International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-1, October, 2019

Image Captioning for Affine Transformed Images using Image Hashing

M. Nivedita, Asnath Victy Phamila Y

Abstract*: *Image captioning is the process of generating a meaningful textual description to the image. The perfect caption for the image not only consists of objects and their attributes, it also concentrates on the actions involved by the objects. There are two main tasks in Image captioning. The first and foremost task is correctly identifying objects present in the given image. Once all the objects are identified along with their attributes, the dense model is trained in order to identify the correct verbs or the actions in which these identified objects are involved. The second part in Image captioning is generating the syntactically correct natural language sentence which connects all the identified objects along with their attributes and actions. In this paper we have generated the captioning for affine transformed images using Flickr 8K dataset.

Keywords*: *deep learning, Convolution neural networks, Recurrent Neural Networks, dense model, language model.

I. INTRODUCTION

"A picture is worth a thousand words" this is a famous saying everybody knows. We can caption an image in multiple ways. But finalizing the most appropriate caption for an image is most challenging task. Many surveys had been conducted in this Image captioning topic for identifying the best caption generation model. In order to do this survey, the surveyors have to compare the models on different standard datasets. But previously when research starts on this topic, the researchers and surveyors won't have many standard dataset. Gradually research on generating captions increases along with the increase in the availability of standard datasets. Currently, a hand full of diverse datasets is available in order to generate new models or to do comparison among the existing models. Based on the survey papers, Image captioning model can be mainly classified into two categories. One of the most widely using categories is supervised learning. In this category, mainly the training images will come along with the label and these labels will help in generating captions for the test images by making use of input and output pairs. But the disadvantage in using this supervised learning is the model might not identify the new objects which are not present in the training dataset. The second category is the one which overcomes the disadvantage mentioned in the supervised learning. It is the unsupervised learning. This learns from the test data which is not labeled.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

M.Nivedita*, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India. Email: nivedita.m@vit.ac.in

Asnath Victy Phamila Y, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India.
Email: asnathvicty.phamila@vit.ac.in

Even though this unsupervised learning is more appropriate than supervised learning for image captioning, the models are still under development. This image captioning helps in various fields like social media for tagging the locations like beach, café etc and for identifying the famous personalities and many other famous idols, places. Microsoft developed an Image captioning model for identifying all the above mentioned objects along with date and time zone. This image captioning is not only helpful for identifying objects. One more interesting advantage of using this image captioning is it helps in sentence base image search. These all are the advantages and areas where image captioning is used in previous researches. Before proceeding for any enhancement in development and research in Image captioning, one must identify the group of people for whom this image captioning is useful and beneficial. One among those identified groups is visually challenged people. Yes, they can use this image captioning for searching their personal objects and correctly identify where they kept their objects. In one word they can use this as their personal assistant. Yet another application of image captioning is it can be used in self-driving cars and video surveillance systems. In self-driving cars the video can be taken by the camera attached to a car and the description about the image will be used for the movement of the vehicle appropriately. For example to describe about the pedestrian crossing, traffic signal detection etc… It can be used in surveillance where any abnormal incidents can be captured and caption can be generated for that.

II. RELATED WORK

The main work in image captioning lies in connecting both the vision helps in identifying objects and language model helps in generating a caption which connects all the objects and framing a caption correctly without grammatical mistakes. Many models are developed in order to generate an effective caption. But all the models contain only one step encoder and decoder where as Zhihao Zhu[1] implemented a model consists of two encoders and decoders. By the end of first stage encoder in this preview and tell model, a sample caption is generated with the major identified objects. Later finalized caption is generated with the suitable sentence framing in second stage. The major difference in this model compared to other models is this preview and tell model frames a caption not only based on previously generated words. It also refers to future words. Future words can be identified with the help of caption generated at the end of first decoder.

Even though the model for captioning is very effective, it is difficult to describe entire image content in one single sentence. We can only mention the highlighted objects and their attributes but not in fine deeper way. Jonathan Krause [2] presented a model which overcomes this problem. The hierarchical recurrent networks help in identifying the image keenly. This generates multiple captions which cover almost all the pixels in an image. After that dense language model helps in connecting all those captions into a meaningful paragraph. The result of this model has given a very high accuracy compared to all other models. But the only disadvantage in this model is lengthiness. Now a days people show much interest on short and crisp captions rather than explored paragraph. Here come the neural networks into picture in order to overcome this disadvantage. These neural networks can broadly be classified into two types as single point and multi point neural networks.

When coming to the concept of image captioning multi-point neural networks is better than single point since the final caption should contain the entire overview of an image rather than concentrating on single object. Zhongliang Yang [3] uses the language convertor as a base concept. Encoder and decoder are needed to convert a sentence from one language to another. Same way here in image captioning, source is either a image or sentence and it has to be encoded. Based on that caption will be generated using a language decoder. Mostly Convolutional neural network is used for encoding and a recurrent neural network is used for decoding. The location where that particular object is located is also mandatory while generating caption. As a result CNN-RNN model is the most optimized model when compared to all other standard model in the field on captioning images.

Almost all models are either based on supervised or unsupervised learning and multi-point neural networks. Jiuxiang Gu [4] developed a new diverse model called stack captioning based on reinforced learning for decoder. This decodes a caption from dense to crisp. In reinforced learning the outcome from one step decoder will be sent to the upper level decoders and will be processed further. But there a chance of noise occurrence in between, so spatial map is used to highlight the important objects and regions in image. Vikram Mullachery [5] had made use of checkpoints in the image captioning model for developing captions for videos also. Few models generate captions but those captions are entirely deviated from the topic and pay more attention on little attributes. Zhihao Zhu [6] tried to overcome this issue by developed a model which first decides the topic to which this image belongs to. Later caption is generated based on that identified topic. This feedback type model is popularly known as topic guided captioning. Ankit Gupta [7] shows the difference in caption generated when RNN is used along with LSTN. The models in which RNN is used to decode the caption from the list of objects identified, there is a need of LSTM which stores content for longer period of time. The well-known work in the field of image captioning is done by Google [8] and published through a paper work called “show and tell”. MD Zahir Hossain [9] has done a comprehensive survey of deep learning for image captioning. The survey describes all the existing methods and classifications in models. The image captioning is mainly classified into three categories. One among those is fixed template based captioning, second is mainly concentrates on the language model used and the third category is retrieving one most suitable caption among set of captions. As discussed earlier, captions may be of single appropriate sentence or a paragraph which covers minute attributes also. A detailed survey was done on all the existing evaluation metrics and standard datasets available.

III. PROPOSED WORK

A commonly-used choice of visual encoder is traditional Convolutional Neural Network (CNN),which provides limited support for exploring spatial invariant property in input images such as

- Scaling
- Translation
- Rotation
- Shearing etc.

one sample image and different affine transformations of the image is shown in Fig.1.

Fig.1 a.Original Image

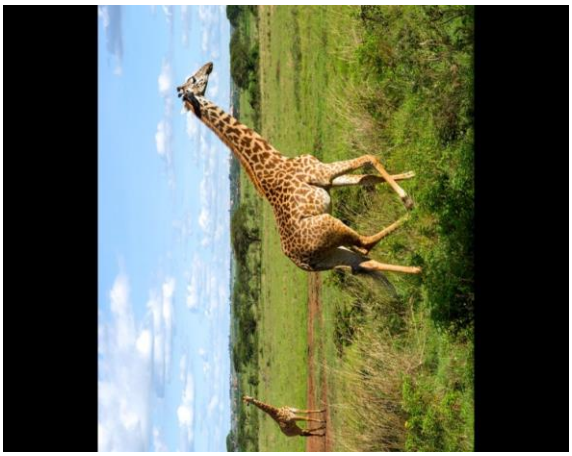

Fig.1 b.90· Rotated Image

Retrieval Number: A2022109119/2019©BEIESP
DOI: 10.35940/ijeat.A2022.109119
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Fig.1 c. Scaled Image

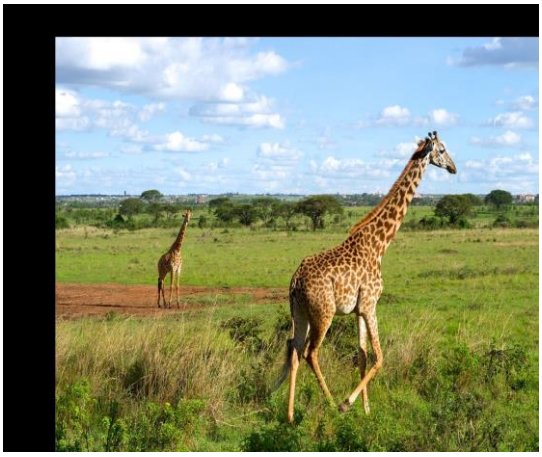

Fig.1 d. Translated Image

Fig.1 e. Sheared Image

By using hashing technique the transformed images has to be matched with the original image and the extracted features has to be given as input to the image captioning model to predict the description about the image. The challenge here is that the features extracted from all the transformed images should be same as that of the original image so that the caption generated will be similar.

We have used dhash (Difference hashing) algorithm along with Hamming distance calculation to assert that the original image and the affine transformed images of the original image are same.

dhash algorithm:

1. Convert the RGB image to a grayscale image which helps us to discard colour information and hash the image faster as there is only one channel for us to examine.

2. Resize the image to 9x8 pixels to ensure that the resulting image hash will match similar photos easily.

3. Compute the difference, this is done by computing the relative gradients between adjacent pixels. So when this is computed for a 9x8 image, we end up with 8 rows of 8 differences which is reciprocated as a 64-bit hash.

4. Build the hash value. If left pixel is brighter than the right pixel, the bit is set as 1, else set the bit as 0

After calculating the hash values for both the original image and the affine transformed images, the hamming distance between them are calculated.If the hamming distance is between 1-3, the images are considered to be same else, the images are different.

IV. DATASETS AND EVALUATION METRIC

New models are evolving day by day for generating the most effective caption for an image. Comparison is needed among the models in order to know which model is effective and robust. Each standard dataset is different and consists of thousands of images. Below listed are the major and vastly using datasets for the image captioning.

MSCOCO: This is the most widely used dataset developed by Microsoft. In any image captioning dataset there will be separate images for training, developing and for testing. Even in this COCO dataset 330k images in which more than 200k are labeled, 1.5 million object instances, 80 object categories and 91 stuff categories. Each image is described in five different captions.

Flickr8k: This is the standard benchmark dataset with 8000 images in image captioning field. This dataset mainly contains images of animals and human. 6000 training images, 1000 developing images and 1000 testing images. Like in COCO dataset each image is described with five different captions. This mainly contains animals and human with generic descriptive captions. Since 8000 images is less number compared to all other datasets, most of the models use this dataset.

Flickr30k: This image captioning dataset contains 30k images. But those images are not divided into training, developing, testing. The one who is using this particular dataset for their model can have their own choice of splitting the images into either training or testing. This mainly concentrates on large objects and the color attributes.

Visual genome dataset: This is the only image captioning dataset which gives separate captions for each and every object, attributes, based on the regions and relations between the objects. This helps in deep understanding of each pixel in the image. But this dataset are not well popular like flicker, coco.

Instagram dataset: This is a pretty different dataset compared to all other datasets. This dataset contains images taken from Instagram which contains mostly celebrities. At the beginning, the idea of this image captioning came into picture to automatically detect the celebrities and locations for the images in social media. This is why researchers prepared a standard benchmark dataset only for social media. Later due to the enhancements in the uses of auto-captioning encouraged researchers to create datasets which concentrates on objects, human, animals, places, colors etc.

There are many other different datasets available for image captioning but the above mentioned are the popularly used datasets. Training one particular model with different datasets can give a broad view of understanding the model working in deeper way. But if the model is generated for only specific group of people like celebrities better automatically go for Instagram dataset. So selecting the correct suitable dataset for the system is very important and necessary since the results are based on that.

Once the model is ready and trained on the suitable standard datasets, there is a need to evaluate the quality of the captions generated from that model. Even though the experts in image-captioning can easily guess the standard of the model by just having a look on the generated captions, there is a need to prove the quality of the captions and difference in captions compared to the captions generated by other models. Below mentioned are the few standard evaluation metrics for image captioning.

BLEU: This is the well-known metric used to evaluate the quality of the generated caption. The BLEU score ranges from 0.0 to 1.0 where 0 is the exact mismatch and 1 is the perfect match. But this works well only for the short captions. When coming for the long paragraph descriptions this BLEU score might not work properly. Using only one evaluation metric is not enough to justify the quality of caption.

SPICE: All other evaluation metrics are overlapped with n-gram. But this SPICE is going to evaluate a system generated caption over a human description for an image. It mainly evaluates the colors and other attributes of an object. This is hard to optimize compared to all other evaluation metrics.

ROUGE: this helps in evaluating the text summary whereas BLEU evaluates short sentences. Different types of ROUGE helps in evaluating different types like adjacent words correlation, summary, entire page etc. But this metric is not going to help in evaluating multiple pages at same time.

V. RESULTS AND DISCUSSION

We have used Flickr 8K dataset to train the model and BLEU score as an evaluation metric to evaluate the caption generated. Our model performed well for translated, scaling and shearing transformation. But it does not work for rotation. The sample of the results generated is shown in Fig.2 and Fig.3.

Original Image	Generated caption	BLEU score
	A person is climbing	60

Fig.2a.. Original Image and Generated Caption

Transformation	Image	Generated caption
Translation		A person is climbing BLUE score 60
Scaling		A person is climbing BLUE score 60
Rotation		erson g down UE re 34

Shearing		A person is climbing BLUE score 60

Fig.2b.Affine Transformed Images and Generated Caption

Original Image	Generated caption	BLUE score
	A kid who is playing	53

Fig.3a.. Original Image and Generated Caption

Transformation	Image	Generated caption
Translation		A kid who is playing BLEU score 53
Scaling		A kid who is playing BLEU score 53

Rotation		A person climbing BLEU score 17
Shearing		A kid who is playing BLEU score 53

Fig.3b.Affine Transformed Images and Generated Caption

VI. CONCLUSION

In this paper, we have discussed the main idea behind the concept of image captioning for original image and its affine transformed images using image hashing technique. Standard benchmark datasets and evaluation metrics used for image captioning are explained. A detailed explanation is given on the proposed system and how our image captioning model predicts the captions for various affine transformed images. It can be useful for visually impaired people, self-driving cars, image search etc. The proposed method works for some of the affine transformed images and relevant captioning is generated. Our model worked for all the transformations except rotation. Further in future the accuracy of the work has to be improved so that the similar captioning will be generated for all the transformed images.

REFERENCES

1. Zhihao zhu, zhan xue, and Zejian Yuan. “Think and Tell: Preview network for Image captioning”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.5561–5570.
2. Shuang Bai and Shan An. 2018. “A Survey on Automatic Image Caption Generation”. Neurocomputing.
3. Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, Barbara Plank, et al. 2016. “Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures”. Journal of Artificial Intelligence Research (JAIR) 55,409–442.
4. XinleiChenandCLawrenceZitnick.2015.Mind’seye:”A recurrent visual representation for image caption generation”. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2422–2431.
5. Bo Dai, Dahua Lin, Raquel Urtasun, and Sanja Fidler.2017. “Towards Diverse and Natural Image Descriptions via a Conditional GAN”. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).2989–2998.
6. Andrej Karpathy and Li Fei-Fei. 2015. “Deep visual-semantic alignments for generating image descriptions”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 3128–3137.
7. Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim.2017. “Attend to You: Personalized Image Captioning with Context Sequence Memory Networks”. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).6432–6440.

8. Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek.2017. “Areas of Attention for Image Captioning”. In Proceedings of the IEEE international conference on computer vision.1251–1259.
9. Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-JiaLi.2017. “Deep Reinforcement Learning-based Image Captioning with Embedding Reward” .In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR).1151–1159.
10. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo.2016. “Image captioning with semantic attention”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.4651–4659.
11. Hanwang Zhang, Zawlin Kyaw,Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, pages 3107–3115, 2017.
12. Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černock`y, and Sanjeev Khudanpur. 2010. Recurrent neural networkbasedlanguagemodel.InEleventhAnnualConferenceoftheInternationalSpeechCommunicationAssociation.
13. MargaretMitchell,XufengHan,JesseDodge,AlyssaMensch,AmitGoyal, AlexBerg,KotaYamaguchi,TamaraBerg,KarlStratos,andHalDauméIII. 2012”Midge:Generating image descriptions from computer vision detections”.InProceedingsofthe13thConferenceoftheEuropeanChapteroftheAssociationforComputationalLinguistics.AssociationforComputationalLinguistics,747–756.
14. Richard Socher,Andrej Karpathy,Quoc VLe,Christopher DManning, andAndrewYNg.2014.“Groundedcompositionalsemanticsforfindinganddescribingimageswithsentences”. Transactions of the Association for Computational Linguistics 2(2014),207–218.
15. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation.InProceedings of the IEEE conference on computer vision and pattern recognition.4566–4575.

AUTHORS PROFILE

M. Nivedita is currently pursuing Ph.D in Computer Science and Engineering, VIT-Chennai Campus, Chennai in the area of Computer Vision and Image Processing. She has completed her M.E. from Anna University, Chennai. She has published 2 papers in journal. Her areas of interest are Image Processing, Computer Vision and Artificial Intelligence.

Asnath Victy Phamila Y, is currently working as Associate Professor in VIT Chennai. Her research area includes Image Processing, Wireless Sensor Networks and Network Security. She has around 13 years of academic and 4 years of industry experience. She has around 20 research papers to her credit. She also serves as reviewer in reputed journals.