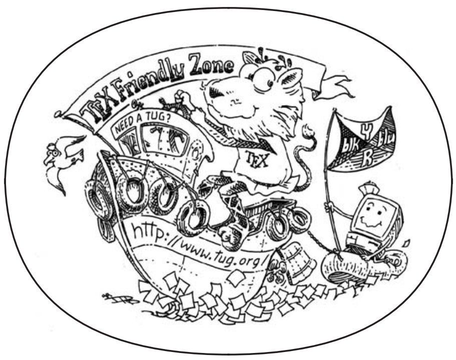
A CLASSIC THESIS STYLE

ANDRÉ MIEDE



An Homage to The Elements of Typographic Style

September 2015 – version 4.2

Ohana means family.
Family means nobody gets left behind, or forgotten.

— Lilo & Stitch

Dedicated to the loving memory of Rudolf Miede.

1939 – 2005

ABSTRACT

Short summary of the contents... a great guide by Kent Beck how to write good abstracts can be found here:

https://plg.uwaterloo.ca/~migod/research/beck00PSLA.html

PUBLICATIONS

Some ideas and figures have appeared previously in the following publications:

Put your publications from the thesis here. The packages `multibib` or `bibtopic` etc. can be used to handle multiple different bibliographies in your document.

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

— **knuth:1974 [knuth:1974]**

ACKNOWLEDGEMENTS

Put your acknowledgements here.

Many thanks to everybody who already sent me a postcard!

Regarding the typography and other help, many thanks go to Marco Kuhlmann, Philipp Lehman, Lothar Schlesier, Jim Young, Lorenzo Pantieri and Enrico Gregorio[1], Jörg Sommer, Joachim Köstler, Daniel Gottschlag, Denis Aydin, Paride Legovini, Steffen Prochnow, Nicolas Repp, Hinrich Harms, Roland Winkler, and the whole LaTeX-community for support, ideas and some great software.

Regarding LyX: The LyX port was initially done by *Nicholas Mariette* in March 2009 and continued by *Ivo Pletikosić* in 2011. Thank you very much for your work and the contributions to the original style.

1 Members of GuIT (Gruppo Italiano Utilizzatori di TeX e LaTeX)

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LISTINGS

ACRONYMS

DRY Don't Repeat Yourself

API Application Programming Interface

UML Unified Modeling Language

Part I

TRANSFER LEARNING IN SUPERVISED LEARNING

1 TRANSFER LEARNING FROM THE NATURAL TO THE NON-NATURAL DOMAIN

Contributions and Outline

This chapter contributes to the field of (Deep) Transfer Learning (TL) by investigating whether popular neural networks that come as pre-trained on datasets containing natural images can perform equally well once they are used on non-natural datasets. Specifically, we explore whether these models can be used for tackling three different art classification problems. Furthermore, assuming this is the case, we also explore whether it is possible to improve on such performance. The chapter is structured as follows: we start by providing the reader with some background information in Section 1.1. In Section 1.2 we present a brief theoretical reminder of 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 2.4 we present and discuss our results. A summary of the main contributions of this work ends the chapter in Section 2.6.

This chapter is based on the publication Sabatelli et al. [42].

1.1 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, 28, 52]. 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 neural networks 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 performance 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 deep neural networks in parallel over several `GPUs` [25], has lead to the development of a large set of different neural architectures that have continued to outperform each other over the years [7, 20, 21, 43, 49]. A promising research field in which the classification performances of such DCNNs can be exploited is that of *Digital Heritage* [37]. 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, 32]. The images constituting these datasets are mostly matched with descriptive metadata which, as presented by Mensink and Van Gemert [32], 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 deep networks 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 popular models that were trained on many large datasets [22, 44], were able to achieve very promising results on classification problems from heterogeneous domains, ranging from medical imaging [50] or gender recognition [55] over plant classification [40] 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 performance 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 [39], simply retrieves the features that were learned by the networks on other datasets and uses them as input for a new classifier. In this scenario the weights of the model 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 network are "unfrozen" and the neural architectures are trained together with the final classifier.

Kornblith, Shlens, and Le [24] have 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 networks that are randomly initialized are able to achieve the same performance as ImageNet pre-trained models. However, none of the results presented in [24] report experiments on datasets containing heritage objects, it is thus still an open question how such pre-trained DCNNs would perform in such a classification scenario. In the rest of this chapter we report results that extensively study the performance of such neural networks; at the same time we also assess whether better TL performance can be obtained when using neural networks that, in addition to the ImageNet dataset, have additionally been pre-trained on a large artistic collection.

1.2 METHODS

We now present the methods that underpin our research. We start by giving a brief formal reminder 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.

1.2.1 *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 $LS_t = \{(x_i, y_i)|i = 1, \dots, N_t\}$ 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 [36]. 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* [36]. While several inductive transfer learning algorithms exist, hereafter we focus on model transfer techniques, where information between the source and target problems is exchanged in the form of a neural network that comes as 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.

1.2.2 *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[1]. 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)[2]. 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 [54]. 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 neural networks 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 that 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),

1 https://staff.fnwi.uva.nl/t.e.j.mensink/uva12/rijks/
2 https://dams.antwerpen.be/

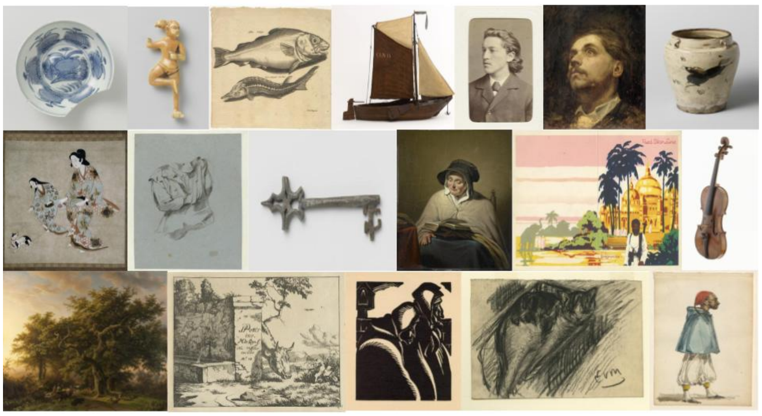


Figure 1: A visualization of the images that are used for our experiments. It is possible to see how the samples range from images representing plates made of porcelain to violins, and from Japanese artworks to a more simple picture of a key.

and that there are some common labels between the two heritage collections when it comes to challenge 2. A visualization reporting some of the images that are present in both datasets is shown in Figure 1.

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	

We use 80% of the datasets for training while the remaining $2 \times 10\%$ 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 [3].

3 https://github.com/paintception/Deep-Transfer-Learning-for-Art-Classification-Problems

1.2.3 *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 [43], Inception-V3 [49], Xception [7] and ResNet50 [57]. We use the implementations of the networks that are provided by the `Keras` Deep Learning library [8] 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 to classify, 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 layer and the *softmax* output. Hence, in this way we are able to obtain a feature vector, 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 all networks minimize the *categorical crossentropy* function loss function.

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 X, which is retrieved from the different models after performing one forward pass of the image through the network [4]. This approach is intended to explore whether the features that are learned by the network on the ImageNet dataset 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. From now on we will refer to these networks with θ^{-}.

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 network 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 models get independently trained on samples coming from the artistic datasets, and therefore their classification predictions will not be restricted to what the networks previously learned on the ImageNet dataset only. Evidently,

4 Please note that 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.

such an approach is computationally more demanding. We refer to these networks as θ^i, where i stands for the source task these models have been trained on, namely the ImageNet dataset.

In order to maximize the performance of all models we follow some of the recommendations presented by Masters and Luschi [30] and train the networks 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 models. 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 [51] which has been initialized with its default hyperparameters (learning rate = 0.001, a *momentum* value $\rho = 0.9$ and $\epsilon = 1e - 08$). 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, the results presented in this chapter are the first ones that systematically asses to which extent models pre-trained on the ImageNet dataset can be used as valuable architectures when tackling art classification problems. Furthermore, we also present the first results that investigate it is also not known whether the fine tuning approach can yield better results than the off the shelf one, and if using such pre-trained models would yield better performance than training the same architectures from scratch as observed by Kornblith, Shlens, and Le [24].

1.3 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 1.2.2. We report these results in Section 1.3.1 where we explore the benefits of using the ImageNet dataset as the TL source data, and how well such pre-trained models generalize when it comes to the artistic domain. We then present the results from classifying the Antwerp dataset, using models that are both pre-trained on the ImageNet dataset and on the Rijksmuseum collection in Section 1.3.3. 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 model from scratch, we also report the results that have been obtained when training a network with weights that have been initially sampled from a "He-Uniform" distribution [19]. Since we take advantage of the work presented by Bidoia et al. [4] we use the Inception-V3 architecture. We refer to it in all figures as Scratch-V3 and always visualize it with a solid orange line. Figures 2 and 3 report the performance in terms of accuracy that the models 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.

1.3.1 *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 1.2.2 for two main reasons. First, the number of possible classes the networks 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 a very similar performance 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.

When comparing the learning curves of the two classification strategies reported in Figure 2, we can 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 network that is trained from scratch. This suggests that features extracted from pretrained ImageNet models are relevant for material classification.

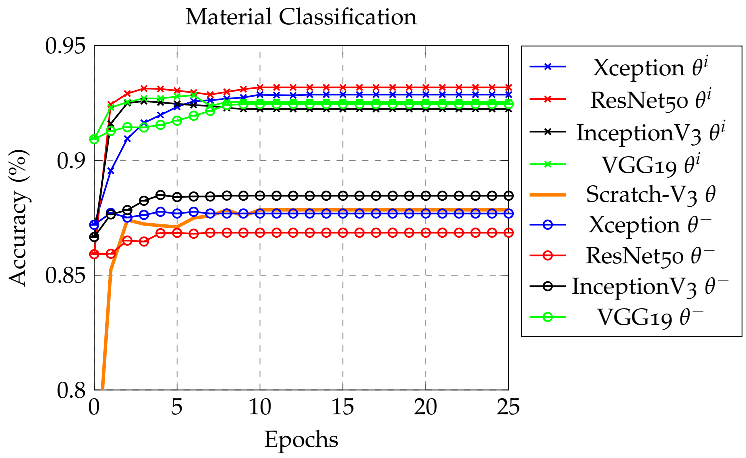


Figure 2: Comparison between the fine tuning approach (θ^i) versus the off the shelf one (θ^-) when classifying the material of the heritage objects of the Rijksmuseum dataset. We can observe that for three out of four neural architectures the first approach leads to significant improvements when compared to the latter one. Furthermore, we can also observe that training a randomly initialized model from scratch (solid orange line) leads to worse results than fine-tuning a network that comes as pre-trained on the ImageNet dataset.

We can also observe that the ResNet50 architecture is the architecture which, when fine tuned, performs overall best when compared to the other three models. This happens despite it being the network that initially performed worse as a simple feature extractor in the off the shelf experiments. As reported in Table 2 we can see that 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 that the performance between the different neural architectures do not strongly differ between each other once they are fine tuned, with all models 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 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 performance of the different neural architectures on the "type" and "artist" classification challenges (respectively the left and right plots reported in Figure 3), we observe that on these problems fine tuning strategy leads to even more significant improvements when compared to what we observed in the previous experiment. The results obtained on the second challenge show again that the ResNet50 architecture is the architecture 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 model 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 network 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 training approach. Similarly to what has been observed before, our results are again not significantly in favor of any fine tuned neural architecture, 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, a final accuracy of 92.12% has 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%), these results still show that models 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.

When considering the third classification challenge we can observe that the Xception, ResNet50, and Inception-V3 architectures all perform extremely poorly if not fine tuned, with the latter two models 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, the performance of each model are again significantly improved when the networks are fine tuned. As already observed in the previous experiments, ResNet50 outperforms the other architectures 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 accuracy that is obtained by all models, and that at first sight might seem disappointing. While it is true that these

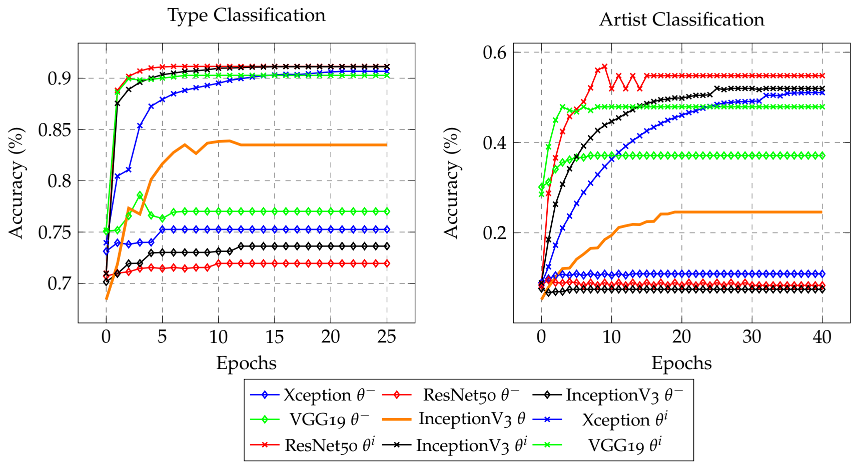


Figure 3: A similar analysis as the one which has been reported in Figure 2 but for the second and third classification challenges (left and right figures respectively). The results show again the significant benefits that fine tuning (reported by the dashed line plots) has when compared to the off the shelf approach (reported by the dash-dotted lines) and how this latter strategy miserably underperforms 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).

classification rates are significantly lower when compared to the ones obtained in the previous two experiments, 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 models to learn all the features that are necessary for successfully dealing with this particular classification challenge. In order to do so, we believe that more training data is required. Moreover, it is worth pointing out that 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 a model that is trained from scratch. This suggests that these networks do learn potentially representative features when it comes the classification of artists, but in order to properly classify them, the model need to be fine tuned.

1.3.2 *Discussion*

In the previous section, we have investigated whether four different architectures pre-trained on the ImageNet dataset can be successfully used to address three art classification problems. We have observed that this is particularly the case when it comes to classifying the material and the type, where in fact, the off the shelf approach already yielded satisfactory results. However, most importantly, we have also shown that the performance of all models can be significantly im-

Table 2: An overview of the results obtained by the different models on the testing set when classifying the heritage objects of the Rijksmuseum. The overall best performing architecture is reported in a green cell, while the second best performing one is reported in a yellow one. The additional columns "Params" and "X" report the amount of parameters the networks have to learn and the size of the feature vector that is used as input for the softmax classifier.

Challenge	**DCNN**	**off the shelf**	**fine tuning**	**Params**	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

proved when the networks are fine tuned, and that an ImageNet initialization is beneficial when compared to training a randomly initialized network from scratch. Furthermore, we have discovered that the pre-trained DCNNs fail if used as simple feature extractors when having they are used for attributing the authorship to the different heritage objects. In the next section, we explore the performance of fine tuned models that are trained for 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.

1.3.3 *From One Art Collection to Another*

Table 3 compares the results that we obtained on the Antwerp dataset when using ImageNet pre-trained models (θ^i) versus the same architectures that were fine tuned on the Rijksmuseum dataset (θ^r). While looking at the performance of the different neural architectures two interesting results can be highlighted. First, models 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 networks 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

1.2.2, the types corresponding to the heritage objects of the two collections partially overlap. This is more surprising 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 presented in the previous section, revolves around the observation that it is always beneficial to fine tune the networks over just using them as off the shelf feature extractors. Once the models get fine tuned on the Antwerp dataset, these DCNNs, which have also been fine tuned on the Rijksmuseum dataset, outperform the architectures that were 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 the models 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 models that have been initially pre-trained on ImageNet (θ^i) and the same architectures which have been fine tuned on the Rijksmuseum dataset (θ^r). Our results show that the latter pre-trained networks yield better results both if used as off the shelf feature extractors and if fine tuned.

Challenge	DCNN	θ^i + off the shelf	θ^r + off the shelf	θ^i + fine tuning	θ^r + 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%

1.3.4 *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 models 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 networks. We do this by using the "VisualBackProp" algorithm presented by [5], which is able to identify which feature maps of the networks are the most informative ones with respect to their final prediction. 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 models 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 model pre-trained on ImageNet and the same neural architecture which has been fine tuned on the Rijksmuseum collection. In Figure 4 we visualize which sets of pixels allow the fine tuned model to successfully classify an artist of the Rijksmuseum collection that the same architecture was not able to initially recognize. It is possible to notice that the saliency maps of the latter architecture either correspond to what is more similar to a natural image, as represented by the central image of the first row of plots, or even to what appear to be non informative pixels at all, as shown by the second image in the second row. However, when considering the fine tuned model we clearly observe that these saliency maps change. In this case the network attends towards the set of pixels that represent people in the bottom, suggesting that this is what allows the model to appropriately recognize the artist of the considered artwork.

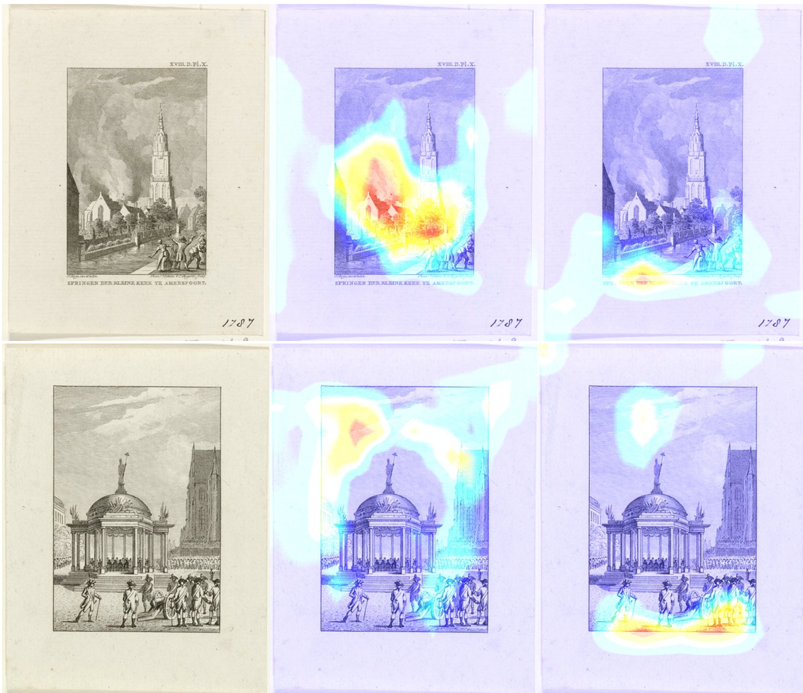


Figure 4: Add caption

These observations can be related to parallel insights in authorship attribution research [12], 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 [56]. 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.

1.4 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 [21] 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.

2 TRANSFER LEARNING WITH NATURAL LOTTERY WINNERS

2.1 INTRODUCTION

The "Lottery-Ticket-Hypothesis" (LTH) [14] states that within large randomly initialized neural networks there exist smaller sub-networks which, if trained from their initial weights, can perform just as well as the fully trained unpruned network from which they are extracted. This happens to be possible because the weights of these sub-networks seem to be particularly well initialized before training starts, therefore making these smaller architectures suitable for learning (see Fig 5 for an illustration). These sub-networks, i.e., the pruned structure together with their initial weights, are called winning tickets, as they appear to have won the initialization lottery. Since winning tickets only contain a very limited amount of parameters, they yield faster training, inference, and sometimes even better final performance than their larger over-parametrized counterparts [14, 16]. So far, winning tickets are typically identified by an iterative procedure that cycles through several steps of network training and weight pruning, starting from a randomly initialized unpruned network. While simple and intuitive, the resulting algorithm, has unfortunately a high computational cost. Despite the fact that the resulting sparse networks can be trained efficiently and in isolation from their initial weights, the LTH idea has not yet led to more efficient solutions for training a sparse network, than existing pruning algorithms that all also require to first fully train an unpruned network [11, 18, 26, 33, 59].

Since the introduction of the idea of the LTH, several research works have focused on understanding what makes some weights so special to be the winners of the initialization lottery. Among the different tested approaches, which will be reviewed in Sec. 2.5, one research direction in particular has looked into how well winning ticket initializations can be transferred among different training settings (datasets and optimizers), an approach which aims at characterizing the winners of the LTH by studying to what extent their inductive biases are generic [34]. The most interesting findings of this study are that winning tickets generalize across datasets, within the natural image domain at least, and that tickets obtained from larger datasets typically generalize better. This opens the door to the transfer of winning tickets between datasets, which makes the high computational cost required to identify them much more acceptable practically, as this cost has to be paid only once and can be shared across datasets.

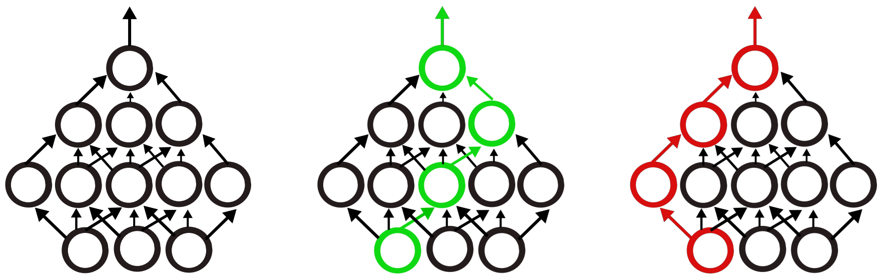


Figure 5: A visual representation of the LTH as introduced in [14]. Let us consider a simplified version of a two hidden layer feedforward neural network as is depicted in the first image on the left. The LTH states that within this neural network there exist multiple smaller networks (represented in green), which perform just as well as their larger counterpart. Training these sparse models from scratch successfully is only possible as long as their weights are initialized with the same values that were also used when the larger (black) model was initialized. As can be seen by the blue curve of the last plot the performance of such pruned models gets barely harmed even when large pruning rates are reached. These models are considered as the winners of the initialization lottery and also perform better than the same models re-initialized randomly (orange line). Results obtained on the MNIST dataset that replicate the findings presented in [14].

In this paper, we build on top of this latter work. While Morcos et al. [34] focused on the natural image domain, we investigate the possibility of transferring winning tickets obtained from the natural image domain to datasets in non natural image domains. This question has an important practical interest as datasets in non natural image domains are typically scarcer than datasets in natural image domains. They would therefore potentially benefit more from a successful transfer of sparse networks, since the latter can be expected to require less data for training than large over-parametrized networks. Furthermore, besides studying their generalization capabilities, we also focus on another interesting property that characterizes models that win the LTH, and which so far has received less research attention. As originally presented in [14], pruned models which are the winners of the LTH can yield a final performance which is better than the one obtained by larger over-parametrized networks. In this work we explore whether it is worth seeking for such pruned models when training data is scarce, a scenario that is well known to constraint the training of deep neural networks. To answer these two questions, we carried out experiments on several datasets from two very different non natural image domains: digital pathology and digital heritage.

Research questions and contributions: this work investigates two research questions. First, we aim at better characterizing the LTH phe-

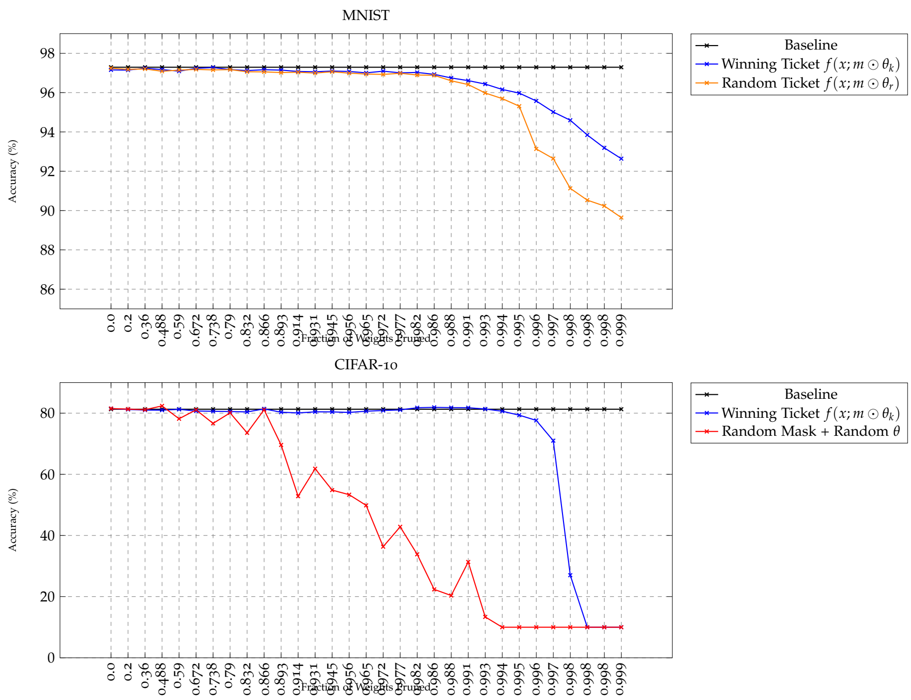


Figure 6

nomenon by investigating whether lottery winners that are found on datasets of natural images contain inductive biases that are strong enough to allow them to generalize to non-natural image distributions. To do so, we present to the best of our knowledge the first results that study the transferability of winning initializations in this particular training setting. Second, we thoroughly study for the first time whether pruned models that are the winners of the LTH can consistently outperform their larger over-parametrized counterparts in conditions with scarce training data.

2.2 DATASETS

i We consider seven datasets that come from two different, unrelated sources: histopathology and digital heritage. Each dataset comes with its training, validation and testing splits. Furthermore the datasets change in terms of size, resolution, and amount of labels that need to be classified. We report an overview about the size of these datasets in Table **??** while a visual representation of the samples constituting these datasets in Fig. 7. The Digital-Pathology (DP) data comes from the Cytomine [29] web application, an open-source platform that allows interdisciplinary researchers to work with large-scale images. While Cytomine has collected a large number of datasets over the years, in this work we have limited our analysis to a subset of four datasets that all represent tissues and cells from either human or animal organs: `Human-LBA`, `Lung-Tissues`, and `Mouse-LBA` were originally proposed in [35], while Bone-Marrow comes from [23]. All four datasets have been used in [35], that researched whether neural networks pre-trained on natural images could successfully be re-used in the DP domain. In this paper, we explore whether an alternative to the transfer-learning approaches presented in [35] could be based on training pruned networks that are the winners of the LTH. This will allow us to investigate the two research questions introduced in Sec. 2.1: we will explore whether winning initializations that are found on datasets of natural images do generalize to non-natural domains, and whether sparse models winners of the LTH can perform better than larger unpruned models that get trained from scratch. Regarding the field of Digital-Humanities (DH) we have created three novel datasets that all revolve around the classification of artworks. We consider two different classification tasks that have already been thoroughly studied by researchers bridging between the fields of Computer Vision (CV) and DH [32, 42, 45]. The first task consists in identifying the artist of the different artworks, while the second one aims at classifying which kind of artwork is depicted in the different images, a challenge which is usually referred to in the literature as type-classification [32, 42]. When it comes to the artist-classification task we have created two different datasets, which purpose will be better explained in Sec. 2.4.3. All images are publicly available as part of the WikiArt gallery [38] and can also be found within the large popular OmniArt dataset [46]. Albeit in DH it is actually easier to find large datasets than in histopathology, it is worth mentioning that we have kept the size of these datasets intentionally small in order to fit the research questions introduced in Sec. 2.1. Furthermore, it is also worth noting that there are several additional challenges that need to be overcome when training deep neural networks on artistic collections, which therefore motivate the use of this kind of datasets in this work. The size, texture, and resolution of the images coming from the DH are usually representative of different time periods, artistic movements and might have gone through different digitization processes, which are all reasons that make these datasets largely varied and challenging.

2.3 EXPERIMENTAL SETUP

We follow an experimental set-up similar to the one that was introduced in [34] (and that has been validated by [17]). Let us define a neural network $f(x;\theta)$ that gets randomly initialized with parameters $\theta_0 \sim \mathcal{D}_{\subseteq}$ and then trained for j iterations over an input space $\mathcal{X}$, and an output space $\mathcal{Y}$. At the end of training a percentage of the parameters in θ_j gets pruned, a procedure which results in a mask m. The parameters in θ_j which did not get pruned are then reset to the values they had at θ_k, where k represents an early training itera-

Table 4: A brief overview of the seven different datasets which have been used in this work. N_t corresponds to the total amount of samples that are present in the dataset, while Q_t represents the number of classes.

Dataset	Training-Set	Validation-Set	Testing-Set	N_t	Q_t
`Human-LBA`	4051	346	1023	5420	9
`Lung-Tissues`	4881	562	888	6331	10
`Mouse-LBA`	1722	716	1846	4284	8
`Bone-Marrow`	522	130	639	1291	8
`Artist-Classification-1`	3103	389	389	3881	20
`Type-Classification`	2868	360	360	3588	20
`Artist-Classification-2`	2827	353	353	3533	19

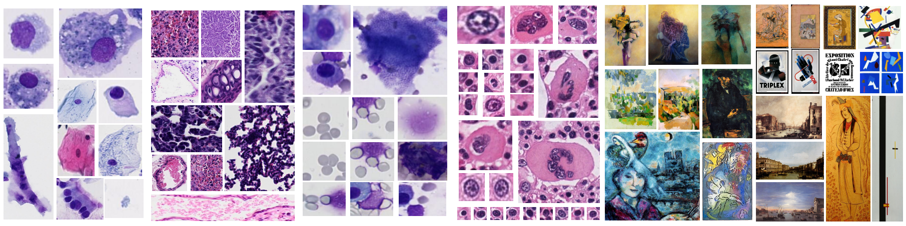


Figure 7: Some image samples that constitute the non-natural image datasets which have been used in this work. From left to right we have the `Human-LBA`, `Lung-Tissues`, `Mouse-LBA` and `Bone-Marrow` datasets, while finally we report some examples that represent artworks which come from the field of digital heritage.

tion. A winning ticket corresponds to the combination between the previously obtained mask, and the parameters θ_k, and is defined as $f(x; m \odot \theta_k)$ [1]. Constructing a winning ticket with parameters θ_k, instead of θ_0, is a procedure which is known as late-resetting [16], and is a simple but effective trick that makes it possible to stably find winning initializations in deep convolutional neural networks [16, 34]. In this study $f(x; \theta)$ comes in the form of a ResNet-50 architecture [18] which gets trained on the three popular CV natural image datasets CIFAR-10/100 and Fashion-MNIST. Following [18, 34], 31 winning tickets $f(x; m \odot \theta_k)$ of increasing sparsity are obtained from each of these three datasets by repeating 31 iterations of network training and magnitude pruning with a pruning rate of 20%. More precisely, at each pruning iteration, the network is trained for several epochs (using early stopping on the validation set as described in the appendix) and then the 20% of weights with the lowest magnitudes are pruned. The parameters θ_k that define each of the 31 tickets are then taken as the weights of the corresponding pruned networks at the kth

1 Note that this formulation generalizes the original version of the LTH [14] that we have represented in Fig. 5, where a winning ticket is obtained after resetting the unpruned parameters of the network to the values they had right after initialization, therefore defining a winning ticket as $f(x; m \odot \theta_0)$.

epoch of the first pruning iteration. Once these pruned networks are found we aim at investigating whether their parameters θ_k contain inductive biases that allow them to generalize to the non-natural image domain. To do so we replace the final fully connected layer of each winning ticket with a randomly initialized layer that has as many output nodes as there are classes to classify. We then fine-tune each of these networks on the non-natural image datasets considered in this study. At the end of training, we study the performance of each winning ticket in two different ways. First, we compare the performance of each network to the performance of a fully unpruned network that gets randomly initialized and trained from scratch. Second, we also compare the performance of winning tickets that have been found on a natural image dataset to 31 new sparse models that are the winners of the LTH on the considered target dataset. Since it is not known to which extent pruned networks that contain weights that are the winners of the LTH on a natural image dataset can generalize to target distributions that do not contain natural images, we report the first results that investigate the potential of a novel transfer-learning scheme which has so far only been studied on datasets from the natural image domain. Moreover, testing the performance of sparse networks that contain winning tickets that are specific to a non-natural image target distribution also allows us to investigate whether it is worth pruning large networks with the hope of finding smaller models that might perform better than a large over-parametrized one. As mentioned in Sec. 2.1, pruned networks that are initialized with the winning weights can sometimes perform better than a fully unpruned network. Identifying such sparse networks leads to a very significant reduction of model size, which can be a very effective way of regularization when training data is scarce.

2.4 RESULTS

The results of all our experiments are visually reported in the plots of Fig. **??**. Each line plot represents the final performance that is obtained by a pruned model that contains a winning ticket initialization on the final testing-set of our target datasets. This performance is reported on the y-axis of the plots, while on the x-axis we represent the fraction of weights that is pruned from the original ResNet-50 architecture. As explained in the previous section the performance of each winning ticket is compared to the performance that is obtained by an unpruned, over-parametrized architecture that is reported by the black dashed lines. The models that are the winners of the LTH on a natural image dataset are reported by the green, red and purple lines, while the winners of the LTH on a non-natural target dataset are reported by the blue lines. Furthermore, when it comes to the latter lottery tickets, we also report the performance that is obtained by winning tickets that get randomly reinitialized ($f(x; m \odot \theta_0')$ with $\theta_0' \sim \mathcal{D}_\theta$). These results are reported by the orange lines. Shaded areas around all line plots correspond to ± 1 std. that has been obtained after repeating and averaging the results of our experiments over four different random seeds.

2.4.1 *On the Importance of Finding Winning Initializations*

We can start by observing that pruned models which happen to be the winners of the LTH either on a natural dataset, or on a non-natural one, can maintain a good final performance until large pruning rates are reached. This is particularly evident on the first three datasets, where models that keep only $\approx 1\%$ of their original weights barely suffer from any drop in performance. This gets a little bit less evident on the last three datasets, where the performance of winning ticket initializations that are directly found on the considered target dataset starts getting harmed once a fraction of ≈ 97 of original weights are pruned. These results show that an extremely large part of the parameters of a ResNet-50 architecture can be considered as superfluous, therefore confirming the LTH when datasets contain non-natural images. More importantly, we also observe that pruned models winners of the LTH, significantly outperform larger over-parametrized models that get trained from scratch. This can be very clearly seen in all plots where the performance of pruned models is always consistently better than what is reported by the black dashed line. To get a better sense of how much these pruned networks perform better than their larger unpruned counterparts, we report in Table 5 the performance that is obtained by the best performing pruned model, found over all 31 possible pruned models, and compare it to the performance of an unpruned architecture. The exact fraction of weights which is pruned from an original ResNet-50 architecture is reported in Table 6 for each configuration. We can observe that no matter which dataset has been used as a source for finding a winning ticket initialization, all pruned networks reach a final accuracy that is significantly higher than the one that is obtained after training an unpruned model from scratch. While in most cases the difference in terms of performance is of $\approx 10\%$ (see e.g. the `Human-LBA`, `Lung-Tissues` and the `Type-Classification` datasets), it is worth highlighting that there are other cases in which this difference is even larger. This is the case for the `Mouse-LBA` and `Artist-Classification-1` datasets where a winning ticket coming from the CIFAR-10 dataset performs more than 20% better than a model trained from scratch. These results show that in order to maximize the performance of deep networks it is always worth finding and training pruned models which are the winners of the LTH.

Table 5: The results comparing the performance that is obtained on the testing-set by the best pruned model winner of the LTH, and an unpruned architecture trained from scratch. The overall best performing model is reported in a green cell, while the second best one in a yellow cell. We can observe that pruned models winners of the LTH perform significantly better than a larger over-parametrized architecture that gets trained from scratch. As can be seen by the results obtained on the `Mouse-LBA` and `Artist-Classification-1` datasets the difference in terms of performance can be particularly large ($\approx$ 20%).

Target-Dataset	Scratch-Training	CIFAR-10	CIFAR-100	Fashion-MNIST	Target-Ticket
`Human-LBA`	$71.85_{\pm 1.12}$	$79.17_{\pm 1.85}$	$76.97_{\pm 0.73}$	$77.32_{\pm 1.85}$	$81.72_{\pm 0.39}$
`Lung-Tissues`	$84.75_{\pm 0.81}$	$88.90_{\pm 1.97}$	$87.61_{\pm 0.90}$	$87.61_{\pm 0.11}$	$90.48_{\pm 0.16}$
`Mouse-LBA`	$48.17_{\pm 1.18}$	$74.20_{\pm 2.04}$	$57.42_{\pm 0.48}$	$52.27_{\pm 1.73}$	$68.20_{\pm 3.79}$
`Bone-Marrow`	$64.66_{\pm 1.36}$	$71.75_{\pm 3.36}$	$69.87_{\pm 0.39}$	$68.77_{\pm 0.39}$	$72.55_{\pm 0.46}$
`Artist-Classification-1`	$45.88_{\pm 0.42}$	$66.58_{\pm 1.54}$	$65.55_{\pm 1.79}$	$63.88_{\pm 0.12}$	$58.74_{\pm 1.92}$
`Type-Classification`	$41.36_{\pm 2.31}$	$58.63_{\pm 2.97}$	$60.56_{\pm 0.44}$	$58.92_{\pm 0.59}$	$50.44_{\pm 2.23}$

Table 6: Some additional information about the lottery winners which performance is reported in Table 5. For each winning ticket we report the fraction of weights that is pruned from an original ResNet-50 architecture and that therefore characterizes the level of sparsity of the overall best performing lottery ticket. The results in the Scratch-Training column are not reported since these are unpruned models that are trained from scratch.

Target-Dataset	Scratch-Training	CIFAR-10	CIFAR-100	Fashion-MNIST	Target-Ticket
`Human-LBA`	-	0.945	0.79	0.886	0.832
`Lung-Tissues`	-	0.977	0.977	0.672	0.965
`Mouse-LBA`	-	0.972	0.893	0.738	0.931
`Bone-Marrow`	-	0.866	0.988	0.931	0.914
`Artist-Classification-1`	-	0.972	0.993	0.991	0.931
`Type-Classification`	-	0.991	0.931	0.995	0.963

2.4.2 *On the Generalization Properties of Lottery Winners*

We then investigate whether natural image tickets can generalize to the non-natural setting. Findings differ across datasets. When considering the datasets that come from the field of DP, we can see that, in three out of four cases, winning tickets that are found on a natural image dataset get outperformed by sparse winning networks that come after training a model on the biomedical dataset. This is particularly evident in the results obtained on the `Human-LBA` and `Lung-Tissues` datasets where the highest testing-set accuracy is consistently reached by the blue line plots. When it comes to the `Bone-Marrow` dataset the difference in terms of performance between the best natural image ticket, in this case coming from the CIFAR-10 dataset, and the one coming from the biomedical dataset, is less evident (see Table 5 for the exact accuracies). Furthermore, it is worth highlighting that on the `Bone-Marrow` dataset, albeit natural image models seem to get outperformed by the ones found on the biomedical dataset, the performance

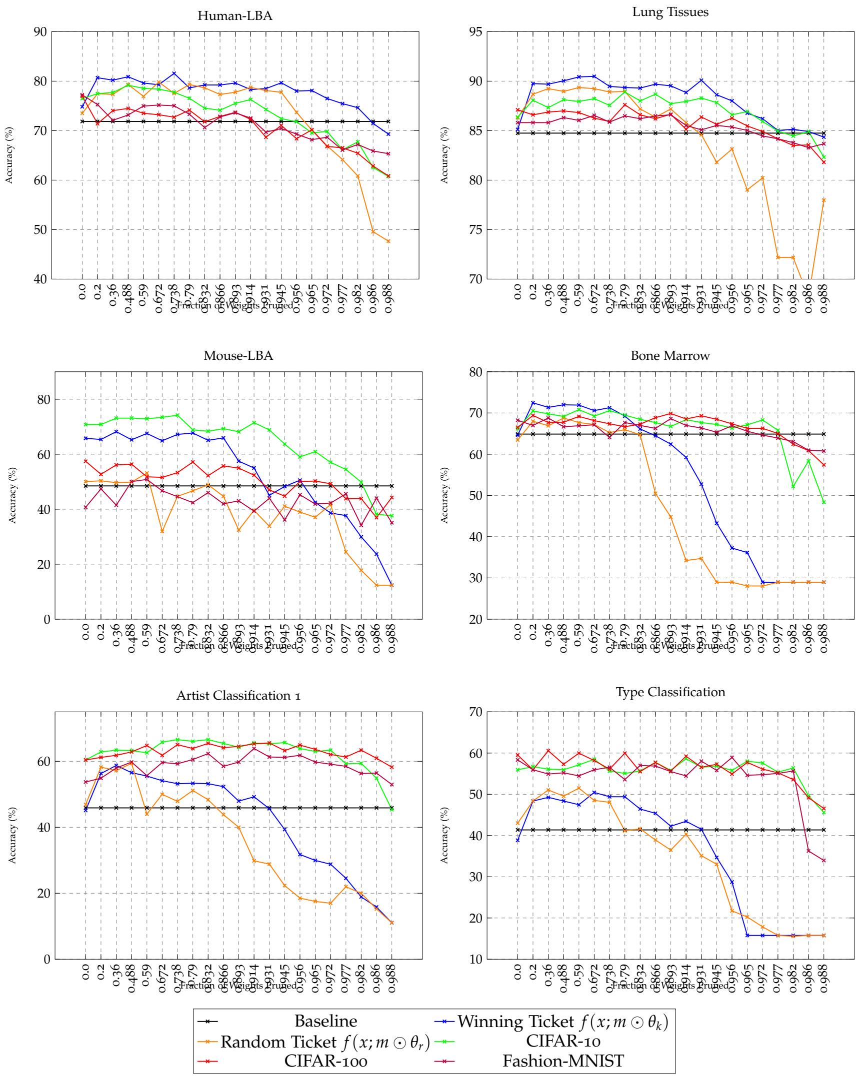


Figure 8: An overview of the results showing that sparse models that are the winners of the LTH (represented by the coloured lines) significantly outperform unpruned networks which get randomly initialized and trained from scratch (dashed black line). This happens to be the case on all tested datasets, no matter whether a winning initialization comes from a natural image source or not. It is however worth mentioning that, especially on the biomedical datasets, natural image tickets get outperformed by sparse networks that that are the winners of the LTH on a biomedical dataset. On the other hand this is not the case when it comes to the classification of arts where natural image tickets outperform the ones which are found within artistic collections.

of the latter ones appears to be less stable once extremely large pruning rates are reached. When it comes to the `Mouse-LBA` dataset these results slightly differ. In fact, this dataset corresponds to the only case where a natural image source ticket outperforms a non-natural one. As can be seen, by the green line plot, pruned models coming from the CIFAR-10 dataset outperform the ones found on the `Mouse-LBA` dataset. When focusing our analysis on the classification of arts, we see that the results change greatly from the ones obtained on the biomedical datasets. In this case, all of the natural image lottery winners, no matter the dataset they were originally found on, outperform the same kind of models that were found after training a full network on the artistic collection. We can see from Table 5 that the final testing performance is similar among all of the best natural image tickets. Similarly to what has been noticed on the `Bone-Marrow` dataset we can again observe that tickets coming from a non-natural data distribution seem to suffer more from large pruning rates.

These results show both the potential and the limitations that natural image winners of the LTH can offer when they are fine-tuned on datasets of non-natural images. The results obtained on the artistic datasets suggest that winning initializations contain inductive biases that are strong enough to get at least successfully transferred to the artistic domain, therefore confirming some of the claims that were made in [34]. However, it also appears that there are stronger limitations to the transferability of winning initializations which were not observed by [34]. In fact, our results show that on DP data the best strategy is to find a winning ticket directly on the biomedical dataset, and that winning initializations found on natural image datasets, albeit outperforming a randomly initialized unpruned network, perform worse than pruned models that are the winners of the LTH on a biomedical dataset.

2.4.3 *Additional Studies*

To characterize the transferability of winning initializations even more, while at the same time gaining a deeper understanding of the LTH, we have performed a set of three additional experiments which help us characterize this phenomenon even better.

2.4.3.1 *Lottery Tickets VS fine-tuned pruned models*

So far we have focused our transfer-learning study on lottery tickets that come in the form of $f(x; m \odot \theta_k)$, where, as mentioned in Sec. 2.3, θ_k corresponds to the weights that parametrize a neural network at a very early training iteration. This formalization is however different from more common transfer-learning scenarios where neural networks get transferred with the weights that are obtained at the end of the training process [35, 42]. We have therefore studied whether there is a difference in terms of performance between transferring and fine-tuning a lottery ticket with parameters θ_k, and the same kind of pruned network which is initialized with the weights that are obtained once the network is fully trained on a source task. We define these kind of models as $f(x; m \odot \theta_i)$ where i stays for the last training iteration. We report some examples of this behaviour in the first row of plots presented in Fig. **??**, where we consider $f(x; m \odot \theta_i)$ models which were trained on the CIFAR-10 and CIFAR-100 datasets, and then transferred and fine-tuned on the `Human-LBA` dataset. We found that these models overall perform worse than lottery tickets, while also being less robust to pruning. This also shows that on this dataset, the slightly inferior performance of the natural image tickets with respect to the target tickets is not due to the weight re-initialization.

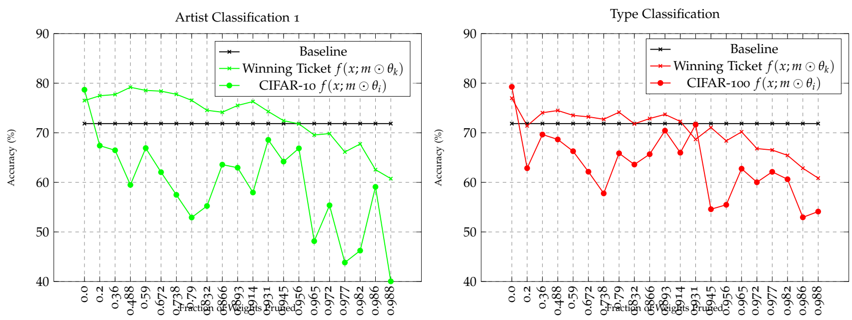


Figure 9

2.4.3.2 *Transferring tickets from similar non-natural domains*

We investigated whether it is beneficial to fine-tune lottery winners that, instead of coming from a natural image distribution, come from a related non-natural dataset. Specifically we tested whether winning tickets generated on the `Human-LBA` dataset generalize to the `Mouse-LBA` one (since both datasets are representative of the field of Live-Blood-Analysis), and whether lottery winners coming from the `Artist-Classification-1` dataset generalized to the `Artist-Classification-2` one. We visually represent these results in the central plots presented in Fig. **??**. As one might expect, we found that it is beneficial to transfer winning tickets that come from a related source. Specifically, `Human-LBA` tickets can perform just as well as winning tickets that are generated on the `Mouse-LBA` dataset, while at the same time also being more robust to large pruning rates. When it comes to lottery winners found on the `Artist-Classification-1` dataset we have observed that these tickets can even outperform the ones generated on the `Artist-Classification-2` one.

2.4.3.3 *On the size of the training set*

We have observed from the blue line plots of Fig. **??** that there are cases in which lottery winners are very robust to extremely large pruning rates (see as an example the first and second plots), while there are other cases in which their performance deteriorates faster with respect to the fraction of weights that get pruned. The most robust performance is obtained by winning tickets that are generated on the Human-LBA and Lung-Tissues datasets, which are the two target datasets that contain the largest amount of training samples. We have therefore studied whether there is a relationship between the size of the training data that is used for finding lottery winners, and the robustness in terms of performance of the resulting pruned models. We generated lottery winners after incrementally reducing the size of the training data by 75%, 50% and 25%, and then investigated whether we could observe a similar drop in performance as the one which can be seen by the last three blue line-plots of Fig. **??** once a large fraction of weights got pruned. Perhaps surprisingly, we have observed that this was not the case, and as can be seen by the plots represented in the last row of Fig. **??**, the performance of lottery winners that are found when using only 25% of the training set is just as stable as the one of winning tickets which are generated on the entire dataset. It is however worth mentioning that, albeit the performance of such sparse models is robust, their final performance on the testing set is lower than the one that is obtained by winning tickets that have been trained on the full training data distribution.

2.5 RELATED WORK

The research presented in this paper contributes to a better understanding of the LTH by exploring the generalization and transfer-learning properties of lottery tickets. The closest approach to what has been presented in this work is certainly [34], which shows that winning models can generalize across datasets of natural images and across different optimizers. As mentioned in Sec. 2.3, a large part of our experimental setup is based on this work. Besides the work presented in [34], there have been other attempts that aimed to better understand the LTH after studying it from a transfer-learning perspective. However, just as the study presented in [34], all this research limited its analysis to natural images. In [53] the authors transfer winning tickets among different partitions of the CIFAR-10 dataset, while in [31] the authors show that sparse models can successfully get transferred from the CIFAR-10 dataset to other object recognition tasks. While these results seem to suggest that lottery tickets contain inductive biases which are strong enough to generalize to different domains, it is worth highlighting that their transfer-learning properties were only studied after considering the CIFAR-10 dataset as a possible source for winning ticket initializations, a limitation which we overcome in this work. It is also worth mentioning that the research presented in this paper is strongly connected to the work presented in [16]. While the first paper that introduced the LTH limited its analysis to relatively simple neural architectures, such as multi-layer perceptrons and convolutional networks which were tested on small CV datasets, the presence of winning initializations in larger, more popular convolutional models such as [48] and [20] trained on large datasets [41] was only first presented in [16]. Since in this work we have used a ResNet-50 architecture [20], we have followed all the recommendations that were introduced in [16], for successfully identifying the winners of the LTH in larger models. More specifically we mention the late-resetting procedure which resets the weights of a pruned model to the weights that are obtained after k training iterations instead of to the values which were used at the beginning of training (as explained in Sec. 2.3), a procedure which has shown to be related to *linear mode connectivity* [15]. While the work presented in this paper has limited its analysis to networks that minimize an objective function that is relevant for classification problems, it is worth noting that more recent approaches have identified lottery winners in different training settings. In [58] the authors show that winning initializations can be found when neural networks are trained on tasks ranging from natural language processing to reinforcement learning, while in [47] the authors successfully identify sparse winning models in a multi-task learning scenario. As future work, we want to study whether lottery tickets can be found on different neural architectures, and also when neural networks are trained on CV tasks other than classification. More specifically we aim at studying whether winners of the LTH, which are found on popular natural image datasets such as [27] and [13] when tackling image localization and segmentation tasks, can generalize to non-natural settings which might include the segmentation of biomedical data, or the localization of objects within artworks.

2.6 CONCLUSION

We have investigated the transfer learning potential of pruned neural networks that are the winners of the LTH from datasets of natural images to datasets containing non-natural images. We have explored this in training conditions where the size of the training data is relatively small. All of the results presented in this work confirm that it is always beneficial to train a sparse model, winner of the LTH, instead of a larger over-parametrized one. Regarding our study on the transferability of winning tickets we have reported the first results which study this phenomenon under non-natural data distributions by using datasets coming from the fields of digital pathology and heritage. While for the case of artistic data it seems that winning tickets from the natural image domain contain inductive biases which are strong enough to generalize to this specific domain, we have also shown that this approach can present stronger limitations when it comes to biomedical data. This probably stems from the fact that DP images are further away from natural images than artistic ones. We have also shown that lottery tickets perform significantly better than fully trained pruned models, that it is beneficial to transfer lottery winners from different, but related, non-natural sources, and that the performance of lottery tickets is not dependant on the size of the training data. To conclude, we provide a better characterization of the LTH while simultaneously showing that when training data is limited, the performance of deep neural networks can get significantly improved by using lottery winners over larger over-parametrized ones.

Part II

TRANSFER LEARNING IN REINFORCEMENT LEARNING

Part III

APPENDIX

A APPENDIX TEST

Aliquam lectus. Vivamus leo. Quisque ornare tellus ullamcorper nulla. Mauris porttitor pharetra tortor. Sed fringilla justo sed mauris. Mauris tellus. Sed non leo. Nullam elementum, magna in cursus sodales, augue est scelerisque sapien, venenatis congue nulla arcu et pede. Ut suscipit enim vel sapien. Donec congue. Maecenas urna mi, suscipit in, placerat ut, vestibulum ut, massa. Fusce ultrices nulla et nisl.

Etiam ac leo a risus tristique nonummy. Donec dignissim tincidunt nulla. Vestibulum rhoncus molestie odio. Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero. Suspendisse in elit quis nisl aliquam dapibus. Pellentesque auctor sapien. Sed egestas sapien nec lectus. Pellentesque vel dui vel neque bibendum viverra. Aliquam porttitor nisl nec pede. Proin mattis libero vel turpis. Donec rutrum mauris et libero. Proin euismod porta felis. Nam lobortis, metus quis elementum commodo, nunc lectus elementum mauris, eget vulputate ligula tellus eu neque. Vivamus eu dolor.

A.1 APPENDIX SECTION TEST

Nulla in ipsum. Praesent eros nulla, congue vitae, euismod ut, commodo a, wisi. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Aenean nonummy magna non leo. Sed felis erat, ullamcorper in, dictum non, ultricies ut, lectus. Proin vel arcu a odio lobortis euismod. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Proin ut est. Aliquam odio. Pellentesque massa turpis, cursus eu, euismod nec, tempor congue, nulla. Duis viverra gravida mauris. Cras tincidunt. Curabitur eros ligula, varius ut, pulvinar in, cursus faucibus, augue.

Nulla mattis luctus nulla. Duis commodo velit at leo. Aliquam vulputate magna et leo. Nam vestibulum ullamcorper leo. Vestibulum condimentum rutrum mauris. Donec id mauris. Morbi molestie justo et pede. Vivamus eget turpis sed nisl cursus tempor. Curabitur mollis sapien condimentum nunc. In wisi nisl, malesuada at, dignissim sit amet, lobortis in, odio. Aenean consequat arcu a ante. Pellentesque porta elit sit amet orci. Etiam at turpis nec elit ultricies imperdiet. Nulla facilisi. In hac habitasse platea dictumst. Suspendisse

LABITUR BONORUM PRI NO	QUE VISTA	HUMAN
fastidii ea ius	germano	demonstratea
suscipit instructior	titulo	personas
quaestio philosophia	facto	demonstrated

Table 7: Autem usu id.

viverra aliquam risus. Nullam pede justo, molestie nonummy, scelerisque eu, facilisis vel, arcu.

A.2 ANOTHER APPENDIX SECTION TEST

Curabitur tellus magna, porttitor a, commodo a, commodo in, tortor. Donec interdum. Praesent scelerisque. Maecenas posuere sodales odio. Vivamus metus lacus, varius quis, imperdiet quis, rhoncus a, turpis. Etiam ligula arcu, elementum a, venenatis quis, sollicitudin sed, metus. Donec nunc pede, tincidunt in, venenatis vitae, faucibus vel, nibh. Pellentesque wisi. Nullam malesuada. Morbi ut tellus ut pede tincidunt porta. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam congue neque id dolor.

Donec et nisl at wisi luctus bibendum. Nam interdum tellus ac libero. Sed sem justo, laoreet vitae, fringilla at, adipiscing ut, nibh. Maecenas non sem quis tortor eleifend fermentum. Etiam id tortor ac mauris porta vulputate. Integer porta neque vitae massa. Maecenas tempus libero a libero posuere dictum. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Aenean quis mauris sed elit commodo placerat. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Vivamus rhoncus tincidunt libero. Etiam elementum pretium justo. Vivamus est. Morbi a tellus eget pede tristique commodo. Nulla nisl. Vestibulum sed nisl eu sapien cursus rutrum.

There is also a useless Pascal listing below: Listing 1.

Listing 1: A floating example (`listings` manual)

```
for i:=maxint downto 0 do
begin
{ do nothing }
end;
```

BIBLIOGRAPHY

[1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. "TensorFlow: A System for Large-Scale Machine Learning." In: *OSDI*. Vol. 16. 2016, pp. 265–283.

[2] Sandro Ackermann, Kevin Schawinksi, Ce Zhang, Anna K Weigel, and M Dennis Turp. "Using transfer learning to detect galaxy mergers." In: *Monthly Notices of the Royal Astronomical Society* (2018).

[3] Nancy Allen. "Collaboration through the Colorado digitization project." In: *First Monday* 5.6 (2000).

[4] Francesco Bidoia, Matthia Sabatelli, Amirhossein Shantia, Marco A. Wiering, and Lambert Schomaker. "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*. 2018, pp. 27–36.

[5] Mariusz Bojarski, Anna Choromanska, Krzysztof Choromanski, Bernhard Firner, Larry Jackel, Urs Muller, and Karol Zieba. "VisualBackProp: efficient visualization of CNNs." In: *arXiv preprint arXiv:1611.05418* (2016).

[6] Rich Caruana, Steve Lawrence, and C Lee Giles. "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping." In: *Advances in neural information processing systems*. 2001, pp. 402–408.

[7] François Chollet. "Xception: Deep learning with depthwise separable convolutions." In: *arXiv preprint* (2016).

[8] François Chollet et al. *Keras*. 2015.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE. 2009, pp. 248–255.

[10] Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. "Decaf: A deep convolutional activation feature for generic visual recognition." In: *International conference on machine learning*. 2014, pp. 647–655.

[11] Xin Dong, Shangyu Chen, and Sinno Pan. "Learning to prune deep neural networks via layer-wise optimal brain surgeon." In: *Advances in Neural Information Processing Systems*. 2017, pp. 4857–4867.

[12] Stamatatos Efstathios. "A survey of modern authorship attribution methods." In: *Journal of the American Society for Information Science and Technology* 3 (2009), pp. 538–556. DOI: 10.1002/asi.21001.

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge." In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[14] Jonathan Frankle and Michael Carbin. "The lottery ticket hypothesis: Finding sparse, trainable neural networks." In: *arXiv preprint arXiv:1803.03635* (2018).

[15] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. "Linear mode connectivity and the lottery ticket hypothesis." In: *arXiv preprint arXiv:1912.05671* (2019).

[16] Jonathan Frankle, G Karolina Dziugaite, DM Roy, and M Carbin. "Stabilizing the Lottery Ticket Hypothesis." In: *arXiv preprint arXiv:1903.01611* (2019).

[17] Varun Gohil, S Deepak Narayanan, and Atishay Jain. "One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers." In: *ReScience-C* (2020).

[18] Song Han, Huizi Mao, and William J Dally. "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding." In: *arXiv preprint arXiv:1510.00149* (2015).

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification." In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[21] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. "Densely connected convolutional networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Vol. 1. 2. 2017.

[22] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *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.

[23] Philipp Kainz, Harald Burgsteiner, Martin Asslaber, and Helmut Ahammer. “Training echo state networks for rotation-invariant bone marrow cell classification.” In: *Neural Computing and Applications* 28.6 (2017), pp. 1277–1292.

[24] Simon Kornblith, Jonathon Shlens, and Quoc V Le. “Do Better ImageNet Models Transfer Better?” In: *arXiv preprint arXiv:1805.08974* (2018).

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks.” In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[26] Ji Lin, Yongming Rao, Jiwen Lu, and Jie Zhou. “Runtime neural pruning.” In: *Advances in Neural Information Processing Systems*. 2017, pp. 2181–2191.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. “Microsoft coco: Common objects in context.” In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[28] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. “Multimodal convolutional neural networks for matching image and sentence.” In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2623–2631.

[29] Raphaël Marée, Loïc Rollus, Benjamin Stévens, Renaud Hoyoux, Gilles Louppe, Rémy Vandaele, Jean-Michel Begon, Philipp Kainz, Pierre Geurts, and Louis Wehenkel. “Collaborative analysis of multi-gigapixel imaging data using Cytomine.” In: *Bioinformatics* 32.9 (2016), pp. 1395–1401.

[30] Dominic Masters and Carlo Luschi. “Revisiting Small Batch Training for Deep Neural Networks.” In: *arXiv preprint arXiv:1804.07612* (2018).

[31] Rahul Mehta. “Sparse Transfer Learning via Winning Lottery Tickets.” In: *arXiv preprint arXiv:1905.07785* (2019).

[32] Thomas Mensink and Jan Van Gemert. “The rijksmuseum challenge: Museum-centered visual recognition.” In: *Proceedings of International Conference on Multimedia Retrieval*. 2014, pp. 451–454.

[33] Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. “Pruning convolutional neural networks for resource efficient inference.” In: *arXiv preprint arXiv:1611.06440* (2016).

[34] Ari Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. “One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers.” In: *Advances in Neural Information Processing Systems*. 2019, pp. 4933–4943.

[35] Romain Mormont, Pierre Geurts, and Raphaël Marée. "Comparison of deep transfer learning strategies for digital pathology." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2262–2271.

[36] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning." In: *IEEE Transactions on knowledge and data engineering* (2010), pp. 1345–1359.

[37] Ross Parry. "Digital heritage and the rise of theory in museum computing." In: *Museum management and Curatorship* (2005), pp. 333–348.

[38] Fred Phillips and Brandy Mackintosh. "Wiki Art Gallery, Inc.: A case for critical thinking." In: *Issues in Accounting Education* 26.3 (2011), pp. 593–608.

[39] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. "CNN features off-the-shelf: an astounding baseline for recognition." In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE. 2014, pp. 512–519.

[40] Angie K Reyes, Juan C Caicedo, and Jorge E Camargo. "Fine-tuning Deep Convolutional Networks for Plant Recognition." In: *CLEF (Working Notes)*. 2015.

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. "Imagenet large scale visual recognition challenge." In: *International journal of computer vision* 115.3 (2015), pp. 211–252.

[42] Matthia Sabatelli, Mike Kestemont, Walter Daelemans, and Pierre Geurts. "Deep transfer learning for art classification problems." In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 631–646.

[43] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." In: *arXiv preprint arXiv:1409.1556* (2014).

[44] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. "The German traffic sign recognition benchmark: a multi-class classification competition." In: *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE. 2011, pp. 1453–1460.

[45] Gjorgji Strezoski and Marcel Worring. "Omniart: multi-task deep learning for artistic data analysis." In: *arXiv preprint arXiv:1708.00684* (2017).

[46] Gjorgji Strezoski and Marcel Worring. "Omniart: a large-scale artistic benchmark." In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.4 (2018), pp. 1–21.

[47] Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. "Learning Sparse Sharing Architectures for Multiple Tasks." In: *arXiv preprint arXiv:1911.05034* (2019).

[48] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.

[49] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2818–2826.

[50] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. "Convolutional neural networks for medical image analysis: Full training or fine tuning?" In: *IEEE transactions on medical imaging* (2016), pp. 1299–1312.

[51] Tijmen Tieleman and Geoffrey Hinton. "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude." In: *COURSERA: Neural networks for machine learning* (2012), pp. 26–31.

[52] Denis Tomè, Federico Monti, Luca Baroffio, Luca Bondi, Marco Tagliasacchi, and Stefano Tubaro. "Deep convolutional neural networks for pedestrian detection." In: *Signal Processing: Image Communication* (2016), pp. 482–489.

[53] Ryan Van Soelen and John W Sheppard. "Using winning lottery tickets in transfer learning for convolutional neural networks." In: *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2019, pp. 1–8.

[54] Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. *Dublin core metadata for resource discovery*. Tech. rep. 1998.

[55] Jos van de Wolfshaar, Mahir F Karaaba, and Marco A Wiering. "Deep convolutional neural networks and support vector machines for gender recognition." In: *Computational Intelligence, 2015 IEEE Symposium Series on*. IEEE. 2015, pp. 188–195.

[56] R. Wollheim. *On Art and the Mind. Essays and Lectures*. Allen Lane, 1972.

[57] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. "Aggregated residual transformations for deep neural networks." In: *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE. 2017, pp. 5987–5995.

[58] Haonan Yu, Sergey Edunov, Yuandong Tian, and Ari S Morcos. "Playing the lottery with rewards and multiple languages: lottery tickets in RL and NLP." In: *arXiv preprint arXiv:1906.02768* (2019).

[59] Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. "Discrimination-aware channel pruning for deep neural networks." In: *Advances in Neural Information Processing Systems*. 2018, pp. 875–886.

DECLARATION

Put your declaration here.

Saarbrücken, September 2015

André Miede

colophon

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of February 17, 2021 (`classicthesis` version 4.2).