Comparison of Different CNN Model used as Encoders for Image Captioning

1st MD Sahrial Alam
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
sahrialalam@gmail.com

2nd MD Sayedur Rahman
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
srsayed2677@gmail.com

3rd MD Ikbal Hosen
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
mdiqbalhossain6463@gmail.com

4th Khairul Anam Mubin
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
anam.mubin1999@gmail.com

5th Sharif Hossen
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
sharifhossen317@gmail.com

6th M. F. Mridha
Department of Computer Science and Engineering
Bangladesh University Of Business And Technology - BUBT
Dhaka-1216, Bangladesh
firoz@bubt.edu.bd

***Abstract*—Image captioning is the process of automatically describing the content of an image which connects computer vision and natural language processing. In this paper, we compare five popular convolutional neural networks architecture. They are Vgg16, InceptionV3, Resnet50, Densenet201 and Xception Model. By using these preprocessing model for same image captioning model. Encoder-Decoder model is a part of recurrent neural networks for sequence-to-sequence prediction problems. Encoders are usually used pre-trained convolutional neural networks for large datasets. There are many different types of Encoder-Decoder architecture used for generating caption. But it is very complicated to evaluate the performance of the architecture. In this paper, we used categorical-crossentropy for loss function, RMSprop for optimizer in Vgg16, Resnet50, InceptionV3, Densenet201 and Xception model.**

***Index Terms*—encoder-decoder framework, image captioning, CNN models.**

I. Introduction

Image caption generation is one of the most important research fields in computer vision and natural language processing [18]. The target of image captioning is identifying the relationship between what objects are presented. Generate sentences is must be human-like base on the information of the given image. Image captioning systems can be used for practical tasks, such as help to visually-impaired people, human-computer interaction and image search [1], [3].

Nowadays most of the case, encoder-decoder paradigm used for architecture [13], [19], [20]. To representing image features, an encoder take an image as input and transform it to a vector. By using that image feature, the decoder generates a sequence of words. Resnet50 [9], Vgg16 [17], InceptionV3 [14], Densenet201 [22] and Xception [2] are usually used as encoders to pre-trained a large image dataset.

LSTM [10] save the previous generated sequence so far and use it for next word generation. LSTM is used as a decoder to generate sequence wise caption.

The main problem of image captioning is different paper used different encoders. Some difficulties increase here to identify the better encoders which should be taken for image captioning.

In this paper, compares five image captioning model to see the effect of performance by changing encoders.

II. Literature Review

Image Caption Generation analysis has inducted significant in the area of deep learning researchers attention during the past few years. In computer vision, generating descriptions in natural language is one of the problem from visual data has long been studied [7], [21], [5]. To encode the image, Vinyals et al. [19] proposed a CNN pre-trained which is based on ImageNet [4], then LSTM [8] generates the sequence of words and make caption for the image. Xu et al. [20] introduced an attention mechanism base image caption generation which generate caption during the generation of each word. It works with the previous generated word and their model. To understand the weight of encoded feature map, the model generates a matrix and then observe the previous generated word and the weighted feature map to generate the next word of the sentence. To generate Bengali caption Al Momin Faruk et al. [6] proposed an architecture model which is based on CNN and bidirectional GRU. To generate captions they use encoder-decoder approach. They used InceptonV3 as a pre-trained model where classification and BGRU layer is used to generate captions. Matiur Rahman et al. [16] developed a model which name is 'Chittron' and it is an automatic image captioning system in Bangla. They work with 16,000 Bangladeshi contextual images. VGG16 image embedding model with stacked LSTM layers used the dataset to train the model. This model is trained to predict the caption of an input image.

III. Existing Models

A. Vgg16

VGG16 is a convolution neural network architecture. In 2014, it was used to win ILSVR (Imagenet) competition. Till now it is one of the best model architecture. It has classify the new images. It has 2 fully connected layers. VGG16 achives 92.7 percent accuracy where usrd 14 million images with 1000 classes. VGG16 refers because it has 16 layers. And Vgg16 has about 138 million (approximate) parameters. Vgg16 is in vector form which is caused by obtaining high-quality representation of the image.

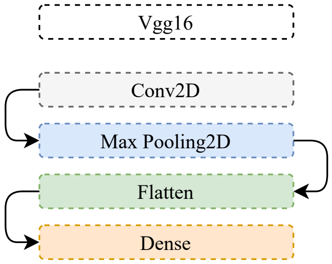

Fig. 1. Process of Vgg16 Encoder

B. Resnet50

ResNet50 is updated version of ResNet model. The ResNet50 architecture is designed to solve "vanishing gradients". ResNet50 is based on shortcut connection idea. In neural network, ResNet50 increase the number of layers. ResNet50 has great advantages because it increases accuracy by adding more layers and is very easy to optimize.

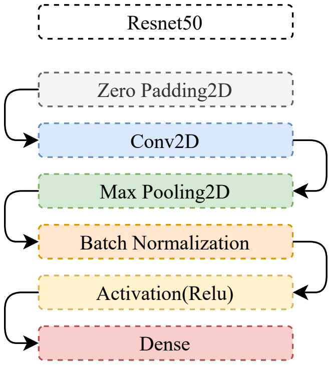

Fig. 2. Process of Resnet50 Encoder

C. InceptionV3

Inception-v3 is a convolutional neural network architecture. It is factorized by 7 x 7 convolution, and it makes the improvements by using Label Smoothing and To propagate label information lower down the network, it uses auxiliary classifier with the batch normalization. Inception v3 is a widely-used image recognition model.

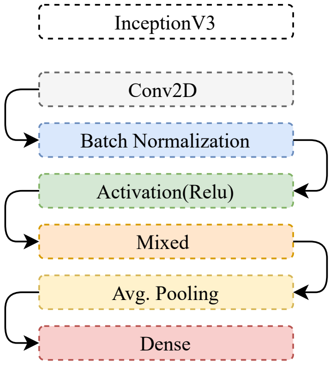

Fig. 3. Process of InceptionV3 Encoder

D. Densenet201

Densenet201 is one of the new discoveries architecture in neural networks.Densenet201 is quite similar to ResNet except some minor difference. To merges the previous layer with the future layer, ResNet uses additive method (+) whereas, Densenet201 concatenates (.) the output of the previous layer with the future layer. Densenets201 have some advantages which are smoothly vanishing-gradient problem, encourage feature reuse, reduce the number of parameters and strengthen feature propagation. [12].

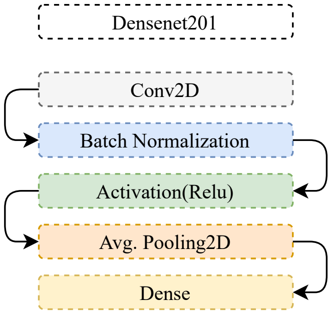

Fig. 4. Process of Densnet201 Encoder

E. Xception

Xception stands for Extreme version of Inception. It's a deep convolutional neural network architecture that involves depth wise separable convolution and it has 37 convolutional layers. It has two main working principle that are Depth Wise Separable Convolution and Shortcuts Between convolution blocks as in Resnet.

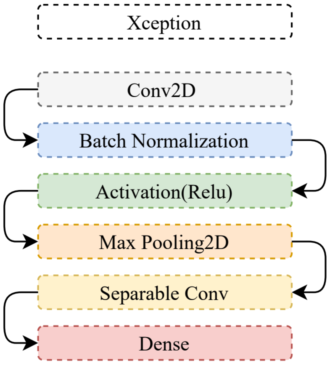

Fig. 5. Process of Xception Encoder

IV. Methodology

We evaluated our comparison on a Flicker dataset [11]. It contains images with 5 describing it caption for each image. In this paper we discuss five different pre-processing encoders architecture are VGG16, Resnet50, InceptionV3 Densenet201 and Xception. We used 1500 images for training and selected randomly image from 8121 images which is not include in training images. We convert all caption to lower-case and removing Punctuation, find unique word from text and we find 3973 unique word. Then we used LSTM model for generating caption. We used categorical cross-entropy for measure loss function, RMSProp for optimize for model. The Gradients are accumulated into an exponentially weighted average [23]. RMSProp maintains recent gradient information [15] and its variants and also RMSProp discards the history .

A. Architecture Of Image Captioning (Par-Inject)

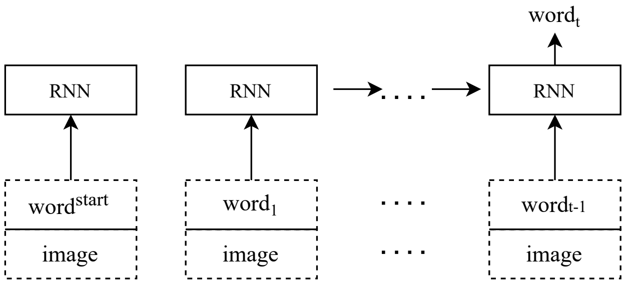

Fig. 6. Architecture Of Image Captioning (Par-Inject)

In this paper, Par-inject architecture is used for image captioning system.At first preprocessing image and text then at the same time input both word and image into RNN. Then RNN uses for making a caption the figure 2 dashed box points to an encoding subset(text or image). The solid boxes suggest a decoding subset.

V. Experiments

A. Dataset and Evaluation

We used Flickr8k dataset from Kaggle. It contains 8121 images with five descriptions for each image. We use 1500 images for training. For text pre-processing [20], we convert all the text to lowercase, removing punctuation and taking unique word from that text.

B. Implementation Details

We used Resnet50, Vgg16, InceptionV3, Densenet201 and Xception as they are most popular encoder architecture. LSTM is a popular decoder architecture. Images are pre-processed by InceptionV3 with 299x299 pixel size, Resnet50 with 224x224 pixel size, Vgg16 with 224x224 pixel size, Densenet201 with 224x224 pixel size and Xception with 299x299 pixel size. By using those encoder architecture, we extract features from images. By removing punctuation, convert all the letters to lowercase, taking unique word we pre-process our text. We used RMSprop optimizer for Resnet50, Vgg16, InceptionV3, Densenet201 and Xception. We trained our model by giving 100 epochs with 5000 batch size.

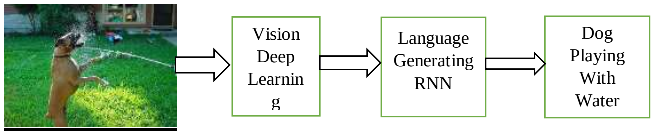

Fig. 7. Caption Generation Methodology

C. Performance Comparison and Analysis

TABLE I
The Table Shows The Accuracy and Value loss of Five Encoders

Encoders	Accuracy	Value Loss
Vgg16	0.8365	0.5666
Resnet50	**0.8768**	**0.4146**
InceptionV3	0.8012	0.7071
Densenet201	**0.8720**	**0.4550**
Xception	0.8113	0.7142

After testing and training, accuracy of Resnet50 encoder is 0.8768 and value loss is 0.4146, accuracy of Vgg16 encoder is 0.8365 and value loss is 0.5666 and accuracy of InceptionV3 is 0.8012 and value loss is 0.7071 and accuracy of Densenet201 is 0.8720 and value loss is 0.4550 and accuracy of Xception is 0.8113 and value loss is 0.7142. Here, speed of testing and training of Resnet50 as encoder is better than Densenet201 as encoder and speed of testing and training of Densenet201 as encoder is better than Vgg16 as encoder and speed of testing and training of Vgg16 as encoder is better than Xception

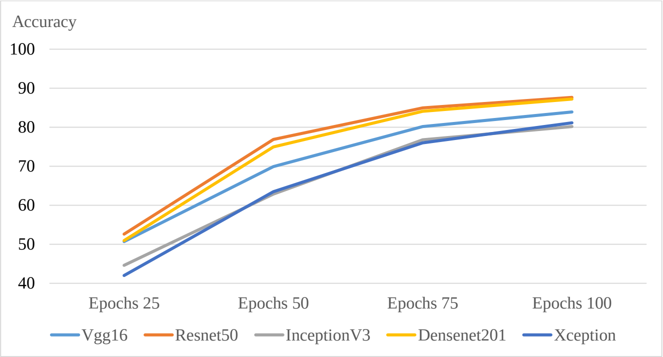

Fig. 8. Accuracy Graph For Vgg16, Resnet50, InceptionV3, Densenet201, Xception

as encoder and speed of testing and training of Xception as encoder is better than InceptionV3 as encoder.

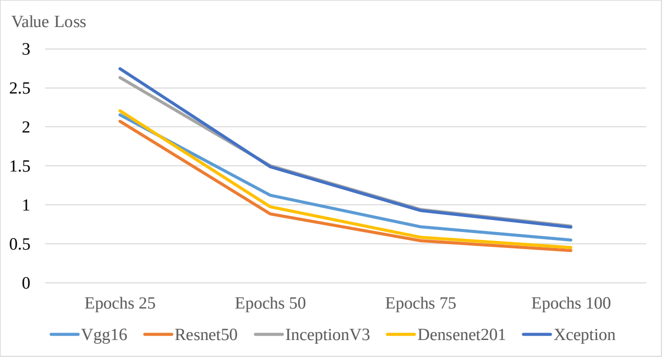

Fig. 9. Value Loss Graph For Vgg16, Resnet50, InceptionV3, Densenet201, Xception

Based on the result, Resnet50 is better than Densenet201 and Densenet201 is better than Vgg16 and Vgg16 is better than Xception and Xception is better than InceptionV3. For solving many problems associated with image processing ,ResNet is a stronger architecture . That's why accuracy of Resnet50 is better than Densenet201, Vgg16, InceptionV3 and Xception as an encoder.

VI. Conclusion

In this paper, we compared five different image encoders such as ResNet50, InceptionV3, Vgg16, Densenet201 and Xception for image captioning. We showed the accuracy and value loss for training. Par-inject architecture is used here for image captioning. LSTM decoder architecture is used to generate sentences based on image feature. After evaluated all encoders performance, Resnet50 gives better accuracy according to our research.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[3] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer, 2010.

[6] Al Momin Faruk, Hasan Al Faraby, Md Azad, Md Fedous, Md Morol, et al. Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit. *arXiv preprint arXiv:2012.12139*, 2020.

[7] Ralf Gerber and N-H Nagel. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In *Proceedings of 3rd IEEE international conference on image processing*, volume 2, pages 805–808. IEEE, 1996.

[8] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. *Neural computation*, 12(10):2451–2471, 2000.

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

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

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

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

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

[14] Varsha Kesavan, Vaidehi Muley, and Megha Kolhekar. Deep learning based automatic image caption generation. In *2019 Global Conference for Advancement in Technology (GCAT)*, pages 1–6. IEEE, 2019.

[15] Mahesh Chandra Mukkamala and Matthias Hein. Variants of rmsprop and adagrad with logarithmic regret bounds. In *International Conference on Machine Learning*, pages 2545–2553. PMLR, 2017.

[16] Matiur Rahman, Nabeel Mohammed, Nafees Mansoor, and Sifat Momen. Chittron: An automatic bangla image captioning system. *Procedia Computer Science*, 154:636–642, 2019.

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

[18] Raimonda Staniūtė and Dmitrij Šešok. A systematic literature review on image captioning. *Applied Sciences*, 9(10):2024, 2019.

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

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

[21] Benjamin Z Yao, Xiong Yang, Liang Lin, Mun Wai Lee, and Song-Chun Zhu. I2t: Image parsing to text description. *Proceedings of the IEEE*, 98(8):1485–1508, 2010.

[22] Feyza Yilmaz, Onur Kose, and Ahmet Demir. Comparison of two different deep learning architectures on breast cancer. In *2019 Medical Technologies Congress (TIPTEKNO)*, pages 1–4. IEEE, 2019.

[23] Raniah Zaheer and Humera Shaziya. A study of the optimization algorithms in deep learning. In *2019 Third International Conference on Inventive Systems and Control (ICISC)*, pages 536–539. IEEE, 2019.