Visual Question Answering System for Indian Regional Languages

Dhiraj Amin
Dept. of Computer Engineering
Pillai College of Engineering
India
amindhiraj@mes.ac.in

Dr. Sharvari Govilkar
Dept. of Computer Engineering
Pillai College of Engineering
India
sgovilkar@mes.ac.in

Sagar Kulkarni
Dept. of Computer Engineering
Pillai College of Engineering
India
skulkarni@mes.ac.in

***Abstract*—Questions are asked over the image within the Visual question answering system (VQAS). VQAS takes input as a user's natural language question and relative image. VQAS are mostly trained with a huge dataset with image, question and its correct answer. Many datasets like VQA, Visual7W and others have questions and answers over the image in English but there is scarcity of resources for many regional languages like Hindi and Marathi. In this paper we investigate different methods over the Hindi and Marathi version of the easy-VQA dataset.**

Keywords—visual question answering, question answering

I. INTRODUCTION

Humans can answer any question from its common sense ability if an image is shown of some children playing cricket on the ground to the human. Natural Language Processing (NLP) can be combined together with Computer Vision (CV) to replicate human capability in computers to answer any natural language question asked over the image or video. Here NLP is used to understand natural language questions and various computer vision methods can be used to understand and analyze images. To get relevant answers for the input questions over the input image Visual Question Answering System (VQAS) can be constructed. Image and User question and abstract or real image is submitted as input. The VQAS system provides the output as the correct answer as a word or number or it can be a phrase describing the answer to the question. Normally the answer is binary yes or no or it can be a multiple choice answer where each choice is presented with a level of accuracy that the answer is correct as compared to other similar choices. VQAS are mostly trained with a huge dataset with image, question and its correct answer.

II. VISUAL QUESTION ANSWERING DATASETS

For the VQA system to be accurate the system should be trained by varied deep learning models over the huge dataset. The dataset for VQAS should have a large number of images and multiple related questions over it with precise answers. If a dataset has more details like concept, scene, object or other entity then there are chances of a trained model to be highly accurate. Microsoft Common Objects in Context (COCO) is used for images in most of the dataset developed for VQA [1]. MS COCO dataset has around 3,28,000 counts of images having overall 2.5 million associated labels for the images [1].

DAQUAR [2] and COCO-QA [3] datasets are built over MS COCO [1]. Visual Question Answering (VQA) dataset [8] consists of 248,349 Training questions, where all the questions for COCO-VQA are built over MSCOCO [1] 82,783 training images. Here answers are in the form of a phrase or one word and for every question 10 accurate and 3 answers are provided without looking into images. Further the dataset consists of 248,349 testing questions for 81,434 testing images and around 121,512 Validation questions for respective 40,504 Validation images. Visual Genome [4] dataset consists of seventeen pairs of question and answer for every image; the size of the complete dataset consists of around 1.7 million Q-A pairs over the images. The dataset consists of 108,249 images having questions of all the Wh type of questions and it's the largest dataset among all VQA task datasets. In the dataset no yes/no binary form questions provided and complex Wh type questions are prioritized. In Visual7W [5] dataset each question in the dataset has one correct answer out of four candidate answers where a total 47,300 images are present. The questions are of 7 types mainly where, what, why, where, who, when and how. Most of the questions are of the form of telling and pointing form and no binary questions.

Visual Madlibs [6] is built over COCO dataset where it consists of 360,001 fill in the blank type of questions over 10,738 images having 12 varied types of fill in the blanks form of question. CLEVR dataset [7] provided by Stanford consists of synthetic images and the questions having 15,000 test and validation images. Images in the dataset have information regarding the location and other attributes of each object related to another object within the image. The objects in the image are in the form of cube, cylinder and sphere. Objects of different shapes, colors, sizes are seen in the images. FigureQA dataset [8] are synthetic and mainly consist of graphs, plots and charts related questions. The questions examined various characteristics like minimum, area under the curve, intersection, smoothness and maximum also its specified relation among plot elements in the image. SHAPES [9] dataset consist of all binary yes/no questions and the dataset does not have any bias. Question ranges from finding position, relationship and attribute of the different shapes.

VQA 2.0 [10] dataset, number of questions were increased to 443,757 for the same existing images in VQA dataset. For the pairs of answer, image and question another image was found where the previously asked question was related to the new image and answer for the new image would be different for that particular question. It helped to reduce bias in the dataset and made the model generation more accurate. The Fact-based Visual Question Answering (FVQA) dataset [11] provides supporting facts to the question answer pairs. The facts are represented as a triplet which are extracted from DBpedia, ConceptNet related to the visual concept in the image. FVQA has 4,608 questions. 580 visual concepts consist of 193,005 candidates supporting facts. Easy Visual Question Answering easy-VQA [12] has 48k questions, 5k geometric images and 13 probable answers of type Yes/No and Shapes and Colors. Most of the dataset for the VQA task is available with English text. The English

2022 5th International Conference on Advances in Science and Technology (ICAST) | 978-1-6654-9263-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICAST55766.2022.10039528

version of easy-VQA dataset was translated to Hindi text for experimental purposes.

III. VARIOUS APPROACHES FOR DEVELOPMENT OF VISUAL QUESTION ANSWERING SYSTEM

Most of the research started from the year 2015 with the release of Visual Question Answering (VQA) dataset. Various VQA systems differ on the way features are created from question and image and the way in which the same are combined together and classified. Bilinear pooling was used over traditional element based approach to allow more deeper concatenation by performing outer product at lower dimension space. Attention based approach was used to improve existing systems where attention was given to specific regions of the image and question instead of generating complete features. Semantic information related to the object and its position are extracted in Bayesian models to provide accurate results for specific objects and its position based question. Compositional models are designed to split the question into a series of sub parts to solve complex questions.

A. Models using joint embedding approach

Many authors have proposed varied approaches to improve joint embedding approach where the vector of image and the question together are embedded as one vector. Antol, Stanislaw, et al. BoW Q + I model has a bag of words as question features from top unique 1000 words from the questions and VGGNet CNN model is used as image embedding which is concatenated together. In another LSTM Q +I model [3], LSTM was used to generate word embedding and the same VGGNet for image embedding. To merge image embedding with question embedding the image is transformed to 1024 dimension and merged with question by doing element wise multiplication. The combined features are then transferred to Multi-Layer Perceptron accompanied by softmax to get the top answer class for the question. Ren, Mengye et al. proposed VIS+LSTM model [13] where image was considered as one of the words in question which was encoded using LSTM. 4096 dimension image feature vectors from VGGNet is converted to a 300 or 500 dimensional vector to sync with dimension of the word embeddings. Bolei Zhou et al. created a model [14] to find the probability of the correct answer class by extracting an image feature using GoogLeNet and the question is one hot encoded and transformed into word embedding. The combined features are fed to multiple class logistic regression models to find the correct answer class. Noh, Hyeonwoo et al. using a parameter prediction network (DPPnet) and classification network where the classification network a CNN has a fully connected dynamic parameter layer [15] where the weights of the layer is filled through weights of candidates obtained from the parameter prediction network. Allan, Jabri et al. model [16] verifies that the image answer question pair is correctly predicted or not by taking the answer as input. The question and answer are represented using BoW. The features of images are pulled out using the last layer of Resnet-101. The question answer features and images features are combined and transferred to logistic regressors and multilayer perceptrons (MLP).

B. Models using attention approach

Attention based models use most relevant image and question regions instead of using the complete image and question embedding. Yang, Zichao, et al. [17] created a three model Stacked Attention Network (SAN) where the attention model takes the image and question features and extracts the most important region in the image. In two-layer image attention map image regions are again merged with the vector of question extracting attention map locating the main regions of image. Kazemi et al [18] used LSTM for question features and image features where features are l2 normalized extracted using ResNet-152. Attention based features of image are extracted by a concatenation of question and image features and passed through multiple convolutional layers to create weighted average of image attention based features. Lu, Jiasen, et al. Hierarchical Question Image Co-Attention (HieCoAtt) model [19] ensures both image and question are taken into scheme by alternating and parallel co-attention where a question at word / phrase level is attended through image and image through question. Teney et al. initiated the use of object detection on VQA models [20] which narrows down the features and apply better attention to images using R-CNN architecture. To model dense inter model interaction multiple blocks of self attention units are used and guided attention units are used for intra model interaction. Yu, Zhou, et al. propose a Deep Modular Co-Attention Networks (MCAN) [21] where Modular Co-Attention (MCA) layer performs the self attention of the questions and image also performs question guided attention of the input images. MCAN is a deep cascade of multiple MCA layers. Bottom up attention visual features concatenated with features from a Faster R-CNN mode are extracted for image. Question features are extracted by passing GloVe word embeddings of question to single layer LSTM. Multi-label classifiers are fed with multimodal features derived from multimodal fusion models to find answers from the image and question.

C. Models using external knowledge base

In Wu, Qi, et al. Ask Me Anything (AMA) [22] model the image vector, the caption vector and the DBpedia KB are combined and provided as input to the LSTM which generates the answers from the questions. To generate attribute based image vectors initially nouns, verbs and adjectives are extracted from the MSCOCO captions. The aggregated caption vector through average pooling of 5 captions each for image was extracted by feeding the LSTM network with the earlier generated image attributes. The top most predicted image attributes are passed as a query to the DBpedia where Doc2Vec is applied to get the semantic KB vector. Wang, Peng, et al. model used ConceptNet, DBpedia and WebChild where important part of the questions is identified and the question is classified into categories which is mapped to database queries over KB. Further to get a precise answer matching score of question and KB extracted facts are calculated. Marino, Kenneth, et al. proposed model [23] retrieves external data from Wikipedia for every question image pair which is then used for training the network to get the answer from articles. The input to the model is image features extracted using ResNet, the question, sentences in the retrieved article and the title of the article. The model is trained to predict where the actual answers appear in the sentence or article and get the score of the sentence and for each word in the sentence. Ma, Chao, et al. model [24] used augmented memory which helps in maintaining long term memory. Co-attention approach is used to attend relevant regions of image and words of questions. In memory augmented network (MAN) concatenated attention based image and question vectors are

passed. In MAN, the time frame is extracted to read or write from the outer memory and is set on by the controller LSTM. To predict the answer the MAN final embedding of the image and question is passed to the classifier.

IV. Model Architectures for Hindi and Marathi VQA

Image and question features are extracted for most of the VQA task, passed to a diversified deep learning based neural network to create a codependent/joint embedding from which response as answer is predicted. In most of the systems, image features are extracted from Convolution neural networks (CNN) and Recurrent Neural Network (RNN) is used to get question features. VGG Net, ImageNet or googleNet is mostly used to generate image features as these models are widely used models for transfer learning which helps in decreasing the training time and also generating much more precise features of the image. For generating questions features, Word2Vec, GloVe and FastText are mostly used to generate word embeddings.

We have tried to create and test different model architectures over Easy Visual Question Answering (easy-VQA) dataset translated to Hindi text and Marathi text from English. A generic architecture is displayed in Figure 1

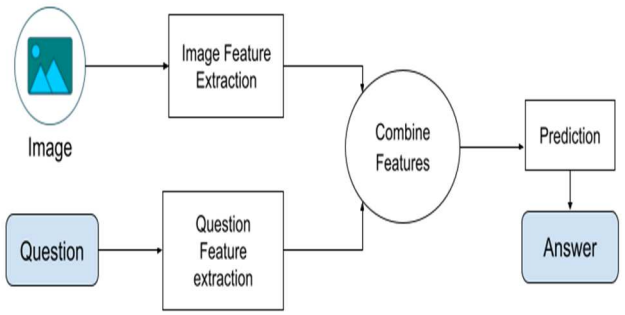

Fig 1. Baseline Visual Question Answering architecture

A VQA system built over Easy Visual Question Answering (easy-VQA) dataset will give the answer “blue” “नीला” in Hindi and “निळा” in Marathi to the question “what color is the triangle?” “त्रिकोण क्या रंग है?” in Hindi and “त्रिकोणाचा रंग कोणता आहे?” in Marathi over the image listed below.

Fig 2. Image from Easy Visual Question Answering (easy-VQA) dataset having blue color triangle

To solve the problem the VQA system looks for shapes and colors in the image where we can use CNN and can be trained to classify whether our image contains a triangle or circle and the color of the shape. In the VQA system both image and question play an important role in getting answers so from the input question, the main task is to understand the question type and other important objects. Here “what color” or “क्या रंग” or “रंग कोणता” in Hindi and Marathi helps us to understand that the answer is related to the color of the shape. As VQA is a classification problem of multi class type where the correct answer will have the largest probability. Here answer matching probability for each answer is generated based on the input features and the largest probability answer is the final choice of answer provided by the system. So, for the question त्रिकोण क्या रंग है? The answer should be the color “Blue” or “नीला” in Hindi as now the model knows that the image has a blue shape and the answer presented to the user should provide color of that shape.

We need to create multiple models for the VQA system which involves a mixture of NLP and computer vision to process text and image to get an exact answer. As the image data was not complex or Easy Visual Question Answering (easy-VQA) dataset we could manage the image feature generation part by creating a simple CNN model without using any transfer learning.

Image model is constructed to extract image features using Convolutional Neural Network (CNN). In CNN an image is passed over multiple Convolutional layers where filters which are applied over the convolution layer generates features from the image which are passed to one or more convolution layers until the final prediction is made. To reduce the parameters, pooling is applied in between the convolution layers and finally a fully connected layer, the input of existing layers are flattened before passing to the output layer. We tried with different numbers of filters to improve the accuracy with limited dataset size.

In the Question Model the question is processed to generate question features. In most of the VQA system question models consist of Recurrent Neural Network (RNN) to generate features from questions. In RNN the input words of the question are related to each other as RNN remembers some information about the sequence of words through its internal memory. In RNN the same operation is done over each input. Long Short Term Memory (LSTM) helps to resolve vanishing gradient problems of RNN and works better for long sequences of data. Through LSTM is able to predict using all past context of the input instead of only the last input which is exhibited by the RNN.

Three gates are present in LSTM input, forget and output gates. Not always the last input will need details of the far way first input for predictions, this is managed by the forget gate which decides what to keep and what to delete in the memory state sequence. Bidirectional LSTMs help to keep past and future information of the input which helps to understand context in a much more efficient way.

LSTM and Bidirectional LSTMs are other alternatives to the Bag of Word Approach, while creating the question model. Attention mechanism helps to focus on the main, essential part of the input which enables efficient learning and better predictions. The RNN encoder decoder architecture encoder provides a single output vector over the input word of the sentence. The output vector is passed as a single input to the decoder which causes loss of information. RNN with attention mechanics placed between the encoder and decoder prevents this loss by generating a sequence of context vectors from the encoder which helps the decoder to use the most important parts of the input and predict better results. Accuracy of different type of image model (CNN, Big CNN) and question model (LSTM,GRU) combined

which are trained over Hindi and Marathi (easy-VQA) dataset is described in Table 1.

TABLE I. ACCURACY OF DIFFERENT MODELS OVER HINDI AND MARATHI (EASY-VQA) DATASET

Model	Accuracy Marathi		Accuracy Hindi	
	Train	*Validation*	*Train*	*Validation*
CNN (Image Model) + Bag of Word (Question Model)	87.19	87.80	85.47	83.18
Big - CNN (Image Model) + Bag of Word (Question Model)	99.57	99.88	99.56	99.96
Big - CNN (Image Model) + LSTM (Question Model)	98.24	97.39	99.27	99.90
Big - CNN (Image Model) + Bidirectional LSTM (Question Model)	99.82	99.39	98.80	99.68
Big - CNN (Image Model) + LSTM with attention (Question Model)	87.18	78.30	86.36	82.91
Big - CNN (Image Model) + Bidirectional LSTM with attention (Question Model)	82.87	80.54	84.65	85.80
Big - CNN (Image Model) + GRU (Question Model)	99.14	99.61	99.27	99.82
Big - CNN (Image Model) + Bidirectional GRU (Question Model)	99.00	99.47	97.72	98.72
Big - CNN (Image Model) + Bidirectional GRU with attention (Question Model)	93.88	95.90	84.60	85.56

GRU (Gated Recurrent Unit) also helps to solve vanishing gradient problems like LSTM. Compared to LSTM here GRU consists of only the update gate and reset gate. Here information is available for a long time. Here the update gate helps specify the amount of information which needs to be kept from the past for future processing and which all past information needs to be forgotten is provided by reset gate.

A simple Question model can be constructed using the Bag of Word Approach, without use of any forms of RNN, where the question text is converted into a vector and fed to the dense network. For a dataset having questions of small length a simple Bag of Word Approach works well to generate question features.

A. CNN (Image Model) + Bag of Word (Question Model)

Image is passed in the image model through a Convolutional layer of 8 filters with a 3x3 kernel size, 1 as stride and padding as same accompanied by Max-Pooling layer and another 16 filters Convolutional layer accompanied by Max-Pooling layer followed by yet another layer of 32 filters. The final layer of the image model consists of a fully connected layer of output dimensions 32 with activation function as tanh. The question model receives a vector generated using the Bag of Word approach followed by two fully connected layers of 32 output dimensions with tanh activation function. The feature vector of image and question are merged together through element wise multiplication followed by fully connected layers of 32 output dimensions with tanh activation function and finally a softmax function used to classify and predict answers.

B. Big - CNN (Image Model) + Bag of Word (Question Model)

Big - CNN (Image Model) has CNN (Image Model) architecture and parameters with extra 64 and 128 filters. The final layer of the image model consists of a fully connected layer of output dimensions 512 with activation function tanh, preceded by 0.5. The question model receives a vector generated using the Bag of Word approach followed by two fully connected layers of 512 output dimensions with tanh activation function. The image and question feature vector are merged together through element wise multiplication followed by fully connected layers of 512 output dimensions with tanh activation function and finally a softmax function used to classify and predict answers.

C. Big - CNN (Image Model) + LSTM (Question Model)

Big - CNN (Image Model) has CNN (Image Model) architecture and parameters with extra 64 and 128 filters. The final layer of the image model consists of a fully connected layer of 512 output dimensions with tanh activation function preceded by 0.5. The question model receives a vector generated using the Bag of Word approach which is passed through LSTM of size 512 followed by a fully connected layer of 512 output dimensions with tanh activation function. The image and question feature vector are merged together through element wise multiplication followed by fully connected layers of 512 output dimensions with tanh activation function and finally a softmax function used to classify and predict answers.

D. Big - CNN (Image Model) + Bidirectional LSTM (Question Model)

Big - CNN (Image Model) has CNN (Image Model) architecture and parameters with extra 64 and 128 filters. The final layer of the image model consists of a fully connected layer of 512 output dimensions with tanh activation function preceded by 0.5. The question model receives a vector generated using the Bag of Word approach which is passed through a Bidirectional LSTM of size 512 followed by a fully connected layer of 512 output dimensions with tanh activation function. The image and question feature vector are merged together through element wise multiplication followed by fully connected layers of

512 output dimensions with tanh activation function and finally a softmax function used to classify and predict answers.

E. Big - CNN (Image Model) + Bidirectional LSTM with attention (Question Model)

Big - CNN (Image Model) has CNN (Image Model) architecture and parameters with extra 64 and 128 filters. The final layer of the image model consists of a fully connected layer of 512 output dimensions with tanh activation function preceded by 0.5. The question model receives a vector generated using the Bag of Word approach which is passed through a Bidirectional LSTM of size 512. After this attention is applied to give the priority to the most important words from the question, which is followed by LSTM of size 512 and fully connected layer of 512 output dimensions with tanh activation function. The image and question feature vector are merged together through element wise multiplication followed by fully connected layers of 512 output dimensions with tanh activation function and finally a softmax function used to classify and predict answers.

F. Big - CNN (Image Model) + GRU (Question Model)

Here we have used the same architecture and parameters as Big - CNN (Image Model) as specified in the earlier model. Here we have made changes in the question model where the LSTM (Question Model) is replaced with GRU and we have used GRU of 512 size.

G. Big - CNN (Image Model) + Bidirectional GRU (Question Model)

Here we have used the same architecture and parameters as Big - CNN (Image Model) as specified in the earlier model. Here we have made changes in the question model where the Bidirectional LSTM (Question Model) is replaced with Bidirectional GRU where we have two GRUs and we have used a GRU of 512 size.

H. Big - CNN (Image Model) + Bidirectional GRU with attention (Question Model)

Here we have used the same architecture and parameters as Big - CNN (Image Model) as specified in the earlier model. Here we have made changes in the question model where the Bidirectional LSTM (Question Model) with attention is replaced with Bidirectional GRU with attention where we have two GRUs and we have used each GRU of 512 size. Through the attention mechanism we have a question model where we focus on important parts of questions.

V. CONCLUSION

Many scientists and researchers have gained interest in solving the visual question answering problem over the past years. VQA dataset was the baseline dataset which provided the researchers enough data to devise solutions to solve the problem. Many different datasets are derived over time with different types of question answer pairs over the image and more real life related information which helps to improve the accuracy of VQA models. Various deep learning models are developed to provide answers by exploring the question and answer features in different forms. Attention based models with knowledge are providing better results over the earlier simple joint embedding approach but still the algorithms can't provide human level accuracy. Development of a bigger dataset with external information for the question can aid the future models to have better results. Also incorporating various NLP modules can help in creating a benchmark VQA model in the future.

REFERENCES

[1] Lin, Tsung-Yi, et al, "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

[2] Malinowski, Mateusz, and Mario Fritz. "A multi-world approach to question answering about real-world scenes based on uncertain input." Advances in neural information processing systems. 2014.

[3] Ren, Mengye, Ryan Kiros, and Richard Zemel, "Image question answering: A visual semantic embedding model and a new dataset," Proc. Advances in Neural Inf. Process. Syst1.2 (2015): 5.

[4] Krishna, Ranjay, et al, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision 123.1 (2017): 32-73.

[5] Zhu, Yuke, et al, "Visual7w: Grounded question answering in images," Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[6] Yu, Licheng, et al, "Visual madlibs: Fill in the blank description generation and question answering," Proceedings of the ieee international conference on computer vision. 2015.

[7] Johnson, Justin, et al, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[8] Kahou, Samira Ebrahimi, et al, "Figureqa: An annotated figure dataset for visual reasoning," arXiv preprint arXiv:1710.07300 (2017).

[9] Andreas, Jacob et al,"Deep Compositional Question Answering with Neural Module Networks." ArXiv abs/1511.02799 (2015).

[10] Goyal, Yash, et al,"Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[11] Wang, Peng, et al, "Fvqa: Fact-based visual question answering." IEEE transactions on pattern analysis and machine intelligence 40.10 (2018): 2413-2427.

[12] https://visualqa.org/external.html Last accessed on June 20.2022

[13] Ren, Mengye, Ryan Kiros, and Richard Zemel, "Exploring models and data for image question answering," Advances in neural information processing systems. (2015)

[14] Zhou, Bolei, et al,"Simple baseline for visual question answering," arXiv preprint arXiv:1512.02167 (2015)

[15] Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han, "Image question answering using convolutional neural network with dynamic parameter prediction," Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).

[16] Jabri, Allan, Armand Joulin, and Laurens Van Der Maaten, "Revisiting visual question answering baselines," European conference on computer vision. Springer, Cham, (2016).

[17] Yang, Zichao, et al, "Stacked attention networks for image question answering," Proceedings of the IEEE conference on computer vision and pattern recognition. (2016).

[18] Kazemi, Vahid, and Ali Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," arXiv preprint arXiv:1704.03162. (2017)

[19] Lu, Jiasen, et al. "Hierarchical question-image co-attention for visual question answering." Advances In Neural Information Processing Systems. (2016).

[20] Teney, Damien, et al, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018).

[21] Yu, Zhou, et al, "Deep Modular Co-Attention Networks for Visual Question Answering," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019).

[22] Wu, Qi, et al, "Ask me anything: Free-form visual question answering based on knowledge from external sources," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016).

[23] Marino, Kenneth, et al, "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019).

[24] Ma, Chao, et al, "Visual question answering with memory-augmented networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018).