Exploring Diverse Methods in Visual Question Answering

1st Panfeng Li
Department of Electrical and Computer Engineering
University of Michigan
Ann Arbor, USA
pfli@umich.edu

2nd Qikai Yang
Department of Computer Science
University of Illinois Urbana-Champaign
Urbana, USA
qikaiy2@illinois.edu

3rd Xieming Geng
Department of Computer Science
Chongqing University
Chongqing, China
xieminggeng@foxmail.com

4th Wenjing Zhou
Department of Statistics
University of Michigan
Ann Arbor, USA
wenjzh@umich.edu

5th Zhicheng Ding
Fu Foundation School of Engineering and Applied Science
Columbia University
New York, USA
zhicheng.ding@columbia.edu

6th Yi Nian
Department of Computer Science
University of Chicago
Chicago, USA
nian@uchicago.edu

***Abstract*—This study explores innovative methods for improving Visual Question Answering (VQA) using Generative Adversarial Networks (GANs), autoencoders, and attention mechanisms. Leveraging a balanced VQA dataset, we investigate three distinct strategies. Firstly, GAN-based approaches aim to generate answer embeddings conditioned on image and question inputs, showing potential but struggling with more complex tasks. Secondly, autoencoder-based techniques focus on learning optimal embeddings for questions and images, achieving comparable results with GAN due to better ability on complex questions. Lastly, attention mechanisms, incorporating Multimodal Compact Bilinear pooling (MCB), address language priors and attention modeling, albeit with a complexity-performance tradeoff. This study underscores the challenges and opportunities in VQA and suggests avenues for future research, including alternative GAN formulations and attentional mechanisms.**

***Index Terms*—Visual Question Answering; Generative Adversarial Networks; Autoencoders; Attention**

I. Introduction

The Visual Question Answering (VQA) task [1], represents a pivotal challenge within the field of Artificial Intelligence (AI), striving to bridge the gap between visual perception and natural language understanding. At its core, VQA endeavors to imbue machines with the capacity to comprehend and respond to inquiries pertaining to the content of images, thereby mirroring a fundamental aspect of human cognition. Indeed, the ability to seamlessly integrate visual information with linguistic input is regarded as a hallmark of human intelligence, facilitating the formation of abstract concepts and fostering deeper comprehension of the surrounding environment.

In essence, the VQA task epitomizes the interdisciplinary nature of contemporary AI research [2], drawing upon insights from computer vision [3], natural language processing [4], and machine learning [5] to realize the ambitious goal of imbuing machines with human-like perceptual and cognitive abilities. By synthesizing information gleaned from visual stimuli with linguistic cues, VQA systems aspire to emulate the nuanced understanding and reasoning capabilities exhibited by human agents when confronted with multimodal inputs.

In light of the aforementioned considerations, this study embarks on a journey to unravel the intricacies of the VQA conundrum, leveraging insights from one of the most prominent and widely utilized datasets in the field [6]. By delving into the depths of this seminal dataset, we endeavor to shed light on the underlying challenges and opportunities inherent in the VQA paradigm, with the ultimate aim of advancing the frontier of AI research and fostering the development of more intelligent and perceptive machines.

II. GAN BASED MECHANISM

Our rationale for applying Generative Adversarial Networks stemmed from several considerations:

1) VQA is fundamentally a generative task—given an image and question, we must generate an answer. We thus reasoned that GANs, due to their generative nature, may be a good fit for the problem.
2) Conditional GAN [7] made us reason that, by inputing question and image embeddings as 2-tuples to the generator, we could learn to accurately generate answer embeddings as the outputs. This is because the generator will learn the probability distribution of the answer embeddings conditioned on the question-image embedding 2-tuple. By thinking of the discriminator term as an adaptive loss function, we also reasoned that, with proper training, this may be able to determine a more appropriate loss for the VQA classification problem.

Since the goal of our GAN-based approach was, in part, to simply investigate the application of GANs to classification problems, we focused more on experimenting with and testing a wide variety of possible applications of GANs to this problem than on beating the state-of-the-art results.

A. Architecture

Similar to most existing approaches to the VQA problem, we first passed the image data through a pre-trained ResNet to generate a feature vector for each image. In parallel, we also pre-processed the questions and pass them through an RNN in order to obtain feature vectors for each question. While we do not train the ResNet, we do train the RNN while training the GAN. After generating the feature vectors, we encode them together into a single vector. We attempted two approaches to combining them. Our first approach, motivated by [8, 9], first passed the image and question embeddings through separate $tanh$ layers, then performed element-wise multiplication of these output vectors, and concatenated them with a vector of the element-wise attention of the outputs of the $tanh$ layers. This vector is finally passed through another $tanh$ nonlinearity. Our second approach was much simpler, utilizing separate $tanh$ layers to encode the image and question feature vectors and then combining these embeddings by element-wise multiplication.

Once we had generated the combined embeddings of the features, we fed them into our generator network. These feature vectors can be thought of as the conditional input to the generator, allowing it to generate an output conditioned on the image and question. In addition to feeding in the feature embeddings as conditions for the output, we also inputed random noise into the generator. We tested doing this in three different ways—by concatenating the noise to the feature vector (**N1**), by adding the noise to the feature vector to essentially create a non-zero mean noise input (**N2**), and by feeding in only the features with no noise (**N0**). While the latter approach may not be a true GAN, it can nonetheless be trained as a GAN and gave us a baseline to compare against. We tested with two different architectures for the generator. The first utilized three fully connected layers with ReLu nonlinearities followed by a linear fully connected layer. The second simply utilized a single linear layer. Each of the generators output a vector of length 1000 encoding the likelihood of each of the 1000 most common answers.

We coupled the simple, single layer generator with the simpler of our embedding methods—for future reference we denote this network as $\mathbf{G_{simp}}$—and the more complex generator with our more complex embedding method—which we denote as $\mathbf{G_{full}}$.

We tested with training generators without the discriminator portion and also, in order to produce the full GAN, we fed the generators' outputs into a discriminator network. We attempted several different architectures of the discriminator portion but each was ultimately several fully connected ReLu layers which output a single number into a sigmoid activation to scale it between 0 and 1. In addition to taking the output of the generator as an input, we also fed the features associated with each question and image into the discriminator by concatenating them to the output produced by the generator. We attempted several variations on this, testing with feeding in the raw features produced by the ResNet and RNN and feeding in combined, embedded features. We denote the coupling of the simple generator with a discriminator as $\mathbf{GAN_{simp}}$ and the coupling of the more complex generator with a discriminator as $\mathbf{GAN_{full}}$ A diagram of this architecture is given in Figure 1.

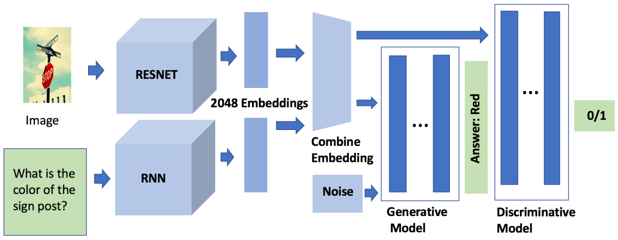


Fig. 1: High Level Architecture of GAN Based System

B. Training

Our basic training algorithm followed the GAN-CLS with step size α, outlined by Reed *et. al.*:

```
Input: minibatch answer embeddings $x$, matching image-question embedding 2-tuple $\psi(t)$, number of training batch step $S$
for $n = 1$ to $S$ do
    $h \leftarrow \psi(t)$ (Encode image-question tuple that matches with the answer)
    $\hat{h} \leftarrow \psi(\hat{t})$ (Encode image-question tuple that doesn't match with the answer)
    $z \sim \mathcal{N}(0,1)^Z$ (Draw random noise)
    $\hat{x} \leftarrow G(z,h)$ (Forward through Generator)
    $s_r \leftarrow D(x,h)$ (correct answer, correct tuple)
    $s_w \leftarrow D(x,\hat{h})$ (correct answer, wrong tuple)
    $s_f \leftarrow D(\hat{x},h)$ (output of Generator, correct tuple)
    $\mathcal{L}_D \leftarrow log(s_r) + \frac{1}{2}(log(1-s_w) + log(1-s_f))$
    $D \leftarrow D - \alpha\frac{\delta\mathcal{L}_D}{\delta D}$ (Update Discriminator)
    $\mathcal{L}_G \leftarrow log(s_f)$
    $G \leftarrow G - \alpha\frac{\delta\mathcal{L}_G}{\delta G}$ (Update Generator)
end for
```

We attempted several variations of this. We experimented with pre-training both the Discriminator and the Generator. When pre-training the Generator we simply trained it as a softmax classifier with noise added to the inputs. This allows us to initialize the weights to values that will produce relatively good results at the start of training the full GAN [10].

We found that pre-training the Discriminator to optimality could be detrimental to the actual training process of the GAN and could worsen the updates of the Generator over time [11]. [11] shows that if the probability densities are either disjoint or lie on a low-dimensional manifold then the Discriminator can distinguish between them perfectly. This happens to be the case for the loss functions proposed by [12]. So instead of pre-training our Discriminator to optimality, we follow the suggestion in [13] and add noise to its inputs.

We tested adding normalization to the outputs of the layers in our generator and discriminator modules. In general, however, we found that this yielded worse results than when using unnormalized layers. We utilized a small amount of dropout in training our generator and discriminator modules.

We attempted initializing the weights under several different distributions, noticing altered results when we did so. We first initialized all weights to follow a Gaussian distribution with large values clipped (we denote this initialization method **I1**) and also tested with initializing weights to follow a uniform distribution (which we denote **I2**).

III. AutoEncoder Based Mechanism

We modified the initial GAN technique to give us an Autoencoder based technique, wherein the concatenated features are passed through an autoencoder to generate low dimensional embeddings. Most existing approaches (such as MCB [14]) utilize a fixed method to embed the question and image features together. By employing an autoencoder, we hoped to learn how to best embed the question and image features into a low-dimensional space. We use this encoding, after passing it through several fully connected layers, to generate the answer.

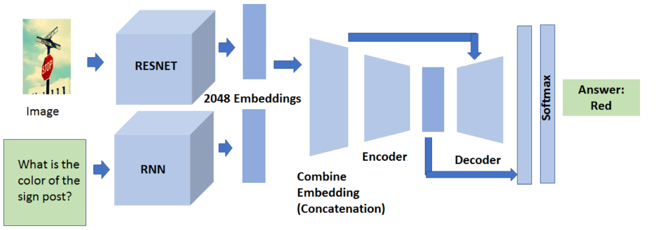


Image
RESNET
2048 Embeddings
What is the color of the sign post?
RNN
Combine Embedding (Concatenation)
Encoder
Decoder
Softmax
Answer: Red

Fig. 2: High Level Architecture of AutoEncoder Based System

IV. Attention Based Mechanism

To answer a question according to an image, it is critical to model both "where to look" and model "what words to listen to", namely visual attention and question attention [15].

However, [6] shows that the language priors make the VQA dataset [1] unbalanced, whereas simply answering "tennis" and "2" will achieve 41% and 39% accuracy for the two types of questions— "What sport is" and "How many". These language priors bring to light the question of whether machines truly understand the questions and images or if they tend to give an answer which has higher frequency in the dataset.

Inspired by the strength of Multimodal Compact Bilinear Pooling (MCB) at efficiently and expressively combining multimodal features [14], we use the MCB operation to replace the simple addition operation used in co-attention mechanism [15] when combining the features learned from the images and questions together, which may help to learn more information from visual part [16, 17].

Method	All	Yes/No	Number	Other
Baseline Methods				
$\mathbf{G_{simp} - N0 - I1}$	11.58	23.86	5.32	1.48
$\mathbf{G_{full} - N0 - I1}$	18.65	40.56	0.25	7.84
$\mathbf{G_{full} - N1 - I1}$	23.41	55.46	3.49	0.61
$\mathbf{G_{full} - N2 - I1}$	14.76	35.77	1.25	0.27
$\mathbf{G_{full} - N2 - I2}$	23.06	54.65	3.50	0.51
Novel Methods				
$\mathbf{GAN_{simp} - N0 - I1}$	25.27	62.49	0.32	0.59
$\mathbf{GAN_{simp} - N0 - I2}$	25.57	56.67	9.05	0.61
$\mathbf{GAN_{full} - N0 - I1}$	27.51	51.90	22.41	0.08
$\mathbf{GAN_{full} - N1 - I1}$	28.81	57.28	19.13	0.54
$\mathbf{GAN_{full} - N2 - I1}$	34.57	65.38	27.36	0.71
Autoencoder	37.65	64.01	24.37	15.77
Attention	44.32	66.64	**32.15**	26.72
Attention + MCB	**47.58**	**67.60**	31.47	**36.98**

TABLE I: Results on VQA 1.9 Validation Dataset. Legend: $\mathbf{G_{simp}}$ - simple, single-layer generator trained as classifier; $\mathbf{G_{full}}$ - full, multi-layer generator trained as classifier; $\mathbf{GAN_{simp}}$ - simple, single-layer generator trained with discriminator; $\mathbf{GAN_{full}}$ - full, multi-layer generator trained with discriminator; **N1** - noise concatenated to generator conditioning input; **N2** - noise added to generator condition input; **N0** - no noise inputed to generator; **I1** - weights initialized via Gaussian distribution; **I2** - weights initialized via uniform distribution

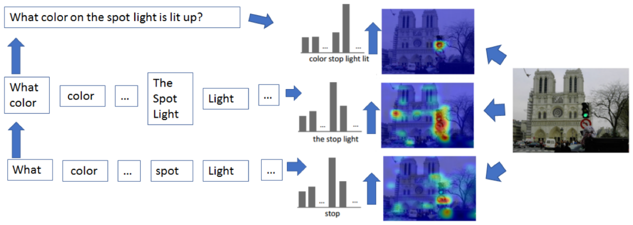
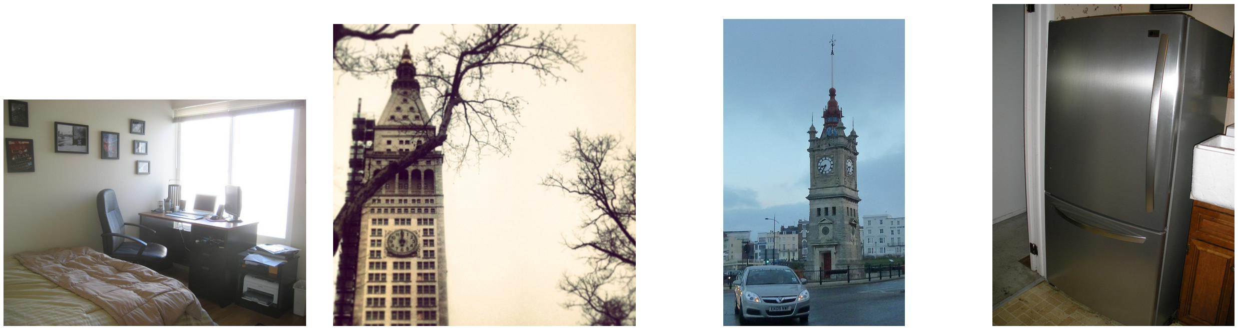


Fig. 3: High Level Architecture of Attention Based System

V. Results

Table I gives the numerical results obtained by every method we tested. The metric used is the one presented in [1] where an answer is considered correct and given a score of one if at least three of the ten human given responses match that answer. As this table illustrates, the baseline methods in general do worse than all novel approaches we attempted.

Figure 4 illustrates the qualitative performance of one of our baseline approaches, one of the GAN-based approaches, and the attention-based model on several images. While the baseline model is able to correctly answer the simpler questions, it fails miserably at more complex questions such as (c) and (d). For (d) it seems to capture some of the meaning of the question (relating a fridge to a meal) yet still incorrectly answers the prompt. As for the GAN-based approach and the attention-based approach, both are able to correctly capture the meaning of the questions while the attention-based approach is slightly better than the GAN-based approach - in (d) our attention-based approach gets the correct answer while our GAN-based approach returns an almost correct answer.



(a) Question: How many chairs are in the photo? Baseline Answer: 3. GAN Answer: 1. Attention Answer: 1

(b) Question: Is it an overcast day? Baseline Answer: yes. GAN Answer: yes. Attention Answer: yes

(c) Question: What year is the car? Baseline Answer: scarf. GAN Answer: 2010. Attention Answer: 2010

(d) Question: What color is the fridge? Baseline Answer: dinner. GAN Answer: gray. Attention Answer: silver

Fig. 4: Qualitative Results of Visual Question Answering

Method	All	Yes/No	Number	Other
Neither Pretrained	0.10	0.08	0.19	0.10
G Pretrained	**21.25**	**54.65**	**3.50**	**0.51**
D Pretrained	0.05	0.00	0.09	0.08
Both Pretrained	13.76	35.47	2.16	0.30

TABLE II: Varying Pretraining on GAN with $\mathbf{N2}$ (noise added) on VQA 1.9 Validation Dataset

Table II illustrates the effect various pretraining combinations have on the results produced by the GAN. There results are given on the GAN utilizing the full, multi-layer generator model $\mathbf{GAN_{full}}$ with noise added to the conditioning inputs $\mathbf{N2}$ and weights initialized according to a Gaussian distribution $\mathbf{I1}$. From this it is clear that the optimal results are obtained when the generator is pretrained but the discriminator is not. When we do not pretrain the generator at all, we found that, given the amount of time for which we trained, the generator is unable to learn the distribution of the data at all. In contrast, discriminator pretraining alone doesn't enhance performance.

VI. Conclusion

This research has embarked on a comprehensive exploration of advanced techniques in VQA, using GANs, attention mechanisms, and autoencoders to enhance the performance of VQA systems. Based on the results from the balanced VQA dataset, we can draw several important conclusions about the effectiveness of these methods in improving VQA systems.

Firstly, our experiments with GANs, especially those with full, multi-layer generator models (GAN_{full}), showed a marked improvement over baseline methods when pretraining is applied selectively—specifically, pretraining the generator but not the discriminator. This approach allowed the GANs to better capture and reproduce the data distribution needed for accurate VQA, although it still presents challenges in task specificity and learning the correct answers to more complex questions, aligned with previous studies [18].

Secondly, our investigation into autoencoder-based techniques revealed that while they are effective at learning optimal embeddings for the question and image data, they achieved slightly better results compared to the GAN-based approach. This is largely due to the much better results on more complex questions. On the contrast, our optimized GAN_{full} method yields better on Yes/No and Number questions.

Lastly, the attention mechanisms, particularly those employing MCB, have demonstrated a substantial benefit in addressing the inherent language priors and improving the modeling of attention over both the textual and visual inputs. This method outperformed both the GAN-based and autoencoder-based approaches in complex question answering scenarios, as seen in the qualitative performance comparisons. However, this comes with the cost of increased computational complexity, which poses a trade-off that future research will need to address.

Overall, the research indicates that while the innovative approaches explored—GANs, autoencoders, and attention mechanisms—show promise, they each come with unique challenges that need to be overcome. Future research should focus on refining these methods, perhaps through alternative GAN formulations and enhanced stability techniques for more complex tasks. Additionally, further exploration into hybrid models combining these techniques could yield improvements in both performance and efficiency for VQA systems. This study thus not only advances our understanding of complex VQA systems but also outlines clear paths for future developments in the field.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433, 2015.

[2] Z. Li, P. Zhou, Y. Zhang, and L. Gao, "Joint coverage and power control in highly dynamic and massive uav networks: An aggregative game-theoretic learning approach," *ArXiv*, vol. abs/1907.08363, 2019.

[3] A. Zhu, J. Li, and C. Lu, "Pseudo view representation learning for monocular rgb-d human pose and shape estimation," *IEEE Signal Processing Letters*, vol. 29, pp. 712–716, 2021.

[4] H. Chen, W. Huang, Y. Ni, S. Yun, F. Wen, H. Latapie, and M. Imani, “Taskclip: Extend large vision-language model for task oriented object detection,” *CoRR*, 2024.

[5] C. Wang, Y. Yang, R. Li, D. Sun, R. Cai, Y. Zhang, C. Fu, and L. Floyd, “Adapting llms for efficient context processing through soft prompt compression,” *arXiv preprint arXiv:2404.04997*, 2024.

[6] 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,” *arXiv preprint arXiv:1612.00837*, 2016.

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, “Generative adversarial text to image synthesis,” *Proceedings of The 33rd International Conference on Machine Learning*, vol. 3, May 2016.

[8] K. Saito, A. Shin, Y. Ushiku, and T. Harada, “Dualnet: Domain-invariant network for visual question answering,” *arXiv preprint arXiv:1606.06108*, 2016.

[9] L. Lin, W. Wu, Z. Shangguan, S. Wshah, R. Elmoudi, and B. Xu, “Hpt-rl: Calibrating power system models based on hierarchical parameter tuning and reinforcement learning,” in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020, pp. 1231–1237.

[10] T. Deng, G. Shen, T. Qin, J. Wang, W. Zhao, J. Wang, D. Wang, and W. Chen, “Plgslam: Progressive neural scene represenation with local to global bundle adjustment,” *arXiv preprint arXiv:2312.09866*, 2023.

[11] M. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” *NIPS 2016 Workshop on Adversarial Training. In review for ICLR*, vol. 2016, 2017.

[12] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.

[13] J. Yao, C. Li, K. Sun, Y. Cai, H. Li, W. Ouyang, and H. Li, “Ndc-scene: Boost monocular 3d semantic scene completion in normalized devicecoordinates space,” in *ICCV*, 2023, pp. 9421–9431.

[14] A. Fukui, D. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, “Multimodal compact bilinear pooling for visual question answering and visual grounding,” *arXiv preprint arXiv:1606.01847*, 2016.

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

[16] W. Ding, S. Li, G. Zhang, X. Lei, and H. Qian, “Vehicle pose and shape estimation through multiple monocular vision,” in *IEEE International Conference on Robotics and Biomimetics*, 2018, pp. 709–715.

[17] H. A. Arief, M. Arief, G. Zhang, Z. Liu, M. Bhat, U. G. Indahl, H. Tveite, and D. Zhao, “Sane: Smart annotation and evaluation tools for point cloud data,” *IEEE Access*, vol. 8, pp. 131 848–131 858, 2020.

[18] X. Song, D. Wu, B. Zhang, Z. Peng, B. Dang, F. Pan, and Z. Wu, “Zeroprompt: Streaming acoustic encoders are zero-shot masked lms,” in *Proc. INTERSPEECH 2023*, 2023, pp. 1648–1652.