Assignment 2: Image classification

Victor Busa

victor.busa@ens-paris-saclay.fr

I Training and testing an Image Classifier

I.A Data preparation and feature extraction

QA1: Why is the spatial tiling used in the histogram image representation?

spatial tiling is used in the histogram image representation in order to add spatial information to the bag of visual words. It provides information about the spatial position of features (sky in upper area of an image, car tyres under the car,...) . Moreover, spatial tiling can be used to do spatial pyramid matching that has proven to be very effective.

I.B Train a classifier for images containing aeroplanes

QB1: Show the ranked images in your report.

See Figure 1

QB2: In your report, show relevant patches for the three most relevant visual words (in three separate figures) for the top ranked image. Are the most relevant visual words on the airplane or also appear on background?

The most relevant visual words appear both on the airplane and on the background, because airplane photos often show the airplane in its environment (on a airstrip or in the sky). See Figure 2

I.C Classify the test images and assess the performance

QC1: Why is the bias term not needed for the image ranking?

The bias term is not needed for the image ranking because it corresponds to a constant scalar (here 1.73) added to all the images. Hence, if we get rid of the bias term, the ranking will not change but the score of each image will be reduce by a constant.

I.D Learn a classifier for the other classes and assess its performance

QD1: In your report, show the top ranked images, precision-recall curves and APs for the test data of all the three classes (aeroplanes, motorbikes, and persons). Does the AP performance for the different classes match your expectations based on the variation of the class images?

I would have expected that the AP performance for the person class would be higher than for the other classes because the images for the person classes varies a lot and hence the features descriptors won't focus

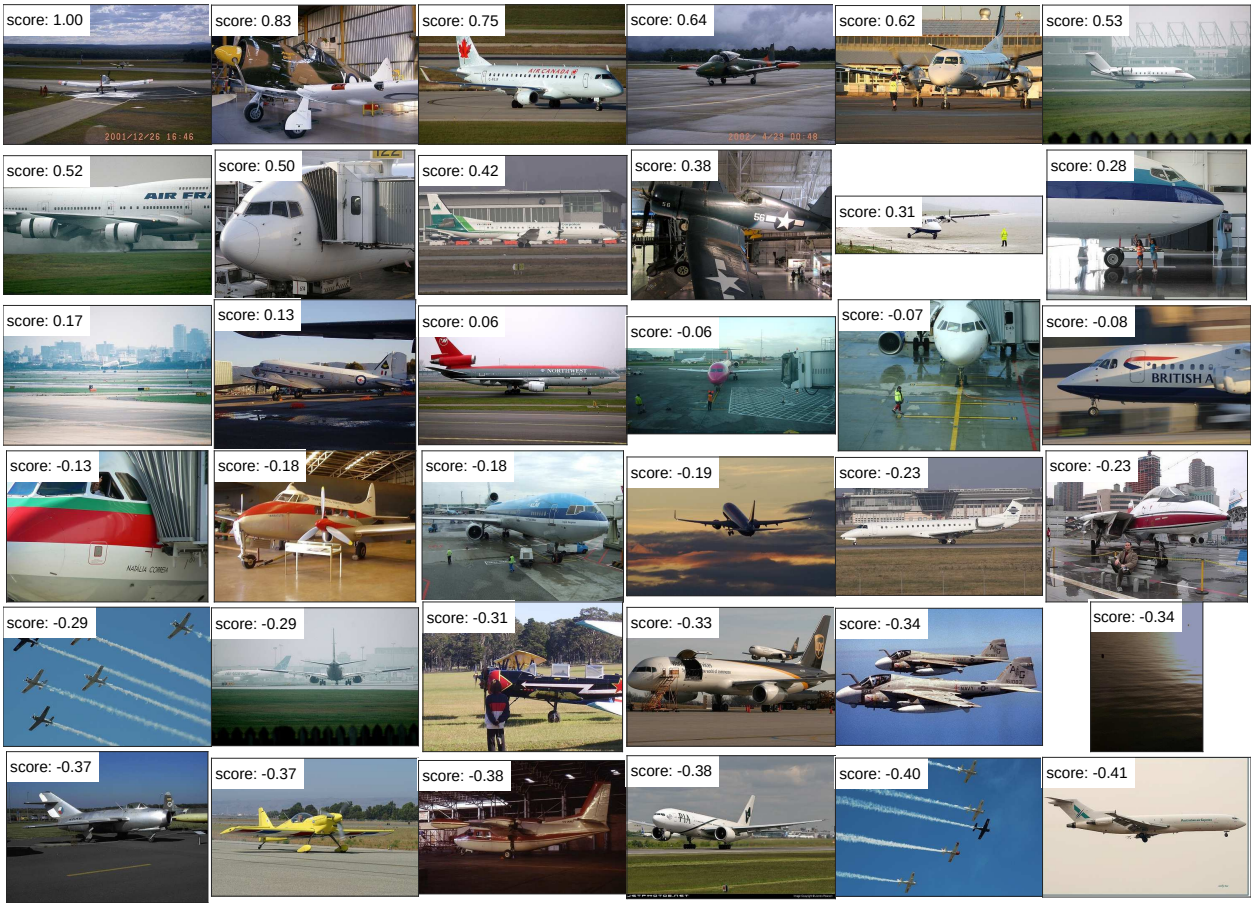
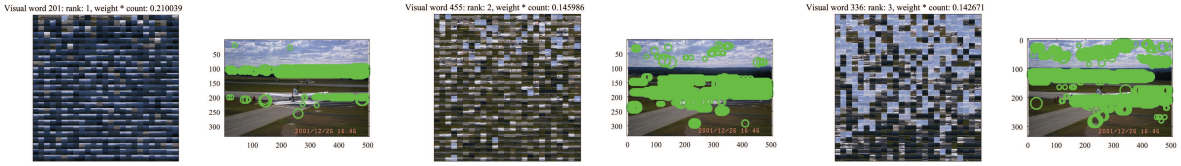


Figure 1: Ranked images for the training dataset



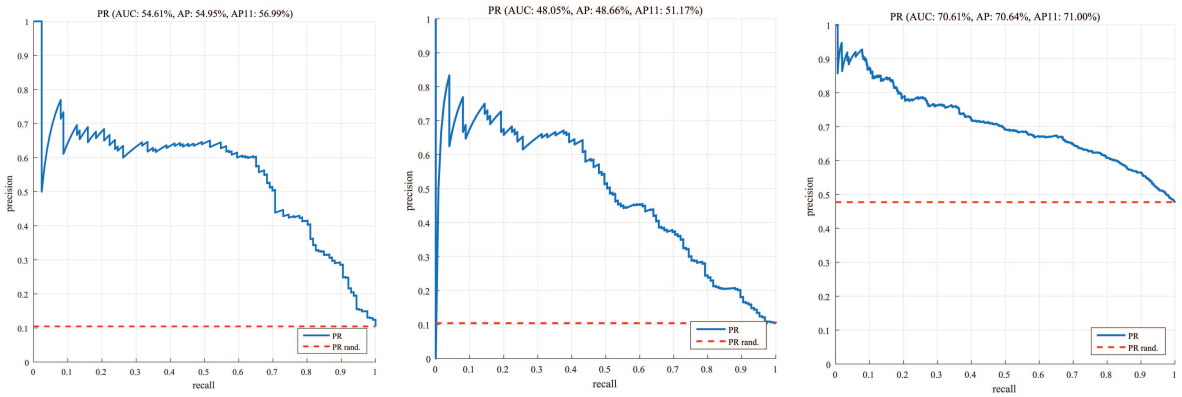
(a) Best visual word (b) Second best visual word (c) Third best visual word

Figure 2: Three most relevant visual words for the top ranked image

on the background of the images belonging to the person class and hence the performance will be higher. On the other hand, both aeroplane and motorbike classes have little variations and hence the performance of the classifier will be reduce because the features descriptors will take into account the background (often airstrips for planes and street or grass for motorbikes) as a characteristic of the class. Moreover the training dataset contains 10 times more data for the person class than the motorbike or the aeroplane classes. More detailed pictures are provided in Annexe 9

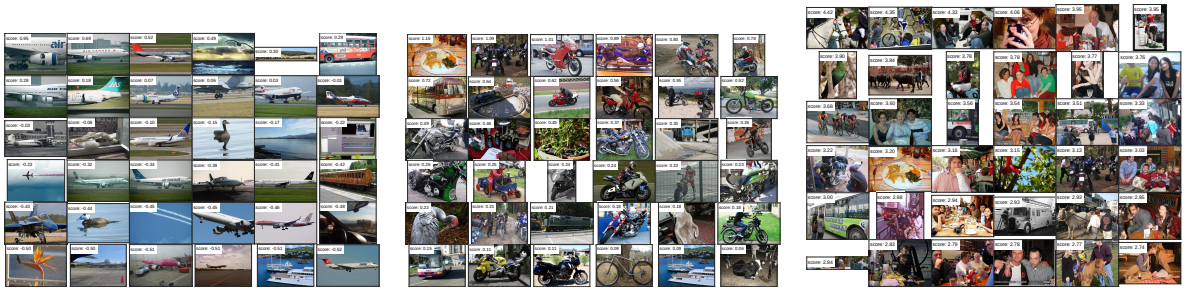
QD2: For the motorbike class, give the rank of the first false positive image. What point on the precision-recall curve corresponds to this first false positive image? Give in your report the value of precision and recall for that point on the precision-recall curve.

The first false positive image for the motorbike class is the first ranked image. This first false positive image corresponds to the point that has: **recall=0** and **precision=0**.



(a) Aeroplane AP = 54.95% (b) Motorbike AP = 48.66% (c) Person AP = 70.64%

Figure 3: Precision-recall curves and APs for the test data



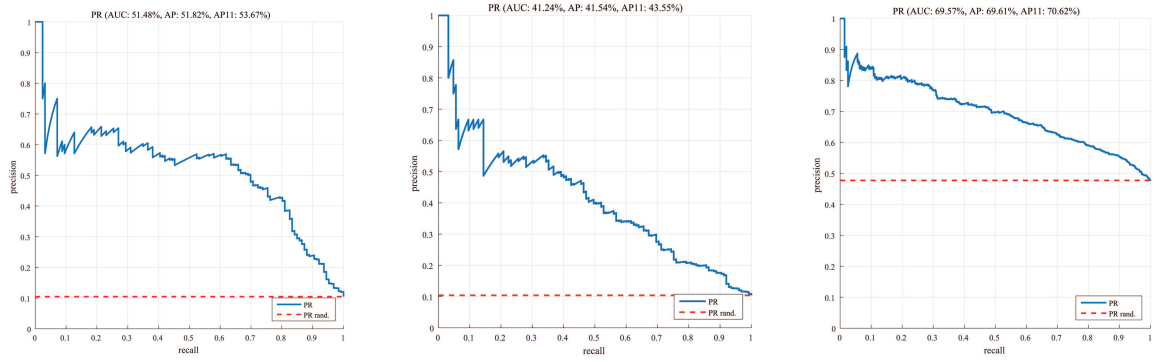
(a) Aeroplane (b) Motorbike (c) Person

Figure 4: Top ranked images for the test data on all the three classes (aeroplanes, motorbikes, persons)

I.E Vary the image representation

QE1: Include in your report precision recall-curves and APs, and compare the test performance to the spatially tiled representation in stage D. How is the performance changing? Why?

When we don't use histograms with spatial tiling, the overall performance decrease. This is due to the fact that spatial tiling provide another piece of information about the position of the feature descriptors.



(a) Aeroplane AP = 51.82% (b) Aeroplane AP = 41.54% (c) Aeroplane AP = 69.61%

Figure 5: Precision-recall curves and APs for the test data with spatial tiling off

QE2: Modify exercise1.m to use L1 normalization and no normalization and measure the performance change.

	Aeroplane	Motorbike	Person
No Normalization	**62.45**	**48.73**	67.20
L1-Normalization	51.68	26.00	56.54
L2-Normalization	54.95	48.66	**70.64**

Table 1: Benefits of descriptor normalization (in % of Average Precision)

QE3: What can you say about the self-similarity, K(h,h), of a BoVW histogram h that is L2 normalized? Hint: Compare K(h,h) to the similarity, K(h,h́), of two different L2 normalized BoVW histograms h and h́. Can you say the same for unnormalized or L1 normalized histograms?

for an histogram h $L2$ normalized we have:

$$k(h,h) = \sum_{i=1}^{d} h_i^2 = 1$$

Moreover, as h can be regarded as a discrete probability distribution, we have, $h_i \geq 0$, $\forall i \in [1,d]$, and by Cauchy-Schwartz:

$$\begin{aligned} k(h,h') = \sum_{i=1}^{d} h_i h'_i &\leq |\sum_{i=1}^{d} h_i h'_i| \\ &\leq \left(\sum_{i=1}^{d} h_i^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^{d} {h'_i}^2\right)^{\frac{1}{2}} = \sqrt{k(h,h)k(h',h')} = 1 \end{aligned}$$

For L1 normalized histograms we have:

$$\begin{aligned} k(h,h') = \sum_{i=1}^{d} h_i h'_i &\leq |\sum_{i=1}^{d} h_i h'_i| \\ &\leq |\sum_{i=1}^{d} h_i| \max_i |h'_i| \\ &\leq \sum_{i=1}^{d} |h_i| \sum_{i=1}^{d} |h'_i| \\ &\leq 1 \times 1 = 1 \end{aligned}$$

Yet, The maximum is not attained for the self-similarity for L1 normalization and No normalization.

QE4: Do you see a relation between the classification performance and L2 normalization?

According to QE3. L2 normalization is better because it respects the fact that the maximum should be attained when $h = h'$ (self-similarity) while if we are using L1 normalization the maximum of $K(h,h')$ is not necessarily attained when $h = h'$.

I.F Vary the classifier

QF1: Based on the rule of thumb introduced above, how should the BoVW histograms h and h' be normalized? Should you apply this normalization before or after taking the square root?

A good rules of thumb is to L_2 normalize any vectors after any processing. As h can be seen as a discrete probability distribution, we have, $h_i \geq 0$ and so:

$$h_i \xrightarrow{\sqrt{\cdot}} \sqrt{h_i} \xrightarrow{L_2} \frac{\sqrt{h_i}}{\sqrt{\sum_{i=1}^{d} \sqrt{h_i}^2}} = \frac{\sqrt{h_i}}{\sqrt{\sum_{i=1}^{d} |h_i|}} \tag{1}$$

$$h_i \xrightarrow{L_1} \frac{h_i}{\sum_{i=1}^{d} |h_i|} \xrightarrow{\sqrt{\cdot}} \frac{\sqrt{h_i}}{\sqrt{\sum_{i=1}^{d} |h_i|}} \tag{2}$$

Hence, doing (1) operations is equivalent to doing (2) operations. So we can apply L_1 normalization **before** taking the square root.

QF2: Why is this procedure equivalent to using the Hellinger kernel in the SVM classifier?

Let K_H denote the Hellinger kernel and K_L be the linear kernel. If we use the square root of the histograms and the L_1-normalization, then it is equivalent to using the Hellinger kernel. Indeed:

$$\bar{h} \leftarrow \sqrt{h}$$

$$K_L(\bar{h}, \bar{h'}) = \sum_{i=1}^{d} \bar{h_i} \bar{h_i}' = \sum_{i=1}^{d} \sqrt{h_i h_i'} = K_H(h, h')$$

$$K_H(h, h) = \sum_{i=1}^{d} |h_i| = 1$$

$$K_L(\bar{h}, \bar{h}) = \sum_{i=^1}^{d} \bar{h_i}^2 = \sum_{i=^1}^{d} |h_i| = 1$$

QF3: Why is it an advantage to keep the classifier linear, rather than using a non-linear kernel?

It is advantageous to use a linear kernel because it is much faster to compute and tractable.

QF4: Try the other histogram normalization options and check that your choice yields optimal performance. Summarize your finding in the report (include only mAP results, no need to include the full precision-recall curves).

Here, all normalizations are taken **before** using the Hellinger kernel.

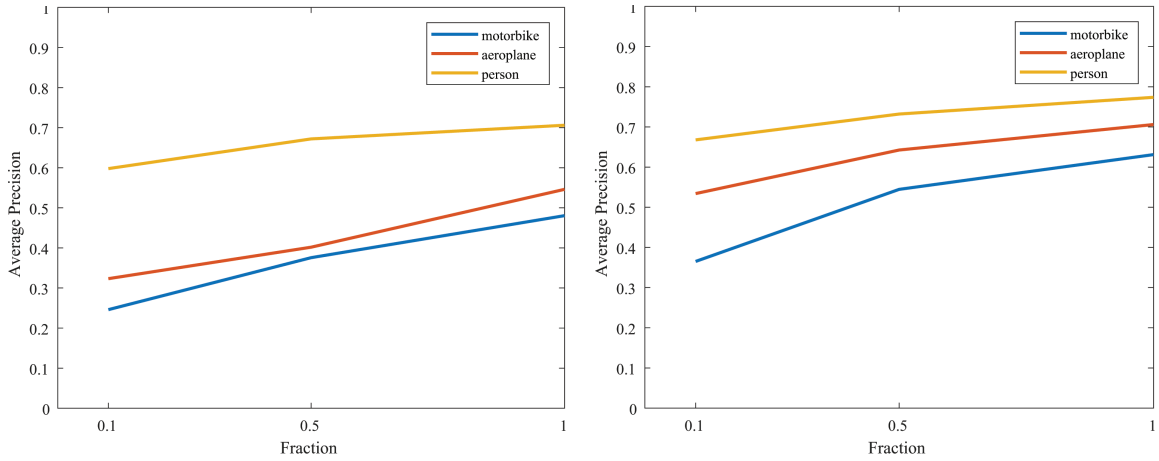
	Aeroplane	**Motorbike**	**Person**
No Normalization	65.87	55.64	69.71
L1-Normalization	**70.72**	**63.25**	**77.39**
L2-Normalization	63.94	52.81	71.45

Table 2: Performance of normalization for the Hellinger Kernel (in % of Average Precision)

Note: Applying L_2 Normalization after the Hellinger Kernel is equivalent to applying L_1 normalization before the Hellinger kernel.

I.G Vary the number of training images

QG1: Report and compare performance you get with the linear kernel and with the Hellinger kernel for the different classes and proportions of training images (10%, 50% and 100%). You don't have to report the precision-recall curves, just APs are sufficient. Plot the APs for one class into a graph, with AP on the y-axis and the proportion of training images on the x-axis. You can use the matlab function plot Plot three curves (one curve for each class) into one figure. Produce two figures, one for the linear kernel and one for the Hellinger kernel. Make sure to properly label axis (use functions xlabel and ylabel), show each curve in a different color, and have a legend (function legend) in each figure. Show the two figures in your report.



(a) linear kernel

(b) Hellinger kernel

Figure 6: Performance measure for the three classes and different proportion of the training images (10%, 50%, 100%)

QG2: By analyzing the two figures, do you think the performance has 'saturated' if all the training images are used, or would adding more training images give an improvement?

The performance has not 'saturated'. However we would likely need a lot more training images in order to increase the performance significantly as the slop is slightly positive and will likely decrease towards 0 (saturation).

II Training and testing an Image Classifier

QP2.1: For the horse class, report the precision at rank-36 for 5 and 10 training images. Show the training images you used. Did the performance of the classifier improve when 10 images were used?

See Table 3 for the precision at rank-36 for 5 and 10 training images. The images used for the horse class are shown in Figure 7

Figure 7: Training images for horse class

QP2.2: What is the best performance (measured by precision at rank-36) you were able to achieve for the horse and the car class? How many training images did you use? For each of the two classes, show examples of your training images, show the top ranked 36 images, and report the precision at rank-36. Compare the difficulty of retrieving horses and cars.

The top ranked 36 images for both horses and cars are shown in Annexe 12

	Number of training images	**5**	**10**
Horses	Precision at rank 36 (AP %)	5/36 (10.85)	17/36 (27.13)
Cars	Precision at rank 36 (AP %)	27/36 (55.55)	34/36 (65.59)

Table 3: Performance for horse and car classes

The images used for the car class are shown in Figure 8

Figure 8: Training images for car class

It is much more difficult to retrieve horses than cars. This is due to the fact that it is difficult to distinguish characteristics of horses from other species (cow, dogs, ...), while, it is relatively easy to recover very precise features of a car (like the rims)

III Advanced Encoding Methods

III.A First order methods

QH1: Compare the dimension of VLAD and BoVW vectors for a given value of K. What should be the relation of the K in VLAD to the K in BoVW in order to obtain descriptors of the same dimension?

On one hand, the total dimension of a VLAD vector is $K_{VLAD} \times D$ where K is the number of centroids and D is the dimension of each descriptor. On the other hand, if we denote by K_{BoVW} the number of SIFT vectors associated with each visual words of $BoVW$, then in order to obtain descriptors of the same dimension, we must have: $K_{VLAD} \times D = K_{BovW}$ for each visual words.

QH2: Replace the encoding used in exercise1 with the VLAD encoding, and repeat the classification experiments for the three classes of Part I (Both linear and Hellinger kernel). How do the results compare to the BoVW encoding? Report mAP results in a table. No need to report all precision-recall curves.

	Kernels	motorbike	Aeroplane	Person
BoVW	Linear L_2	48.66	54.95	70.64
	Hellinger	63.25	70.72	77.39
VLAD	Linear	68.82	74.62	75.54
	Hellinger	75.42	75.56	78.86
FV	Linear	72.62	70.64	77.44
	Hellinger	**81.14**	**78.13**	**82.11**

Table 4: VLAD and FV Average Precision in % on the three classes for linear and Hellinger kernels

VLAD outperforms BoVW for all 3 classes.

III.B Second order methods

QI1: Replace the encoding used in exercise1 with the FV encoding, and repeat the classification experiments for the three classes of Part I. Report the results in the same table as QH2 so that you can see the performance of the three encoding methods side by side.

See Table 4. FV outperforms VLAD for all 3 classes.

QI2: What are the advantages or disadvantages of FV compared to VLAD in terms of computation time and storage/memory footprint - especially for a large number (hundreds of millions) of images.

The Fisher Vector needs to store K centroids of dimension d plus K times a matrix of covariance of dimension $d \times d$. In general, The total size of a Fisher vector is then $K \times (d + d \times d)$, and if the covariance is restricted to the diagonal, the fisher vector has only $2K \times d$ components in memory. In this particular case, the Fisher Vector takes 2 times more memory. Also, the computational cost of FV is higher as it is more difficult to compute a GMM for each K centers. In spite of these drawbacks, FV generally outperforms VLAD as it uses second order information.

IV Annexes

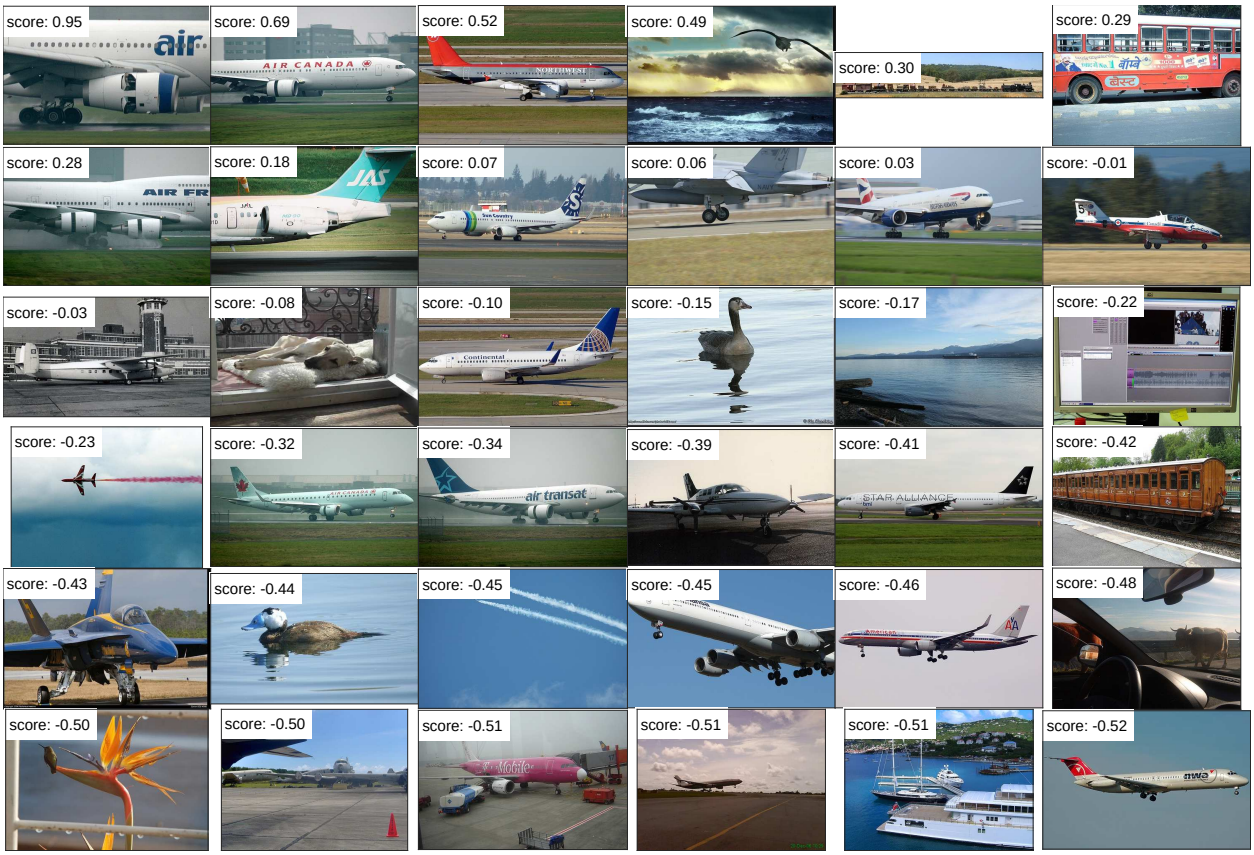


Figure 9: Top ranked images for the aeroplane class on the test data

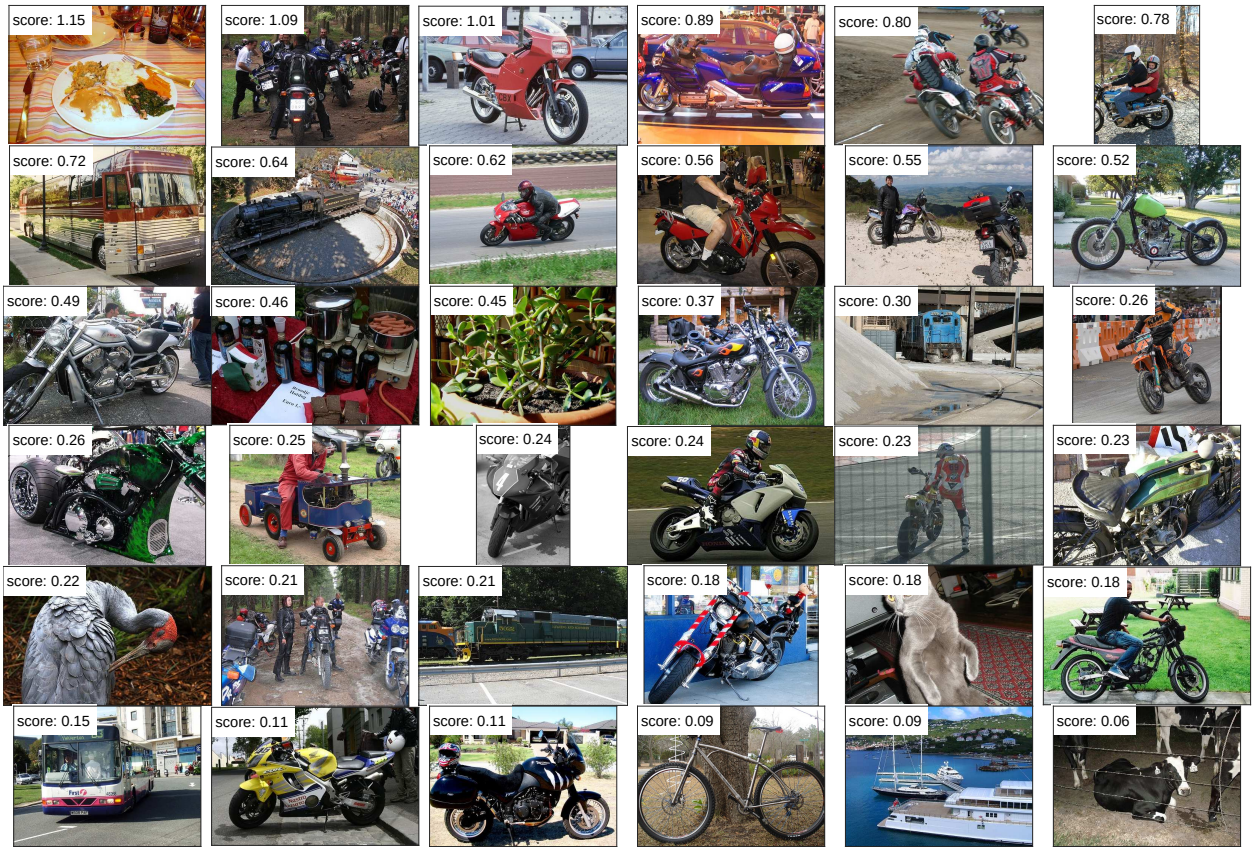


Figure 10: Top ranked images for the motorbike class on the test data

Figure 11: Top ranked images for the person class on the test data

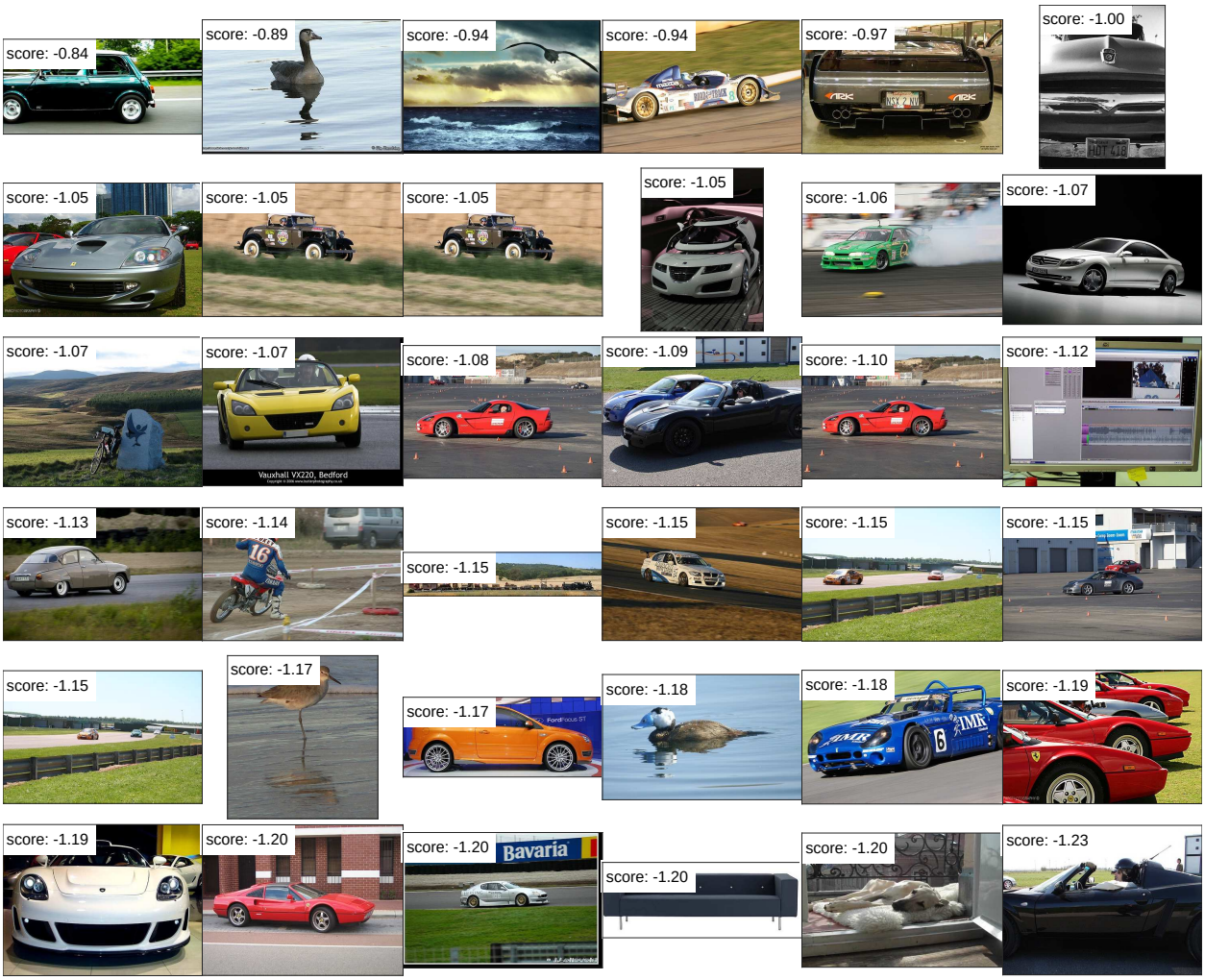


Figure 12: Top ranked 36 images using 5 images of cars

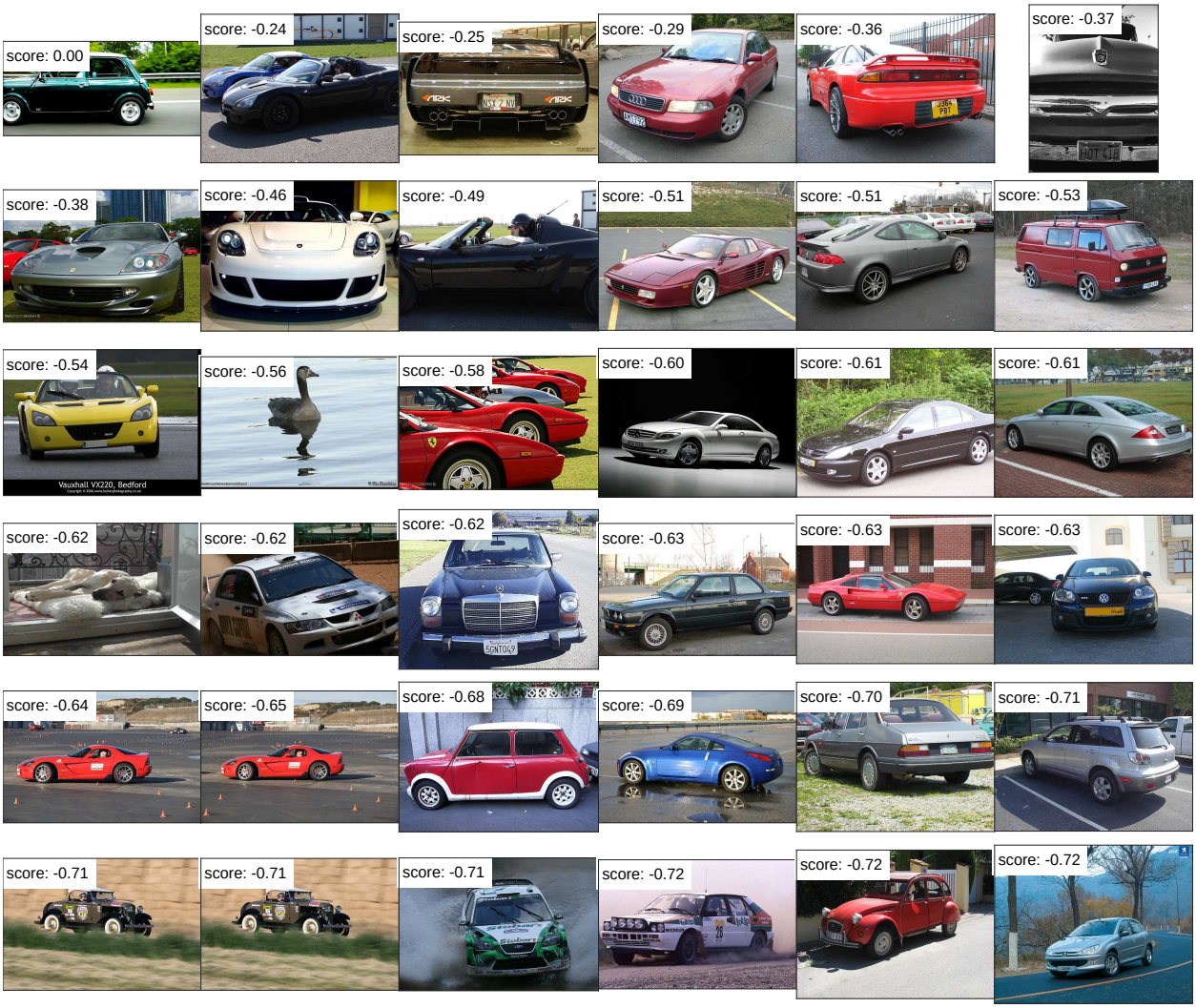


Figure 13: Top ranked 36 images using 10 images of cars

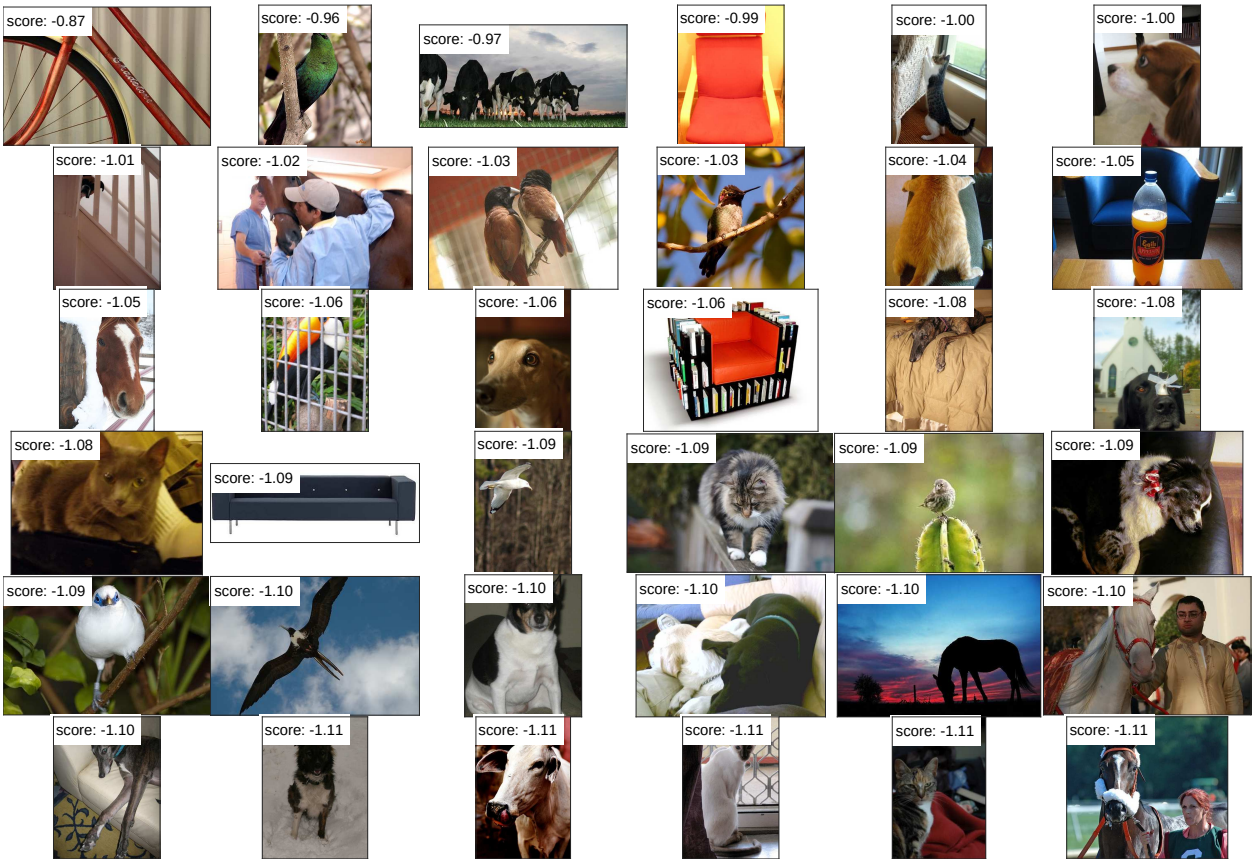


Figure 14: Top ranked 36 images using 5 images of horses

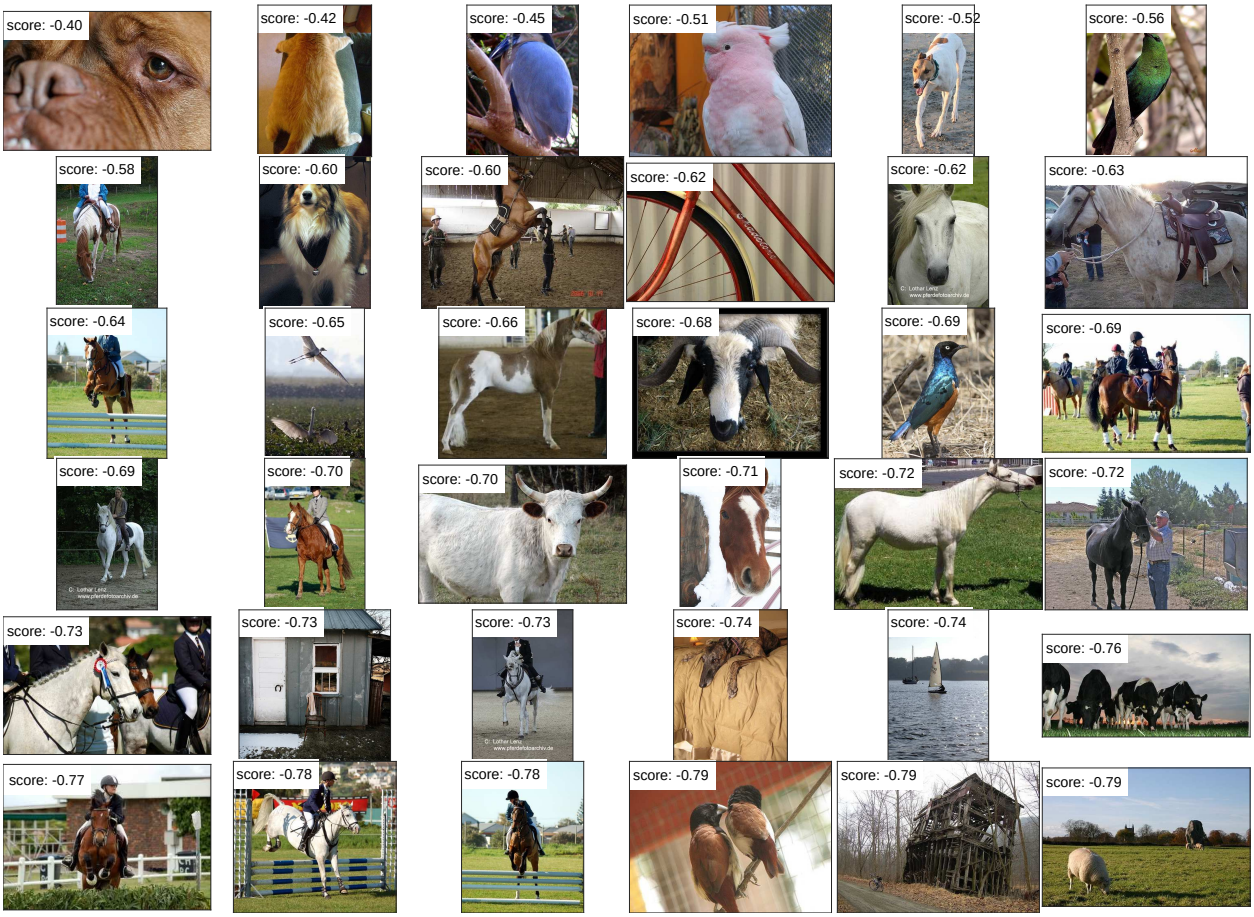


Figure 15: Top ranked 36 images using 10 images of horses