IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⪿ Impact Factor 7.39 ⪿ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11810

Comparison of Different Encoder Techniques in Image Caption

Pooja Negi[1], Sanjay Buch[2]

Phd. Scholar, College of Computer Applications, Bhagwan Mahavir University, Surat, India[1]

Dean, College of Computer Applications, Bhagwan Mahavir University, Surat, India[2]

Abstract: Image captioning, has been one of the most intriguing topics in deep learning. It incorporates the knowledge of both image processing and natural language processing. Most of the current approaches integrate the concepts of neural network. Many pre-defined convolutional neural network (CNN) models are used for extracting features of an image and bi-directional or uni-directional recurrent neural network (RNN) for sentence creation as decoder. This paper discusses about the commonly used models that are used as image encoder, such as VGG16, VGG19, Inception-V3 and InceptionResNetV2 while using the uni-directional LSTMs for the sentence generation. The comparative analysis of the result has been obtained using the BLEU score on the Flickr8k dataset.

Keywords: Image Captioning, CNN, LSTM, BLEU

I.INTRODUCTION

Many different approaches have been proposed in deep learning for, extracting features of an image, and natural language processing [1][2]. Generally convolutional neural network models are used as the encoder to extract features from the images; moreover, the recurrent neural network is used as decoder, to decode the extracted features into a description [2][3].

There can be many possible captions for a single image, and different models can give different captions. This paper compares the encoding models on the basis of BLEU score on the same dataset and language model.

There are many datasets available for the English image captioning task, such as Flickr30k [4],Flickr8k and MS-COCO. This paper uses Flickr8k dataset, that has all different kinds of images.

II.RELATED WORK

This section gives the relevant background on image caption generation. There have been many approaches regarding the image captioning task, [1] proposed a method of learning the mapping between images and captions using a graphical model using features engineered by humans. Chen et al. [5] studied and update an image captioning model LRCN. They decomposed the method to CNN, RNN, and sentence generation for understand the method in deep. they replaced the elements to check the changes on the final result and used the updated techniques to evaluate the COCO caption collection. their results presents that initially the VGGNet outperforms the AlexNet and GoogLeNet in BLEU score calculation; next, the simple GRU model gets equivalent results with more difficult LSTM methods; and then, raising the beam size raise the BLEU marks in general but their model does not satisfied the full need of human based caption.

Sak et al. [6] compare and evaluate the performance of the architecture of LSTM with RNN on the set of large vocabulary speech recognition task. First time they shows LSTM RNN models can be trained by ASGD distributed training. And analyze the deep LSTM to propose deep LSTMP in which multiple LSTM layers are used for each recureent layer. Whereas [7] and [8] uses the modern convolutional neural networks for encoding and recurrent neural networks for language modelling, this is end to end training.

The alternative method, used in [9] and [10], involves initially training a convolutional and bi-directional recurrent neural network that tries to learn how to map caption fragments to the same embedding as well as image fragments. Then it uses a different recurrent neural network that learns to combine the inputs from various object fragments detected in the original image to form captions. Then, in order to create captions, it employs a separate recurrent neural network that is trained to merge the inputs from numerous object fragments found in the original image.

By including an attention mechanism, the aforementioned approaches will function even better. By paying attention, the model can ignore certain information while concentrating on others. This has been covered in [3], and [11] discusses the use of variational auto-encoders for this task.

DOI: 10.17148/IJARCCE.2022.11810

Dataset

The Department of Computer Science at the University of Illinois at Urbana-Champaign offers the Flickr8k dataset [2]. It includes 8k photos, 6k images for training, 1k images for validation, and 1k images for testing.

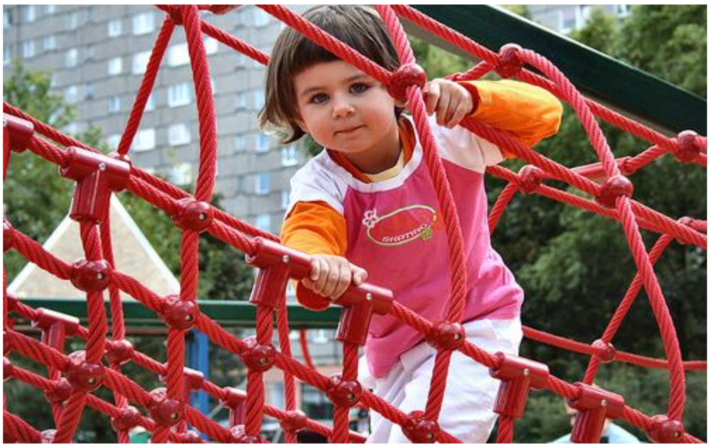

Fig. 1

The dataset's sample image is displayed in Fig. 1, and the five captions that go with it are as follows:

(i) A child playing on a rope net .
(ii) A little girl climbing on red roping .
(iii) A little girl in pink climbs a rope bridge at the park .
(iv) A small child grips onto the red ropes at the playground .
(v) The small child climbs on a red ropes on a playground

III. IMAGE CAPTIONING

This study compares the operation of various convolutional models in order to produce results using the BLEU metric. Various models are used to encode images into a vector of features. Furthermore, the common decoding model makes use of these properties.

3.1 Language Processing

A map is made in the form of a data-frame for the language processing task, with five captions—numbered 1, 2, 3, 4 and 5—for each image. Before sending the caption to the network, the frequently occurring words are removed so that the test set provides a more general description and does not over fit any term.

3.2 Encoders: Convolutional Models

Four encoders are addressed in this section. Three subsections are used to describe these encoders. Encoders VGG16 and VGG19, also known as Inception-V3 and Inception-ResNet-V2, are discussed in Sections 1, 2, and 3, respectively.

3.2.1 VGG16 and VGG19: The 224 by 224 RGB image is the fixed size input for VGG16 and VGG19 [12]. With a 3 x 3 kernel size, an image is transmitted through the various convolutional network layers. Three fully connected layers follow the convolutional layers (first two have 4096 channels and the third has 1000 channels). VGG16 and VGG19 both have 13 convolutional layers, however VGG16 has 16 convolutional layers. Figs. 2 and 3 depict the architectural layers of VGG16 and VGG19, respectively.

DOI: 10.17148/IJARCCE.2022.11810

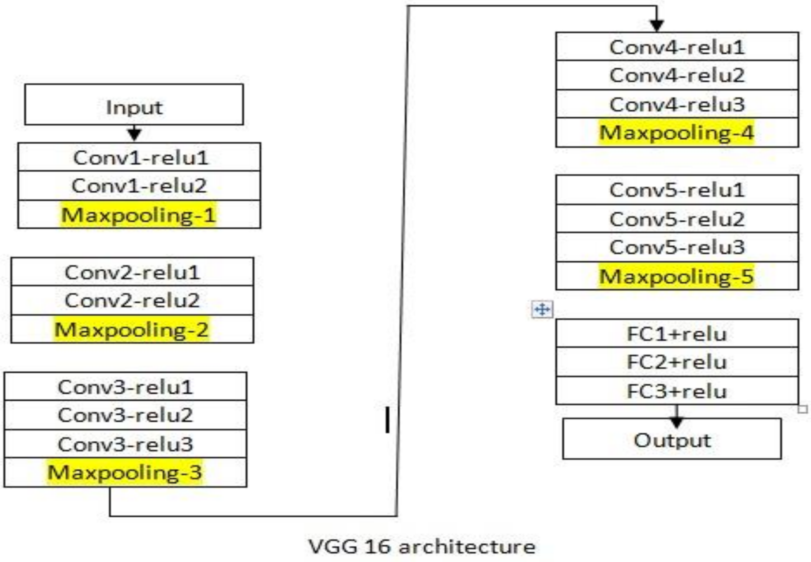

Fig. 2 VGG16 architecture layers

Inception-V3: Choosing appropriate kernels for various images can be challenging because the information that is accessible in an image can vary in size and position. The Inception-V3[13] model employs several kernels with sizes of 1x1, 3x3, and 5x5. Inception-v3 uses batch normalisation in the auxiliary classes and accepts a fixed input of a 299 × 299 RGB image. The sum of the depths of each layer of kernels represents the overall depth. The network's last layer is a linear layer with dimensions of 1x1x1000. Fig. 4 displays the Inception-architectural V3's layers.

Inception-ResNetV2: Inception-ResNetV2[14] accepts inputs that are 299 × 299 RGB pictures, which is comparable to that of the Inception-V3 model. Filter expansion layers are used to scale up the dimensionality of the filter bank before addition to match the depth of the input after each Inception block. Batch-normalization is only used by Inception-ResNetV2 on top of the conventional layers, not on top of the summation. Figure 5 displays several stages of the Inception-ResNetV2 model.

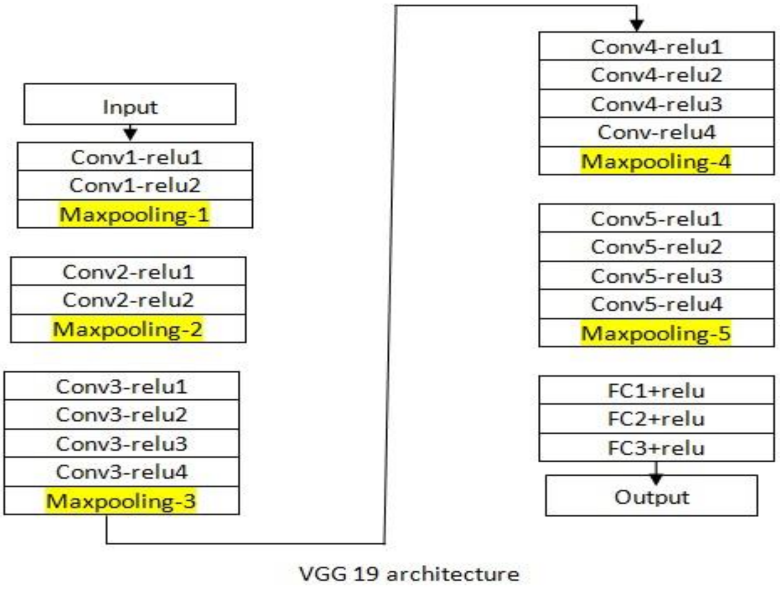

Fig. 3 VGG19 architecture

DOI: 10.17148/IJARCCE.2022.11810

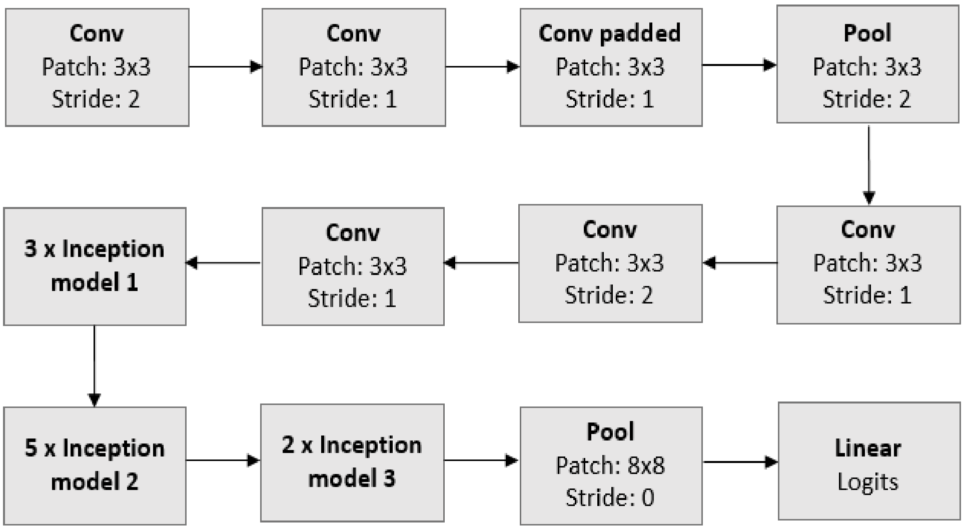

Fig. 4 Inception-V3 architecture

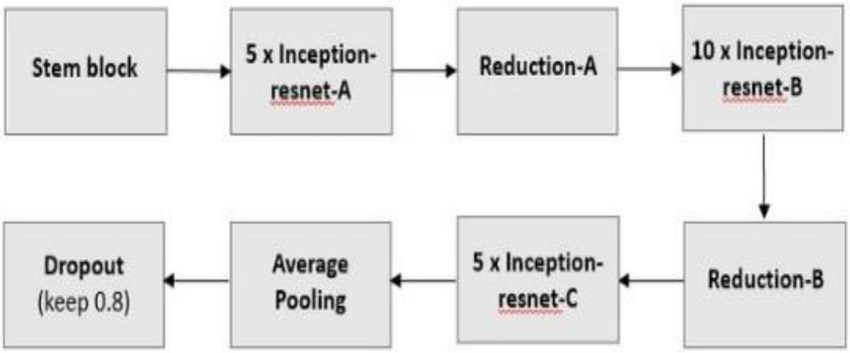

Figure 5. Inception- ResNetV2 model

Decoder: LSTM

A single layer LSTM followed by hidden neural layers is employed as the common decoder for all the aforementioned encoders in order to obtain an accurate comparison between the feature extraction models. LSTMs are typically employed to retain data from prior output. The main concept behind LSTM usage is how it manages the information. In LSTM, the cell serves as the transport highway, allowing data from previous time steps to go to subsequent time steps. The condition of the information, including whether it is added or withdrawn, is determined by the cell's gates. The cells can choose how much information to transfer to the next state of the layer by setting the gates' sigmoid functions (which range from 0 and 1) to 0 or 1. For a value of 1, the entire amount of information is passed to the following cell.

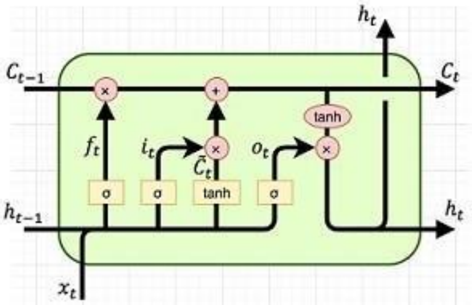

Fig. 6 Information Flow in LSTM Cell.

DOI: 10.17148/IJARCCE.2022.11810

The three types of gates are forget gate, input gate, and output gate. The information's status, including whether it should be destroyed or maintained, is determined by the forget gate. The input gate receives the previous concealed state as well as the current state to update the cell's state. The output gate determines the state of the following cell.

Below are the LSTM equations:

$i = \sigma(x_tU^i + s_t{-}1W^i)$	(1)
$f = \sigma(x_tU^f + s_t{-}1W^f)$	(2)
$o = \sigma(x_tU^o + s_t{-}1W^o)$	(3)
$s_t = \tanh(c_t)o$	(4)

Parameters of equations (1) to (5) are as defined:
1- i is the input gate.
2-f is the forget gate.
3-o is the output layer.
4-W is the connection of the previous.
5-U connects the input to current hidden layer.

IV.TRAINING

Experiment

Assessment Metric The algorithm's generated caption quality is assessed using the BLEU score, which is based on the actual captions. Scores are determined for each segment that has been translated by contrasting it with a collection of accurate reference translations.

Method

Out of the five captions, only one is used because of equipment restrictions. The dataset's breakdown remains unchanged at 6,000 for training, 1,000 for validation, and 6,000 for testing.

The output of the encoding models, which in this research are convolutional models, is limited to the second-to-last layer of the model. Table 1 contains a list of training parameters.

Table 1 Training Arguments

Adaptive Learning Rate	0.07
Optimizer	Adagrad
Sample Size	512

V.RESULT

In part III, many encoding models' structures are covered. These models are used to extract features from photos as vectors. The LSTMs' decoding layer, which produced the captions for the images, then made use of the retrieved features.

On the same photos, different encoding models perform differently. In order to evaluate the outcomes, we used the models' best and worst BLEU scores, which may have been generated on several images.

VGG16

Since VGG16 lacks the depth of VGG19 and the variety of filters found in Inception-V3 and Inception-ResNetV2, it receives the lowest best BLUE score. In Figs. 7 and 8, respectively, the best and worst BLEU scores for the VGG16 layer are displayed.

DOI: 10.17148/IJARCCE.2022.11810

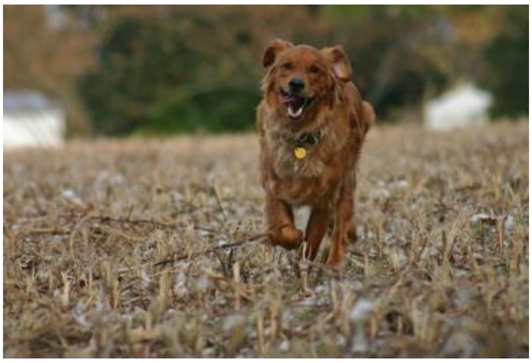

Fig. 7 VGG16 best BLEU score:0.84053816

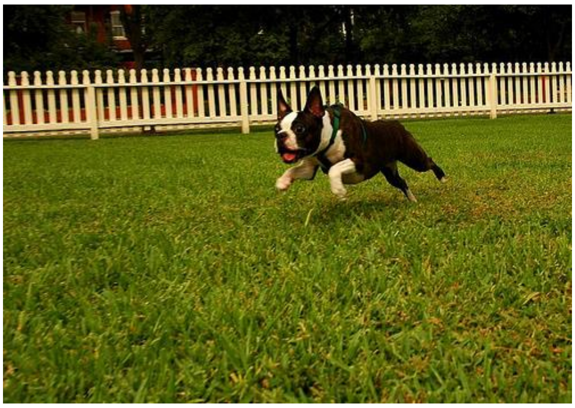

Fig. 8 VGG16 worst BLEU score:5.11287490e-232

VGG19

Because VGG19 has deeper convolutional layers than VGG16 (which led to significantly better feature extraction from the image), the VGG19 model had a higher BLEU score. In Figs. 9 and 10, respectively, the best and worst BLEU scores for the VGG19 layer are displayed.

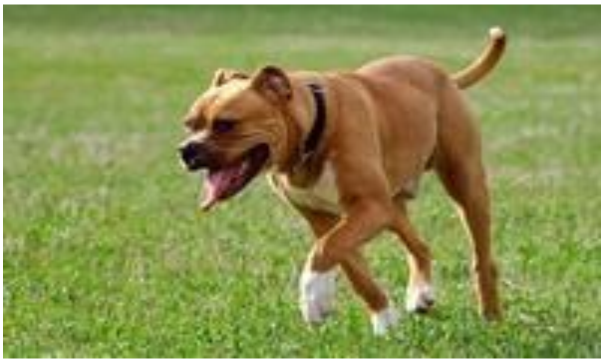

Fig. 9 VGG19 best BLEU score:0.88241298834

Inception-V3 and Inception-ResNetV2

Due to the various filters available in the model, which may extract more features than the VGG16 and VGG19, the range of BLEU score is lowest for Inception-V3. Figs. 11 and 12 display the best BLEU scores for Inception-V3 and Inception-ResNetV2, respectively. Fig. 13 displays the image with the worst BLEU score for both Inception-V3 and Inception-ResNetV2.

It was noted in the results that the models didn't do any better for some challenging photographs. It was challenging for encoding models to extract features from the photos because there were either too many different information types (pixel values) or too many identical information types.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 8, August 2022

DOI: 10.17148/IJARCCE.2022.11810

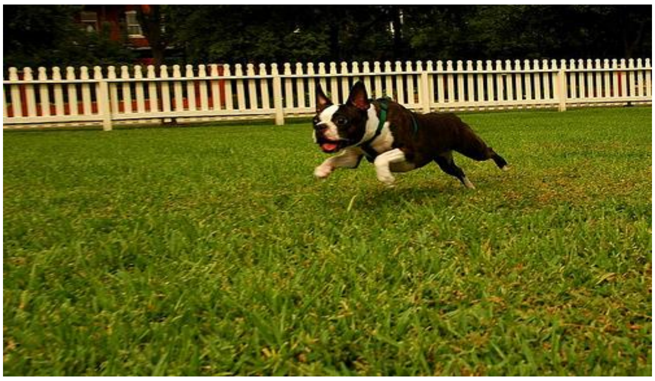

Fig. 10 VGG19 worst BLEU score:2.77554679e-78

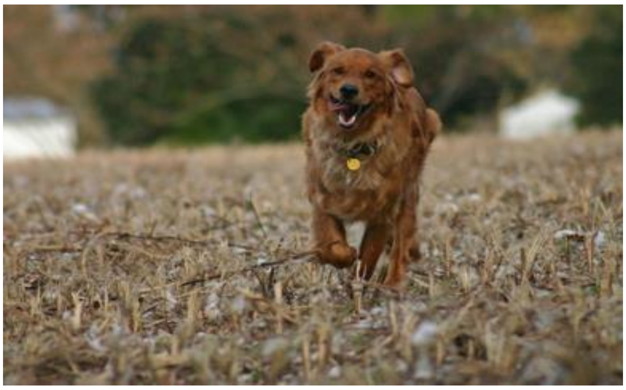

Fig. 11 Inception-V3 best BLEU score:0.895487043

Fig. 12 Inception-ResNetV2 best BLEU score:0.882496

Table 2 Evaluation of Techniques

Techniques	BLEU score
VGG16	0.84053816
VGG19	0.88241298834
Inception-V3	0.895487043
Inception-ResNetV2	0.882496

DOI: 10.17148/IJARCCE.2022.11810

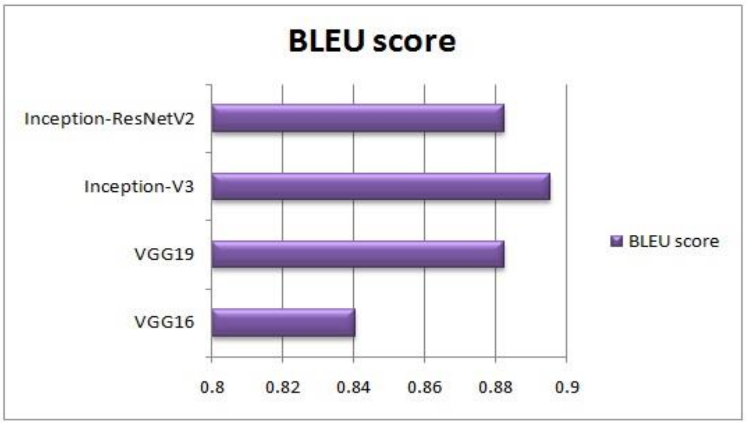

Fig. 13 Graphical Evaluation of Techniques

VI. CONCLUSION

The Flickr8k dataset is trained in this study utilizing several encoding models with the aid of Keras and Tensor flow to produce the desired outcome. The range of the BLEU score for Inception-V3 is the smallest of any of the other models covered in Section III, making it the superior model. These are all fundamental models that can be applied to the majority of image captioning tasks since these tasks entail extracting characteristics from images and applying those features to generate text.

REFERENCES

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Cyrus Rashtctian, Julia Hockenmaier and David Forsyth Every Picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15-29, 2010.

[2] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan 2015 Show and Tell: A Neural Image Caption Generator arXiv:1411.4555

[3] Fang Fang, Hanli Wang, Pengjie Tang 2018 Image Captioning with Word Level Attention 25th IEEE International Conference on Image Processing (ICIP), pages 1278-1282

[4] Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, Svetlana Lazebnik ICCV, 2015 Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Imageto-Sentence Models.

[5] J. Chen, W. Dong, M. Li. "Image Caption Generator Based On Deep Neural Networks", (2017) [Online] Available:https://www.seas.upenn.edu/~minchenl/doc/ImgCapGen.pdf

[6] H. Sak, A.W Senior, & F. Beaufays "Long short-term memory recurrent neural network architectures for large scale acoustic modeling". (2014)

[7] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell 2014 Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389.

[8] Andrej Karpathy and Fei-Fei Li 2014 Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306.

[9] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dolla´r, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Ge- offrey Zweig 2014 From captions to visual concepts and back. CoRR, abs/1411.4952.

[10] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra 2015 DRAW: A recurrent neural network for image generation. CoRR, abs/1502.04623.

[11] Parth Shah, Vishvajit Bakrola, Supriya Pati 2017 Image captioning using deep neural architectures International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pages 1-4

[12] Karen Simonyan and Andrew Zisserman 2014 Very deep convolutional Networks for large-scale image recognition arXiv:1409.1556.

[13] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens and Zbigniew Wojna 2016 Rethinking the Architecture for Computer Vision IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2818-2826

[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi 2016 Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning arXiv:1602.07261.