

ℓ_0 Regularized Structured Sparsity Convolutional Neural Networks

Kevin Bui

Department of Mathematics
University of California, Irvine

kevinb3@uci.edu

Shuai Zhang

Qualcomm AI

shuazhan@qti.qualcomm.com

Fredrick Park
Department of Mathematics & Computer Science
Whittier College

fspark@whittier.edu

Yingyong Qi

Qualcomm AI

yqi@uci.edu

Jack Xin

Department of Mathematics
University of California, Irvine

jxin@math.uci.edu

December 18, 2019

Abstract

Deepening and widening convolutional neural networks (CNNs) significantly increases the number of trainable weight parameters by adding more convolutional layers and feature maps per layer, respectively. By imposing inter- and intra-group sparsity onto the weights of the layers during the training process, a compressed network can be obtained with accuracy comparable to a dense one. In this paper, we propose a new variant of sparse group lasso that blends the ℓ_0 norm onto the individual weight parameters and the $\ell_{2,1}$ norm onto the output channels of a layer. To address the non-differentiability of the ℓ_0 norm, we apply variable splitting resulting in an algorithm that consists of executing stochastic gradient descent followed by hard thresholding for each iteration. Numerical experiments are demonstrated on LeNet-5 and wide-residual-networks for MNIST and CIFAR 10/100, respectively. They showcase the effectiveness of our proposed method in attaining superior test accuracy with network sparsification on par with the current state of the art.

1 Introduction

Deep neural networks (DNNs) have proven to be advantageous for numerous modern computer vision tasks involv-

ing image or video data. In particular, convolutional neural networks (CNNs) yield highly accurate models with applications in image classification [21, 36, 14, 44], semantic segmentation [25, 7], and object detection [32, 16, 31]. These large models often contain millions or even billions of weight parameters that often exceed the number of training data. This is a double-edged sword since on one hand, large models allow for high accuracy, while on the other, they contain many redundant parameters that lead to overparametrization. Overparametrization is a well-known phenomenon in DNN models [8, 3] that results in overfitting, learning useless random patterns in data [45], and having inferior generalization. Additionally, these models also possess exorbitant computational and memory demands during both training and inference. As a result, they may not be applicable for devices with low computational power and memory.

Resolving these problems requires compressing the networks through sparsification and pruning. Although removing weights might affect the accuracy and generalization of the models, previous works [26, 12, 39, 29] demonstrated that many networks can be substantially pruned with negligible effect on accuracy. There are many systematic approaches to achieving sparsity in DNNs.

Han *et al.*[13] proposed to first train a dense network, prune it afterward by setting the weights to zero if below a fixed threshold, and retrain the network with the

remaining weights. Jin *et al.*[18] extended this method by restoring the pruned weights, training the network again, and repeating the process. Rather than pruning by thresholding, Aghasi *et al.*[1] proposed Net-Trim, which prunes an already trained network layer-by-layer using convex optimization in order to ensure that the layer inputs and outputs remain consistent with the original network. For CNNs in particular, filter or channel pruning is preferred because it significantly reduces the amount of weight parameters required compared to individual weight pruning. Le *et al.*[24] calculated the sums of absolute weights of the filters of each layer and pruned the ones with the smallest weights. Hu *et al.*[15] proposed a metric called average percentage of zeroes for channels to measure their redundancies and pruned those with highest values for each layer. Zhuang *et al.*[48] developed discrimination-aware channel pruning that selects channels that contribute to the discriminative power of the network.

An alternative approach to pruning a dense network is learning a compressed structure from scratch. A conventional approach is to optimize the loss function equipped with either the ℓ_1 or ℓ_2 regularization, which drives the weights to zero or to very small values during training. To learn which groups of weights (e.g., neurons, filters, channels) are necessary, group regularization, such as group lasso [42] and sparse group lasso [35], are equipped to the loss function. Alvarez and Salzmann [2] and Scardapane *et al.*[34] applied group lasso and sparse group lasso to various architectures and obtained compressed networks with comparable or even better accuracy. Instead of sharing features among the weights as suggested by group sparsity, exclusive sparsity [47] promotes competition for features between different weights. This method was investigated by Yoon and Hwang [41]. In addition, they combined it with group sparsity and demonstrated that this combination resulted in compressed networks with better performance than their original. Non-convex regularization has also been examined. Louizos *et al.*[26] proposed a practical algorithm using probabilistic methods to perform ℓ_0 regularization on neural networks. Ma *et al.*[28] proposed integrated transformed ℓ_1 , a convex combination of transformed ℓ_1 and group lasso, and compared its performance against the aforementioned group regularization methods.

In this paper, we propose a group regularization method that balances both group lasso and ℓ_0 regularization: it is

a variant of sparse group lasso that replaces the ℓ_1 penalty term with the ℓ_0 penalty term. This proposed group regularization method is presumed to yield a better performing, compressed network than sparse group lasso since ℓ_1 is a convex relaxation of ℓ_0 . We develop an algorithm to optimize loss functions equipped with the proposed regularization term for DNNs.

2 Model and Algorithm

2.1 Preliminaries

Given a training dataset consisting of N input-output pairs $\{(x_i, y_i)\}_{i=1}^N$, the weight parameters of a DNN are learned by optimizing the following objective function:

$$\min_W \frac{1}{N} \sum_{i=1}^N \mathcal{L}(h(x_i, W), y_i) + \lambda \mathcal{R}(W), \quad (1)$$

where

- W is the set of weight parameters of the DNN.
- $\mathcal{L}(\cdot, \cdot) \geq 0$ is the loss function that compares the prediction $h(x_i, W)$ with the ground-truth output y_i . Examples include cross-entropy loss function for classification and mean-squared error for regression.
- $h(\cdot, \cdot)$ is the output of the DNN used for prediction.
- $\lambda > 0$ is a regularization parameter for $\mathcal{R}(\cdot)$.
- $\mathcal{R}(\cdot)$ is the regularizer on the set of weight parameters W .

The most common regularizer used for DNN is $\|\cdot\|_2^2$, also known as weight decay. It prevents overfitting and improves generalization because it enforces the weights to decrease proportional to their magnitudes [22]. Sparsity can be imposed by pruning weights whose magnitudes are below a certain threshold at each iteration during training. However, an alternative regularizer is the ℓ_1 norm $\|\cdot\|_1$, also known as lasso [37]. ℓ_1 norm is the tightest convex relaxation of the ℓ_0 norm and it yields a sparse solution that is found on the corners of the 1-norm ball. Unfortunately, element-wise sparsity by ℓ_1 or ℓ_2 regularization

in CNNs may not yield meaningful speedup as the number of filters and channels required for computation and inference may remain the same [40].

To determine which filters or channels are relevant in each layer, group sparsity using group lasso is considered. Suppose a DNN has L layers, so the set of weight parameters W is divided into L sets of weights: $W = \{W_l\}_{l=1}^L$. The weight set of each layer W_l is divided into N_l groups (e.g., channels or filters): $W_l = \{w_{l,g}\}_{g=1}^{N_l}$. Group lasso applied to W_l is formulated as

$$\begin{aligned} \mathcal{R}_{GL}(W_l) &= \sum_{g=1}^{N_l} \sqrt{|w_{l,g}|} \|w_{l,g}\|_2 \\ &= \sum_{g=1}^{N_l} \sqrt{|w_{l,g}|} \sqrt{\sum_{i=1}^{|w_{l,g}|} w_{l,g,i}^2}, \end{aligned} \quad (2)$$

where $w_{l,g,i}$ corresponds to the weight parameter with index i in group g in layer l , and the term $\sqrt{|w_{l,g}|}$ ensures that each group is weighed uniformly. This regularizer imposes the ℓ_2 norm on each group, forcing weights of the same groups to decrease altogether at every iteration during training. As a result, groups of weights are pruned when their ℓ_2 norms are negligible, resulting in a highly compact network compared to element-sparse networks.

To obtain an even sparser network, element-wise sparsity and group sparsity can be combined and applied together to the training of DNNs. One regularizer that combines these two types of sparsity is sparse group lasso, which is formulated as

$$\mathcal{R}_{SGL}(W_l) = \mathcal{R}_{GL}(W_l) + \|W_l\|_1, \quad (3)$$

where

$$\|W_l\|_1 = \sum_{g=1}^{N_l} \sum_{i=1}^{|w_{l,g}|} |w_{l,g,i}|.$$

Sparse group lasso simultaneously enforces group sparsity by having $\mathcal{R}_{GL}(\cdot)$ and element-wise sparsity by having $\|\cdot\|_1$.

2.2 Proposed Regularizer: Sparse Group L_0 asso

We recall that the ℓ_1 norm is a convex relaxation of the ℓ_0 norm, which is non-convex and discontinuous. In addition,

any ℓ_0 -regularized problem is NP-hard. These properties make developing convergent and tractable algorithms for ℓ_0 -regularized problems difficult, thereby making ℓ_1 -regularized problems better alternatives to solve. However, the ℓ_0 -regularized problems have their advantages over their ℓ_1 counterparts. For example, they are able to recover better sparse solutions than do ℓ_1 -regularized problems in various applications, such as compressed sensing [27], image restoration [4, 6, 10, 46], MRI reconstruction [38], and machine learning [27, 43]. Used to solve ℓ_1 minimization, the soft-thresholding operator $\mathcal{S}_\lambda(c) = \text{sign}(c) \max\{|c| - \lambda, 0\}$ yields a biased estimator [11].

Due to the advantages and recent successes of ℓ_0 minimization, we propose to replace the ℓ_1 norm in (3) with the ℓ_0 norm

$$\|W_l\|_0 = \sum_{g=1}^{N_l} \sum_{i=1}^{|w_{l,g}|} |w_{l,g,i}|_0, \quad (4)$$

where

$$|w|_0 = \begin{cases} 1 & \text{if } w \neq 0 \\ 0 & \text{if } w = 0. \end{cases}$$

Hence, we propose a new regularizer called sparse group ℓ_0 asso defined by

$$\mathcal{R}_{SGL_0}(W_l) = \mathcal{R}_{GL}(W_l) + \|W_l\|_0. \quad (5)$$

Using this regularizer, we expect to obtain a better sparse network than from using sparse group lasso.

2.3 Notation

Before discussing the algorithm, we summarize notations that we will use to save space. They are the following:

- If $V = \{V_l\}_{l=1}^L$ and $W = \{W_l\}_{l=1}^L$, then $(V, W) := (\{V_l\}_{l=1}^L, \{W_l\}_{l=1}^L) = (V_1, \dots, V_L, W_1, \dots, W_L)$.
- For $V = \{V_l\}_{l=1}^L$, $V_{<l} = (V_1, \dots, V_{l-1})$ and $V_{>l} = (V_{l+1}, \dots, V_L)$. Both $V_{\leq l}$ and $V_{\geq l}$ are defined similarly.
- $V^+ := V^{k+1}$.
- $\tilde{\mathcal{L}}(W) := \frac{1}{N} \sum_{i=1}^N \mathcal{L}(h(x_i, W), y_i)$.

2.4 Numerical Optimization

We develop an algorithm to solve (1) with the sparse group ℓ_0 asso regularizer (5). So, with $W = \{W_l\}_{l=1}^L$, the minimization problem we solve is

$$\begin{aligned} \min_W \tilde{\mathcal{L}}(W) + \lambda \sum_{l=1}^L \mathcal{R}_{SGL_0}(W_l) \\ = \tilde{\mathcal{L}}(W) + \lambda \sum_{l=1}^L (\mathcal{R}_{GL}(W_l) + \|W_l\|_0). \end{aligned} \quad (6)$$

Throughout this paper, we assume that \mathcal{L} is continuously differentiable with respect to W_l for each $l = 1, \dots, L$. Because finding the subderivative of the objective problem is difficult due to the ℓ_0 norm, we need to figure out a method to solve it. By introducing an auxiliary variable $V = \{V_l\}_{l=1}^L$, we have a constrained optimization problem

$$\begin{aligned} \min_{V, W} \tilde{\mathcal{L}}(W) + \lambda \sum_{l=1}^L (\mathcal{R}_{GL}(W_l) + \|V_l\|_0) \\ \text{s.t. } V = W. \end{aligned} \quad (7)$$

The constraint can be relaxed by adding a quadratic penalty term with $\beta > 0$ so that we have

$$\begin{aligned} \min_{V, W} F_\beta(V, W) \\ := \tilde{\mathcal{L}}(W) + \sum_{l=1}^L \left[\lambda (\mathcal{R}_{GL}(W_l) + \|V_l\|_0) \right. \\ \left. + \frac{\beta}{2} \|V_l - W_l\|_2^2 \right]. \end{aligned} \quad (8)$$

With β fixed, (8) can be solved by alternating minimization:

$$W_l^{k+1} = \arg \min_{W_l} F_\beta(V^k, W_{<l}^+, W_l, W_{>l}^k) \quad (9a)$$

for $l = 1, \dots, L$

$$V^{k+1} = \arg \min_V F_\beta(V, W^{k+1}). \quad (9b)$$

We explicitly update W_l by gradient descent and V_l by hard-thresholding:

$$\begin{aligned} W_l^{k+1} = W_l^k - \gamma \left(\nabla_{W_l} \tilde{\mathcal{L}}(W) \right. \\ \left. + \lambda \partial \mathcal{R}_{GL}(W_l^k) - \beta (V_l^k - W_l^k) \right) \text{ for } l = 1, \dots, L \end{aligned} \quad (10a)$$

$$V^{k+1} = \mathcal{H}_{\sqrt{2\lambda/\beta}}(W^{k+1}), \quad (10b)$$

where γ is the learning rate, $\partial \mathcal{R}_{GL}$ is the subdifferential of \mathcal{R}_{GL} , and $\mathcal{H}_{\sqrt{2\lambda/\beta}}(\cdot)$ is the element-wise hard-thresholding operator:

$$\mathcal{H}_{\sqrt{2\lambda/\beta}}(w_i) = \begin{cases} 0 & \text{if } |w_i| \leq \sqrt{2\lambda/\beta} \\ w_i & \text{if } |w_i| > \sqrt{2\lambda/\beta}. \end{cases} \quad (11)$$

In practice, (10a) is performed using stochastic gradient descent (or one of its variants) with mini-batches due to the large-size computation dealing with the amount of data and weight parameters that a typical DNN has.

After presenting an algorithm that solves the quadratic penalty problem (8), we now present an algorithm to solve (6). We solve a sequence of quadratic penalty problems (8) with $\beta \in \{\beta_j\}_{j=1}^\infty$ such that $\beta_j \uparrow \infty$. This will yield a sequence $\{(V^j, W^j)\}_{j=1}^\infty$ such that $W^j \uparrow W^*$, a solution to (6). This algorithm is based on the quadratic penalty method [30] and the penalty decomposition method [27]. The algorithm is summarized in Algorithm 1.

2.5 Convergence Analysis

To establish convergence for the proposed algorithm, we show that the accumulation point of the sequence generated by (9a)-(9b) is a block-coordinate minimizer, and an accumulation point generated by Algorithm 1 is a sparse feasible solution to (6). Unfortunately, this feasible solution may not be a local minimizer of (6) because the loss function $\mathcal{L}(\cdot, \cdot)$ is nonconvex. However, it was shown in [9] that a similar algorithm to (1), but only for ℓ_0 minimization, generates an approximate global solution with high probability for a one-layer CNN with ReLu activation function.

Theorem 1. *Let $\{(V^k, W^k)\}_{k=1}^\infty$ be a sequence generated by the alternating minimization algorithm (9a)-(9b).*

Algorithm 1: Algorithm for Sparse Group L_0 asso Regularization

```

1 Initialize  $V^1$  and  $W^1$  with random entries; learning
  rate  $\gamma$ ; regularization parameters  $\lambda$  and  $\beta$ ; and
  multiplier  $\sigma > 1$ .
2 Set  $j := 1$ .
3 while stopping criterion for outer loop not satisfied
  do
4   Set  $k := 1$ .
5   Set  $W^{j,1} = W^j$  and  $V^{j,1} = V^j$ .
6   while stopping criterion for inner loop not
     satisfied do
7     Update  $W^{j,k+1}$  by Eq. (10a).
8     Update  $V^{j,k+1}$  by Eq. (10b).
9      $k := k + 1$ 
10  end
11  Set  $W^{j+1} = W^{j,k}$  and  $V^{j+1} = V^{j,k}$ .
12  Set  $\beta := \sigma\beta$ .
13  Set  $j := j + 1$ .
14 end

```

If (V^*, W^*) is an accumulation point of $\{(V^k, W^k)\}_{k=1}^\infty$, then (V^*, W^*) is a block coordinate minimizer of (8) that is

$$V^* \in \arg \min_{V^l} F_\beta(V, W^*)$$

$$W_l^* \in \arg \min_{W_l} F_\beta(V^*, W_{<l}^*, W_l, W_{>l}^*) \text{ for } l = 1, \dots, L$$

Proof. By (9a)-(9b), we have

$$F_\beta(V^k, W_{\leq l}^+, W_{>l}^k) \leq F_\beta(V^k, W_{<l}^+, W_l, W_{>l}^k) \quad (12)$$

$$F_\beta(V^+, W^+) \leq F_\beta(V, W^+) \quad (13)$$

for all $W_l, l = 1, \dots, L$, and V , so it follows after some computation that

$$F_\beta(V^+, W^+) \leq F_\beta(V^k, W^k) \quad (14)$$

for each $k \in \mathbb{N}$. Hence, $\{F_\beta(V^k, W^k)\}_{k=1}^\infty$ is non-increasing. Since $F_\beta(V^k, W^k) \geq 0$ for all $k \in \mathbb{N}$, its limit $\lim_{k \rightarrow \infty} F_\beta(V^k, W^k)$ exists. From (12), we have

$$F_\beta(V^k, W_{\leq l}^+, W_{>l}^k) \leq F_\beta(V^k, W_{<l}^+, W_{\geq l}^k)$$

for each l . Because F_β is continuous with respect to W_l , applying the limit gives us

$$\lim_{k \rightarrow \infty} F_\beta(V^k, W_{\leq l}^+, W_{>l}^k) = \lim_{k \rightarrow \infty} F_\beta(V^k, W_{<l}^+, W_{\geq l}^k). \quad (15)$$

Since (V^*, W^*) is an accumulation point of $\{(V^k, W^k)\}_{k=1}^\infty$, there exists a subsequence K such that $\lim_{k \in K \rightarrow \infty} (V^k, W^k) = (V^*, W^*)$. If $\lim_{k \in K \rightarrow \infty} V^k = V^*$, there exists $k' \in K$ such that $k \geq k'$ implies $\|V_l^k\|_0 \geq \|V_l^*\|_0$ for each $l = 1, \dots, L$. As a result, we obtain

$$\begin{aligned} F_\beta(V, W^{k+1}) &\geq F_\beta(V^{k+1}, W^{k+1}) \\ &\geq \tilde{\mathcal{L}}(W^{k+1}) \\ &\quad + \sum_{l=1}^L \left[\lambda (\mathcal{R}_{GL}(W_l^{k+1}) + \|V_l^*\|_0) \right. \\ &\quad \left. + \frac{\beta}{2} \|V_l^{k+1} - W_l^{k+1}\|_2^2 \right] \end{aligned}$$

for $k \geq k'$ from (13). Using continuity, except for the ℓ_0 term, and letting $k \in K \rightarrow \infty$, we obtain

$$F_\beta(V, W^*) \geq F_\beta(V^*, W^*). \quad (16)$$

For notational convenience, let

$$\tilde{\mathcal{R}}_{\lambda, \beta}(V, W) := \lambda \mathcal{R}_{GL}(W) + \frac{\beta}{2} \|W - V\|_2^2. \quad (17)$$

From (12), we have

$$\tilde{\mathcal{L}}(W_{<l}^+, W_l, W_{>l}^k) + \sum_{j < l} \tilde{\mathcal{R}}_{\lambda, \beta}(W_j^+, V_j^k) + \tilde{\mathcal{R}}_{\lambda, \beta}(W_l, V_l^k) \quad (18)$$

$$+ \sum_{j > l} \tilde{\mathcal{R}}_{\lambda, \beta}(W_j^k, V_j^k)$$

$$= F_\beta(V, W_{<l}^+, W_l, W_{>l}^k) - \lambda \sum_{l=1}^L \|V_l^k\|_0$$

$$\geq F_\beta(V, W_{\leq l}^+, W_{>l}^k) - \lambda \sum_{l=1}^L \|V_l^k\|_0$$

$$= \tilde{\mathcal{L}}(W_{\leq l}^+, W_{>l}^k) + \sum_{j \leq l} \tilde{\mathcal{R}}_{\lambda, \beta}(W_j^+, V_j^k)$$

$$+ \sum_{j > l} \tilde{\mathcal{R}}_{\lambda, \beta}(W_j^k, V_j^k)$$

for all $k \in K$. Because $\lim_{k \in K \rightarrow \infty} V^k$ exists, the sequence $\{V^k\}_{k \in K}$ is bounded, which implies that $\{\|V^k\|_0\}_{k \in K}$ is bounded as well. Hence, there exists a further subsequence $\bar{K} \subset K$ such that $\lim_{k \in \bar{K} \rightarrow \infty} \|V^k\|_0$ exists. As a result, For each $l = 1, \dots, L$, we have that $\lim_{k \in \bar{K}} \|V_l^k\|_0$ exists. So, we obtain

$$\begin{aligned}
& \lim_{k \in \bar{K} \rightarrow \infty} \tilde{\mathcal{L}}(W_{\leq l}^+, W_{> l}^k) + \sum_{j \leq l} \tilde{\mathcal{R}}_{\lambda, \beta}(W_j^+, V_j^k) \quad (19) \\
& + \sum_{j > l} \tilde{\mathcal{R}}_{\lambda, \beta}(W_j^k, V_j^k) \\
& = \lim_{k \in \bar{K} \rightarrow \infty} F_{\beta}(V, W_{\leq l}^+, W_{> l}^k) - \lambda \sum_{l=1}^L \|V_l^k\|_0 \\
& = \lim_{k \in \bar{K} \rightarrow \infty} F_{\beta}(V, W_{\leq l}^+, W_{> l}^k) - \lim_{k \in \bar{K} \rightarrow \infty} \lambda \sum_{l=1}^L \|V_l^k\|_0 \\
& = \lim_{k \in \bar{K} \rightarrow \infty} F_{\beta}(V^k, W_{< l}^+, W_{\geq l}^k) - \lim_{k \in \bar{K} \rightarrow \infty} \lambda \sum_{l=1}^L \|V_l^k\|_0 \\
& = \dots \\
& = \lim_{k \in \bar{K} \rightarrow \infty} F_{\beta}(V^k, W^k) - \lim_{k \in \bar{K} \rightarrow \infty} \lambda \sum_{l=1}^L \|V_l^k\|_0 \\
& = \lim_{k \in \bar{K} \rightarrow \infty} F_{\beta}(V^k, W^k) - \lambda \sum_{l=1}^L \|V_l^k\|_0 \\
& = \lim_{k \in \bar{K} \rightarrow \infty} \tilde{\mathcal{L}}(W^k) + \sum_{l=1}^L \tilde{\mathcal{R}}_{\lambda, \beta}(W_l^k, V_l^k) \\
& = \tilde{\mathcal{L}}(W^*) + \sum_{l=1}^L \tilde{\mathcal{R}}_{\lambda, \beta}(W_l^*, V_l^*)
\end{aligned}$$

after applying (15). Taking the limit over the subsequence \bar{K} in (18) and applying (19), we obtain

$$\begin{aligned}
& \tilde{\mathcal{L}}(W_{< l}^*, W_l, W_{> l}^*) + \sum_{j \neq l} \tilde{\mathcal{R}}_{\lambda, \beta}(W_j^*, V_j^*) \quad (20) \\
& + \tilde{\mathcal{R}}_{\lambda, \beta}(W_l, V_l^*) \geq \tilde{\mathcal{L}}(W^*) + \sum_{l=1}^L \tilde{\mathcal{R}}_{\lambda, \beta}(W_l^*, V_l^*)
\end{aligned}$$

Adding $\sum_{l=1}^L \|V_l^*\|_0$ on both sides yields

$$F_{\beta}(V^*, W_{< l}^*, W_l, W_{> l}^*) \geq F_{\beta}(V^*, W^*). \quad (21)$$

By (17) and (21), (V^*, W^*) is a block coordinate minimizer. \square

Theorem 2. Let $\{(V^k, W^k, \beta_k)\}_{k=1}^{\infty}$ be a sequence generated by Algorithm 1. Suppose that $\{F_{\beta_k}(V^k, W^k)\}_{k=1}^{\infty}$ is uniformly bounded. If (V^*, W^*) is an accumulation point of $\{(V^k, W^k)\}_{k=1}^{\infty}$, then $V^* = W^*$ and W^* is a feasible solution to (6).

Proof. Because (V^*, W^*) is an accumulation point, there exists a subsequence K such that $\lim_{k \in K \rightarrow \infty} (V^k, W^k) = (V^*, W^*)$. If $\{F_{\beta_k}(V^k, W^k)\}_{k=1}^{\infty}$ is uniformly bounded, there exists M such that $F_{\beta_k}(V^k, W^k) \leq M$ for all $k \in \mathbb{N}$. After some algebraic manipulation, we should obtain

$$\sum_{l=1}^L \|V_l^k - W_l^k\|_2^2 \leq \frac{2}{\beta_k} M, \quad (22)$$

where M is some positive constant equals to the total number of weight parameters in W . Taking the limit over $k \in K$, we have

$$\sum_{l=1}^L \|V_l^* - W_l^*\|_2^2 = 0,$$

which follows that $V^* = W^*$. As a result, W^* is a feasible solution to (6). \square

3 Experiments

We compare the proposed sparse group ℓ_0 asso regularization against four other methods as baselines: group lasso, sparse group lasso, combined group and exclusive sparsity (CGES) proposed in [41], and the group variant of ℓ_0 regularization proposed in [26]. For the group terms, the weights are grouped together based on the filters or output channels, which we will refer to as neurons. We apply these methods on the following image datasets: MNIST [23] using the LeNet-5-Caffe [17] and CIFAR 10/100 [20] using wide residual networks [44]. Because the optimization algorithms do not drive most, if not all, the weights and neurons to zeroes, we have to set them to zeroes when their values are below a certain threshold. In our experiments, if the absolute weights are below 10^{-5} , we set them to zeroes. Then, **weight sparsity** is defined to be the

percentage of zero weights with respect to the total number of weights trained in the network. If the normalized sum of the absolute values of the weights of the neuron is less than 10^{-5} , then the weights of the neuron are set to zeroes. **Neuron sparsity** is defined to be the percentage of neurons whose weights are zeroes with respect to the total number of neurons in the network.

3.1 MNIST Classification

The MNIST dataset consists of 60k training images and 10k test images. It is trained on Lenet-5-Caffe, which has four layers with 1,370 total neurons and 431,080 total weight parameters. All layers of the network are applied with strictly the same type of regularization. No other regularization methods (e.g., dropout and batch normalization) are used. The network is optimized using Adam [19] with initial learning rate 0.001. For every 40 epochs, the learning rate decays by a factor of 0.1. We set the regularization parameter $\lambda = 0.1/60000$. For sparse group l_0 asso, we set $\beta = 2.5/60000$, and for every 40 epochs, it increases by a factor of 1.25. The network is trained for 200 epochs across 5 runs.

Table 1 reports the mean results for weight sparsity, neuron sparsity, and test error obtained at the end of the runs. The ℓ_0 regularization method barely sparsifies the network. On the other hand, CGES obtains the lowest mean test error with the largest mean weight sparsity, but its mean neuron sparsity is not as high as group lasso, sparse group lasso, and sparse group l_0 asso. The largest mean neuron sparsity is attained by sparse group lasso, but its corresponding test error is worse than the other methods. Sparse group l_0 asso attains comparable mean weight and neuron sparsity as group lasso and sparse group lasso but with lower test error. Therefore, the proposed regularization is able to balance accuracy with both weight and neuron sparsity better than the baseline methods.

3.2 CIFAR Classification

CIFAR 10/100 is a dataset that has 10/100 classes split into 50k training images and 10k test images. The dataset is trained on wide residual networks, specifically WRN-28-10. WRN-28-10 has approximately 36,500,000 weight parameters and 10,736 neurons. The network is optimized using stochastic gradient descent with initial

learning rate 0.1. After every 60 epochs, learning rate decays by a factor of 0.2. Strictly the same type of regularization is applied to the weights of the hidden layer where dropout is utilized in the residual block. We vary the regularization parameter $\lambda = \alpha/50000$ by training the model on $\alpha \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$. For sparse group l_0 asso, we set $\beta = 25\alpha/50000$ initially and it increases by a factor of 1.25 for every 20 epochs. The network is trained for 200 epochs across 5 runs. Note that we exclude ℓ_0 regularization by Louizos *et al.*[26] because we found the method to be unstable when $\alpha \geq 0.1$. The results are shown in Figures 1 and 2 for CIFAR 10 and CIFAR 100, respectively.

According to Figure 1, CGES outperforms the other methods when $\alpha = 0.01$ for both sparsity and test error. However, sparsity levels stabilize after when $\alpha = 0.1$. Sparse group lasso attains the highest mean weight and neuron sparsity when $\alpha \geq 0.01$. Group lasso and sparse group l_0 asso have comparable mean weight and neuron sparsity levels, but sparse group l_0 asso outperforms the other methods in terms of test error when $\alpha \geq 0.05$.

Figure 2 shows that the results for CIFAR 100 are similar to the results for CIFAR 10. CGES has better weight sparsity when $\alpha \leq 0.1$, but it has the least neuron sparsity when $\alpha \geq 0.2$. Sparsity levels for CGES appear to stabilize after when $\alpha = 0.1$. Sparse group lasso attains the highest mean weight and neuron sparsity for $\alpha \geq 0.2$. The proposed method sparse group l_0 asso has comparable weight and neuron sparsity as group lasso, but it has the lowest mean test error when $\alpha \in \{0.05, 0.1, 0.2\}$.

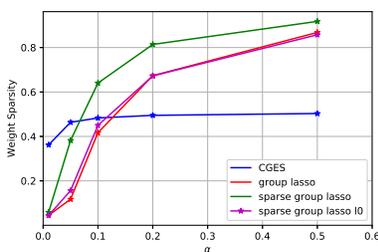
Overall, the results demonstrate that sparse group l_0 asso maintains superior test accuracy with similar sparsity levels as group lasso when trading accuracy for sparsity as α increases.

4 Conclusion and Future Work

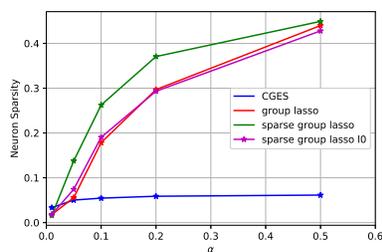
In this work, we propose sparse group l_0 asso, a new variant of sparse group lasso where the ℓ_1 norm on the weight parameters is replaced with the ℓ_0 norm. We develop a new algorithm to optimize loss functions regularized with sparse group l_0 asso for DNNs in order to attain a sparse network with competitive accuracy. We compare our method with various baseline methods on MNIST and CIFAR 10/100 on different CNNs. The experimental re-

Method	Mean Weight Sparsity (%) [Std (%)]	Mean Neuron Sparsity (%) [Std (%)]	Test Error (%) [Std (%)]
ℓ_0 [26]	0.02 [< 0.01]	0 [0]	0.69 [0.02]
CGES	94.12 [0.26]	39.33 [1.61]	0.65 [0.04]
group lasso	88.38 [0.49]	69.39 [0.64]	0.76 [0.02]
sparse group lasso	93.50 [0.13]	73.52 [0.49]	0.77 [0.03]
sparse group ℓ_0 asso (proposed)	89.27 [0.46]	68.25 [0.49]	0.67 [0.02]

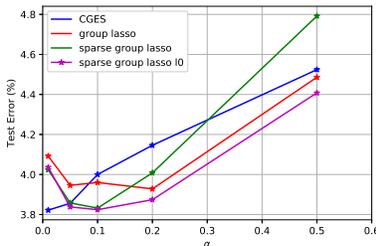
Table 1: Comparison of the baseline methods and sparse group ℓ_0 asso regularization method on LeNet-5-Caffe trained on MNIST. Mean weight sparsity, mean neuron sparsity, and mean test error across 5 runs after 200 epochs are shown. $N = 60000$, the number of training points.



(a) Mean weight sparsity



(b) Mean neuron sparsity



(c) Mean test error

Figure 1: Mean results for CIFAR-10 on WRN-28-10 across 5 runs when varying the regularization parameter $\lambda = \alpha/50000$ when $\alpha \in \{0.01, 0.02, 0.1, 0.2, 0.5\}$.

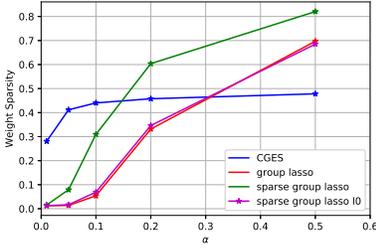
sults demonstrate that in general, sparse group ℓ_0 asso attains similar weight and neuron sparsity as group lasso while maintaining competitive accuracy.

For our future work, we plan to extend our proposed variant to other nonconvex penalties, such as $\ell_1 - \ell_2$, transformed ℓ_1 , and $\ell_{1/2}$. We will examine these nonconvex sparse group lasso methods on various experiments, not only on MNIST and CIFAR 10/100 but also on Tiny Imagenet and Street View House Number trained on different networks such as MobileNetv2 [33]. In addition,

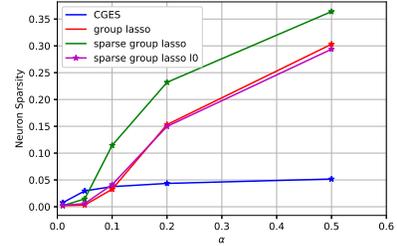
we might investigate in developing an alternating direction method of multipliers algorithm [5] as an alternative to the algorithm developed in this paper.

5 Acknowledgement

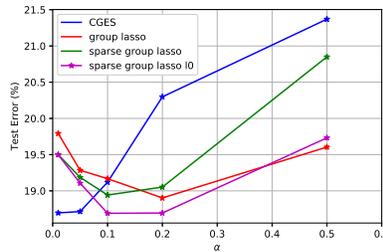
The work was partially supported by NSF grant IIS-1632935, DMS-1854434, a Qualcomm Faculty Award, and Qualcomm AI Research.



(a) Mean weight sparsity



(b) Mean neuron sparsity



(c) Mean test error

Figure 2: Mean results for CIFAR-100 on WRN-28-10 across 5 runs when varying the regularization parameter $\lambda = \alpha/50000$ when $\alpha \in \{0.01, 0.02, 0.1, 0.2, 0.5\}$.

References

- [1] A. Aghasi, A. Abdi, N. Nguyen, and J. Romberg. Net-trim: Convex pruning of deep neural networks with performance guarantee. In *Advances in Neural Information Processing Systems*, pages 3177–3186, 2017.
- [2] J. M. Alvarez and M. Salzmann. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems*, pages 2270–2278, 2016.
- [3] J. Ba and R. Caruana. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662, 2014.
- [4] C. Bao, B. Dong, L. Hou, Z. Shen, X. Zhang, and X. Zhang. Image restoration by minimizing zero norm of wavelet frame coefficients. *Inverse problems*, 32(11):115004, 2016.
- [5] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [6] R. H. Chan, T. F. Chan, L. Shen, and Z. Shen. Wavelet algorithms for high-resolution image reconstruction. *SIAM Journal on Scientific Computing*, 24(4):1408–1432, 2003.
- [7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
- [8] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in neural information processing systems*, pages 1269–1277, 2014.
- [9] T. Dinh and J. Xin. Convergence of a relaxed variable splitting method for learning sparse neural networks via ℓ_1, ℓ_0 , and transformed- ℓ_1 penalties. *arXiv preprint arXiv:1812.05719*, 2018.
- [10] B. Dong and Y. Zhang. An efficient algorithm for l0 minimization in wavelet frame based image restoration. *Journal of Scientific Computing*, 54(2-3):350–368, 2013.
- [11] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

- [12] S. Han, H. Mao, and W. J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*, 2015.
- [13] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. In *Advances in neural information processing systems*, pages 1135–1143, 2015.
- [14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [15] H. Hu, R. Peng, Y.-W. Tai, and C.-K. Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.
- [16] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7310–7311, 2017.
- [17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.
- [18] X. Jin, X. Yuan, J. Feng, and S. Yan. Training skinny deep neural networks with iterative hard thresholding methods. *arXiv preprint arXiv:1607.05423*, 2016.
- [19] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [20] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- [21] 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.
- [22] A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.
- [23] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [24] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.
- [25] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.
- [26] C. Louizos, M. Welling, and D. P. Kingma. Learning sparse neural networks through l_0 regularization. *CoRR*, abs/1712.01312, 2017.
- [27] Z. Lu and Y. Zhang. Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478, 2013.
- [28] R. Ma, J. Miao, L. Niu, and P. Zhang. Transformed ℓ_1 regularization for learning sparse deep neural networks. *arXiv preprint arXiv:1901.01021*, 2019.
- [29] D. Molchanov, A. Ashukha, and D. Vetrov. Variational dropout sparsifies deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning—Volume 70*, pages 2498–2507. JMLR. org, 2017.
- [30] J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
- [31] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *bmvc*, page 6, 2015.
- [32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.
- [33] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.
- [34] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.
- [35] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.
- [36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.
- [37] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
- [38] J. Trzasko, A. Manduca, and E. Borisch. Sparse mri reconstruction via multiscale l_0 -continuation. In *2007 IEEE/SP 14th Workshop on Statistical Signal Processing*, pages 176–180. IEEE, 2007.
- [39] K. Ullrich, E. Meeds, and M. Welling. Soft weight-sharing for neural network compression. *stat*, 1050:9, 2017.

- [40] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li. Learning structured sparsity in deep neural networks. In *Advances in neural information processing systems*, pages 2074–2082, 2016.
- [41] J. Yoon and S. J. Hwang. Combined group and exclusive sparsity for deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3958–3966. JMLR. org, 2017.
- [42] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.
- [43] X.-T. Yuan, P. Li, and T. Zhang. Gradient hard thresholding pursuit. *Journal of Machine Learning Research*, 18:166–1, 2017.
- [44] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.
- [45] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.
- [46] Y. Zhang, B. Dong, and Z. Lu. l0 minimization for wavelet frame based image restoration. *Mathematics of Computation*, 82(282):995–1015, 2013.
- [47] Y. Zhou, R. Jin, and S. C.-H. Hoi. Exclusive lasso for multi-task feature selection. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 988–995, 2010.
- [48] Z. Zhuang, M. Tan, B. Zhuang, J. Liu, Y. Guo, Q. Wu, J. Huang, and J. Zhu. Discrimination-aware channel pruning for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 875–886, 2018.