A New Image Captioning Approach for Visually Impaired People

Conference Paper · November 2019

DOI: 10.23919/ELECO47770.2019.8990630

2 authors, including:

Volkan Kilic
University of Surrey
46 PUBLICATIONS **412** CITATIONS

SEE PROFILE

A New Image Captioning Approach for Visually Impaired People

Burak Makav, Volkan Kılıç

Electrical and Electronics Engineering Graduate Program, Izmir Katip Celebi University, Izmir, Turkey
y180230004@ogr.ikc.edu.tr, volkan.kilic@ikc.edu.tr

Abstract

Automatic caption generation in natural language to describe the visual content of an image has attracted an increasing amount of attention in the last decade due to its potential applications. It is a challenging task to generate captions with proper linguistics properties as it requires an advanced level of image understanding that goes far beyond image classification and object detection. In this paper, we propose to use the Stanford CoreNLP model to generate a caption after images are trained using VGG16 deep learning architecture. The visual attributes of images are extracted with the VGG16, which conveys richer content, and then they are fed into the Stanford model for caption generation. Experimental results on the MSCOCO dataset show that the proposed model significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics.

1. Introduction

The problem of generating natural language descriptions of an image to describe the visual content has received much interest in the fields of computer vision and natural language processing, driven by applications such as image indexing or retrieval, virtual assistants, image understanding and support of the visually impaired people.

Although the visually impaired people use other senses such as hearing and touch to recognize the events and objects around them, the life quality of those people can be dramatically lower than standard level. For this reason, studies such as "guide dog" [1], "smart glasses" [2] and "image captioning" [3] are reported in order to improve the life quality of visually impaired. In this study, a new captioning approach is reported to describe visual content of an image which can be integrated to hardware platforms such as smartphone and smart glass in order to make their life not simply accessible but a socially meaningful and enjoyable experience. To generate a natural language description of an image, sophisticated algorithms are required that goes beyond image classification and object detection which attracts the interest of two major areas of artificial intelligence (AI): computer vision and natural language processing (NLP) [4]. NLP is defined as the automatic exchange of natural language such as general speech and text by software [5], and is a collective term referring to the automatic computational processing of human languages. This term includes both algorithms that take human-generated text as input and algorithms that produce natural-looking text as output [6]. Earlier techniques were designed to use statistical methods in NLP studies. However, theoretical and algorithmic advances together with the increasing capability in computer processing have led to the emergence of more sophisticated techniques like neural networks replaced by statistical methods [7]. Neural networks consist of extremely complex structures, however, deep learning methods provide an effective solution for the processing of data in these structures. In captioning, deep learning architectures are used to extract visual attributes of images. Then, they are fed into the NLP for caption generation. In that sense, deep learning architecture plays a key role in captioning performance as the NLP generates caption based on visual attributes extracted in deep learning side. There are numerous architectures reported in the literature like ZFNet [8], AlexNet [9], GoogLeNet [10] and VGGNet [11]. Here, the VGG16, a popular member of VGGNet family, is employed due to the success of VGGNet over other architectures. To generate caption from visual attributes, models like "Nearest Neighbor" (NN) [12], "Recurrent neural network (RNN)" [13], "Random" [14], "1NN fc7" [15], "Human" [16] and "Stanford" [17] have been proposed. In this study, we propose to use the VGG16 deep learning architecture followed by the Stanford model to generate a caption. We show in our experiments that incorporating the VGG16 architecture and Stanford model in this way improves the captioning performance significantly.

The rest of this paper is organized as follows: the next section introduces the proposed approach for caption generation. Section 3 presents dataset and performance metrics used in comparisons of existing and proposed methods. Closing remarks are given in Section 4.

2. Proposed Captioning Approach

In this section, we describe our approach using the VGG16 deep learning architecture and Stanford model. Our goals are twofold. First, the visual attributes of images need to be extracted with richer content. Second, feeding these attributes to the NLP model to generate the most human-like captions. To accomplish both of these tasks we propose to incorporate the VGG16 architecture and Stanford model as presented in following subsections.

2.1. VGG16 Architecture

The VGG16 deep learning architecture is member of VGGNet [11] family which was the first runner-up in image classification and winner in localization category in ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2014. It is a convolutional neural network model and the VGG16 is deeper using 16 weight layers including thirteen convolutional layers, two fully connected layers and one output later with softmax activation. The convolutional layers are categorized into 5 groups with a max-pooling layer at the end. Illustration of the VGG16 architecture is given in Figure 1. Outstanding performance of the VGG16 over the previous generation of models like AlexNet and GoogLeNet leads us to employ in our captioning approach.

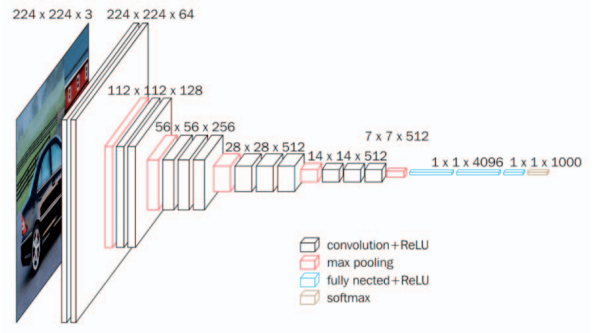

Figure 1. The VGG16 deep learning architecture [18]

2.2. Stanford Model

The Stanford CoreNLP (Stanford native language processing kernel) [17] is a library of object-oriented programming language written in Java and used for tagging text fragments. The overall working algorithm of the model is shown in Figure 2.

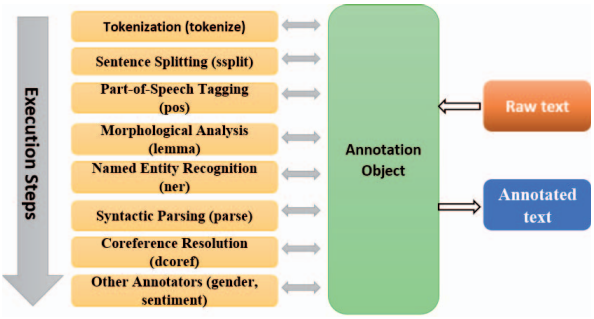

Figure 2. Overall system of the Stanford model [17]

The Stanford model uses the hidden markov chain approach (HMM) which is basically a kind of probabilistic approach [19]. In this approach, the structural features of the languages are statistically extracted. Statistical inference is controlled by random variables. In natural language processing, random variables are words, and in the probabilistic approach of these random variables, it is important to know language-dependent structures [20]. The approach employed in the evaluation of the sentences is the sentence structure of English that a subject followed by a predicate and then an object. The training approach on the basis of this sentence structure consists of the production of some kind of HMM according to human-labeled texts. When the contents of the library are examined, it is stated in the English language that a word with 40% adjective, 40% noun and 20% number will be followed by a word that is known to be an adverbial function, i.e. a word that cannot be followed by a verb. Such examples are used to train human-labeled texts unlikely to be confused by coexisting vocabulary.

When producing natural-looking text, human-labeled text is called reference text, and this reference text is treated as a "annotation object". The first of the application flow process is the tokenization step, which compiles the text into an array and holds it to be processed in subsequent steps. Upon completion of this step, the text allocated to the token becomes easily divided. After dividing, the tokens in the text, such as special names, are identified with high determinants and basic forms are formed in the sentence by using a tagger for these tokens. Apart from these basic forms, another token with high deterministic status is possible gender information. After the gender information is learned, the token in the text including person, position, institution, and numerical information related to money, number, date, time or time period are determined. This determination is found by a sequence function of conditional random fields, one of the structural approaches commonly used in machine learning and deep learning fields for verbal toxins [21].

The proper expression is determined using the Temporal Expression Recogniton system [22]. After the numerical and verbal token diagnosis is done, the process is the process of recognizing the expressions used in daily speech such as ideology, nationality and religion in the strings. After that, the analysis of compound phrases is performed using a kind of probabilistic analyzer [23, 24], and finally the entire basic graph of the text is shown in annotation. The annotations provided with Stanford CoreNLP produce output in plain text format or UTF-8 format containing ASCII characters in XML forms, a language for marking data in a document.

3. Experimental Evaluations

The previous section described proposed captioning approach based on the VGG16 and Stanford model. In this section, datasets and performance metrics used in image captioning are presented, and then the comparative results between existing and proposed models are discussed.

3.1. Dataset

Flickr [25] and MSCOCO [26] are the most popular data sets used in image captioning. For the evaluation of proposed approaches for captioning, selection of dataset is critical as the overall system needs to be trained in advance with the dataset. The number of images in the dataset and the number of reference captions per image are the key factors. Considering these factors, MSCOCO dataset is prominent as there are approximately thirty thousand pictures in Flickr, while MSCOCO contains approximately one hundred sixty thousand pictures. MSCOCO also contains five reference captions per image. In Figure 3, a sample image from MSCOCO dataset is given with reference captions below:

Figure 3. Sample image from MSCOCO

- A train going back to its coarse filled with people.
- A blue commuter train traveling towards a train tunnel.

- Blue train car sitting on a train track near tunnel.
- A train with people inside is about to go into a tunnel.
- A blue train on some train tracks about to go under a bridge

These reference captions are used for both training the system and performance evaluation of the proposed models. The performance metrics used in this study are explained in the next subsection.

3.2. Performace metrics

CIDEr (Consensus-Based Image Description Assessment) [27] and BLEU (Bilingual Evaluation Understudy) [28] metrics, are used to test the success of image subtitle studies. Each of these commonly used metrics evaluates performance from different perspectives.

3.2.1. CIDEr

CIDEr [27] measures the similarity between the generated caption and reference caption. The similarity of captions provides to evaluate captions in terms of grammar, clarity and importance (sensitivity and remembering). The score "CIDErn" for an N-length sentence is calculated using the average cosine similarity between the generated and reference captions valid for both precise and recall:

$$CIDEr_n\left(C_i, S_i\right) = \frac{1}{m}\sum_j \left(\frac{g^n\left(c_i\right).g^n\left(S_{ij}\right)}{\left\|g^n\left(c_i\right)\right\|\left\|g^n\left(S_{ij}\right)\right\|}\right) \quad (1)$$

where $g^n\left(c_i\right)$ is the N length vector formed by $g_k\left(c_i\right)$ and $\left\|g^n\left(c_i\right)\right\|$ is the magnitude of the vector as in $\left\|g^n\left(c_i\right)\right\|$. To capture grammatical features and more schematics, a higher order of N is used [27]. The final version of the equation is as follows:

$$CIDEr_n\left(C_i, S_i\right) = \sum_{n=1}^{N} w_n\left(CIDEr_n\left(C_i, S_i\right)\right) \quad (2)$$

3.2.2. BLEU

BLEU is a popular machine translation metric that analyses an N length caption according to generated and reference captions. A level between captions is calculated with the sensitivity of the truncated N-length: for $j \in m$;

$$CP_n\left(C, S\right) = \frac{\sum_i \sum_k min\left(h_k\left(C_i\right), maxh_k\left(S_{ij}\right)\right)}{\sum_i \sum_k h_k C_i} \quad (3)$$

When K indexes the word set in length N, the truncated precision metric limits that a sentence of length N can be counted up to the maximum number observed in a single reference caption [28]. It is observed that CP_n is a sensitive score and shows favorable captions in short sentences. So a shortness penalty is also used:

$$b\left(C, S\right) = \begin{cases} 1, & \text{if } l_s < l_c \\ e^{1-\frac{l_s}{l_c}}, & \text{if } l_s \geq l_c \end{cases} \quad (4)$$

The total length of the l_c generated captions, l_c and c_i, is the effective reference length at the subject level. When there is more than one reference for a candidate sentence, the closest reference length is used for the shortness penalty. The overall BLEU score is calculated using a weighted geometric mean of the individual N-length precision:

$$BLEU_n\left(C, S\right) = b\left(C, S\right) exp\left(\sum_{n=1}^{N} w_n log CP_n\left(C, S\right)\right) \quad (5)$$

BLEU performed well for subject-level comparisons with multiple N-length mappings. However, a sentence level matches to higher N, resulting in performing poorly when comparing individual sentences [28].

3.3. Results

In this study, proposed approach was tested on the MSCOCO dataset. First, pre-trained VGG16 model is used to extract visual attributes of an image. Then they are fed to the Stanford model to generate a caption. In Figure 4, two examples are shown with generated and reference captions. Caption on the left says "Group of people sitting at a table with plates of food.". Even the subject "Group of people" is not correct, last part of the sentence is meaningful and acceptable. On the other hand, the generated caption for the right image is very similar to the reference captions.

In order to perform a quantitative evaluation of the proposed approach, different metrics are employed by comparing generated caption with reference captions from the MSCOCO dataset. The results are given in Table 1 which compares the proposed approach with state-of-the-art approaches using five metrics.

		Performance Metrics (%)				
	Methods	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**CIDEr**
[12]	NN	48	28.1	16.6	10.0	38.3
[13]	RNN+VGG	8.4	-	-	-	-
[14]	Random	-	-	-	4.6	5.1
[15]	1NN fc7 base (c5)	46.23	26.39	15.07	8.73	-
[16]	Human	51	-	-	-	-
Proposed Method	**Stanford +VGG16**	**57.9**	**40.4**	**27.9**	**19.1**	**60.0**

Table 1. Performance comparison

Table 1 shows that proposed approach significantly outperforms the state-of-the-art approaches.

4. Conclusion

In this study, we presented a new image captioning approach based on the VGG16 deep learning architecture and Stanford model. Our proposed approach was tested on the MSCOCO dataset and showed significantly improved captioning performance over the state-of-the-art approaches. This approach has the potential to be integrated for hardware platforms like smartphones or smart glasses which can ease the problems of visually impaired people having in daily life. Using a NLP embedded smartphone application with narrator options, visually impaired can understand the events in surrounding which turns their life enjoyable experience. This could be interesting directions for future work.

Reference captions:

- A lady sitting at anenormous dining table with lots of food.
- A woman with eye glasses sitting at a table covered with food.
- Several plates of food on a dining table.
- A guest looks over the plates of fruit on the table.
- A woman standing near a table with plates covered in food.

Generated Caption:

Group of people sitting at a table with plates of food.

Reference captions:

- There are many baseball players greeting each other on the field.
- Baseball players high-fiving each other on a baseball field.
- Baseball players are celebrating with each other after a win.
- A number of baseball players on a field with each other.
- A group of men on a baseball field giving each other high fives.

Generated Caption:

Group of baseball players are standing in a baseball game.

Figure 4. Results of proposed captioning approach.

5. References

[1] L. S. Batt, M. S. Batt, J. A. Baguley, and P. D. McGreevy, "Factors associated with success in guide dog training," *Journal of Veterinary Behavior*, vol. 3, no. 4, pp. 143–151, 2008.

[2] J. Bai, S. Lian, Z. Liu, K. Wang, and D. Liu, "Smart guiding glasses for visually impaired people in indoor environment," *IEEE Transactions on Consumer Electronics*, vol. 63, no. 3, pp. 258–266, 2017.

[3] J. Lu, C. Xiong, D. Parikh, and R. Socher, "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*, 2017, pp. 375–383.

[4] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4651–4659.

[5] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[6] Y. Goldberg, "Neural network methods for natural language processing," *Synthesis Lectures on Human Language Technologies*, vol. 10, no. 1, pp. 1–309, 2017.

[7] P. Kantor, "Foundations of statistical natural language processing," *Information Retrieval*, vol. 4, no. 1, pp. 80–81, 2001.

[8] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[13] X. Chen and C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2422–2431.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[15] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[17] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 2014, pp. 55–60.

[18] n. neurohive. VGG16-convolutional network for classification and detection. Internet: https://neurohive.io/en/popular-networks/vgg16/, [Sep. 13, 2019].

[19] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model," in *Proceedings of Computer Society conference on computer vision and pattern recognition*. IEEE, 1992, pp. 379–385.

[20] C. D. Manning, C. D. Manning, and H. Schütze, *Foundations of statistical natural language processing*. MIT press, 1999.

[21] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[22] A. X. Chang and C. D. Manning, "Sutime: A library for recognizing and normalizing time expressions." in *Lrec*, vol. 2012, 2012, pp. 3735–3740.

[23] D. Klein and C. D. Manning, "Fast exact inference with a factored model for natural language parsing," in *Advances in neural information processing systems*, 2003, pp. 3–10.

[24] M.-C. De Marneffe, B. MacCartney, C. D. Manning *et al.*, "Generating typed dependency parses from phrase structure parses." in *Lrec*, vol. 6, 2006, pp. 449–454.

[25] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[26] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[27] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[28] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "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, 2002, pp. 311–318.