Evaluating the four datasets

The previous exercise included four datasets, which you will have evaluated. I hope you now have a good idea of why some worked better than others, but I will summarise.

Dataset 1: Cats and dogs

This dataset was taken carefully to ensure that it would work well. The toys used in the images looked very different. We took many photos of each from different directions and perspectives, but always on the same background to ensure that the background didn't confuse things. The test set was chosen at random to make sure that they were similar to the training set.

This dataset should have worked well, though it might not generalise too well to other backgrounds.

Dataset 2: Cats, dogs and dolphins

This is the same as dataset 1 with one extra class: a dolphin. Adding an extra class shouldn't have affected the performance too much.

Dataset 3: Rebecca and Marco

This was a dataset of people. We humans are extremely good at recognising other humans. Our perceptual system is specialised for recognising faces and we have a lot of experience. But that doesn't mean it is an easy task. Quite the opposite, the visual differences between people are really quite subtle. In particular, while we are very good at telling the difference between women and men, the differences are very small from the point of view of a computer. How does a computer know that the subtle differences in eyes, mouth and face shape are the important ones to pay attention to?

This dataset was designed to confuse the computer. While we were trying to tell the difference between two people, there were a lot of other, more obvious visual features for the computer to see. All of the training images of me were in a blue hoodie and Rebecca was wearing black. We were also on different backgrounds. She was in front of a window while I was against a blank wall. These are much more obvious features to learn than our faces. To demonstrate this, we swapped in the test set.

You should remember that the machine learning algorithm was doing its job: it was learning the difference between two sets of images. If you look at the majority of pixels in

each image, they were images of a blue hoodie and a large window. The human beings involved were only minor details. We pay much more attention to humans because we are human, but the computer doesn't know we are important!

Dataset 4: More toys

This dataset returns to the toys from examples 1 and 2, but puts them in more realistic images. The backgrounds are a lot more varied, and the size of the toys in each image is more varied. The results are likely to be a lot more mixed. In cases where the toy is one element in a very distracting background (for example, the doll's house or lego bucket) the classifier is likely to work less well as it is less obvious what the algorithm should learn. Other examples are likely to work a lot better. Try mixing this dataset up: change what is in the training and test set and see which get learned well.

Conclusions

I hope this exercise has given you an idea of some of the things that are easy and hard to learn. It should have also helped you understand how to design good train and test sets.

Many of the examples that do not work well could be made to work better. Dataset 4 has relatively few examples given that the images are very varied. With more data it is likely to work better. A different algorithm might also help. It used a neural network for feature extraction and a k-nearest neighbour. Using only a neural network might work better (but would need more data). Also a feature extractor that was specialised on faces might make even the very badly designed dataset 3 work well.