A CASCADED LONG SHORT-TERM MEMORY (LSTM) DRIVEN GENERIC VISUAL QUESTION ANSWERING (VQA)

Iqbal Chowdhury, Kien Nguyen, Clinton Fookes, Sridha Sridharan

Queensland University of Technology
{m2.chowdhury,k.nguyenthanh, c.fookes, s.sridharan}@qut.edu.au

ABSTRACT

A cascaded long short-term memory (LSTM) architecture with discriminant feature learning is proposed for the task of question answering on real world images. The proposed LSTM architecture jointly learns visual features and parts of speech (POS) tags of question words or tokens. Also, dimensionality of deep visual features is reduced by applying Principal Component Analysis (PCA) technique. In this manner, the proposed question answering model captures the generic pattern of question for a given context of image which is just not constricted within the training dataset. Empirical outcome shows that this kind of approach significantly improves the accuracy. It is believed that this kind of generic learning is a step towards a real-world visual question answering (VQA) system which will perform well for all possible forms of open-ended natural language queries.

Index Terms— Visual Question Answering (VQA), Long Short-term Memory (LSTM), scene understanding.

1. INTRODUCTION

VQA is a task to reason over free-form natural language queries and to answer relevant details from visual contents. Large-scale image dataset e.g. ImageNet [1] and MSCOCO [2] enable deep machine learning models to achieve much higher accuracy in image understanding tasks. However, it is not vividly evident how the models for VQA task will behave for arbitrary questions made on the available datasets which are totally out of model training. Targeting such gap this work proposes a generic approach for the VQA task and experiment it on manually created questions considering the context of relevant image content.

To enable VQA models to answer arbitrary natural language questions, the model needs to learn the generic pattern of question sentence structure. The model needs to learn how to associate predicted generic concept output with specific visual and textual feature. This work suggests to use a compact and discriminant representation of deep visual features. This discriminant representation is obtained by applying PCA [3] to the layers of a Convolutional Neural Network (CNN). Hence the model solves two stages of problems: (1) how to learn the generic structure of question sentence and respective generic output, and (2) how to associate this generic output with specific actual natural language word and discriminant visual features.

This research proposes to deploy a cascaded LSTM architecture. The purpose of cascading is to simultaneously learn the original question and answer sequence along with the respective generic parts of speech tag pattern. Also discriminant visual features are extracted by applying PCA over the outputs of convoutional neural network (CNN) [4] layers. These outputs are made available to both part of LSTM [5] cascade. Through LSTM the model learns to associate sequence of visual features and generic parts of speech (POS) tag with actual word via recurrence.

1.1. Related Works

VQA incorporates techniques from both computer vision, natural language processing and machine learning. Given an image with a natural language question VQA models generate natural language answers. VQA models are consist of three parts namely vision part, a question understanding part and an answer generation part. Task of the vision part is to extract visual features using either deep convolutional neural network or any other low level feature extraction technique. In case of VQA the question to be answered is determined at the run time. Also the required operations to answer that question is also question-specific and determined at run time. This depicts the challenge of general image interpretation. Processing of both visual and textual information is necessary for VQA. VQA is more complex than image captioning because VQA has to use common sense and subject knowledge information which may or may not be directly available in the image.

VQA problem was first studied by Mlinowski et al. [6] which combines semantic parsing and image segmentation. Geman et al. [7] used an automatic query generator which was trained on annotated images. Earlier approaches are limited on the form of questions that can be answered. Deep neural network architecture combined of CNN [4] and recurrent neural network (RNN) has become popular to learn the mapping from images to sentences. Question understanding part encodes question semantics by using either Bag-of-words model or RNN. The answer generation part uses multi-class classifier to create single word answer. Full sentences are generated by an additional RNN decoder. A linear/non-linear joint projection is used to integrate global visual features and dense question embedding.

Combining LSTM for the question sentence, with a CNN for the image features is proposed for VQA task in [8] for the DAQUAR dataset [9]. Single word is preferred as answer in their [8] approach. In this approach, VQA adapt the models as to do image captioning. A single recurrent network is used to perform both encoding and decoding. Attention-based [10] VQA approaches uses question-guided attention maps to find corresponding attention region in the image. This kind of approach may not always capture the overall context of the whole image with reasoning of different types of questions. So, the proposed methodology considers the whole image with whole generic-pattern of question sentence to capture the overall context of reasoning.

Existing VQA research approaches trains on available dataset with fixed range of annotated question-answer pairs; but the proposed methodology demonstrates that, it is possible to learn the generic pattern of natural language query for any given image features. Thus any free form natural language query can be answered by this model. In this way, VQA model is generalized as an open-ended question answering model.

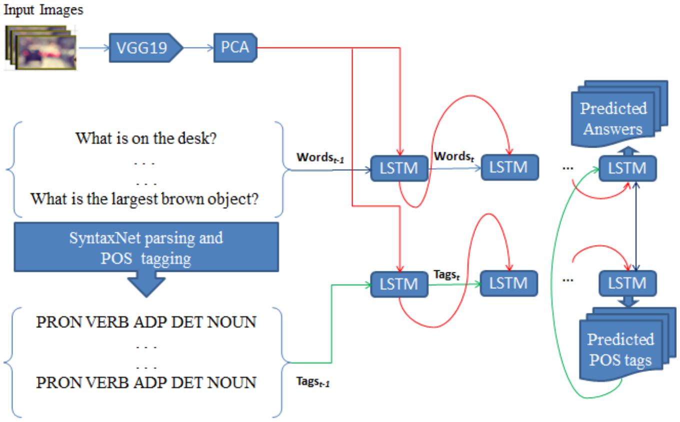


Fig. 1. Proposed VQA framework with cascaded LSTM.

2. METHODOLOGY

The proposed VQA framework is sketched in Figure 1. Input for questions are first tokenized and parsed with respective POS tags. Later question words and POS tags with ground truth answers are trained through LSTM. Predicted tag from LSTM is again fed into another LSTM of the cascade. In this manner LSTM model learns to predict actual answer word by associating respective question, POS tag sequence and deep representation of discriminant visual features.

The VGGNet-19 [11] pretrained model is used for the purpose of deep visual feature extraction from the convolution layers. It is indeed true that an open ended question may involve query about minute details of the image content. Though most of the discriminant parametes are available in the fully connected portion of the convolutional neural network, a significant amount of low level image features are still available in the earlier convolution layers of VGGNet.

$$co\hat{n}v5 = \{conv5_{eig1}, conv5_{eig2}, ..., conv5_{eign}\} \qquad (1)$$

$$f\hat{c}7 = \{fc7_{eig1}, fc7_{eig2}, ..., fc7_{eign}\} \qquad (2)$$

The VGGNet model contains learned kernel weights from the convolutional layer and connection weights from the fully connected layers. In order to preserve and take advantage of both informamtion PCA [3] is applied on convolutional layers and fully connected layers separately. As seen in Equation 1 and Equation 2 PCA is applied on the $\mathbf{conv5}$ features and $\mathbf{fc7}$ features to create training vectors for the cascaded LSTM. $\mathbf{conv5}$ features are chosen because high level visual semantics and features are available in these layers; also $\mathbf{fc7}$ holds class specific information within the object categories of the ImageNet [1] repository. It is expected that taking into account both of these layer information gives the VQA model specialization to answer generic open-ended free form question with specific image content.

Even though enormous types of question can be asked within one available image context, it is infeasible to train a model with all possible kinds of questions. For this reason, the proposed VQA framework includes the parts-of-speech POS tagging of corresponding question words of the empirical dataset. Syntaxnet [12] is used for this purpose. By using POS tag sequences the proposed VQA model will be able to learn the syntactic relationships which are related to the underlying meaning of the question sentence. Syntaxnet [12] uses a globally normalized transition- based feed-forward neural network model. Pre-trained Parsey McParseface [12] is used for this POS tagging task. Word2vec [13] is used to create textual feature vectors of equal length.

A cascaded LSTM architecture simultaneously learns question words with their respective POS tag sequences. Equation 3 to Equation 8 hold the update formulas for LSTM over input sequences. Each LSTM cascade updates at every time step $\mathbf{t}$ for given inputs, $\mathbf{x_t}$ and $\mathbf{h_{t-1}}$. As seen in Figure 1, lower part of the cascaded LSTM deals only with $\mathbf{W_i h_{t-1}^{tag}}$, while the upper part deals with both $\mathbf{W_i h_{t-1}^{word}}$ and $\mathbf{W_i h_{t-1}^{tag}}$; where $\mathbf{W_i}$ stands for respective weight updates. In this way the proposed model learns the structure of questions that could be asked on the respective image contents. Main benefit is that, through this architecture the training becomes more generalized and just not restricted to few possible question available on the training dataset.

$$f_t = \sigma(W_f h_{t-1}^{word} + W_f h_{t-1}^{tag} + W_f x_t + b_f) \quad (3)$$

$$i_t = \sigma(W_i h_{t-1}^{word} + W_i h_{t-1}^{tag} + W_i x_t + b_i) \quad (4)$$

$$g_t = tanh(W_c h_{t-1}^{word} + W_c h_{t-1}^{tag} + W_c x_t + b_g) \quad (5)$$

$$c_t = f_t c_{t-1} + i_t g_t \quad (6)$$

$$o_t^{tag} = \sigma(W_o h_{t-1}^{word} + W_o h_{t-1}^{tag} + W_o x_t + b_o) \quad (7)$$

$$h_t = o_t tanh(c_t) \quad (8)$$

The first LSTM portion learns the sequence of original question words with ground truth answer available in the DAQUAR dataset. The second portion of this cascaded LSTM learns the POS tag sequences for the similar question and answer. Training in this manner allows the model to learn general structure of question that could be asked within the given visual context. Figure 1 shows this LSTM architecture. Visual features of the respective images are fed into this architecture form pre-trained VGGNet(19 layer) [11] model. Word2vec encoding is used to feed fixed length of input to the cascaded LSTMs. Answer word tag at time step t is fed into the respective LSTM of the actual question part. Thus this LSTM model learns how to achieve necessary correspondence between general pattern of tags, actual questions and image features. The proposed LSTM architecture follows similar way as described in [14] for a single portion of LSTM. Later predicted tag is passed to the other portion of LSTM as a linear embedding to predict the corresponding answer word.

Benefit of using cascaded LSTM is that transfer of respective output of one LSTM to another is possible, which allows to simultaneously learn the question word and question tag sequences. To enable cascading the proposed model allows to transfer tag sequence to be passed as input to the latent hidden state. In the same way predicted answer from the LSTM is passed to the hidden state of other LSTM. As seen in Figure 1, output of upper LSTM portion becomes linear embedding of corresponding answer word and tag. It happens the same for the lower LSTM unit. Thus each LSTM learns to make association of visual features, question word and respective tag sequence. In Equation 9, $\mathbf{t}$ denotes the predicted tag by LSTM, $\mathbf{x}$ is the respective image and $\mathbf{q_t}$ is the tagged sequences of the question words which are fed into the LSTM.

$$t = \arg\max \Pr(t \,|\, x, q_t; \theta) \quad (9)$$

VQA approaches on DAQUAR	Acc. of previous methods	Accuracy of the proposed method on DAQUAR	Acc. based on 'Augmented Questions' Only
Neural-Image-QA [8]	19.43	43.05%	37.19%
Multimodal-CNN [17]	23.40		
Attributes-LSTM [18]	24.27		
QAM [10]	25.37		
DMN+ [19]	28.79		
Bayesian [16]	28.96		
DPPnet [20]	28.98		
ACK [21]	29.16		
ACK-S [15]	29.23		
SAN [22]	29.30		

Table 1. Accuracy of VQA methods [23] on DAQUAR dataset

In Equation 10, $\mathbf{I}$ is the respective image features, $\mathbf{t}$ is the predicted tag from the other LSTM part and $\mathbf{q_w}$ are the tokenized question words. The LSTM model is trained to predict the original answer word after it has seen an image, open-ended question and respective POS tag of question words. This formulated in Equation 10.

$$\log \Pr(a \,|\, I, t, q_w) = \sum_{t=0}^{N} \log \Pr(a \,|\, I, t, q_w) \quad (10)$$

3. EXPERIMENTAL RESULTS

Experiment is carried out on the DAQUAR dataset which consists 1449 images with 12468 question-answer pairs. DAQUAR is the first benchmark dataset for the VQA task and experimented by [8], [15], [16] and many other authors. The proposed framework needs accurate tagging of the respective parts of speech (POS) of every question word. For the purpose of tagging Syntaxnet [12] is used which is a state-of-the art parser released by Google. Along with tokenized question words POS tags are also used for training with cascaded LSTM architecture. Figure 2 demonstrates some of the question-answering example of the trained model. The proposed method outperfomrs existing approaches on DAQUAR because those methods are rigidly trained on fixed set of question with visual features. On the other hand, the proposed method trains the VQA model to learn association of general question pattern and visual features, hence it broadens the scope reasoning of the model for any unseen question types with respective visual features.

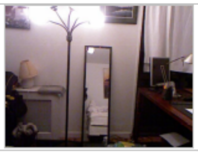
Example of Original and Augmented Questions with Images		**Some Augmented & Original Questions**
		**Original**: how many chairs are on the right side of the table? **Augmented**: what is near the table? **Original**: what is in front of the monitor? **Augmented**: how many monitors are there in the picture? **Original**: what colour is the wall behind the projector screen? **Augmented**: what is behind the projector screen? **Original**: what is in front of the closed door? **Augmented**: what is the colour of the door?
Original question: Which item is red in colour? **Ground truth:** Toaster **Proposed model:** Toaster (right answer)	**Original question:** How many lamps are there? **Ground truth:** 4 **Proposed model:** 2 (wrong answer)	
Augmented question: What is the colour of the toaster? **Ground truth:** Red **Proposed model:** Red (right answer)	**Augmented question:** What is on the table? **Ground truth:** Lamp, monitor **Proposed model:** Lamp (partially right answer)	

Fig. 2. Example of VQA task on DAQUAR and list of manually created *augmented questions*.

The proposed framework is trained on original DAQUAR dataset. Each image is later associated with another manually created open-ended question. These question are created within the object categories provided by the original DAQUAR dataset. These augmented questions consists of 1586 different nouns and all of these nouns are also found in the original dataset. Rightmost side of Figure 2 shows a list of original and manually created questions (Augmented Question) from the DAQUAR dataset. E.g. in the first image, the original question is about the colour of the toaster. And the augmented question is also asked about the same object but with different reasoning. The main fact is that these augmented questions could also be annotated and trained with the model; but numerous combinations of these kind of questions are possible and annotating all of those are infeasible. Thus, the proposed methodology is applied to bolster the generic learning of all possible types of questions.

10-fold cross validation technique is applied to calculate the mearn accuracy of the model. It is evident that the model performs well on the variant of questions which are directly augmented on the content and context of the training dataset. Table 1 and Figure 3 shows accuracy comparison of other methods [23] applied on original DAQUAR dataset with the proposed technique. Again trained model achieved an accuracy of 43.05% on the testing questions of the original DAQUAR dataset. The model is also tested on the augmented set of questions and shows an accuracy of 37.19%.

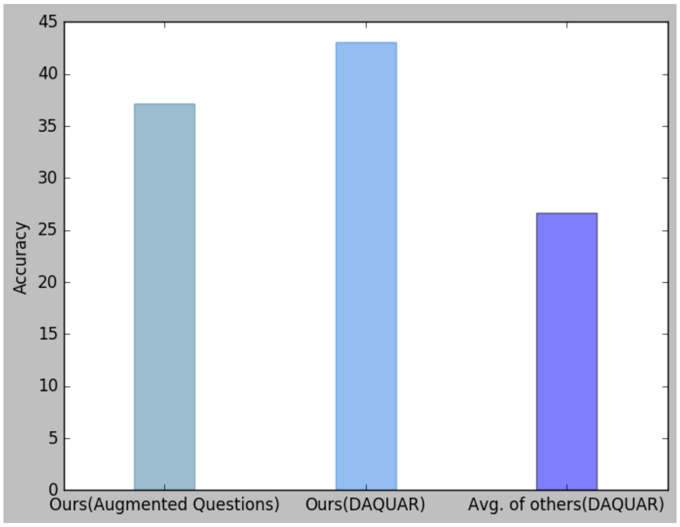


Fig. 3. Performance Comparing Bar Chart

4. CONCLUSION AND FUTURE WORKS

It is infeasible to train a VQA model with all possible types of annotated question-image pairs. A generalization of learning procedure is necessary so that VQA models can achieve the goal of answering any type of free-form natural language question for a given image. State of the art techniques largely depends on training over the fixed form available dateset with explicit annotation of question-image pairs. This paper investigates a way of generalization by learning POS tag of question sentences and deep convolutional features with cascaded LSTM. Experimental outcome demonstrates the feasibility of the proposed technique. In future, other VQA dataset will be augmented with manually created arbitrary question and will be experimented with the proposed technique. Also, cascaded LSTM architecture will be experimented to capture temporal information for video based question-answering.

5. REFERENCES

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[2] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014.

[3] Ian Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.

[4] O. Vinyals J. Hoffman N. Zhang E. Tzeng J. Donahue, Y. Jia and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014, pp. 647–655.

[5] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[6] Mateusz Malinowski and Mario Fritz, "Towards a visual turing challenge," *CoRR*, vol. abs/1410.8027, 2014.

[7] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes, "Visual turing test for computer vision systems," *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015.

[8] M. Rohrbach M. Malinowski and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *ICCV*, 2015, pp. 1–9.

[9] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *NIPS*, 2014, pp. 1682–1690.

[10] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia, "ABC-CNN: an attention based convolutional neural network for visual question answering," *CoRR*, vol. abs/1511.05960, 2015.

[11] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[12] Daniel Andor, Chris Alberti, David Weiss, Aliaksei Severyn, Alessandro Presta, Kuzman Ganchev, Slav Petrov, and Michael Collins, "Globally normalized transition-based neural networks," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016, pp. 2442–2452.

[13] Yoav Goldberg and Omer Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[14] Wojciech Zaremba and Ilya Sutskever, "Learning to execute," *CoRR*, vol. abs/1410.4615, 2014.

[15] Qi Wu, Chunhua Shen, Anton van den Hengel, Peng Wang, and Anthony R. Dick, "Image captioning and visual question answering based on attributes and their related external knowledge," *CoRR*, vol. abs/1603.02814, 2016.

[16] Kushal Kafle and Christopher Kanan, "Answer-type prediction for visual question answering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[17] Lin Ma, Zhengdong Lu, and Hang Li, "Learning to answer questions from image using convolutional neural network," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, 2016, pp. 3567–3573.

[18] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel, "What value do explicit high level concepts have in vision to language problems?," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Caiming Xiong, Stephen Merity, and Richard Socher, "Dynamic memory networks for visual and textual question answering," in *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, 2016, pp. 2397–2406.

[20] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han, "Image question answering using convolutional neural network with dynamic parameter prediction," *CoRR*, vol. abs/1511.05756, 2015.

[21] Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick, "Ask me anything: Free-form visual question answering based on knowledge from external sources.," *CoRR*, vol. abs/1511.06973, 2015.

[22] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola, "Stacked attention networks for image question answering," *CoRR*, vol. abs/1511.02274, 2015.

[23] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel, "Visual question answering: A survey of methods and datasets," *CoRR*, vol. abs/1607.05910, 2016.