Deep Convolutional Computation Model for Feature Learning on Big Data

Amit Kore[1], Yogesh Gaikwad[2], Reshma Ganjure[2], Deepika Kunwar[2], Mrunal Gadhave[2]

[1]Assistant Professor AISSMS's Institute of Information Technology, Pune ,India
[2]Student AISSMS's Institute of Information Technology, Pune ,India

Abstract
Image caption generation has wide range of applications and uses in the field of science and technology,experiments and also in day to day lives. This application can be beneficial for the specially abled if implemented in a creative and thoughtful way. The visually challenged can be greatly benefitted if the captions generated to the images in their surrondings are read out loud to them.In this paper, we present a deep recurrent architecture that creates captions to the images that are provided.We have used a convolutional neural network (CNN) to extract features from an image. These extracted features by the CNN are then given as a input to the recurrent neural network (RNN) or a Long Short-Term Memory (LSTM) network to generate a caption. The captions generated by these models are of high accuracy and speed.
***Index Terms*- Deep Learning, Image Captioning, Convolutional Neural Network, Recurrent Neural Network, Long Short-Term Memory**

I.Introduction

Visually challenged people are not gifted with the natural ability to see the surrounding with their eyes.Due to this they face many challenges in their day to day life. To make life easy and to give them the fruits of today's growing sceintific developments one way is through image captioning.With the help of image caption generation,the blind can be able to know their surroundings and also the obstacles in the way. The visually impaired can make use of camera phones to take images of their surroundings. These images can then be used to generate descriptions of the images that can be read out loud.Also the already existing images in the mobiles can also be captioned for their convenience.In this way ,Image Captioning can be of greater help.

II.Literature Survey

This paper[1] discusses briefly about the model based on convolutional networks and recurrent neural network which has dominated in recent image caption generation tasks.It presents a novel parallel-fusion RNN-LSTM architecture which obtains better results than a dominated one and improves the efficiency as well.This approach divides the hidden units of RNN into several same-size parts and lets them work together in parallel.Then the output is merged with corresponding ratios to generate the final results.

The author proposes a parallel-fusion RNN-LSTM architecture that contains two major structures without additional parts compared to the general model Based on CNN the part of image is represented while the caption generation parts is based on RNN.

The author is using NeuralTalk released by Karpathy as experimental platform.The training dataset used is Flickr8k containing 8000 images and each is annotated with 5 sentences. 6000 images are used for training ,1000 for testing and the rest for validation.The proposed algorith adopts fully recurrent network as an example to illustrate the examples.This parallel-fusion model spends less than half of resources and are still able to improve the performance.

This paper[2] proposes an attention based Recurrent Neural Network(RNN)model for camera Document Image Quality Assessment(DIQA).CNN and RNN are integrated in this model to capture spatial features for several glimpse regions step by step within an image patch. The proposed attention based RNN architecture, which could be roughly separated into glimpse network, RNN and action network.

The whole network is trained with SGD algorithm. As it uses supervised learning with OCR accuracy as ground truth, the MSE is defined as the Euclidean distance between the output of the linear regression network yi and the ground truth li. A hybrid loss is defined to combine MSE criterion and variance reduction reward criterion for optimization.

The author has adopted two datasets to train and test DIQA model.First one is Sharpness-OCR-Correlation (SOC) dataset. This dataset includes a total of 175 color images with resolution 1840 × 3264.The second one is Smartdoc-QA dataset.This dataset includes a total of 4260 color images with resolution 3096×4128. Every image from both datasets is in high resolution, so it is nontrivial for neural network to process each sample as a whole. In the training phase, samples are grouped into batches to do optimization. In the testing phase, the predicted accuracy scores of patches that belong to each document image are averaged to obtain a document accuracy score.

The experimental results show that the proposed method is superior to the conventional methods including unsupervised feature extraction based methods and CNN based methods for DIQA.

In this paper[3] the method uses restricted Boltzmann machine to discover a set of hierarchical features from the auxiliary data. It then selects from these features a subset that are helpful for the target learning, using a selection criterion based on the concept of kernel-target alignment. Finally, the target data are extended with the selected features before training.

The proposed transfer framework is based on feature discovery and transfer. The experiments show that this transfer framework is very effective. In some cases, transfer increases classification accuracy by more than 10%. This is significant because the auxiliary data are from categories very different from the ones involved in the target task.

In this paper[4] an IDS model is constructed with deep learning approach. Long Short Term Memory (LSTM)is applied to Recurrent Neural Network (RNN) and use it for an Intrusion Detection System(IDS) model. Recurrent Neural Network (RNN) is extension of a convention feed-forward neural network.For making them powerful for modeling sequences RNN have cyclic connections.

The author uses KDD Cup 1999 dataset which is used to measure a performance of IDS in many researches. In the dataset there are 4,898,431 network traffics and each traffic has 41 features. And according to their characteristic 22 attacks are categorized.Because there are too many data records in the original dataset, KDD Cup 1999 is used ,10 percent data for training and testing. However, towards DoS attack the data ratio of the attacks is weighted. And the others ratio is only 1 percent.

Hyper-parameters are parameters for model initiation. Depending on the value of hyper-parameter, the performance is changed. First, they change the learning rate from 0.0001 to 0.1. For more evaluating precisely,they calculate the efficiency. Although the DR is the lowest value, they get the best efficiency when they set the learning rate 0.01.The authors sets the the hyperparameters for training the IDS model. For testing, they generated 10 test datasets which are selected from kdddup.data.txt. In order to an objective evaluation, they compare their result to other classifier algorithms.Percentages of the DR and Accuracy are the best, even though the FAR is a little bit higher than other algorithms.

In this paper[5], deep sentiment representation based on convolutional neural network and long short-term memory recurrent neural network capturing model is proposed.As input the model uses the pre-trained word vectors and employs convolutional neural network to gain significant local features of the text,then to two-layer LSTMs features are fed, which can extract context-dependent features and generate sentence representation for sentiment classification

In this paper, a new approach based CNN and LSTM are proposed in order to capture deep sentiment features,which has one-layer CNN and two-layer LSTMs that are stacked in order. The word2vec toolkit provided by Google is used to generate word vectors, and then the pre-trained word vectors are connected in series to represent each sentence. In this model, they adopt two-layer LSTMs stacked on CNN, each LSTM is corresponds to the version designed by Zaremba and Sutskever.

LSTM processes the input vectors by recursive execution of cell block which is dependent on the old hidden state as well as the current input. They apply the designed model for deep representation based on CNN and LSTM in sentiment classification to evaluate its effectiveness. The dataset is the comments about sina micro-blog provided by the seventh COAE (Chinese Orientation Analysis and Evaluation) conference which are denoted as D1. The dataset is in Chinese language and its percentages of positive and negative are all 50%. The D1 contains 7226 comments, 7/8 of which are training sets and 1/8 of which are test sets. The model designed is compared with four methods, including CNN, LSTM, CNN-LSTM, and SVM. It is clear that this model achieves good results in extracting deep sentiment representation for sentiment classification due to its special network structure.

This paper [6] it introduces a novel visualization technique. It gives insight into the function of intermediate feature layers and the operation of the classifier.
It is used in a diagnostic role, these visualizations allow us to find model architectures that outperform Krizhevsky et al. on the ImageNet classification benchmark.
It also performs an ablation study to discover the performance contribution from different model layers. When the softmax which is one classifier is trained again the ImageNet generalizes well to the other datasetst, it beats the current state-of-the-art results on Caltech-101 and Caltech-256 which are the 2 datasets.

Paper introduces a visualization technique that reveals the input stimuli that excite individual feature maps at any layer in the model. It also allows us to observe the evolution of features while training and to diagnose potential problems with the model.
The visualization technique uses a multi-layered Deconvolutional Network.

This paper[7] presents a new dataset with the goal of advancing the state-of-the-art in object recognition by placing the question of object recognition in the context of the broader question of scene understanding. This can be achieved by gathering images of complex everyday scenes contains common objects in their natural context. Objects are labeled using per-instance segmentations to help in precise object localization. This dataset contains photos of 91 objects types that can be easily recognizable by anyone. It contains a total of 2.5 million instances which are labeled in 328k images. The creation of this dataset drew upon extensive workers involvement through novel user interfaces for instance spotting , instance segmentation and category detection

It presents a detailed statistical analysis of the dataset in comparison to PASCAL, ImageNet, and SUN. Finally, It provides baseline performance analysis for bounding box and segmentation detection results using a Deformable Parts Model. Datasets related to object recognition can be split into three groups of image classification, object detection and semantic scene.

Image Classification The task of object classification requires binary labels indicating whether objects are present in an image. Early datasets of this type comprised images containing a single object with blank backgrounds, such as the MNIST handwritten digits or COIL household objects. Caltech 101 and Caltech 256 marked the transition to more realistic object images retrieved from the internet while also increasing the number of object categories to 101 and 256, respectively.

Object detection : Detecting an object needs both specifying that an object belonging to a specified class is present, and localizing it in the image. The location of an object is generally represented by a bounding box.
As the detection of many objects like cellphones or chairs ,sunglasses are highly dependent on contextual information, it is very important that detection datasets contain objects in their natural environments.
Semantic scene : For labeling semantic objects in a scene it is necessary that each pixel of an image be labeled as belonging to a category such as chair, floor, street, sky etc.
In difference to the detection task, individual instances of an objects does not need to be segmented.

In this paper[8], a unified tensor model is proposed to represent the unstructured, semi structured, and structured data. Variety and veracity are two distinct characteristics of large-scale and heterogeneous data .It has been a great challenge to efficiently represent and process big data with a unified scheme .Although many studies have been done on big data processing, very few have addressed the following two key issues: (1) how to represent the various types of data with a simple model; (2) how to extract the core data sets which are smaller but still contain valuable information, especially for streaming data. The purpose of this paper is to explore the above raised issues which are closely related to the variety and veracity characteristics of big data.

With tensor extension operator, different types of data are represented as sub tensors and then this sub tensors are combined to a unified tensor. Core tensor are small but contains valuable information. For extraction of core tensors an incremental high order singular value decomposition (IHOSVD) method is given . By recursively applying the incremental matrix decomposition algorithm, IHOSVD can update the orthogonal bases and compute the new core tensor. Analyzes in terms of time complexity, memory usage, and approximation accuracy of the proposed method are provided in this paper. A case study demonstrates that approximate data reconstructed from the core set containing 18% elements can guarantee 93% accuracy in general. Theoretical analyzes and experimental results demonstrate that the proposed unified tensor model and IHOSVD method are efficient for big data representation and dimensionality reduction.

This paper presents a unified tensor model for big data representation and an incremental dimensionality reduction method for high-quality core set extraction. Data with different formats are employed to illustrate the representation approach, and equivalent theorems are proven to support the proposed reduction method. The major contributions are summarized as follows:
Unified Data Representation Model: It proposes a unified tensor model to integrate and represent the unstructured, semi-structured, and structured data. The tensor model has extensible orders to which new orders can be dynamically appended through the proposed tensor extension operator.
Core Tensor Equivalence Theorem: To tackle the recalculation and order inconsistency problems in big data processing with tensor model, It proves a core tensor equivalence theorem which can serve as the theoretical foundation for designing incremental decomposition algorithms.
Recursive Incremental HOSVD Method: It presents a recursive Incremental High Order Singular Value Decomposition method for streaming data dimensionality reduction. Detailed analyses in terms of time complexity, memory usage and approximation accuracy are also investigated.

This paper[9], presents a visual analytics approach for better understanding, diagnosing, and refining deep CNNs. Deep convolutional neural networks (CNNs) have achieved great performance in many pattern recognition tasks such as image classification. It formulates a deep CNN as a directed acyclic graph.From this formulation, a hybrid visualization is developed. It is used to disclose the multiple facets of every neuron and the interactions among them. In particular, It introduces a hierarchical rectangle packing algorithm and a matrix reordering algorithm to show the derived features of a neuron cluster. It also proposes a bi clustering-based method which is one edge bundling method used to decrease visual clutter occurred by a large number of connections between neurons. It evaluates its method on a set of CNNs and the results are generally favourable

There are basically two technical challenges to understand and analyze deep CNNs. First, a CNN may consist of tens or hundreds of layers (depth), with thousands of neurons (width) in each layer, as well as millions of connections between neurons. Such large CNNs are hard to study because of the sizes involved. Second, CNNs consist of many functional components whose values and roles are not well understood either as individuals or as a whole [5]. In addition, how the non-linear components interact with each other and with other linear components in a CNN is not well understood by experts. In most cases, it is difficult to summarize reusable

knowledge from a failed or successful training case and transfer it to the development of other relevant deep learning models.

In this paper[10], the authors proposes a hybrid system employing the use of multilayer Convolutional Neural Network (CNN) to generate vocabulary describing the images and a Long Short Term Memory (LSTM) to accurately structure meaningful sentences using the generated keywords .In the past few years, the problem of generating descriptive sentences automatically for images has gained a rising interest in natural language processing and computer vision research .For image captioning it is required the semantic understanding of images and ability to generate the description of sentences with correct and proper structure. The convolutional neural network compares the target image with a large dataset of training images and then generates an accurate description of the target image using the trained captions.

It showcases the efficiency of the proposed model using the Flickr8K and Flickr30K datasets and show that their model gives superior results compared with the state-of-the-art models utilizing the Bleu metric. The Bleu metric is an algorithm which evaluates the performance of a machine translation system by grading the quality of text translated from one natural language to another. The performance of the model is evaluated using standard evaluation matrices, which outperform previous benchmark models.
For image captioning there are several image datasets are available but most common datasets are Pascal VOC dataset, Flickr 8K and MSCOCO Dataset. Flickr 8K Image captioning dataset [9] is used here . Flickr 8K is a dataset consisting of 8,092 images from the Flickr.com website. This dataset contains collection of day-to-day activity pictures with their related captions. First each object in image is labeled and after that description is added based on objects in an image. It splist 8,000 images from this corpus into three disjoint sets. The training data (DTrain) has 6000 images and the development and test dataset has 1000 images each.

The model consists of 3 phases:

A. Image Feature Extraction
The features of the images from the Flickr 8K dataset is extracted using the VGG 16 model due to the performance of the model in object identification. The VGG is a convolutional neural network which consists of 16 layer which has a pattern of 2 convolution layers followed by 1 dropout layers until the fully connected layer at the end.

B. Sequence processor
The function of a sequence processor is to handle the text input by acting as a word embedding layer. The embedded layer contains the rules to extract the required features of the text and consists of a mask to ignore padded values. For the final phase of the image captioning the network is then connected to a LSTM

C. Decoder
In the final phase of the model it combines the input from the Image extractor phase and the sequence processor phase by using an additional operation and then fed to a 256 neuron layer and at last to the final output Dense layer which produces a softmax prediction of the next word in the caption over the entire vocabulary which was formed from the text data .That text data was processed in the sequence processor phase.

IV.Future Scope
The great time challenge of image captioning is the successful automation of interpreting any image. This however comes in handy whenever text is used for the images and can be generated or inferred from pictures itself. There are a range of applications that help in extracting description or in- sights from any given source of text. The similar process can be followed here but a little differently. This might also help people who have issues such as visual impairment and will therefore rely on the captions and thus it may also help in social media. Robots can caption the images and can describe the contents of any picture in a more accurate manner. There is a training procedure that is needed to be undertaken in order to enable smooth implementation. The image captioning procedures are the forte of the top machine learning companies in India. These are companies who have the skills and systems required for smoothly generating captions for particular images in a smaller period of time. Image Captioning has diverse applications across multiple areas of operation for multifarious organizations and digital companies.

V.Conclusion
As image captioning is commonly used in almost every domain, these systems need to remain up to date with changing technology. With deep learning being one of the emerging machine learning technology, it is being applied in image captioning for improvement. In this paper we discussed in brief about image captioning and deep learning and reviewed many deep learning-based image captioning models. These techniques overcome limitations of traditional systems but still have many challenges to tackle. We discuss limitation and future scope for each of these systems. We hope our study paper help others understand current deep learning-based image captioning models and direction for future research and work.

References

[1]A Parallel-Fusion RNN-LSTM Archtecture For Image Caption Generation- Minsi Wang∗†, Li Song∗†, Xiaokang Yang∗†, Chuanfei Luo‡

[2] Attention Based RNN Model for Document Image Quality Assessment -Pengchao Li∗, Liangrui Peng∗, Junyang Cai∗, Xiaoqing Ding∗,Shuangkui Ge†

[3]DeepTransferLearningviaRestrictedBoltzmannMachineforDocument Classification- Jian Zhang

[4] Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection-Jihyun Kim, Jaehyun Kim, Huong Le Thi Thu, and Howon Kim

[5] Deep Sentiment Representation Based on CNN and LSTM- Qiongxia Huang*, Xianghan Zheng*, Riqing Chen*, Zhenxin Dong

[6] Visualizing and Understanding Convolutional Networks ,Mathew D.Zeiler,Rob Fergus,ECCV.

[7] Microsoft COCO: Common Objects in Context ,Tsung-yi lin,micheal matire ,serge ,jan-18,arXiv

[8] A Tensor Based Approach for Big Data Represenation and Dimensionality Reduction liwei kuang ,fei hao,laurance t.,yang man li,IEEE,2014

[9] Towards Better Analysis of Deep Convolutional Neural Network ,nenchen liu,IEEE,2017

[10] Image Captioning - A Deep Learning Approach , Lakshminarasimhan Srinivasan , Dinesh Sreekanthan , Amutha A.L