Pattern Recognition Letters 128 (2019) 420–426

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Webly-supervised zero-shot learning for artwork instance recognition

Riccardo Del Chiaro*, Andrew D. Bagdanov, Alberto Del Bimbo

Media Integration and Communication Center, University of Florence, Florence, Italy

ARTICLE INFO

Article history:
Received 22 July 2019
Accepted 28 September 2019
Available online 3 October 2019

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

This paper describes experiments on supervised approaches to webly-labeled artwork instance recognition and zero-shot learning for unseen artwork instance recognition. We build on our earlier work on webly-supervised learning using the *NoisyArt* dataset. The dataset consists of more than 90,000 images and in more than 3,000 webly-supervised classes, and a subset of 200 classes with verified test images. Document embeddings are provided for short descriptions of all artworks. *NoisyArt* is designed to support research on webly-supervised artwork instance recognition, zero-shot learning, and other approaches to visual recognition of cultural heritage objects. We report results of experiments on artwork instance recognition using the *NoisyArt* dataset of webly-labeled images as well as on the CMU-Oxford Sculptures dataset. In addition, we perform extensive experiments on zero-shot learning using webly-labeled training images for unseen artwork recognition. Our results demonstrate the benefits and limitations of zero-shot learning for instance recognition over webly-supervised data.

1. Introduction

Cultural patrimony and exploitation of its artifacts is an extremely important economic driver internationally. This is especially true for culturally dense regions like Europe and Asia who rely on cultural tourism for jobs and important industry. Tourist travelers from the Unites States alone represent nearly 130 million adults spending approximately $171 billion annually on leisure travel [1]. Museums are massive, distributed repositories of physical and digital artifacts. For decades museums have been digitizing their collections in an effort to render their content more available to the general public. Initiatives like EUROPEANA [2] and the European Year of Cultural Heritage[1] have advanced the state-of-the-art in cultural heritage metadata exchange and promoted coordinated valorization of cultural history assets, but have had limited impact on diffusion and dissemination of each collection.

The state-of-the-art in automatic recognition of objects, actions, and other visual phenomena has advanced by leaps and bounds in the last five years. This visual recognition technology can offer the potential of linking cultural tourists to the (currently inaccessible) collections of museums. Researchers and entrepreneurs have long dreamed of building mobile applications capable of recognizing cultural landmarks and artifacts and subsequently delivering rich multimedia content to users in a way that is contextualized and personalized. Such application scenarios are realistic only if we have ways of recognizing a broad range of artworks. The challenges and barriers to this type of recognition technology have been studied in the past in the multimedia information analysis community [3].

Recent breakthroughs in visual media recognition offer promise, but also present new challenges. One key challenging factor in the application of state-of-the-art classifiers is the data-hungry nature of modern visual recognition models. Even modestly sized Convolutional Neural Networks (CNNs) can have tens (or even hundreds) of millions of trainable parameters. As a consequence, they can require millions of *annotated* training examples to be effectively trained. The real problem then becomes the *cost* of annotation. Museum budgets are already stretched with *classical* curation requirements, adding to that the additional costs of collecting and annotating example media is not feasible.

Webly-supervised learning offers one solution to the data annotation problem by exploiting abundantly available media on the web. In our application scenario, for example, there are abundant images, videos, blog posts, and other multimedia assets freely available on the web. If the multimedia corresponding to specific instances of cultural heritage items can be retrieved and verified in some way, it can be exploited as (noisy) training data. The problem then turns from one of a lack of data, to one of mitigating the effects of various types of noise in the training process that derive from its Webly nature [4,5].

The availability of high-quality, curated *textual* descriptions for many works of art opens up the possibility of zero-shot Learning

* Corresponding author.
E-mail address: riccardo.delchiaro@unifi.it (R. Del Chiaro).

[1] https://europa.eu/cultural-heritage/.

https://doi.org/10.1016/j.patrec.2019.09.027
0167-8655/

(ZSL) in which visual categories are acquired without *any* training samples. ZSL relies on alignment of semantic and visual information learned on a training set [6]. Zero-shot recognition is an extremely challenging problem, but it is particularly appealing for artwork recognition because museums normally have at least one curated description for each artwork in their collections.

In the case of artwork recognition we must solve an *instance recognition* problem using zero-shot learning, while all ZSL work to date has been on zero-shot *class recognition*. We designed the *NoisyArt* dataset specifically to target cultural heritage artifacts and their recognition [7]. *NoisyArt* is designed to support research on multiple types of webly-supervised recognition problems. Included in the database are document embeddings of short, verified text descriptions of each artwork in order to support development of models that mix language and visual features such as zero-shot learning [6] and automatic image captioning [8]. We believe that *NoisyArt* represents the first benchmark dataset for webly-supervised learning for cultural heritage collections.

In addition to extending our work on the *NoisyArt* dataset, in this paper we focus on improving recognition results on webly-labeled data by compensating for domain shift between real-world photos and photos of artworks on the web. We also report on experiments evaluating zero-shot artwork recognition and we show how fully webly-labeled classes can significantly improve zero-shot recognition performance. As far as we know we are the first to consider the problem of webly-supervised, zero-shot learning for *instance* recognition.

The rest of this paper is organized as follows. In the next section we review recent work related to our contribution. In Section 3 we briefly describe the NoisyArt dataset [7] designed specifically for research on webly-supervised learning in museum contexts, and in Sections 4 and 5 we describe our approaches to supervised and zero-shot recognition of artworks on webly-labeled data respectively. We report on a range of experiments in Section 6, and conclude in Section 7 with a discussion of our contribution.

2. Related work

In this section we review work from the literature related to the *NoisyArt* dataset and webly-supervised learning.

Visual recognition for cultural heritage. The Mobile Museum Guide was an early attempt to build a system to recognize instances from a collection of 17 artworks using photos from mobile phone [4]. More recently, the Rijksmuseum Challenge dataset was published which contains more than 100,000 highly curated photos of artworks from the Rijksmuseum collection [9]. The PeopleArt dataset, on the other hand, consists of high-quality, curated photos of paintings depicting people in various artistic styles [10]. The objectives of these datasets vary, from person detection invariant to artistic style, to artist/artwork recognition. The UNICT-VEDI dataset [11] focuses on localization of visitors in a cultural site via wearable devices. A unifying characteristic of these datasets is the high level of curation and meticulous annotation invested.

Another common application theme in multimedia analysis and computer vision applied to cultural heritage is personalized content delivery. The goal of the MNEMOSYNE project was to analyze visitor interest *in situ* and to then select content to deliver on the basis of similarity to recognized content of interest [12].

Webly-supervised category recognition. Early approaches to webly-supervised learning were the decontamination technique of [13], and the noise filtering approach of [14]. A more recent approach is the noise *adaptation* approach of [5]. This approach looks at two specific types of label noise – labelflip and outliers – and modifies a deep network architecture to absorb and adapt to them. A very recent approach to webly-supervised training of CNNs is the representation adaptation approach of [15]. The authors first fit a CNN to "easy" images identified by Google, and then adapt this representation to "harder" ones by identifying category relationships in the noisy data.

The majority of work on webly-supervised learning has concentrated on category learning. However, the *NoisyArt* dataset is an instance-based, webly-supervised learning problem. As we will describe in Section 3, instance-based learning presents different sources of label noise than category-based.

Landmark recognition. The problem of landmark recognition is similar artwork classification in that they are both *instance* recognition problems. It is also one of the first problems to which webly-supervised learning was widely applied. The authors of [16] use webly-supervised learning to acquire visual models of landmarks by identifying *iconic views* of each landmark in question. Another early work merged image and contextual text features to build recognition models for large-scale landmark collection [17]. In [18] the authors extend the UNICT-VEDI dataset with annotations of points of interests using an object detector.

Artwork recognition differs from landmark recognition, however, in the diversity of viewpoints recoverable from web search alone. As we will show in Section 3, the NoisyArt dataset suffers from several types of label bias and label noise which are particular to the artwork recognition context.

Zero Shot Learning. Techniques for zero-shot learning (ZSL) attempt to learn to classify never-before seen classes for which semantic descriptions (but no images) are available. Recent advances in ZSL use techniques that directly learn mappings from a visual feature space to a semantic space. In some cases a linear mapping is used to learn a compatibility between visual e semantic features [19–21], in other cases a non-linear mapping is used [22], and in others a metric learning approach is used instead of compatibility [23,24].

3. The *NoisyArt* dataset

The work in this paper builds upon our earlier work on the *NoisyArt* dataset [7]. The goal of *NoisyArt* is to support research on webly-supervised and zero-shot artwork recognition for cultural heritage applications. Webly-supervision is an important feature, since in the cultural applications data can be acutely scarce. Thus, the ability to exploit abundantly available imagery to acquire visual recognition models would be a tremendous advantage. Here we briefly summarize the salient features of *NoisyArt* and refer the interested reader to the original paper or website for more information.[2]

To collect the *NoisyArt* dataset we exploited a range of publicly available data sources on the web. We used public knowledge bases like DBpedia [25,26] and Europeana [2] to select artworks for *NoisyArt*. The result is a set of 3,120 artworks with Wikipedia entries and ancillary information for each one. We retrieved metadata from DBPedia, which for some artworks it also contains an image which we call a *seed image* because it is unequivocally associated with the artwork. To support reproducible research on webly-supervised learning, we provide a subset of 200 classes with manually verified test images (i.e. with *no label noise*).

For images we queried Google Images using the title of each artwork and the artist name. For each query we downloaded the first 20 retrieved images, which tend to be very clean – especially for paintings, which are similar to scans or posters. The diversity of examples can be poor for images acquired in this way. Finally, we used the Flickr API to retrieve a small set of images more similar to real-world pictures taken by users. Images retrieved from

[2] https://github.com/delchiaro/NoisyArt.

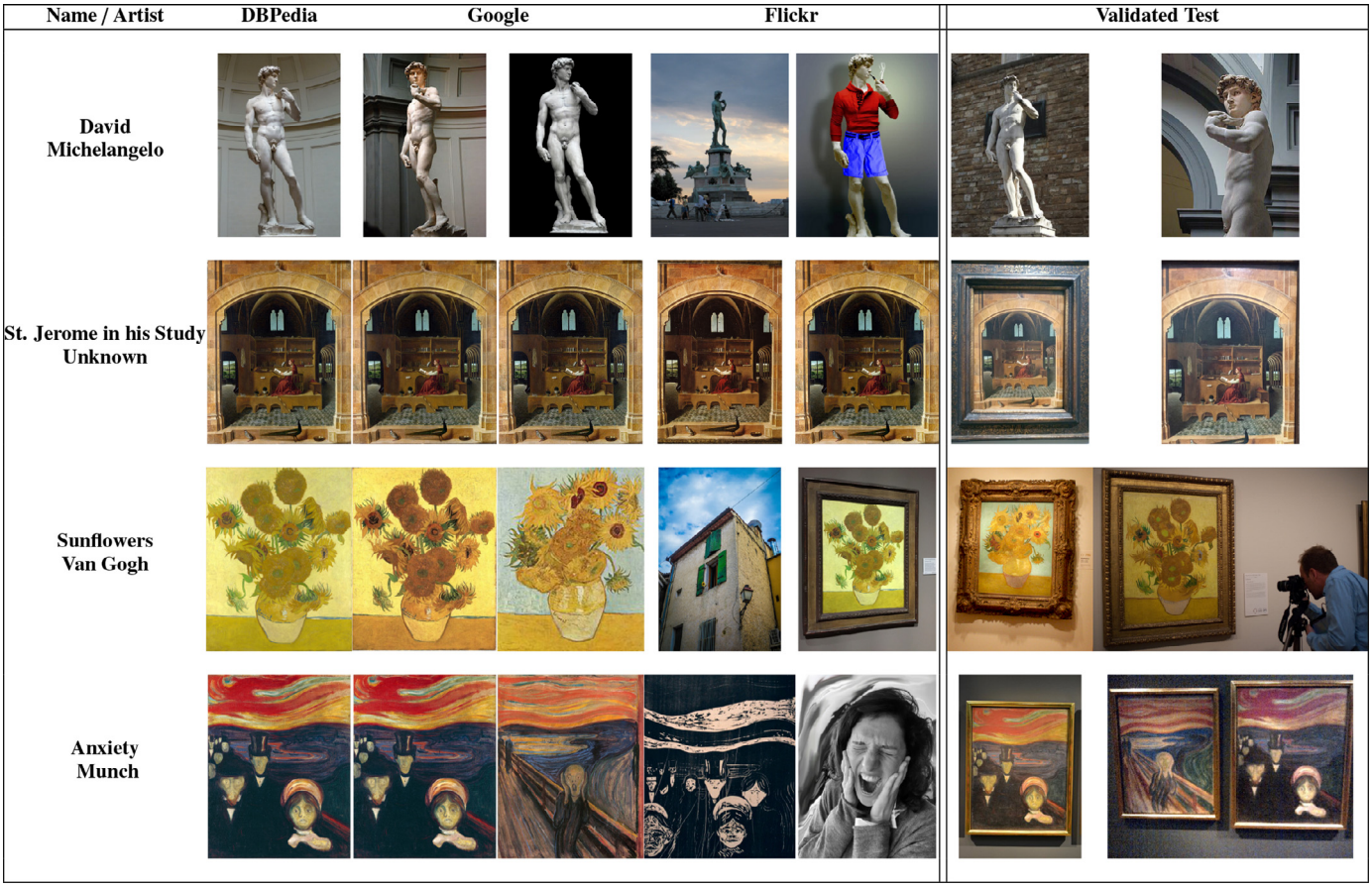

Fig. 1. Sample classes and training images from the *NoisyArt* dataset. For each artwork/artist pair we show the seed image obtained from DBpedia, the first two Google Image search results, and the first two Flickr search results.

Flickr tend to be more noisy: the only supervision is by end-users, and many images (specially for famous and iconic artworks) do not contain the expected subject.

From these sources to collected 89,095 images for the 3,120 classes. Fig. 1 illustrates some sample classes and images from both the webly training set and verified test set. Note the strong domain shift in these images, in particular for paintings, with respect to those in the training set shown in the first columns of the table. Table 1 summarizes the data and annotations provided in *NoisyArt*.

Finally, each artwork has a description and metadata retrieved from DBpedia, from which a single textual document was created for each class. These short descriptions were then embedded using doc2vec [27] in order to provide a compact, vector space embedding for each artwork description. These embeddings are included to support research on zero-shot learning and other multi-modal approaches to learning over weakly supervised data.

4. Webly-supervised artwork recognition

In this section we describe our approach to artwork identification using webly-supervised training data. We first briefly describe the baseline classifier model used in all experiments and then introduce a simple normalization technique that significantly boosts performance by compensating for domain shift in our instance recognition problem.

4.1. Baseline classifier model

Our baseline network for artwork identification is the same described in [7]. It consists of a shallow classifier which is a Multilayer Perceptron (MLP) with a single hidden layer of 4,096 units, followed by a ReLU non-linearity and a linear output layer. This network is trained using image features extracted using ResNet

Table 1
Characteristics of the *NoisyArt* dataset for artwork recognition.

			webly images		verified images
Split Type	split name	classes	training	validation	test
Classification	fully-webly	2,920	65,759	17,368	0
	verified-webly	200	4,715	1,253	1,379
	totals:	3,120	70,474	18,621	1,379
			webly images		verified images
Zero Shot	unseen	50	0		355
	3-fold-seen	150	4,459		1,024
	webly-seen	2,920	83,127		0
	seen totals:	3,070	87,586		1,024

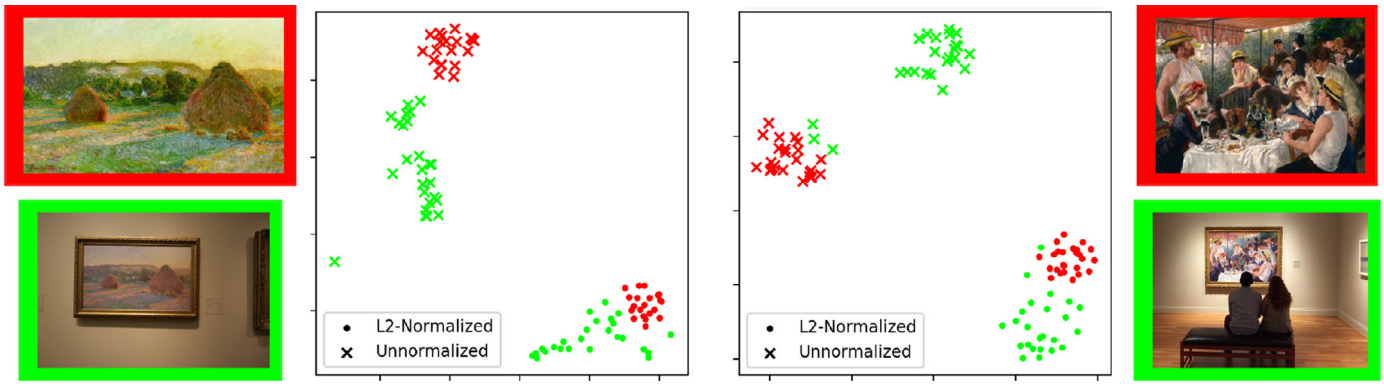

Fig. 2. t-SNE plots of features from examples of single artworks extracted using networks trained on *NoisyArt*. Dots come from a L_2 normalized network and crosses from the baseline network. Green indicates verified test images and red webly-labeled training images. Note how when L_2 normalization is used the training and test image clusters approach one another. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

[28] and VGG [29] CNNs pretrained on ImageNet. The new weights in the shallow network are randomly initialized and the network is trained using a cross-entropy on the webly-supervised training labels.

In the original NoisyArt paper [7] we improved over the baseline network using different techniques to cope with label noise in the dataset. We experimented with decreasing weight on noisy examples in the loss function (entropy scaling) and correcting class predictions for commonly confused classes via labelflip noise mitigation. Nevertheless, we noticed a large difference in performance on the validation set (with noisy, webly-supervised labels) and performance on the test set (validated and thus noise-free labels). This suggested that a source of error could be the domain shift between the training/validation and test images. This domain shift is particularly evident for painting classes, where images returned by Google are often similar to *scans*, while test images acquired with mobile cameras exhibit high degrees of perspective and other distortions.

4.2. Domain shift and L_2 normalization

Because of domain shift and differences observed between test and validation performance, we investigated the use of an L_2 normalization layer [30] inserted before the output layer in our shallow recognition network. The authors of [30] proposed this strategy for face recognition problems, observing that normalization helps create similar representation for images with different visual characteristics (e.g. picture quality) because the magnitude of features is ignored by the final classification layer.

When using this technique we simply modify our baseline model and replace the ReLU activation after the hidden layer with an L_2 normalization layer. This layer simply normalizes the features so that the they have unit norm:

$$f(\mathrm{x}) = \frac{\alpha \mathrm{x}}{||\mathrm{x}||_2}, \tag{1}$$

where x is the output of the last hidden layer and α is a parameter used to rescale the radius of the unit hypersphere. Using $\alpha = 1$ we project each feature x in the hypersphere with unit radius, increasing α we are increasing the radius, and with that the surface area of the hypersphere.

Our intuition is that this should help mitigate domain shift and in general reduce the distance of the features given by images of different quality. Moreover, it should force features from the same class to be closer, while keeping features from different classes far from each other in the normalized space. In Fig. 2 we illustrate the difference in features extracted from images of paintings from the test set (i.e. real world photos) compared to those from the training set (webly-supervised, lacking in variety, and similar to scans). In the first case features cluster together and it is easy to confuse the two different kinds of images from the same class. Without L_2 normalization scans and photographs of paintings tend to cluster in different regions of the space, rendering the classification task much harder. We found this simple technique to be much more helpful to final network performance than to all the other techniques reported in [7].

5. Zero-shot artwork recognition

We embed text descriptions provided for each artwork in *NoisyArt* doc2vec [27] pretrained on Wikipedia. These 300-dimensional vectors become the semantic descriptions classes for zero-shot learning. In the following subsections we describe several baseline ZSL techniques that we implemented and tested on *NoisyArt*, as well as extensions to a known ZSL technique that we propose.

5.1. Compatibility models

Compatibility models learn mappings from a visual embedding space (e.g. CNN features) to the semantic space (e.g. doc2vec embeddings). Training usually consists of pair or triple sampling and a loss function that balances distances between positive and negative image examples to their semantic class embeddings.

Linear compatibility. Linear models rely on mapping visual features into the semantic space through a linear mapping trained to maximize a compatibility function for pairs of visual/semantic features coming from the same class.

For EsZSL [31] we used an open source implementation available online,[3] while for the other comparisons we adapted linear compatibility approaches to our task.

The authors of DEVISE [20] used a dedicated language model trained together with a linear embedding of visual features into semantic space. Instead, we use fixed doc2vec [27] semantic features and a margin of 0.5 instead of 0.1.

ALE [19] is another linear compatibility approach, however from the paper it is unclear which decreasing γ_k function the authors use weight examples in the ranking. We used a modified sigmoid function:

$$\gamma_k = 1 + \frac{\alpha}{1 + e^{\beta k}} - \frac{\alpha}{2}, \tag{2}$$

with α and β fixed to 1.7 and 0.02, respectively.

[3] https://github.com/chichilicious/embarrassingly-simple-zero-shot-learning.

Non-linear compatibility. These models learn a non-linear mapping of image features into the semantic space. For our investigation we implemented variants of a non-linear compatibility model from the literature [22]. These models use a shallow MLP network that embeds visual features in a semantic space. During training we randomly pick a negative label $y' \neq y_n$ for each visual feature x_n with label y_n, and we compute the following loss to train the embedding network using stochastic gradient descent:

$$\mathcal{L}(\mathbf{x}_n, y_n, y') = [m + F(\mathbf{x}_n, y'; W) - F(\mathbf{x}_n, y_n; W)]_+ \quad (3)$$

where $m > 0$ is a margin, W are the network weights and $F(\mathbf{x}, y; W)$ is the cosine distance between the embedding $\mathbf{x}$ and the corresponding semantic embedding of class label y.

To experiment with non-linear compatibility We implemented a simplified version of CMT [22] that follows our framework. We use the same network architecture described in [22], but with using pre-computed doc2vec text embeddings. We refer to this model as CMT* in what follows.

5.2. Zero-shot learning with webly-labeled data

We propose three extensions of ZSL techniques using non-linear compatibility for our instance recognition problem.

COS. The COS model is a modification of CMT* using three hidden layers with 2048, 1024 and 512 units instead of the single hidden layer in CMT. Moreover, ReLU activations is used instead of tanh.

COS+NLL. This model is inspired by [24]. We want visual features embedded in the semantic space by the COS model to be good for classification, and to encourage this we add a new linear layer acting as a classifier connected to the output of the last layer of the COS model. Then we add an additional negative log-likelihood loss (NLL) weighted with a factor 0.1 before adding it to the original margin loss described in Eq. (3).

COS+NLL+L2. In this model we add an L_2 normalization layer as we explained in Section 4.2 before the classifier in the COS+NLL model. This forces all visual features embedded in the doc2vec space onto a hypersphere, simplifying the work of the classifier as shown in Section 4.2.

6. Experimental results

In this section we report on a range of experiments we performed to evaluate the effectiveness of webly-labeled data for both supervised and zero-shot recognition of artwork instances.

6.1. Datasets

We used two datasets for our experiments on supervised and zero-shot instance recognition with webly-labeled data.

NoisyArt. Most experiments were performed on the *NoisyArt* dataset described in Section 3. This dataset was designed specifically to experiment with webly-labeled data for both supervised instance recognition and zero-shot recognition scenarios.

CMU-Oxford Sculptures. In addition to *NoisyArt*, we also experimented on CMU-Oxford Sculptures [32], It about 143K images of 2,197 different sculptures. We chose this dataset because it is another artwork instance recognition problem, although in this case without label noise. We use this dataset only for supervised instance recognition experiments and generated splits different from the original: training, validation, and test sets now contain all the classes, but different images. Our new splits for CMU-Oxford Sculptures has about 74K, 33K and 37K images for the training, validation and test, respectively.

6.2. Supervised artwork recognition with webly-labeled data

For the artwork recognition task we trained the shallow networks described in Section 4 with the Adam optimizer [33] for 500 epochs on the 200-class training set using a learning rate of 10^{-4}. Visual features are extracted using ResNet-50/101/152 and VGG16/19 on NoisyArt and CMU-Oxford Sculptures after resizing the shortest edge to be 255 and extracting the center 224×224 pixel crop and normalizing RGB channels to have zero mean and unit standard deviation. Features are passed to the shallow network which consists of a single hidden layer with 4096 hidden units. When active, the L_2 normalization layer is used instead of the hidden layer activation.

Classification results are given in Table 2 in which we also report results from [7]. Due to space limitations we only report results for a single α value. Each experiment is repeated 5 times with different seeds, in the table the mean and standard deviation per each metric are reported.

From the table we can draw few conclusions. The performance gap between *NoisyArt* test and validation when using the baseline is evidence of domain shift, while for CMU-Oxford Sculptures – which has no domain shift – they achieve similar performance. Moreover, for *NoisyArt* using L_2 normalization yields huge improvement, while in CMU-Oxford Sculptures the improvement is much lower (but still significant). Finally, notw how the performance gap for the best performing model on *NoisyArt* and

Table 2
Recognition accuracy (acc) and mean average precision (mAP) on *NoisyArt*. BL refers to the baseline network (Section 4.1) and BS refers to the noisy label mitigating approach of [7]. Numbers are averages over 5 runs with tandard in parenthesis when available.

	NoisyArt				CMU-Oxford-Sculptures			
	test		validation		test		validation	
	acc	mAP	acc	mAP	acc	mAP	acc	mAP
RN50 BL	64.80 (.31)	51.69 (.35)	76.14 (.22)	63.08 (.28)	83.32 (.19)	66.78 (.53)	83.39 (.09)	66.91 (.52)
RN50 BS [7]	68.27	57.44	75.98	62.83				
RN50 $\alpha = 0.4$	74.89 (.48)	62.86 (.89)	77.14 (.14)	63.71 (.42)	86.02 (.04)	71.78 (.40)	86.01 (.05)	71.05 (.43)
RN101 BL	64.96 (.47)	52.21 (.50)	75.37 (.13)	62.10 (.19)	83.76 (.18)	66.87 (.36)	83.86 (.04)	67.90 (.38)
RN101 BS [7]	68.27	57.41	76.78	63.46				
RN101 $\alpha = 0.4$	74.53 (.12)	62.55 (.22)	77.05 (.14)	63.56 (.31)	86.34 (.16)	71.81 (.32)	86.66 (.07)	72.80 (.44)
RN152 BL	64.28 (.76)	52.05 (.74)	75.31 (.19)	62.37 (.23)	84.12 (.10)	68.11 (1.1)	84.25 (.05)	68.53 (.45)
RN152 BS [7]	67.38	55.81	76.22	62.90				
RN152 $\alpha = 0.4$	75.04 (.23)	62.75 (.36)	79.03 (.14)	66.55 (.17)	86.85 (.06)	73.66 (.56)	86.90 (.02)	73.36 (.31)
VGG16 BL	64.37 (.32)	50.71 (.21)	74.25 (.27)	60.10 (.41)	78.19 (.13)	58.31 (.80)	78.25 (.12)	58.23 (.62)
VGG16 BS [7]	66.27	52.52	74.38	60.07				
VGG16 $\alpha = 1.0$	68.47 (.81)	55.32 (.69)	74.94 (.18)	61.34 (.32)	82.59 (.18)	66.15 (.53)	82.47 (.07)	65.93 (.26)
VGG19 BL	62.07 (.50)	48.14 (.73)	73.73 (.12)	59.62 (.30)	78.51 (.25)	59.72 (1.1)	78.53 (0.18)	58.98 (1.0)
VGG19 BS [7]	63.99	51.14	72.63	58.21				
VGG19 $\alpha = 1.0$	66.25 (.79)	53.05 (.42)	74.49 (.14)	60.42 (.17)	82.29 (.08)	64.91 (.42)	82.50 (.06)	65.48 (.48)

Table 3
Zero-Shot recognition accuracy for NoisyArt.

	Accuracy				Mean Average Precision			
Images	V	W	VW	VWC	V	W	VW	VWC
ResNet50								
SJE [21]	10.70	15.49	7.04	13.52	17.55	14.13	16.06	14.72
EsZSL [31]	12.11	15.21	14.37	25.63	20.48	18.68	22.29	29.89
ALE [19]	14.08	14.08	15.49	22.54	21.43	16.90	18.28	34.99
DEVISE [20]	16.62	14.93	16.90	24.79	22.63	19.18	20.95	31.90
CMT* [22]	19.44	13.24	15.21	21.13	21.53	19.09	24.02	43.72
COS	20.56	18.03	16.62	26.48	26.02	17.84	26.05	43.94
COS+NLL	14.65	15.77	16.06	26.20	27.10	23.53	25.36	44.70
COS+NLL+L2	18.31	8.45	18.03	34.93	24.81	21.26	25.36	45.53
ResNet152								
SJE [21]	10.70	15.49	7.04	13.52	17.55	14.13	16.06	14.72
EsZSL [31]	20.28	14.08	17.75	26.48	24.19	19.52	23.94	29.36
ALE [19]	17.18	13.52	14.37	21.69	24.54	19.65	20.79	33.93
DEVISE [20]	14.37	15.77	17.18	22.54	23.02	19.32	22.44	32.49
CMT* [22]	21.13	12.39	15.77	22.82	25.23	19.63	20.41	37.02
COS	17.75	11.55	16.34	27.04	26.51	17.9	23.04	40.18
COS+NLL	20.56	14.93	18.59	27.32	24.32	25.48	25.67	41.82
COS+NLL+L2	18.31	12.96	17.75	29.58	27.80	20.61	29.14	48.17

CMU-Oxford Sculptures is high (about 10%). We think this gap is due to the small number of examples per class in *NoisyArt* compared to CMU-Oxford Sculptures along with the intrinsic noise in webly-labeled images.

6.3. Zero-shot recognition with webly-labeled data

We trained the models from Section 5 on *NoisyArt* using three-fold cross validation: we split the 200 verified classes into 150 for training/validation and 50 for zero-shot test classes. Test and validation sets only contain human-verified images, while training set can exploit webly-labeled images.

For testing, we again train each network from scratch on the combined training and validation sets (150 classes) using the early stopping epoch computed during cross validation. Each experiment is repeated four times with different training data:

- **V**: verified images from the training classes;
- **W**: webly-labeled images from the training classes;
- **VW**: both verified and webly-labeled images; and
- **VWC**: all the images of **VW** together with all the images from the 2,920 webly-labeled classes.

The results for zero-shot recognition are shown in Table 3. Note how adding webly-labeled images to the fully-verified classes does not always improve recognition performance. However, adding new classes containing only webly-labeled images (together with a single semantic vector for each class) greatly improves results, especially for non-linear techniques.

One of our goals was to understand if the additional webly-labeled images and classes containing only webly-labeled images can help zero-shot recognition performance. We trained the COS+NLL+L2 and CMT* networks several times, gradually increasing the number of webly-labeled classes in each run. For this experiment we used the test set as validation, computing the performance for the best-performing epoch. Results are shown in Fig. 3.

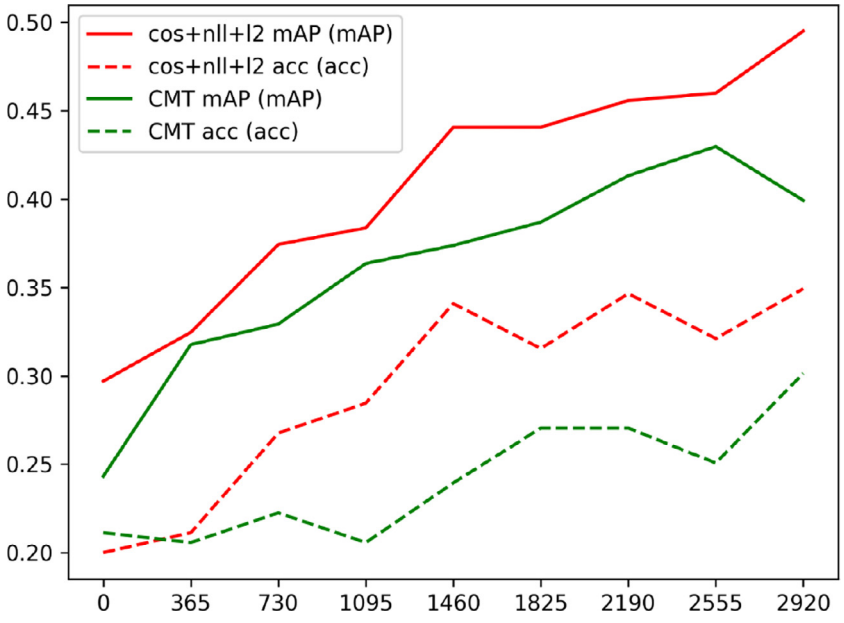

Fig. 3. Performance of COS+NLL+L2 and CMT* with increasing numbers of fully webly-labeled classes. Performance is an upper since we report the best performing epoch on the test set.

Note the rapid increase in mAP values for both models in the first half of the runs (until about 1460 additional classes) passing from 0.24 to 0.37 for CMT* and from 0.30 to 0.44 for COS+NLL+L2. The next 1,460 additional classes increase performance, but the growth is slower.

7. Conclusions and future work

In this paper we described two approaches to exploit web data for artwork instance recognition. Our results on artwork recognition show that shallow classifiers trained on features extracted with pretrained CNNs over webly-labeled images can be effective at artwork instance recognition. Using relatively simple networks and compact image features, our classifiers achieve nearly 80% classification accuracy – a significant improvement over previous results. Key to achieving this performance is treating webly-supervised artwork recognition as an *instance* recognition problem and using L_2 normalization layer before classification. This simple technique, in the case of both *NoisyArt* and CMU-Oxford Sculptures, leads to significant improvement even over more complicated noise mitigation techniques.

Cultural heritage applications involving artwork recognition have the advantage that semantically rich, textual descriptions are abundantly available. These can be exploited with a minimal effort using webly-labelled data and a Zero-shot learning approaches. Experiments show how, despite the noisy supervision, a large set of additional classes can improve zero-shot recognition for this kind of problem – especially when using L_2 normalization to compensate for domain shift.

Declaration of Competing Interest

The authors have no affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgments

The authors wish to thank NVIDIA for the generous donation of GPUs. This work was partially supported by the project ARS01_00421: “PON IDEHA - Innovazioni per l’elaborazione dei dati nel settore del Patrimonio Culturale.”

References

[1] H. Chen, I. Rahman, Cultural tourism: an analysis of engagement, cultural contact, memorable tourism experience and destination loyalty, Tour. Manag. Perspect. 26 (2018) 153–163.
[2] B. Valtysson, Europeana: the digital construction of Europe’s collective memory, Inf. Commun. Soc. 15 (2) (2012) 151–170.
[3] R. Cucchiara, C. Grana, D. Borghesani, M. Agosti, A.D. Bagdanov, Multimedia for cultural heritage: key issues, in: Multimedia for Cultural Heritage, Springer, 2012, pp. 206–216.
[4] F. Temmermans, B. Jansen, R. Deklerck, P. Schelkens, J. Cornelis, The mobile museum guide: artwork recognition with eigenpaintings and surf, in: Proceedings of the 12th International Workshop on Image Analysis for Multimedia Interactive Services, 2011.
[5] S. Sukhbaatar, R. Fergus, Learning from noisy labels with deep neural networks (2014). arXiv: 1406.2080.
[6] Y. Xian, C.H. Lampert, B. Schiele, Z. Akata, Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly, IEEE Trans. Pattern Anal. Mach. Intell. (2018).
[7] R.D. Chiaro, A. Bagdanov, A.D. Bimbo, {NoisyArt}: a dataset for webly-supervised artwork recognition, in: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2019.
[8] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: lessons learned from the 2015 mscoco image captioning challenge, IEEE Trans. Pattern Anal. Mach. Intell. 39 (4) (2017) 652–663.
[9] T. Mensink, J. Van Gemert, The Rijksmuseum challenge: museum-centered visual recognition, in: Proceedings of International Conference on Multimedia Retrieval, ACM, 2014, p. 451.
[10] N. Westlake, H. Cai, P. Hall, Detecting people in artwork with CNNs, in: Computer Vision – ECCV 2016 Workshops, 2016, pp. 825–841.
[11] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G.M. Farinella, Egocentric visitors localization in cultural sites, J. Comput. Cult. Herit. (JOCCH) 12 (2) (2019) 11.
[12] S. Karaman, A.D. Bagdanov, L. Landucci, G. DAmico, A. Ferracani, D. Pezzatini, A. Del Bimbo, Personalized multimedia content delivery on an interactive table by passive observation of museum visitors, Multimed. Tools Appl. 75 (7) (2016) 3787–3811.
[13] R. Barandela, E. Gasca, Decontamination of training samples for supervised pattern recognition methods, in: Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Springer, 2000, pp. 621–630.
[14] C.E. Brodley, M.A. Friedl, Identifying mislabeled training data, J. Artif. Intell. Res. 11 (1999) 131–167.
[15] X. Chen, A. Gupta, Webly supervised learning of convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1431–1439.
[16] R. Raguram, C. Wu, J.-M. Frahm, S. Lazebnik, Modeling and recognition of landmark image collections using iconic scene graphs, Int. J. Comput. Vis. 95 (3) (2011) 213–239.
[17] Y. Li, D.J. Crandall, D.P. Huttenlocher, Landmark classification in large-scale image collections, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1957–1964.
[18] F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. Farinella, Egocentric point of interest recognition in cultural sites, in: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2019, pp. 381–392, doi:10.5220/0007365503810392.
[19] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for image classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (7) (2015) 1425–1438.
[20] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al., Devise: a deep visual-semantic embedding model, in: Advances in Neural Information Processing Systems, 2013, pp. 2121–2129.
[21] Z. Akata, S. Reed, D. Walter, H. Lee, B. Schiele, Evaluation of output embeddings for fine-grained image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2927–2936.
[22] R. Socher, M. Ganjoo, C.D. Manning, A. Ng, Zero-shot learning through cross–modal transfer, in: Advances in Neural Information Processing Systems, 2013, pp. 935–943.
[23] M. Bucher, S. Herbin, F. Jurie, Improving semantic embedding consistency by metric learning for zero-shot classification, in: European Conference on Computer Vision, Springer, 2016, pp. 730–746.
[24] N. Hussein, E. Gavves, A.W. Smeulders, Unified embedding and metric learning for zero-exemplar event detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1096–1105.
[25] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, DBpedia-a crystallization point for the web of data, Web Semant. 7 (3) (2009) 154–165.
[26] P.N. Mendes, M. Jakob, A. García-Silva, C. Bizer, DBpedia spotlight: shedding light on the web of documents, in: Proceedings of the 7th International Conference on Semantic Systems, ACM, 2011, pp. 1–8.
[27] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International Conference on Machine Learning, 2014, pp. 1188–1196.
[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, (2014) arXiv:1409.1556.
[30] R. Ranjan, C.D. Castillo, R. Chellappa, L2-Constrained softmax loss for discriminative face verification, (2017) arXiv:1703.09507.
[31] B. Romera-Paredes, P. Torr, An embarrassingly simple approach to zero-shot learning, in: International Conference on Machine Learning, 2015, pp. 2152–2161.
[32] D.F. Fouhey, A. Gupta, A. Zisserman, 3d shape attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1516–1524.
[33] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization (2014). arXiv: 1412.6980.