2020 International Conference for Emerging Technology (INCET)
Belgaum, India. Jun 5-7, 2020

Vısual questıon answerıng models Evaluatıon

Sarath.S
Department of Computer Science and Engineering,
Amrita School of Engineering, Bengaluru,
Amrita Vishwa Vidyapeetham, India
sarathsreesailam@gmail.com

Amudha.J
Department of Computer Science & Engineering,
Amrita Vishwa Vidyapeetham, Bengaluru.
j_amudha@blr.amrita.edu

Abstract- Visual question answering (VQA), visual dialogs, visual chat bot are multi-discipline exploration problems, which is a blend of Natural Language Processing (NLP), Image feature extraction and Knowledge Reasoning (KR). Rather than captioning, which is naïve approach of computer vision, VQA problems enhances the perspective by providing interactivity to ask domain specific as well as open ended questions to images and give us the insights based on image features or characteristics.Our research is tho evaluate the performance of VQA on counting problems. Given an image, VQA model is expected to answer "how many" question type. We have used few pre-trained models for VQA and visual dialog and tabulated the findings of accuracy of predicted answer with the pre-defined ground truth.

Keywords- VQA , Visual Question answering , Pythia , CNN , LSTM , NLP

I. INTRODUCTION

Visual question answering research problem heads up to asking questions to an image to get meaningful insight based on the image features. There are multiple question types which are asked to an image. Most commonly seen types are 'yes/no', 'object identification', 'counting the objects', 'color identification' etc. From the existing researches and testcase, VQA has performed well in object and color identification and yes/no questions. However, counting number of objects in an image was less explored and existing models found underperform to answer this question type.

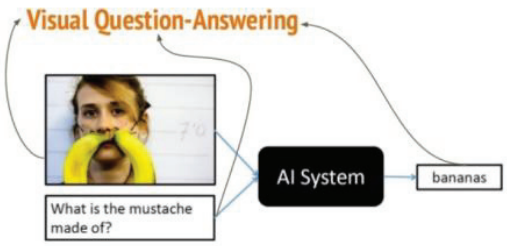


Fig. 1. VQA Model (Source:https://visualqa.org/challenge.html)

For the present study, we are trying to evaluate popular VQA models specifically for counting problems which less explored area is when compared with other question types in VQA domain. We are comparing on simple model based out on CNN-LSTM, Pythia model, Co-attention model and visual dialog model to evaluate their performance on counting problems.

In this paper, we start with introduction of VQA, its use cases and domain specific applications. Literature survey findings are depicted in the next section on the current trends and different models which does VQA. This section will detail on what are the different approaches on model design and best models derived so far for VQA tasks.

II. LITERATURE SURVEY

In Recent times lot of progressions has been published in VQA task. Before advancement of VQA models, there was existing image captioning models which analyses the image features and reproduces a caption which is relevant to the context. Recently numerous researchers have come up with different approaches, where all approaches involve below set up

- Featurization of image
- Feature extraction from question
- An algorithm to combine both image and question feature and generate predictions

Existing literatures are taken into study and . In [3] CNN approach for VQA is considered and this model model uses ResNet to compute image features and ByteNet to compute question embeddings . On top of this a soft attention mechanism is used to focus on particular areas/features in image. In [5], depicts co-attention model which considers both image and question attention.In [13] visual dialog model is bulit on top of VisDial data set which holds an AI agent to collect dialog exchanges between 2 bots. Pythia implementation paper depicts a new architecture which uses res net and inbuilt tokenizer as initial steps. Later hyper parameters are tweeked to obtain optimized result.

III. SYSTEM DESIGN

For this proposed work, we have set up below 4 models to run either from local system or via google colab. We fed the same input image and question to all 4 models and tabulated the results to determine which the best model is suited for counting problems or "how many" question type.

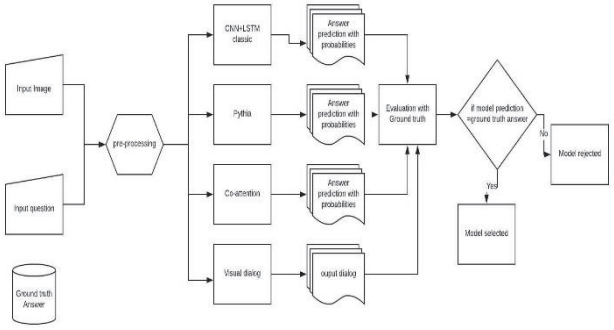


Fig. 2. Flow chart of VQA model evaluations

All VQA model's base design is to take image and question as input. Image featurization and question tokenization will be done via various techniques. Later image and question will be combined in vector space are summation techniques done to obtain the prediction results along with confi-

978-1-7281-6221-8/20/$31.00 ©2020 IEEE

dence . COCO-QA , VQA1.0 , VQA2.0 are the main data sets used for training.

IV. System Implementation

For this research, we have implemented 4 pre-trained models. Each model in detail on the specifications is explained below.

A. CNN-LSTM classic model

This is classic model where language features and image features are calculated separately and joint together and on the combined features, a multi-layer perceptron is trained to obtain the results. Pre-trained VGG16 (VGG net) model is used for image feature extraction by having some modifications by eliminating the last 2 layers.

For word embeddings , we use glove which is Stanford standard for word2vec. LSTM network is used for question feature extraction.

VQA model is simple which it combines image model and question model and runs a multi layer perceptron

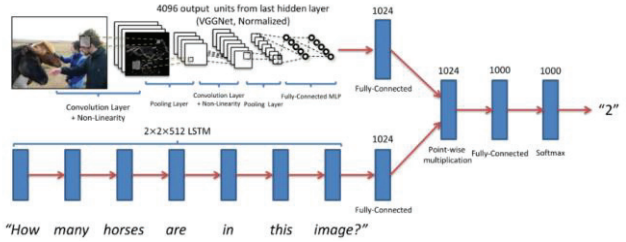


Fig. 3. Classical CNN-LSTM model (Source:https://arxiv.org/pdf/1505.00468v4.pdf)

B. Pythia VQA Model

Pythia which is built on pytorch is a deep learning framework that helps in language and vision domain.This support multi-tasking with modular plug and play support in research and build fast. Pythia is a general-purpose framework where VQA problem can also be addressed.

Pythia VQA model uses resnet 152 model for image feature extraction and pythia's inbuilt word tokenizer is used for question tokenization. Pythia base line model is used for combining both image and question.

Pythia features

- MultiTasking: Provision for multi-tasking which lets training on multiple dataset together.
- Datasets: Comprises support for many datasets built-in including VQA, VizWiz, , VisualDialog ,TextVQA and COCO Captioning.
- Modules: Provides applications for many usually used layers in vision and language area
- Distributed: Support for distributed training based on DataParallel as well as DistributedDataParallel.
- Unopinionated about the dataset and model executions built on top of it.
- Customization: Custom losses, tensorboard, optimizers ,metrics, scheduling, , suits all custom needs.

C. Hierarchical Question-Image Co- Attention model

Co-attention one of the most interesting method used in VQA. General attention models look for visual attention by focus only on the areas of interest in an image. However co-attention focuses on both image and question's areas of interest rather than traversing through all features.

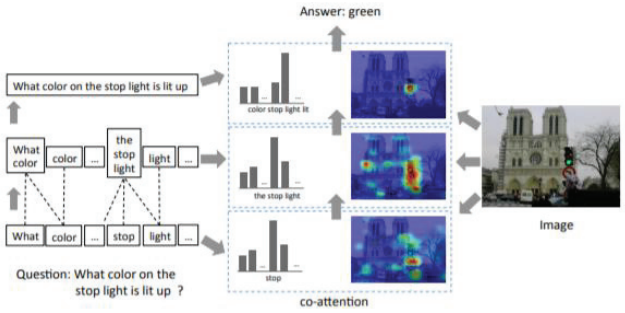


Fig. 4. Hierarchical Question-Image Co-Attention modell(Source:arXiv:1606.00061v5 [cs.CV] 19 Jan 2017)

For a question, extracts its word level, question level and phrase level embeddings. co-attention applied on each level on image as well as question. The final prediction is based on all co-attended question and image features.

D. Visual Dialog Model

Visual dialog is an extended version of VQA which address a sequence of questions given an input image. This uses a novel encoder decoder model for the task which needs an AI agent to hold a dialog like natural human. Precisely, given an image, a question about the image, a dialog history the agent has to ground the question in image, gather context from history, and answer the question correctly.

Fig. 5. Visual Dialog GUI (Source:https://visualdialog.org/)

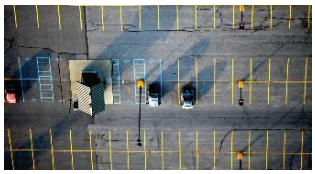
This model consists of 3 encoders and 2 decoders. Late Fusion, Memory Network and Hierarchical Recurrent Encoder are user along with generative(LSTM) and discriminative (softmax) decoders. Since VQA is considered as classification, here 'chopping' the final VQA-answer softmax from the models,then feed these activations to our discriminative decoder and train end-to-end on VisDial

V. System Testing and Testing Cases

We have limited our question type to counting problems alone to understand the capability of VQA models in answering this question type.We have taken 15 images which comprises from different domains but the question asked is to count the number of any specific object present in the image. Here we used pre-trained models for Classic VQA, pythia VQA, co-attention VQA and visual dialog. We have run the model with GPU to gain faster response.

Below are the details of images and the corresponding question asked. Top most 5 predictions are taken as output, but top most prediction is only taken for evaluation.

(1)



(2)

(3)

(4)

(5)

(6)

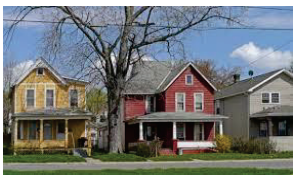
(7)

(8)

(9)

(10)

(11)



(12)

(13)

(14)

(15)

Fig. 6. VQA test images

TABLE I. VQA TEST QUESTIONS AND GROUND TRUTH

image no	Question	Ground truth
1	How many cars are there in this picture	12
2	How many cars are there in this picture	3
3	How many cars are there in this picture	5
4	How many cars are there in this picture	5
5	How many humans are there in this picture	5
6	How many strawberries are there in this picture	3
7	How many monkeys are there in this picture	2
8	how many runners are in this picture	4
9	how many cycles are there in this picture	3
10	how many flowers are in this picture	6
11	how many chairs in this picture	4
12	how many house in this picture	3
13	how many elephants in this picture	3
14	how many clocks in this picture	4
15	how many candle	3

VI. Results and Discussion

We have fed the test images to all the 4 selected models and below is the prediction summary. We have used simple accuracy measure to evaluate the performance of VQA model on counting question type. Even though we have set to display top 5 predictions, top most prediction is only taken for accuracy calculation and if it is matching with the ground

truth, a score of 1 is added for the model else a score of 0 will be given.

At the end, total score is determined to evaluate how good the model performed for counting problems. We have classified the images into below 4 broad categories.

TABLE II. IMAGES GROUPED BASED ON SOURCE OF CAPTURE

Image no	Image Source
1,2,3,4,12	Surveillance Camera
5,9,11	Stage still photography
6,10,14,15	Close up shots
7,8,13	face/candid

TABLE III. CONSOLIDATED RESULTS OF PREDICTION

Image no:	Pythia	HieCoAtt	CNN+LSTM	Visual Dialog
1	No	No	No	No
2	Yes	No	No	No
3	No	No	No	No
4	Yes	No	No	No
5	yes	No	No	No
6	yes	Yes	No	No
7	No	Yes	Yes	No
8	Yes	No	No	No
9	Yes	No	No	No
10	Yes	Yes	No	No
11	No	Yes	No	Yes
12	No	No	No	No
13	Yes	No	No	No
14	Yes	Yes	No	No
15	Yes	No	Yes	No

From the above tabulation and by testing with 15 images, we arrive at below evaluation score

TABLE IV. SIMPLE MODEL ACCURACY COMPARISON

MODEL	accurate predictions	Wrong predictions	Accuracy
CNN+LSTM	2	13	13.33%
PYTHIA	10	5	66.60%
HieCoAtt	5	10	33.33%
Visual Dialog	1	14	6.60%

We arrived at the above accuracy by using 'simple accuracy' method where we considered only the top most prediction and assigned score 1 when system predicted the ground truth. From this research we could see that Pythia based model (Resnet + pythia tokenizer) outperformed all other models in counting question type. Even though existing models don't claim accurate for counting problems, Pythia has significant performance over other models. Co-attention model could possess 34% accuracy where visual dialog turned out to be just 6% accurate.

As per detailed verification, none of the models exhibit significant accuracy for Surveillance Camera captured pictures. Pythia VQA model could able to predict correct answer for all most all close up shots and face/candid images.

We have looked at another perspective where we would consider top 5 predictions and scores will be assigned if the ground truth is predicted as any of the 5 top predictions with a confidence of at least 20%. Below is the result tabulated.

TABLE V. CONSOLIDATED RESULTS OF PREDICTION BASED ON 20% CONFIDENCE

Image no	Pythia	HieCoAtt	CNN+LSTM	Visual Dialog
1	F	F	F	F
2	P	F	F	F
3	F	F	F	F
4	P	F	F	F
5	P	F	F	F
6	P	P	F	F
7	P	P	P	F
8	P	P	F	F
9	P	P	F	F
10	P	F	F	F
11	F	P	F	P
12	F	F	F	F
13	P	P	F	F
14	P	P	F	F
15	P	P	F	F

TABLE VI. MODEL ACCURACY BASED ON DEFINED CONFIDENCE

MODEL	accurate predictions	Wrong predictions	Accuracy
CNN+LSTM	1	14	6.60%
PYTHIA	11	4	73.33%
HieCoAtt	8	7	53.33%
Visual Dialog	1	14	6.60%

Based on above evaluation, Pythia VQA model still outperforms with 73% accuracy than other models. We could see Hierarchical co-attention model has improved drastic performance when considering top 5 prediction list. Classic CNN+LSTM model and visual dialog model could not be considered for counting question types as the accuracy is negligibly low.

VII. CONCLUSION AND FUTURE WORK

VQA has been popular in recent times. Visual question answering is a blend of NLP , Computer vision as well as knowledge reasoning. In our research, we have evaluated the performance of 4 popular VQA models on the counting ("how many") question types. As per our evaluation, none of the VQA models performs outstanding results in counting problems but could see Pythia model outperforms others significantly with 66% accuracy in prediction.

Specifically, all 4 models exhibit poor accuracy for Surveillance Camera captured pictures. Pythia VQA model

could able to predict correct answer for all most all close-up shots and face/candid images.

As a future work, we plan to address this limitation of VQA which is underperforming on numbering questions. Model Training to be done specifically for addressing this context so that pythia model can be improved to have better accuracy in this regard.

REFERENCES

[1] Xu, X., Song, J., Lu, H., He, L., Yang, Y., & Shen, F. (2018). Center for Future Media & School of Computer Science and Engineering University of Electronic Science and Technology of China , China Kyushu Institute of Technology , Japan Qualcomm Technologies , Inc ., San Diego , CA, (16809746).

[2] Shih, K. J., Singh, S., & Hoiem, D. (2016). Where to look: Focus regions for visual question answering. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 4613–4621. https://doi.org/10.1109/CVPR.2016.499

[3] Scholarworks, S., & Koduri, L. A. (2018). A Convolutional Neural Network Based Approach For Visual Question Answering, 9(12), 166–170. https://doi.org/10.1007/s00705-007-1072-4

[4] Masuda, I., de la Puente, S. P., & Giro-i-Nieto, X. (2016). Open-Ended Visual Question-Answering, (June). Retrieved from http://arxiv.org/abs/1610.02692

[5] Lu, J., Yang, J., Batra, D., & Parikh, D. (2016). Hierarchical Question-Image Co-Attention for Visual Question Answering, (c), 1–11. Retrieved from http://arxiv.org/abs/1606.00061

[6] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8693 LNCS(PART 5), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[7] Leonel, E. D. (2009). Phase Transition in Dynamical Systems: Defining Classes of Universality for Two-Dimensional Hamiltonian Mappings via Critical Exponents. Mathematical Problems in Engineering, 2009, 1–22. https://doi.org/10.1155/2009/367921

[8] Kafle, K., & Kanan, C. (2016). Answer-type prediction for visual question answering. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December, 4976–4984. https://doi.org/10.1109/CVPR.2016.538

[9] Sowmya V., P., S. K., and Deepika, J., "Image Classification Using Convolutional Neural Networks", International Journal of Scientific & Engineering Research , vol. 5, no. 6, p. 06/2014, 2014.

[10] R. Suresh and Prakash, P., "Deep learning based image classification on amazon web service", Journal of Advanced Research in Dynamical and Control Systems, vol. 10, pp. 1000-1003, 2018.

[11] K. Radhika, Bindu, K. R., and Parameswaran, L., "A text classification model using convolution neural network and recurrent neural network", International Journal of Pure and Applied Mathematics, vol. 119, pp. 1549-1554, 2018.

[12] Kavikuil K., Amudha J. (2019) Leveraging Deep Learning for Anomaly Detection in Video Surveillance. In: Bapi R., Rao K., Prasad M. (eds) First International Conference on Artificial Intelligence and Cognitive Computing. Advances in Intelligent Systems and Computing, vol 815. Springer, Singapore

[13] Gottimukkala B., Praveen M., Lalita Amruta P., Amudha J. (2019) Semi-automatic Annotation of Images Using Eye Gaze Data (SAIGA). In: Bapi R., Rao K., Prasad M. (eds) First International Conference on Artificial Intelligence and Cognitive Computing. Advances in Intelligent Systems and Computing, vol 815. Springer, Singapore

[14] Amudha J., Chadalawada R.K., Subashini V., Barath Kumar B. (2013) Optimised Computational Visual Attention Model for Robotic Cognition. In: Abraham A., Thampi S. (eds) Intelligent Informatics. Advances in Intelligent Systems and Computing, vol 182. Springer, Berlin, Heidelberg