IMAGE CAPTIONING USING DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs)

G.Geetha*, T.Kirthigadevi, G.Godwin Ponsam, T.Karthik and M.Safa

Department of Information Technology, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India 603203.

Corresponding author e-mail: *geethag@srmist.edu.in

Abstract. Earth is challenging to label satellite image clips with atmospheric conditions and various classes of land cover and land use. We proposed an algorithms to help the global community for a better understanding that where, how, and why deforestation take place all over the world. Upcoming development in satellite imaging technology have set grow to new opportunities for more precise investigation of both broad and minute changes occurring on Earth, including deforestation. Since 40 years, almost a fifth of the Amazon rain forest has been cut down. To estimate and analysis the forest this application is developed. Satellite images are trained on deep convolutional neural networks (CNNs) to learn image features and used multiple classification frameworks including gate recurrent unit captioning and sparse_cross_entropy to predict multiclass, multi-label images. By fine-tuning an architecture consisting of the encoder of pre-trained VGG-19 parameters trained on ImageNet data together with the GRU decoder.

1. Introduction
Labeling the satellite picture with atmospheric conditions and various captions of land cover or land use is challenging. The results of used algorithms will enable the worldwide community for a better understanding of what, how, and why deforestation is happening everywhere over the globe - and the ultimate way to reply. Furthermore, existing methods generally can’t differentiate between man causes of forest loss and natural one. Higher resolution imagery has already been shown to be exceptionally good at this, but robust methods haven't yet been developed for Planet imagery. To overcome this problem our aim is developing a combination of CNN and RNN algorithm encoder decoder architecture to caption these satellite images. The data images were carried out from Earth’s full frame analytic scene products using 4 class satellites in sun synchronously orbit and International artificial satellite orbit. Each contains a few bands of information: green, red, blue, infrared and therefore the set of chips for this project uses an actual pattern. The precise spectral responses of the satellites used for images are found within the Planet documentation.

Each of those channels is in a 16-bit digital number format that meets the specification of the world. An inventory of training file names and their labels, the labels are space-delimited.

The captions can be divided into three types of categories:

- Atmospherically change.(conditions)
- Commons shelter or terrestrial use phenomena
- Care land cover or land use phenomena
High resolution of images have already shown the proof of exceptionally better performance at this, but the robust methodologies haven't yet been developed for earth imagery. Overcoming this the problem our aim is developing a combination of algorithm, encoder decoder architecture to caption these satellite images.

1. Review the info, which has detailed information about the labels and therefore the labeling process.
2. Downloading a sub-sample of the info to urge acquainted with how it's. Explore the sub-sample using python and exploratory data analytics.
3. Motivated by the burgeoning commercial and research interest in satellite images of Earth, we developed various models that are able to efficiently and accurately distinguish the content of such images.

Specifically, we trained deep convolutional neural networks (CNNs) to find out image features and used multiple classification frameworks including long short-term memory (LSTM) label captioning and binary cross entropy to predict multi-class, multi-label images.

2. Literature Study

Results and Implications of a study of 15 years of spatial picture distinction experiments” [1]. The effort of this paper promotes the distinction of images along with the goal of creating high-qualities thematic maps with accurate creation of satellite image class. Few researches have pressed upon the betterment of the distinguishing process, another one is on the verge of using famous distinction architecture in certain kinds of remote sensing fields. The distinction is regarding a basic structure in remote sensing, that is found to be at the depth of conversion to spatial image classification.

Spatial Picture Classification functions and techniques: A Review. Global journal of computer applications” [2]. This paper focuses the spatial image distinction process that includes combining the pixel attributes of images to an appropriate class. Various picture distinction ideology or methods are present. According to this paper, the spatial image distinction functions are widely distinguished into 3 categories 1) hybrid 2) manual and 3) automatic. Widely the spatial picture class functions come under 1st class. Image distinction demands to choose the certain distinction criteria made on the needs. This paper is the field that consists a study on spatial image classification methods.

Supervised the distinction of spatial pictures. Conference on Advances in Signal Processing” [3]. Research in this paper focuses on the process of producing thematic from remote sensing of imagery for distinguishing images. Spectral bands non-analog integers are made to show spectral data. The data is made for non-analog distinguishing of pictures. In this paper, each pixel is distinguished through this spectral-data. Supervised and unsupervised are used for distinguishing images. This particular paper deals with the machine learning supervised distinguish mainly support vector machine, minimum distance, parallelepiped, and maximum likelihood.

ANN distinction using a minimal training set. Comparison to conventional supervised classification. Photogrammetric Engineering and Remote Sensing.”[4]. This paper deals with the strength of applying to NN computation to spatial image processing. The other AIM is to give a primary connecting of learning data in and normalize land area distinction outputs for conventional supervised and artificial neural net classes. ANN is trained to do land area classification of spatial clips of every dominant in the same way of supervised algorithms. This research is the base for creating applying weights for the future idea of software implications for ANN in the spatial image, earthly data preparation.

Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering” [5]. In this paper, they propose a noble technique for unsupervised algorithms to detect changes in multitemporal spatial images. They use PCA and k means clustering. here, the different images are parted in different times non-overlapping partitions. In this every pixel in the different picture is presented on a few-dimensional features array that is a project image data on the created Eigen vector area. The difference is acquired by the partition of the features array space into the different unsupervised clusters using the k-means clustering technique with k value is two, after
which assignment of every pixel-value to 1 of the two k mean clusters by distance formula called Euclidean method.

ArcGIS. What Is Image Classification? ArcGIS 10.5 Help Site” [6]. This is software that is a full combination of needs in the multivariate to do supervised and unsupervised distinction. The distinction process is a work flow; the image distinction toolbar is created to give a suitable area to do classifications. These tools help with the flow for doing unsupervised and supervised distinction.

Multi label text distinction with a mixing model trained by electronic machine “[7]. This paper focuses on a Bayesian classification where the multi-classes that consist of information are presented by the mixing model. The supervised learning info shows which classes results for creating data, it could not indicate which classes were results for creating every word. Therefore we use electronic machines to complete this missing data, learning of both the distribution over combination parameters and word is distributed in every section's mixture part. They describe the advantages of this model and the current primary outputs.

A a unified framework for multi-label image classification” [8]. in this paper, they have utilized recurrent neural network to deal with captioning problem and combined it with CNN, the CNN cum RNN frame trained over a joint image and its labels embedded to characterize the relation of non-independency as well as the pics output relevance and it can be learned end to end from basic. The experimental outputs on community benchmarks data show that the given architecture acquires good prediction over the other state of the art label architectures.

Andrej Karpathy. Transfer Learning, 2017 [9]. CS-231n is a deep learning class by Andrej on computer-vision with deep neural network labels as CNNs for computer recognition, it is recorded at Stanford University, the US in the Engineering School.

Re-thinking the inception architecture for. Computer Vision and Pattern Recognition” [10]. This research paper is used on the image captioning using inception model architecture of CNN. According to this paper, the increased dimensions, computation price results to convert the instant quality gains for most tasks, computationally the accuracy efficient and fewer weights counts are yet allowing the condition of certain use cases such as drone vision and big data areas.

Planet: Understanding the Amazon from Space” [11]. This paper gave us the idea of using the encoder-decoder model for predicting the caption in the satellite images this paper used the inception_v2 model as encoder and decoder is Long Short term Memory and the result is caption generated at the end.

Fig 1. User Interface for Image Captioning.

This displays the whole design of our working model including the components and the states occurring during the execution of the process. It displays the very initial process of image feeding followed by the parsing and breaking down of the image into vector, where all the data regarding the image is stored and fed to the model. The LSTM is used in the encoder-decoder architectures which play the image again and aging develop the caption with the help of language processing and data(trained) stored, thus provide generated caption as output.
Fig 2. Image processing model

Fig 3. Satellite Image captioning Software

This requires the following prerequisite:

1) Anaconda Python installation.
2) Keras with tensor-flow 1.14 backend.
3) Image Model VGG19 model weights.
4) Decoder Model GRU model weights.
5) Create an environment in conda with keras.

3. **Proposed Methodology**

We have well trained deep convolutional neural network (CNN) to obtain image features and used multiple classification frameworks including long short term memory or GRU label captioning and binary cross-entropy to predict multi-class, multi-label images. Satellite images are trained on computer vision to learn image features and used classification captions including GRU label captioning and sparse_cross_entropy to predict multiclass, multi-label images. By fine-tuning an
architecture consisting of the encoder of pre-trained VGG-19 parameters trained on ImageNet data together with the LSTM decoder.

![Decoder Components](image1)

**Fig 4. Decoder Components**

![Encoder and Its layers](image2)

**Fig 5. Encoder and Its layers**

3.1. *Encoder Decoder Architecture*

The Encoder and Decoder architecture is utilized for a kind of setting where a variation of length of input-sequence of the sentence is mapped over the variation length provides output-sequence. The same model can also be trained for image caption or classification. In image captioning, the strong ideology is to utilize VGG19 as an encoder and a normal GRU as decoder multiple classification frameworks including long-short-term-memory (LSTM) or GRU (gated-recurrent-unit). Recurrent Neural Network is used for a variety of applications including machine language translation and chatbot model creation. The Encoder Decoder architecture is used for such practices where the varied insertion procedure is plotted over the varying length of output-array. The same network can be used under the image caption project.

**GRU:** GRU- The Gated Recurrent Unit strives to resolve the gradient disappearing problem in the backpropagation that tags along with basic RNN. GRU is a variation on the LSTM (GRU came after the Long Short Term Memory) and the reason is the similar structure and, in a few instances produce similar awesome outputs in case of machine translations.
4. Model Creation And Training

4.1. Data Pre-Processing: Captions
In Machine Learning, data preprocessing is the key step to clean the data in order to get unified and error free data, or encode, to bring it in a certain form that the system can easily form it. To define in other way words, features, and characteristics of the data can be easily processed and interpreted by certain algorithms. One must note that the captions are the thing that everyone wants it to be predicted. So while training time, captions are the target variables or expected outputs Y that the model is training to predict.

One can predict the output word by each word. Thus, we'd like to encode words in a hard and fast sized list or array. However, this part is going to be seen later once we check out the model design, except for now, we'll create two Python Dictionaries namely “word-to-ix” (pronounced as word to index) and “ix-to-word” (pronounced as an index to word). Stating that one will be representing every distinct word in created vocabulary dictionary by a number index. As seen above, we have 10000 distinct words in the dictionary and hence each word is represented by a number between 1 to 10000. The Python dictionaries are used as follows:

- word-to-ix [‘abc’] -> returns index of the word as ‘abc’
- ix-to-word [p] -> returns the word whose index is as ‘p’

We trained deep convolutional neural network CNN to learn image features and used multiple classification frameworks including long-short-term-memory LSTM label captioning and binary cross-entropy to predict multi-class, multi-label images. Satellite images are trained on computer vision to learn image features and used classification captions including GRU label captioning and sparse_cross_entropy to predict multiclass, multi-label images. By fine-tuning an architecture consisting of the encoder of pre-trained VGG-19 parameters trained on ImageNet data together with the LSTM decoder.

4.2. Data Pre Processing: Images
Non-analog image preprocessing is the use of algorithms in order to do image pre-processing on the picture before feeding them directly to the model. As a result a subarea of digital signal processes, image process contains many merits over no analog image processor. It gives permission to a much wide angle of architecture application to the input info, the main goal of non-analog image
pre-processing is or advancement of the image features by not accepting undesired distortions and promotion of few important images features so that our artificial intelligence computer Vision model can be benefited from this improved image features.

Images are not a thing but input X to the encoder and decoder model. As you may already know that any X to MODEL must be given in a certain sequence of a matrix. One should transform every image into particular sized vectors that can be fed as input X to the particular net. For this to be done, one can go for transfer learning by using the VGG 19 model Convolution Neural Network. VGG 19 was trained on Image net-datasets for image classification on thousands of different images classes. However, our purpose is to generate a caption and not to classify the image. Obtaining an informative vector for each and every picture. The process is known as feature extraction

5. Trained Model
This is the model representation of the complete architecture which show how the model is trained with some images being fed to the architectures and storing the output from the LSTM and CNN as the convolutional neural network and long shorter memory LSTM processes and stores the data of image in vector and reiterate the process until the final meaningful captions are received and also the whole images are processed.

![Fig 7. Training CNN and LSTM flow Diagram.](image)

![Fig 8. Block Model of implementation of encoder decoder architecture.](image)
Using pre-trained state of the art models like VGG-19 architecture our team is ready to create architectures that exploit the structure of our dataset in multiple ways and achieves strong performance accuracy. Still, moving forward, there are still various milestones we wish to pursue further. Specifically, we are currently acting on exploiting the labeling (i.e. hierarchically predictions which exploit the weather label, leaf type, then rare land type natural ordering), assembling multiple optimized models including transfer models using GRU in the decoder and other pre-trained deep RNN algorithms, and leveraging the knowledge within the .tiff files (specifically the Near-IR channel which tends to be very informative and used widely in remote sensing applications).

6. RESULT

Digital pic preprocessing is the use of algorithms in order to do image pre-processing on the picture before feeding them directly to the model. As a result a subarea of digital signal processes, image process contains many merits over no analog image processor. It gives permission to a much wide angle of architecture application to the input info, the main goal of non-analog image pre-processing is or advancement of the image features by not accepting undesired distortions and promotion of few important images features so that our artificial intelligence computer Vision model can be benefited from this improved image features.

Results depends on different CNN: In order to get the results of the Attention-based-method is based on the variation of different kinds of convolutional neural networks. Here is the result of a few experiments. The attention-based-methods are based on the convolutional feature of convolutional neural networks these features are applied in attention_based_method all over most of CNN features are take out by different CNN architecture.

For VGG16 these features maps of conv5 are of 3 sized 14 × 14 × 512 are applicable;
For VGG19 and AlexNet these features are of conv_5 are of size 13 × 13 × 256 are applicable;
For GoogLe_Net, these features are of inceptions 4 c / 3 × 3 sized 14×14×512 are used. The outcomes of CNN features are de rooted by differential models. Our team can view that the outcomes of attention_mechanism are far better than the outcomes of soft_attention_mechanism in most fields. The hard_attention_mechanism are based on the CNN feature generated by GoogLeNet gets the best outcome. But for the captioning dataset, the results of soft_attention_mechanism based on CNN feature generated by VGG16 get the best outcomes. The software we made processes these images and finally provides us with the output as the caption of the fed image using all the stored acquired data from training and the algorithms such as LSTM and GRU. The main role is played by the encoder-decoder architecture which successfully implements the algorithms and provides a better accuracy rate.

Deployment displays the whole design of our working model including the components and the states occurring during the execution of the process of caption generation to storing it. It displays the very initial process of image feeding followed by the parsing and breaking down of the image into vector, where all the data regarding the image is stored and fed to the model. The LSTM (Long Short Term Memory) or its updated version the GRU - Gated Recurrent unit is used in the decoder architectures which plays the image, again and again, to develop the caption with the help of language processing and trained data stored and thus provide the generated caption as output. Overall, experimenting with and optimizing our suite of model frameworks served to be an illuminating and exciting final project. The final output is generated as caption and automatically is stored in the database in the form of the .csv file.

Expected results for the current doing, we applied different stages of deep learning detection architecture to spatial images. Our team trained these architectures using the technical ways of transfer learning, which provide advantages of already trained features to extract on big datasets then stretched the resulting image distinction layer to historically non defined classes. Knowledge of the objects in the image, our team generated captions applying a recurrent neural network with long-short-term-memory. All over the method relates the words and provides captions architecture to those known
animates. Comparing it to historic work, our team’s approach is to compare the picture architecture’s all oversize, therefore, creating onboard spatial processing prospectively feasible while also correcting and influencing the vocab.

Culturally included for large caption training steps. Our work also shows over the benchmark text word in vocabulary to include a similar sentence structure. There are two un-expected results of the newly related structure of the early captioning vocabulary follow from built-in annotations or a ruling sensitivity to visual descriptions, and its non-ability to acquire the earth knowledge that man expertise might not offer such as the description of the physical relationship in between forest habitation and other things. Our team views the application of those two results in more depth.

Predicted caption: clear primary water eeee

generate_caption("2.jpg")

Predicted caption: clear primary water eeee
generate_caption("3.jpg")

Predicted caption: clear primary eeee

generate_caption("4.jpg")

Predicted caption: agriculture clear cultivation primary road eeee
7. Conclusion

With the present boom in satellite earth-imaging companies the apparent challenge lies in accurate and automatic interpretation of the huge datasets of accumulated images. During this project, we tried to tackle the challenge of understanding one subset of satellite images – those capturing images of the Amazon rainforest – with the actual goal of aiding in characterization and quantification of the deforestation of this area. Using pre-trained state-of-the-art models like VGG-19 architecture we were ready to create architectures that exploited the structure of our dataset in multiple ways and achieved strong performance accuracy. Still, moving forward, there are still various milestones we wish to pursue. Specifically, we are currently acting on exploiting the labeling (i.e. hierarchically predictions which exploit the weather label, lea type, then rare land type natural ordering), assembling multiple optimized models including transfer models using Res-Net and other pre-trained deep CNN algorithms, and leveraging the knowledge within the .tiff files (specifically the Near-IR channel which tends to be very informative in remote-sensing applications). We tried to build the software in such a way that it not only generates the caption of a particular image but also tore’s it in a result.csv file that is generated automatically along with the path of the file.

Deployment displays the whole design of our working model including the components and the states occurring during the execution of process of caption generation to storing it. It displays the very initial process of image feeding followed by the parsing and breaking down of image into vector, where all the data regarding the image is stored and fed to the model. The LSTM (Long Short Term Memory) or its updated version the GRU - Gated Recurrent unit is used in the decoder architectures which plays the image again and again to develop the caption with the assist of language processing and learned data stored, thus gives the generated caption as output. Overall, experimenting with and optimizing our suite of model frameworks served to be an illuminating and exciting final project.

At last during the training the model loss was decreasing with each epoch and at last the training loss was 0.05. This was the totally different method in deep learning our team has come across. Our team will still be working on to reduce the loss and increase accuracy with high GPU and Tensor flow Object detection API using other advanced algorithms such as Faster RCNN, SSD, MASK RCNN etc. to detect and segment the forest land covers on the image.
8. Future Enhancements

Further our team would adore to figure on real-time object detection and segmentation of the satellite for that we'd want to use the drone for activity and check functions and would love to publish a look paper.

Algorithms is used for the item detection and segmentation:

- quick RCNN - The Region-based-Convolutional-Network technique (Fast R-CNN) use of entity recognition. It shapes prior effort proficiently categorize object offers pattern profound convolutional networks. Compared to older, It works many innovations for boosting up the employment and challenging haste whereas else cumulative uncovering precision. It trains the extremely profound VGG16 net 9x quicker in comparison with R-CNN, it is 213 x quicker at testing-moment and attains succeeding map on PASCAL VOC 2012. Against to SPPnet, quick R-CNN trained VGG16x quicker.
- faster RCNN - The RPN is used for user to user for coming back up with high-quality region proposals, that unit of activity utilized by quick R-CNN for detection. We have a tendency to tend to any combine RPN, quick R-CNN to one system by the division of the convoluted features-employing the new common word of neural-net with 'attention' mech employability. SSD – Gift of the way for investigation objects in footage victimization one profound neural-net. The maneuver known as solid-state drive, overcome the result home of leaping box into a bunch of evasion box over. Altogether entirely absolutely utterly totally different facet fraction and scaling for every feature to locate mappings. throughout predicting the network’s generating scores for the presence for every object’s class in every default container and performs updates to the container to the raise of the match the matter kind
- MASK RCNN - Mask R-CNN is more faster and efficient than the R-CNN with the help of addition of an aspect to predict which is associated with object’s masking beside the existing subdivision for leaping case identification. It is very usual to coach and it also adds an absolutely little overhead to quicker R-CNN, working on five Federal protective Service. Further, It is simple to standardize [to fully to utterly to totally] completely totally different tasks, e.g., allowing u. s. of America to estimate human poses among identical framework.
- YOLO –It processes as rates up to 45 per second of the image's footage in this quantity. The smaller model version of Yolo network i.e. fast Yolo, functions honor amazing a 155 per second while still achieving the double of the map of assorted quantity indicator. Against with progressive uncovering systems, It marks supplementary localization faults though it’s much fewer potential to foresee untrue findings wherever unknown happens.
- RFCN - entirely convoluted net for correct and economical entity recognition. In distinction to former region-based indicators like Fast/Faster R-CNN that smear an upscale for every region in the sub-net persistently.
- BILTZNET - the amount of your time scene understanding has become crucial in several applications like autonomous driving. throughout this paper, we have got an associate inclination to propose a deep vogue, mentioned as BlitzNet, that place on performs object detection and linguistics segmentation in one play, permitting the quantity of jiffy computations. Besides the strategy gain of getting one network to perform many tasks, we have got an associate inclination as an associate instance that object detection and linguistics segmentation feel in one another in terms of accuracy. These are the few algorithms that we'll be a pattern through tensor-flow object detection API and can plan to notice the short and better accuracy one and would write a the paper supported that.
References

[1] G.G. Wilkinson. “Results and implications of a study of fifteen years of satellite image classification experiments.” IEEE Transactions on Geoscience and Remote Sensing (Vol. 43, No. 3). 2005.

[2] Sunitha Abburu, Suresh Babu Golla. “Satellite Image Classification Methods and Techniques: A Review”. International Journal of Computer Applications, (Vol.119, No. 8). 2015.

[3] Sayali Jog, Mrudul Dixit. “Supervised classification of satellite images”. Conference on Advances in Signal Processing (CASP), 2016.

[4] George F. Hepner. “Artificial neural network classification using a minimal training set. Comparison to conventional supervised classification”. Photogrammetric Engineering and Remote Sensing. (Vol. 56, No. 4). 1990.

[5] Turgay Celik. “Unsupervised Change Detection in Satellite Images Using Principal Component Analysis and k-Means Clustering”. IEEE Geoscience and Remote Sensing Letters, (Vol. 6, No. 4). 2009.

[6] ArcGIS. “What Is Image Classification? ArcGIS 10.5 Help Site, 2017.

[7] A. McCallum. “Multi-label text classification with a mixture model trained by EM. AAAI99 "Workshop on Text Learning. 1999.

[8] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, Wei Xu. “CNN RNN: A Unified Frame-work for Multi-Label Image Classification”. The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2285-2294.

[9] Andrej Karpathy. Transfer Learning, 2017. CS231n: Transfer Learning.

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, Zbigniew Wojna. “Rethinking the Inception Architecture for Computer Vision”. Computer Vision and Pattern Recognition. 2015.

[11] Scott Wallace. Amazon Rainforest, Deforestation, Forest Conservation. National Geographic. Farming the Amazon

[12] Robinson Meyer. Terra Bella and Planet Labs’ Most Consequential Year Yet. The Atlantic, 2016. Terra Bella and Planet Labs’ Most Consequential Year Yet Planet: Understanding the Amazon from Space. Kaggle. Challenge link

[13] Meenakshi K, Safa M, Krthick T, Sivaranjani N.” A novel study of machine learning algorithms for classifying health care data”, Research Journal of Pharmacy and Technology, Research Journal of Pharmacy and Technology, 2017.

[14] Meenakshi K, Maragatham, G, Agarwal, N., Ghosh, I, “ A Data mining Technique for analysing and predicting the success of movie”, Journal of Physics: Conference Series, Vol 1000, Issue 1, 2018

[15] Saranya, G., & Pravin, A. (2020). A comprehensive study on disease risk predictions in machine learning. International Journal of Electrical and Computer Engineering (IJECE), 10(4), 4217.

[16] Saranya, G., Geetha, G., & Safa, M. (2017). E-antenatal assistance care using decision tree analytics and cluster analytics based supervised machine learning. 2017 International Conference on IoT and Application (ICIOT).

[17] G Geetha, M Safa, C Fancy, D Saranya “A hybrid approach using collaborative filtering and content based filtering for recommender system”, Journal of Physics: Conference Series Vol 1000,Issue 1, 2018.

[18] M Srivastava, S Pallavi, S Chandra, G Geetha,”Comparison of optimizers implemented in Generative Adversarial Network (GAN)"International Journal of Pure and Applied Mathematics Volume 119 Issue 12,2018