# DEEPHOYER: LEARNING SPARSER NEURAL NETWORK WITH DIFFERENTIABLE SCALE-INVARIANT SPARSITY MEASURES –REPRODUCE CHALLENGE

**Yameng Li, Yuyue Guo, Shenhui Guo, Shurui Guo**
School of Electronics and Computer Science
University of Southampton, UK
{yl12u20,yg3u20,sg3m20,sz2r20}@soton.ac.uk

## ABSTRACT

The purpose of this paper is to reproduce the Deephoyer regularization method. This method is a set of sparse-induced regularizers, which is both almost everywhere differentiable and scale-invariant. This regularization method includes HS regularization and Group-HS regularization, which are used for element pruning and structure pruning, respectively. In this paper, we use the data set in the original paper and a new data set to reproduce this regularization method. The reproducibility of this method is verified, and some problems and phenomena in the process of reproducibility are analyzed.

## 1 INTRODUCTION

Modern DNN models, such as AlexNet(Krizhevsky et al., 2012) or ResNet(He et al., 2016), introduce a large number of model parameters and computation. It is because of this that the research of model compression technology is strengthened. In the original paper, it is the method of increasing the sparsity of the weight matrix that is discussed by the author to reduce the memory consumption and computational cost of DNN modelss(Yang et al., 2020). On the basis of Hoyer regularization (Hoyer, 2004) the author proposes DeepHoyer, a DNN sparseness regularization method , which combines the advantages of $L_1$ regularization and $L_0$ regularization. This method and is differentiable almost everywhere, easy to optimize, invariable in scale, and intuitively reflects the real sparseness. This method improves sparsity with little loss of accuracy.
In this paper, we repeat the Deephoyer regularization method proposed in Yang et al. (2020) through repeated experiments, and verify its conclusions.

## 2 BACKGROUND

Measuring sparsity with the Hoyer measure, such that the possible sparsest vectors with only one nonzero element has $S = 1$, and a vector with all equal elements has $S = 0$(Hoyer, 2004). The Hoyer measure is as follows(Yang et al., 2020):

$$S(X) = \frac{\sqrt{n} - (\sum_i |x_i|)/\sqrt{\sum_i x_i^2}}{\sqrt{n} - 1}, 1 \leq \frac{\sum_i |x_i|}{\sqrt{\sum_i x_i^2}} \leq \sqrt{n}, \forall X \in \mathbb{R}^n \tag{1}$$

Based on Hoyer regularization, the authors propose two types of DeepHoyer regularization: Hoyer-square regularization (HS) for element pruning and Group-hs regularization for structural pruning(Yang et al., 2020). HS regularization is to square Hoyer regularization:

$$H_S(W) = \frac{(\sum_i |x_i|)^2}{\sum_i x_i^2} \tag{2}$$

More weights can be close to zero to expand the pruning range.
The Group - HS regularizer for structure pruning use HS regularizer to replace the $L_1$ regularizer in the group lasso formulation(Yang et al., 2020).

## 3 EXPERIMENT

We only have reproduced the part of the experiment on the MNIST dataset described in the original paper, because the network structure used on other datasets was relatively complex and could not highlight the description of the Hoyer-Square Regularizer in the original paper.We firstly obtain the Element-wise Pruning Model and Structural Pruning Model trained by several different regularizers and these are dense models. Then we prune the dense models obtained in the previous step based on a fixed threshold and based on a fixed ratio to the std of each layer. The parameters used in the experiment are those provided in the original paper, and those not given are the default parameters.

### 3.1 EXPERIMENTAL DATA AND RESULTS

Table 1 shows the accuracy and pruning ratio of the obtained model after applying different regularizers to the LeNet-300-100 (MLP) model and using different pruning strategies.

Table 1: The result of pruning the LeNet-300-100 MLP model

| | **Pruning based on a fixed threshold** | | **Pruning based on a fixed ratio to the std of each layer** | |
|---|---|---|---|---|
| | Accuracy | Pruning ratio | Accuracy | Pruning ratio |
| | **Element-wise pruning results** | | | |
| No regularizer(None) | 98.71% | 0.76% | 98.44% | 34.69% |
| L1 regularizer | 96.21% | 97.55% | 96.21% | 97.97% |
| Hoyer regularizer | 97.47% | 89.70% | 97.10% | 98.75% |
| **Hoyer-Square regularizer** | **97.97%** | **74.02%** | **97.99%** | **96.28%** |
| Transformed L1 regularizer | 98.15% | 88.57% | 98.24% | 96.91% |
| | **Structural pruning results** | | | |
| No regularizer(None) | 98.59% | 70.30% | 98.51% | 78.23% |
| Group Lasso regularizer | 97.80% | 92.85% | 97.78% | 93.81% |
| **Group-HS regularizer** | **97.79%** | **91.07%** | **97.80%** | **98.64%** |

Table 2 shows the accuracy and pruning ratio of the obtained model after applying different regularizers to the LeNet-5 (CNN) model and using different pruning strategies.

Table 2: The result of pruning the LeNet-5 CNN model

| | **Pruning based on a fixed threshold** | | **Pruning based on a fixed ratio to the std of each layer** | |
|---|---|---|---|---|
| | Accuracy | Pruning ratio | Accuracy | Pruning ratio |
| | **Element-wise pruning results** | | | |
| No regularizer(None) | 99.37% | 1.06% | 99.40% | 32.01% |
| L1 regularizer | 98.69% | 99.52% | 98.59% | 99.53% |
| Hoyer regularizer | 99.11% | 87.04% | 97.32% | 99.58% |
| **Hoyer-Square regularizer** | **99.18%** | **82.98%** | **99.17%** | **98.59%** |
| Transformed L1 regularizer | 99.20% | 96.99% | 99.16% | 98.95% |
| | **Structural pruning results** | | | |
| No regularizer(None) | 99.36% | 89.05% | 99.35% | 92.51% |
| Group Lasso regularizer | 98.54% | 99.31% | 98.65% | 99.40% |
| **Group-HS regularizer** | **97.25%** | **99.76%** | **97.33%** | **99.74%** |

### 3.2 EXPERIMENTAL ANALYSIS AND CONCLUSION

1. Conclusions related to the description of the paper:in the original paper, the author proposed that Hoyer-square regularizer can let the models have the highest sparsity in our experimental results(Yang et al., 2020). But from the "Element-wise Pruning results" parts in Table 1 and 2, the results of the experiment do not support this conclusion. By comparing them, it can be found that although the sparsity of the model processed by Hoyer-Square Regularizer is not as high as that processed by Hoyer Regularizer, the accuracy of the former is greatly improved. However, the Group-

HS Regularizer is indeed effective when doing structure pruning. As can be seen from the results in Table 1 and 2, it greatly improves the sparsity of the model and has little effect on the accuracy.

2. Extra conclusion obtained through experiment: the groups with no regularization were the basic control group. In these groups, the comparison of the pruning ratio of structural pruning and element-wise pruning shows that structural pruning can help pruning to some extent. Hoyer-square Regularizer and Group-HS Regularizer are compared using the results of element pruning and structure Pruning. The MNIST experiment uses Group-HS Regularizer based on a fixed ratio to the STD of each layer can greatly improve the pruning ratio without loss of precision.

### 3.3 THE PERFORMANCE OF HOYER-SQUARE REGULARIZER

From the figure below, we can find that whether Hs is used in CNN or MLP, good pruning results can be achieved. In CNN, the FC1 layer has a 93.3% weight of 0, and the pruning rate of the FC1 and FC2 layers in the MLP also exceeds 90%.

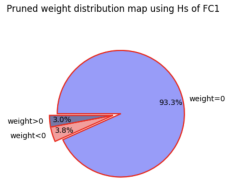


Fig. 1: FC1 layer in CNN

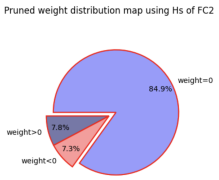


Fig. 2: FC2 layer in CNN

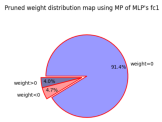


Fig. 3: FC1 layer in MLP

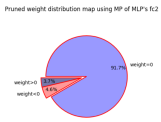


Fig. 4: FC2 layer in MLP

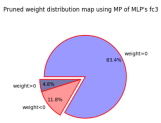


Fig. 5: FC3 layer in MLP

## 4 APPLIED TO THE FANSHIONMNIST DATASET

On the basis of reproducing the conclusion of the original theory, we tried a new dataset to study whether the content of the paper can be applied to other datasets.

### 4.1 EXPERIMENT CONTENT

Table 3: DeepHoyer regularization results on the FanshionMnist dataset

| model | **LeNet-5** | | **LeNet-300-100** | |
|---|---|---|---|---|
| Method | Accuracy | total weight | Accuracy | total weight |
| No regularizer | 90.68% | 431,080 | 88.68% | 266,610 |
| | **element pruning result** | | | |
| Hoyer | 91.47% | 176,702 | 89.31% | 47,015 |
| **Hoyer-Square** | **91.17%** | **24,026** | **89.06%** | **23,318** |
| | **structural pruning result** | | | |
| **Group-HS** | **90.82%** | **8,025** | **89.22%** | **25,617** |

Based on the network model provided in the original paper, they were tested on LeNet-5 model and LeNet-300-100 model.
After many adjustments to the parameters, the training of the dense model still uses the default values. In LeNet-5, the Hoyer regularizer uses decay=0.001 to obtain an accuracy rate of 90.89%; for the Hoyer-Square regularizer uses decay=0.0001, an accuracy rate of 90.37% is obtained; in

LeNet-300- In 100, for the decay=0.002 used by the Hoyer regularizer, an accuracy rate of 89.56% is obtained; for the decay=0.0002 used by the Hoyer-Square regularizer, an accuracy rate of 88.11% is obtained;

### 4.2 Result analysis

From the data in the Table 3, we can see that in the two network models, Hoyer-Square and Group-HS regularizers are effective and perform well. But in fact, we also applied the Transformed L1 regularizer to the LeNet-5 model, and found that its effect is better. In the case of an accuracy of 90.47%, the non-zero weight number is 4,156.
From the results, the model compression effect obtained by the Hoyer-Square method is obviously better than that of the Hoyer method. In the experiment, we spent a long time on parameter adjustment, because we found that the number of effective weights reflected when the accuracy fluctuates in a small range varies greatly, so it is difficult to determine the parameters. The data cannot clearly conclude that the model after being regularized by the DeepHoyer regularizer is optimal in terms of accuracy and model size.

## 5 Final conclusion

In this experiment, the reproduction part of the original paper is to train the MNIST data set on two network model. The reproduction results of this part cannot fully support the author's conclusion that the DeepHoyer regularizer is always superior to previous work in terms of element pruning and structural pruning. However, what we can confirm is that it can indeed take into account the accuracy and model compression rate during structure pruning by using Group-HS, and get better results for both.
We believe that the performance of the DeepHoyer regularizer is still very good. When the accuracy requirements are not strict, after using a fixed ratio of pruning for each layer, the reproduced model results are better than the previous technology. In addition, the results of using the DeepHoyer regularizer on the new data set also show the effectiveness of the method. However, when this method is applied to new data sets, it needs to go through a large and complex parameter adjustment process, which is very time-consuming and increases the calculation amount, which is not mentioned in the original paper by the author. Because under different combinations of decay and threshold values, the difference in accuracy obtained is small but the corresponding pruning ratio is different. This makes us guess the reason why the results cannot be fully reproduced. It may be because there are many parameters and comparison methods that need to be tested during the reproduction process. In the case of small changes inaccuracy, a model with the same parameters cannot guarantee a stable prediction accuracy.

Github repository: https://github.com/COMP6248-Reproducability-Challenge/DeepHoyer

## References

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Patrik O Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(9), 2004.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pp. 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

Huanrui Yang, Wei Wen, and Hai Li. Deephoyer: Learning sparser neural network with differentiable scale-invariant sparsity measures. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rylBK34FDS`.