ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 11, November 2018

Fast, Accurate and Natural Caption Generation Using CNN and LSTM

P. Lavanya[1], J. Sarada Lakshmi[2]

M.Tech Student, Department of CS and SE, Andhra University, Visakhapatnam, AP, India[1]

Research Scholar, Department of CS & SE, Andhra University, Visakhapatnam, AP, India[2]

Abstract: Automatically generating a natural language description of an image is a task close to the heart of image understanding. In recent times most of the work have been published on detecting the objects but failed to detect image at the long distance. In this proposed a multi-model neural network method closely related to the human visual system that automatically learns to describe the content of images. This model consists of an object detection applying a CNN and LSTM which extract the information of objects and their spatial relationship in images; besides, a deep recurrent neural network (RNN) based on long short-term memory (LSTM) units with attention mechanism for sentences generation. Each word of the description will be automatically aligned to different objects of the input image when it is generated. This is similar to the attention mechanism of the human visual system. I evaluated my proposed method in python with real-time COCO dataset and derived the results.

KEYWORDS: CNN, Object detection, image captioning, deep learning

I.INTRODUCTION

Now a days machine learning is a widely used term that encompasses many types of programs that you'll run across in big data analytics and data mining. At the end of the day, the "brains" actually powering most predictive programs including spam filters, product recommenders, and fraud detectors are machine learning algorithms. Data scientists are expected to be familiar with the differences between supervised machine learning and unsupervised machine learning as well as ensemble modeling, which use a combination of approaches techniques, and semi-supervised learning, which combines supervised and unsupervised approaches. Humans looking at an image can instantly identifies what objects are in the image, where they are, and how they interact. The human visual system is fast and accurate, allowing us to perform complex tasks like driving with little conscious thought. Fast, accurate, algorithms for object detection would allow computers to drive cars in any weather without specialized sensors, enable assistive devices to con- vey real-time scene information to human users, and unlock the potential for general purpose, responsive robotic systems.

Object detection is a computer technology related to computer vision and image processing that deals with detecting instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos. Well-researched domains of object detection include face detection and pedestrian detection. Humans can easily detect and identify objects present in an image. The human visual system is fast and accurate and can perform complex tasks like identifying multiple objects and detect obstacles with little conscious thought. With the availability of large amounts of data, faster GPUs, and better algorithms, we can now easily train computers to detect and classify multiple objects within an image with high accuracy. An image classification or image recognition model simply detect the probability of an object in an image. In contrast to this, object localization refers to identifying the location of an object in the image. An object localization algorithm will output the coordinates of the location of an object with respect to the image. In computer vision, the most popular way to localize an object in an image is to represent its location with the help of bounding boxes.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 11, November 2018

Image Captioning is the process of generating textual description of an image. It uses both **Natural Language Processing** and **Computer Vision** to generate the captions. The task of image captioning can be divided into two modules; one is an image based model which extracts the features and nuances out of our image, and the other is a language based model which translates the features and objects given by our image based model to a natural sentence. For image based model (viz encoder) – we usually rely on a Convolutional Neural Network model and for language based model (viz decoder) – we rely on a Recurrent Neural Network.

II.RELATED WORK

In Literature survey providing the related background and work on each contribution listed as:
In 2016 Joseph Redmon proposed a system called You Only Look Once (YOLO), a unified model for object detection. This model is simple to construct and can be trained directly on full images. Unlike classifier-based approaches, YOLO is trained on a loss function that directly corresponds to detection performance and the entire model is trained jointly. Fast YOLO is the fastest general-purpose object detector in the literature and YOLO pushes the state-of-the-art in real-time object detection. YOLO also generalizes well to new domains making it ideal for applications that rely on fast, robust object detection.
In 2013 Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik proposed a system called R-CNN: Rich feature hierarchies for accurate object detection and semantic segmentation, a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012---achieving a mAP of 53.3%. Their approach combines two key insights. One can apply high-capacity convolutional neural networks (CNNs) to bottom-up region proposals in order to localize and segment objects and when labelled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significant performance boost. Since they combine region proposals with CNNs, they call our method R-CNN: Regions with CNN features.

III.PROPOSED SYSTEM

The proposed system generates layouts of object locations from real images and combine it with an encoder-decoder image captioning which uses a convolutional neural network to encode the image sequence-to-sequence model that encodes an input object layout as a sequence, and decodes a textual description by predicting the next word at each time step. 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. I present a deep recurrent architecture that automatically generates brief explanations of images. Our models use a convolutional neural network (CNN) to extract features from an image. These features are then fed into a Long Short-Term Memory (LSTM) network to generate a relevant description of the image in valid English.

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. Advantages of this model is it looks at the whole image at test time so its predictions are informed by global context in the image. This makes it more relevant.

The system consists of following modules

1. Data pre-processing module
2. Train with CNN
3. Encode- Decode with LSTM
4. Predict captioning
5. Comparison

Data pre-processing

Pre-process textual descriptions to reduce the vocabulary of words to generate, and in turn, the size of the model, input sequences to a fixed length; this is in fact a requirement of vectorising your input for deep learning

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 11, November 2018

libraries. The dataset contains multiple descriptions for each photograph and the text of the descriptions requires some minimal cleaning. We can pre-compute the "photo features" using the pre-trained model and save them to file. We can then load these features later and feed them into our model as the interpretation of a given photo in the dataset.

Train with CNN

To pretrain our convolutional layers on the ImageNet 1000-class competition dataset. For pretraining we use the first 20 convolutional layers followed by an average-pooling layer and a fully connected layer. We then convert the model to perform detection.

To improve performance we add both convolutional and connected layers to pretrained networks:

1. Add four convolutional layers and two fully connected layers with randomly initialized weights.
2. To achieve fine-grained visual information we increase the input resolution of the network from 224 × 224 to 448 × 448.
3. Our final layer predicts both class probabilities and bounding box coordinates.
4. Then normalize the bounding box width and height by the image width and height so that they fall between 0 and 1.
5. Lastly, increase the loss from bounding box coordinate predictions and decrease the loss from confidence predictions for boxes that don't contain objects.

Encode Decode with LSTM

An encoder CNN model:

A pretrained CNN is used to encode an image to its features. In this implementation VGG16 model is used as encoder and with its pretrained weights loaded. The last softmax layer of VGG16 is removed and the vector of dimension (4096,) is obtained from the second last layer. To speed up my training, I pre-encoded each image to its feature set. This is done in the prepare_dataset.py file to form a resultant pickle file encoded images p. In the current version, the image model takes the (4096,) dimension encoded image vector as input. This can be overrided by uncommenting the VGG model lines in caption generator py. There is no fine tuning in the current version but can be implemented.

A word embedding model: Since the number of unique words can be large, a one hot encoding of the words is not a good idea. An embedding model is trained that takes a word and outputs an embedding vector of dimension (1, 128). Pretrained word embeddings can also be used.

A decoder RNN model:

A LSTM network has been employed for the task of generating captions. It takes the image vector and partial captions at the current time-step and input and generated the next most probable word as output.

Predict Captions

Predicting detections for a test image only requires one network evaluation. The network predicts bounding boxes per image and class probabilities for each box. The grid design enforces spatial diversity in the bounding box predictions. Often it is clear which grid cell an object falls in to and the network only predicts one box for each object. Some large objects or objects near the border of multiple cells can be well localized by multiple cells. The extracted high level image features are fed into a LSTM to generate caption.

Comparison

Comparative analysis is done between R-CNN, Fast-R-CNN and our system with respect to accuracy and speed. I used Long Short Term Memory Network for caption generation.

Dataset

I used COCO dataset for experimental purpose to show the efficiency of proposed method, which includes 123287 color images. For each image, there are at least five captions given by different AMT workers. To make the results

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 11, November 2018

comparable to other methodologies, we use the commonly adopted splits of the dataset, which assigns 5000 for validation, 5000 for testing and keep the rest for training. For each sentence, we limit the length size to 50 words and truncate the rest. Each sentence ends with a character "<*end*>". Words that appear less than five times are marked as a character "<*unk*>". When dictionary is build, resulting in the final vocabulary with 10020 unique words in COCOdataset. The Flickr8k dataset has become a standard benchmark for sentence-based image description. This paper presents Flickr8k Entities, which augments the 158k captions from Flickr8k with 244k co-reference chains, linking mentions of the same entities across different captions for the same image, and associating them with 276k manually annotated bounding boxes. Such annotations are essential for continued progress in automatic image description and grounded language understanding.

The statistics of the datasets are as follows:

	Flickr 8k DataSet	Flickr 100k DataSet	Coco DataSet
Training Time Taken for YoloCap (80% data)	2 hours	20 hours	18 hours
Testing Time Taken for YoloCap (20% data)	20 mins	3 hours	2.5 hours
Accuracy for YoloCap	85%	88%	87%
Training Time Taken for Yolo with LSTM	3 hours	25 hours	20 hours
Testing Time Taken for Yolo with LSTM	15 mins	2 hours	1.5 hours
Accuracy for YOLO with LSTM	88%	91%	92%

IV. EXPERIMENTAL SETUP

Python 3.6 version with NLTK library for Natural Language Processing (NLP) Techniques and Artificial Neural Networks (ANN) including CNN and LSTM are used. It also uses pandas, numpy for data processing, matplotlib library for plotting the graphs, and sklearn library for the machine learning algorithms. The system trains on the sample images dataset as input from the user. The system creates and predicts the caption for the uploaded test image.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

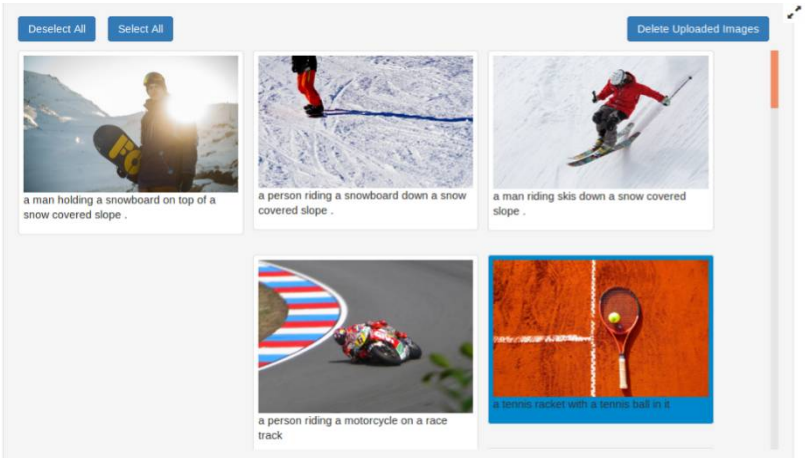
(A High Impact Factor, Monthly, Peer Reviewed Journal)

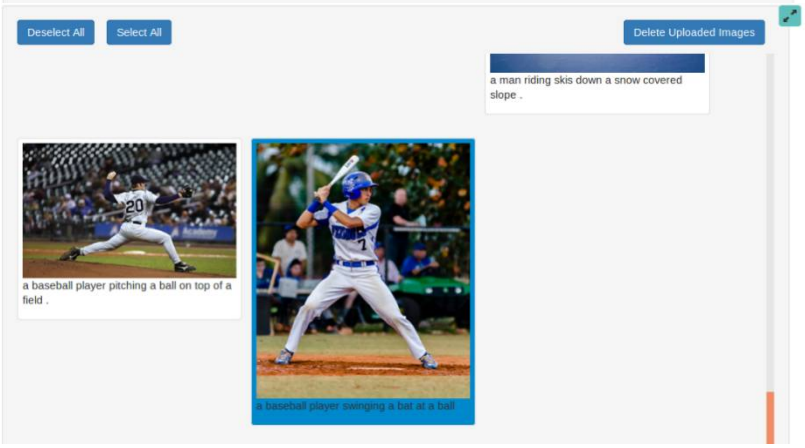
Visit: www.ijirset.com

Vol. 7, Issue 11, November 2018

V.RESULTS

Figures shows the results of text captioning to the selected images. It identifies the images and gives an accurate and natural caption to the selected images.





VI. CONCLUSION

This project Fast, Accurate and Natural Caption Generation Using CNN and LSTMconvolutional neural network and LSTM can be combined to generate captions to an image. It is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. The model is trained to maximize the likelihood of the sentence given the image. Utilize the beam search algorithm to consider multiple captions and select the most probable sentence.

REFERENCES

1. M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. In Computer Vision– ECCV 2008, pages 2–15. Springer, 2008.
2. L. Bourdev and J. Malik. Poselets: Body part detectorstrained using 3d human pose annotations. In International Conference on Computer Vision (ICCV), 2009.
3. H. Cai, Q. Wu, T. Corradi, and P. Hall. The crossdepiction problem: Computer vision algorithms for recognizing objects in artwork and in

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 11, November 2018

photographs. arXiv preprint arXiv:1505.00110, 2015.

4. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005.
5. T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, J. Yagnik, et al. Fast, accurate detection of 100,000 object classes on a single machine. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 1814–1821. IEEE, 2013.
6. J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531, 2013.
7. S. Gidaris and N. Komodakis. Object detection via a multiregion& semantic segmentation-aware CNN model. CoRR, abs/1505.01749, 2015.
8. S. Gould, T. Gao, and D. Koller. Region-based segmentation and object detection. In Advances in neural information processing systems, pages 655–663, 2009.
9. J. Redmon and A. Angelova. Real-time grasp detection usingconvolutional neural networks. CoRR, abs/1412.3128, 2014.
10. He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for imagerecognition. *arXiv preprint arXiv:1512.03385* (2015).
11. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.
12. Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption genera-tor[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.