FINE TUNING CNNS WITH SCARCE TRAINING DATA – ADAPTING IMAGENET TO ART EPOCH CLASSIFICATION

Christian Hentschel *Timur Pratama Wiradarma* *Harald Sack*

Hasso Plattner Institute for Software Systems Engineering,
Potsdam, Germany
christian.hentschel@hpi.de,
pratama.wiradarma@student.hpi.uni-potsdam.de,
harald.sack@hpi.de

ABSTRACT

Deep Convolutional Neural Networks (CNN) have recently been shown to outperform previous state of the art approaches for image classification. Their success must in parts be attributed to the availability of large labeled training sets such as provided by the ImageNet benchmarking initiative. When training data is scarce, however, CNNs have proven to fail to learn descriptive features. Recent research shows that supervised pre-training on external data followed by domain-specific fine-tuning yields a significant performance boost when external data and target domain show similar visual characteristics. In this paper, we evaluate the performance of fine-tuned CNNs when the target domain is visually different from the dataset used to pre-train the model. Specifically, we address the problem of transfer learning from ImageNet domain to the task of classifying paintings into art epochs. Furthermore, we analyze the impact of training set sizes on CNNs with and without external data and compare the obtained models to linear models based on Improved Fisher Encodings. Our results underline the superior performance of fine-tuned CNNs but likewise propose Fisher Encodings in scenarios were training data is limited.

***Index Terms*—** image classification, painting classification, CNN, IFV, comparison

1. INTRODUCTION

Since the first large-scale application of Deep Convolutional Neural Networks (CNN) for image classification in [1] research in computer vision has seen a significant shift away from *shallow*, engineered image descriptors (usually based on aggregation of local histograms of gradients) towards *deep* features trained and fitted on a large corpus of images. Typically, this large corpus of images is provided by the ImageNet database [2], an image dataset of more than 14 million images organized into more than 100,000 concepts. Evidence has been given, that at least parts of the success of Deep Convolutional Neural Networks should be attributed to the availability of large hand-labeled datasets such as ImageNet in order to avoid overfitting models with up to 144 million parameters [3, 4]. Thus, in scenarios where training data is limited, CNNs often fail to learn discriminative features. In [5] we compared and visualized the performance of CNNs and Improved Fisher Vector (IFV, [6]) representations as a function of the training set size using the ILSVRC-2012 benchmarking data. We observed that if the amount of training images is restricted, IFV-based approaches can outperform CNNs (up to 60% of the original training data was required for the CNN-based solution to achieve similar or better results than the IFV-based models). Since assembling large amounts of reliably annotated ground truth data is a costly (sometimes even prohibitively costly) and time consuming process, various efforts have been made to improve the performance of CNNs in scenarios where training data is limited.

Probably the most fruitful and therefore most often applied approach is to pre-train a CNN on a large corpus of outside data and to *fine-tune* parameters on the target data. The authors in [7] have trained deep representations on the ILSVRC-2012 challenge dataset and used PASCAL VOC [8] as well as Caltech [9] datasets to evaluate the learned models. Similarly in [3] a method is presented, which reuses layers trained on the ImageNet dataset to compute mid-level image representation for images in the PASCAL VOC dataset. In [10] the authors keep the first layers of a CNN trained on the ILSVRC-2012 and train a new softmax classifier and SVM classifier on top using the training images of PASCAL VOC and Caltech datasets respectively. Presumably the first to propose fine-tuning a CNN to a new domain were the authors in [11] who suggest to replace the CNN's ImageNet-specific 1000-way classification layer with a randomly initialized layer and to continue backpropagation to adapt the new layer to a new task (object detection on VOC data).

All of the aforementioned scenarios adapted the final CNN models to target data that should be considered visually similar to the ILSVRC data used to train the initial CNN.

While the number and actual set of target classes is different, ImageNet as well as PASCAL VOC and Caltech data represent real-world scenes or objects. Fine-tuning CNN models on data visually similar to the data used for pre-training, however, is more likely to succeed than trying to adapt a CNN model to visually different data as it can be assumed that the respective low- and mid-level features are correlated.

In this paper we therefore analyze whether transfer learning of features from a CNN pre-trained on ImageNet data and fine-tuned to the visually completely different domain of *paintings* is feasible. We approach the task of automatic classification of paintings into their respective art epochs such as *Baroque*, *Renaissance* or *Impressionism*. Furthermore, following the protocol in [5], we compare the performance of a fine-tuned CNN to a CNN trained without outside data as well as to previous state of the art feature representations (i.e. Improved Fisher Vectors). Especially, we analyze the impact of varying training set size on the overall model accuracy.

This paper is structured as follows: In the following section we present the dataset and the evaluation protocol. In section 3 we detail the three different image representations evaluated: CNN, CNN with fine-tuning and Improved Fisher Vectors. Section 4 discusses experimental results and section 5 concludes the paper and gives some outlook to future work.

2. DATASET AND EVALUATION PROTOCOL

The authors in [12] present a dataset of 85,000 paintings manually annotated with 25 style/genre labels (see Fig. 1 for some examples). The paintings as well as the annotations contributed by a community of experts have been collected from the WikiArt.org – Encyclopedia of fine arts[1] website (we follow the notation of the authors in [12] and refer to the dataset as Wikipaintings collection). As can be seen, paintings show visual characteristics such as strong strokes and color compositions that are highly different from the ones of the real-world objects and scenes available in the ImageNet dataset.

We used the collection to generate ground truth data by selecting 1,000 images for training and 50 images for testing (following the number of images provided per class in the ILSVRC-2012 benchmarking data). Since the number of images in the Wikipaintings collection varies a lot (e.g., for some categories less than 100 images are available while others provide more than 10,000 images), only those classes that are supported by at least 1,050 example images were chosen from the entire collection. Thus, our dataset consists of paintings from 22 different art epochs[2].

Based on these classes, we generated training subsets by randomly selecting 5, 10, 20, 40, 60, 80 and 100% of the entire training data (evenly spread over all classes) while keeping the test data fixed. For each subset, we train classifiers and test their performance on the test data by reporting mean average precision scores.

We present results for three different approaches that represent current or previous state of the art methods for image classification. Following our findings in [5] we train Support Vector Machine classifiers based on Improved Fisher Vector representations. Furthermore we evaluate two different approaches for training CNNs: First, we train CNNs using the individual training subsets taken from the Wikipaintings dataset only. Since the data provided in the Wikipaintings collection is rather small when compared to ImageNet, we likewise present results for a CNN pre-trained on the entire ILSVRC-2012 training data and fine-tuned using the different Wikipaintings training data splits.

3. IMAGE REPRESENTATIONS

This section presents the three types of image representations compared in our experiments. We provide detailed information for model parameters and learning strategies.

3.1. CNN training

Our CNN-based classifiers for painting style recognition follow the architecture as proposed by Krizhevsky, et al. in [1] with some minor modifications. These modifications address the sequence of pooling and normalization layers (compared to the original model we flip the order, i.e. pooling is applied before normalization) and mainly help to speed up the forward run without sacrificing the accuracy. The remainder of the architecture is left unchanged: The network consists of five convolutional layers activated by a Rectified Linear Unit (ReLU) and followed by a max pooling layer (applied to $1^{st}, 2^{nd}$ and 5^{th} convolutional layer). Local Response Normalization is applied after the 1^{st} and 2^{nd} pooling layer. Layers 6 to 8 are fully connected layers and a softmax layer computes a probability for each target class.

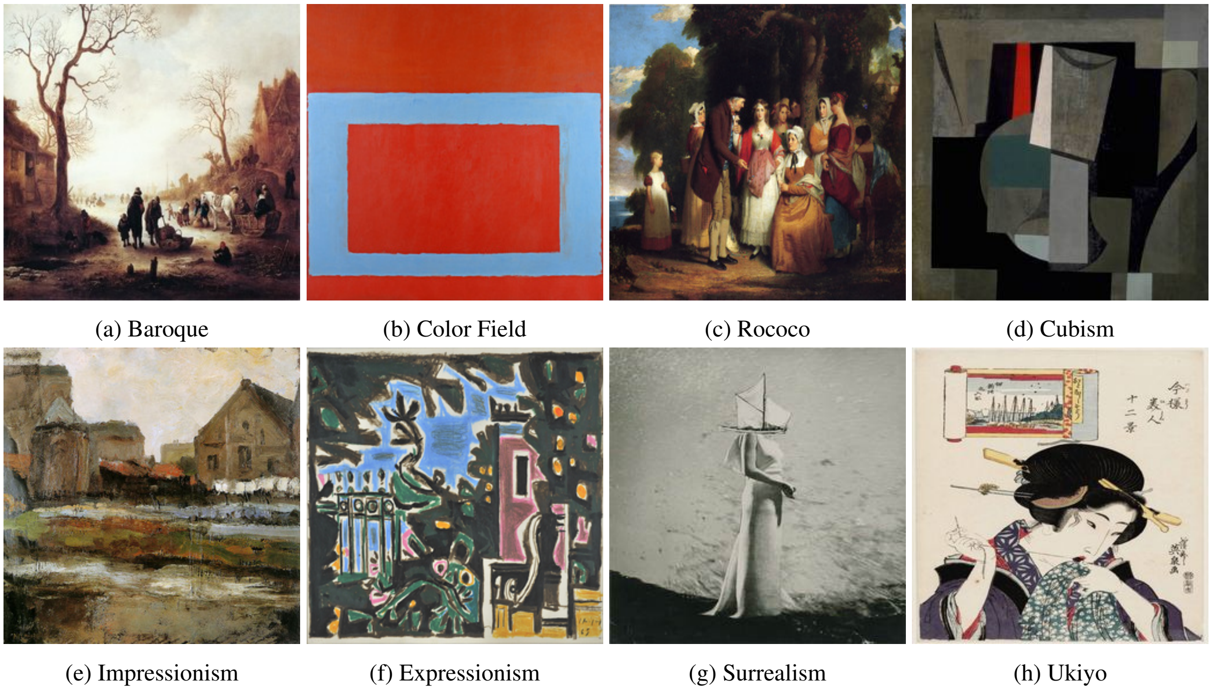
Our implementation uses the the open source Caffe CNN library presented in [13]. Following Krizhevsky, et al., every image is resized to 256×256 pixels and the center crop (224×224) is used as input image for the model. Additionally, mean subtraction – obtained by averaging the pixel values from all training images – is carried out for each input image.

3.2. CNN fine-tuning

Despite the fact that we use a similar amount of images per class and due to the limited overall number of classes, the Wikipaintings dataset is significantly smaller than the

[1] WikiArt.org – Encyclopedia of fine arts, www.wikiart.org

[2] Art epochs: *Mannerism (Late Renaissance), Pop Art, Rococo, Early Renaissance, Ukiyo-e, Art Nouveau (Modern), Northern Renaissance, Abstract Expressionism, Symbolism, Neoclassicism, Color Field Painting, Impressionism, Expressionism, Naive Art (Primitivism), Minimalism, Baroque, Post-Impressionism, Realism, Romanticism, Surrealism, High Renaissance, Cubism*



(a) Baroque (b) Color Field (c) Rococo (d) Cubism

(e) Impressionism (f) Expressionism (g) Surrealism (h) Ukiyo

Fig. 1: Examples of paintings taken from the Wikipaintings collection with manually assigned genre labels.

ILSVRC-2012 dataset (22K images in Wikipaintings as opposed to 1.2M in ILSVRC-2012). The authors in [11] have presented supervised pre-training on a large external dataset, followed by domain-specific fine-tuning on a small dataset, as an effective approach for learning CNNs when data is scarce. Here, we follow that approach by firstly training a CNN on ILSVRC-2012 data and fine-tuning the obtained model on the individual Wikipaintings subsets. Pre-training is again performed using Caffe library and by employing exactly the same CNN architecture as presented in the previous section. The CNN therefore shows a similar performance to the one presented in [1] when tested on the ILSVRC-2012 validation data.

In the presented architecture the last fully-connected layer ($fc8$) of the trained CNN has an output dimensionality equal to the number of classes – 1,000 in case of ILSVRC-2012 data. In order to adapt the CNN to new target data, we replace layer $fc8$ with randomly initialized weights for the 22 target outputs (according to the 22 classes chosen from the Wikipaintings collection). In order to focus learning on the new layer and having the rest of the model change very slowly we multiply the learning rate of the replaced layer by a factor of 10 while at the same time decreasing the overall learning rate (see [13]). Moreover, based on some experimental results, pre-trained models converge faster when using smaller epochs (i.e., 30 epochs). The reminder of the training parameters – such as the initial learning rate, momentum or batch sizes – are left unchanged.

3.3. Improved Fisher Encodings and linear classifiers

As discussed in the introduction, next to comparing the performance of different CNN training strategies on scarce training data, we also evaluate the performance of deep and shallow features by computing linear models on local image encodings. In [14] the authors compared different local feature encodings in a large scale experiment and conclude the superior performance of Fisher Vector (FV) encodings which we therefore adopted in our experiments as well. Consistent with the FV implementations proposed in [6], our approach starts by extracting SIFT descriptors [15] at a dense grid with a stride of 4 pixels at 7 different scales. We use the implementation provided by [16], which uses triangular feature reweighting. Following [6], we decorrelate and reduce the original feature dimensions from $d = 128$ to $d = 80$ by means of Principal Component Analysis (PCA). We further enhance the local descriptors by spatially extending the features with the (normalized) sampling point's coordinates, yielding a $d = 82$ dimensional descriptor.

A FV encoding is then obtained by first computing the Gaussian Mixture Model (GMM) with $k = 256$ components on a random subset of $n = 256,000$ local descriptors equally selected from all training images. Subsequently, each local descriptor of an image is soft-quantized using the obtained mixtures and first and second order statistics between the descriptor and its Gaussian cluster are accumulated. Finally, the *improved* version of FV (IFV, as suggested by the authors in [6]) applies signed square-rooting to the individual compo-

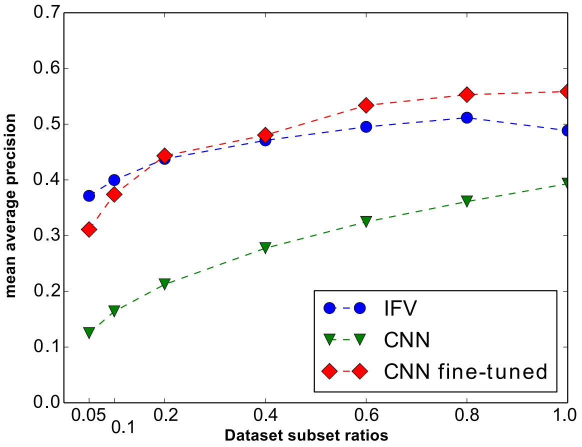


Fig. 2: A comparison of IFV, CNN (w/o pre-training) and CNN, fine-tuned (pre-trained using ILSVRC-2012 data, fine-tuned to Wikipaintings dataset) on the Wikipaintings dataset. Reported scores are MAP scores as a function of increasing training subsets ratios.

nents of the encoding followed by a $\|\cdot\|_2$ normalization.

Usually, visual concept models are trained based on these global feature representations using Support Vector Machines. Our implementation learns a linear SVM per image class (using a *one-vs-rest* pattern) by minimizing the hinge loss function. While the regularization C hyperparameter should be optimized using cross validation, we fix it to $C = 10$ in order to reduce training time.

4. EXPERIMENTAL RESULTS

In order to evaluate the results achieved by the different approaches, we have computed the mean average precision scores (MAP, averaged over all classes) and plotted them as a function of the training dataset size used to train the model. Fig. 2 shows the complete results on Wikipaintings.

Similar to our previous findings comparing CNNs and IFVs on ImageNet data with increasing training set sizes in [5], our results show that adding more training data in general helps to increase the obtained classifier accuracy. The highest score (i.e., MAP=55.9%) is achieved by the fine-tuned CNN model adapted using the entire training set. Moreover, while the IFV models converged at 80% and even dropped at 100% (from 51.2% to 48.9% MAP score) both CNN approaches improve with increasing number of training images.

Furthermore, our findings show that surprisingly little data is required (i.e. 20% of the original training set, in total 4,400 images) for the CNN pre-trained on ILSVRC data to adapt to the new target task of painting classification. The respective models clearly outperform our linear models based on IFV representations for training set ratios > 0.4. Thus, we conclude that in fact fine-tuned CNNs are a powerful concept even when adaptation to visually completely different target data is required.

On the other hand, when considering scenarios where training data is scarce, we can state that aggregated local features such as IFV should be considered a competitive alternative. When reducing the training data to 20% and less of the original size (i.e., 200 paintings per class) the linear IFV-based models outperform the corresponding fine-tuned CNNs (biggest difference appearing at 5% corresponding to 6% difference in achieved MAP score).

Considering the CNN models trained without auxiliary data we find that even the entire dataset is too small to compete with any of the other models. Considering the MAP curve, however, we observe that the model seems to be the least saturated from all of the evaluated representations which suggests that the CNN would benefit most from adding more training data. This is in line with our findings from [5] where for the ILSVRC classes most difficult to detect a CNN learned without external data needed more than 150,00 training images to outperform IFV-based solutions.

Finally, we observe that the achieved average precision scores are significantly lower when compared to classification of real-world objects and scenes as in ImageNet. Whereas CNNs as well as IFV-based approaches were able to achieve near perfect results (AP=100%) for the most easy categories of the ImageNet dataset none of the classification models could achieve near perfect results for any of the available classes. Again, this could be attributed to the lack of sufficient training data but likewise could be an indicator of the difficulty of the task at hand.

5. CONCLUSIONS

In this paper we have compared three state of the art image classification approaches for automatic detection of art epochs on paintings taken from the WikiArt.org project. More precisely, we evaluated Convolutional Neural Networks trained with and without outside data as well as linear classifiers applied on Improved Fisher Encodings. Reported results underline the superior performance of fine-tuned CNN models when 200 or more training images are available per class. In scenarios with fewer training data, IFV representations have proven to outperform CNN-based approaches by up to 6% MAP.

Future work will focus on the impact of data augmentation techniques. CNNs have been reported to benefit from simple data manipulation tricks such as rotating, flipping and blurring training images. By that means, the available training data can be increased at no additional labeling cost. Data augmentation can be conducted for all of the evaluated approaches.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Advances In Neural Information Processing Systems*, pp. 1097—-1105, 2012.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision*, 2015.

[3] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic, “Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks,” *Cvpr*, pp. 1717–1724, 2014.

[4] Karen Simonyan and Andrew Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recoginition,” in *Intl. Conf. on Learning Representations (ICLR)*, 2015, pp. 1–14.

[5] Christian Hentschel, Timur Pratama Wiradarma, and Harald Sack, “If we did not have imagenet: Comparison of fisher encodings and convolutional neural networks on limited training data,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9475, pp. 400–409, 2015.

[6] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, “Improving the Fisher kernel for large-scale image classification,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6314 LNCS, no. PART 4, pp. 143–156, 2010.

[7] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman, “Return of the Devil in the Details: Delving Deep into Convolutional Nets,” in *Proceedings of the British Machine Vision Conference*, Michel Valstar, Andrew French, and Tony Pridmore, Eds. 2014, BMVA Press.

[8] Mark Everingham, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman, “The pascal visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona, “Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories,” *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, apr 2007.

[10] Matthew D. Zeiler and Rob Fergus, “Visualizing and understanding convolutional networks,” *Computer VisionECCV 2014*, vol. 8689, pp. 818–833, 2014.

[11] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. jun 2014, pp. 580–587, IEEE.

[12] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller, “Recognizing Image Style,” in *Proceedings of the British Machine Vision Conference*, Tony Valstar, Michel and French, Andrew and Pridmore, Ed. 2014, BMVA Press.

[13] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, “Caffe: Convolutional Architecture for Fast Feature Embedding,” in *Proceedings of the ACM International Conference on Multimedia - MM ’14*, 2014, pp. 675–678.

[14] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi, and Andrew Zisserman, “The devil is in the details: an evaluation of recent feature encoding methods,” in *Procedings of the British Machine Vision Conference 2011*. 2011, number 1, pp. 76.1–76.12, British Machine Vision Association.

[15] David G. Lowe, “Object recognition from local scale-invariant features,” in *Proceedings of the Seventh IEEE International Conference on Computer Vision*. 1999, vol. 2, pp. 1150–1157 vol.2, IEEE.

[16] Andrea Vedaldi and Brian Fulkerson, “VLFeat: An open and portable library of computer vision algorithms,” in *Proceedings of the international conference on Multimedia*. 2010, pp. 1469–1472, ACM.