ECSE 551 Fall 2020 Mini-project 3 Report

Fei Peng
260712440
fei.peng@mail.mcgill.ca

Yukai Zhang
260710915
yukai.zhang@mail.mcgill.ca

Yi Zhu
260716006
yi.zhu6@mail.mcgill.ca

Abstract

Image identification and classification are two of the most popular research areas in machine learning as well as computer vision. In this mini-project, we chose to investigate and evaluate the performance of different Convolutional Neural Network (i.e., CNN) models on a large image dataset. To have a better understanding of the structure of CNN, we implemented a model from scratch and analyzed its behavior. In addition, we tested different CNN layer arrangements and examined the effect of hyperparameters (e.g., batch size, number of epochs, and learning rate) before the optimum combination was selected, which yields the highest accuracy in the validation set. The performance of other CNN implementations provided by PyTorch and TensorFlow were also investigated, from which we found that VGG has the highest accuracy without any pre-trained data. According to our testing result, correctly choosing a proper CNN layer arrangement can greatly boost the result accuracy, avoid allocating too much memory as well as make sure the training process converges after reasonable iterations. Details will be presented in the following parts of this report.

1 Introduction

The increasing popularity of computer vision and the advance of machine learning technique have brought the opportunities to perform apparel classification [1]. The main objective of this mini-project is to apply machine learning methods on classification of clothes from five different categories where each has a distinctive price from $1 to $5. There are 60,000 monochrome images provided in the training dataset for this mini-project, and each image consists of three clothes. Therefore, their total price lies between $5 and $13 which has 9 different values. After applying machine learning methods on the training dataset, an additional 10,000 images are used to test the accuracy of the classification techniques and the results are submitted on Kaggle competition.

In this project, we are allowed to freely choose any machine learning methods, so as to explore their performance. However, the three clothes could appear in any random place in each image, and most of the classification methods we have been using so far in this course are sensitive to this position shift. Thus, a fully connected layer analysis (e.g., fully connected neural networks) is preferred as it can be used to achieve shift invariance, but performing this method requires a large number of training instances (i.e., weight parameters) as well as a long learning period, whereas we have limited computational resources and we are forbidden to import any pre-trained data [2]. Convolutional Neural Network, on the other hand, conserves the benefit of shift invariant by applying the same weight configuration across the image space. A convolution neural network consists of one or multiple convolutional layers, which are followed by fully connected layers. This network arrangement allows CNN to process/analyze an image in a similar manner as the biological neural systems.

Since the CNN method is wildly applied to solve many image-recognition problems, different types of network designs are available. Choosing the one that best fits this mini-project is essential. Thus, we investigated the performance of the CNN we implemented ourselves, and ResNet18 [3], AlexNet [4], VGG16 as well as VGG19 (both versions provided by PyTorch and TensorFlow) [5].

After comparing the validation accuracy of each implementation of CNN, VGG19 (by TensorFlow) was found to show a promising result. It yielded an accuracy of 99% in the validation dataset and 97% in the Kaggle competition.

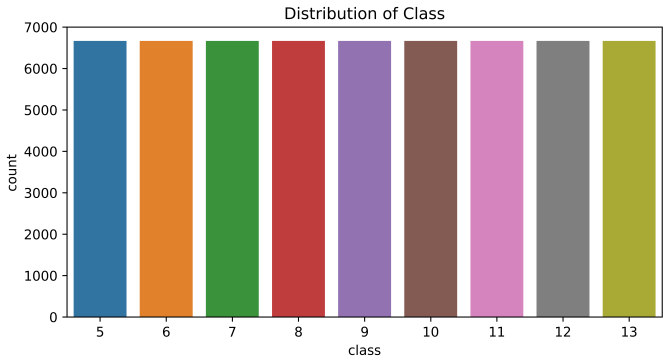

Figure 1: Class distribution of train data

2 Datasets

Before performing any investigation on the supervised machine learning and image classification models, it is important that the characteristics of the dataset (i.e., both features and classes) are properly analyzed and understood. Aside from that, data preprocessing lays the groundwork for the following data analysis, as its product is the final training set [6].

The training dataset we used is in $Train.pkl$ which contains 60,000 grayscale image samples, and their corresponding classes are imported from $TrainLabels.csv$. The test dataset is 10,000 grayscale image samples contained in $Test.pkl$, from which we predicted the classes and submitted the result to Kaggle competition.

In this project, the following data analyzation and preprocessing had been conducted.

2.1 Feature Analysis

The dataset is a modified version of the Fashion-MNIST dataset, which consists of samples with 641281 pixel values as features representing grayscale images. Therefore, the total number of features in each sample is 8192. As the number of samples in this dataset is numerous (i.e., 60,000 training samples), we considered using deep learning to achieve optimum accuracy with large dataset [7], from which we chose Convolutional Neural Networks (i.e., CNN) instead of fully connected neural networks for fewer memory requirements and lower latency [2]. Moreover, based on our previous experience, the images are relatively large and complex, therefore we anticipated to use deeper neural networks (i.e., more layers of neurons) for better accuracy. Details will be discussed later in Section 3.

2.2 Target Analysis

In this dataset, each image contains three articles, and our goal is to predict the total price of all articles presented in the image. The prices (i.e., classes) range from 5 to 13, therefore, it is a nine-way classification problem. The distribution of each class is shown in Figure 1. The uncertainty in prediction can be quantified by the entropy of the dataset with Equation 1, and the calculated data entropy is 3.17 which stands for an even distribution. Therefore, the number of samples in each class is equal. Moreover, before feeding the dataset to any training process, we need to make sure that the class values are encoded. Since the prices are integer numbers range from 5 to 13, we could simply encode them by subtracting the values by 5, thus the obtained target values would range from 0 to 8.

$$H(X) = -\sum_{x \epsilon X} p(x) log_2 p(x) \tag{1}$$

2.3 Normalization

The original pixel value of the images in the dataset is between 0 and 255, to centralize the value distribution while keeping its precision, we normalize the data with mean and standard division both equal to 0.5. Therefore, after normalization, the pixel values range from -1 to 1, which removes any scale factors that might influence the classification result [8].

3 Proposed Approach

In this section, we are going to discuss the models we implemented and the methods we used for this project, as well as the model and method selection in terms of accuracy, robustness/consistency, and speed. However, due to time constraint of this project, the reported accuracy and time might not be based on the absolute optimum hyperparameters, but rather try and error.

3.1 Training/Validation Splits

As mentioned previously in Section 2, the training dataset contains 60,000 samples, while the testing dataset contains 10,000 samples, however, only the training dataset contains the corresponding target classes. Therefore, for model selection, we chose to split the training dataset to 50,000 training samples and 10,000 testing samples, in order to perform validation. In this way, we could ensure consistency between the validation and testing process, and increase the possibility to reach higher accuracy on predicting the testing dataset. Due to the fact that deep learning methods are relatively more time consuming than normal machine learning algorithms, k-fold validation technique is not as common, thus, we compared the accuracy of different models based only on those 10,000 testing samples that are separated from the training dataset.

3.2 Hyperparameters

It is common practice to decay the learning rate in order to achieve better accuracy. However, one can usually obtain the same learning curve on both training and test sets by instead increasing the batch size during training [9]. As will be explained detaily later in Section 4, a smaller learning rate leads to a longer training time to approach the same accuracy. In CNN, it means that more epochs are required for training the network to preserve the same accuracy. Since the time consumed in each epoch is relatively constant, there is a linear relationship between number of epochs and total training time. However, increasing the learning rate reduces the number of parameter updates, thus, leads to greater parallelism and shorter training times. To compensate the influence of increasing the learning rate (e.g., by a factor of ϵ), we could increase the batch size by a factor (e.g., B) that is linearly proportional to the increase of learning rate (i.e., $B \propto \epsilon$) [9]. However, increasing the batch size alone could not ensure the same accuracy [10], whereas in our case decrease the accuracy as will be explained later in Section 4. Therefore, by tuning the learning rate and batch size together, we could select the most suitable combination to achieve higher accuracy and shorter training time. Moreover, we could increase the momentum coefficient (i.e., β) in gradient descent and scale $B \propto \frac{1}{1-\beta}$, while maintaining the same accuracy [9].

On the other hand, although increasing the number of epochs leads to longer training time, it also helps with achieving higher accuracy. In real life, it is usually a trade-off between choosing shorter training time and higher accuracy. In our case, since this is a competition on test accuracy, we would choose higher accuracy by sacrificing training time. However, accuracy will eventually converge, and continuing to train the model for more epochs will cause overtraining.

3.3 Algorithm Selection

At the beginning of this project, we started with building a Convolutional Neural Network (i.e., CNN) from scratch, in order to explore the structure and characteristics of the network. At the beginning, we configured a rather simple network, which starts with only four 2D convolution layers, each followed by a ReLU activation function as well as a 2D max pooling, and ends with two fully connected layers. However, we underestimated the requirement of the feature size (i.e., number of pixels in each image) on the depth of the network. The accuracy did not advance with the increasing number of epochs, but remained no difference from random classification, and could easily end up with data underflow/overflow, no matter how we changed the gradient descent and image normalization parameters. Therefore, we decided to increase the depth of our CNN model by learning and imitating one of the existing successful models provided by PyTorch [11], which is ResNet18 (i.e., residual network) [3], and tweaked the structure to suit our use. As a result, our CNN model contains twenty 2D convolution layers, followed by batch normalisation and ReLU activation functions, as well as one fully connected layer at the end. With this model, the running loss (i.e., cross entropy loss) started to converge as the number of epochs increased, and the accuracy began to rise.

Next, we continued to investigate the behavior of CNN models by performing experiments on ResNet18 [3], AlexNet [4], VGG16, VGG19 as well as VGG19 with batch normalization (i.e., VGG-19 BN) [5] provided by PyTorch [11], as well as VGG19 from TensorFlow (i.e., VGG-19 TF) [12], with minor adjustment on the models to fit our situation (i.e., input color channel equals one, and number of output class equals nine). AlexNet is one of the first Deep Convolutional Networks to achieve considerable accuracy, which consists of five convolutional layers and three fully connected layers, with ReLU as the activation function. VGG16 and VGG19 are both VGGNet, which differ only in the total number of layers in the network (i.e., 16 and 19 layers respectively). The idea of VGG is to use fixed size kernels so that all the variable size convolutional kernels used in AlexNet can be replicated by making use of multiple 3x3 kernels as building blocks. As a result, VGG reduces the number of parameters and the training time. ResNet18 are 18-layer residual nets, which have shortcut connection added to each pair of 3x3 filters compared to plain nets that are inspired by VGGNet. Thus, ResNet is able to find simpler mappings when they exist. Later in the next section (i.e., Section 4), we are going to discuss in detail the experiment results.

4 Results

When performing machine learning with CNN, multiple epochs are needed in order to obtain the highest accuracy. After evaluating the time and accuracy with respect to the number of epochs for the CNN implemented by ourselves, we found that the time required for each epoch is approximately constant and the total time is linearly proportional to the number of epochs. Accuracy, on the other hand, has an abrupt increase at the first several epochs. The slope of the curve gradually reduces as the accuracy approaches 1.0. Details are shown in Figure 2. After training for 32 epochs, our model achieved an accuracy of 84% and the training time is 995.74 seconds, which are shown in Table 1.

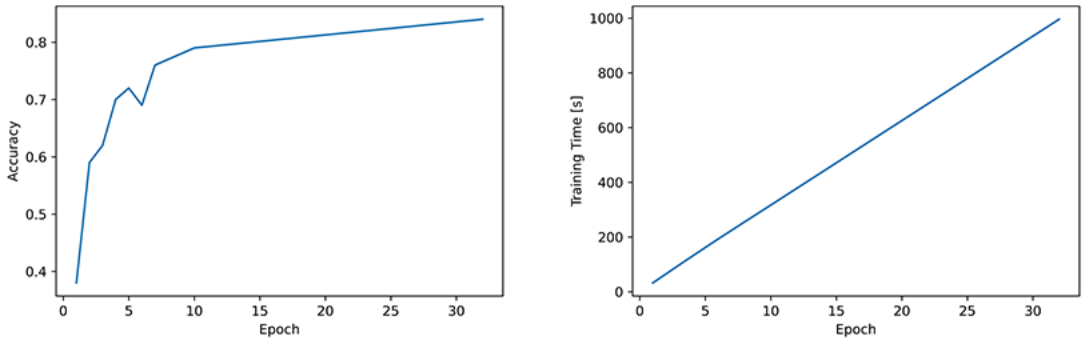

Figure 2: Accuracy (left) and training time (right) with respect to number of epochs

Table 1: Accuracy and training time for seven CNN models (32 epochs)

	CNN	ResNet-18	AlexNet	VGG-16	VGG-19	VGG-19 BN	VGG-19 TF
Accuracy	84%	87%	84%	94%	98%	98%	99%
Time (s)	995.74	1598.56	1275.41	5399.81	6151.02	6758.24	6085.63

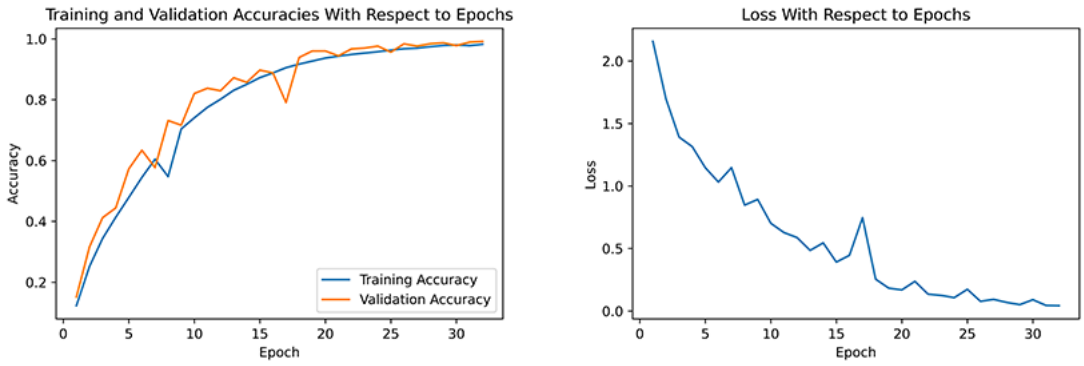

Figure 3: Training/validation accuracy (left) and loss (right) with respect to number of epochs

We then conducted experiments with other CNN models. The performances are recorded in Table 1 where 32 epochs of training are executed for each model. Despite having a slightly lower accuracy (84%), the CNN implemented by ourselves has demonstrated an incredible efficiency which takes only 15% of the training time compared to more complex models such as VGG. The VGG19 provided by TensorFlow, on the other hand, can yield the highest accuracy (99%) with a proper combination of hyperparameters. In our case, we chose SGD as our gradient descent method with a learning rate of 0.001 and a momentum of 0.7. We also set the loss function to $sparse\_categorical\_crossentropy$ and the metrics we use to evaluate the model is $accuracy$. We disabled the usage of pre-train data by setting the weight of VGG19 to be $None$ and we specified a batch size of 16. Its result was used in Kaggle competition and achieved a leaderboard accuracy of 0.98366.

To have a better understanding of this model (i.e., VGG-19 TF), we did further investigation. We picked 10,000 of the 60,000 training dataset as the validation set and the relationship between training/validation accuracies as well as loss with respect to epochs are shown in Figure 3. It is noticeable that the curve of validation accuracy keeps rising in the last few epochs and we can infer that overtraining has not occurred within the first 32 epoches. The loss decreases and approaches 0 as the number of epochs increases.

The accuracies and convergence time depend on not only the training models, but also the hyperparameters. The effects of the learning rate/momentum as well as batches size are investigated. We used VGG19 and VGG19 with batch normalization to test the effect of learning rate and momentum. Their accuracies after 32 epochs are shown in Table 2 and their loss curves are plotted in Figure 4. It appears that with a higher learning rate, the loss will have a more rapid drop in the first several epochs. The loss difference between different learning rates will decrease as the number of epochs gets larger.

Table 2: Accuracy performance of two CNN models under different hyperparameter sets (32 epochs)

	VGG-19	VGG-19 BN
Learning Rate = 0.001 & Momentum = 0	56%	68%
Learning Rate = 0.005 & Momentum = 0.5	98%	98%

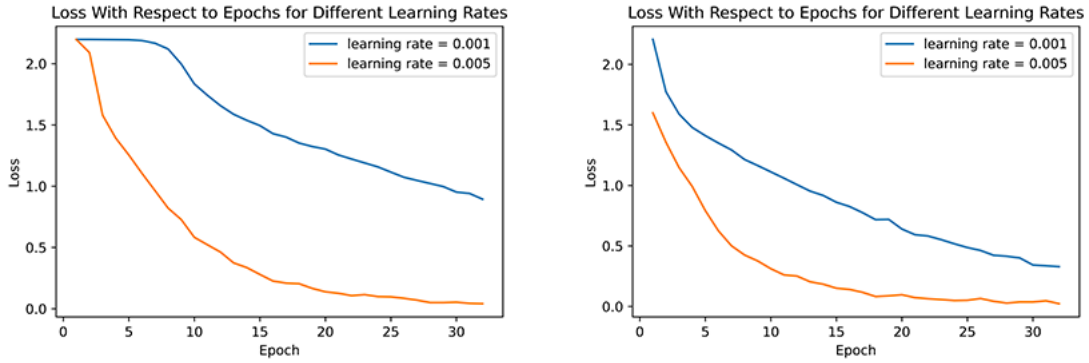

Figure 4: Loss with respect to epochs in different learning rates in VGG19 (left) and VGG19 BN (right)

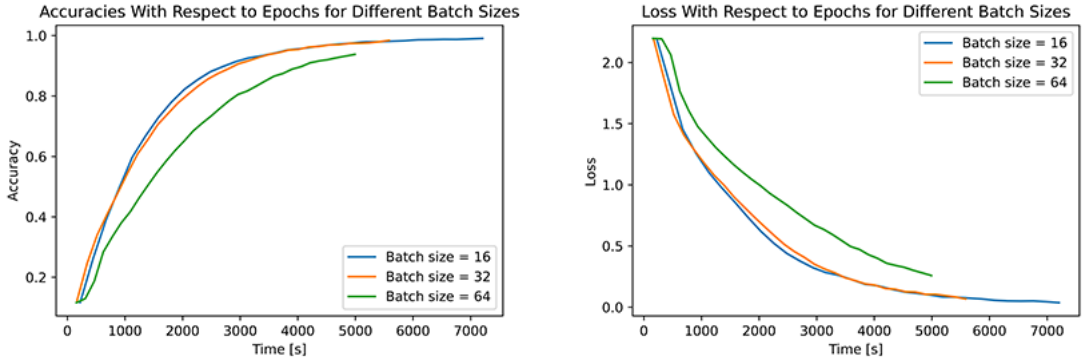

Figure 5: Accuracy (left) and loss (right) with respect to number of epochs for different batch sizes

5 Discussion and Conclusion

In conclusion, the Convolutional Neural Network model we implemented from scratch in this project performed as expected for classifying the clothes images. It achieved a validation accuracy of 84% after training for 32 epochs and the training process takes 995.74 seconds. However, the CNN models provided by TensorFlow yields the highest validation accuracy (99%) without using any pretrainined data. The validation accuracy is found to rise at a decreasing rate as the number of epochs grows. There does not exist a sign of overtraining within the first 32 epochs when the learning rate is set to be no larger than 0.005.

The future investigation could be to improve the layer arrangement of our self-implemented CNN and to find the best combination of hyperparameters in order to achieve a higher accuracy with less training time/epochs.

On the other hand, we achieved a leaderboard accuracy of 0.98366 using the VGG19 provided by TensorFlow without pre-trained data. 64 epochs of training were performed with a batch size of 16 in order to achieve this result as the validation accuracy barely changes (i.e., converges) at this point. In the future, we will take a closer look at the hyperparameters of the VGG19 algorithm to find the best combination that suits this machine learning problem.

6 Statement of Contributions

Yukai Zhang was in charge of researching the structure of the Convolutional Neural Network and implementing a CNN from scratch. Fei Peng browsed the models in PyTorch and TensorFlow, modified and trained the selected base models, as well as investigated the relationship between accuracy and number of epochs. Yi Zhu was responsible for analyzing the correlation between time and number of epochs, as well as proofreading the report.

References

[1] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel Classification with Style," Berlin, Heidelberg, 2013: Springer Berlin Heidelberg, in Computer Vision – ACCV 2012, pp. 321-335.

[2] S. Hijazi, R. Kumar, and C. Rowen, "Using Convolutional Neural Networks for Image Recognition By," 2015.

[3] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[4] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," *ArXiv,* vol. abs/1404.5997, 2014.

[5] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *CoRR,* vol. abs/1409.1556, 2015.

[6] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Data Preprocessing for Supervised Leaning," *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 1, pp. 4104-4109, 2007.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature,* vol. 521, no. 7553, pp. 436-444, 2015/05/01 2015, doi: 10.1038/nature14539.

[8] G. Finlayson, B. Schiele, and J. Crowley, "Comprehensive Colour Image Normalization," in *ECCV,* 1998.

[9] S. L. Smith, P. Kindermans, and Q. V. Le, "Don't Decay the Learning Rate, Increase the Batch Size," *ArXiv,* vol. abs/1711.00489, 2018.

[10] Y. You, I. Gitman, and B. Ginsburg, "Scaling SGD Batch Size to 32K for ImageNet Training," *ArXiv,* vol. abs/1708.03888, 2017.

[11] A. Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library," *ArXiv,* vol. abs/1912.01703, 2019.

[12] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *ArXiv,* vol. abs/1603.04467, 2016.

Appendix

(Please see following pages attached for code implementations)