MVA - Deep Learning In Practice
Pratical session on Image Retrieval

Reda BAHI SLAOUI - Souhaib ATTAIKI

March, 2019

Question 1

Each row of the matrix represents the 4096 features for each image in the database computed using AlexNet network.

Question 2

These features are extracted from the penultimate layer of the AlexNet network, which is the layer just before the classification layer. The size of this layer is 4096, which explains the size of the features. To extract the features, an image is processed by the pre-trained network, and the weights of the penultimate layer are used as features.

Question 3

It can be noticed that the features used do not distinguish between the different classes correctly. The class that is well clustered is "radcliffe_camera" and among the classes that are not well clustered, we can mention "christ_church".

Question 4

Since the first network is training on ImageNet images, which are images of many different classes (animals, everyday objects, nature, etc.), the features extracted are general and allow to classify these different classes, whereas with finetuning, the network will focus on images that are practically the same classes (building image), so the features learned will be more discriminated and more representative of this kind of images. As a result, better results are to be expected.

Question 5

The number of neurons in the output layer represents the number of classes on which the network is trained. And since in finetuning, we freeze the weights of the first layers, and we learn on another dataset, which does not always have the same number of classes, it is normal that the last layer is changed to adapt to the new dataset.

Question 6a

We initialize the layers of model_1b for finetuning by using the weights of the trained model_1a on ImageNet.

Question 6b

The results have improved a little (AP=24 versus AP=20), but the clustering is still not perfect. The same classes that cluster well before are the same using the new features, as well as for bad clusters.

Question 7

Images need to be resized to 224x224 before they can be fed to AlexNet because the input size of the fully connected layers is fixed, so if the image is bigger, the size of the obtained feature map just before FC layers will be bigger than expected, and this will cause dimension mismatch error. This downsampling degrades the resulats as a lot of information about the texture, which can be quite discriminative, is lost.

Question 8

The size of the feature representation changes because we use the output of the generalized mean pooling layer, which outputs the spatial average of the feature maps from the last convolutional layer, which has 256 dimension.

Question 9

Dimensionality reduction is very important in image retrieval, since we are usually quering for the nearest neighbors in very big datasets, and thus for this to be computationally efficient, the features dimension shouldn't be too high.

Question 10

First of all, we can see that the results obtained are better than those obtained before (AP=64), and that clustering is more discriminating. It is clear (except for a few outliers) that many classes are condensed in a region of space, which is a sign of good clustering.

Question 11

It can be seen that similarly labeled images are concentrated in a some part of the space, but there is no particular structure (unless we want to consider an elliptical shape, but this is not always true, and the variance is very large).

Question 12

We use generalized mean pooling because the pooling parameter p can be manually set or learned since this operation is differentiable and can be part of the back-propagation. The latter option has proven to give better results which is intuitive.

Question 13

The operation that the layer model_1d.adpool doing is a power-average adaptive pooling, where for a signal X, the output is: $(\sum_i X_i^p)^{\frac{1}{p}}$. MaxPooling and AveragePooling are special cases where

p=∞ and p=1 respectively. The feature vector finally consists of a single value per feature map, i.e. the generalized-mean activation, and its dimensionality is equal to K (number of feature maps).

Question 14

This model gives AP=11.38, which is lower than the result of model 1c (AP=69). The clustering in the t-sne is also not improved, as many classes are not well distinguished.

Question 15

It also performs worse than model 1b, which had AP=24.67.

Question 16

Once the unlabeled features are included, the data doesn't seem to be separable in the t-sne plane. Therefore it would be useful to train a model to first separate labeled from unlabeled data.

Question 17

The unlabeled features seem to occupy an ellipse in the t-sne plane uniformly.

Question 18

One way to train a model to separate labeled from unlabeled data could be to use a triplet loss, or to use other "community detection" techniques.

Question 19

For the model trained for retrieval (with triplet loss), the different classes are more separable and are well clustered. The unlabeled dataset now occupy a circle uniformly.

Question 20

Without data augmentation, we have AP=55.24, and with it, we have AP=58.36, thus data augmentation techniques such as cropping, pixel jittering, rotation, and tilting, actually improves the result, as expected.

Question 21

It also improves the clustering. Now the clusters are better separated, except for some of the classes, which have not been adequately clustered yet, and this is due to the fact that the used data augmentation techniques are not useful and effective on them, so this can be solved by other data augmentation techniques.

Question 22

We can use other data augmentation techniques such as scaling, affine transformations, horizontal and vertical flips. We can also use more advanced data augmentation techniques such as GAN (Generative Adversarial Networks) or VAE (Variational AutoEncoders).

Question 24

A larger architecture resulted here in a better AP (AP=64.59), because our network is big enough to model complex distributions. However, this will not always lead to better result, as we may start overfitting to the training data (bias-variance dilemma).