Mathematical Review

Report written by
Robert McNicol

Team Name:
Keeping Up Appearances

Team Members:
Vilius Chockevicius - w18007679
Robert McNicol - w18007336
Dawid Michniuk - w18016997
Dominykas Patinskas - w18016919

Programme of Study:
Computer Science with Artificial Intelligence

Introduction

Review overview:
Firstly, this mathematical review aims to look at an existing formulation of the problem of image colourisation, to understand where a mathematical approach may be useful in the development of our own solution. Next, this review explores the mathematical foundation and approach of a proposed solution of which the baseline is modelled off to provide a mathematical underpinning of the workings and performance of the initial model. From this, the aim of the review is to support the further development of the model throughout the iterative development stage and provide further support and mathematical reasoning for design decisions made as the project progresses.

A note concerning COVID-19
Unfortunately, this review is unable to meet the full potential as outlined in the overview above due to the dropping of the iterative development mission because of the effects the Coronavirus crisis has had on the individuals involved in this project.

The review therefore sets out the mathematical underpinning of the project and how this might have been used for the benefit of deriving better solutions to the problem at hand.

Problem Formulation – an existing approach

“Colorful Image Colorization”, a paper written by Richard Zhang, Phillip Isola, and Alexei A. Efros, quite helpfully formalises the problem of image colourisation for their suggested approach to tackle the issue. Their approach to image colourisation is one that proposes to suggest a possible plausible coloured version of the greyscale image, rather than one that recreates the image colour perfectly. The outcome of the paper is a set of results revealing whether or not their image colourisation method was able to fool the human eye into thinking that the recreated image was in fact the real image, when compared up against the ground truth. It manages to achieve this successfully on 32% of their “colourisation Turing test” trials, which is suggested to be significantly higher than previous existing methods. This paper is therefore deemed appropriate to review in the context of our project and perhaps provide a basis of mathematical findings to keep in mind as the project progresses.

The objective function:
This colourisation method takes place in the Lab colour space, which is a 3-axis colour system with dimension L for lightness and a and b for the colour dimensions, where a is the red/green space and b is the yellow/blue space. Given an input

lightness channel $\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$, the objective of their model is to learn a mapping $\widehat{\mathbf{Y}} = \mathcal{F}(\mathbf{X})$ to the two associated colour channels $\mathbf{Y} \in \mathbb{R}^{H \times W \times 2}$, where $H, W$are image dimensions and the colour dimensions are represented by the integer following H, W. This means that the input image will have lightness channel only (being a greyscale image), and that the aim is to make a prediction $\widehat{\mathbf{Y}}$ from mapping $\mathbf{X}$ (making a mathematical association between the input set, the lightness channel, and the output sets, a and b colour channels) onto the two associated colour channels a and b. Essentially this means deriving the a and b colour channels from the features of the lightness channel.

Due to the fact that the CIE Lab colour space is used, the distances in this space are perceptual distance, meaning a natural objective function used is the Euclidean loss between the predicted and ground truth colours. Euclidean loss is a method which outputs the amount of error, as it measures the distance between two points. In this case, the distance between the colour predicted and the real colour is used to see how close the predicted colour is to the real colour and therefore the performance of the developed method. The Euclidean Loss for the proposed model is given:

$$\mathrm{L}_2(\widehat{\mathrm{Y}}, \mathrm{Y}) = \frac{1}{2} \sum_{h,w} \left\| \mathrm{Y}_{h,w} - \widehat{\mathrm{Y}}_{h,w} \right\|_2^2$$

This is showing that the loss calculated between the predicted $\widehat{\mathbf{Y}}$ and the real $\mathbf{Y}$ is an average over the summation of all the distances between the $\mathbf{Y}_{h,w}$ ($\mathbf{Y}$ height and width) values and the predicted $\widehat{\mathbf{Y}}_{h,w}$ ($\widehat{\mathbf{Y}}$ height and width) values. However, it is suggested due to the nature of the problem that this will not be robust enough for their system. This is because of the ambiguity involves as the aim of the system isn't to recreate the actual colours, but rather suggest a realistic colourisation that is believable to be the original image. It is discussed that if any given object can take on a set of distinct ab values, then the optimal solution that the Euclidean loss will output will be the mean of this set. This essentially means that the nature of the problem allows multiple colour solutions for any given space, and this loss function will take an average of all these solutions. Unfortunately, this means that the overall outcome will result in desaturated images and a loss of detail. Further, given the multimodal nature of the problem, plausible colourations may in fact lie in a non-convex set, meaning that there are multiple feasible regions where the colours could be, but being able to find an optimal solution is also the global optimal solution is incredibly difficult to calculate.

Instead, the problem is no longer considered from a regression point of view, but rather a multinomial classification. The paper goes on to redesign their objective function to fit classification. This mathematical review may cover this approach later on, knowing the dangers of treating the colourisation problem as a regression one, however with the implementation of the baseline underway with a regression solution, this review will not cover this until we discover the performances of the baseline and how we want to continue and develop during the iterative development mission of the project.

Our proposed solution – the mathematical foundations

An outline:

The baseline of the project is based on the alpha version of an existing tutorial titled: "How to colorize black and white photos with just 100 lines of neural network code". This tutorial's core logic is centred around the fact that black and white images can be represented in grids of pixels, where each pixel has a value that corresponds to its brightness level – from black to white (these brightness's represented by values 0-255). Further, colour images are made up of three layers (red, green, and blue layers), where each layer works similarly with corresponding colour values (0-255) that combine together to give the overall pixel colour.

The tutorial sets out to use neural networks to find existing traits in greyscale images that link them with coloured images. Their problem formulised is over-simplified for the reader to: f(B&W) = [R], [G], [B] meaning the following: given an input (which is a black and white image), function f() (which is the neural network) is applied to the input to produce an output of three layers (red, green, and blue) to combine to form an RGB image. It is, however, more complicated than this, as this is just a brief surface level explanation.

The full model:

The full model created in the tutorial after gaining knowledge from the experimentation and learning stages explored throughout the alpha and beta versions is a reproduction of the findings from a paper titled "Deep Koalarization: Image Colorization using CNNs and Inception-Resnet-v2" written by Federico Baldassarre, Diego Gonzalez Morin, and Lucas Rodes-Guirao. Although our baseline is based off of the tutorial's alpha version only, this paper gives some mathematical insight into the thinking behind the workings of the model implemented in their approach, and therefore also our approach.

Images of size $H \times W$are considered in the CIE Lab colour space. This colour space is necessary for the model as it ensures a high detail level in the output image. This is because the colour characteristics are separated from the luminance (which holds the main image features), which when combined result in the final reconstructed image having maintained image quality. The model starts with the "luminance component": $\mathrm{X_L} \in \mathbb{R}^{H\times W\times 1}$ (this is a vector with real coordinates as denoted by $\mathbb{R}^{H\times W\times 1}$, and can be described as the Lightness channel in the Lab colour space). Knowing this information, the model's purpose is to make an estimation of the remaining components, i.e. the a and b components of the Lab colour space, for the prediction and generation of a fully coloured image $\widetilde{\mathrm{X}} \in \mathbb{R}^{H\times W\times 3}$.

The paper defines in short, that there is an assumption that there will be a mapping $\mathcal{F}$ such that $\mathcal{F} : \mathrm{X_L} \rightarrow \left(\widetilde{\mathrm{X}}_\mathrm{a}, \widetilde{\mathrm{X}}_\mathrm{b}\right)$, where $\widetilde{\mathrm{X}}_\mathrm{a}, \widetilde{\mathrm{X}}_\mathrm{b}$ are the a and b components of the reconstructed image in the Lab colour space. This means, in relation to the problem at hand, that for every component in the greyscale image input, there will be a mapping (i.e. there is a mathematical association between the input set and the output sets) to two predicted components: the a and b components of the Lab colour space. When these predicted components are combined with the given luminance component which was used as the input there is the creation of the estimated colour image: $\widetilde{\mathrm{X}} = \left(\mathrm{X}_{\mathrm{L}}, \widetilde{\mathrm{X}}_{\mathrm{a}}, \widetilde{\mathrm{X}}_{\mathrm{b}}\right)$. To be absolutely clear here: This is saying that the predicted colour image $\widetilde{\mathrm{X}}$, is **equal to** (i.e. it has been derived from) the **combination** of the input component X_{L} , the predicted a component $\widetilde{\mathrm{X}}_{\mathrm{a}}$ derived from the mapping $\mathcal{F}$ above, **and** the predicted b component $\widetilde{\mathrm{X}}_{\mathrm{b}}$ also derived from the mapping $\mathcal{F}$.

The paper goes on to define its objective function for the purposes of finding the optimal model parameters, as these are found by the minimisation of the objective function defined over the estimated and target outputs. The mean squared error, or the MSE, is used between the estimated pixel colours in the a*b* space of Lab and the ground truth (or real) values for the quantification of the model loss. To better understand the given MSE for picture $\mathbf{X}$, let's first look at what MSE is. The general formula for MSE is given by:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

This error is calculated by taking all the points in an array from $i = 1$ to $i = n$, and for each of these points subtracting the $\tilde{y}$ coordinate from the y coordinate and squaring the result (to remove the possibility of negatives), and finally taking the sum of all these values and dividing by n, where n is the number of data points. The paper defines that for a picture $\mathbf{X}$**,** the MSE is given by:

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

where $\boldsymbol{\theta}$ represents all the model parameters, $X_{k_{i,j}}$ and $\tilde{X}_{k_{i,j}}$denote respectively the ij:th pixel value of the k:th component of the target and reconstructed image. This means, following the description of general MSE formula to put this into context, that the mean squared error is calculated by taking all the points in the image, every width for each and every height, and for each of these points subtracting the reconstructed (model predicted) image pixel value from the corresponding target (ground truth) image pixel value, squaring this value to remove negatives, and finally taking the sum of all these values and dividing by the number of data points which is $2HW$ (Hight*Width*2 – the 2 being representative for the 2 layers of the colour space a*b*). The value obtained reveals the total loss between the predicted and real ab values, and therefore how closely the reconstructed image matches the original/target image. The paper further describes that the loss over a batch ($\mathcal{B}$) of images can be easily derived by averaging the cost amongst all the images in the batch, providing a good indication to the model performance. This can be done simply by $1/|\mathcal{B}|\sum_{\mathbf{X}\in\mathcal{B}} C(\mathbf{X}, \boldsymbol{\theta})$ I.e. taking the summation of every MSE for each image $\mathbf{X}$ in batch $\mathcal{B}$ and dividing that by the total number of images in the batch. Finally, it is worth noting from the paper that the model parameters $\boldsymbol{\theta}$ are updated using an Adam Optimizer (adaptive learning rate optimization algorithm) as the loss is back propagated through the training phase of the neural network, and this is how the optimum parameters are ultimately determined.

Conclusion

This review seems somewhat unfinished due to the unfortunate circumstances that removed the possibility to implement mathematically design features into the model throughout the iterative development stages. However this review has provided the mathematical foundation of the colourisation problem, which did support the understanding of the baseline and would have been used to improve our model through the correct use of an objective function, whether designed by ourselves or using the loss functions supported by the academia covered in this review.

References

[1] Colorful Image Colorization, Richard Zhang and Phillip Isola and Alexei A. Efros, 2016, 1603.08511

[2] Deep Koalarization: Image Colorization using CNNs and Inception-ResNet-v2, Federico Baldassarre and Diego González Morín and Lucas Rodés-Guirao, 2017, 1712.03400