Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8909458, 7 pages
https://doi.org/10.1155/2020/8909458

Research Article

Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention

Yan Chu,[1] **Xiao Yue**,[2] **Lei Yu**,[1] **Mikhailov Sergei**,[1] **and Zhengkui Wang**[3]

[1]*Harbin Engineering University, Harbin 150001, China*
[2]*Zhongnan University of Economics and Law, Wuhan 430073, China*
[3]*Singapore Institute of Technology, Singapore 138683*

Correspondence should be addressed to Xiao Yue; yuexiao@zuel.edu.cn

Received 18 January 2020; Accepted 24 September 2020; Published 21 October 2020

Academic Editor: Yin Zhang

Captioning the images with proper descriptions automatically has become an interesting and challenging problem. In this paper, we present one joint model AICRL, which is able to conduct the automatic image captioning based on ResNet50 and LSTM with soft attention. AICRL consists of one encoder and one decoder. The encoder adopts ResNet50 based on the convolutional neural network, which creates an extensive representation of the given image by embedding it into a fixed length vector. The decoder is designed with LSTM, a recurrent neural network and a soft attention mechanism, to selectively focus the attention over certain parts of an image to predict the next sentence. We have trained AICRL over a big dataset MS COCO 2014 to maximize the likelihood of the target description sentence given the training images and evaluated it in various metrics like BLEU, METEROR, and CIDEr. Our experimental results indicate that AICRL is effective in generating captions for the images.

1. Introduction

With the rapid development of digitalization, there are a huge amount of images, accompanied with a lot of related texts [1]. Automatic image captioning has recently attracted much research interest. The objective of automatic image captioning is to generate properly formed English sentences to describe the content of an image automatically, which is of great impact in various domains such as virtual assistants, image indexing, recommendation in editing applications, and the help of the disabled [2, 3]. Although it is an easy task for a human to describe an image, it becomes very difficult for a machine to perform such a task [4]. Image captioning does not only need to detect the objects contained in an image but also capture how these objects related to each other and their attributes as well as the activities involved in. Moreover, the semantic knowledge should be expressed in a natural language, which requires a language model to be developed based on the visual understanding.

Much research effort has been devoted to automatic image captioning, and it can be categorized into template-based image captioning, retrieval-based image captioning, and novel image caption generation [5]. Template-based image captioning first detects the objects/attributes/actions and then fills the blanks slots in a fixed template [1]. Retrieval-based approaches first find the visually similar images with their captions from the training dataset, and then the image caption is selected from similar images with captions [6]. These methods are able to generate syntactically correct captions but are unable to generate image-specific and semantically correct captions. Differently, the novel image caption generation approaches are to analyze the visual content of the image and then to generate image captions from the visual content using a language model [7]. Compared to the first two categories, novel caption generation can generate new captions for a given image that are semantically more accurate than previous approaches. Most of the works in this category rely on machine learning and deep learning, which is also the approach adopted in this paper. One common framework used in this category is the encoder-decoder framework for image captioning [8]. This framework was first introduced to describe a multimodal

log-bilinear model for image captioning with a fixed context window by Kiros et al. [9]. Recent research works have used the deep convolutional neural network (CNN) as the encoder and the deep recurrent neural network (RNN) as the decoder, which is proven to be promising [8, 10, 11]. However, it still remains challenging to identify the proper CNN and RNN models for the image captioning.

In this paper, we investigate one single-joint mode, AICRL, for automatic image generation using ResNet50 (a convolutional neural network) and LSTM (long short-term memory) with soft attention mechanism. AICRL consists of an encoder and a decoder. We adopt ResNet50 as the encoder to create an extensive representation of an input image by embedding it into a vector. Meanwhile, we utilize the LSTM with a soft attention as the decoder which selectively focuses the attention over a certain part of an image to predict the next sentences. Furthermore, we conduct extensive experiment and empirically determine the structure of the model and fine-tuned the model hyperparameters. Our experimental evaluation indicates that AICRL is effective to generate proper captions for the images.

The rest of the paper is organized as follows. Section 2 introduces the related work. In Section 3, we present the proposed AICRL model. Section 4 and Section 5 provide the experimental evaluation and conclusion, respectively.

2. Related Work

Much research has been devoted on the automatic image captioning recently. The research can be briefly categorized into three different categories including the template-based approaches, retrieval-based approaches, and novel image caption generation approaches.

The template-based approach is aimed at generating captions by using fixed templates with a number of blank slots, in which way different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled. For example, Farhadi et al. [1] use a triplet of scene elements to fill the template slots for generating image captions. Li et al. [12] extract the phrases related to detected objects, attributes, and their relationships for this purpose. Kulkarni et al. [13] adopt a conditional random field (CRF) method to infer the objects, attributes, and prepositions before filling in the gaps. Template-based methods can generate grammatically correct captions. However, templates are predefined and length of captions cannot be variable.

The retrieval-based approach tries to generate description for an image by selecting the most semantically similar sentences from sentence pool or directly copying sentences from other visually similar images. For example, Gong et al. [6] utilize stacked auxiliary embedding method to generate image descriptions from millions of weakly annotated images. Ordonez et al. [14] find similar images in the Flickr database and return the descriptions of these retrieved images to query based on millions of images and their corresponding descriptions. Sun et al. [15] use semantic similarity and visual similarity scores to cluster similar terms and images together first and then retrieve caption of target image from captions of similar images in the same cluster. Hodosh et al. [16] establish a ranking-based framework to treat sentence-based image description as the task of ranking a set of captions for each test image. These methods generate general and syntactically correct captions. However, it is difficult for them to generate image-specific and semantically correct captions.

Different from the mentioned two categories, novel caption generation approaches mainly use deep learning and machine learning to generate the new captions. A general implementation of this method is to analyze the visual content of the image first and then generate image captions from the visual content using a language model. For instance, Vinyals et al. use CNN as an encoder for image classification and LSTM as a decoder to generate sentence for the description [8]. The main drawbacks of the work are the quick model overfitting, so they use the heavy and expensive GoogLeNet with 22 hidden layers and the absence of attention layer that significantly improved the description accuracy. Karpathy et al. investigate the possibility of generating an image description in natural language [10]. Their approach uses image datasets and their description in natural language and seeks an intermodal correspondence between words from the description and visual data. The first model aligns the fragments of sentences to the visual areas, then forms a single description by multimodal embedding. This description is treated as learning data for a second model of a recurrent neural network that learned to a generate caption. Xu et al. use a convolutional neural network to extract feature maps and LSTM to describe the input image, by processing already extracted feature maps [11]. The limitation of this work is the using of obsolete and expensive Oxford VGGnet, where the quality of image classification is low in the modern CNN [7]. Some researchers have put their attention on classification as Yu et al. [17], who propose a SVM classification-based two-side cross-domain algorithm by inferring intrinsic user and item features (CTSIF-SVMs), a two-side cross-domain algorithm with expanding user and item features via the latent factor space of auxiliary domains (TSEUIF) [18].

3. Model

In this section, we present our proposed model, AICRL, for automatic image captioning based on ResNet50 and LSTM with software attention. The ultimate purpose of AICRL is to generate the proper description for the given images. To do so, the AICRL model is designed with an encoder-decoder architecture based on CNN and RNN. In particular, to extract visual features, we use the ResNet50 network as the encoder to generate a one-dimensional vector representation of the input images. After that, to generate the description sentences, we adopt the LSTM as the language model for the decoder to decode the vector into a sentence. Meanwhile, we utilize the soft attention in the decoder to enable the model to selectively focus the attention over a certain part of an image to predict the next sentence better. We conduct extensive experiments, empirically determine the structure of the model, and fine-tune the model hyperparameters.

The whole model is fully trainable by using a stochastic gradient descent.

In the encoder-decoder method, the most likely description of the image is determined by maximizing the log-likelihood function of the expression S, considering the corresponding image I and the parameters of the model θ.

$$\theta^* = \arg \max_{\theta} \sum (I, S) \log p(S|I; \theta), \tag{1}$$

where θ is the parameter of our model, I is the input image, and S is the correct description. Since S represents a sentence of any length, therefore, a chain rule is usually used to model the joint probability over $S_1, \cdots, S_N$, where N is the length of this particular example.

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \cdots, S_{t-1}), \tag{2}$$

where the dependence on θ is omitted for convenience. The network training is represented by the pair of (S, I), and we optimize the sum of the log likelihood functions, as described in Equation (2), over the entire training set using stochastic gradient descent.

The likelihood $\log p(S_t|I, S_0, \cdots, S_{t-1})$ is modelled by a recurrent neural network, where there is a variable number of words that we define up to $t-1$. The hidden state of RNN (latent memory) h_t is updated after the new input x_t with the nonlinear function f.

$$h_{t+1} = f(h_t, x_t). \tag{3}$$

3.1. Image Feature Extraction. To represent the image, we adopt the convolutional neural network (CNN), ResNet50, which is a very deep network that has 50 layers. The depth of the network is crucial for neural networks, but deeper networks are more difficult to train. The structure of ResNet50 facilitates the training of networks and allows them to be much deeper, which leads to increased performance in different tasks. ResNet50 is much deeper than their "simple" counterparts, but moreover, the number of parameters (weights) of such networks is much smaller. For example, Table 1 indicates the number of parameter comparison between ResNet50 and VGG16. Deep convolutional neural networks have led to a series of breakthroughs for image classification. Recent evidence reveals that network depth is of crucial importance. Many other nontrivial visual recognition tasks have also greatly benefited from the deep models.

With the network depth increasing, the accuracy of networks increases rapidly, which is not surprising and then rapidly degrades (saturated). This degradation is not caused by overfitting, and the addition of even more layers leads to a higher learning error. In a sense, this is strange, since a deeper network has a strictly large representational power. It is possible for ResNet50 to get a deeper model trivially, which is not worse than the less deep network. It can be done by adding several identity layers, that is, levels that simply skip the signal further without changes. ResNet50's deeper levels have to predict the difference between the output of the previous layers and the objective function. They could always drive the weights to 0 and simply skip the signal. Hence, deep residual learning is a good method that makes the network learn to predict deviations from past layers.

Table 1: Comparison of total number of VGG16 and ResNet50 parameters.

CNN	Number of parameters
VGG16	138,357,544
ResNet50	23,587,712

The model takes an image and produces a caption, encoded as a sequence of $1 - K$ coded words.

$$y = \{y_1, y_2, \cdots, y_c\}, y_i \in R^K, \tag{4}$$

where K is the size of the dictionary and c is the caption length. We use CNN in particular, ResNet50, to obtain set annotation vectors like the feature vectors. The extractor produces L-vectors, all of which is a D-dimensional representation of the corresponding part of an image.

3.2. The Language Model. The choice of f in Equation (3) is determined by its ability to cope with vanishing problems and exploding gradients, which are the most common problems in the design and training of RNN. LSTM networks are successfully used to accomplish the tasks of machine translation and sequence generation. In our design, we adopt LSTM as our language model to generate proper caption based on the input vector from the ResNet50 output.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right), \tag{5}$$

where the output vector of the previous cell h_{t-1} with the new element of the sequence x_t is concatenated and passed as one vector through the layer with the sigmoid activation function.

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C_t}. \tag{6}$$

Two created vectors are used to update the state from C_{t-1} to C_t. To do this, we multiply the past state by f_t to "forget" the data recognized as unnecessary in the previous step, then add $i_t * \widetilde{C_t}$.

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i), \\ \tilde{C}_t &= \tanh\left(W_c \cdot [h_{t-1}, x_t] + b_c\right). \end{aligned} \tag{7}$$

The input gate must determine what values will be updated, and the tanh layer creates a vector of new candidates for $\tilde{C}_t$, and values can be added to the cell state.

$$h_t = o_t * \tanh\left(C_t\right). \tag{8}$$

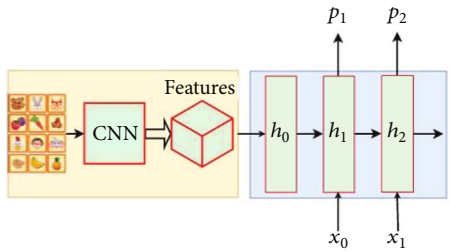

Figure 1: Model without attention.

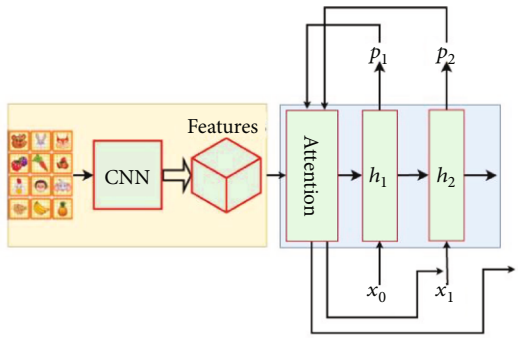

Figure 2: Model with attention.

The obtained values of C_t and h_t are transmitted to the neural network input at time $t + 1$.

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o),$$
$$p_t = \text{soft max}\,(h_t). \tag{9}$$

The multiplicative filters allow to effectively train LSTM, as they are good to prevent the exploding and vanishing gradients. Nonlinearity is provided by the sigmoid $\sigma(\cdot)$ and the hyperbolic tangent $h(\,)$. In the last equation, h_t is fed to the softmax function to calculate the probability distribution p_t over all words. This function is calculated and optimized on the entire training dataset. The word with maximum probability is selected at each time step and fed into next time step input to generate a full sentence.

3.3. Attention Mechanism. To better isolate the image content, we adopt the soft attention mechanism, which has been widely used to solve the problem of image classification, as there is no need to process all pixels of an image. For example, in the classification problem, the background usually plays an insignificant role. Nevertheless, convolutional neural networks, which are the most popular method for solving such a problem, spend the same amount of computational resources on all parts of the image.

Soft attention is implemented by adding an additional input of attention gate into LSTM that helps to concentrate selective attention. The main drawback of the model without attention is that it tries to decode the full image from the last hidden layer of h_0 in Figure 1. It is like an analogy with machine translation in the whole process. To do a translation of the whole text is just from the "last word." So it will lose a lot of useful information from the beginning of the text.

Table 2: Comparison for AICRL with and without attention.

Model	BLEU-4	METEOR	CIDEr
With attention	0.326	0.261	0.872
Without attention	0.262	0.209	0.803

Table 3: Comparison for AICRL with and without attention.

Model	Right choosing of generated description
With attention	71%
Without attention	54%

The attention gate can be represented as an addition input for LSTM in Figure 2. The soft attention depends on the previous output of LSTM p_t and extracted features of input image y_i. Soft attention is differentiable and can be trained by the standard method of the backpropagation algorithm. In the case of model with soft attention, we append an additional a_t in Equation (10).

$$a_t = \sum_{1}^{n} SjYj, \tag{10}$$

where a_t is an attention vector, s_j is a nonlinear function with softmax output, and y_j is the extracted features of the input image.

4. Experiments and Analysis

We perform an extensive set of experiments to evaluate the effectiveness of the proposed model. We have adopted two different datasets in our experiments including the MS COCO 2014 dataset and Flickr8K dataset, which contain the images with their descriptions in English. The MS COCO 2014 dataset contains 102,739 images with their descriptions, five descriptions for each image, and 20,548 testing examples. The Flickr8K dataset is another set of images with their descriptions with 7,000 training examples and 1,000 testing examples. Similar to MS COCO 2014, it also contains five descriptions for each image, but with a much smaller volume. Consider the Flickr8K data has less data than MS COCO 2014. In the training, we first use the Flickr8K dataset to train the model and then use the fine-tuned hyperparameters on MS COCO 2014. All experiments are conducted on NVIDIA GPU GTX-1070.

We evaluate the model using several popular metrics such as BLEU [19], METEOR [20], and CIDEr [21]. BLEU (Bilingual Evaluation Understudy) is an algorithm that measures the precision of an n-gram between the generated and reference captions. BLEU-N ($N = 1, 2, 3, 4$) scores can be calculated based on the length of the reference sentence, the generated sentence, the uniform weights, and the modified n-gram precisions.

METEOR (Metric for Evaluation of Translation with Explicit Ordering) is an evaluation metric which was initially used in machine translation. Besides measuring precision,

Table 4: The performance comparison in the Flick8K dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Mao et al. [22]	0.58	0.28	0.23	—	—	—
Google NIC [28]	0.63	0.41	0.27	—	—	—
Chen and Zitnick [23]	—	—	—	0.141	—	—
Log bilinear [25]	0.656	0.424	0.277	0.177	0.173	—
DVS [26]	0.579	0.383	0.245	0.16	—	—
AICRL-ResNet50	*0.619*	*0.452*	*0.368*	*0.262*	*0.209*	*0.803*
AICRL-VGA16	0.672	0.436	0.338	0.225	0.186	0.743

Table 5: The performance comparison in the MS COCO 2014 dataset.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr
Nearest neighbor [27]	0.48	0.281	0.166	0.1	0.157	0.383
Google NIC [28]	0.666	0.461	0.329	0.246	—	—
LRCN [24]	0.628	0.442	0.304	—	—	—
MS research [29]	—	—	—	0.211	0.207	—
Chen and Zitnick [23]	—	—	—	0.19	0.204	0.141
Log bilinear [25]	0.708	0.489	0.344	0.243	0.2	—
DVS [26]	0.625	0.45	0.321	0.23	0.195	0.66
AICRL-ResNet50	*0.731*	*0.562*	*0.41*	*0.326*	*0.261*	*0.872*
AICRL-VGA16	0.702	0.536	0.398	0.295	0.236	0.857

METEOR places emphasis on the recall between the generated and ground truth captions.

CIDEr (Consensus-based Image Description Evaluation) measures the similarity of generated captions to their ground truth sentences for evaluating image captioning. This measurement takes into account the grammaticality and correctness.

4.1. Training. The first step in the process of generating comments to the image is to create a fixed-length vector that effectively summarizes the content of an image. We use CNN, in particular the ResNet50 architecture. This network is preliminarily trained for 1.2 million images of the ImageNet dataset. Therefore, ResNet50 has a reliable initialization for object recognition and allows reducing training time. For any image from the training set, we get the output vector representation from the last convolution layer. This vector is fed to the LSTM input. Since the training set is a large dataset and each image is represented as a 2048-dimensional vector, the learning will be expensive. Therefore, the principal component method is used to reduce the dimension of the image vector from 2048 to 256. Since the length of the description may differ, the model should know where to start and stop. To do this, we add two tokens $< START >$ and $< END >$, which are the beginning and end of each sign.

The network for generating the captions will have to capture the words between these tokens. In this paper, words are represented as the frequency of occurrence of each word in the dictionary (1-of-N, where N is the power of the dictionary). The LSTM model learns to predict the next word S_t in the commentary based on the vector of visual features and the previous $t-1$ words. $p(S_t|I, S_1, S_2, \cdots, S_{(t-1)})$ is calculated and optimized on the whole training dataset by using stochastic gradient descent. At each time step, the context vector Z_t and the $h_{(t-1)}$ state of the previous step are fed to the LSTM together. After that, LSTM provides the next state vector h_t and next word. The context vector z_t is a concatenation of the feature vector and one hot vector of word representation.

4.2. Experimental Results. To speed up the learning process, we have adopted the method of Adam optimization with a gradual decreasing of learning rate which convergences more quickly. We use Adam optimization with regularization methods such as L_2 and dropout together. Applying the dropout technique in convolutional layers with a value of 0.5 and 0.3 in the LSTM layers helps to avoid overfitting that quickly happens with a small training set like the Flickr8K dataset. A variant with two LSTM layers is selected because we do not find that additional layers improve the quality. Each of the LSTM contains 512 hidden elements in a cell. Batch size equal to 32 and the beam size 3 are empirically found out that values are optimal. The deep models, such as ResNet50, for generating comments to the image increase in efficiency of the whole model. This is especially noticeable in the BLEU metric. Using a large set of MS COCO 2014 dataset avoids the model overfitting, while the overfitting on Flickr8K is achieved very quickly with a large batch size.

First, we study the impact of the soft attention mechanism in AICRL. As Table 2 indicates, the integrating the soft attention mechanism improves the model performances significantly. The soft attention mechanism increases the

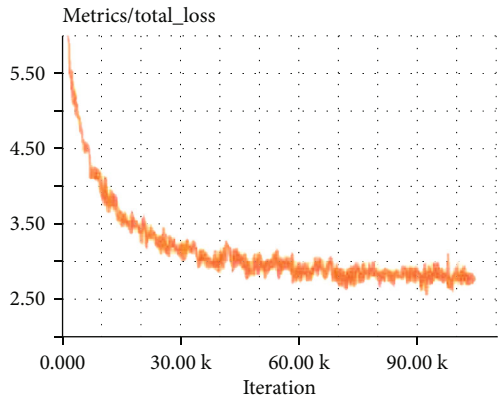

Figure 3: The total loss function.

performance in all metrics like BLEU-4, METEOR, and CIDEr. In addition, after training of the generator model, there are two questions. The first one is whether the model really generates new descriptions, and the second one is that whether they are diverse, qualitative, and understandable for humans. We have also conducted another set of experiments to involve human into the performance evaluation.

A questionnaire is designed with 20 images and the generated descriptions from the two different models. The participants are asked to evaluate whether the generated caption can well describe the images. Table 3 presents the results based on the generated description from the MS COCO 2014 dataset. From the results, we can see that 71% of the captions are well generated for the model with soft attention, while 54% are well generated for the one without soft attention. Based on this, we will use AICRL with the soft attention in the following experiments.

Next, we study the performance comparison between AICRL and other existing image captioning algorithms [22–29]. To make the evaluation complete, we also implemented another algorithm, AICRL-VGA16, by using another CNN network, namely, VGA16, in AICRL. Tables 4 and 5 show the results based on the Flick8K dataset and MS COCO 2014 dataset under six different metrics including BLEU-1, BLEU-2, BLEU-3, BLEU-4, METOER, and CIDEr. From both of the results, we can see that AICRL outperforms other systems in those metrics. The proposed model is able to generate efficient captions and fluent language. Meanwhile, ResNet50 also outperforms the VGA16 network which indicates that ResNet50 is able to capture the image features well. From these experiments, we observe that AICRL achieves good performance by integrating ResNet50, LSTM, and soft attention into a joint model.

Furthermore, we study how the total loss changes during the training. Figure 3 shows that the total loss of the model varies while the training iteration increases. From the results, we can see that the total loss quickly decreases at the beginning of training, but later, the speed of loss changing slows down.

5. Conclusions

In this paper, we have presented one single joint model for automatic image captioning based on ResNet50 and LSTM with software attention. The proposed model was designed with one encoder-decoder architecture. We adopted ResNet50, a convolutional neural network, as the encoder to encode an image into a compact representation as the graphical features. After that, a language model LSTM was selected as the decoder to generate the description sentence. Meanwhile, we integrated the soft attention model with LSTM such that the learning can be focused on a particular part of the image to improve the performance. The whole model is fully trainable by using the stochastic gradient descent that makes the training process easier. The experimental evaluations indicate that the proposed model is able to generate good captions for images automatically.

Data Availability

The data used to support this study are included within the article.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61771155), China MOE Project of Humanities and Social Sciences for Youth (Grant No. 14YJC630181), the Natural Science Foundation of Hubei Province (Grant No. 2017CFB592), Fundamental Research Funds for the Central Universities Harbin Engineering University (Grant No. 3072020CF0608), and Fundamental Research Funds for the Central Universities Zhongnan University of Economics and Law (Grant No. 2722020JCT032 and No. 2722020PY047). This research was also supported in part by Singapore Ministry of Education TIF grant (Grant No. MOE2017-TIF-1-G018), Singapore Institute of Technology MOE Ignition grant (Grant No. R-MOE-E103-D004), and Strategic Initiative Grant on Applied Data Science.

References

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi et al., "Every picture tells a story: generating sentences from images," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds., pp. 15–29, Springer, 2010.

[2] A. Graves, *Generating sequences with recurrent neural networks*, University of Toronto, 2013.

[3] A. Radford, L. Metz, and S. Chintala, *Unsupervised representation learning with deep convolutional generative adversarial networks*, ICLR, 2016.

[4] Y. Zhang, R. Gravina, H. Lu, M. Villari, and G. Fortino, "PEA: Parallel electrocardiogram-based authentication for smart healthcare systems," *Journal of Network and Computer Applications*, vol. 117, pp. 10–16, 2018.

[5] M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, 2018.

[6] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, "Improving image sentence embeddings using large weakly annotated photo collections," in *European Conference on Computer Vision*, pp. 529–545, Springer, 2014.

[7] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Workshop on Neural Information Processing Systems (NIPS)*, 2014.

[8] 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*, pp. 3156–3164, Boston, MA, USA, 2015.

[9] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *Proceedings of the 31st international conference on machine learning (ICML-14)*, pp. 595–603, Beijing, China, 2014.

[10] A. Karpathy and L. Fei-Fei, *Deep visual-semantic alignments for generating image descriptions*, Stanford University, 2017.

[11] K. Xu, J. Ba, R. Kiros et al., "Show, attend and tell: neural image caption generation with visual attention," in *International conference on machine learning*, pp. 2048–2057, Lile, France, 2015.

[12] S. M. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. J. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics*, pp. 220–228, Portland, Oregon, USA, 2011.

[13] G. Kulkarni, V. Premraj, S. Dhar et al., "Baby talk: understanding and generating image descriptions," in *CVPR means IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2891–2903, 2011.

[14] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2Text: Describing images using 1 million captioned photographs," *Advances in Neural Information Processing Systems*, pp. 1143–1151, 2011.

[15] C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2596–2604, Santiago, Chile, 2015.

[16] 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.

[17] X. Yu, Y. Chu, F. Jiang, Y. Guo, and D. Gong, "SVMs Classification based two-side cross domain Collaborative Filtering by inferring intrinsic user and item features," *Knowledge- Based Systems*, vol. 141, pp. 80–91, 2018.

[18] X. Yu, F. Jiang, J. Du, and D. Gong, "A cross-domain collaborative filtering algorithm with expanding user and item features via the latent factor space of auxiliary domains," *Patten Recognition*, vol. 94, pp. 96–109, 2019.

[19] 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 of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, 2002.

[20] S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translationand/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005.

[21] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, Boston, MA, USA, 2015.

[22] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, *Explain images with multimodal recurrent neural networks*, University of California, Los Angeles, 2014.

[23] X. Chen and C. L. Zitnick, *Learning a recurrent visual representation for image caption generation*, Stanford University, 2014.

[24] H. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched Long-Term Recurrent Convolutional Network for Facial Micro-Expression Recognition," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pp. 667–674, Xi'an, China, 2018.

[25] A. Mnih and G. Hinton, "Three new graphical models for statistical language modelling," in *ICML '07: Proceedings of the 24th international conference on Machine learning*, pp. 641–648, Corvalis, OR, USA, 2007.

[26] A. Karpathy and L. Fei-Fei, *Deep visual-semantic alignments for generating image descriptions*, Stanford University, 2015.

[27] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[28] 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*, pp. 2–7, Boston, MA, USA, 2015.

[29] H. Fang, S. Gupta, F. Iandola et al., "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014.