Journal of Positive School Psychology
2022, Vol. 6, No. 3, 6621–6634

Deep Image Captioning system using Attention Two-Layer LSTM Network

Mrs. Priyanka G*, Harsha mol B.P, Sneha R

priyanka@mepcoeng.ac.in, reachharshamol_cs@mepcoeng.ac.in, sneravi126_cs@mepcoeng.ac.in
Department of Computer Science and Engineering
MEPCO Schlenk Engineering College, Sivakasi

Abstract

Image captioning techniques are the algorithmic approach of automatic generation of one or more natural language based textual sentences for an input image. Image captioning is a cross-modal challenge that necessitates the automatic generation of natural sounding phrases to describe the semantic information present in an image. As there is an enormous gap between human visionary understanding and the corresponding natural language description, most of the existing techniques suffer from poor semantic matching between images and generated captions. The generation of caption from an image is intriguingly very challenging as it is the processing of bridging human vision with computer vision. With the invention of deep learning techniques many real-time applications in different modalities including the topic of image captioning have had a lot of success. The work is about a model for image captioning by implementing various pre-trained CNN models for feature extraction with custom defined LSTM layers for captioning. The model was developed using various CNN models such as Resnet-152, ResNet-50, VGG16, Inception V3 to extract features from the images. When compared to other architectures, ResNet-152 performs better. The ResNet-152 acts as a generator block for extracting features from images, while the LSTM acts as a decoder unit, generating words that describe the image. After the caption generation phase, to evaluate the effectiveness of our methodology, the model was evaluated using a variety of metrics like BLUE-n and ROUGE-L. As a result, our method assists the user in obtaining a descriptive caption for the input image with improved performance as compared to the existing state-of-the-art methods.

I. INTRODUCTION

Visual and imaging began to take over social and professional interactions around the world. Caption generation is also becoming a booming industry around the world. Many data annotation companies are making billions as a result of this. The act of identifying the context of a picture and annotating it with relevant captions using deep learning and computer vision is known as image caption generation. Image captioning has a variety of applications, including suggestions in altering software, accessibility for people with visual impairments, image indexing, social platforms, and a number of other Natural Language Processing (NLP) applications. Many scenarios, such as content-based picture retrieval, visual information extraction, and visual conversation can benefit from image captioning.

To annotate visual objects automatically using utterance inNLP has been animportantandhuge challenge for computer vision tasks. With the advancements of deep learning algorithms,the topic of image captioning, in which a single natural sentence is used to describe each image, has gotten a great deal of focus. Researchers have generated a slew of benchmark datasets to go along with this trend (e.g., Microsoft COCO [1] and Flickr 30K [2]), each containing tens of thousands or hundreds of thousands of photos with natural words annotated, to aid in image captioning research.

The following are the two types of image captioning models: 1. A model based on traditional machine learning 2. A model based on deep learning. Because they rely on pre-existing templates or texts [6,7], traditional machine learning models have a number of drawbacks. When compared to traditional models, deep learning-based models [8] outperform them. Both encoder decoder approaches are used in the deep learning picture captioning model. The encoder extracts the image's feature in vector representation, while the decoder generates the phrases. The model may produce many phrases from a single image. These models are developed using a supervised and unsupervised learning technique[9] that necessitates the use of a big dataset. In supervised learning, annotating each image with its accompanying phrase is costly. The majority of these models employ a CNN encoder for image encoding and an RNN decoder [14,15] for phrase generation. RNN has its own set of drawbacks, including low computation and the vanishing and exploding gradient problem.

Most existing methods for generating coherent descriptions start from images by extracting visual cues and then building an RNN model to convert visual cues into meaningful phrases[3,4,5]. However, they overlook the relationship between visual semantic units and their textual equivalents, resulting in erroneous semantic relationships between captions and images.

The work is about an image captioning model with ResNet-152 used here. To encode phrases, GRU is employed along with word embeddings as the encoder. The technique of turning text into numerical representations is known as text vectorization or word embedding. The model uses LSTM with two layers: Visual attention LSTM and Adaptive language LSTM. Image-sentence pairs are generated using LSTM[27]. The LSTM is used to overcome the RNN's restrictions. The Flickr 8K dataset is used to pre-train our model. Finally, the real-time photos will be used to test adaptation models. Non-differentiable metrics such as Bleu-n[10], and ROUGE L[12] are used to evaluate these models.

This paper's contributions are summarized as follows:

- By modelling dialectal forms of the sentences, the work is based on a custom flexible LSTM that can successfully adapt from the training data to the testing data.
- Using Flickr 8K datasets, our strategy beats cutting-edge algorithms in five different circumstances.

The other sections of this paper are organized as follows. In Section 3, there are some connected works. Section 4 explains the proposed work and discusses the system architecture. In Section 5, experimental results and evaluation process of the model has been explained. The paper is concluded in section 6, and the references are listed in section 7.

II. RELATED WORK

Deep Captioning:

Image captioning performance can be improved using deep learning approaches. To create sentences, most deep learning approaches use an encoder-decoder system, the encoder is indeed a CNN, and RNN as the decoder. In [14] for producing new phrase descriptions to describe visual data, they have described a multimodal Recurrent Neural Network (m-RNN) model. This model is made up of two sub-networks: an image-based convolutional neural network and a text-based deep recurrent neural network. These two sub- networks interacting in a multimodal layer build the entire m-RNN model.

[15] generates captions for a given image using a Multimodal RNN architecture. The images are detected and the embedding matrix is formed using RCNN, while the words for the sentences are computed using Bidirectional RNN. The image sentence pair is created using RNN. The captions in [16] are generated using CAM – RNN. The visual and textual features are encoded using CAM, and the captions are generated using RNN. [17] captures global information with RNN and receives feedback and interactivity through bottom-up attributes enforced on both the RNN and the output. This feedback allows [17] to better precisely forecast the description.

In a shared latent space trained using deep canonical correlation analysis, there are some challenges with matching images and captions (DCCA). The computational complexity of the

features poses considerable difficulties in the process of storage and execution complexity when applied in the DCCA framework. [18] proposes strategies to deal with overfitting and uses GPU implementation to overcome these difficulties. [19] employs a paradigm for picture captioning that is based on attributes. Pretrained CNN extracts image features, which are then immediately input into the LSTM.

Krizhevsky, Sutskever, and G. E.[20] has taken non-saturating layers and a GPU version of the convolution technique to speed up training. It has five convolutional operations. To reduce overfitting for fully-connected layers, [20] utilised a freshly proposed regularisation approach called "dropout," which proved to be highly successful.

They proposed an unique unsupervised strategy for training an image translation model without any linked picture-sentence data in [21]. They suggested a technique for creating descriptions for items that aren't found in paired image sentence collections. Quantitative and qualitative results [19], [22] and [32] demonstrate that the proposed algorithm can efficiently integrate new language into produced picture and video summary using certain sight datasets and mismatched textual information.

In our model, we encode the images with different architectures (Resnet-152, ResNet-50, VGG16, Inception V3) and extract the features from the text with GRU. As a decoder, a two-layer LSTM is used. First ever LSTM layer is a visual recognition model that weighs each feature, while the other one is a language adaptation model. Using a two-layer LSTM pseudo image-sentence pair will be generated.

Cross Modal Retrieval:

Cross modal retrieval is the process of just using a single mode to get accurate information in another modality. [23] is a reinforcement-based training method in which the CNN-RNN captioner is taught to generate sentences and critics are used to reward the model based on the sentences. Both the captioner and the critics employ the adversial training technique iteratively. They designed a non-adversarial training technique in order to benefit from the unpaired information in this study. During training, the captioners and critics are pitted against one other.

[24] uses a Cross-modal Relation Guided Network (CRGN) to incorporate image and text in a hidden feature set. The ResNet model is being used to learn the universally guided image feature in the CRGN model, whereas GRU is used to extract text features. Global feature assistance and phrase generation learning can be used to model the relationship between picture areas. The final picture embedding is made using a relation embedding module with an attention method. Cross-modal recovery is focused on cosine similarity and is performed with picture and text embeddings. Picture retrieval along with statement as inquiry and statement retrieval with an image as inquiry are the two sub-tasks of cross-modal recovery between image and text sentence.

Learning connection between distinct data modalities is difficult due to the diverse gaps between them. CCA (canonical-correlation analysis) is an SVD-based way of learning the linkages between two distinct modalities. using linear projections. Deep neural networks are used by DCCA to provide complicated nonlinear features transformations of many data modalities. Because CCA and DCCA primarily focus on pairwise correlation learning, they will fail to gain label information. [25] Supervised-Deep Canonical Correlation Analysis (SDCCA) is used to improve correlation learning for cross-modal retrieval by utilising triplet neural networks (TNN).

Deep supervised cross-modal retrieval[26] is a method for determining a familiar representation space in which multiple modalities can be compared directly. It cuts down on losses of discrimination and modality invariance. Both discriminative and modality-invariant common representations are possible.

[27] combines a cross-modal retrieval model and an adaptable picture captioning model to create an innovative cross-domain picture captioning approach. [27] advocates using an incremental cross-modal retrieval methodology to construct a set of false picture-statement pairs by pre-training a cross-modal retrieval model on

origin datasets before applying it to targeted datasets. By repeatedly fine-tuning the retrieval model and updating the faux picture-statement pairings using the retrieval model, the mock image-sentence combinations are honed even more. CIDer[9] and METEOR[11] are the metrics used.

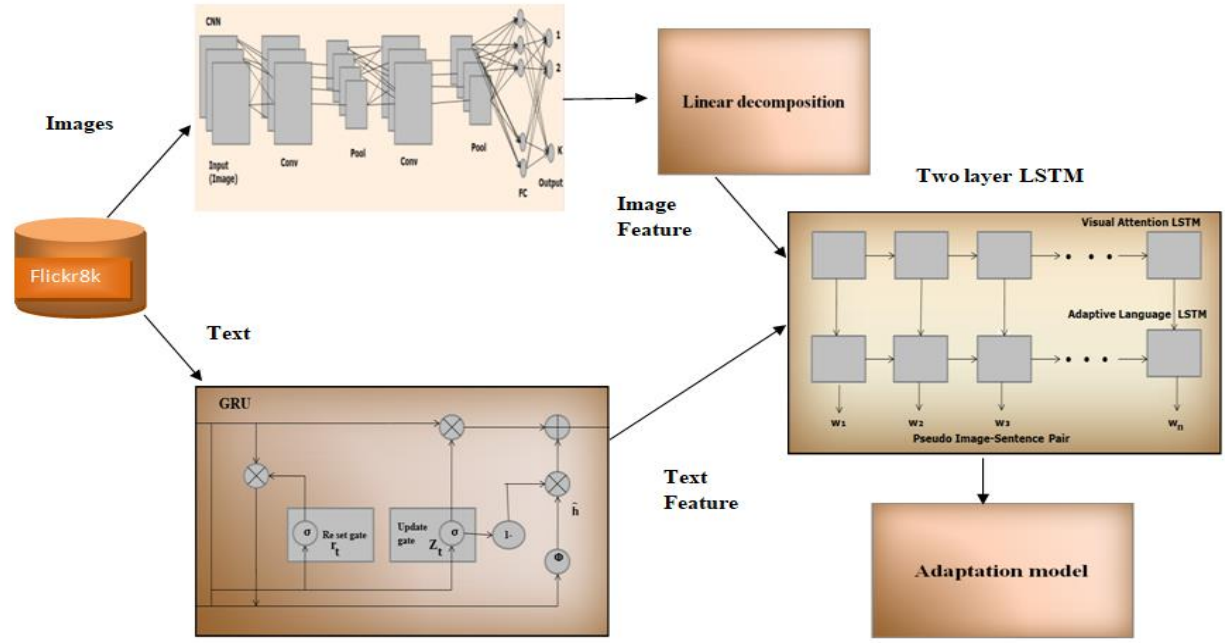
A Deep Visual-Audio Network [28] is forced to detect the relationship between picture and sound right away. [28] describes a novel cross-modal remotely sensed (RS) picture–voice retrieval technique that combines deep features and multimodal learning into a single framework for obtaining statements from pictures. The goal of cross-modal remotely sensed pictures–speech retrieval is to extract relevant distant sensing sounds from pictures. Obtaining useful information is difficult because pattern descriptions of phrases and pattern descriptions of pictures are incompatible.

[29] presents the Cross-modal Center Loss, a new loss function that is specifically tailored to reduce intra-class variation across different modalities. For all modalities, it minimises the distance between multimodal items and their cores in a certain shared feature set. The cross-modal difference between distinct modalities of information can be addressed via cross-modal centre deficit.

For cross-modal retrieval, we use paired image-sentence (Flickr 8K dataset) and unpaired image-sentences in our model. The image features are extracted using different architectures and the best one is used for our model. Thus, the generated sentence will be apt for the image. Also, many of the recent literatures cited are based on the usage of deep learning techniques in the field of text processing [30], [31], [32].

III. Proposed Work

The proposed system consists of five modules: Image feature extraction, text feature extraction, Adaptation model construction, real time deployment of the adoption model. The overall architecture of the proposed system is shown in figure 1.



Testing Phase:

Figure 1. System architecture diagram

The input data for training is provided for image captioning as: $Data_s = \{(x^s_i, y^s_i)|i\}$ x^s_i representing the i-th image and the phrase y^s_i describing x^s_i. Let $I_s = \{x^s_i|i\}$ signifies the set of source images and $C_s = \{y^s_i|i\}$ signifies the sentence set that was used.

The captioning model's vocabulary is determined by $Dict_s = \{w|w \in C_s\}$ $Dict_s$ contain words in the captions present in the dataset. The image captioning model is initially trained using the $Data_s$ with ResNet 152 as image feature extracting architecture. Then various CNN models such as VGG-16, Inception v3, ResNet 50 are replaced in place of ResNet 152 for extracting the image features.

A set of image sentence pairs O^i will be generated as output.

Algorithm: Image Caption Generator

Input: Flickr 8k Dataset

Output: Predicted Captions

```
1. images<- ImageFeaturePreProcessing (Flicks8k images)
2. For i in images:
3.     pretrained_ResNet152(i)
4.     dictionary<- TextFeaturePreProcessing(Flickr8k text)
5.     GRU(dictionary)
6.predicted_captions<- AdaptationModel(images, dictionary)
```

Image Feature Extraction:

To encode the image x_i into a feature map, the various CNN models such as VGG 16, Inception v3, ResNet 50, ResNet 152 were pre trained on ImageNet. These feature vectors are referred as

$F = \{v_1, v_2, ...v_L\}$,

where $v_i \in R^D$**.** The features are stored as a numpy (.npy) file.

Algorithm 1: ImageFeaturePreProcessing

Input: Images

Output: Pre-processed image

```
1. for i=1 to n:
2.     flag←0
3.     for j=1to m:
4.         if(images[i]==image[j])
5.             flag←flag+1
6.         if(flag==0):
7.             image1[m]=images[i]
8.             m=m+1
9. for i in image:
10.    cScale=colorScale(i)
11.    if(cScale==greyScale):
12.        R=G=B=grey
13.    else if(cScale==CMYK):
14.        R=255*(1-C)*(1-K)
15.        G=255*(1-M)*(1-K)
16.        B=255*(1-Y)*(1-K)
17.    else if(cScale==YUV):
18.        R=Y+1.140V
19.        G=Y-0.395U-0.581V
20.        B=Y+2.032U
21.    i←i.resize(299,299)
```

Text Feature Extraction:

Using GRU(Gated Recurrent Unit), text features will be extracted from the Flickr 8K dataset. In the Flickr 8K dataset each image has a minimum of five ground truth captions associated with it. Each captions will be standardized (Special characters will be removed and every character will be converted to lowercase). Each caption will be splitted into small substrings. Substrings will be recombined into index tokens. For every token there will be a separate integer value associated with it. These tokens are used for transforming the captions into a vector representation.

Algorithm 2: TextFeaturePreProcessingInput: Ground Truth Sentences

Output: Dictionary

```
1. caption=[],sentence=[],word=null
2. for sent in dataset(captions):
3.     for i in len(sent):
4.         if(sent[i]>='A'&& sent[i]<='Z')
5.             sent[i]=sent[i]+32
6.         if (sent[i] in [@#$%^&*()_+])
7.             sent[i]=null
8.     for i in len(sent)
9.         if (sent[i]>='a' && sent[i]<='z')
10.            word+=sent[i]
11.        else
12.            sentence.append(word)
13.        word=null
14.    caption.append(sentence)
```

Adaptation Model Construction:

In this module extracted feature vectors from the image and sentences are used to construct the adaptation model. The features are reshaped. For each word there will be an attention point using which the caption will be generated. Meanwhile the attention weights and

the context vectors are calculated for the attention points. The context vector is concatenated with the embedding matrix and it is taken as X. The word will be obtained for the attention point by passing X to the LSTM layer. These words are added to the result. The model produces image-sentence pairs in this module. This adaptation model is used for validation.

Real time deployment of the adaptation model:

The image will be sent as the input to the Image feature extraction in real time deployment. The retrieved features will be sent as input to the fine-tuned adaptation model after the image features have been extracted. The adaptation model has been fine-tuned to produce proper captions for the photographs.

Dataset:

The Flickr 8k Dataset comprises a total of 8092 JPEG photos in various forms and sizes. 6400 are used for training and 3600 are utilized for testing. Text files detailing the train set and test set can be found in the Flickr 8k text folder. Each image in Flickr 8k.token.txt has a maximum of 5 captions, for a total of 40460 captions.

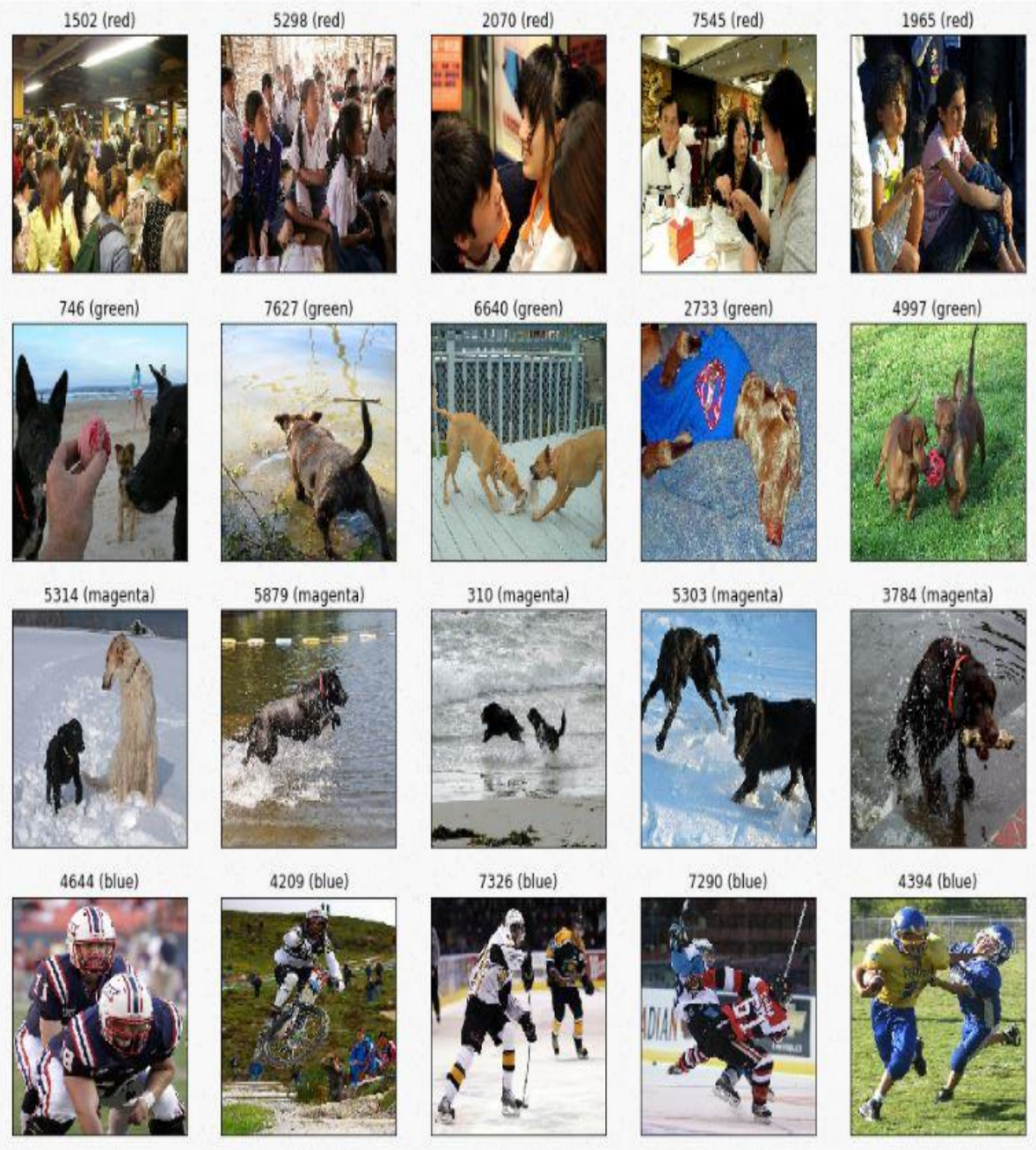


Figure 3: Images used for checking the feature dispersion

IV EXPERIMENTAL RESULTS

The ResNet-152 model pre-trained on Imagenet will be used for extracting the representation of the images. The model is trained using flickr 8k dataset. Unique images were collected from the dataset. The collected images were preprocessed and then the images are used as the input of the ResNet-152.The features are stored in a numpy file.

Fig 3 depicts the images based on particular features such as red color depicts that there are many people in one image, green color depicts that there are dogs are on green yard or on bed, magenta depicts that there are dogs in snow or with water splash, blue depicts that there are guys doing sports with helmet.

Using PCA the image id was plotted based on the extracted image features in Fig4. It shows the dispersion of extracted features with the help of the image id based on the features. The image id that is featured nearby red color id might have the same features of having many people in the image. Similarly for other color ids too. Thus the extracted features are scattered in the space.

Figure 5 shows the image used for testing the model. The real caption for the image is three people prepare table full of food with police car in the background. The image features are extracted using different architectures such as ResNet-152, ResNet-50, VGG16, Inception V3. The predicted caption for each architecture is given in table 1.The model is evaluated using two different metrics - Bleu-n and Rouge-L. The table shows the value for each metric. From the table it is clear that ResNet152 performs better when compared to other architectures.

Figure 5: Image for testing

Table 1: Comparison with different architectures

	ResNet-152	**ResNet-50**	**VGG16**	**Inception V3**
Predicted Caption	two women at food counter in dim lights	two women and group of girls in dim in dim lights	two women and women in black shirt and women in the park at large paper	two people inside the side of food
Prediction time	286ms	260ms	377ms	299ms
Bleu-1	0.8003	0.756	0.6995	0.7708
Bleu-2	0.7452	0.6912	0.6239	0.7092
Bleu-3	0.6405	0.5714	0.4893	0.5941
Bleu-4	0.4102	0.3265	0.2394	0.3529
RougeL	0.9230	0.8976	0.9436	0.9412

Table 2: The real and predicted caption for the given images using ResNet 152

Images	Real Caption	Predicted Caption
	Old man with black coat and hat	An elderly gentleman in red and white polka dot scarf.
	Two people are dressed in animal costumes and are walking down street	Two people dressed in animal costumes entertain the crowd.
	brown dog is chasing tattered soccer ball across low cut field	tan dog chasing ball
	pack of black dogs running in grass.	many black dogs run in grassy area
	some toddlers are playing in foam playpen	two young children are playing with toys.
	black and white dog jumping in the snow.	dog jumps in the snow

Table 3: The real and predicted caption for the given images using Vgg-16

Images	Real Caption	Predicted Caption
	three people in jackets standing at the bottom of large rock formation	three men are standing on rock face
	dog bites an object offered by person	dog is carrying toy
	little boy in blue shorts holding window screen	boy in shorts moving out window screen that has fallen out.
	man is surfing large wave in the ocean	surfer rides in cloud on wave in the background
	young men from two team play game of indoor field hockey	group of players chase ball on field
	one person performing show in front of large audience sitting on the lawn	the children are going on an act in front of large crowd.

Table 4: The real and predicted caption for the given images using Inception-V3

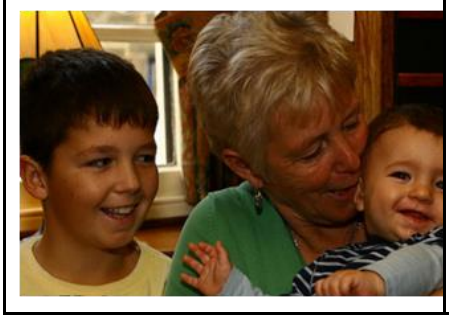
Images	Real Caption	Predicted Caption
	young boy swinging on tire swing	boy swinging on
	Kids in bathing suits with water splashing around them	children are playing in the sprinkler
	picture of campsite with big green tent	large tent.
	boy sits in his seat	toddler boy wearing airplane shirt is in high chair
	an older lady holds toddler while another boy watches	woman holds baby held by eating

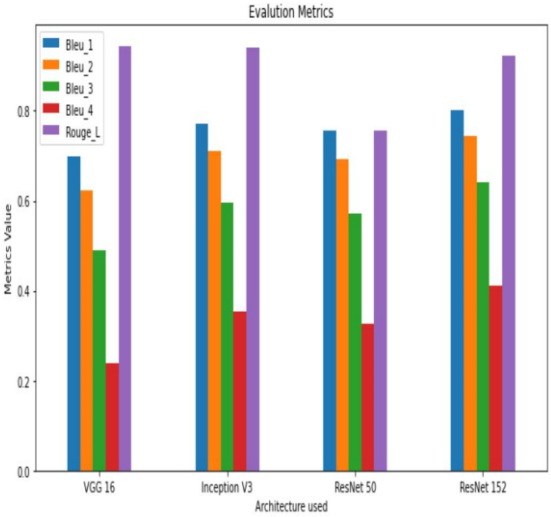
Table 5: The real and predicted caption for the given images using ResNet-50

Images	Real Caption	Predicted Caption
	couple walks past car in parking lot	man and woman walking next to red car with an unusual hood ornament
	two large black dogs are snarling at each other.	two black dogs are bearing
	toddler wearing colorful shorts is standing in the water	young boy in water
	woman plays with long red ribbon in an empty square	two people work in the background
	two workers at grocery store	two people in the camera

Table 2 shows the real and predicted caption for different images in which the features were extracted using ResNet-152.In the same way Table 3,4,5 shows the real and predicted caption for various images in which the features were extracted using ResNet-50, VGG16, Inception V3.

Different metrics were used for evaluating the correctness of the model after the training phase gets completed. In Fig 6 the model is evaluated against the testing dataset.

Figure 6: Evaluation metrics graph



From figure 6 it is clear to infer that our model performs better when the features are extracted using ResNet-152. ResNet-152 performs better when compared to other architectures. Hence the conclusion is that the model can generate more accurate captions if the features were extracted using ResNet-152. In Table 2 it is clearly seen that the model has predicted the caption of the image more accurately. Bleu-n is used to evaluate the closeness of machine translation to human reference translation. Rouge-L is used to evaluate automatic summarization and machine translation software in NLP.

V CONCLUSION

In this paper the image captioning models build using various CNN models for image feature extraction as encoder and Two-layer LSTM based on attention mechanisms as decoder was implemented. Among various CNN models such as VGG 16, Inception v3, ResNet 50, ResNet 152 it is concluded that ResNet-152 and two layer LSTM used to build our model has enhanced the performance of captioning. Our model is trained using Flickr 8k dataset. The experimental results demonstrate that the proposed method can produce quite promising results. Future work includes training the model with different dataset to get more accurate results.

VI REFERENCES

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In Proceedings of ECCV, pages 740–755, 2014. 1

[2] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” Trans. Assoc. for Comput.Linguistics, vol. 2, pp. 67–78, Dec. 2014

[3] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: CVPR, 2017, pp. 6298–6306.

[4] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top down attention for image captioning and visual question answering, in: CVPR, 2018, pp. 6077– 6086.

[5] T. Yao, Y. Pan, Y. Li, T. Mei, Exploring visual relationship for image captioning, in: ECCV, Vol. 11218, 2018, pp. 711–727.

[6] A. Farhadi et al., “Every picture tells a story: Generating sentences from images,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 15–29. [9] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi,

“Composing simple image descriptions using web scale n-grams,” in Proc. 15th Conf. Comput. Natural Lang. Learn., 2011, pp. 220–228.

[7] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” J. Artif. Intell. Res., vol. 47, pp. 853–899, Aug. 2013

[8] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3156–3164

[9] Y. Feng, L. Ma, W. Liu, and J. Luo, “Unsupervised image captioning,” in Proc.

IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 4125–4134.

[10] K. Papineni, S. Roukos, T. Ward, and W.-

J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. 40th Annu. Meeting Assoc.Comput. Linguistics, 2002, pp. 311–318.

[11] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in Proc. ACL Workshop Intrinsic Extrinsic Eval. Measures Mach. Transl. summarization, 2005, pp. 65–72.

[12] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[13] R. Vedantam, C. L. Zitnick, and D. Parikh, “CIDEr: consensus-based image description evaluation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 4566– 4575

[14] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “Explain images with multimodal recurrent neural networks,” 2014, arXiv:1410.1090. [Online]. Available: http://arxiv.org/abs/1410.1090

[15] A. Karpathy and L. Fei-Fei, “Deep visual semantic alignments for generating image descriptions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3128– 3137.

[16] B. Zhao, X. Li, and X. Lu, “CAM-RNN: co attention model based RNN for video captioning,” IEEE Trans. Image Process., vol. 28, no. 11, pp. 5552–5565, Nov. 2019.

[17] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 4651–4659

[18] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[19] Q. Wu, C. Shen, P. Wang, A. Dick, and A. van denHengel, “Image captioning and visual question answering based on attributes and external knowledge,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1367–1381, Jun. 2018

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2012, pp. 1097–1105.

[21] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, “Deep compositional captioning: Describing novel object categories without paired training data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1–10

[22] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, “Captioning images with diverse objects,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1170–1178.

[23] T.-H. Chen, Y.-H.Liao, C.-Y.Chuang, W.-T. Hsu, J. Fu, and M. Sun, “Show, adapt and tell: Adversarial training of cross-domain image captioner,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 521–530.

[24] Y. Zhang, W. Zhou, M. Wang, Q. Tian and H. Li, "Deep Relation Embedding for Cross-Modal Retrieval," in IEEE Transactions on Image Processing, vol. 30, pp. 617-627, 2021, doi: 10.1109/TIP.2020.3038354.

[25]D. Zeng and K. Oyama, "Learning Joint Embedding for Cross-Modal Retrieval," *2019 International Conference on Data Mining Workshops (ICDMW)*, 2019, pp. 1070-1071, doi: 10.1109/ICDMW.2019.00156.

[26] L. Zhen, P. Hu, X. Wang and D. Peng, "Deep Supervised Cross-Modal Retrieval," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 10386-10395, doi: 10.1109/CVPR.2019.01064.

[27] W. Zhao, X. Wu and J. Luo, "Cross-Domain Image Captioning via Cross-Modal Retrieval and Model Adaptation," in *IEEE Transactions on Image Processing*, vol. 30, pp. 1180-1192, 2021, doi: 10.1109/TIP.2020.3042086.

[28] G. Mao, Y. Yuan and L. Xiaoqiang, "Deep Cross-Modal Retrieval for Remote Sensing Image and Audio," *2018 10th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, 2018, pp. 1-7, doi: 10.1109/PRRS.2018.8486338.

[29] L. Jing, E. Vahdani, J. Tan and Y. Tian, "Cross Modal Center Loss for 3D Cross-Modal Retrieval," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3141-3150, doi: 10.1109/CVPR46437.2021.00316.

[30] Dr. Shalini Wadhwa, “Text Mining of Job Profiles: Finding Relevant Job Skills in different Industries”, Journal of Positive School

Psychology, Vol. 6, No. 2, 2022.
[31] N. Durga, Dr. D. Kerana Hanirex &Dr. A. Muthukumaravel, "A Systematic Review on Diabetic Retinopathy and Common Eye Diseases Detection through Deep Learning Techniques", Vol. 6, No. 4, 2022.
[32] L Manjusha, V. Suryanarayana, "Detect /Remove Duplicate Images from a Dataset for Deep Learning", Vol. 6, No. 2, 2022.
[33] G. Priyanka, T. Revathi, K. Muneeswaran, "Automatic Caption Generation from Images based on Facial Emotions", International Journal of Recent Technology and Engineering (IJRTE), Volume-8 Issue-4S5, December 2019.