Visual Question Answering Analysis: Datasets, Methods, and Image Featurization Techniques

Vijay Kumari[1], Abhimanyu Sethi[1], Yashvardhan Sharma[1] and Lavika Goel[2]
[1]*Birla Institute of Technology and Science, Pilani, Rajasthan, India*
[2]*Malaviya National Institute of Technology, Jaipur, Rajasthan, India*

Keywords: Computer Vision, Natural Language Processing (NLP), Visual Question Answering (VQA), Attention Mechanism, Convolutional Neural Networks.

Abstract: Holistic scene understanding is a long-standing objective of core tenets of Artificial Intelligence (AI). Multimodal tasks that aim to synergize capabilities spanning multiple domains, such as visual-linguistic capabilities, into intelligent systems are thus a desideratum for the next step in AI. Visual Question Answering (VQA) systems that integrate Computer Vision and Natural Language Processing tasks into the task of answering natural language questions about an image represent one such domain. There is a need to explore Deep Learning techniques that can help to improve such systems beyond the language biases of real-world priors that presently hinder them from serving as a veritable touchstone for holistic scene understanding. Furthermore, the effectiveness of Transformer architecture for the image featurization pipeline of VQA systems remains untested. Hence, an exhaustive study on the performance of various model architectures with varied training conditions on VQA datasets like VizWiz and VQA v2 is imperative to further this area of research. This study explores architectures that utilize image and question co-attention for the task of VQA and several CNN architectures, including ResNet, VGG, EfficientNet, and DenseNet. Vision Transformer architecture is also explored for image featurization, and a myriad of loss functions such as cross-entropy, focal loss, and UniLoss are employed for training the models. Finally, the trained model is deployed using Flask, and a GUI for the same has been implemented that lets users input an image and accompanying questions about the image to generate an answer in response.

1 INTRODUCTION

There has been an astronomical surge in Computer Vision and Deep Learning research for tasks pertaining the Visual capabilities. Provided with adequate data, Deep Convolutional Neural Networks (CNNs) have come to be at par with human levels in classification tasks (He et al., 2016a). However, a holistic understanding of images, a desideratum for the next step in AI, requires multimodal learning capabilities. Recent research has paved the way to exploring tasks that broaden the spectrum of Artificial Intelligence towards more holistic capabilities, including multimodal learning, which is an essential sub-field for truly AI-complete tasks.

1.1 Background

Visual Question Answering (VQA) is the multimodal task of answering natural language questions about an image. The task of open-ended VQA entails several capabilities across the domain of AI like Activity recognition, Object detection, Spatial awareness, Attribute classification, etc.

A robust VQA system is expected to have the facilities for reasoning about images and answer a wide range of Computer Vision tasks. Recent work has welcomed self-attention-based architectures, specifically Transformers, as the de-facto standard for Natural Language Processing tasks. Due to their computationally efficient architecture, models of extraordinary size can now be trained on an enormous corpus of data. Computer Vision tasks, however, are predominantly based on Convolutional architectures until some recent study in utilizing the Transformer architectures in hopes of transferring their scalability and efficiency over (Dosovitskiy et al., 2020). To that end, Vision Transformers (Dosovitskiy et al., 2020) have been incorporated into the pipeline of a VQA system as image feature extractors in this study.

Kumari, V., Sethi, A., Sharma, Y. and Goel, L.
Visual Question Answering Analysis: Datasets, Methods, and Image Featurization Techniques.
DOI: 10.5220/0011655900003411
In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023)*, pages 281-288
ISBN: 978-989-758-626-2; ISSN: 2184-4313

1.2 Motivation

Apart from immediate applications in assisting the visually impaired, VQA systems serve as an essential component of the Visual Turing Test for image understanding. Multimodal research in the form of image and video captioning is an extensively studied and arguably the most closely related sibling task to VQA that often entails an understanding of complex object relationships and attributes to describe the contents of visual media in natural language. However, visual captioning is often shown to lack the fine-grained scene understanding required of a Visual Turing test; and the lack of a fast, cheap, and reliable automatic evaluation for generated captions further casts doubt upon its capacity as an "AI-complete" task.

1.3 Objectives/Contributions

1. This work explores attention and co-attention mechanisms pertaining to visual and linguistic modalities.
2. Explored Vision Transformers for leveraging the advantages of Transformer architecture for image featurization over Convolutional Neural Networks
3. We formulated a comparative study of different techniques (Transformers, Loss Functions, Multimodal fusion) in the field of VQA over datasets VQA, VizWiz.

2 VQA DATASETS, EVALUATION METRICS AND METHOLOGY

Some of the earliest works in the domain of Visual Question Answering were motivated by the development of a Visual Turing Test for a Computer Vision system and combined visual-linguistic parsing for natural language query answering. These works were, however, limited in their scope to constrained settings due to the unavailability of adequate datasets. The years 2014 and 2015 saw some of the earliest VQA datasets being publicly released, marking VQA research's rise. Some of them are presented below.

2.1 Early VQA Datasets

DAQUAR

One of the smallest yet earliest significant VQA datasets to arrive, the Dataset for Question Answering on Real-world images (DAQUAR) (Malinowski and Fritz, 2014) consisted of 6794 training and 5674 test question-answer pairs. Adding to its small size, DAQUAR was exclusively comprised of indoor scenes, with significant clutter and adverse lighting conditions, which made questions difficult to answer.

VQAv1

The first version of the VQA dataset (Antol et al., 2015) consisted of both 'real' and 'abstract' images. The Real Images portion of the dataset comprises over 80k,40k, and 80k images from the Microsoft Common Objects in Context (MS COCO) (Lin et al., 2014) for training, validation, and test sets, respectively. These images are complex and diverse, hence suitable for the VQA task. The abstract scenes subset containing 50k cartoon scenes were introduced to attract researchers looking to explore the high-level reasoning required for the task of VQA. The dataset comes with innate language biases in that a large proportion of the dataset's questions can be answered without looking at the corresponding images. Additionally, some questions are highly subjective and would not strictly reflect an algorithm's true capability of solving the VQA problem.

COCO-QA

In response to the need to construct a more comprehensive, diverse, and complex dataset, aimed to create a much larger body of question-answer pairs, which were automatically synthesized by converting image descriptions into QA forms. These QA pairs were generated on the MS-COCO dataset (Lin et al., 2014). COCO-QA contains 78,736 training and 38,948 tests QA pairs. The largest potential for limitation in this dataset is the question generation itself, in that it is limited to the objects described in the COCO dataset descriptions. Moreover, some of the questions generated suffer from being grammatically unclear and ambiguous.

Visual Genome

Visual Genome (Krishna et al., 2017) was the largest dataset when it was first released, consisting of over 100k images from the MS-COCO (Lin et al., 2014) and YFCC100M (Thomee et al., 2016) datasets. What is unique about the Visual Genome Dataset is that the questions were constrained to start with one of the Ws with an average of 17 questions per image, resulting in approximately 1.7 million QA pairs.

2.2 Evaluation Metrics

Several automatic evaluation metrics in the form of BLEU (Papineni et al., 2002), METEOR (Banerjee

and Lavie, 2005), and ROUGE (Lin, 2004) originally developed for machine translation evaluation lent to the captioning tasks as well and have their own set of limitations in dealing with natural language which can be highly subjective.

Simple Accuracy

The task of Visual Question Answering faces similar limitations but to a larger degree. VQA is framed either as an open-ended problem, which entails the formulation of a string to generate a natural language answer or as a multiple-choice problem which reduces to a multi-class classification problem. For the latter, simple accuracy is often employed as the evaluation metric of choice, though it may also be utilized for the former. However, this can prove largely inadequate due to the binary nature of the metric. Additionally, this also overlooks the possibility of multiple correct answers.

WUPS

Wu-Palmer Similarity (WUPS) (Wu, 1994) take semantics into consideration by striving to measure the difference between the predicted answer and the ground truth. It assigns a value between 0 and 1 based on the predicted and ground truth similarity. Semantically similar words, such as 'whale' and 'blue whale,' have a higher WUPS score than, say, 'whale' and 'table.'

VQA Challenge

The VQA Dataset contains ten answers per question by different annotators. The evaluation metric utilized by the dataset and associated challenge is:

$$Accuracy_{VQA} = min(\frac{n}{3}, 1) \quad (1)$$

where n denotes the count of annotators who had the same answer as the one predicted. A model is given a full score based on if the predicted answer corresponds to three or more annotators for a question.

Human Evaluation

Finally, there is the method of human evaluation, but it presents a long list of problems that make such a method infeasible, in terms of time, resources, and expenses. Additionally, measuring a system's performance iteratively to strive to improve it greatly adds to the problem. Lastly, judging the quality of an answer requires criteria to be given to the judges. The ideal evaluation metric for a VQA system remains to be an open question.

2.3 Existing VQA Models and Methodology

Some of the earliest VQA pipelines follow the most intuitive forms of VQA pipelines, and models today continue to morph from this fundamental architecture where extracted features of both image and input question are often fed into a multi-layer perceptron (MLP) classifier after they have been fused together (Antol et al., 2015; Kafle and Kanan, 2016; Zhou et al., 2015). Some of the recurring methods of combining features include element-wise addition, product, and concatenation.

Featurization approaches have seen extensive variety over past classification frameworks. The authors of (Zhou et al., 2015) employed the GoogLeNet architecture for extraction of image features, and a bag-of-words model to represent the question features, which were concatenated and then fed to a multi-class logistic regression classifier. Skip-thought vectors were utilized by (Kafle and Kanan, 2016) for question features and ResNet-152 for visual extraction.

The authors of (Antol et al., 2015) utilized an LSTM encoder to represent question features with GoogLeNet for visual features. After equalizing the dimensionality of the two features, they were fused using the Hadamard product, which was propagated to a 2-layer MLP. In (Malinowski et al., 2015), each word and the corresponding CNN features of the image concatenated were sequentially fed into the LSTM. A softmax classifier then predicted the answer.

Several studies have pointed towards the idea that only using global features may overlook the significance of task-relevant regions of the input space. By employing attention mechanisms, models learn to 'attend' to the most relevant regions based on the given task. Attention-based architectures have proved to perform greatly in several NLP and vision tasks, including image captioning, semantic segmentation, object detection, and machine translation. For the task of VQA as well, instead of global features, several models have incorporated spatial attention to formulating region-specific CNN features. The motivation behind this approach is that some specific visual regions in an image, and likewise certain words in a question, contain more information about the task than others. A VQA model consists of the following components or phases:

Image Featurisation

CNN's pre-trained on ImageNet have consistently served as well-performing networks for extracting features for the task of VQA. Two of the most widely

employed are the VGG (Simonyan and Zisserman, 2014) and ResNet (He et al., 2016b) networks. Owing to increasing availability and accessibility to computational resources, recent works have leaned towards utilizing the more complex ResNets, which often achieve better results as well.

Question Featurisation

Skip-thought vectors (Kiros et al., 2015), Bag-of-words (BOW), gated recurrent units (GRU) (Cho et al., 2014), and long short-term memory (LSTM) encoders (Hochreiter and Schmidhuber, 1997) are some of the methods adopted for question featurization.

Feature Integration

Simple mechanisms such as element-wise addition, element-wise multiplication, and concatenation (Antol et al., 2015; Kafle and Kanan, 2016; Zhou et al., 2015) are widely used in VQA research. Additionally, Bilinear pooling (Kim et al., 2016; Saito et al., 2017) has often been employed. Computing spatial attention maps for the visual features based on question features (Yang et al., 2016), and using Bayesian models that utilize the question-image-answer feature distributions (Kafle and Kanan, 2016; Malinowski and Fritz, 2014) are also prevalent.

Answer Generation

The most common is the classification framework, but some have also utilized frameworks to generate multi-word answers sequentially, such as (Malinowski et al., 2015), that used LSTMs for the same.

3 DATA SET(S) USED

The following datasets were employed for experiments with VQA systems (chronologically)

3.1 VizWiz

The dataset contains 20,523 image/question pairs and 205,230 answer/answer confidence pairs for the Training. The Validation set has 4,319 image/question pairs and 43,190 answer/answer confidence pairs. The Test set has 8,000 image/question pairs distributed among 4 question types: Yes/No, number, other and unanswerable types.

3.2 VQA v2

The dataset contains 82,783 images, 443,757 questions, and 4,437,570 answers for the Training. The Validation set has 40,504 images, 214,354 questions, and 2,143,540 answers. The Test set has 81,434 images and 447,793 questions

4 PROPOSED TECHNIQUE(S) AND ALGORITHM(S)

4.1 Proposed Model

The proposed model extends upon the co-attention model first devised in (Lu et al., 2016) illustrated below. The model uses joint attention to simultaneously attend to both image regions and question parts. The co-attention mechanism is implemented hierarchically at three stages (word, phrase, and sentence/question level).

4.1.1 Symbolic Notations

Let the question with M words be denoted as $\boldsymbol{Q} = \{q_1, q_2, ...q_M\}$, q_m being the m-th word's feature vector. Word-level, phrase-level, and question-level embeddings at a position m is represented as q_m^{wo}, q_m^{ph} and q_m^{qu}. Image feature vectors at k spatial locations, on the other hand, are represented as $\boldsymbol{I} = \{i_1, i_2, ...i_K\}$. $\hat{i}^l$ and $\hat{q}^l$ denote co-attention features at each stage of the hierarchy for the image and question, respectively, where l denotes the level in the hierarchy, i.e., $1\ \varepsilon\{wo, ph, qu\}$. $\boldsymbol{W}$ represents weights throughout.

4.2 Word, Phrase, and Sentence Levels

Tokenized and one-hot encoded questions are first fed through an embedding matrix, the output of which is our word-level embeddings, $\boldsymbol{Q}^w = \{q_1^{wo}, q_2^{wo}, ...q_M^{wo}\}$. 1-D convolutions are then applied with three filter sizes denoting trigram, bigram, and unigram scopes.

$$\hat{\boldsymbol{q}}_{s,m}^{ph} = tanh(\boldsymbol{W}^l \boldsymbol{q}_{m:m+l-1}^{wo}), l \in \{1,2,3\} \quad (2)$$

which is then max-pooled to obtain the final phrase-level embeddings at each location:

$$\hat{\boldsymbol{q}}_m^{ph} = max(\hat{\mathbf{q}}_{1,m}^{ph}, \hat{\mathbf{q}}_{2,m}^{ph}, \hat{\mathbf{q}}_{3,m}^{ph}), m \in \{1,2,.....M\} \quad (3)$$

The phrase-level embeddings are encoded through an LSTM to get the sentence or sentence-level embeddings of the entire question.

A detailed illustration of the hierarchical embeddings is represented in fig. 1

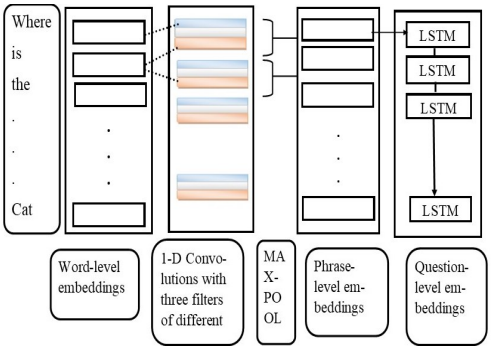

Figure 1: Proposed VQA architecture for hierarchical embeddings.

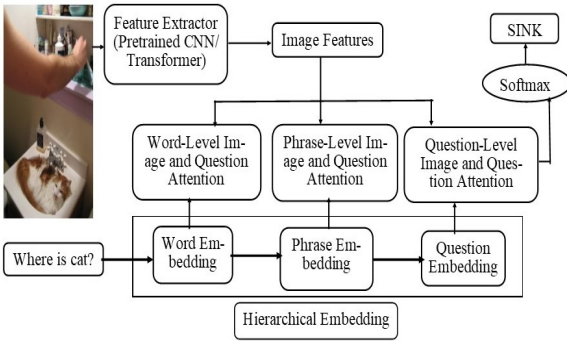

Figure 2: Simplistic representation of proposed VQA pipeline.

4.2.1 Core Mechanism

This mechanism is carried out at each stage of the hierarchy. Similarity is computed between the respective features at each image-question location pair. Concretely, an affinity matrix $A \in R^{M\times K}$ is computed between the image feature map $\boldsymbol{I} \in R^{d\times K}$ and question $\boldsymbol{Q} \in R^{d\times M}$.

$$A = tanh(\mathbf{Q}^T \mathbf{W}_b \mathbf{I}) \tag{4}$$

The image and question attention is learnt as described:

(Image attention:)

$$\begin{aligned} \boldsymbol{Z}^v &= tanh(\boldsymbol{W}_i \boldsymbol{I} + (\boldsymbol{W}_q \boldsymbol{Q})\boldsymbol{A}), \\ \mathbf{a}^v &= softmax(\mathbf{w}_{zi}^T \boldsymbol{Z}^i), \end{aligned} \tag{5}$$

(Question attention:)

$$\begin{aligned} \boldsymbol{Z}^q &= tanh(\boldsymbol{W}_q \boldsymbol{Q} + (\boldsymbol{W}_i \boldsymbol{I})\boldsymbol{A}^T), \\ \mathbf{a}^q &= softmax(\mathbf{w}_{zq}^T \boldsymbol{Z}^q), \end{aligned} \tag{6}$$

where $\mathbf{a}^i \in R^K$ and $\mathbf{a}^q \in R^M$ represent the image region i_k and word q_m attention probabilities respectively. Consequently, the image and question attention are calculated as a weighted sum:

$$\hat{\mathbf{i}} = \sum_{k=1}^{K} a_k^i i_k, \hat{\mathbf{q}} = \sum_{m=1}^{M} a_m^q q_m \tag{7}$$

at each stage of the hierarchy.

4.2.2 Predicting Answers

A multi-layer perceptron consumes the co-attention features of question and image encompassing the hierarchy to predict the answer.

$$\begin{aligned} \mathbf{z}^{wo} &= tanh(\mathbf{W}_{wo}(\hat{\mathbf{q}}^{wo} + \hat{\mathbf{i}}^{wo})) \\ \mathbf{z}^{ph} &= tanh(\mathbf{W}_{ph}[(\hat{\mathbf{q}}^{ph} + \hat{\mathbf{i}}^{ph}), \mathbf{z}^{wo}]) \\ \mathbf{z}^{qu} &= tanh(\mathbf{W}_{qu}[(\hat{\mathbf{q}}^{qu} + \hat{\mathbf{i}}^{qu}), \mathbf{z}^{ph}]) \\ \mathbf{p} &= softmax(\mathbf{W}_z z^{qu}) \end{aligned} \tag{8}$$

where p represents the answer probability.

4.3 Proposed Techniques

A simple representation of the pipeline is represented in fig. 2. An ablation study was carried out over all existing combinations of the below techniques and can be found summarised in table 3.

- Image Featurization: An array of pre-trained image feature extractors were employed:
  1. CNNs: VGG19, Resnet101, EfficientNetB5, DenseNet169 pre-trained on ImageNet
  2. Transformers: Vision Transformers pre-trained on the ImageNet and ImageNet 21k datasets
- Loss Functions: Three different loss functions were employed. The loss function formulas are listed below.
  1. **Categorical Cross Entropy**

$$CCE = -\sum_{i=1}^{N}\sum_{j=1}^{C}(y_{ij} log\hat{y}_{ij}) \tag{9}$$

  2. **Focal Loss**

$$FL = -\sum_{i=1}^{N}\sum_{j=1}^{C}\alpha(1-\hat{y_{ij}})^{\gamma}(y_{ij} log\hat{y}_{ij}) \tag{10}$$

  where α is the weight adjust hyperparameter, and γ is for adjusting the curve. The higher the value of α, the lower the loss for well-classified datasets and vice versa. Upon iterative experimentation, values of $\gamma = 2$ and $\alpha = 1$ yielded best results.
  3. **UniLoss**

$$UL = -\sum_{i=1}^{N}((1-\varepsilon) log\hat{y}_{ik} + \sum_{j=1}^{C}(0+\frac{\varepsilon}{C}) log\hat{y}_{ij}) \tag{11}$$

  where the hyper-parameter, ε, indicates the degree to which the majority class examples' influence should be divided towards other classes; $\hat{y}_{ik}$ is the predicted probability of sample i on its true k^{th} category. Upon iterative experimentation, value of $\varepsilon = 0.5$ yielded the best results.

Vision Transformers

The base variant of Vision Transformers (Vit-Base) with 12 layers and pre-trained weights were employed with a patch size of 32x32. The output of the transformer with the extra learnable token at the beginning discarded was used as the image features for the VQA pipeline.

5 EXPERIMENTS AND ANALYSIS

This section analyzes the effectiveness of different image featurization and error analysis techniques with the Co-attention architecture on the VizWiz and VQAv2 datasets.

5.1 VizWiz Dataset

Architecture

Pre-trained VGG16 and Resnet152 CNN networks were employed for Image Featurization and GloVe Word Embeddings, followed by Bidirectional LSTM layers for question featurization. Visual and Linguistic features were fused using element-wise multiplication. When utilizing only the part of the dataset with the yes/no question type, the top layers consisted of a single dense layer with 512 neurons followed by an output layer with 1 neuron and sigmoid activation for binary classification. When utilizing the entire dataset, the top layers were exactly as specified in the accompanying fig. 3. The loss functions employed were Categorical Cross Entropy, Focal loss, and UniLoss. All models were trained for 25 epochs.

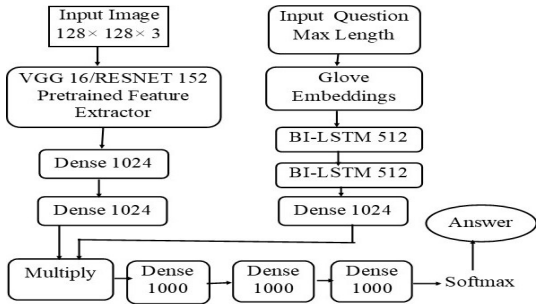

Figure 3: Architecture for the complete VizWiz dataset.

Experimental Setup

A summary of the hyperparameters settings for the VizWiz dataset is provided as follows: The maximum answer value taken is 1000, learning is rate 1e-3, and the number of epochs is 25.

Results

Accuracy of 98.98% and 60.51% were obtained on the training and test set, respectively, for the yes/no part of the VizWiz Dataset. The results on the complete dataset are summarized below in table 1. The table shows that RESNET152 with cross-entropy performs better in comparison to other networks.

Table 1: Results on the entirety of VizWiz Dataset.

Pretrained CNN	Loss function	Training accuracy	Validation accuracy
VGG16	CROSS EN-TROPY	57.89	45.00
VGG16	FOCAL LOSS	57.42	45.72
VGG16	UNILOSS	57.21	45.16
RESNET152	CROSS EN-TROPY	60.41	46.74
RESNET152	FOCAL LOSS	58.92	47.03
RESNET152	UNILOSS	58.85	47.01

5.2 VQAv2 Dataset

Architecture

The hierarchical Question-Image Co-Attention model (Lu et al., 2016) was implemented that utilizes visual-linguistic co-attention for the task of VQA. For the purposes of experimentation, only the training portion of the dataset was utilized, out of which a 75-25 split was maintained for training and validation sets. This meant that the 82,783 training images and associated 443,757 training questions were split between training and validation sets. All models were trained for 60 epochs. Categorical Cross Entropy, Focal Loss, and UniLoss were used for experimentation purposes.

Experimental Setup

A summary of the hyperparameters settings for the VQAv2 dataset is provided in Table 2.

Results

The results of the experiments conducted on the VQA v2 dataset using the proposed model and varying settings can be found summarised in table 3. Experiments were conducted with different combinations of image featurization and error analysis techniques. The Table shows that the performance of the Vision Transformers with cross-entropy loss on the training

Table 2: Summary of main hyperparameters and their values.

HYPERPARAMETER	VALUE
Maximum answers	1000
Maximum words in a sequence	22
Epochs	150
Dimension d	512
Dimension k	256
Learning rate	1e-4
Dropout rate	0.5
Optimizer	Adam
Focal Loss Gamma	2
Focal Loss Alpha	0.25
Focal Loss Epsilon	1e-9

Table 3: Results on the VQAv2 dataset (Training portion [75-25 split]).

Feature Extractor	Loss function	Training F1 Score	Validation F1 Score
VGG19	Cross-Entropy	43.19	40.42
VGG19	Focal Loss	42.34	40.45
VGG19	Uniloss	40.56	40.12
Vision Transformer	Cross-Entropy	55.16	40.13
Vision Transformer	Focal Loss	52.38	39.66
Vision Transformer	Uniloss	50.66	39.94
ResNet101	Cross-Entropy	44.87	41.41
ResNet101	Focal Loss	43.42	41.67
ResNet101	Uniloss	43.26	40.10
EfficientNetB5	Cross-Entropy	52.79	41.68
EfficientNetB5	Focal Loss	50.65	41.58
EfficientNetB5	Uniloss	43.26	40.10
DenseNet169	Cross-Entropy	53.45	41.94
DenseNet169	Focal Loss	51.65	42.08
DenseNet169	Uniloss	43.26	41.10

dataset is good. While on the Validation set, Densenet with Focal loss function performs well.

5.3 VQA Web Application

In order to provide the interface for posing the query, a web application is created. Figure 4 depicts the graphical interface. The Flask framework is used to build the web application's backend, and HTML, CSS, and JavaScript are used to build its front end. The application can accept an image and a text query as input and returns the relevant answer.

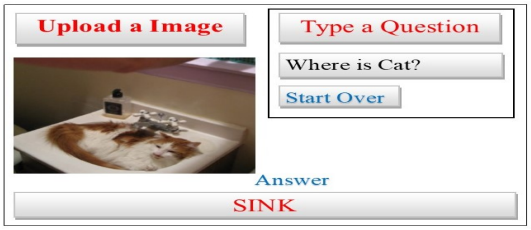

Figure 4: The GUI for the deployed VQA app.

6 CONCLUSIONS

There appears to be an indicative trend regarding attention mechanisms and their benefits to a multimodal problem, such as Visual Question Answering. Although Computer Vision problems have been unable to leverage the advantages that Transformer architectures hold over Convolutional Networks. The research on Vision Transformers to that end has been a considerable step in this regard. Pre-trained Vision Transformers were utilized in our experiments for image featurization. While they offer similar results in terms of accuracy performance, the models train much faster than one CNNs of a similar scale are employed. Co-attention mechanisms incorporated into Transformer architectures seem to be the way forward concerning VQA problems. They could hold great potential in terms of holistically understanding the visual medium as well as learning the correct regions to attend to based on the input question in natural language.

Detailed studies have shown that most models perform slightly worse when they infer answers based on only the question rather than when considering both the input image and question. There is, therefore, a need to focus on visual attention so that the visual information is adequately taken into account for generating an answer to the input space. All of the data and the final deployed model were posted to the "https://github.com/AbhimanyuSethi-98/VQA-Flask-App.git" repository.

ACKNOWLEDGMENT

The authors like to express their sincere gratitude to the Department of Science and Technology (DST/ICPS/CLUSTER/DataScience/2018/Proposal-16:(T-856)) for giving financial support at the department of CSIS, Birla Institute of Technology and Science, Pilani, India.

REFERENCES

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

Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, K., Zhang, X., Ren, S., and Sun, J. (2016b). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Kafle, K. and Kanan, C. (2016). Answer-type prediction for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4976–4984.

Kim, J.-H., Lee, S.-W., Kwak, D., Heo, M.-O., Kim, J., Ha, J.-W., and Zhang, B.-T. (2016). Multimodal residual learning for visual qa. *Advances in neural information processing systems*, 29.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. *Advances in neural information processing systems*, 28.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Lu, J., Yang, J., Batra, D., and Parikh, D. (2016). Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29.

Malinowski, M. and Fritz, M. (2014). A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27.

Malinowski, M., Rohrbach, M., and Fritz, M. (2015). 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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Saito, K., Shin, A., Ushiku, Y., and Harada, T. (2017). Dualnet: Domain-invariant network for visual question answering. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 829–834. IEEE.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., and Li, L.-J. (2016). Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Wu, Z. (1994). Palmer, m.: Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of Association for Computational Linguistics, Las Cruces, New Mexico*.

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*, pages 21–29.

Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., and Fergus, R. (2015). Simple baseline for visual question answering. *arXiv preprint arXiv:1512.02167*.