Data Challenge - Kernel methods

`Team TPT - MVA/MASH 2021-2022`

Florian Lamalle, Rafaël Brutti
{florian.lamalle,rafael.brutti}@telecom-paris.fr

April 9, 2022

Abstract

The goal of the data challenge is to learn how to implement machine learning algorithms, gain understanding about them and adapt them to structural data. The assignment is to create the best model with Kernel Methods to classify images from 10 different classes. The Kernel Methods had to be implemented from scratch. We tried Kernel PCA for dimensionality reduction, Kernel SVC Multiclass "One versus One" and "One versus All" with and without feature descriptors.

Link : Github with the code

1 Introduction

1.1 Presentation of the challenge

The data are from the CIFAR-10 dataset. The training images are stored in Xtr.csv as a matrix of size 5000 x 3072, each row corresponding to a 32 x 32 pixel color image. The labels of the training images are stored in Ytr.csv in the same format of a Kaggle submission file. Similarly, the test set is stored in Xte.csv containing 200 test images without their labels. All images have been pre-processed with several unknown noise methods.

1.2 Data Visualisation

Each row of size 3072 contains 1024 values representing the pixel intensities on the red channel, then the next 1024 represent the green channel, and the last 1024 entries represent the blue channel. We found a way to visualize the images by creating a new dataset of size n_sample $\times$ 1024 $\times$ 3. Moreover, we also checked if the data were well balanced, which was the case.

1.3 Data Split

We split the training images (Xtr.csv) into a training set and a validation set with a ratio 85% / 15% to test our models with an accuracy score before submitting our predictions on Kaggle.

2 Kernels Used

The Kernels used are the Gaussian (RBF), the Linear and the Polynomial Kernel. Most of the time, the Gaussian kernel outperforms the others, so we decided to consider only it in the end.

3 Binary Classification and Multi-class Generalisation

The C-Support Vector Classification SVC algorithm is implemented from scratch with a 'SLSQP' optimizer. Consider a dataset of N pairs (x_i, y_i) where each x_i is a vector of dimension d and y_i is a binary class, i.e. $y_i \in \{-1, 1\}$. This algorithm separates two classes of samples with a separating hyper-surface of equation $f(x_i) + b = 0$ such that $f(x_i) + b < 0$ if x_i belongs to the class $y_i = -1$ and $f(x_i) + b \geq 0$ if $y_i = 1$. Then, one solution to generalise this algorithm to multi-classification with 10 labels is to use two approaches "One vs. One" or "One vs. All" SVC. The main difference between these two approaches is the number of classifiers considered. With $N = 10$ labels, there are $\frac{N(N-1)}{2} = 45$ classifiers with the OvO approach and only $N = 10$ classifiers with the OvA approach. The accuracy of the OvA SVC is slightly better than the OvO approach, and even though there are 4.5 times fewer classifiers, OvA takes significantly more computational time. This is due to the fact that for each classifier the highlighted label (associated with $y = 1$ in the binary classification) is set against the other 9 labels considered as a single class ($y = -1$). Each OvA classifier considers more data than one OvO classifier. We decided to keep the OvO to save computational time because the algorithms are very time consuming.

4 Dimensionality Reduction

PCA is a feature extraction technique widely used for de-noising and classification. Here, its interest is to reduce the dimensionality of the 3072 features into a smaller set of features to speed up the model fitting and the prediction calculation. Unfortunately, it performed poorly on this multi-classification challenge. Non-linearity degradation could be the issue. Therefore, we chose to work with Kernel PCA, a nonlinear variant of PCA that extracts nonlinear features and which is said to outperforms linear PCA in de-noising and image reconstruction. With the PCA Kernel (100 features kept) and a fine-tuning of the Gaussian Kernel parameters, we achieved an accuracy of 24% instead of 15% without it.

5 Feature Descriptors

Histogram of Oriented Gradients (HoG) method is used as a feature descriptors (cf. skimage). The image is divided into 8×8 cells on which the HoG is computed by using the gradients, magnitude and direction are determined for each cell. At the end, each image is represented by 144 features and this enabled us to reach more than 50% accuracy in a faster way.

6 Results

With a grid search, the best model found is the Gaussian kernel with $C = 5$ small enough to limit the over-fitting and $\sigma = 0.029$. Using also HoG features, the validation accuracy is 55.6%, which corresponds to a public score of 55,4% and a private score of 53,9% on Kaggle. Finally, we ranked 27th in the competition.

7 Improvements

Possible improvements include the use of de-noising techniques, the use of other feature descriptors (e.g. SIFT) or the implementation of Kernel K-Means. We tried to implement the CCA kernel without success.

A Illustrations

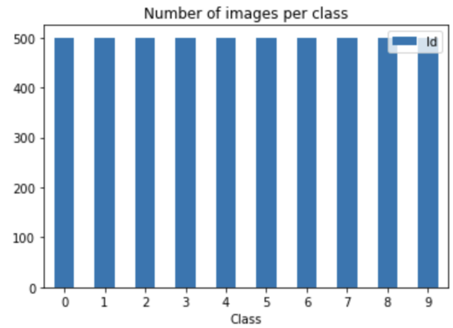

Figure 1: Label distribution in the training images

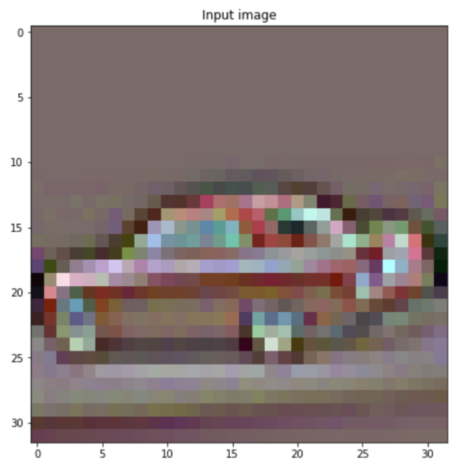

Figure 2: Example of a noisy image from the test set [class = car]

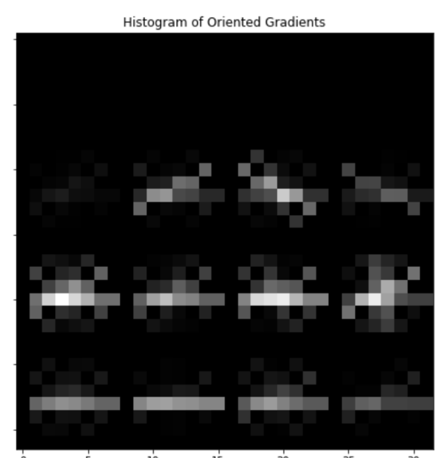

Figure 3: Corresponding HoG features

References

[AKN+18] Taha J. Alhindi, Shivam Kalra, Ka Hin Ng, Anika Afrin, and Hamid R. Tizhoosh. Comparing lbp, hog and deep features for classification of histopathology images. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018.

[FM04] G.M. Foody and A. Mathur. A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing*, 42(6):1335–1343, 2004.

[XX04] Pengcheng Xi and Tao Xu. De-noising and recovering images based on kernel pca theory. pages 197–200, 01 2004.