LEARNING TO SHARE: SIMULTANEOUS PARAMETER TYING AND SPARSIFICATION IN DEEP LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks (DNNs) usually contain millions, maybe billions, of parameters/weights, making both storage and computation very expensive. Moreover, although this complexity allows DNNs to learn complex mappings, it may also allow them to over-fit, which has motivated the use of regularization methods, *e.g.*, by encouraging weight sparsity. Another well-known approach for controlling the complexity of DNNs is parameter sharing/tying, where certain sets of weights are forced to share a common value. Some forms of weight sharing are hard-wired to express certain invariances, with a notable example being the shift-invariance of convolutional layers. However, there may be other groups of weights that may be tied together during the learning process, thus further reducing the complexity of the network. In this paper, we adopt a recently proposed sparsity-inducing regularizer, named GrOWL (*group ordered weighted* ℓ_1), which encourages sparsity and, simultaneously, learns which groups of parameters should share a common value. GrOWL has been proven effective in linear regression, being able to identify and cope with strongly correlated covariates. Unlike standard sparsity-inducing regularizers (*e.g.*, ℓ_1 a.k.a. Lasso), GrOWL not only eliminates unimportant neurons by setting all the corresponding weights to zero, but also explicitly identifies strongly correlated neurons by tying the corresponding weights to a common value. This ability of GrOWL motivates the following two-stage procedure: (i) use GrOWL regularization in the training process to simultaneously identify significant neurons and groups of parameter that should be tied together; (ii) retrain the network, enforcing the structure that was unveiled in the previous phase, *i.e.*, keeping only the significant neurons and enforcing the learned tying structure. We evaluate the proposed approach on several benchmark datasets, showing that it can dramatically compress the network with slight or even no loss on generalization performance.

1 INTRODUCTION

Deep neural networks (DNNs) have recently revolutionized machine learning by dramatically advancing the state-of-the-art performance in many applications, ranging from speech and image recognition to playing video games (Goodfellow et al., 2016). A typical DNN consists of a sequence of concatenated layers of functions, potentially involving millions or billions of parameters; by using extremely large training datasets, these architectures are able to learn extremely complex non-linear mappings, features, and dependencies.

A large amount of research has been devoted to the use of regularization in DNN learning (Goodfellow et al., 2016), aiming at reducing the generalization error. It has been shown that the parametrization of many DNNs is very redundant, in the sense that a large fraction of the parameters can be predicted from the remaining ones, with no accuracy loss (Denil et al., 2013). Several regularization methods have been proposed to tackle the potential over-fitting caused by this redundancy. Arguably, the earliest and simplest choice is the classical ℓ_2 norm, known as *weight decay* in the early neural networks literature (Rumelhart et al., 1988), and as *ridge regression* in statistics. In the past two decades, sparsity-inducing regularization based on the ℓ_1 norm, also known as Lasso (Tibshirani, 1996), and variants thereof, became standard tools in statistics and machine learning, including in deep learning (Goodfellow et al., 2016). In recent work, Scardapane et al. (2017) use group-Lasso (a variant of Lasso that assumes that the parameters are organized in groups and encourages sparsity

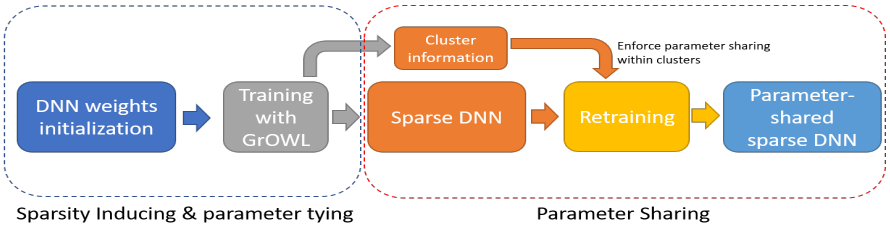

Figure 1: A DNN is first trained with GrOWL regularizer to simultaneously identify the sparse but significant connectivities and the cluster information of the selected connectivities. We then retrain the neuron network only in terms of the selected connectivities while enforcing parameter sharing within each cluster.

at the group level (Yuan & Lin, 2006)) in deep learning. One of the effects of Lasso or group-Lasso regularization in learning a DNN is that many of the parameters may become exactly zero, thus reducing the amount of memory needed to store the model, and lowering the computational cost of applying it.

It has been pointed out by several authors that a major drawback of Lasso (or group-Lasso) regularization is that in the presence of groups of highly correlated covariates/features, it tends to select only one or an arbitrary convex combination of features from each group (Bondell & Reich, 2008; Bühlmann et al., 2013; Figueiredo & Nowak, 2016; Oswal et al., 2016; Zou & Hastie, 2005). Moreover, the learning process tends to be unstable, in the sense that subsets of parameters that end up being selected may change dramatically with minor changes in the data or algorithmic procedure. In DNNs, it is almost unavoidable to encounter correlated features, not only due to the high dimensionality of the input to each layer, but also because neurons tend to co-adapt, yielding strongly correlated features that are passed as input to the subsequent layer (Srivastava et al., 2014).

In this work, we propose using, as a regularizer for learning DNNs, the group version of the *ordered weighted* ℓ_1 (OWL) norm Figueiredo & Nowak (2016), termed group-OWL (GrOWL), which was recently proposed by Oswal et al. (2016). In a linear regression context, GrOWL regularization has been shown to avoid the above mentioned deficiency of group-Lasso regularization. In addition to being a sparsity-inducing regularizer, GrOWL is able to explicitly identify groups of correlated features and set the corresponding parameters/weights to be very close or exactly equal to each other, thus taking advantage of correlated features, rather than being negatively affected by them. In deep learning parlance, this corresponds to adaptive *parameter sharing/tying*, where instead of having to define *a priori* which sets of parameters are forced to share a common value, these sets are learned during the training process. We exploit this ability of GrOWL regularization to encourage parameter sparsity and group-clustering in a two-stage procedure depicted in Fig. 1: we first use GrOWL to identify the significant parameters/weights of the network and, simultaneously, the correlated cluster information of the selected features; then, we retrain the network only in terms of the selected features, while enforcing the weights within the same cluster to share a common value.

The experiments reported in this paper confirm that using GrOWL regularization in learning DNNs does encourage sparsity, by eliminating small weights, and yields parameter sharing, by tying groups of weights to have the same absolute values. We test the proposed approach on two benchmark datasets, MNIST and CIFAR-10, comparing its performance with weight decay and group-Lasso regularization, and exploring the accuracy-memory trade-off. Our results indicate that GrOWL is able to reduce the number of free parameters in the network without degrading the accuracy, as compared to other regularization approaches.

2 Related Work

In order to relieve the burden on both required memory and data for training and storing DNNs, a substantial amount of work has focused on reducing the number of free parameters that need to be estimated, namely by enforcing weight sharing. The classical instance of parameter sharing is found in the convolutional layers of DNNs (Goodfellow et al., 2016). In fact, the use of weight-sharing

as a simplifying technique for neural networks can be traced back to more than three decades ago (LeCun, 1987; Rumelhart & McClelland, 1986).

Recently, with the goal of reducing the storage and communication costs of DNNs, there has been a surge of interest in compressing the description of these models. Various methods have been proposed that approximate or quantize the learned weights after the training process. Denton et al. (2014) have shown that, in some cases, it is possible to replace the original weight matrix with a low-rank approximation. Alternatively, Aghasi et al. (2016) propose retraining the network layer by layer, keeping the layer inputs and outputs close to the originally trained model, while seeking a sparse transform matrix, whereas Gong et al. (2014) propose the use of vector quantization to compress the parameters of DNNs.

Network pruning is another relevant line of work. Early work by LeCun et al. (1989) and Hassibi et al. (1993) uses the information provided by Hessian of the loss function to remove less important weights of the network, which requires expensive computation of second order derivatives. Recently, Han et al. (2016) reduce the number of parameters by an order of magnitude using an iterative algorithm that alternates between learning the parameters and removing those below a certain threshold. Li et al. (2016) propose to prune filters, which seeks sparsity with respect to neurons, rather than connectivities; that approach relieves the burden on requiring sparse libraries or special hardware to deploy the network. All these methods either require multiple training/retraining iterations or a careful choice of pruning thresholds.

There is a large body of work on using sparsity-inducing regularization for training DNNs. For example, Collins & Kohli (2014) exploit ℓ_1 and ℓ_0 regularization to encourage weight sparsity; however, the sparsity level achieved is typically modest, making this approach not competitive in terms of DNN compression. Group-Lasso has also been used in training DNNs, which allows seeking sparsity in terms of neurons (Scardapane et al., 2017; Alvarez & Salzmann, 2016; Zhou et al., 2016; Murray & Chiang, 2015) or other structures, *e.g.*, filters, channels, filter shapes, and layer depth (Wen et al., 2016). However, as mentioned above, both Lasso and group-Lasso can fail at selecting important features if they are strongly correlated with each other. In Section 4 and Appendix B, we will further illustrate this fact on both synthetic and real data.

A recent stream of work has focused on using further parameter sharing in convolutional DNNs. By tying weights in an appropriate way, Dieleman et al. (2016) obtain a convolutional DNN with rotation invariance. On the task of analyzing positions in the game *Go*, Clark & Storkey (2015) showed improved performance by constraining features to be invariant to reflections along the x-axis, y-axis, and diagonal-axis. Finally, Chen et al. (2015) used a hash function to randomly group the weights such that those in a hash bucket share the same value. In contrast, with GrOWL regularization, we aim to learn weight sharing from the data itself, rather than specifying it *a priori*.

Dropout-type methods have been proposed to fight over-fitting and are very popular, arguably due to their simplicity of implementation. For each training example, either activations (Srivastava et al., 2014) or connections (Wan et al., 2013) are randomly dropped and the error is then back-propagated only through the remaining ones. It has been shown that dropout effectively reduces over-fitting and prevents different neurons from co-adapting.

Decorrelation is another popular technique in deep learning pipelines (Bengio & Bergstra, 2009; Cogswell et al., 2015; Rodríguez et al., 2016). Unlike sparsity-inducing regularizers, decorrelation methods try to make full use of the model's capacity by decorrelating the neurons. Although dropout and decorrelation can reduce over-fitting, they do not compress the network, hence do not address the issue of high memory cost. It should also be mentioned that our proposal can be seen as complementary to dropout and decorrelation, in the following sense: whereas dropout and decorrelation can reduce co-adaption of nodes during training, GrOWL regularization copes with co-adaptation by tying together the weights associated to co-adapted nodes.

3 Group-OWL Regularization for Deep Learning

3.1 The Group-OWL Norm

We start by recalling the definition of the group-OWL (GrOWL) regularizer and very briefly reviewing some of its relevant properties (Oswal et al., 2016).

Definition 1. *Given a matrix $W \in \mathbb{R}^{n\times m}$, let $\|w\|_{[i]}$. denote the row of W with the i-th largest ℓ_2 norm. Let $\lambda \in \mathbb{R}^n_+$, with $0 < \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$. The GrOWL regularizer (which is a norm) $\Omega_\lambda : \mathbb{R}^{n\times m} \to \mathbb{R}$ is defined as*

$$\Omega_\lambda(W) = \sum_{i=1}^{n} \lambda_i \|w\|_{[i]}. \tag{1}$$

This is a group version of the OWL regularizer (Figueiredo & Nowak, 2016), also known as WSL1 (*weighted sorted ℓ1* (Zeng & Figueiredo, 2014)) and SLOPE (Bogdan et al., 2015), where the groups are the rows of its matrix argument. It is clear that GrOWL includes group-Lasso as a special case when $\lambda_1 = \lambda_n$. As a regularizer for multiple/multi-task linear regression, each row of W contains the regression coefficients of a given feature, for the m tasks. It has been shown that by adding the GrOWL regularizer to a standard squared error loss function, the resulting estimate of W has the following property: rows associated with highly correlated covariates are very close or even exactly equal to each other (Oswal et al., 2016). In the linear case, GrOWL encourages correlated features to form predictive clusters corresponding to the groups of rows that are nearly or exactly equal. The rationale underlying this paper is that when used as a regularizer for DNN learning, GrOWL will induce both sparsity and parameters tying, as illustrated in Fig. 2 and explained below in greater detail.

3.2 Layer-Wise GrOWL Regularization For Feedforward Neural Networks

A typical feed-forward DNN with L layers can be treated as a function f of the following form:

$$f(x,\theta) \equiv h_L = f_L\left(h_{L-1}W_L + b_L\right), \quad h_{L-1} = f_{L-1}\left(h_{L-2}W_{L-1} + b_{L-1}\right), \quad ..., h_1 = f_1\left(xW_1 + b_1\right)$$

$\theta = \left(W_1, b_1, \ldots, W_L, b_L\right)$ denotes the set of parameters of the network, and each f_i is a component-wise nonlinear activation function, with the *rectified linear unit* (ReLU), the *sigmoid*, and the *hyperbolic tangent* being common choices for this function (Goodfellow et al., 2016).

Given labelled data $\mathcal{D} = \left((x^{(1)}, y^{(1)}), ..., (x^{(m)}, y^{(m)})\right)$, DNN learning may be formalized as an optimization problem,

$$\min_\theta \mathcal{L}(\theta) + \mathcal{R}(\theta), \quad \text{where } \mathcal{L}(\theta) = \sum_{i=1}^{m} L\Big(y^{(i)}, f\big(x^{(i)}, \theta\big)\Big)\,, \tag{2}$$

and $L\big(y, \hat{y}\big)$ is the loss incurred when the DNN predicts $\hat{y}$ for y, and $\mathcal{R}$ is a regularizer. The regularizer herein adopted is a sum of GrOWL penalties, each applied to each layer of the neural network, *i.e.,*

$$\mathcal{R}(\theta) = \sum_{l=1}^{L} \Omega_{\lambda^{(l)}}\big(W_l\big), \quad \lambda^{(l)} \in \mathbb{R}^{N_{l-1}}_+\,, \tag{3}$$

where N_l denotes the number of neurons in the l-th layer and $0 < \lambda^{(l)}_1 \geq \lambda^{(l)}_2 \geq \cdots \geq \lambda^{(l)}_{N_{l-1}} \geq 0$. Notice that $\mathcal{R}(\theta)$ does not depend on $b_1, ..., b_L$, which means these parameters are not regularized, as is common practice.

As indicated in Eq. (3), the number of groups in each GrOWL regularizer equals the number of neurons in the previous layer, *i.e.,* $\lambda^{(l)} \in \mathbb{R}^{N_{l-1}}$. In other words, *we treat the weights associated with each input feature as a group*. For fully connected layers, where $W_l \in \mathbb{R}^{N_{l-1}\times N_l}$, each group is a row of the weight matrix. When considering convolutional layers, where $W_l \in \mathbb{R}^{F_w \times F_h \times N_{l-1} \times N_l}$, with F_w and F_h denoting the width and height, respectively, of each filter, we first reshape W_l to a 2-dimensional array, *i.e.,* $W_l \to W_l^{2D}$, where $W_l^{2D} \in \mathbb{R}^{N_{l-1}\times(F_w F_h N_l)}$, and then apply GrOWL on the reshaped matrix. That is, if the l-th layer is convolutional, then

$$\mathcal{R}(W_l) = \Omega_{\lambda^{(l)}}\left(W_l^{2D}\right). \tag{4}$$

Each row of W_l^{2D} represents the operation on an input channel. The rationale to apply the GrOWL regularizer to each row of the reshaped weight matrix is that GrOWL can select the relevant features of the network, while encouraging the coefficient rows of each layer associated with strongly correlated features from the previous layer to be nearly or exactly equal, as depicted in Fig. 2. The

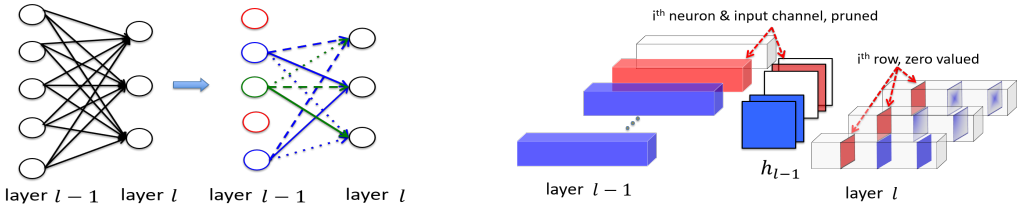

Figure 2: GrOWL's regularization effect on deep nets. **Fully connected layers (Left):** for each layer l GrOWL clusters the input features from the previous layer $l-1$ into different groups, *e.g.,* blue and green. *Within each neuron* of layer l, the weights associated with the input features from the same cluster (input arrows marked with the same color) share the same parameter value. For layer $l-1$, the neurons, corresponding with the zero valued rows of W_h, have zero inputs to the current layer and hence get removed automatically. **Convolutional layers (right):** each group (row) is predefined as the filters that associates with the same input channel, parameter sharing is enforced among the filters within each neuron that corresponds with the same cluster (marked as blue with different effects) of input channels.

goal is to significantly reduce the complexity by: **(i)** pruning unimportant neurons of the previous layer that correspond to zero rows of the (reshaped) weight matrix of the current layer; **(ii)** grouping the rows associated with highly correlated features of the previous layer, thus encouraging the coefficient rows in each of these groups to be very close to each other. As a consequence, in the retraining process, we could further compress the neural network by enforcing the parameters *within each neuron* that belong to the same cluster to share same values.

In the work of Alvarez & Salzmann (2016), each group is predefined as the set of parameters associated to a neuron, and group-Lasso regularization is applied to seek group sparsity, which corresponds to zeroing out redundant neurons of each layer. In contrast, we treat the filters corresponding with the same input channel as a group, and GrOWL is applied to prune the redundant groups and thus *remove the associated unimportant neurons of previous layer*, while grouping associated parameters of the current layer that correspond with highly correlated input features to different clusters. Moreover, shown in Section 4, group-Lasso can fail at selecting all relevant features of previous layers, and for the selected ones the corresponding coefficient groups are quite dissimilar from each other, making it impossible to further compress the DNN by enforcing parameter tying.

3.3 Proximal Gradient Algorithm

To solve (2), we use a *proximal gradient algorithm* (Bauschke & Combettes, 2011), which has the following general form: at the t-th iteration, the parameter estimates are updated according to

$$\theta^{(t+1)} = \text{prox}_{\eta\mathcal{R}}\left(\theta^{(t)} - \eta\nabla_\theta\mathcal{L}(\theta^t)\right) \tag{5}$$

where, for some convex function Q, prox_Q denotes its proximity operator (Bauschke & Combettes, 2011), which is defined as $\text{prox}_Q(\xi) = \arg\min_\theta Q(\nu) + \frac{1}{2}\|\nu-\xi\|_2^2$. In the previous expression, $\|\nu-\xi\|_2^2$ denotes the sum of the squares of the differences between the corresponding components of ν and ξ, regardless of their organization (in our case, a collection of matrices and vectors).

Since $\mathcal{R}(\theta)$, as defined in (3), is separable across the weight matrices of different layers and zero for $b_1, ..., b_L$, the corresponding proximity operator is also separable, thus

$$W_l^{(t+1)} = \text{prox}_{\eta\Omega_\lambda^{(l)}}\left(W_l^{(t)} - \eta\,\nabla_{W_l}\mathcal{L}(\theta^{(t)})\right), \quad \text{for } l = 1,\ldots,L \tag{6}$$

$$b_l^{(t+1)} = b_l^{(t)} - \eta\,\nabla_{b_l}\mathcal{L}(\theta^{(t)}) \quad \text{for } \; l = 1, ..., L. \tag{7}$$

It was proved by Oswal et al. (2016) that the proximity operator of the GrOWL norm can be computed as follows. For some matrix $V \in \mathbb{R}^{N\times M}$, let $U = \text{prox}_{\Omega_\lambda}(V)$, and v_i and u_i denote the corresponding i-th rows. Then,

$$u_i = \frac{v_i}{\|v_i\|}\left(\text{prox}_{\Omega_\lambda}(\tilde{v})\right)_i \tag{8}$$

where $\tilde{v} = [\|v_1\|, \|v_2\|, \cdots, \|v_N\|]$. The proximal operator $\text{prox}_{\Omega_{\lambda^{(l)}}}$ for a vector in $\mathbb{R}^N$ (in which case GrOWL coincides with OWL) can be computed with $O(n \log n)$ complexity, where the core computation is the so-called *pool adjacent violators algorithm* (PAVA (de Leeuw et al., 2009)) for isotonic regression. We provide one of the existing algorithms in Appendix A; for details of the algorithm, the reader is referred to the work of Bogdan et al. (2015) and Zeng & Figueiredo (2015). In this paper, we apply the proximal gradient descent per epoch, which generally yields better performance. The training method is summarized in Algorithm 1.

Algorithm 1

Input: parameters of the OWL regularizers $\lambda^{(l)}, ..., \lambda^{(L)}$, learning rate η
for each epoch T **do**
  for each iteration t in epoch T **do**
    Update the parameter via backpropagation (BP)
  end for
  Apply proximity operator via (6)
end for

3.4 Implementation Details

3.4.1 Setting the GrOWL Weights

GrOWL is a family of regularizers, with different variants obtained by choosing different weight sequences $\lambda_1, \ldots, \lambda_n$. In this paper, we propose the following choice:

$$\lambda_i = \begin{cases} \Lambda_1 + (p - i + 1)\Lambda_2, & \text{for } i = 1, ..., p, \\ \Lambda_1, & \text{for } i = p + 1, ..., n, \end{cases} \tag{9}$$

where $p \in \{1, ...n\}$ is a parameter. The first p weights follow a linear decay, while the remaining ones are all equal to Λ_1. Notice that, if $p = n$, the above setting is equivalent to OSCAR (Bondell & Reich, 2008). Roughly speaking, Λ_1 controls the sparsifying strength of the regularizer, while Λ_2 controls the clustering property (correlation identification ability) of GrOWL (Oswal et al., 2016). Moreover, by setting the weights to a common constant beyond index p means that clustering is only encouraged among the p largest coefficients, *i.e.,* only among relevant coefficient groups.

Finding adequate choices for p, Λ_1, and Λ_2 is crucial for jointly selecting the relevant features of the network and identifying the underlying correlations among them. In practice, we find that with properly chosen p, GrOWL is able to find more correlations of the underlying data than OSCAR. We explore different choices of p in Appendix B.

3.4.2 Parameter Tying

After the initial training phase, at each layer l, rows of W_l that corresponds to highly correlated outputs of layer $l - 1$ have been made similar or even exactly equal. To further compress the DNN, we force rows that are close to each other to be identical. We first group the rows into different clusters [1] according to the *pairwise similarity* metric

$$\mathcal{S}_l(i, j) = \frac{W_{l,i}^T W_{l,j}}{\max\left(\|W_{l,i}\|_2^2, \|W_{l,j}\|_2^2\right)} \in [-1, 1], \tag{10}$$

where $W_{l,i}$ and $W_{l,j}$ denote the i-th and j-th rows of W_l, respectively.

With the cluster information obtained by using GrOWL, we enforce parameter sharing for the rows that belong to a same cluster by replacing their values with the averages (centroid) of the rows in that cluster. In the following retraining process , let $\mathcal{G}_k^{(l)}$ denote the k-th cluster of the l-th layer, then

[1] In this paper, we use the built-in *affinity propagation* method of the scikit-learn package (Buitinck et al., 2013).

centroid $g_k^{(l)}$ of this cluster is updated via

$$\frac{\partial \mathcal{L}}{\partial g_k^{(l)}} = \frac{1}{\left|\mathcal{G}_k^{(l)}\right|} \sum_{W_{l,i} \in \mathcal{G}_k^{(l)}} \frac{\partial \mathcal{L}}{\partial W_{l,i}}. \tag{11}$$

4 Numerical Results

We assess the performance of the proposed method on two benchmark datasets: MNIST and CIFAR-10. We consider two different networks and compare GrOWL with group-Lasso and weight decay, in terms of the compression vs accuracy trade-off. For fair comparison, the training-retraining pipeline is used with the different regularizers. After the initial training phase, the rows that are close to each other are clustered together and forced to share common values in the retraining phase. To evaluate the effect of the different regularizers, we use the following quantities: sparsity $= \frac{\#\text{zero params}}{\#\text{ total params}}$; compression rate $= \frac{\#\text{ total params}}{\#\text{ unique params}}$; parameter sharing $= \frac{\#\text{ nonzero params}}{\#\text{ unique params}}$.

4.1 Fully Connected Neural Network on MNIST

The MNIST dataset contains 28×28 (784) pixels images of handwritten digits (0, 1, ..., 9), each located at the center of the image. Fig 3 (a) shows the (784×784) correlation matrix of the dataset (the margins are zero due to the redundant background of the images). We use a network with a single fully connected layer of 300 hidden units. The network is trained for 200 epochs and then retrained for an additional 100 epochs, both with momentum. The initial learning rate is set to 10^{-4} for both the training and retraining phases, and it is reduced by a factor of 0.96 every 10 epochs. The hyper-parameters are selected through grid search

The pairwise similarities (see Eq. (10)) between the rows of the weight matrices learned with the different regularizers tested are shown in Fig. 3 (b,c,d). For the same level of sparsity (∼47%), GrOWL identifies more correlations than group-Lasso, and the similarity pattern in Fig. 3 (b) is very close to that of the data (Fig. 3(a)). On the other hand, weight decay also identifies correlations between parameter rows, but it does not induce a sparse model. Moreover, as shown in Table 1, GrOWL obtains a higher level of parameter sharing than weight decay.

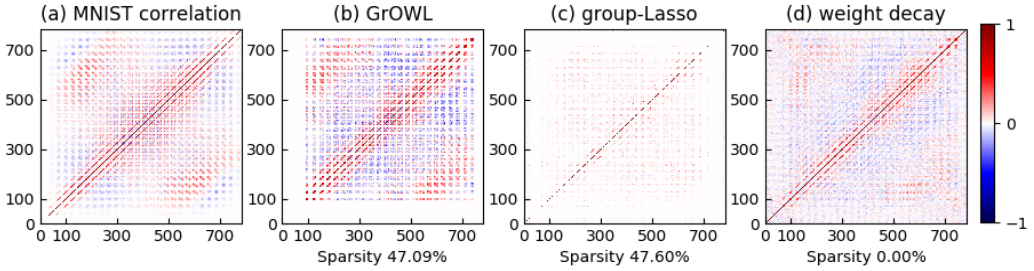

Figure 3: MNIST: comparison of the data correlation and the similarity maps (Eq (10)) of the weight matrices obtained by training the neural network with GrOWL, group-Lasso and weight decay.

The compression vs accuracy trade-off of the different regularizers is summarized in Table 1, which shows that GrOWL achieves the highest compression and accuracy. More specifically, when GrOWL and group-Lasso (1) achieve the same sparsity level after the initial training phase, GrOWL compresses the network ∼ 3 times more than group-Lasso, due to the significant amount of correlation it identifies. On the other hand, when group-Lasso (2) reaches the same level of compression as GrOWL, *i.e.,* the same amount of free parameters, its accuracy is about 1.6% lower than GrOWL. A possible explanation is that group-Lasso suffers from randomly selecting a subset of correlated features. We illustrate this effect in Fig. 4, which plots the indices of nonzero rows, showing that GrOWL stably selects relevant features while group-Lasso does not. The mean ratios of changed in-

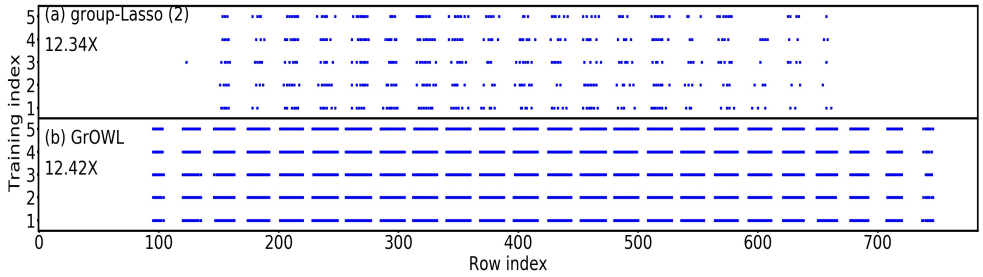

Figure 4: MNIST: sparsity pattern of the trained fully connected layer, for 5 training runs, using group-Lasso (2) and GrOWL. For similar compression rates (12.34 and 12.42), GrOWL is much more stable.

dices [2] are 39.75% and 1.56% for group-Lasso and GrOWL, respectively. This suggests, combined with Table 1, that stably selecting all of the relevant features may be important for achieving a high generalization performance.

Table 1: Sparsity, parameter sharing, and compression rate results on MNIST. Group-Lasso (1) and group-Lasso (2) differ in regularization strength. All values are averaged over 5 trainings.

Regularizers	**Sparsity**	**Parameter Sharing**	**Compression Rate**	**Accuracy**
weight decay	0.00%	1.30	1.30X	96.92%
group-Lasso (1)	47.60%	1.07	2.04X	97.97%
group-Lasso (2)	88.51%	1.43	12.34X	97.07%
GrOWL	47.09%	**6.57**	**12.42X**	**98.12%**

4.2 VGG-16 ON CIFAR-10

To evaluate the proposed method on large DNNs, we consider a VGG-like (Simonyan & Zisserman, 2014) architecture proposed by Sergey (2015) on the CIFAR-10 dataset. The network is summarized in Table 2; comparing with the original VGG of Simonyan & Zisserman (2014), their fully connected layers are replaced with two small ones with 14.3% of the total parameters. A batch normalization layer is added after each convolutional layer and the first fully connected layer. Unlike Sergey (2015), we don't use dropout. We first train under different regularizers over 160 epochs, then retrain for another 40 epochs, using the learning rate decay scheme described by He et al. (2016): the initial rates for the training and retraining phases are set to 0.01 and 0.001, respectively; the learning rate is multiplied by 0.1 every 60 epochs of the training phase, and every 20 epochs of the retraining phase. For GrOWL, we set $p = 0.1\,n$ (see Eq. (9)) for all layers, where n denotes the number of rows of each layer.

The results are summarized in Table 2. In the experiments, we found that it is quite hard to encourage parameter tying in the first 7 convolutional layers. The reason may be that the filters of these first 7 convolutional layers have comparatively large feature maps (from 32×32 to 8×8), which are only loosely correlated. We illustrate this reasoning in Fig. 5, showing the cosine similarity between the vectorized output channels of layers 1, 6, 10, and 11, at the end of the training phase; it can be seen that the outputs of layers 10 and 11 have many more significant similarities than that of layer 6. Although the output channels of layer 1 also have certain similarities, as seen in Table 2, neither GrOWL nor weight decay tends to tie the associated weights. This may mean that the network is maintaining the diversity of the inputs in the first few convolutional layers.

[2]The mean ratio of changed indices is defined as: $\frac{1}{n}\sum_{k=1}^{n} \|I_k - \bar{I}\|_1/\|\bar{I}\|_0$, where n is the number of experiments, I_k is the index vector of kth experiment, and $\bar{I} = \frac{1}{n}\sum_{k=1}^{n} I_k$ is the mean index vector.

Table 2: Sparsity (**S1**) and Parameter Sharing (**S2**) of VGG-16 on CIFAR-10. Layers preceded by * are regularized. Input channel numbers in are equal to the number of rows in the corresponding reshaped parameter matrices.

Layers	Output $w \times h$	#Channels in&out	#Params	GrOWL (**S1**, **S2**)	Group Lasso (**S1**, **S2**)	Weight Decay (**S1**, **S2**)
conv1	32×32	3, 64	1.7E+03	0.00%, 1.00	0.00% 1.00	0.00%, 1.00
*conv2	32×32	*64*, 64	3.7E+04	18.75%, 1.04	25.00% 1.00	$0,00\%$, 1.00
*conv3	16×16	*64*, 128	7.4E+04	29.69%, 1.00	29.69% 1.00	0.00%, 1.00
*conv4	16×16	*128*, 128	1.5E+05	35.94%, 1.00	28.16% 1.00	0.00%, 1.00
*conv5	8×8	*128*, 128	2.9E+05	14.84%, 1.00	9.38% 1.00	0.00%, 1.00
*conv6	8×8	*128*, 256	5.9E+05	44.14%, 1.00	41.80% 1.00	0.00%, 1.00
*conv7	8×8	*256*, 256	5.9E+05	44.92%, 1.00	46.88% 1.00	0.00%, 1.00
*conv8	4×4	*256*, 512	1.2E+06	51.56%, 1.06	47.66% 1.00	0.00%, 1.00
*conv9	4×4	*512*, 512	2.4E+06	76.17%, 1.16	77.73% 1.00	0.00%, 1.02
*conv10	4×4	*512*, 512	2.4E+06	70.51%, 2.51	77.54% 1.00	0.00%, 1.16
*conv11	2×2	*512*, 512	2.4E+06	77.93%, 4.18	82.62% 1.00	0.00%, 1.15
*conv12	2×2	*512*, 512	2.4E+06	78.52%, 2.24	86.91% 1.10	0.00%, 1.87
*conv13	2×2	*512*, 512	2.4E+06	74.61%, 4.81	73.05% 1.05	0.00%, 1.83
*fc	1	*512*, 512	2.4E+06	76.00%, 2.32	74.41% 1.02	0.00%, 4.70
sofrmax	1	512, 10	5.1E+03	0.00%, 1.00	0.00%, 1.00	0.00%, 1.00
Compression rate				13.51X	11.01X	1.25X
Error				7.8%	7.9%	7.0%

Although GrOWL and weight decay both encourage parameter tying in layers 8-13, weight decay does it with less intensity and does not yield a sparse model, thus it cannot significantly compress the network. Li et al. (2016) propose to prune small weights after the initial training phase with weight decay, then retrain the reduced network; however, this type of method only achieves roughly $3X$ compression[3]. As mentioned by Li et al. (2016), layers 3-7 can be very sensitive to pruning; however, both GrOWL and group-Lasso effectively compress them with a minor accuracy loss. Compared with group-Lasso, GrOWL can further compress the network by $2.5X$, due to its ability to identify the correlations and tie the corresponding parameters. The CIFAR-10 dataset is simple enough, so that one could still expect a good performance after strong network compression. We believe this gap in the compression vs accuracy trade-off can be further increased in larger networks on more complex datasets. We leave this question for future work.

5 Conclusion

We have proposed using the recent GrOWL regularizer for simultaneous parameter sparsity and tying in DNN learning. By leveraging on GrOWL's capability of simultaneously pruning redundant parameters and tying parameters associated with highly correlated features, we can significantly reduce the model complexity, with a slight or even no loss in generalization accuracy. We evaluate the proposed method on both a fully connected neural network and a deep convolutional neural network. The results show that GrOWL can compress large DNNs by a factor of more than 13.5, outperforming group-Lasso in terms of the accuracy vs memory trade-off.

The correlation patterns identified by GrOWL are close to those of the input features to each layer. This may be important to reveal the structure of the features, contributing to the interpretability of deep learning models. On the other hand, by automatically tying together the parameters corresponding to highly correlated features, GrOWL alleviates the negative effect of strong correlations that might be induced by the noisy input or the co-adaption tendency of deep neural networks.

Future work will explore the use of GrOWL in the layers corresponding to large output features, either within each neuron, by predefining each 2D convolutional filter as a group, or to obtain new sharing structures.

[3]Although parameter sharing is not considered by Li et al. (2016), according to Table 2, pruning following weight decay together with parameter sharing still cannot compress the network as much as GrOWL.

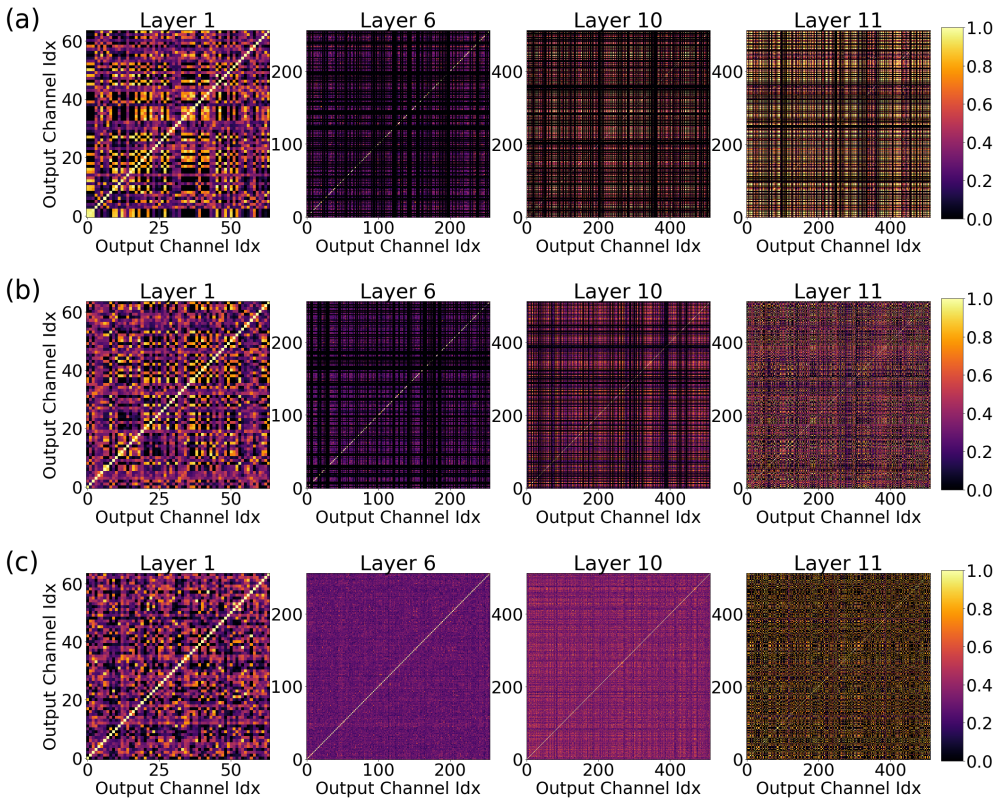

Figure 5: Visualization of the cosine similarities between the vectorized output channels for layers 1, 6, 10 and 11. We pass the whole training dataset (50, 000 samples) to the network trained with different regularizers **(a)** GrOWL, **(b)** group-Lasso, **(c)** weight decay. We then record the corresponding outputs of each layer at the end of the training phase. The cosine similarity is obtained by taking the averages with respect to all 50,000 samples. The above plots correspond the results in Table 2. The channel similarities for all layers are provided Appendix C.

REFERENCES

Alireza Aghasi, Nam Nguyen, and Justin Romberg. Net-trim: A layer-wise convex pruning of deep neural networks. *arXiv preprint arXiv:1611.05162*, 2016.

Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems*, pp. 2270–2278, 2016.

H. Bauschke and P. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.

Yoshua Bengio and James S Bergstra. Slow, decorrelated features for pretraining complex cell-like networks. In *Advances in neural information processing systems*, pp. 99–107, 2009.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J Candès. Slope–adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3): 1103, 2015.

Howard D Bondell and Brian J Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64(1):115–123, 2008.

P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143:1835–1871, 2013.

Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

Wenlin Chen, James Wilson, Stephen Tyree, Kilian Weinberger, and Yixin Chen. Compressing neural networks with the hashing trick. In *ICML*, 2015.

Christopher Clark and Amos Storkey. Teaching deep convolutional neural networks to play go. *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015.

Maxwell Collins and Pushmeet Kohli. Memory bounded deep convolutional networks. In *arXiv:1412.1442*, 2014.

J. de Leeuw, K. Hornik, and Patrick Mair. Isotone optimization in R: Pool adjacent-violators algorithm (PAVA) and active set methods. *Statistical Software*, 32:1–24, 2009.

Misha Denil, Babak Shakibi, Laurent Dinh, Nando de Freitas, et al. Predicting parameters in deep learning. In *Advances in Neural Information Processing Systems*, pp. 2148–2156, 2013.

Emily L Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*, pp. 1269–1277, 2014.

Sander Dieleman, Jeffrey De Fauw, and Koray Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

Mário AT Figueiredo and Robert D Nowak. Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 930–938, 2016.

Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*, 2014.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

Song Han, Huizi Mao, and William Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization, and Huffman coding. In *ICLR*, 2016.

Babak Hassibi, David G Stork, et al. Second order derivatives for network pruning: Optimal brain surgeon. *Advances in neural information processing systems*, pp. 164–164, 1993.

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.

Y. LeCun. *Modèles connexionnistes de l'apprentissage*. PhD thesis, Université Pierre et Marie Curie, Paris, France, 1987.

Yann LeCun, John Denker, Sara Solla, Richard Howard, and Lawrence Jackel. Optimal brain damage. In *Neural Information Processing Systems (NIPS)*, volume 2, pp. 598–605, 1989.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Kenton Murray and David Chiang. Auto-sizing neural networks: With applications to n-gram language models. *arXiv preprint arXiv:1508.05051*, 2015.

Urvashi Oswal, Christopher Cox, Matthew Lambon-Ralph, Timothy Rogers, and Robert Nowak. Representational similarity learning with application to brain networks. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1041–1049, 2016.

Pau Rodríguez, Jordi Gonzàlez, Guillem Cucurull, Josep M Gonfaus, and Xavier Roca. Regularizing cnns with locally constrained decorrelations. *arXiv preprint arXiv:1611.01967*, 2016.

D. Rumelhart and J. McClelland. *Parallel Distributed Processing*. MIT Press, 1986.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

Zagoruyko Sergey. 92.45% on cifar-10 in torch. `http://torch.ch/blog/2015/07/30/cifar.html`, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (B)*, 58:267–288, 1996.

Li Wan, Matthew Zeiler, Sixin Zhang, Yann L Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1058–1066, 2013.

Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 2074–2082, 2016.

Ming Yuan and Yi 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.

Xiangrong Zeng and Mário Figueiredo. The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms. *arXiv*, 1409.4271, 2015.

Xiangrong Zeng and Mário AT Figueiredo. Decreasing weighted sorted l1 regularization. *IEEE Signal Processing Letters*, 21(10):1240–1244, 2014.

Hao Zhou, Jose M Alvarez, and Fatih Porikli. Less is more: Towards compact cnns. In *European Conference on Computer Vision*, pp. 662–677. Springer, 2016.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

APPENDIX A PROXGROWL

Various methods have been proposed to compute the proximal mapping of OWL (ProxOWL) . It has been proven that the computation complexity of these methods is $O(n \log n)$ which is just slightly worse than the soft thresholding method for solving ℓ_1 norm regularization. In this paper, we use Algorithm 2 that was originally proposed in Bogdan et al. (2015).

Algorithm 2 ProxGrOWL Bogdan et al. (2015) for solving $\text{prox}_{\eta,\Omega_\lambda}(z)$

Input: z and λ

Let $\widetilde{\lambda} = \eta\lambda$ and $\widetilde{z} = |Pz|$ be a nonincreasing vector, where P is a permutation matrix.

while $\widetilde{z} - \widetilde{\lambda}$ is not nonincreasing: **do**

Identify strictly increasing subsequences, ***i.e.,*** segments $i : j$ such that

$$\widetilde{z}_i - \widetilde{\lambda}_i < \widetilde{z}_{i+1} - \widetilde{\lambda}_{i+1} < \widetilde{z}_j - \widetilde{\lambda}_j \tag{12}$$

Replace the values of $\widetilde{z}$ and $\widetilde{\lambda}$ over such segments by their average value: for $k \in \{i, i+1, \cdots, j\}$

$$\widetilde{z}_k \leftarrow \frac{1}{j-i+1} \sum_{i \le k \le j} \widetilde{z}_k, \qquad \widetilde{\lambda}_k \leftarrow \frac{1}{j-i+1} \sum_{i \le k \le j} \widetilde{\lambda}_k \tag{13}$$

end while

Output: $\widehat{z} = \text{sign}\,(z)\, P^T(\widetilde{z} - \widetilde{\lambda})_+$.

APPENDIX B DIFFERENT CHOICES OF GROWL PARAMETERS

We first evaluate GrOWL norm regularization on synthetic data. The covariance matrix Σ of the data X is set as a *block diagonal matrix* such that each block represents the covariances of a cluster of correlated features, and there is a gap g between two blocks. Within each cluster, the covariance between two features X_i and X_j is $\text{cov}(X_i, X_j) = 0.96^{|i-j|}$. we assume that the features from different clusters are independent of each other, while In this section, we set $n = 784, K = 10$ and block size equals 50, $g = 28$. We generate 10000 training examples and 1000 testing examples.

We train a neural network with a single fully connected layer that consists of 300 hidden units. In Fig 6 (a) -(c), we visualize the similarities (Eq 10) between rows of the weight matrix learned by different regularizers. As we can see, both Weight Decay and GrOWL can capture the correlations of the data by tying the corresponding rows together. However, as it's implied by Fig 6 (a), (c) and (d), in the highly correlated regime of the data, GrOWL ties the associated rows even tighter than weight decay does. Moreover, GrOWL can capture more correlation information of the data while encouraging sparsity simultaneously. On the other hand, although Group Lasso can prune away the redundant rows, it fails at selecting all relevant features and for the selected features, the associated rows are quite dissimilar with each other. In contrast, GrOWL can select all of the important relevant features and tying them together. This capability of GrOWL can help interpret the model better while allowing more space for parameter sharing in the retraining phase.

In Figure 6, we plot the the first 25000 entries of the sorted similarity matrices obtained by applying GrOWL with different p (Eq 9) values. Although, by setting the weights to a common constant beyond index p means that clustering is only encouraged among the p largest coefficients, i.e., only among relevant coefficient groups, Figure 6 shows that, with probably chosen p, GrOWL is able to encourage more parameter tying than OSCAR ($p = n$). This implies that with probably chosen p, GrOWL is able to identify more correlations of the input features than OSCAR.

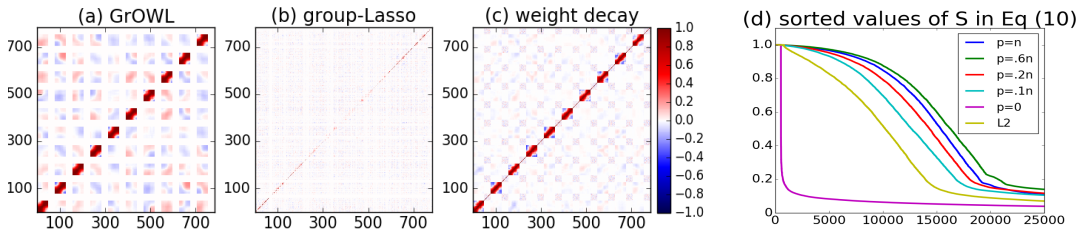

Figure 6: **(a) - (c)**: Visualization of similarity pattern (Eq (10)) between rows of the weight matrices learned by different regularizations. The similarity matrices corresponding with GrOWL and weight decay manifest block diagonal pattern that matches the covariance of the data, while GrOWL performing much better than weight decay. The weight matrices for generating (b) and (c) have the same sparsity level. As we can see, group-Lasso cannot identify the underlying correlations of the data. **(d)**:Regularization effect of GrOWL with different choices of p in Eq (9).

APPENDIX C VGG-16 ON CIFAR-10

In this section, we visualize the similarity pattern between the output channels of each layer. For each layer, we also plot the similarity between the rows of the (reshaped) weight matrices of the next layer. As we can see in Figures 7, 8, and 9 GrOWL and weight decay, the similarity pattern of the output channels of each layer and that of the corresponding weight matrix of the next layer match well with each other; while for Group Lasso, they do not match well with each other.

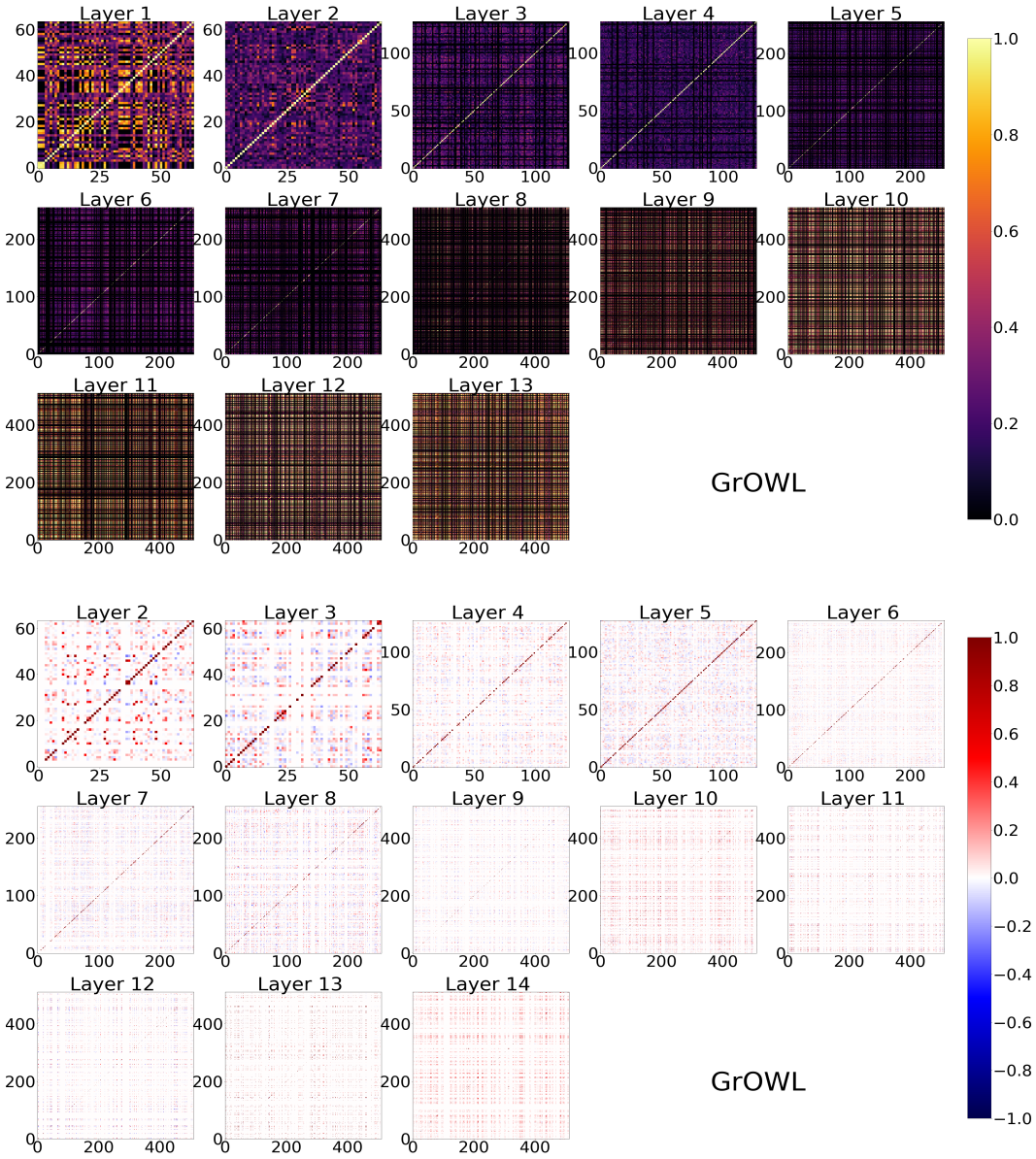

Figure 7: VGG-16: visualization of the pairwise similarity between the vectorized output channels and that between the rows of the (reshaped) weight matrices of each layer. Results are recorded at the end of the training phase. Only GrOWL regularized layer are plotted here. First block are the similarity patterns between the output channels of each layer, and the second block are the similarity patterns of the associated matrix of the next layer.

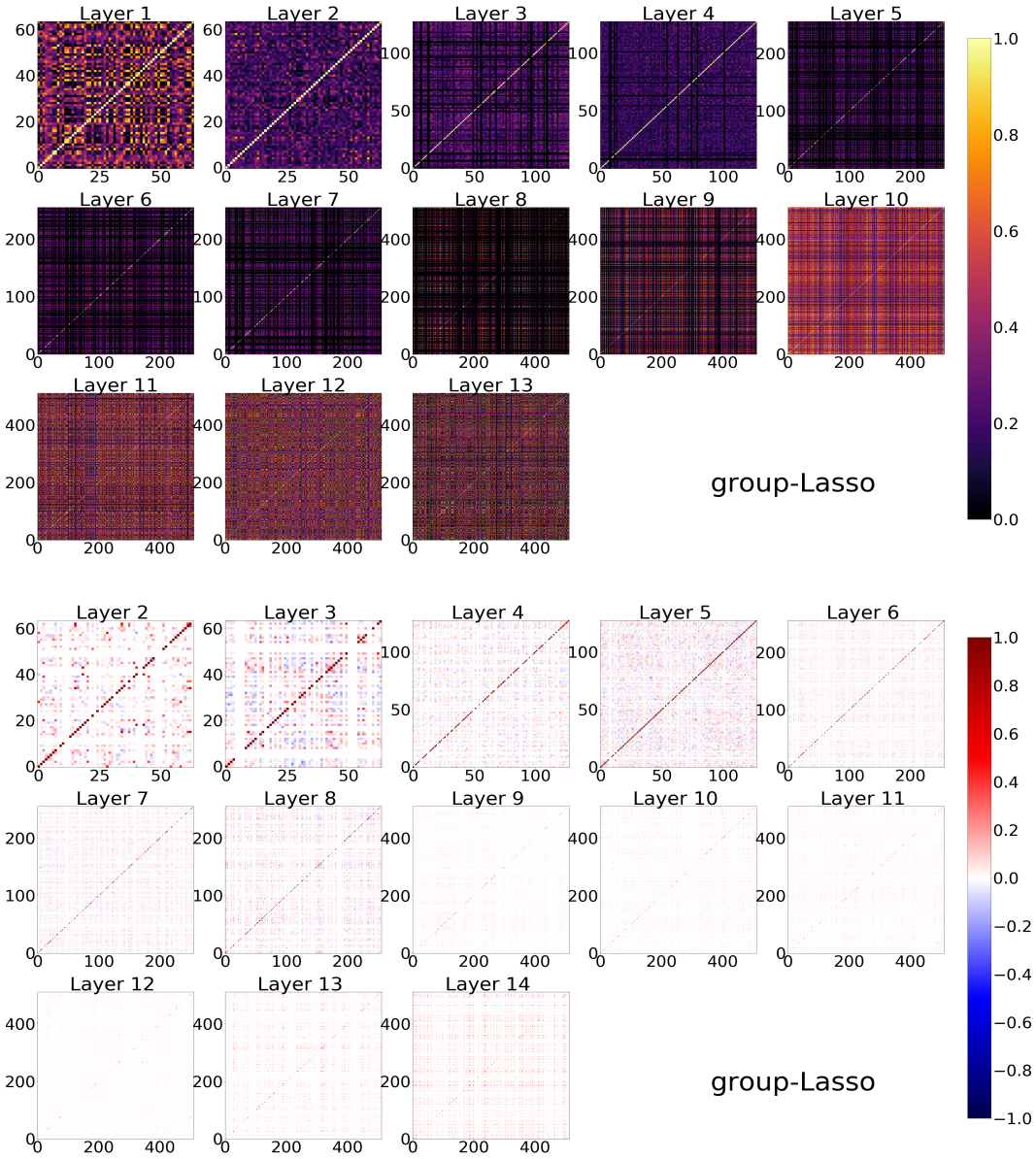

Figure 8: VGG-16: visualization of the pairwise similarity between the vectorized output channels and that between the rows of the (reshaped) weight matrices of each layer. Results are recorded at the end of the training phase. Only group-Lasso regularized layer are plotted here. First block are the similarity patterns between the output channels of each layer, and the second block are the similarity patterns of the associated matrix of the next layer.

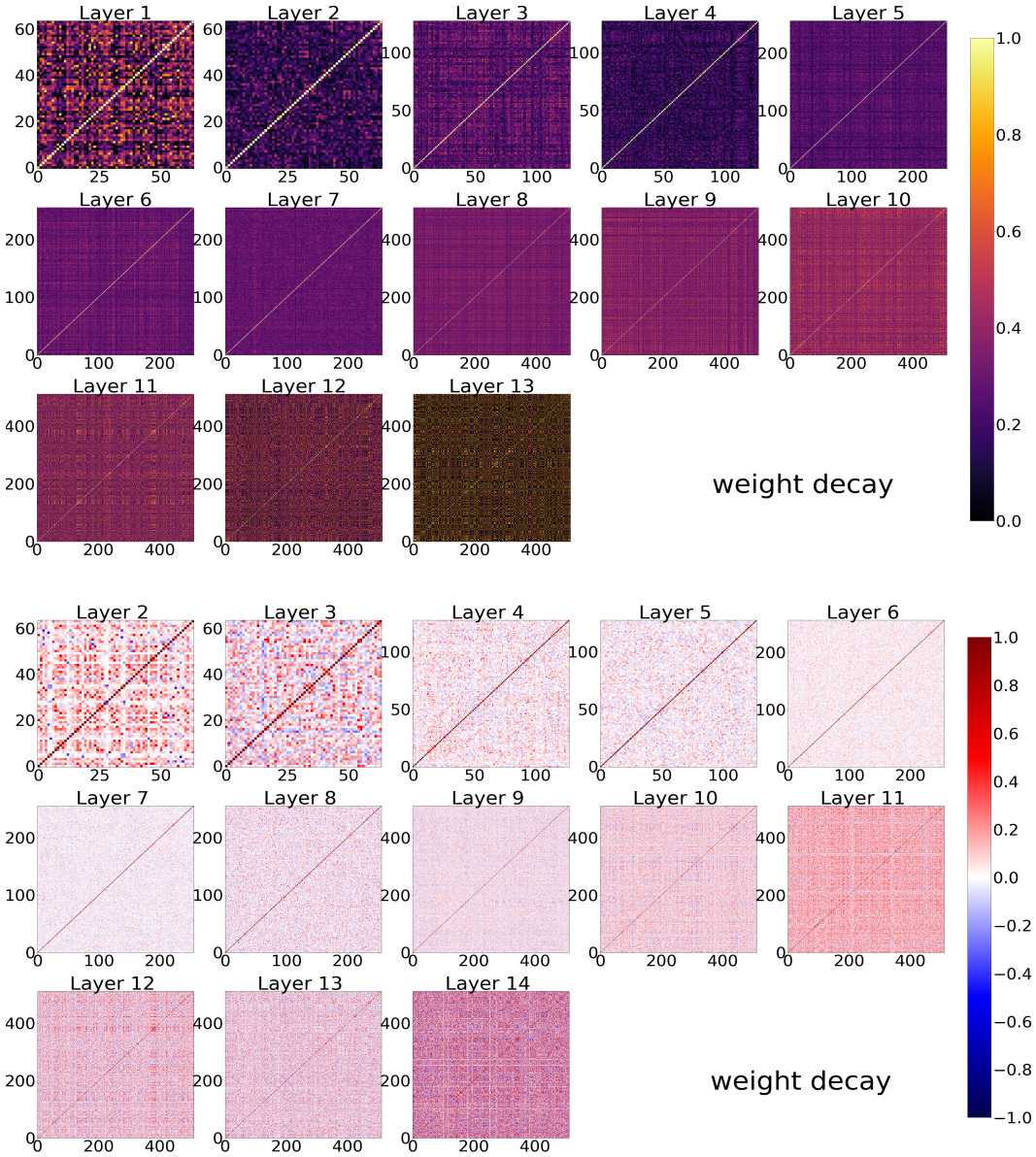

Figure 9: VGG-16: visualization of the pairwise similarity between the vectorized output channels and that between the rows of the (reshaped) weight matrices of each layer. Results are recorded at the end of the training phase. Only weight decay regularized layer are plotted here. First block are the similarity patterns between the output channels of each layer, and the second block are the similarity patterns of the associated matrix of the next layer.