

# Image Caption Generation Using A Deep Architecture

Ansar Hani  
National Engineering School  
Sfax University  
Sfax, Tunisia  
ansar.hani@gmail.com

Najiba Tagougui  
Higher Institute of Computer Science  
and Multimedia,  
Sfax University  
Sfax, Tunisia  
tag.najiba@gmail.com,

Monji Kherallah  
Faculty of Sciences  
Sfax University  
Sfax, Tunisia  
monji.kherallah@fss.usf.tn

**Abstract**—Recently, image captioning is a new challenging task that has gathered widespread interest. The task involves generating a concise description of an image in natural language and is currently accomplished by techniques that use a combination of computer vision (CV), natural language processing (NLP), and machine learning methods.

In this paper, we presented a model that generates natural language description of an image. We used a combination of convolutional neural networks to extract features and then used recurrent neural networks to generate text from these features. We incorporated the attention mechanism while generating captions. We evaluated the model on MSCOCO database. The obtained results are promising and competitive.

**Keywords**— *image captioning, convolutional neural networks, recurrent neural networks, attention mechanism*

## I. INTRODUCTION

The rapid growth of technology with the spread of Internet use has leads to the availability of large collections of digital images that are a real fortune for the research community working on machine vision (MV) field. The major problem with this big amount of data is that it is basically unlabeled and this leads to untapped resources. Therefore, many recent works are actually dealing with the automatic image caption generation with the main purpose of giving meaning to these images to be used in advanced artificial intelligence applications. It is a real challenging task which will have a huge impact in improving research results. The Image captioning field has been around for the last ten years [1,2,3,4,5,6,7] But the efficiency of the used techniques was limited at the beginning and they weren't robust enough. Everything has changed with firstly the free accessibility to huge databases like ImageNet [8], Flickr 8k [9], Flickr 30k[10] and the Microsoft COCO: Common Objects in Context(MS COCO) [11] and secondly the use of encoder-decoder framework [3,5,6,7] inspired from the success of neural networks in MV tasks especially the famous architecture deep Convolutional Neural Networks (CNN) [8,12].

This new framework is adopted by many recent researches and it has proven her effectiveness [1,2,3,4,5,6,7]. The CNNs are used as an encoder to extract images features that are next fed to Recurrent Neural Networks

(RNNs)[13,14,15] for language modeling. The major drawback in this architecture is that it doesn't consider the spatial aspects of the image and automatically generates captions for images by considering the image's scene as a whole. To overcome this main limitation, the attention mechanism [13] has been proposed to be incorporated to the encoder-decoder framework.

We tried among this current study to implement a system that will be able to generate images caption using an encoder/decoder system using CNN and RNN with attention mechanism. Results were promising and competitive.

In this paper, we started by discussing related works in section2 then we exhibited in section 3 our proposed architecture after giving the justification of such choice. Experimental results are detailed in section 4 and finally a conclusion summarizing the all is presented in section5.

## II. RELATED WORKS

Current image captioning has witnessed a rapid growth from initial retrieval-based methods to the current one based on deep neural networks. We can classify these approaches into 3 categories, which are as follows.

### 1. Retrieval -based image captioning:

This sort of methods treats image captioning as retrieval task. By leveraging distance metric to retrieve similar captioned images, then modify and combine retrieved captions to generate caption [16]. But these approaches generally need additional procedures such as modification and generalization process to fit image query.

### 2. Template based image captioning:

The basic idea behind this strategy is the set of predefined objects and actions that are defined as templates. It acts by detecting actions, scenes, objects and attributes in the target image and then to fill the obtained data information into blank spaces of the sentence template to form captions. This technique has been used by Kulkarni et al. Their method [17] uses an object detector to detect objects in the image, and then sends candidate object regions into attribute classifier and prepositional relation function to obtain attribute information

of candidate objects and prepositional relation information between objects. Furthermore, a Conditional Random Field

(CRF) is adopted to predict the best labelling for an image. Template-based methods can generate grammatically correct captions. However, they cannot generate variable-length captions.

### 3. Deep learning based image captioning:

Inspired by recent advances in deep learning, recent works begin to rely on deep neural networks for image captioning. Usually, CNN is used as an encoder to extract

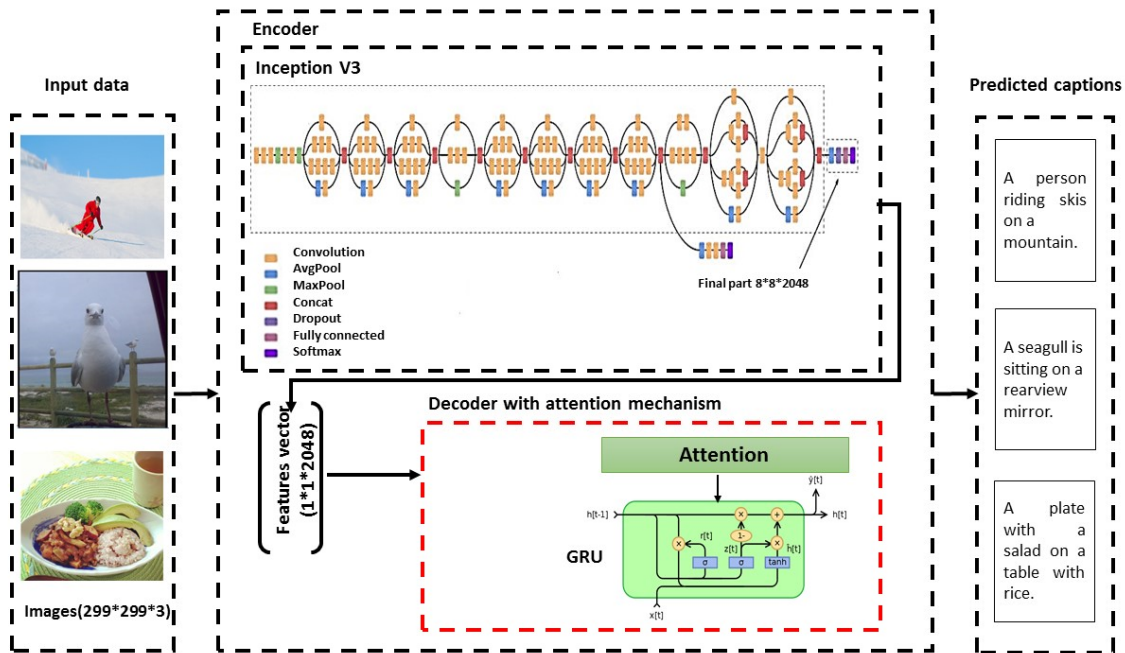


Fig. 1. Proposed architecture

information from images. CNN are widely followed by RNN. RNN is used as decoder to convert this representation into natural language description.

However, the encoder has to compress all the input information into a single fixed length vector that is passed to the decoder. Using this vector to compress long and detailed input sequences may lead to loss of information [18].

Attention mechanism has been proposed to handle this complexity. In the attention algorithm, an attention vector related to time  $t$  is used to replace the fixed length vector obtained from the image encoder CNN.

In fact, Attention mechanism is incorporated to the encoder-decoder image captioning framework to allow the decoding process to focus on the emphasis and details of the input image at each time step while the output sequences are being produced [13]. Xu et al [19] were the first to introduce an attentive encoder-decoder model. Many useful improvements are proposed on encoder-decoder structure such as semantic attention [20] and review network [21]. You et al. proposed a novel algorithm that combines the bottom-up and top-down approaches. This is achieved through a semantic attention model, which combines semantic concepts and the feature representation of the image/encoding.

For yang et al. [21], they propose a novel module called “the reviewer module” in purpose to improve the encoder-decoder framework. This module performs review steps on the hidden states of the encoder and gives a vector at each step while attention mechanism is applied to determine weights assigned to hidden states. The method of Anderson et al.[22] also adapted both bottom-up and top-down attention

mechanism that enables attention to be calculated at the level of objects and other salient image regions. For Su et al. [23], they proposed a hierarchical deep neural network. The proposed architecture consists of the bottom layer and the top layer. The bottom layer extracts the visual and high level semantic information from image and detected regions, while the top layer integrates both of them with attention mechanism for the caption generation.

An overview of Deep learning based image captioning methods is described in table I

In this work, we have proposed an attention based encoder-decoder model. Firstly, we have used a pretrained model to interpret the context of images. We opted for the Inception V3 model [26] to extract features of the image by saving the output of the last hidden layer. Also, we have used Gated Recurrent Unit (GRU)[27] instead of long Short-Term Memory (LSTM) [28] due to its several advantages as described in details in [27].

The visual semantic attention model used has to automatically decrease the weights of irrelevant words to just focus on one single word that is judged the most coherent with the current context. This will lead to a better performance that has been shown via extensive experiments by eliminating the influence of non-relevant attributes word.

### III. PROPOSED ARCHITECTURE

As we have explained in section 2, we started by adopting an encoder-decoder architecture that incorporates visual attention mechanism to generate image captioning. The encoder part is CNN based and the decoder one uses the

visual attention module. The proposed architecture is illustrated within Fig. 1.

Suppose  $\{S_0, \dots, S_{T-1}\}$  is a sequence of words in a sentence of length  $T$ , the model aims to directly maximise the probability of the correct description given an image.

Thus, the optimization problem can be formulated by

$$\theta^* = \arg \max_{\theta} \sum \log p(S_i | I_i; \theta) \quad (1)$$

Where  $\theta$  represents the parameters of the model,  $I_i$  is an image and  $S$  is the generated description.

Table I

Summary of deep learning based image captioning

Deep learning based image captioning				
Method	Ref	Dataset	Image encoder	Decoder
Encoder-Decoder	Karpathy et al. [3]	Flickr8k, Flickr30k, MSCOCO	VGGNet	RNN
	Vinyals et al.[6]	Flickr30k, MSCOCO	GoogleLeNet	LSTM
	Mao et al.[5]	IAPRTC,Flickr8k, Flickr30k, MSCOCO	AlexNet, VGGNet	RNN
	Zhang et al.[7]	MSCOCO	Inception V3	LSTM
Encoder-Decoder with Attention	Xu et al. [19]	Flickr8k, Flickr30k, MSCOCO	AlexNet	LSTM
	You et al. [20]	Flickr30k, MSCOCO	GoogleLeNet	RNN
	Yang et al. [21]	Visual Genome Dataset	VGGNet	LSTM
	Anderson et al. [22]	Visual Genome Dataset, MSCOCO	ResNet	LSTM
	Su et al. [23]	MSCOCO	ResNet	LSTM
	WANG et al. [24]	MSCOCO	VGGNet	CNN
	Tan et al. [25]	MSCOCO	GoogleLeNet	GRU +LSTM

### 1. The Encoder part

Under the encoder-decoder framework for image captioning, a CNN can produce a rich representation of the input image by embedding it to a fixed length vector representation.

Many different CNN can be used, e.g., VGG[29], Inception V3[26], ResNet [30]. In this paper, we use Inception V3 model created by Google Research as encoder. This model was pre-trained on ImageNet dataset where it was the first runner up for image classification in ILSVRC 2015. We have removed the last layer of the model as it is used for classification. We have pre-processed images with the Inception V3 model and have extracted features.

The extractor produces  $L$  vectors, each of which is a  $D$ -dimensional representation corresponding to a part of the image:

$$a = \{a_i, \dots, a_L\}, a \in \mathbb{R}^D \quad (2)$$

Global image feature can be obtained by:

$$a_g = \frac{1}{L} \sum a_i \quad (3)$$

Image Vector and Global Image Vector can be obtained by using a single layer perceptron with rectifier activation function:

$$v_i = \text{ReLU}(W_a a_i) \quad (4)$$

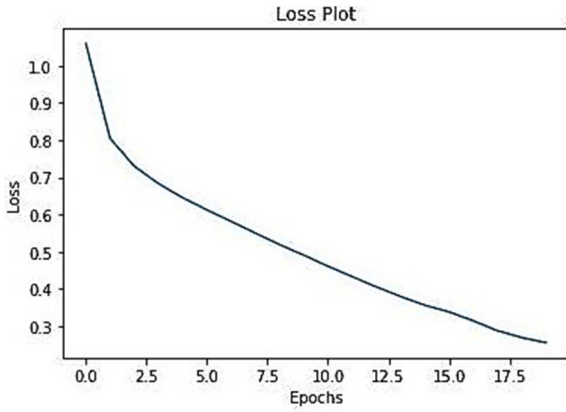
$$v_g = \text{ReLU}(W_g a_g) \quad (5)$$

where  $W_a$  and  $W_g$  are the weight parameters,  $L$

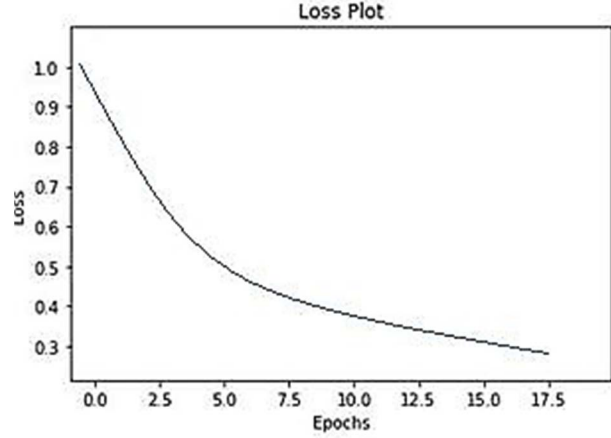
is number of vectors and  $D$  is size of each vector.

The transformed spatial image feature form is  $V =$

$[v_1, \dots, v_L]$



(a) Loss presentation without ADAM



(b) Loss presentation with ADAM

Fig. 2. Loss with and without ADAM

#### Attention mechanism

For image captioning, attention tends to focus on specific regions in the image while generating descriptions.

At time  $t$ , based on the hidden state, the decoder would attend to the specific regions of the image and compute context vector using the spatial image features from a convolution layer of a CNN

$$c_t = g(v, h_t) \quad (6)$$

We feed  $V$  and  $h_t$  through a single layer neural network followed by a softmax function to generate the attention distribution over the  $k$  regions of the image

$$Z_t = W_h^T \tanh(W_v v + (W_g h_t) 1^T) \quad (7)$$

$$\alpha_t = \text{softmax}(Z_t) \quad (8)$$

Where  $1^k$  is a vector with all elements set to 1.  $W_v, W_g \in \mathbb{R}^{L \times D}$  and  $W_h \in \mathbb{R}^{L \times L}$  are parameters to be learnt.  $\alpha \in \mathbb{R}^L$  is the attention weight over features in  $V$ . Based on the attention distribution, the context vector  $c_t$  can be obtained

$$c_t = \sum (\alpha_{ii} v_{ii}) \quad (9)$$

#### 2. The Decoder part

Given the image representations, a decoder is employed to translate the image into natural sentences. A decoder is a RNN which are typically implemented using either LSTM or GRU.

Here we have used GRU as a decoder which has a simpler structure than LSTM.

Also, unlike RNN, GRU does not suffer from the vanishing gradient problem.

$X_t$  is the input vector and we obtain it by concatenating the word embedding vector,  $W_t$ , global image feature vector,  $v_g$ , and context vector,  $c_t$ , to get the input  $v$

$$X_t = [W_t; v_g; c_t] \quad (10)$$

## IV. EXPERIMENTS AND RESULTS

### 1. Data Selection

Generally, getting data for experimental studies is problematic. Blessedly, it is not the case in our context. There are at least three popular open source datasets that can be used to train deep network models like Flickr 8k [9], Flickr 30k[10] and MS COCO[11]. We decided to work with the MS-COCO database which represents a large dataset of images with multiple human-label descriptions of said images to train our model. This choice is not arbitrary. In fact, the images in previous datasets are iconic and do not capture the settings in which these objects usually co-occur. To remedy this problem, MS-COCO annotated real-world scenes that capture object contexts. This dataset with both captions and region-level annotations is containing over 300k images. It has 91 common object categories such as person, car, bus, etc with 82 having more than 50000 labelled instances. It is designed for object detection, segmentation and caption generation. It has instance level segmentation which means that the dataset has every instance of every object category labelled and fully segmented.

The same number of captions is maintained such as the previous datasets. For each image, 5 independent user generated sentences describing the scene are provided. Captions of test images are unavailable publicly. This dataset poses great challenges to the image captioning task. The images create a testbed for image captioning since most images contain multiple objects and significant contextual information [31]

An additional 40k test images were released in the 2015 version of the dataset. However, it was unable to describe all

the objects in the image, since they annotate only 91 objects categories. The dataset is available at [32].

For our model, we trained a relatively small amount of data—the first 30,000 captions for about 24,000 images. This limitation is basically the result of using a single gamer

machine which is not really sufficient to ensure dealing with huge data. We have tried to select a collection of images that


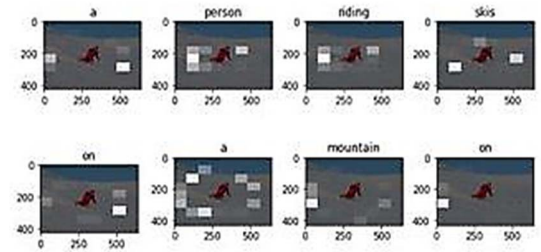

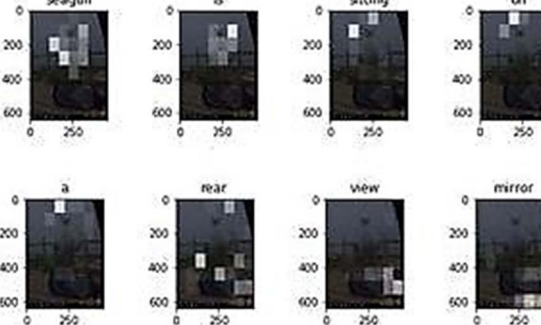

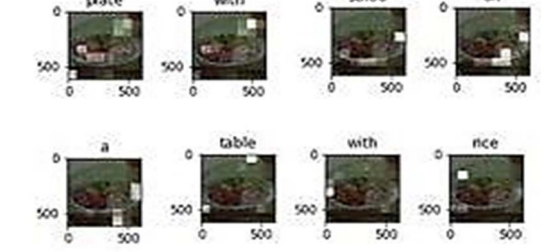
Original image	Plotting attention	Captions
		<p><b>Real caption:</b> A person in a red suit down the mountain.</p> <p><b>CaptionBot:</b> A man riding skis down a snow covered slope</p> <p><b>Ours:</b> A person riding skis on a mountain .</p>
		<p><b>Real caption:</b> A white bird sits on the rearview mirror of a car.</p> <p><b>CaptionBot:</b> A bird standing in front of a body of water</p> <p><b>Ours:</b> A seagull is sitting on a rearview mirror.</p>
		<p><b>Real caption:</b> A close up of a &lt;unk&gt; of food and a drink.</p> <p><b>CaptionBot:</b> Plate of food and a cup of coffee</p> <p><b>Prediction caption:</b> A plate with a salad on a table with rice.</p>

Fig. 3. Some examples of generated captions by our model

represent almost all the categories available in the database to allow the generalization of the obtained results.

## 2. Implementation details

The Model is implemented using TensorFlow 2.0 .We opt for Inception-V3 model with batch normalisation pre-trained on ImageNet as an encoder. The input images are resized to  $299 \times 299$ , before being fed to the CNN. We take the  $8 \times 8 \times 2048$  activations map of the last convolutional layer as annotations. A GRU network with hidden state size of  $r = 512$  is used. The word size is set to  $q = 256$  dimensions.

During training, we adopt ADAM [33] for optimization, with batch size of 64. We trained the model for 20 epochs. The loss values for some epochs are presented in Fig. 2.

Our model was trained by minimizing the loss function using an optimization method called Stochastic Gradient Descent (SGD). To minimize the loss, the optimizer needs the gradient of the loss function which tells the optimizer how much and in which direction it needs to adjust each model's parameter.

## 3 Results

Here are some output results from running the model on the validation part of the MS-COCO database.



The obtained results prove that our system is competitive. For our proposed architecture, ADAM adaptive gradient optimizer is a good choice which is giving us good accuracy with faster rate of convergence.

We compare our method with CaptionBot [34] an online caption generator created by Microsoft. The results that we obtained are better than those given Microsoft caption generator as shown in fig. 3.

## V. CONCLUSION

We have presented an attention-based image captioning model that can generate a description from a given image. Our model is based on a convolution neural network to encode an image into a fixed length vector representation which is further used by attention layer to produce context vector which is later decrypted by the use of the recurrent neural network. The obtained results are promising and competitive compared to results presented in the literature.

For future work, we plan to implement more compact deep convolution network with the use of the BLEU metric [35] to better enhance the performance of the proposed model. More investigations should be done to the natural language processing when dealing with sentence generation to ensure a better caption formulation.

## REFERENCES

- [1] A.Karpathy, A.Joulin and L.F.Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," *In Advances in neural information processing systems*, pp. 1889-1897, 2014.
- [2] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," *In Advances in neural information processing systems*, pp. 3104-3112, 2014.
- [3] A.karpathy, L.F.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, 2014.
- [4] R. Kiros, R. Salakhutdinov and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.
- [5] J. Mao et al. "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.
- [6] 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, 2014.
- [7] L.Zhang et al., "Actor-critic sequence training for image captioning," *arXiv preprint arXiv:1706.09601*, 2017.
- [8] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *In Advances in neural information processing systems*, pp. 1097-1105, 2012.
- [9] M. Hodosh, P. Young, 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.
- [10] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, "From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions," *in: Meeting on Association for Computational Linguistics*, vol. 2 pp. 67-78, 2014.
- [11] T.Y. Lin et al., "Microsoft coco: Common objects in context," *In European conference on computer vision*, Springer, Cham, pp. 740-755, 2014.
- [12] M.D. Zeiler, R. Fergus, "Visualizing and understanding convolutional networks," *In European conference on computer vision*, Springer, Cham, pp. 818-833, 2014.
- [13] D. Bahdanau, K. Cho and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [14] T. Mikolov, M. Karafiát, L. Burget, J. Černocký and S. Khudanpur, "Recurrent neural network based language model," *In Eleventh annual conference of the international speech communication association*, 2010.
- [15] T. Mikolov, S. Kombrink, L. Burget, J. Černocký and S. Khudanpur, "Extensions of recurrent neural network language model," *In 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 5528-5531, 2011.
- [16] P. Kuznetsova, V. Ordonez, T.L. Berg and Y. Choi, "Treetalk: Composition and compression of trees for image descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 351-362, 2014.
- [17] G.Kulkarni et al., "Baby talk: Understanding and generating image descriptions," *In Proceedings of the 24th CVPR. Citeseer*, 2011.
- [18] J. K.Chorowski, D. Bahdanau, D.Serdyuk, K.Cho and Y.Bengio, "Attention-based models for speech Recognition," *Advances in neural information processing systems*, pp. 577-585, 2015.
- [19] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," *In International conference on machine learning*, pp. 2048-2057, 2015.
- [20] Q. You, H. Jin, Z. Wang, C. Fang and J. Luo, "Image captioning with semantic attention," *In: CVPR*, pp. 4651-4659, 2016.
- [21] L. Yang, K.D. Tang, J. Yang and, L.J. Li, "Dense captioning with joint inference and visual context," *In: CVPR*, pp. 1978-1987, 2017.
- [22] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," *In CVPR*, 6077-6086, 2018.
- [23] Y. Su, Y. Li, N. Xu, and A.A. Liu, "Hierarchical Deep Neural Network for Image Captioning," *Neural Processing Letters*, pp. 1-11, 2019.
- [24] Q. Wang, A.B. Chan, "CNN+ CNN: convolutional decoders for image captioning," *arXiv preprint arXiv:1805.09019*, 2018.
- [25] J. H. Tan, C. S. Chan and J. H. Chuah, "Image Captioning with Sparse Recurrent Neural Network," *arXiv preprint arXiv:1908.10797*, 2019.
- [26] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the inception architecture for computer vision," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826, 2016.
- [27] K. Cho et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [28] S. Hochreiter, I. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [29] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778, 2016.
- [31] H. Fang et al., "From captions to visual concepts and back," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1473-1482, 2015.
- [32] <http://cocodataset.org/>
- [33] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [34] K. Tran et al., "Rich image captioning in the wild," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 49-56, 2016.
- [35] 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, pp. 311-318, 2002.