eaVQA: An Experimental Analysis on Visual Question Answering Models

Souvik Chowdhury[1] and **Badal Soni**[1]
[1]National Institute of Technology, Silchar
souvik21_rs@cse.nits.ac.in, badal@cse.nits.ac.in

Abstract

Visual Question Answering (VQA) has recently become a popular research area. VQA problem lies in the boundary of Computer Vision and Natural Language Processing research domains.Various details about each dataset are given in this paper, which can help future researchers to a great extent. In this paper, we discussed and compared the experimental performance of the Stacked Attention Network Model (SANM) and bidirectional Long Short Term Memory (LSTM) and Multimodal Tucker Fusion (MUTAN) based fusion models. As per the experimental results, MUTAN accuracy and loss are 29% and 3.5, respectively. SANM model is giving 55% accuracy and a loss of 2.2, whereas the VQA model is giving 59% accuracy and 1.9 loss.

1 Introduction

The visual Question Answer model should have the ability to understand both visual and linguistic capabilities. These models must possess some capabilities to function. They are locating an object (finding the location of any object within any given image mostly based on coordinate position along with height and width), finding object attributes (retrieval of attributes of any given object, e.g., color, shape, or any other traits of an object), activity being performed by an object (e.g., sport being played, running, walking, etc.), understanding of any given scene which is basically providing a high-level representation of the environment. The VQA models are prone to bias to remove this problem; multiple answers are normally being provided for any question.

This paper has been structured in the following way. Section 2 will give related existing work or literature study. This will be followed by Section 3, which will give a study of existing datasets available which is relevant for this analysis along with dataset details like licensing, year of formation, and other statistics about image, question, answer, etc. Section 4 will give details about the experiment being conducted along with an analysis of the results. Section 5 has conclusion details.

2 Background Study

Assessing the performance of the VQA model is a bit tricky. Regular ML models metrics like Accuracy (for classification problems), MAE (for regression problems), etc., cannot estimate performance properly. To effectively quantify the performance of a VQA model, the metric (Shrestha, R et al., 2017) must have the following capabilities:

a. Consistency: This depicts the ability to provide a consistent answer for a different form of the same question.

b. Grounding/Localization: The main purpose is better model interpretability. This is the ability to localize the region in the image which is relevant with respect to the answer.

c. Plausibility: Providing a justifiable answer, e.g., for a question like, is it a sunny day? The answer will be either yes, no, or cannot say something of this like.

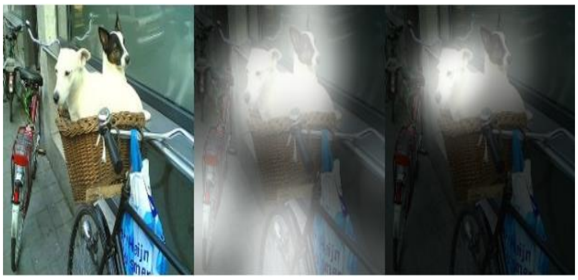
VQA models generate an attention map to perform object localization with respect to the question being asked. The attention maps are nothing but a matrix that identifies a region within an image that is relevant for an image-question pair.A ranking score is generated based on a number of overlaps of bounding regions (Shalini Ghosh et al., 2019).

In a simple attention map mechanism mix of images and questions can derive whether a region is relevant or not with respect to that image. They propose a new architecture that represents the

Proceedings of the 18th International Conference on Natural Language Processing, pages 550–554
Silchar, India. December 16 - 19, 2021. ©2021 NLP Association of India (NLPAI)

image and question mix as a vector representation rather than pairwise confluence. This kind of vector representation can give accurate information about a region that is relevant for a specific context hence making this a context-aware region search (Xi, Y. et al., 2020). The authors proposed a very basic approach towards Visual Question Answer methods (Antol, S. et al., 2015). In this paper, the authors have shown the importance of captioning.

In a paper (Yang, Z. et al.,2016), the authors have adopted a stacked attention mechanism because the reasoning is complex; hence if multiple attention map mechanisms are stacked together, they perform well, as shown below how stacked attention map is able to pinpoint the region of an image which is related to a question. Figure 1 shows a sample example of how the stacked attention layer works for the question "What is sitting on the basket on a bicycle?"



Original Image First Attention Layer Second Attention Layer

Figure 1: Multiple attention layers stacked together.

The stacked attention mechanism is a sequential approach, i.e., the next layer of the attention mechanism attempts to reduce the error caused by the previous attention layer. This approach is like a bagging or bootstrap aggregating algorithm. Another attempt is to try boosting approach, i.e., instead of sequential attention layer using parallel attention layer and at the result aggregates all attention layers data (like boosting concept). The use of multiple layers parallel manner and aggregating the results to get combined data works well and reduces the chance of overfitting.

In the paper (Goyal, Y. et al.,2017), the authors have emphasized on visual part or image part in VQA. The language part can be biased based on various human and other unavoidable biases.

In the paper (Anderson, P. et al., 2018), the authors proposed a mixed top-down and bottom-up approach for the generation of attention maps rather than the existing top-down approach for generating attention maps. The existing top-down approach is generating an attention map based on feature weights. The authors are claiming the attention maps are being generated for each feature. The proposed model generates the attentions map where each bounding box is associated with an attribute followed by an object, as shown in Figure 2.

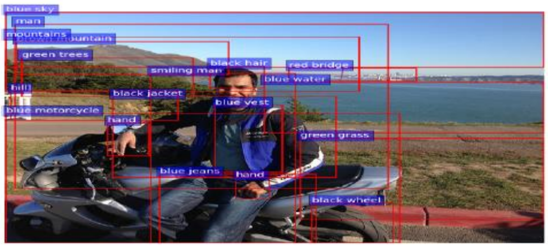


Figure 2: Attention map for a bottom-up approach.

In the above image, the top left bounding box is attached to "object: sky "and the attribute of sky object, "color: blue."

3 Related Datasets

This section is focused on the different dataset descriptions. Dataset is an important part of VQA.

DVQA dataset:

This dataset was developed in the year 2018. This is restricted only to research and educational purpose. This dataset also provides additional supplemental material. Along with the question-answer pair, detailed annotations of every object have been provided in the form of a bar chart. (Kafle, K. et al., 2018).

VQA v1 and v2 Dataset:

This dataset was developed in the year of 2016. This dataset has 256016 images. The images are inherited from the COCO image database. VQAv1 and VQAv2 have around 0.6M and 1.1M questions, respectively. The dataset has 7.9M answers. The dataset has 50K abstract scenes and 15K questions, and 1.9M answers to cater to the need of analysis abstract scenes (Zhang, P. et al., 2016).

VCR Dataset:

The dataset has 290K multiple choice questions along with their answers (290K). The dataset maintains answers are of 7.5 words on average. (Zellers, R. et al., 2019).

GQA Dataset:

This dataset was developed in the year 2019.The dataset has 110K images, where each image is associated with a scene graph of objects and relations. The dataset has 22M multistep questions.

A new metric also has been added to this dataset which is a combination of Accuracy, consistency, validity, and plausibility (Hudson, D. A. et al., 2019).

CLEVR:

This dataset was developed in the year 2018. The train set has 70K images and 700K questions, validation, and test sets have 15K images and 150K questions. Scene graph annotations are additional supplemental material available to boost the performance (Johnson, J. et al., 2018).

4 Experimental Result and Performance Analysis

For the experiment, we have used 30 epochs with minibatch gradient descent with a batch size of 128. Each epoch took almost 1.5 hours for the stacked area network model and roughly 2 hours for the VQA model.Now we will discuss internal model details.

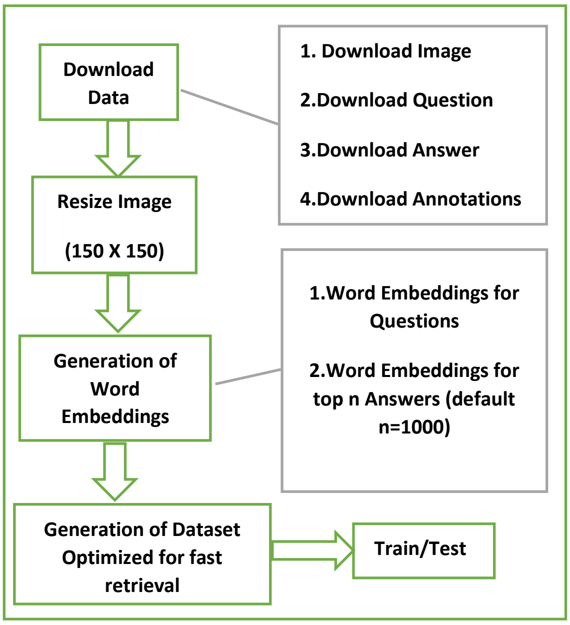

Download Data
1. Download Image
2.Download Question
3.Download Answer
4.Download Annotations
Resize Image
(150 X 150)
1.Word Embeddings for Questions
2.Word Embeddings for top n Answers (default n=1000)
Generation of Word Embeddings
Generation of Dataset Optimized for fast retrieval
Train/Test

Figure 3: Execution phases for VQA models.

Visual Question Answer (VQA) Model:

As stated earlier, in any VQA model, image and question-answer have been trained or processed separately, and at a later point in time, the feature vectors from image and text have been combined in a different manner. In the VQA model, we have used the pre-trained CNN model VGG19 to retrieve the image feature vector. For question-answer processing, we have used the Embedding layer and followed by a bidirectional LSTM layer to understand word embedding sequences pattern. During the experiment both image and text features are processed separately however we need to combine both the features so that the combined output can be fed into subsequent Neurons in Deep Learning frameworks. The image and text feature vectors have been combined using a torch tensor multiplication (like matrix multiplication). We have also ensured bypassing the argument to image and feature vector to have the same size. Followed by this step, we have used two hidden layers, and we have used tanh activation function in our hidden layer. The use of tanh activation function ensures we have bounded the output within -1 to 1 range. We have tried with relu and leaky relu but did not do well.

Multimodal Tucker Fusion for Visual Question Answering (MUTAN):

This model is similar to the VQA model with a small difference. In VQA, the image and text features are mixed using torch multiplication. However, in this model, image and text feature passed through separate deep neural network. Then the output from both the neural networks has been multiplied and forwarded to output generation.

Stacked Attention Network Model (SANM):

In this model also image and question-answer pairs have been processed in the same way as with the previous VQA model. However, there is a small change in combining image and text features. During the VQA model, we are doing simple torch tensor multiplication of text and image features, however, in this model, we are first creating a list of attention layers, and for each attention layer, we are combining image features and text features related to that attention map. Then all attention maps are stacked together, and they are further passed through a deep neural network with a single hidden layer where both image and text feature vectors have been passed with tanh as an activation function. Further, in the final layer, we have used softmax to extract which attention maps are relevant for that image-question pair. Then the combination of an image feature, word feature, and combined output from the stacked attention layer is passed through a shallow neural network for further processing.

Model Performance:

We have performed the experient on VQAv2 dataset. Table 1 shows the experimental results for models. As stated, earlier the VQA model performance is slightly better than the SANM model.

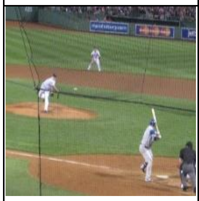
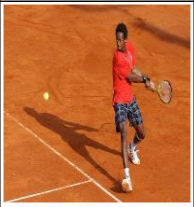
Image	Question	Output		
		SANM Model	VQA Model	Mutan
	Is the ball flying towards the batter?	'no' - 0.2597 'yes' - 0.2281 'both' - 0.1156	'yes' - 0.1507 'no' - 0.1015 'ground' - 0.0380 'floor' - 0.0192	'yes' - 0.2160 'no' - 0.1992 'night' - 0.1057 'red' - 0.0292
	What sport is being played?	'yes' - 0.2171 'no' - 0.1455 'old' - 0.0626 'both' - 0.0529 'new' - 0.0351	'tennis' - 0.1301 'baseball' - 0.0569 'unknown' - 0.0158 'none' - 0.0141	'yes' - 0.2466 'no' - 0.1710 'red' - 0.0520 'night' - 0.0375

Table 1: Experimental Results on test dataset

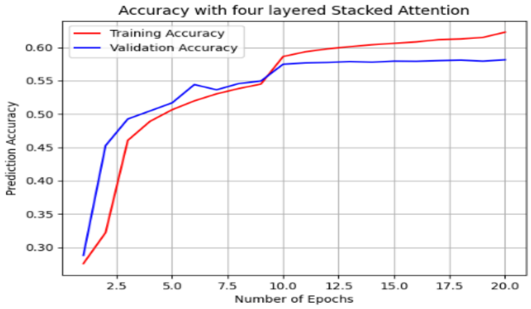


Figure 5: SANM model accuracy.

From Figure 5 we can see during epoch 8 or 9, the difference between training and validation loss was almost 0. We have taken that model as our best model since it reduces overfitting and generalizes well. The Accuracy is around 55%, and the loss is around 2.2.

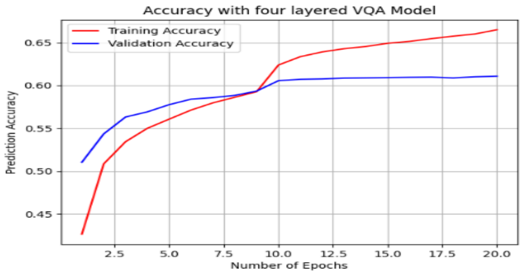


Figure 6: VQA model accuracy.

From Figure 6, we can see during epoch 8 or 9, the difference between training and validation loss is almost 0. We have taken that model as our best model since it reduces overfitting and generalizes well. The Accuracy is around 59%, and the loss is around 1.9. It is quite evident that the 2nd model is performing better than the 1st model.

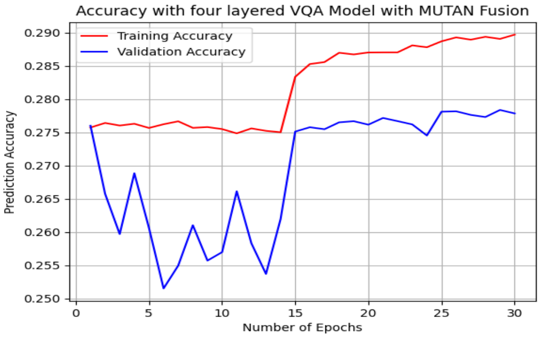


Figure 7: MUTAN model accuracy.

From Figure 7, we can see during epoch 14 or 15, the difference between training and validation loss is almost 0. We have taken that model as our best model since it reduces overfitting and generalizes well. The Accuracy is around 27%, and the loss is around 3.8. This model performance is not up to the mark. Overall VQA model is giving best accuracy among the 3 models evaluated.

5 Conclusions and Future Scope

Loosely speaking, so far, visual question answering models and datasets follow a specific pattern. Visual Question Answering dataset normally contains multiple images, which could be natural or synthetic, and along with this, each image contains a pair or combination of question and answer. Along with these in a few datasets, additional supplemental materials have been provided to support models as a feature to make better decisions. As per the VQA model is concerned with preprocessing image and text (question-answer) part preprocessing is going separately, i.e., no relation or dependency on each other. After preprocessing, the preprocessed data from image and text have been combined, which varies from model to model and then fed into Neural network models. This is overall architecture in simple terms. In this experiment, we have tested some of the VQA models successfully in our lab environment and produced results. We have also ensured to reduce overfitting by selecting the best model where training and validation loss difference is minimum.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). *Bottom-up and top-down attention for image captioning and visual question answering.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). *Vqa: Visual question answering.* In Proceedings of the IEEE international conference on computer vision (pp. 2425-2433).

Cadene, R., Ben-Younes, H., Cord, M., and Thome, N. (2019). *Murel: Multimodal relational reasoning for visual question answering*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1989-1998).

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. (2017). *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 (pp. 6904-6913).

Hedi Ben-younes, Rémi Cadene, Matthieu Cord, Nicolas Thome. (2017). *MUTAN: Multimodal Tucker Fusion for Visual Question Answering.* In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV).

Hudson, D. A., and Manning, C. D. (2019). *GQA: A new dataset for real-world visual reasoning and compositional question answering*. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 6700-6709).

Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). *CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2901-2910).

Kafle, K., Price, B., Cohen, S., and Kanan, C. (2018). Dvqa: *Understanding data visualizations via question answering*. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5648-5656).

Kafle, K., & Kanan, C. (2017). *An analysis of visual question answering algorithms.* In Proceedings of the IEEE International Conference on Computer Vision (pp. 1965-1973).

Shalini Ghosh, Giedrius Burachas, Arijit Ray, Avi Ziskind: *Generating Natural Language Explanations for Visual Question Answering using Scene Graphs and Visual Attention*: arXiv:1902.05715v1, 2019.

Shrestha, R., Kafle, K., and Kanan, C. (2019). *Answer them all! Toward universal visual question answering models*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10472-10481).

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., and Rohrbach, M. (2019). *Towards vqa models that can read.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8317-8326).

Suhr, A., Lewis, M., Yeh, J., and Artzi, Y. (2017, July). *A corpus of natural language for visual reasoning.* In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 217-223).

Xi, Y., Zhang, Y., Ding, S., & Wan, S. (2020). *Visual question answering model based on visual relationship detection.* Signal Processing: Image Communication, 80, 115648.

Yang, Z., He, X., Gao, J., Deng, L., and Smola, A. (2016). *Stacked attention networks for image question answering.* In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 21-29).

Yi, K., Wu, J., Gan, C., Torralba, A., Kohli, P., and Tenenbaum, J. B. (2018). *Neural-symbolic vqa: Disentangling reasoning from vision and language understanding*. arXiv preprint arXiv:1810.02338.

Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. (2019). *From recognition to cognition: Visual commonsense reasoning*. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6720-6731).

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., and Parikh, D. (2016). Yin and yang: *Balancing and answering binary visual questions*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5014-5022).