University of Amsterdam

Final Lab Report: Part 1

Bag-of-Words Image Classification

October 22, 2022

Students:
Roni Kremer
14564807

Tsatsral Mendsuren
14530775

Didier Merk
11037172

Tutor:
Qi Bi, Jiayi Shen

Group:
Practicals Group E

Course:
Computer Vision 1

1 Introduction

In this part of the final lab for computer Vision 1, a system for image classification will be implemented. Image classification can be described as the task of assigning a label or *class* to a specific image. In this assignment the system proposed will, given an image as input, classify an image as one of five classes. The algorithm used for the classification is based on the Bag-Of-Words concept, in which a visual vocabulary of an image is created, using the features of images and a histogram representation. This visual vocabulary can then be used to represent images by their visual word frequencies and train a classifier.

For this assignment the STL-10 dataset will be used, a dataset commonly used for developing unsupervised feature learning and deep learning (Adam Coates 2011). It contains 5000 train images and 8000 test images, which can be divided into 10 classes: airplane, bird, car, cat, deer, dog, horse, monkey, ship and truck. In this report the classifier is trained to assign an image to any of the five classes: airplane, bird, car, horse or ship.

2 Bag-of-Words Image Classification

2.1 Feature Extraction and Description

The Bag-Of-Words approach towards image classification consists of multiple steps. The first step of the pipeline is to extract and describe the features of images. This can be done in multiple ways, such as from SIFT densely sampled regions or keypoints, or using the HOG feature descriptor algorithm. The results of the different feature extraction methods used in this project are displayed per class in figure 1 below and figure 7, which can be found in the appendix.

The keypoints are generated with SIFT from the OpenCV library in Python. In the visualizations, the keypoint sizes are represented by the size of the circles. In the second method, the keypoints of the densely sampled regions are generated in a grid-like fashion with a window size of 7. Finally, in a similar grid-wise approach with a cell size of 4 pixels, the histograms of the local gradients from 8 directions are represented by the vectors (arrows of magnitude and direction) in the visualizations.

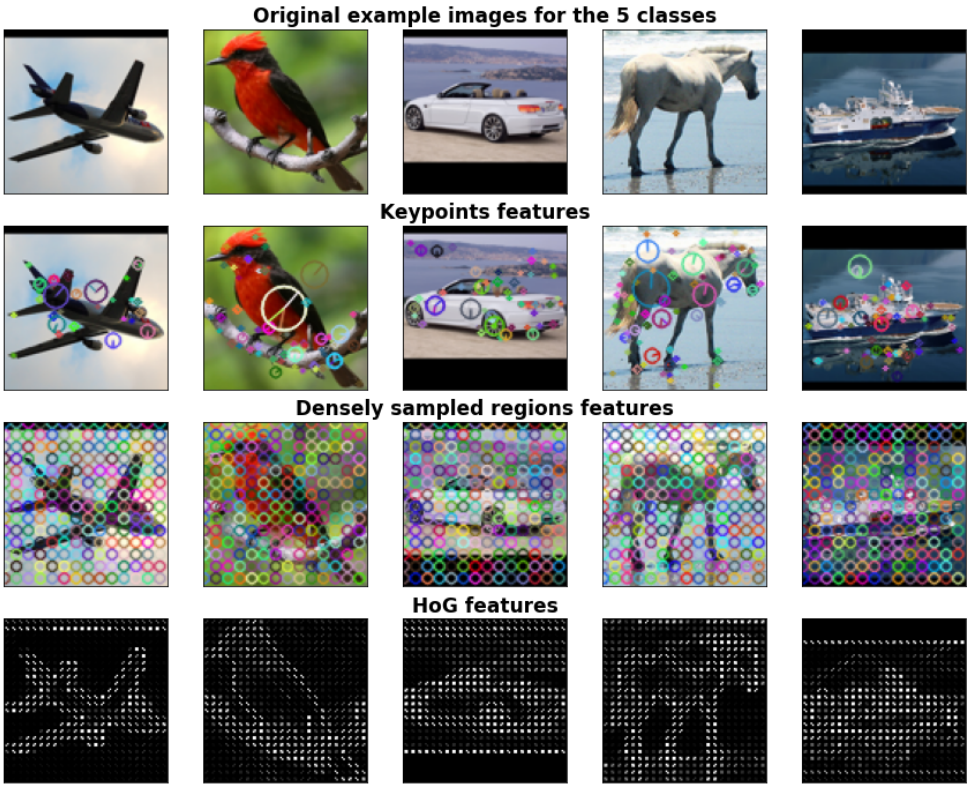

Figure 1: Visualization of the multiple different methods used to extract features, represented by images from all 5 classes (from left to right: airplane, bird, car, horse, ship).

2.2 Building Visual Vocabulary

In this step of the Bag-of-Words image classification pipeline we use the previously found image features to build a visual vocabulary using the K-means clustering algorithm from sklearn. All the features of a subset of the training images were used by the K-means algorithm to find a 1000 clusters, serving as our visual vocabulary.

Now, given this visual vocabulary, each image can be represented as a collection of 'visual words'. This is done by extracting the features of an image as before and assigning it to its closest clusters given by the K-means algorithm. After this, each image can be represented by a histogram of its visual words. For this assignment this was executed using different subset sizes of the training images, namely: 30%, 40%, 50% and 60%. The resulting histogram representations are displayed per class in figure 2.

Figure 2: The visual vocabularies displayed per class for all of the different sizes of the training images' subset. The K-means algorithm was trained with a fixed vocabulary size of a 1000 clusters. On the y-axis these 1000 clusters are displayed and on the x-axis we can see the normalised frequency of a specific cluster.

These visual vocabularies can now be used as some way to represent their class, and when a new image is received it can be compared to these different visual vocabularies to analyze which one it is most similar to. In figure 3 the difference between classes are displayed more clearly. The figure shows the visual vocabulary made using a subset size of 50%

and a 1000 clusters. It can be seen that the different classes are clearly separable by their visual vocabularies, when comparing for example the *plane* and *bird* classes. However, some classes are more similar to each other than others, such as the *bird* and the *horses* classes.

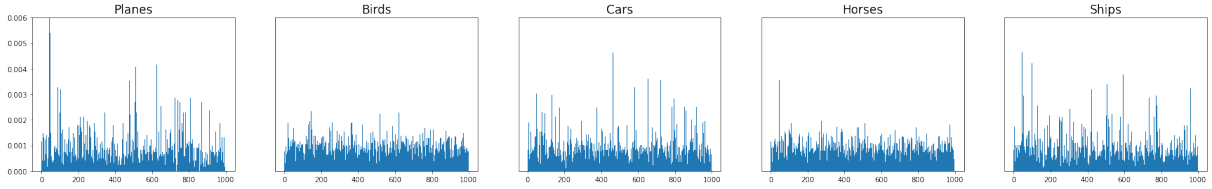

Figure 3: The visual vocabularies displayed per class trained using a subset size of 50%. The K-means algorithm was trained with a fixed vocabulary size of a 1000 clusters. On the y-axis these 1000 clusters are displayed and on the x-axis we can see the normalised frequency of a specific cluster.

2.3 Classification

The goal of the assignment has always been to be able to accurately classify an image. This will be done by training a Suport Vector Machine model (by sklearn) on the visual vocabularies found in section 2.2 and 2.1. This training resulted in five binary classifiers (one for each class: airplane, bird, car, horse and ship) that each output the probability of an image belonging to one class or to any of the other four classes.

Each binary SVM classifier was trained using the same image subset of 1500 training images that were not used to find the visual vocabulary. This resulted in 150 images with a 'positive' label for each class and 600 images with a 'negative' label for each class (the other 4 classes), to be used for the training of the model. The five resulting classifiers could then be used for evaluation.

3 Evaluation

A new image can now be put 'into' the five SVM classifiers and the model which outputs the highest probability of belonging to its 'positive class', is the class we assign to an image. This also means a set of new test images can be used on each of the five classifiers and be sorted on their probability of belonging to that specific class.

In addition we can measure the performance of the system quantitatively using the Mean

Average Precision (mAP). The Average Precision for a single class c is defined as follows:

$$AP_c = \frac{1}{m_c} \sum_{i=1}^{n} \frac{f_c(x_i)}{i}$$

In this equation m_c is the number of image of a certain class, n is the total number of images, x_i is the i^{th} image in the ranked list as constructed by the SVM classifiers and f_c is a function which returns the number of images of class c in the first i images if x_i is of class c and 0 otherwise. The Mean Average Precision is then calculated by taking the mean of the average precision of all classes. The mAP allows us to give a value to the performance of our classifying system and select the model that performs the best.

Four steps were taken to optimize this system of Bag-of-Words image classification and eventually select the best performing settings.

- Step 1: The K-means clustering algorithm used to create the visual vocabulary, is trained on different sizes of the training images subset. However, it is trained under the fixed vocabulary size of 1000 clusters. The SVM classifiers will then be trained on these different K-means models and using SIFT Keypoints as the image features extractor. The mAP of these different classifying systems can then be calculated and the best one can be selected.

- Step 2: The size of the training images subset is set to the one that performed best in step 1. This time the K-means is trained on said size, however the vocabulary sizes are varied from 500 to 2000. The SVM models will once again be trained on these different K-means models using SIFT Keypoints as the image features extractor. The mAP of these different systems can be calculated and the best one can be selected.

- Step 3: Now the best performing subset size and vocabulary size have been found, the different image feature extracting algorithms are tested. The K-means algorithm is trained using the vocabulary and subset size found in step 1 and 2, and the SVM-models will be trained in two different ways using the SIFT Keypoints and the HOG algorithm as image feature extracting methods. Once again the mAPs can be calculated and the best performing system can be selected.

- Step 4: Finally, the most effective vocabulary and subset size have been found in addition to the most effective feature extracting algorithm. A hyperparameter optimization can be performed on the Support Vector Machine model.

3.1 Fixed Vocabulary Size

In this section the results of training the K-means clustering algorithm with different training image subset sizes and a fixed vocabulary size of 1000 are displayed. In table 1 the mAPs of the different subset sizes are displayed. In figure 4 the top-5 and bottom-5 ranked test images for the classifier with the highest mAP are shown. In this case this is the classifying system trained on an image subset of 40%. The visualizations of the top-5 and bottom-5 of the other classifiers can be found in the appendix (figures [810].

Class	30%	40%	50%	60%
Plane	0.392	0.382	0.398	0.416
Bird	0.061	0.056	0.074	0.034
Car	0.222	0.260	0.245	0.221
Horse	0.094	0.095	0.099	0.078
Ship	0.304	0.294	0.246	0.269
mAP	**0.215**	**0.217**	**0.212**	**0.204**

Table 1: The APs of the SVM classifiers for all classes. Each classifier was trained with visual vocabularies clustered using different subset sizes (30%, 40%, 50%, 60%) and a fixed vocabulary size of 1000. In bold we have the mAPs of each classifier; the highest mAP (using an train image subset size of 40%) is highlighted in green.

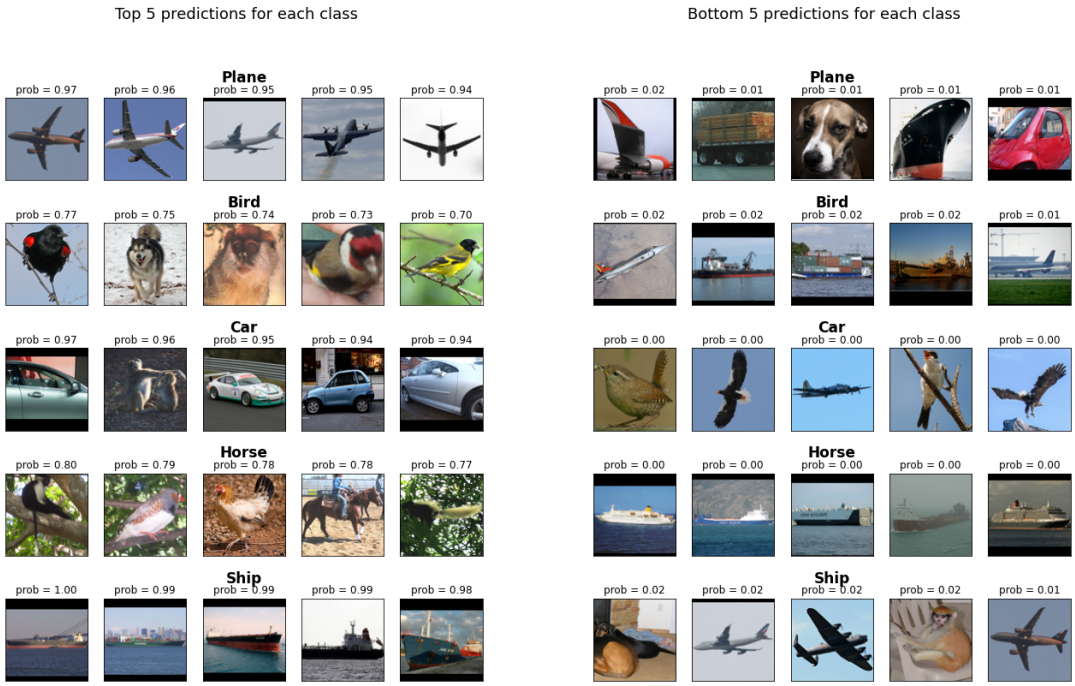

Figure 4: The top- and bottom five predictions of each class, predicted by the best performing SVM classifier (subset size of 40%). Images are ranked by their predicted probabilities of belonging to each class (displayed above the images). As mentioned a train images subset size of 40% and a vocabulary size of 1000 was used to train the K-means clusters. SIFT Keypoints was used as the images features for both SVM and K-means classification.

In table **??** it is shown that using a training images subset of 40% led to the highest mAP on the 2500 test images. However, all the mAP values are relatively close together. Similarly, the differences of the APs per class are not significant either, being mostly within 1-2 percent for all three sizes. In figure 4 and figures 8-10 it is visible that all the top five predictions had ships, cars and planes predicted correctly but for the remaining two other classes had wrongly predicted images in them. The decision was made to use 40% training images to train the visual vocabulary in further optimization because this is a moderate size which balances both the model's reliability and training complexity.

Here, it is important to make a general remark about the mAP values in this report. The classifying system was trained to classify on five different classes. However, the test images subset contains 2500 images from 10 different classes. The choice was made to keep these images in the test subset, since it was not mentioned anywhere it was allowed to remove these images. A consequence of this, however, is that half of the test images will inherently be classified wrong, resulting in a lower overall mAP.

3.2 Fixed Image Subset Size

In this section the evaluations of classifying systems using different visual vocabulary sizes (500, 1000, 1500 and 2000), but a fixed train image subset size (40%), are discussed. In table 2 the APs and mAP with respect to each vocabulary size are displayed and in figure 5 the top- and bottom-5 ranked test images for the best performing classifier (visual vocabulary of size 2000) are shown. The others are found in the appendix (figures [11-12]).

Class	500	1000	1500	2000
Plane	0.382	0.382	0.413	0.440
Bird	0.046	0.056	0.066	0.070
Car	0.255	0.260	0.241	0.195
Horse	0.112	0.095	0.092	0.083
Ship	0.252	0.294	0.255	0.297
mAP	**0.209**	**0.217**	**0.213**	**0.217**

Table 2: Average Precision of all the SVM classifiers for all classes. Each classifier was trained using visual vocabularies of size 500, 1000, 1500 or 2000. In bold we have the mAPs of each classifier. In blue the previously found highest mAP and in green a new highest mAP are displayed.

As shown in table 2, the mAPs of the classifiers again are very similar. The systems trained using a vocabulary size of 1000 and 2000 even have the same combined highest mAP. This indicates that an increase of visual vocabulary size does not significantly improve the classification performance. Hence, it is decided to use a visual vocabulary size of 1000 when further optimizing the image classification system, due to its shorter

training time and it having the combined highest mAP.

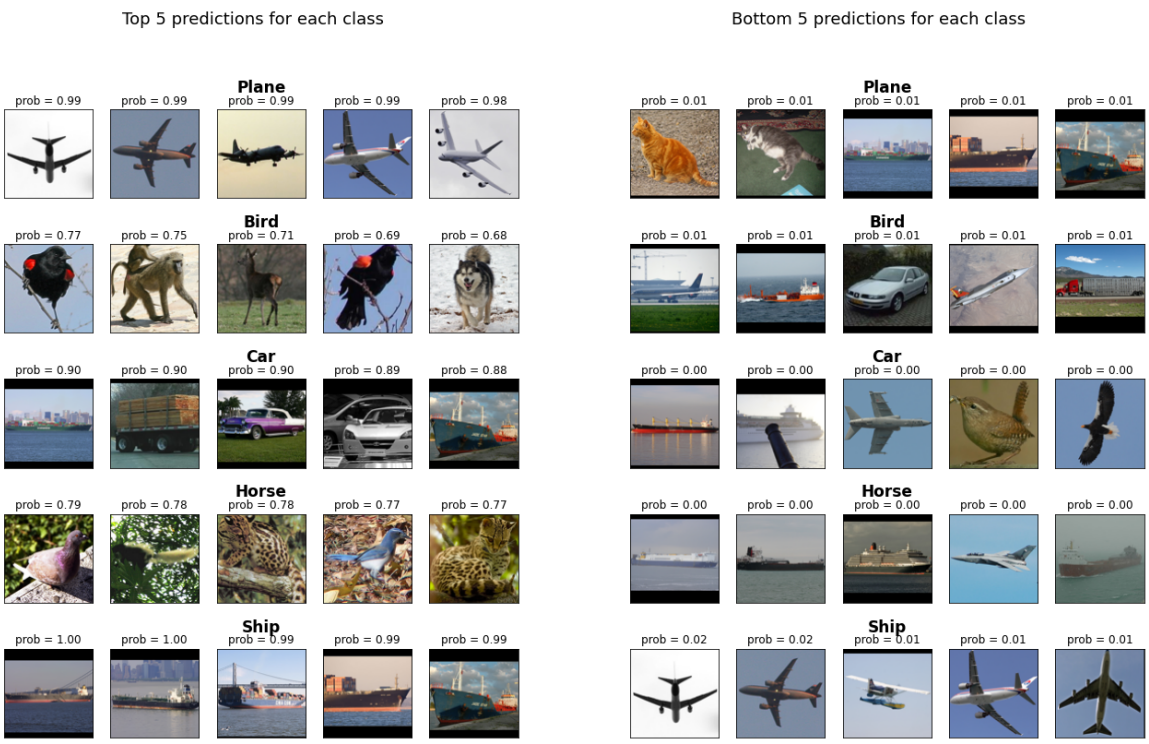

Figure 5: The top- and bottom five predictions of each class, predicted by the best performing SVM classifier (vocabulary size of 2000). Images are ranked by their predicted probabilities of belonging to each class (displayed above the images). As mentioned a train images subset size of 40% and a vocabulary size of 2000 was used to train the K-means clusters. SIFT Keypoints was used as the images features for both SVM and K-means classification.

3.3 Feature Extractors

In previous experiments, the classifying system was trained using SIFT Keypoints as the image feature extraction algorithm. In this part of the assignment the performance is compared to using a different feature extraction algorithm: *Histogram of Oriented Gradients* (HOG). The feature output of this algorithm can be seen in figure 1 and 7.

To continue optimizing the Bag-of-Words classifying system, the settings with the current highest mAPs are chosen: a training image subset size of 40% (section 3.1) and a vocabulary size of 1000 (see section 3.2). The mAPs of the classifiers trained using these optimal settings and the SIFT and HOG feature extractors are displayed in table 3. The top-5 and bottom-5 ranked images of the system with the highest mAP, which is the HOG feature extractor, are displayed in figure 6.

Class	SIFT	HOG
Plane	0.382	0.447
Bird	0.056	0.333
Car	0.260	0.368
Horse	0.095	0.181
Ship	0.294	0.433
mAP	**0.217**	**0.352**

Table 3: In this table the Average Precision (AP) for each class and the Mean Average Precision (mAP) over all classes is displayed, for both the SIFT keypoints and HOG feature extractor. The system was trained using a vocabulary size of 1000 and image subset size of 40%. Shown in blue is the mAP value that until now had been the highest and in green it is shown that a new highest mAP value has been discovered.

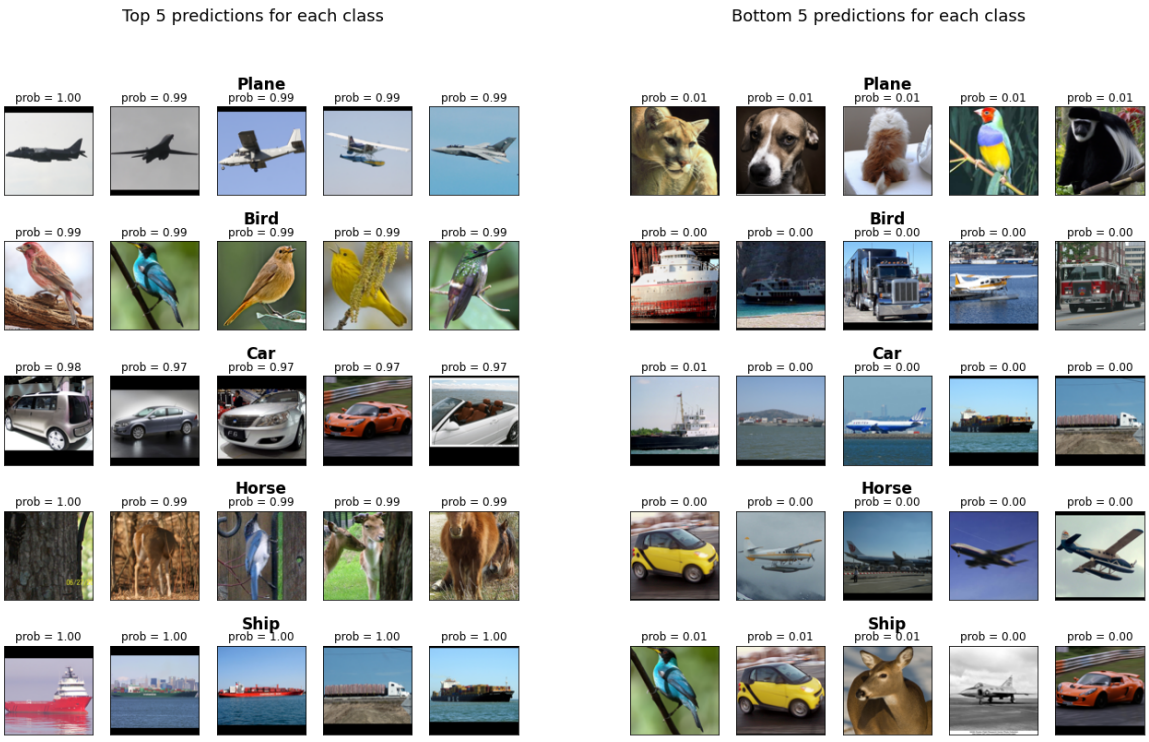

Figure 6: Here the top- and bottom five predictions of each class as predicted by the SVM classifiers using the HOG feature extractor are displayed. Images are ranked by their probabilities of belonging to each class (as displayed above each image). The visual vocabulary was trained by the K-means classifier using a image subset size of **40%** and vocabulary size of **1000**. And as mentioned the both the SVM and K-means were trained using the HOG algorithm as the feature extractor.

The previous highest mAP found was 0.217, in table 3 it can be shown that the HOG algorithm performs significantly better. In addition when comparing the top 5 predictions of figure 4 and 6, the HOG descriptor classifies the images more accurately. This means the best performing system found so far, trains using a vocabulary size of 1000, train image subset size of 40% and the HOG feature extractor.

3.4 Hyper-parameter analysis

The training of the SVM model depends on certain *hyper-parameters*, which can be described as a set of parameters that are not learned by the algorithm. In this assignment the *kernel* parameter of the `svm.SVC()` model has been chosen to be optimized. As the name suggest, this parameter specifies the type of kernel used in the algorithm. Previously the default kernal option *'rbf'* has been used; for this part of the assignment the other three non-precomputed options ('linear', 'poly' and 'sigmoid') will be evaluated. The results of using these different kernel types are shown in table 4.

Class	*linear*	*poly*	*rbf*	*sigmoid*
Plane	0.379	0.299	0.447	0.005
Bird	0.218	0.189	0.333	0.062
Car	0.346	0.286	0.368	0.124
Horse	0.171	0.163	0.181	0.146
Ship	0.271	0.314	0.433	0.044
mAP	**0.277**	**0.250**	**0.352**	**0.076**

Table 4: Result of training the SVM on different kernel types. The K-means for this algorithm was trained using the previously found optimal vocabulary size of 1000 and train image subset size of 40% and both the K-means and the SVm model were trained using the HOG feature extractor.

From this table it can be concluded that using a different kernel type than the default *rbf* kernel that was previously used, does not improve the precision of the classifier. The top- and bottom 5 predictions of this are displayed in the previous section (figure 6) and the remaining can be found in the appendix (figures 13-15).

4 Conclusion

In this assignment, the settings of the Bag-Of-Words image classification method on 5 classes of the STL-10 dataset were optimized. The best setting found is to use HOG as the feature extractor (8 directions, cell size of 4), a K-means algorithm with vocabulary size of 1000, trained with 2000 images (40% of the total 5000 training images in STL-10 dataset). In addition to SVM models with the *rbf* kernel as the image classifiers.

Though achieved ascertained conclusion, the optimization process may be extended in future experiments. For example, different thresholds on the size of keypoints may be tried to filter out the less relevant features found by SIFT Keypoints; the HoG feature extractor may be tested on more settings instead of only the best setting found for SIFT Keypoints extractor; and other hyperparameters of the SVM model may be altered for a more completed optimization process.

References

Adam Coates Honglak Lee, Andrew Y. Ng (2011). "STL-10 dataset". In: URL: http://cs.stanford.edu/~acoates/stl10.

5 Appendix

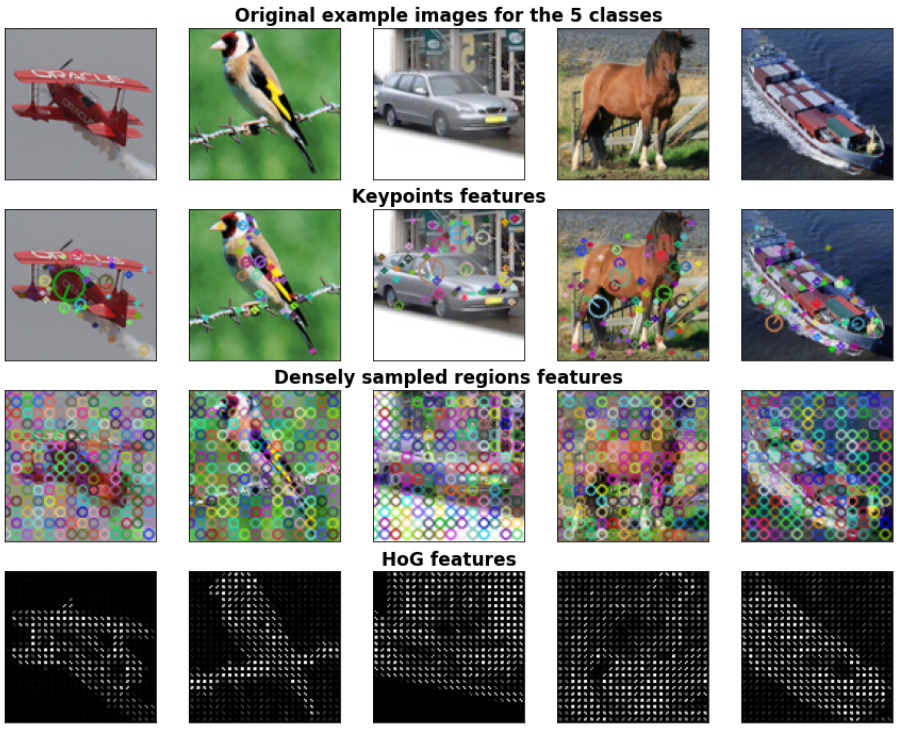

Figure 7: Visualization of the multiple different methods used to extract features, represented by images from all 5 classes (from left to right: airplane, bird, car, horse, ship).

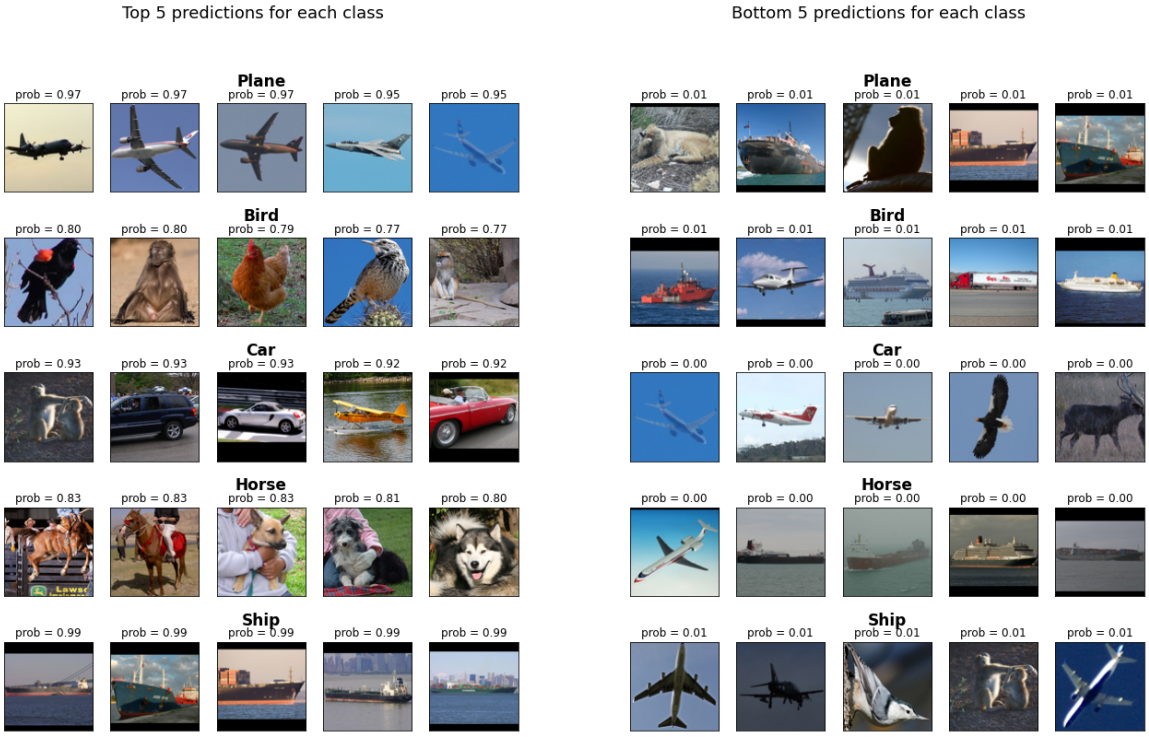

Figure 8: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 1000, train image subset size of 30% and SIFT keypoints as feature extractor.

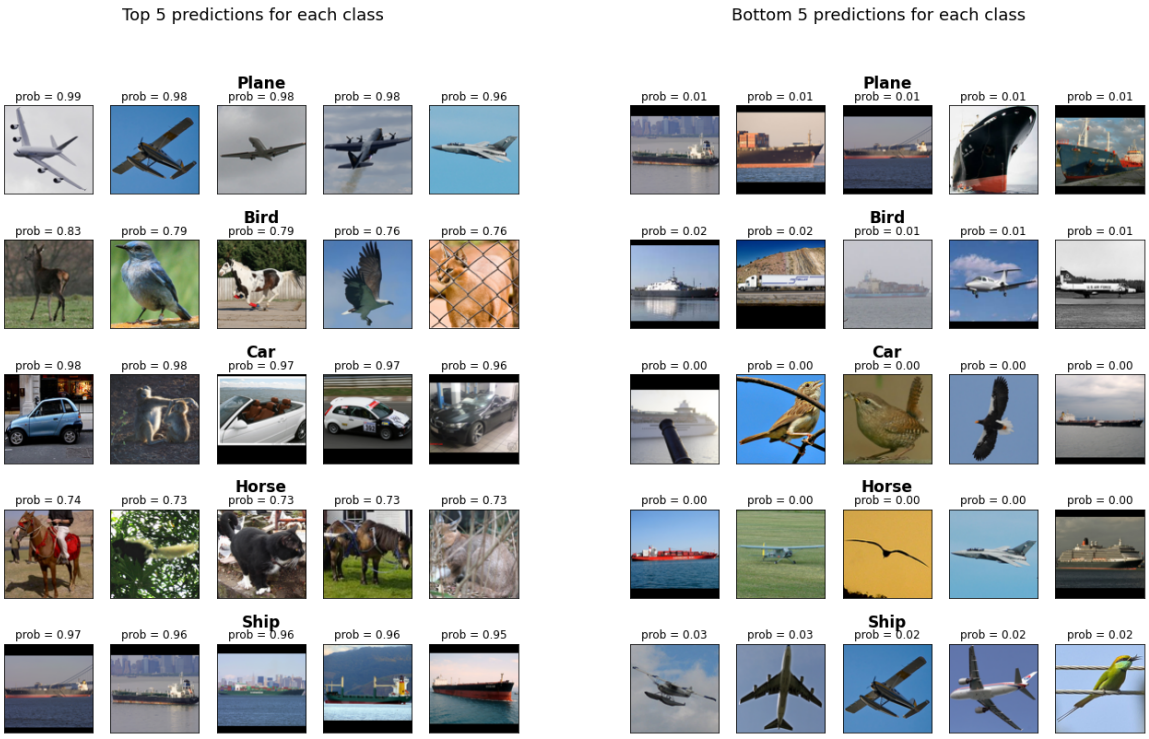

Figure 9: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 1000, train image subset size of 50% and SIFT keypoints as feature extractor.

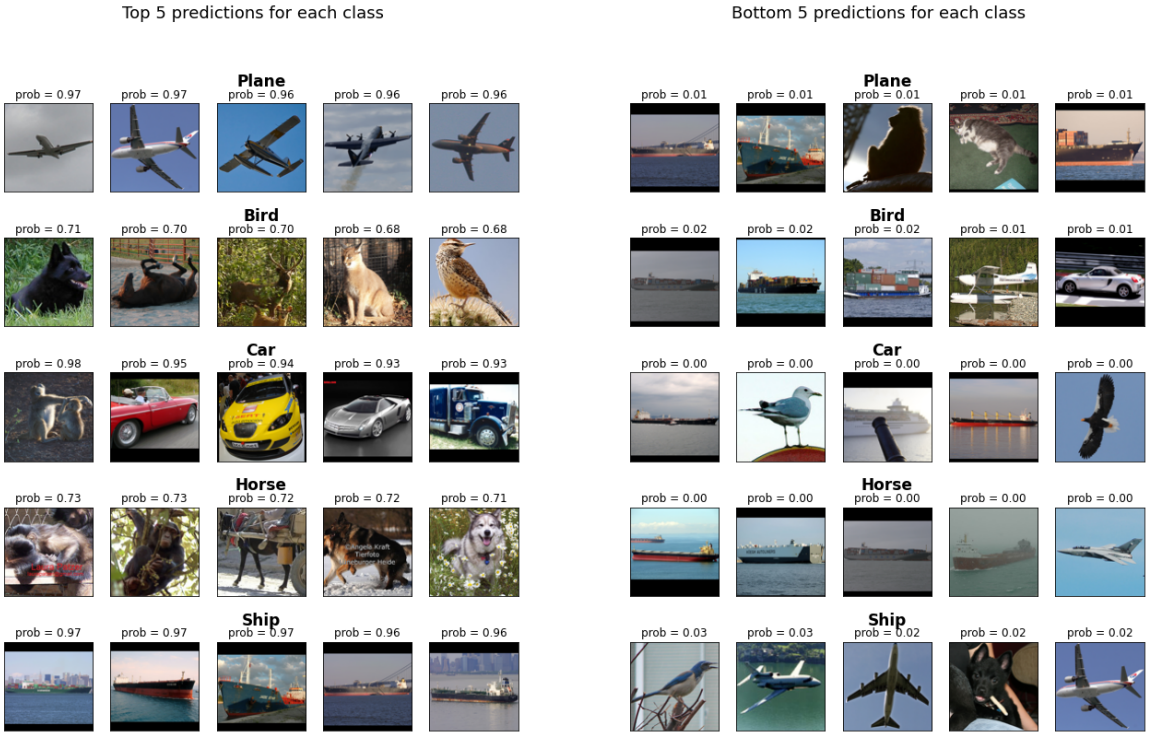

Figure 10: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 1000, train image subset size of 60% and SIFT keypoints as feature extractor.

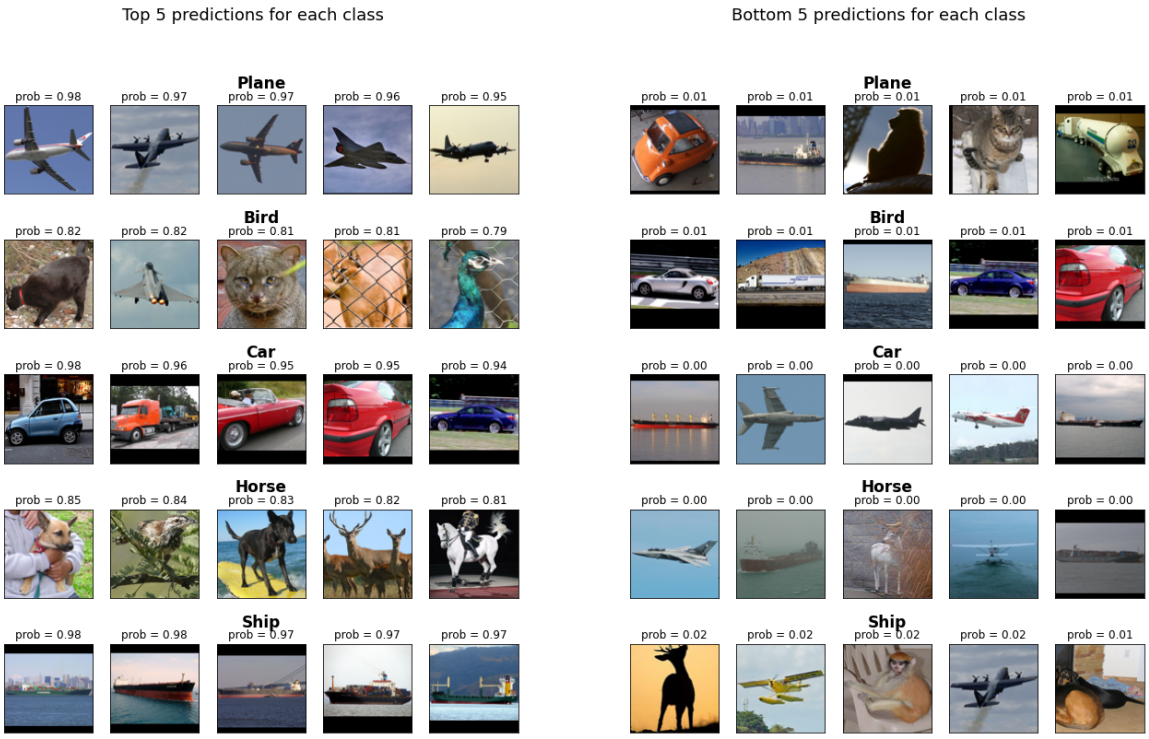

Figure 11: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 500, train image subset size of 40% and SIFT keypoints as feature extractor.

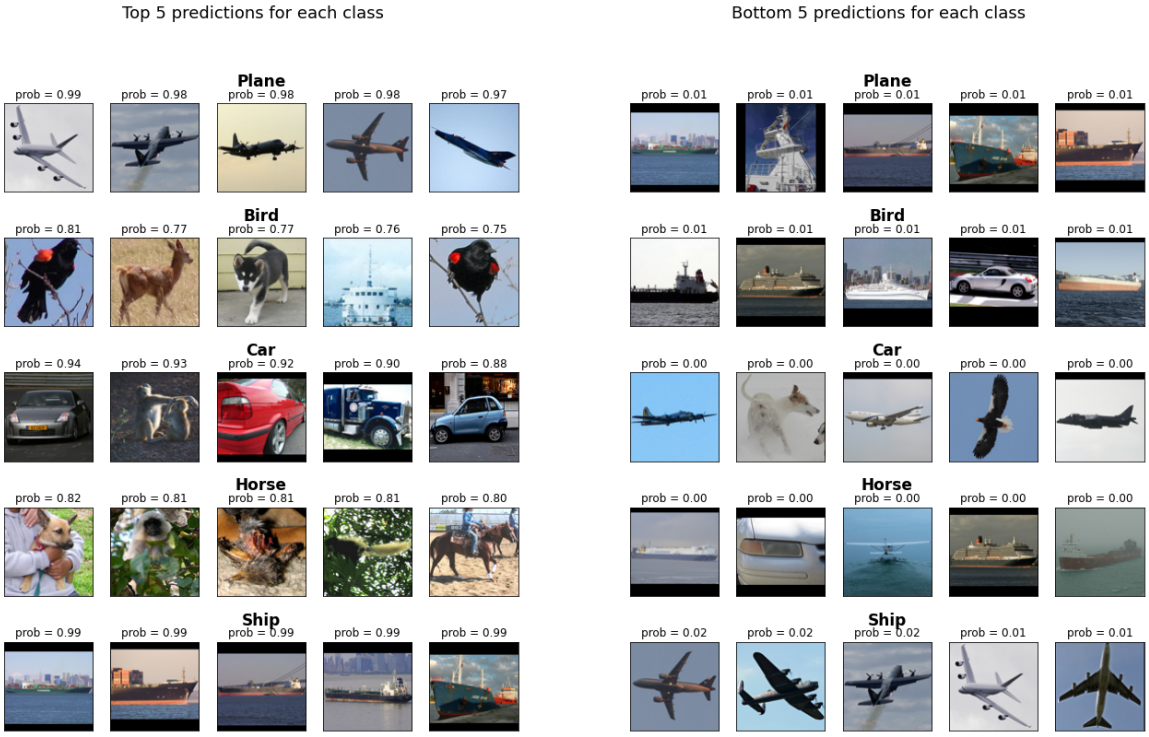

Figure 12: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 1500, train image subset size of 40% and SIFT keypoints as feature extractor.

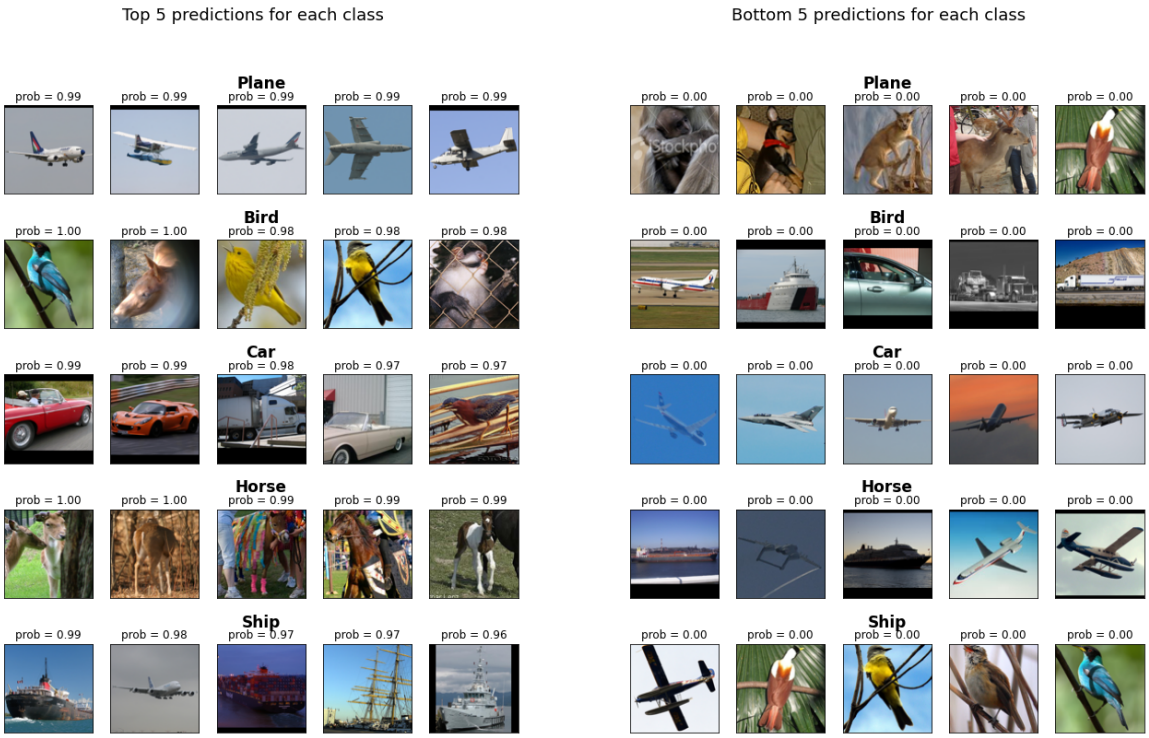

Figure 13: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 1000, train image subset size of 40% and HOG with a linear kernel as feature extractor.

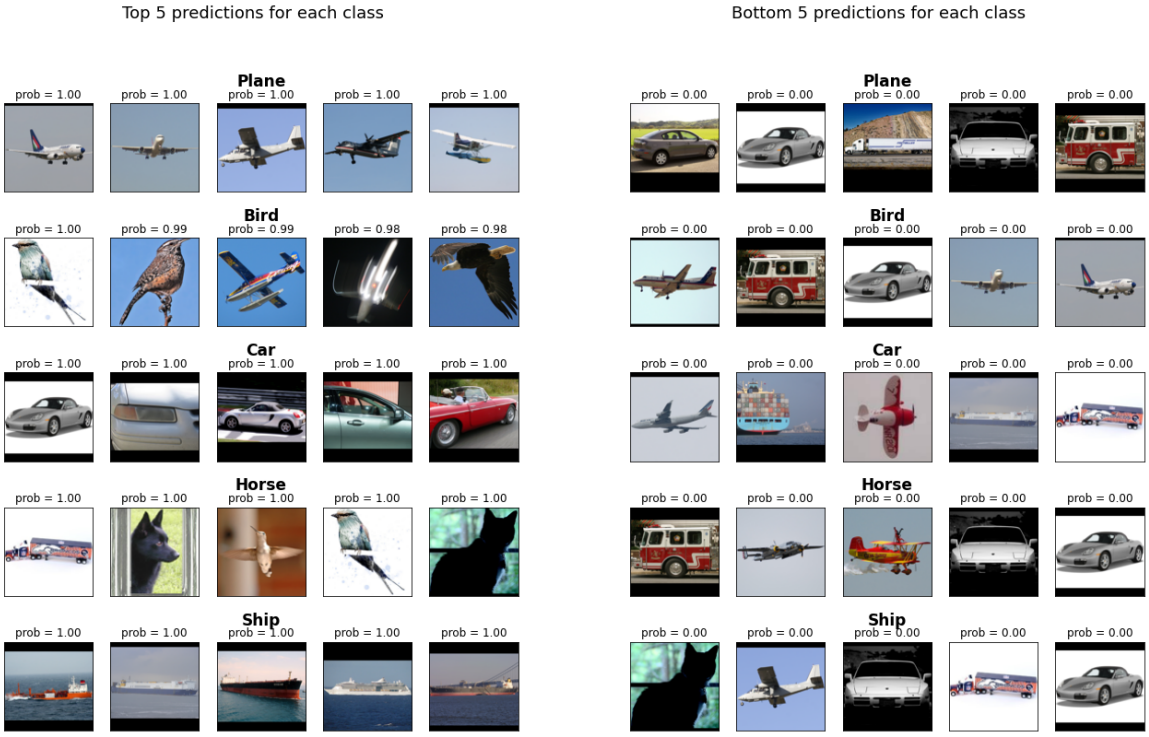

Figure 14: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 1000, train image subset size of 40% and HOG with a poly kernel as feature extractor.

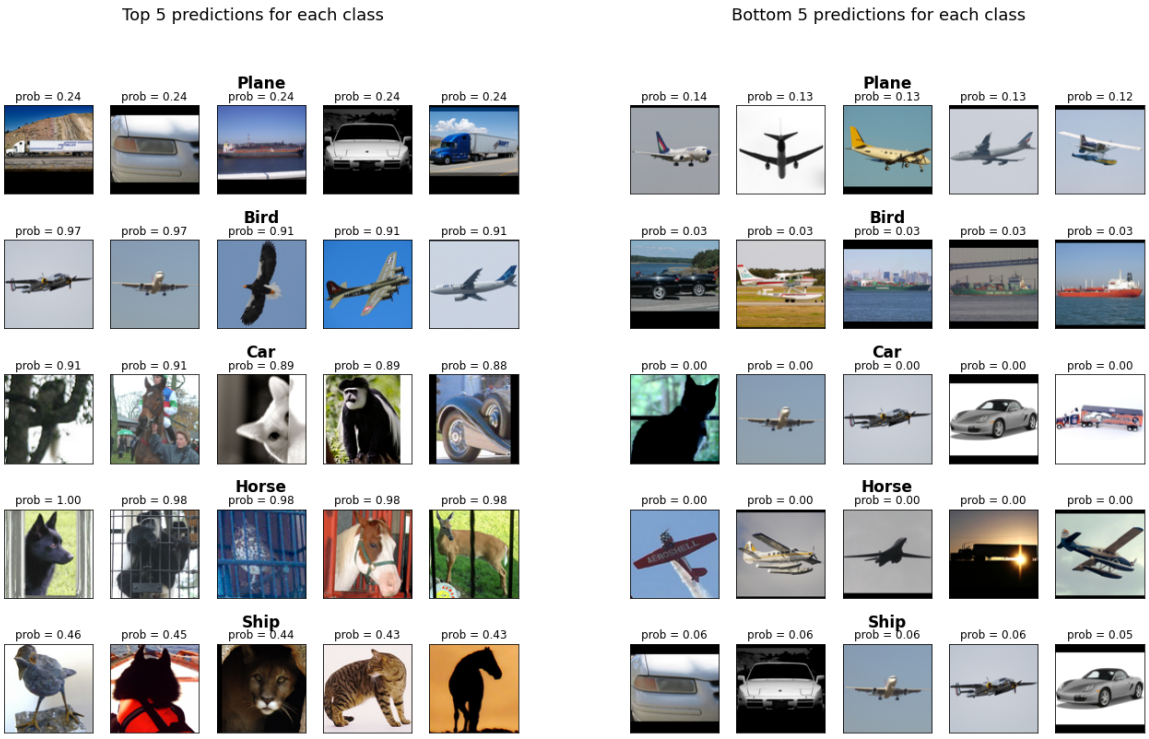

Figure 15: Top- and bottom-5 prediction of the SVM classifiers trained using a vocabulary size of 1000, train image subset size of 40% and HOG with a sigmoid kernel as feature extractor.