Journal of Engineering and Applied Sciences 14 (2): 610-614, 2019
ISSN: 1816-949X


# A Novel Method for Image Captioning Based on Attributes and External Knowledge

[1]Maram Adil Ali Alaziz and [2]Suhaam Adnan Abdul Kareem
[1]General Directorate of Education in Basrah, Ministry of Education, Baghdad, Iraq
[2]Department of Graduate Studies, University of Baghdad, Baghdad, Iraq

**Abstract:** Modern development in visualization-to-verbal communication troubles have been attained through a combination of deep neural networks and Recurrent Neural Networks (RNNs). This mechanism is employed to combine in exterior skill which is seriously significant to answer advanced visual questions. A visual question answering model is calculated which merge an interior depiction of the subject matter of an picture with data obtained from a common skill foundation to respond a wide point of picture-based queries. It especially, permits questions to be raised where the picture only does not have the data necessary to pick the suitable respond.

**Key words:** Visual question, image, recurrent neural networks, deep neural network, RNNs, queries

## INTRODUCTION

Visualization-to-verbal communication troubles donate a specific challenge in PC visualization, since, they need transformation among twofold unlike variety of data. Here, in logic the trouble is alike to that of machinery transformation among verbal communications. In machinery, words transformation present have been a sequence of outcome displaying that fine quality functioning may be accomplished with no increasing a superior-point model of the position of the globe. Here, Sutskever *et al*. (2014), Simonyan and Zisserman (2014) and Cho *et al*. (2014) in spite of the presumed equality among a picture. Person verbal communication is planned fully, so that, convey data amongst persons where as similar the most cautiously created picture is the end of a compound set of material methods more which persons hold less monitor. Known the inequalities among these two kind of informations, it appears amazing that techniques cheered through machinery words conversion have been so, victorious. These RNN-based techniques which interpret openly from picture attributes to content, with no increasing a advanced model of level of the planet, symbolize the present position of the skill.

## MATERIALS AND METHODS

**Image captioning:** This problem of translating the pictures with normal words on the prospect stage have long been learned at together PC visualization then usual speech processing, Mao *et al*. (2014) planned to enfold sentence-supported picture explanation as the mission of position a known set of captions. Likewise, Szegedy *et al*. (2015), Krizhevsky *et al*. (2012) and Vinyals *et al*. (2014) pretensed the challenge as a recovery trouble, although, supported on coimplantingof pictures and content into the similar universe. Lately, Krizhevsky *et al*. (2012) utilize neural networks to co-insert picture and sentences mutually and Bahdanau *et al*. (2015) cofixed picture produces and subsentences. Characteristics contain employed to various image captioning systems to fulfil the cracks in pre ascertained caption patterns. For example, employed discoveries to deduce a triple of view components which is changed to content utilizing created a pattern. LeCun *et al*. (1998) and Krizhevsky *et al*. (2012) created picture depictions arranged PC visualization based enters for instance perceived things, modifiers with positions utilizing network-dimension n-Grams. Zhu *et al*. (Chen and Lawrence, 2015) changed picture word parsing outcomes keen on a semantics demonstration in the appearance of net physiology verbal communication which is renewed to person understandable content. A additional difficult CRF based technique utilize of feature discoveries further than triples (Donahue *et al*., 2015). Confront with a shift toward namely as lot as it utilizes a diagram model finding course in the presence of producing judgments. They preliminary cultured 1000 in reliant detectors pro visualize lexis based on a manifold case to knowing frame plan and functional a utmost entropy thoughts design trained on the firm of imaginably distinguished lexis straight forwardly to produce headers. Here, disparity toward the a for mentioned twofold-step approaches, the up to date leading style in V two L is to make apply of a structural

**Corresponding Author:** Maram Adil Ali Alaziz, General Directorate of Education in Basrah, Ministry of Education, Baghdad, Iraq

design which joins a DNN to an RNN to gather the plan as of imagery toward sentences straight. By Mao *et al.* (2014) for case, planned a manifold modal RNN (m-RNN) to deduce the probability allocation of consequent utterance given preceding utterances and the depth CNN characteristic of a picture at each time step. Alike, built a combined manifold modal embedding place utilizing influential deep DNN Model and an LSTM that take to mean content. Karpathy *et al.* (2014) too recommended a manifold modal RNN creative design, although, in compare to Mao *et al.* (2014), their RNN is hardened at the picture data lone at the initial clock step. Vinyals *et al.* (2015) united depth CNNs used for picture classifying through a LSTM for series styling, to generate a solo set of connections that creates depictions of images. Chen and Lawrence (2015) gather a bi-way charting among pictures which permits to rebuild image attributes given an figure depiction. Yao *et al.* (2015) planned a design founded on image notice. Chen and Lawrence (2015) functional extra regained sentences to lead the LSTM in producing descriptions. Fascinatingly, this terminate-to-end DNN-RNN methodology pay no attention to the figure-to-utterance charting which was an necessary pace in various of the prior figure captioning methods notify above (Szegedy *et al.*, 2015; Krizhevsky *et al.*, 2012; Chen and Lawrence, 2015; Donahue *et al.*, 2015). The DNN-RNN methodology has the merits that is to say up to produce a wider type of captions may be educated terminate-to-end and outputs the earlier tactic on the yardsticks. This is not clear, though what is the collision of by transitory the intermediary advanced representation is and mainly to what level design could be reimbursing. Donahue *et al.* (2015) and Devlin *et al.* (2015) showed an experimentation for instance, utilizing label and CRF designs as a middle coating illustration for videotape to create metaphors, except it was planned outputs an SMT-based methodology (Chen and Lawrence, 2015). It remainder unsure which the middle coating description or the LSTM directs to the victory. Belonging to us study present many well planned experimentations to reply that question. We thus, here, demonstrate not lone a technique to introduce a advanced illustration into the CNN-RNN structure and that doing so get better occurrence except, we also look into the worth of high-level data further generally in V2L assignments. Dangerous significance at this clock because V2L has a extended route to go, mainly in the generalization of the picture s and text it is appropriate to visual question answering (Malinowski *et al.*, 2015) can be the initial to review the VQA complication. They projected a system that group essemantic parsing and picture apportionment with a Bayesian shift towardto exampling from nearby fellow citizens in the coaching''' set. Tu *et al.* built a quiz responding technique based on a combined parsegraph from content and videotapes. Geman *et al.* planned a recaped 'inquiry originator' that is coached on explained images and constructs a series of two fold querions from any prearranged trial figure. Each of these techniques locations noteworthy boundaries on the type of query that can be responded. Most freshly instinctive by the noteworthy progressachieved applying depth neural network exemplars in mutually computervision and natural language processing an architecture which join togethers a DNN and RNN to study the charting from metaphors to sentences has become the foremost trend. Both Gao *et al.* (2015) and Malinowski *et al.* (2015) applied RNNs to command the query and yield the respond. Whereas, Gao *et al.* (2015) applied two fold networks, a detach cryptography and decipher, pre-owned a solitary set of contacts for mutually crypto graphing and decoding. Karpathy *et al.* (2014) and Malinowski *et al.* (2015) paying attention on queries with a solo utterance answer and planned the assignment as a categorization complexity utilizing an LSTM. Antol *et al.* (2015) suggested a larger scale open terminated VQA information set supported on COCO which is called VQA. Motivated by Xu *et al.* who cipher image attention in the image captioning, suggest to use the dimensional heed to help responding image queries. Devise the VQA as a categorization issue and limit the answeronly can be drawn from a predetermined respond space. Our framework also, create use of both DNN, RNNs but in contrast to prior methods which use only image attributes extracted from a DNN in replying a question, we use manifold foundation, generated image captions and mined outside knowledge, to offer for an RNN to answer questions. DNN architecture is displays in Fig. 1, larger-scale Knowledge Bases (KBs) such as independent base and DB pedia have been employed victoriously in a lot of normal language Question Answering (QA) systems. However, VQA systems utilizing KBs are even nowrelatively infrequent.

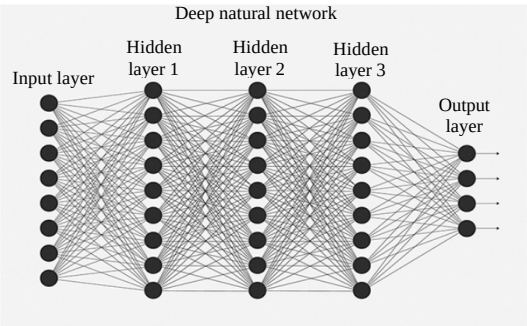


Fig. 1: Depth neural network

## RESULTS AND DISCUSSION

**Using attributes for image captioning:** Our model holds a picture evaluate section and a caption design section. In the picture evaluate section, we primarily make utilize of superintended training to forecast a group of characteristics, found on words usually establish in image captions. We reply this as a manifold tag cataloging issue and instruct a correlating DNN by reduceing a component wise organization defeat task. Further more, a immobile dimension vector Vatt (I) is formed for every one Image I whose dimension is the size of the characteristic group. Every measurement of the vector have the prognosis possibility for a specific characteristic. In the captioning creation section.

**Top 3 attributes:** Wild mount gorilla, photographer and camera.

**Generated captions:**

- A wildlife photographer enclosed by wild mountain gorilla
- A wild life photographer keeping a camera
- A wild mountain gorilla calming following the photographer
- A little one gorilla playing with the photographer

**Top 3 attributes:**

- Girl, horse, polo
- Girl playing polo
- Girl wavering polo bat
- The girl traveling on a trained polo horse

**Top 4 attributes:**

- Baby, play home, playthings generated captions
- A baby in a play household
- A baby be seated on green matress
- A toy dangles upon from the play household

Figure 2-4 explain some instances of the predicted attribute sand created captions.

**A VQA Model with external knowledge:** Discriminator of our VQA replica is that it is capable to expediently merge figure data with that haul out from a tradition base, within the LSTM substructure. The originality lies in the information that this is executed by epitomizing together of these disparity sorts of that as wording ahead inclusion them. Disposed an image, a quality based depiction Vatt (I) is primary engendered. The secondary enter source are those captions produced in this study. We supplementally accomplished a query carried wisdom

Fig. 2: Predicted attribute of tourist

Fig. 3: Predicted attribute of polo player

Fig. 4: Predicted attributes of playhouse

assortment system to rule out the din data, since, we noticed that a few mined tradition are not required for responsive the certain query. For instance, if the queryis talking about the 'Fog' in the appearance, we primary utilize our pre practiced Doc2Vec replica to get out the phonological feature V(Q) of the query and the

characteristic $V(K_i)$ for eachone alone tradition passages where i 2 n. Next, we locate the k nighest wisdom passages to the inquiry based on the cosine synonymity amid the V(Q) and $V(K_i)$.

## CONCLUSION

In this manuscript, we primarily scrutinized the significance of introducing an intermediary characteristic forecast coat into the preponderant DNN -LSTM structure which was abandoned by almost all earlier chore. The tradition bases which are presently accessible do not have mucho of the data which would be useful to this procedure but nevertheless may immobile be utilized to considerably develop functioning on queries needing exterior wisdom (like 'Why' queries). The method that we suggest is very generic, yet and will be appropriate to much more communicative tradition bases ought they become accessible. We further executed a tradition collection schematic which mirrors twain of the contentment of the query and the appearance in demand to take out extra exactly correlated data. Subord more chore contains engendering tradition base inquiries which mirror the contentment of the inquiry and the figure in demand to remove further exactly correlated data. The tradition base on its own also can be get better. For example, open IE affords many wide-ranging commonsense tradition such as 'dogs eat fish'. Such tradition will assist respond advanced enquiries.

## REFERENCES


Antol, S., A. Agrawal, J. Lu, M. Mitchell and D. Batra *et al.*, 2015. VQA: Visual question answering. Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, IEEE, Santiago, Chile, ISBN:978-1-4673-8390-5, pp: 2425-2433.

Bahdanau, D., K. Cho and Y. Bengio, 2015. Neural machine translation by jointly learning to align and translate. Proceedings of the 2015 International Conference on Learning Representations, May 7-9, 2015, CBLS, San Diego, California, USA., pp: 1-9.

Chen, X. and Z.C. Lawrence, 2015. Mind's eye: A recurrent visual representation for image caption generation. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, IEEE, Boston, Massachusetts, USA., ISBN:978-1-4673-6964-0, pp: 2422-2431.

Cho, K., B. Van Merrienboer, C. Gulcehre, D. Bahdanau and F. Bougares *et al.*, 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. J. Comput. Lang., 1: 1-15.

Devlin, J., H. Cheng, H. Fang, S. Gupta and L. Deng *et al.*, 2015. Language models for image captioning: The quirks and what works. J. Comput. Lang., 1: 1-6.

Donahue, J., H.L. Anne, S. Guadarrama, M. Rohrbach and S. Venugopalan *et al.*, 2015. Long-term recurrent convolutional networks for visual recognition and description. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR 2015), June 7-12, 2015, IEEE, New York, USA., pp: 2625-2634.

Gao, H., J. Mao, J. Zhou, Z. Huang and L. Wang *et al.*, 2015. Are you Talking to a Machine? Dataset and Methods for Multilingual Image Question. In: Advances in Neural Information Processing Systems, Cortes, C., N.D. Lawrence, D.D. Lee, M. Sugiyama and R. Garnett (Eds.). Curran Associates, Inc., Red Hook, USA., pp: 2296-2304.

Karpathy, A., A. Joulin and L.F. Fei-Fei, 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. In: Advances in Neural Information Processing Systems, Ghahramani, Z., M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger (Eds.). Curran Associates, New York, USA., pp: 1889-1897.

Krizhevsky, A., I. Sutskever and G.E. Hinton, 2012. Image net classification with deep convolutional neural networks. Proc. Neural Inf. Process. Syst., 1: 1097-1105.

LeCun, Y., L. Bottou, Y. Bengio and P. Haffner, 1998. Gradient-based learning applied to document recognition. Proc. IEEE, 86: 2278-2324.

Malinowski, M., M. Rohrbach and M. Fritz, 2015. Ask your neurons: A neural-based approach to answering questions about images. Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, IEEE, Santiago, Chile, ISBN:978-1-4673-8390-5, pp: 1-9.

Mao, J., W. Xu, Y. Yang, J. Wang and Z. Huang *et al.*, 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). Proceedings of the International Conference on Learning Representations, June 11, 2015, ICLR, New Orleans, Louisiana, USA., pp: 1-17.

Simonyan, K. and A. Zisserman, 2014. Very deep convolutional networks for large-scale image recognition. J. Comput. Vision Pattern Recognit., 1: 1-14.

Sutskever, I., O. Vinyals and Q.V. Le, 2014. Sequence to Sequence Learning with Neural Networks. In: Advances in Neural Information Processing Systems, Ghahramani, Z., M. Welling, C. Cortes, N.D. Lawrence and K.Q. Weinberger (Eds.). Curran Associates, Inc., Red Hook, New York, USA., pp: 3104-3112.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet and S. Reed *et al*., 2015. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA., pp: 1-9.

Vinyals, O., A. Toshev, S. Bengio and D. Erhan, 2015. Show and tell: A neural image caption generator. Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, IEEE, Boston, Massachusetts, USA., ISBN:978-1-4673-6963-3, pp: 3156-3164.

Yao, L., A. Torabi, K. Cho, N. Ballas and C. Pal *et al*., 2015. Describing videos by exploiting temporal structure. Proceedings of the IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, IEEE, Santiago, Chile, ISBN:978-1-4673-8390-5, pp: 4507-4515.