MLP CW 4: Colouring black and white images using conditional Generative Adversarial Networks

G71: s1570541, s1725186, s1773394

Abstract

This project studies conditional Generative Adversarial Networks (cGANs) and their application as a photograph colouring technique as well and proposed an improvement to the algorithm proposed by Philip Isola (2017).

This final report proposes a new model based on the idea "divide and conquer" where individual colouring models are trained for specific categorised data and two metrics are selected to evaluate its performance. It is concluded that an assembled model not only reduces considerably the training time of the model, but also outperforms the general cGAN model. There are still improvement to be made and the good results obtained so far give motivation to continue investigating this approach further.

1. Introduction

1.1. Motivation

The objective of this project is to colourise grayscale images using conditional Generative Adversarial Networks (cGANs) based on the model proposed by Philip Isola (2017) to obtain an image colourisation that looks realistic.

In the interim report we studied the cGAN model proposed by Philip Isola (2017) and its implementation as a grayscale image colouring technique, exploring some of the model's parameters and comparing its performance against the TensorFlow project developed by Dahl (2016) that adapted the structure of the VGG16 classification model Karen Simonyan (2014). When comparing our best model to the VGG16 colourisation one we discovered that our model outperformed it, providing a wider variety of tints and colours closest to the original images instead of the persistent green and brown tints that the VGG16 model insisted on applying to all images alike.

As expected, we discovered that the bigger the size of the training set, the better the colourisation of the images was, but even when using our biggest training set (1000 images), the colourisation was still very pale in comparison to the real colours of the image, regardless of the number of epochs the model was run for, and since we had shown that the training time of the model was not linear with respect to the size of the training set, it made it unfeasible to train a better model with the current resources and time limitations we had.

To address this problem, on this final report we implement the modification to the original cGAN colouring algorithm proposed in the objective number four of the interim report and restated here in Section 1.2 to improve the model's training time without affecting its performance. The main idea behind this new model is 'divide and conquer':

Based on one of the hypothesis formulated on the interim report which states that when restricting a model to a certain type of images with similar colours and shapes, its performance on this type of images will be better than other more general models, we propose to build an assembled model composed of several smaller and more specialised models, and use an image classifier model to send each image to the colouring model that best fits its characteristics.

This new approach allows us to train the models in parallel because they are independent from each other and in just a fraction of the time it would take to train one model with all the images, not only because of the parallel implementation of each part, but because of the computing time scaling problem mentioned before.

To maintain the most continuity possible with the methods and datasets used from the interim report, we decided to use the VGG16 classification method Karen Simonyan (2014) as the image classifier model and obtained the subsets of images for each model from the ImageNet dataset Olga Russakovsky (2015), this time form the group of predefined classes or *synsets* identified by a unique WordNet ID (wnid), instead of selecting them from the general image archive.

To evaluate the performance of this model we will compare it to two baseline models: the first one will be the same baseline from the interim report, the VGG16 colourisation model, and the second will be the original cGAN colourisation model, without the different categories. Performance evaluation is not a straightforward task, as we commented on the interim report before we reviewed the most popular methods used, so based on that review, for this last part of the project we decided to use two metrics that we believe complement each other, which are visual inspection and a semantic classifier.

1.2. Research questions and objectives

Can segmenting images into classes based on similarity improve the colourisation quality of the model? Exploiting the different modalities for downloading data from the ImageNet dataset, can training models specialised in different classes and using a classification model to send the images to their corresponding model improve the colourisation of the images? How does the performance of the models compare to a general model? How much does an error of the classification model alter the colourisation of the image?

When dividing models by classes, do different classes need

different model specifications? Does the similarity of the images within a class determine the number of epochs needed to obtain the best colourisation? Is an increase in the stochasticity of the outcome of the model prejudicial to some models and beneficial to others?

Can we alter the model's parameters to reduce the training time? Can we modify the original model to receive smaller sized images reducing the number of operations needed to process them? How much does the batch size affect training time?

Are the results obtained from the two metrics chosen consistent? The nature of the two performance metrics chosen, visual inspection and a semantic classifier, are very different, since one depends on people's opinion and the other on the performance of a model when receiving the colourisation as input so, will the metrics obtained by them be consistent? Will they complement or downright contradict each other? Is the semantic classifier's sensitivity to different colourisations enough to affect the performance of the model?

2. Methodology

2.1. Dataset

Given a WordNet ID, it is fairly simple to download its corresponding images programatically, we need to retrieve the list of image urls from *http://www.image-net.org/api/text/imagenet.synset.geturls?wnid=* adding the corresponding wnid, randomly select a sample of rows of the desired size and download the corresponding images. After this, a search has to be done for corrupted or no longer available images to replace them with newly sampled urls from the remaining list of unused urls until the desired sample size is reached with valid images. This has to be repeated for all the desired classes.

The difficult part of creating this group of datasets is selecting these classes, since just a small portion of the ImageNet dataset is labelled so only very general classes have enough images for our model, and at the same time, the classification model has a different and much smaller set of labels, so equivalences between both sets of labels need to be defined in a way that each class contains enough images and is as relevant as possible.

This was especially troublesome because most of the labels the classification model has correspond to animals or objects, turning our original plan of including a 'landscapes' class unfeasible and a second class corresponding to 'people' less straightforward as we had thought it would be.

We believe having a 'people' class is important for the model because most of the images we had sampled for our previous model contained people, proving the dominance of pictures of people on the web, and also because it's an especially relevant category in the field of image colouring, since both black and white movies and photographs contain images of people predominantly. So in order to create this category we selected all the VGG16 model's labels related to items of clothing and analysed how well they detected images with people on them.

To test how well this equivalence in labels was between ImageNet and VGG16 we needed to calculate the type 1 and type 2 errors for this class, which meant that we needed to choose the other classes we would use for the model. We decided we would implement this model with only three classes, using this project as a small proof of concept, and that the other two classes would be 'animals' and 'vehicles', providing three very different but relevant classes, we then created a mapping from the VGG16 labels to their corresponding class, removed all the labels that did not belong to any of these and adapted the classification model to output the class corresponding to the label with the highest score.

100 images were downloaded for the 'people' class using the procedure explained above and were sent to the classifier model, from these, only 60 images were classified as people, but analysing the misclassified ones, it was easy to notice that the reason why they hadn't been classified as human was because there weren't any items of clothing visible on the images (in some cases it was a picture of just a face and in some others, the image of the person was very far away), luckily, in these cases the probability the model assigned to all of the classes was very low because it couldn't detect any other labels as well, so we decided to incorporate a threshold into the classifier, and if all the classes were given a probability less than 0.05, then the image would be classified as person. With this new classification criteria the percentage of people classified correctly increased from 60% to 85%.

Finally, to make sure that the model wasn't classifying as people more images than it should, we downloaded 100 random images from each of the other classes and obtained the following confusion matrix with their classification results, getting a type 1 error of 15%, a type 2 error of 13% and a 90.6% accuracy in the whole classification model.

	Animal	Person	Vehicle
Animal	100	0	0
Person	11	85	4
Vehicle	0	13	87

One problem we found for this classes was that, although they were some of the more general classes, they each contained less than 1000 images when removing all the corrupted ones, limiting the size of the models we could train with them.

2.2. GANs and cGANs (recap)

Generative Adversarial Networks were first proposed by Ian Goodfellow (2014) as models that learns a mapping from a random noise vector z to an output image y, i.e. $G : z \rightarrow y$ by building on the concept from a zero-sum game; a generator G is competing against a discriminator D, where the task for G is to generate images as similar as the real ones as possible, and at the same time, the task of D is to be able to distinguish if the input it receives is real or not.

Conditional GANs (cGANs) receive both the observed image x and a random noise vector z as input and learns a mapping to an output image y, i.e. $G : \{x, z\} \rightarrow y$. The objective is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y\sim p_{\text{data}}(x,y)}[\log D(x, y)] \\ + \mathbb{E}_{x\sim p_{\text{data}}(x),z\sim p_z(z)}[\log(1 - D(x, G(x, z))]\end{aligned} \tag{1}$$

L1 regularisation is applied, since the goal is not only to trick D but also to stay close to the ground truth and together, G tries to

minimise and D tries to maximise the objective given, therefore we have:

$$G^* = \arg\min_G \max_D \mathcal{L}_{\text{cGAN}}(G, D) + \lambda \mathcal{L}_{\text{L1}}(G). \quad (2)$$

2.2.1. Architecture of the general model (recap)

The implementation used for what we denote as the General model and for the three individual class models was obtained from Chenlin (2016–2017), which in turn was based on the cGAN architecture proposed by Philip Isola (2017). The architecture of both the Generator and Discriminator, the general model and parameter values are described more formally and with more detail in the interim report.

2.3. VGG-16 as a colourisation method (recap)

The VGG classification model proposed by Karen Simonyan (2014) modifies the traditional structure of Convolutional Neural Networks by decreasing the size of the convolution filters and increasing the depth of the network, obtaining networks with 16 to 19 layers and filters of size 3x3. This architecture has been widely used in image recognition tasks with very good results.

The implementation of the VGG-16 model as a colourisation method was proposed and developed by Dahl (2016) extracting information about image structure from the original, previously trained VGG-16 model's intermediate feature layers, using the *hypercolumns* of the model as pixel descriptors and training a model on top of this to reconstruct the colour of an image using YUV colour encoding.

2.4. Architecture of the Assembled model

For this implementation we are using labelled images both in the training and validation sets, so the task of sending each image to their correct colourisation model is straightforward, nonetheless, we want to study a more general approach to this problem in which unlabelled images can be coloured as well. Because of this, our colourisation method needs to have an image classification model to process the incoming images and send them to their corresponding model. A diagram illustrating this architecture can be found in figure 1.

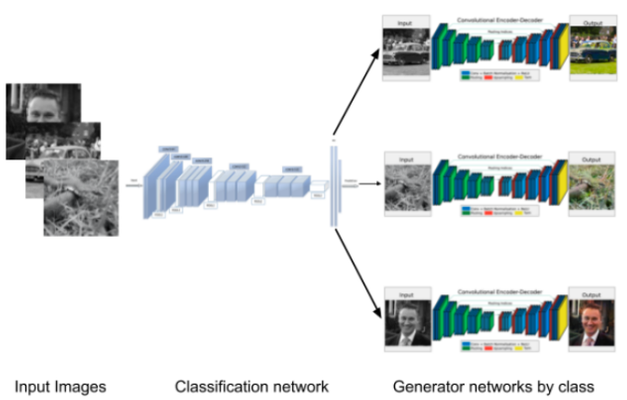

Figure 1. Architecture of our assembled model. An image classification model receives the images and sends them to their closest related colourisation algorithm.

The classification model selected was the VGG-16 pre-trained model, which has been reported to achieve a 92.7% accuracy in the ImageNet dataset (Karen Simonyan, 2014), using the classes previously described in section 2.1.

2.5. Evaluation Metrics

The overall objective of automating colourisation of black and white images is to make the colourised images credible and natural looking from a human perspective. This is therefore the essence of what we want to capture when evaluating the performance of the resulting images, however, finding an evaluation metric that accomplishes this in a concise, robust and cost free manner is complicated.

In the interim report, the most popular metrics used for the colourisation task were discussed and their advantages and disadvantages were weighted, but the only one implemented was an informal version of the visual inspection metric carried out by the three of us in an attempt to understand the colourisation performed by each method, discovering trends, their strengths and weaknesses, and allowing the tuning of different parameters. Although this approach is prone to a certain level of subjectivity, it was good enough for tuning the different parameters of the model, finding patterns the model incurred in when overfitting and selecting what we thought was the best performing model.

For this final report, this informal visual inspection will be used to select the best number of epochs and images for each individual model, as done previously, but a more rigorous approach will be taken to perform the method comparison using two performance metrics: the first one will be a formalisation of the visual analysis and the second one will be the semantic classifier. Both metrics and their implementations are described in the section below.

2.5.1. Visual Inspection Survey

This more formal approach to the visual inspection metric was used by Iizuka et al. (2016) to evaluate their colourisation algorithm by conducting a user study asking if the coloured images were considered natural or not. Inspired by this, we created an internet survey with the aim of evaluating the performance of the assembled model, the VGG-16 colourisation model and the general cGAN model against each other.

The main advantage of this approached is the direct way in which it answers the main aspect we are trying to measure; do the colourisations look natural to people? However, this strength also becomes a disadvantage, since because of its nature, it relies on people's perception, which makes it subjective.

This disadvantage can be regulated through the number of people included in the study and the number of images they are asked to inspect. This is because of the Law of large numbers of probability theory, which states that when an experiment is repeated a large number of times, the average of the results will be close to the expected value of the true distribution underlying the data, and as the number of experiments (in this case people taking part of the study and the number of questions in the study) increases, the better the approximation becomes.

Nonetheless, a new disadvantage arises from this, which is that performing a big enough number of experiments is costly and many times the number of times the experiment needs to be

repeated to attain the desired level of certainty is not feasible.

We created the survey using the free version of the online platform Survey Monkey www.surveymonkey.com, and although it was the best option we found, it came with two disadvantages: the first one was that only ten questions with images were allowed per survey and the second one was that only the first 100 answers are recorded. We considered 100 answers per question were enough to obtain reliable results but that ten questions were not enough, so we created two separate surveys with nine question in each in order to have the questions balanced between the three classes, and got 100 people to answer both.

Each question of the survey had the same format: the three different colourisations of a single image were presented and the question asked was "*Which image do you think looks most natural?*", to avoid any kind of bias the images were shuffled in every display of the survey. The 18 images are balanced between the three classes selected for the assembled model and each of these images were randomly selected from the corresponding test set, correcting the sample only in cases were the colouring of the image was especially bad for all the models since we thought that kind of cases would not provide any useful information regarding which model is better.

The surveys were distributed through our social media accounts and it was answered by a wide variety of people from various nationalities and backgrounds, although predominantly by people between 20 and 30 years old. [1]

The surveys are still open, although no more information is being recorded, and can be found on the following links (survey 1 and survey 2.)

2.5.2. Semantic classifier

The idea behind this pseudo-metric developed by Jonathan Long (2015) is that the performance of an image semantic classifier depends on the quality of the colourisation of an image, so the score obtained by the classification method can be used as a score to measure the quality of the colourisation too. The problem with this metric is that images need to be labelled, but luckily, for this second part of the assignment we have segmented the images into pre established classes and even implemented a classification model to send the test images to their corresponding colourisation model, so taking advantage of this, we are going to use the VGG-16 classification method as semantic classifier, modifying the model to receive the corresponding class of the image as part of the input and instead of outputting the most probable class, it will output the highest score given by all the labels related to the original class of the image.

Performing this task for all three colourisations of the images and then grouping them by image ID, it is possible to compare the colourisation quality of each method, either by comparing which method obtained the highest score in the most amount of images, or comparing each models statistics, such as mean, median or variance, to analyse the average performance of the model. As a baseline against to compare how good or bad a colourisation is, the original grayscale images were also processed and their classification score obtained.

3. Experiments

The experiments are divided in two phases: on the first phase the three class models are trained individually and the aim is to find and justify the best model for each class, and on the second phase experiments are carried out to compare the performance of the assembled model with our two baseline models.

3.1. Experiments and results on the class specific cGAN models

This section reports the experiments conducted on the three individual class models for *Animals*, *People* and *Vehicles*. Some of the results are class dependent while other generalise to all the class models.

Unless stated otherwise, all models were trained from scratch using all available labelled images (excluding the ones for the validation and test sets) with the following setup: weights are initialised from a Gaussian distribution with zero mean and standard deviation 0.02, L1 regularisation parameter is $\lambda = 100$, dropout rate is $p = 0.5$, batch size of 5 and input images of size 256x256 pixels.

The best models for the three classes where: 747 images trained for 350 epochs for Animals, 884 images trained for 200 epochs for Vehicles and 780 images trained for 300 epochs for People.

3.1.1. Reducing the number of pixels

One intuitive approach to reduce training time is to reduce the number on pixels in the input images, so we modified the original implementation of pix2pix Chenlin (2016–2017) to receive different sizes of images and resized the training images to 128x128 pixels.

While this modification in fact reduced the training time by $\sim 50\%$, it also lowered the quality of the colourisation. For example, for the vehicle class two patterns were detected: a high proportion of images were colourised with sepia tones while another high proportion were colourised with red and blue tones. Figure 2 shows samples of these patterns for a model trained with 500 images for 250 epochs. This patterns suggest there might be some kind of mode collapse, and makes this faster model undesirable.

Figure 2. Sample of validation images for the vehicle class when models are trained with images of dimensions 128x128 pixels. *Top*: samples with sepia tones. *Bottom*: samples with red and blue tones.

3.1.2. Varying batch size

Experiments were run on the animal class model on three different batch sizes: 1, 5 and 10 in order to investigate the impact on training time as well as how this influences the colourisation.

For each batch size, the model was evaluated every 50 epochs and the best colourisation was selected performing visual inspection. Due to the space constraint in this report we are only going to show images of the best model for each number of epochs.

Batch size 1 When training the model with a batch of a single image, the best colourisation seems to be reached at epoch 50, even though the image still has a predominantly sepia colouring, because when the epochs to increase, the colourisation becomes less stable, presenting green hues in random regions of the images at 100 epochs and red ones at 150. This model took just over 8 hours to train 50 epochs. [2]

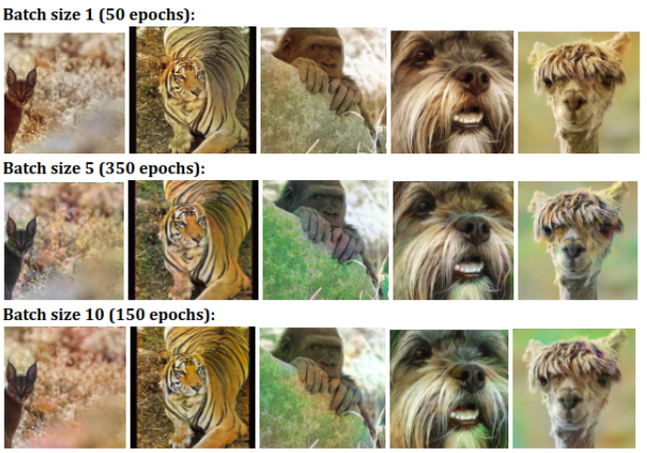

Figure 3. Best models for batch size 1, 5 and 10

Batch size 5 Experiments on batch size 5 were trained from 50 to 350 epochs, the reason this experiment was prolonged until much bigger epochs than the last one is that it continued to show signs of improvement through the whole training process. The first thing to notice with this batch size is that the training time was reduced approximately by half, taking approximately 4 hours to train 50 epochs, and the second thing to notice is the improvement in the colourisation: regardless of the amount of epochs, the images have stronger and more varied colours. The best model was the one with 350 epochs.

This more intense colourisation comes with the risk that images or regions of images that are coloured incorrectly are going to have a stronger (incorrect) colourisation and the errors will be easier to notice.

Batch size 10 The best number of epochs when training a model with a batch size of 10 is 150, in general there doesn't seem to be a big difference between the results of the model with batch size 5 and this model, however, when examining the validation set in more detail, it was noticed that the model had become less stable in terms of strong, wrongfully coloured sections of the image (like the green hues on the chin of the dog), so even though using a batch size of 10 reduces the training time down another 25% (approximately 50 epochs in three hours), the loss in stability of the colourisation makes the model with batch size 5 the best model of the three.

3.1.3. Epochs and Training time

Each of the three class models were run for different number of epochs and the best model was selected, analysing the results we discovered that strong colourisations were accomplished from as little as two epochs, but as the epochs increased, the colourisation would become more stable and predictable, and would also begin to converge to a specific colouring, slowing the learning process as the batches increase, as we can see in the background of the first image in 4.

This similarity between the results obtained during the first few epochs and a much larger number of epochs made us consider selecting the smaller model as the best one, especially when taking into account the difference in training time these two presented, the first model needing less than an hour and the second one over 30 hours, but we decided to select the bigger model because we believe a reliable and stable colourisation is very important, especially when incorporating the models into our classification model and colouring a wide variety of new, possibly missclassified images.

Figure 4. Colourisations after different number of epochs for the People model with 780 images

3.1.4. Failures in the models

As mentioned in Section 3.1.2, using a batch size of five gave stronger colourisations but also presented some strong, incorrect colours in some areas of the images. Some examples of this phenomena are presented in Figure 5.

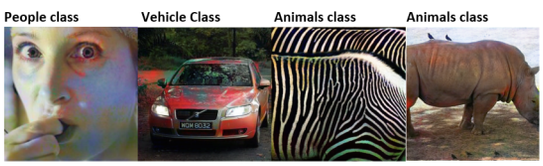

Figure 5. Intensely unnatural coloured areas using batch size of 5 in the different class models

Another problem we discovered with the colourisations was that the models were too specialised in their corresponding classes, so when a 'hybrid' image was presented, such as people in the animal class, the models performed poorly, as can be seen in Figure 6, and other problem we believe can be related is that they tend to leave the image background with a grayish colour, especially with sky and sometimes grass, perhaps because it didn't have enough training images with these elements.

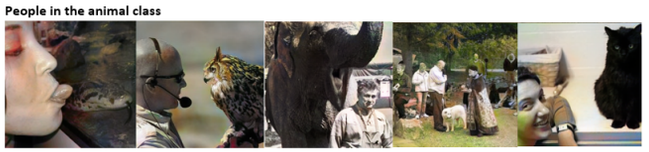

Figure 6. The colourising of people in the animal class

3.2. Comparison of assembled model with baselines

For the model comparison, a General model was trained with all the images used for the three classes (2,411) for a total of 120 epochs, using the same parameters used for each of the individual models, including a batch size of 5, and using as test set the union of the three class test sets. This general model took 32 hours to train. The VGG16 pre-trained model was the same that was used for the interim report.

Figure 7 shows colourings performed by the VGG-16 model in eight of the images used for the visual inspection survey, all colourisations look good, with natural colours and no evident misscolourisations, but at the same time it can be noticed that the colouring is generally quite weak, using only green and brown hues and no strong colours.

Figure 7. Example on images from the VGG16 model used in the visual survey

The colourisation of the general model for the same set of images is presented on Figure 8 where we can notice some of the problems we had previously observed like the red and green hues present near some borders of the image, nonetheless, this model uses a stronger and more varied colourisation than the VGG-16 model, including red, blue and yellow colours to its palette and using more intense colours.

Figure 8. Example on images from the general cGAN model used in the visual survey

Figure 9 shows the colourisation of the assembled model, this model presents much stronger colourisations, especially with vehicles, which seems reasonable, since it has a specific model dedicated to them, but seems to have problems colouring faces with more than a light pink tint and also fails to recognise grass and the green corresponding to plants in most images.

Figure 9. Example on images from the Assembled model used in the visual survey

3.2.1. Visual inspection survey results

This results were obtained from the two surveys performed each with 9 images and 100 responses, obtaining six images per class in total. The median time spent on answering each survey was 1.46 minutes and the completion rate was of 99%.

	VGG-16	**Assembled**	**General**
Votes	59.97%	29.93%	10.11%
Win count	11	7	0
Loss count	1	3	13

Table 1. Results from the visual inspection survey summarised by model

A summary of the performance of each model can be seen in Table 1, which shows that the best performing model was VGG-16, followed by the Assembled model and the General model in the end, however, this absolute count of best models does not include the distribution of the votes that happened underneath, so to study that,Table 2 shows the mean and standard deviation of each of model.

	VGG-16	**Assembled**	**General**
Mean	59.7	29.8	10.1
SD	25.0	23.3	13.7

Table 2. Mean and standard deviation on the votes in the survey

Splitting the votes between the different classes to analyse if there were patterns specific to a certain class we discovered that although the VGG-16 model got the majority of votes in all classes, the votes were more balanced in the Vehicle class, with an almost equal share of the votes between the assembled model and the VGG-16, and this model was also the one with the most amount of votes given to the General model. The People class had the most similar behaviour to the general behaviour with almost 30% of the votes given to pipeline and less than 10% to General.

3.2.2. Semantic classifier results

A similar approach to the visual inspection survey analysis was taken for the semantic classifier results. Figure 11 displays a raw count of the model with the highest score for each question and then a set of boxplots, each corresponding to a different model, display the median and variance of each model. As a baseline, the original uncoloured images were included as a

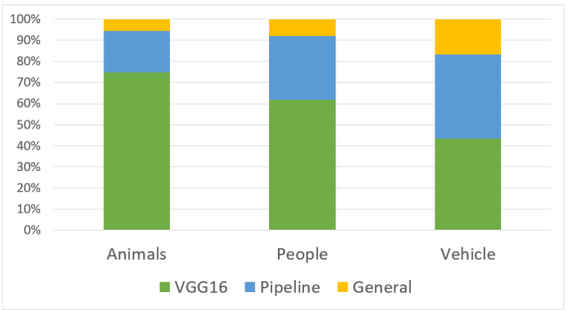

Figure 10. The accumulated responses grouped by label in the visual insepction survey

fourth model (which could be interpreted as an identity) to be able to see if the differences between models were significant.

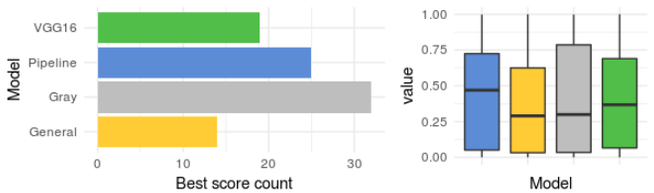

Figure 11. Count of the number of times a specific model was given the highest classification score of all colourisation models for a specific image (left) and boxplot of the classification scores given to each model (right)

Interestingly, the best performing model was the Gray one, which means that the semantic classifier is extremely sensitive to incorrect colourisations, followed by the assembled model, then the VGG-16 and in the end the General model, which backs up our theory of incorrect colourisation sensitivity since the General model was the most prone to add patches of strong, not related to the image, colours and VGG-16 coloured everything with similar tones.

	Assembled	General	VGG16	Grayscale
Mean	0.4254	0.3665	0.3962	0.4155
Median	0.4694	0.2897	0.3676	0.2992
SD	0.3136	0.3074	0.3339	0.3120
Best count	25	14	19	32

The distributions are generally symmetric, so the mean and median are similar, it is interesting to notice that the different models' standard deviations are also very similar. When comparing the best model by image segmented by classes we observe that they all have a very similar behaviour, always the Gray model outperforms the others, and the General model has the worst performance.

3.2.3. Discussion on the results form the two metrics

The visual inspection survey showed that the VGG-16 classification is preferred to the assembled model and these two methods greatly outperform the general model, this contradicts our results from the interim report and we believe this happened for two reasons, the first one is that a lot of images displayed on the survey had grass or plants around, which is the things that this model colours the best, and the second reason is that

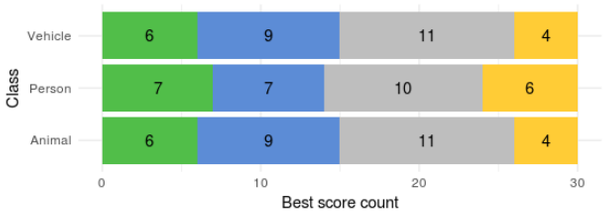

Figure 12. Count of the number of times a specific model was given the highest classification score of all colourisation models for a specific image segmented by image class

when analysing multiple images colourised by the VGG-16 model together, it is easy to notice the colouring patterns and the strong reliability on the green and brown palettes, but when analysing images separately, this patterns are not noticeable and the colouring seems very natural.

The semantic classifier metric proved to be extremely sensitive to bad colourisation, so models where the colourisation is not very stable, like the General model, will perform very poorly, but also light colourisations with incorrect colours, like the VGG-16 will perform poorly, so a balanced colourisation without extreme colours but a balanced palette seems to provide the best results. Nonetheless, it is important to notice that the original images outperformed all the others, leaving a lot of room for improvement.

Combining the results obtained from both metrics we can conclude that the assembled model clearly outperforms the general model, which proves that our hypothesis of "divide and conquer" is a way to improve not only the training time of the model, but also its performance.

4. Related Work

In this project we have followed the work developed in Philip Isola (2017) by using cGANs to perform image-to-image translation tasks. The results reported there to the specific task of colourisation are limited in the sense that one single a experiment is reported over the entire ImageNet training set. We have focused on this specific task and we have extended their work in four directions: (the objective numbers refer to the ones stated in the interim report)

1. While a fully convolutional model and specifically engineered to do well on colourisation developed by Richard Zhang (2016) was used by Philip Isola (2017) as baseline, we have used a different baseline provided by the VGG-16 transfer learning based model by Dahl (2016), by means of objective 1 of this project.

2. We have performed more experimentation and analysed the behaviour of cGANs for the colourisation task by means of the objectives 2 and 3.

3. We have proposed a simple but effective approach following objective 4 by taking advantage of parallelism and by stacking a powerful pre-trained classification network to our assembled model.

4. While the **FCN-8s** architecture developed by Jonathan Long (2015) was used by Philip Isola (2017) to score the colourisation of an image by the classification accuracy of the model, we have

explored two different evaluation metrics for our models and baseline by formalising the visual inspection and by adapting the FCN-8s score using the VGG-16 classification method as semantic classifier.

While GANs are best known for generating very realistic images, the frequently used datasets together with GANs are MNIST [3], CIFAR10 [4], CelebA [5], and SVHN [6]. On the other hand, ImageNet is one of the standard de facto datasets in visual object recognition research. Nevertheless, works that use GANs (or its multiple variations) together with the ImageNet datasets are not very common. As noted by Goodfellow [7], making GANs work on thousands of categories is challenging and works where GANs work well on ImageNet are very new. Examples of those works are Takeru Miyato (2018) and Toshiki Kataoka (2018). Regarding Neural Networks based approaches to colourisation tasks, convolutional approaches are the most frequent. Examples of those approaches are Richard Zhang (2016) and Dahl (2016), which are the baselines used by Philip Isola (2017) who was one of the first ones to address the colourisation task by using an adversarial approach. By proposing an assembled model constituted of models dependent on class labels, the novelty of our approach is to address the challenge of working with a high diversity of categories in the ImageNet dataset and combining an adversarial approach, even if we were only able to provide a small proof of concept. After completing this project, we highlight the following ideas as possible lines of future research:

1. Extend our proposed assembled model would be a straightforward task to allow for more ImageNet class labels to be considered. On the other hand, while we have used the VGG-16 network for the classification stage, other award-winning pretrained models can be considered for this stage, such as AlexNet (Alex Krizhevsky (2012)) or Inception (Christian Szegedy (2014)).

2. Authors in Philip Isola (2017) pointed out noise ignorance while training the generator and provided noise only in the form of dropout applied on several layers of the generator at both training and test time. Nevertheless, even with dropout, low stochasticity is present in all our models. Ideally, we would like to have enough stochasticity for each realisation to produce a plausible image, nevertheless, for those few samples where some degree of stochasticity is observed, we noticed that one realisation was most plausible than the others. How to effectively add noise under the architectures considered in this work to allow for enough diversity remains as an open research question. Yun Cao (2017) constitutes a work related to that goal, where they follow ideas from Philip Isola (2017) by using cGANs for the colourisation task, but they add noise in the standard way trough random noise vectors and using simpler architectures. While they showed that their approach accomplish highly diverse and realistic colourisations, their approach seem to be suitable for items in indoor scenes, such as bedrooms [8]. Therefore, it would be interesting to gather the ideas presented there with the architectures considered in our approach to allow for higher diversity colourisations when working with datasets richer in categories, such as ImageNet.

3. While in this work we have explored a colourisation approach based on the class prediction of each (full) image, using an object-based approach remains as an open and difficult research question. Following our ideas presented here about an assembled model, an object detection model could be used together with colouring models per each recognised object. Large colorful datasets designed for object recognition such as Microsoft's COCO [9] (Lin (2015)) would be suitable for such task.

4. Adversarial Machine Learning (Alexey Kurakin (2016), Florian Tramer (2017)), is a recent and exciting research area. In this context, adversarial examples are malicious inputs designed to fool machine learning models and adversarial training, the process of explicitly training a model on adversarial examples in order to make it more robust for reducing its test error on clean inputs. ImageNet is also the most frequently used dataset for this kind of problems, so the use of test samples generated by models following our approach could be a potential research area.

5. Conclusions

This report has carried out the plan laid out in the interim report; building and experimenting with an assembled model and testing the results using two different metrics. The first phase of experiments showed among other things the impact of increasing the batch size from 1 to 5, lowering the training time and giving stronger colourisations but with the cost of an increase in miss-colourisations too, we also studied the effects of training the models with different numbers of epochs, obtaining a more stable and reliable colourisation with big numbers of epochs but also at a cost of an increase in training time.

During the second phase of experiments we noticed there were big differences between the colourisation patterns followed by each of the models, VGG-16 providing soft but accurate colouring mainly with green and brown hues, the assembled model a stronger and more varied colourisation and the general model an even stronger colourisation but with many errors.

The visual inspection survey showed that, when displayed different colourisations of a single image, without any more context of the general behaviour of the models, people tend to consider the VGG-16 colourisation as the most natural one, followed by the assembled model and in the end the general model, for which we even received some comments saying that 'it had a curious bubble effect', which we found quite accurate. On the other hand, the semantic classifier gave a better performance to the assembled method than to the VGG-16, we believe this was because of its big sensibility to colourisation, preferring more balanced colourisations that some other biased towards a specific hue of colours. The selection of these two metrics proved to be a good decision, since they seem to complement each other, one focusing on the human perception of the image and the other on the balance and accuracy of the colours used.

Independent of the metric used, the assembled method unmistakably outperformed the general model, even though the two models used the same images, which makes the assembled model not only a superior classifier in terms of training time, but also in performance.

Notes

[1]We did not answer the survey since we thought our answers might be biased after working with the different models for so long.

[2]The training times we are reporting were taken at the beginning of the month, when the cluster was relatively free, this times increased as the deadline came closer.

[3]http://yann.lecun.com/exdb/mnist/

[4]https://www.cs.toronto.edu/~kriz/cifar.html

[5]http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[6]http://ufldl.stanford.edu/housenumbers/

[7]https://tinyurl.com/y9ufozwn

[8]http://lsun.cs.princeton.edu/2017/

[9]http://cocodataset.org/

References

Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*, 2012.

Alexey Kurakin, Ian Goodfellow, Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv: 1611.01236*, 2016.

Chenlin, Yen. pix2pix-tensorflow. https://github.com/yenchenlin/pix2pix-tensorflow, 2016–2017.

Christian Szegedy, Yangqing Jia, Pierre Sermanet. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

Dahl, Ryan. Automatic colorization. http://tinyclouds.org/colorize/, 2016.

Florian Tramer, Alexey Kurakin, Ian Goodfellow. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv: 1705.07204*, 2017.

Ian Goodfellow, Aaron Courville, Yoshua Bengio. Generative adversarial nets. pp. 2672–2680, 2014.

Iizuka, Satoshi, Simo-Serra, Edgar, and Ishikawa, Hiroshi. Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. *ACM Transactions on Graphics (TOG)*, 35(4):110, 2016.

Jonathan Long, Evan Shelhamer, Trevor Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, pp. 3431/3440, 2015.

Karen Simonyan, Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Lin, Tsung-Yi. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2015.

Olga Russakovsky, Jia Deng, Hao Su. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

Philip Isola, Jun-Yan Zhu, Tinghui Zhou. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Richard Zhang, Phillip Isola, Alexei A. Efros. Colorful image colorization. *arXiv preprint arXiv:1603.08511*, 2016.

Takeru Miyato, Masanori Koyama. Conditional generative adversarial networks with projection discriminator. *International Conference on Learning Representations*, 2018.

Toshiki Kataoka, Takeru Miyato, Masanori Koyama. Spectral normalization for generative adversarial networks. *International Conference on Learning Representations*, 2018.

Yun Cao, Zhiming Zhou, Weinan Zhang. Unsupervised diverse colorization via generative adversarial networks. *arXiv preprint arXiv:1702.06674*, 2017.