Coursework: Big Data and Machine Learning

Nicolas Bataille
2131029

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Bachelor of Science

Department of Computer Science
Swansea University

December 9, 2021

Contents

1 Introduction

Since the creation of the Internet, the amount of data processed online has been growing exponentially in the past few years, and especially for images. As of 2020, an estimated amount of 210 millions of pictures has been taken on a daily basis on the mobile application Snapchat.[1] Image recognition in the Artificial Intelligence field is a domain in which progress is still made in the recent years, as there is many fields where image recognition is revealed to be much useful, like in facial recognition or intelligent camera software. In this paper, we will answer the two following questions : does hog data provides better results than plain data, and which types of algorithms are better between Machine Learning and Deep Learning algorithms.

For the first problem, we will generate the hog datasets from the already existing training and testing CIFAR-100 datasets provided, which contains a total of 60000 images defined in 32 by 32 pixels with RGB coloring and normalized values. The dataset has 100 classes of images with each 600 images. Those classes are then regrouped in 20 'superclasses', to provide a coarse and a fine labelling of the dataset. The dataset's samples is divided into a training set of 50000 samples and 10000 samples for the testing set. We will use a single Dense Neural Network architecture on hog and plain datasets, with a model for both fine and coarse data for each types.

As for the second problem, we will use the Logistic Regression Algorithm on the hog dataset and a Convolutionnal Neural Network on the plain dataset, also with a model for both fine and coarse data types. An evaluation followed by predictions on the Testing dataset will be done for each model, with a confusion matrix used to present the results on the prediction as well as a percentage of global accuracy. We will in further details that even though the hog data allows to have a slightly better accuracy than with the plain dataset, we can affirm that a Machine Learning algorithm like the Logistic Regression is far less accurate than a Deep Learning Model like a Convolutionnal Neural Network which is very efficient on image processing.

[1] https://seedscientific.com/how-much-data-is-created-every-day/

2 Method

2.1 Comparing the results on the Hog dataset to the plain dataset with a Dense Neural Network

For this part of the coursework, we first had to perform the HOG (Histogram Of Gradient) on the dataset to extract an array of features from the plain dataset. In this coursework, for the sake of having a reasonnable computing time, we generate a feature vector of 324 features for the training and the testing dataset. Then, we perform a dimensionnality reduction on these feature vectors with a PCA object instanciated to keep every features as principle components, in order to have the best accuracy possible while reducing the computing time. We have performed some tests to compare predictions made with data projected by a PCA object compared to data projected with a LDA object, and those that used the PCA objects gave slightly better results. Then, we instanciate a Dense Neural Network, consisting of the following architecture [1] :

Then, we proceed to train a model for each types of

layer type	no of neurons	parameter
Input Layer	Input Size	none
Dense	128	ReLu
Dropout	none	0.5 (dropout rate)
Dense	Output size	softmax

Dense Neural Network Architecture

labels, on both hog and plain datasets, though the plain dataset needs to be flattened in order to be processed. The optimizer used for this model is the Adam optimizer, considered to perform well, with a learning rate of 0.0001. The modification of the learning rate as well as the dropout layer are much needed to reduce overfitting as much as possible [2]. The models are trained with 100 epochs and a validation rate of 0.2. Finally, we make predictions on the testing datasets with both types of labels for both types of data, and evaluate the average accuracy of each model.

2.2 Comparing a Machine Learning algorithm to a Deep Learning model

As for comparing a Machine Learning algorithm to a Deep Learning Model, we had to use the best ones for images processing. Support Vector Machine is considered as one of the best performing Machine Learning Algorithm for image processing, but it is also very time and power consuming, so we opted for the Logistic Regression algorithm instead. We kepts the default parameters from the sklearn library, except for the maximum number of iteration that we doubled in order to process the whole dataset. The LR (Logistic Regression) algorithm also used the hog and dimensionnaly reduced datasets, for faster computation results. The instanciated LR objects were fitted to their respective types of labels and dataset, and then evaluated with an average accuracy percentage and a confusion matrix of their predictions. For the CNN (Convolutionnal Neural Network), we made an architecture that is commonly used in image recognition [3] :

layer type	size	activation/parameter
Conv2D	32, kernel size of 3	activation : ReLu
MaxPool2D	pool size : (2,2)	none
Conv2D	96, kernel size of 3	activation : ReLu
MaxPool2D	pool size : (2,2)	none
Dropout	none	0.5 (dropout rate)
Flatten	none	none
Dense	256	ReLu
Dropout	none	0.5 (dropout rate)
Dense	128	ReLu
Dense	number of classes	Softmax

Convolutionnal Neural Network Architecture

This model is instantiated for each type of labels, and also uses the Adam optimizer with a learning rate of 0.0001 and dropout layers [2] to prevent overfitting as much as possible. The models are trained on the plain training set with 100 epochs, a validation split of 0.2 and a default batch size of 32. After fitting, the models are evaluated on the testing datasets to give an average accuracy, as well as a confusion matrix.

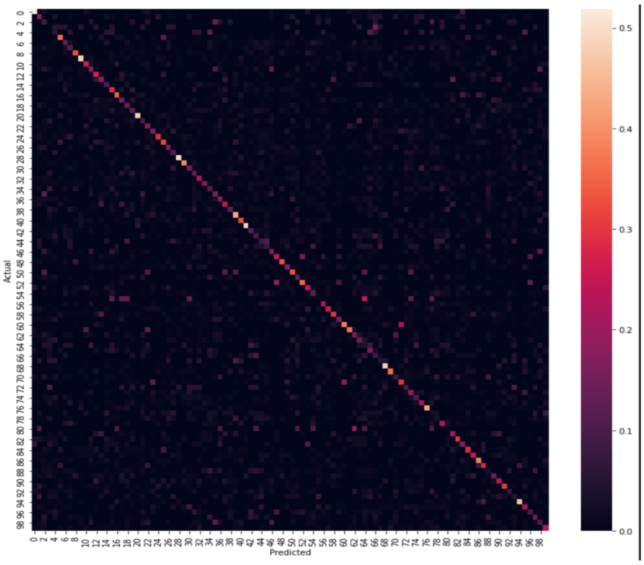
3 Results

3.1 Hog features versus Plain features

After performing the predictions on every classifiers, we can finally answer the problems stated at the beginning of this document. The accuracy provided in the following table is calculated as

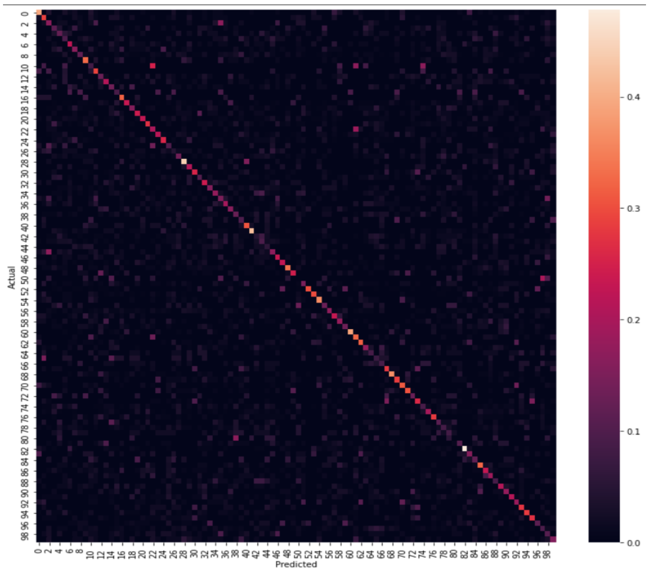
$$Accuracy = \frac{Number of correct predictions}{Total number of predictions}$$

As we can see on the table regrouping the accuracy results for every classifier, the DNN(Dense Neural Network) is slightly more precise on both coarse and fine predictions with the Hog and projected data than the plain data. This can also be seen on the confusion matrices.



Confusion matrix of the DNN prediction on the Hog Fine dataset

When we compare this matrix with the one using the plain dataset, we can notice that a few more classes are mislabeled with the plain data. The Coarse Confusion Matrices are in appendices in order to have a more readable document. Those results proves that the Hog feature vectors allows to have fewer feature to process while get-



Confusion matrix of the DNN prediction on the Plain Fine dataset

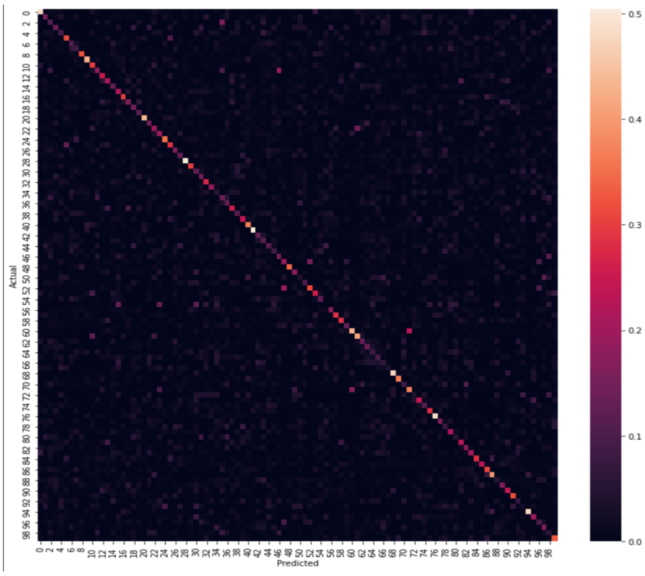
ting better accuracy overall, though the difference is very slight, but can be good to consider when we would try to compute images.

3.2 Machine Learning versus Deep Learning

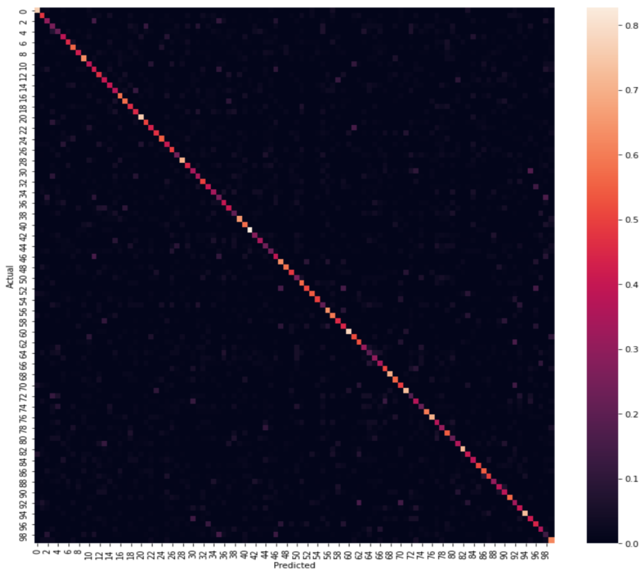
When we compare the LR classifier to a CNN model, we have a clear distinction of performance between the two classifiers. We almost attain with the CNN an accuracy the double of the LR's one. The disparity is even more evident on the confusion matrices for the Fine labels :

Classifier	Coarse Accuracy	Fine Accuracy	Data
Logistic Regression	30.17%	22.96%	Hog + PCA Projection
Dense Neural Network	33.67%	23.15%	Hog + PCA Projection
Dense Neural Network	31.79%	21.22%	Plain data (flattened)
Convolutionnal Neural Network	56.68%	44.18%	Plain data
Referencial accuracy	39.43%	24.49%	unknown

Accuracy for each Classifier on both labels types



Confusion matrix of the LR prediction on the Hog Fine dataset



Confusion matrix of the CNN prediction on the Plain Fine dataset

The fine confusion matrix for the CNN has a very few misslabeling compared to the LR's matrix.

4 Conclusion

After comparing the accuracy of a same DNN model on Hog data and Plain data, and opposing the LR classifier to a CNN model, the obtained results yields a solid evidence that using DL(Deep Learning) models and more precisely CNNs is the wisest choice to achieve image processing with accurate results. Although there is much more efficient models existing [2], a simple CNN model can easily gives very good accuracy scores. Some classes in the Fine labels are shown to be less predicted than other, taking for example the class 73, "sharks", that is poorly recognized among every classifier used in this work, compared to the class 82, "sunflower", that belongs to the best fine classes results. A good reason behind that would be that the images of the "shark" class are much more different from each other than the sunflowers'.

Globally, every classifiers used here could have produced better results with much more tweaking of their hyperparameters, like the batch size for the DL models for example that if augmented, can increase the computing time but also the accuracy and the chance of overfitting for the models. This can be explained by the fact that a larger batch size tends to lead to a poor generalization and thus a poor global accuracy[3]. Data augmentation is also a good way to get more training material for the models, making them more accurate with less overfitting, by creating slightly different images from the already existing one, affecting the RGB values by a small percentage for instance.

[2]https://paperswithcode.com/sota/image-classification-on-cifar-100

[3]https://medium.com/mini-distill/effect-of-batch-size-on-training-dynamics-21c14f7a716e

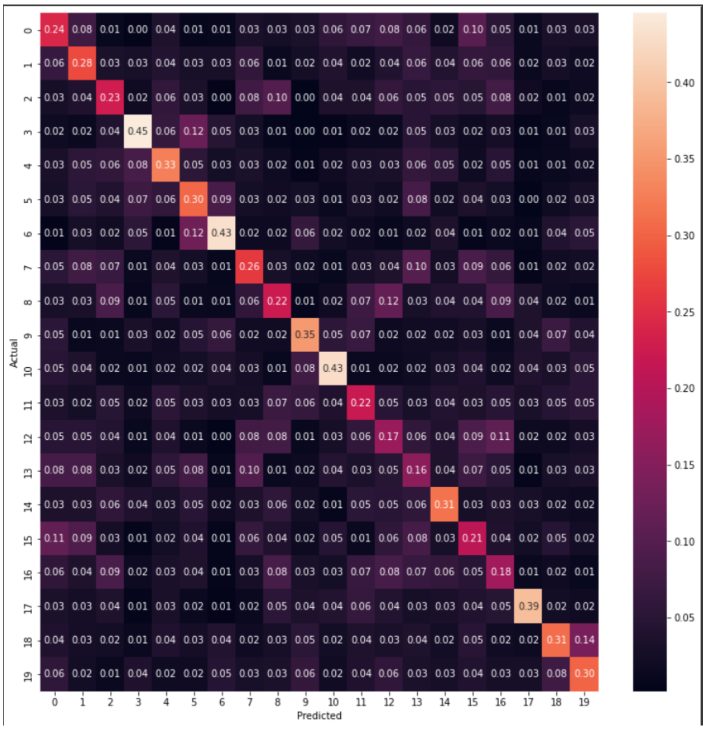
Bibliography

[1] K. Gnana Sheela, S. N. Deepa, "Review on Methods to Fix Number of Hidden Neurons in Neural Networks", Mathematical Problems in Engineering, vol. 2013, Article ID 425740, 11 pages, 2013. https://doi.org/10.1155/2013/425740

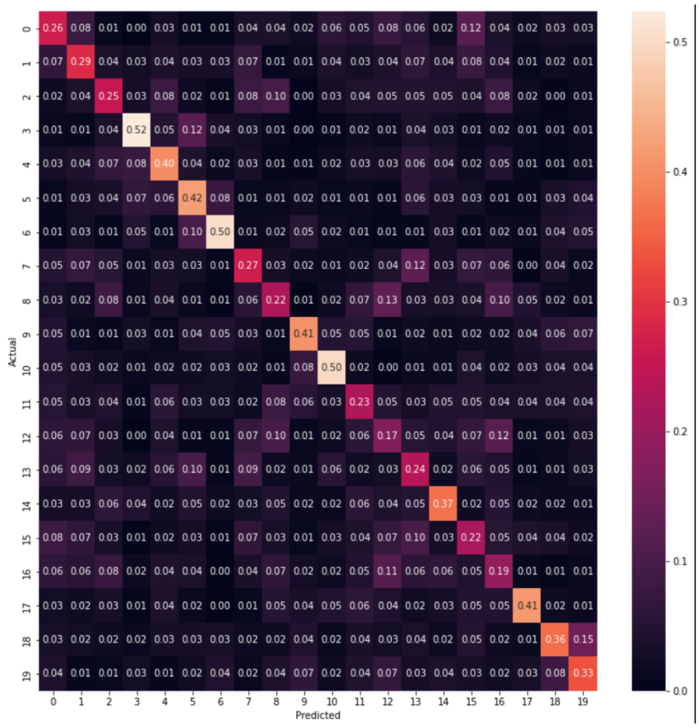
[2] Park S., Kwak N. (2017) Analysis on the Dropout Effect in Convolutional Neural Networks. In: Lai SH., Lepetit V., Nishino K., Sato Y. (eds) Computer Vision – ACCV 2016. ACCV 2016. Lecture Notes in Computer Science, vol 10112. Springer, Cham. https://doi.org/10.1007/978-3-319-54184-6_12

[3] Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9

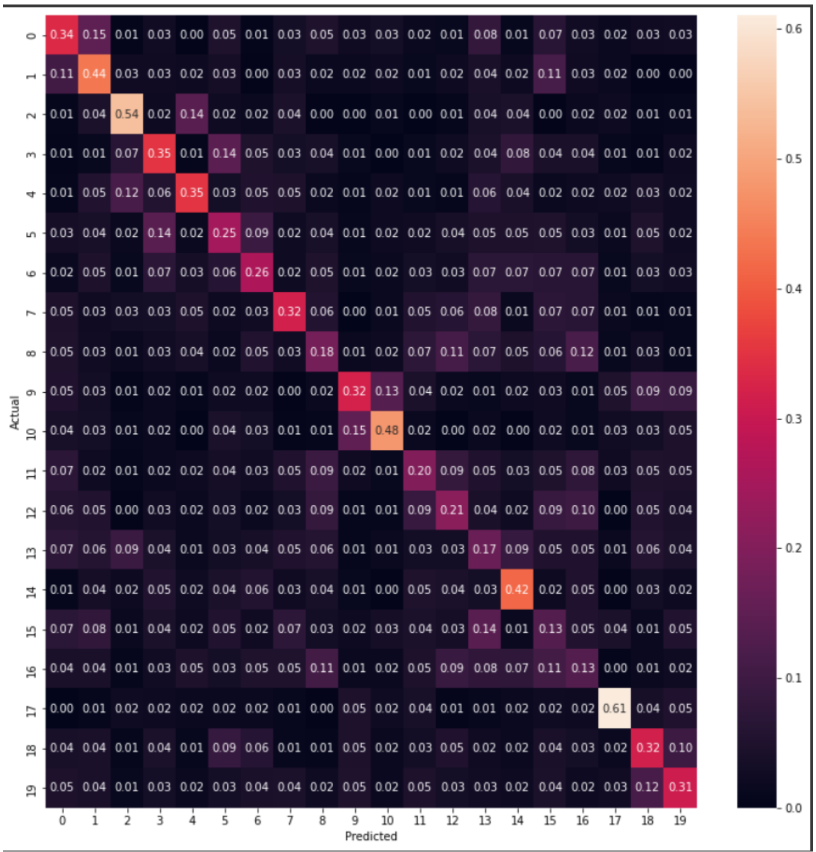
Appendices



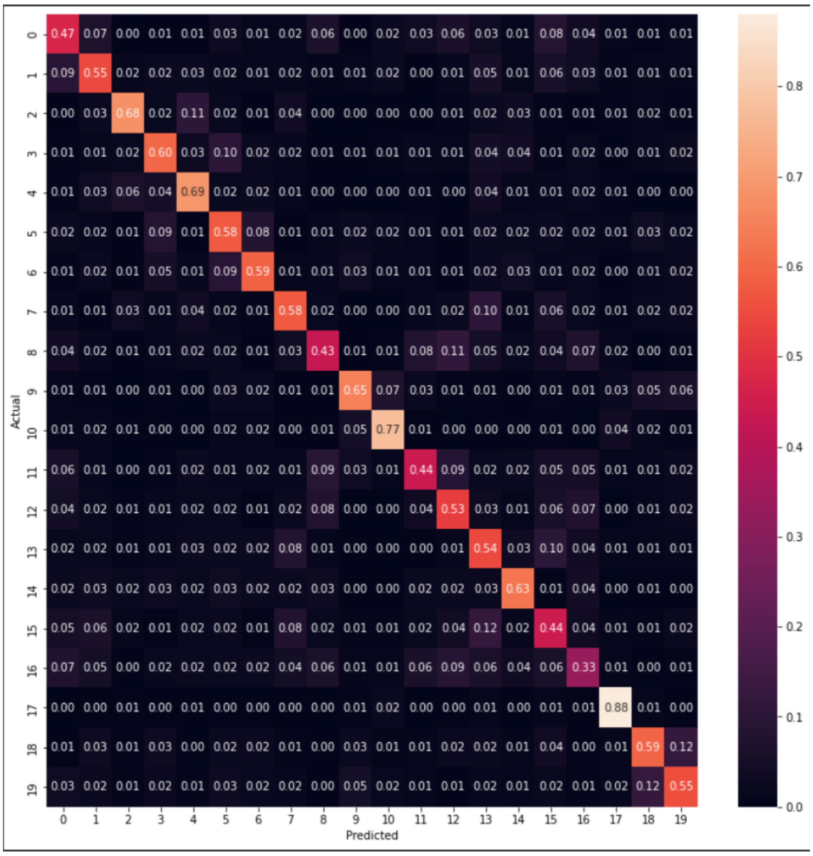
Confusion matrix of the LR prediction on the Hog Coarse dataset



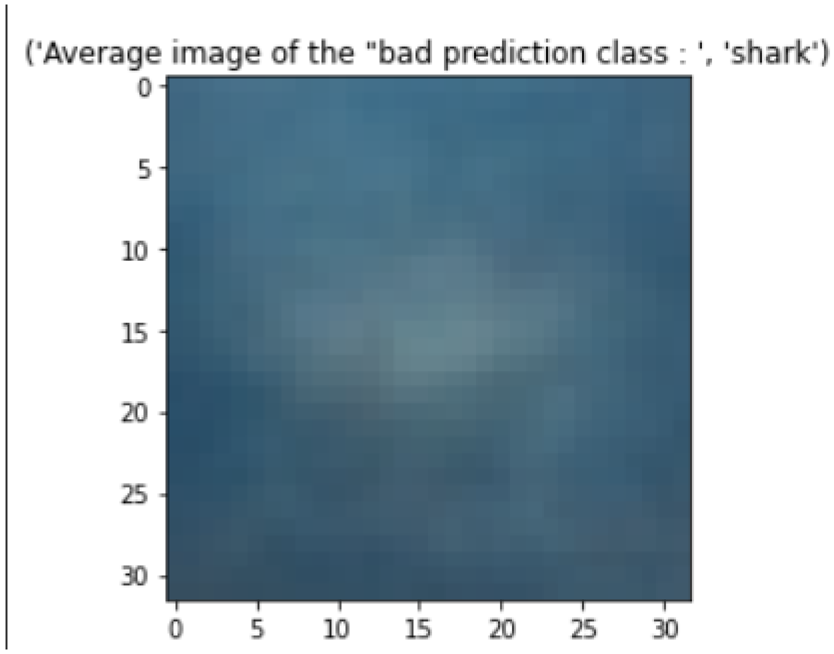
Confusion matrix of the DNN prediction on the Hog Coarse dataset



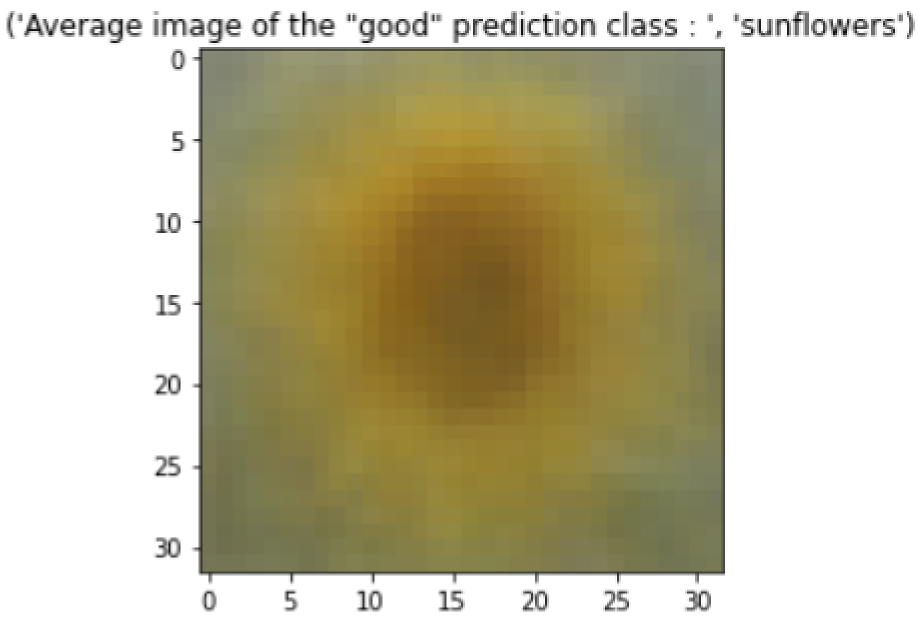
Confusion matrix of the DNN prediction on the Plain Coarse dataset



Confusion matrix of the CNN prediction on the Plain Coarse dataset



Average Image of the "shark" class, that has bad accuracy



Average Image of the "sunflower" class, that has good accuracy