Deep learning for object detection in fine-art paintings

Stanislav Smirnov
Signal and System Theory Group
University of Paderborn
Germany
ssmirnov@campus.uni-paderborn.de

Alma Eguizabal
Signal and System Theory Group
University of Paderborn
Germany
alma.eguizabal@sst.upb.de

***Abstract*—We propose deep learning and neural networks to automatically detect objects in digital pictures of fine-art paintings. This automatic annotation of digitized artwork provides innovation for content analysis, and therefore enhances the process of documenting and managing cultural heritage. Deep neural networks have outperformed all previous machine learning techniques in computer vision and achieve the highest accuracy in object detection. However, a very big amount of labeled training samples are required for such good performance. Typically, this big data is collected from everyday natural images, which is possible because millions are generated each day. Unfortunately there are not such big datasets of digitized fine-art paintings. In this contribution we present a set of strategies to overcome the lack of labeled data and hence make use of the promising deep learning in this application.**

***Index Terms*—automatic annotation, deep learning, digitized fine-art paintings, object detection.**

I. Introduction

The generation of digitized fine-art collections has been growing fast recently [1]. These digital archives are challenging to manage but also are a potential source for documenting research and museums' narrative innovation [2]. An automatic annotation provides a powerful tool that saves time to art historians, for example, to find how an object evolved during an artistic period. Also, this automatic annotation can enhance virtual reality experiences in musems, as well as the access to museums' online data sources [2].

Deep learning (DL) is a machine learning method that allows learning features directly from data. Unlike the classic machine learning approaches that require some human guidance to design hand-crafted features, DL is capable of determining these features itself. With recent breakthroughs in computational power, DL methods have taken the computer vision community by storm, establishing a new state-of-the-art accuracy detecting objects, as well as in many other applications. Such an exceptional performance in DL would not be possible without the vast amount of training data available. The ImageNet challenge [3] has become instrumental in providing this data for the image classification task. Most of the datasets consist of natural images (i.e., everyday photos taken with a regular camera), which are labeled with objects and readily available, such as in [4] and [5]. Unfortunately, in many application fields it is very challenging to obtain a sufficient amount of labeled data. This introduces the main drawback and challenge to DL: to deal with a limited amount of annotated training samples. That is the case with respect to digitized fine-art paintings, where the datasets are limited and these are lacking annotations about the objects in them. Object detection is typically performed with DL using deep convolutional neural networks (CNN), which have contributed with a great improvement [6]. However, it is not straightforward to consider a CNN that was trained with natural images in a detection over artistic paintings, since these differ from the natural images in many low-level features, such as their texture statistics or color histograms. Even the painting images differ between them, since there exist many artistic styles in which they can be depicted.

A noticeable amount of DL approaches have been presented to classify style, artist or genre in artistic paintings, such as the techniques presented in [1] and [7]. Also, there are machine learning studies about the usability of brushstrokes to identify an artist, such as [8] and [9]. In [1], the authors looked for the most appropriate combination of visual features to achieve maximum accuracy in artist, style, and genre classification. With respect to DL and object detection in digitized fine-art painting, the authors in [10] have presented an approach with CNN and transfer learning to perform object retrieval in the paintings. However, in light of the results, their technique still has room for improvement.

In this paper, we propose a novel method for overcoming the difficulties of using DL with relatively few training samples. We perfom a dataset augmentation over the natural images using artistic style transfer [11]. Then, we make use of training images from paintings that are labeled for different classification tasks, such as style recognition, to train two paralel CNNs and fuse their output features in a support vector machines (SVM) classifier [12].

II. Datasets

Here we briefly present the datasets that we have used to train our proposed approaches.

A. PASCAL VOC

PASCAL VOC 2012 [5] is one of the most well-known publicly available datasets of natural images with object an-

(a) class *train* from Paintings dataset, in II-B

(b) class *train* from PASCAL VOC dataset

Fig. 1. Example class images from the Paintings dataset. Two images of class *train* are presented.

notations to be used in recognition and segmentation. The source of their images is flickr.com, and it contains more than $10,000$ samples, with significant variability in terms of object sizes, positions, illumination, orientation, etc. Some examples of object classes in this dataset are: aeroplane, boat, cow or dog. An example is shown in Fig. 1.

B. Painting dataset

This dataset of digitized fine-art paintings is presented in [10] and consists of $10,000$ annotated images with the PASCAL VOC labels. We show an example of class "train" in Fig. 1.

C. WikiArt

The WikiArt Paintings dataset [1] is one of the largest publicly available dataset of digitized paintings, and contains a collection of more than $80,000$ samples, from more than 1000 artists, 27 different styles, and 45 different genres. They do not have, unfortunately, annotations about objects.

III. Deep Learning

DL is a machine learning approach that considers a very non-linear structure to extract and learn features from the data. DL has recently achieved great success in different areas, such as computer vision [6], speech processing [13], and biomedicine [14].

Fig. 2. VGG-19 [16] architecture. It consists of 16 convolutional layers ("conv"), separated by max pooling layers ("pool"), and ending in 2 fully connected layers ("fc") . The first "conv" layer contains 64 filters of size 3×3; the last output layer is "fc" with 4096 nodes. Image source: https://devblogs.nvidia.com/,

We consider in this contribution deep neural networks, particulary CNN [15]. These are a class of deep, feed-forward neural networks that are specially designed for image-recognition applications. Since the input of the network is an image, their convolution-based architecture allows to encode certain properties. We specifically choose VGG-19 architecture (depicted in Fig. 2), which was trained using ImageNet dataset. This dataset contains more than one million natural images.

A. Transfer learning

The object recognition features learnt by a CNN can generalize along different datasets [17], [18], [19]. The transfer learning technique allows that a source CNN and a target CNN, whose domains are related [20], share these learnt features. Thus, the target CNN saves training time, or can even avoid the need of such big training data. The target CNN uses the features from the source CNN as as reference or a starting point during the training process.

B. Data augmentation

Data augmentation is a strategy to artifially increment the training dataset by applying certain transformation to the training images. The goal is that the resulting CNN using these images is invariant to these transformations [21]. Such transformations are generally geometric (affine, cropping) or photometric (relative to the RGB channels).

C. Evaluation metrics

We consider the average precission to evaluate the performance of the object detection. The average precision is the precision across all values of possible recalls. The mean averaged precision is the same metric averaged over all samples in the test dataset. We choose this metric to test our performance, since it is the one used in PASCAL VOC challenge and in the competing strategy in [10].

To evaluate intermediate steps we also consider the F1 scores and accuracy.

IV. Proposed solutions

We present the set of techniques that we have considered to design our approach.

A. Augment datasets with artistic style

PASCAL VOC dataset contains annotated natural images. The digitized painting datasets are, however, lacking of annotation and do not have enough samples. We propose to generate a new dataset out of PASCAL VOC after a style transformation

(augmented dataset). We use the strategy proposed in [11] to transfer artistic style to the natural images in PASCAL VOC dataset. This allows the creation of artificial artistic images with high perceptual quality. We show examples in Fig. 3. The goal is to multiply each image in the PASCAL VOC dataset by a selection of reference artistic styles. As a result, we obtain a set of images that contain the same original content (object), but not necessarily the same texture and style (e.g. in Fig. 3 (e) - (g)). The resulting dataset consists of 91650 new images, under eight painting styles: Abstract, Expressionism, Baroque, Cubism, Expressionism, Impressionism, Post Impressionism, and Romanticism.

The style transfer technique in [11] consists in a trade-off between a content and a style fit. A loss function is minimized (L), and is determined as a weighted sum of the loss due to the content fit, $L_{content}$, and the loss due to style fit, L_{style}, as

$$L_{total}(\mathbf{c}, \mathbf{b}, \mathbf{s}) = \alpha L_{content}(\mathbf{c}, \mathbf{b}) + \beta L_{style}(\mathbf{s}, \mathbf{b}), \quad (1)$$

where the vector $\mathbf{c}$ represents the content image, $\mathbf{s}$ the style image, and $\mathbf{b}$ the target image, which is initially blank. We illustrate this scheme in Fig. 4.

We propose $\frac{\alpha}{\beta} = 10^{-1}$, since we observed empirically that this combination of parameters leads to better results. If we consider more weight in the loss of the content ($\frac{\alpha}{\beta} > 10^{-1}$), the style may not contribute sufficiently in the augmentation. Whereas if we consider more weight in the style ($\frac{\alpha}{\beta} < 10^{-1}$), the contents may be blurred and the object detection accuracy compromised.

B. Fusion of features from two CNNs: object and style

We also propose a different alternative to deal with the lack of annotated objects in the training data. Our motivation is to test whether having additional information about a predicted style in a painting enhances the object detection accuracy.

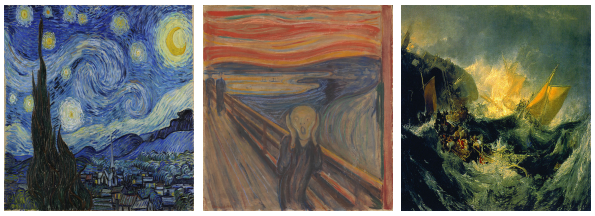
The WikiArt dataset contains annotations for style. We propose to use an additional CNN that is trained with WikiArt images to perform a style prediction [1]. Then, both CNNs in parallel (object and style detection) generate a collection of features.

We concatenate these features and consider it the input of a SVM classifier. We show a flowchart describing this approach in Fig. 5.

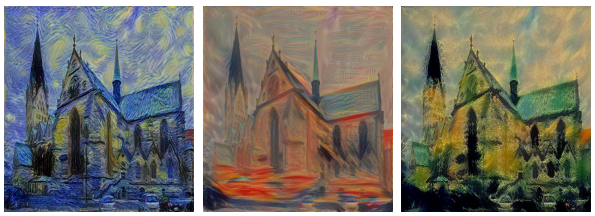
C. Bayesian hyperparameter optimization

Hyperparameter optimization is one of the main challenges during the training of these complex, multilayer models [22]. During the training process, the weights in the convolutional layers are set. However, there are many more hyperparameters of design that do not change during the training process and nevertheless make an important impact on the resulting accuracy. Some of these are the number of layers, the activation function or the number of hidden units. We propose a bayesian optimization strategy [23] to efficiently find a good combination of hyperparameters.

(a) Paderborn cathedral



(b) The starry night (c) The scream (d) The shipwreck of the minotaur



(e) Cathedral with style in (b) (f) Cathedral with style in (c) (g) Cathedral with style in (d)

Fig. 3. The examples of Style Transfer method, where (a) correspond to the content image, (b),(c),(d) are style images and (e),(f),(g) the resulting images. Style images were chosen according to [11].

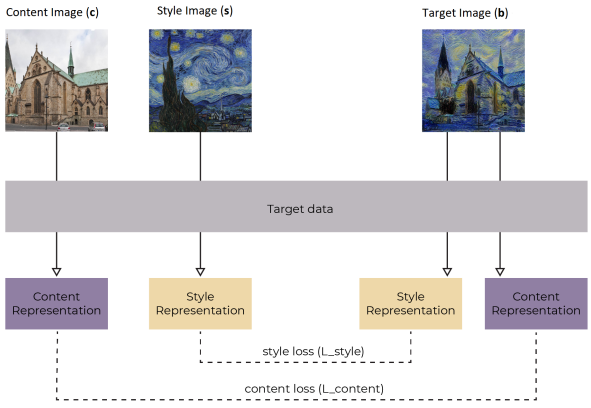


Fig. 4. Scheme of the style transfer strategy in [11].

V. Experiments and Results

We overcome the problem of lacking data by enhancing the use of the available images and labels. Firstly, we augment the dataset of natural images PASCAL VOC by transferring artistic styles, and use this augmented dataset to train a CNN

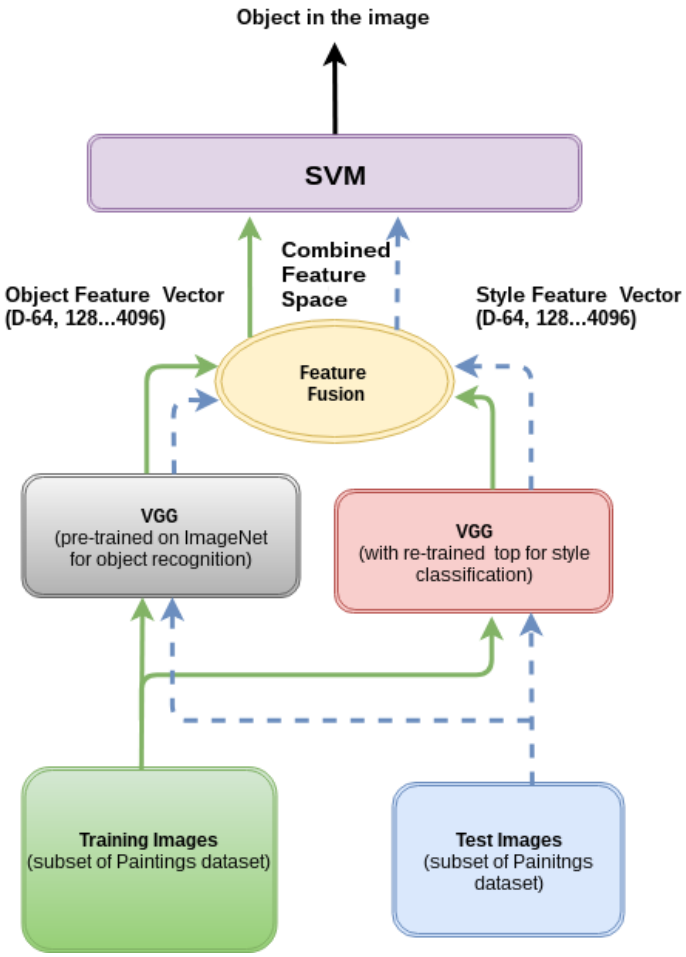


Fig. 5. Flowchart for proposed CNN fusion. Two paralell CNNs are trained to perform different tasks (object and style detection) and their output features are the input of an SVM classifier.

to detect objects. Secondly, we combine the output features obtained from it with the output features of a different CNN that has been trained with WikiArt to detect artistic styles.

Finally, we compare the obtained performance with the results in [10]. The authors in [10] also used a VGG architecture, but considered regular data augmentation transfer learning to deal with the lack of annotated data.

A. CNN trained with augmented dataset

In this step, we use the architecture VGG-19 together with Bayesian optimization and transfer learning. The transfer learning we propose consists in using the original VGG-19 features, except for the the last fully connected layer (output), where we train new weights. As a comparative baseline, we use the original PASCAL VOC dataset of natural images as training set. Then, we repeat the same procedure but using the proposed augmented dataset as training set. We perform the performance tests over the Paintings dataset. We compare the results in Table I, where we observe the average precision of the CNN when it is trained with natural images (natural) and with the proposed augmented dataset (augmented), for some

TABLE I
MEAN AVERAGE PRECISION FOR OBJECT DETECTION, TESTING IN PAINTINGS DATASET AND TRAINING IN AUGMENTED DATASET WITH TRANSFERRED STYLE (AUGMENTED) AND NATURAL TRAINING IMAGES (NATURAL).

data / object	plane	boat	cow	dog	horse	sheep	train
augmented	0.51	0.82	0.51	0.46	0.66	0.47	0.77
natural	0.43	0.74	0.39	0.39	0.57	0.32	0.61

TABLE II
AVERAGE PRECISSION OVER ALL CLASSES. WE COMPARE THE RESULTS IN [10] WITH OUR PROPOSED TECHNIQUE.

Strategy	**average precision (mAP)**
proposed, as in Fig. 5	0.58
strategy in [10]	0.52

of the object classes in PASCAL VOC. We see that there is a considerable improvement after the dataset augmentation.

B. CNN trained with WikiArt style labels

We now consider a CNN that is trained with WikiArt images to detect artistic style. We consider the arquitecture proposed in [1]. We also consider a new style, natural image, to account for the original images in PASCAL VOC, since these are also considered in the training set. We enhance the performance selecting the hyperparameters with Bayesian optimization. Our resulting accuracy is 64%. The styles with better performance are Baroque and Cubism (with 0.74 and 0.72 F1 score), whereas the least performance is observed in Post Impressionism and Expressionism (0.55 and 0.51 F1 score). Notice that this CNN is, however, only detecting artistic styles and not objects.

C. Enhancing detection with two CNNs fused with SVMs

Our goal is to detect objects in the paintings. We expect to gain detection performance considering a style detector too. In order to fuse the CNN to detect styles, described in V-B, with the CNN trained with the augmented dataset, described in V-A, we concatenate the output features from the last fully connected layers of both CNNs (object and style features). This is the Feature Fusion step in Fig. 5. This combination of features is the input of the SVM.

We train the SVM using the images in the Painting dataset, i.e., true digitized fine-art painting images. We perform a cross-validation test to evaluate the performance of the classification. The Painting dataset is divided in 5 sets of equal size. Then, four of these sets (that is, 80% of the images) are used to train the SVM, whereas the remaining set (20% of the images) is used to test it. The process is repeated five times, to account for every combination of train/test sets. The final performance is the average over the five cross-validation splits.

Out techique outperforms the competing strategy in [10], improving the performance by 5%. Table II shows the obtained results averaged over all object classes.

VI. Conclusions

Object detection in paintings is a challenging task. DL and CNN are very promising machine learning approaches to deal with this problem. However, they require very large labeled training datasets, which are not available for digitized fine-art paintings. We presented a technique that overcomes this problematic.

We proposed a training data augmentation based on transferring styles from representative artworks to natural images. To the best of our knowledge, this is the first paper where an artistic style is transferred to enhance object detection in digitized fine-art. We showed that this is a promising strategy of data augmentation. We also incorporated in the training a set of fine-art paintings that have style labels only, and not object labels. In order to make use of the style information we proposed a fusion of two CNNs (object + style) with a SVM. We showed that including the style information increases the performance of the overall object classification. Our proposed technique outperformed the competing strategy, providing a better average precision measured in a cross-validation test with fine-art painting images.

The proposed approach allows detecting objects on digitized paintings in real-time, which provides innovation in cultural heritage, opening new paths in the online museums' resources, and enhancing the cultural visits.

VII. Acknowledgment

The authors would like to thank Prof. Peter Schreier for providing the resources to develop this contribution. Also, the support of Aaron Pries, who gave technical guidance to the authors.

References

[1] B. Saleh and A. M. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," *CoRR*, vol. abs/1505.00855, 2015.

[2] L. Bordoni, F. Mele, and A. Sorgente, *Artificial Intelligence for Cultural Heritage*, Cambridge Scholars Publishing, 2016.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.r C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, Dec 2015.

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

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun 2010.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, p. 2012.

[7] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka, "Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification," in *2016 IEEE International Conference on Image Processing (ICIP)*, Sept 2016, pp. 3703–3707.

[8] C. R. Johnson, E. Hendriks, I. J. Berezhnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang, "Image processing for artist identification," *IEEE Signal Processing Magazine*, vol. 25, no. 4, pp. 37–48, July 2008.

[9] S.i Lyu, D. Rockmore, and H. Farid, "A digital technique for art authentication," *Proceedings of the National Academy of Sciences*, vol. 101, no. 49, pp. 17006–17010, 2004.

[10] E. J. Crowley and A. Zisserman, "In search of art," in *Workshop on Computer Vision for Art Analysis, ECCV*, 2014.

[11] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *CoRR*, vol. abs/1508.06576, 2015.

[12] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July 1998.

[13] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[14] A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vanhaelen, K. Khrabrov, and A. Zhavoronkov, "The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology," *Oncotarget*, vol. 8, no. 7, pp. 10883, 2017.

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[17] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 818–833, Springer International Publishing.

[18] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Washington, DC, USA, 2014, CVPRW '14, pp. 512–519, IEEE Computer Society.

[19] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1717–1724.

[20] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, Oct 2010.

[21] L. Taylor and G. Nitschke, "Improving deep learning using generic data augmentation," *CoRR*, vol. abs/1708.06020, 2017.

[22] A. Coates and A. Y. Ng, "The importance of encoding versus training with sparse coding and vector quantization," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, USA, 2011, ICML'11, pp. 921–928, Omnipress.

[23] E. Brochu, V. M. Cora, and N. de Freitas, "A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning," *CoRR*, vol. abs/1012.2599, 2010.