Locale-Specific Image Colourisation

Rory Ward
College of Science and Engineering
National University of Ireland, Galway
Galway, Ireland
R.Ward15@nuigalway.ie

Prof. John Breslin
College of Science and Engineering
National University of Ireland, Galway
Galway, Ireland
john.breslin@nuigalway.ie

***Abstract*—This paper compares the performance of locale-specifically trained image colourisers to their generically trained counterparts. Two models need to be created, one acting as the locale-specifically trained image colouriser and the other as the generically trained colouriser. The locale-specifically trained image colouriser will be trained on a dataset including Irish domain specific images, and the generically trained image colouriser will be trained on generic images. The popular, open-source, automatic image colourisation system, DeOldify will form the basis of the two colourisation models. The architecture of this system will be investigated. The creation of both the training and testing datasets will be explained, the sources cited and the reason for their inclusion outlined. Both models will be tested on the same test set and the results will be compared. This will allow for like-for-like comparison between the performance of the models quantitatively and qualitatively.**

***Index Terms*—Colorization, Deep learning, Vision for graphics, Video**

I. Introduction

Image colourisation is an area that has received plenty of attention recently in academia as well as in industry. This project plans to explore how the impact of what an image colourisation model is trained on affects the output that it produces. Specifically, whether training the model on locale-specific data will give its resulting images more photo-realistic qualities or not.

II. Related Work

Although the task of hand colouring photographs is quite old, dating back to a swiss painter named Johann Baptist Isenring in 1839 [1], the notion of doing this automatically through the use of Machine Learning is relatively new, and evolving at a very fast rate. Leading the race in this area is a technology called DeOldify [2]. This technology is based on the idea of GAN (Generative Adversarial Network) which has only been introduced since 2014 [3].

This technology is going to be leveraged in this project to construct both the locale-specifically trained image colouriser and the generically trained colouriser. One of the main drawbacks of this technology is its struggle to colourise locale-specific images, for example Irish specific images, as examined in this paper. Similar projects have been undertaken in different locales. Singaporean specific images were tested to see if training the model on locale-specific images improved the quality of the produced images. The next evolution of this technique is colourising black and white videos. This is quite exciting, but also challenging as videos have particular technical issues associated with them. These issues include, greater compute require and generally more complex system architectures such as RNNs (Recurrent Neural Network) [4] or transformer based models [5] instead of the conventional CNN (Convolutional Neural Network) [6] based models which are used in image colourisation. This is outside the scope of this paper though and will not be examined in detail.

The rest of this journal has the following structure. Section 3 describes the methodology followed. Experimental results are shown in Section 4 and conclusions are drawn in Section 5. This is followed by future work recommendations in Section 6.

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

III. Methods

Publicly available sources needed to be found to both train and test the models. There are a plethora of high quality sources available for generic images, these include ImageNet [7], OpenImages [8], OI (Open Images) Extended, OI with Bounding Boxes and COCO (Common Objects in Context). Sources of Ireland specific images were scarcer but some sources used were Ireland's Content Pool, NLI (National Library of Ireland) and the NFC (National Folklore Collection). There were many factors that needed to be considered when deciding which sources would be included and which would not. Some of these include image quality, quantities, and relevance to the desired outcome images.

When the images had been selected and compiled, they needed to be split into two datasets, one for testing and one for training. These datasets were split into a reasonable percentage of 90% training and 10% testing. The images were also split at random to avoid introducing any biases to the results, for example if all the testing images were very similar and not included in the training data, it might unintentionally skew the results. With the testing set and the training set defined it was time to move onto the training stage of the process.

The testing was split up into three main categories, to get an accurate, fair representation of how each of the models perform. These categories are survey, visual comparison, and automated performance evaluation. Both models are tested on the test dataset. This allows for like-for-like comparison

between the performance of the models based upon how photo-realistic the resultant images are.

How many images to use for each model would be crucial to the success of this project. Enough needed to be used to make a reasonable impact on each model, but not so much that the hardware available will not be able to manage the computational demand. The original specifications of the data used for the DeOldify Model is:

Open Images Extended (478,000 images) (90GB)

Open Images Subset (200,000 images) (60GB)

MSCOCO (330,000 images) (19GB)

The additional data for Our model is:

Ireland's Content Pool (101GB)

It was decided that Irish images needed to account for more than a third of the total dataset if they were to make a reasonable impact on the model. Therefore the 101 GB offered by Ireland's Content Pool satisfied this condition at 37.4%, (See Fig. 1).

$$\frac{ICP}{ICP + OIE + OIS + MSCOCO} = Proportion\ of\ Dataset\ made\ up\ of\ ICP\ images$$

Where ICP is the amount of data contributed from Ireland's Content Pool,

OIE is the amount of data contributed from Open Images Extended,

OIS is the amount of data contributed from Open Images Subset,

MSCOCO is the amount of data contributed from MSCOCO

Fig. 1. Calculating the proportion of the dataset contributed by Locale-Specific Images relative to all the images in the dataset.

The testing was split up into three main categories, to get an accurate, fair representation of how each of the models perform. These categories are survey, visual comparison, and automated performance evaluation. Both models are tested on the test dataset. This allows for like-for-like comparison between the performance of the models based upon how photo-realistic the resultant images are.

The first means of comparing a selection of the resultant images was to compile a survey and ask participants which image they thought looked more authentic. Some considerations in compiling the survey included, what questions to ask, what and how many images to include, how to build in randomness and avoid bias, and how to distribute the survey. A simple "which picture looks better" question was decided upon. This was decided upon because the survey was designed to be easy to answer, and therefore more likely to get many participants to take it, and a more diversified and accurate representation as a result. This survey is also meant for a general audience, so any questions that had any pre-requisite understanding of the project were avoided.

A range of images were selected covering a wide range of categories, from portraits of people to landscapes and buildings. This was to get a broader understanding of how the models performed on a wide range of images. The number of images was kept reasonably low, this was to keep the time to complete low and hopefully up the number of participants. Our model and the DeOldify model's image's order were alternated to avoid bias and introduce some randomness into the survey.

The second means of comparing a selection of the resultant images was to use a personal visual comparison. Colourisations of the same image by the different models were placed beside each other and they were compared visually. Obvious mis-colourisations were observed and any trends in the data were investigated.

The third means of comparing a selection of the resultant images was to use compare their PSNR values. To do this the corresponding pixel value of each image was compared to its ground truth value. The resultant PSNR alue was then thresholded to see to what significance the difference between the pictures was. The number of pixels outside of this range is then summated and compared to see which model got the pixel value within the range more often and thus gave a more accurate colourisation. Certain parameters i.e., fuzz factor had to be chosen for the comparison, how this was chosen will now be discussed.

Some testing parameter settings had to be chosen to provide fair, repeatable testing. The two main ones are the render factor used by the models to colourise the images and the fuzz factor used when comparing the image's PSNR values.

The first parameter that needed to be decided was what render factor the models should use to colour the images. Generally, higher render factor results in better precision in the resultant images, (See Fig. 2), but it is more computationally demanding, requiring more RAM. The freely available DeOldify model is limited to below 40, so a value of 25 was chosen for both models to ensure that neither had an unfair advantage in this regard.

The second parameter that needed to be decided was what fuzz factor that ImageMagick would use when comparing the colourised images to the ground truth image. Generally, the lower the fuzz factor the better, as the closer the pixels must be to correct, but too low and neither image gets in the range very often and the results are not that useful.

An important consideration when comparing the colourised images to their ground truths was what performance metric to use. The two main ones considered were Peak Signal to Noise Ratio and Mean Absolute Error.

Power Signal to Noise Ratio is the ratio of the maximum possible pixel value in an image to the mean squared error, where mean squared error is the squared difference between the colourised pixel value and its ground truth value.

Mean Absolute Error is the difference between the colourised pixel value and its ground truth value, normalized by the average error of the whole image. This was the performance metric that was chosen for this comparison.

The results will be discussed in the following section.

IV. Results

The first method of evaluation was the use of a human visual comparison test through the use of a survey. There was good

Fig. 2. Comparing the effect of the render factor on the colourisation results qualitatively.

participation in the survey as reflected in the numbers, (See Fig. 3).

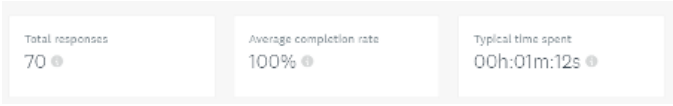

Fig. 3. Showing the survey participation statistics.

The votes for each model were tallied for each image and the results were graphed, (See Fig. 4). As can be seen from the graph, the DeOldify model is generally deemed to have performed better on the images selected in this survey on most pictures.

The votes for each model for each image were then summated and the results were plotted (See Fig. 5). Overall, the DeOldify model is deemed to have performed better than my model throughout the survey.

In addition to the survey, the images were also compared manually, and trends were observed. The main classes of images that were compared were Landscape, Cottage, People and High-Resolution. Each of these headings will now be investigated in greater detail.

Our model exaggerates blue sky and green grass, it is optimistic whereas DeOldify's are dimmer and less vibrant.

Fig. 4. Results of the survey carried out. DeOldify's tallied results per image in orange and our model's tallied results per image in blue.

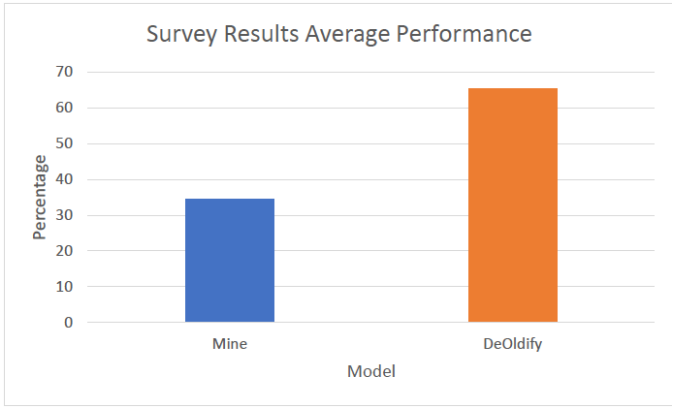

Fig. 5. Results of the survey carried out. DeOldify's total tallied results in orange and our model's total tallied results in blue.

In our model the green bleeds into the white of the rock more than in DeOldify's, (See Fig. 6).

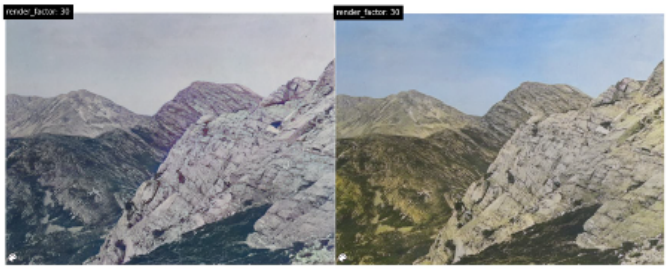

Fig. 6. Qualitative comparison of the two models on a mountain landscape.

The DeOldify Model has very sharp transitions from one object to another whereas our model is more bled. Our model seems more authentic with the spots of purple Heather and golden fur bushes, (See Fig. 7).

DeOldify tends to colour vibrant navy coats. Our model tends more towards softer natural colours and fabrics. In attempting to be very colourful, DeOldify can have inconsistencies such as one the body of the cart which has a much more colourful area, (as highlighted below), just under the

Fig. 7. Qualitative comparison of the two models on an image of a cottage.

child's arm than anywhere else. Our model's image uses more conservative colours but is relatively consistent, (See Fig. 8).

Fig. 8. Qualitative comparison of the two models on an image containing people.

Both models struggled with high-resolution images. DeOldify's colourisation was more monochrome and less vibrant. Our model used bolder colours on the sky and the rock. It did a relatively good job with the sky but struggled with transitions, stopping the sky too far from the lion and bleeding some of the green from the rock into the lion's mane, (See Fig. 9).

Fig. 9. Qualitative comparison of the two models on a high quality image.

The third and final way that the images were compared was using Automated Performance Evaluation. The colourised images were compared against their respective ground truths, (See Fig. 10, Fig. 11).

The pixels in each image that were outside of the range were then overlayed onto each respective Difference Graph, (See Fig. 14).

The pixels out of the range for each model were then summed up and graphed, as well as a selection of other image's differences, calculated in the same way, (See Fig. 13).

Fig. 10. Qualitative comparison of DeOldify to it's respective ground truth.

Fig. 11. Qualitative comparison of our model to it's respective ground truth.

The differences for each image were then summated to give the differences for each model and plotted, (See Fig. 14).

From the graphs, the DeOldify model outperformed our model, but by a far smaller margin than what was seen in the survey results.

With the results described, it is now time to conclude.

V. Conclusion

Three main conclusions have been drawn from this project, they are related to, how Locale-Specific deep learning colouristation compares to generic deep learning colourisation, the effect of training data on colouriser performance, and the specificity of machine learning models. The first main conclusion drawn is that locale-specific deep learning colourisation can perform as well if not better than their generically trained counterparts. This is justified by the performance of our model in the survey, the visual inspection and particularly in the automated performance evaluation. Across these three areas of comparison, there has been little difference in the results obtained from both models. The second main conclusion drawn is that what a machine learning model is trained on has a huge effect on how the model will perform during testing. The variance of a model is also very important when characterising it as well as its possible bias. The third and final conclusion drawn is that machine learning models trained specifically to perform a specific task can perform better in that specific task, e.g., colourise locale-specific images, then a general model which may perform better at the more general task, e.g., colourise generic images. With the Conclusions described, it is now time to outline the recommendations for future work.

VI. Future Work

There are two main areas that the project scope could be expanded into, these are the area of video colourisation and the idea of building a bespoke model using different architectures

Fig. 12. Difference Graph between our model and DeOldify.

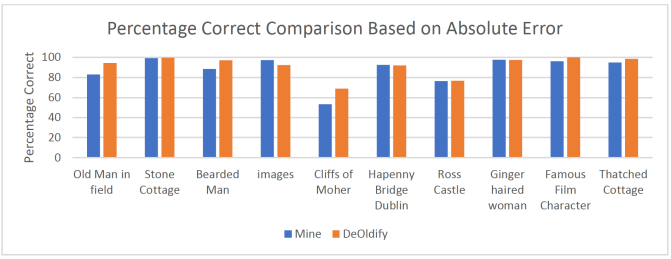

Fig. 13. Percentage Correct Comparison Based on Absolute Error of the two models per image.

to see if more performant models can be created. Firstly, video colourisation. This project has been mainly focused on image colourisation, but what about old black and white videos. There are multitudes of old video archives which could also be colourised to give some of the same advantages as image colourisation, such as greater understanding and feeling of connection towards old content. Video colourisation would add some additional complexity in that the models would also have to deal with the concept of time and progression. Video can be decomposed into a collection of images or frames but as most of the time a lot of the image stays constant between consecutive frames, the model would need to be able to remember past frames when colourising new ones to ensure consistency. This is generally achieved using Recurrent Neural Networks as opposed to Convolutional Neural Networks used in image colourisation. Secondly, building bespoke models using different architectures. For this project, the DeOldify model was chosen to be the base architecture, and the investigation was of the effect of training data on this model, but what if other models may perhaps perform better than this model. More investigation could be done into what makes a good model and perhaps what could be improved to create an even more accurate model.

Acknowledgment

We would like to thank the Irish Centre for High End Computing (ICHEC) for their computing resources.

References

[1] H. K. Henisch and B. A. Henisch, *The Painted Photograph, 1839-1914: Origins, Techniques, Aspirations*. Heinz K. Henisch and Bridget A. Henisch. University Park, Pa: Pennsylvania State University Press, 1996.
[2] J. Antic, "Deoldify," 2019.
[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

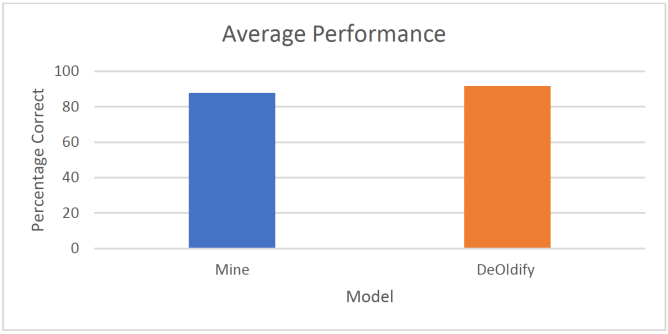

Fig. 14. Percentage Correct Comparison Based on Absolute Error of the two models summated.

[4] Y. Qian, K. Chen, J. Nikkanen, J.-K. Kämäräinen, and J. Matas, "Recurrent color constancy," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5459–5467.
[5] M. Kumar, D. Weissenborn, and N. Kalchbrenner, "Colorization transformer," 2021. [Online]. Available: https://arxiv.org/abs/2102.04432
[6] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," *CoRR*, vol. abs/1603.08511, 2016. [Online]. Available: http://arxiv.org/abs/1603.08511
[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[8] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, mar 2020. [Online]. Available: