ISSN (Online): 2455-3662
EPRA International Journal of Multidisciplinary Research (IJMR) - Peer Reviewed Journal
Volume: 8| Issue: 8| August 2022|| Journal DOI: 10.36713/epra2013 || SJIF Impact Factor 2022: 8.205 || ISI Value: 1.188

IMAGE CAPTION GENERATOR

Akash Shetty[1], Abhiram Srivathsa K H [2], O S Sumukh [3], Kavitha S N [4]
Department of Information Science and Engineering, R V College of Engineering
Mysore Road, Bengaluru - 560059

ABSTRACT

Generating accurate captions for an image has remained one of the major challenges in Artificial Intelligence with plenty of applications ranging from robotic vision to helping the visually impaired.Long term applications also involve providing accurate captions for videos in scenarios such as security system.The aim is to build an optimal system which can generate semantically and grammatically accurate captions for an image and also to suggest captions which can be used on social media platforms.In this system, images are preprocessed and captions are generated. Then the image features are extracted using Resnet50.Then the captions are generated word by word using LSTM. The application hopes to be useful for visually impaired people and to be useful for generating the social media captions which can be used on various social media platforms.

KEYWORDS-***CNN, RNN, LSTM, Resnet50***

I. INTRODUCTION

Making a computer system detect objects and describe them using natural language processing (NLP) in an age-old problem of Artificial Intelligence. This was considered an impossible task by computer vision researchers till now. With the growing advancements in Deep learning techniques, availability of vast datasets, and computational power, models are often built which will generate captions for an image. Image caption generation is a task that involves image processing and natural language processing concepts to recognize the context of an image and describe them in a natural language like English or any other language.Making a computer system detect objects and describe them using natural language processing (NLP) in an age-old problem of Artificial Intelligence. This was considered an impossible task by computer vision researchers till now. With the growing advancements in Deep learning techniques, availability of vast datasets, and computational power, models are often built which will generate captions for an image. Image caption generation is a task that involves image processing and natural language processing concepts to recognize the context of an image and describe them in a natural language like English or any other language.

Our model is based on a deep learning neural network that consists of a vision CNN followed by a language generating RNN. It generates complete sentences as output captions or descriptive sentences.In recent years a lot of attention has been drawn towards the task of automatically generating captions for images. However, while new datasets often spur considerable innovation, benchmark datasets also require fast, accurate, and competitive evaluation metrics to encourage rapid progress. Being able to automatically describe the content of a picture using properly formed English sentences may be a very challenging task, but it could have an excellent impact, as an example by helping visually impaired people better understand the content of images online. This task is significantly harder, for instance, than the well-studied image classification or visual perception tasks, which are a main focus within the computer vision community. Deep learning methods have demonstrated advanced results on caption generation problems. What is most impressive about these methods is that one end-to-end model is often defined to predict a caption, given a photograph, rather than requiring sophisticated data preparation or a pipeline of specifically designed models. Deep learning has attracted a lot of attention because it's particularly good at a kind of learning that has the potential to be very useful for real-world applications. The ability to find out from unlabeled or unstructured data is a huge benefit for those curious about real-world applications.

II. METHODOLOGY

The modules are divided into following sections:

Preprocessing of Image

For image detection, we are using a pre-trained model ResNet50 which is trained on image net dataset. For feature extraction, the image input is 244*244*3 size and will give an output vector of length 2048.The features of the image are extracted just before the last layer of classification as this is the model used to predict a classification for a photo.

Creating the vocabulary for the image

First cleaning of the text is done by splitting it into words and handling punctuation and case sensitivity issues. As computers do not understand English words, they are represented with numbers and they map each word of the vocabulary with a unique index value, and the encoding of each word into a fixed sized vector is done and it represents each word as a number.In order to achieve the mentioned objectives we do-Loading the data,creating a descriptions dictionary that maps images,removing punctuations, converting all text to lowercase and removing words that contain numbers,separating all the unique words and creating vocabulary from all the descriptions,creating a file to store all the captions.

Training the model

In our dataset we have a file Flickr_8k.trainImages.txt file that contains a list of 6000 image names that will be used for training purposes.First creation of a dictionary that contains captions for each photo from the list of photos is done.Next tokenization of vocabulary is done using keras.Keras library provides a function that will be used to create tokens from vocabulary and then to a tokenizer.pkl pickle file.To make this a supervised learning task, we have to provide input and output to the model for training. The model was trained on 6000 images and each image will contain a 2048 length feature vector and the corresponding caption for the image is also represented as numbers.This large volume of data generated for 6000 images is not possible to hold in memory, so generator method that will yield batches was used.To train the model,6000 training images will be used by generating the input and output sequences in batches from the above data generation module and fitting them to the model. Training the model was done with 200 epochs.

Testing on individual images

Testing the model on images is done on random images.The predictions contain the max length of index values so we will use the same tokenizer.pkl to get the words from their index values.

III. MODELING AND ANALYSIS

Although it is sometimes not clear whether a description should be deemed successful or not given an image, priorart has proposed several evaluation metrics. The most reliable (but time consuming) is to ask for raters to give a subjective score on the usefulness of each description given the image.

The most commonly used metric so far for image caption generators has been the BLEU score, which is a form of precision of word n-grams between generated and reference sentences.

BLEU (BiLingual Evaluation Understudy) is a metric for automatically evaluating machine-translated text. The BLEU score is a number between zero and one that measures the similarity of the machine-translated text to a set of high quality reference translations. A value of 0 means that the machine-translated output has no overlap with the reference translation (low quality) while a value of 1 means there is perfect overlap with the reference translations (high quality). It has been shown that BLEU scores correlate well with human judgment of translation quality. Note that even human translators do not achieve a perfect score of 1.0.

Testing should also be done on the GUI

Table 1: Test Case -1: Generate caption by clicking on caption image button

Test scenario	Generating caption by clicking on caption image button			
Test case description	The caption image button is clicked after uploading the image		**Test priority**	High
Prerequisite	Image should be uploaded		**Postrequisite**	The caption is generated
Action	**Input**	**Expected output**	**Actual output**	**Test result**
Click on Caption Image button	A jpeg image	Caption relevant to the image should be generated	Caption was generated on clicking the caption image button	Pass

Table 2: Test Case -2: Recommend social media captions by clicking on Recommend Captions button

Test scenario	Recommend social media captions by clicking on Recommend Captions button			
Test case description	The recommend caption button should generate relevant social media captions based on the label generated		**Test priority**	High
Prerequisite	The label should be generated beforehand		**Postrequisite**	A list of social media captions should be generated
Action	**Input**	**Expected output**	**Actual output**	**Test result**
Click on Recommend Caption button	A label defining the image	A list of social media captions based on the label	5 captions based on the label was generated	Pass

IV. RESULTS AND DISCUSSION

For caption generators , the quantitative evaluation metric is BLEU (Bilingual Evaluation Understudy)scores:

The BLEU score is a number between zero and one that measures the similarity of the machine-translated text to a set of high quality reference translations. A value of 0 means that the machine-translated output has no overlap with the reference translation (low quality) while a value of 1 means there is perfect overlap with the reference translations (high quality)

Cumulative BLEU scores capture the weighting of the different n-grams during the calculation of BLEU scores .

Individual N-Gram Scores : An individual N-gram score is the evaluation of just matching grams of a specific order, such as single words (1-gram) or word pairs (2-gram or bigram). The weights are specified as a tuple where each index refers to the gram order.

Cumulative N-Gram Scores : Cumulative scores refer to the calculation of individual n-gram scores at all orders from 1 to n and weighting them by calculating the weighted geometric mean.

The cumulative and individual 1-gram BLEU use the same weights, e.g. (1, 0, 0, 0). The 2-gram weights assign a 50% to each of 1-gram and 2-gram and the 3-gram weights are 33% for each of the 1, 2 and 3-gram scores.

It is common to report the cumulative BLEU-1 to BLEU-3 scores when describing the skill of a text generation

Fig 4.3 shows the cumulative scores obtained for the model but there are few weaknesses of BLEU scores which are that the BLEU metric performs badly when used to evaluate individual sentences. For example, both example sentences get very low BLEU scores even though they capture most of the meaning. Because n-gram statistics for individual sentences are less meaningful, BLEU is by design a corpus-based metric; that is, statistics are accumulated over an entire corpus when computing the score. Note that the BLEU metric defined above cannot be factorized for individual sentences.

Hence we can define a human evaluation system with four categories-Describes without errors,describes with minor errors,somewhat related to image,unrelated to the image

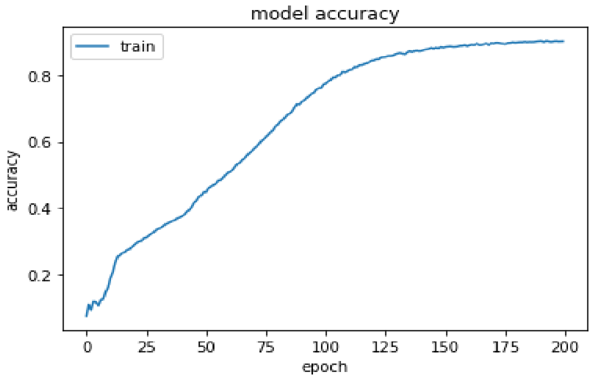


Fig 1: A graph of accuracy vs epoch showing training accuracy.

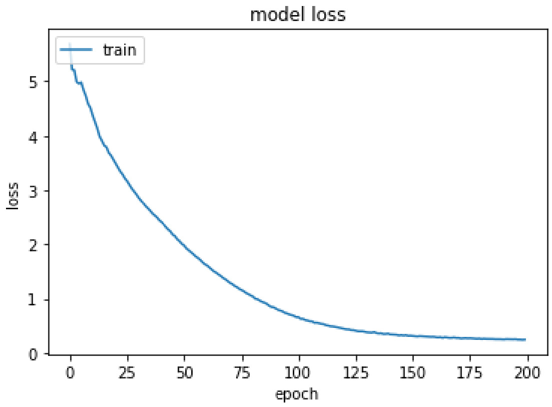
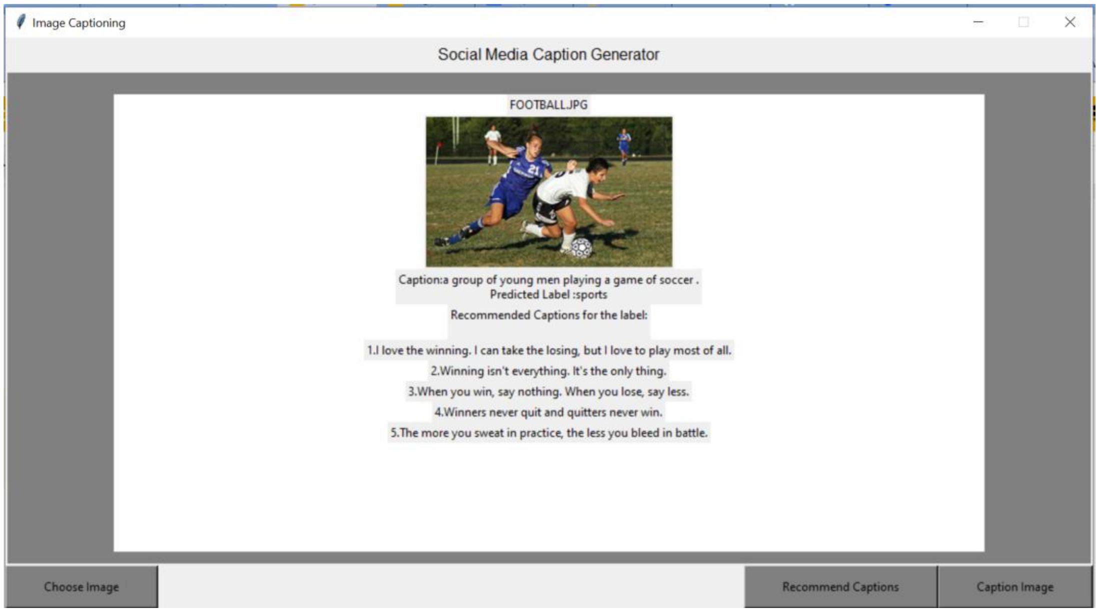


Fig 4.2: A graph of loss vs epoch showing training loss.



Caption:a group of young men playing a game of soccer .
Predicted Label :sports

Recommended Captions for the label:

1.I love the winning. I can take the losing, but I love to play most of all.
2.Winning isn't everything. It's the only thing.
3.When you win, say nothing. When you lose, say less.
4.Winners never quit and quitters never win.
5.The more you sweat in practice, the less you bleed in battle.

Fig 3:Captions based on the label is predicted

Table 3: BLEU score for different metrics

Cumulative 1-gram	0.700000
Cumulative 2-gram	0.557773
Cumulative 3-gram	0.430509

V. CONCLUSION

A CNN model combined with the LSTM model has been presented which gives a reasonable description of an image in English language. We also extract a label based on the image and generate captions which can be used on social media. The model has been trained on Flickr 8k dataset and its BLEU scores have been found to match the state of art results. As the size of the dataset and epochs increases, the performance of the model will improve. Based on the captions labels were generated using neural networks and these labels were then used to look up catchy social media captions . A GUI was then developed which can make the model easy to use.

VI. REFERENCES

1. *Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, & Vrinda Mathur. (2021). Image Caption Generator. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 10(3), 87–92.*
2. *Sharma, Grishma and Kalena, Priyanka and Malde, Nishi and Nair, Aromal and Parkar, Saurabh, Visual Image Caption Generator Using Deep Learning (April 8, 2019). 2nd International Conference on Advances in Science & Technology (ICAST) 2019 on 8th, 9th April 2019 by K J Somaiya Institute of Engineering & Information Technology, Mumbai, India.*
3. *Rashid khan , M Shujah Islama , Khadija Kanwala , Mansoor Iqbal, Md. Imran Hossaina & Zhongfu Ye.National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230026, Anhui, China*
4. *Tanti, Marc & Gatt, Albert & Camilleri, Kenneth. (2017). What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?.Proceedings of the 10th International Conference on Natural Language Generation*
5. *P. Shah, V. Bakrola and S. Pati, "Image captioning using deep neural architectures," 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), 2017, pp. 1-4.*
6. *Oriol Vinyals Google,Alexander Toshev Google,Samy Bengio Google,Dumitru Erhan Google ,2015 IEEE*
7. *Rashid khana , M Shujah Islama , Khadija Kanwala , Mansoor Iqbal, Md. Imran Hossaina & Zhongfu Yea National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230026, Anhui, China*
8. *Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics*
9. *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.*
10. *Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In CVPR, 2016*