Review on Latest Approaches used in Natural Language Processing for Generation of Image Captioning

M. A. Bhalekar , Dr. M. V. Bedekar
Department of Computer Engineering, MAEER'S Maharashtra Institute of Technology, Pune, Maharashtra, India.

Abstract- *Recently the area of image captioning has received a lot of attention from researchers and academia. Image caption generation area has received attentions since the development of Deep Learning. Automatically generating caption from an image is done by integrating the domain of computer vision and natural language processing. Describing the content of an image is inherently a natural language processing and computer vision task. Many image captioning systems have shown that it is possible to describe the most salient information conveyed by images with accurate and meaningful sentences. This paper gives a survey about different recent approaches that have been used for image captioning with discussing the datasets been used.*

Index terms – *Image captioning, Computer Vision, Natural language processing, Deep Learning.*

I. INTRODUCTION

A good image description is often said to "paint a picture in your mind's eye." The creation of a mental image may play a significant role in sentence comprehension in humans. An image is not only just a collection of objects but also is very specific about the properties and relations among objects that represent the true meaning of the image. Thus, captions are a central component in image posts that communicate the background story behind photos. Captions can enhance the engagement with audiences and are therefore critical to campaigns or advertisement. Captions are a vital part of image posts in social media because they convey a richer semantic representation which can tell a story about a photo and express user's experiences including why/when/where a photo was captured.

In today's world, automatically generating caption from an image is one of the important tasks of computer vision. Automatic image captioning is the process by which a computer system automatically assigns metadata in the form of captions or keywords to a digital image. In more technical terms it can also be called as Automatic image tagging or Automatic image annotation.

This application of computer vision techniques is used in image retrieval systems to organize and locate images of interest from a database. In simple words, Automatic Image Captioning is a system which details about image. This paper presents a detail survey about the different approaches that have been used for image captioning and the datasets has been used for experimentations.

Generating of description of images is a very popular and interesting Machine Learning and Artificial Intelligence (AI) challenge. It a challenging task as it requires the understanding of the images and the translation to a sentence description. To capture the correlation between two modalities Visual and Natural language we need to map them to a common space.

To generate an image captioning system, the steps required are,

a) To understand the image,
b) Recognizing the objects in it,
c) Reasoning about the relationship among those objects , and
d) Focusing on the more salient parts of the image.

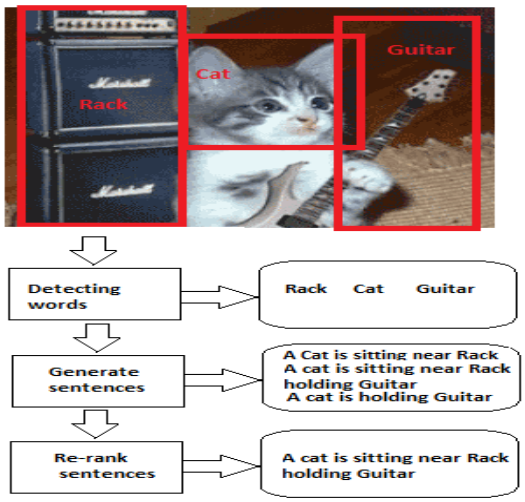

Fig. 1 General work flow of Image Captioning system

Image captioning is the task of assigning phrases to images describing their visual content. Fig. 1 shows the general work flow of a Image captioning system.

The main commonly used approaches for image captioning are: traditional approaches in which captions are assign from the most similar images to the image query. On the other hand, recent methods generate captions by sentence generation systems that learn a joint distribution of captions-images relying on a training set. The main limitation is that both approaches require a great number of manually labeled captioned images.

The most commonly used approaches for image caption generation are Convolutional Neural Network (CNN), Recurrent Neural Networks (RNN), Long Short-term Memory (LSTM). They are explained in the next section.

II. OVERVIEW OF COMMON APPROACHES

Image captioning is challenging task as it requires the understanding of images and the translation to a sentence description. Therefore a Caption Generator must identify and interpret the interplay of the different elements of a image. And must be able to coherently put this interpretation in a Target Language creating meaningful descriptions in the form of sentences.

The approach which is widely used for current Visual/Image recognition tasks is Convolutional Neural Network (CNN) a popular deep learning technique [29]. And with recent advances in Machine Translation using Recurrent Neural Networks (RNN) is used to transform a sentence written in a source language, into its translation in the target language. In machine translation, you give a sentence in a language and then the system translates it into another language. In caption generation, you give an image and then the system translates it into the image's description.

For many years, machine translation was achieved by a series of separate tasks (translating words individually, aligning words, reordering, etc), but recent work has shown that translation can be done using encoder-decoder pair in a much simpler way using Recurrent Neural Networks (RNNs) and still reach state-of-the-art performance. An encoder RNN reads the source sentence and transforms it into a rich fixed-length vector representation, which in turn in used as the initial hidden state of a decoder RNN that generates the target translated sentence.

A) *Convolutional Neural Network (Cnn)*

A Convolutional Neural Network (CNN) is a type of feed-forward artificial neural network in which the connectivity pattern between its neurons is inspired by the organization of the animal visual cortex. Individual cortical neurons respond to stimuli in a restricted region known as the receptive field. The receptive fields of different neurons partially overlap such that they tile the visual

field. The response of an individual neuron to stimuli within its receptive field can be approximated mathematically by a convolution operation. Using convolution, with minimal preprocessing on pixel image visual patterns can be directly recognize.

[29] Neural networks, as its name suggests, is a machine learning technique which is modeled after the brain structure. It comprises of a network of learning units called neurons that exchange messages between each other.

Neurons respond to different combinations of inputs and convert it to corresponding output to generates automated recognition.

The process of determining whether a picture contains a particular object involves an **activation function**. If the picture resembles prior similar images the neurons have seen before, the label would be activated. Therefore

To train the neuron it is better to provide more labeled images to neurons so that it learns to identify other unlabelled images.

The four key steps of a CNN are Convolution, Subsampling (also known as Pooling), Activation and Fully Connected as shown in Fig. 2.

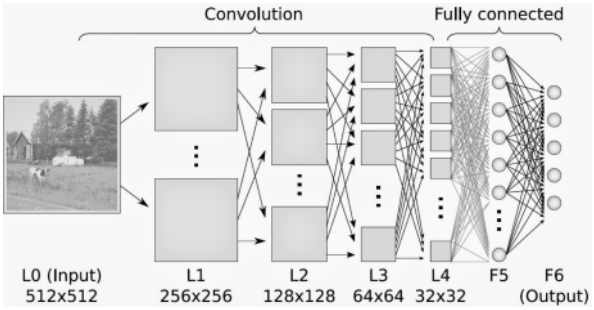

Fig. 2 Typical structure of Convolutional Neural Network [29].

CNN consist of multiple convolutional layers including subsampling step followed by one or more fully connected layers.

Step 1: Convolution

Convolution is a process where the network tries to label the input signal by referring to what it has learned in the past.

In Convolution layer, it receives the input parameters which consist of a set of learnable filters. that initially Convolution is applied on an image. A kernel (usually called as neuron.) is passed over the entire image, then

the features are extracted from the kernel and passes to the next layer.

Step 2: Subsampling

Subsampling layer is used in convulational neural networks to reduce the spatial size of the imager representation. Which further reduces the amount of features and the computational complexity of the network. It also prevent the problem of overfitting in convolutional model.

Step 3: Activation

The activation layer controls how the signal flows from one layer to the next.

[33] After each convolutional layer, it is convention to apply a nonlinear layer (**activation layer**) immediately afterward. The purpose of this layer is to introduce nonlinearity to a system that basically has just been computing linear operations during the convolutional layers. In the past, nonlinear functions like TanH and sigmoid were used, but now most common function being the *Rectified Linear Unit (ReLU)* is used, which is favored for its faster training speed.

Step 4: Fully Connected

The last layers in the network are fully connected, meaning that neurons of preceding layers are connected to every neuron in subsequent layers. This mimics high level reasoning where all possible pathways from the input to output are considered.

CNN is predominantly used in applications like - Face Recognition, Speech Recognition, Scene Labelling, Image Classification, Document Analysis etc.

B) *Recurrent Neural Network (Rnn)*

In a traditional neural network we assume that all inputs (and outputs) are independent of each other. But whenever there is need of prediction e.g. to predict what will be the next word in a sentence, it requires the knowledge of previous word. [30] The idea behind RNNs is to make use of sequential information.RNN can use in such situation where output is depends on previous computation. RNNs can use their internal memory to process arbitrary sequences of inputs.

They are networks with loops in them, allowing information to persist. These loops feed output at time *t - 1* to input at time step t of the same cell. This allows a Recurrent neural network to use it's reasoning about previous events to inform later ones. It stands as a promising solution to tackling the problem of learning sequences of information. RNNs are built on the same computational unit as the feed forward neural net, but differ in the architecture of how these neurons are connected to one another.

Fig.3 shows a typical structure of RNN which represents RNN being *unrolled* (or unfolded) into a full network.

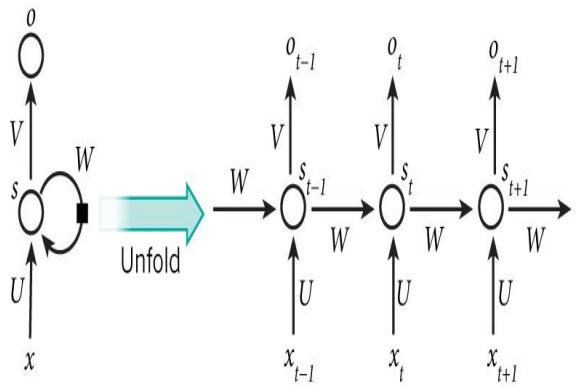

Fig. 3 Typical structure of a recurrent neural network and the unfolding in time of the computation involved in its forward computation [30].

For example, if the sequence we care about is a sentence of 5 words, the network would be unrolled into a 5-layer neural network, one layer for each word. The formulas that govern the computation happening in a RNN are as follows:

1) x_t is the input at time step *t*. For example, x_1 could be a one-hot vector corresponding to the second word of a sentence.

2) s_t is the hidden state at time step *t*. It's the "memory" of the network. s_t is calculated based on the previous hidden state and the input at the current step: $s_{t=f(U\ x_t\ +W\ s_{t-1)}}$. The function *f* usually is a nonlinearity such as tanh or ReLU. s_{t-1}, which is required to calculate the first hidden state, is typically initialized to all zeroes.

3) o_t is the output at step *t*. For example, if we wanted to predict the next word in a sentence it would be a vector of probabilities across our vocabulary. $o_{t=SOFTMAX}\ (V\ s_t)$.

Unlike a traditional deep neural network, which uses different parameters at each layer, a RNN shares the same parameters (*U,V,W* in Fig. 3) across all steps. This reflects the fact that we are performing the same task at each step, just with different inputs. This greatly reduces the total number of parameters we need to learn.

There are wide range of applications of RNN such as Language modeling and generating text, Machine Translation, Speech Recognition, Generating Image Descriptions etc.

C) ***Long Short-Term Memory (LSTM)***

Numerous researchers now use a deep learning RNN called the long short-term memory (LSTM) network [1]. It is a deep learning system that unlike traditional RNNs does not have the vanishing gradient problem (compare the section on training algorithms below). LSTM is normally augmented by recurrent gates called forget gates. LSTM RNNs prevent back-propagated errors from vanishing or exploding. Instead errors can flow backwards through unlimited numbers of virtual layers in LSTM RNNs unfolded in space. That is, LSTM can do "Very Deep Learning" tasks that require memories of events that happened thousands or even millions of discrete time steps ago. Problem-specific LSTM-like topologies can be evolved. LSTM works even when there are long delays, and it can handle signals that have a mix of low and high frequency components.

LSTMs have been used to achieve state-of-the-art performance in several tasks such as handwriting recognition, sequence generation speech recognition and machine translation [7] among others.

III. LITERATURE SURVEY

A detailed study of various approaches used in image captioning methods, was conducted with primary focus on different methods, and datasets been used.

A large body of work has been done on learning multimodal representations of images and text. Popular approaches include learning joint image-word embeddings [25, 26] as well as embedding images and sentences into a common space [27,28].

Based on the above mentioned ideas Ryan Kiros et al.[6] introduce an encoder-decoder pipeline that learns:

- a multimodal joint embedding space with images and text and
- a novel language model for decoding distributed representations.

[6] proposed system model as shown in Fig. 4 in which it make the use of : Encoder: A deep convolutional network (CNN) and long short-term memory recurrent network (LSTM) for learning a joint image-sentence embedding. And Decoder: A new neural language model that combines structure and content vectors for generating words one at a time in sequence.

It consists of the structure-content neural language model that disentangles the structure of a sentence to its content, conditioned on representations produced by the encoder. The encoder allows one to rank images and sentences while the decoder can generate novel descriptions from scratch.

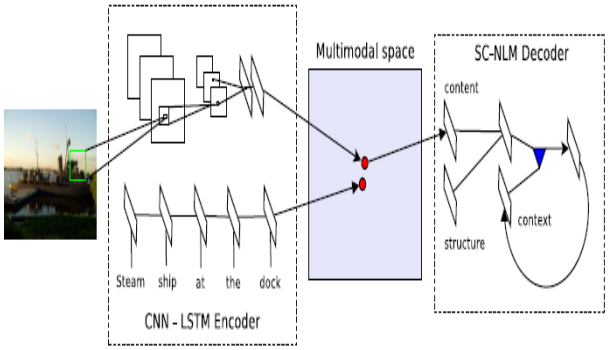

Fig. 4 Encoder and Decoder system for generating image caption [6].

[6] Gives the summary for the approaches for generating descriptions of images:

A) ***Template-based methods:*** Template-based methods involve filling in sentence templates, based on the results of object detections and spatial relationships [9,10,11]. While these approaches can produce accurate descriptions, they are often more 'robotic' in nature and do not generalize to the fluidity and naturalness of captions written by humans.

B) ***Composition based methods:*** These approaches aim to harness existing image-caption databases by extracting components of related captions and composing them together to generate novel descriptions [26,27]. The advantage of these approaches is that they allow for a much broader and more expressive class of captions that are more fluent and human-like then template-based approaches.

C) ***Neural network methods:*** These approaches aim to generate descriptions by sampling from conditional neural language models. The initial work in this area, based off of multimodal neural language models [14], generated captions by conditioning on feature vectors from the output of a deep convolutional network. These ideas were recently extended to multimodal recurrent networks with significant improvements. The methods described in this paper produce descriptions that at least qualitatively on par with current state-of-the-art composition-based methods.

The datasets use by [6] are Flickr8K [8] and Flickr30K. These datasets come with 8,000 and 31,000 images respectively with each image annotated using 5 sentences by independent annotators.

Xinlei Chen et al. [31] explore the bidirectional mapping between images and their sentence-based descriptions. Proposed RNN model is bi-directional that is, it can generate image features from sentences and sentences from image features. To evaluate its ability to do both, authors measure its performance on two retrieval tasks i.e. retrieve images given a sentence description, and retrieve a description given an image. This approach is using only RNN model not explore the use of LSTM. So in future work to achieve more impressive captioning results these RNN's model can be replace with LSTM model to learn bidirectional models.

For generating image caption requires to capture the correlation between two modalities Visual and Natural language. This is achieved by modeling a end-to-end system made of encoder-decoder pair wherein the encoder encodes the image into a vector. O. Vinyals et al.[5] proposes the system in which the decoder decodes this representation and forms the sentence description sequentially. A Convolutional Neural Network (CNN) is used as the encoder as it is used for understanding Images. A Recurrent Neural is used for Sequence to Sequence model, wherein language is generated in temporal sequence to make coherent and descriptive sentences describing the given Image. The proposed model of [5] is shown in Fig. 5 in which the unrolled connections between the LSTM memories are horizontally laid out and they correspond to the recurrent connections. They share the same parameters.

The model is trained to maximize the likelihood of the target description sentence given the training image. Authors proposed a Neural Image Caption (NIC) model which is based end-to-end on a neural network consisting of a vision CNN followed by a language generating Recurrent Neural Network (RNN). It generates complete sentences in natural language from an input image. To make the RNN more concrete authors used LSTM-based Sentence Generator model.

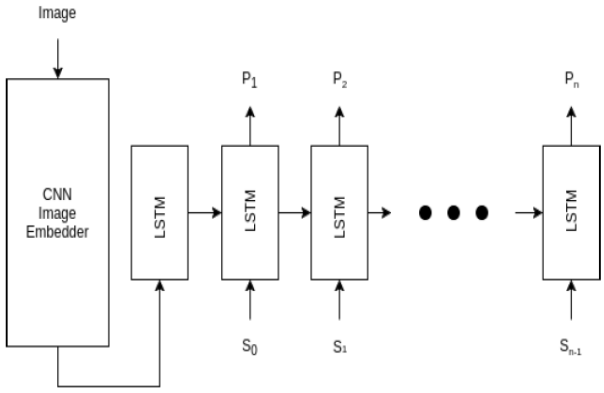

Fig. 5 LSTM model combined with a CNN image embedder and word embeddings [5].

While training the model [5] indicated of facing the problem of overfitting. By building a very complex model, it's quite easy to perfectly fit a given dataset. But when the model is evaluated on new data, it performs very poorly. In other words, the model does not generalize well. This becomes an even more significant issue in deep learning, where neural networks have large numbers of layers containing many neurons. The number of connections in these models is astronomical, reaching the millions. Purely supervised approaches require large amounts of data and are better performing, but the datasets that are of high quality have less than 100000 images which makes generalisation a problem. So to deal with this problem authors explored several techniques and they find the most obvious way to not overfit is to initialize the weights of the CNN component of their proposed system to a pretrained model (e.g., on ImageNet).

This approach NIC 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.

O. Vinyals et al.[5] propose the use of a pre-trained model called Google LeNet which won the ImageNet Large Scale Visual Recognition Competition(ILSVRC) 2014 for Classification. The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) evaluates algorithms for object detection and image classification at large scale to allow researchers to compare progress in detection across a wider variety of objects - taking advantage of the quite expensive labeling effort and to measure the progress of computer vision for large scale image indexing for retrieval and annotation.

Johnson et al.[15] propose a dense captioning task which requires a computer vision system to both localize and describe salient regions in images in natural language.

In dense captioning task each region of an image is detected and a set of descriptions is generated for each region. Hence object detection task is a special case when descriptions are one word and image captioning when detected region is the whole image. Their main contribution is the introduction of the dense localization layer. Authors introduces Fully Convolutional Localization Network (FCLN) architecture which is based on recent CNN-RNN models developed for image captioning but includes a novel, differentiable localization layer that can be inserted into any neural network to enable spatially localized predictions.

Visual Genome (VG) [22] region captions dataset been used for experimentation which comprises 94,000 images and 4,100,000 region-grounded captions. Images were taken from the intersection of

MS COCO [24] and YFCC100M [23], and annotations were collected on Amazon's Mechanical Turk [32].

Kun Fu, Junqi Jin et al.[2] proposes an image captioning system that exploits the parallel structures between images and sentences. The process of generating the next word, given the previously generated ones, is aligned with the visual perception experience where the attention shifts among the visual regions – such transitions impose a thread of ordering in visual perception. This alignment characterizes the flow of latent meaning, which encodes what is semantically shared by both the visual scene and the text description. It also proposes a novel model by introducing scene-specific contexts that capture higher-level semantic information encoded in an image. The contexts adapt language models for word generation to specific scene types.

The system for image captioning is composed of the following components: visual feature representation of the image with localized regions at multiple scales, an LSTM-based neural network that models the attention dynamics of focusing on those regions as well as generating sequentially the words, And a visual scene model that adjusts the LSTM to specific scenes.

Girish Kulkarni et al.[20] proposed system to automatically generate natural language descriptions from images. It consists of the following steps:

1. Felzenszwalb et al. [17] detectors are used to detect objects like things (e.g., bird, bus, car, person, etc.) and To detect stuff (e.g., grass, trees, water, road, etc.) classifiers are trained to detect regions corresponding to non-part-based object categories using linear Support Vector Machine (SVM).
2. Each candidate object (either thing or stuff) region is processed by a set of attribute classifiers. An Radial Basis Function (RBF) kernel SVM is used to learn a classifier for each visual attribute term
3. Each pair of candidate regions is processed by prepositional relationship functions.
4. A conditional random field (CRF) is constructed to predict a labeling for an input image and base on the labeling Sentences are generated.

The nodes of the CRF correspond to several kinds of image content:

a) Objects - things or stuff
b) Attributes which modify the appearance of an object
c) Prepositions which refer to spatial relationships between pairs of objects.

These potentials are the probability of various attributes for each object (given the object) and the probabilities of particular prepositional relationships between object pairs (given the pair of objects). The conditional probabilities are computed from counts of word co-occurrence. The output of the CRF is a predicted labeling of the image. This labeling encodes three kinds of information: objects present in the image (nouns), visual attributes of those objects (modifiers), and spatial relationships between objects (prepositions). UIUC PASCAL [21] sentence dataset was used for Evaluation, which contains up to five human-generated sentences that describe 1000 images. From this set results were evaluated on 847 images.

A passenger aircraft with landing gear down.
A passenger jet flies through the air.
A passenger plane fly through the sky.
The Austrian plane soars in the sky.
The white jet with its landing gear down flies in the blue sky.

(a)

A girl on a chopper bicycle.
A girl with a helmet on a bicycle near a river.
A girl with a helmet riding a bicycle.
The girl is riding her bicycle down the road.
Young girl riding bicycle on bike path.

(b)

Fig. 6 (a) and (b) Sample Images from the dataset PASCAL sentences –UIUC

Andrej Karpathy et. al. [3] focus mainly on a model that generates natural language descriptions of

images and their regions. It informs about a novel combination of Convolutional Neural Networks over image regions, Bidirectional Recurrent Neural Networks (BRNN) over sentences, and a structured objective that aligns the two modalities through a multimodal embedding for caption generation. Provided approach strive to take a step towards the goal of generating dense descriptions of images. They develop a deep neural network model that infers the latent alignment between segments of sentences and the region of the image that they describe with multimodal Recurrent Neural Network architecture which generates its description in text. One contribution of proposed system [3] is a model that aligns the process of generating captions and the attention shifting among the visual regions. Another advantage is to introduce the scene-specific contexts to LSTM that adapts language models for word generation to specific scene types.

The image captioning datasets used by the authors for experiments are the Flickr8K [8], Flickr30K and MSCOCO [24]. These datasets contain 8,000, 31,000 and 123,000 images respectively and each is annotated with 5 sentences using Amazon Mechanical Turk (AMT)[32]. BRNN approach features a novel ranking model that aligned parts of visual and language modalities through a common, multimodal embedding.

Farhadi et al. [10] suggested that there is a space of Meanings that comes between the space of Sentences and the space of Images. The similarity between a sentence and an image is then evaluated by mapping each to the meaning space and then comparing the results. The intermediate space learns the mapping from images (respective sentences) to meaning discriminatively from pairs of images and assigned meaning representations. Mapping Image to Meaning is achieved by presentation of meaning in the form of a triplet as <object, action, scene>.

It provides a holistic idea about what the image is about and what is most important. The mapping from images to meaning is reduced to learning to predict triplet for images. The problem of predicting a triplet from an image involves solving a (small) multi-label Markov random field. The potentials to predict the triplets are computed as linear combinations of feature functions.

Image features are constructed using: Felzenszwalb et al.[17] detector to predict the confidence scores on all the images. A threshold is set such that all of the classes get predicted, at least once in each image and then the authors consider the max confidence of the detections for each category, the location of the center of the detected bounding box, the aspect ratio of the bounding box, and it's scale.

Hoiem et al.[18] classification responses classify based on geometry, Histograms of Oriented Gradients (HOG) features, and detection responses and Gist-based scene classification responses[19] where the Global information of images is encoded using gist. Node Features are built by fitting a discriminative classifier (a linear SVM) to predict each of the nodes independently on the image features. Although the classifiers are being learned independently, they are well aware of other objects and scene information. .K-nearest neighbor is then used to match image features.

Sentence representation is done by computing the similarity between the sentence and triplets. For this to done a notion of similarity for objects, scenes and actions in text authors used Curran & Clark parser [16] to generate a dependency parse for each sentence. These dependencies were used to generate the (object, action) pairs for the sentences.

For experimentation image caption data set consisting of roughly 8,000 images is taken from six Flickr groups. PASCAL Sentence data set [21], and Amazon's Mechanical Turk to generate five captions for each image.

IV. CONCLUSION

Generating description of images is a very popular and interesting Machine Learning and Artificial Intelligence challenge. It's a challenging task as it requires the understanding of the images and then the translation to a sentence description. This can be achieved by using Computer Vision and Natural Language Processing domain.

By analyzing the different approaches we can conclude that for object recognition CNN method can be used further with RNN and LSTM as machine translation. Neural network methods give much better descriptions as compared to Template based methods and Composition based methods.

Even Convolutional Neural Networks over image regions and Bidirectional Recurrent Neural Networks (BRNN) over sentences, can be used as modalities through a multimodal embedding for caption generation. Another approach followed is of Fully Convolutional Localization Network (FCLN) architecture which is based on recent CNN-RNN models developed for image captioning includes a novel, differentiable localization layer that can be inserted into any neural network to enable spatially localized predictions.

LSTM model can be uses as a bidirectional model to map between images and their sentence based descriptions to achieve more impressive captioning results.

REFERENCES

[1] Sepp Hochreiter and Jürgen Schmidhuber. "Long short-term memory", Neural computation, 9(8):1735–1780, 1997.

[2] Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, Changshui Zhang, "Aligning Where to See and What to Tell: Image Captioning with Region-based Attention and Scene-specific Contexts" IEEE Transaction on Pattern analysis and machine intelligence, 2016.

[3] Andrej Karpathy, Li Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions", in journal of Latex class files, Vol. 14, No. 8, August 2015

[4] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S-C. Zhu. "I2T: Image parsing to text description", Proceedings of the IEEE, 98(8), 2010.

[5] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator", Computer Vision and Pattern Recognition, 2015.

[6] Ryan Kiros, Ruslan Salakhutdinov, Richard S. Zemel, "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models", arXiv:1411.2539v1 [cs.LG]10 Nov 2014,

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks", NIPS, 2014.

[8] M. Hodosh, P. Young, and J. Hockenmaier. "Framing image description as a ranking task: data, models and evaluation metrics", Journal of Artificial Intelligence Research, 2013.

[9] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg, "Baby talk: Understanding and generating simple image descriptions", CVPR, 2011.

[10] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth, "Every picture tells a story: Generating sentences from images", In ECCV, 2010.

[11] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi, "Composing simple image descriptions using web-scale grams", In CONLL, 2011.

[12] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi, "Collective generation of natural image descriptions", ACL, 2012.

[13] Polina Kuznetsova, Vicente Ordonez, Tamara L. Berg, and Yejin Choi, "Treetalk : Composition and compression of trees for image descriptions", TACL, 2014.

[14] Ryan Kiros, Richard S Zemel, and Ruslan Salakhutdinov., "Multimodal neural language models", ICML, 2014.

[15] Justin Johnson, Andrej Karpathy, and Li Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning" InarXiv:1511.07571, 2015.

[16] Curran, J., Clark, S., Bos, J, "Linguistically motivated large-scale NLP with C&C and Boxer", Proceedings of the ACL 2007 Demo and Poster Sessions, pages 33–36, 2007.

[17] Felzenszwalb, P, Mcallester, D, Ramanan, D., "A discriminatively trained, multi-scale, deformable part model", in CVPR, 2008.

[18] Hoiem, D., Divvala, S., Hays, J., " Pascal voc 2009 challenge. In: PASCAL challenge workshop" in ECCV, 2009.

[19] Oliva, A., Torralba, " Building the Gist of a scene: the role of global image features in recognition" In: Progress in Brain Research, 2006.

[20] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Alexander C. Berg, Tamara L. Berg' "BabyTalk: Understanding and Generating Simple Image Descriptions", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 35, No. 12, December 2013

[21] http://vision.cs.uiuc.edu/pascal-sentences [PASCAL sentences – UIUC: Available data Set]

[22] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei., "Visual genome: Connecting language and vision using crowd sourced dense image annotations", 2016.

[23] B. Thomee, B. Elizalde, D. A. Shamma, K. Ni, G. Friedland, D. Poland, D. Borth, and L.-J. Li. "YFCC100M: The new data in multimedia research", Communications of the ACM, 59(2):64–73, 2016.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context." arXiv preprint arXiv:1405.0312, 2014.

[25] Jason Weston, Samy Bengio, and Nicolas Usunier. "Large scale image annotation: learning to rank with joint word-image embeddings", Machine learning, 2010.

[26] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, and Tomas Mikolov MarcAurelio Ranzato. Devise: A deep visual-semantic embedding model. NIPS, 2013.

[27] Richard Socher, Q Le, C Manning, and A. Ng. "Grounded compositional semantics for finding and describing images with sentences", In TACL, 2014.

[28] Andrej Karpathy, Armand Joulin, and Li Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping", NIPS, 2014.

[29] https://algobeans.com

[30] http://www.wildml.com/2015/09/recurrent neural-networks-tutorial-part-1introduction-tornns/

[31] Xinlei Chen, C. Lawrence Zitnick, "Mind's Eye: A Recurrent Visual Representation for Image Caption Generation", The 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[32] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting Image Annotations Using Amazon's Mechanical Turk," Proc. NAACL HLT Workshop Creating Speech and Language Data with Amazon's Mechanical Turk, 2010.

[33] http://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks-Part-2/