Karnataka Law Society’s

GOGTE INSTITUTE OF TECHNOLOGY

Udyambag Belagavi -590008

Karnataka, India.



A Course Project Report on

**Image Colouring and Captioning**

Submitted for the requirements of 6th semester B.E. in ISE

for **“ARTIFICIAL INTLLIGENCE (18IS61)”**

**Submitted by**

**NAME USN**

**1) John Nixon 2GI19IS016**

**2) Hrutuja Patnekar 2GI19IS017**

**Under the guidance of**

**Prof . Shuba Sanu**

**Academic Year 2021-2022 (Even Semester)**

Karnataka Law Society’s

GOGTE INSTITUTE OF TECHNOLOGY

Udyambag Belagavi -590008

Karnataka, India.

**Department of Information Science and Engineering**



**Certificate**

This is to certify that the Course Project work titled **“Image Colouring and Captioning”** carried out by **Students: John Nixon, Hrutuja Patnekar** bearing **USNs: 2GI19IS016, 2GI19IS017** is submittedin partial fulfilment of the requirements for 6th semester B.E. in **INFORMATION Science and Engineering,** Visvesvaraya Technological University, Belagavi. It is certified that all corrections/ suggestions indicated have been incorporated in the report. The course project report has been approved as it satisfies the academic requirements prescribed for the said degree.

Date: Signature of Guide

Place: Belagavi

Prof . Shuba Sanu

KLS Gogte Institute of Technology, Belagavi

Name of the Examiners Signature of the Examiners

1.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 1.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ 2.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

CONTENTS:

1. [INTRODUCTION: 3](#_Toc104802303)
2. [METHODOLOGY: 4](#_Toc104802304)
3. [LIMITATIONS: 9](#_Toc104802312)
4. [CONCLUSION: 11](#_Toc104802313)
5. [REFERENCES: 12](#_Toc104802315)

# TITLE:

Image Colouring and Captioning

# ABSTARCT:

We present a convolutional-neural-network-based system that faithfully colorizes black and white photographic images without direct human assistance. We explore various network architectures, objectives, colour spaces, and problem formulations. The final classification-based model we build generates colorized images that are significantly more aesthetically-pleasing than those created by the baseline regression-based model, demonstrating the viability of our methodology and revealing promising avenues for future work.

Artificial Intelligence is also doing great progress in all factors, so automatically detecting the content of an image is a basic and important problem in AI fields that deals with computer sight and natural language processing. The paper is a study about the model based on a deep recurrent architecture that interacts with the recent advances in done in image captioning in computer sectors and machine translation fields and that can be used to produce natural sentences that gives detailed information of an image. Image captioning is a piece of work that requires the understanding of images and the awareness of producing correct description sentences with proper and suitable structure by extracting the features of image. In this study, we try to understand a hybrid system describing the use of Convolutional Neural Network (CNN) to generate an accurate description of the images and make the use of an LSTM to accurately arrange the meaningful sentences using the removed or extracted keywords. CNN checks the similarity in the target image with a large dataset of training images and tries to generate an accurate description using the trained captions.

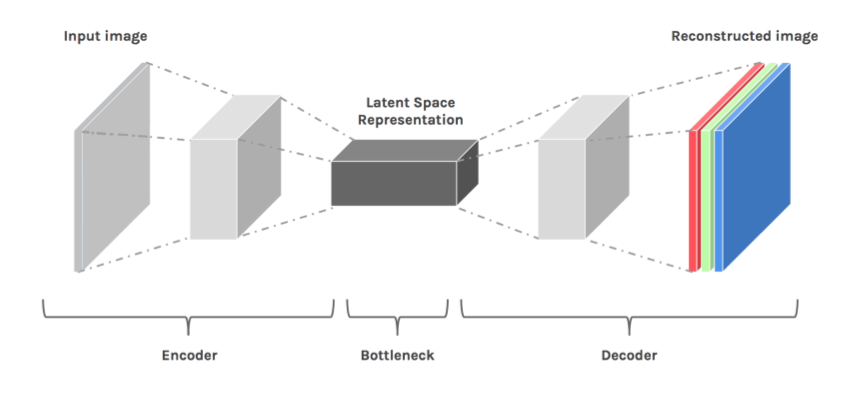
# OBJECTIVES:

* For any given image as a input our Model should be able to detected weather the image has colours.
* For any uncoloured image our Model should be able to generate a new image with accurate colours.
* Finally, our Model should be able to generate caption’s for the final coloured image by extracting the features present in the image and running searching algorithm’s to generate the sentence appropriate to the Inputted image

# METHODOLOGY:

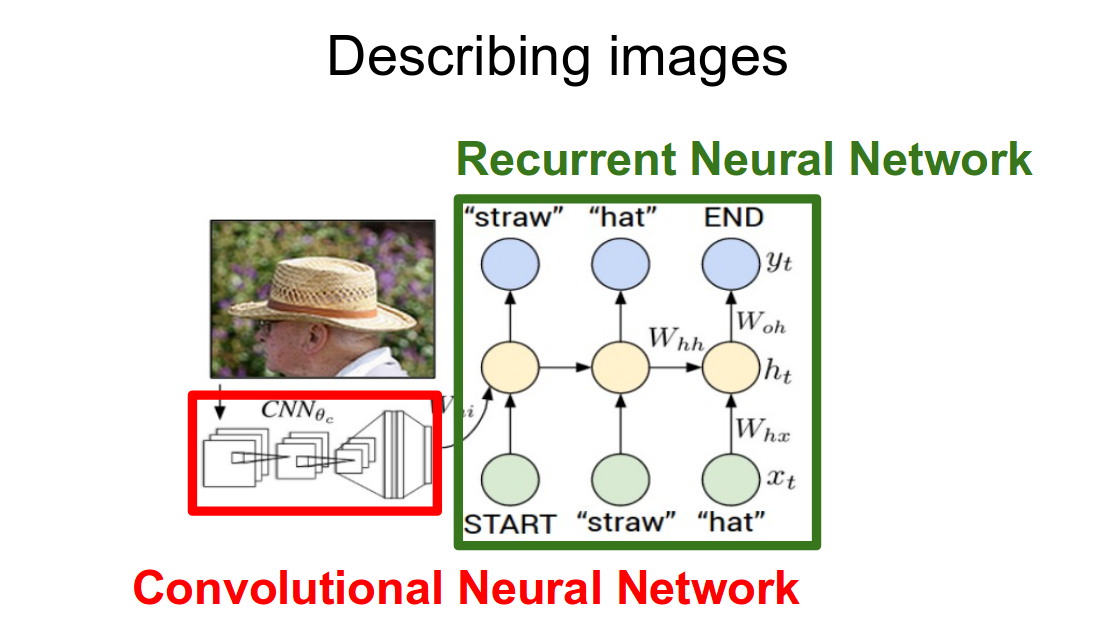
We can use Auto-encoder for the reconstruction of the image, in other words, we would say that it has the ability to generate and that’s exactly what we want to do, we want to generate the three channels of **RGB**.

One approach is to make two copies of your image, one to be a grayscale image and it will act as your input to the encoder which is responsible for extracting the features of the image “Latent Space Representations” that can be used as input for the decoder to reconstruct the image, the other copy will be the same image but colorized as your target to the decoder ([Supervised Learning](https://en.wikipedia.org/wiki/Supervised_learning)) so that it can minimize the error between the original colored image and the generated one. Auto-encoder architecture would be something like in the figure below.



The task of image captioning can be divided into two modules logically – one is an **image based model** – which extracts the features and nuances out of our image, and the other is a **language based model** – which translates the features and objects given by our image based model to a natural sentence.

For our image based model (viz encoder) – we usually rely on a Convolutional Neural Network model. And for our language based model (viz decoder) – we rely on a Recurrent Neural Network. The image below summarizes the approach given above.

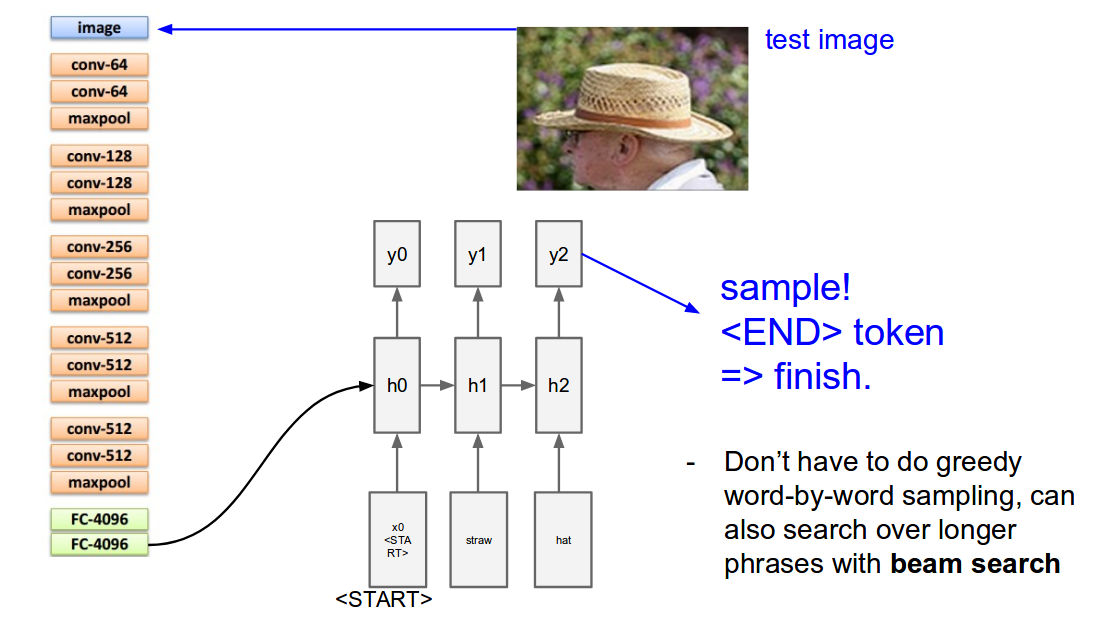


Usually, a pretrained CNN extracts the features from our input image. The feature vector is linearly transformed to have the same dimension as the input dimension of the RNN/LSTM network. This network is trained as a language model on our feature vector.

For training our LSTM model, we predefine our label and target text. For example, if the caption is “A man and a girl sit on the ground and eat.”, our label and target would be as follows –

*Label – [ <start>, A,  man,  and,  a, girl,  sit,  on,  the,  ground,  and,  eat,  . ]*  
*Target – [ A, man,  and,  a,  girl,  sit,  on,  the,  ground,  and,  eat,  ., <end> ]*

This is done so that our model understands the start and end of our labelled sequence.



# IMPLEMENTATION:

Our model is a convolutional neural network. We first apply a number of convolutional layers to extract features from our image, and then we apply deconvolutional layers to upscale (increase the spacial resolution) of our features.

Specifically, the beginning of our model will be ResNet-18, an image classification network with 18 layers and residual connections. We will modify the first layer of the network so that it accepts grayscale input rather than colored input, and we will cut it off after the 6th set of layers.

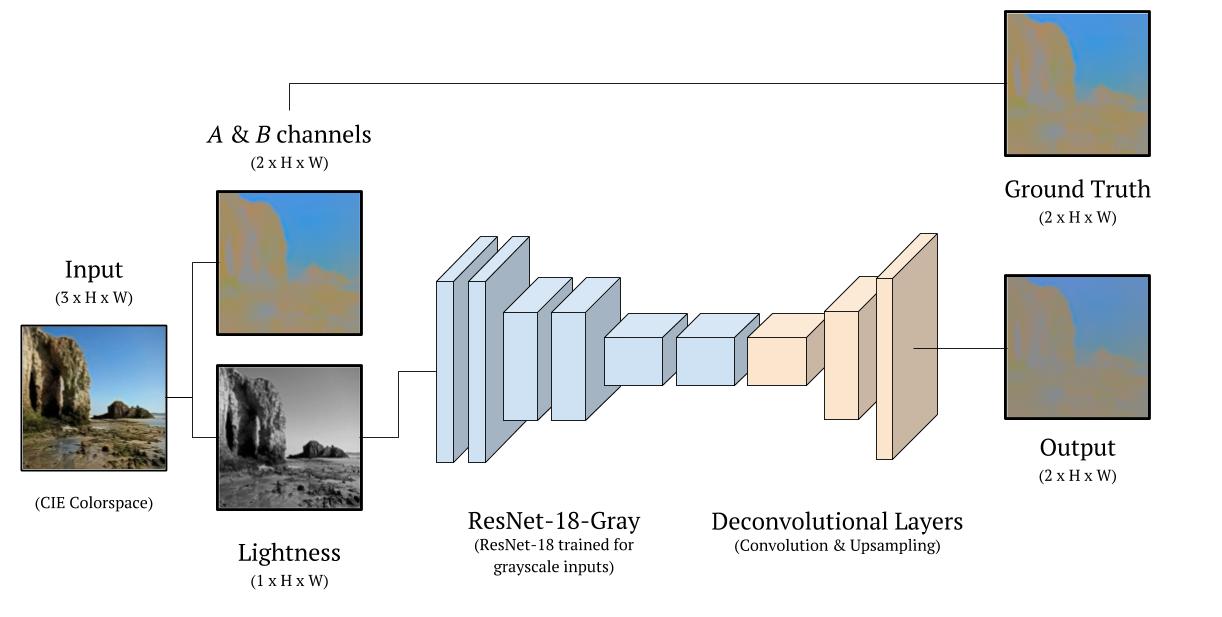
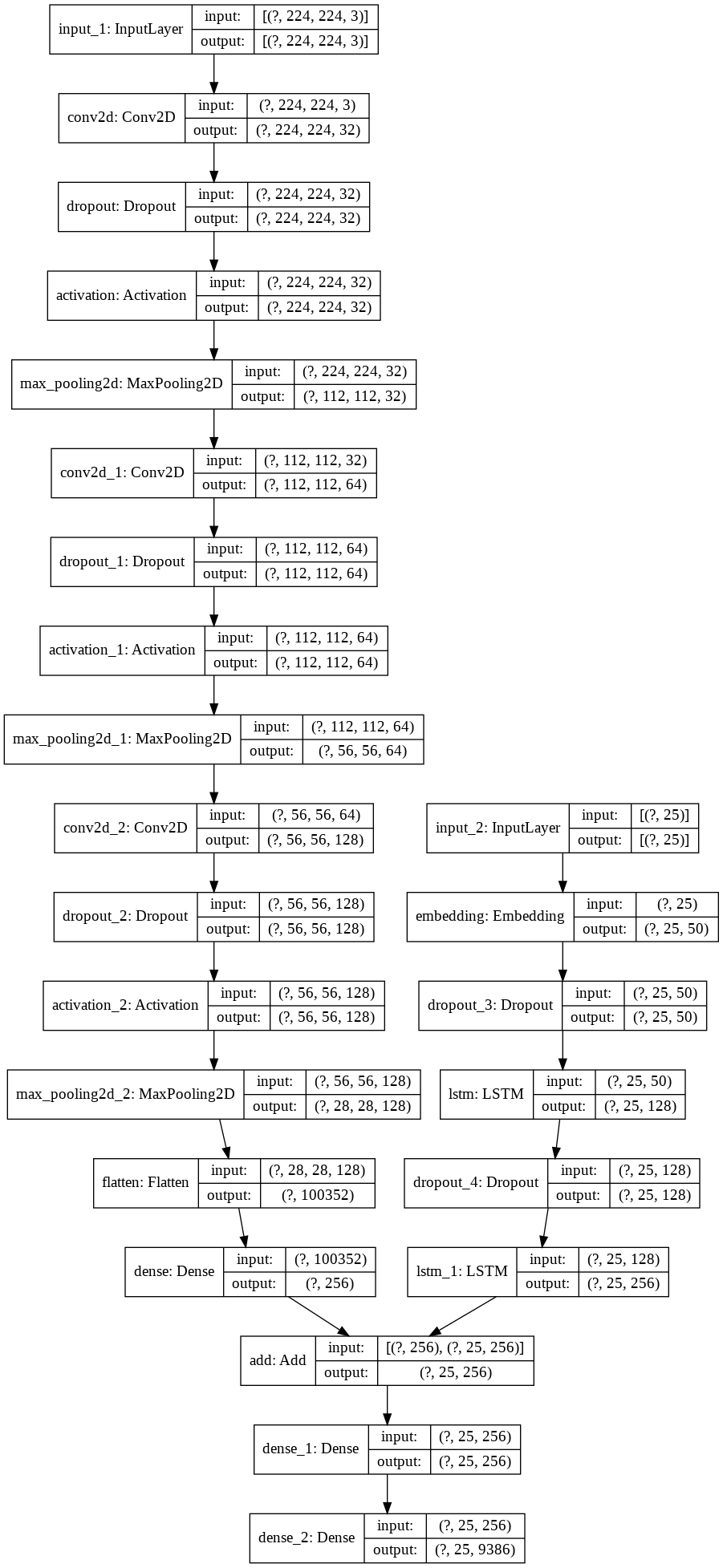


Image Captioning:

The Flickr 8k dataset has 8091 images and for each image, there are 5 descriptions. The dataset can be found at the[University of Illinois site](https://forms.illinois.edu/sec/1713398).

We have used merging architecture and We have created our own convolutional network though normally the image feature extractions are done using pre-trained CNN architectures using transfer learning.



# 

# CONCLUSION:

For any given sample image our model was able to predict the colors as well the captions and properly convert into sentences .

We have also run beam searches on the features to get a faster approach on the captioning part.

Thus the Models perform well as it was trained for .

There are still some room for improvement as the captions are generated in a generic way which is not always correct in some cases

# REFERENCES:

* Comparison of Architectures: <https://arxiv.org/pdf/1703.09137.pdf>
* Andrej Karpathy’s Architecture: <https://cs.stanford.edu/people/karpathy/cvpr2015.pdf>
* Google’s Architecture: <https://arxiv.org/pdf/1411.4555.pdf>
* Microsoft’s Architecture: <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/06/ImageCaptionInWild-1.pdf>