Project 1 Report: Feature Extraction and Transfer Learning

1 Team members

The team members and their corresponding responsibilities are listed below:

- Ziyu Wang: data pre-processing and visualization
- Shichao Li: wavelet scattering feature extraction and training for MNIST
- Zhenzhen Huang: CNN feature visualization for Raphael paintings and training

2 Wavelet scattering experiments on MNIST

2.1 Feature Extraction

We use scale parameter of 2 and obtain 81 coefficients for each output position. The output feature map is of size 7 by 7. After flattening an image is represented by a $81 * 7 * 7 = 3969$ long feature vector.

As shown in Fig. 1, we use t-sne to learn a two dimension manifold of 5000 training samples for both raw pixels and their corresponding feature vectors. Compared to the raw pixels, the feature vectors are much easier to separate for different classes. This means wavelet scattering transform extracts useful information that can help distinguishing different classes.

2.2 Classification

After the feature extraction, we use a simple logistic regression classifier. We use stochastic gradient descent for optimization and set learning rate to 0.01. We train for 30 epochs and a momentum of 0.9 is used. The final testing accuracy is 98.49. Without any learning from data, the pre-computed wavelet filters already achieve high performance on this dataset.

3 Transfer Learning for Raphael paintings

Here we use a pre-trained convolutional neural network (CNN) as a fixed feature extractor for classifying Raphael paintings,

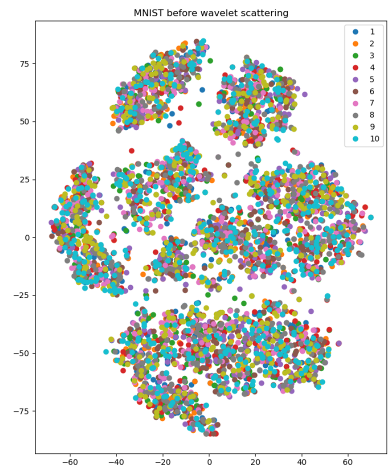
3.1 Feature Extraction

We choose a VGG19 model pre-trained on ImageNet to extract features. Removing the dense layers, the features are extracted from the last convolution layer $(7, 7, 512)$ for transfer learning. And we flatten extracted features to (25088) for visualization.

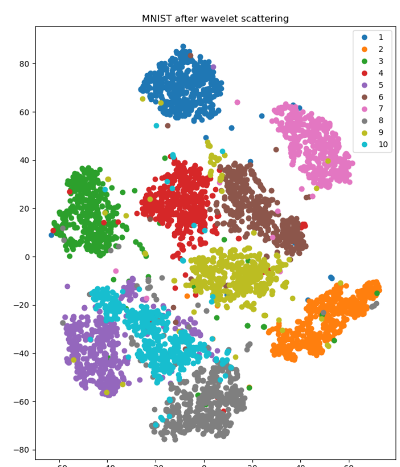
3.2 Visualization

Since the number of features from pre-trained CNN19 is 25088, we would like to reduce the features to low-dimension and then visualize the important features. The first method is using PCA method

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.



(a) Before wavelet scattering transformation



(b) After wavelet scattering transformation

Figure 1: T-SNE visualization of 5000 random images from MNIST training set before and after wavelet scattering transform.

and examining the explained variance ratio, we could have an idea of how many linear features are required to describe the data. And the other method is using nonlinear manifold embeddings to output a two-dimensional projection of all the images, which give us an intuition about the layout of features in the first two dimensions.

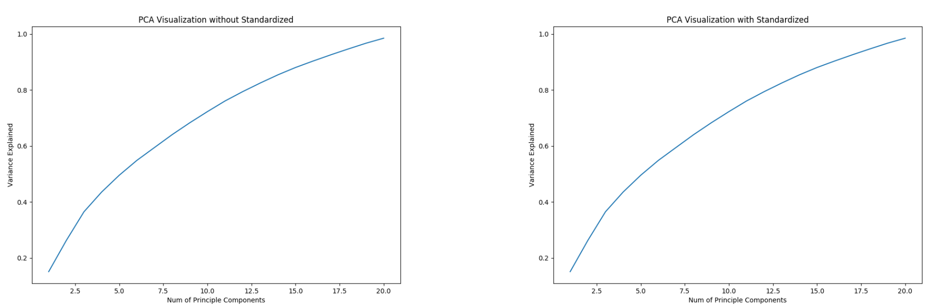
Before implementing visualization methods, We compare the performance of original feature data and standardized feature data in Table 1. The variance of original feature data is significant large comparing with the mean, which means that the fluctuation of original feature data can not be ignored. The statistics of standardized feature data show stability and less fluctuation.

	mean	var
Original Feature	1.047	25.945
Standardized Feature	0.045	0.033

Table 1: The mean and variance of original features and standardized features. The standardized features is scaled and translated such that it is in the given range on the training set, that is, between zero and one.

3.2.1 PCA

By PCA We explore the change of explained variance with the number of principle components. The result is shown in Figure 2. There is insignificant different from original feature data and standardized feature data. As the number of principle components increases, the explained variance also increases. The figure shows that we can reduce from 25088 to 19 dimensions while retaining 95% of variance. A quarter of principle components can explain 55.00% of variance and half of principle components can explain 76.00% of variance.



(a) PCA for Original Feature

(b) PCA for Standardized Feature

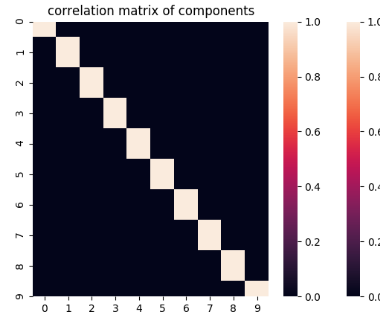
Figure 2: The relationship between the number of principle components and explained variance by PCA method

Figure 3 gives the relationship between different principle components. From (a) we can see that the different between original feature data and standardized feature data is insignificant. And the effect of principle component tend to cluster together to the same category while separate with the other category. The principle component are independent with each other by PCA method, which can be shown in (b).

3.2.2 Manifold Learning

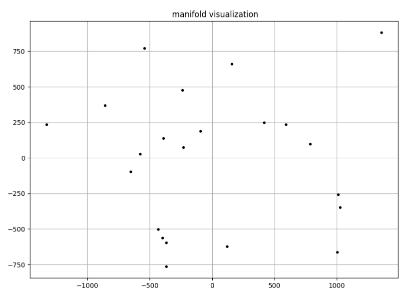
Because the feature data is intrinsically very high dimensional, it is hardly to be described linearly with just a few components. We implement nonlinear manifold embeddings like lsomap to give a two-dimensional projection of all the input, and output the image thumbnails at the locations of the projections. The result is given in Figure 4, which shows that the first two lsomap dimensions describe global image features. This gives us a nice visual indication of some of the fundamental features in data.

(a) 1st versus 2nd principle components

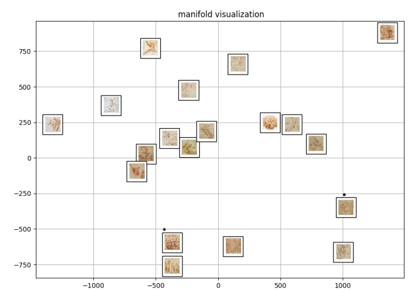


(b) Correlation Matrix for top ten components

Figure 3: The relationship between principle components. (a) shows the influence of the first principle and the second principle to different categories, that is, Raphael and non-Raphael; (b) shows the correlation between first ten principle components.



(a) without image



(b) with image

Figure 4: The visualization of manifold learning to features in first two dimensions.

3.3 Train

After feature extraction, Our dataset of images can be trained using a Logistic Regression model. We use flattened extracted features as an input to Logistic regression classifier, and train our model on all but two labeled samples, then test using the left-out two. By repeat the cross validation ten times, the mean of predict accuracy is 0.62.