2020

Architectural heritage images classification using deep learning with CNN

Mohammed Hamzah Abed

Muntasir Al-Asfoor

Zahir M. Hussain
Edith Cowan University

Architectural Heritage Images Classification Using Deep Learning With CNN

Mohammed Hamzah Abed
Faculty of Computer Science and
Information Technology,
University of Al-Qadisiyah , Iraq
mohammed.abed@qu.edu.iq

Muntasir Al-Asfoor
Faculty of Computer Science and
Information Technology,
University of Al-Qadisiyah , Iraq
muntasir.al-asfoor@qu.edu.iq

Zahir M Hussain
Faculty of Computer Science and
Mathematics, kufa University, Iraq
School of Engineering, Edith Cowan
University, Joondalup ,Australia
 zmhussain@ieee.org

Abstract

Digital documentation of cultural heritage images has emerged as an important topic in data analysis. Increasing the size and number of images to be processed making the task of categorizing them a challenging task and may take an inordinate amount of time. This research paper proposes a solution to the mentioned challenges by classifying the subject of the image of the study using Convolutional Neural Network. Classification of available images leads to improve the management of the images dataset and enhance the search of a specific item, which helps in the tasks of studying and analysis the proper heritage object. Deep learning for architectural heritage images classification has been employed during the course of this study. The pre-trained convolutional neural networks GoogLeNet, resnet18 and resnet50 proposed to be applied on public dataset Cultural Heritage images. Experimental results have shown promising outcomes with an accuracy of "87.91", "95.47" and "95.57" respectively.

Keyword Classification, deep learning, Convolutional Neural Network and Architectural Heritage Images

1 Introduction

One of the most important aspects in the study of architectural cultural heritage documents is the diagnosis, analysis and classification of the state of monuments and buildings and thus contribute effectively to the conservation and restoration, so these documents must accurately reflect the information that can be extracted from these heritage images [Llamas 17]. The number of digital documentation of cultural heritage is increased daily because of different available sources and technology's ability to help who is working in digital documentation analysis and understanding. In addition, some of these images are taken by non-professional people and taken by the phone's camera. This kind of sources is easy to download and use but suffers from the lack of clearness without source caption and categorization, contrariwise the expert photographer. In this paper deep learning technique is proposed for digital documentation (usually colored images) classification. In this study, convolutional neural network CNN to extract the useful information and features from the images has been used to help the classification process. Nowadays, in academia and industry there is more focus on different applications of deep learning and

convolutional neural network based on images classification, such as medical images [N 16], license plate and vehicle recognition based on local tiled convolutional neural network as proposed by Yongbin Gao et.al [Gao 16], satellite and aerial images based on CNN by M. A. Kadhim et. al [Kadhim 20] and many more. Furthermore, literature has focused on cultural heritage images classification using pre-trained CNN like Alexnet, Resnet and Inception v3 [Llamas 17], finally the research proposed Gabor filter for features extraction and support vector machine for automatic architecture style recognition [Mathias 11]. In this work multi-label images [Llamas 17] are used in the experimental analysis and results to train our model to classify new test images. Also, the pre-trained CNN for high-level features extraction and classification are suggested.

2 Deep Learning

Deep structured learning is a discipline within the field of machine learning [Abed 19] based on an artificial neural network [Mustafa 19] that used multi-layer to extract high-level features from raw input data or images. Depending on the problem domain of study, the features have been selected. for instance, in image processing, the low-level features can help to identify edges while high-level features can help to identify the semantic concept of images [Llamas 17]. Most of the modern deep learning strategies are based on Convolutional Neural Network [Abed 19].

2.1. Convolutional Neural Network (CNN)

The convolutional neural network is a class of deep learning network [Abdelaziz 19], and it is inspired by an artificial neural network [Abdelaziz 19]. CNN is a typical structure designed as a series of stages formed by the layers. the first phases consist of two kinds of layers: convolutional layers and assembly layers [Harangi 18], at the end of the network's structure the classification performance of the features that extracted using fully connected layers [Harangi 18]. In this work, many convolutional neural networks have been suggested and proposed such as GoogLeNet [Szegedy 15], resnet18 [He 16] and resnet50 [He 16], tested on different. CNN is a multi-layer network structure basically consist of five layers starting from the input layer, convolutional layer, pooling layer, fully connected layer and finally output layer as shown in figure 1.
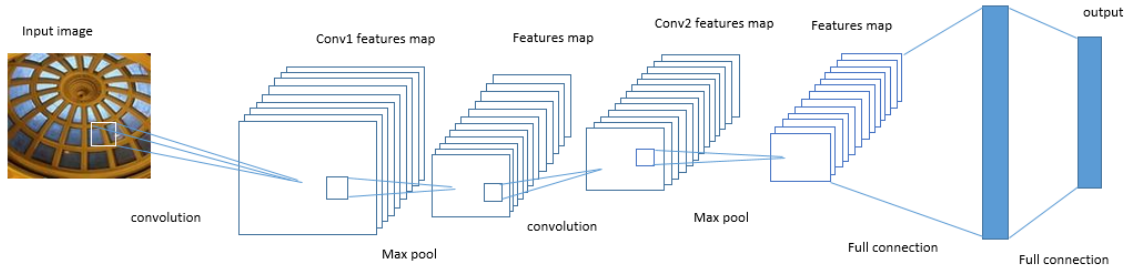

Figure 1: The basic structure of CNN

1- Input layer: it is the first layer of the Convolutional Neural Network, in general, it is the input of the whole CNN which is represented as a matrix of image's pixels [Zhang 19] .
2- Convolutional layer: this layer is responsible for the features extraction from the input matrix. The earlier convolutional layer extracts low features like edge, lines and corner. And deep level convolutional layer used for semantic features and high features extractions. The deep features extraction depends on the learning of the previous low features of entire matrix [Belhi 18] . The convolutional layer holds multiple features maps by convolving the convolution kernel of a previous layer as shown in equation 1.

$$X_j^l = f\left[\sum_{i \in M_j} (X_i^{l-1} * K_{ij}^l + b_j^l)\right] \quad \dots 1$$

Where

M_j : represents the input image.

X_i^l : represents the jth features map of the l th layer.

* : represents the convolution operation.

X_i^{l-1}: is the ith features map of the *l-1* layer

K_{ij}^l : represents the filter connecting the jth feature map of the l th layer and i^{th} features map of the *l-1* layer.

b_j^l: is the bias.

The most common activation function was sigmoid, ReLU and Tanh [Zhang 19] . The equation as shown

Sigmoid $\qquad f(x) = \frac{1}{1+e^{-z}} \qquad .... \quad 2$

ReLU $\qquad f(x) = \max(0 \cdot x) \quad .... \qquad 3$

Tanh $\qquad f(x) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \qquad .... \quad 4$

3- Pooling layer is used to reduce the dimensional features maps based on the limitations of the human visual system [Zhang 19] [Belhi 18] by downsampling the conventional maps. The efficiency of pooling layer to features reduction helps the CNN to speed up the computation process. In general, there are two kinds of pooling operations: maximum and average pooling.
4- The fully connected layer is usually used in the last layers of the CNN structure as shown in figure(1), to combine the features together in the former layers [Zhang 19] .
5- The output layer is the final layer of CNN architecture [Belhi 18] , is finally passed through the classifier, the most commonly used classifier in this part is binary classification problems and multi-classification problems such as softmax classifier as shown in equation no 5.

Softmax $\qquad \sigma(x_j) = \frac{e^{xj}}{\sum_i e^{xi}} \qquad ... \qquad 5$

Three different convolutional neural network methods have been used through this study, namely:

2.1.1. GoogLeNet: is a pre-trained convolutional neural network [Szegedy 15]. Basically, it has 22 layers deep [Szegedy 15]. It is trained over two different online datasets ImageNet [Img net] and Places365 [Zhou 16] ,[Places] . The GoogLeNet version which trained on ImageNet can classify images into 1000 categories, and the other network which trained by Places365 can classify into 365 different place categories. Both of two network has an input image size of 224 by 224. The network version that has been used in our experimental study was 144 layers starting from the input layer, convolution layer, ReLU and Max pooling this structure repeated until reach the fully connected layer, then the classifier layer occurs to give the output classification result.

2.1.2. ResNet-18 [He 16] is a pre-trained convolutional neural network, that has been trained on more than a million images from images dataset ImageNet [Img net]. the network consists of 18 layers deep which can classify images up to 1000 categories. The Network has learned rich features extracted from a wide range of images. The input images size of 224 by 224. The network design used in this research has 72 layers.

2.1.3. ResNet101 [He 16] is another version of the Resnet network architecture. Same as the previous version ResNet -18 is a pre-trained convolutional network has been trained on the same dataset ImageNet. The network depth is 347 layers, which can classify images up to 1000 categories.

The setting of structure for a pre-trained convolutional network that we used in experimental shown in table 1.

Table 1: Setting of pre-trained CNN that used.

Pre-trained CNN algorithm	No of layers	No of connection
GoogLNet	144× 1	170×2
ResNet 18	72× 1	79× 1
ResNet 101	347× 1	349×2

3. Proposed Model reused pre-trained CNN

The most commonly used in deep learning applications is transfer learning. The deep Convolutional Neural Network was training based on large scale images dataset is applied to architecture heritage images. It is focused on retrained the pre-trained convolutional neural network on a new task and on new images. Transfer learning with fine-tuning is much faster and easier than training blinded network with random initializing weight. figure 2 shows the proposed model architecture that suggested for this work.
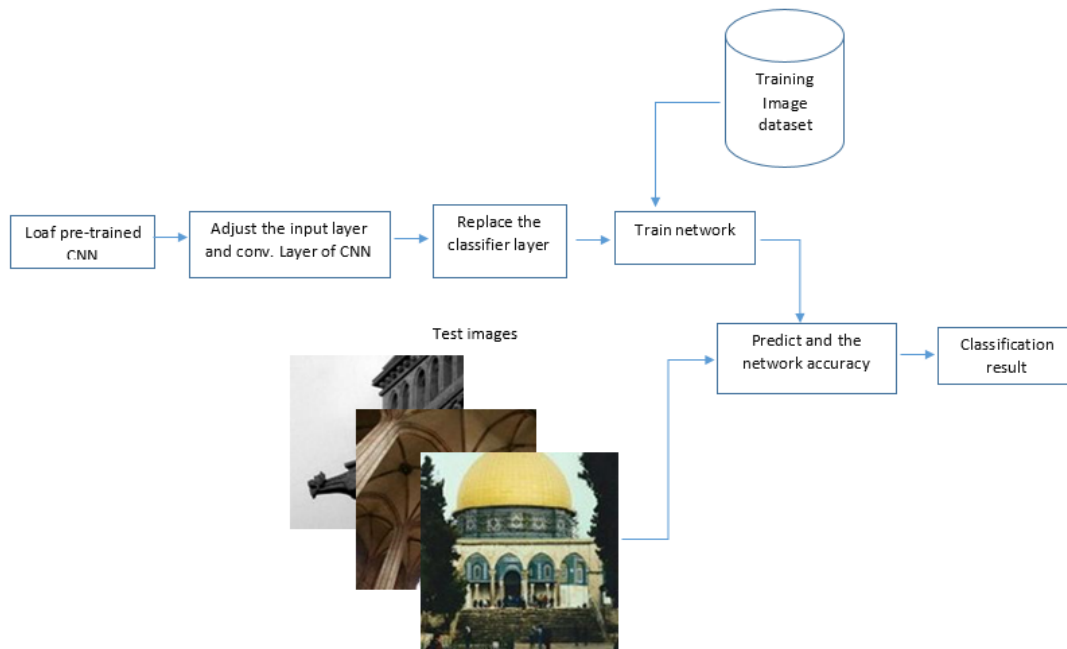
Figure 2: The proposed model

1- Load Architectural Heritage Images dataset for start training phase. in this work, the training images ratio was 70%, and testing was 30% randomly distribution from the dataset that we used in our experiment.

2- The second step in the proposed model was how to adjust and fitting the layer's filter with the image size that we used as an experimental case. Starting from the input layer by selecting the size and number of channels or color then convolutional layer till the fully connected layers. The information contains who to combine the features that extracted from all layers to be trainable features.

3- Replace the classification layer to be suitable to the number of final categories. The adaptation occurs to loss2-classifier layer and output layer. in our case ten classes in all three experimental study.

4- Training phase by load the editing pre-trained Convolutional Neural Network (GoogLNet, ResNet-18 and ResNet-101). By extracting all the features from a fully connected layer by combining the low and high features, to train the network based on these features. After train the CNN the test phase occurs to check and predicate class category of the testing images as the output, and check the accuracy of the model.

4. Experimental Result

In this section the experimental setting is introduced first, to establish the basic idea of the work.

4.1 The experimental setting

In this paper the setting of the experiment was as the following:

1- Architectural Heritage Elements Dataset (AHE_Dataset), which generated in three versions: Originally the dataset was published in two versions, first one contains images of different sizes and the second was scaled into 128×128 pixels, as well as the small dataset was created with a small size of 64×64 and 32×32. The third version was selected as a subset of each class which consists of 500 images that scaled into 224 × 224 pixels to be compatible with the pre-trained CNN. Table 2 illustrates the details of the dataset.

Table 2: Architecture heritage dataset details

Dataset	No of images	Size of images
AHE_Dataset small version	50× 10 class	64×64
AHE_Dataset original version	10235 for all classes	128× 128
Sub of AHE_Dataset	500× 10 class	224×224

    The dataset was partitioned randomly into 70% for training and 30% for validation.

2- The pre-trained Convolutional Neural Network (GoogLNet, ResNet 18 and ResNet 101) were used during the course of this research to classify the Architectural Heritage Elements Dataset. All the CNN that used were modified to be suitable for those kinds of images. The modification starts from the input layer to be able for multi-channel images, convolutional layer to select size of filters proper the size of the entire images, fully connected layer was edited to be suitable the number of categories in the used dataset, finally the output classification layer was edit based on number of classes of fully connected layer.

3- All the editing, designing and coding of pre-trained CNN training and validation are implemented by Matlab R2019a and Deep Network Designer tools.

4.2 Full training of Convolutional Neural Network

Our experimental study divided into three major parts based on the dataset used in this work. First one training AHE_Dataset small version, second sub of AHE_Dataset select 500 images from each class, finally the AHE_Dataset original version was trained. All of the versions of dataset are trained fully by three pre-trained Convolutional Neural Networks. Figure(3) shows the accuracy and loss of AHE_Dataset small version based on GoogLNet.
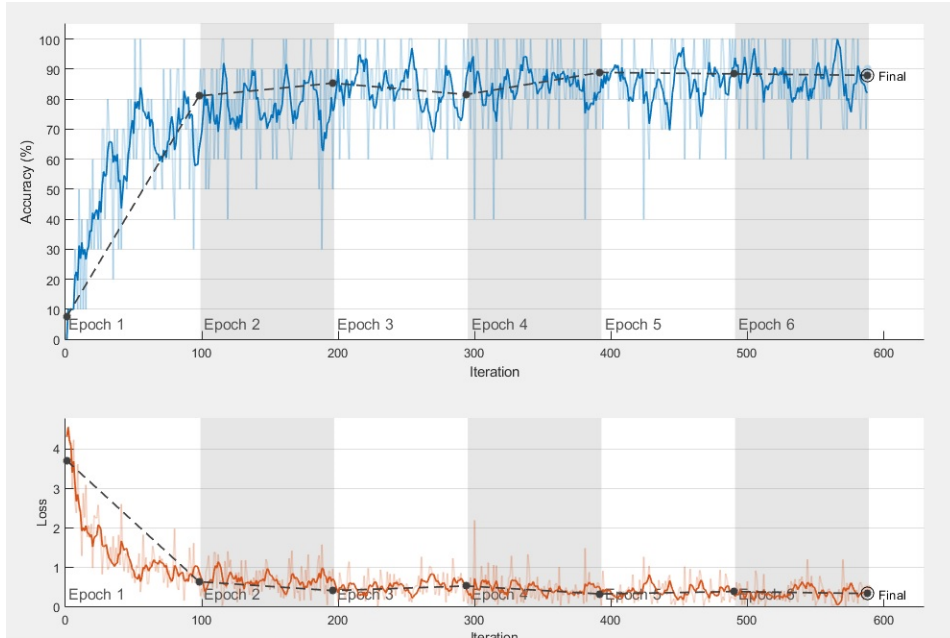
Figure 3: Accuracy and loss of AHE _Dataset small version based on GoogLNet

And the confusion matrix of the same version shows in figure (4). Furthermore, the accuracy and loss of the second version of sub-images based on ResNet-101 are shown in figure (5). The accuracy of the confusion matrix was calculated based on condition positive and condition negatively as shown in equation number 5. Finally, the confusion matrix of the training for sub-AHE_Dataset based on ResNet-101 shows in figure (6).

$$accuracy\ of\ confusion\ matrix = \frac{\sum true\ positive + \sum true\ negative}{\sum total\ population} \quad \dots \text{equation 5}$$

Figure 4: Confusion matrix of AHE_Dataset small version based on GoogLNet

6
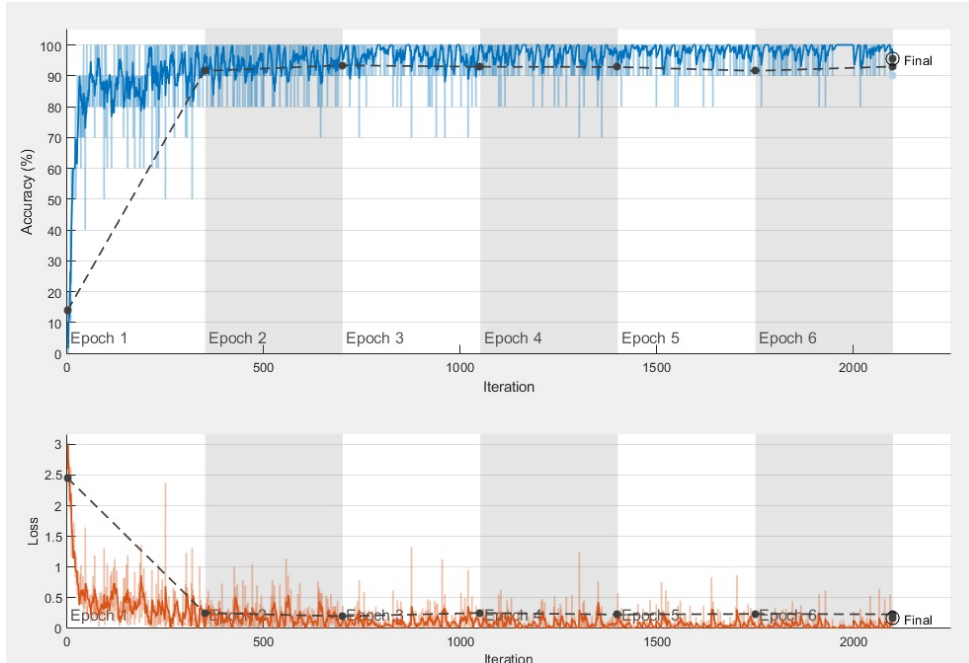
Figure **5** : Accuracy and loss of sub-AHE_Dataset 2nd version based on ResNet 101

Figure 6: Confusion matrix of sub-AHE_Dataset 2nd version based on ResNet 101

The third version of dataset which is the original version of dataset the highest accuracy and loss function was based on ResNet-18. Figure 7 shows the accuracy and loss function, and figure 8 shows the confusion matrix of it.
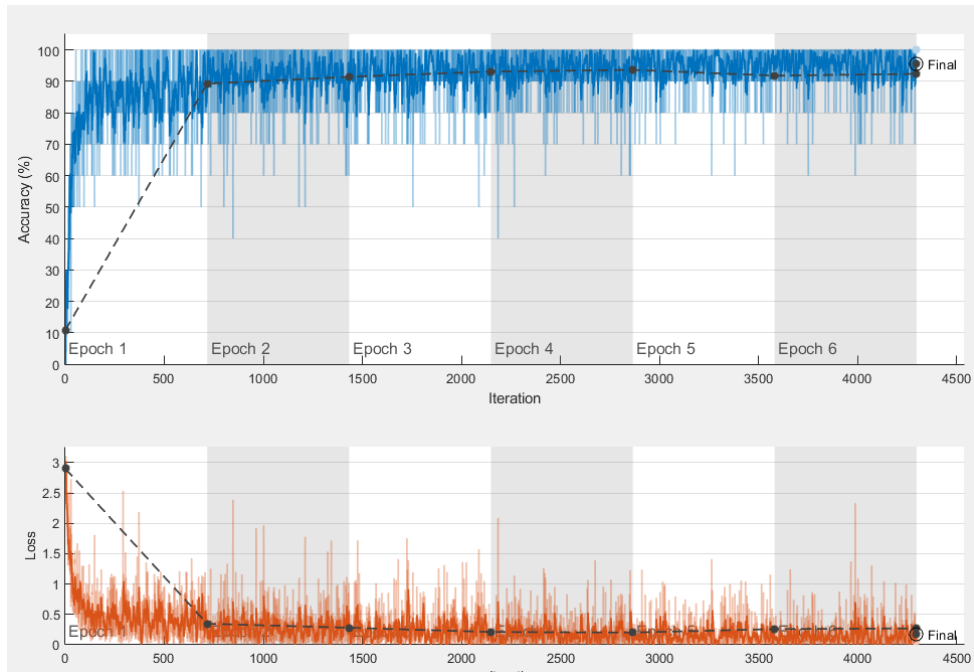
7

Figure 7 : .Accuracy and loss function of the original version of AHE_Dataset based on ResNet-18

Figure 8: Confusion matrix of the original version of AHE_Dataset based on ResNet-18

4.3 Result Analysis

Table 3 shows a summary of the results based on a different test performed. Consider the dataset with 64×64 image size, the GoogLNet achieved the highest accuracy of 87.91 at the same learning rate and iteration in comparison to other techniques. In another case, the highest accuracy achieved is 95.57 based on ResNet-18 under the same conditions for images with a size of 128×128. Finally, the accuracy of sub dataset with 224×224 image size is 95.47 based on ResNet-101.

Table 3: Comparing the accuracy result obtained in different tests performance

Pre-trained CNN algorithm	Image size	Epoch	validation iteration	Training iteration	Learning rate	Accuracy
GoogLNet	64×64	6	98	588	0.0003	**87.91**
ResNet 18	64×64	6	98	588	0.0003	85.07
ResNet 101	64×64	6	98	588	0.0003	87.44
GoogLNet	128× 128	6	716	4296	0.0003	95.47
ResNet 18	128× 128	6	716	4296	0.0003	**95.57**
ResNet 101	128× 128	6	716	4296	0.0003	95.49
GoogLNet	224×224	6	350	2100	0.0003	95.00
ResNet 18	224×224	6	350	2100	0.0003	95.27
ResNet 101	224×224	6	350	2100	0.0003	**95.47**

Comparing the highest result that obtained based on the proposed system with a different algorithm used in [Llamas 17]. Table 4 shows the summary of the result based on a different test performed tested on the same dataset and dataset of architectural style with 25 categories [Xu 14].

Table 4: Summary of test performance based on different algorithms

Pre-trained CNN algorithm	Image size	accuracy
GoogLNet	64×64	87.91
ResNet 18	128× 128	95.57
ResNet 101	224×224	95.47
ResNet (Full Training) [Llamas 17]	32×32	89.6
ResNet (Full Training) [Llamas 17]	64×64	93
Inception-ResNet-v2 (Fine Tuning) [Llamas 17]	128× 128	93.19
MLLR + SP [Xu 14]	Different sizes (typically 800 × 600)	46.21

The best result that achieved based on 64×64 dataset was by ResNet (Full Training) [Llamas 17], and the test based on 128× 128 dataset the highest accuracy achieved was by the proposed model based on ResNet 18 95.57.

Figure 9, Figure 10 and Figure 11 show samples of a testing image with Predicted accuracy for each sample of an original dataset based on GoogLNet, ResNet-18 and ResNet-101 respectively.
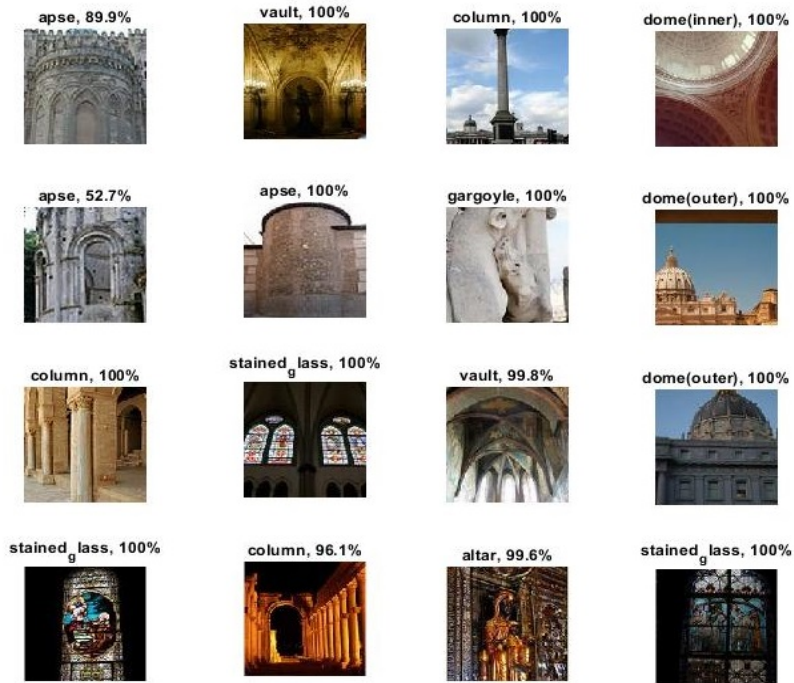
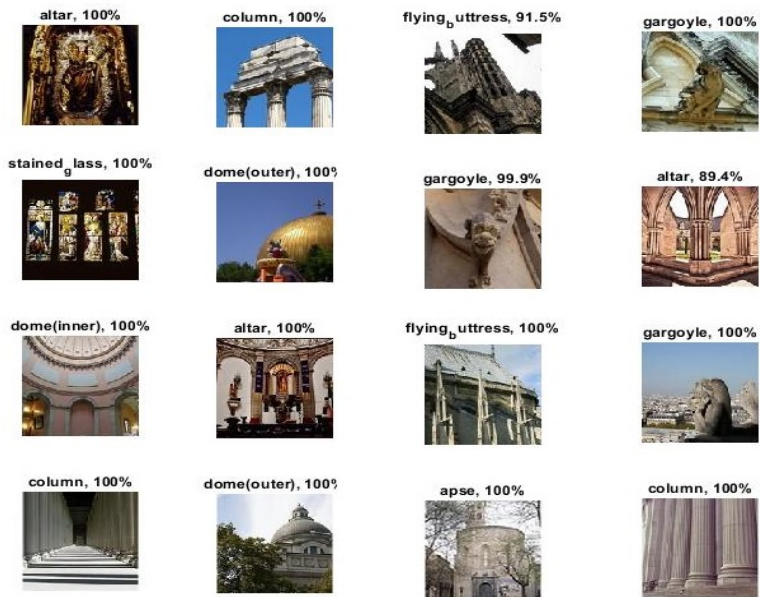Figure 10: Sample of Predicted accuracy of an original dataset based on GoogLNet

Figure 11: Sample of Predicted accuracy of an original dataset based on ResNet-18

Figure 12: Sample of Predicted accuracy of an original dataset based on ResNet-101

5. Conclusion

This work presents a model for architecture heritage image classification using deep learning with a convolutional neural network. In this model, CNN uses a pre-trained structure (GoogLNet , ResNet-18 and ResNet-101). All of them tested on three versions of the dataset. The classification results which are achieved by GoogLNet overperformed the other techniques in comparison with other CNN for the small version of the dataset. However, ResNet-18 produced better classification results compared with other pre-trained CNN. Finally, sub-images of the dataset based on ResNet-101 was the highest accuracy 95.47.

References

[Llamas 17]   Llamas J, M. Lerones P, Medina R, Zalama E, Gómez-García-Bermejo J. Classification of Architectural Heritage Images Using Deep Learning Techniques. *Applied Sciences*. 2017; 7(10):992. doi.org/10.3390/app7100992

[N 16] Tajbakhsh N, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Jianming Liang. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?; IEEE Trans Med Imaging. 2016 May;35(5):1299-1312. doi: 10.1109/TMI.2016.2535302.

[Gao 16] Gao Y, Lee HJ. Local Tiled Deep Networks for Recognition of Vehicle Make and Model. Sensors (Basel). 2016;16(2):226. Published 2016 Feb 11. doi:10.3390/s16020226

[Kadhim 20] Kadhim M.A., Abed M.H. Convolutional Neural Network for Satellite Image Classification. In: Huk M., Maleszka M., Szczerbicki E. (eds) Intelligent Information and Database Systems: Recent Developments. ACIIDS 2019. Studies in Computational Intelligence, vol 830. Springer, Cham. 2020 doi https://doi.org/10.1007/978-3-030-14132-5_13

[Mathias 11] Mathias, M.; Martinovic, A.; Weissenberg, J.; Haegler, S.; Van Gool, L. Automatic Architectural Style Recognition. In Proceedings of the 4th ISPRS International Workshop 3D-ARCH 2011, Trento, Italy, 2–4 March 2011; Volume XXXVIII-5/W16, pp. 171–176.

[Abed 19] Mohammed Hamzah Abed, Atheer Hadi Issa Al-Rammahi and Mustafa Jawad Radif, REAL-TIME COLOR IMAGE CLASSIFICATION BASED ON DEEP LEARNING NETWORK, Journal of Southwest Jiaotong University, vol 54 no 5 .2019. http://www.jsju.org/index.php/journal/article/view/384

[Mustafa 19]   Hafiz Tayyab Mustafa Jie Yang Masoumeh Zareapoor , Multi-scale convolutional neural network for multi-focus image fusion , Image and Vision Computing Volume 85, , Pages 26-35 May 2019.

[Abdelaziz 19]   Abdelhak Belhi,Abdelaziz Bouras, Taha Alfaqheri, Akuha Solomon Aondoakaa and Abdul HamidSadka, Investigating 3D holoscopic visual content upsampling using super-resolution for cultural heritage digitization , Signal Processing: Image Communication Volume 75, July 2019, Pages 188-198 https://doi.org/10.1016/j.image.2019.04.005

[Harangi 18] Balazs Harangi, Agnes Baran and Andras Hajdu, Classification of skin lesions using an ensemble of deep neural networks, 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2018. DOI: 10.1109/EMBC.2018.8512800

[Szegedy 15]   Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich , Going Deeper with Convolutions, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015. DOI: 10.1109/CVPR.2015.7298594

[He 16] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, 2016, pp. 770-778. doi: 10.1109/CVPR.2016.90

[Zhang 19] Zhang, X.; Wang, Y.; Zhang, N.; Xu, D.; Chen, B. Research on Scene Classification Method of High-Resolution Remote Sensing Images Based on RFPNet. *Appl. Sci.* **2019**, *9*, 2028. **https://doi.org/10.3390/app9102028**

[Belhi 18] Belhi, A, Bouras, A & Foufou, S , 'Leveraging known data for missing label prediction in cultural heritage context', *Applied Sciences (Switzerland)*, vol. 8, no. 10, 1768. 2018 https://doi.org/10.3390/app8101768

[Img net] ImageNet. http://www.image-net.org.

[Zhou 16] Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. "Places: An image database for deep scene understanding." *arXiv preprint arXiv:1610.02055* (2016).

[Places] Places. http://places2.csail.mit.edu/

[Xu 14] Xu, Z.; Tao, D.; Zhang, Y.; Wu, J.; Tsoi, A.C. Architectural Style Classification Using Multinomial Latent Logistic Regression. In *Computer Vision—ECCV 2014*; Springer: Cham, Switzerland, 2014; Volume 8689, pp. 600–615.