

Architectural Heritage Classification

Eleonora Ocello

ELEONORAOCELLO@ICLOUD.COM

1. Model Description

The main objective of this study is the application of techniques based on deep learning for the classification of images of architectural heritage, specifically through the use of convolutional neural network. The aim is to verify the real usefulness of these networks that set the current state of art for their application in the classification of images. The deep model consists of, firstly, a 2D Convolution layer which learns features from input images ($3 \times 128 \times 128$). This layer is made of 3 input and 32 output channels, kernel size 5×5 , stride 1×1 and padding 0, followed by a ReLU function useful to introduce the non linearity property. The second layer is composed by another 2D Convolutional layer, with 32 input and 64 output channels, kernel size 3×3 padding 0 and stride 1, followed also by a ReLU function and MaxPool with kernel size 2 and stride 2 in order to reduce the dimensionality and preserve spatial invariance. Generally, CNNs increase channels and reduce image size. At this stage the output features have size 61×61 and before passing to the last layer of the model, the Fully Connected one, we have to flatten them. So, in sequence, FC layer has 238144 input features (the output of CNNs) and 1024 output features, ReLU layer and, at the end, the Classifier layer with 1024 input features and 10 output features.

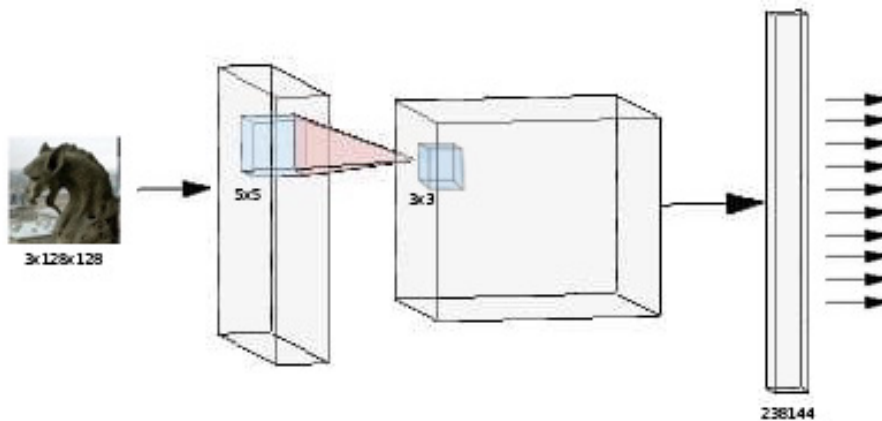
















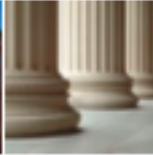










Figure 1: A simple representation of the proposed model.

2. Dataset

The dataset was provided by MDPI Institute in Switzerland. It is a set of more than 10,000 images classified in 10 types of architectural elements of heritage buildings, mostly churches and religious temples. The original dataset contained images of different sizes, respectively test set 64x64 and train, val 128x128, that's why all the images have been resized to 128x128. In total, there are specifically 10,245 images, 80% of the total (8196) were used for the training phase and the remaining 20% (2049) for the validation phase. The validation set has been created by randomly selecting 20% of the images from each class. In addition, 1434 images have been compiled which form an independent set of tests. The figure 2 shows some example images of each of the ten categories considered (in brackets the number of images used in the train and validation of each category). The dataset created has been called *Architectural Heritage Elements Dataset* and has been made publicly available. For the selection of the dataset categories it was consulted the cataloguing of Getting Art & Architecture Thesaurus. The use of this well-known and controlled vocabulary allows consistency in the classification of current and future elements, as well as a more efficient retrieval of information in a standardized way. To increase the amount of relevant images in the dataset so, to reduce the amount of irrelevant features, I flipped the images in the existing dataset horizontally such that they face the other side.

Category	Examples				
Altar (829 images)					
Apse (514 images)					
Bell tower (1059 images)					
Column (1919 images)					
Dome (inner) (616 images)					

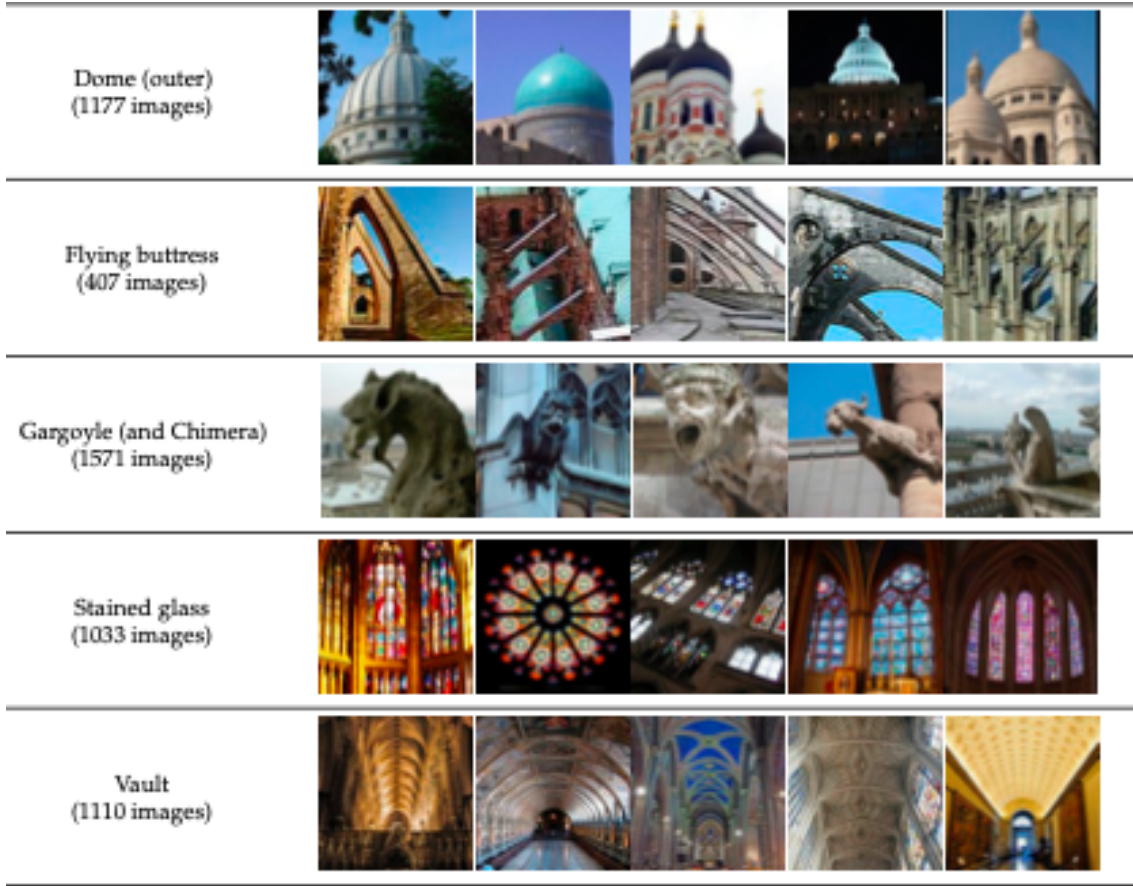


Figure 2: Dataset samples of Cultural Heritage images used.

3. Training procedure

As we know, classification relates to building models that separate images into different classes. The best model, among others (trained), is defined when all the parameters are adjusted. The goal is to increase the validation accuracy and to minimize the global cost function. This function, in this case the CrossEntropy, is an average of the loss functions for each image in the dataset. It is intended to evaluate the error between the output obtained from the network and the desired output. The learning rate is setted to 0.01, it determines the amplitude of the jump to be made by the optimization technique in each iteration. In this case, this technique is the Stochastic Gradient Descent, the gradient calculation requires the errors of the last layer to previous layers to be backpropagated. Since the other possible values trained (64) got worse performance, the batch size is 32. Generically, the batch uses the complete dataset to update the values (this is slow and requires a lot of memory). On GPU CUDA, all the models are fully trained with 20 epochs.

4. Experimental Results

Several experiments were conducted to test the effectiveness of the proposed Architectural Heritage Classification model. Models were trained as described in the previous section. Table 1 shows validation results for the proposed architectures as well as of ablation studies (i.e., different variants of the final architecture when adding or removing layers).

Model	Val. Accuracy	Val. Loss
– + Layer 2	70.3%	11%
– + Layer 3	69.7%	10%
– + Layer 4	68.9%	9.7%
Your final model	70.3%	11%

Table 1: Performance of the models

Some of possible errors could be the presence of other elements in the images and moreover, that the element of the image to be classified is similar to another element. Probably better results could be obtained with fine tuning or off the shelf procedure.