DEEP CONVOLUTIONAL NEURAL NETWORKS FOR SENTIMENT ANALYSIS OF CULTURAL HERITAGE

M. Paolanti[2]*, R. Pierdicca[1], M. Martini[2], A. Felicetti[2], E. S. Malinverni[1], E. Frontoni[2], P. Zingaretti[2]

[1]Universitá Politecnica delle Marche, Dipartimento di Ingegneria Civile, Edile e dell'Architettura, 60100 Ancona, Italy
(r.pierdicca, e.s.malinverni)@staff.univpm.it
[2]Universitá Politecnica delle Marche, Dipartimento di Ingegneria dell'Informazione, 60100 Ancona, Italy
(m.martini, a.felicetti)@pm.univpm.it, (m.paolanti, e.frontoni, p.zingaretti)@staff.univpm.it

Commission II, WG II/8

KEY WORDS: Cultural Heritage, Sentiment Analysis, Deep Convolutional Neural Network

ABSTRACT:

The promotion of Cultural Heritage (CH) goods has become a major challenges over the last years. CH goods promote economic development, notably through cultural and creative industries and tourism. Thus, an effective planning of archaeological, cultural, artistic and architectural sites within the territory make CH goods easily accessible. A way of adding value to these services is making them capable of providing, using new technologies, a more immersive and stimulating fruition of information. In this light, an effective contribution can be provided by sentiment analysis. The sentiment related to a monument can be used for its evaluation considering that if it is positive, it influences its public image by increasing its value. This work introduces an approach to estimate the sentiment of Social Media pictures CH related. The sentiment of a picture is identified by an especially trained Deep Convolutional Neural Network (DCNN); aftewards, we compared the performance of three DCNNs: VGG16, ResNet and InceptionResNet. It is interesting to observe how these three different architectures are able to correctly evaluate the sentiment of an image referred to a ancient monument, historical buildings, archaeological sites, museum objects, and more. Our approach has been applied to a newly collected dataset of pictures from Instagram, which shows CH goods included in the UNESCO list of World Heritage properties.

1. INTRODUCTION

A monument is a structure that was built to commemorate a person or an event and that takes on historical, cultural, religious, identity importance of a people (Osborne et al., 2001). Furthermore, the monuments are among the main tourist destinations of a place, also affecting the economy of the country. In Italy, the Cultural Heritage (CH) is huge and there is an urgent need to preserve, protect, maintain security, manage without waste and above all enhance this heritage, for Italians, foreigners and future generations. For this purpose it is essential to have an evaluation of the cultural heritage and its state, which cannot be built without good documentation (Szmelter, 2013). We can include the sentiment related to the monument in the evaluation, considering that this affects its public image and increases its value. Then it is necessary that the aforementioned information has a specific structure in order to be correctly maintained, successfully searched, and used with ease. The digital documentation is the process that creates the documentation and is divided into two steps: first data acquisition and storage and subsequent data classification in order to create the structuring (Remondino and Rizzi, 2010). Despite the pressure from international organizations, a standardization regarding the documentation and the process has not yet been reached. In any case, among the absolute most common sources of documentation there are photographs, drawings, images of all kinds, and coming from every sensor or technology, in enormous quantities. The many images made by non-professionals but easily available on Social can be a source of complementary documentation useful from various points of view, especially if structured through Sentiment Analysis (SA).
The protection and promotion of Cultural Heritage (CH) goods are major challenges of these years (Foradori, 2017). In this respect, monuments and historical buildings constitute primary means by which tangible and intangible testimonies of nature and human cultures are safeguarded (Ferretti and Comino, 2015). These represent cultural transmission, learning, intercultural dialogue, discussion and training, also these play an important role in education (formal, informal, and lifelong learning), social cohesion and sustainable development. CH goods promote economic development, notably through cultural and creative industries and tourism. These aspects draw the attention to the importance of the promotion of CH goods and collections. Thus, an effective planning of archaeological, cultural, artistic and architectural sites within the territory make CH goods easily accessible (Colace et al., 2015).
A way of adding value to these services is making them capable of providing, using new technologies, a more immersive and stimulating fruition of information (Lops et al., 2009). In particular, tourists, visitors, citizens equipped with appropriate devices easily create and share their ideas, opinions and contents with millions of other people around the world with social networks.
In this light, an effective contribution can be provided by sentiment analysis (Pang et al., 2008). The sentiment related to a monument can be used for its evaluation considering that if it is positive, it influences its public image by increasing its value. The main purpose of SA is the identification of the agreement or dis-agreement pictures that deal with positive or negative feelings. Besides, Instagram provides a rich repository of images and captions that are associated with users sentiments (Wang and Li, 2015, Wang et al., 2015). These images do not only reflect people social lives, but also express their opinions about places, monuments and territory. Social media pictures represent a rich source of knowledge to understand users opinions.

*r.pierdicca@staff.univpm.it

This work introduces an approach to estimate the sentiment of social media pictures CH related. The sentiment of a picture is identified by an especially trained Deep Convolutional Neural Network (DCNN).
The DCNN is based on a VGG16 network architecture (Simonyan and Zisserman, 2014) and it is trained by fine-tuning a model pretrained using the ImageNet dataset (Krizhevsky et al., 2012), (Pierdicca et al., 2018). Fine tuning, which is a very simple transfer learning method, is implemented by exploiting the pretrained neural network for any other task to initialize the weights of the network, except the last ones, instead of the random initialization. Furthermore, we compare the performance of the VGG16 network with other DCNNs used for image classification: ResNet (He et al., 2016) and InceptionResNet (Szegedy et al., 2017). The evaluation has been performed in Keras, a TensorFlow-based python module, which allows the implementation of neural network models and training them. Since the templates are pre-trained on 1000 classes, it was necessary to modify the last layers of the networks before proceeding with the fine tuning.
The approach has been applied to a newly collected dataset "Cultural HeRItage Sentiment" (CHRIS) Dataset of daily news pictures from Instagram, which show CH goods included in the UNESCO list of World Heritage properties. To the best of our knowledge, it is the first study on sentiment analysis of CH-related pictures on Instagram. We utilized the captions of the Instagram posts to pre-select images that have detectable sentiment content about well-known CH good (such as Tour Eiffel, Colosseum and so on). Typically, the image captions indicate the users sentiment for the uploaded images. To obtain the ground truth of the collected pictures, the true sentiment has been manually estimated by human annotators, thus providing a more precise and less noisy dataset compared to automatically generated labels from image captions or hashtags. Since sentiment estimation is a subjective task where different people may assign different sentiments to images, we asked two persons to judge the sentiment of the images and measured their agreement. The inter-annotator-agreement is a common approach to determine the reliability of a dataset and the difficulty of the classification task (Bhowmick et al., 2008), (Paolanti et al., 2017).
The paper is organised as follows: Section 2 is an overview of the research status of sentiment analysis approach for cultural heritage field; Section 3 introduces more specifically our approach, describing the dataset purposely created, the CNNs visual classifier and also the performance evaluation metrics used for analysing the ability of the three different CNNs employed. Section 4 shows the results obtained; and Section 5 discusses the conclusions and future works.

2. RELATED WORKS

The sentiment analysis, or mathematical/statistical methods that analyze information, aims to quantify the intensity (positive, negative and neutral) of a sentiment described, in our case, in an image posted in a social network. The added value of sentiment analysis with respect to the usual customer satisfaction techniques lies in the fact that it is about listening to emotions that are spontaneously provided and for this reason they reflect the real expectations and moods of users. SA makes it possible to transform the immense amount of data generated by social media into knowledge. The analysis of the perception of a CH by citizenship becomes, therefore, an unparalleled tool of territorial marketing and a political indicator that can guide and direct choices and actions. The data sources analyzed during the study are innovative and alternative to those on which the classic customer satisfaction methods are based.
According to (Campos et al., 2017), the visual sentiment analysis is a growing area of research since images by capturing moments related to deep feeling have became an important component of our digital social life. In the work, they used the Twitter dataset collected, called DeepSent, to train and evaluate a CaffeNet CNN architecture to recognize the polarity (positive and negative) of a visual sentiment.
In (Zheng et al., 2017) the authors investigate whether and which types of objects are most responsible for evoking emotions and feelings within an image. It occurs if the sentiment corresponding to the "regions of salience" agrees with the general sentiment expressed by the entire image. The dataset is partitioned by dividing the image according to meta-attributes at the scene level such as indoor-outdoor, natural-manmade, face-noface, and more, before providing the sentiment classification in only two positive-negative categories. The value of accuracy is over 82% using pre-trained CNN for image recognition. In conclusion, the attributes that tend to dominate the perception of sentiment by the subject are first of all the faces, meaning that in many cases we can consider facial expressions without paying attention to other objects within the image.
In (You et al., 2015), the authors have explored different application of CNN for visual sentiment prediction and then have presented a CNN for the same purpose, demonstrating that their proposed is better than the state-of-art architectures. They propose a training method called progressive learning, which addresses the problem of data noise, filtering them from the training set. The basic idea is that in the SA each class contains images that are so different that it is extremely difficult to discover features that can characterize the classes, and more people could have opposing sentiments for the same image. For this it becomes necessary a supervised learning engine able of tolerating a significant level of noise in the training set, such as the progressive one.
In (Baldoni et al., 2012), the idea is to use the "game with a purpose" paradigm as a source of crowdsourcing annotation, in which users, as a side effect of the game, perform the annotation work. This strategy goes alongside the more common one that exploits the famous Amazon Mechanical Turk crowdsourcing platform, where workers can be took as annotators. It also refers to the problem of background knowledge: the high level of abstraction required by the SA may require basic knowledge, context and history, which go beyond the strict visual content.
In (Yao et al., 2016), a fundamental point of the CNN is asserted: they can well perform the direct mapping between vision and sentiment, by deducing directly the sentiment from the visual content, without the need to construct mid-level representations, which are provided in automatic way from the networks themselves. Using the same dataset, they compare the performance of three architectures: GoogleNet, VGGNet and RESNet, demonstrating that the latter works better than the first two, even if the specific dataset plays a fundamental role.
In (Jindal and Singh, 2015) a sentiment rating based on seven "votes" which include the neutral and different degrees of intensity of sentiment of the same polarity was proposed, on the idea that the strength of sentiment is as important as polarity, beyond that better than the fine-grained categorization. The authors have trained an Alexnet with 5 convolutional layers and 3 fully connected through a strategy that has led to an accuracy of 53.5%.
From the point of view of visual sentiment for cultural heritage we cited some works that have been an useful guide for our work. In (Saini et al., 2017), the authors achieved an accuracy of 92.7% by training a network to recognize 100 particular Indian mon-

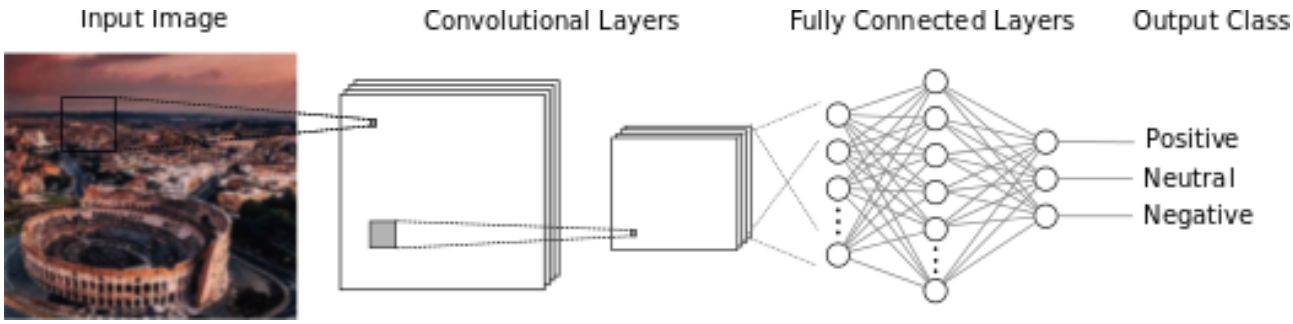

Figure 1. DCNNs for sentiment classification.

uments through a dataset of 5,000 image. The great variety of background images, points of view, monuments architecture, and more, is a noise that degrades the accuracy. To solve this problem, the authors used clean images that are manually cut to minimize noise. To capture the different characteristics of the data, they then trained 3 CNNs of Alex-Net architecture, concatenating the outgoing representations from all 3 into a single final vector. The best result was achieved by excluding the FC7 layer from the architecture.

In (Llamas et al., 2017) the authors illustrate the importance of digital documentation for architectural cultural heritage. They have collected a dataset of 10,000 images to be grouped according to the constituent elements of the monuments, and therefore they have created others with the same images but scaled according to different dimensions to estimate the compromise between performance an d time. Comparing the performance of full training and fine tuning on different networks, using the fine tuning of the Inception-ResNet-v2 network the best accuracy of 93% was achieved.

In (Campos et al., 2017) the authors show views of the local patterns learned from the network associated with sentiment in order to see how sentiment is perceived by the model. They have purposely introduced ambiguous annotations in order to develop a model that is reliable even with this kind of noise on the labels. They also analyzed the impact of initialization of weights by varying the source domain within transfer learning.

3. MATERIALS AND METHODS

The huge amount of images shared on Instagram, together with its recent widespread expansion in Italy has led us to use images of this social network for the CH. Users upload an image, write a caption to their liking that often describes an experience or opinion, and finally tag it with some hashtags and publish it on their profile. There are also generic profiles that present themselves as disseminators of a particular theme. The hashtags are now very popular and there is no shortage of hashtags inherent in cultural heritage, such as hashtags related to emotions and feelings. They can be generic as well as specific to the single monument. Finally there are also locations of the monuments, which translate into the unique IDs that Instagram assigns to all the images tagged with a certain geotag. In this section, we introduce the framework as well as the novel social media CHRIS dataset collected for this work and used for evaluation. The framework is depicted in the Figure 1. Further details are given in the following subsections.

3.1 The CNNs for visual sentiment classifier

Convolutional neural network (CNN or ConvNet) is one of the most common algorithms for deep learning, a type of machine learning in which a computer model learns to perform classification activities directly from images, videos, text or sounds. CNNs are particularly useful for finding patterns in images to recognize objects, faces and scenes. They learn directly from image data, using patterns to classify images and eliminating the need for manual feature extraction. Moreover, CNNs can be re-trained for new recognition activities, allowing existing networks to be exploited. A CNN can have tens or hundreds of layers, each of which learns to detect the different features of an image. For each image, in different resolutions, filters are applied and the output of each image is used as input for the next layer. Filters can be initially very simple features, such as brightness and edges, to take on increasingly complex shapes that uniquely define the object. Like other neural networks, a CNN is composed of an input layer, an output layer and various layers hidden in the middle. These layers perform operations that alter the data in order to learn the specific features of the data. Three of the most common layers are: convolution, activation or ReLU and pooling.

- Convolution: applies a series of convolutional filters to the input images, each of which activates certain features of the images.
- Rectified Linear Unit (ReLU): enables faster and more effective training by mapping negative values to zero and maintaining positive values. Only activated features are passed to the next layer.
- Pooling: simplifies output by performing non-linear subsampling, reducing the number of parameters the network needs to learn.

These operations are repeated on tens or hundreds of layers, each of which learns to identify different features. After learning the features in various layers, the architecture of a CNN moves to classification. The penultimate layer is a fully connected layer that generates a vector of K dimensions where K is the number of classes that the network will be able to predict. This vector contains the probabilities for each class of any classified image. The final layer of the CNN architecture uses a classification layer like softmax to provide the classification output.

3.1.1 VGG-16 network The first CNN we considered is a VGG-16 net (Simonyan and Zisserman, 2014). The VGG-16 network is chosen because easy to implement and its success in the ILSVRC-2014 competition where it placed first in the *2a* challenge. Originally VGG-16 is trained on the ImageNet database consisted of of labeled images in 1000 classes (Krizhevsky et al., 2012) and is a very deep, 16-convolutional-layer network. The VGG-16 network consists of 5 convolutional blocks where corresponding output filter have [54, 128, 256, 512] dimensions and

a fully-connected classifier. A characteristic element of the VGG architecture is also represented by the introduction of 1x1 size convolution filters, which in general are used to make a reduction in the space of the channels while leaving unaltered the dimension of the output; the meaning of this operation lies in the fact that since each convolution layer is followed by a ReLU layer, the insertion of these filters allows to model further non-linear effects. We developed the VGG16 model in Keras, a high-level neural networks library and written in Python. Our implementation for VGG16 net refers to the work in (Simonyan and Zisserman, 2014). The image is resized to 224×224 pixel. Since there is not batch normalization layer in VGG16, input images are been normalized. We use the stochastic gradient descent (SGD) optimizer with a batch size of 5. After preliminary experiment the learning rate is very small and fixed to 10^{-5}. The network is trained with a binary cross entropy (BCE) by training for 100 epochs.

3.1.2 ResNet network The second network are the residual networks (ResNet) (He et al., 2016), an innovation in the field of convolutional neural networks, winner of the ImageNet competition in 2015, with a top-5 error of 3.75%. The key question that the developers asked themselves, thinking about the model of this network was: why does each deep network have a worse performance as layers are added? The hypothesis that the authors of ResNet did, was that direct mappings are difficult to learn. So they proposed a change: instead of trying to estimate a function $G(\cdot)$ that given an x returns $G(x)$, it is better to learn the difference between the two, also called residual, hence the name of the network. Consequently, to calculate $G(x)$ starting from x its residue must be added: $F(x) = G(x) - x$. $F(x)$ is the residual, and therefore the network instead of directly learning $G(x)$, will try to learn $F(x) + x$. Thus the ResNet blocks (residual network) were introduced into the network. Each ResNet block consists of a series of layers and an identity mapping that adds the input of the block to its output. This "addition" operation is done element by element and if the input and output have different sizes, zero-padding or projection techniques (through 11 convolutions) can be used to create corresponding dimensions. Therefore, ResNet provides layers with a reference point x from which to learn, rather than starting from zero with an identity transformation. Before this idea, in the deep neural networks there was a problem of cancellation of the gradient, whose descent, given by the minimization of the error function, is reduced exponentially through the retropropagation of the previous layers. In essence, the long road through the previous layers made the signs of errors so small that they did not allow the network to learn. Thanks to the innovation introduced by ResNet it is now possible to build networks of innumerable layers (potentially even more than a thousand) with a high degree of accuracy. In our work, after preliminary experiment ResNet has a learning rate small and fixed to 10^{-6}. The network with a batch size of 16 is trained for 50 epoch.

3.1.3 Inception network If ResNet focuses on depth, Inception network is instead focused on the extension. For Inception, the authors were interested to the computational efficiency of the larger networks learning. The original work concerned a component known as the Inception model. An Inception module processes multiple and different transformations on the same input data simultaneously, linking the results into a single output. In other words, each Inception module layer consists of a 55 convolution, a 33 and a max-pool. The selection of the most relevant features is left to the next layer. However, the greater information density of the architecture in this model had a great problem, namely the drastic increase in the necessary computational capacity. Not only the wider convolutional filters (55) are demanding and expensive to calculate, but also the overlapping of different filters increases the number of feature maps for each layer. This is a real impediment for the model. The authors of Inception have used 11 convolutions to filter (or reduce) the depth of the outputs. These convolutions take into account one value at a time, but through multiple channels, can extract spatial information and compress them to a smaller size. By reducing the number of input maps, the authors were able to overlap different transformations in parallel of layers, with the result of having networks that were at the same time deep (with numerous layers) and wide (with parallel operations). The first version of Inception, called "GoogLeNet" is composed by 22 layers and won the 2014 ILSVRC competition. Inception v2 and v3, developed a year later, are better than the previous version for several factors: the most relevant it is the refactoring of large convolutions into smaller and consecutive ones, which are therefore easier to learn.

3.2 Cultural HeRItage Sentiment (CHRIS) Dataset

First of all we chose to classify the sentiment according to the most common subdivision into 3 classes: positive, neutral and negative. The definition of the sentiment has been affected by the characteristics of the dataset that has been achieved. Normally we look for images according to the characteristics we want to analyze, while in this case we have adapted the definitions of positive and negative based on the majority of the images obtained. The next two phases were: the **subjectivity classification** and the **polarity classification**, both carried out by manually selecting the images.
The first phase, once the dataset has been acquired, has resulted in the attempt to define the "neutral", or to separate neutral examples from those containing any kind of sentiment. Through this phase we have tried to limit the problem of the subjectivity of the researcher in the assessment of sentiment: not having access to crowdsourcing platforms, in fact each phase was completed by a single person and the images that are in uncontrolled conditions, embedded, and non-iconic scenes were not an advantage. In the second phase, positive examples were finally separated from negative examples, among those that were most common in the images. The results have been divided into three categories and also figure 2 shows three examples of pictures in the CHRIS Dataset:

- positive: selfies, tricolor arrows, fireworks, individuals and groups posing for the photo or in the act of photographing, soap bubbles, bright Christmas decorations, fresh flowers in the foreground, flags deployed or flying, large crowd (to the edge image), light projections on the monument, kisses, rainbows, objects that "imitate" the monument.

- negative: rain and therefore crowd with umbrellas or reflections of the monument in the puddles on the ground, snowfalls, smoke, the "head in the clouds" phenomenon for the towers or when the point disappears above the clouds or the fog, planes flying near the towers, images with tram wires in the foreground, scaffolding on the monument or on adjacent ones, queues of people or traffic, military parades and law enforcement in general, red cross, barriers, bottles of alcohol, dirt on the ground, demonstrations and disorderly protests. A special case for the tower of Pisa: poses that portray kicks or obscene behavior towards the tower.

- neutral: close-up fragments of the monument, large-scale images, regular life scenes around the monument, and above all the co-presence of positive and negative sentiment in the

same image. A case apart this time it was considered for the Victorian: the monument guards were considered negative if armed, but neutral if not armed.

(a) Positive (b) Neutral (c) Negative

Figure 2. Images from CHRIS Dataset. Figure 2a is an example of positive image, Figure 2b represents an image with neutral sentiment, and Figure 2c is a picture with negative sentiment.

The following phase of preparing the dataset concerned the process of acquiring and labeling the dataset. Before defining the guidelines stably, it was decided to verify the set and the types of images circulating on Instagram by downloading from profiles, hashtags and locations. For hashtag search most popular on the social network, the help of the keywordtool[1] online tool was used. At first, images were downloaded via generic hashtags, then generalizing to hashtags that have the name of the monument many images were obtained, but with a problem related to the difficulty of finding negative examples. To overcome this problem, geotags have been downloaded for some monuments, obtaining better results. the advantage is due to the fact that with this strategy it is possible to obtain photos that do not have hashtags of the monument and therefore otherwise unavailable. We initially obtained a collection of more than 17500 images, considered suitable, filtered only from those examples that had a multiframe structure and then the annotative phase was started, carried out as yet said in a first part of subjectivity classification and in a second one polarity classification. The biggest problem in the annotative phase was to consider very detailed guidelines, which led to the loss of consistency between the classes. The biggest problem in the annotative phase was to consider very detailed guidelines, which led to the loss of consistency between the classes. The resulting dataset was rich but strongly unbalanced. By slightly changing the guidelines, many images ranged from positive to neutral sentiment and vice versa but the real problem remained the negative examples. Only negative examples were acquired through hashtags and keywords in the caption. At this point some considerations were made:

- Training is better with smaller and more reliable dataset than large and ambiguous ones.
- Often the cost of missing a minority class is much greater than missing a majority class.
- Performance decreases when the task becomes too fine grained, which happens when the discriminating elements of the sentiment are too small.
- The problem of subjectivity also lies in the influence that the photo has on the human being, or when it is not classified according to objective events, but by the quality and aesthetic impact of the image.

As a result we have considered appropriate not to use all the images for training, but it was considered correct to cut out a more reliable dataset from the one acquired, discarding the examples considered more ambiguous. So in conclusion we have filled the neutral and positive class so as to balance the negative one. In this way the final dataset obtained is composed as in the table 1. The final dataset is comprised of a total amount of 8472 pictures, including 2960 pictures with overall positive sentiment, 2891 pictures with overall neutral sentiment and 2626 pictures with overall negative sentiment. We perform the experiments by splitting the labeled dataset into a training set and a test set.

Positive	Neutral	Negative	Total
2960	2891	2626	**8477**

Table 1. Final dataset.

The dataset had to be pre-processed in order for the images to adapt to the input sizes of the various networks. Some images have been deleted because, although well viewable through the operating system, not correctly recognized and processed.
Three versions of the dataset have been created, respectively from 224x224, 299x299 and 331x331 pixels. However with the larger dimensions, probably due to the memory limits imposed, the execution environment crashed so all the networks were fed with 224x224 pixels inputs.

3.3 Performance Evaluation Metrics

To evaluate the performance of the algorithms the following quantities were defined (Khoshelham et al., 2010):

- t_p (True Positive), the number of images correctly classified in the class of belonging;
- t_n (True Negative), the number of images correctly classified in another class;
- f_n (False Negative), the number of images incorrectly classified in other classes;
- f_p (False Positive), the number of images incorrectly classified in the class;

We have employed the following metrics to compare the performance of the algorithms that used for the evaluation of the image sentiment:

- *Accuracy*: indicates the effectiveness of the algorithm by showing the probability of the true value of the class (positive, neutral, negative):

$$Accuracy = \frac{t_p + f_n}{t_p + t_n + f_p + f_n} \quad (1)$$

where t_p is the number of true positives and f_n the number of false negatives.

- *Recall*: is a function of its correctly classified examples (true positives) and its incorrectly classified examples (false negatives).

$$recall = \frac{t_p}{t_p + f_n} \quad (2)$$

[1] https://keywordtool.io/instagram

- *Precision*: is a function of true positives and examples incorrectly classified as positives (false positives).

$$precision = \frac{t_p}{t_p + f_p} \tag{3}$$

- *F1-score*: is a measure of a test's accuracy.

$$F1 - score = \frac{(\beta^2 + 1) * precision * recall}{\beta^2 * precision + recall} \tag{4}$$

The F1-score is evenly balanced when $\beta = 1$. It favours precision when $\beta > 1$, and recall otherwise. The F1-score can be interpreted as a weighted average of the precision and recall.

- *Support*: is the number of occurrences of each class in ground truth (correct) target values.

The confusion matrix is used With the aim of schematizing the results of the model. In fact, the confusion matrix depicts information about actual and predicted classifications done (Provost and Kohavi, 1998). In the field of artificial intelligence, the confusion matrix, also called the wrong classification table, returns a representation of the statistical classification accuracy. Each column of the matrix represents the predicted values, while each row represents the real values.

4. RESULTS AND DISCUSSION

In this section, we report the results of the experimental phase conducted on CHRIS Dataset. The dataset comprises 8477 images containing visual elements. In particular, the dataset includes “embedded scenes”, i.e. images that in addition to the object of interest (e.g. a monument) contain other objects that directly influence sentiment (e.g. happy people, waste, vandalism, and so on). The true sentiment is not automatically judged by the accompanying texts or hash-tags but has been manually estimated by human annotators, thus providing a more precise dataset.
The experiments are based only on these images of the dataset, where both annotators have agreed on the sentiment. By removing pictures with ambiguous sentiment, we increase the quality of the dataset and ensure the validity of the experiments. The dataset is split into 80% training and 20% test images.

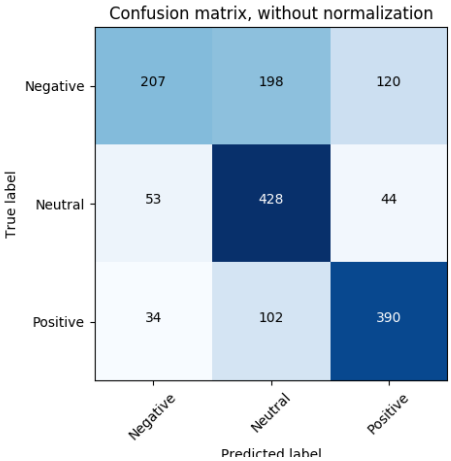

Figure 3. Confusion matrix for VGG16 network (learning rate=0.00001, batch=5, epochs=100).

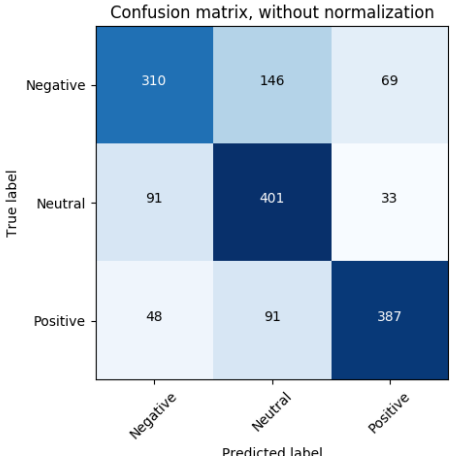

Figure 4. Confusion matrix for ResNet network (learning rate=0.000001, batch=16, epochs=50.

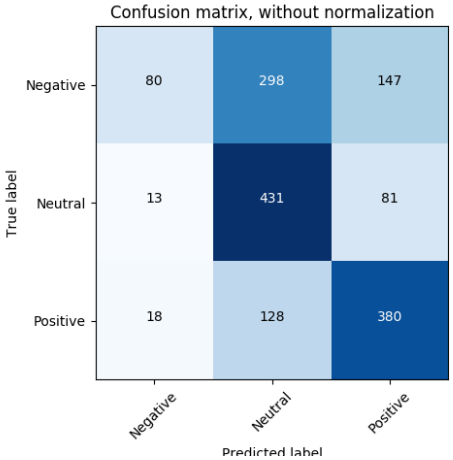

Figure 5. Confusion matrix for ResNetInceptionV2 network.

The application of our approach to this dataset yields good results in terms of precision, recall and F1-score and demonstrates the effectiveness of the proposed approach (Table 2).

DCNNs	Accuracy	Precision	Recall	F1-score
VGG16	0.65	0.67	0.65	0.66
ResNet	**0.70**	**0.70**	**0.70**	**0.70**
IncResNet	0.57	0.62	0.57	0.52

Table 2. Overall classification results comparing the three different DCNNs.

Looking at the confusion matrices in Tables 3,4 and 5 and the final accuracy in Table 2, the first consideration concerns the fact that the ResNet50 with 70% provided the best result. It must be remembered that, given its speed, more tests have been conducted on this network. However, an interpretation could be the usefulness of the skip connections in transferring less “advanced” features to the following levels. A second consideration concerns the fact that the VGG16 network has reached 65% accuracy, with a more precise dataset, albeit with a smaller number of images. This confirms the above: better smaller and more reliable datasets that are larger and noisier. The third consideration concerns preprocessing and as a demonstration shows some of the various examples that were really compromised by the crop and resize procedure, which effectively changed the sentiment. In fact, it can

happen that figures have a positive sentiment but they were classified as neutral since the pre-processing phase has filtered the discriminating property of the sentiment. The fourth consideration concerns the category that has been most highly misclassified: from all three confusion matrixes on the final dataset it is clearly seen from the highlighted cases that the majority of the errors was predicting neutral sentiment for a negative image, see for example Tables 3 and 5.
Generally, positive examples have individuals and objects in the foreground much more than the other two classes, while neutrals are distinguished mainly by not having them. Finally, the negative examples are characterized by having many discriminatory attributes for sentiment small in terms of pixels (for example barriers and non-central scaffolding in the image) of which the resizing and cropping operations have further compromised visibility, to the eye human as to the artificial network. Finally we can say that overall the networks generalize quite correctly: only in one case we saw a clear example of overfitting as figure 6 shows. Therefore, the experiment of introducing a dropout layer into the architecture did not bring any noteworthy results.

(a) VGG16 (b) ResNet (c) InceptionV2

Figure 6. Accuracy for training and validation phases.

5. CONCLUSIONS AND DISCUSSIONS

The promotion of Cultural Heritage (CH) goods is an important challenge in the last years. Many are the aspects that must be considered: cultural transmission, learning, intercultural dialogue, discussion and training, that play an important role in education (formal, informal, and life-long learning), social cohesion and sustainable development. CH goods promote economic development, notably through cultural and creative industries and tourism. A way of adding value to these services is making them capable of providing, using new technologies, a more involving and stimulating use of information.
In this paper, we introduce a deep learning approach for recognizing the sentiment of cultural heritage pictures by taking only visual information into account. The sentiment of a picture is identified by a machine learning classifier based on visual features extracted from especially trained DCNNs. The experiments on a purposely created dataset compared the performances of three DCNNs (VGG16, ResNet and ResNetInceptionV2) on the sentiment recognition. For this purpose an efficient contribution is provided by sentiment analysis. The main purpose of sentiment analysis is the identification of the agreement or disagreement pictures that deal with positive or negative sentiment. For example, instagram images also express their opinions about places, monuments and territory and so they represent a rich resource to understand users opinions. The experiments on the purposely created CHRIS Dataset yield high accuracies and demonstrate the effectiveness and suitability of our approach. To briefly summarize, the main contributions of this work are: (i) a demonstration that Deep Learning architectures can be applied for sentiment analysis evaluation of Social Media Pictures CH related; (ii) a challenging new dataset of images collected by Instagram of CH goods in the List of UNESCO, hand-labelled with ground truth; (iii) performance comparison of different DCNNs for image classification; (iv) a system that ensures the management of data with a multimedia and multidisciplinary approach through sentiment analysis techniques, to allow a CH analysis finalized to planning and distribution needs.

ACKNOWLEDGEMENTS

The authors would like to thank Nancy Amichetti who made an invaluable work for her thesis.

REFERENCES

Baldoni, M., Baroglio, C., Patti, V. and Rena, P., 2012. From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale* 6(1), pp. 41–54.

Bhowmick, P. K., Mitra, P. and Basu, A., 2008. An agreement measure for determining inter-annotator reliability of human judgements on affective text. In: *proceedings of the workshop on human Judgements in Computational Linguistics*, Association for Computational Linguistics, pp. 58–65.

Campos, V., Jou, B. and Giro-i Nieto, X., 2017. From pixels to sentiment: Fine-tuning cnns for visual sentiment prediction. *Image and Vision Computing* 65, pp. 15–22.

Colace, F., De Santo, M., Lemma, S., Lombardi, M., Rossi, A., Santoriello, A., Terribile, A. and Vigorito, M., 2015. How to describe cultural heritage resources in the web 2.0 era? In: *2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, IEEE, pp. 809–815.

Ferretti, V. and Comino, E., 2015. An integrated framework to assess complex cultural and natural heritage systems with multi-attribute value theory. *Journal of Cultural Heritage* 16(5), pp. 688–697.

Foradori, P., 2017. Protecting cultural heritage during armed conflict: the italian contribution to cultural peacekeeping. *Modern Italy* 22(1), pp. 1–17.

He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Jindal, S. and Singh, S., 2015. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In: *2015 International Conference on Information Processing (ICIP)*, IEEE, pp. 447–451.

Khoshelham, K., Nardinocchi, C., Frontoni, E., Mancini, A. and Zingaretti, P., 2010. Performance evaluation of automated approaches to building detection in multi-source aerial data. *ISPRS Journal of Photogrammetry and Remote Sensing* 65(1), pp. 123–133.

Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. pp. 1106–1114.

Llamas, J., M Lerones, P., Medina, R., Zalama, E. and Gómez-García-Bermejo, J., 2017. Classification of architectural heritage images using deep learning techniques. *Applied Sciences* 7(10), pp. 992.

Lops, P., de Gemmis, M., Semeraro, G., Musto, C., Narducci, F. and Bux, M., 2009. A semantic content-based recommender system integrating folksonomies for personalized access. In: *Web Personalization in Intelligent Environments*, Springer, pp. 27–47.

Osborne, B. S. et al., 2001. Landscapes, memory, monuments, and commemoration: Putting identity in its place. *Canadian Ethnic Studies* 33(3), pp. 39–77.

Pang, B., Lee, L. et al., 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval* 2(1–2), pp. 1–135.

Paolanti, M., Kaiser, C., Schallner, R., Frontoni, E. and Zingaretti, P., 2017. Visual and textual sentiment analysis of brand-related social media pictures using deep convolutional neural networks. In: *International Conference on Image Analysis and Processing*, Springer, pp. 402–413.

Pierdicca, R., Malinverni, E., Piccinini, F., Paolanti, M., Felicetti, A. and Zingaretti, P., 2018. Deep convolutional neural network for automatic detection of damaged photovoltaic cells. *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*.

Provost, F. and Kohavi, R., 1998. Glossary of terms. *Journal of Machine Learning* 30(2-3), pp. 271–274.

Remondino, F. and Rizzi, A., 2010. Reality-based 3d documentation of natural and cultural heritage sitestechniques, problems, and examples. *Applied Geomatics* 2(3), pp. 85–100.

Saini, A., Gupta, T., Kumar, R., Gupta, A. K., Panwar, M. and Mittal, A., 2017. Image based indian monument recognition using convoluted neural networks. In: *2017 International Conference on Big Data, IoT and Data Science (BID)*, IEEE, pp. 138–142.

Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A. A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In: *Thirty-First AAAI Conference on Artificial Intelligence*.

Szmelter, I., 2013. New values of cultural heritage and the need for a new paradigm regarding its care. In: *CeROArt. Conservation, exposition, Restauration dObjets dArt*, Association CeROArt asbl.

Wang, Y. and Li, B., 2015. Sentiment analysis for social media images. In: *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, IEEE, pp. 1584–1591.

Wang, Y., Wang, S., Tang, J., Liu, H. and Li, B., 2015. Unsupervised sentiment analysis for social media images. In: *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Yao, J., Yu, Y. and Xue, X., 2016. Sentiment prediction in scene images via convolutional neural networks. In: *2016 31st Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, IEEE, pp. 196–200.

You, Q., Luo, J., Jin, H. and Yang, J., 2015. Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Zheng, H., Chen, T., You, Q. and Luo, J., 2017. When saliency meets sentiment: Understanding how image content invokes emotion and sentiment. In: *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, pp. 630–634.