DeepArt: Learning Joint Representations of Visual Arts

Hui Mao
HKUST-NIE Social Media Lab
The Hong Kong University of Science and Technology
hmaoaa@connect.ust.hk

Ming Cheung
HKUST-NIE Social Media Lab
The Hong Kong University of Science and Technology
cpming@ust.hk

James She
HKUST-NIE Social Media Lab
The Hong Kong University of Science and Technology
eejames@ust.hk

ABSTRACT

This paper aims to generate a better representation of visual arts, which plays a key role in visual arts analysis works. Museums and galleries have a large number of artworks in the database, hiring art experts to do analysis works (e.g., classification, annotation) is time consuming and expensive and the analytic results are not stable because the results highly depend on the experiences of art experts. The problem of generating better representation of visual arts is of great interests to us because of its application potentials and interesting research challenges—both content information and each unique style information within one artwork should be summarized and learned when generating the representation. For example, by studying a vast number of artworks, art experts summary and enhance the knowledge of unique characteristics of each visual arts to do visual arts analytic works, it is non-trivial for computer. In this paper, we present a unified framework, called DeepArt, to learn joint representations that can simultaneously capture contents and style of visual arts. This framework learns unique characteristics of visual arts directly from a large-scale visual arts dataset, it is more flexible and accurate than traditional handcraft approaches. We also introduce Art500k, a large-scale visual arts dataset containing over 500,000 artworks, which are annotated with detailed labels of artist, art movement, genre, etc. Extensive empirical studies and evaluations are reported based on our framework and Art500k and all those reports demonstrate the superiority of our framework and usefulness of Art500k. A practical system for visual arts retrieval and annotation is implemented based on our framework and dataset. Code, data and system are publicly available[1].

KEYWORDS

Visual arts; DeepArt; image representations; deep learning; Art500k

1 INTRODUCTION

Visual arts have great values in terms of heritage, culture and history. Historians study human origin through cave frescoes, ordinary people can take a glimpse of the artists' lives by appreciating their works. With the development of the computer science and the explosive growth of digital copies of visual arts, the advanced computer science algorithms and large-scale digital copies of visual arts have opened both opportunities and challenges to computer science researchers and art community researchers. It is a very important interdisciplinary research field in that the computer science and the art community can boost the development of each other. On one hand, new art theories can be explored by the art community to provide computer science researchers with more theoretical support for algorithm design, and on the other hand, new automatic analyzing techniques and tools can be developed by computer science researchers to help the art community understanding visual arts further. In recent years, many museums and galleries have made their collections publicly available, people can browse, learn and buy visual arts on-line. In order to fulfill the requirements of on-line exhibitions or selling, heavy categorizing and indexing works need to be conducted by art experts. For example, WikiArt[2] featured all digital artworks by art movements (e.g., Abstract Art, Cubism.) or artists, however the visual arts are highly related to visual, it is not a good way to capture users' expectations using category based or keyword based search methods. We think that it is better to solve those problems based on the visual perceptions. In this paper, we propose an powerful framework to learn visual representations of visual arts to facilitate the different kinds of visual arts related

Figure 1: The main difference between visual art and nature image is shown: the contents of (a) visual art and (b) nature image is almost the same, but (a) contains the style that formed by Vincent van Gogh. For generating representations of visual arts, considering both the contents and style is essential.

[1] http://deepart.ece.ust.hk

[2] https://www.wikiart.org

MM '17, October 23–27, 2017, Mountain View, CA, USA

ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123405

applications, for example, visual arts retrieval and recommendation systems, visual arts analysis and annotation tools and forged artworks detecting tools.

In the art domain, both low level visual characteristics (e.g., color, texture) and high level visual characteristics (e.g., genre, content, brushwork) provide cues for visual arts analysis [3]. High level visual characteristics are usually constructed by multiple low level visual characteristics and it can convey more information from one visual art. From human perception, we consider two most important informations within visual arts for generating representations, one is the contents information, another is the style information. The contents within one visual art is very important for representing what things have been involved in one visual art, it can convey the main topic of the visual art to audiences. For example, if there is one person in the visual art, it is high probability to identify this visual art as one Portrait painting. The style concept is very abstract in visual arts, even those artworks that created by the same artist may have very different style, to explain it more concretely, style is something like *Brushwork* and *Strokes*, which is a characteristic way an artist creates the artwork [6]. Many impressionism masters formed their styles by special strokes [13]. The notion of style has long been the art historian's principal mode of classifying works of visual art, any piece of visual art is in theory capable of being analysed in terms of style; neither periods nor artists can avoid having a style and conversely natural objects or sights cannot be said to have a style, as style only results from choices made by a maker. To some extent, the style of one visual art is more important than the contents of one visual arts, because it defines the unique of one visual art and attaches emotions to visual arts. Figure 1 shows the main difference between visual arts and nature images, (a) is an image of artwork (*The Church at Auvers*) that created by Vincent van Gogh, (b) is a nature image of church at Auvers, the contents of two image is almost the same, but the image of artwork contains the style that formed by Vincent van Gogh. Because of adding the style, the image of artwork conveys totally different information than the nature image. This phenomenon indicates that both the contents and the style are important for visual arts analysis; Sometimes this two information can complement each other, for example, in different artworks, there may be many different styles showing the same contents, style sensitive representation s are more likely fulfill the humans' expectations.

In this paper, we present a unified framework, named DeepArt, to learn joint representations that can simultaneously capture contents and style of visual arts. Dual feature representation paths construct the whole framework and the outputs of the dual paths are linearly embedded as joint representations for different tasks of artworks analysis. In the framework, a VGG-16 architecture [31] is employed to capture the contents of visual arts, the performance with respect to classification and feature extraction for nature images of this architecture has been proved by many research works. Inspired by [14, 15], we profile the style of visual arts by adopting a Gram matrix to the filter responses in certain layers of the VGG-16 network. There are many optimization methods that can be used to conduct the learning process, for example, if using this framework to do classification, one softmax layer can be added to the top of the framework and using cross-entropy method to make the framework learn appropriate weights. We use a triplet-based deep ranking method [37] to learn these joint representations mainly for adapting the retrieval task of our system. A ranking loss defined on a set of triplet artwork samples is used as a surrogate loss to solve the optimization problem that is caused by non-smooth and multivariate ranking measures, and then the stochastic gradient descent (SGD) with momentum can be used to optimize model's parameters.

To evaluate our framework, implement the system and facilitate further research, we introduce the collected large-scale visual arts dataset (Art500k) that contains detailed labels (artist, genre, place, etc.). This dataset has three advantages over other digital artwork datasets. (1) The dataset has over 500,000 digital artworks with rich annotations, more than two times larger than the previous largest digital artworks dataset. (2) There are a wide variety of labels of digital artworks, apart from some general labels (e.g., artist, genre, art movement), and some special labels (e.g., event, historical figure, place) are included. (3) The dataset is well organized and can be accessed publicly under the use of research propose. Further researches related to visual arts will benefit from this dataset. One visual arts retrieval and annotation system is also implemented based on the presented framework and dataset. The experiment results show the superiority of the framework and the effectiveness of the dataset in terms of visual arts retrieval task and annotation task.

The **contributions** of this paper include the following.

(1) Inspired by the concepts of art theory and the characteristics of visual arts, a unified framework for learning the contents and style of visual arts is proposed. The joint representations are learned by a triplet-based deep ranking method, an efficient two-stage triplet sampling method is proposed for sampling triplets from visual arts dataset. This work offers insight into the possible connections between the art community and deep learning techniques.
(2) We build a large-scale visual arts dataset (Art500k) of over 500,000 digital artworks, which is richly annotated by detailed labels. To our best knowledge, it is the largest visual arts dataset for research.
(3) We adopt our framework to two important real-world visual arts analysis tasks: retrieval and annotation and through extensive experiments with our method as well as some other baselines, we demonstrate the superiority of our framework. A practical system is also implemented on the Internet, which can provide tools for art lovers and experts and some references for organizations that want to architecture similar systems.

2 RELATED WORK

In this section, we review some previous works related to digital artworks analysis, visual arts datasets, deep ranking learning methods and content-based retrieval systems. We choose important works rather than perform a board survey.

The past tendency is to use some computer vision methods (e.g., SIFT [22], GIST [26] and 2-D MHMMs [21]) to model some low-level features (e.g., color, texture), and those features can be used by machine learning methods (e.g., SVM [9]). For example, previous works [1, 4, 17, 18, 21, 24, 30, 33–35, 40, 41] utilized handcrafted

Table 1: The comparison of different visual arts datasets.

	PrintART [5]	Painting-91 [19]	Rijksmuseum [25]	VGG Paintings [10]	**Art500k**
# of Visual Arts	998	4,266	112,039	18,523	**554,198**
# of Big Classes	75	2	4	1	**10**
Contain Special Classes (Yes/No)	Yes	No	No	No	**Yes**
Public Availability (Yes/No)	No	No	Yes	Yes	**Yes**
Contain Eastern Artworks (Yes/No)	No	No	No	No	**Yes**

features according to some of those artistic concepts and appropriate machine learning methods to achieve automatic analysis for artworks annotation, retrieval and forgery detection. Despite the success of these works, the drawbacks are obvious: the handcrafted features are not flexible enough, it is also very hard to design a good handcrafted feature for certain task. In recent years, learning features from a large number of data has shown great promise by taking advantage of deep neural networks (DNNs). As convolutional neural networks (CNNs) [20] had great success on ImageNet Large Scale Visual Recognition Competition (ILSVRC) [12], some works [2, 10, 11, 27, 28, 36] used CNN-based methods to automatically find objects in artworks, identifying artists of artworks and performing artwork categorization. Promising results have been got using CNN, however most of them do not incorporate characteristics of visual arts as discussed in Section 1. Furthermore with the development of computer vision or statistics methods being used change fast. Proposing a unified framework associated with some state-of-the-art techniques is essential.

A comparison of previous datasets [5, 10, 19, 25] and Art500k is summarized in Table 1. In Table 1, Big classes means some class annotations like artist, genre, pose and composition, which also contains some subclasses; and Special classes refers to some uncommon Big classes annotations like composition, pose, event and historical figure. From the summary in Table 1, Art500k has obvious advantages over previous datasets.

Recently, there have been many feature learning methods investigated. Most of them [20, 31] require big data with labels and some explicit objectives (e.g., category-based classification). For artworks, there are so many big classes and overlaps between each class, the scope of visual variability within the same class is usually very large. Sometimes, we do not have very explicit objectives or cannot define appropriate objective functions, and more flexible learning methods should be used. Deep ranking models provide such a flexible way for learning some powerful representations. Previous works [8, 16, 37] about a Siamese-based ranking network and Triplet-based ranking network have achieved good performance on different tasks, especially retrieval and verification tasks.

Although content-based image retrieval (CBIR) systems for nature images databases have been studied extensively [32, 38, 39], however researches related to visual arts retrieval is not many. Traditional CBIR systems aim to locate some relevant images in a database according to a query concept, this query concept is explicit, for example, if the query image contains a dog, the system needs to find all images that containing dog. Visual arts retrieval focuses on another related but different topic, the criterion of results are more complex than traditional CBIR system, for example, if the query artwork contains a dog, the system not only needs to find all visual arts containing dogs, but also consider the style of these visual arts, because different styles can convey different emotions, it is not good to return results just based on the contents. Previous work [42] implemented an artwork retrieval system that fill the gap by considering users' preference profiles, however getting users' preference profiles is impossible. Our system is implemented based on the contents and style and can fill this gap and provide good experiences for users.

3 THE ART500K DATASET

We construct Art500k, a large visual arts dataset to facilitate the research in both art community and computer science community. Currently, the Art500k contains 554,198 images of visual arts, which is more than two times larger than the previous largest digital artworks dataset. The images of visual arts are labeled with a wide range of categories in the art domain, there are ten big classes in Art500k: origin, artist, art movement, genre, media, technique, school and three uncommon classes: event, place and history figure, some examples are shown in Figure 6. Some statistical results of top-10 classes in Artist, Genre, Medium, History Figure and Event are shown in Figure 7. From the comparison items summarized in Table 1, we see that Art500k surpasses the existing datasets in terms of scale, richness of annotations, as well as availability.

3.1 Data Collection

The images of visual arts were mainly scraped from four websites: WikiArt[3], Web Gallery of Art[4], Rijks Museum[5] and Google Arts & Culture[6]. All download images are low resolution copies of the original artwork and are unsuitable for commercial use and follow the copyright term[7]. Other small number of images of visual arts are collected from Google Search Engine. All websites are processed through a pipeline that download information through API or extracted relevant text from the raw HTML, downloaded linked images, and insert the data into MySQL database in which each datum was uniquely identified.

3.2 Data Preparation

After finishing collecting all data from the website, we conduct data cleaning, missing labels completing and easy accessing on the dataset. We encode all images of visual arts by the MD5 hashing

[3]https://www.wikiart.org
[4]http://www.wga.hu
[5]https://www.rijksmuseum.nl
[6]https://www.google.com
[7]The copyright term is based on authors' deaths according to U.S. Copyright Law, that is 70 years.

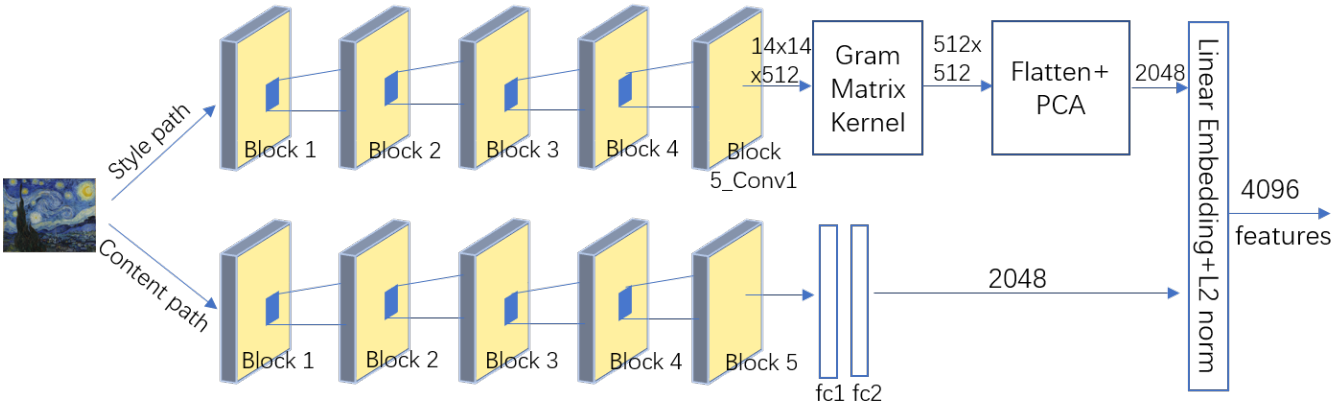

Figure 2: The architecture of DeepArt framework. It contains dual paths that can extract style feature and content feature respectively. The five convolutional blocks in the network are the same as VGG-16 [31]. Meaningful weights of the framework can be learned via appropriate learning methods. The number shown on the top of an arrow is the size of the output feature.

method, and remove digital artworks that have repeated MD5 hashing. The missing labels will be completed by program automatically. The logical is described as below: Firstly, the collected images usually have multiple labels (title, artist, art movement, etc.), if the title is not missing, we will search title in Google Search Engine, for example the *Impression, Sunrise*, then we can get the Wiki page[8] of visual art, we can complete other labels by extracting information from this page. If title is not available, we will search this image by Google Image Engine, then extracting fully-connected layer feature for the first image of the result to check if it has above 90% similarity with the query image, if they are the same image, utilizing the pages that we got from the results to complete other labels. To make the dataset easier to access, we format the data list to .csv files, .sql files, .txt files, etc.

4 DEEPART FRAMEWORK

Our goal is to design a unified framework that can learn joint representations containing both the contents and style of visual arts from a large number of digital artworks with multi-labels. The architecture of framework is shown in Figure 2, which is constructed by dual feature extraction paths. Each input digital artwork goes through the dual paths: the top network for extracting style information and the bottom network for extracting content information. The outputs of two paths are linearly embedded for generating the joint representations. This framework is very flexible, because the performance of joint representations can be improved with the development of computer vision or machine learning techniques, the content feature extraction method and style feature extraction method can be replaced by state-of-the-art methods easily.

4.1 Content Representation

The content representations presented here are generated on the basis of a VGG-16 network [31]. We directly use all five convolutional blocks of this network and change the dimension of two fully-connected layers from 4,096 to 2,048 (see the bottom path in

[8]https://en.wikipedia.org/wiki/Impression,_Sunrise

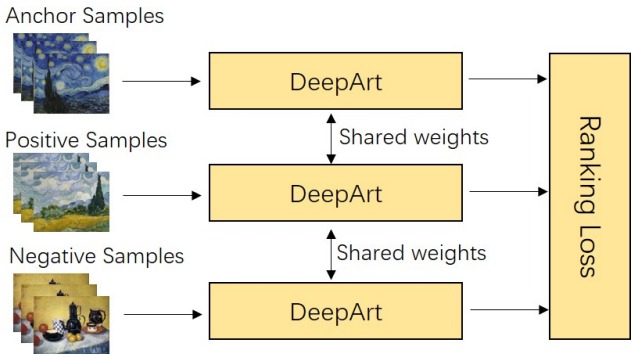

Figure 3: Triplet-based deep ranking model. It contains three same DeepArts (architecture is shown in Figure 2) that shared weights with each other. Given some triplet training samples, we get features from the last layer of DeepArt by forward propagation and compute the ranking loss.

Figure 2). This network was trained to perform classification and object recognition—more details can be found in work [31]. When the network is trained to do object recognition, a representation of the image that makes the object information increasingly explicit with the deepening of the hierarchy can be obtained. After the fully-connected layer has learned the non-linear combinations of features from previous layers, the output representations capture the high-level content information. Therefore, the input image will be encoded to a feature representation that is very sensitive to the content or object of the image. For the above reasons, we refer to the output of the second fully-connected layer as the content representation.

4.2 Style Representation

As discussed in last Section 4.1, the representation directly extracted by the deep neural network can capture content features well, but it cannot handle the style information of artworks, which plays a key

role in digital artworks analysis tasks. Inspired by [14, 15], a feature space that can represent the style of visual arts can be built on the top of the filter response in certain layers of CNN. Here, we adopt a Gram matrix kernel to filter maps of the first convolutional layer in the fifth convolutional block to construct a feature space (see the top path in Figure 2). The output of the Gram matrix contains the correlations between these filter maps, which can capture the style feature of visual arts well, but for nature images, the result is not significant. The Gram matrix $G^l \in R^{N_l \times N_l}$ in layer l is defined by Eq. (1), where F^l are the filter maps in layer l and F^l_{ik} is the value in the k^{th} position of the flattened i^{th} filter map.

$$G^l_{ij} = \sum_k F^l_{ik} F^l_{jk} \tag{1}$$

The output size of the Gram matrix is 512×512, which is a symmetrical matrix, therefore we choose the independent values from the symmetrical matrix and flatten the matrix to a vector. The dimension of this vector becomes 13, 1328 ($512 \times 513/2$), and then principal component analysis (PCA) is used to reduce the dimension of this vector. We tested some dimensions (e.g., 256, 512, 1024) of the vector, by the tradeoff between the representation performance and the computational efficiency, we choose the 2048-dimension output feature vector.

4.3 Joint Embedding

As it is expected, we get two output features from the dual paths that can represent the content and style of visual arts. The embedded methods can also be investigated in terms of unequal weights embedding, non-linearly embedding, etc. Here, we embedded the two output features equal weights and linearly, the final output joint representation is the normalized embedded feature. The output joint representations can be used to do different automatic digital artworks analysis tasks. The embedding feature visualization is shown in Figure 4, and we can find that those relevant digital artworks are close to each other.

5 LEARNING VIA VISUAL ARTS

We have introduced the architecture of DeepArt framework, given the reasons that learned features are better than handcrafted features, and explained the flexibility of using triplet-based deep ranking method to do learning. In this section, we will introduce how to adopt the triplet-based deep ranking method for our goal and the reliable two-stage triplet sampling method for visual arts dataset.

5.1 Triplet Network

The architecture of the triplet network is shown in Figure 3, which takes triplet samples as the input. The triplet samples contain one anchor sample, one positive sample and one negative sample, which are feed into three of the same subnetworks. The architecture is same as previous works [16, 37], we replace the subnetworks with the DeepArt architecture. The three subnetworks share the same weights and architectures. A ranking layer is built on the top of the three subnetworks, which is in charge of computing the ranking loss of the triplet samples. By forward pass, the distance between triplet samples can be evaluated, and by backward pass, the gradients are propagated to lower layers, where a lower layer can adjust

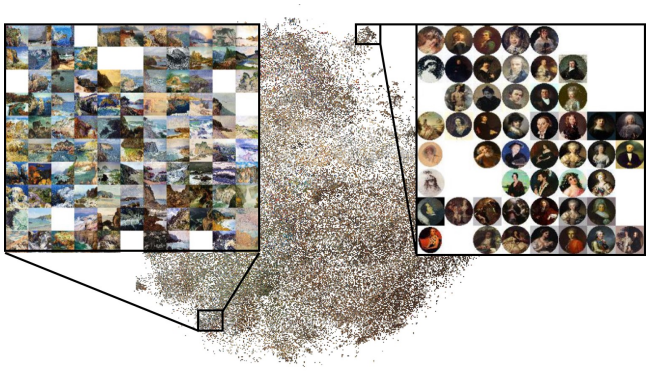

Figure 4: A t-SNE [23] visualization demo. We pick parts of images in the Art500k dataset and encode them by the DeepArt framework, then mapping the joint representations to two dimensions feature space by t-SNE.

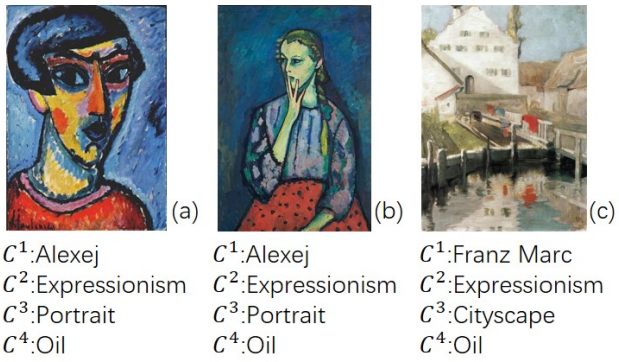

Figure 5: The illustration of calculating relevance. Based on Eq. (5) and the threshold the relevance of (a) and (b) is 4, the relevance of (a) and (c) is 2.

the weights to minimize the ranking loss. By using this kind of learning method, the weights in DeepArt can be learned, therefore the joint representations that are extracted from DeepArt can capture meaningful information.

5.2 Ranking Loss Optimization

For a set of visual arts $\mathcal{P}$, one anchor sample x^a_i, one positive sample x^p_i and one negative sample x^n_i can format a triplet sample $t_i = (x^a_i, x^p_i, x^n_i)$, where $x_i \in \mathcal{P}$. After the forward pass, those triplet samples are mapped by $f(.)$ to a new feature space. We use cosine similarity $D(.,.)$ as the distance in this new feature space. We want to train the network to find $f(.)$ that can make the distance between the anchor sample and negative sample much larger than the distance between the anchor sample and positive sample, which is defined in Eq. (2):

$$D(f(x^a_i), f(x^n_i)) > D(f(x^a_i), f(x^p_i)) \tag{2}$$

Figure 6: Examples of visual arts categories in the Art500k dataset.

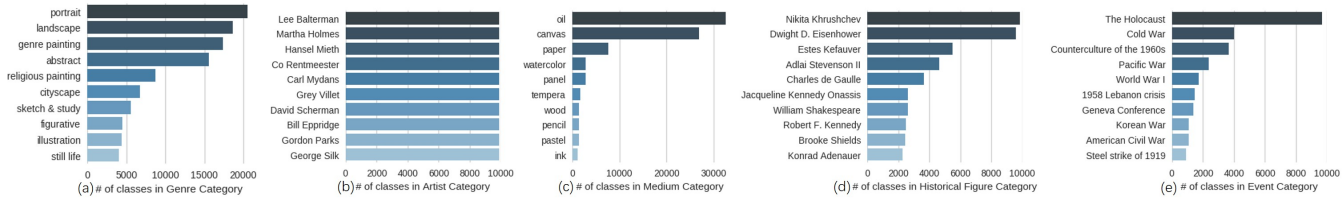

Figure 7: The number of top-10 classes in each category. (a), (b) and (c) are Genre, Artist and Medium respectively, which are very general categories in all datasets. The Art500K dataset also contains some uncommon categories (d) history figure, and (e) event that other datasets usually do not contain.

In order to achieve this goal, we employ hinge loss as the ranking loss function, which is defined in Eq. (3):

$$l(x_i^a, x_i^p, x_i^n) = \max\{0, m + D(f(x_i^a), f(x_i^p)) - D(f(x_i^a), f(x_i^n))\} \tag{3}$$

where m is a margin between two distances. To minimize this loss function, which is defined by Eq. (4):

$$\min_W \lambda \|W\|_2^2 + \sum_i \max\{0, m + D(f(x_i^a), f(x_i^p)) - D(f(x_i^a), f(x_i^n))\} \tag{4}$$

We can obtain the optimal weights W^* of $f(.)$, which also represents the subnetworks. In Eq. (4), to avoid overfitting, we add a L_2 regularization to the loss function. In the subnetworks, the dropout value is set to 0.5; in loss function weight decay λ is set to 5×10^4; margin m is set to 0.6; during the stochastic gradient descent process, the learning rate is set to 0.01; the decay is set to 10^{-6}; and the momentum is set to 0.9. The architecture of DeepArt and triplet-based deep ranking learning model are implemented in *Keras* [7].

5.3 Two-Stage Triplet Sampling

For training a triplet network, it is crucial to select triplet samples. The number of triplet samples increases cubically with the coming of digital artworks. For example, if you have two categories and each category contains 100 samples, you can still get 10^6 combinations. It is impossible to enumerate all triplet samples and some samples are not helpful for minimizing the loss, therefore a more efficient method should be employed in this task. In this work, we divide the process of sampling triplet samples from our large-scale visual arts dataset into two stages. The first stage is called fast sampling, as the triplet samples can be sampled very quickly based on the categories they belong to. We select four categories—Artist (1,000 classes), Art Movement (55 classes), Genre (42 classes) and Medium (112 classes) to evaluate the relevance of each digital artwork. For

example, artist class contains some labels like *Baade Knud*, *Baba Corneliu* and *van Gogh Vincent*; art movement class contains some labels like *impressionism*, *post impressionism* and *realism*; genre class contains some labels like *history painting*, *portrait* and *landscape*; medium class contains some labels like *oil*, *wash* and *pen*. We define the relevance r by the overlapping of the classes (C^1, C^2, C^3, C^4) of the four categories:

$$r_{i,j} = \{C_i^1, C_i^2, C_i^3, C_i^4\} \cap \{C_j^1, C_j^2, C_j^3, C_j^4\} \quad (5)$$

where i and j mean the i^{th} and j^{th} digital artworks. As illustrated in Figure 5, if $\sum r > 2$, two digital artworks will be regarded as relevent. The second stage is called hard sampling, the sampled relevant triplet samples will be evaluated by cosine similarity. Extracting fc7 features from those relevant samples by a pre-trained VGG-16 network, and then calculating the similarity between them. The top ten similar samples will be retained. After 20 epochs of training using samples that are sampled by stage one, samples generated in stage two will be used for the rest of the training.

6 EXPERIMENTS

In this section, we evaluate the proposed joint representations on Art500k, and the evaluation metrics, the competing methods, the competing architectures and the results that we obtained in both retrieval task and annotation task will be reported.

6.1 Evaluation Methods

6.1.1 Metrics. After finishing the two-stage triplet sampling, the relevant relationships between each digital artwork are determined according to the Eq. (5) and the threshold. For the retrieval task, we use cosine similarity to find the nearest neighborhoods of the query image, and then based on the nearest neighborhoods and the relevant relationships, we can calculate the Precision at rank k (Pre@k) and the normalized discounted cumulative gain (nDCG) to evaluate the performance of our method. Considering the real scenarios in life, when we are using the artworks retrieval application, we usually focus on the top 60 results. In this paper, the max k is set to 60. For annotation task, we compute the top-1 and top-3 classification accuracy by comparing the annotated label with the true annotations of digital artworks. We select 5 important categories to evaluate the annotation performance using the joint representations. The 5 categories are Origin (West/East), Art Movement (55 classes), Artist (1000 classes), Genre (42 classes) and Medium (112 classes).

6.1.2 Competing Methods. We compare the joint representations with hand-craft visual features and learned visual features. For each hand-crafted feature and learned features, we report its performance using its best experimental settings. The hand-craft features are Color visual features (generate using LAB histogram), GIST visual features [26], SIFT-like visual features [22] and SIFT-like Fisher visual features [29]. The learned features are these features that are extracted from the fully-connected layer of the convolutional neural network. In this paper, we select the VGG-16 with the ImageNet pre-trained weights as the baseline neural network, the output fully-connected layer feature is named fc7 (ImageNet). Then we use the triplet samples and the triplet training method to fine-tune the VGG-16 network based on the ImageNet pre-trained weights, the output fully-connected layer feature is named fc7 (Triplet).

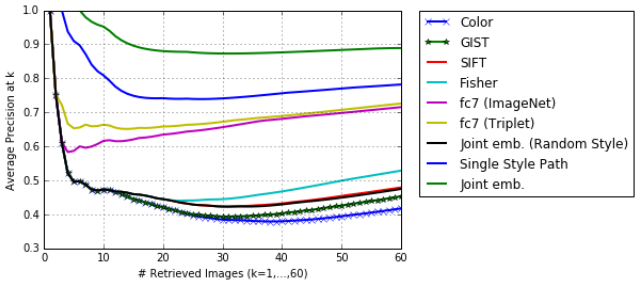

Figure 8: The average precision at k of different methods under comparison.

Table 2: The performance of different methods in a retrieval task

Methods	Pre@60 (%)	AP@60 (%)	nDCG
Color	56.7	41.7	0.778
GIST [26]	61.7	45.3	0.779
SIFT [22]	63.3	47.9	0.804
Fisher [29]	70.0	52.9	0.823
fc7 (ImageNet)	81.7	71.5	0.893
fc7 (Triplet)	83.3	72.6	0.901
Joint emb. (Random Style)	63.3	47.5	0.803
Single Style Path	85.0	78.2	0.933
Joint emb.	**90.0**	**88.9**	**0.970**

6.1.3 Competing Architectures. We compare two structures that all based on the VGG-16 network architecture. The first structure is that we just using the style feature extraction path in the DeepArt, not including the content features. We train this single style path network by triplet ranking method. The second architecture is same as the DeepArt architecture, but the weights of the style path are assigned randomly and set to non-trainable, we named it as Joint emb. (Random Style).

6.2 Retrieval Task

The Table 2 and Figure 8 summarizes the performance of different methods in a visual arts retrieval task. We can see that these hand-craft features without learning does not perform very well. For these learned visual features, the features are extracted from the network that have been trained using our triplet samples perform better than the features that are extracted from the ImageNet pre-trained network. We find that the architecture of single style path can perform better performance than traditional CNN architecture, it is because for visual arts retrieval, style features are more sensitive than the content features. For the Joint emb. (Random Style) architecture, we can find that the performance is not well, it means that the style feature path plays a key role in generating the joint representations. The joint representations that are extracted from the DeepArt perform the best in terms of average precision and the normalized discounted cumulative gain. It cannot only find

Table 3: The performance of different features in the annotation task

Methods	Origin	Art Movement		Genre		Artist		Medium	
	Top-1 (%)	Top-1 (%)	Top-3 (%)	Top-1 (%)	Top-3 (%)	Top-1 (%)	Top-3 (%)	Top-1 (%)	Top-3 (%)
Color	91.5	5.1	10.1	9.2	14.5	3.4	9.5	7.3	13.2
GIST [26]	82.3	10.5	16.5	12.4	19.2	6.7	10.5	11.3	19.6
SIFT [22]	90.1	10.7	22.1	15.2	19.6	9.6	13.4	12.5	21.3
Fisher [29]	93.2	20.1	25.5	21.3	30.1	14.6	19.3	27.4	35.5
fc7 (Triplet)	98.1	37.2	45.4	38.9	45.2	20.5	27.8	40.0	55.2
Joint emb.	**99.3**	**39.2**	**47.3**	**39.2**	**50.0**	**30.2**	**39.1**	**53.5**	**60.0**

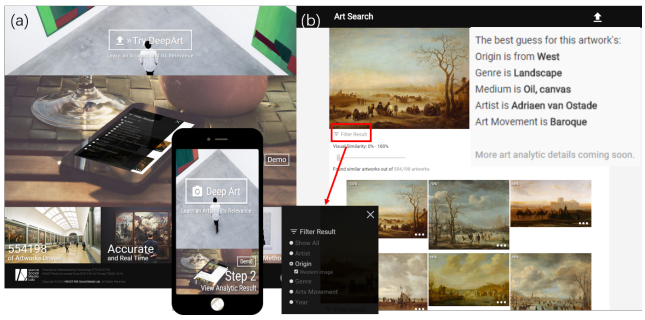

Figure 9: The illustration of the DeepArt retrieval and annotation system. Users can upload one digital artwork or take a photo via (a) and the results will be shown in (b).

more relevant results than other features but also rank the most relevant results in the front of the ranking list. This is because the generated joint representations consider both the content features and the style features of one visual art.

6.3 Annotation Task

The Table 3 summarizes the performance of different methods in a visual arts annotation task. As mentioned in Section 6.1.1, we select 5 categories to evaluate the artworks annotation performance. Firstly, we encode all images in each category using the joint representations, and secondly we train a multi-class SVM to predict the class labels of artworks. The performance of all these features are not very well, one reason is that we do not train the framework based on category-based objective function, we only use the network to extract the joint features. We can still find that the style feature provides extra gain for representing visual arts. One possible way to improve is to fine-tune the DeepArt based on the category-based objective function.

7 REAL-WORLD SYSTEM

We found that on the Internet there are so many content-based image search engines like Google Images[9], TinEye[10], but no art domain content-based search engine. Art related content-based search engine is useful because it does not have any language barriers, users just need to upload one visual art. It is more consistent for users to use, because the visual information is hard to describe using words. Museums, galleries and schools all need this kind of system to do exhibition, selling or education.

Based on the DeeArt framework and the Art500k dataset, a visual arts retrieval and annotation system, namely DeepArt Search is implemented on the Internet, which is shown in Figure 9. Users can upload the query visual arts via the web page or take a picture via their phones, as shown in Figure 9 (a), and then the uploaded image will be encoded by the DeepArt framework, the joint representations will be passed to the computing server. The joint representations are used to find the nearest neighborhoods and predict the categories of the query image on the server, then the results will be shown on the result page, from where users can get some useful information, as shown in Figure 9 (b), you can find that the annotation results of the query image are listed and users can filter the results by visual similarity or categories. The data preparation model is running on the server, duplicate images can be removed. We also train a Siamese network [8] using a large-scale of digital artworks and non-artworks images to verify if the collected or uploaded image is digital artworks, the performance is satisfactory.

8 CONCLUSION

This paper presents a unified framework, namely DeepArt, to learn joint representations that can simultaneously capture content feature and style feature of visual arts. One important advantage of this framework is that it is very flexible, the performance of joint representations can be improved with the development of computer vision or machine learning techniques. To boost the research in the art community and computer science, we contribute a large-scale visual arts dataset—Art500k with comprehensive annotations. By conducting different experiments on the large-scale visual arts dataset, we show the powerful of the joint representations and the usefulness of the Art500k dataset. We also propose an efficient two-stage triplet sampling method that enable us to learn the DeepArt framework from very large amount of training data. A real-world system based on the DeepArt framework and the Art500k dataset is implemented, providing art lovers and art experts with power tools, improving the online exploration experiences. This work also offers insight into the possible connection between the art domain and deep learning techniques.

ACKNOWLEDGMENTS

This work is supported by the HKUST-NIE Social Media Lab., HKUST. The author would like to thank Carmen Ng for the art domain knowledge consultation and Jing, Xizi for data preparation.

[9]https://images.google.com
[10]https://www.tineye.com

REFERENCES

[1] Patrice Abry, Herwig Wendt, and Stéphane Jaffard. 2013. When Van Gogh meets Mandelbrot: Multifractal classification of painting's texture. *Signal Processing* 93, 3 (2013), 554–572.

[2] Rao Muhammad Anwer, Fahad Shahbaz Khan, Joost van de Weijer, and Jorma Laaksonen. 2016. Combining holistic and part-based deep representations for computational painting categorization. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 339–342.

[3] Rudolf Arnheim. 1956. *Art and visual perception: A psychology of the creative eye*. Univ of California Press.

[4] Igor E Berezhnoy, Eric O Postma, and H Jaap van den Herik. 2005. Authentic: computerized brushstroke analysis. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 1586–1588.

[5] Gustavo Carneiro, Nuno Pinho da Silva, Alessio Del Bue, and João Paulo Costeira. 2012. Artistic image classification: An analysis on the printart database. In *European Conference on Computer Vision*. Springer, 143–157.

[6] CC Chen, Alberto Del Bimbo, Giuseppe Amato, Nozha Boujemaa, Patrick Bouthemy, Joseph Kittler, Ioannis Pitas, Arnold Smeulders, Kirk Alexander, Kevin Kiernan, et al. 2002. Report of the DELOS-NSF working group on digital imagery for significant cultural and historical materials. *DELOS-NSF Reports* (2002).

[7] François Chollet. 2015. Keras. (2015).

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 539–546.

[9] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

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

[11] E. J. Crowley and A. Zisserman. 2016. The Art of Detection. In *Workshop on Computer Vision for Art Analysis, ECCV*.

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

[13] Bert Dodson. 1990. *Keys to drawing*. North Light Books.

[14] Leon Gatys, Alexander S Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. In *Advances in Neural Information Processing Systems*. 262–270.

[15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. 2016. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2414–2423.

[16] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Springer, 84–92.

[17] James M Hughes, Daniel J Graham, and Daniel N Rockmore. 2010. Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proceedings of the National Academy of Sciences* 107, 4 (2010), 1279–1283.

[18] C Richard Johnson, Ella Hendriks, Igor J Berezhnoy, Eugene Brevdo, Shannon M Hughes, Ingrid Daubechies, Jia Li, Eric Postma, and James Z Wang. 2008. Image processing for artist identification. *IEEE Signal Processing Magazine* 25, 4 (2008).

[19] Fahad Shahbaz Khan, Shida Beigpour, Joost van de Weijer, and Michael Felsberg. 2014. Painting-91: a large scale database for computational painting categorization. *Machine vision and applications* 25, 6 (2014), 1385–1397.

[20] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation* 1, 4 (1989), 541–551.

[21] Jia Li and James Ze Wang. 2004. Studying digital imagery of ancient paintings by mixtures of stochastic models. *IEEE Transactions on Image Processing* 13, 3 (2004), 340–353.

[22] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, Vol. 2. Ieee, 1150–1157.

[23] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[24] Bangalore S Manjunath and Wei-Ying Ma. 1996. Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence* 18, 8 (1996), 837–842.

[25] Thomas Mensink and Jan Van Gemert. 2014. The rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of International Conference on Multimedia Retrieval*. ACM, 451.

[26] Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision* 42, 3 (2001), 145–175.

[27] Kuan-Chuan Peng and Tsuhan Chen. 2015. Cross-layer features in convolutional neural networks for generic classification tasks. In *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 3057–3061.

[28] Kuan-Chuan Peng and Tsuhan Chen. 2015. A framework of extracting multi-scale features using multiple convolutional neural networks. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*. IEEE, 1–6.

[29] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. 2010. Large-scale image retrieval with compressed fisher vectors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 3384–3391.

[30] Lior Shamir, Tomasz Macura, Nikita Orlov, D Mark Eckley, and Ilya G Goldberg. 2010. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Transactions on Applied Perception (TAP)* 7, 2 (2010), 8.

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

[32] John R Smith and Shih-Fu Chang. 1997. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the fourth ACM international conference on Multimedia*. ACM, 87–98.

[33] Robert P Taylor, R Guzman, TP Martin, GDR Hall, AP Micolich, D Jonas, BC Scannell, MS Fairbanks, and CA Marlow. 2007. Authenticating Pollock paintings using fractal geometry. *Pattern Recognition Letters* 28, 6 (2007), 695–702.

[34] H Jaap van den Herik and Eric O Postma. 2000. Discovering the visual signature of painters. In *Future Directions for Intelligent Systems and Information Sciences*. Springer, 129–147.

[35] LJ Van der Maaten and Eric O Postma. 2010. Texton-based analysis of paintings. *SPIE Optical Engineering+ Applications* (2010), 77980H–77980H.

[36] Nanne van Noord, Ella Hendriks, and Eric Postma. 2015. Toward Discovery of the Artist's Style: Learning to recognize artists by their artworks. *IEEE Signal Processing Magazine* 32, 4 (2015), 46–54.

[37] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1386–1393.

[38] James Z Wang, Kurt Grieb, Ya Zhang, Ching-chih Chen, Yixin Chen, and Jia Li. 2006. Machine annotation and retrieval for digital imagery of historical materials. *International Journal on Digital Libraries* 6, 1 (2006), 18–29.

[39] James Ze Wang, Jia Li, and Gio Wiederhold. 2001. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on pattern analysis and machine intelligence* 23, 9 (2001), 947–963.

[40] Marchenko Yelizaveta, Chua Tat-Seng, and Aristarkhova Irina. 2005. Analysis and retrieval of paintings using artistic color concepts. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 1246–1249.

[41] Marchenko Yelizaveta, Chua Tat-Seng, and Jain Ramesh. 2006. Semi-supervised annotation of brushwork in paintings domain using serial combinations of multiple experts. In *Proceedings of the 14th ACM international conference on Multimedia*. ACM, 529–538.

[42] Kai Yu, Wei-Ying Ma, Volker Tresp, Zhao Xu, Xiaofei He, HongJiang Zhang, and Hans-Peter Kriegel. 2003. Knowing a tree from the forest: art image retrieval using a society of profiles. In *Proceedings of the eleventh ACM international conference on Multimedia*. ACM, 622–631.