COGS260 Assignment 2

Yan Sun
University of California, San Diego
La Jolla, California, USA
yas108@ucsd.edu

Abstract

Image classification is one of the most important research interests nowadays, for which MNIST dataset is a well-known data set that contains hand-written images and their labels for related study. In this assignment, several models and methods have been tried to classify the images in MNIST dataset, including Convolutional Neural Network, K-Nearest Neighbors, Support Vector Machine and Spatial Pyramid Matching. During the process of implementing these methods or models, different parameters are tried in order to optimize the classification performance. Finally, optimized parameters for these models are able to make the classification accuracy as high as 99%.

1. Method

In this assignment, several methods or models have been implemented to classify the images in MNIST dataset[9], including Convolutional Neural Network, K-Nearest Neighbors, Support Vector Machine and Spatial Pyramid Matching.

1.1. Convolutional Neural Network

Convolutional Neural Network (CNN)[7] is one type of feed-forward neural network that consists of input layers, hidden payers and output layers. In the hidden layers, there exist convolution layers, pooling layers and fully connected layers. The various combinations of these layers are able to create many types of CNN models, among which LeNet[10], VGG[14], AlexNet[7] and ResNet[3] are well-known models for image classification.

Figure 1 shows the detailed structure of LeNet-5 model, which contains one input layer, two convolution layers, two pooling layers, two fully connect layers and one output layer. This is the basic and classic LeNet-5 model. In order to get better results for different datasets, the hyperparameters in the model should be optimized.

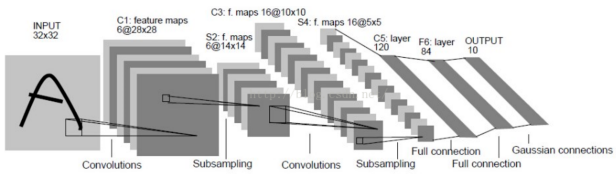

Figure 1. LeNet-5 Architecture[9]

In this assignment, various models with modified hyperparameters based on LeNet-5 have been explored in order to get higher classification prediction accuracy.

1.2. K-Nearest Neighbors

K-Nearest Neighbors (KNN) algorithm[18] is trying to find the K-closest neighbors for selected sample point in the feature space based on specific distance definition. Then the predicted label will be classified as the majority vote of these neighbors. There also exist parameters that are needed to be optimized such as the value of K and the definition of distance. Different combination of these parameters have been tried in the experiment of this part.

1.3. Support Vector Machine

Support Vector Machine (SVM)[15] is a supervised learning method that is able to find a hyperplane in the multiple dimension or high dimension space to classify the data points. If the data points can be represented in n dimension space, there could be many hyperplanes in (n-1) dimension that separate these points. One method to choose the best one is to select the hyperplane that represents the largest separation between two classes. In this situation, the classifier could be considered as maximum margin classifier. In this assignment, SVM is implemented to classify the image in MNIST dataset, during which the pixel of the image is utilized as input for SVM and the prediction will be compared with their true labels.

1.4. Spatial Pyramid Matching

Spatial Pyramid Matching[8] is one method based on Bag of Words method[19] and SIFT feature extraction algorithm[12] to build the appropriate feature for pictures for the further classification task. First, SIFT descriptor is used to extract the vectorized feature for different patches for pictures. Then K-means method[5] is utilized to find the corresponded group information for the patches. Then a codebook could be built so that the patch can be assigned to its group based on the codebook. Then the pyramid matching kernel combined with pyramid pooling process is utilized to build the ultimate feature for the final classification by SVM method.

2. Experiment

2.1. Convolutional Neural Network

In the experiment for CNN, two main aspects optimization are explored for LeNet-5 model. As for the following exploration process, the pixel information of image is utilized as input to be fed into the model. The total training step is set up to be 20000 maximum. Compared to the original model, the dropout process is also added to avoid overfitting.

2.1.1 Neural Network Architecture Exploration

The first type of exploration is about CNN architecture hyperparameter. The number of channels for the convolution layers and pooling layers is one factor that has influence for the prediction accuracy since more channels generally tend to have more ability to learn more complex subliminal relation in the picture.

Index	lr	dropout	1st Channels	2nd Channels	fc	Test Loss	Test Acc
1	1e-4	0.5	6	16	1024	0.026	0.9908
2	1e-4	0.5	16	32	1024	0.027	0.9918
3	1e-4	0.5	32	64	1024	0.026	0.9916

Table 1. Results for differect Architectures

The Table 1 shows the result of the exploration for different architectures modification based on LeNet-5 and corresponded prediction result. In Table 1, lr means the learning rate, 1st Channels and 2nd Channels are the number of channels for corresponded convolution layers and pooling layers, fc is the number of neurons for the only fully connected layer.

Compared to original LeNet-5 neural network, several modification have been added to the CNN model. First, the number of channels for convolution layers and pooling layers are changed. Second, there is only one fully connected layer. Finally, the dropout process is added for the fully connected layer. Figure 2 shows the optimum neural network structure obtained in this part of experiment ultimately.

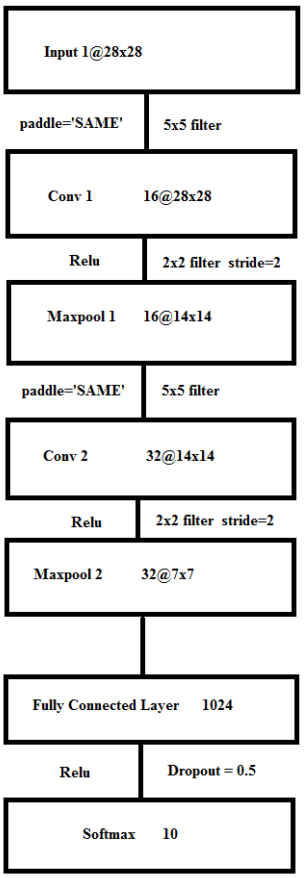

Figure 2. Final Optimized Neural Network Architecture

2.1.2 Learning Parameter Exploration

In this part of exploration, different combinations of optimizer and learning rate have been tried and compared. Based on Table 1, the best configuration is the architecture with index 2 so that this architecture is utilized for next part experiment.

Table 2 shows the results of different learning parameters configuration, among which the experiment with index 1 has the best result. Figure 3 and 4 are the related curve plot.

2.2. K-Nearest Neighbors

In this part of experiment, different combinations of distance types and values of K have been tried to find the optimum configuration for K-Nearest Neighbors methods to get higher accuracy for MNIST hand-written images classification. The detailed result data is in Table 3 and Figure 5 for this best configuration. During the selection process, 5000

Index	Optimizer	learning rate	Test Loss	Test Acc
1	Adam	1e-4	0.026	0.9916
2	Adam	1e-5	0.054	0.9818
3	Adam	1e-6	0.230	0.9323
4	SGD	1e-4	0.038	0.9863
5	SGD	1e-5	0.112	0.9652
6	SGD	1e-6	0.404	0.9002
7	RMSprop	1e-4	0.032	0.9905
8	RMSprop	1e-5	0.066	0.9777
9	RMSprop	1e-6	0.233	0.9315

Table 2. Model Performance with different learning configuration

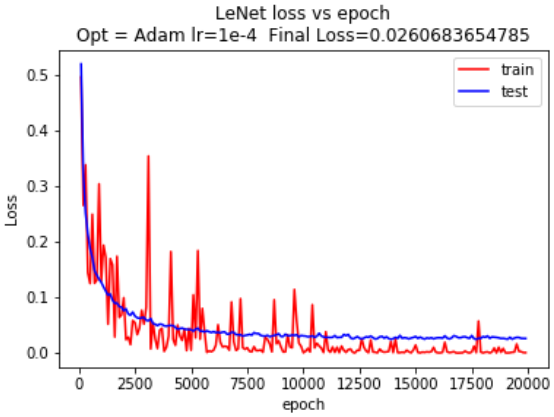

Figure 3. Loss vs Epoch Curve

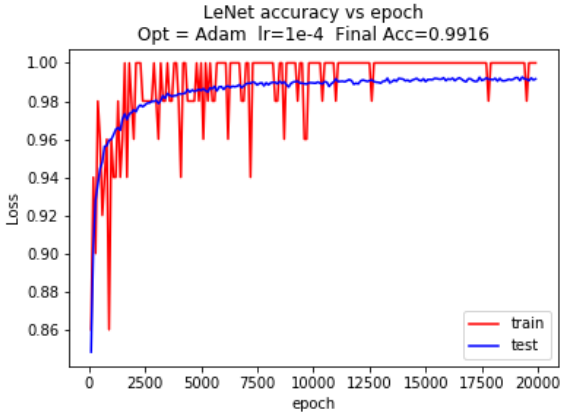

Figure 4. Accuracy vs Epoch Curve

data samples selected from training set are considered as validation set for the model selection. The classification result for first 1000 samples of test data is shown in confusion matrix (Figure 5).

Index	distance	K	valid accuracy	test accuracy
1	eluclidean	1	0.97	0.956
2	eluclidean	2	0.963	0.954
3	eluclidean	3	0.965	0.959
4	manhattan	1	0.962	0.940
5	manhattan	2	0.963	0.944
6	manhattan	3	0.967	0.953

Table 3. KNN model selection

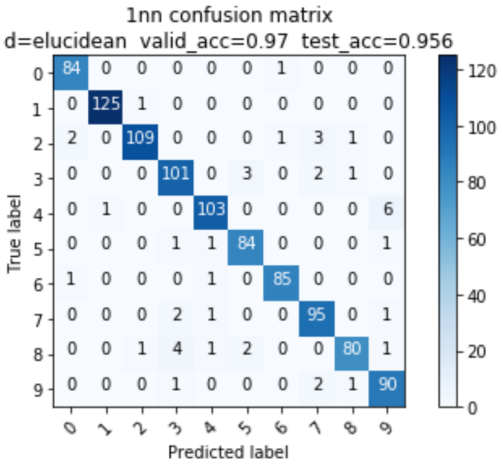

Figure 5. 1nn Confusion Matrix

2.3. Support Vector Machine

Different combinations of distance types and values of K have been tried to find the optimum configuration for K-Nearest Neighbors methods to get higher accuracy for MNIST hand-written images classification. Table 4 shows all the results obtained for SVM models.

During the selection process, 5000 data samples selected from training set are considered as validation set for the model selection. The classification result for first 1000 samples of test data is shown in confusion matrix. Specifically, the model with index 1,2,3,4,6 is implemented with manually selected parameters and the model with index 5 is selected by 3-fold cross validation. The model 5 has the best performance, which has 97.9% classification prediction accuracy. The best model's prediction confusion matrix is shown in Figure 6.

2.4. Spatial Pyramid Matching

As for Spatial Pyramid Matching (SPM) method, the parameters needed to be optimized is the number of clusters during the grouping process for featured patches and the level of spatial pyramid encoding procedure. As the explanation for SPM method in Section 1.4, there exist parameters in this method needed to be optimized: the number of clusters in K-Means procedure and the level of pyramid.

Index	C	gamma	Kernel	Valid Accuracy	Test Accuracy
1	1	1.276e-3 (1/number of features)	Linear	0.928	0.925
2	10	1.276e-3 (1/number of features)	Linear	0.913	0.914
3	0.1	1.276e-3 (1/number of features)	rbf	0.899	0.874
4	1	1.276e-3 (1/number of features)	rbf	0.928	0.925
5	2	0.0296	rbf	0.979	0.979
6	10	1.276e-3 (1/number of features)	rbf	0.951	0.95

Table 4. Result of SVM models

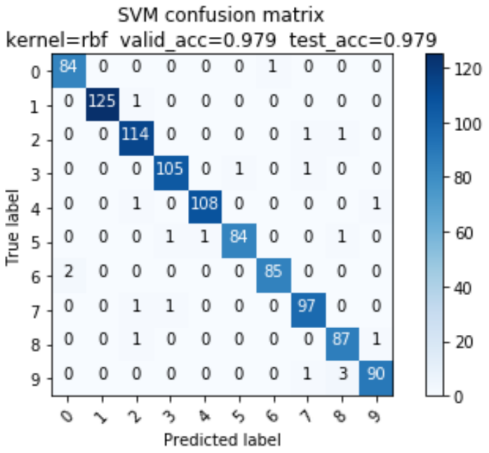

Figure 6. SVM Confusion Matrix

The code for this part of experiment is based on one publicized project[2]. During the selection process, 5000 data samples selected from training set are considered as validation set for the model selection. The classification result for first 1000 samples of test data is shown in confusion matrix. Due to the computing resources and computing time limit, only 2 models are tried for the SPM model selection. The result data is in Table 5 and the confusion matrix for model 2 is in Figure 7.

Index	VOC_SIZE	DSIFT_STEP_SIZE	C	gamma	Valid Accuracy	Valid Accuracy
1	100	4	100	10	0.916	0.846
2	100	4	10	1	0.983	0.908

Table 5. Result of SPM Model Selection

3. Discussion

3.1. Convolutional Neural Network

During the exploration of CNN on MNIST dataset, several aspects of experiment have been tried and the comparably better configuration has been found.

First, traditional LeNet-5 model has two fully connected layers while in this assignment the modified LeNet model with only one fully connected layer has been tried and the

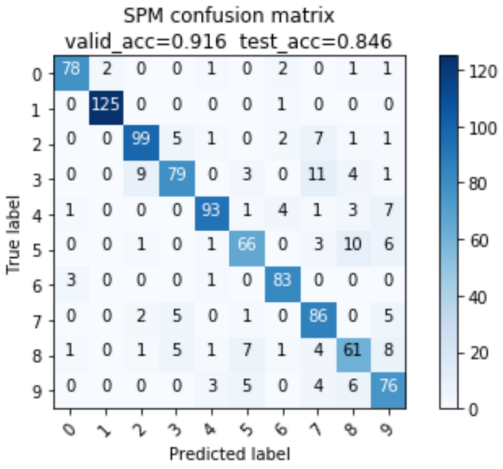

Figure 7. SPM Confusion Matrix

prediction accuracy can reach as high as 99%. This indicates that only one fully connected layer with 1024 neurons is already enough to finish the remained process work after the convolution and pooling layers. However, compared to the original LeNet-5 model, the number of weights between second pooling layer and fully connected layer is much more in this experiment but the ultimate result is good and the computing process is not slow.

Second, the number of channels for the model's convolution layer and pooling layer have been changed. The number of channels for convolution and pooling layers has influence on the model performance, which affects the model's ability to recognize the pattern of pictures. The number could not be too small or too large. When it is too small, the classification ability is weak. When the number is too large, the computing time needed will grow. After the experiment, when first convolution layer and pooling layer have 16 channels, second convolution and pooling layer have 32 channels, the model has best performance among all the experiments results in consideration of prediction accuracy and computation efficiency.

Third, as for the selection optimizer and learning rate, the results show that the Adam optimizer [6] combined with 10^{-4} learning rate has the best performance, which could make test data prediction accuracy as high as 99.16% and the value of test data loss as low as 0.026. Apart from the comparison of learning rate, the comparison of optimizer shows that Adam is the best selection for MNIST dataset classification. Adam optimizer computes adaptive learning rate for each parameter, it stores an exponentially decaying average of past squared gradients similar to RMSprop [16]. Furthermore, Adam optimizer also has an exponentially decaying average component for past gradients, which is similar to momentum that can help training process to go across local minimum area.

In summary, the CNN model has wonderful performance (99% accuracy) for MNIST hand-written digit dataset image classification, very close to human mind level.

3.2. K-Nearest Neighbors

As for KNN experiment, the combination of two types of distance and K = 1,2,3 is tried for looking for better performance of KNN model. The result of the experiment shows that when the distance is defined by euclidean distance and K=1 the model performs best and the corresponded validation set accuracy and test set accuracy is 97% and 95.6%, respectively.

Euclidean distance is the L2 norm of the distance between two data points in their feature space while Manhattan is the sum of absolute value of difference in each dimension. The difference in their definition make the KNN method performs different results. Although this part of result shows euclidean distance gives better result, there is argument that states if the dimension becomes higher the Manhattan distance will behave better[1]. With the increase of dimension, a curious phenomenon will appear: the distance for the nearest and farthest points will get closer. In this way, the points dramatically become uniformly distant from each other. This type of circumstance can be observed for a large variety of distance metrics, which is more obvious for Euclidean distance metric than Manhattan distance metric. In this way, the distance will not work well for classification so that Euclidean distance metric does not perform as good as Manhattan distance metric.

Specifically in this assignment, the Euclidean distance still performs better, which is possibly that the dimension is not high enough or more optimized combinations for values of K and distance type remained to be founded.

3.3. Support Vector Machine

There exist many parameters that could be optimized for SVM classifier for MNIST dataset. Specifically, the value of kernel coefficient gamma, penalty parameter C of the error term have been explored to get higher classification prediction accuracy[17]. In Table 4, it can be concluded that rbf kernel SVM with C=2 and gamma=0.0296 has the best result, whose accuracy can reach as high as 97.9%.

The Radial Based Function kernel (RBF) has obtained much success in several research projects[20][4][11]. It has the ability to project the feature into high dimension space, which means the linear kernel will be one special case for RBF kernel. In addition, compared to polynomial kernel, it has fewer parameters to be optimized so it has more widely application.

3.4. Spatial Pyramid Matching

For Spatial Pyramid Matching method, the most significant parameter needed to be optimized are the level of pyramid for the feature generation step. Section 5 shows that SPM model with level 2 is better than model with level 1. In this way, it can be concluded that the pyramid pooling procedure does assist in the feature extraction for the images in MNIST dataset, which is similar to the pooling layer in CNN models.

First, the pooling procedure is able to assist in dimensional reduction since the image data always contains a large number of pixel information while pooling could help reduce the computing time. Second, pooling procedure is able to extract rotational and position invariant feature, which could be considered as one method for subsampling[13].

Due to the limitation of computing resources and computing time, models with higher value of levels configuration are not tried at the end of experiment, which deserves future exploration for better performance for MNIST classification based on SPM method.

4. Conclusion

There exist many methods for image classification, among which optimized Convolutional Neural Network model has the best classification prediction performance (99% Accuracy), which is very close to human recognition level. Furthermore, more exploration needed to find better configurations for these models to get better classification results.

References

[1] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pages 420–434. Springer, 2001. 5

[2] CyrusChiu. Cyruschiu/image-recognition. 4

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009. 1

[4] T. Handhayani, J. Hendryli, and L. Hiryanto. Comparison of shallow and deep learning models for classification of lasem batik patterns. In *Informatics and Computational Sciences (ICICoS), 2017 1st International Conference on*, pages 11–16. IEEE, 2017. 5

[5] J. A. Hartigan and M. A. Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979. 2

[6] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[8] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, volume 2, pages 2169–2178. IEEE, 2006. 2

[9] Y. LeCun, C. Cortes, and C. Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010. 1

[10] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, et al. Learning algorithms for classification: A comparison on handwritten digit recognition. *Neural networks: the statistical mechanics perspective*, 261:276, 1995. 1

[11] Y. Liu and K. K. Parhi. Computing rbf kernel for svm classification using stochastic logic. In *Signal Processing Systems (SiPS), 2016 IEEE International Workshop on*, pages 327–332. IEEE, 2016. 5

[12] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 2

[13] D. Scherer, A. Müller, and S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010. 5

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[15] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. 1

[16] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012. 4

[17] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001. 5

[18] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006. 1

[19] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007. 2

[20] B. Yekkehkhany, A. Safari, S. Homayouni, and M. Hasanlou. A comparison study of different kernel functions for svm-based classification of multi-temporal polarimetry sar data. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(2):281, 2014. 5