Sci

MDPI

Article

Siamese Deep Networks for Ancient Coin Matching

Zhongliang Guo and Ognjen Arandjelović *

School of Computer Science, University of St Andrews, Scotland KY16 9AJ, UK; ognjen.arandjelovic@gmail.com
* Correspondence: ognjen.arandjelovic@gmail.com; Tel.: +44-(0)1334-48-26-24

Received: 29 February 2020; Accepted: 18 March 2020; Published: date

Abstract: Computer vision and numismatics have been cooperating closely in recent years to solve the problems of ancient coins such as classification, identification. The current popular area of research is coin classification. However, due to the limitations of the quantity and coverage of existing datasets (e.g., AnCoins-12, RRC-60), pre-existing classification models cannot classify samples from unknown issues. To overcome such insufficiency of the coin classification method in the status quo, the project has attempted to define the labelling problem as coin matching - rather than coin classification. The project proposes a new ancient coin matching methodology that uses a set of deep network architectures, called Single Siamese ViT (Vision Transformer) and Double Siamese ViT. To elaborate, coin issues and the image semantics of the obverse and reverse sides of coins are matched via this methodology. The two models are obtained by training the Single Siamese ViT on the obverse and reverse sides of a coin, respectively. Afterwards, the two models are loaded into the Double Siamese ViT in combination with transfer learning, where the network requires only a small amount of data (566 images) to achieve a good evaluation result on a dataset of 14,820 images. Furthermore, the Double Siamese ViT model achieved an overall accuracy of 81% in matching each emperor's issues. Among the 150 emperors, 135 emperors have reliable performance results (over 70%). Finally, the project has also analysed the potential reasons for the inferior performance of this model on some emperors with.

Keywords: siamese neural network; ancient coin matching; deep learning; computer vision; machine learning; low-shot learning

1. Introduction

Ancient numismatics focus on studying ancient coins, paper currency and tokens. This subject has piqued the interests of academics, antique dealers, and collectors throughout history and in different continents. However, the field is still not well-known to the general public and it only has attracted the attention of specialised segment of the market and academia. Therefore, some terms[1] of numismatics shall be clarified for the ease of understanding for readers who are not familiar with ancient coins and their peculiarities.

- **coin**
  This term refers to a single coin, a specific and unique physical existence. It is different from the term **"issue"**.
- **issue**
  This term refers to a kind of coin. For example, if the head and tail of two coins illustrate the same objects or attributes (same human, clothes, characters, etc.), they are of the same issue.

  For example, Figure 1 illustrate two different coins which are the same issue.

(**a**) Sample 1 from RIC439 of Aelius (**b**) Sample 2 from RIC439 of Aelius

Figure 1. Samples from the same issue

Computer vision and ancient numismatics have been cooperating closely in recent years to solve the problems such as classification, identification, matching, etc. The study in this field was drawn and valued by collectors and museums globally due to the historical meaning and beauty of such ancient treasures. To facilitate the reader's understanding, the terms related to this project are explained below.

- **classification**
  Classification refers to the division of the given data into groups according to certain criteria. In this project, we consider issues as groups, and the task of the classifier is to group coins of the same issue into the same class as much as possible.
- **matching**
  Matching means exploring whether two things have the same appearance, design, etc. If they do, they match, otherwise they don't. In this project, matching means how similar are the two coins, usually the matching result is a number to indicates similarity.

The vast majority of the literature pieces on this field focus their attention on the classification of ancient coins with traditional classification methods[2–5]. However, for samples from the issues which the dataset doesn't cover or exist, the model will clearly not work. Although doing matching can solve this problem, most of the current ancient coin matching models are sensitive to the image quality[6–8]. Deep learning has the potential to solve this complex problem[3], but few researchers have designed methods based on deep learning in ancient coin matching. Therefore, to address these limitations, this project is aim to come up with a new ancient coin matching method using a dataset with a wide coverage of issues, and deep learning. Hence, the model can be utilised for more diverse issues of coins.

2. Related work

Researchers[6–8] in this field have been utilising conventional image classification methods to achieve the tasks such as classification of different issues and establishing the relationship between two different coins. This helped to save human resources and create academic and commercial advancements. An extensive analysis of recent research for coin classification is available in Zaharieva et al's work "Image based recognition of ancient coins"[5]. Their work illustrates that existing approaches can be extended from modern coins to ancient coins. Moreover, Wei et al[4] proposed a novel approach to extract textual information for classification and recognition of ancient coins using Tree-structured Wavelet Transform (TWT) and Ant Colony Optimisation (ACO) based algorithms. According to the authors, an energy map containing energy value and energy channels is constructed accordingly[4]. To elaborate on the term "energy", if "a channel contains significantly less energy than others, it implies that the channel usually contains less information and further decomposition should not be applied to it"[4]. In order to develop the recognition system of ancient coins which had not been studied before, researchers have attempted to extend some methods for local image descriptions of ancient coins. The selection of similar images based on feature matching has proved feasible[7], yet

the performance is not considered optimal to cover a wide-ranging variety of ancient coins. There are benefits of utilising local entropy and grey value range. As a simplistic and fast solution, it can be used to separate most of the coins from images[9]. Computer vision has also made its mark on heritage crime. Shape descriptors formed by local features allow algorithms to automatically identify stolen ancient coins on the Internet[8]. Fine-grained classification of counts has been proposed for identifying ancient coins and has received some applications due to its good performance[2]. This model can recognise and classify coins in a data set with a large number of classes as well as individual objects of high levels of similarities, but it is unable to be applied on samples from unknown issues. A vision-based approach to ancient coins' identification can achieve 99% accuracy[6], and the plain Bayesian algorithm is responsible for the fusion of features. To obtain more implicit information in the images and make the process more conducive for the algorithm to extract useful information in high-dimensional space, the human expertise and the knowledge of Numismatics are also incorporated into the design of the algorithm[10]. Such combined algorithm is demonstrated in the Bag of Visual Words (BoVWs) method which combines the computer vision with Numismatics into a framework. Subsequently, Anwar, Zambanini, and Kampel[11] improved on the BoVWs method by integrating the spatial information to the BoVWs model in a rotation-invariant way by encoding the triangular relationship among the positions of identical visual words in the 2D image space. To explore the application of semi-supervised learning on ancient coins, a new data set[12] was also extended to enable a wider variety of tasks in terms of the number of issues, and make the model more capable of learning intra-issue variations. At the same time, the researchers developed a Game Theory model on the data set, which allowed the model to achieve better performance even with a limited amount of labelled data. The above research is based on those data sets with less variety[3,12]. Nonetheless, these data sets are not enough when compared to the thousands of categories of coins in the real world. Thus we need to design and implement more robust algorithms to support ancient coin research.

The above algorithms and models[2–5] primarily focus on classifying ancient coins with conventional classification methods. Traditional classification algorithm[3–5] is difficult to be applied to such a complex task, because there are thousands of categories of ancient coins, resulting classification algorithms cannot deal with the unseen issues. For example, one cannot easily classify a coin that has never been observed with no category to be labelled readily, i.e., the classifier can only classify coins that are previously observed and labelled in named categories. Therefore, focusing on ancient coin matching rather than classification would not only solve this problem, but also make the task simpler and facilitate the development of the study. Thus far, some studies on ancient coin matching based on visual matches have been conducted[6–8]. They performed relatively well in specific situations, however, performing well with diverse backgrounds, illumination, and coin placement directions still remain challenging tasks. This is not feasible in practical applications because the quality of coin photos is affected by a series of factors such as photographic equipment, lighting intensity, etc[1]. It is significantly challenging to have a uniform standard to process to bring all images to a standard that is acceptable to classification algorithms. This critical problem can be solved by deep learning-based matching[3], but so far, there aren't many of deep-learning-based studies for ancient coin matching. The usual deep learning frameworks require enormous amounts of data to prevent over-fitting. Therefore, herein, the project's key idea is to design an ancient coin matching method rather than classification, which is aim to solve the difficulty that classification cannot deal with samples from uncovered issues. To elaborate, a powerful model is used to obtain the feature maps of coin images, and matching is achieved by comparing the similarities between feature maps, thus solving the limitations of traditional classifiers.

3. Data

RIC-ACS is the dataset I used for this project, which is from Ancient Coins Search Engine (https://www.acsearch.info/). This kind of data has high-quality images obtained in controlled environments, such as uniform background, lighting angle, height, normalised reversal angle, etc.

This dataset compressed 200,000 coins images which cover most coins issued by 228 ancient Roman emperors. For this project, I only took 100,000 coins images that cross all 228 ancient Roman emperors who issued various coins at different periods from the whole dataset. Each image corresponds to a txt file containing the emperor's name and the different type(s) of coins the emperor issued. The information that the dataset provided has a variety of coin conditions intervals from badly damaged so well preserved. RIC-ACS also has demonstrated the scarcity of a highly differentiated coin offering from different emperors (Ruling period, issuing authorities).

For splitting the data, firstly, I checked if the txt file contains the name of an emperor. If it has only one emperor name, then I check if the content of the txt file contains "RIC XX" via the regular expression "RIC.*?\d+". Afterwards, I put the coin image into "emperor/RIC XX" as its path.

Then for each image (width$\times$height), I take the obverse = image[0:$\frac{1}{2}$width, 0:height], reverse = image[$\frac{1}{2}$width:width, 0:height].

4. Methodology

Compared to classifying ancient coins, matching ancient coins is easier to adapt to uneven data distribution and a dataset with thousands of categories[13–20]. Meanwhile, so far, there is no project based on Siamese Neural Network architecture for ancient coin matching to make more training image pairs for reducing overfitting. Additionally, recent research shows that Vision Transformer provided excellent performance and fantastic potential to reach state of the art on many computer vision tasks[21–30], a potential backbone network for the ancient coin matching task. In summary, this project will focus on ancient coin matching rather than classifying. It will be designed to achieve ancient coin matching by proposing a deep network architecture based on Siamese Neural Network and Vision Transformer.

In order to achieve the task of ancient coin matching on the given dataset, this project will be divided into three sub-tasks to be performed sequentially:

- Design and train a matching model on the obverse of coins.
- Design and train a matching model on the reverse of coins.
- Design and train a matching model on both obverse and reverse of coins. For boosting training speed, transfer learning may be used to transfer the prior knowledge learned in the two previous sub-tasks to this sub-task.

4.1. Organised data

For Siamese Neural Network, its dataset is the combination of many pairs of images with labels. However, the label only has two types of values which are 0 or 1. To elaborate, if the pair of images from the same issue, the label is 1. If the pair of images from the different issues, the label is 0.

For this project, it is hard to organise a dataloader because there are around 20,000 images in over 7,000 issues. If we take all the combinations of 2 images, there will be over $C^2_{20000} = 199,990,000$ combinations. Even when we take 1 sample for each issue, there will be over $C^2_{7000} = 24,496,500$ combinations, which is almost untrainable.

Therefore, this study took a creative route and only took the issues with over 20 samples to participate in the training. Samples in each issue have then divided these issues into train set, validation set and test set in the ratio of 14:3:3. The remaining issues that did not participate in the training were called unseen issues, which were used with the test set when evaluating the model. Only then can the model be evaluated on every emperor.

After the above steps, the train set has 542 images in 24 issues, and the final model will be evaluated on a dataset consisting of 196 emperors with a total of 7605 issues. Figure 2 shows the data distribution of the train set.

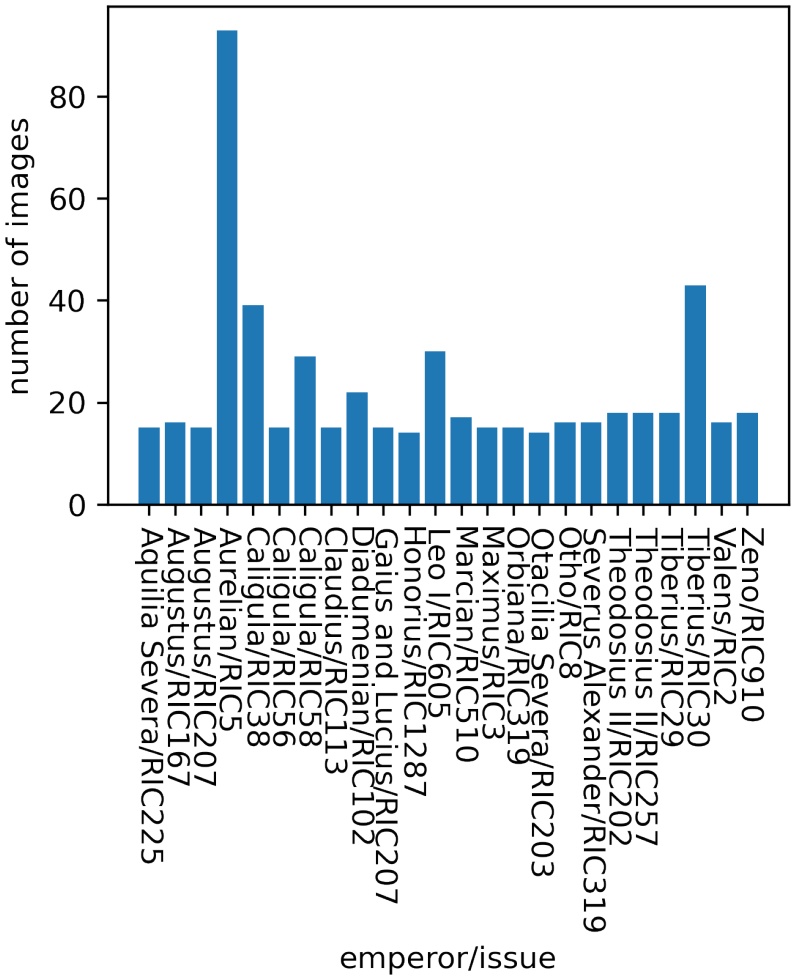


Figure 2. The data distribution of the train set

In organising the dataset, given the data distribution is relatively uneven, if two images are randomly taken to form a pair, the image pairs with label 0 in the whole dataset will far outnumber the image pairs with label 1. This will easily lead the network to tend to produce mismatched results. Therefore, we need to design a more balanced data organisation method to make the number of image pairs with label 0 and label 1 more evenly distributed. Therefore, I randomly select three images (img_1, img_2, img_3) from the same issue, and afterwards take one image(img_4) from a different issue. Then, for img_1 and img_2, they are a pair of images for training; the label is 1. And for img_3 and img_4, they are another pair of images for training; the label is 0. The flow chart of this process is in Figure 3.

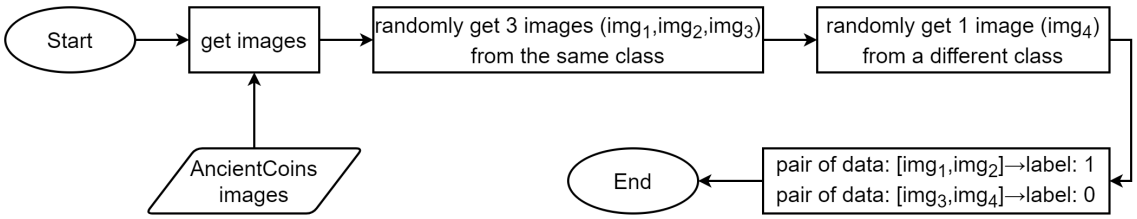


Figure 3. The flow chart for organising dataset

4.2. Proposed network architectures

For this project, I proposed two architectures for three sub-tasks based on pre-trained Vision Transformer and Siamese Neural Network, i.e., an architecture for training only based on the obverse or reverse coins, the other for training based on both obverses and reverse coins.

I chose the Vision Transformer Base (ViT-Base) model which was pre-trained on imagenet-1k.

4.2.1. Single Siamese ViT

For sub-task 1 and 2, I proposed a network architecture called Single Siamese ViT. First, I replaced the backbone network of the general Siamese Neural Network with pre-trained ViT, then flattened the semantic layer of output x_1 and x_2 of two ViT models, and calculated the absolute distance d between x_1 and x_2. Finally, using 3 Linear layers and 1 batch normalisation layer to reduce the dimension of d to a size of 1×1. For processing the output of the network, I used the *Sigmoid* function to get a number between 0 and 1 as the probability. The architecture of this network can be seen in Figure 4.

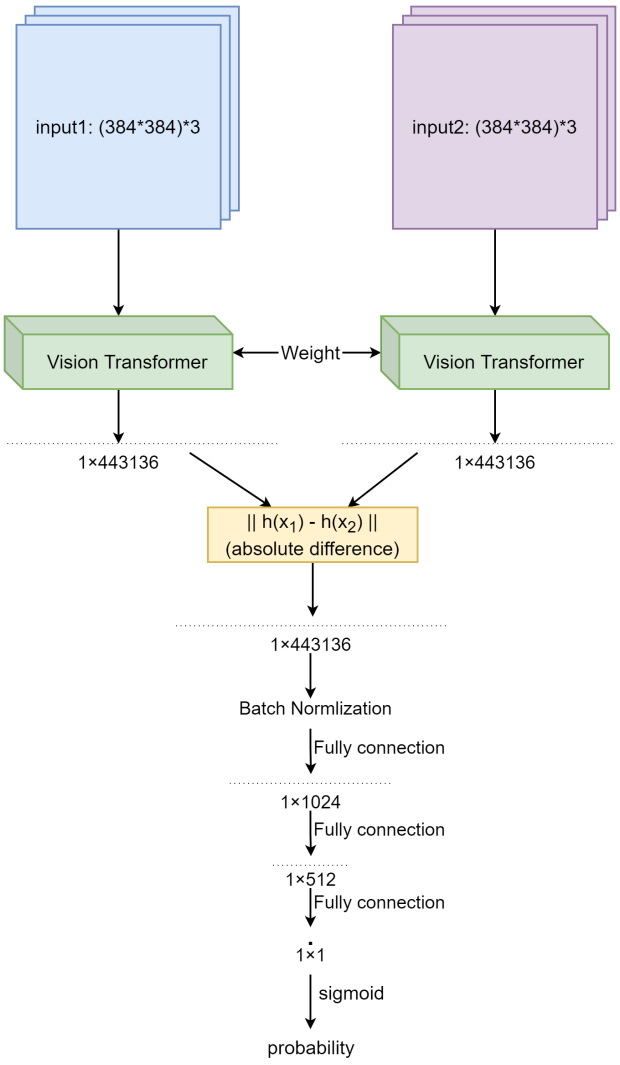


Figure 4. The architecture of Single Siamese ViT

4.2.2. Double Siamese ViT

For sub-task 3, I proposed a network architecture called Double Siamese ViT. This network architecture allows for four-branch inputs (obverse of coin 1, the obverse of coin 2, the reverse of

coin 1, and the reverse of coin 2), containing a total of two sub-networks. Until getting the absolute distance, both sub-networks have the same architecture as the Single Siamese ViT. Then, after a layer normalisation, the network gets the output via two fully connected layers. Sigmoid is used as the activation function for getting the probability. The architecture of this network can be seen in Figure 5.

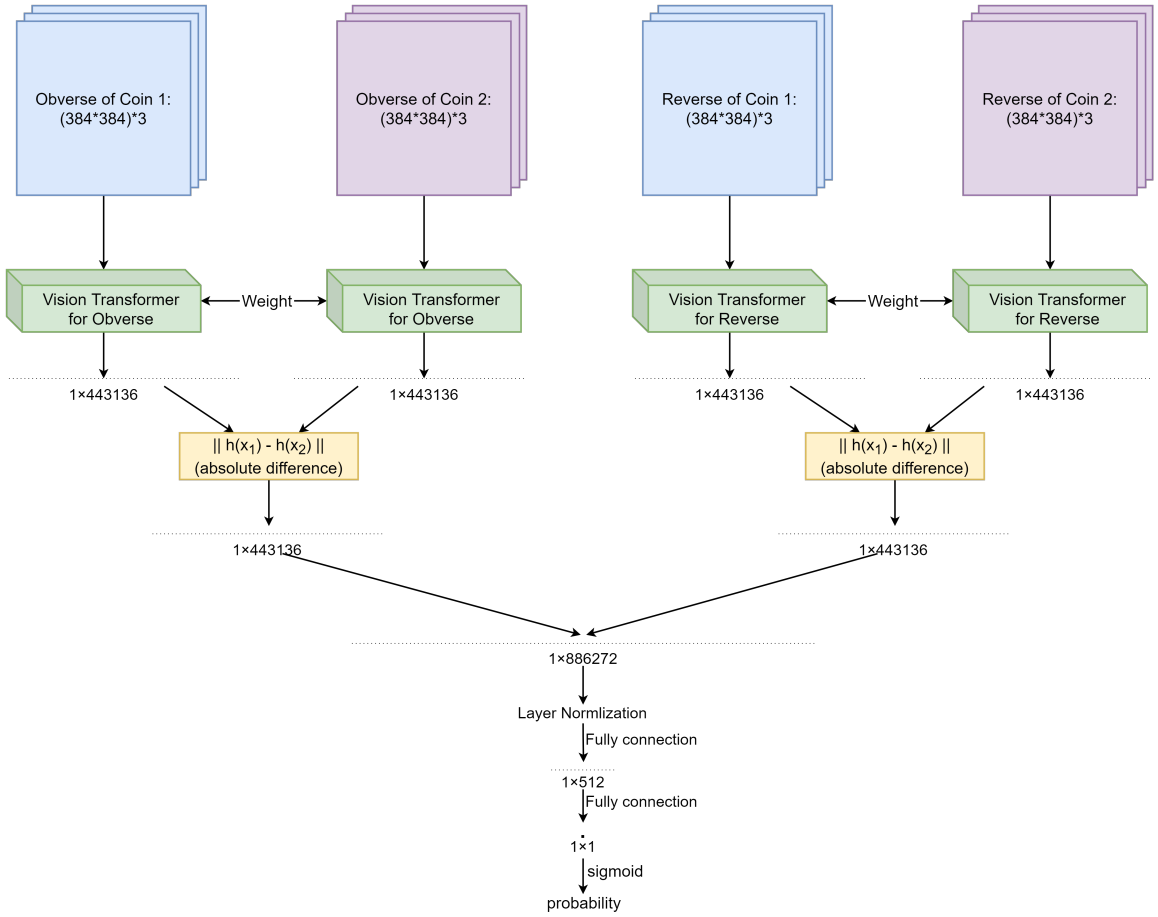


Figure 5. The architecture of Double Siamese ViT

5. Results and evaluation

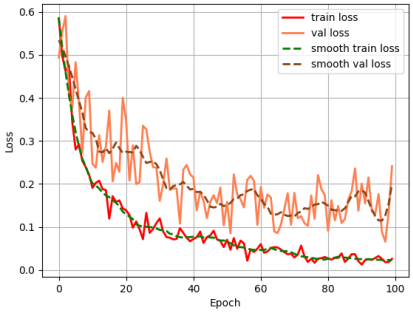
Using the proposed networks and data, we have the following results to report. The results will be presented according to the two networks.

5.1. Single Siamese ViT

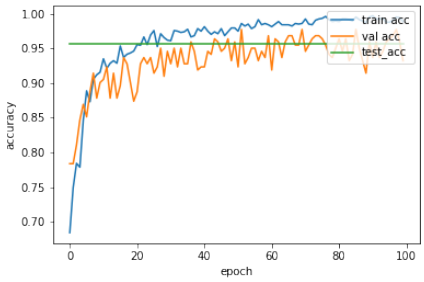
The first proposed network is Single Siamese ViT, which was trained on obverse and reverse of coins respectively. The final model performance will be illustrated according to the side of coins used for training.

5.1.1. Obverse of coins

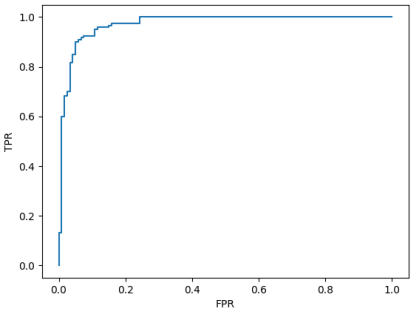
As the Figure 6 shows, During the training proceeds, the training loss continues to decrease. Although the oscillation of validation loss is relatively large, its overall trend keeps decreasing. We can assume that the model fits well based on the loss curve. The model achieved an accuracy of 95.73% on the test set and 74.32% on the unseen issues testset based on the same method of organising the data as the training set. The ROC curves of the model on the two test sets are also shown in the figure.



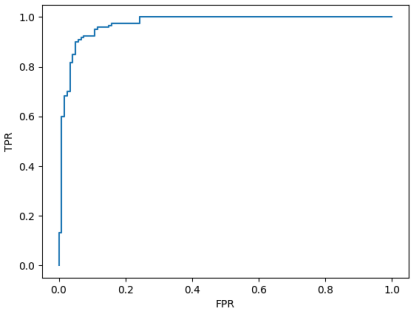
(**a**) Loss curve



(**b**) Accuracy curve



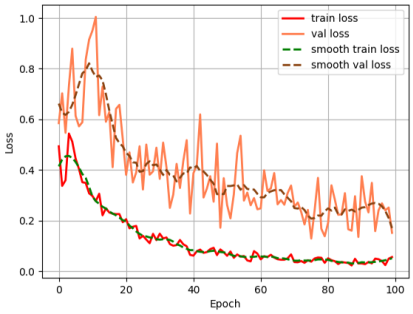
(**c**) ROC curve on testset



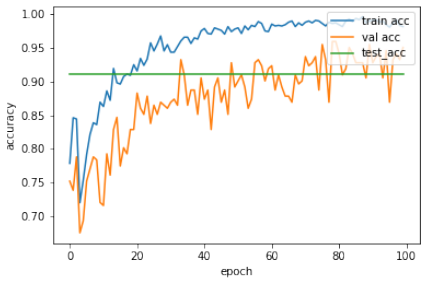
(**d**) ROC curve on unseen issues testset

Figure 6. Training results for Single Siamese ViT on the obverse of coins

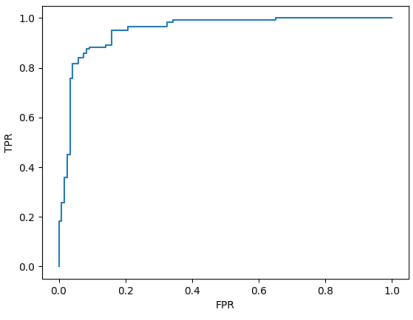
5.1.2. Reverse of coins



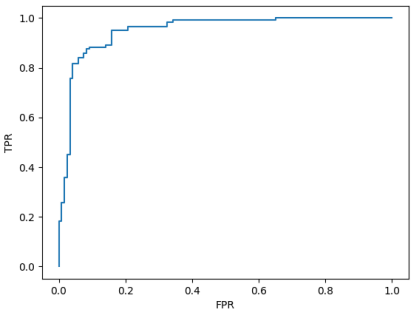
(**a**) Loss curve



(**b**) Accuracy curve



(**c**) ROC curve on testset



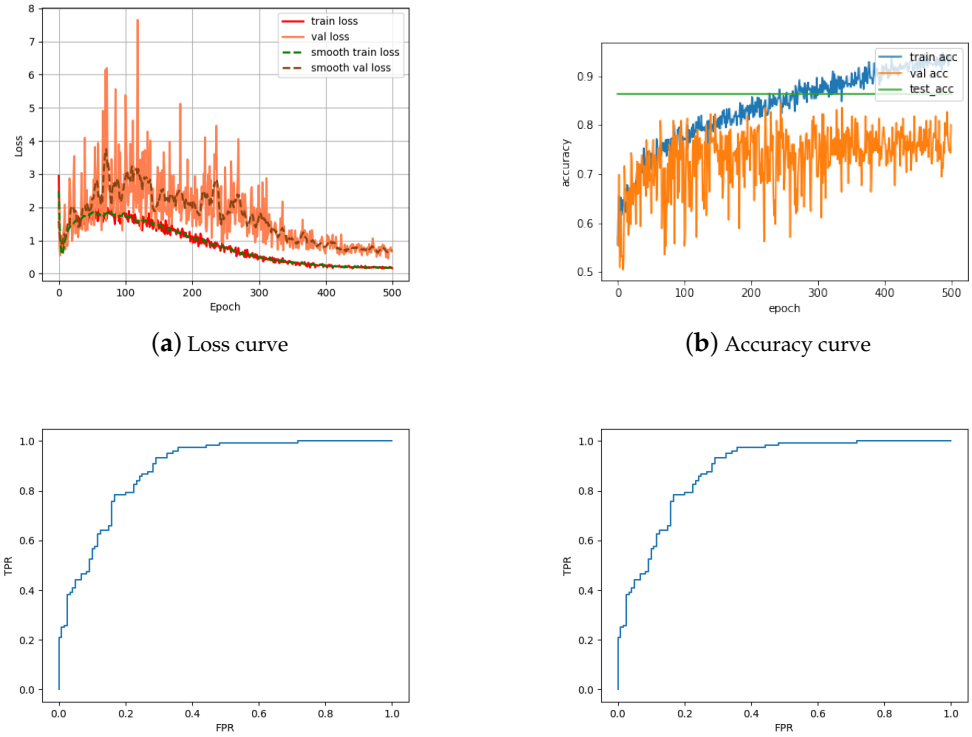
(**d**) ROC curve on unseen issues testset

Figure 7. Training results for Single Siamese ViT on the reverse of coins

The curves in Figure 7 for training on reverse of coins are quite similar as obverse's, the model achieved an accuracy of 91.03% on the test set and 75.57% on the unseen issues testset based on the same method of organising the data as the training set.

5.2. Double Siamese ViT

During the training process of the Double Siamese ViT, if the two sub Single Siamese ViT are not frozen, the model will be severely overfitting regardless of the hyperparameter settings, so we freeze the weights of the two sub-ViT networks in the subsequent training, and the final result is shown in Figure 8. The model achieved an accuracy of 86.36% on the test set and 61.18% on the unseen issues testset based on the same method of organising the data as the training set.



(**a**) Loss curve

(**b**) Accuracy curve

(**c**) ROC curve on testset

(**d**) ROC curve on unseen issues testset

Figure 8. Training results for Double Siamese ViT

For each emperor, I randomly take one image from every issue. Afterwards, all combinations of these images are taken. For each issue, I retake a new sample to get the combination with another one from this issue (those two images are not the same image), then to evaluate the performance of the obverse Single Siamese ViT model, reverse Single Siamese ViT model, and both obverse and reverse Double Siamese ViT model for each emperor.

For example, we assume now we are evaluating emperor A, which has 3 issues RIC1, RIC2, RIC3 respectively, each of issues has 3 images, img_a^b denotes the b_{th} image in "RIC a". So we have:

$emperorA = \{RIC1, RIC2, RIC3\}$, $RIC1 = \{img_1^1, img_1^2, img_1^3\}$, $RIC2 = \{img_2^1, img_2^2, img_2^3\}$, $RIC3 = \{img_3^1, img_3^2, img_3^3\}$.

$img_a^?$ denotes a random image taken from "RIC a", which is different from $img_a^{?'}$.

After taking combination, we have pairs of images to evaluate as:

$\{(img_1^?, img_1^{?'}), (img_1^?, img_2^?), (img_1^?, img_3^?), (img_2^?, img_2^{?'}), (img_2^?, img_3^?), (img_3^?, img_3^{?'})\}$

Specifically, if the "RIC a" just has 1 image, i.e., $img_a^? = img_a^{?'}$, we just drop this pair.

As an extreme example, $emperorB = \{RIC1, RIC2\}$, $RIC1 = \{img_1^1, img_1^2, img_1^3\}$, $RIC2 = \{img_2^1\}$, After taking combination, we just have 1 pair of images to evaluate: $\{(img_1^?, img_2^1)\}$.

For those emperors which has fewer than ten pairs of images, I will ignore them because the sample size is too small, and the statistical results were not representative.

We consider an accuracy rate greater than 85% as an excellent matching result, greater than 70% as a good matching result, and less than 70% as an undesirable matching result.

Finally, the performance of each of the three models in matching different issues among the 150 emperors. Figure 9 shows the visualisation results. By comparing the performance of the three models, we can find that Double Siamese ViT has a huge improvement in matching accuracy after combining the obverse and reverse Single Siamese ViT models.

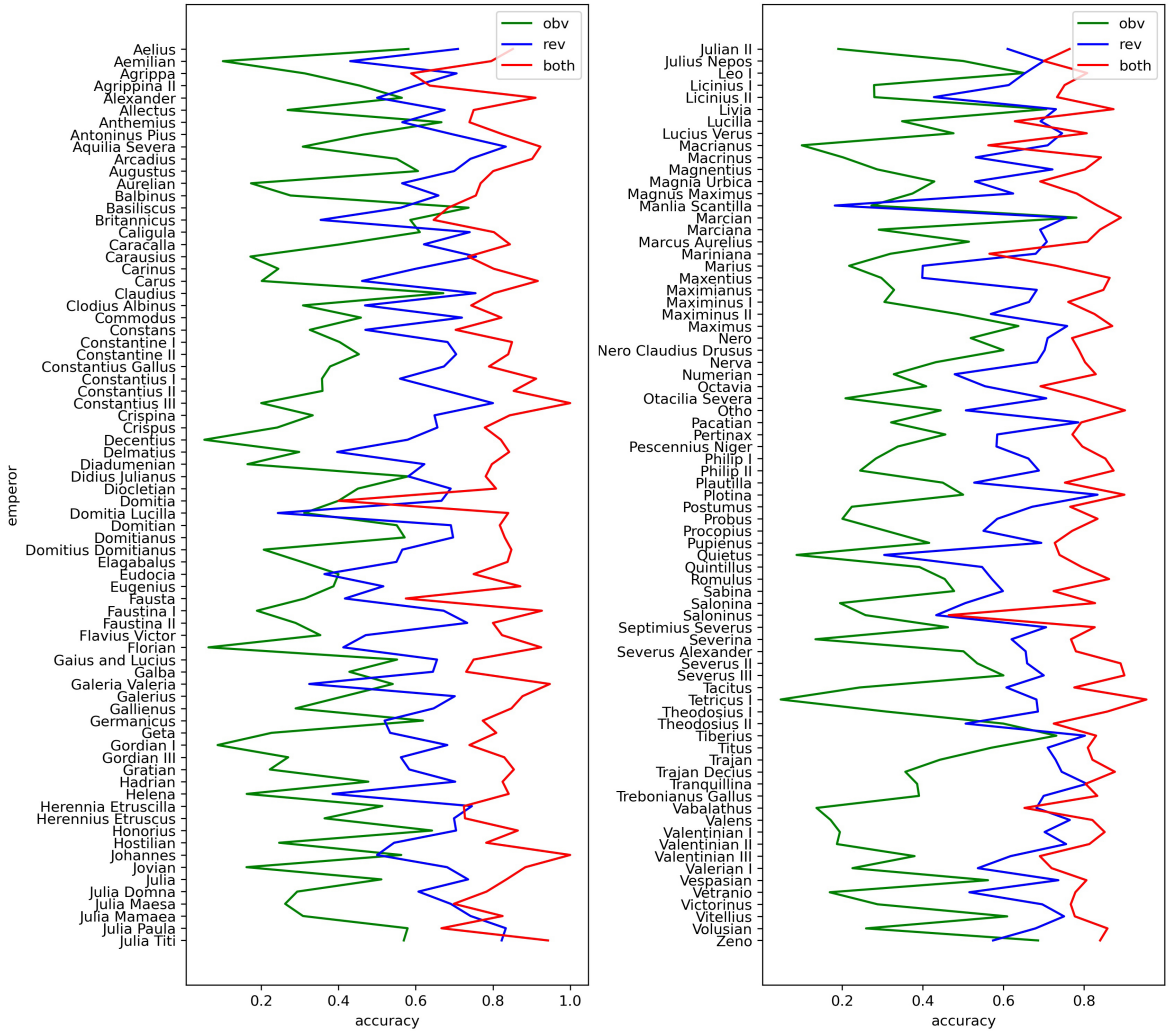


Figure 9. The evaluation result in each emperor

6. Conclusion

This project has attempted to solve the three proposed sub-tasks. To elaborate, those three tasks are the following: A. To design and train a matching model on the obverse of coins; B. To design and train a matching model on the reverse of coins; and finally, C. To design and train a matching model on both obverse and reverse of coins. One of the key objectives of this project is to have a meaningful contribution to the field of ancient coin research. Because the quantity and quality of datasets are not enough, researchers in this field is difficult to utilise classification models and deep learning architectures to group coins to their corresponding coin issues. In order to meet this objective, the project has done the following key processes. First, the project has illustrated the inefficiencies of classifying ancient coins with traditional image classification methods and explained the lack of

practical value. Instead, the project claimed that further efforts should be concentrated on ancient coin matching. Subsequently, the project has explained the new approach for ancient coin matching and demonstrated satisfactory results in combination with a deep learning framework. More specifically, the project proposed an original method that can be trained on small samples (566 images) and managed to achieve comparatively superior evaluation results based on Siamese Neural Network, Vision Transformer and Transfer Learning. The evaluation of a test set of more than 14,820 coin images demonstrated the effectiveness of the proposed method in matching each emperor's issues, making this experiment one of the successful current experiments that can balance computational cost and accuracy of results. In addition to the statistical analysis, the project has also analysed those emperors where the model performs poorly and discussed, and speculated some of the reasonable causes of such phenomenon. This project might give future researchers new perspectives and methods to design and conduct their research on ancient coin matching.

Author Contributions: All authors have contributed to all aspects of the described work.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Arandjelović, O.; Zachariou, M. Images of Roman imperial denarii: A curated data set for the evaluation of computer vision algorithms applied to ancient numismatics, and an overview of challenges in the field. *Sci* **2020**, *2*, 91.
2. Huber-Mörk, R.; Nölle, M.; Rubik, M.; Hödlmoser, M.; Kampel, M.; Zambanini, S. Automatic coin classification and identification. *Advances in Object Recognition Systems* **2012**, *127*.
3. Kiourt, C.; Evangelidis, V. AnCoins: Image-Based Automated Identification of Ancient Coins Through Transfer Learning Approaches. International Conference on Pattern Recognition. Springer, 2021, pp. 54–67.
4. Wei, K.; He, B.; Wang, F.; Zhang, T.; Ding, Q. A novel method for classification of ancient coins based on image textures. Second Workshop on Digital Media and its Application in Museum & Heritages (DMAMH 2007). IEEE, 2007, pp. 63–66.
5. Zaharieva, M.; Kampel, M.; Zambanini, S. Image based recognition of ancient coins. International Conference on Computer Analysis of Images and Patterns. Springer, 2007, pp. 547–554.
6. Huber-Mörk, R.; Zambanini, S.; Zaharieva, M.; Kampel, M. Identification of ancient coins based on fusion of shape and local features. *Machine vision and applications* **2011**, *22*, 983–994.
7. Kampel, M.; Zaharieva, M. Recognizing ancient coins based on local features. International Symposium on Visual Computing. Springer, 2008, pp. 11–22.
8. Kampel, M.; Huber-Mörk, R.; Zaharieva, M. Image-based retrieval and identification of ancient coins. *IEEE Intelligent Systems* **2009**, *24*, 26–34.
9. Zambanini, S.; Kampel, M. Robust Automatic Segmentation of Ancient Coins. VISAPP (1), 2009, pp. 273–276.
10. Anwar, H.; Zambanini, S.; Kampel, M. Supporting ancient coin classification by image-based reverse side symbol recognition. International Conference on Computer Analysis of Images and Patterns. Springer, 2013, pp. 17–25.
11. Anwar, H.; Zambanini, S.; Kampel, M. Encoding spatial arrangements of visual words for rotation-invariant image classification. German Conference on Pattern Recognition. Springer, 2014, pp. 443–452.
12. Aslan, S.; Vascon, S.; Pelillo, M. Two sides of the same coin: Improved ancient coin classification using Graph Transduction Games. *Pattern Recognition Letters* **2020**, *131*, 158–165.
13. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems* **1993**, *6*.
14. Chicco, D. Siamese neural networks: An overview. *Artificial Neural Networks* **2021**, pp. 73–94.
15. Berlemont, S.; Lefebvre, G.; Duffner, S.; Garcia, C. Siamese neural network based similarity metric for inertial gesture classification and rejection. 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE, 2015, Vol. 1, pp. 1–6.

16. Zhang, C.; Liu, W.; Ma, H.; Fu, H. Siamese neural network based gait recognition for human identification. 2016 ieee international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2016, pp. 2832–2836.
17. Long, T. Research on application of athlete gesture tracking algorithms based on deep learning. *Journal of Ambient Intelligence and Humanized Computing* **2020**, *11*, 3649–3657.
18. Ostertag, C.; Beurton-Aimar, M. Matching ostraca fragments using a siamese neural network. *Pattern Recognition Letters* **2020**, *131*, 336–340.
19. Kim, M.; Alletto, S.; Rigazio, L. Similarity mapping with enhanced siamese network for multi-object tracking. *arXiv preprint arXiv:1609.09156* **2016**.
20. Ichida, A.Y.; Meneguzzi, F.; Ruiz, D.D. Measuring semantic similarity between sentences using a siamese neural network. 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018, pp. 1–7.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
22. Lin, T.; Wang, Y.; Liu, X.; Qiu, X. A survey of transformers. *arXiv preprint arXiv:2106.04554* **2021**.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; others. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
24. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. International Conference on Machine Learning. PMLR, 2021, pp. 10347–10357.
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
26. Yuan, L.; Chen, Y.; Wang, T.; Yu, W.; Shi, Y.; Jiang, Z.H.; Tay, F.E.; Feng, J.; Yan, S. Tokens-to-token vit: Training vision transformers from scratch on imagenet. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.
27. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.
28. Wu, H.; Xiao, B.; Codella, N.; Liu, M.; Dai, X.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 22–31.
29. Ranftl, R.; Bochkovskiy, A.; Koltun, V. Vision transformers for dense prediction. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 12179–12188.
30. Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; Feichtenhofer, C. Multiscale vision transformers. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6824–6835.