Gen-VQA: Generative Visual Question Answering with Abstract Scenes

Yifan Yang
MIT
yifany@mit.edu

Zhen Guo
MIT
zguo0525@mit.edu

Abstract

One of the most challenging multi-discipline/multimodal task is arguably Visual Question Answering (VQA), which combines Computer Vision (CV), Natural Language Processing (NLP) and Knowledge Representation & Reasoning (KR) in one. Completing such task in a single model may further enable the ability of a machine to perform general intelligent action, achieving AI-complete task with a well-defined evaluation metric. However, producing VQA datasets with high-quality, which the existing VQA models heavily rely on, takes a large amount of human effort. As the result, the potential of existing VQA models is severely constrained by the scale of available VQA datasets.

In this paper, we propose Gen-VQA[1]*, a novel generative VQA model, designed from the ground up, based on Text-to-Image generative adversarial networks (GANs) and basic VQA model (LSTM + CNNs). The proposed generative VQA (Gen-VQA) augments the training process by feeding training data as well as synthetic data to the basic VQA module consecutively, aiming to improve its robustness to noises and accuracy to answers. Our experiments show that the Gen-VQA architecture is more efficient than basic VQA during the process of inference, using a resnet18 inspired network, it can achieve comparable accuracy to the basic VQA using a deeper and more complex resnet50.*

1. Introduction

Multi-discipline research which combines Computer Vision (CV), Natural Language Processing (NLP) and Knowledge Representation & Reasoning (KR) could be a big leap towards the next generation AI algorithms. Such belief is supported by the argument that capturing multi-modal knowledge beyond a single sub-domain may further enable the ability of a machine to perform general intelligent action, achieving AI-complete task with a well-defined evaluation metric. To this end, the task of Visual Question Answering (VQA) has been proposed in 2015 [4]. As shown in Figure 1, it takes an image and natural language question as inputs, and produces a natural language answer as the output. However, there are many challenges and difficulties of in solving the VQA task: It requires a model to understand and reason about visual-linguistic concepts, and demands numerous capabilities compared to single-modal model, including object localization, attribute detection, activity classification, scene understanding, reasoning, counting, and etc [24]. More importantly, high-quality labeled VQA datasets are scarce, as it take a large amount of human effort to label them. Although most of the current VQA models have large model capacity, *there is not enough data for training*. For example, the VQAv2 [1] datasets contains 82783 real-world training images for VQA task. However, datasets such as ImageNet [14] have more than 1.2 million training images for classification. Therefore, the input and output of various VQA models generally limit to a small set where the probability distribution of the predicted answers are on a fixed space made by the most common answers of the used datasets [25]. Up to date, a high accurate and efficient VQA model is not yet to be realized [32].

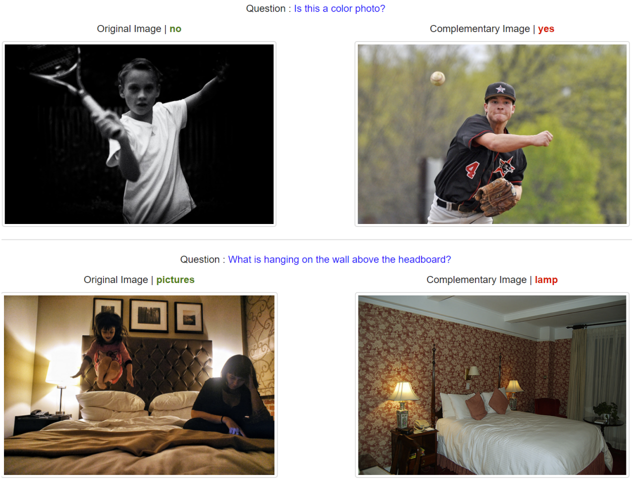
Question : Is this a color photo?
Original Image | no
Complementary Image | yes
Question : What is hanging on the wall above the headboard?
Original Image | pictures
Complementary Image | lamp

Figure 1: Examples of VQA datasets collected via Amazon Mechanical Turk [4].

In this paper, we introduce a generative learning strategy to overcome the issue of data scarcity, which is motivated by the previous success of generative adversarial learning in many computer vision problems [31, 5]. We particularly frame the task as a generative learning process as we si-

[1] https://github.com/Yang-YiFan/vqa-gan

multaneously train the generator and VQA module. In our model, the generative module produces synthetic images based on the question, answer and images from the VQA training datasets, and the VQA module leverages the synthetic image from the generator, the real image from the datasets, as well as training questions to yield an answer.

To simplify our model, we constrain the questions to be multiple-choice which only requires the model to pick from a predefined list of possible answers, rather than giving open-ended and free-form response as it was originally proposed [4]. Furthermore, we constrained our training data within abstract scene of VQA datasets for faster preprocessing and faster training giving the limited computation resources on our hands.

Our main contribution in this work is to develop a generative VQA model that enables compelling and effective question answering with given abstract scenes. The proposed generative VQA (Gen-VQA) is designed from the ground up, based on Text-to-Image generative adversarial networks [20] and basic VQA model [4]. Our Gen-VQA augments the training process by feeding training data as well as synthetic data to the basic VQA module consecutively, and our experiments show that the Gen-VQA architecture is much more efficient than basic VQA, as it has equivalent accuracy using resnet18, compared to the basic VQA model with a much deeper and more complex resnet50.

2. Related Work

2.1. VQA Datasets

VQA is initially proposed as a ”Visual Turing Test” [8], and its format has very soon been accepted as the basis for many datasets and benchmarks since 2015. A number of general datasets, based on MSCOCO images, have been introduced over the past four years, including COCOQA [21], Baidu-FM-IQA [7], VQA [4], Visual7W [34], and etc. Here, we briefly review some of the most popular datasets previously used for training and testing.

VQAv1/VQAv2. VQAv1 [4] is one of the earliest, open-ended form of VQA datasets collected from human labeling, but it has multiple kinds of language biases, including few reasoning questions and more detection questions. Therefore, VQAv2 [10] was proposed to mitigate the biases by collecting complementary images per question that result in different answers. Though the biases are not completely resolved, both dataset is the de facto benchmark for natural image VQA task.

CLEVR [11] is the synthetic generated dataset using ground-truth programs within modular networks for compositional language and elementary visual reasoning. It's similar in spirit to the SHAPES dataset [3], but more complex and varied both in terms of visual content and question variety and complexity. The images have associtaed ground-truth object locations and attributes, and the questions have an associated machine-readable form, testing abilities such as counting, comparing, logical reasoning, and storing information in memory. Later, **CLEVR-Human** [12] and **CLEVR-CoGenT** [22] were created to expand the original datasets.

2.2. VQA Algorithms/Models

We briefly review the progress and recent studies on VQA, paying special attention to the model/architecture development for better answering accuracy.

Multimodal Fusion Model. Many basic VQA models combine CNNs (convolutional neural networks) and Long Short-Term Memory (LSTM) networks to extract the global features/patterns in the image and question, and then fuse the features to output the answer [2, 7, 16]. Some models introduce a more complex model to learn faster and better question's representations with LSTM and Tanh networks [1], or a better multimodal fusion with residual networks [13].

Image Attention VQA. Recently proposed models, therefore, also take into the account of attention mechanisms for text-guided analysis of images, where the attention is learned by using neural networks that predict which regions of the image are useful, only extracting features from those regions, and then performing multimodal feature fusion to obtain the accurate prediction [27, 29, 23, 26, 32].

Co-attention VQA. Beyond understanding the visual contents of the training image, VQA also requires to understand the semantics of the natural language question. Therefore, it is also necessary to learn the textual attention for the question, as well as the visual attention for better accuracy and performance of the VQA task. Co-attention model is then introduced, where the neural networks not only predict which regions of the image, but also which word in the sentence are useful or not for accurate answering [15, 17, 18].

2.3. Generative Adversarial Text to Image Synthesis

In recent years, generic and powerful recurrent neural network (RNN) have been developed to process sequences of inputs and thus to learn text feature representations discriminately [30]. Meanwhile, deep convolutional generative adversarial networks (GANs) have begun to generate highly realistic images that are at least superficially authentic to human observers [6]. GANs consist of a generator G and a discriminator D that compete in a two-player minimax game: The discriminator tries to distinguish real training data from synthetic images, and the generator tries to fool the discriminator. Specifically, D and G play the minimax game in the following manner [9]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \quad (1)$$

It has been proved that this minimax game has a global optimium when $p_g = p_{data}$, and when G and G have enough capacity p_g converges to p_{data}. Novel architecture using deep convolutional generative adversarial network (DCGAN) has been proposed to generate images conditioned on text features [28, 33, 20], which inspired us to use a generative training strategy for the VQA task.

3. Approach

Inspired by the generator-discriminator architecture of GANs, our generative visual question answering (Gen-VQA) is also comprised of two major modules: the generator and the VQA. The architecture of Gen-VQA is shown in Figure 2. The generator takes the embeddings of the question-answer pair, the high level feature of the image, and also random noise as input, and then generates a synthetic image that corresponds to the question-answer pair. The VQA module is trained not only with real image-question pair but also with the synthetic/fake images. This real/synthetic image is fused with the embedding of the question, and fed toward the VQA module to produce an answer label. After training, we turn the generator module off, and then use the VQA module for testing and validation. This generative approach can be seen as data-augmentation that augments the original datasets with question information and noise perturbation. The following sections will be focusing on laying out the details of our architecture.

For the generator, we first use word2vec that takes a large corpus of text in the VQA datasets as input and produces a vector space as word embeddings. Each unique sentence in the corpus is being assigned a corresponding vector in the high dimensional embedding space. Note that the text from question and answer will go through two separate word2vec projections to differentiate their vector space. After the word embeddings for the text in question and answer are produced and concatenated to one vector, we feed this vector into a LSTM block, a fully-connected layer with Tanh activations, as the question-and-answer encoder. With the question-and-answer feature from the encoder, random gaussian noise, and the image features from the resnet18 feature extractor in VQA model, which is pretrained on ImageNet, are concatenated together and sent to the image generator block that we built. Inspire by resnet, the image generator architecture from low-dimensional vector to high resolution images consists of a series of resnet block and upsample layer. De-convolution and regular convolution is used at the first and last layer to expand and shrink the number of channel of the features. The image generator will upsample the input and produce a 224×224 synthetic image as an output. The detailed architecture of our generator is shown in Figure 3.

The generator module's loss function has three terms [20]: the first term is the regular cross entropy loss between the right answer and the answer generated by the synthetic image; the second term is feature matching loss, which measures the distance between the real and generated images' statistics by comparing intermediate layers activations; the third term is L1 pixel-wise distance between the generated and real images, which further constrains the generated image scene to be close to the source image.

For our VQA module, it's similar to the Multimodal Fusion Model (basic VQA model [4]) except we are feeding it with real image and synthetic image consecutively. The image is forwarded to the pretrained resnet18 feature extractor which produces the global features of that image. The embeddings of the question are directly connected to a newly initiated LSTM and Tanh block from the generator to encode the question. This is because the relationship among words may be different in the case of individual question and question-answer pair. Later, the question features and image features are fused by element-wise matrix multiplication, passing through a fully-connected layer, the VQA module then outpus the answer label corresponding to the question based on the image.

The loss for the VQA module is calculated based on the difference between the produced answer (by both the real image and the synthetic image) and the correct answer in the training datasets using cross entropy loss. With such loss calculation, the VQA model constrains the generator to generate images that can help the VQA module in producing correct answer, while the VQA module can be regarded as a basic VQA model except it also takes the synthetic images generated by the generator as data.

4. Experimental Results

4.1. Resnet Inspired Image Generator

Generating high-resolution images of ImageNet size (224x224) is quite challenging and relies on deep convolutional neural networks. However, the deep CNNs are hard to train because of gradient vanishing and exploding. Therefore, we tackle the problem by introducing resnet block into the image generator in replacement of conventional DCGAN architecture to train a deep convolutional generator.

To test that resnet is indeed a working neural network architecture for text-to-image generation, we first borrow an existing pytorch implementation of Text-to-Image GANs [20] and change the generator and discriminator network to a resnet inspired network. The generator network is built upon several resnet BasicBlocks, Upsample, BatchNorm, ReLU, Conv2d, and Tanh layers. While the discriminator network is built like a mirror architecture except it em-

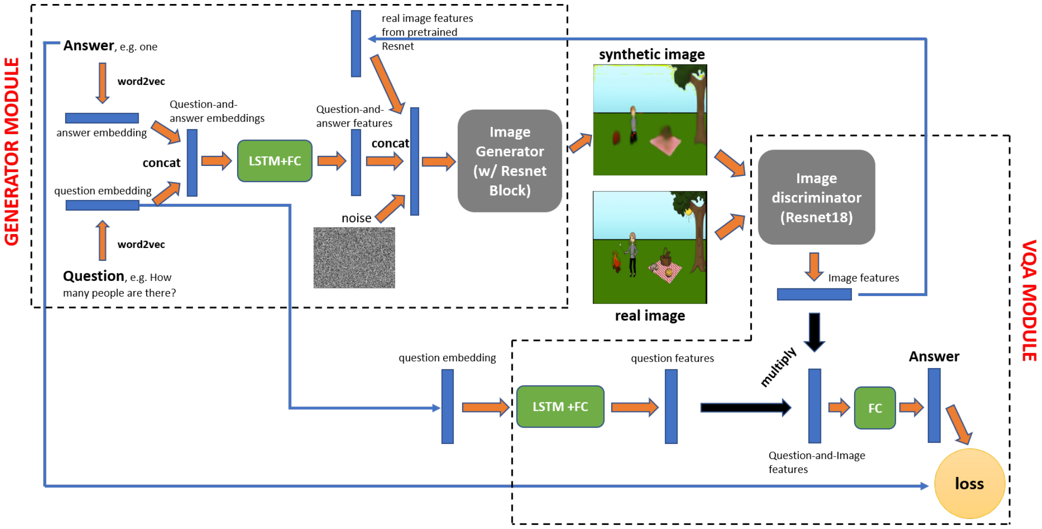
GENERATOR MODULE
Answer, e.g. one
word2vec
answer embedding
Question-and-answer embeddings
concat
question embedding
word2vec
Question, e.g. How many people are there?
LSTM+FC
real image features from pretrained Resnet
Question-and-answer features
concat
noise
Image Generator (w/ Resnet Block)
synthetic image
real image
Image discriminator (Resnet18)
Image features
VQA MODULE
question embedding
LSTM +FC
question features
multiply
FC
Answer
Question-and-Image features
loss

Figure 2: Architecture for Generative Visual Question Answering.

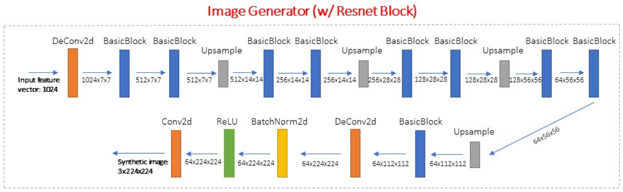

Figure 3: Image Generator Block Details

ploys Downsample instead of Upsample layers in-between BasicBlocks, where the Downsample is the standard stride 2 3x3 convolution in resnet. Using Oxford-102 Flowers datasets [19] and training for 80 epochs, we have the result (in 64x64 size) in Figure 4.

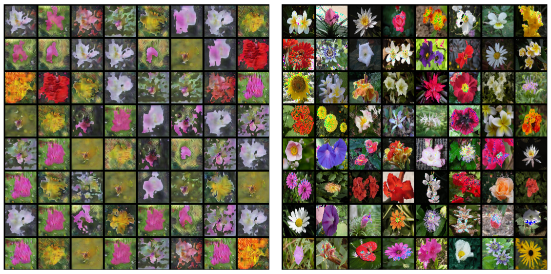

Figure 4: On the left are the synthetic flower images generated by the resnet inspired Text-to-Image GANs, on the right are the real flower images in the datasets

Our resnet inspired Text-to-Image GANs suffers slightly on model collapse, where the generator produces limited modes of synthetic flower images although the inputs are different. The reason of which, we suspect, are threefolded: First, resnet is a more complex and deep network architecture, using small image datasets like Flowers (which is 64×64 pixels for each image) may leads to model overfitting and thus model collapse. Second, the last layer of the discriminator doesn't have convolutions, which forbids the information in different channels to mix. Third, though the generator and discriminator have mirror-like resnet structures, their loss functions are defined differently in the model. The generator has cross entropy loss, feature matching loss, as well as L1 distance loss between real and fake images, while the discriminator only has loss between outputs and labels. The better defined loss function with more strict constraints for the generator leads to its faster training and, therefore, to its domination in the minmax game before our GANs find the global optimal.

In order to solve the model collapse issue, we don't further make the generator model deeper or wider when the generation target is 224x224. In this way the generator is less likely to suffer from overfitting. We also change the learning strategy of generator and discriminator. The learning rate of discriminator is twice as the generator's, making the learning process asymmetric.

4.2. Gen-VQA Optimizations

Now we implement the Gen-VQA model as illustrated in Figure 2 and 3. First, we only use embedding and features of the question-and-answer pairs in the VQA datasets to train the generator module in the Gen-VQA, after 10 epoch we have the result in Figure 5.

We found that all the generated abstract scenes are identical, and we suspect this is because the word2vec networks in our model are randomly initialized, and there is no explicit loss term to capture this deviation. Thus, even with different question-and-answer pairs, word2vec networks will always have similar output features thus identical generated abstract scenes. This suggests that the question-answer constraints are insufficient for the text to image GANs. To overcome such issue, we concatenate not only

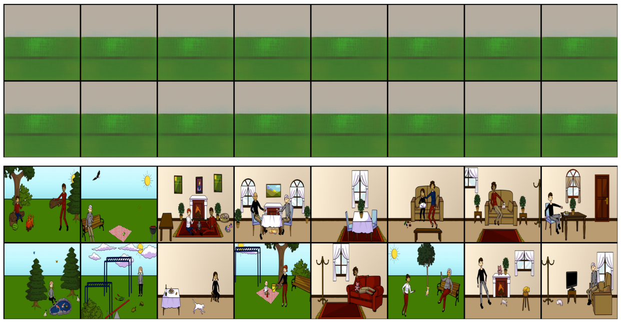

Figure 5: On the top are the synthetic abstract scenes generated by the resnet based generator in Gen-VQA with question-and-answer pair embeddings, on the bottom are the real images in the datasets.

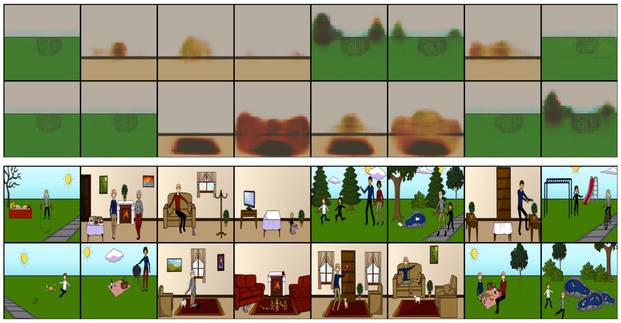

Figure 6: On the top are the synthetic abstract scenes generated by the resnet18 based generator in Gen-GANs, where the question-and-answer pair embeddings as well as real image features are the inputs, on the bottom are the real images in the datasets

the question-and-answer pair's features, but also the real abstract scene features from a pretrained resnet18 (which is always fixed during training). The image features serve as a better initialization point and constraint to the generated image.

With image features, we were able to generate synthetic abstract scene images that are much closer to the input abstract scenes in VQA datasets in Figure 6. However, the upper half of the synthetic images have the same grey-like color regardless of the input image features. We print the RBG values of generated images and real images and find out that the real images have pixel value outside of $(-1, 1)$ range. The Tanh layer in the image generator, however, clips the generated output image to range $(-1, 1)$, which causes the color of the generated image to set in a fix region. Therefore, we decided to remove the Tanh layer in the generator, and indeed, after doing that the upper half images become discriminative with respect to different input images as illustrated in Figure 7.

Another observation from Figure 7 is that since the feature from the resent18 is very high level (features from the last layer of resnet before going through fully-connected layer), the generated synthetic images only contain the low-frequency general structure of the input images, e.g. trees, road, couch, which isn't ideal. The high frequency information, which is often the core of VQA questions, is ignored.

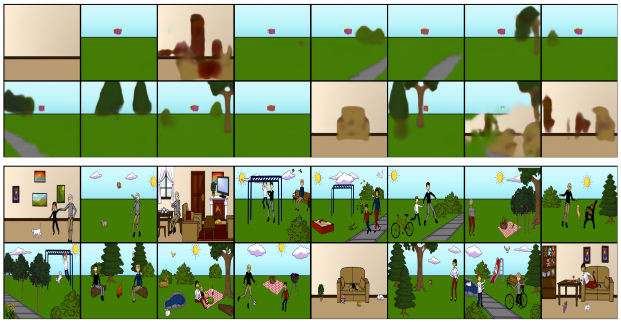

Figure 7: On the top are the synthetic abstract scene VQA images generated without Tanh layer in generator, on the bottom are the real images in the VQA datasets

Therefore, we further concatenate the second last layer features (before the global pooling layer) from the pre-trained resnet18 feature extractor, along with the last layer, question-and-answer pair's features, before send it into the generator module. Our intuition is that the image feature before global average pooling contains more high frequency details and spatial information of objects. By concatenating more low-level, high-frequency features, we expect to be able to produce images with more details. And indeed, as we can see from the Figure 8, the generated images are able to obtain more details, e.g. people, pond, window, of the real images. Besides, the input gaussian noise effectively modify the synthetic images. We can see from the figure that perturbations such as color shift and shape distortion are added to the synthetic image. This significantly augments the original VQA datasets by adding variations to the images and could improve model accuracy.

Figure 9 shows two concrete examples of the visual images and answers. We can see that Gen-VQA correctly renders out the bush and pictures in the image, which helps VQA module to better understand the abstract scene.

4.3. Performance

With the generator network working properly, we train the whole Gen-VQA system together with 30 epochs using Amazon EC2 Nvidia V100 GPU instance. The batch size is set to 64 and each epoch of training takes 15 minutes. We compare our Gen-VQA model (based on resnet18 feature extractor) with the basic VQA models in Figure 10 in terms of training/validation loss and accuracy curve. We train the basic VQA model with resnet18, resnet34 and resnet50 feature extractor with 30 epochs.

GenVQA achieves the validation accuracy of 64.0% on abstract scene dataset. Using the same CNN structure, Gen-VQA significantly outperforms the basic-VQA with resnet18 by 5.1%. More surprisingly, our Gen-VQA with

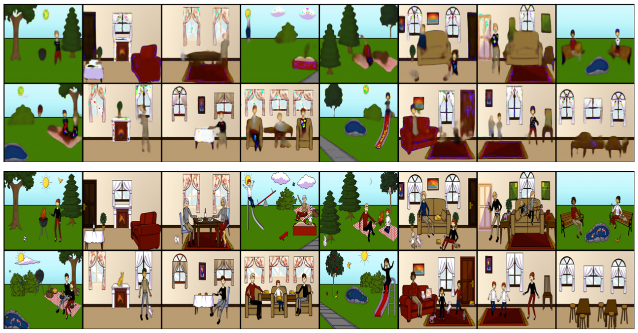

Figure 8: The synthetic (top) and real (bottom) abstract scene VQA images after 30 epochs of training. Compared to Figure 7, all the synthetic images obtains more high-frequency information from the real/input images as well as perturbations such as colors and shapes.

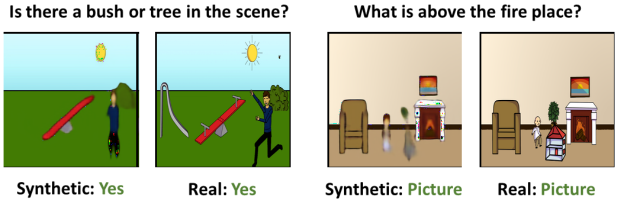
Is there a bush or tree in the scene?
What is above the fire place?
Synthetic: Yes
Real: Yes
Synthetic: Picture
Real: Picture

Figure 9: Success case examples

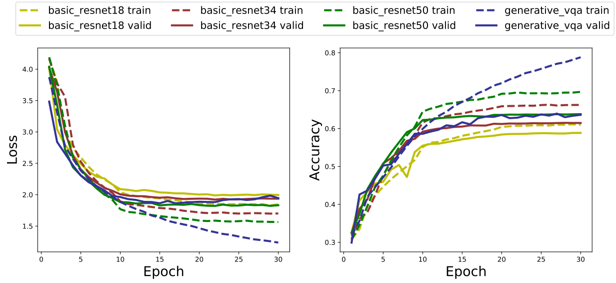

Figure 10: Benchmarking Gen-VQA and basic VQA with different resnet architectures on abstract scene

resnet18 has higher accuracy performance (0.2%) with basic VQA with resnet50 network structure. This shows that our generative approach significantly augments the VQA datasets, which effectively feeds more data to the network.

Figure 11 shows the model complexty (parameter size) and accuracy of the 3 baselines and our Gen-VQA. The pareto optimal curve of basic VQA in terms of accuracy and parameter size is marked by dotted line. Gen-VQA significantly stretch the pareto optimal curve with higher accuracy and lower model complexity. This shows that resnet18 has sufficient capacity to extract features for VQA abstract scene. The scarcity of data limits its potential. With our generative learning/training strategy, a shallow network can perform as well as a much deeper network.

5. Limitations and Future Works

We show two cases where our generative VQA fails in Figure 12. In the first case the generator generates two people and thus misleading the VQA module. This suggests that if the distortion is added to the interested part of the image, the answer could be affected. In the second case, the generator fails to synthesize the football so that the answer is incorrect. This indicates the need to improve the generator to restore more high frequency details.

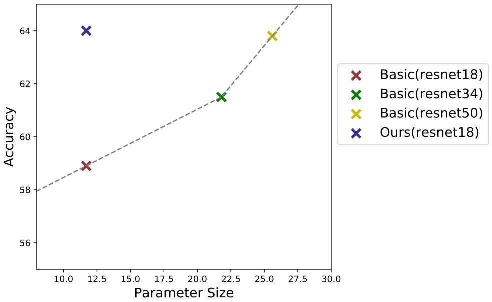

Figure 11: Parameter Size-Accuracy Plot

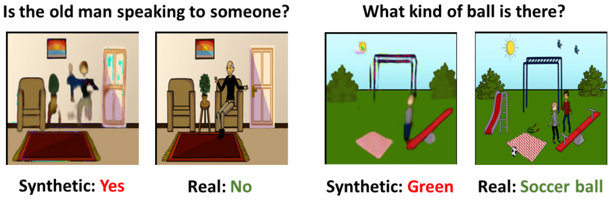
Is the old man speaking to someone?
What kind of ball is there?
Synthetic: Yes
Real: No
Synthetic: Green
Real: Soccer ball

Figure 12: Failure cases

In future work, we aim to further scale up the model to higher resolution real-world images beyond abstract scene and expand the VQA task to include open-ended, free-form questions that are more complex. We also plan to introduce more noises or inputs to the generator to generate more diversified images, such as different weather, number of trees etc. Training the VQA module only on the synthetic images with questions would be an interesting direction to pursue.

6. Conclusions

In this work, we developed a generative and effective VQA model based on Text-to-Image GANs and basic VQA model. We demonstrated that the generator in our model can synthesize many plausible visual images of a given abstract scene sets. Our generator module substantially improved the VQA module on the efficiency of inference in the VQA task, and we showed that our Gen-VQA using resnet18 performs equivalently with a basic VQA using resnet50 (a much deeper and more complex network structure).

Contribution - Zhen Guo

Z.G. and Y.Y. together came up with the idea of working on generative learning/training strategy with VQA task. Z.G. did a comprehensive survey on existing VQA datasets and VQA algorithms/models, re-wrote and tested the basic VQA model on the local host, setup the virtual machines on AWS EC2, proposed ideas during the implementation of the Gen-VQA model, and summarized the results and figures in the logical order.

Contribution - Yifan Yang

Y.Y. and Z.G. together came up with the idea of working on generative learning/training strategy with VQA task. Y.Y. re-designed/re-engineered the generator module using resnet inspired building-block, combined the generator module with the basic VQA module, implemented the Gen-VQA model and optimizations, performed VQA task benchmarking on the AWS virtual machines, and pointed out limitations of our Gen-VQA model.

References

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. Vqa: Visual question answering. *International Journal of Computer Vision*, 123(1):4–31, 2017.

[2] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*, 2016.

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[6] L. Dang, S. Hassan, S. Im, J. Lee, S. Lee, and H. Moon. Deep learning based computer generated face identification using convolutional neural network. *Applied Sciences*, 8(12):2610, 2018.

[7] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015.

[8] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[11] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[12] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2989–2998, 2017.

[13] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *Advances in neural information processing systems*, pages 361–369, 2016.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[15] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[16] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[17] H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307, 2017.

[18] D.-K. Nguyen and T. Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018.

[19] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[20] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[21] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[22] J. Shi, H. Zhang, and J. Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8376–8384, 2019.

[23] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4613–4621, 2016.

[24] R. Shrestha, K. Kafle, and C. Kanan. Answer them all! toward universal visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10472–10481, 2019.

[25] T. Tommasi, A. Mallya, B. Plummer, S. Lazebnik, A. C. Berg, and T. L. Berg. Combining multiple cues for visual madlibs question answering. *International Journal of Computer Vision*, 127(1):38–60, 2019.

[26] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Fvqa: Fact-based visual question answering. *IEEE*

transactions on pattern analysis and machine intelligence, 40(10):2413–2427, 2018.

[27] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[28] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[29] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[30] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

[31] Y. Yu, Z. Gong, P. Zhong, and J. Shan. Unsupervised representation learning with deep convolutional neural network for remote sensing images. In *International Conference on Image and Graphics*, pages 97–108. Springer, 2017.

[32] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

[33] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[34] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.