Visual Question Answering

Pratyush Jain
IIIT, Delhi
pratyush20396@iiitd.ac.in

Daksh Gulati
IIIT, Delhi
daksh20048@iiitd.ac.in

Kritarth
IIIT, Delhi
kritarth20384@iiitd.ac.in

Abstract

Visual Question Answering is a field in machine learning that has seen rapid growth in the past few years. This is mainly due to the advancements in computation and algorithms along with the motivation behind the task. This problem caters the visually-abled individuals, who can leverage assistance from such solutions in their daily life. Here we present the baseline models chosen for the problem along with the dataset details.

1 Introduction

We aim to work towards solving the problem of Visual Question Answering. It is an existing field that has areas for improvement and has seen rapid development in recent years, but the results still haven't been exceptional. Deep Learning has been instrumental in solving complex problems such as cancer prediction but tasks that are trivial for humans, such as basic VQA, are still a great challenge for DL techniques. We endeavour to help understand and solve this problem better. Our aim is to be able to answer open-ended questions from images of a variety of scenarios. This would require the use of a vast set of techniques from different disciplines. For example, objects would need to be identified, colours identified, and natural language processing techniques will be used to understand and create sentence structure. The answers would range from simple one-word answers like yes and no to simple phrases. In our final model, we use multi-modal transfer learning and build upon pre-trained existing models to improve accuracy and reduce the memory and time required to achieve this goal.

2 Dataset

We used the VQA dataset 2017 v2.0 (Goyal et al., 2017). The dataset consists of 82,783 training images, 40,504 validation images. 443,757 training questions, 214,354 validation questions and 4,437,570 training answers, 2,143,540 validation answers. Most of the questions were 4-8 words in length. Most of the answers were single words, with most common being yes, followed by no. Subject confidence was used to gauge the correctness of the answer and assign its weights accordingly.

The dataset was balanced in 2016, which is the dataset that we used as the earlier dataset was found to be extremely biased towards some particular answers. This dataset put greater emphasis on the role of images in the prediction of the answer as they identified the same question pairs with different answers and slightly differing images, thus, requiring the decision to have a greater influence on the image.

This similarity exercise did not yield image pairs for all images. But, this balancing did reduce the bias in the dataset by a big margin.

3 Related Works

3.1 VQA

This paper(Goyal et al., 2017) looks into one-ended visual question answering, their aim is to answer a question posed about a given photo. Their dataset is composed of photos from MSCOCO along with photos from other sources, a question, its answers and the confidence level in the answer. They try to combine different fields of study to achieve this. They use techniques from computer vision, natural language processing and knowledge representation. They presented a variety of baseline models such as random, prior, per-Q-type-prior and nearest neighbour. Their DL based baseline models include using activation from the last layer of VGGNet, using the l2 normalized version of this method and they also used a Bag of words as well as an LSTM. They also combined these various DL methods to achieve varying results.

3.2 VizWiz

This paper(Gurari et al., 2018) is focused on using Visual Question Answering to help blind people. They work on the premise that smartphone cameras are widely available to everybody and they can play a vital role in helping alleviate the day-to-day problems of visually impaired individuals. They also aim to help prevent an unintentional breach of privacy of blind people due to the photos they take and share on social media. Their data set consists of images taken solely by blind people so some of them are blurry and unclear, a question and its answer, which can range from one word to a phrase to unanswerable. Their baseline model included fine tuning Res-Net-50, they benchmarked variants of the same using different variants of their dataset.

3.3 Attention on Attention

This paper (Singh et al., 2018) describes an architecture that was used for the VQA challenge. Their model is based on RNN architecture. GloVe, along with GRU, is used to generate question word embeddings. Image vectors are generated using Faster R-CNN. Attention layer uses pairwise multiplication of the word and image vectors. And a special function is introduced to compute the attention weights. They experimented with several different structures of the attention layer and settled on the one that gave the best accuracy.

3.4 Prismer

The basic idea behind prismer (Liu et al., 2023) is to maximize the use of existing knowledge in the field of vision and language. They aim to circumvent the need for large datasets and huge processing time and power by using existing frameworks. It tries to take transfer learning using multi-modality to the extreme while trying to reduce the number of parameters that need to be trained. Weights are frozen so that learning is retained across pre-training and training, and adaptors are used to ensure that the learning translates well to the problem it is being applied to. It basically works as a sequence of attention block, adaptor, and feed-forward block, with most of the new learning happening in the adaptor blocks. Pre-training was done on COCO, Visual Genome, Conceptual captions, SBU captions, and Captions 12M dataset. It is found that it performs better or at least on par with models that were trained with as much as 100 times more data.

4 Methodology

Here we provide the architecture of the two baseline models, along with our final model.

4.1 Model 1: 1-LSTM Q and VGGNet16 I

For the questions, we fed the tokenised question into an embedding layer. The whole question vocabulary was passed to the embedding layer. The output from the embedding layer is then fed to the LSTM. We concatenate the hidden state and cell state representations of LSTM. We pass this vector to a fully-connected layer with tanh non-linearity to give us a 1024-dimensional vector as output.

We extracted the features of the images from VGGNet16 to give us 4096-dimensional vector representations. We then fed this to a fully-connected layer with tanh non-linearity to obtain a 1024-dimensional vector.

We then perform element-wise multiplication between the question and image representation. Lastly, we feed this 1024-dim vector to two fully-connected layers having tanh non-linearity. The outputs are the probabilities of the top-3 most frequent answers.

4.2 Model 2: 2-LSTM Q and VGGNet16 I

This model is similar to that of model-1. The only difference is that we use 2 layers of LSTM instead of one and concatenate both of their cell states and hidden states. This output is fed to a fully-connected layer, giving us the 1024-dimensional vector. The rest of the model is the same.

4.3 PrismerZ-base Model with MLP Head

Our final model is inspired from the work done in this paper (Liu et al., 2023). The design of the model is illustrated in Fig. 1. Our model is an encoder-decoder transformer model. The model has a vision encoder and a language decoder. The vision encoder is based on the architecture of the vision transformer, while the text decoder inherits the architecture of the RoBERTa model. The processed image input is fed to the vision encoder that outputs image features. The language decoder is then conditioned on these features with the help of cross attention, and then it outputs the combined multi-modal features. The features or the hidden state of the classification token from the last decoder layer is then fed to a classification head that outputs the probabilities for the top-3000 answers seen in the training set. This step is the main point

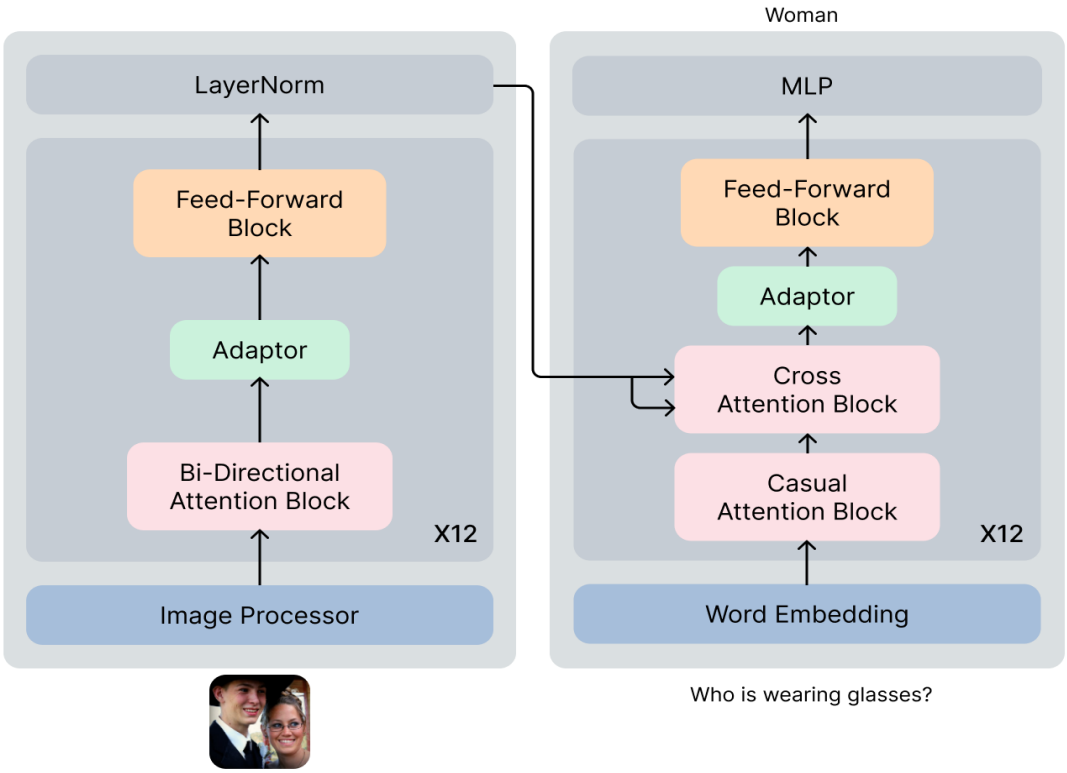

Figure 1: The Model architecture.
Source: Adapted from (Liu et al., 2023)

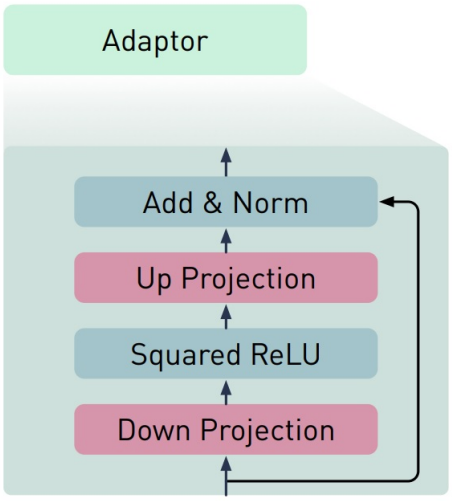

Figure 2: The Adaptor architecture.
Source: (Liu et al., 2023)

of distinction from the PrismerZ-base model (Liu et al., 2023) that produces a sequence of text tokens as output.

We choose "google/vit-base-patch16-224-in21k" as the backbone for our vision encoder and "roberta-base" as the backbone of our language decoder. We freeze the weights of these backbone layers to maintain the integrity of their learned knowledge and prevent catastrophic forgetting (Kemker et al., 2018; Kirkpatrick et al., 2017). In addition, each encoder and decoder layers of the transformer have a special trainable component: Adaptor, that assists the pre-trained layers from the vision and language backbones to better adapt to new modalities and challenges. The design of the Adaptor can be seen in Fig 2. The adaptor first down-projects the input features into a smaller dimension, applies squared ReLU as nonlinearity to improve stability, and then up-projects the features back to the original input dimension. With the residual connection, we initialise all adaptors with near-zero weights to approximate the identity function. Combined with a standard cross-attention block in the language decoder, the model is able to smoothly transition from the domain-specific vision-only and language-only backbones to a vision-language model. Since most of the network parameters are frozen, and only a fraction of the weights are trainable, our model becomes efficient for training purposes.

We utilise AutoImageProcessor and AutoTokenizer for "google/vit-base-patch16-224-in21k" and "roberta-base" respectively from HuggingFace library in order to process our images and questions

Hyperparameter	Value
Vocab Size	50265
Hidden Size	768
Max Question Length	40
Dropout	0.1
Attention Heads	12
Intermediate Size	3072
Hidden Layers	12
Epochs	20
Batch Size	32
Learning Rate	5e-5
Criterion	CE Loss
Optimizer	AdamW

Table 1: Hyperparameters for the final model.

Model	Accuracy			
	Yes/No	**Number**	**Other**	**Overall**
1-LSTM	62.30%	0.0%	0.0%	19.90%
2-LSTM	62.30%	0.0%	0.0%	19.90%
Prismer	75.06 %	35.30%	42.72%	53.90%

Table 2: Performance of all the models on the validation set.

Accuracy			
Yes/No	**Number**	**Other**	**Overall**
74.82 %	35.79%	42.92%	54.00%

Table 3: Performance of the final model on the small validation set.

and convert them in a format feedable to the model.

5 Experimental Setup

Here we provide the experimental setup and hyperparameters for the baseline models and the final model.

5.1 Baseline Models

The dimension of the embedding layer is 300. The vocabulary size for the embedding layer is 13683. The hidden layer size for the LSTMs is 70. The two fully-connected layers have 100 and 256 hidden neurons respectively with dropout (p=0.5). The model output is a 3-dimensional vector for top-3 most frequent answers.

Both the baseline models were trained for 10 epochs that took around 2 hours for each model. The machine that was used for training was Windows-based and had Nvidia RTX 2060 GPU with 6 gigabytes of VRAM. The machine had 16 gigabytes of RAM storage.

5.2 Final Model

The hyperparameters for training are detailed in Table 1. The model was trained on a linux-based virtual machine, acquired using Google Cloud Platform. The machine had 26 gigabytes of RAM, 150 gigabytes of disk storage, 4 CPU cores and one Nvidia V100 GPU with 16 gigabytes of VRAM. The model was trained for 20 epochs which took 40 hours, clocking it around 2 hours per epoch.

6 Results

We report the results of the baseline as well as the final models on the validation set in Table 2. We also report the results of our final model on the truncated small validation set to be used during the project demo in Table 3. We have divided the accuracy into the answer types seen in the dataset.

Since our baseline model was made to train to output the probabilities of the top-2 answers that were indeed "Yes" and "No", the accuracy of other answer types automatically becomes zero. We can see that our final model outperformed the baseline models in "Yes/No" answer types by about 13%. The overall accuracy of the final model comes out to be 53.90% on the validation set and 54.00% on the smaller version of the validation set.

7 Error Analysis

From the results reported in Table 2, it becomes clear that the model has learned to generalize "Yes/No" type answers well compared to the number-related answers and other open-ended answers. The accuracy of the number-type answers is severely worse compared to other categories. We provide some sample question-image pairs along with their generated answers in Fig. 3. Perhaps the model requires more data with more diverse questions about each answer-type category to perform better. The current dataset consists of only around 450,000 questions about 82,000 images. This is obviously not enough when compared to real-world scenarios. Hence, we conclude that more data might be the key to improve performance of the model, along with better architectures with more parameters to accommodate for greater learning capacity.

Figure 3: Some erroneous predictions from the model.

8 Contribution

All the members of the team contributed equally throughout the project. We held daily meetings to discuss the progress and approach to the problem. Every member shared the workload equally and no part of the project was done in isolation by any individual.

References

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. 2018. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Shikun Liu, Linxi Fan, Edward Johns, Zhiding Yu, Chaowei Xiao, and Anima Anandkumar. 2023. Prismer: A vision-language model with an ensemble of experts. *arXiv preprint arXiv:2303.02506*.

Jasdeep Singh, Vincent Ying, and Alex Nutkiewicz. 2018. Attention on attention: Architectures for visual question answering (vqa). *arXiv preprint arXiv:1803.07724*.