Image Caption Generation and Comprehensive Comparison of Image Encoders

Shitiz Gupta[1], Shubham Agnihotri[2], Deepasha Birla[3], Puneet Singh Lamb[4], Achin Jain[5,*], Thavavel Vaiyapuri[6]

[1]Bharati Vidyapeeth's College of Engineering, New Delhi, India;guptashitiz17@gmail.com

[2]Bharati Vidyapeeth's College of Engineering, New Delhi, India;skagnihotri1@gmail.com

[3]Bharati Vidyapeeth's College of Engineering, New Delhi, India;birladeepasha99@gmail.com

[4]Bharati Vidyapeeth's College of Engineering, New Delhi¡ India;singhs.puneet@gmail.com

[5]Bharati Vidyapeeth's College of Engineering, New Delhi, India;achin.mails@gmail.com

[6]College of computer engineering and sciences, Prince Sattam bin abdulaziz University, Saudi Arabia; t.thangam@psau.edu.sa

* Corresponding:achin.mails@gmail.com

Abstract

Image caption generation is a stimulating multimodal task. Substantial advancements have been made in the field of deep learning notably in computer vision and natural language processing. Yet, human-generated captions are still considered better, which makes it a challenging application for interactive machine learning. In this paper, we aim to compare different transfer learning techniques and develop a novel architecture to improve image captioning accuracy. We compute image feature vectors using different state-of-the-art transfer learning models which are fed into an Encoder-Decoder network based on Stacked LSTMs with soft attention, along with embedded text to generate high accuracy captions. We have compared these models on several benchmark datasets based on different evaluation metrics like BLEU and METEOR.

Keywords: Image Captioning, Transfer Learning, CNN (Convolutional Neural Network), RNN (Recurrent neural network)and LSTM (Long Short Term Memory).

1 Introduction

One of the human's abilities is to express the conditions they are present in. Whenever an image is given, it is easy for humans to tell all the things about the image with just a glance. [1]. Developing machines with the ability to understand and interpret the real world are one of the driving forces for researchers in the domain of artificial intelligence. Two main methods used in the previous literature in modelling the pronunciation -variation [7][8] Knowledge-based approach, that uses phonetic and linguistic knowledge[46] to write phonological rules that handle variants in pronunciation. Data-driven approach uses a corpus from real speech to derive the variation in speech. The chosen approach depends on the type of variation you need to handle in your work and the purpose of handling these variations [6]. The pronunciation variation modelling should be considered in three levels: the pronunciation dictionary, acoustic model, and the language model [9].

Even though extensive research has been made in different computer vision problems, such as object recognition [2], [3], attribute classification [4], [5], action classification [6], [7], image classification [8], and scene recognition [9], [10], making a computer automatically describe an image with human-like sentences is comparatively a new task. Utilizing a machine to automatically create a natural language description for an image, named image captioning, is challenging. Combining both the research communities of computer vision and NLP, captioning an image not only needs a notable understanding of the visual contents of an image but also needs to turn these understandings into a human-like sentence. Determining behaviors, attributes,

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

and connections of objects in an image is not an easy task. Converting the visual understandings into human readable sentences makes this task even more difficult. Since natural languages constitute most of the human interaction, whether written or vocalized, enabling machines to describe the visual world will lead to a substantial number of feasible applications, like building natural human-robot interactions, new childhood learning, information retrieval, and visually impaired aid, and more.

Being both challenging and significant, the image captioning field is getting widespread recognition all around the globe. Image captioning has a wide range of application including self-driving cars and aid to blinds.

Figure 1: Caption: A dog swimming in the water

Provided an image, the motive of image captioning is to form a sentence that is grammatically credible and semantically valid to the content of the image as shown in Figure 1. This process involves two steps: Visual processing and linguistic processing. To assure that the generated captions are grammatically and semantically correct and to deal with problems arising from the corresponding modality and integrated competently, techniques of computer vision and NLP are utilized So, by this end, many methods discussed below.

Though the image captioning task is complex, the latest breakthroughs in deep neural networks [11–16], used extensively in the domain of computer vision [17–20] and NLP [21–24], made it easier and hence image caption generating machines based on deep neural networks came into existence. Robust deep neural networks implement effective solutions to visual and language modelling. Therefore, they are used to supplement existing systems and design many new approaches. Engaging deep neural networks to handle the image captioning task resulted in state-of-the-art outcomes [25–30].

With the recent progress in transfer learning and image captioning, we propose a novel architecture that compares multiple transfer learning models based on different metrics includes BLEU Score and others.

2 Related Work

Progress in the field of machine learning has opened new avenues of using deep neural networks instead of hand-engineered features and shallow models used earlier.

To generate a descriptive caption for a given image, Socher et al. applied dependency-tree recursive neural networks to convert phrases and sentences into compositional vectors. They used another deep neural network [31] to convert images into a feature vector.

Ma et al. proposed a multimodal convolutional neural network [32] to measure the relationship between images and captions based on different levels of interactions between them. This framework included CNN's to encode the image [33], [34], a matching CNN to relate visual and textual data [35-38] and multilayer perceptrons for scoring compatibility of image and caption data. The author used various modifications of matching CNN's to establish the correct relationship between images and captions. An ensemble of the multimodal convolutional neural network determines the final matching score.

By taking into consideration, recent advances in neural machine translation [22], [39], [40], the encoder-decoder framework is applied to generate captions for images. Kiros et al. introduced an encoder-decoder framework in the field of image captioning to merge joint image-text embedding models and multimodal neural language models so that a sentence output is generated word by word[33] for a given image like language translation. For the purpose of encoding the data, Kiros et al. used Long Short-Term Memory (LSTM) [41] and a Convolutional Neural Network(CNN) to encode the given image. Then, by reducing the pairwise

DOI: 10.5281/zenodo.5196025

Received: March 01, 2021 Accepted July 27, 2021

ranking loss, this encoded visual data is extended into an embedding space spanned by LSTM hidden states that encode textual data. Finally, a structure-content neural language model is used to decode visual features based on context word feature vectors to generate captions word by word.
Inspired by the human visual attention mechanism [42], [43], utilized attention mechanism to guide the image caption generation. By adding an attention mechanism to the encoder-decoder framework, caption generation depended on the values generated by the hidden states as well as on different parts of the image served by attention mechanism.

3 Methodology

In this paper, we proposed a novel architecture for image caption generation which extracts image feature vectors using different pretrained models and encoder-decoder model for caption generation. Following are the tasks performed in chronological order to achieve the goal.

3.1 Data Collection

In this paper, we compare the results of a different pre-trained model on benchmark datasets like Flickr 8k and Flickr 30k.
Flickr8k [26] contains 8,000 images obtained from Flickr. The images on this dataset include people and animals. There are five sentences corresponding to every image collected by a crowdsourcing service from Amazon Mechanical Turk. During the image annotation process, workers are instructed to just focus on the image ignoring the context behind them.
Flickr30k [44] is an extended Flickr8k dataset. There are in total 31,783 image data points in the dataset. Every image is annotated with five captions deliberately written for it. The images in this dataset are about but not limited to people involved in normal activities and daily events.

3.2 Text Cleaning

For text cleaning, some basic cleaning operations are performed like converting all the words to lowercase (otherwise “hello” and “Hello” will be considered as two different words), deleting special tokens (like ‘%, ‘$, ‘#, etc.), eliminating words which are alphanumeric (like ‘hey199’, etc.).
A vocabulary is created of all the words present in all the captions. A vocab is created which stores all the word and their frequency, respectively. For creating a caption predictive model, words with higher occurence rate present in the vocabulary are the ones which are more likely to occur or which are quite frequent are chosen by deciding the threshold i.e. the minimum frequency of a word in the entire dataset. This removes the model’s dependency on outliers and make model more robust. The maximum length of the caption among all the captions is calculated i.e. the count of words present in the caption respectively. Total number of distinct word are called vocab size. Vocab size decides the total number of neurons in the output layer of the merge model.

3.3 Data Preprocessing

Data available from datasets need to preprocessed before feeding them to train the models. There are 2 types of data available to us i.e. images and their corresponding captions. Following are the detailed explanation of preprocessing each of the inputs individually.

3.3.1 Captions

The main motive is to generate appropriate captions for every image. So, during the training phase, The model is learning to generate the captions as the target variables (Y). The prediction of caption for the given image is not done at once. The prediction of the caption is done iteratively word by word. Thus, each word is needed to be encoded into a fixed size word vector. Two tables are maintained word to idx and idx to word. Word to idx is the table in which all the words are stored, and each word is given a integer value. Similarly index to word is also a table which is storing the reverse mapping of the word to index table. Lets say word to index table stores a word “abc” and is mapped with a integer value K then the index to word table will have a integer K which will be mapped with the word “abc”.
The preprocessing of the caption is done to maintain some basic similarity between all the caption present. So, in order to achieve the above we modify out caption and add a special token at the starting and ending in each

DOI: 10.5281/zenodo.5196025

Received: March 01, 2021 Accepted July 27, 2021

of the caption. This will help the model to remember whether a caption is starting or ending. These special starting("<seqstart>") and ending("<seqend>") tokens are also added to both the tables word to index and index to word, respectively. All the captions are made of similar length so to not waste the import data that is collected, all the captions are made to have the same number of words as the caption with maximum number of words. This is done with the help of padding the captions which have less number of words is padded with a word and that word is stored in the word to index table with an integer 0 and in index to word at an index 0.

3.3.2 Images

All the datasets consist of images with real world entities. To make things easier, we can make use of transfer learning as the pre-trained models such as ResNet, Inception, and EfficientNet are available which are already trained on the millions of similar images and have a remarkable accuracy in classifying the images. These models can be used to extract feature vectors from images.
Datasets contains multi-channel images stored the form of 3-d matrix having the red, green and blue color channels values. The values are in the range from 0 to 255. To speed up the process, computation speed scaling is performed, and the values of each pixel is made to lie in range of -1 to 1. After removing the final layer from the pretrained model, a bottleneck layer is used to produce the information vector. Image features vector is then passed as an input to the final merge model.

4 Model Architecture

In this paper, we propose a novel architecture for image caption generation as shown in Figure 2. The model firstly consists of a transfer learning model which converts the input image to a feature vector and an embedding layer that embeds the captions corresponding to each image. This part of the model is called Feature Extraction Model. Then the embedded data is sent over an encoder-decoder network with soft attention which generates a next word given the image and the partial sentence. This model is known as the Merge model.
For a given image, the model generates a sequence of words as an output caption y encoded in the form of $1 - of - K$ vectors,

$$y = \{y_1, y_2, y_3, ..., y_n\}, y_i \in R^K \qquad (1)$$

Here the vocabulary size is K and n represents the caption with maximum length.
Following are the detailed explanation of these models with different techniques used.

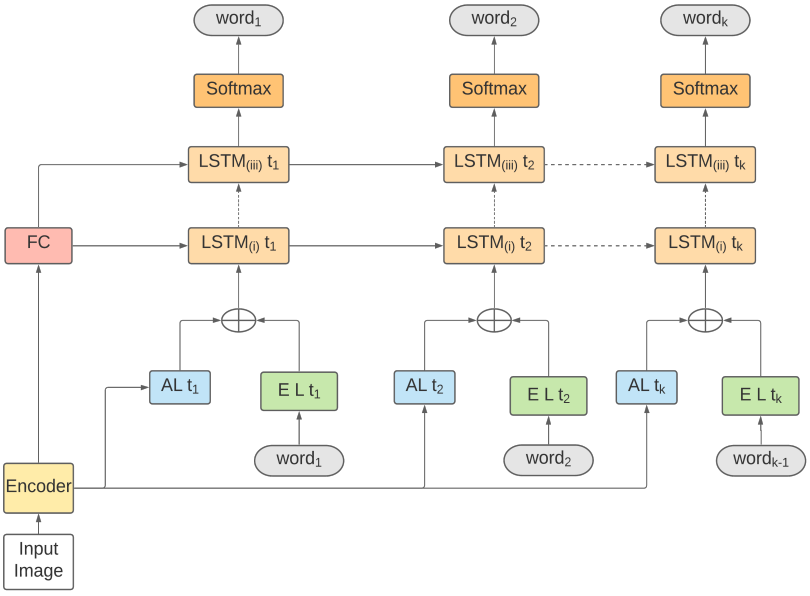

Figure 2: Proposed model architecture

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

4.1 Feature Extraction Model

This model consists of a pre-trained deep learning model to convert images into feature vectors and a word embedding layer to convert captions into feature vectors. Image feature vectors are extracted by slicing the last softmax layer from the transfer learning model.
Using the different pre-trained model to encode the image, we extracted V vectors of D dimensions each from the lowest convolution layer i.e global average pooling layer, such that each of them represents a part of the image. Stated formally, we extract an image representation p such that,

$$p = \{p_1, p_2, p_3, ..., p_V\}, p_i \in R^D \qquad (2)$$

Different pre-trained models were used to extract feature vectors from the images.

4.1.1 ResNet

A residual neural network (ResNet) [48] is a deep learning model whose construction is based on science behind the pyramidal cells in the cerebral cortex. ResNet models achieve this capability by making the use of skip connections. Generally double or triple layer skips are used with non-linearities (ReLU) and batch normalization to implement general ResNet models. HighwayNets are a special Resnet model which utilizes a separate weight table to learn the skip weights. DenseNets are models with several parallel skips.
With a considerably increased depth, Residual networks are easier to optimize and achieve high accuracies. Residual networks with a depth of 152 layers have lower complexity irrespective of being 8 times deeper than VGG nets. An ensemble of these ResNets achieved 3.57% error on the ImageNet test-set.

4.1.2 Inception

The Inception network [49] was a principal invention in the evolution of Convolutional classifiers. This invention removed the general idea of stacking up convolutional layers to make CNN deeper to get better performance. The authors of the inception net proposed to make the network wider rather than deeper. The inception net use three different sizes of filters i.e 1x1, 3x3 and 5x5 to perform convolution operation with an additional max-pooling layer. Their outputs are then concatenated and sent to the next inception module. Inception net has evolved iteratively with each version having a significant improvement over the previous one.
The inception net v3 inherited all updates from inception net v2 with the addition of rmsprop optimizer, factorize 7x7 convolution, batch normalization in the auxiliary classifier and label smoothing.

4.1.3 Efficient Net

Efficient Nets [50] are the new series of models which utilizes neural architecture search to obtain and scale new baseline networks. Efficient Nets have achieved much better accuracy and efficiency than previous Convolutional Networks. In particular, Efficient Net-B0 being 4.9x smaller and 4x faster on inference achieved state-of-the-art 77.1.3% top-1 accuracy on ImageNet.
Efficient Nets uses a multi-objective neural architecture search that optimizes both accuracy and FLOPS. Using the compound scaling method, Efficient Net models can scale up effectively, surpassing state-of-the-art accuracy having fewer training parameters and FLOPS.

We use pre-trained Word Embeddings for the input sequence of words. For each word, we get an embedding vector of length m where $m = 300$ in this case. Hence for a sequence of n words, we get a nm matrix as an input to the LSTM Here m represents the number of features in the input and n represents the length of the sequence.
More formally, we can state that :

$$X = x1, x2, ..., xn \qquad (3)$$

Here n represents the length of the sequence, the embedding layer generates a sequence :

$$E = e1, e2, ..., en \qquad (4)$$

Here n represents the length of the sequence and $e_i \in R^{m \times K}$. Here K represents the size of the vocabulary and m represents the embedding dimensions.

These image feature vectors and embedded captions are then fed to merge model to generate captions.

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

4.2 Merge Model

Up till now, we have encoded images and captions into feature vectors. The output of the image encoder and the encoded form of the caption generated so far are merged in merge model to predict the caption.
Figure 3 shows the visual representation of the working of the merge model.

X1,	X2(text sequence),	y(word)
photo	startseq,	little
photo	startseq, little,	girl
photo	startseq, little, girl,	running
photo	startseq, little, girl, running,	in
photo	startseq, little, girl, running, in,	field
photo	startseq, little, girl, running, in, field,	endseq

Figure 3: Data Flow during training

The combination of these two encoded inputs is then used to generate the next word in the sequence.
In merge architectures, we used Three Stacked Deep Long Short Term Memory(LSTM) layers with soft attention to combine both the encoded image input with the caption generated so far.

4.2.1 Long Short Term Memory Network

We developed a Deep Long Short Term Memory (LSTM) network by stacking multiple LSTM layers on top of each other such that the input of an LSTM layer is the output of the previous LSTM layer. The basic description of an LSTM cell is depicted in Figure 4. In an LSTM cell, there is a single cell state and three gates which are the input, output, and forget gates.
During each time-step t, the generated cell state c_{t1} and the hidden state h_{t1} which were generated during previous time-step $t1$, are forwarded back to the LSTM. The input u_t is received at present time t. If we use $f_{LSTM}(\cdot)$ to represent a feed-forward function of LSTM, we can say that the LSTM updates its state by :

$$h_t, c_t = f_{LSTM}(u_t, h_{t1}, c_{t1)} \qquad (5)$$

The internal computations of the LSTM on its gates and memory cells are enumerated as follows :

$$i_t = \sigma(W_i u_t + R_i h_{t1} + b_i) \qquad (6)$$

$$f_t = \sigma(W_f u_t + R_f h_{t1} + b_f) \qquad (7)$$

$$o_t = \sigma(W_o u_t + R_o h_{t1} + b_o) \qquad (8)$$

$$z_t = tanh(W_z u_t + R_z h_{t1} + b_z) \qquad (9)$$

$$c_t = i_t * z_t + f_t * c_{t1} \qquad (10)$$

$$h_t = o_t * tanh(c_t) \qquad (11)$$

Here R represents the recurrent weight matrix, W represents the input weight matrix learned during training and b represents the bias vector. Also, σ denotes the sigmoid function which is expressed by $\sigma(x) = 1/(1 + exp(x))$. It has a squashing effect and condenses the input into the range of (0, 1). $Tanh$ is a hyperbolic tangent function and produces values in the range (-1,1) to avoid the explosive growth of values over time. Both the functions are computed in an element-wise manner. i_t, o_t and f_t denote the input, output and forget gates respectively. To compute them we add the linear projections of u_t and h_{t1} followed by the output of the sigmoid function. The input transformation z_t, the cell value of the previous state c_{t1} and the output of the element-wise multiplication output which is denoted by $*$ are modulated by the input, forget and output gates, respectively.

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

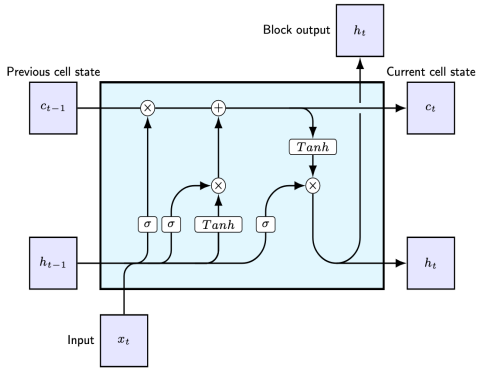

Figure 4: Basic LSTM Cell

4.2.2 Stacked LSTM

We use a Three-Layer stacked LSTM where the input of one layer is the output for the previous one and the output of the last layer is the output of the LSTM stack. The initial hidden and cell states,h_0^1 and c_0^1 for the first LSTM layer, h_0^2 and c_0^2 for the second LSTM layer, h_0^3 and c_0^3 for the third LSTM layer are calculated by first calculating an average of the set of vectors p and then feeding it through two separate Multi-Layer Perceptrons represented as $f_{init\_c}^i$ and $f_{init\_h}^i$,$i = 1, 2, 3$:

$$h_0^i = f_{init\_h}^i(\frac{1}{L}\Sigma a_i) \qquad (12)$$

$$c_0^i = f_{init\_c}^i(\frac{1}{L}\Sigma a_i) \qquad (13)$$

where the annotation vectors $a_i, i = 1, ..., L$ are the features that correspond to different image sub-regions. Also, since we use a multi-layer LSTM stack, the first layer $LSTM^1$ takes the word embeddings of the input sequence as the input. For LSTM layers $LSTM^2$ and $LSTM^3$, the input is the output of the previous layer.

4.2.3 Soft Attention

In the soft attention approach, in addition to the word embeddings of the input sequence, we also use context vector s_t i.e. a representation of the image for that time-step which provides information about the relevant portion of the image for that time-step. Hence the LSTM outputs would be calculated according to the following equations :

$$h_t, c_t = f_{LSTM}(u_t, z_t, h_{t1}, c_{t1}) \qquad (14)$$

where,

$$i_t = \sigma(W_i E_{y_{t1}} + R_i h_{t1} + S_i s_t + b_i) \qquad (15)$$

$$t = \sigma(W_f E_{y_{t1}} + R_f h_{t1} + S_f s_t + b_f) \qquad (16)$$

$$o_t = \sigma(W_o E_{y_{t1}} + R_o h_{t1} + S_c s_t + b_o) \qquad (17)$$

$$z_t = tanh(W_z E_{y_{t1}} + R_z h_{t1} + S_c s_t + b_z) \qquad (18)$$

$$c_t = i_t * z_t + f_t * c_{t1} \qquad (19)$$

$$h_t = o_t * tanh(c_t) \qquad (20)$$

For the computation of the context vector s_t, we adopt the same mechanism, where for each annotation vector a_i corresponding to a location in the image, a positive weight α_i calculated by the attention mechanism which denotes the relative importance of the location for generating the word at the present time-step. The attention model m_{att} is a function of α_i and h_{t1}, i.e, the annotation vector and the hidden state at time-step $t1$.

$$x_{ti} = m_{att}(a_i, h_{t1}) \qquad (21)$$

$$\alpha_{ti} = \frac{exp(e_{ti})}{\Sigma exp(e_{tk})} \qquad (22)$$

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

From the weights associated with each image region, the context vector s_t can be calculated as :

$$s_t = \phi(a_i, \alpha_i) \qquad (23)$$

where ϕ returns a single vector for the image.

5 Result

We perform extensive experiments to evaluate the proposed models. We report all the results on Flickr8k and Flickr30k dataset using BLEU and METEOR.

5.1 Evaluation Metrics

5.1.1 BLEU

BLEU [45] is an evaluation metric that is used to match variable length phrases of a predicted or generated sentence to original sentences which are written by humans to measure their correlation. BLEU score is calculated by comparing a predicted or generated sentence with original sentences in n-grams. Categorically, BLEU-1 is calculated by comparing the predicted sentence with original sentences in uni-gram, while BLEU-2 is calculated by using bigram for matching. The best correlation with human judgement is obtained by empirically determining BLEU with a maximum order of four. In BLEU higher n-gram scores are responsible for fluency and the uni-gram scores are responsible for adequacy.

5.1.2 METEOR

METEOR [47] is an automatic machine translation evaluation metric. It has two steps of the calculation, first one being performing generalized uni-gram matches between a predicted sentence and original sentence which are written by humans and the second one computing a score based on the results matched. The computation associates calculation of recall, precision and alignments of words matched. While calculating for more than one original sentence, the highest score among all the scores that are calculated independently is considered as the final result for the predicted sentence. This metric was introduced to address the shortcoming of BLEU metric, which is based only on the precision of the n-grams matched.

5.2 Performance on Flickr 8k

We calculated different evaluation metrices on different models to find out the best pretrained model for image captioning.
Model based on Resnet 50 performed well on Flickr8k dataset with BLEU-1 Score of 0.62 and METEOR Score of 0.153. Inception Model performed better than Resnet 50 with BLEU-1 Score of 0.627 with difference of 0.07 from Resnet. Inception model scored 0.157 on METEOR
EfficientNet performed best out of the 3 models compared with a BLEU-1 Score and METEOR score of 0.636 and 0.16 respectively. EfficientNet being 4 times lighter than the others achieved an increase of 0.01 in BLEU-1 metrices.
Table 1 and Figure 5 depicts detailed comparison between these models over different metrics on Flickr8k dataset.

Table 1: Result from Flickr 8k Dataset.

Image Encoder	**BLEU-1**	**METEOR**
ResNet	0.62	0.153
InceptionNet	0.627	0.157
EfficientNet	0.636	0.162

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

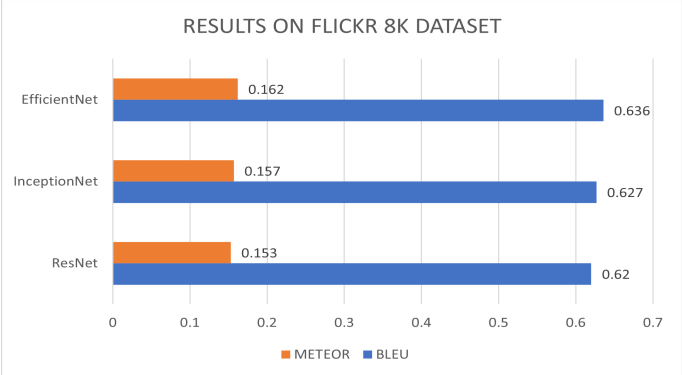

Figure 5: Comparison of different image encoders on Flickr 8k dataset

5.3 Performance on Flickr 30k

The available models performed better on Flickr 30k dataset due to availability of 4 times more training samples. Model based on Resnet 50 performed well on Flickr30k dataset with BLEU-1 Score of 0.651. and METEOR Score of 0.172. Inception Model performed better than Resnet 50 with BLEU-1 Score of 0.657 with difference of 0.006 from Resnet. Inception model scored 0.179 on METEOR metrices. EfficientNet performed best out of the 3 models compared with a BLEU-1 Score and METEOR score of 0.665 and 0.184 respectively. EfficientNet being 75% lighter than the others achieved an increase of 0.01 in BLEU-1 metrices.
Table 2 and Figure 6 depicts the detailed comparison between these models over different metrics on Flickr30k dataset.

Table 2: Result from Flickr 30k Dataset.

Image Encoder	BLEU-1	METEOR
ResNet	0.651	0.172
InceptionNet	0.657	0.179
EfficientNet	0.665	0.184

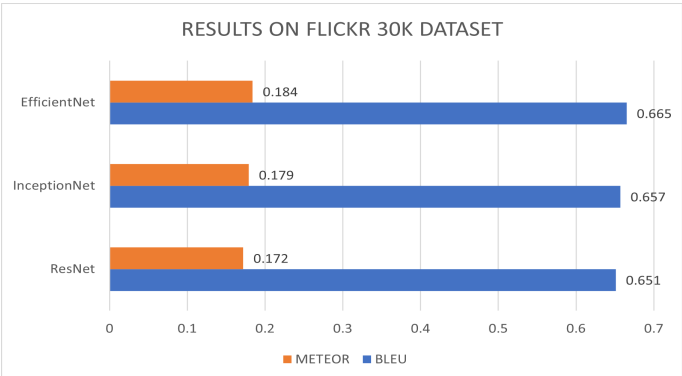

Figure 6: Comparison of different image encoders on Flickr 30k dataset

5.4 Quantitative Results

We compare our methods with the comparable methods proposed in the literature in Table 3 and Figure 7. Images with their generated captions as predicted by the system shown in Figure 8-13

DOI: 10.5281/zenodo.5196025

Received: March 01, 2021 Accepted July 27, 2021

Table 3: Quantitative Result from Flickr 30k Dataset.

Method Name	**BLEU**	**METEOR**
Karpathy et al.[51]	0.573	–
LRCN[53]	0.5872	–
Mao et al.[29]	0.5479	–
Soft-Attention[52]	0.667	0.1849
Our Method	0.665	0.1840

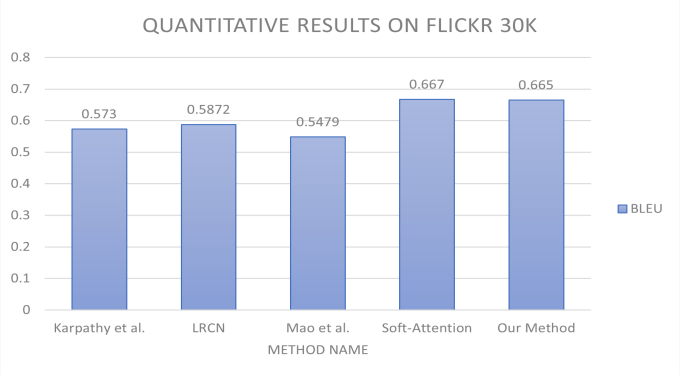

Figure 7: Quantitative comparison of different methods on Flickr 30k dataset

Figure 8: Generated Caption: A black and white dog is running through field of grass

Figure 9: Generated Caption: A young boy swims in the water

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

Figure 10: Generated Caption: A dog is running through the snow

Figure 11: Generated Caption: Woman in white dress and hat sits on bench and plays in bench.

Figure 12: Generated Caption: Group of wedding party at the beach.

Figure 13: Generated Caption: Man and dog are walking down path lined terrain.

DOI: 10.5281/zenodo.5196025
Received: March 01, 2021 Accepted July 27, 2021

6 Conclusion

After evaluating different pre-trained models on available benchmark datasets, we conclude that the latest pre-trained model Efficient Net performed best for image captioning. Being 75 % lighter than the other existing models, training time for the model was reduced to 12 sec per epoch and BLEU score increased upto 0.015 as compared to others. METEOR Score for efficient net model surpassed other models by a value of 0.12 making it best image feature extractor so far.

References

[1] L. Fei-Fei, A. Iyer, C. Koch, P. Perona., What do we perceive in a glance of a real-world scene? J. Vis. 7 (1) (2007) 1–29.

[2] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[3] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580–587.

[4] C.H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between class attribute transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 951–958.

[5] C. Gan, T. Yang, B. Gong, Learning attributes equals multi-source domain generalization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2016, pp. 87–97.

[6] L. Bourdev, J. Malik, S. Maji, Action recognition from a distributed representation of pose and appearance, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, 2011, pp. 3177–3184.

[7] Y.-W. Chao, Z. Wang, R. Mihalcea, J. Deng, Mining semantic affordances of visual object categories, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 4259–4267.

[8] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Twenty Fifth International Conference on Neural Information Processing Systems, 2012, pp. 1097–1105.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Proceedings of the Advances in Neural Information Processing Systems (NIPS), 2014, pp. 487–495.

[10] Y. Gong, L. Wang, R. Guo, S. Lazebnik, Multi-scale orderless pooling of deep convolutional activation features, in: Proceedings of the European Conference on Computer Vision, 2014, pp. 392–407.

[11] H. Goh, N. Thome, M. Cord, J. Lim, Learning deep hierarchical visual feature coding, IEEE Trans. Neural Netw. Learn. Syst. 25 (12) (2014) 2212–2225.

[12] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.

[13] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, DeCAF: a deep convolutional activation feature for generic visual recognition, in: Proceedings of The Thirty First International Conference on Machine Learning, 2014, pp. 647–655.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, arXiv:1408.5093v1 (2014).

[15] N. Zhang, S. Ding, J. Zhang, Y. Xue, Research on point-wise gated deep networks, Appl. Soft Comput. 52 (2017) 1210–1221.

DOI: 10.5281/zenodo.5196025

Received: March 01, 2021 Accepted July 27, 2021

[16] J.P. Papa, W. Scheirer, D.D. Cox, Fine-tuning deep belief networks using harmony search, Appl. Soft Comput. 46 (2016) 875–885.

[17] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling., IEEE Trans. Pattern Anal. Mach. Intell. 35(8).

[18] E.P. Ijjina, C.K. Mohan, Hybrid deep neural network model for human action recognition, Appl. Soft Comput. 46 (2016) 936–952.

[19] S. Wang, Y. Jiang, F.-L. Chung, P. Qian, Feedforward kernel neural networks, generalized least learning machine, and its deep learning with application to image classification, Appl. Soft Comput. 37 (2015) 125–141.

[20] S. Bai, Growing random forest on deep convolutional neural networks for scene categorization, Expert Syst. Appl. 71 (2017) 279–287.

[21] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv:1409.0473v7 (2016).

[22] K. Cho, B.V. Merrinboer, C. Gulcehre, Learning phrase representations using RNN encoder–decoder for statistical machine translation, arXiv:1406.1078v3 (2014).

[23] R. Collobert, J. Weston, A unified architecture for natural language processing:deep neuralnetworks with multitask learning, in: Proceedings of the Twenty Fifth International Conference on Machine Learning, 2008, pp. 160–167.

[24] A. Mnih, G. Hinton, Three new graphical models for statistical language modelling, in: Proceedings of the Twenty Fourth International Conference on Machine Learning, 2007, pp. 641648.

[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in Neural Information Processing Systems, 2013.

[26] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, J. Artif. Intell. Res. 47 (2013) 853–899.

[27] H. Fang, S. Gupta, F. Iandola, R. Srivastava, From captions to visual concepts and back., in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.

[28] J. Donahue, L. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2625–2634.

[29] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep captioning with multimodal recurrent neural networks, in: Proceedings of the International Conference on Learning Representation, 2015.

[30] M.R.R.M.S. L A Hendricks, S. Venugopalan, Deep compositional captioning: describing novel object categories without paired training data, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1–10.

[31] A. Karpathy, A. Joulin, F. Li, Deep fragment embeddings for bidirectional image sentence mapping, in: Proceedings of the Twenty Seventh Advances in Neural Information Processing Systems (NIPS), 3, 2014, pp. 1889–1897.

[32] L. Ma, Z. Lu, Lifeng, S.H. Li, Multimodal convolutional neural networks for matching image and sentences, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2623–2631.

[33] R. Kiros, R. Salakhutdinov, R. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, arXiv:1411.2539(2018).

[34] Q.V. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A.Y. Ng, Building high-level features using large scale unsupervised learning, in: Proceedings of the International Conference on Machine Learning, 2012.

DOI: 10.5281/zenodo.5196025

Received: March 01, 2021 Accepted July 27, 2021

[35] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556v6 (2015).

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermannet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich., Going deeper with convolutions, arXiv:1409.4842 (2018).

[37] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Proceedings of the Twenty Seventh International Conference on Neural Information Processing Systems, 2014, pp. 2042–2050.

[38] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, arXiv:1404.2188v1 (2014).

[39] N. Kalchbrenner, P. Blunsom, Recurrent continuous translation models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2013.

[40] I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2014.

[41] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.

[42] R. A.Rensink, The dynamic representation of scenes, Vis. Cognit. 7 (1) (2000) 17–42.

[43] M. Spratling, M.H. Johnson, A feedback model of visual attention, J. Cognit. Neurosci. 16 (2) (2004) 219–237.

[44] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, in: Proceedings of the Meeting on Association for Computational Linguistics, 2014, pp. 67–78.

[45] K. Papineni, S. Roukos, T. Ward, W. Zhu, BLEU: a method for automatic evaluation of machine translation, in: Proceedings of the Meeting on Association for Computational Linguistics, vol. 4 (2002).

[46] C.-Y. Lin , F.J. Och , Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in: Proceedings of the Meeting on Association for Computational Linguistics, 2004 .

[47] A . Lavie , A . Agarwal , METEOR: an automatic metric for MT evaluation with improved correlation with human judgments, in: Proceedings of the Second Workshop on Statistical Machine Translation, 2007, pp. 228–231 .

[48] He, Kaiming, 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, pp. 770-778. 2016.

[49] Szegedy, C., W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going deeper with convolutions. arXiv 2014." arXiv preprint arXiv:1409.4842 1409 (2014).

[50] Tan, Mingxing, and Quoc V. Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." arXiv preprint arXiv:1905.11946 (2019).

[51] Karpathy, A., Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[52] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).

[53] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2625-2634).

DOI: 10.5281/zenodo.5196025

Received: March 01, 2021 Accepted July 27, 2021