IMAGE CAPTIONING USING DEEP LEARNING MODELS

Md.Thousif Ahmed thousifahmed19@ece.iiitp.ac.in
K.Karthik kommalapatikarthik19@cse.iiitp.ac.in
K.Rohit Reddy rohitreddy19@cse.iiitp.ac.in
B.Karthik Reddy karthikreddy19@ece.iiitp.ac.in

Abstract. In recent years, with the rapid development of artificial intelligence, image caption has gradually attracted the attention of many researchers in the eld of artificial intelligence and has become an interesting and arduous task. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. The application of image caption is extensive and significant, for example, the realization of human-computer interaction. This paper summarizes the related methods and focuses on the attention mechanism, which plays an important role in computer vision and is recently widely used in image caption generation tasks. Furthermore, the advantages and the shortcomings of these methods are discussed,providing the commonly used datasets and evaluation criteria in this field.To generate the caption for the image , we are going to use Flikr_8K dataset. It contains images obtained from the Flickr website. Some other big datasets like Flickr_30K and MSCOCO dataset are also available but it takes a lot of time to train the system, hence we are going to use this small Flikr_8K dataset .The dataset has 8000 images in JPEG format and each image has 5 captions and #(0 to 4) number is assigned for each caption.we have used three different feature extractors i.e Inception - V3, ResNet, and VGG16 in which Inception V3 has given better results compared to ResNet and VGG16. we have also used some different Language models(language model provides context to distinguish between words and phrases that sound similar) i.e; CNN(Convolutional neural network), RNN(Recurrent Neural Network), LSTM(Long short term memory).

Keywords: Image Captioning, Artificial Intelligence,Natural Language.

1 Introduction

Image captioning is a process which contains computer vision and natural language processing concepts to recognize the context of an image and generate the textual description of the image in common language such as English . We, human beings can look at an image and write the information about the image , in the same way by using Artificial Intelligence & Deep Learning models we can write a program which makes the system to detect the context of the image provided by the user and display the caption for the image as output . So the system predicts the information present in the image and provides a caption. This is called Image Captioning .Image Captioning is a multi model technique. Image captioning process can be divided into 2 modules , one is image based & the other is language based model .It involves Convolution (CNN) and Recurrent Neural Networks (RNN). Here, the image based model that is the CNN model extracts the features from the image and the language based model that is the RNN model translates the features into textform .

We can create a product for the blind which can guide them travelling on the roads without the support of anyone else. we will try this by first converting the scene into text then the text to voice. Both are now famous applications of Deep Learning. Automatic Captioning can help make Google Image Search nearly as good as Google Search, as then every image might be first converted into a caption so searches are often performed supporting the caption.

Deep Learning

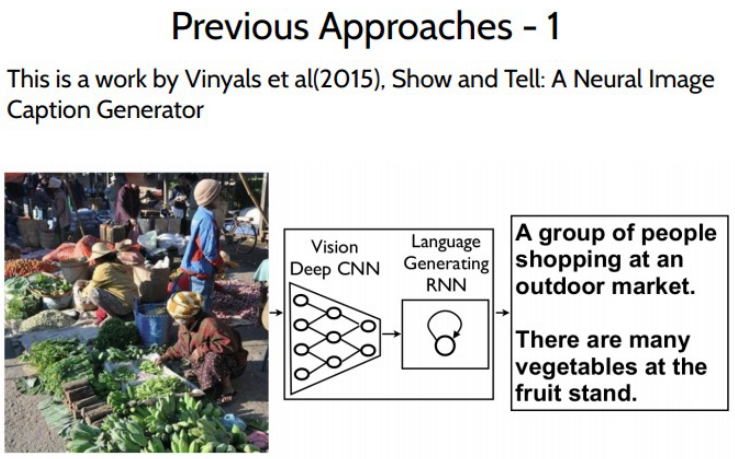
To know about Deep learning we need to first understand about artificial intelligence. Artificial intelligence (AI) is a very vast branch in Computer Science and it is used for constructing smart machines. The smart machines are capable of executing a task that requires human intelligence. Deep learning and machine learning are learning techniques in Artificial Intelligence. We have achieved many advancements in machine learning and deep learning. Now , as we are discussing Deep Learning. Deep learning is an Artificial Intelligence (AI) function that works almost similar to human brain in terms of performing tasks such as processing data and creating patterns in terms of decision making.Deep learning is a part of machine learning in artificial intelligence which consist networks such as deep neural networks which are capable of learning unsupervised data.Consider a human brain: it consists of millions of neurons. A deep neural network (DNN) has several layers in between input and output layers. Each layer consists of several nodes and are similar to neurons of the human

brain. To solve a task the system has to process layers of data between input and output. Stimulus must be passed at the input layer in order to execute a task.

Importance of Deep Learning

- Deep learning has a very significant importance in terms of handling big data.
- Deep learning includes a large number of features and its corresponding process.
- However , the access of the vast amount of data is required for deep learning algorithms to be effective.
- But deep learning models will be overfitted if data is too simple or incomplete.
- Deep learning is very useful for real life applications because of its algorithm's potential at learning.

2 Literature Survey

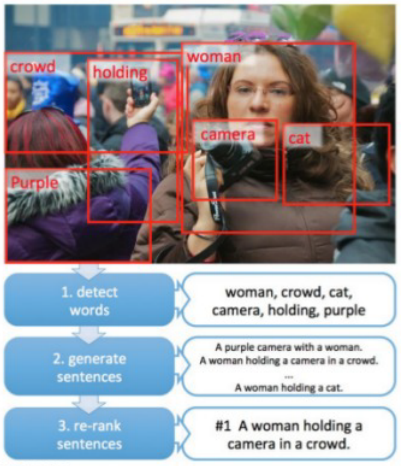


- Generate a fixed length vector representation of the image .say V_i, This representation is taken by extracting the features from a hidden layer of a pretrained CNN.
- Convert the caption to a vector representation,say Vc.
- Vi is fed to an long short term memory RNN(LSTM) as auxiliary input.
- Now Vc is fed ,element by element to the RNN to train it.

S.NO	Author Name	Work Description	Published Year
1.	VINYALS et al	A Neural image caption generator using vision deep CNN and language.	2015
2.	FANG et al	Analyzing visual concepts and generating captions, This process contains 3 stages, ● Word detection ● Sentences generation ● Ranking generated	2015

Previous Approaches - 2

This is a work by Fang et al(2015), From Captions to Visual Concepts and Back

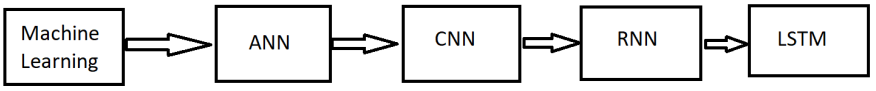


Word Detection Stage

They filter out the most common words in the training set and then use noisy-OR Multiple instance Learning -taking a sliding windows on the image as the bags - to determine the regia for each word for each image .The need for bounding boxes is eliminated by taking the representation generated by a CNN trained on ImageNet for the base of the MIL.

3 Methodology

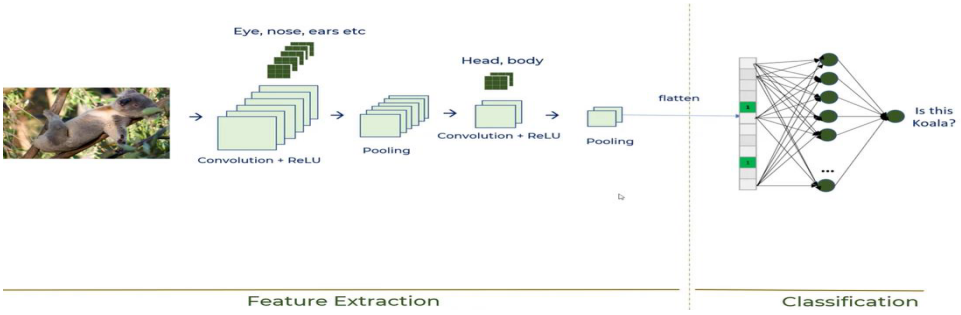
The below figure shows the models used in our project and their evolution.



CNN(Convolutional neural network)

As the name says this model uses Convolution in at least one of it's layers. Here in the first part, we use convolution operation for feature extraction. In this feature extraction we extract important points or features from the image known as Filters.

1. **Pooling**: Pooling layers reduce the dimensions of the data by combining the outputs of neuron clusters at one layer into a single neuron in the next layer. It also helps with the position invariant feature detection which means in this example no matter where eyes, ears, nose etc. are present, it will detect them. It also makes the model tolerant towards small distortions.
2. **ReLU**: It speeds the training of the model.
3. **Padding**: padding means giving additional pixels at the boundary of the data.
4. **Flattening**: In this process we convert the data into one-dimensional array to give it to the final classification model.



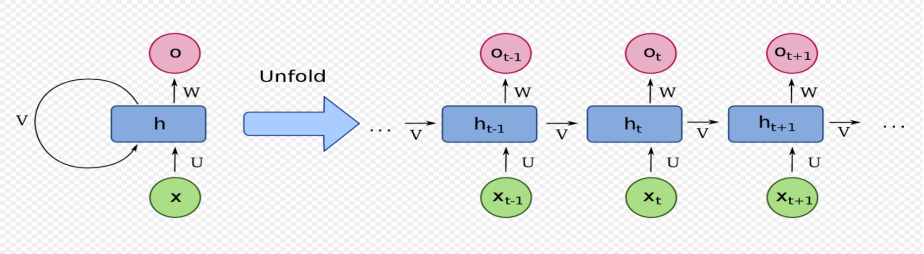
RNN(Recurrent Neural Network)

The main purpose to use RNN is that it is helpful in modelling sequential data. In a RNN model there are multiple ANN's in which we not only pass the data from the hidden layer to it's output but also pass the data to the hidden layer of the next ANN. This gives a benefit that there can be a clear outlook of the whole output produced.

There are different types of RNN's

1. **ONE to ONE:** Here there is a single input and we receive a single output.
2. **ONE to MANY:** Here there can be single input and multiple outputs for a single input.
3. **MANY to ONE:** Here there can be multiple inputs and single output. Eg: Telling the language of a given sentence.
4. **MANY to MANY:** Here there can be multiple inputs and multiple outputs.

Eg: Language Transition.



LSTM(Long short term memory)

RNN's cannot understand the context of the given input. They cannot recall the data which was said long before.

RNN's remember the data for a small duration of time but when a lot of words are fed, the information gets lost. Here is where Long Short Term Memory is used.

Here the every cell has three gates

1. **Forget gate:** Here the useless information or less important information

which is no longer required by the model is dropped.

2. **Input gate:** Here new information is added or updated.
3. **Output gate:** Here selecting useful information from the current cell state and showing it as an output is done.

This is used in speech recognition, Handwriting recognition, etc

Natural Language Processing

1.Research on Natural Language Processing has been started about 50 years ago and now came into work as technological evolution in terms of computers is huge.

2. Natural Language processing (NLP) is a branch of artificial intelligence. It deals with interaction between computers and human language.

3. It uses computer as a medium to read and understand the language irrespective of spoken or written.

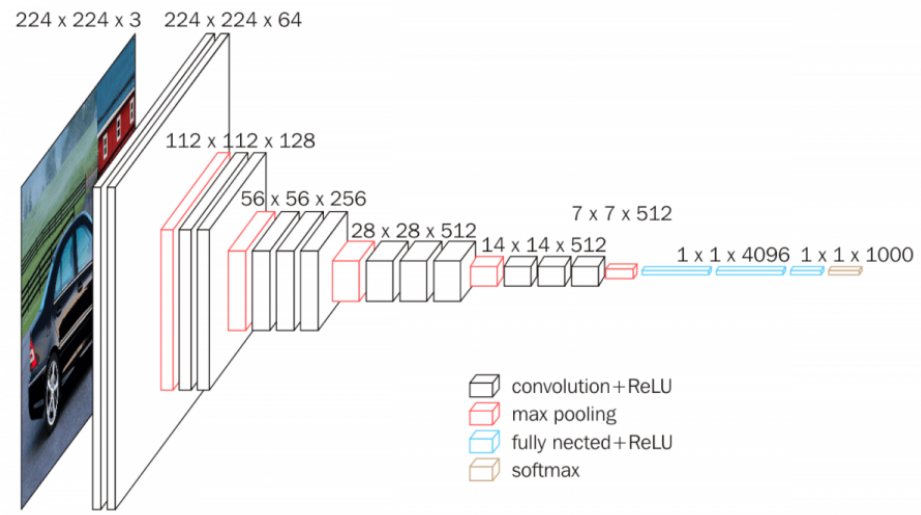
4. This also includes translation of one language into another language.

5. Natural Language Processing helps computer to communicate with humans in human Language.

6. It is because of NLP the computers are able to read text, hear speech,interpret the language.
7. It deals with various components of a given language. The five main components of Natural Language Processing are:
   - Semantic analysis.
   - Syntactic analysis.
   - Morphological and lexical analysis.
   - Pragmatic analysis.
   - Discourse integration.

VGG16

VGG16 (also known as OxfordNet) is a convolutional neural network architecture named after the Visual Geometry group from Oxford, who developed it. ... By only keeping the convolutional modules, our model is adapted to discretionary input sizes. The model loads a group of weights pre-trained on ImageNet.



The input to the convolutional layer is of fixed size 224 x 224 RGB image. The image is passed through a stack of convolutional layers, wherever the filters were used with a really tiny receptive field: 3×3 (which is that the smallest size to capture the notion of left/right, up/down, center). In one amongst the configurations, it additionally utilizes 1×1 convolution filters, which might be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to one pixel; the spatial padding of convolutional. layer input is such the spatial resolution is saved after convolution, i.e. the padding is 1-pixel for 3×3 convolutional. layers. spatial pooling is meted out by 5 max-pooling layers, that follow a number of the convolutional. layers (not all the convolutional. layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window, with stride two.Three Fully-Connected (FC) layers follow a stack of convolutional layers (which features a different depth in several architectures): the primary 2 have 4096 channels each, the 3rd layer performs a thousand-way ILSVRC classification and hence contains 1000 channels (one for every class). The ultimate layer is the soft-max layer. The configuration of the fully connected layers is the same for all networks.All hidden layers are provided with the rectification (ReLU) non-linearity. it's additionally noted that none of the networks (except for one) contain Local Response Normalisation (LRN), such normalisation doesn't increase the performance on the ILSVRC dataset, however results in augmented memory consumption and computation time.

INCEPTION-V3

Inception-v3 is a CNN architecture that belongs to the inception family that creates many enhancements including using Label Smoothing, Factorized 7 x 7 convolutions, and therefore the use of an auxiliary classifier to generate label data lower down the network (along with the utilization of batch normalization for layers within the side head).

Inception v3 Architecture.

The architecture of an Inception v3 network is gradually built, step-by-step, as explained;

1. Factorized Convolutions: This helps to decrease the computational efficiency because it decreases the amount of parameters involved in an exceedingly network. It also checks the network efficiency.

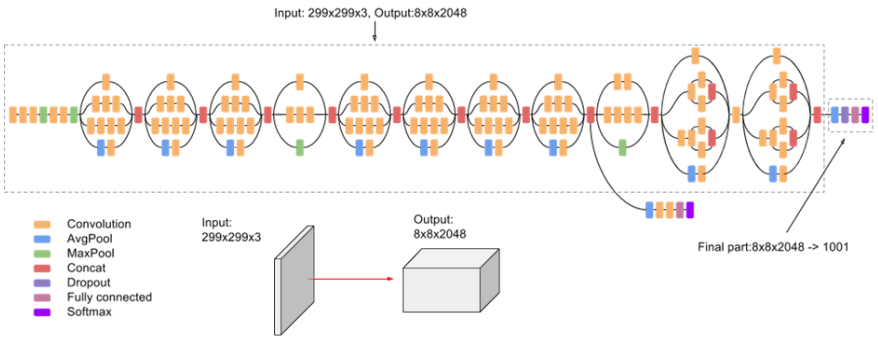
2. Smaller convolutions: Replacing bigger convolutions with smaller convolutions definitely ends up in faster training. Say a 5 × 5 filter has 25 parameters; two 3 × 3 filters replacing a 5 × 5 convolution has only 18 (3*3 + 3*3) parameters instead.

3. Asymmetric convolutions: A 3 × 3 convolution may well be replaced by a 1 × 3 convolution followed by a 3 × 1 convolution. If a 3 × 3 convolution is replaced by a 2 × 2 convolution, the quantity of parameters would be slightly more than the asymmetric convolution proposed.

4. Auxiliary classifier: An auxiliary classifier could be a small CNN inserted between layers during training, and therefore the loss incurred is added to the most network loss. In GoogLeNet auxiliary classifiers were used for a deeper network, but in Inception v3 an auxiliary classifier acts as a regularizer.

5. Grid size reduction: Grid size reduction is sometimes done by pooling operations. However, to combat the bottlenecks of computational cost, a much more efficient method is proposed.

All the above concepts are combined into the final architecture.



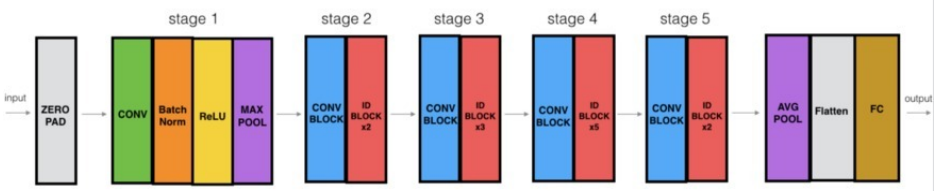
ResNet

ResNet also known as Residual Networks are classic neural networks which are used as backbone for several computer vision tasks. In the ImageNet challenge in 2015 this model was the winner. The main use with ResNet was it allowed users to train very deep neural networks with 150+layers successfully. Due to the matter of vanishing gradients, before ResNet training very deep neural networks was troublesome .

Increasing network depth doesn't work by simply stacking layers together. Due to the vanishing gradient problem as said above, Deep networks are hard to train. Vanishing gradient problem means the gradient is back-propagated to earlier layers, repeated multiplication may make the gradient extremely small. Due to the result of this problem, as the network goes deeper, its performance gets saturated or even starts degrading rapidly.

ResNet is able to train very deep neural networks because of skip connection. The concept of skip connection was first introduced by ResNet. The diagram below illustrates skip connection. The left figure shows the stacking of convolution layers together one after another while in the right figure, like before we stack convolution layers but now we additionally add the input we originally received to the output of the convolution block. This is called skip connection.

The figure here is of a ResNet-50 model.The ResNet-50 model has 5 stages each has a convolution and Identity block. Each convolution block consists of three convolution layers and each identity block also consists of three convolution layers. The ResNet-50 consists of over 23 million trainable parameters.

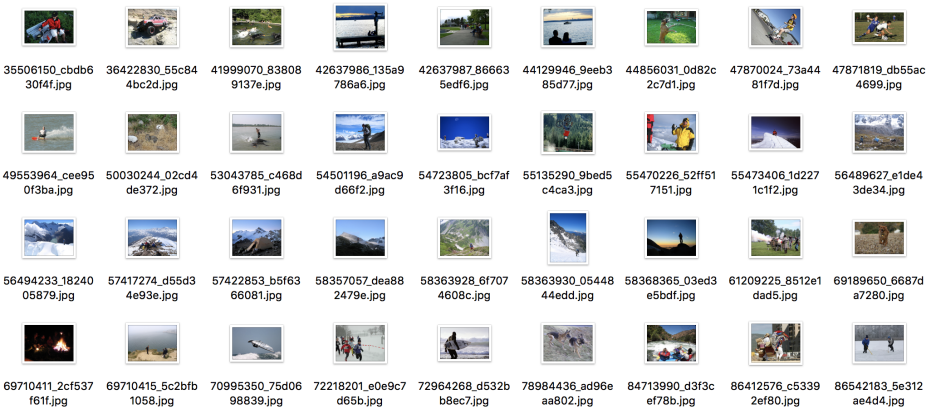


4 Experiments and Results

Dataset Description;

Source:- Flickr 8k dataset from Kaggle

To generate the caption for the image , we are going to use Flikr_8K dataset. It contains images obtained from the Flickr website. Some other big datasets like Flickr_30K and MSCOCO dataset are also available but it takes a lot of time to train the system, hence we are going to use this small Flikr_8K dataset .The advantage of big datasets is to build better models. The Flickr 8k.token file is present inside the Flickr_8K_text folder , this token is the main file of our dataset . This file contains the image name and respective captions for each image. The dataset has 8000 images in JPEG format and each image has 5 captions and #(0 to 4) number is assigned for each caption. The images present in the flicker_8k dataset have been selected from different Flickr groups, which doesn't contain any well known people or locations, but are manually picked to represent a variety of scenes and situations.



Model Performance metrics

BLEU

BLEU stands for Bilingual Evaluation Understudy.
It is an algorithm, which has been used for evaluating the quality of machine translated text. We can use BLEU to check the quality of our generated caption.

- BLEU is language independent
- Easy to understand
- It is easy to compute.
- It lies between [0,1]. Higher the score better the quality of caption
- It's fast and easy to calculate, especially compared to having human translators rate model output.
- It's ubiquitous. This makes it easy to compare your model to benchmarks on the same task.

Trying to compare BLEU scores across different corpora and languages is strongly discouraged. Even comparing BLEU scores for the same corpus but with different numbers of reference translations can be highly misleading. However, as a rough guideline, the following interpretation of BLEU scores (expressed as percentages rather than decimals) might be helpful.

BLEU SCORE	INTERPRETATION
<10	Almost useless
10 - 19	Hard to get the gist
20 - 29	The gist is clear, but has significant grammatical errors
30 - 40	Understandable to good translations
40 - 50	High quality translations
50 - 60	Very high quality, adequate, and fluent translations
>60	Quality often better than human

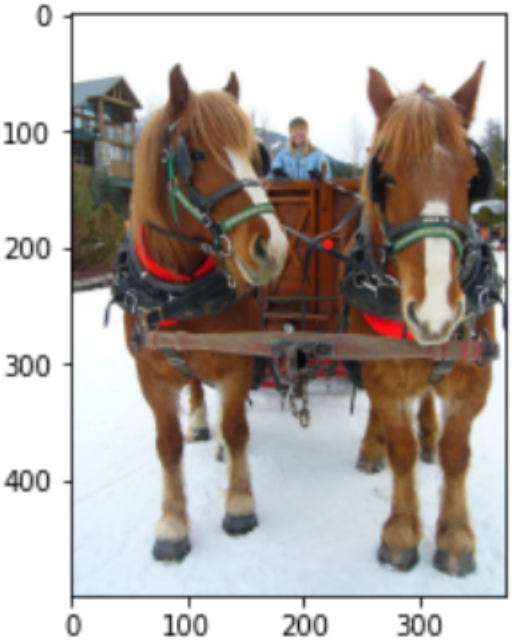
METEOR

Metric for Evaluation of Translation with Explicit ORdering is an automatic metric that evaluates translation hypotheses. It is based on a generalized concept of unigram matching between the machine-produced translation and human-produced reference translations

The Meteor automatic evaluation metric scores machine translation hypotheses by aligning them to one or more reference translations. Alignments are based on exact, stem, synonym, and paraphrase matches between words and phrases. Segment and system level metric scores are calculated based on the alignments between hypothesis-reference pairs.

MODELS	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Inception-V3	50.62	29.98	23.23	17.41	16.86
ResNet	48.26	27.39	21.65	15.32	15.14
VGG16	45.78	26.16	21.03	14.94	14.13

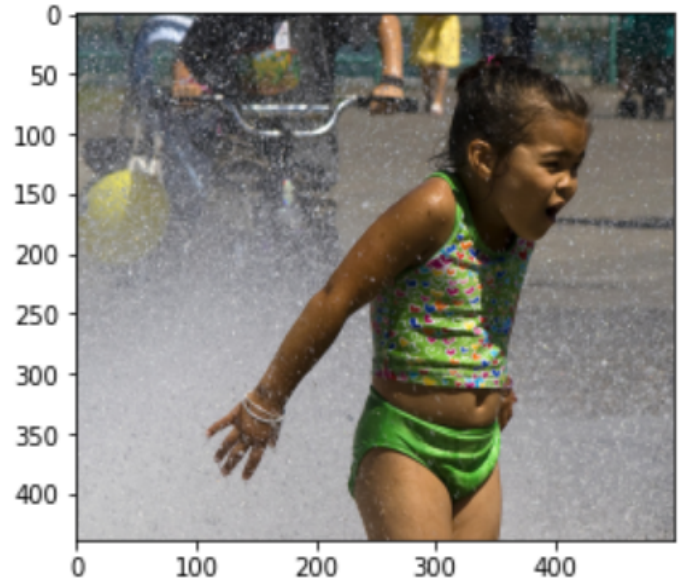
Two horses are pulling a woman in a cart .



A dog is jumping in the air to catch an item .

A man is riding a bicycle on the beach .

A child in a swimsuit walks among large waves .



DISCUSSION

The project proposed by us can predict captions of images up to an extent. As the data set we are using is small it is possible that our model will not be able to predict captions of some images and it may predict wrong captions for some images. In the captions of some images it may be possible that there are grammatical errors as the words are also limited. In some images the model may not be able to differentiate between the colors. But for some images shown above it can predict captions accurately with no mistakes

5 Conclusion and Future Scope

We have introduced a model that generates natural language descriptions of image regions based on obtained labels in form of a dataset of images and their respective sentence description, and with few assumptions. experimental results show that this task still has better performance systems and improvement. It mainly faces the following three challenges: rst, how to generate complete natural language sentences like a human being; second, how to make the generated sentence grammatically correct; and third, how to make the caption semantics as clear as possible and consistent with the given image content Some key points to note are that our model depends on the data, so it cannot predict the words that are out of its vocabulary. We used a small dataset consisting of 8000 images. For production-level models, we need to train on datasets larger than 100,000 images which can produce better accuracy models. A lot of modifications can be made to improve this solution like using a larger dataset, changing the model architecture, e.g. include an attention module, doing more hyper parameter tuning (learning rate, batch size, number of layers, number of units, dropout rate, batch.

Acknowledgments

This project would not have been possible without the help and cooperation of many.I would like to thank the people who helped me directly and indirectly in the completion of this project work.First and foremost, I would like to express my gratitude to our beloved director, Dr.Anupam Shukla, for providing his kind support in various aspects. I would like to express my gratitude to my project guide Dr.Sanjeev Sharma, Assistant Professor,Department of CSE, for providing excellent guidance, encouragement,inspiration, constant and timely support throughout this B.Tech project.I would like to express my gratitude to the head of department Dr.Tanmoy Hazra, Assistant Professor, Department of CSE, for providing his kind support in various aspects. I would also like to thank all the faculty members in the Dept. of CSE and my classmates for their steadfast and strong support and engagement with this project.

References

1. Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares, Image Captioning: Transforming Objects into Words Yahoo Research San Francisco, CA, 94103 .
2. Yin Cui1,2 Guandao Yang1 Andreas Veit1,2 Xun Huang1,2 Serge Belongie1,2,Learning to Evaluate Image Captioning ,Department of Computer Science, Cornell University 2Cornell Tech,2018.
3. R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In CVPR, 2015
4. A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In CVPR, 2015.
5. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.
6. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In OSDI, 2016
7. M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. J. Artif. Int. Res., 47(1):853–899, May 2013.
8. Vedantam, R.; Bengio, S.; Murphy, K.; Parikh, D.; Chechik, G. Context-aware captions from context-agnostic supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu,2017.
9. S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In ICCV, 2017
10. Wu, Y.; Zhu, L.; Jiang, L.; Yang, Y. Decoupled novel object captioner. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Korea, 22–26 October 2018; ACM: New York, NY, USA, 2018.
11. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, http://arxiv.org/abs/1409.0473 Computer Science.
12. L. Minh-Thang, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015.
13. T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion using RNN pre-trained by recurrent temporal restricted Boltzmann machines," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 3, pp. 580–587, 2015.
14. C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," in *Proceedings of the 9th ISCA Speech Synthesis Workshop*, pp. 146–152, Sunnyvale, CA, USA, September, 2016.
15. G. Kulkarni, V. Premraj, V. Ordonez et al., "Babytalk: understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.
16. Karpathy A., and Fei-Fei L. (2015) Deep visual-semantic alignments for generating image descriptions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), USA, 3128-3137.

17. Villegas M., M¨uller H., Gilbert A., Piras L., Wang J., Mikolajczyk K., de Herrera A.G.S., Bromuri S., Amin M.A., Mohammed M.K., Acar B., Uskudarli S., Marvasti N.B., Aldana J.F., del Mar Rold´an Garc´ıa M. (2015).
18. Farhadi A., Hejrati M., Sadeghi M.A., Young P., Rashtchian C., Hockenmaier J., and Forsyth D. (2010) Every picture tells a story: Generating sentences from images.
19. Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description{C}// Computer Vision and Pattern Recognition. IEEE, 2015:677.
20. Lecun Y, Boser B, Denker JS et al (1989) Back propagation applied to handwritten zip code recognition.Neural Comput 1(4):541–551.
21. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. International conference on neural information processing systems. Curran Associates Inc: 1097–1105
22. Bin J, Gardiner B, Liu Z et al (2019) Attention-based multi-modal fusion for improved real estate appraisal:a case study in Los Angeles. Multimed Tools Appl: 1–22. doi:https://doi.org/10.1007/s11042-019-07895-5.
23. Zhou Y, Zhenzhen H, Ye Z, Liu X, Hong R (2018) Enhanced text-guided attention model for image captioning. 2018 IEEE fourth international conference on multimedia big data (BigMM).
24. Sadeghi MA, Sadeghi MA, Sadeghi MA, et al (2010) Every picture tells a story: generating sentences from images. European conference on computer vision. Springer-Verlag.
25. Liu, C.; Sun, F.; Wang, C. MAT: A multimodal translator for image captioning. In Proceedings of the Artificial Neural Networks and Machine Learning, Pt II, Alghero, Italy, 11–14 September 2017; Lintas, A.Rovetta, S., Verschure, P., Villa, A.E.P., Eds.; Springer International Publishing Ag: Cham, Switzerland, 2017;
26. Fu, K.; Jin, J.; Cui, R.; Sha, F.; Zhang, C. Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2321–2334.
27. Weaver, Lex and Tao, Nigel. The optimal reward baseline for gradient-based reinforcement learning. In Proc. UAI'2001, pp.538–545, 2001.
28. Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard. Unifying visual-semantic embeddings with multimodal neural language models. arXiv:1411.2539, November 2014
29. Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever,Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. JMLR, 15, 2014
30. Kilickaya, M.; Akkus, B.K.; Cakici, R.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N. Data-driven image captioning via salient region discovery. IET Comput. Vis. 2017, 11, 398–406.
31. Shetty, R.; Rohrbach, M.; Anne Hendricks, L.; Fritz, M.; Schiele, B. Speaking the same language: Matching machine to human captions by adversarial

training. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4135–4144.
32. Bai, S.; An, S.Asurvey on automatic image caption generation. Neurocomputing 2018, 311, 291–304.
33. Yan, S.; Xie, Y.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image captioning based on a hierarchical attention mechanism and policy gradient optimization. J. Latex Cl. Files 2015.