Visual Question Answering

Keshav Bharadwaj Vaidyanathan, Bhanu Sai Simha Vanam[*]
Department of Computer Engineering
Northeastern University
Boston, MA 02115
bharadwajvaidyanat.k@northeastern.edu,vanam.b@northeastern.edu

1 Introduction

Visual question answering (VQA) has become one of the most active research areas, due to the advancements in computer vision and natural language processing (NLP). The task of answering a natural language question concerning a picture is known as visual question answering (VQA). VQA encompasses various challenges in language representation and grounding, recognition, common sense reasoning, and specialized activities such as counting and reading. The goal of this project is to create an AI system that accepts an image and a free-form, open-ended, or natural language question regarding the image as inputs and outputs a natural language answer.

The basic philosophy would be to use CNN and RNN for the input picture and text combination, and then combine these two feature sets to generate an answer using the RNN decoder module or a simple NN classifier if the answer is binary. First, we'll strive to comprehend and build the above general template, and then we'll concentrate on understanding and improving them by identifying the architecture's fall backs.We will deal with problems of language priors in training data, introducing relation networks by object pairing with the question. Other strategies include improvements to the ConvNet backbone architecture, a feature merging approach for textual and visual feature maps, attention, and classification of intermediate question types. For the task of VQA, we will also use Transformer models like BERT, which extends the classical attention process to several layers and heads. Parallel to this, we will refer to other studies in a related topic and attempt to incorporate a few of the relevant methodologies.

2 Related Work

The task of free-form and open-ended Visual Question Answering (VQA), where given an image and a question associated with an image the task was to implement a model which would give a natural language answer was first proposed and implemented by paper [1]. They experimented with 2 different models, an MLP with 2 hidden layers with 1000 neurons in each layer and a tanh non-linearity and an Fully connected layer followed by softmax. The embedding for questions was created using LSTM with a dimension. The image embeddings were extracted from the last layer of VGGNet consisting of 4096 dimensional features. They trained said models with one or more combinations of input embeddings consisting of questions,images and answers.The model performance for both open-Answer and Multiple choice question was tabulated. The best performing model was LSTM Q+I which had an overall net accuracy of 57.75%.

As studies on VQA continued The use of a convolutional neural network (CNN) for visual

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

question answering (VQA) was proposed in paper [2]. The convolutional architecture discussed provides an end-to-end framework for learning not only the picture and question representations, but also their intermodal interactions to generate the required answer. The architecture consists of three CNNs, one image CNN for encoding picture content, one sentence CNN for composing question words, and one multimodal convolution layer for learning their joint representation for classification in the space of candidate answer words. The model was trained on DAQUAR and COCO-QA dataset.

Several research papers have begun to study VQA, however, they started moving from baseline LSTM Q + I to invest more in feature representation of Image, Question, and Multi-modular fusion. Using bottom-up attention, the paper [3] employs Faster RCNN to produce region recommendations. We take the final output of the model and perform non-maximum suppression for each object class using an IoU threshold to provide an output set of image features for or VQA. The locations where any class detection probability above a confidence threshold are then selected. A mean pooling convolutional feature is done for each selected region, resulting in a 2048-dimension image feature vector, which is subsequently weighted using a 'soft' top-down attention mechanism, with the question representation as context. Finally, the well-known joint multimodal embedding of the question and the image is implemented, followed by a regression score prediction over a collection of candidate answers.

Implementation of efficient VQA systems required the examination of natural language questions in VQA. Paper [4] Studied the natural questions in VQA dataset in detail, finding that they necessitate a different set of text representations than previous natural language processing tasks. The paper proposed a CNN architecture which consists of a CNN and gate for learning these text representations. The CNN + gate model achieves accuracy of 61.33% which is higher than other baseline models.

In the paper [5] image features are extracted using Faster-RCNN Bottom-up attention, and the output of GRU for each time step gives the question embeddings at each token. Following that, a Bilinear Attention map is created with inputs of question feature vectors and an Image feature vector, followed by a Bilinear Attention network. Finally, the joint representation is represented using a two-layered Classifier. When compared to [3], this method yielded somewhat better outcomes.

The research in the publication [6] focuses on creating a unified image-text representation, in which multi-modal inputs are processed simultaneously for combined visual and textual understanding. Masked Language Modeling, Mask Region Modeling, Image Text Matching, and Word Region Alignment are the four pre-trained tasks. Transformer was chosen as the model's core to make use of its elegant self-attention method for learning contextualised representations. We pre-train UNITER using four pre training tasks, including Masked Language Modelling conditioned on the image, Masked Region Modelling conditioned on text, Image-Text Matching, and Word-Region Alignment, as inspired by BERT, which has successfully applied Transformer to NLP tasks through large-scale language modelling. Visual Question Answering is one of the downstream jobs that uses this Universal representation.

3 Methods

3.1 Baseline Model

We experimented by building the baseline LSTM Q+I model with accuracy as our prime performance metrics.

The first step was to create both the question and the image embeddings:-

1. Question embeddings - An LSTM with one hidden layer is used to obtain 1024-dim embedding for the question. The embedding obtained from the LSTM is a concatenation of the last cell state and last hidden state representations (each being 512-dim * 2 bidirectional) from

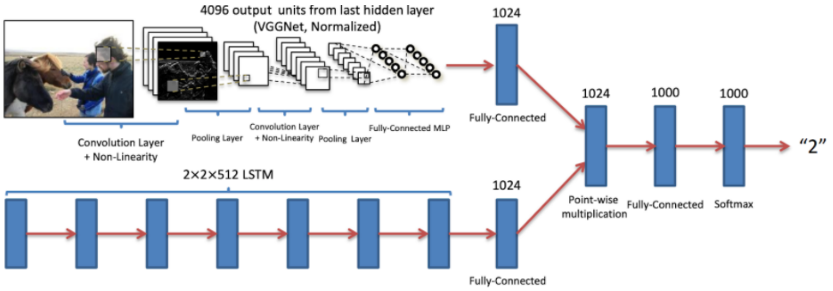


Figure 1: *Figure from paper [1].*Baseline model architecture

the hidden layer of the LSTM. Each question word is encoded with 300-dim embedding by a fully-connected layer + tanh nonlinearity which is then fed to the LSTM. The input vocabulary to the embedding layer consits of all words present in the questions of the training set.

2. Image embeddings - 4096-dim Image embedding is generated using a pre-trained VGG network

The combined image and question embedding was then passed to an MLP – a fully connected neural network classifier with 2 hidden layers and 1000 hidden units (dropout 0.5) in each layer with tanh nonlinearity, followed by a softmax layer to obtain a distribution over K answers. The entire model is learned end-to-end with a cross-entropy loss. VGGNet parameters are frozen to those learned for ImageNet classification and not fine-tuned in the image channel.

3.2 Transformer based model

Visual conceptions, language semantics, and, most critically, the alignment and linkages between these two modalities are all required for vision-and-language understanding. Transformer models, in essence, use attention and cross attention mechanisms to capture this critical alignment and interactions between the two modalities.

Learning Cross-Modality Encoder Representations from Transformers is a transformer framework that consists of a large-scale Transformer model with three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. To increase its capabilities it has been pretrained with masked language modeling using image caption data and cross modality matching to match caption and images which is similar to next sentence prediction in BERT.

The embedding layers converts the inputs (i.e., an image and a sentence) into two sequences of features:-

1. Word-Level Sentence Embeddings - The sentence is split into a sequence of words $\{w_1, ..., w_n\}$ using a WordPiece tokenizer.The word w_i and its index i are projected to vectors by embedding sub-layers, and then added to the index-aware word embeddings:

$$\hat{w}_i = WordEmbed(w_i)$$
$$\hat{u}_i = IdxEmbed(i)$$
$$h_i = LayerNorm(\hat{w}_i + \hat{u}_i)$$

2. Object-Level Image Embeddings - Faster R-CNN is used to generate region proposals. If m region are detected in the image denoted by $\{o_1, ..., o_m\}$, Each region or object o_j is represented by its position feature (bounding box coordinates) $p_j$and its 2048-dimensional region-of-interest (RoI) f_j. Position aware embedding. In our case we use top 36 region

proposals. v_j is learnt by adding outputs of 2 fully-connected layers.

$$\begin{aligned}\hat{f}_j &= LayerNorm(W_F f_j + b_F)\\ \hat{p}_j &= LayerNorm(W_p P_j + b_p)\\ v_j &= \frac{(\hat{f}_j + \hat{p}_j)}{2}\end{aligned}$$

Encoders are built on the basis of two kinds of attention mechanisms: self-attention layers and cross attention layers:-

1. Single-Modality Encoders - After the embedding layers, two transformers encoders namely language encoder and object-relationship encoder are applied to respective modality. A self-attention sub-layer and a feed-forward sub-layer are included in each layer of a single modality encoders, with the feed-forward sub-layer consisting of two fully-connected sub-layers. N_L and N_R layers in the language encoder and the object-relationship encoder, are taken respectively.

2. Cross-Modality Encoder - Two self-attention sub-layers, one bi-directional cross-attention sublayer, and two feed-forward sub-layers make up each cross-modality layer in the cross-modality encode. N_X of these cross modality encoders are implemented. . Inside the $k-th$ sublayer the bi-directional cross-attention sub-layer is first applied, which contains two unidirectional cross-attention sub-layers: one from language to vision and one from vision to language. The query and context vectors are the outputs of the $(k-1)-th$ layer. This encoder captures the alignment and linkage between the two modalities.

$$\begin{aligned}h_i^k &= CrossAtt_{L\rightarrow R}\left(h_i^{k-1}, \left\{v_1^{k-1}, \ldots, v_m^{k-1}\right\}\right)\\ v_j^k &= CrossAtt_{R\rightarrow L}\left(V_j^{k-1}, \left\{h_1^{k-1}, \ldots, h_n^{k-1}\right\}\right)\end{aligned}$$

Further building internal connections, the self-attention sub-layers are then applied to the output of the cross attention sub-layer.

$$\begin{aligned}h_i^k &= SelfAtt_{L\rightarrow L}\left(h_i^{k-1}, \left\{h_1^{k-1}, \ldots, h_n^{k-1}\right\}\right)\\ v_j^k &= SelfAtt_{R\rightarrow L}\left(v_j^{k-1}, \left\{v_1^{k-1}, \ldots, v_m^{k-1}\right\}\right)\end{aligned}$$

Residual connection and layer norms are added after each sublayer and the finally $k-th$ layer output is fed to a Feed forword layer.

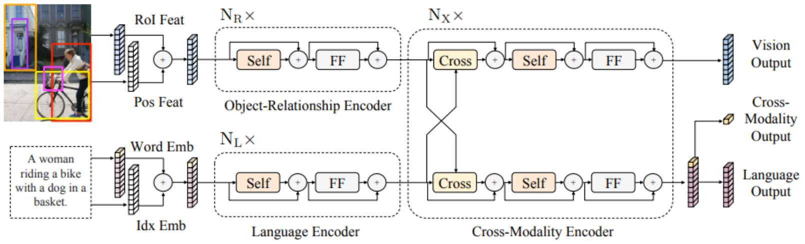


Figure 2: *Figure from paper [8]*The LXMERT model for learning vision-and-language cross-modality representations

The model has three outputs for language, vision, and cross-modality, respectively. The language and vision outputs are the feature sequences generated by the cross-modality encoder. We then combine both the coss-modality outputs and pass them through 2 Fully connected layers. we then use softmax to generate distribution over k answers. In our case k is 3128.

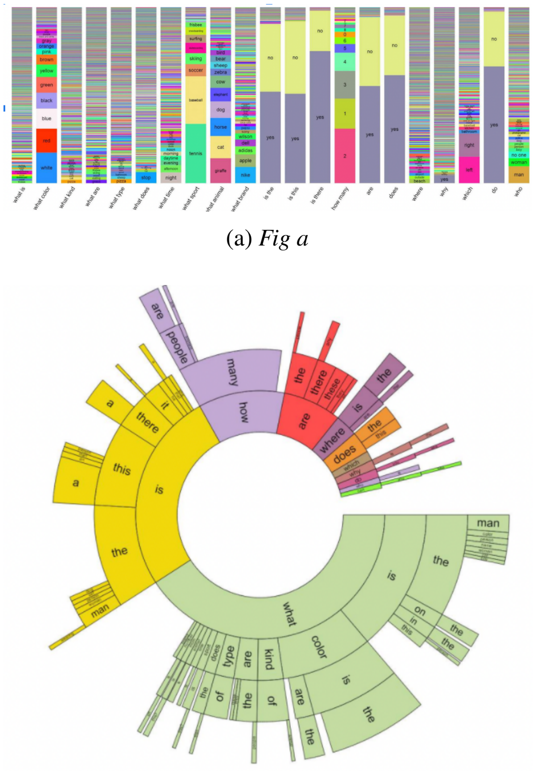
4 Experiments

4.1 Dataset

We will be using VQA dataset which contains binary classification questions, count questions, and few open-ended questions which require some domain knowledge and common sense to address the above mentioned problem.

VQA dataset has broadly two sets of images one comprises real images and the other is abstract images. Our models will be trained on real images dataset. Real Image comprises 123,237 training and validation images. The dataset includes 614,163 questions and 7,984,119 answers (including answers provided by workers with and without looking at the image) for 204,721 images from the MS COCO dataset and 150,000 questions with 1,950,000 answers for 50,000 abstract scenes. Each set is further bifurcated into Open-ended and Multiple Choice Questions. There are also set question types which are Yes/No, Number and others.

So summarizing it, VQA contains at least 3 questions (5.4 average) per image, 10 ground-truth answers power question, and 3 plausible but likely incorrect answers for questions.



(a) *Fig a*

(b) *Fig b*

Figure 3: *Figure from paper [1]. a and b shows distribution of different types of question present in the dataset*

The vqa dataset has seprate train dataset ,val dataset and test dataset. All our models use the same split as shown in *table*1.

4.2 baseline

The baseline model was trained on discovery cluster with NVidia Tesla P100 GPU. Hyperparameter of our baseline model can be seen in *table*2. The results are show in *table*4

	Train	Validation	Test
Number of Questions	443,757	214,354	447,793
Number of Images	82,783	40,504	81,434

Table 1: VQA train test split

Hyper parameters	
Learning rate	$1e^{-3}$
Batch Size	64
output features	2048
epochs	50
preprocess batch size	64
dropout	0.5

Table 2: Hyper parameters of our baseline model

4.3 Transformer Model

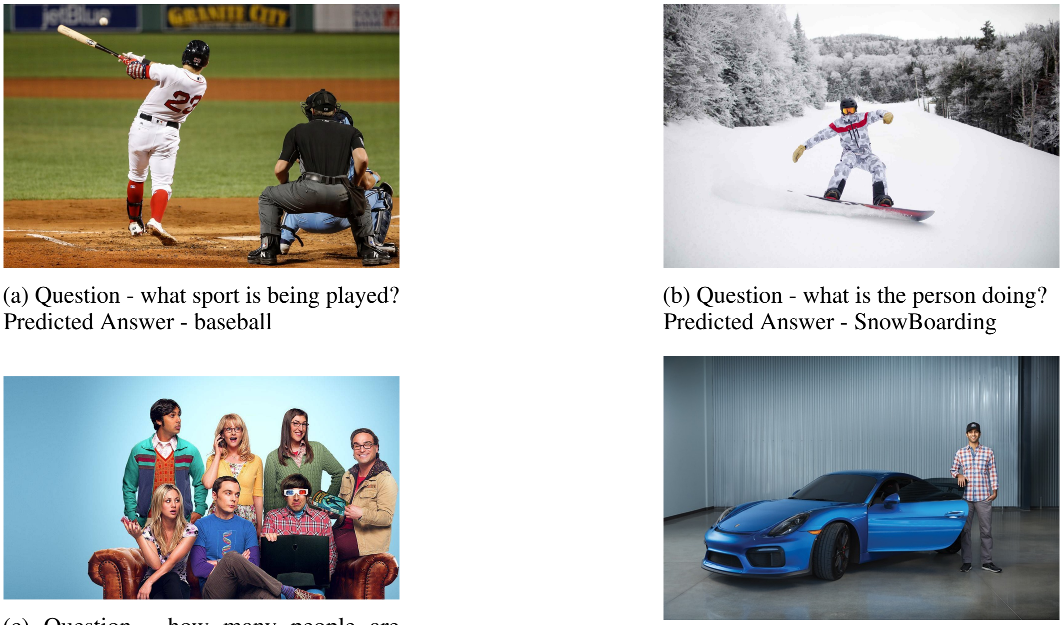
The transformer model was trained on discovery cluster with NVidia Tesla P100 GPU. Training time was around 9 hours. Hyper parameters of the model are shown in *table*3

Hyper parameters	
Learning Rate	$3e^{-5}$
Learning rate halflife	50000
Batch Size	64
epochs	25
dropout	0.1

Table 3: Hyper parameters of our Transformer model

Running inference on a few example is show in *Figure*4

We calculated the accuracy for different segments of the dataset like, Yes/No, Numbers and other types of questions. We compare the performance of our models with the state of the art models. This is shown in *table*4



(a) Question - what sport is being played?
Predicted Answer - baseball

(b) Question - what is the person doing?
Predicted Answer - SnowBoarding

(c) Question - how many people are present?
Predicted Answer - 7

(d) Question - is the person standing beside the car
Predicted Answer - yes

Figure 4: Examples of questions,images and answers predicted by our model

open ended				
	All	Yes/No	Count	Other
snubi-naverlabs	60.60	82.23	38.22	46.99
MM$_{PaloAlto}$	60.36	80.43	36.82	48.33
nearest neighbour	42.70	71.89	24.36	21.94
deeper LSTM Q + norm I	57.75	80.50	36.77	43.08
Baseline model	**52.71**	**76.32**	**34.68**	**36.12**
Transformer based model	**70.68**	**87.4**	**53.5**	**61.8**
Reniassance	79.78	93.27	68.4	70.65

Table 4: Accuracy of different models for task of VQA

5 Conclusion

In conclusion, We took up the task of Visual Question Answering (VQA) where Given an image and an open-ended, natural language question about the image the task is to provide an accurate natural language answer. We used the VQA2.0 dataset and built a baseline model which consisted of an LSTM network to generate language embeddings and a pretrained VGG network to generate image embeddings. This embeddings were then passed through an MLP with 2 fully connected layers and softmax in order to generate probability distribution over K answers. In order to better capture the inter modality fusion between language and image we developed a transformer based model which has single as well as cross modality encoders to generate self and cross attention respectively. The output of the transformer model was then combined and passed through an MLP with 2 fully connected layer followed by layer norm and softmax to generate probability distribution over top k answers.

Future work includes :-

1. Increasing the diversity of the dataset as we observed that limitation in the dataset causes inappropriate prediction example:- The image of a person playing cricket was predicted as baseball as both the sports have similar features.
2. Incorporating common sense and openended questions in the training corpus.
3. Incorporating a decoder module to predict sentence answer rather than a single predefined set of tokens.

References

[1] https://doi.org/10.48550/arxiv.1505.00468Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C., Batra, D. & Parikh, D. VQA: Visual Question Answering. (arXiv,2015), https://arxiv.org/abs/1505.00468

[2] https://doi.org/10.48550/arxiv.1506.00333Ma, L., Lu, Z. & Li, H. Learning to Answer Questions From Image Using Convolutional Neural Network. (arXiv,2015), https://arxiv.org/abs/1506.00333

[3] https://doi.org/10.48550/arxiv.1707.07998Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S. & Zhang, L. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. (arXiv,2017), https://arxiv.org/abs/1707.07998

[4] Wang$_2018 Wang, Z.\&Ji, S.LearningConvolutionalTextRepresentationsforVisualQuestionAnswering.Procee$
$602(2018, 5), https://doi.org/$10.1137

[5] https://doi.org/10.48550/arxiv.1805.07932Kim, J., Jun, J. & Zhang, B. Bilinear Attention Networks. (arXiv,2018), https://arxiv.org/abs/1805.07932

[6] https://doi.org/10.48550/arxiv.1909.11740Chen, Y., Li, L., Yu, L., Kholy, A., Ahmed, F., Gan, Z., Cheng, Y. & Liu, J. UNITER: UNiversal Image-TExt Representation Learning. (arXiv,2019), https://arxiv.org/abs/1909.11740

[7] https://doi.org/10.48550/arxiv.1908.03557Li, L., Yatskar, M., Yin, D., Hsieh, C. & Chang, K. VisualBERT: A Simple and Performant Baseline for Vision and Language. (arXiv,2019), https://arxiv.org/abs/1908.03557

[8] https://doi.org/10.48550/arxiv.1908.07490Tan, H. & Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. (arXiv,2019), https://arxiv.org/abs/1908.07490