Survey on Feature Extraction of Images for Appropriate Caption Generation

[1]Aswathy K S, [2]Prof. (Dr.) Gnana Sheela K
[1] Department of Electronics and Communication, Toc H Institute of Science and Technology, Kerala, India
Email: aswathysudhan91@gmail.com
[2] Department of Electronics and Communication, Toc H Institute of Science and Technology, Kerala, India

Abstract- In many industrial, medical and scientific image processing applications, various feature and pattern recognition techniques are used to match specific features in an image with a known template. Despite the capabilities of these techniques, some applications require simultaneous analysis of multiple, complex, and irregular features within an image as in semiconductor wafer inspection. In wafer inspection discovered defects are often complex and irregular and demand more human-like inspection techniques to recognize irregularities. By incorporating neural network techniques such image processing systems with much number of images can be trained until the system eventually learns to recognize irregularities. The aim of this project is to develop aframework of a machine-learning system that can produce caption to accurately describe images. Such a system finds application in semiconductor industry, biomedical field where microscopy for identifying different cell types and analysis of tumour growth etc. Also such systems help visually impaired people in understanding pictures. They can be used for providing alternate text for images in parts of the world where mobile connections are slow and making it easier for everyone to search on Google for images are also possible.

Keywords: Artificial Intelligence, Neural Networks, Computer Vision, Learning, Bag of words, Caption, Convolution;

I. INTRODUCTION

Automatically describing the content of an image is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. In machine learning and cognitive science, artificial neural networks (ANNs) are a family of statistical learning algorithms inspired by biological neural networks (the central nervous systems of animals, in particular the brain) and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which send messages to each other. The connections have numeric weights that can be tuned based on experience, making neural nets adaptive to inputs and capable of learning. For example, a neural network for handwriting recognition is defined by a set of input neurons which may be activated by the pixels of an input image. After being weighted and transformed by a function (determined by the network's designer), the activations of these neurons are then passed on to other neurons. This process is repeated until finally, an output neuron is activated. This determines which character was read. Like other machine learning methods - systems that learn from data - neural networks have been used to solve a wide variety of tasks that are hard to solve using ordinary rule-based programming, including computer vision and speech recognition.

II. LITERATURE REVIEW

Automatically generating captions of an image is a task very close to the heart of scene understanding. This requires, identifying and detecting objects, people, scenes etc., reasoning about spatial relationships and properties of objects, combining several sources of information into a coherent sentence. Hence it is a complex task to define an image or a scene; which is an important problem in the field of computer vision. Even though it is a challenging one, a lot of research is going on which explores the capability of computer vision in the field of image processing and it helps to narrow the gap between the computer and the human beings on scene understanding. The purpose of this survey is to analyze various techniques used for an image caption generation using the neural network concepts.

Jim Mutch et al (2006) proposed a biologically inspired model of visual object recognition to the multiclass object categorization problem. First applied gabor filters on every positions and scale of the image. The template matching and max pool operations are utilized to built the feature complexity and position or scale invariance. Utilizes S-layer called the simple layer which uses convolution with local filters to compute higher order features. C-layer stands for the complex layer which increases invariance by pooling units of same type in previous layers over limited ranges [20].

Eric Nowak et al (2006) described about various sampling methods for BoF for image classification. Representing images as a collection of independent local patches has been a great aid for object recognition or image classification; but it raises the question of which patch to choose. There lies the importance of sampling. Dense sampling processes every pixel at every scale, thus captures the most information, but it is also memory and computation intensive as it spends time on processing relatively featureless regions [19].

Juan C Caicedo et al (2009) proposed a model for the evaluation of different representations obtained from the bag of features approach to classify histopathology images. The process involved feature detection and description, construction of a visual vocabulary and image representation building through visual word occurrence analysis. The obtained image descriptors are processed using appropriate kernel functions for Support Vector Machines classifiers. Medical imaging applications are challenging because they require effective and efficient content representations to manage large image collections. The first stage for medical image analysis is modeling image contents by defining an appropriate representation, which is a fundamental problem for all image analysis tasks such as image classification, automatic image annotation, object recognition and image retrieval, which require discriminative representations according to the application domain. In this work two feature detection strategies with their corresponding feature descriptor have been evaluated. The first strategy is dense random sampling and the other one is the raw pixel descriptor which is more computationally efficient [18].

Ahmet Aker et al (2010) describes about the automatic captioning of geo-tagged images. The approach was based on the summarization of multiple web documents that holds information related to an image's location. It is based on dependency patterns modeled towards sentences which contains features typically provided for different types of images such as church, bridges etc. Such a model is said to have high scores than any other methods proposed earlier. Such dependency patterns also lead to more readable summaries than those generated without dependency patterns. This method is applied only to images of static features of the built or natural landscape, i.e. objects with persistent geo-coordinates, such as building sand mountains, and not to images of objects which move about in such landscapes, e.g. people, cars, clouds, etc. The summarizer is an extractive, query-based multi-document. It is given two inputs: a toponym associated with an image and a set of documents to be summarized which have been retrieved from the web using the toponym (topographical feature) as a query. The summarizer creates image descriptions in a three step process. First, it applies shallow text analysis, including sentence detection, tokenization, lemmatization and POS-tagging to the giveninput documents. Then it extracts features from the document sentences. Finally, it combines the features using a linear weighting scheme to compute the final score for each sentence and to create the final summary [17].

Xiaoli Yuan et al (2011) proposed a novel image retrieval system based on bag-of-features (BoF) model by integrating scale invariant feature transform (SIFT) and local binary pattern(LBP) which leads to a patch-based integration and an image-based integration. Based on thier experimental studies, this image-based integrationgave the best performance compared to other existing models utilizing a codebook size of N = 200 and K-means weight of w=0.6 [16].

Stephen O'hara et al (2011) studied about the growing importance of Bag of Features (BoF) approaches to many computer vision tasks, including image classification, video search, robot localization, and texture recognition. The main reason is its simplicity and described it as a method based on order-less collections of quantized local image descriptors and they discard spatial information and are therefore conceptually and computationally simpler than many alternative methods.There are two common perspectives for explaining the BoF image representation. Thefirst is by analogy to the Bag ofWords representation. With Bag of Words, one represents a document as a normalized histogram of word counts. Commonly, one counts all the words from a dictionary that appear in the document. This dictionary may exclude certain non-informative words such as articles like "the", and it may have a single term to represent a set of synonyms. The term vector that represents the document is a sparse vector where each element is a term in the dictionary and the value of that element is the number of times the term appears in the document divided by the total number of dictionary words in the document. The term vector is the Bag of Words document representation called a bagbecause all ordering of the words in the document have been lost [15].

Yezhou Yang et al (2011) proposed a sentence generation strategy that describes images by predicting the most likely nouns, verbs, scenes and prepositions that make up the core sentence structure. The inputs are initial noisy estimates of the objects and scenes detected in the image using state of the art trained detectors.It used a language model trained from the English Gigaword corpus to obtain their estimates; together with probabilities of co-located nouns, scenes and prepositions.Then used these estimates as parameters on a HMM that models the sentence generation process, with hidden nodes as sentence components and image detections as the emissions.The most natural thing would be to describe it using words: using speech or text. This description of an image is the output of an extremely complex process that involves: 1) perception in the Visual space, 2) grounding to World Knowledge in the Language Space and 3) speech/text production, and introduced a computational framework that attempts to integrate these components together. Here the hypothesis is based on the assumption that natural images accurately reflect common everyday scenarios which are captured in language.For example, knowing that boats usually occur over water will enable us to constrain the possible scenes a boat can occur and exclude highly unlikely ones – street, highway etc.It also enables us to predict likely actions (verbs) given the current object detections in the image: detecting a dog with a person will likely induce walk rather than swim, jump, fly.Key to this approach is the use of a large generic corpus such as the English Gigaword as the semantic grounding to predict and correct the initial and often noisy visual detections of an image to produce a reasonable sentence that succinctly describes the imageThe input is a test image where objects and scenes are detected using trained detection algorithms [14].

Siming Li et al (2011) presented an approach to automatically compose image descriptions given computer vision based inputs and using web-scale n-grams. Experimental results indicate that it is viable to generate simple textual descriptions that are pertinent to the specific content of an image, while permitting creativity in the description – making for more human-like annotations than previous approaches.These work contrasts to most previous approaches in four key aspects; First composed fresh sentences from scratch, instead of retrievingor summarizing existing text fragments associated with an image.Second, generated textual descriptions that are truthful to the specific content of the imagework in automatic caption generation creates news-worthy textor encyclopedic textthat is contextually relevant to the image, but not closely pertinent to the specific content of the image. Third one aimed to build a general image description method as compared to work that requires domain specific hand-written grammar rules. The proposed approach consists of two steps: (n-gram) phrase selection and (n-gram) phrase fusion.The first step – phrase selection – collects candidate phrases that may be potentially useful for generating the description of a given image. This step naturally accommodates uncertainty in image recognition inputs aswell as synonymous words and word re-ordering to improve fluency. The second step – phrase fusion finds the optimal compatible set of phrases using dynamic programming to compose a new phrase that describes the image [13].

Misha Denil et al (2011) proposed an attention model for simultaneous object tracking and recognition that is driven by gaze data. Motivated by theories of perception, the model consists of two interacting pathways: identity and control, intended to mirror what and where pathways in neuroscience models. The identity pathway models object appearance and perform classification using deep Restricted Boltzmann Machines. At each point in time the observations consist of foveated images, with decaying resolution toward the periphery of the gaze. The control pathway models the location, orientation, scale and speed of the attended object. The posterior distribution of these states is estimated with particle filtering.This approach gives good performance in the presence of partial information and allows us to expand the action space from a small, discrete set of fixation points to a continuous domain [12].

Chih-Fong Tsai et al (2012) described about Content based image retrieval (CBIR) which require users to query images by their low-level visual content and this not only makes queries, but also can lead to unsatisfied retrieval results. Thus an image annotation was proposed. The aim of image annotation is to automatically assign keywords to images, so image retrieval users are able to query images by keywords. Image annotation can be regarded as the image classification problem either image is represented by some low- level features and high-level concepts. The bag-of-words is one of the most widely used feature representation method.

Image annotations is an automatic classification of images by labeling images intoone of a number of predefined classes or categories, whereclasses have assigned keywords or labels which can describethe conceptual content of images in that class [11].

Bharathi S et al (2014) proposed a BoF framework for remote sensing image classification. The most representative features are selected using Gabor convolution, Scale-invariant Feature Transform (SIFT) key points and Random Sample Consensus (RANSAC) method**.** The feature vector forms the classifier under K-means and Support Vector Machines (SVM) for semantic annotation. In the testing stage key points are extracted from every image, fed into the visual dictionary to map them with one feature vector and it was finally fed into the multi-class SVM training classifier model to recognize the category of an image. The time complexity of the classification is not very complex; it took 3mins for given dataset. And they came to a conclusion that BoF is one of the best methods for content based image classification. To extract Gabor features, a set of Gabor filters tuned to several different frequencies and orientations are utilized.SIFT keypoints are used for feature extraction. It has rich information and is suitable for fast and accurate features in huge data set and will produce a large number of feature vectors even though there are a few objects [10].

Razvan Pascanu et al (2014)in this paper describes the different ways to extend a recurrent neural network (RNN) to a deep RNN.The three important points of an RNN includes (1) input-to-hidden function, (2) hidden-to hidden transition and (3) hidden-to-output function.Two novel architectures of a deep RNN which are orthogonal to an earlier attempt of stacking multiple recurrent layers to build a deep RNN are proposed.The proposed deep RNNs are empirically evaluated on the tasks of polyphonic music prediction and language modeling. The experimental result supports that the proposed deep RNNs benefit from the depth and outperform the conventional, shallow RNNs. Four types of RNN are described in this paper and based on them two deep RNN architectures are defined [9].

Andrej Karpathy et al (2014) presented a model that generates natural language descriptions of images and their regions.This model is based on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks over sentences, and a structured objective that aligns the two modalities through a multimodal embedding.Then used a Multimodal Recurrent Neural Network (MRNN) architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. Used the Flickr8K, Flickr30K and MSCOCO datasets for the experiment.The Multimodal RNN model is subject to multiple limitations. First, the model can only generate a description of one input array of pixels at a fixed resolution. A more sensible approach might be to use multiple saccades around the image to identify all entities, their mutual interactions and wider context before generating a description [8].

Junhua Mao et al (2014) proposed a multimodal Recurrent Neural Network (m-RNN) model for generating novel image captions. It directly models the probability distribution of generating a word given previous words and an image. Image captions are generated according to this distribution. The model consists of two sub-networks: a deep recurrent neural network for sentences and a deep convolutional network for images. These two sub-networks interact with each other in a multimodal layer to form the whole m-RNN model [7].

Jeff Donahue et al (2014)developed a novel recurrent convolutional architecture suitable for large-scale visual learning which is end-to-end trainable, and demonstrate the value of these models on benchmark video recognition tasks, image description and retrieval problems, and video narration challenges.In contrast to existing models which assume a fixed spatio-temporal receptive field or simple temporal averaging for sequential processing, recurrent convolutional models are “doubly deep” in that they can be compositional in spatial and temporal “layers”. Such models may have advantages when target concepts are complex and/or training data are limited.Long-term RNN models are appealing in that they directly can map variable-length inputs (e.g., video frames) to variable length outputs (e.g., natural language text) and can model complex temporal dynamics; yet they can be optimized with back propagation [6].

Kyunghyun Cho et al (2014) propose a novel neural network model called RNN Encoder– Decoder that consists of two recurrent neural networks (RNN). One RNN encodes a sequence of symbols into a fixed length vector representation, and the other decodes the representation into another sequence of symbols. The encoder and decoder of the proposed model are jointly trained tomaximize the conditional probability of a target sequence given a source sequence. The performance of a statistical machine translation system is empirically found to improve by using the conditional probabilities of phrase pairs computed by the RNN Encoder–Decoder as an additional feature in the existing log-linear model. Qualitatively the proposed model learns a semantically and syntactically meaningful representation of linguistic phrases maximize the conditional probability of a target sequence given a source sequence [5].

Dzmitry Bahdanau et al (2015) The models proposed recently for neural machine translation often belong to a family of encoder–decoders and encode a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, it is concluded that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft) search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation has been achieved. Furthermore, qualitative analysis reveals that the (soft) alignments found by the model agree well with this intuition [4].

Jimmy Lei Ba et al (2015) proposed an attention-based model for recognizing multiple objects in images.The proposed model is a deep recurrent neural network trained with reinforcement learning to attend to the most relevant regions of the input image.The model learns to both localize and recognize multiple objects despite being given only class labels during training.Finally evaluated the model on the task of transcribing multi-digit house numbers from publicly available Google Street View imagery and showed that it is more accurate than the state-of-the-art convolutional networks and uses fewer parameters and less computation.Processing an image x with an attention based model is a sequential process with N steps.At each step n, the model receives a location l(n) along with a glimpse observation x(n) taken at location l(n).The model uses the observation to update its internal state and outputs the location l(n+1) to process at the next time-step. Usually the number of pixels in the glimpse x(n) is much smaller than the number of pixels in the original image x, making the computational cost of processing a single glimpse independent of the size of the image [3].

Oriol Vinyals et al (2015) In this paper, a generative model based on a deep recurrent architecture that combines recent advances in computer vision and machine translation and that can be used to generate natural sentences describing an image is presented. The model is trained to maximize the likelihood of the target description sentence given the training image. Experiments on several datasets show the accuracy of the model and the fluency of the language it learns solely from image descriptions. This model is often quite accurate, when verified both qualitatively and quantitatively. For instance, while the current state-of-the-art BLEU-1 scores on the Pascal dataset is 25, this approach yields 59, to be compared to human performance around 69. Also showed BLEU-1 score improvements on Flickr30k, from 56 to 66, and on SBU, from 19 to 28. Lastly, on the newly released COCO dataset, it achieved a BLEU-4 of 27.7, which is the current state-of-the-art [2].

Kelvin Xu et al (2015) described an approach to caption generation that attempt to incorporate a form of attention with two variants: a hard attention mechanism and a soft attention mechanism. The hard stochastic attention mechanism is trainable by maximizing an approximate variation lower bound while the soft deterministic attention mechanism is trainable by standard back-propagation methods. The main attention of the framework is the visualization of ‘Where’ and ‘What’ the attention is focused on. Here a CNN act asan encoder and it extracts a set of features called convolution features of the input image.In order to obtain a correspondence between the feature vectors and portions of the 2-D image, features are extracted from a lower convolutional layer. This allows the decoder to selectively focus on certain parts of an image by selecting a subset of all the feature vectors.Then used a long short-term memory (LSTM) networkthat produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words. Two alternative mechanisms are used for learning as stochastic attention and deterministic attention. Stochastic hard attention represents location variables as where the model decides to focus attention when generating a particular word.Learning stochastic attention requires sampling the attention location while taking the direct expectation of the context vector can formulate deterministic soft attention model.Finally, quantitatively validated the usefulness of attention in caption generation with state of the art performance on three benchmark datasets: Flickr8k, Flickr30k and the MS COCO dataset [1].

III. CONCLUSION

The major part lies in the decision of choosing the better method for feature extraction. There are different types of features as well. There are various methods of feature extraction in image processing. Upon survey it was found that most of the former methods are concentrating on a single feature alone, which would not aid for my purpose. Hence after working on various available methods the SURF features were found to be better as it is independent of the scale and orientation of an image. The Bag-of-Words utilizes the SURF feature extraction method.

REFERENCES:

[1] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, YoshuaBengio, “Show, Attend and Tell: Neural Image CaptionGeneration with Visual Attention”, February 2015.

[2] OriolVinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator”, Google, 2015.

[3] Jimmy Lei Ba, Volodymyr Minho, Koray Kavukcuoglu, “Multiple object recognition with visual attention”, Google, April 2015.

[4] Dzmitry Bahdanau, Kyung Hyun Cho, Yoshua Bengio, “Neural machine translation by jointly learning to align and translate”, April 2015.

[5] Kyunghyun Cho, Bart van, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, Yoshua Bengio, “Learning Phrase Representations using RNN Encoder–Decoderfor Statistical Machine Translation”, September 2014.

[6] Jeffrey Donahue, Lisa Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, Trevor Darrell, “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”, University of California at Berkeley, Technical Report No. UCB/EECS-2014-180, November 2014.

[7] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan L. Yuille, “Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)”, Published as a conference paper at ICLR 2015,July 2014.

[8]Andrej Karpathy, Li Fei-Fei, “Deep Visual-Semantic Alignments for Generating Image Descriptions”, Stanford University, 2014.

[9] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Yoshua Bengio, “How to Construct Deep Recurrent Neural Networks”, April 2014.

[10] Bharathi S, Karthik Kumar S, P Deepa Shenoy, Venugopal K R, L M Patnaik, “Bag of Features Based Remote Sensing Image Classification Using RANSAC And SVM”, Proceedings of the International Multi Conference of Engineers and Computer Scientists Vol I, IMECS, March 2014.

[11] Chih-Fong Tsai, F. Camastra, “Bag-of-Words Representation in Image Annotation: A Review”, International Scholarly Research Network ISRN Artificial Intelligence Volume 2012.

[12] Misha Denil, Loris Bazzani, Hugo Larochelle, Nando de Freitas, “Learning where to Attend with Deep Architectures for Image Tracking”, September 2011.

[13] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi, “Composing Simple Image Descriptions using Web-scale N-grams”, Stony Brook University, 2011.

[14] Yezhou Yang , ChingLikTeo, Hal Daume, Yiannis Aloimonos, “Corpus-Guided Sentence Generation of Natural Images”, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pages 444–454,Scotland, UK, July, 2011.

[15] Stephen O’hara AND Bruce A. Draper , “Introduction to the bag of features paradigm for image classification and retrieval”, January 2011.

[16] “A SIFT-LBP image retrieval model based on Bag-Of-Features”, 18th IEEE International Conference on Image Processing, 2011.

[17] Ahmet Aker, Robert Gaizauskas¸”Generating image descriptions using dependency relational patterns”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1250–1258, Uppsala, Sweden, July 2010.

[18] Juan C. Caicedo, Angel Cru, Fabio A. Gonzalez, “Histopathology Image Classification using Bag of Features and Kernel Functions”, Bioingenium Research Group, National University of Colombia, 2009.

[19] Eric Nowak, FredericJurie, Bill Triggs, “Sampling Strategies for Bag-of-Features Image Classification”, Springer-Verlag Berlin Heidelberg, ECCV Part IV, LNCS 3954, pp. 490–503, 2006.

[20] Jim Mutch, David G. Lowe, “Multiclass Object Recognition with Sparse, Localized Features”, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), 2006.