Most of the approaches focus on developing end-to-end models or adopting various training techniques on the VIST dataset (Kim et al. 2018). Since there is only one existing VIST dataset, these end-to-end VIST works often limit the knowledge to this dataset. To avoid overfitting on the dataset, some research leverage external sources to enrich story contents. Since knowledge graphs have shown beneficial on language modeling, most of the VIST work use knowledge graphs (KG) to enrich stories (Hsu et al. 2020). Among all the KG-based VIST, KG-Story (Hsu et al. 2020) designed a three-stage framework, which uses Visual Genome knowledge graph (Krishna et al. 2017) to add semantic relations between two adjacent images, generated more interesting and coherent stories. Since stories can generate more creative content from visual inputs than captions, our framework takes visual stories to produce response-provoking questions.

Methods

The overall framework contains two stages, as described in Figure 2. The system first generates stories based on input vision as intermediate perception. From the generated stories, the system then performs question generation and outputs an response-provoking question.

Stage 1: Visual Storytelling (KG-Story)

We implement KG-Story[1] (Hsu et al. 2020), a three-stage story generation framework, as our story generation model. KG-Story extracts representative terms from input images, enriches terms by knowledge graph, and eventually generates stories based on enriched terms.

1. **Extracting representative terms:** Given a sequence of images, KG-Story uses a pre-trained Faster-RCNN (Ren et al. 2015) as the object detection model. To reduce computational complexity, only the objects within the top 25 confidence scores are used. Since objects lack semantic meaning, KG-Story designs a Transformer based model that transforms objects to terms (*e.g.*, objects and actions). From VIST dataset, they use SpaCy[2] and Open-SASEME (Swayamdipta et al. 2017) to parse stories into object nouns and semantic frames as ground truth terms. Taking predicted objects from Faster-RCNN as input and parsed terms as output, a Transformer encoder (Vaswani

[1] KG-Story: https://github.com/zychen423/KE-VIST

[2] SpaCy: https://spacy.io/

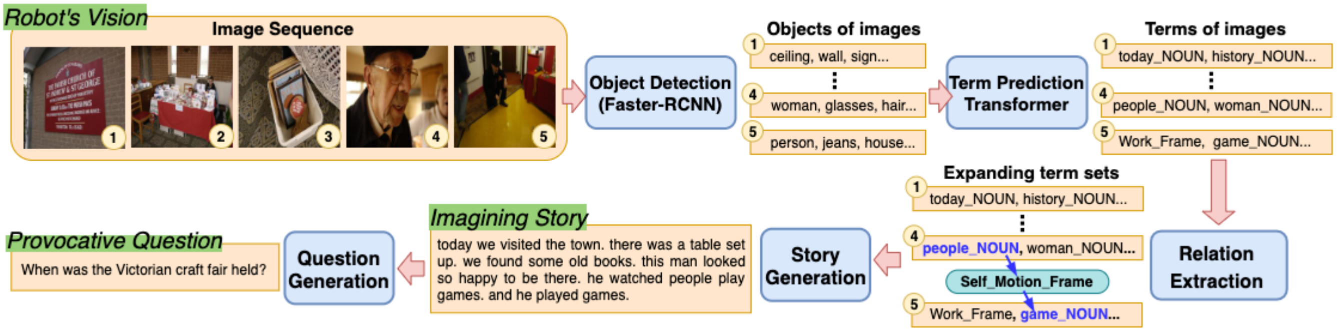

Robot's Vision
Image Sequence
1
2
3
4
5
Object Detection (Faster-RCNN)
Objects of images
1 ceiling, wall, sign...
4 woman, glasses, hair...
5 person, jeans, house...
Term Prediction Transformer
Terms of images
1 today_NOUN, history_NOUN...
4 people_NOUN, woman_NOUN...
5 Work_Frame, game_NOUN...
Relation Extraction
Expanding term sets
1 today_NOUN, history_NOUN...
4 people_NOUN, woman_NOUN...
Self_Motion_Frame
5 Work_Frame, game_NOUN...
Story Generation
Imagining Story
today we visited the town. there was a table set up. we found some old books. this man looked so happy to be there. he watched people play games. and he played games.
Question Generation
Provocative Question
When was the Victorian craft fair held?

Figure 2: The pipeline of proposed framework. We first ask the system to think by generating stories based on given image sequence. The system then produces provocative question based on the story.

et al. 2017) and a GRU decoder with an attention mechanism (Bahdanau, Cho, and Bengio 2014) are used as term prediction model.

2. **Knowledge enrichment:** Since previous end-to-end models tend to generate caption-like incoherent stories which are relatively boring, KG-Story enriches stories by introducing knowledge graph. Knowledge graph serves as the source of ideas that connects two images and ensures the coherence of logic. KG-Story link terms in two adjacent images using the relations provided by Visual Genome knowledge graph (Krishna et al. 2017).

   Given 5 images, the extracted terms from Stage 1 are represented as $\{m_1^1, ..., m_i^t, ..., m_{N_5}^5\}$, where $\{m_1^1, ..., m_h^1\}$ denotes first image's term set, m_i^t denotes the i-th term from image t and N_k is the number of terms from image k. From consecutive images, KG-Story explores all possible relations (m_i^t, r, m_j^{t+1}) and $(m_i^t, r_1, m_{middle}, r_2, m_j^{t+1})$, while m_{middle} denotes a knowledge graph entity that bridges m_i^t and m_j^{t+1}. With all possible relations, KG-Story uses a RNN-based language model to obtain a relation with lowest perplexity. The chosen relation is inserted to original term sequence expanding the number of term sets from 5 to 6.

3. **Story generation:** To generate stories, KG-Story leverages Transformer (Vaswani et al. 2017) with expanded term sets from Stage 2 as input. Three modifications are made for the original Transformer model. (1) The length-difference positional encoding is adopted to perform variable-length story generation. Since all the samples of VIST dataset contain five images, this mechanism allows KG-Story to generate additional sentence. (2) Anaphoric expression generation is used for the unification of anaphor representation. To enable the use of pronouns, KG-Story uses a coreference resolution tool [3] on the stories to find the original mention of each pronoun. (3) A designed repetition penalty for inter- and intra-sentence with beam search are adopted to reduce redundancy. After feeding term sets into designed Transformer, the model generates a knowledgeable story with 6 sentences. We then use stories as robot's perception and generate an response-provoking question in the next step.

Stage 2: Response-Provoking Question Generation from the Story

We utilize a Transformer-based end-to-end model (Lopez et al. 2020) as our question generation model. For the pretrained model, we use Hugging-Faces implementation (Wolf et al. 2020) of the 60 million parameters T5, the smallest of the five available T5 model sizes.

As pre-trained T5 has shown a strong ability to solve the text-to-text problem, we finetune it by taking the paragraphs of Stanford Question Answering Dataset (SQuAD) as input and the question as output. The entire dataset is firstly transformed into a continuous body of text; each training sample consists of a context paragraph and associated question(s) transformed into a single continuous sequence with a "delimiter" in between. During training, the delimiter enables the model to successfully distinguish between context paragraph and corresponding question(s), while during inference, the delimiter acts as a marker at the end of context to invoke question generation behavior of the model.

After pretraining and finetuning, the model then takes the generated stories as input and generates a question. Higher Temperature values give more randomness to question generation, while lower values approach greedy behavior. We set Temperature to 0.6.

The generation process use the top-p nucleus sampling method with a value $p = 0.9$, which allows for more diverse generations than a purely greedy scheme, and minimizes the occurrence of certain tokens or token spans repeating indefinitely in the generated text. For each context paragraph input, the question generation stops when the model reaches the generation length of 26 tokens or the model generates a newline character \n. We set the maximum length to terminate the question generation of some context that don not reach the \n on their own.

Experimental Setups

In this section, we first introduce the datasets we use for model training and performance evaluation. We then provide details of the baseline models we compare to and all models'

[3] NeuralCoref 4.0: Coreference Resolution in spaCy with Neural Networks. https://github.com/huggingface/neuralcoref

hyperparameter settings. Lastly, we crowdsource our human evaluation on Amazon Mechanical Turk: ask the workers to rank the performance of generated questions and conduct user study.

Data Preparation

Four datasets are used in this paper: Visual Genome, ROC-Stories, SQuAD, and VIST. For story generation part, VIST is used to extract terms from images (Stage 1) and fine-tune the story generation model (Stage 3). Visual Genome knowledge graph is used for relation linking (Stage 2) between the extracted terms. ROCStories supplies a large quantity of pure textual stories for pre-training the story generator (Stage 3). For question generation part, Stanford Question Answering Dataset (SQuAD) is used to train the question generation model. The test data of VIST are used to examine the performance of our proposed framework. The detail of each dataset is described bellow.

- **Visual Genome:** Visual Genome (Krishna et al. 2017) is a knowledge-based dataset that connects images concepts to language. It has 108,077 images, 3.8 million object instances, and 2.3 million relationships. The knowledge graph we utilize covers nouns and relations, categorized into semantic frames, provided by the scene graph of Visual Genome. Compared with most image-to-text works that focus on objects nouns and generate static stories, adding logical relation and activities frames makes our stories reasonable and vivid.
- **ROC-Stories Corpora:** We use the ROC-Stories (Mostafazadeh et al. 2016a), which contains 98,159 pure textual stories, to pre-train our story generator. As the annotators were asked to write five-sentence stories given a prompt, ROC-Stories focuses on specific, everyday topics.
- **SQuAD:** SQuAD (Rajpurkar et al. 2016) is a reading comprehension dataset consisting of 100,000+ question-answer pairs posted by crowdworkers on a set of Wikipedia articles, which contain 23,215 paragraphs covering a wide range of topics. We use the paragraphs and questions in SQuAD to train the question model.
- **VIST:** VIST (Huang et al. 2016) is a sequential vision-to-language dataset that moves visual understanding from basic perspective to more human-like understanding of grounded event structure. We train KG-Story model and conduct experiments on VIST, which includes 10,117 Flicker albums with 210,819 unique images. We follow the standard split setting as previous work, with 40,098 samples for training, 4,988 for validation, and 5,050 for testing. Each sample contains one story that describes five images from a photo stream.

Hyperparameter Configuration

In all of our experiment, we use the same hyperparameters the authors mentioned in the KG-Story paper. For term prediction (Stage 1) and story generation (Stage 3), the hidden size is set to 512. The number of head and layer of the Transformer encoder are 2 and 4. All KG-Story models are trained with Adam optimizer (Kingma and Ba 2015) with initial learning rate 1e-3.

For the training parameters of question generation model, the model is trained for 3 epochs using general language modeling loss. Adam optimizer was also applied with an initial learning rate of 5e-5 and a linearly decreasing learning rate scheduler with warm up for 10% of total training steps.

Baselines

We compare our proposed story-to-question concept with two state-of-the-art frameworks. The first framework is an end-to-end image question generation model (Mostafazadeh et al. 2016b), which aims to generate an engaging question given an image. The second framework is an image captioning model (Xu et al. 2015), which generates a sentence through convolutional neural network and recurrent neural network to describe the content of an image. We then concatenate captions and perform question generation on captions. The purpose is to examine the performance of questions generated by different perception, captions or stories.

- **Image Question Generation:** We use Gated Recurrent Neural Network[4] (Mostafazadeh et al. 2016b) as our question generation baseline model. This model is trained on the VQA dataset [5] (Antol et al. 2015), which contains 204,721 COCO images and at least 3 questions per image. The model uses a pre-trained 19-layer VGG Net (Simonyan and Zisserman 2015) for encoding image features. It transforms the 4096-dimensional output of the VGG-19's last fully connected layer (fc7) to a 512-dimensional vector that serves as the initial state of a long-short term memory unit (LSTM) to generate the corresponding question.
- **Image Captioning:** We use the model Xu et al. (2015) proposed [6] as the image captioning baseline. The model consists of a 101-layer ResNet pre-trained on the ImageNet classification task, a soft attention network, and a 512-dimensional LSTM. A linear layer is used to map the encoded images to the initial hidden and cell states for the LSTM. The Attention network considers the sequence generated thus far and attends to the part of the image that needs describing next. The LSTM is used to produce the output caption one word at a time conditioned on the context vector, the previous hidden state, and the previously generated word. The whole model is trained on the MS COCO '14 Dataset [7].

The baseline models, either image question generation model or image captioning model, take 5 images as input and generate 5 sentence, either 5 questions or 5 captions. To align all settings, we concatenate 5 sentences from the baseline models and feed them into the same question generation model we use to generate questions for comparison.

[4] https://github.com/chingyaoc/VQG-tensorflow

[5] https://visualqa.org/

[6] https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning

[7] https://cocodataset.org/

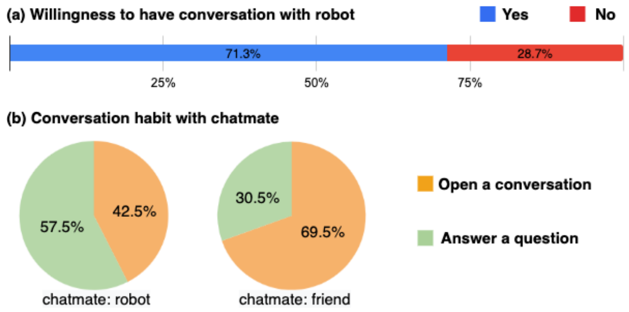


Figure 3: The result of survey question (1) that asks (a) “Will annotators have conversation with robots?”, and questions(2) & (3) that ask (b) “Which conversation habit annotators prefer? with robot or friends as chatmate”.

Human Evaluation

We conduct human evaluation using crowd workers from Amazon Mechanical Turk (MTurk) to evaluate the quality of generated questions. We randomly selected 250 photo sequences from the test set of VIST dataset and use three models (our model and two baselines) to generate questions for each photo sequence. In the MTurk tasks, we show the photo sequence and three generated questions to the workers. Each worker is asked “which question is the best to start a conversation or keep the conversation continue” and instructed to rank three questions. We collect five response from five different workers for each photo sequence.

We also design questionnaires to explore potential directions of future work, in which we ask annotators three questions: *(1)* If the chatmate is a **robot**, will you have a conversation with it? (Yes/No); *(2)* Which one you would prefer if your chatmate is a **robot**? start a conversation yourself/wait for the robot to start the conversation; and In Question *(3)* we ask again the same questions as (2), but change the chatmate to **your friend**.

Results and Analysis

User Survey The results of the designed questionnaires are shown in Figure 3. Among 3,750 annotators, exceeding 70% annotators accept having a conversation with a robot, which validates chatbots’ validity. Figure 3(b) shows the result of the questionnaire that asks users’ conversation habits. If the chatmates are annotators’ friends, approximately 70% of them tend to open a conversation actively. However, if their chatmates are robots, they would wait for the robots to start a conversation and answer a question. This result shows the importance of the questions’ quality if the chatbot demands to open a conversation with a human.

Quantity Analysis In Table 1, we show the rank of the questions generated from images (Img2Q), from captions (Caption2Q), and from stories (Story2Q). Questions generated by three methods are ranked from 1 to 3 (lower is better). Story2Q receives the best rank among baselines with the highest percentage (44.9%) in rank 1 and the lowest percentage in rank 3 (23.2%). The fact that both Story2Q and Caption2Q outperform Img2Q shows that generating intermediate ingredients can have better understanding of images for generating questions than directly using image features as input. In summary, Story2Q performs the best in generating response-provoking questions.

Method	1st	2nd	3rd	Avg rank
img2Q	19.4%(243)	30.6%	49.9%(624)	2.30
caption2Q	35.7%(447)	37.2%	26.9%(337)	1.91
story2Q	**44.9%(560)**	32.2%	**23.2%(289)**	**1.78**

Table 1: Human ranking evaluation between our proposed framework and two methods. First three columns indicate percentage of workers’ ranking for each method, and last column denotes average rank (1 to 3, lower is better). Numbers in brackets indicates the quantity of the best and the worst stories for each method. The questions generated by our proposed framework is significantly better than all baseline methods with $\rho < 0.05$.

Method	What	Where	When	Why	Who	How	Other
img2Q	70.7%	1.4%	5.1%	3.0 %	0%	13.8%	6.0%
caption2Q	79.2%	1.1%	2.7%	3.7%	0%	7.6%	5.7%
story2Q	55.5%	5.6%	11.6%	15.3%	1.0%	6.3%	4.7%

Table 2: The question distribution of different methods.

In Table 2, we also show the composition of questions generated by each method by counting the 5W1H. We find that Story2Q generates more diverse questions compared to Caption2Q and Img2Q. Story2Q tends to generate various questions beginning with “What”, “Why”, “When”, “Where”, while Img2Q and Caption2Q mostly generate the questions begin with “What”.

Quality Analysis Img2Q prefers to generate literal questions that ask the quantity and information in the images, which is not suitable for starting a conversation. Take the right half of Figure 4 as an example, Img2Q asks “How many people are shown?”, which is considered a literal question that can not make the conversation lasting. Img2Q directly uses features extracted from images to generate questions. The failure of this end-to-end method lets us consider developing a multi-stage model and using extra methods like object detection and relation extraction to support engaging question generation.

Since captioning is good at summarizing all the contents in an image with the attention mechanism’s help, the information in the question strongly matches with the image sequence. However, the questions generated by Caption2Q lack creativity and diversity, since captions are only designed to be faithful to the original images. In the left example of Figure 4, Caption2Q asks “What is the name of a person sitting on a bench?”, since all of the contents can be easily found in the image sequence, the question may be considered boring and obvious by human.

Fortunately, Story2Q generates response-provoking questions with high quality because it firstly generates interesting stories by object detection and relation extraction. Since

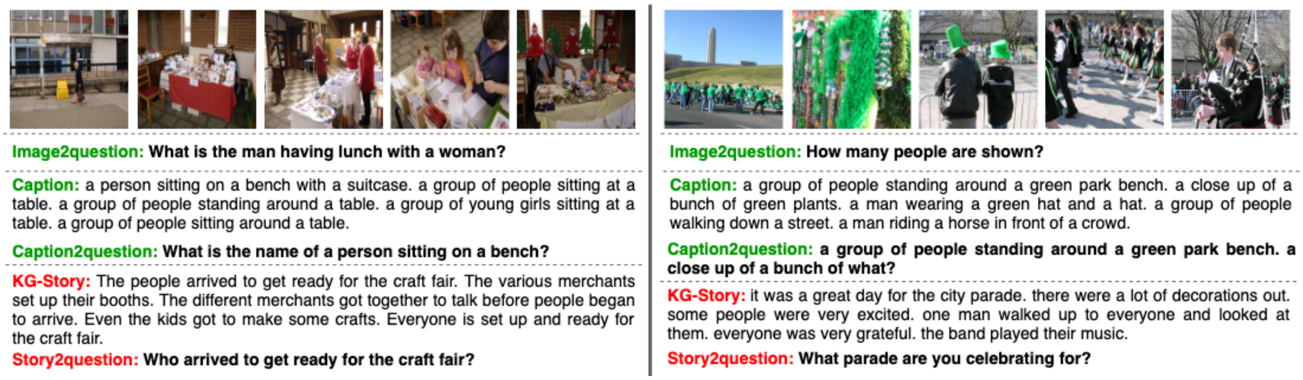


Image2question: What is the man having lunch with a woman?

Caption: a person sitting on a bench with a suitcase. a group of people sitting at a table. a group of people standing around a table. a group of young girls sitting at a table. a group of people sitting around a table.

Caption2question: What is the name of a person sitting on a bench?

KG-Story: The people arrived to get ready for the craft fair. The various merchants set up their booths. The different merchants got together to talk before people began to arrive. Even the kids got to make some crafts. Everyone is set up and ready for the craft fair.

Story2question: Who arrived to get ready for the craft fair?

Image2question: How many people are shown?

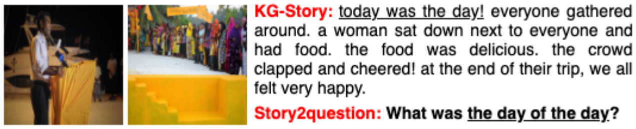
Caption: a group of people standing around a green park bench. a close up of a bunch of green plants. a man wearing a green hat and a hat. a group of people walking down a street. a man riding a horse in front of a crowd.

Caption2question: a group of people standing around a green park bench. a close up of a bunch of what?

KG-Story: it was a great day for the city parade. there were a lot of decorations out. some people were very excited. one man walked up to everyone and looked at them. everyone was very grateful. the band played their music.

Story2question: What parade are you celebrating for?

Figure 4: Examples of questions generated by different methods.



KG-Story: today was the day! everyone gathered around. a woman sat down next to everyone and had food. the food was delicious. the crowd clapped and cheered! at the end of their trip, we all felt very happy.

Story2question: What was the day of the day?

Figure 5: Repetition error example of question generation model.

the story contains deeper relations among different objects in the image sequence, the generated question is profound. In Figure 4, Story2Q asks "Who arrived to get ready for the craft fair?". It takes care of the "arrive" relation between human and craft fair. Paying attention to relationships among different objects is similar to human thinking, so it is more natural and reasonable to start a conversation with this kind of question.

Error Analysis Since the story generation model and question generation model are trained on different datasets(*e.g.*, VIST and SQuAD), we find that question generation model tends to copy the same sentence or predict repetitive words when meeting the sentence it did not learn before. Taking the question generated from a caption in the right example of Figure 4 as an example, the story generation model directly copies the whole sentence of the caption and generates a strange question. In Figure 5, KG-Story sometimes generates sentence that beyond images, like "today was the day!" for this example. This kind of sentences is hard to exist in the SQuAD dataset, thus resulting in the chaos of the question generation model and yielding the model to generate repetitive words.

Conclusion and Future Work

We have introduced a new scenario of visual question generation, in which, when given a sequence of images, the system should generate a provocative question in order to start a conversation. Instead of generating question directly from image features, we ask the model imagine first by generating creative stories to have a better understating of an image sequence, and then produce an engaging question based on stories to start a conversation. Human evaluation results show that our proposed framework significantly outperforms two baseline models, Img2Q and Caption2Q, on the VIST dataset. This provides evidence that thinking before asking can enhance the question's quality and make people want to communicate with the system.

There are several potential future research directions. Our model is currently only evaluated on one vision-to-language dataset (*e.g.*, VIST), and thus we want to explore the generalization of our idea to other datasets. Besides, according to the result of experiments and human feedback, we see this vision-to-question task's potential. We can collect a dataset, which will launch a new challenge to the community and further invoke interests in studying the importance of asking a response-provoking question.

Acknowledgement

This research is partially supported by *(i)* Ministry of Science and Technology, Taiwan under the project contract 108-2221-E-001-012-MY3 and *(ii)* College of IST Seed Fund at Penn State University. We also thank the workers on MTurk who participated in our studies.

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. VQA: Visual Question Answering. *CoRR* abs/1505.00468. URL http://arxiv.org/abs/1505.00468.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. URL http://arxiv.org/abs/1409.0473. Cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

Bhar, S. S. 2014. Reminiscence therapy: A review. .

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Duan, N.; Tang, D.; Chen, P.; and Zhou, M. 2017. Question Generation for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 866–874. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/

D17-1090. URL https://www.aclweb.org/anthology/D17-1090.

Hsu, C.-C.; Chen, Z.-Y.; Hsu, C.-Y.; Li, C.-C.; Lin, T.-Y.; Huang, T.-H. K.; and Ku, L.-W. 2020. Knowledge-Enriched Visual Storytelling. In *Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Huang, K.; Yeomans, M.; Brooks, A. W.; Minson, J.; and Gino, F. 2017. It doesn't hurt to ask: Question-asking increases liking. *Journal of personality and social psychology* 113(3): 430.

Huang, T.-H. K.; Ferraro, F.; Mostafazadeh, N.; Misra, I.; Agrawal, A.; Devlin, J.; Girshick, R.; He, X.; Kohli, P.; Batra, D.; et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1233–1239.

Kettip Kriangchaivech, A. W. 2019. Question Generation by Transformers. *arXiv preprint arXiv:1909.05017* .

Kim, T.; Heo, M.-O.; Son, S.; Park, K.-W.; and Zhang, B.-T. 2018. GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation. *arXiv preprint arXiv:1805.10973* .

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL http://arxiv.org/abs/1412.6980.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. 2014. Multimodal Neural Language Models. volume 32 of *Proceedings of Machine Learning Research*, 595–603. Bejing, China: PMLR. URL http://proceedings.mlr.press/v32/kiros14.html.

Krishna, R.; Bernstein, M.; and Fei-Fei, L. 2019. Information Maximizing Visual Question Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; and et al. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123(1): 32–73. ISSN 1573-1405. doi:10.1007/s11263-016-0981-7. URL http://dx.doi.org/10.1007/s11263-016-0981-7.

Lopez, L. E.; Cruz, D. K.; Cruz, J. C. B.; and Cheng, C. 2020. Transformer-based End-to-End Question Generation.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016a. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696* .

Mostafazadeh, N.; Misra, I.; Devlin, J.; Zitnick, L.; Mitchell, M.; He, X.; and Vanderwende, L. 2016b. Generating Natural Questions About an Image. *CoRR* abs/1603.06059. URL http://arxiv.org/abs/1603.06059.

Pan, B.; Li, H.; Yao, Z.; Cai, D.; and Sun, H. 2019. Reinforced Dynamic Reasoning for Conversational Question Generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2114–2124. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1203. URL https://www.aclweb.org/anthology/P19-1203.

Patro, B. N.; Kumar, S.; Kurmi, V. K.; and Namboodiri, V. P. 2018. Multimodal Differential Network for Visual Question Generation. *CoRR* abs/1808.03986. URL http://arxiv.org/abs/1808.03986.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *CoRR* abs/1606.05250. URL http://arxiv.org/abs/1606.05250.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 91–99.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015 . *arXiv preprint arXiv:1409.1556* .

Swayamdipta, S.; Thomson, S.; Dyer, C.; and Smith, N. A. 2017. Frame-Semantic Parsing with Softmax-Margin Segmental RNNs and a Syntactic Scaffold. *arXiv preprint arXiv:1706.09528* .

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2014. Show and Tell: A Neural Image Caption Generator. *CoRR* abs/1411.4555. URL http://arxiv.org/abs/1411.4555.

Wang, Y.; Liu, C.; Huang, M.; and Nie, L. 2018. Learning to Ask Questions in Open-domain Conversational Systems with Typed Decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2193–2203. Melbourne, Australia: Association for Computational Linguistics. doi:10.18653/v1/P18-1204. URL https://www.aclweb.org/anthology/P18-1204.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A. C.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *CoRR* abs/1502.03044. URL http://arxiv.org/abs/1502.03044.

Zhou, Q.; Yang, N.; Wei, F.; Tan, C.; Bao, H.; and Zhou, M. 2017. Neural Question Generation from Text: A Preliminary Study. *CoRR* abs/1704.01792. URL http://arxiv.org/abs/1704.01792.