The Role of Attention Mechanism and Multi-Feature in Image Captioning

Tien X. Dang
School of Electronics and Computer Engineering
Chonnam National University
South Korea
dxtien95@gmail.com

Aran Oh
School of Electronics and Computer Engineering
Chonnam National University
South Korea
dhdkfks9@naver.com

In-Seop Na
Software Convergence Education Institute
Chosun University
South Korea
ypencil@hanmail.net

Soo-Hyung Kim
School of Electronics and Computer Engineering
Chonnam National University
South Korea
shkim@jnu.ac.kr

ABSTRACT

Up to now, caption generation is still a hard problem in artificial intelligence where a textual description must be generated for a given image. This problem combines both computer vision and natural language processing. Generally, the CNN - RNN is a popular architecture in image captioning. Currently, there are many variants of this architecture, where the attention mechanism is an important discovery. Recently, deep learning methods have achieved state-of-the-art results for this problem. In this paper, we present a model that generates natural language descriptions of given images. Our approach uses the pre-trained deep neural network models to extract visual features and then applies an LSTM to generate captions. We use BLEU scores to evaluate our model performance on Flickr8k and Flickr30k dataset. In addition, we carried out a comparison between the approaches without attention mechanism and attention-based mechanism.

CCS Concepts

• **Computing methodologies→Natural language generation**
• **Computing methodologies→Computer vision**

Keywords

Image captioning, attention mechanism, CNN, RNN, LSTM.

1. INTRODUCTION

Image captioning combines two main fields of artificial intelligence include computer vision and natural language processing. Thus, it is a truly challenging problem in artificial intelligence. Automatically describing the content of an image is a challenging task, but the effect is great, for example, it helps the visually impaired people can understand the content of images and motivate human-robot department is developed. In order to produce high-quality captions, the model must understand visual clues from the image. Generally, the images are encoded by convolutional neural networks and recurrent neural networks like Long Short-Term Memory network (LSTM) [1] or Gated Recurrent unit network (GRU) [2], is used to generate captions. Recently, deep learning is expanded, especially in deep convolutional neural networks or CNN for short. There are have many state-of-the-art, namely VGG networks [3], GoogLeNet (Inception) [4, 5, 6], ResNet [7].

In this paper, we follow a general framework: CNN for features extractor combines with LSTM for caption generator. We apply three approaches 1) single feature approach which apply one pre-trained CNN model to extract image visual features, 2) multi-features approach which combines two kinds of image features are extracted by two pre-trained CNN models, 3) using multi-features and adding the attention mechanism. After extracting features, we push all of them to an LSTM layer to generate captions. We evaluate model performance on two datasets namely Flickr8k [8] and Flickr30k [9]. Each dataset we split into three parts includes training, validation, and testing. The bilingual evaluation understudy (BLEU) [10] scores are used to evaluate the skill of the model. Our BLEU-1 scores on Flickr8k and Flickr30k test dataset are 0.51 and 0.51, respectively.

In this research, the main contributions are implementation the multi-feature (feature fusion) in image captioning and comparison among our approaches single feature, multi-feature without attention and multi-feature with attention mechanism.

The rest of this paper can be organized as follows. In Section 2, we first revise current works in image captioning. We introduce the method in Section 3. All experiments are presented in Section 4. Finally, we give our conclusion and further work in the last two sections.

2. RELATED WORK

Nowadays, image captioning which attract a considerable amount of attention has many important applications in several research fields. There are have been many approaches and datasets related to the image captioning problem.

Commonly, a large of all method [11, 12, 13, 14, 15, 16] follow a CNN-RNN architecture: the image features are extracted by a CNN trained on the large dataset, such as ImageNet [17], in images classification task, and then a recurrent neural network is applied to predict descriptions of an image based on image features. Moreover, some method [11, 12, 18] add the result of object detection from R-CNN or its variant into the CNN-RNN framework to obtain the higher performance. Otherwise, there are have several methods [19, 20, 21] adopt a multi-modal framework

ICMLSC 2019, January 25–28, 2019, Da Lat, Viet Nam

ACM ISBN 978-1-4503-6612-0/19/01…$15.00
https://doi.org/10.1145/3310986.3311002

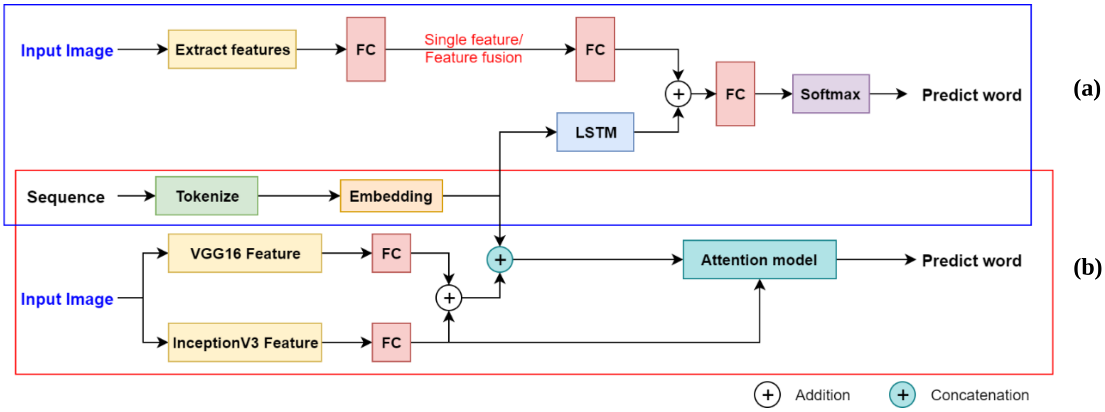


Figure 1. Proposed approaches, (a) The single and multi-feature approach in blue area and (b) The multi-feature with attention approach in red area.

that predicts the caption word by word based on the multi-modal embedding in which language features and image features are embedded in a multi-modal space.

Latterly, there are have many approaches [22, 23, 24, 25] apply attention mechanism to learn a latent alignment from scratch when generating corresponding words.

For image captioning dataset, we have three popular datasets involve: Flickr8k[8], Flickr30k[9] and MSCOCO[26] for image captioning.

- **Flickr8k**, Hodosh et al. [8] introduced a dataset called Flickr8k which contains 8000 images and contains sets accommodating different sources of descriptions for each image.
- **Flickr30k**, P. Young et al. [9] in their paper they present Flickr30k dataset contains 31783 images. Most of these images illustrate humans activities in real-life. Each image in both of dataset is paired with five descriptions.
- **MSCOCO**, Microsoft also released a big dataset with name MSCOCO [26]. MSCOCO is large-scale object detection, segmentation, and captioning dataset. For image captioning, the dataset contains 83000, 41000, 41000 images for training, validation, and testing respectively.

3. PROPOSED METHOD

We first describe the feature extractors using CNN trained models in Sec. 3.1, then introduce the language model using LSTM and evaluation metrics bilingual evaluation understudy (BLEU) [10] in Sec. 3.2 and Sec. 3.3.

3.1 Feature Extractors

In this research, we used two famous CNN trained models namely: VGG 16-layer [3] and InceptionV3 [5]. These are powerful models which are trained on the large dataset, ImageNet [17]. Each model, we remove the last layer to receive the high-level features. The high-level features of VGG16 are specific and represent for the object in an image. It is needed in exactly determine what is description related to. However, we not only need specific features but also need low-level or middle-level features that are a reason why we chose InceptionV3. InceptionV3 with module architecture allow it to keep the information in lower levels. Each image, the feature vectors receive by applying VGG16 and InceptionV3 have 4096 dimensions and 2048 dimensions, respectively. Feature vector:

$$\text{VGG16}\left(x_1^{(1)}, x_2^{(1)}, \dots, x_{4096}^{(1)}\right)$$

$$\text{InceptionV3}\left(x_1^{(2)}, x_2^{(2)}, \dots, x_{2048}^{(2)}\right)$$

In the single feature approach, the features computed by a fully-connected layer are combined with the output of an LSTM layer. In the feature fusion method, we merge both types of feature VGG16 and InceptionV3 after we fed them into the first fully connected layers. The formula of addition is shown in Figure 1

$$\left(x_1^{(1)}, x_2^{(1)}, \dots, x_n^{(1)}\right)^T + \left(x_1^{(2)}, x_2^{(2)}, \dots, x_n^{(2)}\right)^T = \left(x_1^{(1,2)}, x_2^{(1,2)}, \dots, x_n^{(1,2)}\right)^T \qquad (1)$$

Where n is the dimension of vector, in this case $n = 512$. After fusion step, we push them into language model.

3.2 Language Model

3.2.1 *Single LSTM language model*

Each word in a dataset is embedded with zero elements to creating vector the same length. Then, it is fed to an LSTM layer and merged with image features. After that, we obtain the blending features between image and word.

$$\left(x_1^{(F)}, x_2^{(F)}, \dots, x_n^{(F)}\right)^T + \left(x_1^{(L)}, x_2^{(L)}, \dots, x_n^{(L)}\right)^T = \left(x_1^{(F,L)}, x_2^{(F,L)}, \dots, x_n^{(F,L)}\right)^T \qquad (2)$$

Where F is image features, L is the output of the LSTM layer, n is the dimension of vector, in this case $n = 512$. After that, the blending feature is pushed into a fully-connected layer and then through a softmax layer to predict word. (See Figure 1a)

3.2.2 *Attention model*

The vector, VT, is a vector of concatenated multi-feature with embedded words. We have:

$$VT = [V, T] \qquad (3)$$

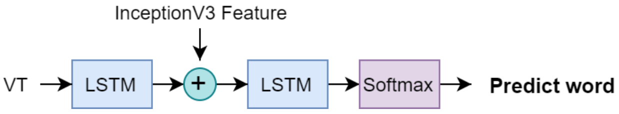


Figure 2. The attention model.

Where, V is a vector of multi-feature and T is a vector of embedded word. The attention model is described in the Figure 2.

The first LSTM layer tries to generate a coarse caption based on the input, VT. The input of the second LSTM layer is a concatenation of the InceptionV3 feature and the output of the first LSTM layer. The role of the second LSTM like an advisor, which adjust the coarse caption to a better one. For example, the first LSTM layer is a learner and the second LSTM layer is a teacher. When the learner told the teacher what it saw and then the teacher fixes the mistake in sentences for the learner. Finally, the learner can make better sentences.

3.3 BLEU score

The bilingual evaluation understudy, or BLEU for short, was proposed by Kishore et al. in their paper "BLEU: a Method for Automatic Evaluation of Machine Translation" [10]. It is the most common metric for evaluating a generated sentence to a reference sentence based on n-grams ($n = [1, 2, 3, 4]$). The BLEU metric ranges from 0 to 1 and the higher, the better. Their approach is counting matching n-grams in the generated sentence to n-grams in the reference text, the word order is neglected. The formular:

$$BLEU_n(gen, des) = \frac{\sum_{w_n \in gen} \min\left(C_{gen}(w_n), \max_{j=1,\dots,k} C_{des_j}(w_n)\right)}{\sum_{w_n \in gen} C_{gen}(w_n)} \quad (3)$$

Where gen is generated caption, des is set of reference sentences, k is number of sentences in des, w_n is n-grams and $C_x(y_n)$ is count of n-grams y_n in sentence x.

4. IMPLEMENTATION

4.1 Datasets

The datasets used for all experiments were Flickr8k [8] and Flickr30k [9]. All two datasets consist of images combined with five manually written captions per image. We split them into three set namely training set, validation set and testing set (The details are shown in Table 1).

Table 1. Detailed of dataset

Dataset name	Size		
	Train	Validation	Test
Flickr8k	6000	1000	1000
Flickr30k	25700	3000	3000

4.2 Experiments

4.2.1 Pre-processing dataset

We prepare image data and text data before training model.

- **Prepare image data.** Image features are extracted by pre-trained VGG16 layers and InceptionV3 model, which are available in Keras [29]. Both these models are trained on the ImageNet [17] dataset.
- **Prepare text data.** Tokenize, convert all words to lowercase, remove punctuation, remove all character word (e.g. 'a'), remove all words containing number.

4.2.2 Training detail

We trained our networks using Keras with TensorFlow backend. Each model has trained in around 1000 iterations on the training datasets. The weights were initialized randomly before updating by Adam algorithm [30] with minibatch size of 64 or 1024. The learning rate started from 3×10^{-4} and then reduced to 10^{-6}. The learning rates are changed to avoid overfitting. After training, we used our models to generate image captions. Finally, the BLEU metrics are applied for performance evaluation.

4.3 Results

After training, we evaluate the performance on testing sets by BLEU score. The best results are shown in Table 2 and 3. Additionally, we will depict some examples in Figure 3. The multi-feature with attention mechanism approach produce meaningful and rich captions. Therefore, the performance of this approach better than the others.

Table 2. Performances on Flickr8k testing set

Model name	BLEU score			
	B-1	B-2	B-3	B-4
VGG16 + LSTM	0.47	0.28	0.19	0.08
InceptionV3 + LSTM	0.48	0.27	0.18	0.08
Multi-feature	0.50	0.30	0.20	0.09
Attention model	**0.51**	**0.31**	**0.22**	**0.10**

Table 3. Performances on Flickr30k testing set

Model name	BLEU score			
	B-1	B-2	B-3	B-4
VGG16 + LSTM	0.49	0.30	0.20	0.10
InceptionV3 + LSTM	0.48	0.29	0.20	0.09
Multi-feature	0.49	0.28	0.20	0.09
Attention model	**0.51**	**0.30**	**0.20**	**0.10**

5. CONCLUSION

In this paper, we have presented a potential approach for image captioning. We have conducted experiments using the single feature approach, multi-feature approach and multi-feature with attention mechanism approach. The model which use attention mechanism get better results than other ones. In the future, we will improve the model performance by applying some another state-of-the-art CNN model such as Xception [27], ResNet [7] and in the language model, we will consider implementing Word2Vec [28] for word embedding.

6. ACKNOWLEDGMENTS

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A3A03000947, NRF-2017R1A4A1015559)

7. REFERENCES

[1] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8, 1735–1780.

[2] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

[3] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[4] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception

architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[6] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning.. In AAAI, Vol. 4. 12.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[8] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47, 853–899.

[9] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2, 67–78.

[10] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, 311–318.

[11] Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*. 1889–1897.

[12] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[13] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2018. Where to put the image in an image caption generator. *Natural Language Engineering* 24, 3, 467–489.

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[15] Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2017. What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator? *arXiv preprint arXiv:1708.02043*.

[16] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 248–255

[18] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, Vol. 3. 6.

[19] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–10.

[20] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *International Conference on Machine Learning*. 595–603.

[21] Junhua Mao, Xu Wei, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan L Yuille. 2015. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE International Conference on Computer Vision*. 2533–2541.

[22] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 6. 2.

[23] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.

[24] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[25] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV*. 22–29.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740 755

[27] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. arXiv preprint, 1610–02357.

[28] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efcient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[29] Keras Model API. https://keras.io/models/model/

[30] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

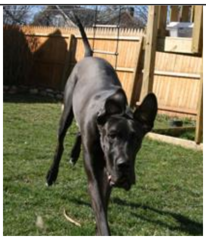
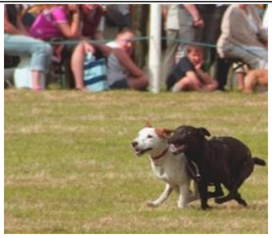
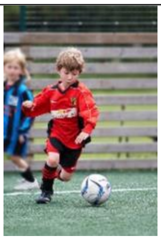
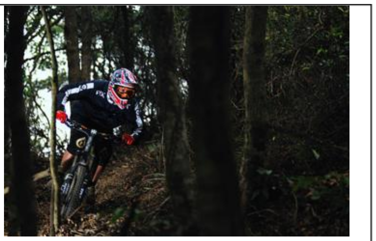
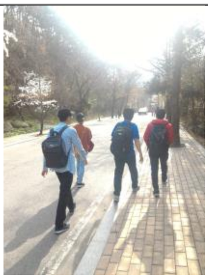
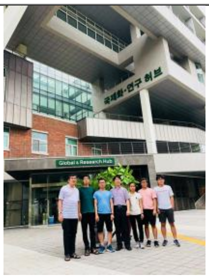
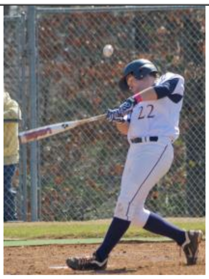
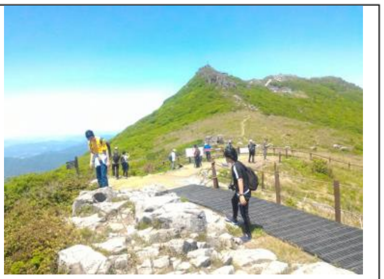
				
(a)	**GT: black dog running in backyard** **VGG16 + LSTM:** brown dog is running through the grass **InceptionV3 + LSTM:** **black** dog is running through the grass **Multi-feature:** dog is running through the grass **Attention model:** **black** dog is running on the grass	**GT: black and white dog are running on the grass** **VGG16 + LSTM:** two dogs are running through the grass **InceptionV3 + LSTM:** two dogs are playing in the grass **Multi-feature:** two dogs are running through the grass **Attention model:** two dogs are running in field	**GT: boy soccer player running down the field** **VGG16 + LSTM:** two men are playing in the air **InceptionV3 + LSTM:** two boys playing soccer in the air **Multi-feature:** two boys are playing soccer **Attention model:** two soccer players are playing soccer	**GT: mountain biker in red striped helmet rides through the trees** **VGG16 + LSTM:** man is jumping through the air **InceptionV3 + LSTM:** person in red helmet is its bike through the onto **Multi-feature:** man in red helmet rides his bike **Attention model:** man in red helmet **is riding** bike
				
(b)	**Multi-feature:** two people are walking down the street **Attention model:** group of people are walking down path	**Multi-feature:** group of people are standing in front of building **Attention model:** group of people are standing in front of face	**Multi-feature:** baseball player is playing baseball **Attention model:** baseball player **in white uniform** is playing baseball	**Multi-feature:** man in blue shirt and black pants is walking down the hill **Attention model:** man is standing on top of mountain

Figure 3. Some examples generated by our models: the single models, multi-feature without attention and multi-feature with attention (Attention model). (a) The images which have the ground-truth (GT) caption are from testing dataset, and (b) The images from other sources without ground-truth caption.