Automatic colorization of gray-scale images using deep attention models

Vrishabh Lakhani
Rochester Institute of Technology
Rochester, N.Y., U.S.A
val3917@rit.edu

Abstract

There has been various architectures to colorize gray-scale images using deep learning methods. Improvising over this, we propose a model which combines the attention mechanisms in deep learning to the current state-of-art models which are trained with high-level features extracted from Inception-ResNet-v2 pre-trained models over ImageNet dataset. While the current model uses a Seq2Seq model which consist of encoder-decoder networks, where we encode the grayscale image into latent space using the encoder layer and we output the color values for the image using the decoder network. Further, such a Seq2Seq helps us to train the dataset over image of any sizes and aspect ratio, in this paper we show for the first time that adding attention mechanisms to this will help the model distinguish better between the objects inside the images and thus improve the overall results. Since image colorization requires the model to understand the local features, adding an attention mechanism gives the model a better idea of the local features over which it should infer the color values by fundamentally assigning different weights to the local features. The resulting model outperforms all the previous deep fully automatic colorization methods as well as non deep-learning based approaches. Not only that, attention mechanism reduces the total number of trainable features in a deep learning model thus making it more computationally efficient. We also discuss the implications and the advantage of solving such a problem using attention mechanism as such kinds of models thus are useful for real-life applications which works on any type of image.

1. Introduction

One of the most challenging tasks in Image-to-Image translation is translation between different domains of the image. One of such application is colorization of grayscale images to a color image space. This in itself has variety of application like colorizing historical pictures and enhancing all types of grayscale images for a better visual aesthetic and understanding of the image.

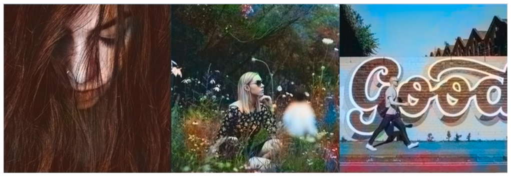

Figure 1. Rendered color images using Seq2Seq model with Attention mechanism

Inspired by recent successful Image-to-Image translation models using deep Seq2Seq models, this papers tries to improve this architecture. Seq2Seq models consists of two networks, the encoder and the decoder network. The encoder network takes in the input image vector as it's input and translates the image vector into a latent space. In case of this paper, this encoder network takes in gray-scale images as input. This latent space translation vector is then turned into the input of the output decoder network. The output decoder network converts this latent representation into a sequence of outputs as vectors, which could be in the form of 1D, 2D or n-dimensional vectors, thus it can be a representation in terms of text, audio, image, video, etc. In this paper, the decoder networks translates the latent representation of the input image channel into the chromatic output of the image which is used to color the image. This is possible using CNNs which are present in encoder and also in decoder layers as deconvolutional layers.

In order to provide such an input to the model that it outputs the chromatic values of each pixels and nothing apart from it, we have to provide use the La*b* colorspace, where L provides the Luminance channel and a* and b* provides the color channel. This can be used such that the input to the encoder layer will be an image of the L channel and the output of the decoder network will consists of two channels, the a* and b* channels. Thus, we will be translating one channel of the input to two channel of the output.

Recent research in this field is based on the deep learning approaches to the problem which are based on training deep Convolutional Neural Networks and/or using pretrained weights from models such are Inception, ResNet, or VGG which are trained over images in RGB colorspace. Based on previous research, we observe that instead of applying these networks directly to the grayscale images, they give much better results when we use some sort of prior to find out the chrominance of the grayscale image. However, doing so comes up with it's challenges and thus in order to make realistic looking colored images, there is a space for improvement over this.

Thus, we try to propose a model that is expected to be able to do this task better than the State-of-the-art model. The State-of-the-art model combines a Deep Convolutional Neural network with the Inception-ResNet-v2 which is pretrained over the ImageNet dataset. Our idea is to improve this model by using attention mechanism over the global prior of the network as attention mechanism have been proven to increase the accuracy as well as improve the performance of image classification task. They are also useful in successfully differentiating between different objects in an image and thus, this could be an effective technique for the application of automatic colorization of grayscale images. Additionally, we are adding the co-ordinates of each pixels as input with the raw image as input, this has known to help in translation according to recent studies. [13]

2. Related Work

Traditionally, works in the field of colorization have been based on statistical approaches where the image pixel values are approximated given the context of the picture by the user or using data driven approaches. There has also been for semi-automated approaches where the model is data driven and also at the same time expects helps over some pixel on the image from the user or from other contextual images.

2.1. Sequence to Sequence Learning with Neural Networks

In this [5] , the researchers give a novel approach of introducing Sequence to Sequence learning which helps deep learning models to learn sequences from other sequences, which is not possible through traditional DNN approach, which are usually map input sequence vector to an output target label.

This model uses the idea of encoder and decoder networks which translate an input of fixed input vector to an output of different fixed length output vector. This model can be used for text, images, audio and video sequences.

In this paper, they work on text dataset, WMT '14, to translate English to French.

Here, they use LSTM networks over the text dataset and achieve a BLEU score of 34.8, and on further improving they got a BLEU score of 36.8. This paper is used for all kinds of translations including caption generation, text-to-image, etc.

2.2. Neural Machine Translation by Jointly Learning to Align and Translate

In this [1] paper, the researchers propose a new approach to machine translation models whereby they try to improve over the encoder-decoder network to remove the bottleneck of having a fixed-length vector.

The idea here is to make a model that can automatically search which part of the encoding vector are relevant and search them so that they don't to explicitly make a hard segmentation of the input vector of a specific length.

They are working over the text dataset WMT' 14, to translate English to French. The paper gets the State-of-the-art results for Neural Machine translation.

Such kinds of model can be used for any sort of sequences, like text, images, video, audio, etc.

2.3. Deep patch-wise colorization model for grayscale images

In this [7] paper, the researchers propose deep patch-wise colorization model for grayscale image to RGB reproduction. Their model consist of three parts which consists of low-level feature networks, color fusion and refinement.

The model is a vectorized convolutional neural network (VCNN) which works on a YUV colorspace which is outputting from the two VCNNs. The output are then fused together along with alginment and reshaping of these images. This further undergoes a guided filtering techniques based on priors for the constructive color mapping models and this helps to reduce boundary artifacts.

They are working over smaller datasets since deep colorization techniques and historically shown to over generalize over huge datasets and thus working on smaller dataset with low variance keeps the consistency between the images.

2.4. Deep Colorization

In this [3] paper, the researchers propose a deep learning approach to image colorization where they are approaching the problem as a regression problem and solving it using a regular deep neural network.

Their approach is one of more earlier approaches where they are making systems to extract feature descriptor at each pixel location and these feature descriptors acts as inputs to a neural network which is used to obtain the corresponding chrominance values for each pixels. These chrominance values are further refined to remove any possible artifacts. The input grayscale image is then combined with the chrominance values to get the final results.

The experiment is run on 2688 images of Sun database and it comprises of 47 object categories.

This was one of the earlier papers which proposed this idea of deep colorization and the idea of colorizing images based on the referencing all the image of the dataset and also finding local features instead of just relying on global features for such tasks.

2.5. Automatic Colorization with Deep Convolutional Generative Adversarial Networks

In this [6] proceeding paper, there is an experimental approach to automatic colorization using Deep Convolutional Generative Advarsarial Networks, DCGANs. It is an effort to use GANs over this problem to reduce the generalization that occurs while dealing with this problem where the results seem to be averaged out.

This paper compares the results with a baseline Convolutional Neural network which maps the grayscale images to color image space in a regression problem fashion.

The entire model is trained and evaluated over CIFAR-10 dataset. The fundamental issue with this paper is that it isn't using state-of-the-art networks for comparison with the DCGAN approach, but nonetheless, it does show that GANs can be applied to this problem.

2.6. Real-Time User-Guided Image Colorization with Learned Deep Priors

In this [8] paper, the researchers are using semi-supervised approach to automatic colorization of images which is user-guided.

The priors are based on the user input and the model which they are using is a Convolutional Neural Network which has two types of inputs to it, Global Hints, and Local Hints which are used to guide the network for the output colorization.

The Local Hints are the one's which are guided by the user while the Global Hints are posteriors which the network learns after being trained on the entire dataset. For the experiment, they are using the ImageNet dataset.

2.7. Deep Koalarization: Image Colorization using CNNs and Inception-Resnet-v2

In this [2] paper, the researchers come up with a novel approach of using a fushion layer in order to merge global feature priors with local feature priors.

The high level features in this model are extracted using the pre-trained Inception-ResNet-v2 model and this is fused with the local feature prior which is based on CNNs.

The pre-trained network is trained on the ImageNet and so the local feature encoding network is also trained on a feature of the ImageNet. This simplifies the training of the model. The experiment is evaluated using a user-based study of the output image to check how realistic they look. And they test it by colorizing historical pictures which are grayscale.

2.8. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification

In this [4] paper, the researcher come up with a novel technique for the problem of automatic colorization using a series of deep neural networks which consists of global and local priors as well as a colorization model. This paper is based on the baseline model as described in [3].

The global and the local priors are two different convolutional neural networks which are used to find local features and global features and they resulting outputs are fused together into a fusion vector. This fusion vector is sent to the colorization network which outputs the chrominance of the input image and this could be combined with the original grayscale image in order to get a colored image.

The local and global networks share the weights but the global prior network is trained separately and it uses its own loss function. For the experiment the researchers work on Place scene dataset and as for the evaluation, they ask for user reviews and calculate the median of what each user tells about their output images. Comparing it with the baseline and state-of-the-art results, the model seems to be outperforming severely.

3. Method and Approach

3.1. General Pipeline

For the training of the network, our model reads the images of the pixel dimension 256x256x1 which is a grayscale image which we derive from the luminance channel of the Lab colorspace. We are using the Lab colorspace because it is closer to how the human eye's perceives objects and secondly it gives color channels which gives us to chrominance based channels which we will be using as our output prediction channels and the Luminance channel which can be used as the input channel.

The advantage of using the Luminance channel as the input channel is that, it retains the luminance values of the image pixels and human eyes are more sensitive to detecting edges by seeking contrast in brightness as compared to seeking contrast between color channel, thus this gives proves to be a good channel to work on as it retains the resolution of the image.

Apart from this, during the training time, we also provide images of the size 224x224x1 RGB image as the embedding image vector to the Inception-ResNet-V2 pre-trained network which uses this vector for classification task and this output and concatenated to the output of the main network in the fusion layer.

The output of the fusion layer is then manipulated using the attention mechanism before providing it as an input to the decoder network. The decoder network outputs two color

channel values for the a* and b* colorspace. Since the output from the deep learning model is normalized between -1 to 1, and the Lab colorspace has the a* and b* color channel ranging from -128 to 128, we multiply the output vector result to 128. This gives us the final vector of the two color channels.
Finally, in order to get the correct colored output of the image, we combine the two color channels along with the Luminance channel which we had provided as input. The resulting image is the colored image of the grayscale version which we had provided to the network.

3.2. Activation function

In this model, since the encoder and the decoder network is predominantly based on convolutional neural network layers, we are using the rectified linear unit as the non-linearity that follows after each of the convolutional as well as dense layers in the network. Mathematically, the rectified linear unit (ReLU) is defined as

$$f\left(x\right) = \max\left(0, x\right)$$

Since there is empirical evidence which has been shown in previous State-of-the-art Image classification papers that (ReLU) provides faster convergence in training Convolutional Neural networks it is the go-to activation function for such application.
Since we require an output which is supposed to range between -128 to 128, we need a normalized output which is ranging between -1 to 1. Thus for the last layer of the decoder network, we use a hyperbolic tangent activation function. Tanh activation function is defined as

$$f(x) = 1 - \tanh^2\left(z\right)$$

3.3. Batch Normalization

In our experiments we notice that using Batch Normalization greatly reducing the convergence time and decrease the loss during training. In order to properly apply Batch Normalization, we are using the Batch Normalization layer before every layer showing non-linearity apart from the last few layers in the decoder network. Doing so drastically improves the training rate of the systems.

3.4. Baseline Regression Model

For the baseline model, we use a regression-based model for colorization the image dataset based on the basic Seq2Seq Encoder-Decoder network. This model is used in the paper [3].
In this model, the input encoder network takes the input of 256x256x1 size and converts the input image array into a latent space vector. It is described as 'summarizing' the input image and this 'summarization' is fed to the input of the decoder network, where it is 'creating' the colorized imagery. In the encoder network, after each of the convolutional network, we choose to avoid max-pooling the network because that can lead to loss of details in the image and we may not be able to colorize the images in finer details. Similarly, during the decoder network process, in order to increase the size of the image, we use upscaling of the image. This progressive upscaling of the image from the latent representation given by the encoder helps the network learn propagate the global spatial features to more local features to define the colors in context of the nearby pixels and over a higher abstraction, even over objects and surrounding Markov blanket.

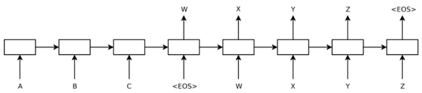

Figure 2. An example of an encoder and a decoder network as shown in [5]

3.5. Final regression-based model

For any sequence-to-sequence translation, the approach to be used is using some modification of the Seq2Seq [5] model, this model helps us translate inputs of a specific fixed length vector to an output of a different sized fixed length vector, thus this form of architecture is perfect for any sequenced inputs and outputs. This is thus used in places like caption generation, audio-to-text, etc.

The fusion mechanism is a novel idea which is used for automatic colorization application in [2]. We use this to combine two networks which are meant to be for local and global priors respectively. We require an approach which is based on local and global priors because the global prior acts as very high-level abstract feature extractor which is useful for the network to understand the image contents, that is, it helps in differentiating the objects in the image from the other objects and the background. The local prior is the main encoding layer of the Seq2Seq model and we fuse the output of the encoding layer with the output of this global prior network to get a better result.

For the fusion layer, we have to attach the feature volume output of the encoder layer and the feature vector from our pre-trained Inception network after replicating it's feature vector a few times. Finally we will use some convolutional kernels to this product so that we get an output of out appropriate size.
As for our own input in this project, we are planning to add an attention mechanism to this network, Attention in term of

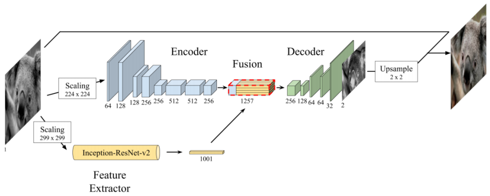

Figure 3. Model to combine global and local prior networks with fusion layer before inputting to the colorization layer as shown in [2]

image-to-image translation is a vector, which is usually the output of dense layer using softmax function. Without attention, the entire image would have to be compressed into a fixed-length vector which is bound to have some information loss, losing important features, etc. The way attention mechanism works is that it comprises of a context vector of all the encoder's output to compute the probability distribution of each element of the source image's vector for each element of the output (decoder's) vector. The math is given by:

Attention weight: $\alpha_{\text{ts}} = \frac{\exp\left(\text{score}\left(h_s, \overline{h_t}\right)\right)}{\sum_{s'=1}^{S} \exp\left(\text{score}\left(h_s, \overline{h'_s}\right)\right)}$
Context vector: $\sum_s \alpha_{ts} \overline{h_s}$
Attention vector: $f\left(c_t, h_t\right) = \tanh\left(W_c\left[c_t; h_t\right]\right)$

3.6. Datasets

In order to train the network, we have considered a few dataset, Table 1 shows the datasets that we considered.

The FloydHub Public dataset consists of 9500 images taken by professional photographers with a validation set of 500 images. Such kind of a dataset is useful in understanding how the network would work in case of real pife photography are thus it seems like a good choice to test the dataset. Further, it has been used by one the the State-of-the-Art methods and it's helpful to compare results with them.

The MIT CVCL Urban and Natural Scene Categories dataset contains thousands of images categorized into 8 categories. For the purpose of this project, we are going to use the images in the ”Open Country” category and measure the model's performance.

To experiment the model with how vast it can generalize we also test it on a subsection of the ImageNet dataset. We randomly sample 60000 images of the ImageNet dataset and 500 for the testing set. We will use these to colorize the image and it can give a strong metric about how well the model is performing.

Most deep automatic colorization models tend have to an issue with high variance and have a problem with generalizing over really huge dataset with high variance among categories. Thus in order to evaluate the model perfectly, we will be testing the model in two ways, first, in which we will be training and testing the model over images which are similar, and second, where we will check how it performs over the entire dataset and that is how we can check it's generalization abilities.

Table 1. Datasets

Due Date	Training set	Testing set
ImageNet Subset	60000	500
FloydHub Unsplash public dataset	9500	500
MIT CVCL	361	50

We have to first preprocess the image before feeding it to the network and reshape each image to the size of 224x224x3 and a grayscale version of 256x256x1 which is converted from the RGB colorspace to the Lab colorspace. Since the embedding model is pre-trained over the ImageNet, it expects the images to have be zero-centered and in order to do so we have to subtract the mean R, G and B values from all the images on the ImageNet dataset in our image dataset.

4. Experiments to be performed

4.1. Experimental setup and model structure

As described in the section above, we use this deep learning model along and train it using a subset of ImageNet. Since it is not possible to train it on the entire ImageNet due to time and computational resource limitation, training it on some of the image should give a good enough result. Apart from this, past research shows that training a deep colorization layer on a smaller dataset which has lower variance gives a better result since training the colorization network over a large network makes it generalize and underfit which leads to gray or brownish images.

To start evaluating the model, we have to first check whether our model is working. In order to do that, we train the model over a very small amount of images and try to overfit it. Further we try to choose the learning rate to start out and realise that the learning rate of 0.001 was giving the best results. We chose the Adam optimizer to train our network since this optimizer converged faster than all other methods to do so. From Hwang et al.[9] we deduce that we should use Xavier initializers for the decoder network and thus we use that for initialization.

4.2. Evaluation Metrics

For training the networks and the model, we shall use the Mean Square Error between each estimated pixel colors in the a*b* space and the target values. Thus MSE is given by,

$$C\left(X, \theta\right) = \frac{1}{2HW} \sum_{k \in a,b} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(X_{ki,j} - \tilde{X}_{ki,j}\right)^2$$

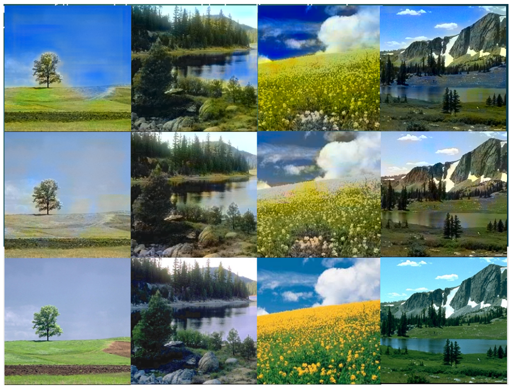

Figure 4. Final model compared to the baseline model and ground truth for the given gray-scale images. The first row is result of our model, the second row is the baseline model. And the third row is the ground truth.

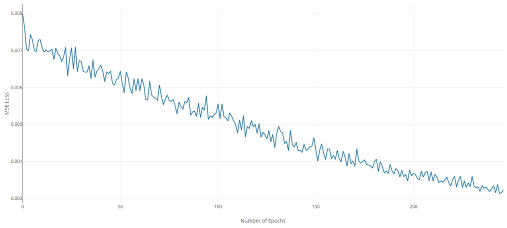

Figure 5. MSE loss over the training dataset

where X is the Image vector and θ represents parameters of the model and $X_{k,j}$ and $\tilde{X}_{k,j}$ represents the pixel of the input and the output image respectively.

5. Results and discussion

In the following figure, we see some examples of how our models works in colorizing gray-scale images as compared to the baseline model. On the extreme right column we have the results of our model and on the middle we have the baseline model. These are the outputs to the gray-scale images which are on the left-most column.

And in the Figure 5 we see the MSE loss over the training data, and Fig 6 gives us the MSE loss for the baseline model,

We notice that our model gives slightly better results as compared to the baseline model, it is more aesthetically pleasing and the color does smear across objects. Even though the baseline model is good at keeping the colors in the right places, there are some defects, which are not so strong in our model.

Note that however, that are still some smear and noise in the final image of our model. This may be a result of several issues like difficulty in understanding the context of the image. For example in the open country image examples, we notice that the grass usually is a combination of patches of green and yellow color. This may be an issue of understanding the season from the image and thus it creates such artifacts.

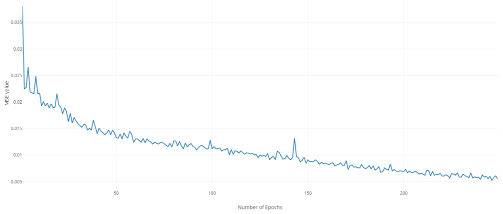

Figure 6. MSE loss over the training dataset

Further, we notice that there are many images in the validation set which have very low saturation or are biased towards brown color since brown color tends to give the lowest MSE values on Lab color-b spaces. Apart from that, the model lags behind in understanding the different colors within a particular image. Thus colored clothes where usually colored mono-chromatically and so on.

6. Conclusion and future work

Through this experiment, we have shown the effectiveness and potential of using attention mechanism based deep learning models for automation colorization of gray scale images. In particular, we have shown that attention mechanism which is traditionally used for sequential data like text or Image-to-text translation can also be used for Image-to-Image translation for such a task and it can yield colorized images that are definitely better than the one's produced by the baseline model, and thus there seems to be a potential promise for further development in this direction for automatic colorization of images as well as Image-to-Image translation using attention mechanism.

Therefore, our work lays a foundation for future work.

Further, we can improve the results by using some probabilistic graphical model based post-processing such as using conditional-random-fields and total variance minimization models to further reduce the amount of color inconsistency within the colorized images. Further, there can be a space to improve in this field by looking into the works of generative adversarial networks which are designed to do similar tasks by putting some efforts into the direction to reduce the loss in such networks since we notice that GANs for such tasks have very high losses and only work on severely small samples of the test dataset. That being said, using our encoder model along with the attention mechanism could have the potential to turn out to be a good generator for a generative adversarial network for such tasks.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014.

[2] F. Baldassarre, D. G. Morín, and L. Rodés-Guirao. Deep koalarization: Image colorization using cnns and inception-resnet-v2. *CoRR*, abs/1712.03400, 2017.

[3] Z. Cheng, Q. Yang, and B. Sheng. Deep colorization. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 415–423, 2015.

[4] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Trans. Graph.*, 35:110:1–110:11, 2016.

[5] Q. V. L. Ilya Sutskever, Oriol Vinyals. Sequence to sequence learning with neural networks. *NIPS*, 2:3104–3112, 2014.

[6] S. Koo. Automatic colorization with deep convolutional generative adversarial networks. 2016.

[7] X. Liang, Z. Su, Y. Xiao, J. Guo, and X. Luo. Deep patch-wise colorization model for grayscale images. In *SIGGRAPH Asia Technical Briefs*, 2016.

[8] R. Zhang, J.-Y. Zhu, P. Isola, X. Geng, A. S. Lin, T. Yu, and A. A. Efros. Real-time user-guided image colorization with learned deep priors. *ACM Trans. Graph.*, 36:119:1–119:11, 2017.

[9] Jeff Hwang. Image Colorization with deep convolutional Neural Networks In ASIAGRAPH 2016, 2016.

[10] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. Rethinking the inception architecture for computer vision. In CoRR, abs/1512.00567, 2015.

[11] He, K., Zhang, X., Ren, S., Sun, J. Deep residual learning for image recognition. In CoRR, abs/1512.03385, 2015.

[12] Simonyan, K., Zisserman, A. In CoRR, abs/1409.1556, 2014.

[13] Rosanne Liu, Joel Lehman Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, Jason Yosinski in CoRR, abs/1807.03247, 2018.