Visual Question Generation using Transformer Decoders and Image Pretraining

---

By Jacob Fernandez, Grant Zhao

github/CS4782-FinalProj

1. Introduction

Our project replicates and expands upon the work presented in *Generating Natural Questions About an Image* (Mostafazadeh et al., 2016), which introduced the task of Visual Question Generation (VQG). Previous tasks like Visual Question Answering (VQA) or image captioning do not emphasize common sense or inferential reasoning in the questions the way VQG does. The questions we aim to generate should be natural and engagin questions, not just descriptive inquiries. The paper suggests that questions should both reference objects in a scene and imploke implicit events or states.

In the process, the main obstacles were the datasets at hand. The datasets used in the original paper were image-link datasets that were over 10 years old, in which many of these links were expired. Over half the data was lost from each data source, with the MS COCO dataset losing over 75% of its original size. It was at this point we knew we'd have to get fancy with our approach, particularly data sourcing and model training, if we wanted to replicate the results.

2. Chosen Result

The original paper benchmarked various generative models, including retrieval-based methods, maximum entropy pipelines, and a gated recurrent neural network (GRNN) trained to ask questions. Their evaluation demonstrated that GRNN models most effectively captured human-like question patterns, particularly on their event-centric Bing and Flickr datasets.

We aimed to replicate this result: specifically, the ability of an end-to-end generative model (in our case, a Transformer-based decoder) to generate abstract, context-aware questions about an image. Our central goal was to replicate their qualitative insight: **"models can generate plausible questions, but there's still a gap to human naturalness."**

3. Methodology

Due to dataset failures, we constructed our own large-scale VQG dataset using multiple sources:

- **GQA**: 22M questions over 3M+ real-world images. We parsed and filtered image-question pairs based on image availability and relevance, to create a subset of this dataset, grabbing up to 50k unique images and all questions linked to each image. While its questions are more literal and structured, they offer raw volume for training. We used GQA as a *pretraining source*, mapping images to their associated questions by parsing their val_all_questions.json and aligning them with local image files.

- **MS COCO**: ~1.5k images and 7.5k questions. We extracted images and image IDs from train2014 and val2014, mapping these to questions from complementary VQA-style datasets. Almost 75% of the data was corrupted.

- **Flickr/Bing**: ~2.5k images each and 12.5k questions. Leveraged existing .pt tensor embeddings with matched questions from prior VQG tasks. For both of these data sources, almost half the data was corrupted.

3.1 Data Preprocessing

Before training, the original VQG sources were unified. First, the images were run through a preprocessing pipeline that standardized all tensors to ResNet50 or ViT-B/16 image features. We then use back-translation to augment the text-space by 2.5x. The images were then processed into a centralized directory (resnet_data/), with dimension checking and name reformatting. From there one comprehensive .csv was built, containing image-question pairs and paths to precomputed embeddings for train-test splitting.

3.2 Modeling

Due to concerns of limited data we upped the model architecture to aTransformer-based decoder conditioned on image embeddings. Our architecture accepts 2048-dim (ResNet) or 512-dim (ViT) vectors and uses a vocabulary generated via utils.py based on token frequency. Training used cross-entropy loss on teacher-forced sequences with custom padding/mask handling. Evaluation uses both BLEU/METEOR and qualitative inspection via a generate_question() function that samples real images and visualizes the generated questions.

Automatic Evaluation	$Human_{consensus}$	$Human_{random}$	$Transformer_{gqa.pretrain}$	$GRNN_{all}$
Bing	87.1	83.7	**12.4**	11.1
COCO	86.0	83.5	**13.8**	14.2
Flickr	84.4	83.6	**10.0**	9.9
Bing	62.2	58.8	**16.0**	15.8
COCO	60.8	58.3	**18.2**	18.5
Flickr	59.9	58.6	**14.1**	14.9
Bing	63.38	57.25	**11.5**	10.8
COCO	60.81	56.79	**12.3**	12.46
Flickr	62.37	57.34	**9.4**	9.55

4. Results and Analysis

Our results come with the caveat that we could not recreate the results fully under the conditions outlined in the paper. Aside from this, our model outperformed the original paper in almost every metric, on a test set from every source. We attribute this to the use of the GQA pre-training, as this allowed us to first develop a model that could discernably generate language in the form of questions and from there fine tune a network on our unified data sources. Our findings accurately represent two qualities of the paper, firstly the ability to reasonably generate questions from images, and secondly that the gap between human and visual QG are still large despite the introduction of new NLP techniques. Our findings highlight the need for more reliable data in the field to achieve significant results.

5. Reflections

Throughout this re-implementation effort we gained several important insights. One of them was how to handle data quality challenges; working with the image datasets revealed many issues with corrupted images and we had to work around this problem as it wasn't an original issue of the paper. Furthermore,

while our model produces reasonable questions, achieving the full diversity of human-generated questions remains challenging as the model generates safer, more generic questions. This result made us look further into the unobtainable gap between computer and human creativity, as well as if there are ways to bridge this gap. Finally, we also realized that the evaluation metrics for this paper are very complex and even the ones we used for our results, such as BLEU, cannot fully capture the quality or naturalness of our generated question. Future directions in our work revolve around developing more sophisticated models to increase generation diversity and metrics that better capture question quality. We can also go in a different but related direction of incorporating question generation within a conversational system to create more contextually aware questions.

6. References

Mostafazadeh, N. et al. (2016) Generating natural questions about an image, arXiv.org. Available at: https://arxiv.org/abs/1603.06059 (Accessed: 05 May 2025).

Hudson, D. A., and Manning, C. D. *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. https://cs.stanford.edu/people/dorarad/gqa/. Accessed 5 May 2025.

He, K., et al. *Deep Residual Learning for Image Recognition*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. GitHub Repository: https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py. Accessed 5 May 2025.

Radford, A., et al. *Learning Transferable Visual Models From Natural Language Supervision*. arXiv, 2021. https://github.com/openai/CLIP. Accessed 5 May 2025.

Sennrich, R., Haddow, B., and Birch, A. *Improving Neural Machine Translation Models with Monolingual Data*. arXiv, 2015. https://arxiv.org/abs/1511.06709. Accessed 5 May 2025.

Howard, J., and Ruder, S. *Universal Language Model Fine-tuning for Text Classification*. arXiv, 2018. https://arxiv.org/abs/1801.06146. Accessed 5 May 2025.