Deep Transfer Learning for Art Classification Problems

Matthia Sabatelli[1], Mike Kestemont[2], Walter Daelemans[3], and Pierre Geurts[1]

[1] Montefiore Institute, Department of Electrical Engineering and Computer Science, Université de Liège, Belgium {m.sabatelli, p.geurts}@uliege.be
[2] Antwerp Center for Digital Humanities and Literary Criticism (ACDC), Universiteit Antwerpen, Belgium
[3] CLiPS, Computational Linguistics Group, Universiteit Antwerpen, Belgium
{mike.kestemont, walter.daelemans}@uantwerpen.be

Abstract. In this paper we investigate whether Deep Convolutional Neural Networks (DCNNs), which have obtained state of the art results on the ImageNet challenge, are able to perform equally well on three different art classification problems. In particular, we assess whether it is beneficial to fine tune the networks instead of just using them as off the shelf feature extractors for a separately trained softmax classifier. Our experiments show how the first approach yields significantly better results and allows the DCNNs to develop new selective attention mechanisms over the images, which provide powerful insights about which pixel regions allow the networks successfully tackle the proposed classification challenges. Furthermore, we also show how DCNNs, which have been fine tuned on a large artistic collection, outperform the same architectures which are pre-trained on the ImageNet dataset only, when it comes to the classification of heritage objects from a different dataset.

Keywords: Deep Convolutional Neural Networks, Art Classification, Transfer Learning, Visual Attention

1 Introduction and Related Work

Over the past decade Deep Convolutional Neural Networks (DCNNs) have become one of the most used and successful algorithms in Computer Vision (CV) [10] [18] [30]. Due to their ability to automatically learn representative features by incrementally down sampling the input via a set of non linear transformations, these kind of Artificial Neural Networks (ANNs) have rapidly established themselves as the state of the art algorithm on a large set of CV problems. Within different CV testbeds large attention has been paid to the ImageNet challenge [9], a CV benchmark that aims to test the performances of different image classifiers on a dataset that contains one million natural images distributed over thousand different classes. The availability of such a large dataset, combined with the possibility of training ANNs in parallel over several GPUs [17], has lead to the development of a large set of different neural architectures that have continued to outperform each other over the years [25] [27] [7] [13] [14].

A promising research field in which the classification performances of such DCNNs can be exploited is that of *Digital Heritage* [22]. Due to a growing and rapid process of digitization, museums have started to digitize large parts of their cultural heritage

collections, leading to the creation of several digital open datasets [3] [20]. The images constituting these datasets are mostly matched with descriptive metadata which, as presented in e.g. [20], can be used to define a set of challenging machine learning tasks. However, the number of samples in these datasets is far smaller than those in, for instance, the ImageNet challenge and this can become a serious constraint when trying to successfully train DCNNs from scratch.

The lack of available training data is a well known issue in the Deep Learning community and is one of the main reasons that has led to the development of the research field of Transfer Learning (TL). The main idea of TL consists of training a machine learning algorithm on a new task (e.g. a classification problem) while exploiting knowledge that the algorithm has already learned on a previously related task (a different classification problem). This machine learning paradigm has proved to be extremely successful in Deep Learning, where it has been shown how DCNNs that were trained on many large datasets [15] [26], were able to achieve very promising results on classification problems from heterogeneous domains, ranging from medical imaging [28] or gender recognition [32] over plant classification [24] to galaxy detection [2].

In this work we explore whether the TL paradigm can be successfully applied to three different art classification problems. We use four neural architectures that have obtained strong results on the ImageNet challenge in recent years and we investigate their performances when it comes to attributing the *authorship* to different artworks, recognizing the *material* which has been used by the artists in their creations, and identifying the *artistic category* these artworks fall into. We do so by comparing two possible approaches that can be used to tackle the different classification tasks. The first one, known as off the shelf classification [23], simply retrieves the features that were learned by the DCNNs on other datasets and uses them as input for a new classifier. In this scenario the weights of the DCNN do not change during the training phase, and the final, top-layer classifier is the only component of the architecture which is actually trained. This changes in our second explored approach, known as fine tuning, where the weights of the original DCNNs are "unfrozen" and the neural architectures are trained together with the final classifier.

Recent work [16] has shown the benefits that this particular pre-training approach has. In particular, DCNNs which have been trained on the ImageNet challenge typically lead to superior results when compared to the same architectures trained from scratch. However, this is not necessarily beneficial and in some cases DCNNs that are randomly initialized are able to achieve the same performances as ImageNet pre-trained models. However, none of the results presented in [16] have been applied to datasets containing heritage objects, it is thus still an open question how such pre-trained DCNNs would perform in such a classification scenario. Below, we extensively study the performance of these DCNNs; at the same time we assess whether better TL performances can be obtained when using DCNNs that, in addition to the ImageNet dataset, have additionally been pre-trained on a large artistic collection.

Contributions and Outline: This work contributes to the field of (Deep) TL applied to art classification problems. It does so by investigating if DCNNs, which have been originally trained on problems that are very dissimilar and far from art classification, can still perform well in such a different domain. Moreover, assuming this is the

case, we explore if it is possible to improve on such performances. The paper is structured as follows: in Section 2 we present a theoretical introduction to the field of TL, a description of the datasets that we have used and the methodological details about the experiments that we have performed. In Section 3 we present and discuss our results. A summary of the main contributions of this work together with some ideas for possible future research is finally presented in Section 4.

2 Methods

We now present the methods that underpin our research. We start by giving a brief formal definition of TL. We then introduce the three classification tasks under scrutiny, together with a brief description of the datasets. Finally, we present the neural architectures that we have used for our experiments.

2.1 Transfer Learning

A supervised learning (SL) problem can be identified by three elements: an input space $\mathcal{X}_t$, an output space $\mathcal{Y}_t$, and a probability distribution $p_t(x,y)$ defined over $\mathcal{X}_t \times \mathcal{Y}_t$ (where t stands for 'target', as this is the main problem we would like to solve). The goal of SL is then to build a function $f : \mathcal{X}_t \to \mathcal{Y}_t$ that minimizes the expectation over $p_t(x,y)$ of a given loss function ℓ assessing the predictions made by f:

$$E_{(x,y)\sim p_t(x,y)}\{\ell(y, f(x))\}, \tag{1}$$

when the only information available to build this function is a learning sample of input-output pairs $LS_t = \{(x_i, y_i) | i = 1, \ldots, N_t\}$ drawn independently from $p_t(x,y)$. In the general transfer learning setting, one assumes that an additional dataset LS_s, called the source data, is available that corresponds to a different, but related, SL problem. More formally, the source SL problem is assumed to be defined through a triplet $(\mathcal{X}_s, \mathcal{Y}_s, p_s(x,y))$, where at least either $\mathcal{X}_s \neq \mathcal{X}_t$, $\mathcal{Y}_s \neq \mathcal{Y}_t$, or $p_s \neq p_t$. The goal of TL is then to exploit the source data LS_s together with the target data LS_t to potentially find a better model f in terms of the expected loss (1) than when only LS_t is used for training this model. Transfer learning is especially useful when there is a lot of source data, whereas target data is more scarce.

Depending on the availability of labels in the target and source data and on how the source and target problems differ, one can distinguish different TL settings [21]. In what follows, we assume that labels are available in both the source and target data and that the input spaces $\mathcal{X}_t$ and $\mathcal{X}_s$, that both correspond to color images, match. Output spaces and joint distributions will however differ between the source and target problems, as they will typically correspond to different classification problems (ImageNet object recognition versus art classification tasks). Our problem is thus an instance of *inductive transfer learning* [21]. While several inductive transfer learning algorithms exist, we focus here on model transfer techniques, where information between the source and target problems is exchanged in the form of a DCNN model pre-trained on the source data. Although potentially suboptimal, this approach has the advantage of being more computationally efficient, as it does not require to train a model using both the source and the target data.

2.2 Datasets and Classification Challenges

For our experiments we use two datasets which come from two different heritage collections. The first one contains the largest number of samples and comes from the Rijksmuseum in Amsterdam[4]. On the other hand, our second 'Antwerp' dataset is much smaller. This dataset presents a random sample that is available as open data from a larger heritage repository: DAMS (Digital Asset Management System)[5]. This repository can be searched manually via the web-interface or queried via a Linked Open Data API. It aggregates the digital collections of the foremost GLAM institutions (Galleries, Libraries, Archives, Museums) in the city of Antwerp in Belgium. Thus, this dataset presents a varied and representative sample of the sort of heritage data that is nowadays being collected at the level of individual cities across the globe. While it is much smaller, its coverage of cultural production is similar to that of the Rijksmuseum dataset and presents an ideal testing ground for the transfer learning task under scrutiny here.

Both image datasets come with metadata encoded in the Dublin Core metadata standard [31]. We selected three well-understood classification challenges: (1) "material classification" which consists in identifying the material the different heritage objects are made of (e.g paper, gold, porcelain, ...) ; (2) "type classification" in which the DCNNs have to classify in which artistic category the samples fall into (e.g. print, sculpture, drawing, ...), and finally (3) "artist classification", where the main goal is to appropriately match each sample of the dataset with its creator (from now on we refer to these classification tasks as challenge 1, 2 and 3 respectively). As reported in Table 1 we can see how the Rijksmuseum collection is the dataset with the largest amount of samples per challenge (N_t) and the highest amount of labels to classify (Q_t). Furthermore it is also worth noting that there was no metadata available when it comes to the first classification challenge for the Antwerp dataset (as marked by the $\times$ symbol), and how there are some common labels between the two heritage collections when it comes to challenge 2. A visualization reporting some of the images present in both datasets can be seen in Figure 1.

We use 80% of the datasets for training while the remaining 2 x 10% is used for validation and testing respectively. Furthermore, we ensure that only classes which occur at least once in all the splits are used for our experiments. Naturally, in order to keep all comparisons fair between neural architectures and different TL approaches, all experiments have been performed on the exact same data splits which, together with the code used for all our experiments, are publicly released to the CV community [6].

2.3 Neural Architectures and Classification Approaches

For our experiments we use four pre-trained DCNNs that have all obtained state of the art results on the ImageNet classification challenge. The neural architectures are VGG19 [25], Inception-V3 [27], Xception [7] and ResNet50 [34]. We use the implementations of the networks that are provided in the `Keras` Deep Learning library [8]

[4] https://staff.fnwi.uva.nl/t.e.j.mensink/uva12/rijks/
[5] https://dams.antwerpen.be/
[6] https://github.com/paintception/Deep-Transfer-Learning-for-Art-Classification-Problems

Table 1: An overview of the two datasets that are used in our experiments. Each color of the table corresponds to a different classification challenge, starting from challenge 1 which is represented in yellow, challenge 2 in blue and finally challenge 3 in red. Furthermore we represent with N_t the amount of samples constituting the datasets and with Q_t the number of labels. Lastly, we also report if there are common labels between the two heritage collections.

Challenge	Dataset	N_t	Q_t	% of overlap
Material	Rijksmuseum	110,668	206	None
	Antwerp	×	×	
Type	Rijksmuseum	112,012	1,054	
	Antwerp	23,797	920	≈ 15%
Artist	Rijksmuseum	82,018	1,196	None
	Antwerp	18,656	903	

together with their appropriate `Tensorflow` weights [1] that come from the `Keras` official repository as well. Since all architectures have been built in order to deal with the ImageNet dataset we replace the final classification layer of each network with a new one. This final layer simply consists of a new *softmax* output, with as many neurons as there are classes, which follows a 2D global average pooling operation. We rely on this dimensionality reduction step because we do not add any fully connected layers between the last convolution block and the *softmax* output. Hence, in this way we are able to obtain a feature vector, $\mathcal{X}$, out of the rectified activation feature maps of the network that can be properly classified. Since all experiments are treated as a multi-class classification problem we use the *categorical crossentropy* function as the loss function of the DCNNs.

We investigate two possible classification approaches that are based on the previously mentioned pre-trained architectures. The first one, denoted as off the shelf classification, only trains a final *softmax* classifier on $\mathcal{X}$, which is retrieved from the different DCNNs after performing one forward pass of the image through the network [7]. This approach is intended to explore whether the features that are learned by the DCNNs on the ImageNet challenge are informative enough in order to properly train a machine learning classifier on the previously introduced art classification challenges. If this would be the case, such pre-trained models could be used as appropriate feature extractors without having to rely on expensive `GPU` computations for training. Naturally, they would only require the training of the final classifier without having to compute any backpropagation operations over the entire network.

[7] Please note how instead of a *softmax* layer any kind of machine learning classifier can be used instead. We experimented with both Support Vector Machines (SVMs) and Random Forests but since the results did not significantly differ between classifiers we decided to not include them here.

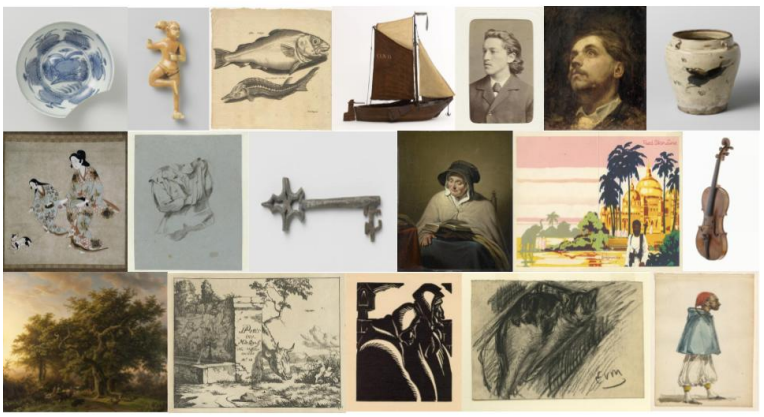

Fig. 1: A visualization of the images that are used for our experiments. It is possible to see how the samples range from images representing plates made of porcelain to violins, and from Japanese artworks to a more simple picture of a key.

Our second approach is generally known as fine tuning and differs from the previous one by the fact that together with the final *softmax* output the entire DCNN is trained as well. This means that unlike the off the shelf approach, the entirety of the neural architecture gets "unfrozen" and is optimized during training. The potential benefit of this approach lies in the fact that the DCNNs are independently trained on samples coming from the artistic datasets, and thus their classification predictions are not restricted by what they have previously learned on the ImageNet dataset only. Evidently, such an approach is computationally more demanding.

In order to maximize the performances of the DCNNs we take the work presented in [19] into consideration and train them with a relatively small batch size of 32 samples. We do not perform any data augmentation operations besides a standard pixel normalization to the $[0,1]$ range and a re-scaling operation which resizes the images to the input size that is required by the different DCNNs. Regarding the stochastic optimization procedures of the different classifiers, we use two different optimizers, that after preliminary experiments, turned out to be the best performing ones. For the off the shelf approach we use the RMSprop optimizer [29] which has been initialized with its default hyperparameters (learning rate = 0.001, a *momentum* value $\rho = 0.9$ and $\varepsilon = 1e-08$). On the other hand, when we fine tune the DCNNs we use the standard (and less greedy) Stochastic Gradient Descent (SGD) algorithm with the same learning rate, 0.001, and a *Nesterov Momentum* value set to 0.9. Training has been controlled by the *Early Stopping* method [6] which interrupted training as soon as the validation loss did not decrease for 7 epochs in a row. The model which is then used on the testing set is the one which obtained the smallest validation loss while training.

To the best of our knowledge, so far no work has been done in systematically assessing to which extent DCNNs pre-trained on the ImageNet dataset could also be used as valuable architectures when tackling art classification problems. Furthermore, it is also not known whether the fine tuning approach would yield better results when compared to the off the shelf one and if using such pre-trained ANNs would yield better performances than training the same architectures from scratch as observed by [16]. In the coming section we present new results that aim to answer these research questions.

3 Results

Our experimental results are divided in two different sections, depending on which kind of dataset has been used. We first report the results that we have obtained when using architectures that were pre-trained on the ImageNet dataset only, and aimed to tackle the three classification problems of the Rijksmuseum dataset that were presented in Section 2.2. We report these results in Section 3.1 in which we explore the benefits of using the ImageNet dataset as the TL source data, and how well such pre-trained DCNNs generalize when it comes to artistic images. We then present the results from classifying the Antwerp dataset, using DCNNs that are both pre-trained on the ImageNet dataset and on the Rijksmuseum collection in Section 3.3. We investigate whether these neural architectures, which have already been trained to tackle art classification problems before, perform better than the ones which have been trained on the ImageNet dataset only.

All results show comparisons between the off the shelf classification approach and the fine tuning scenario. In addition to that, in order to establish the potential benefits that TL from ImageNet has over training a DCNN from scratch, we also report the results that have been obtained when training one DCNN with weights that have been initially sampled from a "He-Uniform" distribution [12]. Since we take advantage of work [4] we use the Inception-V3 architecture. We refer to it in all figures as Scratch-V3 and visualize it with a solid orange line. Figures 2 and 3 report the performances in terms of accuracies that the DCNNs have obtained on the validation sets. While the performances that the neural architectures have obtained on the final testing set are reported in Tables 2 and 3.

3.1 From Natural to Art Images

The first results that we report have been obtained on the "material" classification challenge. We believe that this can be considered as the easiest classification task within the ones that we have introduced in Section 2.2 for two main reasons. First, the number of possible classes the ANNs have to deal with is more than five times smaller when compared to the other two challenges. Furthermore, we also believe that this classification task is, within the limits, the most similar one when compared to the original ImageNet challenge. Hence, the features that might be useful in order to classify the different natural images on the latter classification testbed might be not too dissimilar from the ones that are needed to properly recognize the material that the different samples of the Rijksmuseum collection are made of. If this would be the case we would expect

very close performances between the off the shelf classification approach and the fine tuning one. Comparing the learning curves of the two classification strategies in Figure 2, we actually observe that the fine tuning approach leads to significant improvements when compared to the off the shelf one, for three architectures out of the four tested ones. Note however that, in support of our hypothesis, the off the shelf approach can still reach high accuracy values on this problem and is also competitive with the DCNN trained from scratch. This suggests that features extracted from networks pretrained on ImageNet are relevant for material classification.

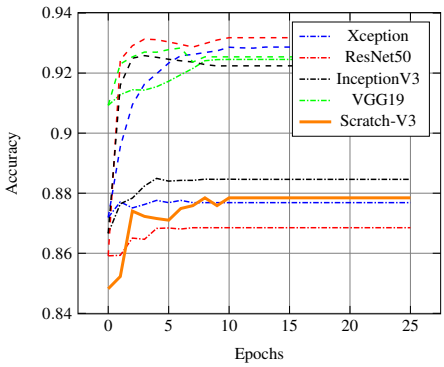

Fig. 2: Comparison between the fine tuning approach versus the off the shelf one when classifying the material of the heritage objects of the Rijksmuseum dataset. We observe how the first approach (as reported by the the dashed lines) leads to significant improvements when compared to the latter one (reported by the dash-dotted lines) for three out of four neural architectures. Furthermore, we can also observe how training a DCNN from scratch leads to worse results when compared to fine-tuned architectures which have been pre-trained on ImageNet (solid orange line).

The ResNet50 architecture is the DCNN which, when fine tuned, performs overall best when compared to the other three ANNs. This happens despite it being the DCNN that initially performed worse as a simple feature extractor in the off the shelf experiments. As reported in Table 2 we can see how this kind of behavior reflects itself on the separated testing set as well, where it obtained the highest testing set accuracy when fine tuned (92.95%), and the lowest one when the off the shelf approach was used (86.81%). It is worth noting how the performance between the different neural architectures do not strongly differ between each other once they are fine tuned, with all DCNNs performing around $\approx 92\%$ on the final testing set. Furthermore, special attention needs to be given to the VGG19 architecture, which does not seem to benefit from the fine tuning approach as much as the other architectures do. In fact, its off the shelf performance on the testing set (92.12%) is very similar to its fine tuned one (92.23%). This suggests that this neural architecture is actually the only one which, in this task, and when pre-trained on ImageNet, can successfully be used as a simple feature extractor without having to rely on complete retraining.

When analyzing the performances of the different neural architectures on the "type" and "artist" classification challenges (respectively the left and right plots reported in Figure 3), we observe how the fine tuning strategy leads to even more significant improvements when compared to what has been observed in the previous experiment. The results obtained on the second challenge show again how the ResNet50 architecture is the DCNN which leads to the worse results if the off the shelf approach is used (its testing set accuracy is as low as 71.23%) and similarly to what has been observed before, it then becomes the best performing ANN when fine tuned, with a final accuracy of 91.30%. Differently from what has been observed in the previous experiment, the VGG19 architecture, despite being the ANN performing best when used as off the shelf feature extractor, this time performs significantly worse than when it is fine tuned, which highlights the benefits of this latter approach. Similarly to what has been observed before, our results are again not significantly in favor of any neural architecture once they are fine tuned, with all final accuracies being around $\approx 91\%$.

If the classification challenges that we have analyzed so far have highlighted the significant benefits of the fine tuning approach over the off the shelf one, it is also important to note that the latter approach is still able to lead to satisfying results. In fact, accuracies of 92.12% have been obtained when using the VGG19 architecture on the first challenge and a classification rate of 77.33% was reached by the same architecture on the second challenge. Despite the latter accuracy being very far in terms of performance from the one obtained when fine tuning the network (90.27%), it still shows how DCNNs pre-trained on ImageNet do learn particular features that can also be used for classifying the "material" and the "type" of heritage objects. However, when analyzing the results from the "artist" challenge, we can see that this is partially not the case anymore.

For the third classification challenge, the Xception, ResNet50, and Inception-V3 architectures all perform extremely poorly if not fine tuned, with the latter two DCNNs not being able to even reach a 10% classification rate. Better results are obtained when using the VGG19 architecture, which reaches a final accuracy of 38.11%. Most importantly, all performances are again significantly improved when the networks are fine tuned. As already observed in the previous experiments, ResNet50 outperforms the others on the validation set. However, on the test set (see Table 2), the overall best performing network is Inception-V3 (with a final accuracy of 51.73%), which suggests that ResNet50 suffered from overfitting. It is important to state two major important points about this set of experiments. The first one relates to the final classification accuracies which have been obtained, and that at first sight might seem disappointing. It is true that these classification rates are significantly lower when compared to the ones obtained in the previous two experiments. However, it is important to highlight how a large set of artists present in the dataset are associated to an extremely limited amount of samples. This reflects a lack of appropriate training data which does not allow the DCNNs to learn all the features that are necessary to successfully deal with this particular classification challenge. In order to do so, we believe that more training data is required. Moreover, it is worth pointing out how despite performing very poorly when used as off the shelf feature extractors, ImageNet pre-trained models do still perform better once they are fine tuned than the DCNN which is trained from scratch. This suggests that

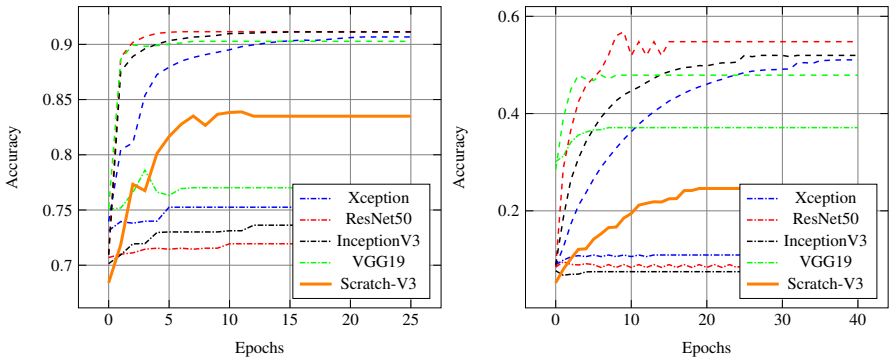

Fig. 3: A similar analysis as the one which has been reported in Figure 2 but for the second and third classification challenges (left and right figures respectively). The results show again the significant benefits that fine tuning (reported by the dashed line plots) has when compared to the off the shelf approach (reported by the dash-dotted lines) and how this latter strategy miserably under-performs when it comes to artist classification. Furthermore we again see the benefits that using a pre-trained DCNN has over training the architecture from scratch (solid orange line).

these networks do learn potentially representative features when it comes to challenge 3, but in order to properly exploit them, the DCNNs need to be fine tuned.

3.2 Discussion

In the previous section, we have investigated whether four different DCNNs pre-trained on the ImageNet dataset can be successfully used to address three art classification problems. We have observed how this is particularly the case when it comes to classifying the material and types, where in fact, the off the shelf approach can already lead to satisfactory results. However, most importantly, we have also shown how these performances are always significantly improved if the DCNNs are fine tuned and how an ImageNet initialization is beneficial over training the networks from scratch. Furthermore, we have discovered how the pre-trained DCNNs fail if used as simple feature extractors when having to attribute the authorship to the different heritage objects. In the next section, we want now to explore if the fine tuned DCCNs can lead to better performances, when tackling two of the already seen classification challenges on a different heritage collection. For this problem, we will again compare the off the shelf approach with the fine tuning one.

3.3 From One Art Collection to Another

Table 3 compares the results that have been obtained on the Antwerp dataset when using ImageNet pre-trained DCNNs (which are identified by θ) versus the same architectures fine tuned on the Rijksmuseum dataset ($\widehat{\theta}$). Similarly to the results presented in the

Table 2: An overview of the results obtained by the different DCNNs on the testing set when classifying the heritage objects of the Rijksmuseum. Bold results report the best performing architectures overall. The additional columns "Params" and "$\mathcal{X}$" report the amount of parameters the ANNs have to learn and the size of the feature vector which is used as input for the softmax classifier.

Challenge	**DCNN**	**off the shelf**	**fine tuning**	**Params**	$\mathcal{X}$
1	Xception	87.69%	92.13%	21K	2048
1	InceptionV3	88.24%	92.10%	22K	2048
1	ResNet50	86.81%	**92.95%**	24K	2048
1	VGG19	**92.12%**	92.23%	20K	512
2	Xception	74.80%	90.67%	23K	2048
2	InceptionV3	72.96%	91.03%	24K	2048
2	ResNet50	71.23%	**91.30%**	25K	2048
2	VGG19	**77.33%**	90.27%	20K	512
3	Xception	10.92%	51.43%	23K	2048
3	InceptionV3	.07%	**51.73%**	24K	2048
3	ResNet50	.08%	46.13%	26K	2048
3	VGG19	**38.11%**	44.98%	20K	512

previous section the first blue block of the table refers to the "type" classification task, while the red one reports the results obtained on the "artist" classification challenge.

While looking at the performances of the different neural architectures two interesting results can be highlighted. First, DCNNs which have been fine tuned on the Rijksmuseum dataset outperform the ones pre-trained on ImageNet in both classification challenges. This happens to be the case both when the DCNNs are used as simple feature extractors and when they are fine tuned. On the "type" classification challenge, this result is not surprising since, as discussed in Section 2.2, the types corresponding to the heritage objects of the two collections partially overlap. This is more suprising on the "artist" classification challenge however, since there is no overlap at all between the artists of the Rijksmuseum and the ones from the Antwerp dataset. A second interesting result, which is consistent with the results in the previous section, is the observation that it is always beneficial to fine tune the DCNNs over just using them as off the shelf feature extractors. Once the ANNs get fine tuned on the Antwerp dataset, these DCNNs, which have also been fine tuned on the Rijksmuseum dataset, outperform the architectures which have been pre-trained on ImageNet only. This happened to be the case for both classification challenges and for all considered architectures, as reported in Table 3. This demonstrates how beneficial it is for DCNNs to have been trained on a similar source task and how this can lead to significant improvements both when the networks are used as feature extractors and when they are fine tuned.

Table 3: The results obtained on the classification experiments performed on the Antwerp dataset with DCNNs which have been initially pre-trained on ImageNet (θ) and the same architectures which have been fine tuned on the Rijksmuseum dataset ($\widehat{\theta}$). Our results show how the latter pre-trained DCNNs yield better results both if used as off the shelf feature extractors and if fine tuned.

Challenge	**DCNN**	**θ + off the shelf**	**$\widehat{\theta}$ + off the shelf**	**θ + fine tuning**	**$\widehat{\theta}$ + fine tuning**
2	Xception	42.01%	62.92%	69.74%	72.03%
2	InceptionV3	43.90%	57.65%	70.58%	71.88%
2	ResNet50	41.59%	**64.95%**	76.50%	**78.15%**
2	VGG19	38.36%	60.10%	70.37%	71.21%
3	Xception	48.52%	**54.81%**	58.15%	58.47%
3	InceptionV3	21.29%	53.41%	56.68%	57.84%
3	ResNet50	22.39%	31.38%	62.57%	**69.01%**
3	VGG19	49.90%	53.52%	54.90%	60.01%

3.4 Selective Attention

The benefits of the fine tuning approach over the off the shelf one are clear from our previous experiments. Nevertheless, we do not have any insights yet as to what exactly allows fine tuned DCNNs to outperform the architectures which are pre-trained on ImageNet only. In order to provide an answer to that, we investigate which pixels of each input image contribute the most to the final classification predictions of the DCNNs. We do this by using the "VisualBackProp" algorithm presented by [5], which is able to identify which feature maps of the DCNNs are the most informative ones with respect to the final predictions of the network. Once these feature maps are identified, they get backpropagated to the original input image, and visualized as a saliency map according to their weights. The higher the activation of the filters, the brighter the set of pixels covered by these filters are represented.

The results that we have obtained provide interesting insights about how fine tuned DCNNs 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. We report the existence of these mechanisms in Figure 4 where we visualize the different saliency maps between a DCNN pre-trained on ImageNet and the same neural architecture which has been fine tuned on the Rijksmuseum collection (specifically renamed RijksNet [8]). On the left side of Figure 4 we visualize which sets of pixels allow the fine tuned DCNN to successfully classify an artist of the Rijksmuseum collection that the same architecture was not able to initially recognize. It is possible to notice how the saliency maps of the latter architecture either correspond to what is more similar to a natural image, as present in the ImageNet dataset (e.g. the buildings of the first and

[8] To show these results we have used the VGG19 architecture since it provided a better integration with the publicly available source code of the algorithm which can be found at https://github.com/raghakot/keras-vis

third images), or even to non informative pixels at all, as shown by the second image. However, the fine tuned DCNN shows how these saliency maps change towards the set of pixels that correspond to the portions of the images representing people in the bottom, suggesting that this is what allows the DCNN to appropriately recognize the artist. Similarly, on the right side of the figure we report which parts of the original image are the most useful ones when it comes to classify the type of the reported heritage object, which in this case corresponds to a glass wall of a church. We can see how the pre-trained architecture only identifies as representative pixels the right area above the arch, which turned out to be not informative enough for properly classifying this sample of the Rijksmuseum dataset. However, once the DCNN gets fine tuned we clearly see how in addition to the previously highlighted area a new saliency map occurs on the image, corresponding to the text description below the represented scene. It turns out that the presence of text is a common element below the images that represent clerical glass windows and as a consequence it is recognized by the fine tuned DCNN as a representative feature.

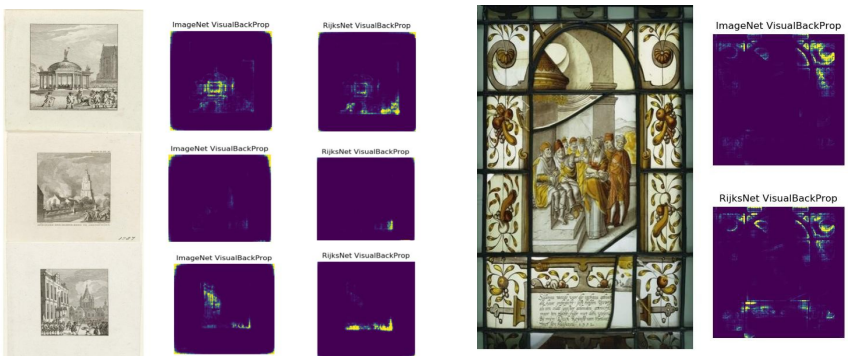

Fig. 4: A visualization that shows the differences between which sets of pixels in an image are considered informative for a DCNN which is only pre-trained on ImageNet, compared to the same architecture which has also been fine tuned on the Rijksmuseum collection. It is clear how the latter neural network develops novel selective attention mechanisms over the original image.

These observations can be related to parallel insights in authorship attribution research [11], an established task from Natural Language Processing that is highly similar in nature to artist recognition. In this field, preference is typically given to high-frequency function words (articles, prepositions, particles etc.) over content words (nouns, adjectives, verbs, etc.), because the former are generally considered to be less strongly related to the specific content or topic of a work. As such, function words or stop words lend themselves more easily to attribution across different topics and genres. In art history, strikingly similar views have been expressed by the well-known scholar Giovanni Morelli (1816-1891), who published seminal studies in the field of artist recognition [33]. In Morelli's view too, the attribution of a painting could not happen on the basis of

the specific content or composition of a painting, because these items were too strongly influenced by the topic of a painting or the wishes of a patron. Instead, Morelli proposed to base attributions to so-called *Grundformen* or small, seemingly insignificant details that occur frequently in all paintings and typically show clear traces of an artist's individual style, such as ears, hands or feat, a painting's function words, so to speak. The saliency maps above reveal a similar shift in attention when the ImageNet weights are adapted on the Rijksmuseum data: instead of focusing on higher-level content features, the network shifts its attention to lower layers in the network, seemingly focusing on insignificant details, that nevertheless appear crucial to perform artist attribution.

4 Conclusion

This paper provides insights about the potential that the field of TL has for art classification. We have investigated the behavior of DCNNs which have been originally pre-trained on a very different classification task and shown how their performances can be improved when these networks are fine tuned. Moreover, we have observed how such neural architectures perform better than if they are trained from scratch and develop new saliency maps that can provide insights about what makes these DCNNs outperform the ones that are pre-trained on the ImageNet dataset. Such saliency maps reflect themselves in the development of new features, which can then be successfully used by the DCNNs when classifying heritage objects that come from different heritage collections. It turns out that the fine tuned models are a better alternative to the same kind of architectures which are pre-trained on ImageNet only, and can serve the CV community which will deal with similar machine learning problems.

As future work, we aim to investigate whether the results that we have obtained on the Antwerp dataset will also apply to a larger set of smaller heritage collections. Furthermore, we want to explore the performances of densely connected layers [14] and understand which layers of the currently analyzed networks contribute the most to their final classification performances. This might allow us to combine the best parts of each neural architecture into one single novel DCNN which will be able to tackle all three classification tasks at the same time.

Acknowledgements

The authors wish to acknowledge Jeroen De Meester (Museums and Heritage Antwerp) for sharing his expertise on the Antwerp dataset. The research for this project was financially supported by BELSPO, Federal Public Planning Service Science Policy, Belgium, in the context of the BRAIN-be project: "INSIGHT. Intelligent Neural Systems as InteGrated Heritage Tools".

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: OSDI. vol. 16, pp. 265–283 (2016)

2. Ackermann, S., Schawinksi, K., Zhang, C., Weigel, A.K., Turp, M.D.: Using transfer learning to detect galaxy mergers. Monthly Notices of the Royal Astronomical Society (2018)
3. Allen, N.: Collaboration through the colorado digitization project. First Monday **5**(6) (2000)
4. Bidoia, F., Sabatelli, M., Shantia, A., Wiering, M.A., Schomaker, L.: A deep convolutional neural network for location recognition and geometry based information. In: Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods, ICPRAM 2018, Funchal, Madeira - Portugal, January 16-18, 2018. pp. 27–36 (2018)
5. Bojarski, M., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., Muller, U., Zieba, K.: Visualbackprop: efficient visualization of cnns. arXiv preprint arXiv:1611.05418 (2016)
6. Caruana, R., Lawrence, S., Giles, C.L.: Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In: Advances in neural information processing systems. pp. 402–408 (2001)
7. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. arXiv preprint (2016)
8. Chollet, F., et al.: Keras (2015)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 248–255. IEEE (2009)
10. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
11. Efstathios, S.: A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology (3), 538–556 (2009). https://doi.org/10.1002/asi.21001
12. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. vol. 1 (2017)
15. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst (2007)
16. Kornblith, S., Shlens, J., Le, Q.V.: Do better imagenet models transfer better? arXiv preprint arXiv:1805.08974 (2018)
17. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
18. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence. In: Proceedings of the IEEE international conference on computer vision. pp. 2623–2631 (2015)
19. Masters, D., Luschi, C.: Revisiting small batch training for deep neural networks. arXiv preprint arXiv:1804.07612 (2018)
20. Mensink, T., Van Gemert, J.: The rijksmuseum challenge: Museum-centered visual recognition. In: Proceedings of International Conference on Multimedia Retrieval. p. 451. ACM (2014)
21. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Transactions on knowledge and data engineering pp. 1345–1359 (2010)

22. Parry, R.: Digital heritage and the rise of theory in museum computing. Museum management and Curatorship pp. 333–348 (2005)
23. Razavian, A.S., Azizpour, H., Sullivan, J., Carlsson, S.: Cnn features off-the-shelf: an astounding baseline for recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. pp. 512–519. IEEE (2014)
24. Reyes, A.K., Caicedo, J.C., Camargo, J.E.: Fine-tuning deep convolutional networks for plant recognition. In: CLEF (Working Notes) (2015)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: The german traffic sign recognition benchmark: a multi-class classification competition. In: Neural Networks (IJCNN), The 2011 International Joint Conference on. pp. 1453–1460. IEEE (2011)
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016)
28. Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J.: Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging pp. 1299–1312 (2016)
29. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning pp. 26–31 (2012)
30. Tomè, D., Monti, F., Baroffio, L., Bondi, L., Tagliasacchi, M., Tubaro, S.: Deep convolutional neural networks for pedestrian detection. Signal Processing: Image Communication pp. 482–489 (2016)
31. Weibel, S., Kunze, J., Lagoze, C., Wolf, M.: Dublin core metadata for resource discovery. Tech. rep. (1998)
32. van de Wolfshaar, J., Karaaba, M.F., Wiering, M.A.: Deep convolutional neural networks and support vector machines for gender recognition. In: Computational Intelligence, 2015 IEEE Symposium Series on. pp. 188–195. IEEE (2015)
33. Wollheim, R.: On Art and the Mind. Essays and Lectures. Allen Lane (1972)
34. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 5987–5995. IEEE (2017)