International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol. 9, Issue.8, August 2021 E-ISSN: 2347-2693

Automatic Image Caption Generation Using CNN, RNN and LSTM

S.S. Pophale[1*], Praveen Mokate[2], Sandip Najan[3], Sandesh Gajare[4], Sanket Swami[5]

[1,2,3,5]Department of Information Technology, D.V.V.P.C.O.E. Ahmednagar, Maharashtra, India
[4]Department of Computational Sciences and Technology, Delhi University, Delhi, India

[*]*Corresponding Author: shr.pophale@gmail.com*

DOI: https://doi.org/10.26438/ijcse/v9i8.6062 | Available online at: www.ijcseonline.org

Received: 25/Jul/2021, Accepted: 17/Aug/2021, Published: 31/Aug/2021

Abstract— The paper aims at generating automated captions by learning the contents of the image. At present images are annotated with human intervention and it becomes nearly impossible task for huge commercial databases. The image database is given as input to a deep neural network (Convolutional Neural Network (CNN)) encoder for generating "thought vector" which extracts the features and nuances out of our image and RNN (Recurrent Neural Network) decoder is used to translate the features and objects given by our image to obtain sequential, meaningful description of the image .In this paper we are going to explain the survey about image captioning and our proposed system.

Keywords— image annotation, deep learning, CNN, RNN, LSTM, python3, flask, etc.

I. Introduction

Automatic image annotation (also known as automatic image tagging or linguistic indexing) is the process by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. This application of computer vision techniques is used in image retrieval systems to organize and locate images of interest from a database. This method can be regarded as a type of multi-class image classification with a very large number of classes - as large as the vocabulary size. Typically, image analysis in the form of extracted feature vectors and the training annotation words are used by machine learning techniques to attempt to automatically apply annotations to new images. The first methods learned the correlations between image features and training annotations, then techniques were developed using machine translation to try to translate the textual vocabulary with the 'visual vocabulary', or clustered regions known as blobs. Works following these efforts have included classification approaches, relevance models and so on. The advantages of automatic image annotation versus content-based image retrieval (CBIR) are that queries can be more naturally specified by the user. CBIR generally (at present) requires users to search by image concepts such as color and texture, or finding example queries. Certain image features in example images may override the concept that the user is really focusing on. The traditional methods of image retrieval such as those used by libraries have relied on manually annotated images, which is expensive and time-consuming, especially given the large and constantly growing image databases in existence. In this paper, we will combine techniques in both computer vision and natural language processing to form a complete image description approach. This will be responsible for constructing computer-generated natural descriptions of any provided images. The idea is to replace the encoder (RNN layer) in an encoder-decoder architecture with a deep convolutional neural network (CNN) trained to classify objects in images. Normally, the CNN's last layer is the softmax layer, which assigns the probability that each object might be in the image. But if we remove that softmax layer from CNN, we can feed the CNN's rich encoding of the image into the decoder (language generation RNN) designed to produce phrases. We can then train the whole system directly on images and their captions, so it maximizes the likelihood that the descriptions it produces best match the training descriptions for each image. A convolutional neural network can be used to create a dense feature vector. This dense vector, also called an embedding, can be used as feature input into other algorithms or networks. For an image caption model, this embedding becomes a dense representation of the image and will be used as the initial state of the LSTM.An LSTM is a recurrent neural network architecture that is commonly used in problems with temporal dependences. It succeeds in being able to capture information about previous states to better inform the current prediction through its memory cell state. An LSTM consists of three main components: a forget gate, input gate, and output gate. Each of these gates is responsible for altering updates to the cell's memory state.

II. Related Work

1. Long Xu, Jia Li, Weisi Lin, Yongbing Zhang, Lin Ma, Yuming Fang, "Yihua Yan, Multi-task Rank Learning for Image Quality Assessment", IEEE Transactions on Circuits and Systems for Video Technology, 2016.In practice, images are distorted by more than one distortions. For image quality assessment (IQA), the existing machine learning (ML) based methods generally established a

unified model for all the distortion types, or each model is trained independently for each distortion type, which is therefore distortion aware. In distortion-aware methods, the common features among different distortions were not exploited. In addition, there were fewer training samples for each model training, which may result in over fitting. To address these problems, authors proposed a multi-task learning framework to train multiple IQA models together, each model is for each distortion type; however all the training samples are associated with each model training. Thus, the common features among different distortion types, and the said underlying relatedness among all the learning tasks are exploited, which would benefit the generalization ability of trained models and prevent over fitting possibly. In addition, pairwise image quality ranking instead of image quality rating is optimized in this learning task, which is fundamentally departed from the traditional ML based IQA methods toward better performance. The experimental results confirm that the presented Multi-task Rank Learning based IQA (MRLIQ) metric is prominent against all state-of-the-art NR-IQA approaches.

2. Z. Li, J. Liu, J. Tang, and H. Lu, Projective Matrix Factorization withunified embedding for social image tagging", Computer Vision and Image. Understanding, 2014.This paper presents a general formulation, named ProJective Matrix Factorization with unified embedding (PJMF), by which social image retagging is transformed to the nearest tag-neighbor search for each image.

Authors solve the proposed PJMF as an optimization problem mainly considering the following issues. First, it attempt to find two latent representations in a unified space for images and tags respectively and explore the two representations toreconstruct the observed image-tag correlation in a nonlinear manner. In thiscase, the relevance between an image and a tag can be directly modeled as the pair-wise similarity in the unified space. Second, the image latent representation is assumed to be projected from its original visual feature representation with an orthogonal transformation matrix. The projection makes convenient to embed any images including out-of-samples into the unified space, and naturally the image retagging problem can be solved by the nearest tag-neighbors search for those images in the unified space. Third, local geometry preservation of image space and tag space respectively are explored as constraints in order to make image similarity (and tag relevance) consistent in the original space and the corresponding latent space. Experimental results on two publicly available benchmarks validate the encouraging performance of this work over the state-of-the-arts.

3. Zechao Li and Jinhui Tang, Unsupervised Feature Selection via Non negative Spectral Analysis and Redundancy Control", IEEE, 2015.In many image processing and pattern recognition problems, visual contents of images are currently described by high-dimensional features, which are often redundant and noisy. Toward this end, authors proposed a novel unsupervised feature selection scheme, namely, non-negative spectral analysis with constrained redundancy, by jointly leveraging non-negative spectral clustering and redundancy analysis. The presented method can directly identify a discriminative subset of the most useful and redundancy-constrained features. Non-negative spectral analysis is developed to learn more accurate cluster labels of the input images, during which the feature selection is performed simultaneously. The joint learning of the cluster labels and feature selection matrix enables to select the most discriminative features. Row-wise sparse models with a general are leveraged to make the proposed model suitable for feature selection and robust to noise. Besides, the redundancy between features is explicitly exploited to control the redundancy of the selected subset. The presented problem is formulated as an optimization problem with a well-definedobjective function solved by the developed simple yet efficient iterative algorithm. Finally, authors conduct extensive experiments on nine diverse image benchmarks, including face data, handwritten digit data, and object image data. The presented method achieves encouraging the experimental results in comparison with several representative algorithms, which demonstrates the effectiveness of the proposed algorithm for unsupervised feature selection.

4. Z. Lin, G. Ding, M. Hu, J.Wang, and X. Ye, Image Tag Completion via Image-Specific and Tag-Specific Linear Sparse Reconstructions", IEEE, 2013. Though widely utilized for facilitating image management, user-provided image tags are usually incomplete and insufficient to describe the whole semantic content of corresponding images, resulting in performance degradation in tag dependentapplications and thus necessitating effective tag completion methods. In this paper, authors proposed a novel scheme denoted as LSR for automatic image tag completion via image-specific and tag-specific Linear Sparse Reconstructions. Given an incomplete initial tagging matrix with each row representing an image and each column representing a tag, LSR optimally reconstructs each image (i.e. row) and each tag (i.e. column) with remaining ones under constraints of sparsity, considering image-image similarity, image-tag association and tag-tag concurrence. Then both image-specific and tag-specific reconstruction values are normalized and merged for selecting missing related tags. Extensive experiments conducted on both benchmark dataset and web images well demonstrate the effectiveness of the proposed LSR.

III. PROPOSED SYSTEM

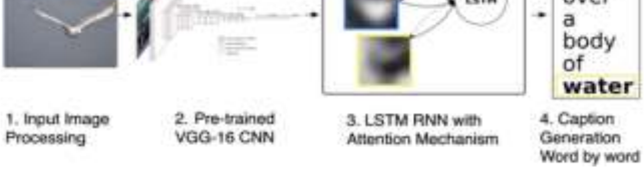
A
bird
flying
over
a
body
of
water
LSTM
1. Input Image Processing
2. Pre-trained VGG-16 CNN
3. LSTM RNN with Attention Mechanism
4. Caption Generation Word by word

Fig: 1. Proposed System

The project aims to generate a brief caption describing the image content, relationship between scenes and identified objects. The system uses attention mechanism which focuses on the region of interest (ROI). It is used to assign weights to the identified objects and more weight is given to the ROI after feature extraction in the encoder. This is then fed into the decoder which decodes the objects according to the weights assigned and uses language model to generate the relevant caption. Inception V3 model is used to accelerate the training as it is pre-trained on Image-net; this helps in classifying the objects easily. After hyper parameter tuning few layers can be extracted from the inception v3 model and used to train the base model. Visual attention mechanism] is used to focus and reassign the weights to the region of interest. This can be manipulated according to the expected output. Visual attention mechanism extracts all the visual words and assigns more weightage using visual mask. Extracted visual words are fed into the language model and using feature similarity. Visual words are used to constrain the attention layer. The decoder uses LSTM (Long short term memory) as the language model to generate the caption. To generate a caption words are sampled to be output at time t. Beam search strategy is used where the pre-determined number called beam size of best by now sentences is computed and expanded with new words. For this experiment the beam size is 10. Greedy search is also experimented and the beam size is set to 1.

Algorithm:

CNN Encoder

Step 1: Dataset containing images along with reference caption is fed into the system

Step 2: The convolutional neural network is used a encoder which extractsimage features f pixel by pixel.

Step 3: Matrix factorization is performed on the extracted pixels. The matrix is of m x n.

Step 4: Max pooling is performed on this matrix where maximum value is selected and again fixed into matrix.

Step 5: Normalization is performed where the every negative value is converted to zero.

Step 6: To convert values to zero rectified linear units are used where each value is filtered and negative value is set to zero.

Step 7: The hidden layers take the input values from the visible layers and assign the weights after calculating maximum probability.

IV. Conclusion

Thus, we have implemented a system which automatically generates annotations for given image. The use of CNN and LSTM along with NLP gives a great enhancement in accuracy for image captioning. So, using deep learning techniques it is possible to generate automatic image annotations. Using this approach, we have achieved an overall efficiency of 93 percent.

References

[1]. Vinyals, Oriol, et al. Show and tell: A neural image caption generator. Proceedings of the IEEE conference on computer vision and pattern recognition, **2015.**

[2]. Deepak A Vidhate, Parag Kulkarni, 2019, International Journal of Computational Systems Engineering, Inderscience Publishers (IEL), **Volume 5, Issue 3, pp 169-178.**

[3]. Fang, Hao, et al. From captions to visual concepts and back. Proceedings of the IEEE conference on computer vision and pattern recognition, **2015.**

[4]. Deepak A Vidhate, Parag Kulkarni, Information and Communication Technology for Intelligent Systems, Springer, Singapore, **pp 693-703, 2019.**

[5]. Y. Bin, Y. Yang, F. Shen, X. Xu, and H. T. Shen, Bidirectional long short term memory for video description, in Proceedings of the 2016 ACM on Multimedia Conference. ACM, **pp. 436440, 2016.**

[6]. Deepak A Vidhate, Parag Kulkarni, Communications in Computer and Information Science, Springer, Singapore, **Volume 905, pp 352-361, 2018.**

[7]. K. Cho, A. Courville, and Y. Bengio, Describing multimedia content using attention-based encoder decoder networks, IEEE Transactions on Multimedia, **vol.17, no. 11, pp. 18751886, 2015.**

[8]. Deepak A Vidhate, Parag Kulkarni, Smart Trends in Information Technology and Computer Communications. SmartCom 2017, **Volume 876, pp 71-81, 2018.**

[9]. B. Qu, X. Li, D. Tao, and X. Lu, Deep semantic understanding of high resolution remote sensing image, in Proc. Int. Conf. Computational., Inf. Telecommunication. Syst., Jul.2016, **pp. 15, 2016.**

[10]. X. Lu, B. Wang, X. Zheng, and X. Li, Exploring models and data for remote sensing image caption generation, IEEE Trans. Geosci. Remote Sens., **vol. 56, no. 4, pp.21832195, Apr., 2018.**

[11]. X. Zhang, X. Wang, X.Tang, H.Zhou , and c.Li, Description generation for remote sensing images using attribute attention mechanism, Remote Sens., **vol. 11, no. 6, p.612, 2019.**