https://doi.org/10.2352/EI.2022.34.8.IMAGE-273

Cultural Assets Identification using Transfer Learning

Simon Bugert, Huajian Liu, Waldemar Berchtold, Martin Steinebach; Fraunhofer SIT / ATHENE; Darmstadt, Hesse/Germany

Abstract

Identifying cultural assets is a challenging task which requires specific expertise. In this paper, a deep learning based solution to identify archaeological objects is proposed. Several additions to the ResNet CNN architecture are introduced which consolidate features from different intermediate layers by applying global pooling operations. Unlike general object recognition, identifying archaeological objects poses new challenges. To meet the special requirements in classifying antiques, a hybrid network architecture is used to learn the characteristics of objects using transfer learning, which includes a classification network and a regression network. With the help of the regression network, the age of objects can be predicted, which improves the overall performance in comparison to manually classifying the age of objects. The proposed scheme is evaluated using a public database of cultural assets and the experimental results demonstrate its significant performance in identifying antique objects.

Introduction

Trafficking of archaeological cultural assets includes the illicit import, export and trade of cultural property, which has been identified as a possible source of criminal financing [1, 2]. However, it is challenging to identify questionable archaeological objects in trade and at customs efficiently, from which region and epoch they probably originate, due to lack of necessary expertise. Investigators or officials often do not have a deep archaeological background and the required expert support may not always be available on site and in time. The goal of this work is to develop an AI- based solution using transfer learning, which is able to recognize archaeological objects from illicit excavations and provide immediate on-site assistance in the initial assessment of the potential origin of an object.

Objective

The goal is to identify looted archaeological objects. The identification is not confined to the type, but more importantly, to the geographical and temporal origin of the object. Unlike stolen objects, the looted objects from illicit excavations are usually unknown and have never previously been photographed and classified by experts. In the reference database, only similar objects may exist, but certainly not identical ones. Moreover, archaeological objects of the same type may possess completely different artistic characteristics due to their origins in different regions and eras. Conversely, different types of objects from the same region and epoch are likely to share similar artistic characteristics.

While image classification is a well-studied problem, the given task imposes special requirements. The identification of cultural assets should not be based only on the object type, but also on its distinctive regional and contemporary characteristics, such as the characters or symbols specific to a particular country or period. The objects to be examined should be linked to the reference objects bearing the same or similar characteristics in the database, even if they belong to different types of objects, with different shapes and appearances. In this way, the metadata of the reference objects can be used to infer the provenance of the unknown objects.

The proposed model is to be used in the context of an application the purpose of which is to assist investigators in the analysis of cultural property using smartphones.

Proposed approach

In this section, a new classification network architecture is introduced which can determine an object's class, representing the temporal and geographical origin. Subsequently, a second approach is introduced which improves the performance by using a hybrid architecture with a classification head to infer the regional class and a regression head to determine the object age.

Classification network

We propose several additions to the ResNet architecture [5] which are based on the notion that not only high-level features are needed to classify objects with the same attributes, but also the intermediate and low-level features from the network are beneficial. In a commonly used convolutional neural network, the classification head is only receiving the output of the last convolutional layer.

After testing features from several different layers for image matching, we concluded that high-level features, such as the ones passed to the classification head, are often not expressive enough for the given task. A possible explanation is, that not only the overall shapes, but also low-level features representing textures and patterns are important. To this end, we introduce several changes to the ResNet architecture, which are illustrated in figure 1.

First, skip connections are introduced after each ResNet block. The feature dimensions are decreasing from 120×120 to 15×15 while the filter depths are increasing from 256 to 2048 from block 1 to block 4. In order to apply a depth-wise concatenation, an average pooling layer is used for upsampling of all feature maps to a common resolution of 120×120.

Subsequently, both global maximum pooling and global average pooling [6] are applied to the feature maps in parallel and again concatenated to a single vector. In addition to group normalization [7] and activation layers, a fully-connected hidden layer is added. In comparison to the standard ResNet classification head, this increases the model capacity which is beneficial to the processing of the additional feature inputs. We call this network *MaxAvg-Cat*.

For data augmentation, MixUp [8] is used. At training time, the input images are randomly cropped and horizontally flipped. For the classification output y, a standard softmax layer is used. Additionally, a feature extraction point a is defined which can be

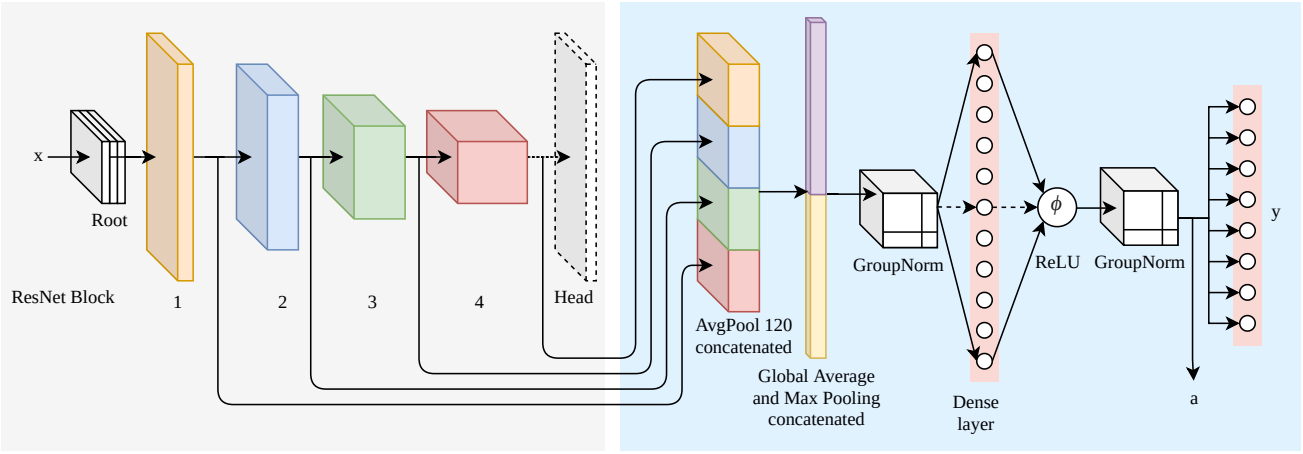

Figure 1. *MaxAvg-Cat architecture: extending ResNet by introducing skip connections and a new classification head. Global max and average pooling is applied and the result is concatenated before being forwarded to a fully-connected layer.*

used for image matching. The size of this feature vector is a hyperparameter and defined by the size of the previous hidden layer.

Hybrid network

Although the geographical origin and the age of reference objects are most of the times tagged in the database, it is not clear in which period of time the objects will bear the same features. It requires very specialized archaeological knowledge and a lot of work to label and classify these features, so this information is not available in most cases.

To identify unknown archaeological objects, we propose a hybrid network architecture to learn the characteristics of objects as shown in figure 2. The proposed architecture consists of three parts: a backbone which is used for feature extraction, a classification head and a regression head. A pre-trained ResNet model is used for feature extraction. The extracted features are then fed into the classification network to classify objects into categories, which are defined based on the available metadata of reference objects.

The regression head in the proposed architecture is used to predict the year of origin of objects as an integer, which achieves better results than manually classifying the age of objects and subsequently improves the overall performance. The classification head is used to predict the geographical origin of objects. We call this architecture *MOCaR* – multi-output classification and regression.

In order to train a neural network with multiple tasks or outputs using SGD, the individual losses have to be summarized to a single number. For the computation of a single loss value, we define three methods. Let $\mathscr{L}_c$ denote the classification loss and $\mathscr{L}_r$ the regression loss of the model.

Then, equation 1 defines the sum $\mathscr{L}_S$, equation 2 the product $\mathscr{L}_P$ and equation 3 the sum of these two methods $\mathscr{L}_{SP}$:

$$\mathscr{L}_S = \mathscr{L}_c + \mathscr{L}_r \tag{1}$$

$$\mathscr{L}_P = \mathscr{L}_c \times \mathscr{L}_r \tag{2}$$

$$\mathscr{L}_{SP} = \mathscr{L}_S + \mathscr{L}_P = \mathscr{L}_c + \mathscr{L}_r + \mathscr{L}_c * \mathscr{L}_r \tag{3}$$

A single data instance is defined by the object image x, a class label corresponding to the geographical origin y_c and an integer corresponding to the object's year of origin y_t. The log-cosh loss function is used for the regression while the classification task is optimized using the cross-entropy loss function.

Evaluation

For the evaluation, we crawled the online database of the Berlin State Museums (*SMB-digital*)[1] consisting of 15 museums and four institutions. The database comprises more than 250k archaeological objects, each with an image and metadata. From this, we were able to choose about 140k suitable objects from 8 different collections. Figure 3 shows a randomly drawn set of 20 objects from the database.

Because the geographical origin is not always labeled in a usable way, we used the museum collection attribute as a substitute for the geographical label. Then, two datasets are created for the introduced approaches *MaxAvg-Cat* and *MOCaR*. For the latter, we used the 8 collections for the classification while the object time will be used directly during the training. For the former, the class labels are formed through concatenation of the collection and a time range representing an epoch.

For binning the objects into epochs, we applied the *Jenks natural breaks optimization* [10]. This method is designed to reduce the variance within classes and maximize the variance between classes. This way, 50 distinct classes are formed while the object collection and year are concatenated to a single class label in the following form:

```
collection, (object time begin, object time end]
```
(4)

The proposed approaches utilize the pretrained ResNet [5] network called BigTransfer [3] for transfer learning. This base model is pretrained on ImageNet21k [4] comprising 14M images from 21k classes. The model head is dropped and our additions are trained from scratch while the convolutional layers are fine-tuned. The proposed training scheme from BigTransfer is adopted

[1] SMB-digital online collections database of the Berlin State Museums `http://www.smb-digital.de/eMuseumPlus?lang=en` (visited on 31/01/2022)

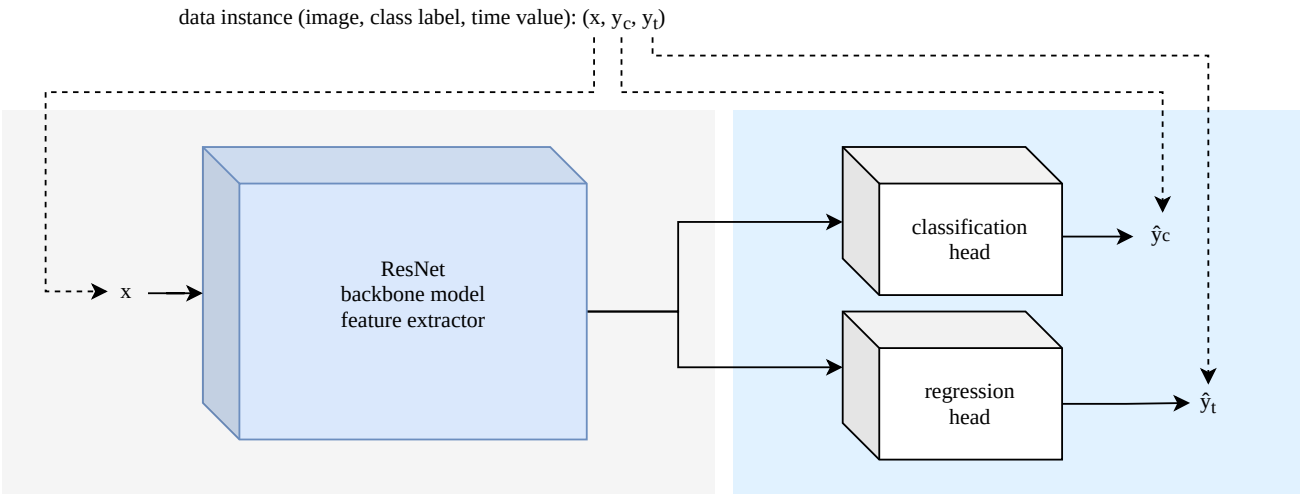

Figure 2. *Multi-output classification and regression (MOCaR): splitting the inference of the temporal and the geographical attributes using two separate model heads.*

as is. After a short warm-up phase, the learning rate is first linearly increased to the target value and then divided by 10 at 30%, 60% and 90% of the training with 10k batches.

In addition to the described schedule, the base learning rate is optimized using a small run during which it is increased after each batch in order to determine the optimal learning rate empirically [9]. During training, after each epoch, a validation run is performed and a checkpoint is saved while the loss is decreasing. The fully-connected layer is parameterized with a size of 1024 and 3840 and dropout is added after this layer for one experiment.

For the training we used a 80–10–10 split into train, validation and test sets. The validation set was used for hyperparameter tuning and the development of the different methods while the test set was only used for a final evaluation.

As baselines, the pretrained ResNet 50x1 and ResNet 152x2 models are fine-tuned with the standard classification head. As shown in table 1, the top-1 accuracy on the test set increased to 0.8224 for *MaxAvg-Cat* with a hidden layer size of 3840 from the baselines of 0.7791 and 0.8079 for R50x1 and R152x2, respectively.

For *MOCaR*, table 2 lists the experimental results. The classification performs best when using the multiplication variant (see equation 2) and the regression performs best for the summed losses (see equation 1). The sum of the two variants (see equation 3) is a compromise as both the classification and regression accuracy is between the accuracy of the individual variants. While the best mean absolute error is about 132 years, the compromise model achieves an error of 180 years. The accuracy decreases likewise from 0.9240 to 0.8817.

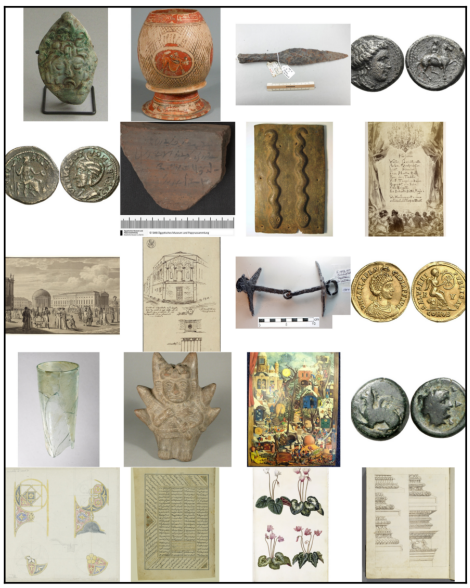

Figure 3. *Example training data from the SMB-digital online database.*

Table 1: MaxAvg-Cat test results with top-1 accuracy, top-5 accuracy and F-score

test val	top1	top5	f1
R50x1	0.7791 0.7824	0.9860 0.9859	0.7176 0.7125
R152x2	0.8079 0.8076	**0.9885** 0.9894	0.7472 0.7393
MaxAvg-Cat dropout	0.7980 0.7925	0.9867 0.9861	0.7241 0.7268
MaxAvg-Cat 1024	0.8032 0.8094	0.9847 0.9874	0.7444 0.7458
MaxAvg-Cat 3840	**0.8224** 0.8267	0.9873 0.9880	**0.7691** 0.7656

Table 2: MOCaR test results

test val		add	mult	mult+add
classification	accuracy	0.7417	**0.9240**	0.8817
		0.7352	0.9172	0.8817
	loss	0.6693	**0.2216**	0.3189
		0.6840	0.2363	0.3323
regression	mae	**132.14**	1063.53	180.29
		141.41	1064.02	167.51
	loss	**131.45**	1062.84	166.82
		140.73	1063.33	179.60
comb	loss	**132.1268**	341.6426	211.54
		141.414	368.617	231.00

Conclusion and future work

The proposed network is trained and tested using the SMB-digital online database, which includes approximately 140k usable tagged objects out of different collections. The data set is split in the ratio 80–10–10 for training, validation and testing respectively. The classification approach is a solution which can classify cultural assets with an accuracy of about 82% and can be of great help for the initial assessment of cultural objects.

One disadvantage with this approach is that it is not always clear from the data in which period the objects will result in similar features. A second solution to identify cultural assets is proposed to this end. Different from general image or object recognition, identifying and classifying archaeological objects poses special requirements and challenges. Therefore, a hybrid network architecture is introduced to learn the features of archaeological objects, in order to classify the archaeological objects according to their unique regional and contemporary characteristics. The regression network in the proposed architecture enables automatic prediction of the age of objects, which improves the accuracy of identification.

Different strategies are applied to combine the losses of the classification and regression networks in the training process. The accuracy rate ranges from 0.742 to 0.924 for the top 1 of best matches. For the top 5 of best matches, the accuracy rate increases to a range between 0.986 and 0.998. The average prediction error of the regression network lies between 132.13 and 180.29 years.

In future work, we aim to improve the balance between the classification and regression optimization during training. Also, the model performance would benefit from incorporating images from multiple perspectives if the given object is three-dimensional.

Acknowledgments

The Federal Government Commissioner for Culture and the Media is funding the project with up to 500,000 euros from funds of the Federal Government's national AI strategy.

References

[1] L. A. Amineddoleh, Cultural heritage vandalism and looting: the role of terrorist organizations, public institutions and private collectors. Santander Art and Culture Law Review, 1(2), 27-62, 2015.

[2] N. Brodie, I. Sabrine, The illegal excavation and trade of Syrian cultural objects: a view from the ground. Journal of Field Archaeology, 43(1), 74-84, 2018.

[3] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Houlsby, Big Transfer (BiT): General Visual Representation Learning, arxiv, 1912.11370, 2020.

[4] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, ImageNet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248-255.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep Residual Learning for Image Recognition, arxiv, 1512.03385, 2015.

[6] Min Lin, Qiang Chen, Shuicheng Yan, Network In Network, arxiv, 1312.4400, 2014.

[7] Yuxin Wu, Kaiming He, Group Normalization, arxiv, 1803.08494, 2018.

[8] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz, mixup: Beyond Empirical Risk Minimization, arxiv, 1710.09412, 2018.

[9] Leslie N. Smith in Cyclical Learning Rates for Training Neural Networks, arxiv, 1506.01186, 2017.

[10] George F. Jenks and Fred C. Caspall. "Error on Choroplethic Maps: Definition, Measurement, Reduction". Annals of the Association of American Geographers, pp. 217–244, 1971.