GENERATING NATURAL QUESTIONS FROM IMAGES FOR MULTIMODAL ASSISTANTS

Alkesh Patel, Akanksha Bindal, Hadas Kotek, Christopher Klein, Jason Williams
Apple, Cupertino, CA, USA

ABSTRACT

Generating natural, diverse, and meaningful questions from images is an essential task for multimodal assistants as it confirms whether they have understood the object and scene in the images properly. The research in visual question answering (VQA) and visual question generation (VQG) is a great step. However, this research does not capture questions that a visually-abled person would ask multimodal assistants. Recently published datasets such as KB-VQA, FVQA, and OK-VQA try to collect questions that look for external knowledge which makes them appropriate for multimodal assistants. However, they still contain many obvious and common-sense questions that humans would not usually ask a digital assistant. In this paper, we provide a new benchmark dataset that contains questions generated by human annotators keeping in mind what they would ask multimodal digital assistants. Large scale annotations for several hundred thousand images are expensive and time-consuming, so we also present an effective way of automatically generating questions from unseen images. In this paper, we present an approach for generating diverse and meaningful questions that consider image content and metadata of image (e.g., location, associated keyword). We evaluate our approach using standard evaluation metrics such as BLEU, METEOR, ROUGE, and CIDEr to show the relevance of generated questions with human-provided questions. We also measure the diversity of generated questions using generative strength and inventiveness metrics. We report new state-of-the-art results on the public and our datasets.

Index Terms— Multimodal assistant, computer vision, visual question generation, long-short-term memory

1. INTRODUCTION

Current voice-based digital assistants do not take visual input into account, so there is no easy way to understand what users would ask if these assistants were able to see the real world. One way to mimic the real scenarios is to show human annotators real-world images and ask them to note natural questions they would ask a multimodal digital assistant.

The field of Visual Question Answering (VQA) has made incredible strides in recent years, including a large number of standard VQA datasets [1][8][11]. However, current VQA datasets [12][13][14] are focused on recognition, and most questions are about simple counting, colors, and other visual detection tasks, so these datasets do not require much association with external knowledge. The most challenging and exciting questions people ask digital assistants generally require knowing more than what the question entails or what information is contained in the images.

Imagine a user comes across a flower shown in Fig. 1 for the first time. The information in the image does not say anything about where this kind of flower is grown, what time of the year they bloom, or where they can be bought. Thus, the image is not complete for answering these kinds of questions. To answer the questions listed in Fig. 1 we need to link the image content to external knowledge sources, such as the excerpts taken from Wikipedia, or a summary snippet from a relevant web search result.

More recent research tries to address the shortcomings of existing VQA datasets by incorporating structured knowledge bases [15][16][2][17] into VQA datasets. The OK-VQA dataset [15] goes a step further and also includes questions that need to reason over unstructured knowledge. In all of these datasets, questions are designed so that the answer cannot be obtained only by looking at the image. We carefully examined the OK-VQA dataset as it is close to our work. We make two main observations about its incompatibility for multimodal assistants: i) the image types in OK-VQA datasets are often not appropriate for inspiring any meaningful questions for the digital assistant, ii) the OK-VQA dataset has many obvious or common-sense questions for its images, as shown in Fig. 2 which are not challenging enough to ask a digital assistant. This paper introduces an effective dataset of natural questions that are more suitable for multimodal assistant use cases.

Once we have a relevant dataset, we build the image captioning style model to generate questions from the given image. Learning to ask meaningful questions by looking at an image is an important task in NLP and vision as it demonstrates the capabilities of machine to understand the scene. Such ability can be an integral component of any digital assistant, either to engage the user to proactively start a conversation, elicit task-specific information, or suggest the question when the user cannot formulate his/her needs in natural language.
Fig. 2. Some selected questions from the images provided in OK-VQA. The questions shown along with images may not require help from digital assistant for a visually abled person as the answers seem obvious.

| Datasets          | # Questions | # Images | Image source  | Goal                          | Avg. Qn. length |
|-------------------|-------------|----------|---------------|-------------------------------|-----------------|
| KB-VQA            | 2402        | 700      | COCO          | Visual reasoning with given KB| 6.8             |
| FVQA              | 926         | 2390     | COCO + ImageNet| Visual reasoning with given KB| 9.3             |
| OK-VQA            | 14055       | 14031    | COCO          | Visual reasoning with open knowledge | 8.1             |
| Ours (VQG-Apple)  | 132214      | 12006    | Flickr        | Visual reasoning with open knowledge | 7.7             |

Table 1. Comparison of various visual QA datasets

We found no previous approaches providing insight into how different image and text encoding schemes can impact the quality of generated questions for a given image. We experiment with various image/text encoding schemes and decoding schemes that generate meaningful and diverse questions from the given image and associated metadata. We compare our results with previous state-of-the-art techniques on standard VQG-COCO and VQG-Flickr datasets and also provide the benchmark results on our dataset.

Our contributions in this paper are multi-fold: i) we introduce visual questions dataset that is relevant for multimodal digital assistants, ii) provide a comparison of various image and text encoding schemes and their impact on question generation, iii) build a state-of-the-art natural question generation model that takes an image and its metadata and outputs diverse and meaningful questions.

2. DATA COLLECTION METHODOLOGY

The image source of the OK-VQA is the COCO image dataset, which is mainly object-centric. We created a more generic dataset from carefully sampled Flickr images. Collected questions from images; however, their task was to generate a natural question that can potentially engage a human in starting a conversation. We adopt a similar methodology, but our guidelines emphasize asking questions about an image that one might ask a digital assistant. Our prompt shows an image and a corresponding keyword/phrase that potentially describes the image. We collected images related to products, arts, plants, flowers, animals, sculptures, places of interest, foreign text, and scenarios where people frequently seek help from a digital assistant.

Fig. 3. Some examples of the questions that annotators provided from sampled images

from annotators using our guidelines. Table 1 shows the salient characteristics of various knowledge-based VQA datasets and compares them with ours. As one can see, our dataset has a significantly larger number of questions from a diverse set of images from Flickr.

3. MODELS

We use a model architecture similar to image captioning [13]. However, we have additional input for representing image metadata. The image metadata can be in the form of text such as a short caption describing the image, search tags/keywords associated with the image, and location in the image. Fig. 4 shows the higher-level architecture of the overall model. In computer vision literature, several pre-trained CNN models such as VGGNet [19], ResNet [20], MobileNet [8], and DenseNet [21] are popular for encoding an image. We use the image vector derived from the last layer of these CNNs to represent the image. We need to represent two types of text inputs fed into the model: i) image metadata/keywords ii) question words/text fed into encoder/decoder in every timestamp. We can represent text using word embeddings like GloVe [9] or sentence embeddings derived from pre-trained networks such as ELMo [24] or transformer networks like BERT [10]. In the default settings, we use greedy decoding scheme. However, we additionally experiment with simple beam search with beam size of 5 and diverse beam search similar to [22] that promotes diversity in the generated questions without compromising the questions’ naturalness.

In our diverse beam search decoding scheme, we first generate question tokens with a beam size of 5 until a minimum number of timestamps, T (we chose T=3). After timestamp T, we start clustering the questions generated so far across beams using token-level text similarity. We randomly pick the cluster head from the formed clusters and keep generating question tokens. We keep track of questions that have hit the (end) token and add them to the result list.
4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

**VQG dataset:** The Visual Question Generation dataset has images from MSCOCO, Flickr, and Bing. Each of these sets contains roughly 5000 images and five questions per image. It contains natural and engaging questions (not necessarily related to multimodal assistant use cases) about images. We noticed that in the Bing search images’ set, most of the image links were broken. So, we do not consider them in our experiments.

**VQG-Apple-Flickr dataset:** We created this dataset using the data collection methodology described in Section 2. Due to customized image selection and annotation guidelines, our images as well as corresponding questions are more tailored to multimodal assistant use cases. There were about 12K images in our dataset. Each image has an average of nine questions, which amounts to more than 100k questions. We split the dataset into training (60%), validation (20%), and test (20%) sets based on the number of images.

Since experimenting with an exhaustive list of image representation strategies is beyond this paper’s scope, we used VGG19, MobileNet(v2), and DenseNet, which are pre-trained on the Imagenet dataset for our experiments to see their impact on generated questions. For representing text, we picked one static word embedding-based representation i.e. GloVe, and one contextual sequence embedding-based representation i.e. BERT. We first evaluate our proposed approach quantitatively on the standard automatic metrics i.e., BLEU, METEOR, ROUGE, and CIDEr. Since none of these metrics truly represents the diversity of the generated questions as they are designed to check whether the generated questions are similar to the ground truth questions or not, we used the generative strength and inventiveness metrics described in Table 4 to measure the naturalness and diversity of the generated questions. Generative strength tells us how many unique questions are generated by the model per image, while inventiveness indicates, out of all the generated questions, how many are new, i.e., never seen in the training data.

4.2. Quantitative Results

Table 2 shows the results of the experiments using the automatic metrics discussed previously. We observe that in terms of the image encoding schemes, DenseNet performs better on VQG-COCO dataset while VGG19 and MobileNet perform better on Flickr images. Further analysis reveals that the VQG-COCO dataset has images with objects, while the Flickr dataset has generic images that contain natural indoor/outdoor scenes. This observation suggests that choice of image representation matters and if we know beforehand that there are recognizable objects in the image, we can choose one type of image representation over the others to get better results. All pre-trained image representations we used in our experiments were using Imagenet weights i.e., trained on the same Imagenet training data. However, the underlying architecture of the network does seem to impact the outcome of the generated questions.

We notice that there is no clear pattern for the effect of text representation on the automatic evaluation metrics. However, in general, GloVe embeddings worked better across the datasets. We further notice that BERT based text encoding tends to generate more grammatical and long questions. However, they are often not grounded or relevant to the image. The relatively low BLEU, METEOR, ROUGE, and CIDEr scores across all datasets for the BERT-based text encoding scheme confirms this phenomenon. For our VQG-Apple-Flickr dataset, we also notice that adding keywords either gave the same results or slightly improved them.

To generate more diverse questions, we experimented with a simple beam search (beam size = 5) and diverse beam search (DBS) on VQG-COCO and VQG-Apple-Flicker datasets. Experiments were fixed to Greedy for these experiments.

**Table 2.** Comparison of various image encoding and text encoding schemes on VQG-COCO, VQG-Flick, and VQG-Apple-Flicker datasets. The decoding strategy was fixed to Greedy for these experiments.

| Experiments | BLEU | METEOR | ROUGE | CIDEr |
|-------------|------|--------|-------|-------|
| **VQG-COCO dataset** | | | | |
| VGG19 + GloVe | 35.9 | 20.4 | 46.9 | 22.9 |
| MobileNet + GloVe | 31.5 | 19.9 | 44.4 | 22.6 |
| DenseNet + GloVe | 32.6 | 20.1 | 45.5 | 21.6 |
| VGG19 + BERT | 32.8 | 21.0 | 46.7 | 22.1 |
| MobileNet + BERT | 33.2 | 20.6 | 47.3 | 21.8 |
| DenseNet + BERT | 30.1 | 19.6 | 44.5 | 24.1 |
| VGG19 + GloVe + Keywords | 33.9 | 20.4 | 46.9 | 22.9 |
| MobileNet + GloVe + Keywords | 31.6 | 19.9 | 44.5 | 22.7 |
| DenseNet + GloVe + Keywords | 32.9 | 20.4 | 45.0 | 29.0 |

| Experiments | BLEU | METEOR | ROUGE | CIDEr |
|-------------|------|--------|-------|-------|
| **VQG-Apple-Flicker dataset** | | | | |
| VGG19 + GloVe + GloVe | 31.0 | 19.2 | 41.8 | 40.1 | 0.7 | 54.7 |
| DenseNet + GloVe + BS | 37.8 | 22.3 | 47.5 | 53.7 | 3.9 | 44.4 |
| DenseNet + GloVe + DBS | 36.8 | 21.7 | 46.4 | 50.4 | 5.3 | 57.5 |
| VGG19 + GloVe + GloVe | 33.9 | 20.4 | 46.9 | 23.0 | 0.6 | 33.9 |
|MobileNet + GloVe + BS | 43.1 | 23.6 | 55.2 | 40.1 | 2.3 | 24.9 |
|VGG19 + GloVe + GloVe + DBS | 41.6 | 23.4 | 53.9 | 39.3 | 3.2 | 35.4 |

**Table 3.** Comparison of various decoding schemes such as greedy(GD), simple beam search(BS) (k=5) and diverse beam search(DBS) on VQG-COCO and VQG-Apple-Flicker datasets.

4.3. Qualitative Results

We perform a qualitative analysis of generated questions in randomly selected images. Table 4 shows the ground truth questions from human annotators and questions generated from various strategies discussed in this paper. If we compare the generated questions in second row of Table 4, we notice that VGG19 & MobileNet generate better questions, such as “what is the temperature of this place,” “when was it built,” “what body of water is that,” “what kind of food is that,” “what kind of design is this,” “how old is this castle,” which
Diverse beam search generates more and innovative questions such as “what’s the name of that body,” “how much does it cost to make,” “how old is the castle building.” However, as mentioned previously, BERT-based questions were grammatically correct and relatively long.

We show the top-4 questions generated by the corresponding decoding schemes (MobileNet image encoding is picked just for relative comparison) in fourth row. A simple beam search (k=5) and diverse beam search can generate many new and diverse questions that are not only different from human-annotated ground truth questions but also not seen in the entire training set (as depicted in blue and red color). Blue-colored questions indicate that newly generated questions do not make sense for the image or not acceptable for digital assistants. Red-colored questions indicate that newly generated questions do not make sense for the image or not acceptable for digital assistants. Diverse beam search generates more and innovative questions such as “what’s the name of that body,” “how much does it cost to make,” and “how old is the castle building.” However, it also generates some ungrammatical quality questions (shown in red). Instead, the beam search generates less but relatively good quality questions with some drop in generative strength and inventiveness scores.
6. REFERENCES

[1] Jain, U., Zhang, Z., & Schwing, A. (2017). Creativity: Generating Diverse Questions using Variational Autoencoders. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. https://doi.org/10.1109/CVPR.2017.575

[2] Wang, P., Wu, Q., Shen, C., Dick, A., & Van Den Hengel, A. (2017). KB-VQA: Explicit knowledge-based reasoning for visual question answering. IJCAI International Joint Conference on Artificial Intelligence, 1290–1296. https://doi.org/10.24963/ijcai.2017/179

[3] Wang, P., Wu, Q., Shen, C., Hengel, A. van den, & Dick, A. (2016). FVQA: Fact-based Visual Question Answering. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2017.2754246

[4] Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. Retrieved from http://arxiv.org/abs/1906.00067

[5] Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., & Vanderwende, L. (2016). Generating natural questions about an image. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, 3, 1802–1813. https://doi.org/10.18653/v1/p16-1170

[6] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

[7] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.

[8] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Adam, H. (2017). MobileNets: Efficient convolutional neural networks for mobile vision applications ArXiv Preprint [arXiv:1704.04861] https://doi.org/arXiv:1704.04861

[9] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. https://doi.org/10.3115/v1/d14-1162

[10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [arXiv:1810.04805] [Cs]. https://doi.org/arXiv:1811.03600v2

[11] Zhang, S., Qu, L., You, S., Yang, Z., & Zhang, J. (2016). Automatic Generation of Grounded Visual Questions. IJCAI International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2017/592

[12] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are you talking to a machine? Dataset and methods for multilingual image question answering. Advances in Neural Information Processing Systems.

[13] Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. Proceedings of the IEEE International Conference on Computer Vision. https://doi.org/10.1109/ICCV.2015.9

[14] Ren, M., Kiros, R., & Zemel, R. (2015). Exploring Models and Data for Image Question Answering. Advances in Neural Information Processing Systems. Retrieved from http://arxiv.org/abs/1505.02074

[15] Marino, K., Rastegari, M., Farhadi, A., & Mottaghi, R. (2019). OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. Retrieved from http://arxiv.org/abs/1906.00067

[16] Narasimhan, M., Lazebnik, S., & Schwing, A. G. (2018). Out of the box: Reasoning with graph convolution nets for factual visual question answering. Advances in Neural Information Processing Systems.

[17] Wu, Q., Wang, P., Shen, C., Dick, A., & Van Den Hengel, A. (2016). Ask me anything: Free-form visual question answering based on knowledge from external sources. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2016.500

[18] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2015.7298935

[19] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

[20] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2016.90

[21] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017.

[22] Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., & Batra, D. (2016). Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. 1–16. Retrieved from http://arxiv.org/abs/1610.02424

[23] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Parikh, D., & Batra, D. (2017). VQA: Visual Question Answering: www.visualqa.org. International Journal of Computer Vision. https://doi.org/10.1007/s11263-016-0966-6

[24] Shahbaz, M., Suresh, L., Rexford, J., Feamster, N., Rottenstreich, O., & Hira, M. (2019). Elmo. https://doi.org/10.1145/3341302.3342066

[25] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pre-training Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. (NeurIPS), 1–11. Retrieved from http://arxiv.org/abs/1908.02265

[26] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, J. D. (2020). VL-BERT: Pre-training of Generic Visual-Linguistic Representations. 1–16. Retrieved from http://arxiv.org/abs/1908.06066