Supplementary material: Structured Compression of Deep Neural Networks with Debiased Elastic Group LASSO

Oyebade K. Oyedotun, Djamila Aouada, Björn Ottersten
Interdisciplinary Centre for Security, Reliability and Trust (SnT),
University of Luxembourg, L-1855 Luxembourg
{oyebade.oyedotun, djamila.aouada, bjorn.ottersten}@uni.lu

Models	Error ↑	FLOPS ↓	Param. ↓
Group-LASSO	1.15%	33.0%	20.0%
Group-LASSO	1.60%	37.4%	28.3%
EGL	1.09%	34.0%	21.2%
EGL	1.94%	37.5%	29.4%
Ours: DEGL1	**-0.26%**	**34.0%**	**21.2%**
Ours: DEGL2	**0.15%**	**37.5%**	**29.4%**
Ours: DEGL3	**1.21%**	**46.9%**	**41.2%**

Reference ResNet-56: Error=30.37%, Param.=0.85M, FLOPS=125M

Table A1: ResNet-56 compression results on CIFAR-100 dataset

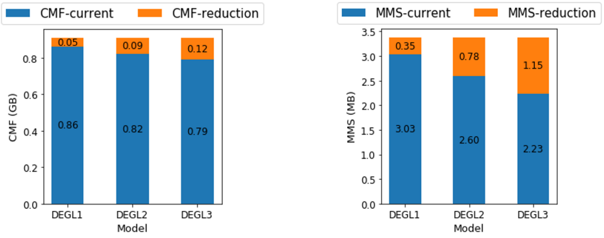

Figure A1: ResNet-56 CMF and MMS results on CIFAR-100

A1. ResNet-56 on CIFAR-100 dataset

The results of DEGL using ResNet-56 on CIFAR-10 is given in Table A1, where DEGL1, DEGL2 and DEGL3 are obtained using the same t_{th} values as with VGG-16 on CIFAR-100 in the main material. We particularly observe that DEGL outperforms conventional group LASSO for ResNet architectures, where skip connections can increase features correlations among different layers, and therefore conventional group LASSO struggles with inconsistent feature selection; see Section 3.2.2 for discussion. Figure A1 shows how pruning impacts CMF and MMS for models reported in Table A1.

A2. Additional ablation studies

A2.1. Pruning threshold values and performance

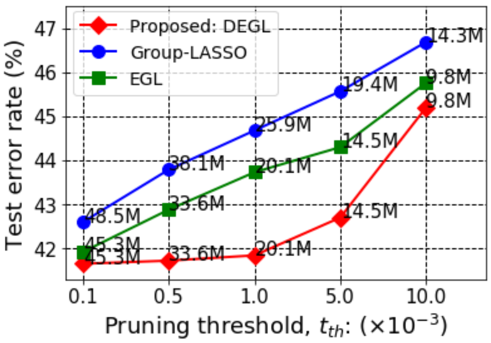

Figure A2: Pruning threshold values and AlexNet performance loss on ImageNet dataset

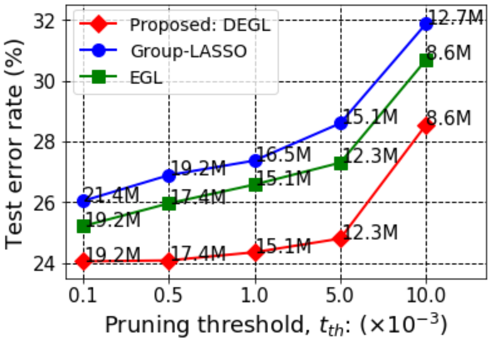

Figure A3: Pruning threshold values and ResNet-50 performance loss on ImageNet dataset

Herein, we perform additional experiments to observe how pruning threshold values (t_{th}) impacts performance loss for compressed models, which include the proposed DEGL, group LASSO and EGL. For this investigation, experiments are carried out on ImageNet using AlexNet and ResNet-50 models. Figure A2 and Figure A3 show obtained results including the current number of model parameters on AlexNet and ResNet-50, respectively; results given are recorded after pruning and retraining all models. It is seen that the proposed model compression approach, DEGL, consistently incurs smaller performance loss as compared to

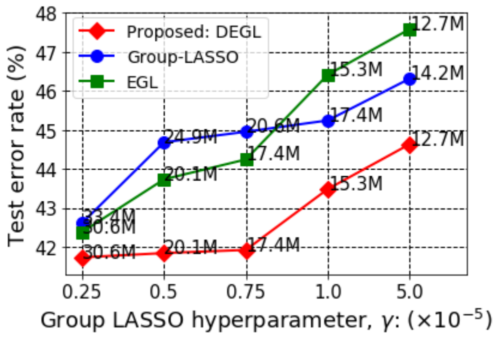

Figure A4: Group LASSO penalty weight and AlexNet performance loss

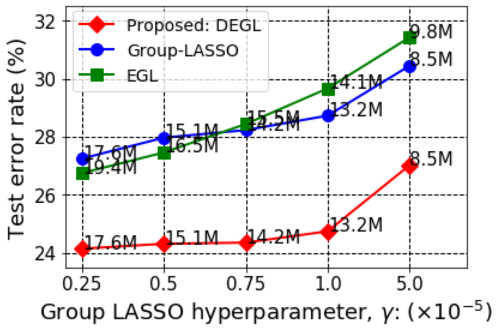

Figure A5: Group LASSO penalty weight and ResNet-50 performance loss

group-LASSO and EGL. Furthermore, it is noted that EGL outperforms group LASSO. Also, given a specified pruning threshold value, t_{th}, DEGL results in smaller number of model parameters than group-LASSO. Overall, it is seen that the performance losses of all the compression methods increase with an increase in pruning threshold values, since the resulting models become progressively smaller.

A2.2. Feature selection regularization hyperparameter and performance

We also observe the impact of group feature selection hyperparameter, γ, on compression results based on performance loss. Experiments are performed on imageNet dataset using AlexNet and ResNet-50 models, and results are shown in Figure A4 and Figure A5, respectively. It is observed that for small values of γ, EGL outperforms group-LASSO. However, for both AlexNet and ResNet-50, the progress increase of γ leads to worse EGL performance than group-LASSO models. This interesting scenario is directly attributed to high model bias for EGL when γ exceeds a certain limit. This follows from the fact, both group LASSO and $l2$-norm penalties are used for retraining after model pruning. Conversely, group-LASSO models use only the group LASSO penalties for retraining after model pruning. Importantly, it is seen that for all values of γ for the compared models on AlexNet and ResNet-50, the proposed DEGL models incur the smallest performance loss. The good performance of DEGL is attributed to the debiasing step after model pruning; the retrained model uses only the $l2$-norm penalty.

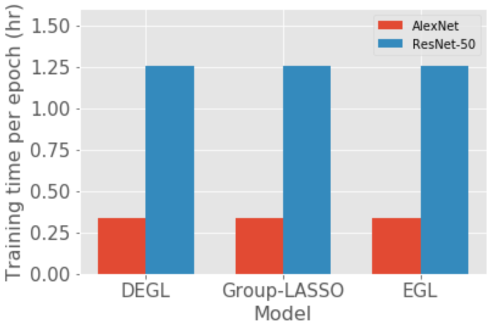

Figure A6: Training time for compression methods on imageNet

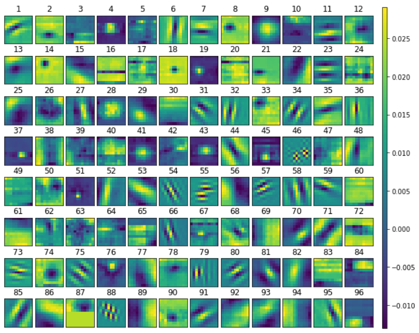

Figure A7: First convolution layer filters in AlexNet trained with DEGL. Filters '3', '20', '40', '49', '60'and '64' are selected for pruning

A2.3. Training time

Figure A6 shows the training times for compression methods DEGL, group LASSO and EGL on AlexNet and ResNet-50. Specifically, the times for the completion of one epcoh for the different models are given; results for each model are averaged over 3 different runs using a training batch size of 256; four V100 GPUS running on a workstation with 128GB of RAM are used. As such, it is seen that the proposed DEGL for compression does not increase training time; all the compression approaches compared require approximately the same training time. The same observation is made on all the other datasets used in this paper.

A2.4. Visualization of filters selected for pruning

The 96 convolution filters of the first layer of AlexNet using DEGL are shown in Figure A7. For compression, the six filters reported are selected for pruning; selected filters are determined as in Section 4.1.2 in the main manuscript.