**Abstract**

Automation of caption i.e. creating the language description of an image using any natural language processing is a diffiult task. It requires to have a good grip in both image processing as well as natural language processing. This report discusses about different available techniques which models for the image captioning. the advancement of technolgy in terms of object recognition and machine learning has very much improved the performance of image captioning model in recent years. In addition to that we will try to approach the different techiques and algorithms to build this model . At the end, model evaluation play an important role which will be useful to consider which algorithm suits best in the scenario. There are centain techniques which can be used to evaluate the performance such as confusion matrix, f1-score but in this case the data is based on natural language so based on that there are certain special type of techniques which can be used such as “Bilingual Evaluation Understudy Score” (BLEU Score), Microsoft COCO and Flickr30K.

**Literature Survey**

[1] Parth Shah, Vishvajit Bakrola, Supriya Pati, “**Image Captioning using Deep Neural Architectures**” in proceedings of “2017 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS)”

**Summary**

This paper discusses uses Deep Neural Network Architecture to generate a reasonable and human readable description of an Image or we can say a caption of an image. The architecture uses encoder and decoder techniques, in which ecoder is a part where the feature extraction takes place and the output of encoder act as an input to the decoder which is nothing but LSTM’s neural network .Instead of using text as input to encoder, Show & Tell model uses image as an input. This image is then converted to word vector and then word vector is translated to caption using recurrent neural networks as decoder after the building of model , it is further evaluated by evaluation matrix called as “Bilingual Evaluation Understudy Score” used for any model that generates natural language sentence ( BLEU Score).Ther more the score the better is the perception with respect to human interpretation, the paper acheived the BLEU average score of 65.5, which is considered as a good score.

**Contributions**

**Algorithms:** The algorithms used in the paper are deep neural networks and language processing techinques which is: CNN, LSTM, NLP

**Benefits:** The deep understanding if supervised deep learning algorithms with respect to natural language processing and understanding of its applications. The Image captioning can generate short summary of any image which can be useful for event handeling and event sumarization.

**Disadvantage:** Bad quality of data may lead to bad classifier which will be useless and It may not be able to interpret all the details in the image insteed of focusing on main regions.

[2] Quanzeng You , Hailin Jin , Zhaowen Wang , Chen Fang , and Jiebo Luo ,“ **Image Captioning with Semantic Attention** ” in proceedings of “2016 IEEE Conference on Computer Vision and Pattern Recognition”

**Summary**

In this Paper, a method for the job of image captioning introduced, which achieves good performance comparatively to standard benchmarks. Different from previous work,this method combines top-down and bottom-up strategies to extract important information from an image, and ccombines them with a RNN that can selectively attend on semantic attributes detected from the image. This method, therefore, exploits not only an overview under- standing of input image, but also abundant fine-grain visual semantic aspects. The real power of model lies in its ability to attend on these aspects and seamlessly fuse global and local information for better caption.

evaluation of algorithm are done on two public benchmarks: Microsoft COCO and

Flickr30K. Experimental results show that our algorithm significantly outperforms the state-of-the-art approaches consistently across different evaluation metrics.

**Contributions**

**Algorithms:** The algorithms used in the papers are: CNN, RNN, Semantic model, Top-Down Approach, Bottom-Up Approach

**Benefits**: It use CNN to extract a top-down visual feature and at the same time detects visual concepts (regions, objects, attributes, etc.). it employ a semantic attention model to combine the visual feature with visual concepts in a RNN that generates the image caption, which reduces the chances wrong interpretation.

**Disadvantage:**

One of the limitations of the top-down paradigm is that it is hard to attend to fine details which may be important in terms of describing the image.

**Correlating Theoretical Concept To Practical Implementation**

**NLP**

Natural language processing is a subfield of computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human languages, in particular how to program computers to process and analyze large amounts of natural language data.

There are sevaral main techniques used in analysing natural language processing. Some of them can be breafly described as follows.

**Pattern matching**

The idea here is an approach to natural language processing is to interpret input utterances as a whole father than builing up their interpretation by combining the structure and meaning of words or other lower level constituents. That means the interpretations are obtained by matching patterns of words against the input utterance. For a deep level of analysis in pattern matching a large number of patterns are required even for a restricted domain. This problem can be ameliorated by hierarchical pattern matching in which the input is gradually canonicalized through pattern matching against subphrases. Another way to reduce the number of patterns is by matching with semantic primitives instead of words.

**Syntactically driven Parsing**

Syntax means ways that words can fit together to form higher level units such as phrases, clauses and sentences. Therefore syntacticaly driven parsing means interpretation of larger groups of words are built up out of the interpretation of their syntacticconstituent words or phrases. In a way this is the opposite of pattern matching as here the interpretation of the input is done as a whole. Syntactic analyses are obtained by application of a grammar that determines what sentenses are legal in the language that is being parsed.

**Semantic Grammars**

Natural language analysis based on semantic grammar is bit similar to systactically driven parsing except that in semantic grammar the catogaries used are defined semantically and syntactically. There here semantic grammar is also envolved.

**Case frame instantiation**

case frame instantiation is one of the major parsing techniques under active research today. The has some very useful computational properties such as its recursive nature and its ability to combine bottom-up recognition of key constituents with top-down instantiation of less structured constituents.

**Computer Vision**

Computer vision is an interdisciplinary field that deals with how computers can be made to gain high-level understanding from digital images or videos. From the perspective of engineering, it seeks to automate tasks that the human visual system can do.

Computer vision tasks include methods for acquiring, processing, analyzing and understanding digital images, and extraction of high-dimensional data from the real world in order to produce numerical or symbolic information, *e.g.*, in the forms of decisions.Understanding in this context means the transformation of visual images (the input of the retina) into descriptions of the world that can interface with other thought processes and elicit appropriate action. This image understanding can be seen as the disentangling of symbolic information from image data using models constructed with the aid of geometry, physics, statistics, and learning theory.

**Deep Learning**

Deep learning is a class of machine learning algorithms that:

* use a cascade of multiple layers of nonlinear processing units for feature extraction and transformation. Each successive layer uses the output from the previous layer as input.
* learn in supervised (e.g., classification) and/or unsupervised (e.g., pattern analysis) manners.
* learn multiple levels of representations that correspond to different levels of abstraction; the levels form a hierarchy of concepts.

Deep Learnng is also classified in two different types of categories such as :

* Supervised Classification
* Unsupervised classificatio

**Supervised Classifications**

is based on the idea that a user can select sample pixels in an image that are representative of specific classes and then direct the image processing software to use these training sites as references for the classification of all other pixels in the image. Training sites (also known as testing sets or input classes) are selected based on the knowledge of the user. The user also sets the bounds for how similar other pixels must be to group them together. These bounds are often set based on the spectral characteristics of the training area, plus or minus a certain increment (often based on "brightness" or strength of reflection in specific spectral bands). The user also designates the number of classes that the image is classified into. Many analysts use a combination of supervised and unsupervised classification processes to develop final output analysis and classified maps.

E.g. ANN, CNN, RNN

**Unsupervsed Clasifications**

Is where the outcomes (groupings of pixels with common characteristics) are based on the software analysis of an image without the user providing sample classes. The computer uses techniques to determine which pixels are related and groups them into classes. The user can specify which algorism the software will use and the desired number of output classes but otherwise does not aid in the classification process. However, the user must have knowledge of the area being classified when the groupings of pixels with common characteristics produced by the computer have to be related to actual features on the ground (such as wetlands, developed areas, coniferous forests, etc.).

E.g. AutoEncoders, Boltzman Machines, PCA

**Flow diagram/Design of the system and Implementation**

# **Recurrent Neural Networks**

Recurrent Neural Networks are the supervised type of Neural Networn whch was first developed in 1980’s and just been recently becoming popular for its advances to the NN designs and improved computational power from GPU’s. They are especially useful with related to the sequential data such as time series analysis, because each perceptron or unit can use its internal memory to maintain information about the previous input. This is really helpful because in cases of language, “I had washed my house” is much more different than “I had my house washed”. This allows the network to gain a deeper understanding of the statement.

Basic architecture of RNN with single perceptron



A RNN has loops in them that allow infromation to be carried across neurons while reading in input.

The above loop can be expanded as below ,



In these diagrams Xt is some input, A is a part of the RNN and Ht is the output. Essentially you can feed in words from the sentence or even characters from a string as Xt and through the RNN it will come up with a Ht.

The goal is to use Ht as output and compare it to your test data (which is usually a small subset of the original data). You will then get your error rate. After comparing your output to your test data, with error rate in hand, you can use a technique called Back Propagation Through Time (BPTT). BPTT back checks through the network and adjusts the weights based on your error rate. This adjusts the network and makes it learn to do better.

**Long Short-Term Memory**

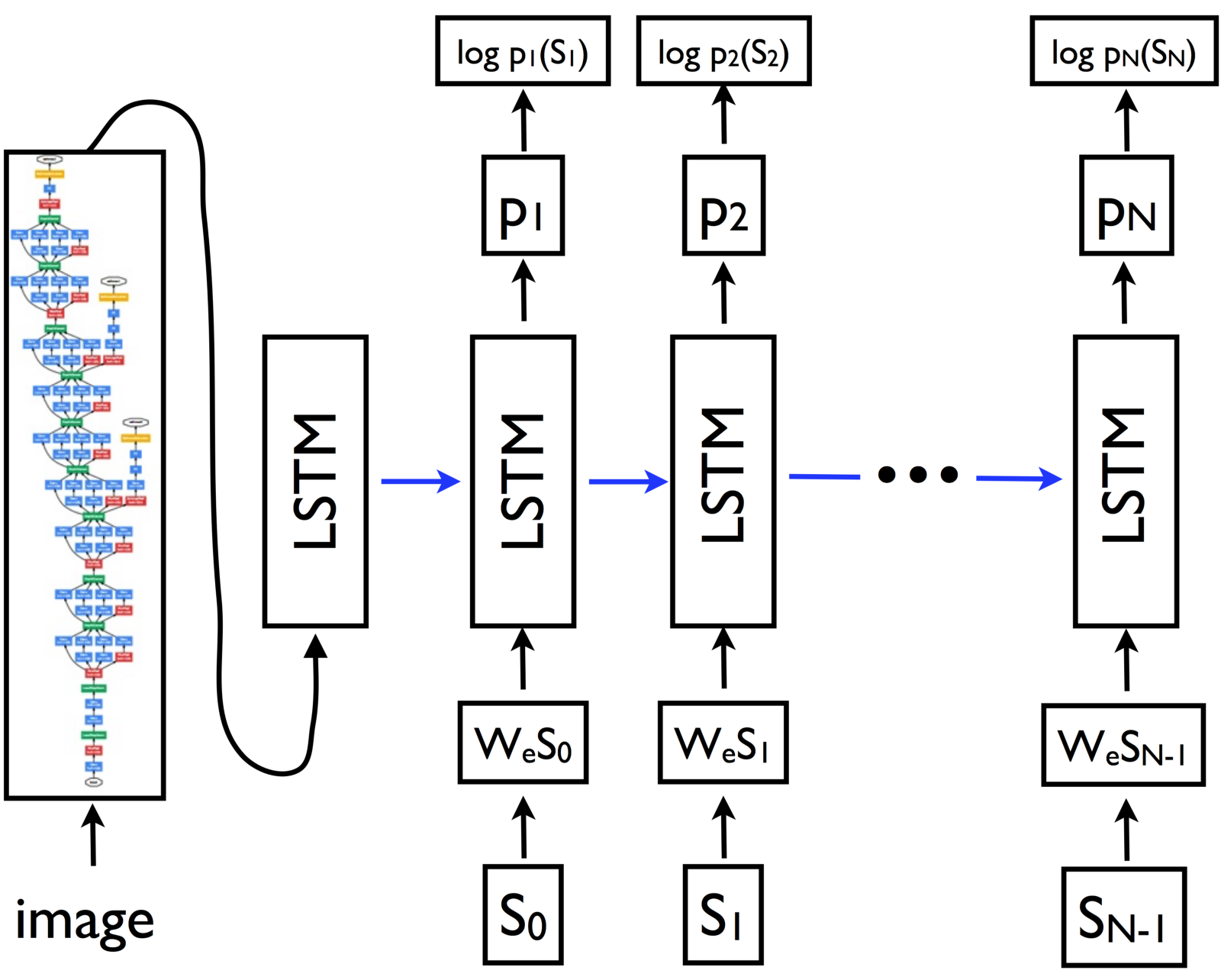
Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. They were introduced by [Hochreiter & Schmidhuber (1997)](http://www.bioinf.jku.at/publications/older/2604.pdf), and were refined and popularized by many people in following work.[1](http://colah.github.io/posts/2015-08-Understanding-LSTMs/" \l "fn1) They work tremendously well on a large variety of problems, and are now widely used.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the form of a chain of repeating modules of neural network. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

Structure of LSTM’s

**Image Captioning System Design and Implementations**

****

#### **Encoder**

The

Convolutional Neural Network(CNN) can be thought of as an encoder. The input image is given to CNN to extract the features. The last hidden state of the CNN is connected to the Decoder.

#### Decoder

The Decoder is a Recurrent Neural Network(RNN) which does language modelling up to the word level. The first time step receives the encoded output from the encoder and also the START vector.

### Training

The output from the last hidden state of the CNN(Encoder) is given to the first time step of the decoder. We set X1 = START vector and the desired label H1 = first word in the sequence. Analogously, we set X2 =word vector of the first word and expect the network to predict the second word. Finally, on the last step, XT = last word, the target label HT = END token.

During training, the correct input is given to the decoder at every time-step, even if the decoder made a mistake before.

### Testing

The image representation is provided to the first time step of the decoder. Set X1 = START vector and compute the distribution over the first word H1. We sample a word from the distribution (or pick the argmax), set its embedding vector as X2, and repeat this process until the END token is generated.

During Testing, the output of the decoder at time t is fed back and becomes the input of the decoder at time t+1