Proceedings of 10th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2021 Month: October Volume: 10

Coloring of Gray Scale Image using U-net Architecture

Bhuwan Acharya [a], Sitaram Pokhrel [b]

[a, b] *Department of Electronics and Computer Engineering, Paschimanchal Campus, IOE, Tribhuvan University, Nepal*
Corresponding Email: [a] acharyabhuwan444@gmail.com

Abstract
Coloring of a gray scale image is extensively chosen for various researches in graphics and computer vision especially with deep learning methods.U-net is crucial for coloring of the gray scale image, which the major part of this research work. The U-net is a type of deep Convolution Neural Network (CNN) which consists of the down sampling and up sampling paths.They are connected through the bottleneck and skip connection between the encoder and decoder which provides the copy and concatenate the feature maps from encoder to decoder. The input to the network is gray image of size 256×256×1 and it produced the RGB color image of same size as input. The Mean Squared Error (MSE) loss function is to distinguished the quality between the generated color image to its corresponding ground truth image.The Peak Signal to Noise Ratio (PSNR) is used for the quality measurement of the predicted color image and the ground truth image.

Keywords
U-Net, CNN, Mean Square Error, Encoder-Decoder,PSNR

1. Introduction

Colorization is the process in which color components are added to the grayscale image. The information contained in the grey-scale image is limited than that of color image. Thus adding the color components can provides more insights about its semantics. However,Colorization is one of the challenging task due to the small data set,variety of images in the training set and the availability of computational resources [1].

Colorization is an ambiguous task because it does not have unique solution. An U-net takes the grayscale and predict the color image .The coloring of the gray scale image has the strong impact on the wide verity of the domains typically color restoration re-design of the historical images [2]. The colorization problem is both complex and interesting as the final output image and input image will be of same dimension [3].

CNNs plays the important role to handle different diversified tasks such as image colorization, classification, image-labeling and so on. In recent years the use of CNN widely used to get the solution of image colorization purposes[3].The U-Net architecture is convolution network architecture for the purpose of image segmentation. U-Net is widely used for image segmentation, and the problem of coloring image is essentially same as segmentation. Thus, it is chosen for image coloring[4]. The U-Net consist of mainly there parts Encoding, Bridge and Decoding. The encoding part is responsible for converting the input image into compact representation called latent space of the input. The decoding process involves the reconstruction of the input image with the same size of the input images by using upsampling and convolution operation. The role of the bridge is to bind the encoding and decoding units. The low level detail features in the encoding part are concatenated with corresponding high level features in the decoding part [4].

The colored image carries the more information than the gray scale images and most of the coloring tasks are based on the auto-encoder i.e. encoder-decoder. A number of proposed solutions are available for colorization of Black and White images. The challenges lies in focusing on accuracy colorizing standard would look natural to human eye. The U-net is basically used for the segmentation of the medical images but the task of coloring is similar to the segmentation because coloring involve the separation of similar region in the image and fill the appropriate color in that segment. Thus the U-Net architecture is used for the coloring because the encoder-decoder suffers from an information bottleneck during the flow of low level information in the network[3]. To reduce this problem features from the contracting path are also connected with the upsampling output layer within the network then more information can be obtained from greyscale image and can be automatically colored with greater accuracy and natural representation [4].

1.1 Motivation

Manual method of coloring of grayscale image takes significant time. Thus to prevent all these efforts and haphazard way of coloring, deep learning model was developed by using U-net architecture. A number of proposed solutions are available for colorization of Black and White images. The U-Net architecture is used for the coloring because the encoder-decoder suffers from an information bottleneck during the flow of low level information in the network[3].To reduce this problem,features from the contracting path are also connected with the upsampling output layer within the network. Therefor,more information can be obtained from greyscale image and can be automatically colored with greater accuracy and natural representation[4].

2. Related work

In paper [5] involves the colorization of the cultural and historical images for the Nepal using the deep learning CNN model combined with Inception-ResnetV2. Two objective functions MSE (Mean Squared Error) and PSNR (Peak Signal to Noise Ratio) are implemented for objective quality assessment between the estimated color image and its ground truth.The accuracy was found to be 75.23%.

In paper[1],U-Net architecture is used for the segmentation of brain tumor images that helps diagnosis of disease, treatment planning and surgical navigation.The U-Net model consists of a downsampling contraction path and corresponding upsampling expansion path. In this paper the data augmentation method is not used because the U-Net model itself requires relatively small amount of data. In paper [6] the CNN is trained in order to map the gray scale image input over distribution of quantized color value outputs. In this paper the model is focus on the design of objective function and inferring point estimates of color from the predicted color distribution.In paper [7] The CNN model is trained using 244×244×3 and the image is converted in to CIE color space. Black and white luminance L* channel is fed to the model as input and the a* and b* channels are extracted which are the target values. The testing process gray scale image 244×244 input to the trained model and generates the corresponding to the a* and b*channels of the CIE color space. These three channels are combined for represent the predicted color image.In paper[1] the coloring of image is done by using the convolution classification approach in which the colorUNet architecture is used. In this approach the problem is like a segmentation, in which correct class is predicted for every segment of the input image among the 32 color bin.In paper[8] the U-Net architecture is used in the cell segmentation task in the biological images. The network is trained using the combination of loss function for image segmentation.In paper [9] the U-net is used for the medical image segmentation. The main advantage of using U-net in comparison with other networks have short training time, simple structure and less sample demand but has slightly depth is insufficient. However the Residual U-net is more efficient than general U-net in which the convolution layer is replaced with the residual network.

3. Research methodology

The purpose model consists of U-Net architecture.The block diagram of the purposed system is shown in figure 1.

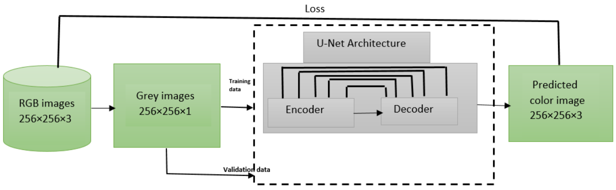

Figure 1: Block Diagram of Proposed system

3.1 U-Net Architecture

It is an approach to use a CNN as an encoder-decoder network, where the encoder decreases the width and the height of the image but increase the depth or the number of features, while the decoder uses transposed convolution operations to increase its size and decrease depth. The transpose convolution operation is a process of moving in the opposite direction of normal convolution. Here the input to the U-net is gray image and output is colored image[3].

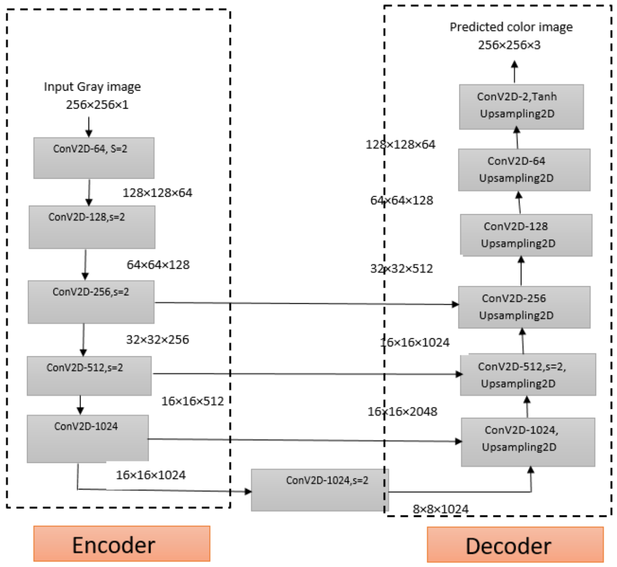

Figure 2: U-Net Architecture

When the image is input to the U-Net, the input image is progressively downsampled using a series of contractive encoding layers and then the reverse process is performed using the series of expansive decoding layers to reconstruct the input. The architecture of U-Net is symmetric with encoding units and decoding units.

In figure 2 the downsampling contraction path on the left and upsampling expansion path on the right. These two paths are structurally symmetric. After each convolution layer, the activation function LeakyReLU is used to ensure nonlinear mapping in in the downsampling path and ReLU is used in the upsampling path. However, the last layer of the decoder uses the tanh activation function. The number of channels in the output layer is 3 because the required output is color image which has three channels. Each convolution layer has the kernel size 3*3 with the stride 2. The function of convolution layer is to extract the features form the input image. The local features such as edge, lines etc. are extracted by the encoder.The coloring model of U-net consists of the six convolution layers in each upsampling and downsampling path.

The contraction part is mainly used for the extraction of the high-dimensional feature information. With each down sampling the image size reduced by ½ and the number of features doubles. Similarly, the expansion network is used for upsampling, the size is double and the number of features is reduced to ½[2]. When the deeper encoder is used to extract the features then it able to extract global feature as well but in this case the dimension of the representation is often too small then the decoder is not able to reproduce the original image[5].

3.1.1 Convolution 2D

The U-net consists of the convolution operation which is a general matrix multiplication in the layers. An image in the computer system is represented by simply 2D array. The gray scale image is represented as 2 dimensional matrix of pixels in which each element contains the pixels intensity. Similarly for the color image there are three such arrays or channels namely Red, Green and Blue. The output of the convolution operation is given by the formula:

$$\left(\frac{N-f-2p+1}{S} \times \frac{N-f-2p+1}{S}\right) \tag{1}$$

Where, N×N and f×f size of image and size of the filter, P is padding and S is stride. When we are not using the padding and the stride is equal to one then output of the feature map is less than the size of the input. It the input is padded with the zero then the output of the feature map is same as the input size. In the above architecture the stride is two that means the output of the is just half as the input size and the padding is adjusted in such a way the output of the feature map is same as the input by using the keyword ’same’.

3.1.2 Up sampling 2D

Function of Up sampling 2D layer is to make double the height and width of the input image by repeating the rows and column values in the matrix of an image. The function of the UP sampling 2D is just the reverses as the pooling.

3.2 Loos/Cost function

The objective of the training is to minimizing the MSE between the estimated pixel colors and their ground truth color image. Since the MSE is quadratic in nature i.e. the graph is a gradient decent with only one global minima. For an image X and its prediction is X of size H×W the MSE is given by the equation.

$$MSE = \frac{1}{H \times W} \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} [\tilde{X}_{i,j} - X_{i,j}] \tag{2}$$

This loss is used to compare across the three RGB channels and the objective is to reduce the MSE. The function of the optimizer is to reduce the loss.

3.3 Image Pre-processing

”balraj98/summer2winter-yosemiteare” is color image data collected from www.kagle.com.The Winter data set was used during training. The data set contains the variety of images like trees, mountains, clouds, houses, etc. The data set contains 1231 training samples and 309 validation samples.These 1540 total color images are placed on the same folder and they are divided in the training and validation with different proportion namely 70:30, 80:20, and 90:10%.Cropping and resizing was done to fix the resolution of the image 256× 256 × 3.Then the RGB images were converted in to the grey scale images.

3.4 Training and Testing

During training the gray scale image is fed to the model as input then RGB color components are extracted as target value. During the back propagation the weights are updated. The optimizer is designed to minimize the loss function during the training process.The Adam optimizer is used during the training process. During testing gray image of size 256× 256 × 1 is feed to the model which produced corresponding RGB image.

4. Results and Discussion

After this research the deep learning model was successfully build, train and tested.The trained model take image 256×256×1 and generate the color image of size 256×256×3. The model was trained on the different value of parameters namely batch size, learning rate of optimizer, training and validation data ratio.

The results obtained on different hyper parameters and their values are as follows:

- Batch size = 64
- Epoch = 100
- Learning rate of Adam optimizer = 0.001(default)
- Training and validation splits in the ratio of 80% and 20%

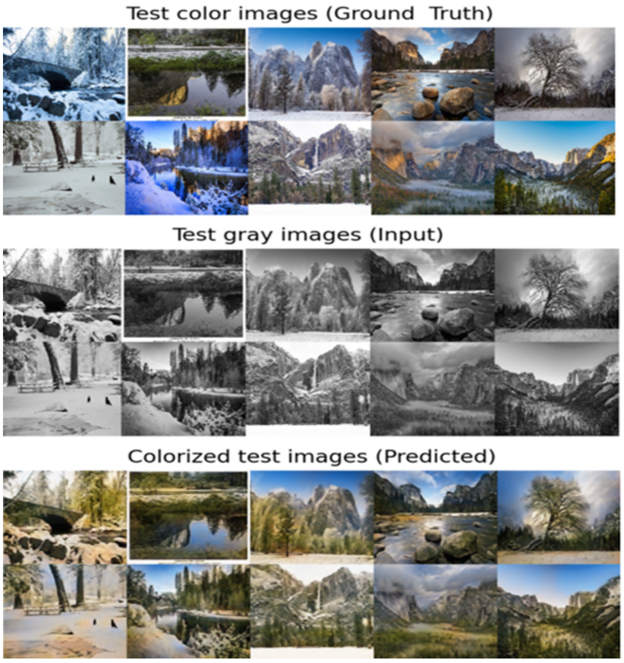
Test color images (Ground Truth)
Test gray images (Input)
Colorized test images (Predicted)

Figure 3: Ground trouth, gray and pridicted color image

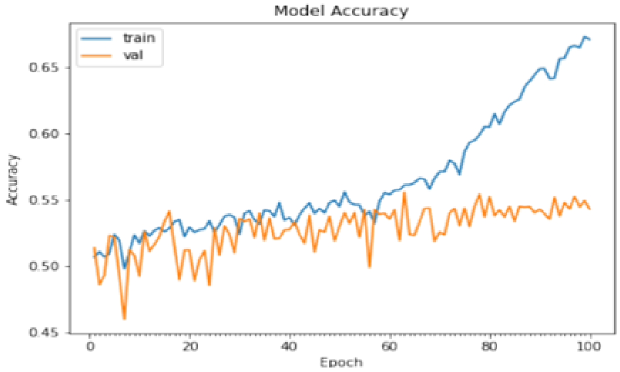

Figure 4: Training and validation accuracy

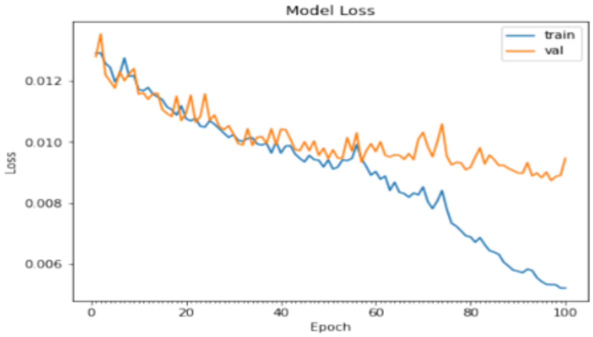

Figure 5: Training and validation loss

In order to fix the batch size 64, a number of experiments were performed between 20 to 100. From the experiment it is found that the value of accuracy

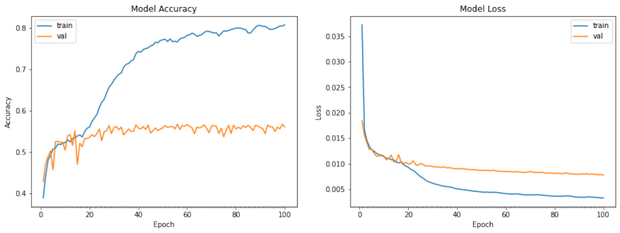

Figure 7: Training and validation loss and accuracy

and loss is optimum when batch size is 64. After selecting the batch size, the learning rate was changed and the best result was found at the rate of 0.0001.

The results obtained on different hyper parameters and their values are as follows:

- Training and validation splits = 70% and 30%
- Learning late of Adam optimizer = 0.0001
- Batch size = 64
- Epochs = 100

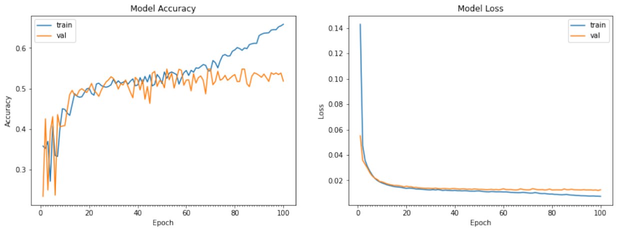

Figure 6: Training and validation loss

The results show that when more images are used to train the model then the training accuracy of the model was increased but the number of validation images becomes less then validation accuracy becomes low. In figure 4 the training accuracy is high but the validation accuracy is not good. In figure 6, 30% of data is used for the validation purpose and 70% used for training the model which results decrease in validation loss and increase in validation accuracy. Furthermore it is found that when the learning rate of Adam optimizer is 0.0001 then the model gives better result.

Figure 7 shows that the plot of training and validation accuracy and loss where training and validation splits 90% and 10% respectively. Batch size and learning rate is 64 and 0.0001 respectively.

The figure shown in 8 is the result obtained on test data set of different location of the pokhara with corresponding gray image.

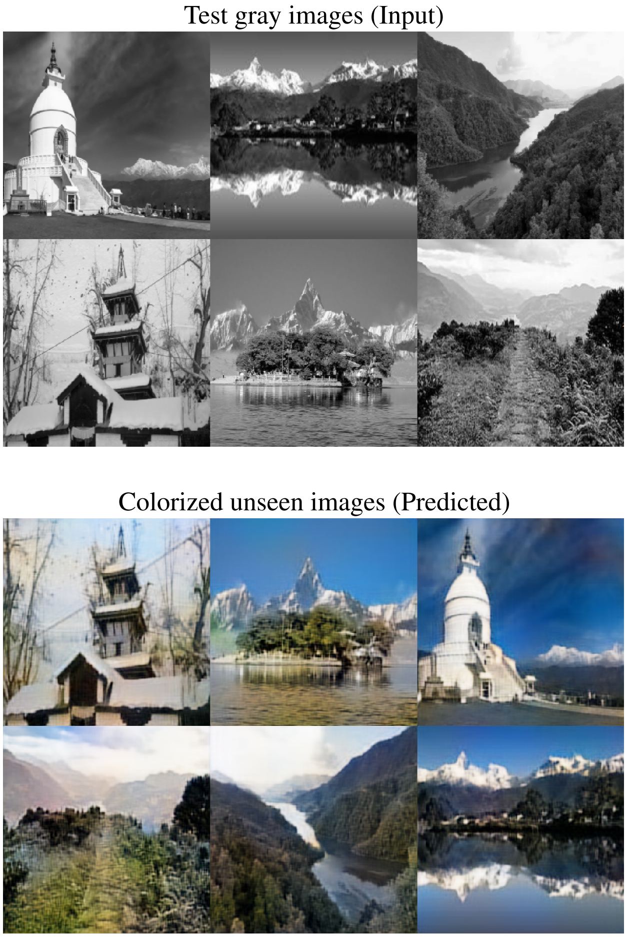

Figure 8: Gray and predicted color images of different locations of pokhara

4.1 Quality Measure Metric

The PSNR (Peak Signal to Noise Ration) is the most commonly and widely used metric for image quality assessment. In this research the PSNR is calculated between the colorized predicted image and the corresponding ground truth image[10]. It is expressed in dB and the mathematical equation of the PSNR is shown in equation:

$$\text{PSNR} = 10 \times \log_{10}\left[\frac{\text{peakval}^2}{\text{MSE}}\right] \tag{3}$$

$$\text{PSNR} = 20 \times \log_{10}(\text{peakval}) - \log_{10}(\text{MSE}) \tag{4}$$

Where, the peakval (peak Value) is the maximum value of the pixel in an image. From the equation 3 it is found that as MSE approaches to zero; then the value of PSNR becomes infinity which ensure the higher quality of the image.

Image quality metric are used to describe the underlying characteristic of image quality. It perform the comparison between the ground truth image and the predicted color image.The PSNR is widely used to measure image quality. Figure 9 shows the value of PSNR of the predicted color image along with the corresponding gray and ground truth image.

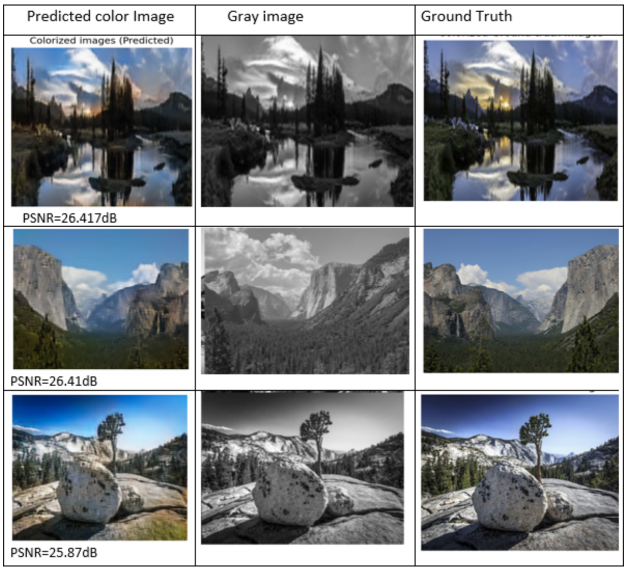

Figure 9: PSNR between the ground truth color image and the predicted color image

5. Conclusion

In this research, the auto-colorization model was developed by using U-net architecture.The model was trained over 1231 images of natural seen. The Adam optimizer is used to minimize the MSE loss between the predicted and ground truth color image. The best learning rate and batch size was found to be 0.0001 and 64 respectively. The model gives good result on the unseen test data in which an image components are sky, forest, mountains etc. The PSNR is used for the estimation of the quality of the predicted color image.

6. Limitation and Future works

To increase the performance of the model should be trained on the LAB color representation rather than RGB color representation.However, the model gives good result in RGB color representation. In the RGB color representation each pixel is represented by how much extent Red, Green and Blue the pixel is. However in L*a*b* color space, use three numbers for each pixels but they have different meanings. The purpose of separating the input RGB image into CIE L*a*b* color space is to separate the color characteristics from the luminance. When L*a*b model is used then L channel is used as input to the model (Gray scale) and want to predict the other two color channels a* and b*. But if we use RGB, first of all, the RGB images are converted in to gray scale image and then feed the grayscale image to the model as input and the task is to predict the three numbers of channel which is more unstable and difficult due to their many more possible combination of three numbers compare to two channels. Suppose we use 256×256×1 as input then the prediction of each pixels involves choosing between 256*256*256 combinations which is more than 16 million but if two channels are predicted having the combination of 256*256 which is 65000 choices. Performance of the model over the unseen images highly depends on the type of their specific domain. To overcome this problem the model should be trained on the variety of large set of data. The coloring of gray scale image can also be done by using Generative Adversarial Network (GAN).

References

[1] Vincent Billaut, Matthieu de Rochemonteix, and Marc Thibault. Colorunet: A convolutional classification approach to colorization. *arXiv preprint arXiv:1811.03120*, 2018.

[2] Federico Baldassarre, Diego González Morín, and Lucas Rodés-Guirao. Deep koalarization: Image colorization using cnns and inception-resnet-v2. *arXiv preprint arXiv:1712.03400*, 2017.

[3] Kamyar Nazeri, Eric Ng, and Mehran Ebrahimi. Image colorization using generative adversarial networks. In *International conference on articulated motion and deformable objects*, pages 85–94. Springer, 2018.

[4] Wei Zhang, Ping Tang, Lijun Zhao, and Qingqing Huang. A comparative study of u-nets with various convolution components for building extraction. In *2019 Joint Urban Remote Sensing Event (JURSE)*, pages 1–4. IEEE, 2019.

[5] Madhab Raj Joshi, Lewis Nkenyereye, Gyanendra Prasad Joshi, SM Islam, Mohammad Abdullah-Al-Wadud, and Surendra Shrestha. Auto-colorization of historical images using deep convolutional neural networks. *Mathematics*, 8(12):2258, 2020.

[6] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[7] Jeff Hwang and You Zhou. Image colorization with deep convolutional neural networks. In *Stanford University, Tech. Rep.* 2016.

[8] Cem Emre Akbaş and Michal Kozubek. Condensed u-net (cu-net): An improved u-net architecture for cell segmentation powered by 4x4 max-pooling layers. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 446–450. IEEE, 2020.

[9] Xueyan Gao and Lijin Fang. Improved u-net semantic segmentation network. In *2020 39th Chinese Control Conference (CCC)*, pages 7090–7095. IEEE, 2020.

[10] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.