


Article

Leveraging Known Data for Missing Label Prediction in Cultural Heritage Context

Abdelhak Belhi ^{1,2,*} , Abdelaziz Bouras ¹ and Sebti Foufou ³¹ CSE, Qatar University, P.O. Box 2713, Doha, Qatar; abdelaziz.bouras@qu.edu.qa² DISP Laboratory, University Lumière Lyon 2, 69500 Lyon, France³ Le2i Lab, University of Burgundy, 21000 Dijon, France; sfoufou@u-bourgogne.fr

* Correspondence: abdelhak.belhi@qu.edu.qa; Tel.: +974-4403-6697

Received: 17 August 2018; Accepted: 27 September 2018; Published: 30 September 2018



Abstract: Cultural heritage represents a reliable medium for history and knowledge transfer. Cultural heritage assets are often exhibited in museums and heritage sites all over the world. However, many assets are poorly labeled, which decreases their historical value. If an asset's history is lost, its historical value is also lost. The classification and annotation of overlooked or incomplete cultural assets increase their historical value and allows the discovery of various types of historical links. In this paper, we tackle the challenge of automatically classifying and annotating cultural heritage assets using their visual features as well as the metadata available at hand. Traditional approaches mainly rely only on image data and machine-learning-based techniques to predict missing labels. Often, visual data are not the only information available at hand. In this paper, we present a novel multimodal classification approach for cultural heritage assets that relies on a multitask neural network where a convolutional neural network (CNN) is designed for visual feature learning and a regular neural network is used for textual feature learning. These networks are merged and trained using a shared loss. The combined networks rely on both image and textual features to achieve better asset classification. Initial tests related to painting assets showed that our approach performs better than traditional CNNs that only rely on images as input.

Keywords: cultural heritage; convolutional neural networks; multimodal classification; digital heritage; digital preservation

1. Introduction

Cultural heritage is the most effective medium for the transfer of historical information between generations and civilizations. Cultural heritage assets are distinguished by their variety and importance. These items are generally priceless as they represent great moral values. Unfortunately, cultural heritage assets face various risks related to physical degradation and information loss. Therefore, many researchers all over the world are placing much effort into finding reliable approaches to increase the value of these assets through preservation and curation using digital tools [1]. Today, new technologies provide cost-effective yet reliable techniques for the documentation and management of cultural heritage. However, several challenges have arisen regarding the curation and completion of missing information for these assets. The classification and annotation of cultural assets is a tedious and labor-intensive task that requires the involvement of highly qualified and experienced art specialists. This is mainly due to the nature and the specificities of cultural assets, as they usually come from various locations, old civilizations, or else their level of degradation simply prevents accurate annotation. To increase the value of their collections, heritage institutions are currently funding numerous research efforts to develop innovative methods for the completion and annotation of cultural data.

Recent advances in data science and machine intelligence are being applied in the cultural context. Many researchers are developing tools that can classify and annotate cultural assets based only on their visual features. Cultural assets are generally partially annotated with textual labels that may be used as additional information for their classification. To our knowledge, leveraging this additional information has not yet been used for the classification of cultural data.

In this paper, we propose a novel approach for the classification of cultural content. Our approach is multimodal, as, in addition to asset visual features, we leverage the available asset metadata. The experimental results for a painting dataset show that leveraging additional data in the classification improves the classification performance.

In the following sections, we present the impact of digital technologies on cultural heritage, and also give a literature review of the work related to cultural data classification and annotation using computer vision and machine learning.

1.1. Art and Culture in the Digital Era

Today, cultural heritage collections are enriched using digital technologies. Technologies such as 2D and 3D capturing, along with their respective visualization tools, have introduced new methods of content consumption and broadcasting. Digital heritage is widely used not only for entertainment and historical transfer, but also for long-term digital preservation and data analytics [2–4]. This is mainly due to the increasing reliability and falling costs of IT systems. As a result, cultural assets are now more accessible for larger audiences than ever before. However, one of the limiting factors is the quality and effectiveness of cultural asset digitalization. An asset with unavailable or incomplete metadata is automatically devalued. Consequently, the research community is focusing on how to leverage recent advances in data science and machine learning to promote and increase the value of cultural assets through an automatic and effective labeling process [4]. Semantic web technologies have been used in the past for cultural knowledge management by linking cultural assets using semantic relations and then inferring missing data using the acquired knowledge [5,6]. However, the use of these techniques is unfortunately limited to certain contexts. The exploitation of 2D or 3D acquisition and visualization technologies by heritage institutions has resulted in big data collections of cultural assets that have presented numerous challenges [4,7–9]. In this paper, our aim is to address the challenge of how to efficiently mine large data collections to effectively label and annotate overlooked cultural assets in order to increase their value.

1.2. Cultural Heritage Annotation and Classification

To raise the value of a cultural asset and boost its social and cultural impact, heritage organizations and researchers are combining efforts to develop methods that can overcome the problems of cultural asset mis-annotation and the associated lack of metadata. The studied approaches are mainly related to big data analytics for which the main goal is to reach a satisfactory accuracy of annotation [10,11].

Much research has been dedicated to the classification and annotation of visual cultural content [8,10,12–15], in which researchers generally try to analyze visual features in the digital 2D capture of a cultural artifact and then leverage these features for the classification of the asset. Usually, the classification is performed using a certain class of supervised machine learning models to categorize the data. There are two main types of approaches distinguished by the manner by which they capture visual features from the data samples. The first type is based on feature extraction techniques, where a feature extractor is used to identify a certain type of known (generally low level) visual features from the data. These features are then used to categorize the data sample at hand (high-level classification). The second type of approach does not require any prior feature extraction; the features are learned directly from the data.

The rapid progress of machine learning through deep learning and image processing has positively impacted cultural heritage [8,12]. The proven performance of these techniques is making the visual recognition of cultural heritage reliable and cost-effective [10,15,16] and many contributions have

been dedicated to the classification and annotation of cultural heritage through visual recognition. Traditional approaches rely on visual feature extractors and other specific features associated with cultural heritage, such as brush strokes in the case of paintings. Unfortunately, even if these techniques work well in some scenarios, their generalization to other domains or to new data samples is limited and very complicated. However, since the rise of deep learning techniques, and with their proven performance, the focus of the computer vision community regarding image recognition techniques moved from conventional techniques (feature extraction) towards ones based on deep learning (feature learning), such as convolutional neural networks (CNNs) [17]. The benefit of these new techniques compared to traditional methods is their very high performance in the context of big data, as well as their capacity for effective generalization [17]. In the following, we present a short review of the methods frequently used among researchers for visual feature extraction and feature learning for cultural heritage images.

1.2.1. Feature Extraction Approaches

In the past, visual recognition and image classification were performed using feature extraction approaches. These approaches are designed to examine the low-level features of an image by analyzing its pixel structure. The approaches used to extract these features include SURE, SIFT, HoG, GIST, Color Features, etc. Once extracted, the low-level features are used as inputs for supervised or unsupervised machine learning models such as SVM, KMeans, Random Forests, etc. The resulting models are then used to classify or cluster the images [16,18–20]. These methods perform very well in some applications and scenarios, but their performance is limited in general classification tasks and big data environments where deep-learning-based approaches outperform feature extraction approaches [21].

Using feature extraction techniques, many methods have been proposed to tackle the challenge of cultural heritage classification. The authors of [11] published a dataset collected from the Rijksmuseum of Amsterdam consisting of more than 100 thousand artifacts. They also provided baseline results of a four-task classification challenge using SIFT features and Fisher vector encoding. The authors of [10] combined HoG, SIFT, GIST and Color Features to classify painting genres, while the authors of [22] proposed a ranking-based method to classify paintings using Random Forests. The authors of [23] presented a histogram-based combination of local and global features for the classification of paintings. Several other researchers have attempted to combine feature extraction and feature learning. A comprehensive study investigating the correlations between feature maps for image style identification is presented in [24], and the authors of [15] combined GIST and CNN features for the classification of paintings. Additionally, in [25], the authors presented a hybrid method dedicated for the identification of copyrighted paintings in TV shows and movies. The technique uses deep learning to identify objects and a local feature detector to identify paintings. Furthermore, in [26], a study comparing CNNs to feature extraction methods clearly shows that CNNs have better performance than handcrafted features in the classification of cultural artwork. Nevertheless, feature extraction methods still perform well in many scenarios related to cultural heritage such as the identification of forged paintings [27].

1.2.2. Feature Learning Approaches

The majority of feature learning approaches related to the visual identification of cultural heritage assets are based on deep learning. CNNs, which represent the state-of-the-art in visual recognition, are used to categorize and annotate images. These approaches are more straightforward than those based on feature extraction as they do not require any prior feature extraction step. Instead, the developed models learn to identify patterns and features directly from the data. However, deep learning approaches often suffer from overfitting, where the model is unable to accurately classify samples that were not used in the training stage. Several deep-learning-based approaches have tackled the challenge of cultural heritage classification and annotation. The most remarkable contribution was made by PigeoNET [28], as it delivered a reliable confirmation that deep learning approaches

outperform handcrafted feature-based approaches for the categorization of paintings. Indeed, the results of this contribution surpassed the baseline results of the Rijksmuseum challenge set in [11]. The authors of [29] studied the use of multiple image patches to improve the accuracy of CNNs for the classification of paintings using transfer learning. In [30], the authors tried to use both painting images and brush stroke information for the identification of paintings using two parallel deep residual networks. The approaches in [31,32] explore Siamese convolutional networks for painting matching and retrieval in large databases. Both solutions can be used to capture visual links between assets and thus cluster assets sharing the same visual features. A large cultural dataset, called OmniArt, consisting of more than half a million assets in addition to a multitask classification approach, is published in [33]. Architectural heritage classification and annotation are addressed in [8,13].

Overall, deep learning approaches are effective, however there are still many challenges related to the effectiveness of model training, such as the need for powerful computing platforms, the time required to train networks and the risk of overfitting. In comparison, feature extraction approaches have lesser requirements in terms of computational resource requirements [21].

1.3. Our Contribution

In this paper, we present a novel approach based on designing and implementing a new multimodal classifier for cultural data that takes both images and textual labels as input. The aim is to leverage the available information related to an asset along with its visual capture when performing a classification. The majority of approaches found in the literature address this problem from a single point of view where a classifier that relies on the visual features of a cultural asset is used for its categorization. To our knowledge, our approach is the only one to study a real-world scenario where a small set of metadata is used as additional input for classification. The use of multimodal classification and multiple outputs is set to boost the performance of the proposed model, as it has been designed to learn more correlations between the input and output labels.

The remainder of this paper is organized as follows. In Section 2, we present the different techniques and tools used to design and implement our approach. These include the various datasets, the machine learning tools such as CNNs and concepts such as transfer learning. In this section, we also describe the architecture of our framework. In Section 3, we depict some of the experimental results using a designed experimental setup with data relating to paintings and compare our multimodal input architecture with single input traditional CNNs. In Section 4, we discuss the results and interpret the pros and cons of our approach. Finally, in Section 5, we draw our conclusion and give some perspectives for future work.

2. Materials and Methods

In this section, we outline the general concept and the tools used to design and implement our multimodal classification approach. This includes a presentation of the datasets collected and used within this work, followed by a description of the design and implementation of the technical approach along with the different key concepts related to deep learning.

2.1. Data Collection and Pre-Processing

Several cultural datasets were collected from various heritage institutions. In the following, we present those datasets, which were used to design, test, tune and validate our approach. The pre-processing step, used before feeding the data to the model, is also described, outlining the encoding in addition to filtering techniques. More information related to the datasets discussed in this section are provided in the Supplementary Materials section.

2.1.1. Data Collection

- The WikiArt Dataset

This dataset is a collection of more than 140,000 paintings which can be accessed from the Wikiart.org website. However, the website does not provide an easy means, such as a public API, to perform data collection. We designed a custom script based on the Python library *beautifulsoup* [34] to capture the most relevant data fields directly from the site's webpages. The data were then inserted into a MySQL database for further filtering while preserving the integrity of the original data scheme. The data fields selected were the artist, the media, the genre, the year and the style. Some samples from the dataset are shown in Figure 1.



Figure 1. Examples of paintings from the WikiArt dataset.

- The Metropolitan Museum of New York (MET) Dataset

The Metropolitan Museum of New York (MET) recently published a collection of more than 200,000 cultural artifacts under the Creative Commons open access license. The data represent images of assets along with their metadata, however the metadata are not fully available for the majority of the assets. The collection is accessible from a CSV file, nevertheless, the downloading of the images required the writing of a custom script to harvest them from the asset webpages on the museum website. The script was also based on the Python library *beautifulsoup* [34]. Some samples from the dataset are shown in Figure 2.



Figure 2. Some artworks from the Metropolitan Museum of New York (MET) dataset.

- The Rijksmuseum Dataset

More than 100,000 assets have been published by the Rijksmuseum of Amsterdam (which is often referred to as the Rembrandt museum, as it hosts a large collection of the artist's works). The collection is accessible via an API and has been the subject of multiple research contributions [11,33]. As a result, many researchers have hosted the collection on archival platforms. However, we noted that the museum had updated its collection with multiple new assets, and thus we relied on the museum's API to collect and organize the harvested assets. In this work, we only focused on paintings. Some samples from the dataset are shown in Figure 3.



Figure 3. Some artworks from the Rijksmuseum dataset.

2.1.2. Data Pre-Processing

The data pre-processing step is very critical for our approach. For this phase, we discarded the non-relevant samples in the aforementioned datasets. Unfortunately, the required labels for the data we collected were incomplete. We found that paintings were the most complete and fully annotated category, especially the samples collected from Wikiart.org. The most relevant labels we selected for the paintings were the artist name, the year of creation, the genre, the style and the medium. We selected a total of 43,594 paintings annotated for all of these labels. To fully evaluate the performance of our approach, we created four datasets regarding the number of paintings per artist, as we found that the artist attribution is the most relevant task for painting categorization (see Table 1). Each of the new sub-datasets retains only paintings of the artists that have a higher artwork count than the dataset threshold. For example, the >50 dataset will only host paintings by artists that painted more than 50 paintings. This was intended to evaluate the impact of the number of paintings per artist on the results, as often artists do not change style or genre, and their paintings are only created within a specific time frame. This can help the model in learning correlations between labels.

Table 1. Selection of the painting datasets used in this study.

Number of Samples Per Artist	Data Samples Included	Number of Artists Involved
all	43,594	917
>50	34,759	227
>100	26,141	107
>300	11,166	19

The five textual labels were encoded using One Hot Encoding and the year label was divided into bins of 10 years. Each entry in the datasets consists of a painting identified by its content ID, which points to the image file and its five annotations. At the training stage, the media, the style and the genre are fed to the model along with the asset image. The artist and the year are the output of the model. Table 2 represents a dataset sample of five entries.

Table 2. Dataset sample.

Content ID	Artist	Creation Year	Media	Genre	Style
323653	Anker Albert	1874	Oil, Canvas	Portrait	Realism
452365	Bierstadt Albert	1858	Oil, Canvas	Landscape	Luminism
179324	Picasso Pablo	1944	Oil, Canvas	Still life	Surrealism
197632	Sisley Alfred	1897	Pastel	Landscape	Impressionism
127801	Leonardo da Vinci	1504	Oil, Panel	Portrait	High Renaissance

At the training stage, for each item in the dataset, the data are split into two sets: The input set X , which consists of the visual capture, the genre, the style and the medium; and the output set Y , which consists of the artist name and the creation year.

For visual data, we chose to resize the images to a minimum of 300 pixels either in length or width (preserving the aspect ratio). This was mainly done to increase the performance of the online data augmentation while training the models. The main purpose of data augmentation is to increase

the validation and generalization performance of the trained CNN by avoiding the phenomenon of overfitting, where the network becomes unable to accurately classify samples not used in the training phase. The data augmentation generally extends the training dataset with an augmented dataset composed of the same images with distortions and adjustments such as flips, zoom, rotations, whitening, contrast and lighting adjustments, etc. This new augmented set will help the network avoid overfitting by training the network with slightly modified samples at each iteration. In our experiments, we initially tested data augmentation with the single input network to evaluate its impact (comparing the training of normal images with augmented images).

We found that, other than the techniques we used (horizontal flips and random crops), any data augmentation will decrease and deteriorate the generalization capability of the network, as generally the visual features found in paintings are the same (see Figure 4). Making distortions or other changes in the images will force the network to learn non-real samples (generally we only feed the painting to the network in its normal orientation, etc.).

Figure 5 represents the distribution of the number of paintings per artist, for artists that painted more than 300 paintings.



Figure 4. Random crops collected from images.

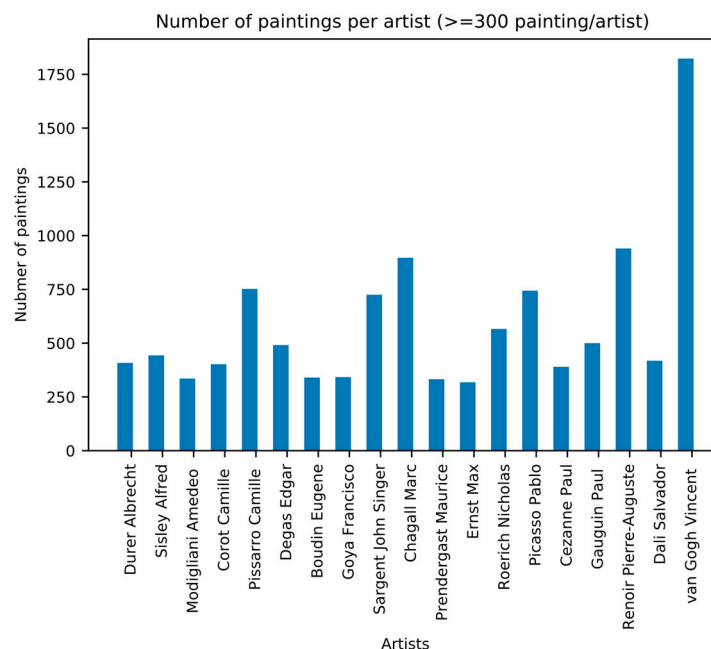


Figure 5. Distribution of paintings (number of paintings per artist).

2.2. Convolutional Neural Networks

Deep learning represents a collection of algorithms used in machine learning to model high-level abstraction or functions. Its architecture consists of several levels, where the initial levels capture low-level features or characteristics and the top levels capture high-level features resulting from the lower-level features in a hierarchical way [8]. Recently, different deep neural network structures

have been proposed such as CNNs, recurrent neural networks (RNNs), long short-term memory (LSTM), etc.

CNNs are an enhanced version of the multi-layer perceptron model which were mainly intended for image processing. Today, CNNs are used in many other applications such as natural language processing (NLP) [35]. They also represent the state-of-the-art in image recognition tasks thanks to their superior performance, as they introduce a new layer called the convolutional layer. This layer plays a role in optimizing the detection of local visual features. Each convolutional unit in a layer is responsible for detecting a certain feature, and many convolutional units can be used in the same layer. To reduce computational cost, a pooling layer is introduced in order to reduce the size of the feature set learned by the convolutional layer. For classification tasks, fully connected layers are attached to the output of convolutional layers in order to aggregate the features learned and to perform the classification. Figure 6 represents a basic CNN with five convolutional layers and three fully connected dense layers. CNNs are trained with the same backpropagation algorithm used in multilayer perceptrons. The main drawbacks of CNNs are their slow convergence and high computational complexity.

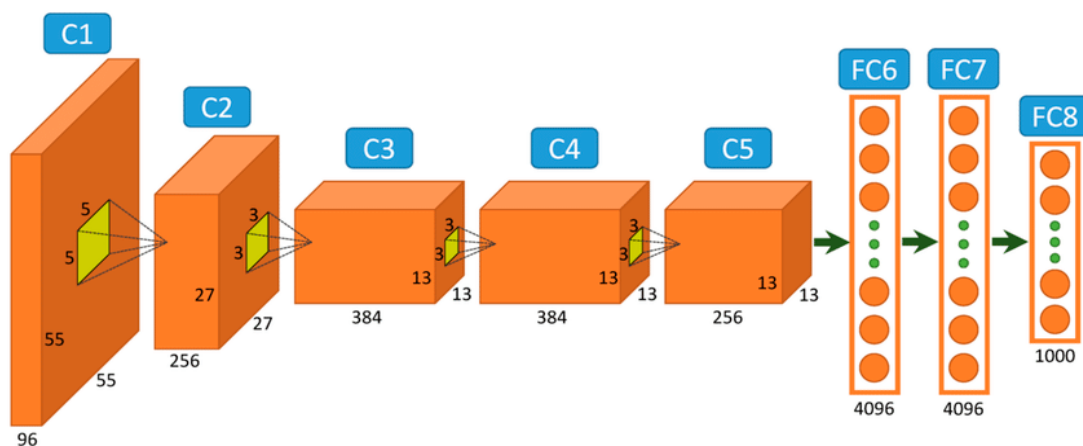


Figure 6. A Convolutional Neural Network (CNN) architecture.

2.3. Transfer Learning

In machine learning, transfer learning is the process of transferring knowledge from a model that performs well at a certain task to a new model used for a different task. Generally, the model from which transfer occurs is trained and tuned in a high-performance computing environment. However, due to the lack of resources, similar performance cannot be achieved using consumer-level hardware. Additionally, the models from which transfer occurs are generally trained on millions of samples over a long period of time. Transfer learning has been proven to be an efficient technique that helps newer models to generalize faster by benefiting from previously acquired knowledge [36].

For CNNs, there are multiple ways to reuse features or transfer knowledge from a trained network to a new model. The most widely used techniques are:

1. **CNN feature extraction:** A fully trained CNN can be repurposed and treated as a feature extractor by removing its output layer. For a network such as the VGG16 [37], the first dense layer of 4096 units can be seen as a CNN features vector. A lightweight linear classifier can be used such as a Support Vector Machine (SVM) that takes this CNN features vector as input and outputs the data classes. In most cases, CNN features can provide a fast way of implementing an accurate image classifier with few data samples on low-end hardware.
2. **Fine-tuning:** Fine-tuning consists of retraining or replacing the classification layer of a CNN (final layer). There are multiple fine-tuning approaches. The first is straightforward, and consists of removing the last layer, freezing the weights of the already trained network and training the last classification layer. The second approach consists of removing the last classification dense

layer, connecting a new output layer and then training the whole network with a smaller learning rate; this helps the model to converge faster as it benefits from the weight initialization of the previously trained network. The third approach consists of freezing only the weights of the first few convolutional layers and then training the remaining layers. Previous work [38] has shown that the first layers perform very accurate universal low-level feature extraction. As a result, transferring these features from a well-trained network is beneficial.

In this work, we used the ResNet50 network [39], which is a residual network with fine-tuning of all its layers. We removed the last dense layers and used average pooling to compress the features of the last convolutional layer. The output layer of the network consists of 2048 ReLU (Rectified Linear Unit) activated units.

In the following section, we present an overview of residual networks and justify their selection in our approach.

2.4. Residual Networks

Residual networks were first introduced by Kaiming He et al. [39]. Theoretically, neural network depth is important, especially for image recognition tasks. However, once reaching a certain threshold for the network depth, adding more layers will lead to problems, such as vanishing gradients, where the network becomes unable to learn from the data and gets saturated. Additionally, the computational complexity associated with optimizing many network parameters makes the training of the network much more difficult. Residual networks provide a new way of connecting neurons using skip connections. These connections are added every two or more layers. Figure 7 shows the normal CNN connections (left) and residual networks or “skip connections” (right). The principal is as follows: Let x be the input and $H(x)$ be the “underlying mapping” representing the function that fits two or more layers (two in the present study). Instead of $H(x)$, we try to approximate a function $F(x) = H(x) - x$. Before computing the output of the second layer, x is added to $F(x)$ and passed to the ReLU activation. We obtain $H(x)$, which is composed of x (the input). This procedure can transfer important information (x) from any layer to the following layers and prevent the vanishing gradient problem [39].

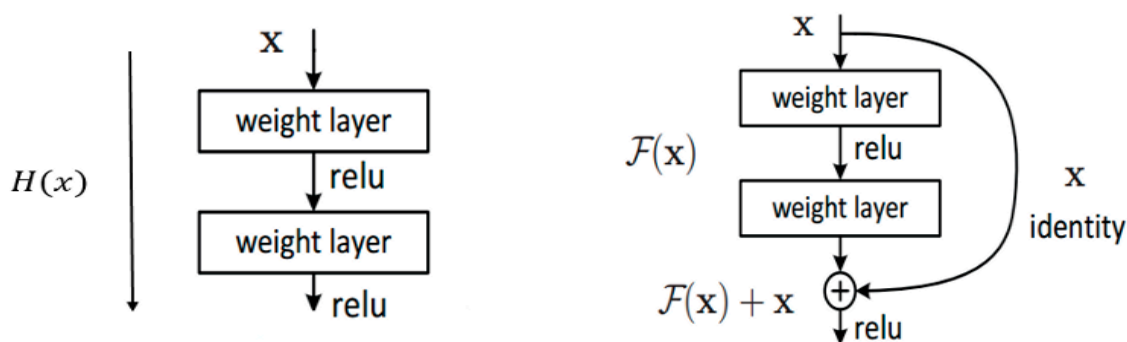


Figure 7. Normal CNN blocks and residual blocks.

Following some preliminary tests of several pretrained CNNs on the ImageNet challenge, such as the VGG16 and the VGG19 [37], we found that ResNet50 has a small advantage in terms of training accuracy and offers a considerable improvement in validation accuracy compared to normal CNNs. This was the main motivation behind the use of residual networks in this work.

2.5. The Proposed Solution

Multimodal classification is a classification technique that requires at least two different data types to represent a cultural asset in order to better perform the sample classification. In cultural heritage, and according to several heritage institutions, the data at hand regarding an “incomplete” asset are not limited to the visual copies (images) of objects. Other information can easily be found along with

the incompletely annotated asset consisting of a small set of metadata. It would be interesting to leverage these metadata for the identification and prediction of the missing data. To our knowledge, the approaches found in the literature do not investigate the impact of adding the available metadata as additional input to the classification models proposed [8,10,12–14,23]. They rather focus on using a 2D digital image of the asset and either apply feature extraction or feature learning approaches to leverage visual features of the asset for its classification.

In this work, we propose a novel approach focused on feeding the model with all of the available data to hand in order to predict the missing content of other data fields. Our approach is based on convolutional neural networks used for visual feature learning. The features learned from the asset image are subsequently merged with textual features which are then fed to the model in parallel. Adding textual features offloads some of the classification work from the CNN and the classification will therefore be joint (text + image). Globally, the model will accept two or more types of input. In the present study, the inputs are a visual input consisting of the asset's 2D capture and textual inputs consisting of arrays that represent the encoding of different textual labels. All of the features will then be concatenated in a shared features layer which will encode a special representation of the asset features. This features layer is used to perform the classification either for a single or multiple tasks (outputs). As a proof of concept and for validation, we performed our tests and evaluations on paintings. In contrast with traditional CNNs, our solution uses multiple inputs while CNNs use a single visual input. In our case, the additional inputs represent some textual features that are found along the cultural asset visual capture. These textual features coupled with visual features will form an aggregated representation of the asset in a higher order feature space. This in fact will result in a more accurate representation of the asset that improves the categorization and the classification.

In our tests with painting data, we concluded that the most relevant labels are the artist, the year of creation, the genre, the style and the media. After a further analysis and several consultations with heritage institutions regarding the metadata for paintings, we found that the missing labels are generally the artist and the creation year. The media, the genre and the style of a painting can be easily attributed with standalone models or manually (directly from perception). Our model takes as input the image, the style, the media and the genre of a given painting. The model is then trained to predict the artist and the creation year in a multitask fashion using hard parameter sharing [40]. The multitask attribution is mainly performed to force the model to learn the correlations between the artist and the painting's creation year. For example, paintings created by an artist who lived between 1835 and 1875 can never be attributed to 1885, as the artist in question was not alive at that time. For a human, such correlations are straightforward after identifying the artist. However, for a machine learning system, these correlations are not so obvious. Our network learns such correlations implicitly by jointly training it to predict the two output labels and thus making predictions more accurate.

Our selected dataset is smaller than large datasets such as ImageNet used to train large-scale CNNs [41]. As a result, we relied on transfer learning from models trained for the ImageNet challenge in order to benefit from their visual feature extraction. A deeper analysis of these networks revealed that they are very good at extracting and learning low-level image features even for data that were not used to train them. In fact, the majority of domain-specific image recognition tasks rely on transfer learning [42]. For the visual features part, we used pretrained CNNs such as ResNet50 with the ImageNet weights as initialization.

We designed a set of deep-learning-based classifiers using transfer learning (for the visual feature learning part). We conducted several tests and compared the classification performance for different base CNNs, including the ResNet50 [39], VGG16, VGG19 [37] and InceptionV3 [43] CNNs (these networks were the winners of the ImageNet challenge).

We designed four networks based on the following CNNs: ResNet50, VGG16, VGG19 and InceptionV3. The InceptionV3 network has an input image resolution of 299×299 pixels, and we therefore increased the patch size in the online data augmentation process while training the network. For each of the networks we conducted the following procedure: We removed the top dense layers

and connected the last convolutional layer to a global average pooling of 2048 units; this layer was regularized with 50% dropout; this average pooling layer was then connected to a ReLU activated dense layer of 1024 units (visual features layer); and finally, we connected the ReLU dense layer to two output SoftMax activated layers to perform the multitask learning (for artist attribution and year estimation). It is worth noting that the convolutional layer weights were initialized to the ImageNet weights and all network layers were fine-tuned. The hyperparameters used are described in Table 3, and the architecture of the networks is outlined in Figure 8.

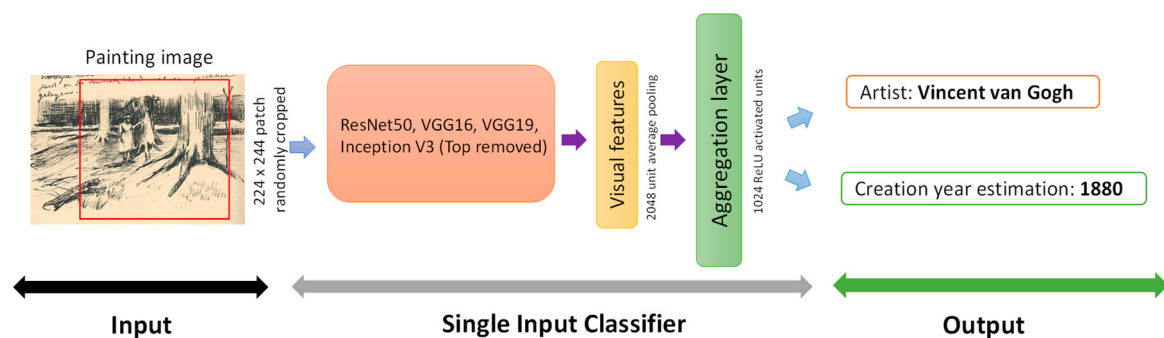


Figure 8. Single input network.

Following our tests, we chose the ResNet50 (residual networks) CNN as our image feature extractor. This network was mainly selected for its training and validation performance compared to the others. The comparison results of the single input network are outlined in Section 3.

2.6. Framework Architecture

Our final network has four inputs and two outputs. The main input are the image data and the remaining inputs are the associated text labels of the paintings (genre, style, and media). The image input is fed to a CNN in order to learn and extract the visual features. We used the ResNet50 network with fine-tuning of all its layers. This was decided after performing several tests with other CNN-based models such as the VGG16 and the VGG19 for the visual extraction task, from which ResNet50 was found to yield the best performance. ResNet50 is a convolutional neural network of 50 layers based on blocks called residual blocks. These blocks were introduced to deal with the problem of the vanishing gradient in the training stage of the network. Due to GPU memory limitations, we could not use the ResNet152 network as a base classifier. The textual inputs are encoded with One Hot arrays regarding the number of samples per class for each of the five textual labels. The textual inputs (three) are then connected to an aggregation layer of 256 units in order to aggregate textual features and reduce their dimension. The CNN output is connected to an average pooling layer of 2048 units. We then merged the 2048 units and the 256 units in a concatenation layer. This concatenation layer was then regularized with 50% dropout and connected to a final aggregation layer of 1024 units which represents the global features (text + image). The output SoftMax layers are then connected to the global features layer. For each of the outputs, the error used is the categorical cross-entropy (CCE) and the global network error is the sum of the errors of each output. Figure 9 outlines the global architecture of our multimodal classification network. The final architecture of the network was established after several tests and assessment of the settings for the network global shape, such as kernel types, layer count, the shared representation layer length, etc. This architecture was validated through many tests to maximize its performance and to reduce its complexity.

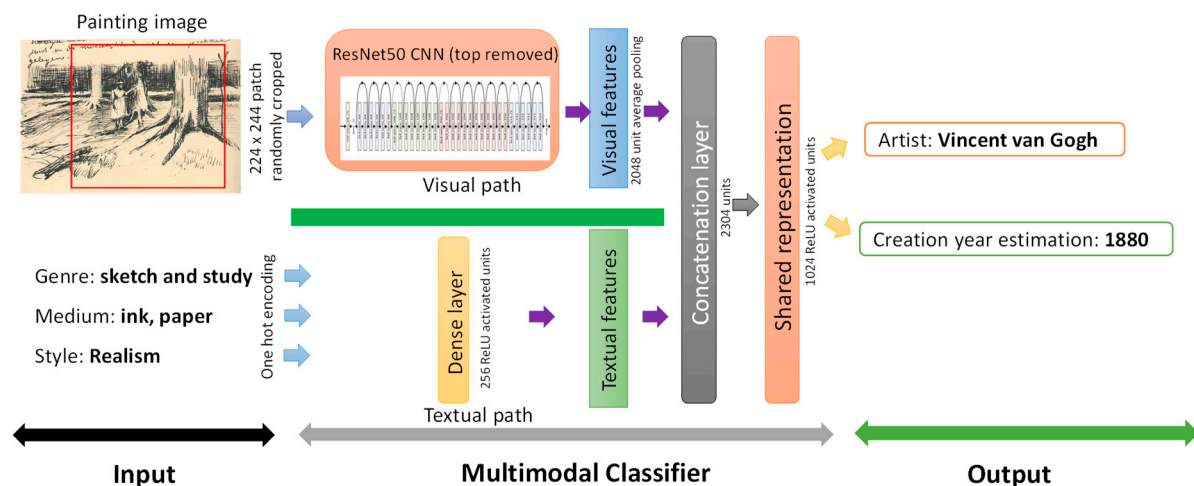


Figure 9. Model architecture.

The training of the networks requires setting several hyperparameters, particularly when using the Stochastic Gradient Descent (SGD) optimization. Several strategies can be followed to establish these hyperparameters [44]. In this study, we initially set their values according to those used in previous works. These hyperparameters were then tuned through several tests leading to the values outlined in Table 3.

To train and validate the model, we used five-fold cross validation. At each fold, we used 80% of the data for training and 20% of the data for validation. The process was performed five times while changing the training and validation sets. We validated our approach using this protocol as a large training set was not available.

Table 3. Training hyperparameters.

Optimizer	LR	Decay	Epochs	Batch Size	Momentum	Nesterov
SGD	0.001	1×10^{-6}	100	64	0.9	True

3. Results

In this section, we outline the evaluation protocol implemented in this study for the purpose of testing and validating our approach. We also present the results of the experimental classification regarding the painting datasets described in the previous section. In particular, we compare the multimodal classification against the single input (image only) classification in order to evaluate the benefits and the additional performance of the former.

3.1. Experimental Setup and Implementation

The models that we compared in this study were implemented using the Keras (2.1.5) deep learning library with Tensorflow GPU (1.7.0) backend [45]. The training was performed on a system with an Intel i7-7700HQ CPU, 16 GB of RAM and a GeForce GTX 1070 GPU with 8 GB of VRAM. The average training time of the models for 100 epochs was 5–7 h. We conducted several tests not only to adjust the hyperparameters but also to tune several of the settings in the architecture of the models to better fit the training data. To evaluate our new method, we used paintings; however, we believe that this approach can also be used with other categories of cultural data.

Table 4 shows the results for the Top-1 accuracy measure for the artist attribution task. These results demonstrate the superior performance of the ResNet50 network compared to the other networks. The tests were performed on a single-input network; an image of a painting was used as input, and the two outputs were the artist and the year of creation. The accuracies shown in Table 4 are the average results after five-fold cross-validation for the dataset of 300 paintings or more per artist.

Table 4. Performance of base image classifiers for the single-input model.

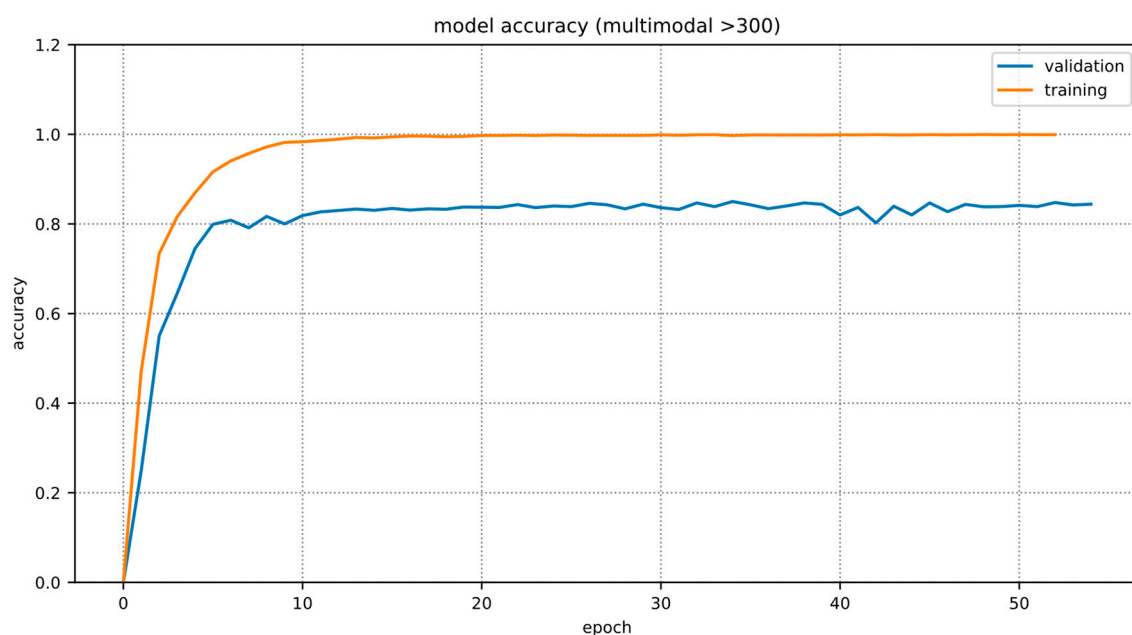
Base Model	Top-1 Accuracy
InceptionV3	0.783
ResNet50	0.802
VGG16	0.752
VGG19	0.775

We found that the difference in performance between the tested networks was not very large. This was also the case in the ImageNet challenge, where we could only observe single digit improvements between these networks from year to year. Nevertheless, the ResNet50 network was found to have the best validation accuracy for our dataset, and fewer training parameters, compared to the other networks.

After the selection of the ResNet50 network as the base CNN for the image path, and for comparison purposes, we trained the model on the four datasets mentioned in the data pre-processing section. We also used the single-input network with the ResNet50 network as the base CNN to compare a single-input CNN with the multimodal network (multiple inputs).

Both multimodal and single-input networks were trained using the SGD optimizer with the hyperparameters outlined in Table 3. Our first tests with the Adam optimizer [46] were mostly comparable to SGD performance, although the SGD was more consistent and yielded slightly better results.

Figure 10 outlines the training and validation accuracy for the multimodal classifier on the >300 dataset.

**Figure 10.** Training and validation accuracy evolution of the multimodal network training on the >300 dataset.

Since we used cross-validation, the training and validation operations were performed five times. Each time, we perform the training for 100 epochs, which takes 5–7 h. The validation is mostly instantaneous as it is a feed-forward computation.

3.2. Approach Evaluation

To evaluate our approach, we trained our model on the datasets discussed in Section 2.1. We mainly evaluated the model prediction accuracy in scenarios where we leveraged the available

information concerning the asset along with its image (visual features + textual features). We also compared the performance of the multimodal classification with other scenarios where only the 2D capture is fed to the model as input while tasked with predicting the same labels. Although we used two outputs in our implemented models (artist and creation year), for our experiments, we mainly focused on predicting the artist class, as it is the most relevant and widely used classification metric for painting categorization. The tests were performed on the four datasets mentioned in Section 2.1 by comparing the multimodal network performance against the single input network. This resulted in four pairs of networks, a pair for each dataset (multimodal, single input).

In the following, we outline the classification results of the multimodal and the single-input classifications. As previously mentioned in the data collection and pre-processing section, we used the selected data to produce four sub-datasets to compare the impact of varying the number of samples per class on the classification. To more clearly see and interpret the results, we present the evaluation details such as the confusion matrix only for the dataset with 19 classes (300 or more paintings per artist). The limited number of classes enables us to better perceive and discuss these results.

We evaluate the performance of our classification by using the following metrics: precision, recall, F1 score and Top-1 Accuracy. Following our test protocol, we performed our evaluations on the four different sub-datasets. For each dataset, we created two networks (single-input and multimodal input) and split the data into 80% training and 20% testing. To feed the data to the multimodal network, we wrote a custom batch generation function that yields more than one input. We performed five-fold cross-validation on the eight models, mainly due to the limited number of samples. Table 5 outlines the average results for the classification metrics of the trained networks on the validation sets (Top-1 accuracy, precision, recall and F1 score).

Figure 11 outlines the confusion matrices of the single-input and multimodal input networks for the >300 dataset. For the other datasets, the confusion matrices would not be clear due to the high number of classes. The results shown are for the best performing networks on the validation set.

Table 5. Classification metrics of our approach (multimodal classification) against a single-input model (image only classification) on the validation datasets (average results after five-fold cross-validation).

Model-Paintings Per Artists	Top-1 Accuracy		Precision		Recall		F1 Score	
	1-Input	Multi	1-Input	Multi	1-Input	Multi	1-Input	Multi
ResNet50, all	0.503	0.603	0.472	0.541	0.433	0.529	0.437	0.534
ResNet50, >50	0.531	0.649	0.534	0.607	0.494	0.593	0.496	0.598
ResNet50, >100	0.601	0.753	0.616	0.703	0.603	0.687	0.600	0.694
ResNet50, >300	0.802	0.873	0.826	0.854	0.823	0.842	0.823	0.851

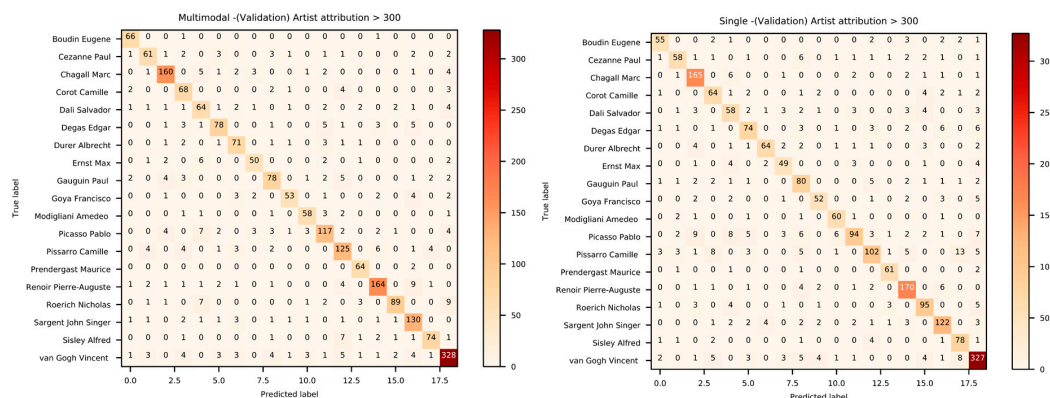


Figure 11. Confusion matrices of the multimodal input (left) and single-input (right) methods for the artist attribution task of the >300 dataset.

3.3. Results Interpretation

Based on the previous results, it can be observed that, in the majority of cases, multimodal classification surpasses the single-input classification in terms of accuracy. Additionally, we can see that increasing the number of samples from a given artist contributes significantly to the classification results. This increase can be explained by the fact that more samples per class help deep-learning-based techniques in the generalization, as the intraclass correlation (features categorizing a specific class) can be learned easily and more effectively when many samples representing that class are used in the training.

The largest difference in accuracy between the multimodal and single-input classifications was observed in the dataset with more than 100 paintings per artist, where the Top-1 accuracy of the multimodal network was 75% and that of the single-input network was 60%. The confusion matrix of the multimodal classification reveals that some false positives and false negatives were eliminated by the use of the additional inputs.

4. Discussion

The main goal of our study was to investigate the impact of adding relevant information describing cultural data samples on their classification. Besides being innovative, our approach yields very good results in terms of real-world performance. The idea of including known data along with images for classification tasks boosted both the classification accuracy and efficiency. Our approach aims to study real-world scenarios that are often faced by art curators in museums and heritage organizations when assigned the task of labeling or annotating an asset. The aim is to achieve an efficient metadata annotation for incompletely labeled assets. Missing labels currently require a deep cultural knowledge and several experts for their completion. To address this issue, we designed and implemented an automated annotation procedure based on deep learning for painting annotation when only a part of an asset's metadata is known. We leveraged visual features of assets, which were learned using a CNN, along with textual features learned using a conventional neural network (multi-layer perceptron). Through this development, we tried several designs and settings, especially for the visual recognition part (CNN base models). We relied mainly on transfer learning and fine-tuning of networks pretrained on the ImageNet challenge. Through several comparisons, we found that residual networks yielded the best performance, a result which validates previous research which considered such networks to be the state-of-the-art in visual recognition [47]. Our painting classification model, as previously described, is designed to take as input multiple types of data, such as image and textual labels. At the training stage, the model receives the data and learns to predict the labels for two tasks, namely the artist attribution and the creation year estimation. By doing so, we forced the model to learn correlations between the input labels and, more importantly, between the output labels. This was possible with hard parameter sharing, as we were using a "shared features" layer that aggregates the features of the inputs (image + text). This layer was shared between the two output layers (artist and year). After deep analysis, we found that the network is implicitly forced to learn several correlations between the input and output layers.

Following our evaluation protocol, we can confidently conclude that adding relevant information to visual features helps to increase the accuracy of the classification and annotation of cultural assets. For the most part, this is not surprising, as it is known that deep-learning-based techniques benefit from large amounts of data. The use of multiple inputs yields better results since, in addition to the visual capture, the textual labels contribute to the feature encoding of the asset. Moreover, instead of predicting a single output, we used multitask learning with hard parameter sharing to predict several output labels at once, leveraging the features extracted from the data samples. This also conditions the output of the network and helps the model to learn correlations between the output labels.

The real-time usability of such a system is straightforward. It is a fact that the system takes considerable time to be trained and validated, but once these steps are completed, the predictions are generated instantly.

In our study, the first limiting factor was the dataset. The datasets we collected were not complete regarding the labels required to classify cultural data of other types such as ceramics, weapons, carpets, etc. After evaluation and testing, another limitation was observed in that some classes were not classified correctly. This is mainly due to the inherent nature of paintings, as visual similarities and characteristics are often observed between artists who were inspired by each other, were known to have the same style or who simply came from the same art school. These artifacts are usually hard to distinguish and categorize, even for humans. Other limitations related to the performance of the computer hardware used to train the model (mainly the GPU memory) could be addressed with the use of high-performance computing systems. Time complexity could also be addressed, either with a high-performance computing architecture or with future CNN designs that require fewer parameters while having the same or better feature learning performance.

Several improvements could form foundations for future work. The use of future CNN designs as base networks could be beneficial, especially with the recent progress achieved in the ImageNet challenge. Additionally, the inclusion of other data sources and the use of completely annotated data samples to train the models could significantly help diversify this approach to other cultural categories. Furthermore, the inclusion of other annotation techniques for textual labels such as the use of name-entity recognition (NER) to extract tags directly from cultural asset descriptions is highly likely to improve means of creating cultural datasets and thus the generalization of the classification.

5. Conclusions

In this paper, we propose a novel multimodal and multitask classification approach for the categorization and annotation of cultural assets. As a preliminary study, we focused on the categorization of paintings collected from multiple museums and heritage organizations. Such organizations often face challenges and difficulties in labeling incomplete and partially annotated assets which could lead to the deterioration of their collections' values. In contrast to the approaches found in the literature, and after several consultations with cultural heritage experts, we found that cultural assets are usually partially annotated, as some additional information can be found along with their visual copy. For these reasons, in addition to visual data, our cultural annotation and classification approach relied on other data types, such as textual labels, that can be obtained from partially annotated assets. We designed and implemented a framework based on convolutional neural networks for visual feature learning and regular neural networks for textual feature learning. Both visual and textual features were aggregated and used to perform asset annotation. Our tests on paintings clearly show that the use of additional textual data boosts the accuracy of asset classification. We compared our method against a single-input design, and found that the average accuracy increase is between 5% and 15%. We conclude that the reason for this increase is related to the constraints that we introduced to the inputs and outputs of the network at the training stage. These constraints condition the model outputs and force the network to learn new types of correlations between the input and output labels. Through this work, we aim to provide heritage specialists with tools and techniques that can help with the annotation and classification of unlabeled and incomplete collections. Future improvements and works were also suggested involving the inclusion of more cultural types and the extraction of textual labels by directly tagging asset descriptions with natural language processing tools.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2076-3417/8/10/1768/s1>.

Author Contributions: A.B. proposed the idea, took care of the technical approach design and its implementation, and wrote the paper. Co-authors A.B. and F.S. supervised the research and monitored its progress. They also contributed to the writing and revision of the paper. All authors read and approved the final manuscript.

Acknowledgments: This publication was made possible by NPRP grant 9-181-1-036 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Parry, R. *Recoding the Museum: Digital Heritage and the Technologies of Change*; Routledge: London, UK, 2007.
2. Wang, Y.W.; Stash, N.; Aroyo, L.; Gorgels, P.; Rutledge, L.; Schreiber, G. Recommendations based on semantically enriched museum collections. *Web Semant. Sci. Serv. Agents World Wide Web* **2008**, *6*, 283–290. [[CrossRef](#)]
3. Van Garderen, P. Archivemata: Using micro-services and open-source software to deliver a comprehensive digital curation solution. In Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria, 19–24 September 2010; Österreichische Computer Gesellschaft: Vienna, Austria, 2010; pp. 145–149.
4. Stanco, F. *Digital Imaging for Cultural Heritage Preservation: Analysis, Restoration, and Reconstruction of Ancient Artworks*, 1st ed.; Battiato, S., Gallo, G., Eds.; CRC Press: Florida, FL, USA, 2011.
5. Doerr, M. Ontologies for cultural heritage. In *Handbook on Ontologies*; Springer: Cham, Switzerland, 2009; pp. 463–486.
6. Belhi, A.; Bouras, A.; Foufou, S. Digitization and preservation of cultural heritage: The ceproqha approach. In Proceedings of the 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Colombo, Sri Lanka, 6–8 December 2017; pp. 1–7.
7. Haslhofer, B.; Isaac, A. Data. Europeana. Eu: The europeana linked open data pilot. In Proceedings of the International Conference on Dublin Core and Metadata Applications, Washington, WA, USA, 26–29 October 2011; pp. 94–104.
8. Llamas, J.; Lerones, P.M.; Medina, R.; Zalama, E.; Gómez-García-Bermejo, J. Classification of architectural heritage images using deep learning techniques. *Appl. Sci.* **2017**, *7*, 992. [[CrossRef](#)]
9. Pavlidis, G.; Koutsoudis, A.; Arnaoutoglou, F.; Tsioukas, V.; Chamzas, C. Methods for 3D digitization of cultural heritage. *J. Cult. Herit.* **2007**, *8*, 93–98. [[CrossRef](#)]
10. Agarwal, S.; Karnick, H.; Pant, N.; Patel, U. Genre and style based painting classification. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2015; pp. 588–594.
11. Mensink, T.; Van Gemert, J. The rijksmuseum challenge: Museum-centered visual recognition. In Proceedings of the International Conference on Multimedia Retrieval, Glasgow, UK, 2–3 April 2014; p. 451.
12. Tan, W.R.; Chan, C.S.; Aguirre, H.E.; Tanaka, K. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, Arizona, 25–28 September 2016; pp. 3703–3707.
13. Obeso, A.M.; Vazquez, M.S.G.; Acosta, A.A.R.; Benois-Pineau, J. Connoisseur: Classification of styles of mexican architectural heritage with deep learning and visual attention prediction. In Proceedings of the 15th International Workshop on Content-Based Multimedia Indexing, Florence, Italy, 19–21 June 2017; ACM: New York, NY, USA, 2017.
14. Banerji, S.; Sinha, A. Painting classification using a pre-trained convolutional neural network. In Proceedings of the International Conference on Computer Vision, Graphics, and Image Processing, IIT Roorkee, India, 19–21 February 2016; pp. 168–179.
15. Saleh, B.; Elgammal, A. A unified framework for painting classification. In Proceedings of the 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 14–17 November 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1254–1261.
16. Arora, R.S.; Elgammal, A. Towards automated classification of fine-art painting style: A comparative study. In Proceedings of the 2012 21st International Conference on Pattern Recognition (ICPR), Tsukuba Science City, Japan, 11–15 November, 2012; pp. 3541–3544.
17. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December, 2012; pp. 1097–1105.
18. Shamir, L.; Macura, T.; Orlov, N.; Eckley, D.M.; Goldberg, I.G. Impressionism, expressionism, surrealism: Automated recognition of painters and schools of art. *ACM Trans. Appl. Percept. (TAP)* **2010**, *7*, 8. [[CrossRef](#)]
19. Puthenpuhussery, A.; Liu, Q.; Liu, C. Color multi-fusion fisher vector feature for fine art painting categorization and influence analysis. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), New York, NY, USA, 7–9 March 2016; pp. 1–9.

20. Li, J.; Yao, L.; Hendriks, E.; Wang, J.Z. Rhythmic brushstrokes distinguish van gogh from his contemporaries: Findings via automated brushstroke extraction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 1159–1176. [PubMed]
21. Ma, D.; Gao, F.; Bai, Y.; Lou, Y.; Wang, S.; Huang, T.; Duan, L.-Y. From part to whole: Who is behind the painting? In Proceedings of the 2017 ACM on Multimedia Conference, California, CA, USA, 23–27 October, 2017; ACM: New York, NY, USA, 2017; pp. 1174–1182.
22. Tseng, T.-E.; Chang, W.-Y.; Chen, C.-S.; Wang, Y.-C.F. Style retrieval from natural images. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1561–1565.
23. Sheng, J.; Jiang, J. Recognition of Chinese artists via windowed and entropy balanced fusion in classification of their authored ink and wash paintings (IWPs). *Pattern Recognit.* **2014**, *47*, 612–622. [CrossRef]
24. Chu, W.-T.; Wu, Y.-L. Deep correlation features for image style classification. In Proceedings of the 2016 ACM on Multimedia Conference, Amsterdam, The Netherlands, 15–19 October 2016; ACM: New York, NY, USA, 2016; pp. 402–406.
25. Hong, Y.; Kim, J. Art painting detection and identification based on deep learning and image local features. *Multimed. Tools Appl.* **2018**, 1–16. [CrossRef]
26. Karayev, S.; Trentacoste, M.; Han, H.; Agarwala, A.; Darrell, T.; Hertzmann, A.; Winnemoeller, H. Recognizing image style. *arXiv* 2013; arXiv:1311.3715.
27. Elgammal, A.; Kang, Y.; Leeuw, M.D. Picasso, matisse, or a fake? Automated analysis of drawings at the stroke level for attribution and authentication. *arXiv* 2017; arXiv:1711.03536.
28. Van Noord, N.; Hendriks, E.; Postma, E. Towards discovery of the artist's style: Learning to recognize artists by their artworks. *IEEE Signal Process. Mag.* **2015**, *32*, 46–54. [CrossRef]
29. Bianco, S.; Mazzini, D.; Schettini, R. Deep multibranch neural network for painting categorization. In Proceedings of the 19th International Conference on Image Analysis and Processing, Catania, Italy, 11–15 September 2017; pp. 414–423.
30. Huang, X.; Zhong, S.-H.; Xiao, Z. Fine-art painting classification via two-channel deep residual network. In Proceedings of the Pacific Rim Conference on Multimedia, Harbin, China, 28 September 2017; Springer: Heidelberg, Germany, 2017; pp. 79–88.
31. Mao, H.; Cheung, M.; She, J. Deepart: Learning joint representations of visual arts. In Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, CA, USA, 23–27 October 2017; ACM: New York, NY, USA, 2017; pp. 1183–1191.
32. Seguin, B.; Striolo, C.; Kaplan, F. Visual link retrieval in a database of paintings. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Heidelberg, Germany, 2016; pp. 753–767.
33. Strezoski, G.; Worring, M. Omniart: Multi-task deep learning for artistic data analysis. *arXiv* 2017; arXiv: 708.00684.
34. Richardson, L. Beautiful Soup Documentation. April 2007. Available online: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (accessed on 28 September 2018).
35. McCann, B.; Bradbury, J.; Xiong, C.; Socher, R. Towards the imagenet-cnn of nlp: Pretraining sentence encoders with machine translation. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6285–6296.
36. Shao, L.; Zhu, F.; Li, X. Transfer learning for visual categorization: A survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 1019–1034. [CrossRef] [PubMed]
37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* 2014; arXiv:1409.1556.
38. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the European conference on computer vision, Zurich, Switzerland, 6–12 September 2014; Springer: Heidelberg, Germany, 2014; pp. 818–833.
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
40. Ruder, S. An overview of multi-task learning in deep neural networks. *arXiv* 2017; arXiv:1706.05098.

41. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 248–255.
42. Hoo-Chang, S.; Roth, H.R.; Gao, M.; Lu, L.; Xu, Z.; Nogues, I.; Yao, J.; Mollura, D.; Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **2016**, *35*, 1285.
43. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 2818–2826.
44. Bengio, Y. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*; Springer: Heidelberg, Germany, 2012; pp. 437–478.
45. Chollet, F. Keras: Deep Learning Library for Theano and Tensorflow. Available online: <https://keras.io> (accessed on 28 September 2018).
46. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* 2014; arXiv:1412.6980.
47. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. *arXiv* 2016; arXiv:1602.07261.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).