Which ancient civilisation does this artefact belong to? A deep learning model that identifies the provenance of the pictured artefact.

Mathilde Martin

Master of Science in Computer Science
The University of Bath
2023-2024

Which ancient civilisation does this artefact belong to? A deep learning model that identifies the provenance of the pictured artefact.

Submitted by: Mathilde Martin

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Master of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

This project involved the creation of a novel cultural-heritage specific dataset which covers 6 classes with a combined total of 223,486 images. After rigorous work and pre-processing, the quality of this dataset was demonstrated using various CNNs pre-trained on ImageNet. ResNet50 and EfficientNetB0 to B2 were the models chosen for this task and all achieved between 84% and 86% global accuracy. Additionally, we investigated how much class imbalance can impact the accuracy of a CNN using class weights, undersampling, and oversampling techniques. Our experiments showed that undersampling and class weights made our models worse. In contrast, oversampling using data augmentation helped our model be more balanced by bringing its precision and recall values closer together; they attained 68.9333% and 63.5916% respectively.

As there is very little research in art classification and no cultural-heritage specific dataset readily available for visual classification tasks, the findings and output of this project opened up new avenues of research for cultural-heritage tasks.

Contents

List of Figures

List of Tables

Acknowledgements

Firstly I would like to express my sincere gratitude to my supervisor Dr Georgios Exarchakis for his continued support, guidance and encouragement throughout this project. Thank you for being as interested and invested as me in this project and in putting together this cultural heritage dataset.

Thank you to my other and better half Ben for his continued love, support and faith. Thank you also to our beautiful cats, Charlie and Perceval.

Thank you to my best friend Salomé and my godson Robin. Knowing that you are always present for me, even thousands of kilometres apart, keeps me going.

Thank you to my work colleagues for always believing in me, making my work/study life balance a breeze, and for their friendship. Namely Johnny, Colin, Neil and Adie.

Thank you to Travis for his encouragements and his support throughout the year.

Finally I would also like to thank my family, in particular my mum, my brothers and my aunts Laurence and Sylvie for their continued support and their unwavering faith in me.

Chapter 1

Introduction

1.1 Personal Motivation

In 2017 I graduated from the University of Winchester with a Bachelor of Drama. My dissertation combined not only the analysis of various plays but also delved deep into their historical context and history as a whole. Art and History have always played a big part in my life and I have always aimed to find ways to bring them to the forefront of what I do. Whilst researching and brainstorming ideas for this Master's dissertation, it was important to me to understand the relationship between computer science, art, and history and see what could come of it.

1.2 Background and Context

As noted in my project proposal, the Oxford Dictionary defines heritage as 'denoting or relating to things of special architectural, historical, or natural value that are preserved for the nation' (Stevenson, 2010). Cultural heritage can therefore be defined as the legacy of artefacts from a society's past to future generations. Those artefacts can be either physical (vases, jewellery, paintings, architecture) or immaterial (languages and dialects, dances, songs) (UNESCO, 2021).

Cultural heritage, whether from our own culture or that of others, is extremely important. Boyd and Timothy (2003, cited by Nilson and Thorell, 2018) argued that the preservation of our cultural heritage was vital for scientific, political, economic and social reasons. Indeed it contributes to our understanding of how we once lived and provides context as to our relationship with other nations. It allows us to increase our knowledge and understanding of the world. It also gives us a 'sense of place and cultural identity' (Nilson and Thorell, 2018) as well as attracting tourists who are eager to learn, and invest time and money in the pursuit of knowledge. Therefore 'The preservation of material culture – objects of art and of daily use, architecture, landscape form – and intangible culture – performances of dance, music, theater, and ritual, as well as language and human memory – are generally regarded as a shared common good by which everyone benefits' (Silverman and Ruggles, 2007).

Through time, however, an unfortunate number of items, sites, and customs have either been lost, damaged or destroyed. According to The United Nations Educational, Scientific and Cultural Organization (UNESCO), 'Natural and cultural heritage sites, museums, libraries and archive collections around the world are increasingly becoming the collateral victims of natural disasters and armed conflicts' (UNESCO, 2021). A solution must be found to recognise, protect and preserve those artefacts for future generations. It is, therefore, worth asking ourselves how, in an increasingly digital world, can technology help the conservation of such artefacts. How can we find what is lost? Repair what is broken? Prevent further damage? Accurately and efficiently share and expand our knowledge? Correctly classify and preserve known artefacts as well as newly-discovered ones?

1.3 Related Works

Art Preservation

Some measures are already in place to help capture, protect and serve our cultural heritage. Virtual Reality (VR) for example, is now used more and more in an attempt to preserve it. As Zhou, Geng and Wu (2012) described, a lot of museums have started to create their own virtual environment in which to share their diverse exhibits. This has many advantages including a reduction in theft and damage of artefacts, no time/route restriction for visitors, various demonstration methods that allow users to absorb knowledge better, and better management for the museum which improves retrieval and research (Zhou, Geng and Wu, 2012). Zhou, Geng and Wu (2012) also dived deep into how Artificial Intelligence (AI) can help digitally restore broken or damaged artefacts, calligraphy, and paintings. Whereas traditional methods are time-consuming and costly, with the development of digital image processing technology, 'repair work can be done by a computer, the cost of repair is reduced greatly, artificial repair experience is simulated, and the efficiency of the repair is largely improved compared to the artificial repair' (Zhou, Geng and Wu, 2012). This led to museums starting to 'digitize large parts of their cultural heritage collections, leading to the creation of several digital open datasets' (Sabatelli et al., 2018).

Art Classification

On the subject of Art Classification, some research has been done regarding both fine arts and architecture. Sabatelli et al. (2018) for example have looked into Transfer Learning for Art Classification Problems by comparing four architectures on three art classification tasks. They found that fine-tuning a pre-trained network on a small artistic dataset leads to improved results as opposed to simply using them as off-the-shelf feature extractors. This is due to the Deep Convolutional Neural Networks (DCNNs) developing new selective attention mechanisms. Lecoutre, Negrevergne and Yger (2017) and Imran et al. (2023) in their respective research have both looked at comparing DCNNs when classifying the artistic style of a given painting. The former focused particularly on comparing AlexNet with deep residual networks such as ResNet50. The latter brought a new technique to the table which saw them combining Deep and Shallow Neural Networks, and evaluating them using 6 different pre-trained models. On the subject of architecture Obeso et al. (2017) worked on classifying the architectural style of Mexican buildings using CNNs and saliency-driven content selection, and compared AlexNet to GoogLeNet.

Dataset Creation

Deng et al. (2009) proposed ImageNet in 2009, with the hopes of providing the most comprehensive coverage in the image world within the next two years so that it could be used further in vision-related tasks. To put together such a large dataset, they used both an automated way to collect candidate data and a manual and lengthy process to verify it. Closer to our research, Kambau, Hasibuan and Pratama (2018) manually put together a dataset covering 5 Indonesian ethnicities. They did so by collecting data in various formats: videos, texts, and audio files.

1.4 Gap in Literature

From those examples, however, it becomes quite apparent that there is a gap in the literature. Indeed, whilst there exist architectural datasets - such as the Modern Architecture dataset available on Kaggle (Paulat, 2023) - as well as many datasets relating to paintings - such as the WikiArt dataset (WikiArt, n.d.) - there still exists a gap when it comes to certain cultural and artistic artefacts. By this we mean images of sculptures, writings, paintings, pottery, jewellery, furniture, weapons, and clothes, among other examples. There is no ready-made, fresh off-the-shelf dataset ready to use for classification tasks.

What's more, to this author's knowledge, no attempt has been made to put together a dataset of artefacts covering distinct civilisations from across the globe, to use for diverse vision classification tasks. On a small scale, an effort has been made by Kambau, Hasibuan and Pratama (2018) who adopted Deep Learning techniques to classify an image, audio, video, or text into one of five Indonesian Ethnic Groups. Using both CNNs and RNN they obtained the following accuracies for image, audio, video and text classification respectively: '77%, 83%, 55%, and 93%' (Kambau, Hasibuan and Pratama, 2018). This research was however focused on a small dataset of 100 data points for each tribe (for a total of 500 data entries) and concerned a single civilisation's tribes. There is therefore room to develop this idea and expand it.

1.5 Novelty

This project's novelty comes from the following points:

1. This project produces a novel Dataset for cultural heritage classification composed of 6 classes and totalling 223,486 images.
2. This project builds on Kambau, Hasibuan and Pratama (2018)'s work by covering a larger geographical region - with civilisations from diverse continents - using a dataset containing 400 times more data.

1.6 Project Aims

The aim of this project has always been, first and foremost, to put together a large dataset of high-quality cultural artefact images. Once the dataset was put together, it was trained and evaluated over several Deep Convolutional Neural Networks. This had two purposes :

- To test models pre-trained on a general dataset such as Imagenet to classify a dataset of cultural heritage images. From there we were able to analyse how efficient the given model was to generalise to cultural artefact classification tasks.
- To demonstrate the quality of the dataset.

The second goal of this research is for this data to be used in further vision tasks, to help preserve our rich and diverse cultural heritage.

1.7 Structure

This thesis begins with a review of the literature surrounding Deep Learning and Image Classification (chapter 2). We first take a broad look at the history and advancement of Artificial Intelligence, Deep Learning and Neural Networks. We then look at and compare various classification models which have been developed and which are still being used to this day. After this, we analyse the most recent advances and identify gaps in the literature which this paper attempts to answer. Finally, we provide a framework for comparing classification models on our new datasets and highlight criteria which will be used for comparison.

This Literature Review is followed by a detailed analysis of the dataset in chapter 3. This section goes into detail into how the data was collected, cleaned, and prepared. It also explores the ethical considerations when collecting data, especially when it comes to copyrights. We also dive into the limitations of our dataset as well as the improvements that could be undertaken in future work.

In chapter 4 we look at the design and implementation of diverse classification models on our cultural heritage classification task. In this section, we detail the choices we made for both our models and their hyperparameters. Additionally, we delve into issues encountered along the way for both CNNs and Transformers.

Finally, in our Experiment and Results chapter (chapter 5) we take a look at our models' performances and compare them against each other. After this, we analyse the results seen when applying three data balancing techniques to our best model. We continue this chapter by acknowledging the limitations and assumptions we made in our experiments. We then compare our findings to what we have seen in the literature and discuss similarities and differences. We end this section by summarising our key findings.

This thesis ends with a discussion and summary of our key findings (chapter 6). We discuss our contributions to the field, before outlining the future work and research that could be undertaken on our dataset and for visual cultural heritage tasks.

Chapter 2

Literature and Technology Survey

2.1 Introduction

This Literature and Technology Survey aims to:

1. Give us a general understanding of Artificial Intelligence whilst focusing on Deep Learning.
2. Investigate various classification algorithms with a focus on Convolutional Neural Networks.
3. Compare different image classification models.
4. Analyse and identify the benefits of transfer learning and fine-tuning.
5. Understand the emerging trends and the future of classification models.

2.2 Overview of Deep Learning

2.2.1 Convolutional Neural Networks

Deep Learning is a subset of Machine Learning which finds its origins in works that tried to mimic networks of neurons in the human brain. Some of the earliest evidence of this is in the work of McCulloch and Pitts (1943) where they describe a network of neurons connected by weighted paths. Deep Learning models can extract features without the need for a human expert to do this manually. The most well-known models are Recurrent Neural Networks (RNNs) - for natural language tasks - and Convolutional Neural Networks (CNNs) - for vision tasks. Those were 'inspired originally by models of the visual cortex proposed in neuroscience' (Russell and Norvig, 2022).

As Krizhevsky, Sutskever and Hinton (2017) explain in the prologue of their updated version of their 2012 paper however, most computer vision researchers - before the availability of big data and processing power of computers - 'believed that a vision system needed to be carefully hand-designed using a detailed understanding of the nature of the task' (Krizhevsky, Sutskever and Hinton, 2017). CNNs were therefore vastly ignored and led to papers such as one written by Yann LeCun and collaborators being rejected by the leading computer vision conference (Krizhevsky, Sutskever and Hinton, 2017).

However, this changed with the emergence of large labeled datasets such as Imagenet (Deng et al., 2009) which 'populates 21,841 synsets of WordNet with an average of 650 manually verified and full resolution images. As a result, ImageNet contains 14,197,122 annotated images organized by the semantic hierarchy of WordNet (as of August 2014)' (Russakovsky et al., 2015). From this dataset comes the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) which sees models competing against each other, trying to outperform each other as well as previous state-of-the-art models. 2012 marked the year when interest in deep learning rose due to Krizhevsky, Sutskever and Hinton (2012) using a deep convolutional neural network, AlexNet, to win the ImageNet competition. They attributed their state-of-the-art performance to the DCNNs' learning capacity as well as the 'strong and mostly correct assumptions about the nature of images' (Krizhevsky, Sutskever and Hinton, 2012) they make.

2.2.2 Residual Neural Networks

With interest in CNNs renewed following AlexNet's win in the 2012 ImageNet competition (Krizhevsky, Sutskever and Hinton, 2012), further research ensued. One notable contribution to further improving the training of deep neural networks is that of He et al. (2016) with their proposal of a residual learning framework: ResNet. Their work addresses the degradation problem: as the depth of the network increases, the accuracy starts to get saturated and then degrades rapidly. The innovation from their residual learning framework comes from their residual module which consists of two convolutional layers, where the output of the second layer is added to the input of the first (this is known as a shortcut connection). 'Instead of hoping each few stacked layers directly fit a desired underlying mapping, we explicitly let these layers fit a residual mapping' (He et al., 2016). In their experiments, they first compared two plain networks, one 18 layers deep and the other 34 layers deep. This exposes the degradation problem as the deeper network has higher validation and training errors than the shallower one. Following this they evaluate an 18-layer and a 34-layer residual network (ResNets) and observe that contrary to their experiments on plain networks, quite the opposite happens: the deeper 34-layer residual network has better validation error by 2.8% and lower training error (He et al., 2016).

This is a major advancement in computer science and to this day residual networks are still some of the most used models. In their 2021 paper, Wightman, Touvron and Jégou (2021) aimed to re-evaluate how ResNet-50 performs when trained using new advances in the field such as novel optimization and data augmentation. To do so they propose three training procedures at inference resolution 224x224. Those three training procedures (A1, A2, A3) correspond to different numbers of epochs (100, 300, 600) with adjustments of hyperparameters and ingredients. Their results found that A1 outperformed the state-of-the-art of the time on ImageNet with a vanilla ResNet-50 architecture. This shows how relevant ResNets still are and this is a reason why ResNet-50 was selected in our image classification task.

2.3 Classification Tasks in Deep Learning

2.3.1 An Overview of Classification

Classification has always been a part of human nature. It is something we constantly do, we take something, name it, and group it with similar things to better make sense of them and the world as a whole. As an example Hampel (2002) cites the start of astrology as man's need to name and separate constellations and planets. He also qualifies the beginning of modern geology as man's need to separate and name geological layers.

Living in a time where data is abundant, it is important to understand it and learn both how to work with it and how to use it to our advantage. The availability of big data and the improvements in the field of Artificial Intelligence (AI), allow deep learning models to analyze data and give results. Those can - in some instances - even surpass human capabilities. This was shown by He et al. (2015) in their 2015 paper where their model achieved 4.94% top-5 test error on the ImageNet 2012 dataset with a 0.15% improvement over human-level performance. It is then quite easy to guess and understand why data classification is so crucial for data analysis. Deep Learning models could help us see patterns that might not seem obvious to us. In the medical field for example, image classification and deep learning have been used in the development of Computer-Aided Diagnostic applications to classify disease and normal patterns, malignant and benign lesions, and prediction of high and low-risk patterns of developing cancer in the future (Chan et al., 2020). On the subject of cultural heritage, and since there is an urge for museums to digitize their collections, image classification can help in automating this process and accelerating it.

Whilst AI and Machine Learning have numerous algorithms capable of classification, Deep Learning is the best, most efficient, and most accurate choice for image classification with so many data points. Indeed, as we can see from figure 2.1, compared to a shallow model with short computation paths (such as linear regression), or to a decision list network that has both long and short paths, a deep learning network 'has longer computation paths, allowing each variable to interact with all the others' (Russell and Norvig, 2022). This is why for this research we focus solely on deep neural networks.

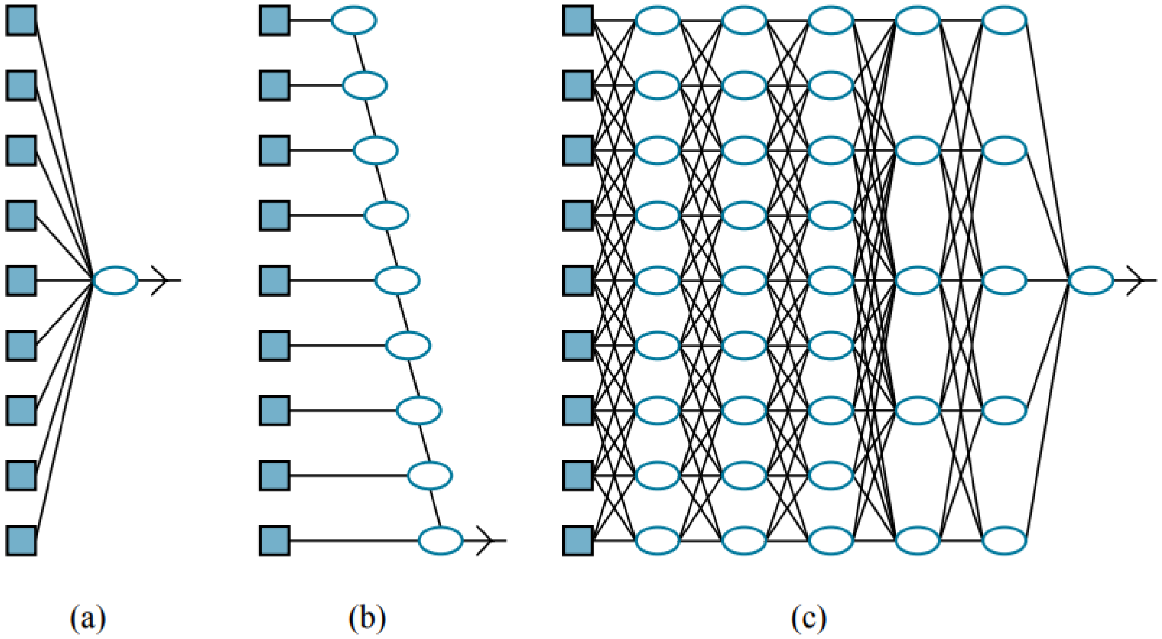


Figure 2.1: Linear Regression (a), Decision List (b), and Deep Learning Network (c) taken from Russell and Norvig (2022)

2.3.2 Model Comparison Studies

As we have already seen, there are many deep learning models in existence since AlexNet (Krizhevsky, Sutskever and Hinton, 2012). One of the most notable contributions to the field is that of Residual Networks introduced by He et al. (2016) to avoid the degradation problem. However, those are not the only notable CNNs worth mentioning. Tan and Le (2020) investigated how scaling can improve performance and lead to better accuracy levels. As we have seen, residual networks scaled networks deeper and deeper whilst fending off the degradation problem (He et al., 2016). The following year, Zagoruyko and Komodakis (2017) offered a new method of scaling up CNNs using a ResNet model, however, this time they chose to go wider instead of deeper. They demonstrated that a wider network only 16 layers deep can outperform all previous residual networks in both accuracy and efficiency on CIFAR, SVHN, and COCO.

Tan and Le (2020) further explored the idea of scaling networks and showed that whilst most methods only scale one dimension, there is a way to scale each dimension with a constant ratio which leads networks to better results. Their research led them to propose ‘a new compound scaling method which uses a compound coefficient to uniformly scale network width, depth, and resolution’ (Tan and Le, 2020) as well as a new family of CNNs called EfficientNet. EfficientNetB0 offers a top-1 and top-5 accuracy improvement of 1.1% and 0.3% respectively over ResNet-50 with 20.7M parameters less. EfficientNetB7, the largest of the family, achieved state-of-the-art top-1 accuracy with 84.3% whilst being faster (by 6.1x) and smaller (with 491M parameters less) than the best ConvNet at the time (GPipe).

2.3.3 Transfer Learning

One of the great improvements that has accompanied the development of deep learning models has been the use of Transfer Learning (TF) to improve the accuracy of models. This improvement comes from the need to reduce both the training process time and cost, as well as overcome the lack of availability of training data (Iman, Arabnia and Rasheed, 2023). It consists of 'training a machine learning algorithm on a new task [...] while exploiting the knowledge that the algorithm has already learned on a previously related task' (Sabatelli et al., 2018).

Fine-tuning in particular has been shown to increase model accuracy by a significant amount as opposed to simply retraining the last layer of a model pre-trained on a dataset. In Lecoutre, Negrevergne and Yger (2017)'s research, they investigated whether deeper retraining is necessary to obtain the best possible performance from models pre-trained on ImageNet. To do so they used ResNet50 pre-trained on ImageNet. They then experimented by retraining an increasing number of layers (from only the last layer, all the way to the entire network) and comparing results. Their experiment showed that peak performance is achieved when retraining about 20 layers out of 106. Training too deep can make the model overfit, leading to accuracy drops. This 'confirms the assumption that deeper layer neurons are not specialized to only one specific task' (Lecoutre, Negrevergne and Yger, 2017) and corroborates what Iman, Arabnia and Rasheed (2023) discussed in their article regarding the catastrophic *forgetting dilemna*. Unfreezing too many layers can lead to drastic changes of weights in the entire model leading to it forgetting previous knowledge. As the earlier layers extract low-level features, that knowledge is crucial to our model and needs to be kept untouched whilst the top layers contain high-level features and can be retrained with the desired dataset in mind (Iman, Arabnia and Rasheed, 2023).

As Lecoutre, Negrevergne and Yger (2017) delved deeper into their results and analysed them, they noted that the classes which had the worst performances tended to be visually and historically very close. Those classes also tended to have little data available. This is something our research came to highlight as our smallest class (between 7-8 thousand images as opposed to around 60 thousand for the largest class) had the highest number of misclassified images. Those were mostly misclassified as being either egyptian or grecoroman which concurs with Lecoutre, Negrevergne and Yger (2017)'s findings as byzantium (our smallest class) was a Greek colony that covered Egypt among other countries.

Sabatelli et al. (2018) went further in their investigation of Transfer Learning and Fine-Tuning. To demonstrate how invaluable Fine-Tuning is, they investigated how useful it is as opposed to simply using Transfer Learning. To do so, they used four neural architectures (VGG19, Inception-V3, Xception, and ResNet50) having all achieved great results on the ImageNet challenge. They then evaluated them on three different classification tasks: classifying by material used (1), classifying by artistic category (2), and classifying by authors (3). The results speak for themselves and show us that it 'is always beneficial to fine-tune the DCNNs over just using them as off-the-shelf feature extractors' (Sabatelli et al., 2018). For the first and second challenges, ResNet50 has seen the biggest accuracy increase by 6.14% for the former and 20.07% for the latter, between using simply Transfer Learning and using Fine-Tuning. For the third challenge, InceptionV3 has seen the biggest increase: + 51.66%. Additional research into those results provided insights into what changed to make those DCNNs perform so much better: they 'develop novel selective attention mechanisms over the images, which are very different from the ones that characterize the networks that are pre-trained on ImageNet' (Sabatelli et al., 2018). This means that the model learns and adapts to the heritage dataset to develop new features better suited to artistic classification than just natural image classification.

2.3.4 Handling Imbalanced datasets

Class imbalance is a common and well-known issue in deep learning and highly imbalanced datasets can impact a model's accuracy and predictions. Most CNNs assume that classes are balanced, thus an imbalanced dataset might lead the model to become biased towards the dominant class and in some instances even ignore the smallest of classes (Johnson and Khoshgoftaar, 2019). Learning to work with class imbalance is therefore fundamental.

Krawczyk (2016) emphasized three approaches to tackling imbalance: Data-level methods, Algorithm-level methods, and Hybrid methods. The first affects the dataset directly by utilising over or undersampling in order to balance out the distribution of samples in each class by adding data to the minority class or removing data from dominating classes. The second concerns the models directly and generally means making use of class weights to diminish and eliminate the bias towards dominant classes. The latter sees a combination of both previous methods.

As important as this research is, however, 'very little statistical work has been done which properly evaluates techniques for handling class imbalance using deep learning and their corresponding architectures' (Johnson and Khoshgoftaar, 2019). Our work will thus aim to investigate this further by directly comparing class weights, oversampling, and undersampling methods on our imbalanced dataset.

2.4 Recent Advances and Trends

2.4.1 Transformers

Vaswani et al. (2017) first proposed Transformers for machine translation tasks. Recurrent Neural Networks (RNNs) were state-of-the-art approaches before Transformers were proposed. The novelty lays in the fact that the Transformer model would rely entirely on 'attention mechanism to draw global dependencies between input and output' (Vaswani et al., 2017). Their research and results saw their models outperform previous state-of-the-arts, and do so at a fraction of the training cost.

Pure Transformers

Taking inspiration from Vaswani et al. (2017), Cordonnier, Loukas and Jaggi (2020) and Dosovitskiy et al. (2020) both explored this further in their own research. The scale of their research is however drastically different. Cordonnier, Loukas and Jaggi (2020) first focused on mathematically proving that self-attention can perform convolution. They then went on to prove this by creating a model that extracts 2x2 patches from input images and applying self-attention to it. They did not aim to attain state-of-the-art results and decided to compare it to a standard ResNet-18 model to prove the significance and validity of their results.

Dosovitskiy et al. (2020) on the other hand went further and their models could handle medium-resolution images as well. Their research was crucial in showing that the application of a pure Transformer directly to sequences of image patches could perform even better than CNNs on classification tasks. In their first experiments, they evaluated the representation learning capabilities of a classic CNN (ResNet), a Vision Transformer (with 3 models ViT-Base, ViT-Large, ViT-Huge), and a hybrid. They pre-trained those models on datasets of varying size (one with 1.3M images, another with 14M images, and the last with 303M images). For their models, their Transformers were inspired by Vaswani et al. (2017). The idea of a Transformer for image classification is that the input image is split into fixed-size patches. Each patch is then linearly embedded and position embedding is added to them. The resulting sequence of vectors is fed to a Transformer Encoder block. The output is then fed to a Multi-Layer Perceptron for classification (Dosovitskiy et al., 2020) (figure 2.2).

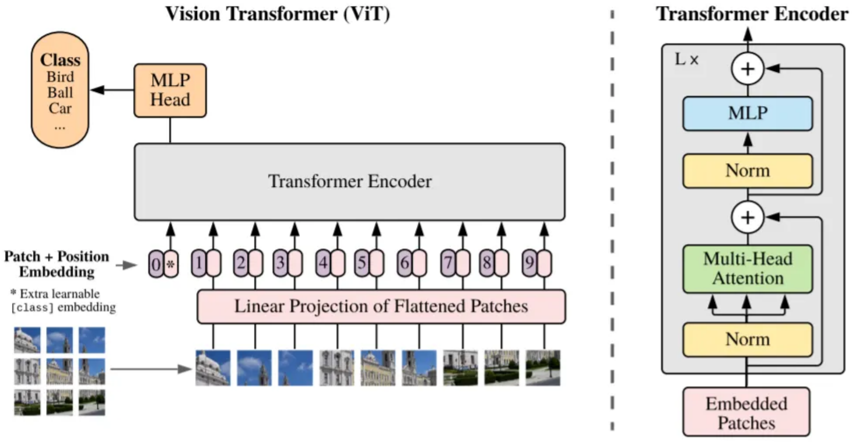


Figure 2.2: Transformer Model Overview (Dosovitskiy et al., 2020)

When comparing ViT-Large and Huge to state-of-the-art CNNs from the literature (Big Transfer and Noisy Student) they noted that the Large model outperformed them both when trained on the bigger dataset (303M images) whilst the Huge model improved performance further. Bhojanapalli et al. (2021) also highlighted the correlation between the size of the training dataset and the robustness of the model. Indeed they showed that on a sufficient amount of data, ViTs are more robust to a variety of perturbations to inputs and model parameters than standard DCNNs (in this case ResNet50 was used as a baseline model for their study). They also noted that ViTs exhibit better scaling when compared to ResNet; this suggests that given a large enough pre-training dataset the gap in robustness between the two will keep expanding significantly.

Dosovitskiy et al. (2020) went further in investigating how important the size of the dataset is by conducting two experiments. They first pre-trained ViT models on datasets of increasing size. From this, they noted that large ViT models are outperformed by BiT ResNets when pre-trained on small datasets, whereas they shine and outperform them when pre-trained on larger datasets. They then trained the models on random subsets of 9M, 30M, 90M, and 300M datasets. From their second experiment, they noted that ViT performed better with larger pre-training, especially on 90M+ subsets.

Hybrid Transformers

Whilst Transformers have shown their potential and strength when combined with sufficient amounts of data, the high computational costs associated with them - due to their millions of parameters - make scaling up Transformers difficult and expensive (Khan et al., 2022). As mentioned previously they also require large amounts of training data, with at least above 90M images (Dosovitskiy et al., 2020) to be able to give strong and accurate results. These are the main reasons why ViTs have not yet been adopted as the de facto models for image classification.

To counter this, hybrid models - which combine aspects of both CNNs and ViTs - have been developed alongside pure Transformers. Dosovitskiy et al. (2020) have touched upon hybrid models. They did see a small improvement from hybrid models over pure Transformers on smaller model sizes, nevertheless, this gap vanishes on larger models. Their work however was mostly focused on pure Transformers and not on creating and configuring strong hybrid models.

In recent years, Tu et al. (2022) have researched this topic and proposed a new hybrid convolutional transformer that aims to counter the lack of scalability of self-attention mechanisms (with respect to image size) in current Transformers. This is done thanks to two features introduced in their model: local and dilated global attention. 'These design choices allow global-local spatial interactions on arbitrary input resolutions with only linear complexity' (Tu et al., 2022). They then applied their new model to diverse vision tasks. Their model became state-of-the-art on ImageNet benchmarks by achieving an ImageNet-1K top-1 accuracy of 86.5% without extra data and an 88.7% top-1 accuracy with ImageNet-21K pre-training. It also performed well on object detection where the MaxViT backbone models outperformed all other backbones by a large margin (and where MaxViT-S outperformed other base-level models with about 40% less computation cost), as well as for visual aesthetic assessment where compared to the state-of-the-art model it shows better linear correlation.

This shows that the strength of hybrid models is not to be undermined as they can bring the best of both Convolution models and Transformers in order and overcome pure Transformers' flaws.

2.4.2 Transformers or Convolutional Networks?

Whilst in recent years the focus has been on Transformers and hybrids, especially due to their results on web-scale datasets, recent trends have seen researchers highlight the competitiveness of CNNs against Transformers.

Both Liu et al. (2022) and Smith et al. (2023) showed that CNNs can rival Transformers at scale. Liu et al. (2022) proposed a new family of ConvNets (ConvNeXt) - by examining design choices not only on their own but also collectively - which can rival Transformers in terms of accuracy and scalability. They then gradually modernized a standard ResNet50 model taking inspiration from Transformers, without introducing any attention-based modules. Their new models reached an impressive 87.8% accuracy on ImageNet-1K, in many cases outperforming Swin Transformers in the Object Detection and Segmentation on COCO and the ADE20K semantic segmentation task with UperNet.

Smith et al. (2023) on their part challenged the belief that Vision Transformers outperform ConvNet architectures by investing the same amount of compute budgets (110k TPU-v4 core hours) to compare both fairly. To do so they used the NFNet model family for the CNN and pre-trained on the JFT-4B dataset whilst fine-tuning on ImageNet. Their research found that when trained with the same budget CNNs can achieve results matching those of ViTs.

This shows that whilst the rise of Transformers is impressive, it is fundamental not to forget the role, strength, and importance of convolution in computer vision.

2.5 Research Questions

Transfer Learning along with CNNs have played an important part in image classification. As we have seen through our literature review, transfer learning is crucial in creating strong and accurate models as opposed to simply building them from scratch. So naturally it makes sense to use pre-trained models to determine the quality of our dataset. Thus the first research question of this paper is:

RQ1 : How do pre-trained CNNs perform on a newly created cultural-heritage specific dataset?

ResNet50 and EfficientNetB0-2 pre-trained on ImageNet are the models that will be used to evaluate our new cultural heritage dataset. The quality of the dataset should be reflected in our model's capacity to classify the images correctly. Therefore precision, recall, f1-score, and accuracy will be analysed to understand how well the model is classifying the images.

However, in our dataset, not all classes are created equal. Our smallest class in particular counts 7,317 images. Our dominant class on the other hand counts 61,185 images. This drastic imbalance might lead the model to create a bias for the dominant classes. As we have seen, class imbalance is an important area of research in Deep Learning and some methods can help in countering this imbalance. Therefore our second research question is naturally:

RQ2 : How much does class imbalance impact the accuracy of a CNN on this task?

To answer this question we will use our best model and compare how the same above criteria (precision, recall, f1-score, and accuracy) are affected after using one of three different methods:

1. Undersampling: The dataset will be reduced so that all classes have the same number of images.
2. Class weights: Weights will be assigned to each class based on its frequency in the dataset to avoid our models becoming biased toward the dominant class(es).
3. Oversampling: Data Augmentation will be used to create more data of the smallest class for the model to use.

Understanding how much imbalance can affect a model's accuracy will allow us to understand better our dataset. This can highlight assumptions or limitations present in the dataset as well as provide insights into imbalanced datasets for wider image classification tasks.

Finally, our last area of research will aim to follow recent trends and advances by looking at Vision Transformers. Whilst we recognise that Transformers could have a whole research paper dedicated to themselves, we feel like it is important to include them and try to understand their strengths and flaws and see how they could perform on our dataset. Whilst this dataset is much smaller than what is required by data-hungry ViTs, we choose to put this fact aside to focus on understanding its architecture and its potential. Our last research questions therefore reflect this :

RQ3 : How would Transformers compare on a similar art-specific classification task?

We will aim to create a ViT from scratch and the same metrics will be used for comparison. Creating a ViT will allow us to better understand their strengths and flaws and provide a good starting point for future work into cultural-heritage classification and Vision Transformers.

Chapter 3

Dataset

3.1 Introduction

As we have seen, very early on in our research it became clear that there were no datasets readily available with the required criteria:

- Hundreds of thousands of images of cultural artefacts.
- A minimum of 5 classes.
- Images of artefacts from distinct civilisations covering most continents and/or different eras.

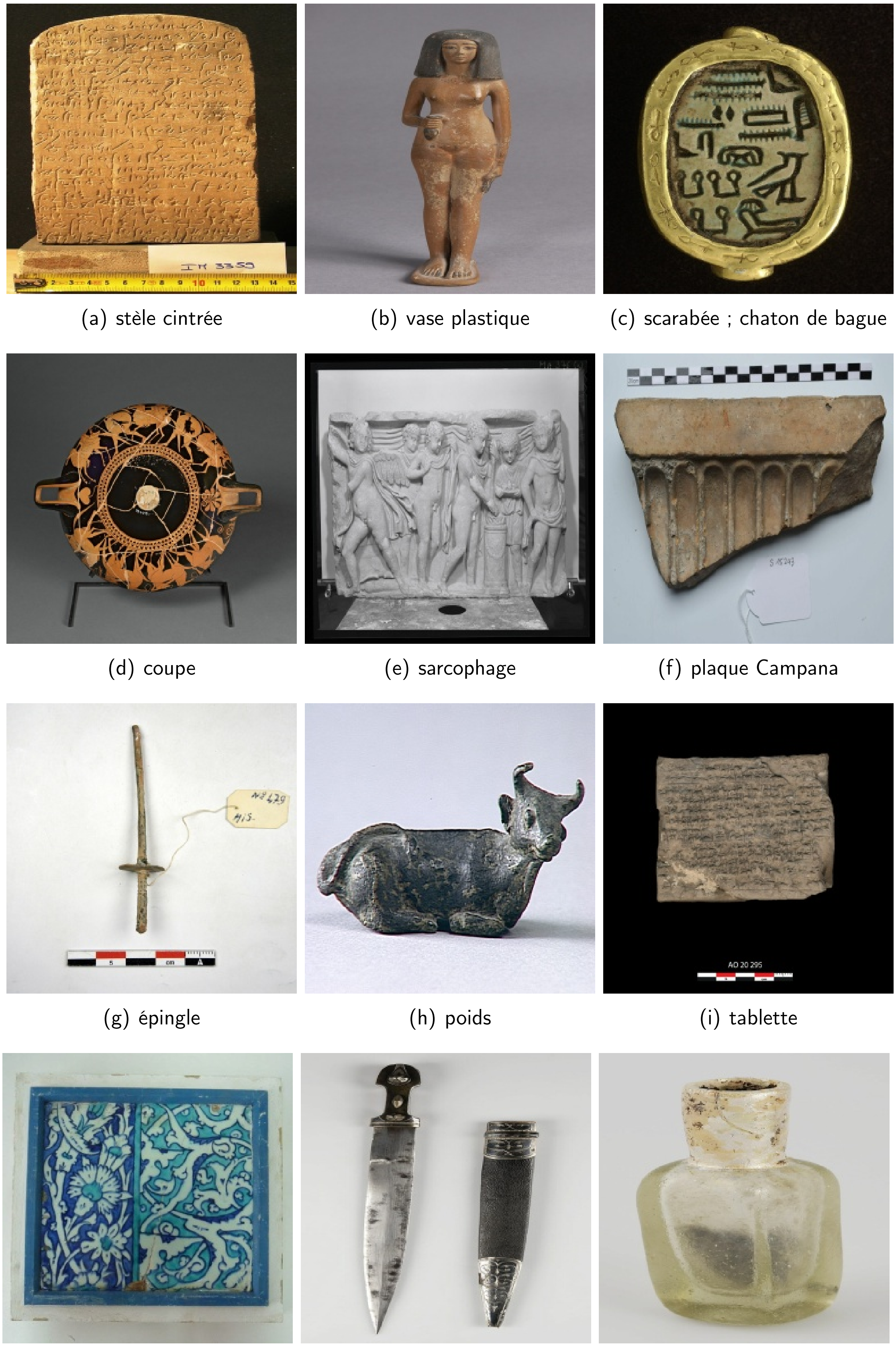
Whilst we mentioned that museums have started to digitise their collections, very few have made it easily accessible for researchers to use. Creating a clean and labelled dataset therefore became the focal point of this study. The creation of this dataset would open the way to cultural heritage image classification on a large scale, enabling the future development of automation for museums.

To put together such a dataset we will be using data available from both Le Louvre and the Metropolitan Museum of Art (MET) as they have both considerable amounts of images on offer. The data there is already divided per class, and images seem professionally taken on a clear background to avoid any kind of noise.

3.2 The Data

The dataset (available to anyone with the following link : here) is made up of 6 classes for a total of 223,486 images. Classes 1-5 were taken from Le Louvre whilst class 6 was taken from the MET. Most images are in colour, however, some were also taken in black and white. To our knowledge, there are no other datasets for cultural heritage classification. The below list indicates how many images were collected for any given class. Figure 3.1 offers a snapshot of the images present in the dataset.

1. Egyptian (44,075 images) - examples (a), (b), (c)
2. Grecoroman (51,696 images) - examples (d), (e), (f)
3. Orient (67,983 images) - examples (g), (h), (i)



(a) stèle cintrée

(b) vase plastique

(c) scarabée ; chaton de bague

(d) coupe

(e) sarcophage

(f) plaque Campana

(g) épingle

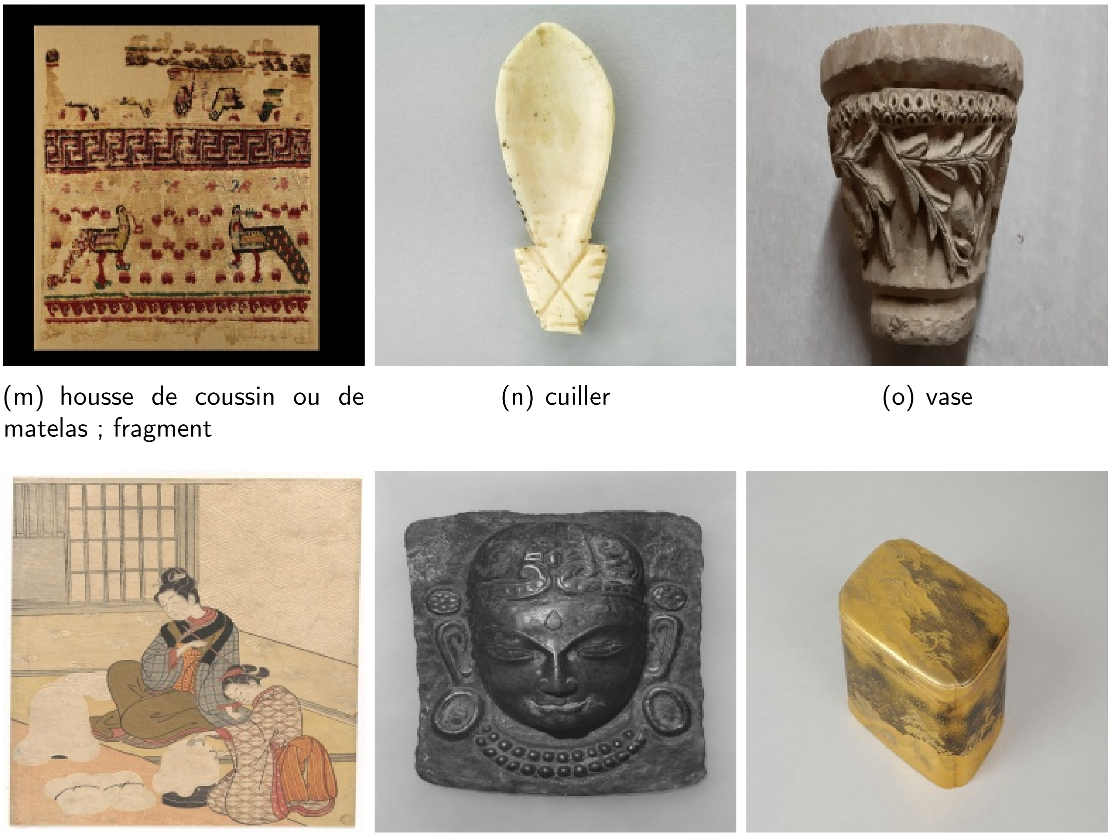
(h) poids

(i) tablette

(j) Carreau à décor de plamettes et d'oeillets

(k) Poignard à longue lame

(l) Flacon quadrangulaire



(m) housse de coussin ou de matelas ; fragment

(n) cuiller

(o) vase

(p) "Evening Snow on the Nurioke, from the series Eight Parlor Views"

(q) "Relief Plaque of Hindu Deity, Probably Processional: Face of a Deity"

(r) Box with Design of Pines Along the Shore

Figure 3.1: Dataset Overview. - (a) to (o) from Le Louvre (n.d.) and accessible from https://collections.louvre.fr/en/recherche - (p) to (r) from The MET (n.d.) and accessible from www.metmuseum.org

4. Islam (18,417 images) - examples (j), (k), (l)
5. Byzantium (8,130 images) - examples (m), (n), (o)
6. Asian (33,185 images) - examples (p), (q), (r)

The data contained in each class represents a multitude of ancient artefacts belonging to diverse categories. The Louvre did a wonderful work in identifying and categorising each particular artefact. Each of them belongs to one of the following categories: Weapons, Jewellry and Adornments, Book Art, Drawings and Engravings, Writings and Inscriptions, Furniture, Coins and Medals, Monument and Funerary elements, Monument and Objects of Worship, Tools and Instruments, Personal Objects, Household Objects, Paintings, Seals and Glyptics, Sculptures, Stelae, Textile and Clothing, and Vases and Tableware (Louvre, n.d.). The MET goes even further in its categorisation by creating deeper and more specific categories. From a higher-level perspective, however, the artefacts contained in images taken from the MET fall under similar categories to the ones from Le Louvre.

3.3 Data Collection

To simplify and automate the process of collecting all the images, Scrapy was used. A spider was created for each class it was collecting data for. The spider's job was to go onto the given website, and for each artefact present on the page, to follow its URL back to its description page. From there, it could take the image's information and URL, and save it to a CSV file, before going back and moving on to the next artefact. This was repeated for each page in that category.

Once all the URLs had been collected, a simple script was created and run which would download each of the images to the correct folder. Each picture was carefully labelled in the following way 'CLASSNUMBER_NAMEOFMUSEUM_UNIQUEID'. We chose to follow this naming convention to clearly establish some important facts about the data.

1. **The class number**: it is needed for the correct classification of the image.
2. **The name of the museum** : it allows us to know the provenance of the artefact and its associated image.
3. **A unique ID number**: taken directly from the image's URL, it allows us to give the artefact a unique ID which can be directly linked back to its source.

3.4 Considerations When Collecting Data

Before any kind of data collection takes place it is important - for both legal and ethical reasons - to understand the data we are trying to obtain, to whom it belongs, whether it is protected by copyrights, and how it can be used. In our case, it meant understanding Le Louvre's and the MET's Terms of Use. In both cases, those can be easily found on their website. For both museums, the use of their data is allowed for academic and research purposes.

The MET (n.d.) has a lot of its data identified as Open Access. This means that those are available under a Creative Commons Zero (CCZ) license and can be used freely without requiring permission from the museum for any commercial or non-commercial purpose. Their collections do however also contain some works believed to be under copyright. For those, 'restrictions are available for limited noncommercial, educational, and personal use only' (MET, n.d.) and users must 'cite the author and source of the Materials as they would material from any work, and the citations should include the URL "www.metmuseum.org" ' (MET, n.d.).

For Le Louvre (n.d.), the use of the images is also subject to certain terms and conditions. As such any 'use of one or more Photograph(s) must include photographic credits such as those provided on the collections website in addition to a permalink to the description of the work' (Louvre, n.d.).

In both cases, the terms of use were respected.

3.5 Data Pre-Processing

3.5.1 Manual check

Once all the data had been collected, we first proceeded to do a manual check of the images. This was important in checking for any errors (duplicates or blank images) or any noise in the pictures (images being mostly background or having the artefact in a corner of the picture for example). This check determined that all the images were of high quality with minimal noise and no errors.

3.5.2 Resizing

Although we could now be fairly reassured of the quality of the data, every image collected had a different size. Therefore the second step taken was to resize the dataset to a chosen standard size. We kept our raw data intact in a separate folder and created a new hierarchy of folders with all the resized data which would later be used in our experiments. As we decided to work with ResNet50 and EfficientNet, the decision was made to resize all images to 256x256 using a simple script.

After resizing all images, another manual check was done to check that this had neither altered nor distorted the original images too drastically.

3.5.3 Creating a Test Set

Finally, we created a script that, ahead of testing our dataset on various models, would create a test set out of 10% of each of the class' images. This test folder was kept separate from the rest of the images and would be used to evaluate the models' accuracy in predicting the provenance of a given artefact.

3.6 Data Validation and Quality Control

To validate our data we trained a few models on our dataset. We then compared the accuracy, precision, recall, and f-1 score of each model on our test set. This showed us that one of our classes (the byzantium class, the one with the most limited amount of data) was getting misclassified. With a simple script, we were then able to copy each image being misclassified to a separate folder for visual verification. Additionally, our script created a table showing us what class the byzantium images were being misclassified as (see table 3.1).

Table 3.1 reveals that out of a total 369 misclassified byzantium images, 43.3% were being misclassified as belonging to the egyptian class and 33.3% to the grecoroman class. This links back to what we read in the literature where Lecoutre, Negrevergne and Yger (2017) identified that classes that performed worse were the ones with the least amount of data and which were visually and historically close to other classes. This is what we can see happening here as the Byzantium Empire was the continuation of the Roman Empire and covered Egypt, Greece, and Italy among other places.

Class	Misclassified Byzantium Images
egyptian	160
grecoroman	123
orient	35
islam	16
asian	35
Total	369

Table 3.1: Number of Misclassified Byzantium Images per Class.

Figure 3.2: Blank Image (Louvre, n.d.)

On top of this, a visual check on our misclassified images revealed that our preliminary manual check missed some data errors. Indeed, some blank images as seen in figure 3.2 were also scraped from the Louvre website. To rectify this, and after checking that all blank images were the same, we built a small script that went into each folder and removed them. In total, roughly 1,000 blank images were found and discarded.

3.7 Data Security and Storage

To keep the data safe, this was stored on an external hard drive. None of the data was kept on this author's laptop or any cloud services. This never leaves this author's house and when not in use is kept away from prying eyes.

The goal of this research, however, is to get permission from both museums to release this dataset under a non-commercial, research-only license, for other researchers to use. If this is indeed possible, Kaggle would be a strong choice of platform to release this onto as this is where the target audience for this dataset would be.

3.8 Limitations

This dataset offers formidable opportunities for work on cultural heritage vision tasks. It is however important to acknowledge the areas where certain limitations can be seen. This will in turn allow us to understand the possible results or biases which we could observe. On top of this, this could highlight areas for which further work could be undertaken. Finally, it could also guide our research when using and validating this dataset.

3.8.1 Imbalance

As we have already mentioned, whilst this dataset contains a large number of images, certain of its classes are highly imbalanced. When training a model on this data, this can in turn create a bias. To better understand the impact that this imbalance can have, we will be comparing techniques aiming to combat class imbalance. The results of those experiments will indicate what potential improvements they can make.

3.8.2 Related Classes

Imbalance is not the only issue affecting our dataset. Certain classes are closely related to each other, particularly the byzantium class with the egyptian and grecoroman classes. Indeed, the former shares geographical ties with the latter two as well as a certain closeness of eras with the grecoroman class. This means that artefacts share similar traits to each other therefore making it harder for an individual as well as a Neural Network to differentiate.

3.8.3 Representation

Finally, it is important to note that this dataset is but a small representation of the rich civilisations of the world. Due to time constraints for this project, the sixth class contains a lot of artefacts of chinese, japanese, and korean heritage under the same class. Ideally and with more time, the aim would be to separate each of those into their own class. What's more, representation of ancient American cultures are missing as well as further African and Eastern European cultures. There is therefore scope to expand this dataset further.

3.9 Future Work

Whilst already being considerable, this dataset can be improved upon in a few different ways. We hope that with the approval of Le Louvre and the MET, this dataset can be made available for people to do so. The main immediate areas of work would be to:

1. Collect more data to create new classes, especially for continents currently not represented (or not enough).
2. Break down the 'Asian' class into separate classes each representing its own country or ancient civilisation.
3. Break down the 'Asian' class further to separate modern from ancient artefacts.
4. Further break down each class per era to identify style and culture changes.

Chapter 4

Design and Implementation

4.1 Overview

In this chapter, we will review the choices made when it comes to our models and their parameters. We will explain which models were chosen and why, and how each parameter was picked. This chapter will also address RQ3 - How would Transformers compare on a similar art-specific classification task? - and explain why this research path came to an end.

4.2 Convolutional Neural Network Models

4.2.1 Models

We aim to understand how pre-trained CNNs can perform on our newly created dataset. Testing multiple models on our dataset will allow us to confirm that our dataset is either of high quality, highlight its flaws, or both.

As we have seen from the literature, certain models have played a major role in the advancement of machine learning and continue to prove themselves today, namely ResNet50 and the EfficientNet family (B0, B1, and B2). Both are still widely used in studies today and act as a point of reference against which to compare new models to. This is why we have chosen to use ResNet50 and three variants of EfficientNet in our research. The EfficientNet variants chosen were B0, B1, and B2 as they were the ones for which the input size was the closest to our image size.

Once we can answer our first research question (RQ1), we will use the model which has overall the best accuracy, precision, recall, and f1-score (EfficientNetB0) and test three data balancing methods (oversampling, undersampling, and class weighting) aimed at improving those criteria when working with an imbalanced dataset.

4.2.2 Loss Function

Categorical cross entropy was chosen as the loss function. In a multi-class classification setting like ours, where only one class is true, this is the best choice. Furthermore, as our true labels are hot-encoded vectors, categorical cross entropy is used as opposed to sparse categorical entropy which would be best suited to cases where true labels are integers.

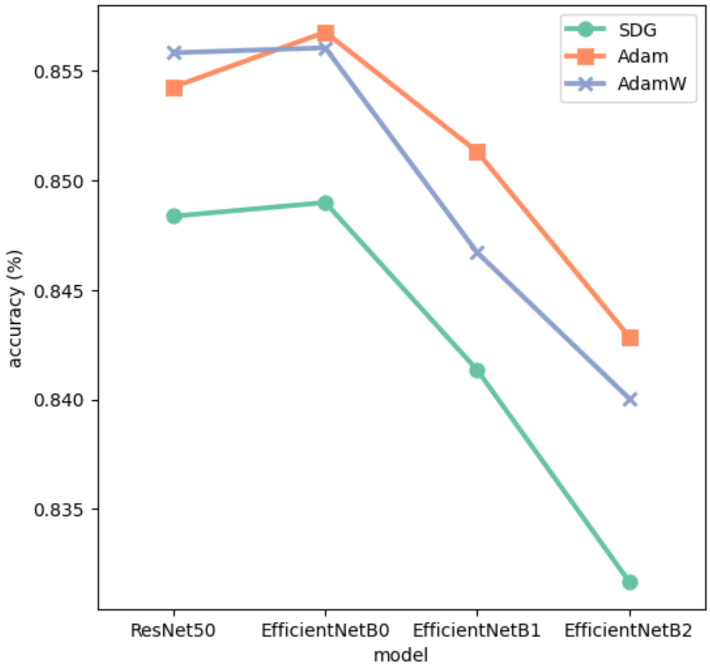


Figure 4.1: Optimizer Comparison per Model

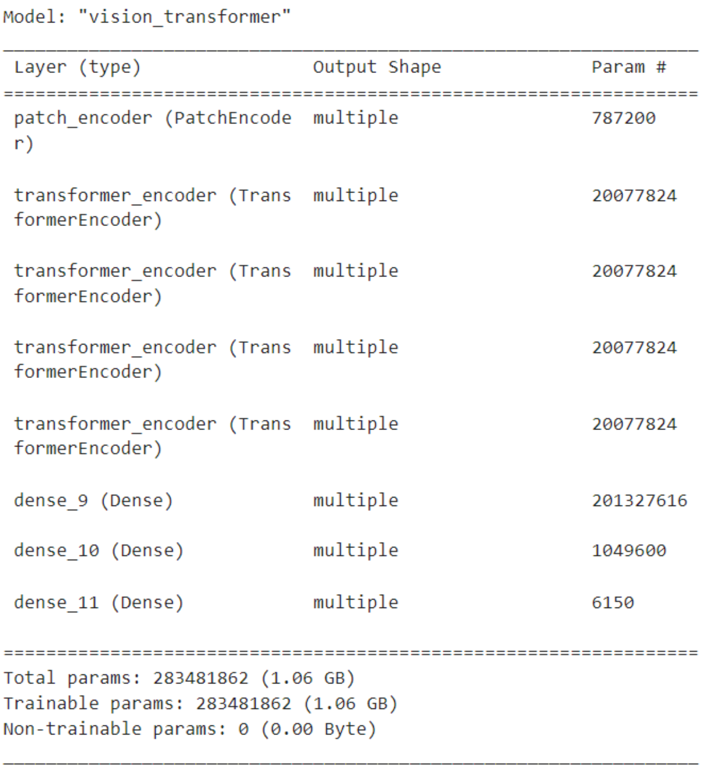
4.2.3 Optimizer

To choose the best optimizer, we compared each model's global accuracy when trained using SDG, Adam, and AdamW. We can see from figure 4.1 that all models apart from ResNet50 reach higher global accuracy when trained using Adam. For EfficentNetB1 for example, Adam improves the accuracy by nearly 0.5% over AdamW and almost 1% over SDG. For ResNet50 on the other hand, AdamW improves its accuracy by almost 0.15% over Adam, and almost 0.75% over SDG. Therefore we decided to use Adam for all the EfficientNet models, and AdamW when using ResNet50.

4.3 Vision Transformer

As we understand it from the literature, Vision Transformers are now all the rage and can obtain state-of-the-art results on vision classification tasks. Therefore comparing them and their performance to CNNs on a new dataset makes for interesting research. In our case, we will aim to create a ViT from scratch following Dosovitskiy et al. (2020)'s description of a Vision Transformer.

To do so, we followed the step-by-step video from Neuralearn (2023) to make the simplest ViT from scratch inspired by Dosovitskiy et al. (2020). Figure 4.2 shows the summary of our model. As we can see, the total number of parameters is above 283M which makes this a very big architecture.



```
Model: "vision_transformer"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 patch_encoder (PatchEncode  multiple                  787200    
 r)                                                              
                                                                 
 transformer_encoder (Trans  multiple                  20077824  
 formerEncoder)                                                  
                                                                 
 transformer_encoder (Trans  multiple                  20077824  
 formerEncoder)                                                  
                                                                 
 transformer_encoder (Trans  multiple                  20077824  
 formerEncoder)                                                  
                                                                 
 transformer_encoder (Trans  multiple                  20077824  
 formerEncoder)                                                  
                                                                 
 dense_9 (Dense)             multiple                  201327616 
                                                                 
 dense_10 (Dense)            multiple                  1049600   
                                                                 
 dense_11 (Dense)            multiple                  6150      
                                                                 
=================================================================
Total params: 283481862 (1.06 GB)
Trainable params: 283481862 (1.06 GB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Figure 4.2: Vision Transformer from Neuralearn (2023)

4.4 Encountered Issues

4.4.1 CNNs

Whilst doing a first test run of some of our models on our dataset we quickly realised that a certain number of blank images (see figure 3.2) were distributed among the classes. Those images slipped through the cracks during the manual checks of the pre-processing stage. This highlights the flaws of tedious manual checks. Using a script that scanned each folder, around 1,000 blank images were found and removed from the dataset. Another manual check confirmed that none other could be found.

4.4.2 Transformers

As we have seen in the literature and our implementation of it, Transformers have millions of parameters and require a lot more implementation power. Unfortunately due to our limited resources, we could not go through with its training. Whilst we could not go further in answering our third research question, it would certainly make for an interesting area of research for further studies. Comparing Transformers built from scratch against pre-trained Transformers could prove quite insightful.

Chapter 5

Experiments and Results

5.1 Overview

In this chapter, we will review how our models have performed and compare them against each other. Seeing how pre-trained CNNs perform on a new dataset will help us understand its flaws and strengths, as well as judge how good of a dataset we put together as per RQ1. We will also select our best-performing model and use it to understand how the imbalance of our dataset can impact the accuracy and other metrics of our model as presented by RQ2. To do so we will be comparing three well-known data balancing techniques: class weighting, undersampling, and oversampling using data augmentation.

5.2 Model Performance

The results of our experiments are presented in figure 5.1 and further details can be found in Annex A.

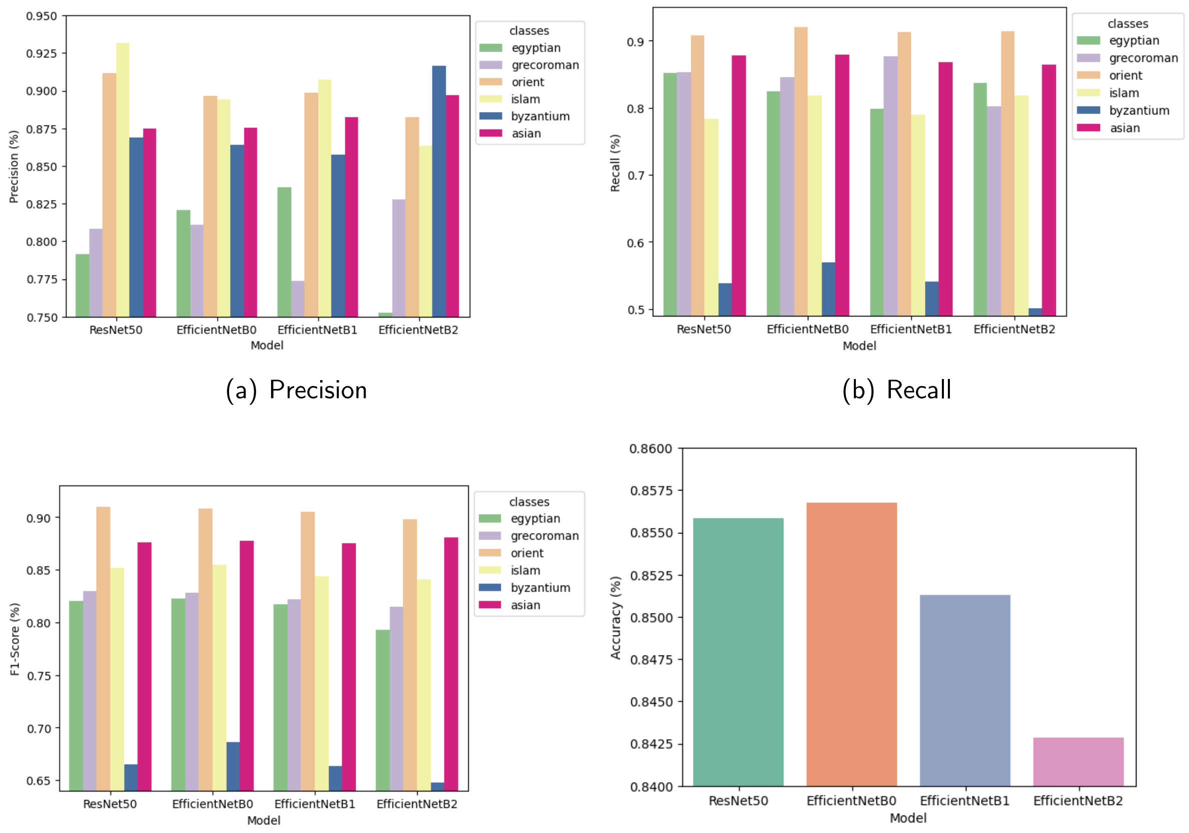
Understanding precision and recall is key to being able to understand the results. Precision and Recall are calculated using the following equations :

$$Precision = TruePositives/(TruePositives + FalsePositives)$$

$$Recall = TruePositives/(TruePositives + FalseNegatives)$$

- True Positives: the number of images of a given class being correctly predicted as belonging to that class.
- False Positives: the number of images belonging to other classes and being predicted as our given class.
- False Negatives: the number of images of a given class being incorrectly predicted as belonging to another class.

Seeing how the egyptian and grecoroman classes showed lower precision scores (75% to 83% for the egyptian class and 77% to 83% for the grecoroman class) and high recall (80% to 86% for the egyptian class and 81% to 88% for the grecoroman class), we can understand that whilst few images from both classes were wrongly predicted as belonging to other classes, many of the other classes were wrongly predicted as belonging to one of these two classes. If



(a) Precision

(b) Recall

(c) F1-Score

(d) Accuracy

Figure 5.1: Model Comparisons

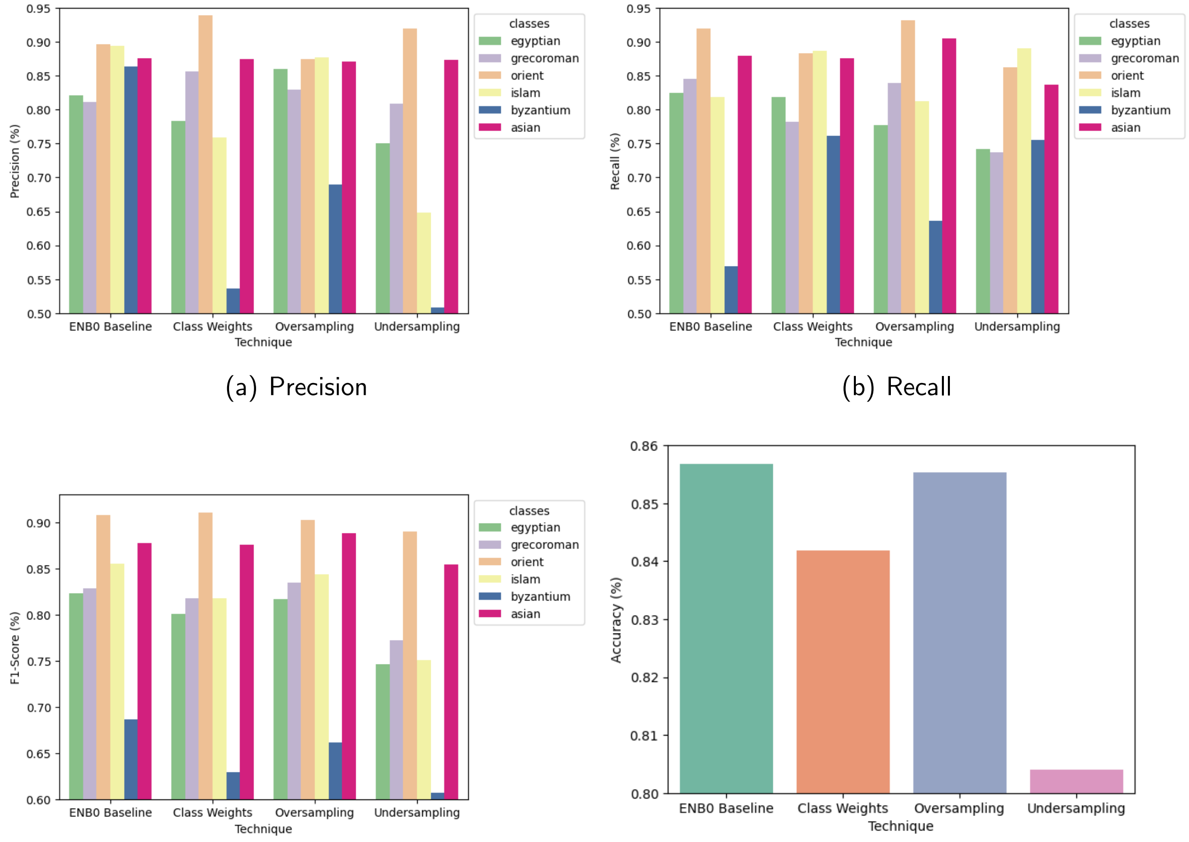
we turn our attention to the byzantium class however, quite the opposite happens : we see high precision (between 85% and 87%), and low recall (between 50% and 57%). This in turns shows that not many images from other classes were predicted as belonging to the byzantium class, whereas many of our byzantium images were wrongly predicted as belonging to one of the other classes.

When it comes to our models themselves, if we turn our attention to figure 5.1 (d), we can see how their global accuracy compares. We see that EfficientNetB0 has the highest accuracy with 85.6772%.

5.3 Imbalance Test

5.3.1 Class Weighting

The first of the three imbalance methods that we used was Class Weighting. This method was implemented using the *compute_class_weight* method which automatically calculates the weights to be attributed to each class in an unbalanced dataset. As we have seen our model is currently quite biased towards our dominant classes. This leads it to ignore the smallest (byzantium). Using this method, higher weights are given to the minority class whilst smaller weights are given to the dominant classes during the training process. This helps the model focus on the smallest classe(s) and, in theory, helps it make better predictions.



(a) Precision

(b) Recall

(c) F1-Score

(d) Accuracy

Figure 5.2: Imbalance Techniques Comparison

The following weights were calculated for our dataset :

- egyptian: 0.8470954291309979
- grecoroman: 0.7188253900176915
- orient: 0.549473403819359
- islam: 2.024650811627029
- byzantium: 4.532110866993409
- asian: 1.118752607425949

As we can see from figure 5.2 (Annex B.1), this technique has allowed an improvement of 19.1882% in the recall value of our byzantium class compared to our baseline model (our best model with no imbalance technique applied to it). This means that few images of our byzantium class are being incorrectly predicted as belonging to another class. If we look at its precision, however, we can observe that it has dropped by 32.7875%. In parallel and compared to our baseline model, if we observe the dominant classes which are correlated to the minority class, we can observe the following:

- The grecoroman class' precision went up by 4.513% whilst its recall went down by 6.3031%.
- The egyptian class' precision and recall went down by 3.7396% and 0.5656% respectively.

5.3.2 Under Sampling

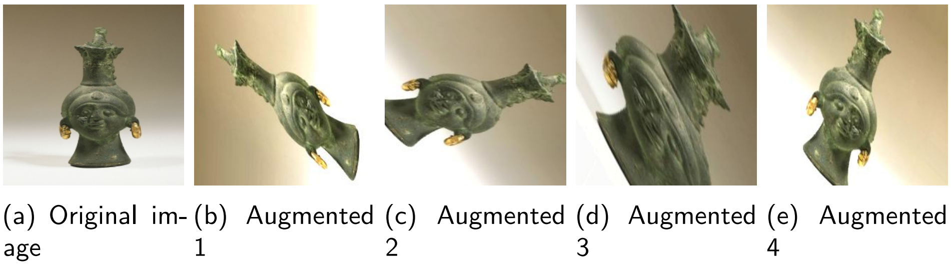
To test out the undersampling method, we modified the dataset to have the exact same number of images in all classes. To do so we created a script which randomly picked 7,317 images from each class. We then trained our model on this artificially balanced dataset.

The first and most obvious point that we can make from figure 5.2 (Annex B.1) is that the overall accuracy of our model significantly dropped from 85.6772% to 80.4051% : about 5% less.

Overall, the model acted in a similar way - if not slightly worse - than when we used class weights. Compared to the baseline :

- The byzantium class' precision went down by 35.5948% whilst its recall went up by 18.5732%.
- The egyptian class' precision and recall went down by 7.0488% and 8.3032% respectively.
- The grecoroman class' precision and recall went down by 0.1593% and 10.8082% respectively.

5.3.3 Manual Data Augmentation



(a) Original image (b) Augmented 1 (c) Augmented 2 (d) Augmented 3 (e) Augmented 4

Figure 5.3: Manual Data Augmentation. Oenochoé (Louvre, n.d.).

Finally our last technique also known as oversampling, had us manually augment the data of our smallest class in order to create more training data for our model. To do so we created a script which used ImageDataGenerator, and which allowed us to create any number of images with different zoom, brightness and distortion levels. Figure 5.3 shows an example of byzantium data with different levels of data augmentation applied.

Compared to our baseline model:

- The byzantium class' precision went down by 17.4473% whilst its recall went up by 6.642%.
- The egyptian class' precision went up by 3.9671% and its recall went down by 4.7964%.
- The grecoroman class' precision and recall went down by 1.9101% and 0.5607% respectively.

5.4 Limitations and Assumptions

Whilst this work provides a window into cultural-heritage specific vision tasks, it does so on a small-scale. Indeed this work focuses on 6 classes and is limited to a few cultures and continents. Certain cultures and continents are absent from it, such as : Oceania, South America (Aztec, Maya, Inca), North America's Native american heritage, and many others. This was due to a strict and limited amount of time, and a lack of available data. Given more time and resources, we believe it necessary to gather and include the data from those many cultures and civilisations from a professional and ethical standpoint.

In addition to this, whilst most classes focus on one particular culture, the asian class groups together japanese, chinese, and korean artefacts, amongst others. This choice was made in order to minimise imbalance and create a class with substantial amounts of data. We however recognise the moral and ethical implications of this choice and we recognise that each of those cultures should be separated if this work was to continue further.

Regarding our experiments and due to time constraints, the imbalance tests were limited to the given three techniques when applied on their own. We did not investigate how models could perform when those techniques are combined. Those techniques were also tested in their simplest and standard form. By this we mean :

- Class Weights : Those were calculated automatically using the *computer_class_weights* method and we did not experiment with choosing our own weights and seeing how the model would react.
- Data Augmentation : We did not experiment with creating various amounts of augmented data. Instead we roughly created enough augmented data so that the smallest class would hold almost as much data as the dominant ones.
- Undersampling : We chose to make each class hold the exact same amount of data and we did not experiment on reducing the dominant class progressively to see how this would affect the results.

Finally, as we previously mentioned, due to a lack of resources we could not test the capacity of Transformers on this dataset. This would however be an interesting direction in which to take this work further.

5.5 Comparison with Previous Studies

Kambau, Hasibuan and Pratama (2018)'s work is the only work - this author has found - which has attempted to take data belonging to different ethnic groups of a population and correctly classify it. They did so using different medium such as images, audio, video and texts. Even with a small dataset, they obtained an impressive 77% accuracy on their images. Our best model in comparison, namely EfficientNetB0 pre-trained on Imagenet, attained 85.6772%. This is an improvement of 8.6772% over their model thanks to the size and quality of our dataset. Our own dataset, when compared to theirs, is at least 400 times bigger and covers cultures and civilisations acrosss various continents (Europe, Asian and African).

5.6 Summary of Key Findings

Based on these results we can say that :

- Our cultural-heritage dataset achieves above 84% accuracy with all our models.
- Data Augmentation helped to create a more balanced model.

- Class weighting and undersampling alone did not improve nor balance our model.

Chapter 6

Discussion

6.1 Summary of Findings

6.1.1 Discussion of Results

Model Comparisons

The results from our models align with our predictions. Classes with very distinct styles and features such as Orient, Islam, and Asian, are well analysed and learned by CNNs. Indeed, they all see high precision and recall levels. On the other hand, correlated classes such as the egyptian, grecoroman and byzantium classes, see quite an imbalance in their results. Whereas the egyptian and grecoroman classes show lower precision and higher recall, the byzantium class follows the opposite trend. This shows that a lot of the byzantium class is being misclassified as being either egyptian or grecoroman. This in turn can be explained by the correlation between the classes as the byzantium class shares geographical ties with the other two classes, as well as being closely related to the ancient grecoroman time period.

This imbalanced pattern can be found in all the models we used. Turning our attention to their global accuracy, we have seen that EfficientNetB0 has the highest percentage. To select the best model, however, accuracy is not the only factor that we must take into account. EfficientNetB2 for example sees better precision levels for our byzantium class. It however accomplishes this by having the lowest precision level for the egyptian class, and the lowest recall level for the byzantium class. EfficientNetB0 is better and more balanced overall whilst having better accuracy. This is why this model was the one used to test our three imbalance techniques.

Class Weights

Whilst we can see that our model is no longer simply predicting everything as being either egyptian or grecoroman (subsection 5.3.1), it seems to still struggle to differentiate the byzantium class from the other two. Due to the class weights, it could be that our model - now focusing on our smallest class - is making broad assumptions in the training stages as to what makes an image belong to the byzantium class. These broad assumptions could in turn lead it to classify a lot of the images from our correlated classes as being byzantium.

Undersampling

By applying undersampling (subsection 5.3.2) and giving the model the same amount of data for each class, it has forced it to focus on each class equally. This seems to have had a similar effect than our class weights : our model does not identify most byzantium images as belonging to either the egyptian or grecoroman class. Instead, our model mostly correctly identifies byzantium images as belonging to the byzantium class. It however seems to struggle to differentiate from the other two as evidence by byzantium's poor precision value.

Regarding the model's accuracy, which dropped of about 5% when compared to the baseline model, this is understandable seeing as we went from a dataset containing 201,123 training images to a dataset of merely 43,902 images.

Oversampling

As we can see from our results (subsection 5.3.2), whilst the precision of our model went down 17.4473% for our byzantium class, it still maintains a good rate when compared to our class weights and undersampling models. Where they can not go past 55% precision, this model reaches 68.9333%. Its recall on the other hand, whilst seeing a 6.642% improvement on our baseline, falls below the other two at 63.5916%.

Overall, however, data augmentation gives us the most balanced results. Indeed when comparing it to our baseline and against the other two data balancing techniques, we can see that the imbalance, whilst still being present, is not as drastic.

6.1.2 Hypotheses

Whilst data augmentation, as we have seen, has helped bring some balance to our model, both class weights and undersampling have given us worse results than our baseline model with no imbalance technique applied to it. This could be the result of one or many factors :

- Our byzantium class is geographically linked to both our egyptian and grecoroman classes, as well as being very close to the grecoroman time period.
- When looking through our dataset, we can observe colour, texture, and shape similarities between the correlated classes. For example in figure 3.1, images (o) and (f) share resemblance in terms of colour, texture, and material. Images (a), (d), and (m) also share similar colours.
- Due to the correlation between the classes, our model is not learning the important and specific features of those three classes. It instead makes broad assumptions as to what defines each class.

It is fair to assume that the balance brought by the data augmentation could be due to our model learning better features thanks to the twisted and distorted images of our smallest class. This could help our model see patterns and features that could not be obvious otherwise.

Understanding what the model is seeing in each image using Saliency maps could help understand its predictions. Whilst this could not be explored at this time, this would make for an interesting avenue of research.

6.1.3 Key Findings

This project's key findings are as follows :

- Our cultural-heritage dataset is of high-quality as shown by all our models achieving above 84%.
- Pre-trained CNNs can generalise and apply well to new high-quality datasets.
- In this context imbalance had little impact on the accuracy of our model.
- Correlation between classes impacted our precision and recall levels more than imbalance.
- Data augmentation helped bring balance to our correlated classes' precision and recall values.

6.1.4 Research Objectives

The work done through this project has allowed us to successfully bring answers to most of our research questions.

RQ1: After putting together an extensive dataset, we showed that pre-trained CNNs could perform really well on cultural-heritage specific dataset. All of our models achieved an accuracy of 84% and above, with our best model - EfficientNetB0 - achieving 85.6772% accuracy.

RQ2 : We showed that, in this instance, trying to counter class imbalance using class weights or undersampling could actually lead to worse results. Oversampling (Data Augmentation) on the other hand was the better option and overall helped balance the model. The differences however were minimal and in our case, class imbalance had little impact on the accuracy of our CNN.

RQ3 : Regarding our last research question however, due to a lack of resources we could not explore this any further. We therefore had to leave this aside for future endeavours.

6.2 Contributions

The first contribution of this project is an extensive literature review focusing on image classification. The review covered deep learning classification and models, transfer learning, imbalance, and recent advances such as Transformers. This served as the foundation for all other contributions.

The main contribution of this project is the creation of a novel cultural-heritage dataset. This dataset contains 223,486 images spread over six classes : egyptian, grecoroman, orient, islam, byzantium, and asian. This new dataset will enable researchers to develop more tools to help the preservation of our heritage. It is our hope that it becomes a key resource for art and cultural-heritage vision research.

The third contribution of this project is the design and implementation of classification models to determine the quality of our dataset. This contribution brings legitimacy to our above mentioned dataset.

The last contribution of this project is the research and analysis of imbalance in the context of this project. We used and compared class weights, oversampling, and undersampling and observed how each of those techniques impacted the precision, recall, and accuracy of our model. Analysing those three techniques allowed us to appreciate how intricate and difficult it is for a model to differentiate highly correlated classes.

6.3 Limitations and Areas for Future Research

6.3.1 Expanding the Dataset

Per Culture and Civilisation

As we discussed in sections 3.8 and 5.4, time constraints impacted the scale of this project. Therefore, a number of cultures and civilisations are currently not represented in this dataset. It is our hope that perhaps through collaboration or via a dedicated individual, this dataset could see the addition of many classes. Initially, the asian class could be divided into the many cultures it represents. After which, classes could be added to provide representation for continents and civilisations it does not currently cover. Finally additional data could be found and included in the smaller classes.

Per Era

To take this work even further, each class could be divided in the various eras it covers. This could help build a framework capable of identifying not only the provenance of an artefact but also giving it an estimated age.

6.3.2 Continuing Research on Imbalance

Our humble research on imbalance lead us to understand that even in perfectly balanced datasets, correlated classes can still create imbalanced results. Creating saliency maps could therefore help us understand what our model is seeing and where it is struggling. Whilst we chose to simply use models pre-trained on ImageNet, Lecoutre, Negrevergne and Yger (2017)'s work shows us that in order to achieve the best performance on an artistic dataset, our model needs to be deeply retrained (about 20 layers) on a similar artistic dataset. It would therefore be interesting to put our own dataset through the test and see whether this would help it learn better and more specific features, especially for our correlated classes.

6.3.3 Using Transformers

As detailed above, due to our limited resources we could not experiment with Transformers, be that from scratch or pre-trained. As we have seen from the literature, Transformers are seeing a huge gain in popularity due to their performance over CNNs. Although there are pros and cons to both, and although Transformers do better with more data, it would be interesting to see how this new dataset could perform when paired with a Transformer. We believe that this would be a beneficial avenue of research. As artistic and cultural-heritage dataset are so different from general dataset such as ImageNet (Lecoutre, Negrevergne and Yger, 2017), it would be interesting to see how Transformers perform on them and how results differ compared to CNNs.

Bibliography

Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T. and Veit, A., 2021. *Understanding robustness of transformers for image classification* [Online]. Cornell University Library. Available from: https://arxiv.org/abs/2103.14586 [Accessed 16 February 2024].

Chan, H.P., Samala, R.K., Hadjiiski, L.M. and Zhou, C., 2020. Deep learning in medical image analysis [Online]. In: G. Lee and H. Fujita, eds. *Deep learning in medical image analysis: Challenges and applications*. Springer, *Advances in Experimental Medicine and Biology*, vol. 1213, pp.3–21. Available from: https://link.springer.com/chapter/10.1007/978-3-030-33128-3_1 [Accessed 06 March 2024].

Cordonnier, J.B., Loukas, A. and Jaggi, M., 2020. *On the relationship between self-attention and convolutional layers* [Online]. Cornell University Library. Available from: https://arxiv.org/abs/1911.03584 [Accessed 01 March 2024].

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database [Online]. *2009 ieee conference on computer vision and pattern recognition*. Miami, FL: IEEE, pp.249–255. Available from: https://ieeexplore.ieee.org/document/5206848 [Accessed 21 February 2024].

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N., 2020. *An image is worth 16x16 words: Transformers for image recognition at scale* [Online]. Cornell University Library. Available from: https://arxiv.org/abs/2010.11929 [Accessed 01 March 2024].

Hampel, F., 2002. Some thoughts about classification [Online]. In: K. Jajuga, A. Sokołowski and H.H. Bock, eds. *Classification, clustering, and data analysis*. Springer, pp.5–26. Available from: https://link.springer.com/book/10.1007/978-3-642-56181-8 [Accessed 04 March 2024].

He, K., Zhang, X., Ren, S. and Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification [Online]. *2015 ieee international conference on computer vision (iccv)*. Santiago, Chile: IEEE, pp.1026–1034. Available from: https://ieeexplore.ieee.org/document/7410480 [Accessed 04 March 2024].

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition [Online]. *2016 ieee conference on computer vision and pattern recognition (cvppr)*. Las Vegas, NV: IEEE, pp.770–778. Available from: https://ieeexplore.ieee.org/document/7780459 [Accessed 01 March 2024].

Iman, M., Arabnia, H.R. and Rasheed, K., 2023. A review of deep transfer learning and recent advancements. *Technologies* [Online], 11. Available from: https://www.mdpi.com/2227-7080/11/2/40 [Accessed 06 March 2024].

Imran, S., Naqvi, R.A., Sajid, M., Malik, T.S., Ullah, S., Moqurrab, S.A. and Yon, D.K., 2023. Artistic style recognition: Combining deep and shallow neural networks for painting classification [Online]. *Mathematics (basel)*. vol. 11, pp.4564–4591. Available from: https://www.proquest.com/publiccontent/docview/2893160636?pq-origsite=primo&sourcetype=Scholarly%20Journals [Accessed 26 February 2024].

Johnson, J.M. and Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. *Journal of big data* [Online], 6, pp.1–54. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0192-5#citeas [Accessed 15 March 2024].

Kambau, R.A., Hasibuan, W.A. and Pratama, M., 2018. Classification for multiformat object of cultural heritage using deep learning [Online]. *2018 third international conference on informatics and computing (icic)*. IEEE, pp.1–7. Available from: https://doi.org/10.1109/IAC.2018.8780557 [Accessed 26 February 2024].

Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S. and Shah, M., 2022. Transformers in vision: A survey. *Acm computing surveys* [Online], 54, pp.1–41. Available from: https://dl.acm.org/doi/full/10.1145/3505244 [Accessed 15 March 2024].

Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions [Online]. *Progress in artificial intelligence*. Springer, vol. 5, p.221–232. Available from: https://link.springer.com/article/10.1007/s13748-016-0094-0#article-info [Accessed 15 March 2024].

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks [Online]. *Proceedings of the 25th international conference on neural information processing systems*. Curran Associates Inc., vol. 1, p.1097–1105. Available from: https://dl.acm.org/doi/10.5555/2999134.2999257 [Accessed 29 February 2024].

Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks [Online]. In: M.Y. Vardi, ed. *Communications of the acm*. ACM Press, vol. 60, pp.84–90. Available from: https://dl-acm-org.ezproxy1.bath.ac.uk/doi/abs/10.1145/3065386 [Accessed 21 February 2024].

Lecoutre, A., Negrevergne, B. and Yger, F., 2017. Recognizing art style automatically in painting with deep learning [Online]. *Journal of machine learning research*. Seoul: Microtome Publishing, vol. 77, pp.327–342. Available from: https://proceedings.mlr.press/v77/lecoutre17a/lecoutre17a.pdf [Accessed 26 February 2024].

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T. and Xie, S., 2022. *A convnet for the 2020s* [Online]. Cornell University Library. Available from: https://www.proquest.com/docview/2618756580?pq-origsite=primo&sourcetype=Working%20Papers [Accessed 15 March 2024].

Louvre, n.d. *Louvre site des collections* [Online]. Available from: https://collections.louvre.fr/en/recherches [Accessed 20 March 2024].

McCulloch, W.S. and Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity [Online]. *The bulletin of mathematical biophysics*. Springer, vol. 5, pp.115––133. Available from: https://link.springer.com/article/10.1007/BF02478259 [Accessed 29 February 2024].

MET, n.d. *The metropolitain museum of art* [Online]. Available from: https://www.metmuseum.org/art/collection [Accessed 20 March 2024].

Neuralearn, 2023. *Building a vision transformers (vit) with tensorflow 2 from scratch - human emotions detection*. Available from: https://www.youtube.com/watch?v=JcuFQdnawuE [Accessed 15 January 2024].

Nilson, T. and Thorell, K., 2018. Introduction [Online]. In: T. Nilson and K. Thorell, eds. *Heritage preservation: The past, the present and the future*. Halmstad, chap. 1, pp.9–20. Available from: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1224014&dswid=-3778 [Accessed 26 February 2024].

Obeso, A.M., Vázquez, M.S.G., Acosta, A.A.R. and Benois-Pineau, J., 2017. Connoisseur: classification of styles of mexican architectural heritage with deep learning and visual attention prediction [Online]. *Cbmi '17: Proceedings of the 15th international workshop on content-based multimedia indexing*. pp.1–7. Available from: https://dl.acm.org/doi/abs/10.1145/3095713.3095730?casa_token=eOP1nnWmLhcAAAAA:M48ltHy_vD6FD1CB_SqnrsW2eSQmILPhSRZzyPDUyYRfW4cqsFgMQlXWr5j2dPJfebz1lIP9fcqU [Accessed 9 April 2023].

Paulat, T., 2023. *Modern Architecture (100k Images)* [Online]. Available from: https://www.kaggle.com/datasets/tompaulat/modernarchitecture [Accessed 29 February 2024].

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy1, A., Khosla, A., Bernstein, M., Berg, A.C. and Fei-Fei, L., 2015. Imagenet large scale visual recognition challenge [Online]. *International journal of computer vision*. vol. 115, pp.211–252. Available from: https://www.proquest.com/docview/1732536709/fulltextPDF?parentSessionId=D8Fo9x12JPsA4ewpKkXgOLGBHMDTwW2UcfJjHTOMybM%3D&pq-origsite=primo&accountid=17230&sourcetype=Scholarly%20Journals [Accessed 01 March 2024].

Russell, S. and Norvig, P., 2022. *Artificial intelligence: A modern approach*. 4th ed. Pearson Education Limited.

Sabatelli, M., Kestemont, M., Daelemans, W. and Geurts, P., 2018. Deep transfer learning for art classification problems [Online]. In: L. Leal-Taixé and S. Roth, eds. *Computer vision – eccv 2018 workshops*. Munich, Germany: Springer Cham, pp.631–646. Available from: https://link.springer.com/chapter/10.1007/978-3-030-11012-3_48 [Accessed 26 February 2024].

Silverman, H. and Ruggles, D.F., 2007. Cultural heritage and human rights [Online]. In: H. Silverman and D.F. Ruggles, eds. *Cultural heritage and human rights*. Springer, chap. 1, pp.3–29. Available from: https://link.springer.com/chapter/10.1007/978-0-387-71313-7_1 [Accessed 26 February 2024].

Smith, S.L., Brock, A., Berrada, L. and De, S., 2023. *Convnets match vision transformers at scale* [Online]. Cornell University Library. Available from: https://www.proquest.

com/docview/2882110506?pq-origsite=primo&sourcetype=Working%20Papers [Accessed 15 March 2024].

Stevenson, A., ed., 2010. *Oxford dictionary of english* [Online]. 3rd ed. Oxford University Press. Available from: https://www-oxfordreference-com.ezproxy1.bath.ac.uk/display/10.1093/acref/9780199571123.001.0001/m_en_gb0374630?rskey=zf7qy6&result=41773 [Accessed 26 February 2024].

Tan, M. and Le, Q.V., 2020. *Efficientnet: Rethinking model scaling for convolutional neural networks* [Online]. Cornell University Library. Available from: https://www.proquest.com/docview/2231642984?pq-origsite=primo&sourcetype=Working%20Papers [Accessed 21 February 2024].

Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A. and Li, Y., 2022. Maxvit: Multi-axis vision transformer [Online]. In: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella and T. Hassner, eds. *Computer vision – eccv 2022*. Springer, *Lecture Notes in Computer Science*, vol. 13684, p.459–479. Available from: https://link.springer.com/chapter/10.1007/978-3-031-20053-3_27 [Accessed 15 March 2024].

UNESCO, 2021. *Preserving our heritage* [Online]. Available from: https://en.unesco.org/content/preserving-our-heritage [Accessed 26 February 2024].

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. *Attention is all you need* [Online]. Cornell University Library. Available from: https://doi.org/10.48550/arxiv.1706.03762 [Accessed 01 March 2024].

Wightman, R., Touvron, H. and Jégou, H., 2021. *Resnet strikes back: An improved training procedure in timm* [Online]. Cornell University Library. Available from: https://www.proquest.com/docview/2578930653?pq-origsite=primo&sourcetype=Working%20Papers [Accessed 21 February 2024].

WikiArt, n.d. *Visual Art Encyclopedia* [Online]. Available from: https://www.wikiart.org/ [Accessed 29 February 2024].

Zagoruyko, S. and Komodakis, N., 2017. *Wide residual networks* [Online]. Cornell University Library. Available from: https://www.proquest.com/docview/2076206202?pq-origsite=primo&sourcetype=Working%20Papers [Accessed 21 February 2024].

Zhou, M., Geng, G. and Wu, Z., 2012. *Digital preservation technology for cultural heritage* [Online]. Springer Berlin. Available from: https://link.springer.com/book/10.1007/978-3-642-28099-3 [Accessed 26 February 2024].

Appendix A

Model Comparisons

Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
egyptian	0.801177	0.831448	0.816032	0.803852	0.821493	0.812577	0.791754	0.851584	0.82058
grecoroman	0.830997	0.825213	0.828095	0.792208	0.849188	0.819709	0.808101	0.852475	0.829695
orient	0.887848	0.919976	0.903627	0.915111	0.894381	0.904627	0.911669	0.907914	0.909788
islam	0.891048	0.82139	0.854802	0.861159	0.814875	0.837378	0.931613	0.783931	0.851415
byzantium	0.847036	0.544895	0.663174	0.747164	0.567036	0.644755	0.869048	0.538745	0.665148
asian	0.87437	0.889391	0.881817	0.881443	0.87613	0.878779	0.874737	0.877637	0.876185
accuracy			0.854268			0.848366			0.855833
macro avg	0.855413	0.805386	0.824591	0.83349	0.80385	0.816304	0.864487	0.802048	0.825468
weighted avg	0.85435	0.854268	0.852846	0.849152	0.848366	0.847972	0.858629	0.855833	0.854945

Table A.1: ResNet50 with Adam - ResNet50 with SDG - ResNet50 with AdamW

Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
egyptian	0.820946	0.824661	0.822799	0.787601	0.836425	0.811279	0.790163	0.846833	0.817517
grecoroman	0.811016	0.845514	0.827906	0.808035	0.824439	0.816155	0.819849	0.838554	0.829096
orient	0.896374	0.919976	0.908022	0.910966	0.90306	0.906996	0.893736	0.919241	0.906309
islam	0.894425	0.818675	0.854875	0.852208	0.817047	0.834257	0.899165	0.818132	0.856737
byzantium	0.863806	0.569496	0.686434	0.831261	0.575646	0.680233	0.871893	0.560886	0.682635
asian	0.87545	0.879144	0.877293	0.876354	0.877939	0.877145	0.906836	0.859554	0.882562
accuracy			0.856772			0.848992			0.856057
macro avg	0.860336	0.809578	0.829555	0.844404	0.805759	0.821011	0.863607	0.8072	0.829143
weighted avg	0.857276	0.856772	0.855656	0.849905	0.848992	0.848404	0.857773	0.856057	0.855164

Table A.2: EfficientNetB0 with Adam - EfficientNetB0 with SDG - EfficientNetB0 with AdamW

Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
egyptian	0.836058	0.798416	0.816804	0.765317	0.83371	0.798051	0.76297	0.848416	0.803428
grecoroman	0.773549	0.87645	0.821791	0.790501	0.83024	0.809883	0.808107	0.840294	0.823886
orient	0.89845	0.912327	0.905335	0.91076	0.897764	0.904215	0.917159	0.89085	0.903813
islam	0.907049	0.789359	0.844122	0.881807	0.773616	0.824176	0.866706	0.801303	0.832722
byzantium	0.8577	0.541205	0.66365	0.843874	0.525215	0.64746	0.878661	0.516605	0.650658
asian	0.882623	0.867993	0.875247	0.874886	0.868294	0.871578	0.878576	0.870102	0.874319
accuracy			0.851317			0.841345			0.846711
macro avg	0.859238	0.797625	0.821158	0.844524	0.78814	0.809227	0.85203	0.794595	0.814804
weighted avg	0.854111	0.851317	0.850223	0.844061	0.841345	0.840646	0.850183	0.846711	0.846052

Table A.3: EfficientNetB1 with Adam - EfficientNetB1 with SDG - EfficientNetB1 with AdamW

Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
egyptian	0.752594	0.836878	0.792501	0.753141	0.813801	0.782297	0.781054	0.815158	0.797742
grecoroman	0.827813	0.802204	0.814808	0.781782	0.826373	0.803459	0.79985	0.823666	0.811583
orient	0.882236	0.914681	0.898166	0.914758	0.887173	0.900754	0.868461	0.915858	0.89153
islam	0.86361	0.818132	0.840256	0.866794	0.738328	0.79742	0.913411	0.761672	0.830669
byzantium	0.916667	0.500615	0.647574	0.830612	0.500615	0.624712	0.904656	0.501845	0.64557
asian	0.896843	0.864678	0.880466	0.844867	0.883062	0.863543	0.883109	0.869801	0.876404
accuracy			0.842865			0.831686			0.840048
macro avg	0.856627	0.789531	0.812295	0.831992	0.774892	0.795364	0.858424	0.781333	0.808916
weighted avg	0.845911	0.842865	0.841497	0.834681	0.831686	0.830771	0.842509	0.840048	0.838304

Table A.4: EfficientNetB2 with Adam - EfficientNetB2 with SDG - EfficientNetB2 with AdamW

Appendix B

Imbalance Comparison

Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
egyptian	0.78355	0.819005	0.800885	0.860617	0.776697	0.816506	0.750458	0.741629	0.746017
grecoroman	0.856146	0.782483	0.817658	0.830117	0.839907	0.834983	0.809423	0.737432	0.771752
orient	0.939418	0.88276	0.910208	0.875069	0.93145	0.90238	0.920082	0.862018	0.890104
islam	0.759182	0.886536	0.817931	0.877491	0.812704	0.843856	0.648734	0.890337	0.750572
byzantium	0.535931	0.761378	0.629065	0.689333	0.635916	0.661548	0.507858	0.755228	0.607319
asian	0.875075	0.87613	0.875602	0.870977	0.905365	0.887838	0.873466	0.836347	0.854503
accuracy			0.841882			0.855297			0.804051
macro avg	0.79155	0.834715	0.808558	0.833934	0.817007	0.824519	0.75167	0.803832	0.770045
weighted avg	0.850291	0.841882	0.84424	0.854656	0.855297	0.854087	0.81671	0.804051	0.807198

Table B.1: Class Weighting - Data Augmentation (Oversampling) - Undersampling

Appendix C

Data Management Plan

Data Management Plan

1 Overview

1.1 Project title

Which ancient civilisation does this artefact belong to? This project will produce a deep learning model that identifies the provenance of the pictured artefact.

1.2 PI name and department

Note: the University of Bath Principal Investigator is the Data Steward for the project.

Dr Georgios Exarchakis, Computer Science

1.3 Project description

This project aims to produce a new dataset of cultural artefacts belonging to 6 different civilisations as well as comparing various models in identifying the correct provenance of a given artefact. All data will be collected using Scrapy. Each of the images will be resized to a standard 256x256. We will then build and compare pre-trained models such as ResNet50 and EfficientNetB0-B2 to a Transformer build from scratch.

2 Compliance

2.1 University policy requirements

This project will comply with the policies listed below.

University policy or guidance

- University of Bath Research Data Policy
- University of Bath Code of Good Practice in Research Integrity
- University of Bath Electronic Information Systems Security Policy
- University of Bath Intellectual Property Policy
- University of Bath Code of Ethics

2.2 Legal requirements

The following legislation is relevant to this project.

- United States copyright laws

UK Legislation or framework

N/A

2.3 Contractual requirements

Name of funder	Data policy URL
Le Louvre	https://collections.louvre.fr/en/page/cgu
The MET	https://www.metmuseum.org/policies/terms-and-conditions

3 Gathering data

3.1 Description of the data

3.1.1 Types of data

Data used by this research will be in the form of images.

3.1.2 Format and scale of the data

This research will generate .JPEG file formats. It is expected that roughly around 200,000 images will be collected amounting to 37GB of raw data.

3.2 Data collection methods

The main data collection method is expected to be Scrapy. The data will come from museum websites such as Le Louvre and The MET.

3.3 Development of original software

This entire project will be written using Python. To collect the data we will use Scrapy. For the image classification task Keras and TensorFlow will be used. The image classification scripts will be for use only for this project however the aim is to make the dataset available to researchers for further projects.

4 Working with data

4.1 Short- and medium-term data storage arrangements

All research data will be stored on the University managed storage (X: or H: Drive): **No**

The data will be stored on my computer as well as an external drive for backup. The CSV files used to store the images' URLs to download them will also be stored on both as a backup.

To submit this paper, the dataset will be stored on GoogleDrive and accessible to anyone using the following link (with Viewer rights only): https://drive.google.com/drive/folders/1H-2idZH1GAuY6YJMPX0wFmASzRG2QJGY?usp=sharing

4.2 Control of access to data and sharing with collaborators

The data will only be accessible to those reading and reviewing the dissertation. This will be done via a link to a GoogleDrive. Their rights will be strictly limited to Viewer rights.

4.3 Documentation that will accompany the data

Each image will be labelled in the following way : CLASSID_MUSEUMNAME_UNIQUEID.JPEG. For example an image from the egyptian class collected from the Louvre might look like this : 1_LOUVRE_26784283.JPEG. The Unique ID will be taken from the source's website so that it is clearly identifiable and can be linked back to its source. The dataset will be clearly described and visualised within a chapter of my dissertation.

5 Archiving data

5.1 Selection of data to be retained and deleted at the end of the project

The data on the GoogleDrive will be deleted at the end of the project (maximum of a year - after all submission requirements are fulfilled). However the copy on my external drive will be kept as the goal is to discuss with both Le Louvre and The MET in order to make this newly put together dataset available for researchers to use. Depending on agreements, it will either be made available under a non-commercial licence for research/academic purposes, or it will be deleted from both drives.

5.2 Data preservation strategy and retention period

The data will be kept on my external drive for a time of at least a year. This drive never leaves the house and no one else has access to it. Preserving it further on a research data archive will be looked into.

5.3 Maintenance of original software

The code will be made available to the University of Bath and will be on GitHub in a private repository. After the end of the project this code is not expected to be developed further and no one else will have the rights to edit it. Further research on the dataset might however be done.

6 Sharing data

6.1 Justification for any restrictions on data sharing

To respect Le Louvre's and The MET's terms and conditions and copyright laws, access to the dataset will be limited and not shared openly. Whilst most images from The MET are under the CC licence, some are still under copyright/restrictions and can therefore not be redistributed further. The copyright/restrictions rules are similar for Le Louvre.

6.2 Arrangements for data sharing

Discussions with both parties (Le Louvre and The MET) will be undertaken in order to be able to make this new dataset available to academics and researchers for future research and works. If both parties agree this dataset could be made available through kaggle with a non-commercial licence.

7 Implementation

7.1 Review of the Data Management Plan

The Data Management Plan will be reviewed every month to ensurance that it is kept up to date.

Appendix D

Ethics Approval Letter

Looks like an interesting project !

UNIVERSITY OF BATH

Research Governance and Compliance

Vice-Chancellor's Office

University of Bath

Bath BA2 7AY

26/02/2024

Dear Mathilde

Ethics application reference number: 3967-3860

Project title: Which ancient civilisation does this artefact belong to? A deep learning model that identifies the provenance of the pictured artefact.

The above application has been considered in line with the University of Bath ethics review processes. Please accept this letter as confirmation that no further review is required by any other Committee or Board.

The following documents were reviewed:

Document Type	File Name	Date	Version
H1 Other documentation	Data_Management_Plan (3)	23/02/2024	1
H1 Other documentation	Louvre email chain	23/02/2024	1

You can view the application and any reviewer comments here: https://ethics.bath.ac.uk/Project/Index/4076

Your project can now start and data collection can commence.

If there are any changes to this project (including amendments to the design, sample, or start/end dates etc.), you will need to submit an amendment.

Kind regards,

Prof Peter Hall