ASTES

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 4, 405-411 (2020)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

Keyword Driven Image Description Generation System

Sreela Sreekumaran Pillai Remadevi Amma*, Sumam Mary Idicula

Artificial Intelligence and Computer vision Lab, Department of Computer Science, Cochin University of Science And Technology, Kerala, 682022, India

ARTICLE INFO

Article history:
Received: 15 May, 2020
Accepted: 07 July, 2020
Online: 09 August, 2020

Keywords:
Image captioning
Keyword Extraction
BLSTM

ABSTRACT

Image description generation is an important area in Computer Vision and Natural Language Processing. This paper introduces a novel architecture for an image description generation system using keywords. The proposed architecture uses a high-level feature such as keywords for generating captions. The important component of caption generation is the deep Bidirectional LSTM network. The space and computational complexity of the system are smaller than that of the CNN feature-based image description generation system. The number of parameters is also small in the keyword-based image description generation system. It generates novel meaningful sentences for images. The systems performance depends on the keyword extraction system.

1 Introduction

Image description generation is an active research area for increasing the performance of an image search engine. The system is also useful for accessing image collections, helping visually impaired people, enhancing the education system, robotics, etc.

An image is described by using objects, attributes, actions, scenes, and spatial relationships. In this process, all the keywords of an image are mapped to a sentence called a caption. The image description is created from image features. The main image features are low level and high-level features. Here the proposed system is focused on high-level features such as keywords. The main objective of the system is to develop an image description generation system from keywords. Keyword-based image description system is a direct generation model which extracts keywords from the visual content of the image and generate sentence from keywords.

To achieve the objective, different phases are developed such as

- Keyword extraction system: The objects, attributes, actions, and scenes are extracted.
- Caption generation system: Three methods are employed for generating captions such as template-based, CFG based, and BLSTM based.

The outline of the paper is as follows. Section 2 describes the related work of the system. Section 3 explains the proposed method of the system. Section 4 discusses the experiments and results of the system. The paper is concluded in section 5.

2 Related works

From the literature, there are three kinds of image description generation models such as Direct generation models, Visual space model, and Multimodal space model.

Direct generation models analyse the visual content of the image and create a sentence reflecting the meaning of the visual content. Examples of these models are Google's Neural Image Caption Generator [1], BabyTalk [2], Midge [3], Karpathy's system [4] etc. The visual space model finds similar images of the query image from a visual space and transfer the description to the query image. Research [5]-[6] follows visual space model. The multimodal space model discovers similar images from visual and textual modal space, and it is treated as a retrieval problem. While, [7]-[9] systems follow Multimodal space model.

Another classification of image description generation system is Template based approach and Deep Neural Network based approach [10]-[13]. Deep Neural Network based Image captioning systems are improved using visual attention mechanisms. The first attention-based image captioning system is done by [14]. While [15]-[17] are also proposed attention-based image captioning system. While [17]-[19] follows attention-based architecture. Examples of keyword-based image captioning systems are Midge, BabyTalk, etc. Midge and BabyTalk have less accuracy because of incorrect object detection, invalid action classification, etc. So a novel system for keyword-based image captioning systems is developed using a new keyword extraction system.

*Corresponding Author: Sreela Sreekumaran Pillai Remadevi Amma, +919497708518 & sreela148@cusat.ac.in

https://dx.doi.org/10.25046/aj050447

3 Proposed Method

An image is processed by Computer vision algorithms can be represented as a quadruple $\langle O_i, Attr_i, A_i, S_i \rangle$. Where O_i is the set of objects in the image, $Attr_i$ is the set of attributes of each object, A_i is the set of actions of the object, and S_i is the set of scenes associated with an image. Similarly, the description of the image can be characterized as a quadruple $\langle N_d, Adj_d, V_d, NSc_d \rangle$. Where N_d is the set of nouns in the description, Adj_d is the set of adjectives in the description, V_d is the set of verbs and NSc_d is the set of nouns associated with a scene. Image description generation is the process of mapping image quadruple $\langle O_i, Attr_i, A_i, S_i \rangle$ to description quadruple $\langle N_d, Adj_d, V_d, NSc_d \rangle$.

The overview of the proposed methodology is explained in figure 1. The methodology is as follows: 1) Detect Objects and scenes in the image. 2) Recognize action in an image 3) Identify attributes such as cardinality, shape, and color of the object.4) Generate text corresponding to the image. The method is divided into two parts, such as keywords generation and caption generation.

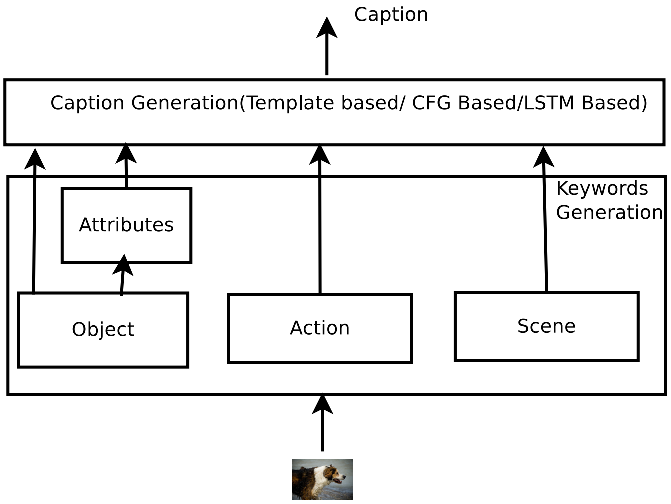


Figure 1: Proposed Architecture

3.1 Keywords Generation

3.1.1 Object detection

Darknet architecture is used for object detection. The images are divided into 7 × 7 grid cells. For each grid cell, we apply a classifier for identifying an object. The number of layers in Darknet is 53. Object classes are obtained in this phase. The object is considered as the noun of the sentence. Top-k objects are selected for sentence construction.

3.1.2 Attributes Identification

This phase identifies the color, shape, and cardinality of the object. Attributes are identified as the adjectives of the sentence.

Color detection The color of the object is calculated from the minimum distance between L*a*b value and the average of the intensity of the image. The color with the smallest distance is treated as the color of the object.

Shape detection The algorithm for shape detection is explained below:

```
Compute the contour perimeter of the image;
Approximate the contour shape to another shape with fewer
  vertices depending upon the precision using the
  Douglas-Peucker algorithm;
if The approximated shape has three vertices then
    it is a triangle;
else
    if the shape has four vertices then
        Compute the bounding box of contour and find the
          aspect ratio;
        if the aspect ratio is approximately equal to one then
            The shape is a square;
        else
            otherwise, it is a rectangle;
        end
    else
        if the shape has more than four vertices then
            it is treated as a circle;
        end
    end
end
```

Cardinality Cardinality is obtained by the count of each object in the image.

3.1.3 Action recognition

Action recognition is done in still images. Action is treated as the verb of the sentence or caption. The architecture of action recognition is explained in figure 2.

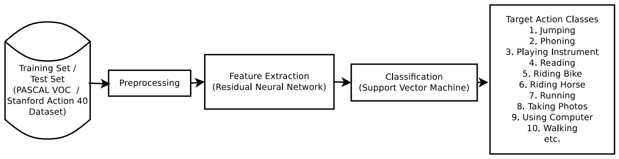


Figure 2: Proposed Architecture of action recognition

It is done using deep convolutional neural network features and Support Vector Machine. Residual Neural Network[20] is used for CNN feature extraction, and classification is done using Support Vector Machine.

3.1.4 Scene Classification

Context or scene is important for caption generation. It is obtained using deep neural network-based learning. The scene is the place where an image is associated. VGG architecture is used for scene classification. Three hundred sixty-five scene categories are used for this classification.

3.2 Caption generation

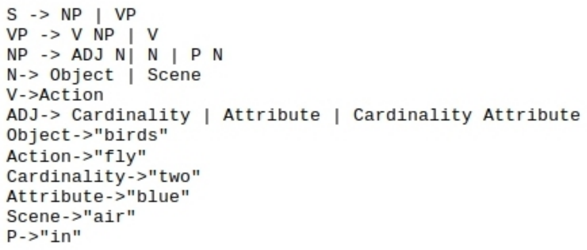
It is a process of generating sentences or captions from a set of keywords like objects, attributes, action, and scene. We implemented three kinds of caption generation, such as template-based, CFG based, and Bidirectional LSTM[21] based caption generation.

3.2.1 Template based

Keywords are extracted as a five-tuple $\langle object, cardinality, attribute, action, scene\rangle$. Captions are generated from the five-tuple. Captions have a common form "There is **cardinality attribute object action** in the **scene**. The object is treated as noun, cardinality, and attributes are adjectives, the action is the verb. An example of a sentence is, "There is two blue birds fly in the sky". The set of words in the meaning representation is fixed, and generation must make use of all given content words; and, generation may insert only gluing words (i.e., function words such as there, is, the, etc.). Advantages of this approach is that it incorporates cardinality and actions in the image captions. Disadvantages of this approach is that it generates similar kinds of captions. And also still image based action recognition generates 10 classes only.

3.2.2 CFG based

We design a CFG for the caption. CFG has the form represented in figure 3. Using this CFG, captions are generated.



```
S -> NP | VP
VP -> V NP | V
NP -> ADJ N| N | P N
N-> Object | Scene
V->Action
ADJ-> Cardinality | Attribute | Cardinality Attribute
Object->"birds"
Action->"fly"
Cardinality->"two"
Attribute->"blue"
Scene->"air"
P->"in"
```

Figure 3: CFG for caption generation

3.2.3 Bidirectional LSTM based

It uses encoder-decoder architecture as Machine translation. Let (k, C) be a keyword and caption sentence pair. Let $k = k_1, k_2, k_3...k_m$ be a sequence of M symbols in keywords and $C = C_1, C_2, C3....C_N$ be the sequence of N symbols in the target caption of the image. The model maximizes the following objective.

$$\theta^* = argmax_\theta \Sigma_{k,C} logp(C/k;\theta) \quad (1)$$

The encoder is simply a function of the following form

$$K_1, K_2, K3...K_M = encoderBLSTM(k_1, k_2, k_3...k_m) \quad (2)$$

Where $K_1, K_2, K_3, ...K_M$ is a list of hidden representation of keywords of size M. $enoderBLSTM$ is a bidirectional LSTM.
Using chain rule, the MAP of the sequence $p(C/k)$ can be decomposed as

$$\begin{aligned} p(C/k) &= p(C/k_1, k_2, k_3, ...k_m) \\ &= \Pi_{i=1}^{N} P(C_i/C_0, C_1, C_2...C_{i-1}; k_1, k_2, k_3...k_m) \quad (3) \\ &= \Sigma_{i=1}^{N} logP(C_i/C_0, C_1, C_2...C_{i-1}; k_1, k_2, k_3...k_m) \end{aligned}$$

where C_0 is a special "beginning of sentence" symbol. During inference, we calculate the probability of next symbol given the source encoding and the decoded target caption so far as $p(C_i/C_0, C_1, ...C_{i-1}; K_1, K_2, ...K_M)$.

Our decoder is organized as a combination of a fully connected neural network(FCN) layer and a softmax layer. The FCN layer produces the hidden representation of the next symbol to be predicted;which then goes through the softmax layer to generate a probability distribution over candidate vocabulary symbols.

The deep neural network architecture is drawn in Figure 4.

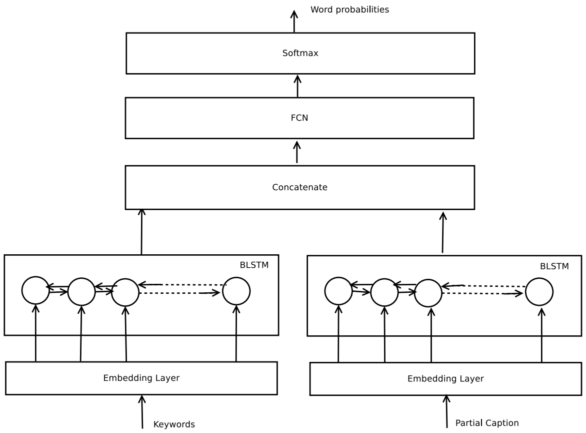


Figure 4: Bidirectional LSTM based caption generation

4 Results

4.1 Keyword Extraction

The Table 1 shows the output of keyword extraction. The table contains objects, attributes such as cardinality, shape,and color, scene, and action of images.

The Table 2 is described the various advantages and disadvantages of various phases of keyword extraction system.

4.2 Caption Generation

The deep neural network based caption generation model is experimented with LSTM[22] and Bidirectional LSTM. From the experiment, the accuracy graphs are plotted in figures 5 and 6. The accuracy plot explained that the accuracy of the model is more when the bidirectional LSTM is used.

The loss of the model is plotted in figures 7 and 8.From the plots, the loss of the model is less when Bidirectional LSTM is used.

Bidirectional LSTM is more suitable for designing the caption generation from keywords. The table 3 describes the captions from keywords using three methods.

The Table 4 explains the various properties of different methods used for caption generation.

4.3 Analysis

The computational complexity of the caption generation from keywords is small. The space complexity of this system is described by

$s_k = |Keywords| \times max\_cap\_len$

Table 1: Keywords Extraction results

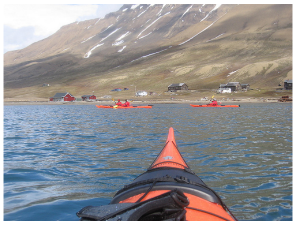
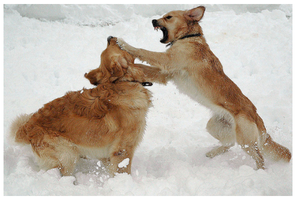
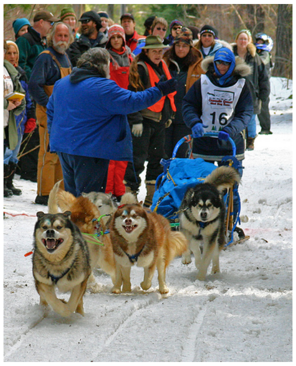
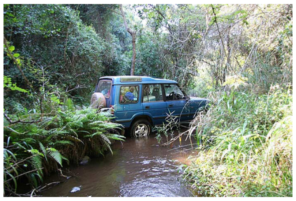
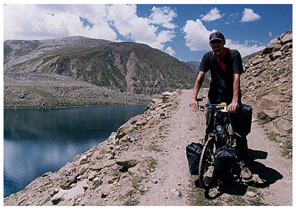
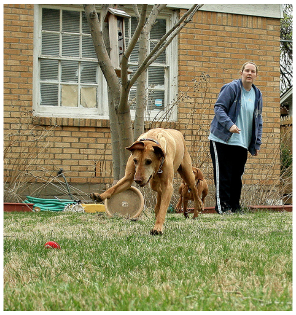
Image	Objects	Attributes	Scene	Action
	Boat, Person	Boat - Color:Olive, Shape: Rectangle Person- Color: Yellow, Shape: Square	Raft, Beach, Lake	Running
	Dog	Dog - Color:Silver, Shape: Square	Sandbox, Archaelogical excavation, Trench	Jumping
	Person	Person - Color:Teal, Shape: Rectangle, Cardinality: 10	Playground, Pet shop, Toy shop	Walking
	Car,Truck	Car - Color:Gray, Shape: Rectangle, Truck - Color:Gray, Shape:Rectangle	Forest path, River	-
	Person, Bicycle, Backpack	Person - Color:Teal, Shape: Rectangle, Bicycle - Color:Teal, Shape:Rectangle, Backpack-Color:Green, Shape: Rectangle	Coast, Beach, Ocean	Riding Bike
	Person, Dog, Frisbee	Person - Color:Gray, Shape: Rectangle, Dog - Color:Gray, Shape:Square, Frisbee- Color:Olive, Shape: Rectangle	Lawn, Yard, Outdoor	Walking

Table 2: The characteristics of different phases of Keywords extraction

Model	Advantages	Limitations
Object detection	Number of objects is 1000.	Vocabulary of objects is limited.
Attributes identification	Color, shape and cardinality are identified.	Only three attributes of objects are detected.
Action recognition	Number of action classes is 10.	The vocabulary of actions is limited.
Scene classification	365 scenes are classified.	Number of scenes is limited.

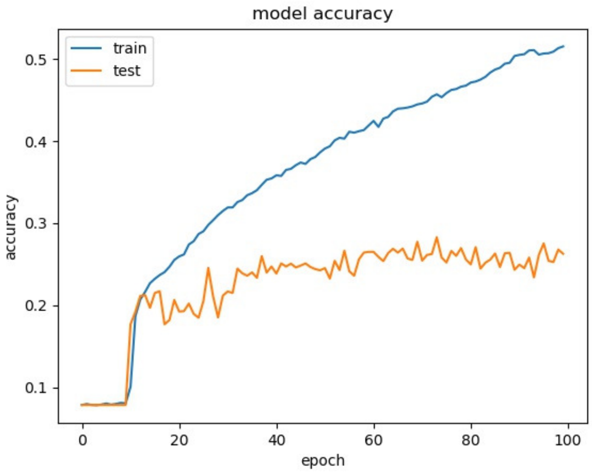


Figure 5: Accuracy plot for LSTM Based Caption generation from keywords

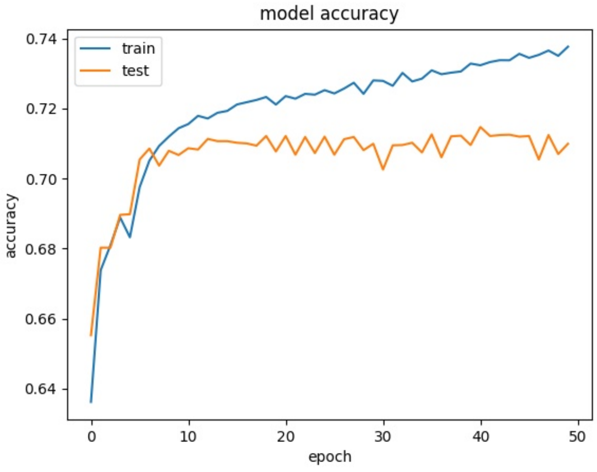


Figure 6: Accuracy plot for Bidirectional LSTM Based Caption generation from keywords

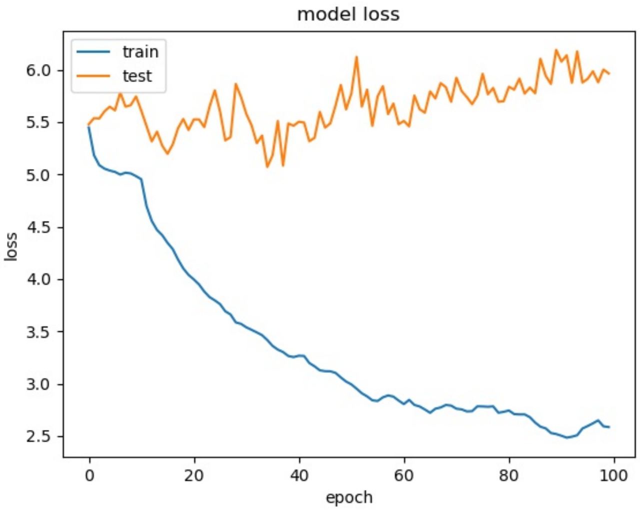


Figure 7: Loss plot for LSTM Based Caption generation from keywords

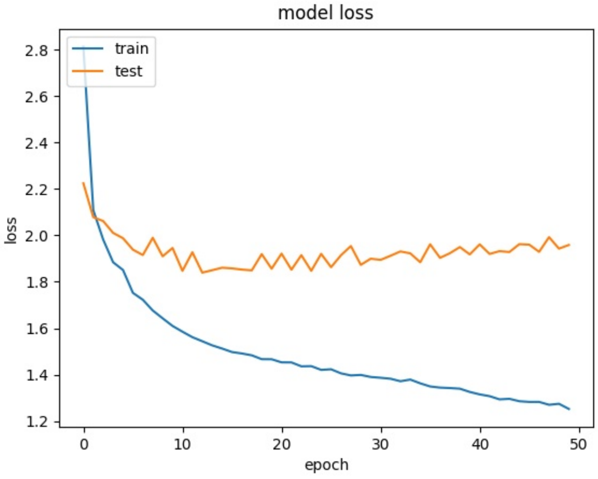


Figure 8: Loss plot for Bidirectional LSTM Based Caption generation from keywords

Table 3: Output of caption generation from keywords using different methods.

Keywords	Template based	CFG based	LSTM based
Person, Boat, Lake, Running	There is a person running in a lake.	A person running in a lake.	A Person is running in the lake with a boat.
Dog, Silver, Sandbox, Jumping	There is a silver dog jumping in sandbox.	Silver dog jumping in sandbox.	Silver dog is jumping in the sandbox.
Person, Ten, Playground, Walking	There is ten persons walking in playground.	Ten persons walking in playground.	Ten persons are walking in the playground.
Truck, Gray, Forest path	There is a gray truck in forest path.	Gray truck in forest path.	Gray truck in the forest path.
Person, Bike, Backpack, Green, Coast, Riding bike	There is a person riding a bike in bike.	Person riding a bike in a bike.	Person is riding bike.

Table 4: The characteristics of different methods of caption generation

Method	Advantages	Disadvantages
Template based	It incorporates cardinality and actions in the image captions.	It generates similar kinds of captions.
CFG based	It generates different types of captions.	It generates caption with same grammar.
BLSTM based	It creates various kinds of captions.	The vocabulary of objects, attributes, actions and scenes is limited.

where $|Keywords|$ is the number of keywords and max_cap_len is the maximum caption length. The $|Keywords|$ is less than the maximum number of keywords. Here we set the maximum number of keywords as 10.

The space complexity of the caption generation in CNN feature-based caption generation system is given by

$s_c = |CNN\_feature| \times max\_cap\_len$

where $|CNN\_feature|$ is the length of the CNN feature, and it is 2208 if the CNN used is Densenet. From the analysis, $s_k < s_c$.

If the number of network parameters of the keyword-based caption generation system is p_k and the number of network parameters of the CNN feature-based caption generation system is p_c, then $p_k < p_c$. The network of keyword-based caption generation system is small.

The limitation of the system is that the performance of the system depends on the keyword extraction system. Keyword extraction is based on object detection, object identification, action recognition, and scene classification. The mean average precision (MAP) of the action recognition system is only 66%. The vocabulary of actions contains only ten words. The vocabulary size of objects is 1000. The number of scenes is 365. So the total vocabulary contains around 1.5k words. But the vocabulary size of the flickr8k dataset is 8256. Hence the vocabulary of the keyword-based caption generation system is small.

In the future, we will concentrate on increasing the performance of the keyword extraction system. Thereby we can improve the functioning of caption generation from keywords.

5 Conclusions

The paper proposed a novel system for keyword-based image caption generation. The main components of the system are the keyword extraction and caption generation. It produces meaningful captions for images. The space and computational complexity of this system are small. The limitation of the system is that the performance of the system mainly depends on the keyword extraction system. So in the future, we will focus on improving the performance of the keyword extraction system.

References

[1] Vinyals, Oriol, 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, pp. 3156-3164. 2015. DOI: 10.1109/CVPR.2015.7298935

[2] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg. "Baby talk: Understanding and generating simple image descriptions.", IEEE Transactions on Pattern Analysis and Machine Intelligence, **35**(12), 28912903, 2013. DOI: 10.1109/TPAMI.2012.162

[3] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, A. Mensch, A. Berg, X. Han, T. Berg, and O. Health., "Midge: Generating Image Descriptions From Computer Vision Detections.", Eacl, pp. 747-756, 2012.

[4] K., Andrej, and L. Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128-3137. 2015. DOI: 10.1109/CVPR.2015.7298932

[5] V. Ordonez, G. Kulkarni, and T.L. Berg., "Im2text: Describing images using 1 million captioned photographs.", Advances in Neural Information, pp. 19, 2011.

[6] P. Kuznetsova, V. Ordonez, A.C. Berg, T.L. Berg, and Y. Choi, "Collective generation of natural image descriptions.", In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers **1**. Association for Computational Linguistics, 359-368, 2012.

[7] H., Micah, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics.", Journal of Artificial Intelligence Research **47** 853-899, 2013. https://doi.org/10.1613/jair.3994

[8] Gong, Yunchao, et al. "Improving image-sentence embeddings using large weakly annotated photo collections.", European Conference on Computer Vision. Springer, Cham, 2014. DOI: https://doi.org/10.1007/978-3-319-10593-2_35

[9] R. Socher., A. Karpathy, Q. V. Le, C. D. Manning, A. Ng, "Grounded Compositional Semantics for Finding and Describing Images with Sentences.", Transactions of the Association for Computational Linguistics (2), 207218, 2014. DOI: 10.1162/tacl_a_00177

[10] Y. H. Tan and C. S. Chan, "phi-LSTM: A Phrase-Based Hierarchical LSTM Model for Image Captioning." Cham: Springer International Publishing, pp. 101-117 , 2017. https://doi.org/10.1007/978-3-319-54193-8_7

[11] Tan, Y. Hua, and C.S. Chan. "Phrase-based Image Captioning with Hierarchical LSTM Model." 2017. DOI: arXiv preprint arXiv:1711.05557

[12] Aneja, Jyoti, Aditya Deshpande, and Alexander G. Schwing. "Convolutional image captioning."Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. DOI: 10.1109/CVPR.2018.00583

[13] Han, Meng, W. Chen, and A.D. Moges. "Fast image captioning using LSTM." Cluster Computing. DOI: https://doi.org/10.1007/s10586-018-1885-9

[14] Xu, Kelvin, et al., "Show, attend and tell: Neural image caption generation with visual attention.", International Conference on Machine Learning, 2015.

[15] Junqi Jin, Kun Fu, Runpeng Cui, Fei Sha, and Changshui Zhang. 2015. "Aligning where to see and what to tell: image caption with region-based attention and scene factorization." arXiv preprint arXiv:1506.06272, .

[16] Z.Y.Y.Y.Y. Wu and R.S.W.W. Cohen, "Encode, Review, and Decode: Reviewer Module for Caption Generation." In 30th Conference on Neural Image Processing System(NIPS), 2016.

[17] Lu, Jiasen, et al. "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). **6** 2017. DOI: 10.1109/CVPR.2017.345

[18] You, Quanzeng, et al. "Image captioning with semantic attention." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016. DOI: 10.1109/CVPR.2016.503

[19] He, Chen, and Haifeng Hu. "Image captioning with text-based visual attention." Neural Processing Letters 49.1: 177-185, 2019. DOI: https://doi.org/10.1007/s11063-018-9807-7

[20] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian, "Deep Residual Learning for Image Recognition". Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2016. DOI: 10.1109/CVPR.2016.90

[21] Wang, Cheng, et al, "Image captioning with deep bidirectional LSTMs", Proceedings of the 2016 ACM on Multimedia Conference, ACM, 2016. DOI: https://doi.org/10.1145/2964284.2964299

[22] Hochreiter, Sepp, and Jrgen Schmidhuber. "Long short-term memory.", Neural computation pp. 1735-1780, 1997. DOI: https://doi.org/10.1162/neco.1997.9.8.1735