

An Efficient Hybrid Architecture for Visual Behavior Recognition using Convolutional Neural Network

Md Javedul Ferdous
Department of Computer
Science
American International
University-Bangladesh
Dhaka, Bangladesh

A. F. M. Saifuddin
Saif
Department of Computer
Science
American International
University – Bangladesh
Dhaka, Bangladesh

Dip Nandi
Department of Computer
Science
American International
University – Bangladesh
Dhaka, Bangladesh

Mashiour Rahman
Department of Computer
Science
American International
University – Bangladesh
Dhaka, Bangladesh

ABSTRACT

The purpose of this research work is to understand of visual behavior from image. Since computer vision is hugely potential research area for researcher, connecting image captioning and detection of an object, visual behavior detection started to fasten researchers' consideration because of its descriptive power and clear structure in terms of accuracy. By the progress of Deep Learning, giving the computer a chance to comprehend an image is by all accounts progressively closer. With the analysis on object recognition slowly getting to develop progressively more scientists put their consideration on more elevated amount comprehension of the scene. Object detection, visual context is now more consideration in scene understanding as a middle stage. The goal of the research is to discover visual relationships in a given image between objects and understand the whole scenario. This research presents a framework to this problem. Proposed approach performs object detection by using convolutional neural network. This research focus on relationships that can be generated by long short term memory (LSTM). The focus was to design the framework to adopt the Convolutional Neural network with LSTM architecture. Proposed framework is validated using COCO dataset and achieved a BLEU-4 of 23.5 shows better efficiency than previous research methods.

Keywords

CNN; Deep learning; LSTM; Object detection; Scene graph; Visual behavior.

1. INTRODUCTION

An image, not only collection of pixels but also expresses key identical characteristics as well as moments. Behind every image there is a countless expression. For instance, by using only two objects 'person' and 'cycle' can make many different scenario shown in figure 1. It could be "a person carrying a cycle", "a person falling off from cycle", "a person pushing a cycle", "a person riding a cycle", "a person standing beside a cycle", "a person walking with a cycle" and so many on. As a human, it is easy to understand the context of specific scenario by eyes. But, it is difficult for a machine to understand context like the way a human does. The recent success of deep learning-based recognition models has surged interest in examining the detailed structures of a visual scene [1]. In computer vision, Understanding the relation of language to its visual personification remains a testing and crucial issue. Both content and image corpus offer generous measures of data about our physical world.



Fig 1: Different action of a person with a cycle

Relating the data between these areas may progress applications in both and prompt new applications. For example, image look is ordinarily performed utilizing content as input. It may move towards additional descriptive and natural sentence-based picture seek since fields of computer vision and Natural Language processing (NLP) has advanced. Having the capacity to consequently define the context of an image by suitably formed English sentences is an enormously problematic task. In fact, a description must take not just the objects contained in an image, but it additionally should express how these items identify with each different and their qualities and their exercises they are associated with. Besides, the above semantic learning must be communicated in a characteristic dialect like English, which implies that a dialect demonstrate is required notwithstanding visual comprehension. The aim of this research is to understand the behavior of visual context. The major challenge could be solving the uncertain anomalies and unexpected behavior. Recognize an object and find the relation of this object in corresponding other object will be key contribution of this research.

2. BACKGROUND STUDY

Predicting visual relationship is not new in the era of artificial intelligence which is links to computer vision and natural language processing. There has been a successions of work correlated to improving the behavior describe occurred. Previous methods was focused on generating annotations (i.e., nouns and adjectives) from images [2], [3], then generate a sentence from the annotations by Gupta [4]. In 2015, Donahue [5] developed a recurrent convolutional architecture which is appropriate for large-scale visual learning, and demonstrated the value of the models on three different tasks. Lu [1] introduced a model that use object with predicates and train visual models to predict countless relationships per image. Although it can model can measure to predict numerous of forms of relationships, but describing like a human does is still a hard unsolved task. Text to Image has also been developed in 2017. Reed [6] has proposed a novel deep architecture and Generative Adversarial Network formulation in text and image modelling, interpreting visual description from character to pixel. Using a phase of triplet (subject, predicate, object) is key formulate inter connected problem, proposed by Yikang Li, Ouyang [7]. They proposed visual

phase convolutional network, although it shows that it still cannot exceed the broadly used multi-class label targets. While executing iterative message passing among the primal and dual sub-graph alongside the physical structure of a scene graph, [8] proposed a model to make better predictions on objects and their relationships. The outcome show that their significantly performed well than previously methods on generating scene behavior. Although it was quite a development, it only experimented in indoor only. Introducing encoder-decoder to generate sentence is seen on Xu [9] work. Learning action and retrieve data by a models with a large number of actions which are linked to each other is tackled is stimulating in a practical setting, according to Ramanathan [10]. These model combined language cues, visual cues and logical consistency to estimate these action relationships [11]. However, they used some predefined semantic graph which a human never does. By statistical patterns of object co-occurrence and spatial layout can be used as understanding visual context [12], [13], [14], [15], [16], [17]. Achieving comparable to state of the art performance, and generate highly descriptive captions by Using CNN and RNN that can potentially greatly improve the lives of visually impaired people. This model was proposed by Elamri and Planque [18]. However, RNN is difficult to train them to learn continuing dynamics and has consist of problem of vanishing and exploding gradient. To analyze modify an image captioning and decomposed the method to CNN, RNN and sentence segmentation, has achieved comparable results with more complicated LSTM model [19]. By directly mapping, turn an images or sentences into a common embedding space, works on both embeds fragments of images (objects) and a finer level and fragments of sentences into a common space can be a way solution. A bidirectional retrieval model is proposed by Karpathy, Joulin and Li [20]. It decay images and sentences into fragments and understand their inter-modal alignment using a ranking objective. However, it need to remove all relation type that less than 1% due to over fitting. While particular object class for each region to match with the relation ground truth is more consistent to judge [21], but faster RCNN requires discrete grid split for the proposal region. Noise can be a significant issue in the model if there are not sufficient data are available [22]. SVM classifiers has some limitation, which could leads to fail the model.

3. PROPOSED RESEARCH METHODOLOGY

A visual behavior recognition is a structured outline of an image, where each particular quota is bounding boxes with their object groups, edges correspond to their pairwise relationships between objects. Objects involving people, stuff.

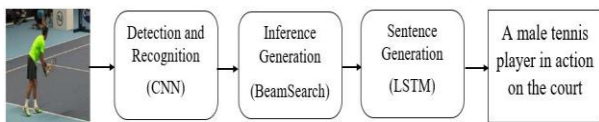


Fig 2: Proposed architecture

Scene etc. With a primary complemented set of images based on global descriptor similarity. It need to comparable to rearrange the selected sentence by combining estimations of image content. Our worked has majority over current state-of-the-art methods, specifically, for out-of-domain images. In order to solve, it needs to present individually main advance of the proposed framework in detail in the following subsection:

3.1 Detection and Recognition

Nowadays, object detection methods have improved remarkably, not only it shows practical performance for a small number of groups of objects but also for a mid-level demonstration for scene recognition. Images. However, it still produces quite noisy results, usually in the form of a large number of false positive detections for consecutively detectors on general web.

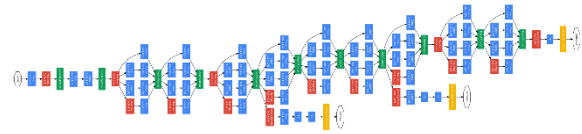


Fig 3: Inception V3 Architecture for detection and recognition.

This becomes even more of a difficulty to content prediction because of the number of object detectors grows. However, it is understandable that if it have some earlier knowledge about the content of an image, then it can utilize even these imperfect detectors. In this work for detection and recognition, it will use the pre-trained tensorflow [23] with Inception-v3 (Figure. 3) as the convolutional neural network (CNN) for visual feature extraction which was tested on MSCOCO data set with over 1 million iteration.

3.2 Inference Generation

There are numerous methodologies that can be used to produce a sentence from an image, but it will be viable to follow with Vinyals [24] work. Here, it sampling first one where it just trial the first word permitting to p_1 , which arrange for the resultant embedding's input and sample, persistent like this until it get one of those: distinct end-of-sentence token or highest length. The second one is Beam Search; it is continuously reflect k (the set of the best sentences) with t (time) as candidates of size t + 1 to generate sentences, and retain only best k result of them. This approach write as follow,

$$S = \operatorname{argmax}_{S'} P(S'|I)$$

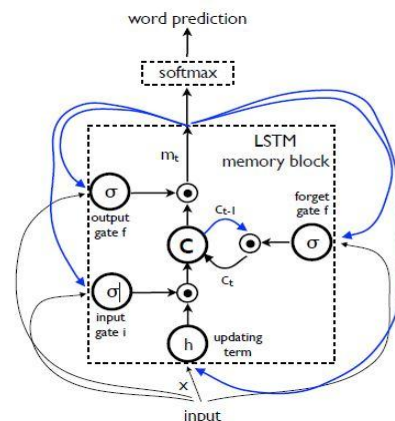


Fig 4: LSTM Block

It is understandable that to get efficient performance, beam size is limited to 20 in the following experiments.

3.3 Sentence Generator

It is one the major challenge on this research. The produced sentences are projected supposed to be further constructive than previous generated by other obtainable research in terms of attribute and labeling details. It needs to use Vinyals [24] LSTM which has shown remarkable performance on sequence jobs such as machine translation. Vanishing and exploding gradients problems are adversary here. To deal with, it needs to use LSTM. The mechanism of LSTM model is consist of a memory cell c which encoding knowledge at every time step of what inputs have been detected to this step (see Fig. 4). “Gates” – layers measured the behavior which are applied multiplicatively. In specific, the purpose use of three gates is to control current cell value so that it never forget and to output the new cell value (output gate o). The equation of the gates and cell update and output as follows:

$$\begin{aligned}
 i_t &= \sigma(W_{ix}x_t + W_{im}x_{t-1}) \\
 f_t &= \sigma(W_{fx}x_t + W_{fm}x_{t-1}) \\
 o_t &= \sigma(W_{ox}x_t + W_{om}x_{t-1}) \\
 c_t &= f_t \odot c_{t-1} + i_t \odot h(W_{cx}x_t + W_{cm}x_{t-1}) \\
 m_t &= o_t \odot c_t \\
 p_{t-1} &= \text{Softmax}(m_t)
 \end{aligned}$$

Where \odot denotes the product of a gate value in addition to the trained parameters represent various w . This multiplicative gates make it possible as it discerning the exploding and vanishing gradients problems by deal with it and train the LSTM robustly. The nonlinearities are sigmoid $\sigma(\cdot)$ and hyperbolic tangent $h(\cdot)$ the last equation is what is used to feed to a Softmax, is to use to determine probability distribution p_t over all words.

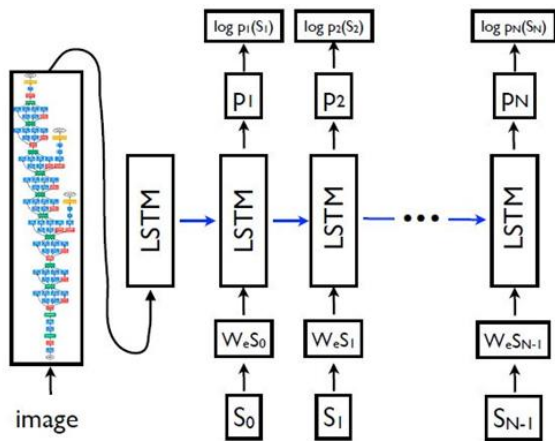


Fig 5: LSTM model joined with a CNN image embedder for sentence Generation

To predict each word of the sentence, LSTM model is used to preceding all words as defined by

$$i_t = p(S_t | I, S_0, \dots, S_{t-1}).$$

For this purpose, it is understandable that, the LSTM in unrolled form – a replica of the LSTM memory by creating the image and it shares the same parameters as well as the output m_{t-1} of the LSTM at time $t - 1$ is fed to the LSTM at

time t for each sentence word such that in Figure 5. All connections that recurrent are changed to feed-forward connections in the unrolled version. In additional point, if it symbolize by I . the image that input and by

$$S = (S_0, \dots, S_n)$$

A true sentence generating this image, the unrolling technique reads:

$$x_{-1} = CNN(I)$$

$$x_t = W_e S_t, t \in \{0 \dots N - 1\}$$

$$p_{t+1} = LSTM(x_t), t \in \{0 \dots N - 1\}$$

Individual word used as a one-hot vector S_t of aspect equivalent to the dictionary size. Here, S_0 is indicated as a distinct ‘start’ word and by S_N a distinct ‘stop’ word where its purpose is the labels the initial and ending of the sentence. In specific by avoiding the ‘stop’ word, the LSTM signals generate a comprehensive sentence. To the same space, for both the image and the words are mapped; vision CNN is used for the image and word embedding W_e for the words. At $t = -1$, image I takes the input once to acknowledge the LSTM about the image subjects. by means of observation, it confirmed that the image is serving at respectively by time step with an further input produces substandard results as well as the network can clearly exploit which leads to reduce noise and over fits in the image more easily.

4. EXPERIMENTAL RESULT AND DISCUSSION

4.1 Experimental Set Up

The fundamental goal of this work is quest of visual intelligent that can interpret an image since it is much more difficult for a computer. Since it used pre-trained model is data driven which was trained by MSCOCO. It wanted to see whether it can able cope up with LSTM model efficiently. The preparation to setup and performed the experiment to get fair and precise result from the proposed framework. Therefore, it load pre-trained model and vocabulary. After that, it sent those data into sentence generator which will able to generate sentence by the corresponding objects.

4.2 Experimental Result

To test the projected framework, it need to compare the result from LSTM based sentence generator. Using respective dataset i.e. flicker8k dataset with annotation. The subset of flicker8k is split into 6091 training images with 1000 images for validation and 1000 images test, respectively used for evaluation of prediction. Some of the result are shown in Figure 6.

The output are matched with the reference sentences linked with specific image of the generated sentences, the evaluation database approved through the MSCOCO evaluation script. This script uses all the necessary metrics and enables comparison between themselves. It will more difficult to work, if each image has only one corresponding sentence, which makes the sentence generation challenging task for all methods to score well with lack of variation.

The CNN + LSTM structure used as skeleton for this visual behavior detection research. While the proposed system services a local based approach worked for both local object

detection and recognition and associated attribute prediction to notify successive encoder–decoder based description generation. By means of observation results indicate that the sentences generated by our system are more descriptive than other with such an environment. Relatively, all baseline approaches generate smaller descriptions using their particular approaches.





| | Image | Result |
|-------------|---|---|
| Good Result |  | A Cat sitting on a top of a window sill |
| |  | A male tennis player in action on the court |
| Bad Result |  | A man in suit and tie on a chair sitting |
| |  | A man lying on a bed with a child |

Fig 6: A selection of results

4.3 Analysis and Discussion

It need to check comparison with other previous experiment such as Show, Attend and Tell [9], NIC [24], Neural Talk [25] and Adaptive Attention [26], to assess the efficiency of the overall system for visual behavior detection. For that, first present the evaluation metrics, applied in this work.

4.3.1 Evaluation Metrics

BLEU [27], ROUGE [28], and CIDER [29] are major popular metrics for sentence generation based evaluation. Those methods use a comparison based portion between what

machine generated and what ground truth supposed to be for a sentences. It needs to present each of these evaluation methods in the following sub-sections in a table 1.

4.3.1.1 BLEU

BLEU is one of the widely used metrics for machine translation. It is used for to determine the similarity among the sentences. The BLEU score 0 to 1. The more its score closer to 1, the more it gives accurate result on a context from a sentence.

4.3.1.2 ROUGE

Recall-Oriented Understudy for Gisting Evaluation, which is known as ROUGE, is a set of used as evaluating summarization and machine interpretation software in natural language processing (NLP). It takes two input; one longest common subsequence and ground truth. The calculation of result produce a summary of a reference or set of reference.

4.3.1.3 CIDER

CIDER is a new metrics for evaluation. This metrics protocol based on by comparison of machine translation approaches based on their “human-likeness”, without making arbitrary calls on content like grammar, weighing content saliency etc.

Since our pre-trained model trained by the MS COCO dataset [30], therefore it is on of key metrics for sentence generation evaluation. Huge improvements in the most recent years, it is understandable that it is more important to report BLEU-4, which is the standard in machine interpretation assertive ahead. In spite of ongoing activities on better assessment measurements in Vedantam [29], our pre-trained model passages unequivocally gradually. The result are charted all metrics comparison on Table 1. In any case, while assessing our subtitles utilizing human raters, it tolls considerably more inadequately, proposing more work is required towards improved measurements. By the official test set, for which designations are just accessible through the authorized site, our generator had a 0.235 BLEU-4.

Table 1: Compare with metrics.

| Project Name | BLEU-4 | Cider | Rouge |
|-----------------------------|--------|-------|-------|
| Neural image Caption | 0.27 | 0.85 | 0.209 |
| Neural Talk | 0.025 | - | 0.222 |
| Show, Attend and Tell | 0.021 | - | 0.204 |
| Adaptive Attention | 0.008 | - | 0.196 |
| Random | 0.046 | 0.051 | - |
| Nearest Neighbor | 0.099 | 0.36 | - |
| Human Evaluation | 0.217 | 0.25 | - |
| Visual Behavior Recognition | 0.235 | 0.78 | 0.38 |

From our experiment, it can understand that, it can be possible to enhance our result by training more data. Since all our experiment done by CPU, it has unavoidable limitation on training and iteration.

With a specific end goal to speak to the past word S_{t-1} as contribution to the translating LSTM creating S_t , by utilizing word embedding vectors [31], which has the upside of being free of the measure of the dictionary (in spite of a less complex one hot-encoding approach). Besides, these word embedding’s can be mutually prepared with whatever remains of the model. It is amazing to perceive how the educated portrayals have caught some meaningful semantic from the insights of the language. In reality, having "horse", "horse",

and "jackass" near each other will urge the CNN to remove includes that are applicable to horse-looking creatures. It guess that, in the outrageous situation where it may see not very many cases of a class (e.g., "unicorn"), its nearness to other word embedding's (e.g., "Zebra") ought to give significantly more data that would be totally lost with more customary words pack based methodologies. Therefore, it will find its capability in other structured prediction problems not only in vision but also in other problem domains. Understanding visual context will boost machine to gain visual intelligence. Although it sounds might not be difficult, it can make great impact on vision field. For instance, if a machine can capable of describe a photo from crime scene then it'll be easy to analyze data to find out whole scenario from crime zone. Besides, it will help in monitoring exam hall to reduce duplicability.

5. CONCLUSION

In this research, the proposed work is one of the popular deep learning architecture for visual behavior detection to understand the visual context and interpret several objects and people within an image. The following work present an end-to-end neural network for machine translation and our generated text is in plain English. The purpose of this work is to give a boost to machine so that can interpret an image like a human. Importance for visual intelligence cannot be describe at all. Proposed architecture involves object detection and recognition, scene classification, LSTM-based attribute prediction. A convolutional neural network that can turn an image into a dense demonstration, monitored by a LSTM based sentence generation. The model is used here is pre-trained, which was limitation due to computational power. For evaluation it used BLEU, CIDER, rouge metrics and our result is quite satisfactory level with the limited computation. In future, it is possible to use customize model to generate more efficient sentence. Proposed research methodology produces interpretable predicate shifts, permitting us to approve demonstrate in reality learning in terms of context. Because of the high quality of the generated image descriptions, accuracy of result could have fluctuated by the using of GPU and number of iteration. Expectation of the proposed methodology and experimental results intend to utilize to localize totally concealed categories by depending on fractional alluding connections and how it can be amplified to perform consideration saccades on scene charts. Enhancements in behavior location might concrete the way for vision calculations to identify inconspicuous substances. Another work may focus on interpreting videos straightforwardly to sentences rather than producing content of images. Static pictures can as it gave blind individuals with data around one particular moment of time, whereas video caption era may possibly give dazzle individuals with ceaseless genuine time data.

6. REFERENCES

- [1] Lu, C., Krishna, R., Bernstein, M. and Fei-Fei, L, Visual relationship detection with language priors. In European Conference on Computer Vision, Springer Cham, (pp. 852-869), 2016.
- [2] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R. and LeCun, Y., Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229, 2013.
- [3] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), pp.211-252, 2015.
- [4] Gupta, A. and Mannem, P., From image annotation to image description. In *International Conference on Neural Information Processing*. Springer, (pp. 196-204), 2012.
- [5] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. and Darrell, T., Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2625-2634), 2015.
- [6] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B. and Lee, H., Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396, 2016.
- [7] Li, Y., Ouyang, W. and Wang, X., Vip-cnn: Visual phrase guided convolutional neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7244-7253), 2017.
- [8] Xu, D., Zhu, Y., Choy, C.B. and Fei-Fei, L., Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (Vol. 2)*, 2017.
- [9] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R. and Bengio, Y., Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048-2057), 2015.
- [10] Ramanathan, V., Li, C., Deng, J., Han, W., Li, Z., Gu, K., Song, Y., Bengio, S., Rosenberg, C. and Fei-Fei, L., Learning semantic relationships for better action retrieval in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1100-1109), 2015.
- [11] Galleguillos, C., Rabinovich, A. and Belongie, S., Object categorization using co-occurrence, location and appearance. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1-8), 2008.
- [12] Ladicky, L., Russell, C., Kohli, P. and Torr, P.H., Graph cut based inference with co-occurrence statistics. In *European Conference on Computer Vision* (pp. 239-253). Springer, 2010.
- [13] Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewiora, E. and Belongie, S., Objects in context. *IEEE 11th international conference Computer vision, ICCV 2007* (pp. 1-8). 2007.
- [14] Salakhutdinov, R., Torralba, A. and Tenenbaum, J., Learning to share visual appearance for multiclass object detection. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on* (pp. 1481-1488), 2011.
- [15] Jia, Z., Gallagher, A., Saxena, A. and Chen, T., 3d-based reasoning with blocks, support, and stability. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (pp. 1-8), 2013.
- [16] Silberman, N., Hoiem, D., Kohli, P. and Fergus, R., October. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision* (pp. 746-760). Springer, 2012.

- [17] Zheng, B., Zhao, Y., Yu, J., Ikeuchi, K. and Zhu, S.C., Scene understanding by reasoning stability and safety. *International Journal of Computer Vision*, 112(2), pp.221-238, 2015.
- [18] Elamri, C. and de Planque, T., Automated Neural Image Caption Generator for Visually Impaired People, 2016.
- [19] Chen, J., Dong, W. and Li, M., Image Caption Generator Based On Deep Neural Networks.
- [20] Karpathy, A., Joulin, A. and Fei-Fei, L.F., Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems* (pp. 1889-1897), 2014.
- [21] Zhang, H., Kyaw, Z., Chang, S.F. and Chua, T.S., Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, No. 3, p. 4), 2017.
- [22] Thomason, J., Venugopalan, S., Guadarrama, S., Saenko, K. and Mooney, R., Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1218-1227), 2014.
- [23] Ordonez, V., Kulkarni, G. and Berg, T.L., Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems* (pp. 1143-1151), 2011.
- [24] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 3156-3164), 2015.
- [25] Karpathy, A. and Fei-Fei, L., Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128-3137), 2015.
- [26] Lu, J., Xiong, C., Parikh, D. and Socher, R., Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 6, p. 2)*, 2017.
- [27] KPapineni, K., Roukos, S., Ward, T. and Zhu, W.J., BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics. Association for Computational Linguistics*, (pp. 311-318), 2002.
- [28] Lin, C.Y., Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004
- [29] Vedantam, R., Lawrence Zitnick, C. and Parikh, D., Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566-4575), 2015.
- [30] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P. and Zitnick, C.L., Microsoft coco: Common objects in context. In *European conference on computer vision*, Springer, Cham, (pp. 740-755), 2014
- [31] Mikolov, T., Chen, K., Corrado, G. and Dean, J., Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013