Reproducibility Report for ML Reproducibility Challenge 2020

Group 80: Pingsheng Li Jacob Wang Xiyuan Feng
McGill University
Montreal, QC H3A 0G4
pingsheng.li@mail.mcgill.ca jingyu.wang@mail.mcgill.ca xiyuan.feng@mail.mcgill.ca

Reproducibility Summary

Scope of Reproducibility

This report aims to verify the conclusion and reproduce a subset of results in the paper *Deep Learning using Linear Support Vector Machines* (Yichuan Tang, 2015) using two datasets (MNIST and CIFAR-10), both of which are also utilized in the original paper. The main claim of this paper is that the objective function of linear L2-SVMs (squared hinge loss) has advantages over cross-entropy loss with softmax activation function in multi-class classification tasks when combined with feed-forward neural network models (Multilayer Perceptrons and Convolutional Neural Networks). Besides, we also performed several ablation studies on the models implemented in the original paper to gain deeper insights about the functionality of each component in the model, and how much they contribute to the performance of the models.

Methodology

Since the author did not provide the code for experiments, in this report, we re-implemented the models mentioned in the original paper using tensorflow package in Python, and the experiments were concurrently conducted on several online machine learning laboratory platforms, specifically Google Colab and Kaggle, with GPU support. Also, we performed multiple ablation studies on the models implemented in the original paper by removing certain components in the model to examine their functions in the model and how significant the effects can be.

Results

Due to the insufficient details provided in the original paper, and since there is no available code offered by the author, some procedures, like data preprocessing & augmentation, may not be the same with the one in the original paper. Therefore, our results are mixed. In MNIST dataset, we reproduce the accuracy to within 1.5% of the reported values (99.01%-99.13%) in the original paper, but it does not support the conclusion that loss function for SVMs outperforms cross-entropy loss with softmax activation. In CIFAR-10 dataset, although we did not achieve the same level of accuracy as the author did (86%-88%), we obtained some evidence that supports the conclusion. Besides, we conducted additional ablation studies, and these studies suggest that in some cases, squared hinge loss function indeed is superior to cross-entropy loss with softmax activation in multi-class classification, supporting the conclusion.

What was easy

The easy part in this reproduction is the re-implementation of the models using tensorflow package in Python since all implemented features are supported by the library.

What was difficult

In this reproduction, the difficult part is that the details in the paper is not comprehensive and sometimes ambiguous. For example, the author did not specify how to normalized the data before PCA, the parameters for stochastic gradient descent with momentum are not given, etc.

Communication with original authors

Our team did not have any form of communication with the original author.

Preprint. Under review.

1 Abstract and introduction

This work attempts to reproduce a subset of results reported in the paper *Deep Learning using Linear Support Vector Machines* (Yichuan Tang, 2015), which claims that the loss function for L2-SVM (squared hinge loss) is superior to cross-entropy loss with softmax activation function when using neural network in multi-class classification problems. To verify this conclusion, we repeated some experiments mentioned in the original paper as precise as possible, with possible minor modifications in certain procedures that are not detailed in the original paper. Also, we performed several ablation studies on the implemented models to examine their functionality and the level of significance of their effects on the models. To summarize, in the repeated experiments, the results we obtained are mixed, some of them are contradictory while others are supportive; but in the additional experiments, our results suggest that in higher dimensional data, squared hinge loss function may have some advantages over the cross-entropy loss function with softmax activation, which gives limited, but supportive evidence of the author's conclusion.

2 Dataset

In this section, we provide the context of our dataset, and the procedures for data preprocessing & augmentation.

2.1 Data Source

Both MNIST and CIFAR-10 are imported via tensorflow package.

2.2 Pre-processing & Augmentation

Although it is not mentioned by the author, normalization was applied to every image in the MNIST dataset, since later it will be subject to Principal Component Analysis (PCA), which requires the data has a mean of 0. Also, each entry of images in CIFAR-10 dataset has been divided by 255, which scales all values into the [0, 1] interval. This operation effectively prevent many numerical problems during the training. Besides, despite the fact that the author applied horizontal reflection and jitters to augmented CIFAR-10 training dataset, the author did not include the details of this augmentation (how jittery the augmented images should be), so we decided not to include this procedure in our reproduction.

2.3 Information

In MNIST dataset, there are 70,000 28x28 images of handwritten digits from 0 to 9, 60,000 of which are for training and the rest is for testing. In CIFAR-10 dataset, 60,000 28x28 images of 10 different objects are partitioned into 50,000 images for training and 10,000 images for testing. Every image is represented by 2-D array with entries filled by integer from 0 to 255 that indicates the brightness of a pixel. Classes in both datasets are relatively evenly distributed.

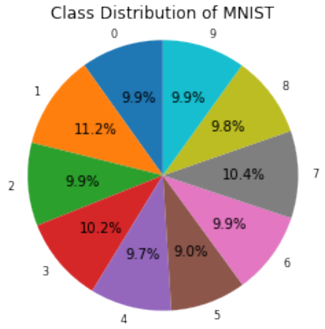


Figure 1: The class distribution of MNIST dataset.

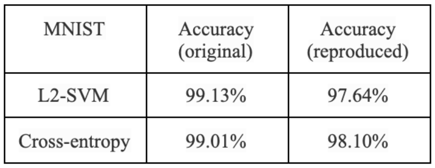
3 Results

In summary, our results that reproduce the original paper are mixed: experiments conducted on MNIST dataset did not support the conclusion of the author, but it was also not strong enough to dispute the conclusion since the difference is not statistically significant; however, for the experiments reproduced on CIFAR-10 dataset, they consolidated the conclusion in the original paper despite the accuracy achieved in the experiments is not at the same level as the author claimed in the original paper. Besides, we discovered that the conclusion in the original paper is supported by the results of our additional experiments, which suggests that in high dimensional data, L2-SVM loss function is more likely to give a better performance cross-entropy loss with softmax activation function. Therefore, in general, we consider our experiments offer more supportive evidence for the conclusion of the original paper.

3.1 Results reproducing original paper

3.1.1 Result on MNIST dataset

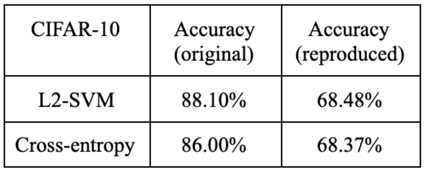
In this experiment, we attempted to follow the exact description provided in the original paper, performing PCA, applying regularization and Gaussian noise (Raiko et al., 2012; Rifai et al., 2011b), etc. However, since some details are not provided in the description, we had to made up the missing part ourselves. The momentum of stochastic gradient descent has been set to 0.9, and the distribution of data was normalized to a standard normal distribution before the application of PCA. The accuracy achieved is very close (within 1.5%) to that of the original paper, but L2-SVM loss function did not outperform cross-entropy with softmax activation as expected. So this experiment did not support the conclusion of the original paper, but it also did not substantially dispute the conclusion because the difference is too small to make a difference.



MNIST	Accuracy (original)	Accuracy (reproduced)
L2-SVM	99.13%	97.64%
Cross-entropy	99.01%	98.10%

Figure 2: A table reporting the accuracy achieved by each model on the MNIST dataset.

3.1.2 Result on CIFAR-10 dataset



CIFAR-10	Accuracy (original)	Accuracy (reproduced)
L2-SVM	88.10%	68.48%
Cross-entropy	86.00%	68.37%

Figure 3: A table reporting the accuracy achieved by each model on the CIFAR-10 dataset.

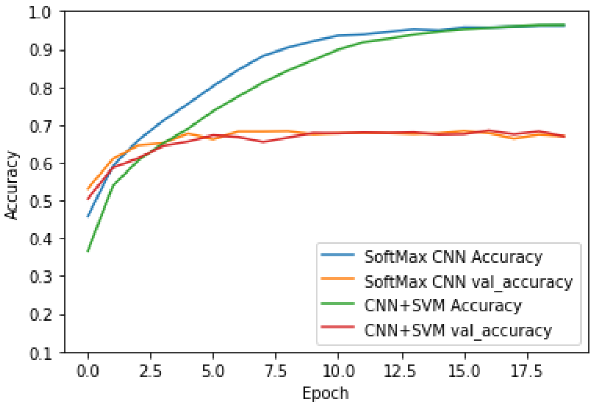


Figure 4: Accuracy vs. Epochs on training and testing set. label "accuracy" is the accuracy achieved on the training set while val_accuracy represents the accuracy on the testing set.

In this experiment, we also tried to strictly follow the description given by the author to build the model, but the parameters for stochastic gradient descent method were not given and the details about procedures of data augmentation is also unclear, so there might be some inconsistencies between our implementation and that of the author. But our results still support the conclusion that L2-SVM loss function sightly outperform cross-entropy loss with softmax activation. In Figure 4, after epoch 8, the CNN with SVM loss consistently achieved sightly higher accuracy than the CNN with cross-entropy loss.

3.2 Results beyond original paper

Besides reproducing the results given by the original paper, we also conducted a few experiments to perform some ablation studies and hyperparameter tuning.

3.2.1 MNIST: optimized number of epochs during training

In the original paper, the author trained the model on MNIST dataset over 400 epochs. However, in our experiment, it suggested that the accuracy (for both candidates models) had already reached its plateau around 100 epochs. Therefore, we decided to train our model over 100 instead of 400 epochs as the author suggested, which greatly boosts the training speed and allows the model to maintain the same level of performance.

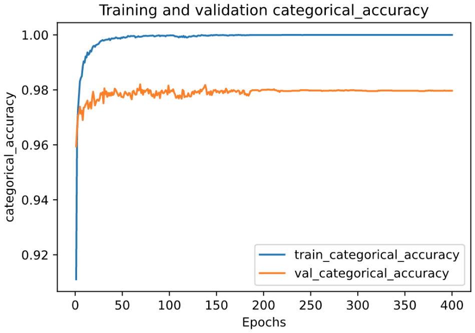


Figure 5: Training/Testing Accuracy vs training epochs of CNN using softmax

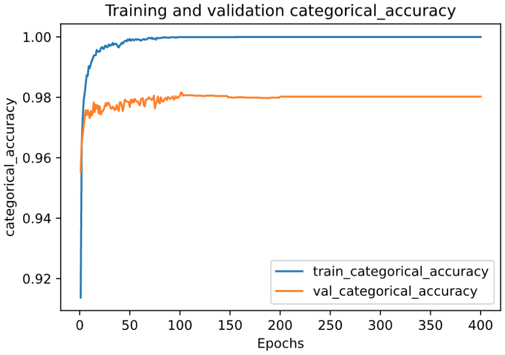


Figure 6: Training/Testing Accuracy vs training epochs of CNN using L2-SVM

3.2.2 MNIST: ablation study on PCA

In the original paper, the author also performed Principal Component Analysis (PCA) on MNIST dataset before the training of the model. What deserves special notice is the author decided to reduce the 784 dimensional data into 70 dimensional ones, which preserves approximately 60 70% of the variance of the original data.

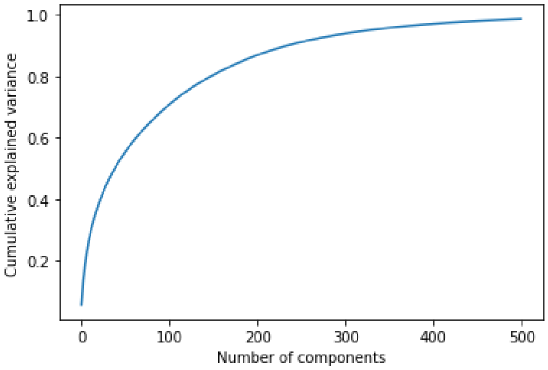


Figure 7: MNIST: variance preserved vs number of components preserved in PCA

Therefore, we were motivated to explore how the number of components in PCA will affect the performance of model, and we can predict the effect of removing PCA on the model performance.

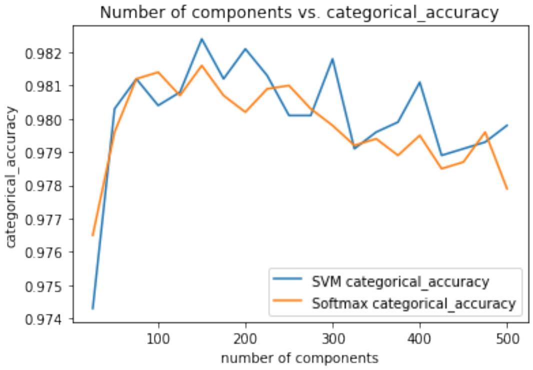


Figure 8: Testing accuracy of both models vs. number of components in PCA

In Figure 8, we can see that if we keep other hyperparameters fixed to the values given in the original paper, the accuracy of both models reaches its maximum when the number of components is around 175. After that, the accuracy of both models continuously goes down as the number of components preserved increases. Therefore, it's predictable that if we completely remove PCA before the training of models, the performance of both models would be undermined. Thus, performing PCA on MNIST dataset is a good practice and we agree with the author. Besides that, what also deserves special attention is that when more the components in the data are preserved, higher the dimension of the data used in training, and more frequently the model using L2-SVM loss outperforms the one using cross-entropy loss with softmax activation. Therefore, this ablation study gives supporting evidence to the conclusion of the original paper: when working with higher dimensional data, L2-SVM loss indeed has advantages over cross-entropy loss in multi-class classification problems.

3.2.3 CIFAR-10: Ablation study on Dropout

We also attempt to remove the dropout layer in the model trained on CIFAR-10 dataset to see the effect. To summarize, without dropout, the model performs poorly (lower 10% accuracy), and the training is futile. So dropout is extremely essential for this experiment.

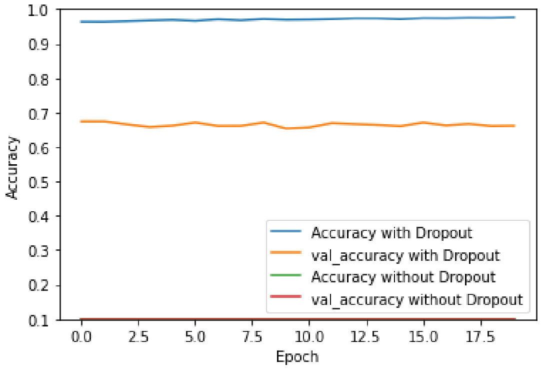


Figure 9: Accuracy of models with or without Dropout

4 Discussion of the details and challenges in this reproduction

In the original paper, the author did not mention whether or not we should perform data normalization before the application of PCA, despite the fact that normalization is a common practice before PCA. Also, the author did not offer sufficient details on the parameters of the stochastic gradient descent with momentum optimizer. Therefore, our main challenge in the reproduction is to understand the exact meaning of the ambiguous description of the experiments. To overcome this challenge, we sometimes used default parameters given by the library, while sometimes we tuned the parameters before application and it turns out to be unnecessary.

5 Summary of conclusion & key takeaways

In conclusion, although we did not reproduce the exact results in the original paper, our reproduction still offers supporting evidence of the main conclusion of the author. In addition, we conducted several additional experiments beyond the original paper, which improves the efficiency of the training, gives deeper insights into the functionality of each components of the implemented models and further supports the conclusion in the context of higher dimensional data. The key takeaway for this report is that a neural network model using L2-SVM loss function indeed shows slight advantages over the baseline model using cross-entropy loss with softmax activation function, and the advantages are more obvious in higher dimensional data.

6 Breakdown of the workload

Pingsheng Li: reproduce the experiments in the original paper & conduct additional ablation studies on the model using MNIST dataset.

Jacob Wang: reproduce the experiments in the original paper & conduct additional studies on the model using CIFAR-10 dataset.

Xiyuan Feng: analysis & interpretation of the experiment results and the compilation of the report.

References

Tang, Y., Deep Learning using Linear Support Vector Machines, *arXiv e-prints*, 2013.

Raiko, Tapani, Valpola, Harri, and LeCun, Yann. Deep learning made easier by linear transformations in per- ceptrons. *Journal of Machine Learning Research - Proceedings Track*, 22:924–932, 2012.

Rifai, Salah, Dauphin, Yann, Vincent, Pascal, Bengio, Yoshua, and Muller, Xavier. The manifold tangent classifier. In *NIPS*, pp. 2294–2302, 2011a.

LeCun, Y. & Cortes, C. (2010). MNIST handwritten digit database