

Natural Language Description Generation for Image using Deep Learning Architecture

Phyu Phyu Khaing¹, Mie Mie Aung², Myint San³

¹Assistant Lecturer, ^{2,3}Lecturer

¹Faculty of Information Science, Myanmar Institute of Information Technology, Mandalay, Myanmar

^{2,3}Faculty of Information Science, University of Computer Studies, Monywa, Myanmar

How to cite this paper: Phyu Phyu Khaing | Mie Mie Aung | Myint San "Natural Language Description Generation for Image using Deep Learning Architecture"

Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019,



IJTSRD26708

pp.1575-1581,

<https://doi.org/10.31142/ijtsrd26708>

Copyright © 2019 by author(s) and International Journal of Trend in Scientific Research and Development Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (<http://creativecommons.org/licenses/by/4.0>)



ABSTRACT

Automatic natural description generation of an image is currently a challenging task. To generate a natural language description of the image, the system is implemented by combining with the techniques of computer vision and natural language processing. This paper presents different deep learning models for generating the natural language description of the image. Moreover, we discussed how the deep learning model, which works for the natural language description of an image, can be implemented. This deep learning model consists of Convolutional Neural Network (CNN) as well as Recurrent Neural Network (RNN). The CNN is used for extracting the features from the image and RNN is used for generating the natural language description. To implement the deep learning model in generating the natural language description of an image, we have applied the Flickr 8K dataset and we have also evaluated the performance of the model using the standard evaluation matrices. These experiments show that the model is frequently giving accurate natural language descriptions for an input image.

KEYWORDS: *natural language description, computer vision, natural language processing, deep learning model*

1. INTRODUCTION

The main communication of people is the words that express the language written or spoken. Another communication of people is images or signs for the physically challenged people.

The natural language description generation of image is also a challenging task [1], but the generation process can get a great impact to understanding the description of images. A good description of an image is often said for 'Visualizing a picture in the mind'. The proper sentence description of an image can play a significant role in artificial intelligence and image processing field.

Human can easily understand the description of image and can also easily describe with natural language description. However, teaching that to the machine is still a difficult task. In [2], machines can recognize the human activities in videos, but the automatic description for visual scenes has remained unsolved. In the community of computer vision, automatic understanding the activities in the complex and continuous activities is still challenging in the action recognition system [3]. Activity recognition is representing with the verb phrases as a linguistic perspective by extracting the semantic similarity from the human actions [4].

We have studied different existing natural language description generation model for an image and how it works to generate the new language description for unknown image. Based on the existing model, we have implemented the deep learning model for generating the natural language description of an image. In deep learning

model, Convolutional Neural Network (CNN) is used to extract the features of images and Recurrent Neural Network (RNN) is used to generate the natural language description from the image features. We have implemented InceptionResNetV2 pre-trained model for CNN and Long-Short Term Model (LSTM) for RNN. We have also described the implementation results of this model along with comparisons.

Literature surveys related with natural language description generation of an image are described in Section 2 of this paper. Section 3 presents the natural language description model. The implementation details contained about dataset and evaluation metrics is showed in Section 4. At the end of this paper, we conclude about natural language description generation system using deep learning model with Section 5.

2. LITERATURE SURVEYS

Vinayals et al. [5] proposed the end-to-end framework that generated the image description. This framework is created by replacing the RNN encoder in the place of CNN encoder that produced a better textural description from the image representation. The proposed model is named as Neural Image Caption model. The input of the RNN decoder is entered from the last hidden layer of the CNN to generate the textual description for the image.

You et al. [6] introduced the new approach by adding the semantic attention model to the combination of top-down and bottom-up approaches. A convolutional neural network is used to extract the visual features of the image and to detect the visual concepts of the image based on the top-down approach. The semantic attention model is combined both the attributes of the image and the visual concepts of the image to generate the sentence description of the image by using RNN. The iteration processes of RNN can change the attention weights for several candidate concepts by using the bottom-up approach.

Gany et al. [7] developed a semantic compositional network that applied the semantic concepts for the textual description generation from the query image. Likelihoods of all tags are used to compose the semantic concept vector to process the LSTM weight matrices in the ensemble. The advantage of this network when the description of the image generates is that it can learn the collaborative semantic concept dependent weight metrics.

Pan et al. [8] proposed the communication framework between CNN and RNN to extract the semantic features

from the video named as LSTM unit with transferred semantic attributes (LSTM-TSA) framework. The sequence that generates the textual description uses the semantic features and these semantic features are interpreted the objects and scenes of the image, but failed to reflect the temporal structure of video. Adding together the image source and video source has improved the system that generated natural language description from the video.

Xu et al. [9] developed a novel Sequential VLAD layer, named as SeqVLAD which generates the better representation of video by combining the VLAD and the RCN framework. This model exploring the fine motion details present in the video by learning the spatial and temporal structure. An improved version of Gated Recurrent Unit of Recurrent Convolutional Network (RCN) named as Shared GRU-RCN (SGRU-RCN) was proposed to learn the spatial and temporal assignment. Overfitting problem is resolved in this model because the SGRU-RCN contains only less parameters and this achieve better results.

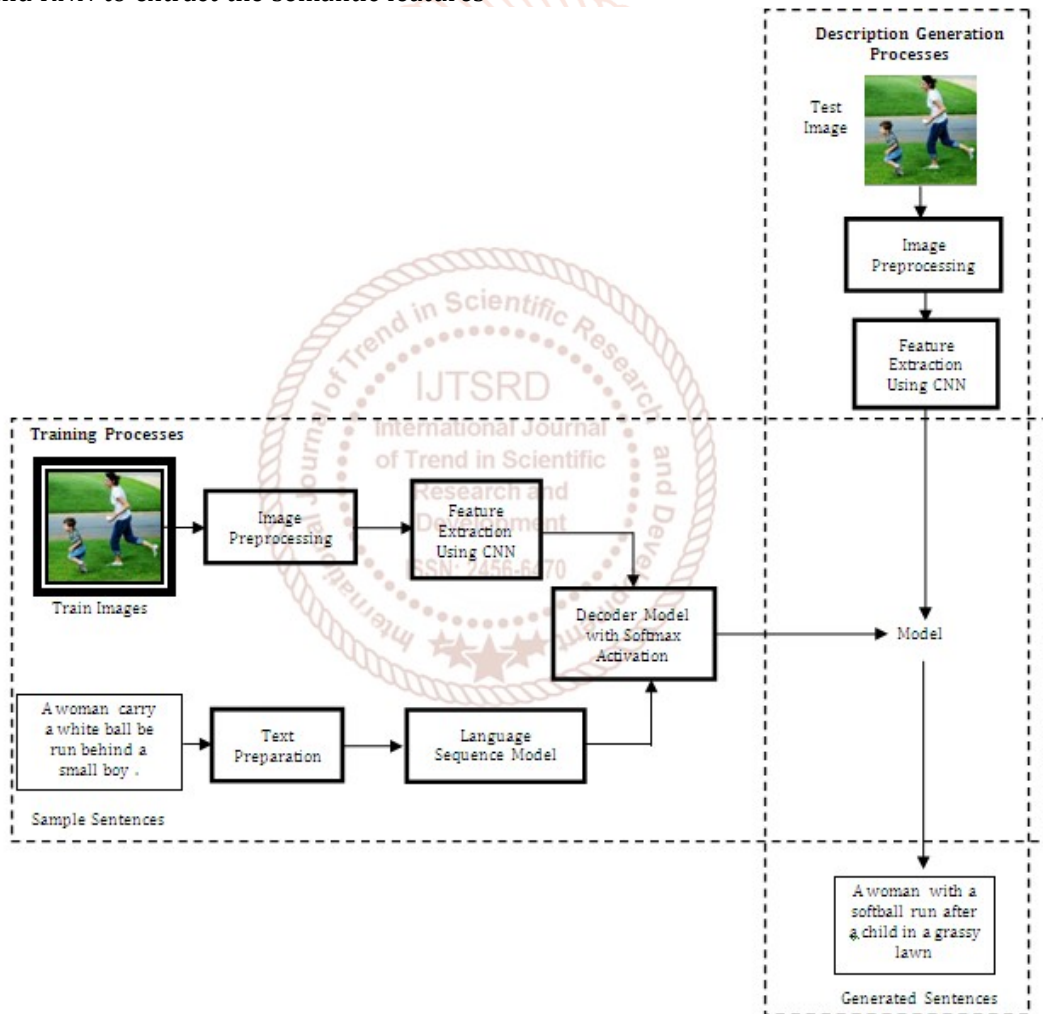


Fig.1 Natural Language Description Generation Framework

3. NATURAL LANGUAGE DESCRIPTION NETWORK

This paper proposed the natural language description framework. At that framework, we have applied the pre-trained CNN for the feature extraction of image and LSTM for the sentence feature extraction. To train the model, we have used decoder network with softmax activation

function. Fig.1 shows the natural language description framework using deep learning.

In the natural language description generation framework, there are two parts: training process and the natural language description generation process. For the training process, we are firstly pre-processed for both

images and the sentence descriptions. And then, pre-processed images are extracted features by applying pre-trained CNN model, namely InceptionResNetV2, and the pre-processed sentences are entered into the language sequence model, that combined with word embedding and the LSTM model. After that, the image features and language features are combined as a single feature vector and a feature vector enters into the decoder model to train the model with softmax activation function. Finally, the training process is extracted the natural language description generation model.

In the natural language description generation process, the image is the input of the process and the sentence is the output. The input image is pre-processed to extract the features and features are extracted by using pre-trained CNN. The output sentence is generated by passing the extracted image features to the natural language description generation model.

A. InceptionResNetV2

InceptionResNetV2 is a convolutional neural network that trained on more than a million of images from the Image Net dataset [10]. The deep of the network is 164 layers and it can classify one thousand categories of objects from images, such as mouse, keyboard, animals, and pencil. The network has learned the from the feature representations for a wide range of images.

InceptionResNetV2 is a costlier hybrid Inception version with significantly improved recognition performance [11]. The default input size for this model is 299x299. This model and can be built both with 'channels_first' data format (channels, height, width) or 'channels_last' data format (height, width, channels). The compressed view of the InceptionResNetV2 is shown in Fig. 2.

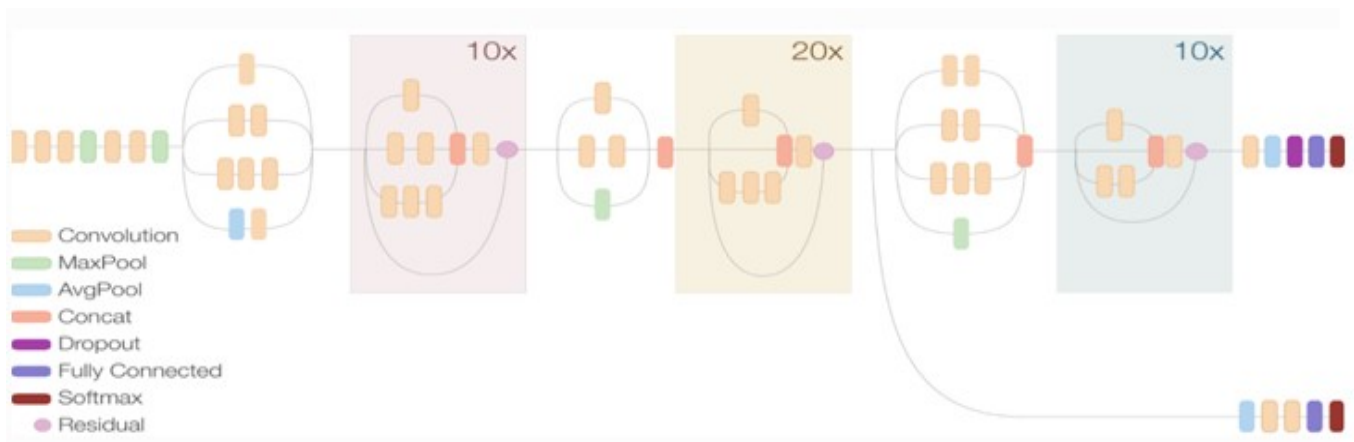


Fig.2 Compress View of InceptionResNetV2

B. Long-Short Term Memory (LSTM)

Recurrent Neural Network (RNN) is used to model the transitory dynamics in a set of things [12]. The ordinary RNN is very difficult to procure the long-term dependencies [13]. To address long-term dependencies problems, Long-Short Term Memory (LSTM) is

implemented. The LSTM cell is illustrated in Fig. 3. The main block of LSTM is the memory cell. The memory cell can store the values for a long period of time. Gates are used to control the updated states of LSTM cell. The number of connections between the memory cell and the gates represent the variants.

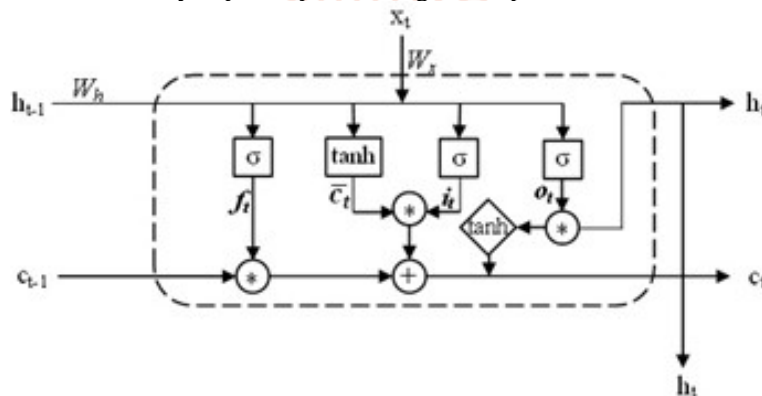


Fig.3 LSTM Cell Structure

The memory cell and gates of LSTM are implementing with the following equations:

$$\begin{aligned}
 f_t &= \sigma(x_t * W_{xf} + h_{t-1} * W_{hf}) & (1) \\
 \bar{c}_t &= \tanh(x_t * W_{xc} + h_{t-1} * W_{hc}) & (2) \\
 i_t &= \sigma(x_t * W_{xi} + h_{t-1} * W_{hi}) & (3) \\
 o_t &= \sigma(x_t * W_{xo} + h_{t-1} * W_{ho}) & (4) \\
 c_t &= f_t * c_{t-1} + i_t * \bar{c}_t & (5) \\
 h_t &= o_t * \tanh(c_t) & (6)
 \end{aligned}$$

4. IMPLEMENTATION

A. Dataset

To implement the natural language description generation of image, we have used Flickr 8k dataset. That dataset is published on 2013 by M. Hodosh et al., with "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics" paper, in Journal of Artificial Intelligence Research [14]. The Flickr 8k dataset contains 8,091 images with five sentence descriptions and includes 8,765 vocabularies. Among all images, the dataset separates 6,000 images for training, 1,000 images for testing, and 1,000 images for validation. The Fig. 4 illustrates the dataset structure that contains five natural language captions of an image.

B. Evaluation Metrics

Several evaluation metrics have been proposed to evaluate the results of image captioning and video captioning [15]. The accuracy of image captioning is calculated by comparing the generated sentence with the

ground truth sentence using the n-gram. The mostly used evaluation metrics for image captioning are BiLingual Evaluation Understudy (BLEU), Recall-Oriented Understudy of Gisting Evaluation (ROUGE), Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Consensus-based Image Description Evaluation (CIDEr).

BiLingual Evaluation Understudy (BLEU) [16] is mostly simple and popular to measure the accuracy of image description generation. It calculates the numerical translation closeness between the generated sentence and the ground truth sentence. BLEU scores can measure the fraction of n-gram (n=1,2,3,4) in common between the references and the generated sentences and it is focus on the precision. This evaluation metric is not considered the small changes or the grammatical errors in the order of words. It is more suitable for the shorter sentence descriptions.




 <p>Caption1: A black and white dog catch.</p> <p>Caption2: A black and white dog be play with a Frisbee outside.</p> <p>Caption3: A black and white dog jump to catch a green Frisbee.</p> <p>Caption4: A black and white dog leap to catch a Frisbee in a field.</p> <p>Caption5: A small black and white dog jump on the grass to catch a Frisbee</p>	 <p>Caption1: A woman carry a white ball be run behind a small boy.</p> <p>Caption2: A woman hold a ball chase a small boy run in the grass.</p> <p>Caption3: a woman hold a small ball chase after a small boy.</p> <p>Caption4: A woman be run after a boy on the grass.</p> <p>Caption5: A woman with a softball run after a child in a grassy lawn.</p>	 <p>Caption1: A man on a motorcycle go around a corner.</p> <p>Caption2: a man ride a green motorcycle around a corner</p> <p>Caption3: A man with a blue helmet lean into a sharp turn on his motorcycle.</p> <p>Caption4: A motorcyclist on a number 52 bike lean in for a sharp turn.</p> <p>Caption5: The number 52 motorcyclist in a blue and black helmet be go around a corner.</p>
---	---	---

Fig4: Sample Dataset Structure

Recall-Oriented Understudy of Gisting Evaluation (ROUGE) [17] was intended for automatically summarization of the documents. ROUGE is similar with BLEU evaluation metrics. The difference is that ROUGE measures with n-gram in the sum of number of human annotated sentences while the BLEU is considered the occurrences of the total summation of generated sentences. There have been separated into four different types, namely ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S(U). Among them, ROUGE-N and ROUGE-L are popular to evaluate the image and video captioning.

Metric for Evaluation of Translation with Explicit Ordering (METEOR) [18] is calculated mean value of precision and recall scores based on the unigram. The main difference of BLEU and METEOR is that it combines both precision and recall metric. METEOR can solve the limitation of strict matching by utilizing the word and synonyms based on unigram while BLEU and ROUGE have the difficulties to solve that limitation.

Consensus-based Image Description Evaluation (CIDER) [19] is used as the evaluation metric to measure the natural language description generation of image. This metric is calculated by measuring the consensus between generated sentence from image and ground-truth sentence. The two sentences, generated sentence and

ground-truth sentence, are compared by using the cosine similarity and the metric works as the extension of TF-IDF method. This evaluation metric is not significant and effective in the evaluation for the accuracy of natural language description generation of image.

Table1: Performance of Natural Language Description Generation Network

Epoch	BLUE-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	ROUGE-L
1 epoch	0.57786	0.310881	0.170631	0.089829	0.20545	0.19774	0.45214
5 epochs	0.553699	0.302392	0.171061	0.09074	0.21512	0.25641	0.45233
10 epochs	0.50827	0.272958	0.155084	0.084381	0.21444	0.24516	0.43794



Fig.5 Generated Descriptions without Error



Fig.6 Generated Descriptions with Error

C. Results

To implement the natural language description generation framework for an image, we have used the machine with Intel Core I7 processor with 8 cores and 8GB RAM running on Window 10 OS. Keras library based on tensorflow is used for creating and training deep neural networks. Tensorflow is a deep learning library developed by Google [20]. Tensorflow uses the graph definition to implement the deep learning network. It can be executed on any supported devices by defining one graph at once.

For the natural language description generation framework, we are using by combining CNN and RNN. Pre-trained CNN is used for the image features extraction task and it acts as an image encoder. For the sentence features extraction, LSTM is used. After the feature extraction, the image features and sentence features are combined for the input of decoder to train the model. After training, the decoder of the model can generate the sentence description of an image. For the training, we have used three types of epochs: 1 epoch, 5 epochs, and 10 epochs. The implementation results are shown in Table 1.

The sentence is generated based on the common descriptions that exist in the dataset. The natural language description generation framework, implemented the InceptionResNetV2 and LSTM, can only generate the simple sentence description. Some generated descriptions without error are shown in Fig. 5. However, the predicted sentences can wrong sometimes than the original sentences of image and the sentence can weakly related to the input image. The generated descriptions with error are described in Fig. 6. The generated results at the first row of Fig. 6 are the generated descriptions with minor error such as places or color and the results at last row are presented the generated descriptions unrelated with image.

5. CONCLUSION

This paper presents the natural language description generation framework used deep learning network. The framework is trained to produce the sentence description from the given image. This framework has been implemented on the Flickr 8k dataset and the performance of this framework has been measured with the standard evaluation metrics. The sentence descriptions obtained from the framework are categorized into the generated description without errors, the generated description with minor error, and the generated description unrelated with image. The performance of the framework is not extremely good. This framework should run on other large datasets and it can add the attention mechanism as the future extension.

Acknowledgment

I would like to deeply thank Prof. G.R.Sinha for guiding me. We would like to give our thanks and gratitude to people who have contributed towards the development of this manuscript from the planning stage to the finish stage.

References

- [1] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Computer Vision and Pattern Recognition (CVPR)*, 2015 IEEE Conference on. IEEE, 2015.
- [2] A. Torralba, K. Murphy, W. Freeman, and M. Rubin, "Context-based vision system for place and object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2003, pp. 273–280.
- [3] C. Wallraven, M. Schultze, B. Mohler, A. Vatakis, and K. Pastra, "The Poeticon enacted scenario corpus—A tool for human and computational experiments on action understanding," in *Proc. 9th IEEE Conf. Autom. Face Gesture Recognit.*, 2011, pp. 484–491.
- [4] N. Mavridis, "Grounded situation models for situated conversational assistants," Ph.D. dissertation, Dept. Archit., Massachusetts Inst. Technol., Cambridge, MA, USA, 2007.
- [5] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: A Neural Image Caption Generator", *IEEE CVPR*, pp. 3156-3164, 2015.
- [6] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo, "Image Captioning with Semantic Attention", *IEEE CVPR*, 2016.
- [7] Z. Gany, C. Gan, X. Hez, Y. Puy, K. Tranz, J. Gaoz, L. Cariny, L. Dengz, "Semantic Compositional Networks for Visual Captioning", arXiv:1611.08002v2, 28 Mar 2017.
- [8] Y. Pan, T. Yao, H. Li and T. Mei, "Video Captioning with Transferred Semantic Attributes", *IEEE CVPR*, pp. 984-992, 2017.
- [9] Y. Xu, Y. Han, R. Hong, Q. Tian, "Sequential Video VLAD: Training the Aggregation Locally and Temporally" in *IEEE Transactions on Image Processing*, Vol. 27, No. 10, October 2018.
- [10] ImageNet. <http://www.image-net.org>
- [11] Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." In *AAAI*, vol. 4, p. 12. 2017.
- [12] Lu, Jiasen, et al., "Knowing when to look: Adaptive attention via a visual sentinel for im-age captioning", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 6, 2017.
- [13] 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)*, Vol. 6, 2017
- [14] Hodosh, M., Young, P. and Hockenmaier, J., 2013, "Framing image description as a ranking task: Data, models and evaluation metrics", *Journal of Artificial Intelligence Research*, 47, pp.853-899.
- [15] J. Park, C. Song, J.-H. Han, "A Study of Evaluation Metrics and Datasets for Video Captioning", *ICIIBMS* 2017.
- [16] 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 (ACL'02), Association for Computational Linguistics, Stroudsburg, PA, USA, 311-318, 2002.

- [17] Lin CY, "ROUGE: a package for automatic evaluation of summaries", in Proceedings of the workshop on text summarization branches out, Barcelona, Spain, (WAS2004) 2004.
- [18] D. Elliott and F. Keller, "Image description using visual dependency representations," in Proc. Empirical Methods Natural Lang. Process, 2013, vol.

13, pp. 1292-1302.

- [19] R. Vedantam, C. L. Zitnick and D. Parikh, "CIDER: Consensus-based image description evaluation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4566-4575.
- [20] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

