****

**PES University, Bangalore**

**(Established under Karnataka Act No. 16 of 2013)**

**Department of Computer Science and Engineering**

**(Jan – May 2022)**

**UE19CS343 – TOPICS IN DEEP LEARNING**

**(ELECTIVE-4 FOR 6TH SEMESTER)**

PROJECT REPORT

On

**“IMAGE CAPTION GENERATOR”**

Under Professor Asha J

**Team - 10**

**Name SRN**

**Akhil PES2UG19CS025**

**Bhavana R PES2UG19CS089**

**Chandra Snehal PES2UG19CS394**

**Problem statement**

IMAGE CAPTION GENERATOR

# **Introduction**

Image caption generator teaches a machine to accurately describe an image or surroundings in the same way that a human does.

They can be used for automatic image indexing. Image indexing is important for Content-Based Image Retrieval and therefore, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching

# **Literature Survey**

In the last 5 years, a large number of articles have been published on image captioning with deep machine learning being popularly used. Convolutional Neural Networks (CNN) are widely used for feature learning, and a classifier such as Softmax is used for classification. CNN is generally followed by Recurrent Neural Networks (RNN) or Long Short-Term Memory Networks (LSTM) in order to generate captions.

In the visual space-based methods, the image features and the corresponding captions are independently passed to the language decoder.

In Multimodal Space, The vision part uses a deep convolutional neural network as a feature extractor to extract the image features. The language encoder part extracts the word features and learns a dense feature for each word. It then forwards the

semantic temporal context to the recurrent layers. The multimodal space part maps the image features into a common space with the word features.

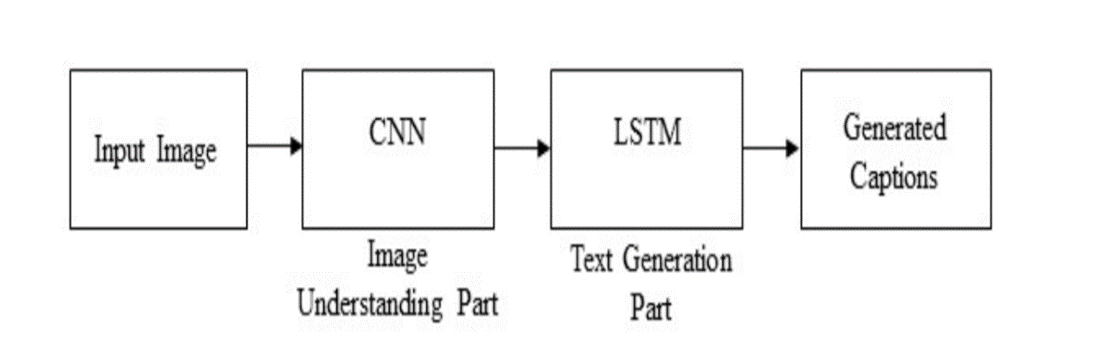
Supervised learning-based networks have successfully been used for many years in image classification, object detection, and attribute learning. This progress makes researchers interested in using them in automatic image captioning.

Recently, researchers are focusing more on reinforcement learning and unsupervised learning-based techniques for image captioning.

ENCODER-DECODER ARCHITECTURE-BASED IMAGE CAPTIONING, In this network, global image features are extracted from the hidden activations of CNN and then fed into an LSTM to generate a sequence of words.

A CNN is used to obtain the scene type, and to detect the objects and their relationships.

The output of this is used by a language model to convert them into words, combined phrases that produce image captions.

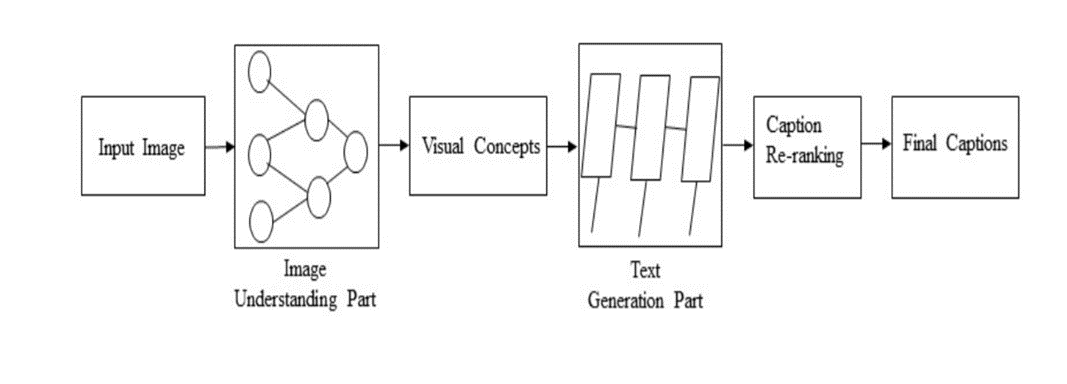


COMPOSITIONAL ARCHITECTURE, Image features are obtained using a CNN.

Visual concepts (e.g. attributes) are obtained from visual features.

Multiple captions are generated by a language model using the information of CNN.

The generated captions are re-ranked using a deep multimodal similarity model to select high-quality image captions.



In recent years, LSTM-based models have dominantly been used in sequence to sequence learning tasks.LSTM networks are a type of RNN that has special units in addition to standard units. LSTMs ignore the underlying hierarchical structure of a sentence. They also require significant storage due to long-term dependencies through a memory cell.

Another network, Gated Recurrent Unit (GRU) has a similar structure to LSTM but it does not use separate memory cells and uses fewer gates to control the flow of information.

recently, convolutional architectures are used in another sequence to sequence tasks.

(Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An empirical study of language CNN for image captioning. In Proceedings of the International Conference on Computer Vision (ICCV). 1231–1240) proposed a CNN language

model-based image captioning method. This method uses a language-CNN for statistical language modeling.

(Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570). network with the language CNN to model the temporal dependencies properly. Aneja et al. [5] proposed a convolutional architecture for the task of image captioning. They use a feed-forward network without any recurrent function.

(Qingzhong Wang and Antoni B Chan. 2018. CNN+ CNN: Convolutional Decoders for Image Captioning. arXiv preprint arXiv:1805.09019.) proposed another CNN+CNN-based image captioning method. it uses a hierarchical attention module to connect the vision-CNN with the language-CNN.

The authors of this method also investigate the use of various hyperparameters, including the number of layers and the kernel width of the language-CNN.

They show that the influence of hyperparameters can improve the performance of the method in image captioning.

# 

# 

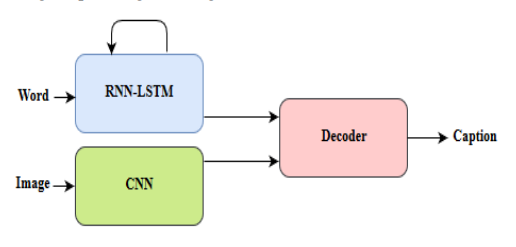
# 

# **Design**

For designing this project, we have used a VGG16 model, a CNN, and an LSTM model. The CNN-LSTM architecture involves using CNN layers for feature extraction on input data combined with LSTMs to support sequence prediction.

Feature extraction is a mandatory step to train any image in deep learning. We do this feature extraction using CNN. CNN is the image-based model.

And we use a language-based model, which translates the features and objects extracted by our image based model to a natural sentence. This is achieved by LSTM.



# **Implementation**

The image caption generator uses a pre-trained CNN model, namely VGG16, and a normal LSTM to generate captions. After conducting a thorough literature survey, the preferred pre-trained model was decided as VGG16.

The Flickr8k data set consists of images and their related captions. The images are fed to VGG16 for feature extraction. The end 2 layers of the pre-trained model have been removed as we wish to use our own classification and dense layer.

The related captions are tokenized and used to construct a vocabulary which is then fed into an LSTM.

An encoder-decoder architecture is being used, where the output of the CNN and the LSTM vector is fixed to a particular size, and a decoder is used to generate the appropriate captions for the image.

In order to run the above model a workstation having around 32GB - 64GB of RAM is required, since the amount of available RAM was not more than 8GB, the Progressive Loading method was used to run the model. A data generator function is used which yields one photo’s worth of data per batch. The trade-off of using a data generator to save RAM caused the code to execute much slower.

# **Results**

After training the CNN-LSTM model on the given dataset, the model outputs a series of captions that describe the test images satisfactorily. It is observed that a few captions generated for some images may not fully accurately fit the image, this is due to the relatively small dataset that has been used to train the model.

*BLEU* is the most widely used evaluation indicator. It is used to analyze the correlation of n-gram between the translation statement to be evaluated and the reference translation statement. The higher the BLEU score, the better the performance.

# **Conclusion and Future Scope**

Visuals and imagery continue to dominate social and professional interactions globally. With a growing scale, manual efforts are falling short on tracking, identifying, and annotating the prodigious amounts of visual data. With the advent of artificial intelligence, multimedia businesses are able to accelerate the process of image captioning while generating significant value.

Once the model is trained in a sufficiently big data set such as Flickr16k/32k we begin to observe much more accurate captions being generated for images. This forms a strong core for building many computer vision-based projects and can further be utilized in the following fields.

### **Recommendations in Editing Applications-**The image captioning model automates and accelerates the close captioning process for digital content production, editing, delivery, and archival. Well-trained models replace manual efforts for generating quality captions for images as well as videos.

### **Assistance for Visually Impaired-**The advent of machine learning solutions like image captioning is a boon for visually impaired people who are unable to comprehend visuals. With an AI-powered image caption generator, image descriptions can be read out to the visually impaired, enabling them to get a better sense of their surroundings.

### **Media and Publishing Houses-**The media and public relations industry circulates tens of thousands of visual data across borders in the form of newsletters, emails, etc. The image captioning model accelerates subtitle creation and enables executives to focus on more important tasks.

### **Social Media Posts-**For social media, artificial intelligence is moving from discussion rooms to underlying mechanisms for identifying and describing terabytes of media files. It enables community administrators to monitor interactions and analysts formulate business strategies.

For future work, we propose the following four possible improvements:

* An image is often rich in content. The model should be able to generate description sentences corresponding to multiple main objects for images with multiple target objects, instead of just describing a single target object.
* For corpus description languages of different languages, a general image description system capable of handling multiple languages should be developed.
* A very real problem is the speed of training, testing, and generating sentences for the model should be optimized to improve performance.