

A APPENDIX

A.1 PROOFS

Lemma 1. *If $\max_{c \in C} P(y = c|\mathbf{x}) > \sigma_l, 1/2 \leq \sigma_l < 1$, then the estimation result is probably approximately correct. Otherwise, if $\max_{c \in C} P(y = c|\mathbf{x}) \leq \sigma_h, 0 < \sigma_h \leq 1/2$, then the estimation result can be probably approximately wrong.*

Proof.

From the study (Corbière et al., 2019), it can be stated that:

If the true probability is over the certain value ($1/2 \leq \sigma_l < 1$) for the certain input instance (\mathbf{x}), then the estimated answer from the model is correct:

$$P(y = y^*|\theta_{task}, \mathbf{x}) > \sigma_l \Rightarrow \hat{y} = y^*. \quad (4)$$

Eq.(4) can be rewritten as:

$$1 - \sum_{c \neq y^*} P(y = c|\theta_{task}, \mathbf{x}) > \sigma_l \quad (5)$$

As $P(y = c|\theta_{task}, \mathbf{x}) \geq 0, P(y = c|\theta_{task}, \mathbf{x}) < 1/2 < P(y = y^*|\theta_{task}, \mathbf{x}) \quad \forall c \neq y^*$ holds.

Therefore, Eq.(6) holds.

$$\hat{y} = \arg \max_{c \in C} P(y = c|\theta_{task}, \mathbf{x}) = y^* \quad (6)$$

From the probably approximately correct (PAC) bound (Valiant, 1984) on the trained model with parameter (θ_{task}) and train dataset (D_{train}) through empirical risk minimization:

$$P(|P(\hat{y} = y^*|\theta_{task}) - P(\hat{y} = y^*|\theta_{task}, D_{train})| \leq \epsilon) \geq 1 - \delta \quad (7)$$

If $\max_{c \in C} P(y = c|\theta_{task}, \mathbf{x}) > \sigma_l$ where $1/2 \leq \sigma_l < 1$, note that $\hat{y} = \arg \max_{c \in C} P(y = c|\theta_{task}, \mathbf{x})$ and by the PAC bound, $P(y = y^*|\theta_{task}, \mathbf{x}) > \sigma_l$ with probability confidence on $1 - \delta$ approximately up to an error ϵ correct.

It means that $\hat{y} = c$ with probability confidence on $1 - \delta$ and approximately correct (up to error ϵ).

The proof for the otherwise case also can be derived in likewise way.

■

Theorem 1. *If $L(\mathcal{M} \odot \mathcal{K}; m_j^i = 0, \mathcal{D}) \geq L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) \geq 0, \forall (i, j) \in \{(i, j) | m_j^i = 1, \forall m_j^i \in \mathcal{M}\}$, pruning a channel by CS_j^i is equal to solving empirical risk minimization (ERM) problem of the neural network in the finite hypothesis space.*

Proof.

Consider the pruning a channel by conduct element-wise product on filters with masking matrix ($\mathcal{M}_i \odot \mathcal{K}_i$) is equal to finding \mathcal{K}_i with which channel to be pruned ($\theta_{j,q}^i = 0, \forall q$) in the finite hypothesis space H where the space consists of all possible channel pruning cases. Therefore, choosing a channel to prune by CS_j^i is written as:

$$\begin{aligned} & \arg \min_{j \in [0, n_i)} CS_j^i \\ &= \arg \min_{\mathcal{K}_i | \theta_{j,q}^i = 0, \forall q \in H} CS_j^i \end{aligned} \quad (8)$$

In the original score formulation transformed for channel pruning from weight-wise score in SNIP (Lee et al., 2019), $g_j^i((\mathcal{M} \odot \mathcal{K}; \mathcal{D}))$ is approximated by the derivative of L_j^i for calculation efficiency in implementation.

$$\begin{aligned} \Delta L_j^i(\mathcal{K}; \mathcal{D}) &= L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) - L(\mathcal{M}|_{m_j^i=0} \odot \mathcal{K}; \mathcal{D}) \\ &\approx g_j^i(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) = \frac{\partial L(\mathcal{M} \odot \mathcal{K}; \mathcal{D})}{\partial m_j^i} \Big|_{\mathcal{M}=1} \end{aligned} \quad (9)$$

Apply this to CS_j^i , obtains

$$CS_j^i = \frac{|g_j^i(\mathcal{K}; \mathcal{D})|}{\sum_{i'} \sum_{j'} |g_{j'}^{i'}(\mathcal{K}; \mathcal{D})|} \approx \frac{|L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) - L(\mathcal{M}|_{m_j^i=0} \odot \mathcal{K}; \mathcal{D})|}{\sum_{i'} \sum_{j'} |L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) - L(\mathcal{M}|_{m_j^i=0} \odot \mathcal{K}; \mathcal{D})|} \quad (10)$$

Therefore, Equation 8 becomes

$$\arg \min_{\mathcal{K}_i | \theta_{j,q}^i=0, \forall q \in H} \frac{|L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) - L(\mathcal{M}|_{m_j^i=0} \odot \mathcal{K}; \mathcal{D})|}{\sum_{i'} \sum_{j'} |L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) - L(\mathcal{M}|_{m_j^i=0} \odot \mathcal{K}; \mathcal{D})|} \quad (11)$$

where denominator term and $L(\mathcal{M} \odot \mathcal{K}; \mathcal{D})$ term in numerator are constant with regard to j (wrapped as $\mathcal{K}_i | \theta_{j,q}^i=0, \forall q$).

As the cross entropy is usually used as loss function $L(\cdot)$, assume $L(\cdot) > 0$. Then, when we denote $L(\mathcal{M} \odot \mathcal{K}; \mathcal{D})$ in numerator and denominator term as constant $\alpha, \beta \in \mathbb{R}$, $\alpha, \beta \geq 0$ respectively, the equation becomes

$$\arg \min_{\mathcal{K}_i | \theta_{j,q}^i=0, \forall q \in H} \frac{|\alpha - L(\mathcal{M}|_{m_j^i=0} \odot \mathcal{K}; \mathcal{D})|}{\beta} \quad (12)$$

This implies that, when $L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) \geq \alpha \geq 0$, the problem becomes equal to empirical risk minimization problem of the neural network in the finite hypothesis space H' where the space consists of all possible channel pruning cases.

$$\begin{aligned} &\arg \min_{\mathcal{K}_i | \theta_{j,q}^i=0, \forall q \in H} \frac{|\alpha - L(\mathcal{M}|_{m_j^i=0} \odot \mathcal{K}; \mathcal{D})|}{\beta} \\ &= \arg \min_{\mathcal{K}_i | \theta_{j,q}^i=0, \forall q \in H} L(\mathcal{M} \odot \mathcal{K}; \theta_{j,q}^i = 0, \forall q, \mathcal{D}) \\ &= \arg \min_{\mathcal{K} \in H'} L(\mathcal{K}; \mathcal{D}) \end{aligned} \quad (13)$$

■

Corollary 1. *As solving ERM guarantees probably approximately correct (PAC) bound, under the same condition in Theorem 1., pruning a channel by CS_j^i also guarantees PAC bound and its estimation error is upper bounded.*

Proof.

According to the theorem from ERM (Vapnik, 1992), it is proven that:

For an ERM solution $\mathcal{K}_{\mathcal{D}}^{ERM} \in \arg \min_{\mathcal{K} \in H} L(\mathcal{K}; \mathcal{D})$, its estimation error can be upper bounded as

$$L(\mathcal{K}_{\mathcal{D}}^{ERM}; \mathcal{X}) - \inf_{\mathcal{K} \in H} L(\mathcal{K}; \mathcal{X}) \leq 2 \sup_{\mathcal{K} \in H} |L(\mathcal{K}; \mathcal{D}) - L(\mathcal{K}; \mathcal{X})| \quad (14)$$

In addition, from the theory from the PAC (Valiant, 1984), it is proven that:

(PAC Learnability of finite hypothesis space) For a finite hypothesis space H , H is PAC-learnable by the empirical risk minimization algorithm.

(PAC Bound in finite hypothesis space) Hypothesis space H finite, dataset \mathcal{D} with M i.i.d. samples, $0 < \epsilon < 1$: for any learned hypothesis h that is consistent on the training data:

$$P(\text{error}_{\mathcal{X}}(h) - \text{error}_{\mathcal{D}}(h) > \epsilon) \leq |H|e^{-2M\epsilon^2} \quad (15)$$

where \mathcal{X} denotes data instance space.

Therefore, under certain condition $L(\mathcal{M} \odot \mathcal{K}; m_j^i = 0, \mathcal{D}) \geq L(\mathcal{M} \odot \mathcal{K}; \mathcal{D}) \geq 0, \forall (i, j) \in \{(i, j) | m_j^i = 1, \forall m_j^i \in \mathcal{M}\}$, as we proved that pruning a channel by CS_j^i is equal to solving empirical risk minimization in the finite hypothesis space H' where the space consists of all possible channel pruning cases, we can say that estimation error can be upper bounded for the pruning problem.

In addition, pruning problem can be the PAC-learnable by the above mentioned theorem in Valiant (1984), therefore, gap between true error and train error is guaranteed to be upper bounded. ■

A.2 FURTHER EXPERIMENTS

Experimental settings. We conducted further experiments on other network models and datasets to verify robustness of accuracy degradation on the proposed pruning scheme over various cases. Including the ResNet-101 for UC Merced land use satellite imagery dataset presented in main content, we further evaluated on VGG-16 for CIFAR-10 dataset and WRN-18 for Caltech101 dataset. The detail settings for each experiment is described as follows.

- **ResNet-101 for UC Merced satellite imagery dataset:** We use ResNet-101 architecture (He et al., 2016) transforming only the last single fully connected layer with 21 output channels for the dataset. The model is trained with 160 epochs in total by using SGD optimizer with momentum of 0.9, batch size of 32, and weight decay rate of 0.0001. The initial learning rate is set to 0.1 and decayed by 1/10 at every 60 epochs. For data augmentation, only resizing the input data to 256×256 size is applied. In the dataset (Yang & Newsam, 2010), 90% of total dataset is split to train set and the other 10% is used for test set.
- **VGG-16 for CIFAR-10 dataset:** In order to see the performance of our pruning scheme on non-residual network, we also evaluate on VGG-16 with CIFAR-10 dataset. We modify the VGG-16 architecture where an average pooling layer is attached after the last convolutional layer, and only a single fully connected layer with 512 input channel is connected at the end of network for CIFAR-10 from the original VGG-16 architecture (Simonyan & Zisserman, 2015). The model is trained with 160 epochs in total by using SGD optimizer with momentum of 0.9, batch size of 128, and the weight decay rate of 0.0001. The initial learning rate is set to 0.1 and decayed by 1/10 at every 60 epochs. The standard data augmentation (i.e., translation up to 4 pixels for fitting to VGG-16 operations, random horizontal flip and normalization) is applied for input.
- **WRN-18 for Caltech101 dataset:** We also try to see the performance on wider residual network by evaluating on WRN-18 with Caltech101 dataset. We use the WRN-18 architecture (Zagoruyko & Komodakis, 2016) where only a single fully connected layer at the end of network is modified with 101 output channels for Caltech101 dataset. The model is trained with 80 epochs in total by using SGD optimizer with momentum of 0.9, batch size of 32, and weight decay rate of 0.0001. The initial learning rate is configured to 0.1 and decayed by 1/10 at 60 epoch. We just resize the input data to 224×224 size for data augmentation. In the dataset (Fei-Fei et al., 2006), 90% of total dataset is randomly split to train set and the other 10% is used for test set.

Results of further experiments. The results of further experiments on VGG-16 for CIFAR-10 dataset and WRN-18 for Caltech101 dataset are presented in Figure 7 and Figure 8 respectively. The results show similar tendency to the result of ResNet-101 for satellite imagery dataset presented on

the evaluation section in main paper. The proposed pruning scheme shows the best robustness to the accuracy degradation by considering the layer-wise sensitivity.

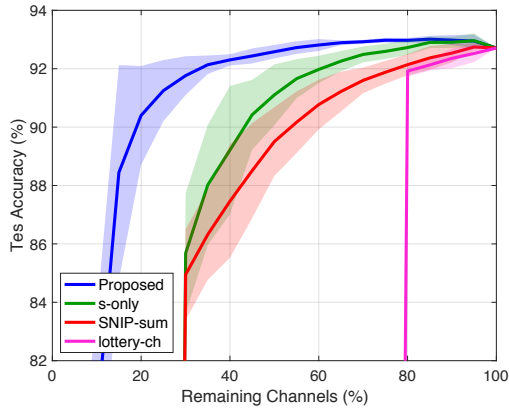


Figure 7: Test accuracy with respect to the percentage of remaining channels over pruning methods on VGG-16 for CIFAR-10 dataset

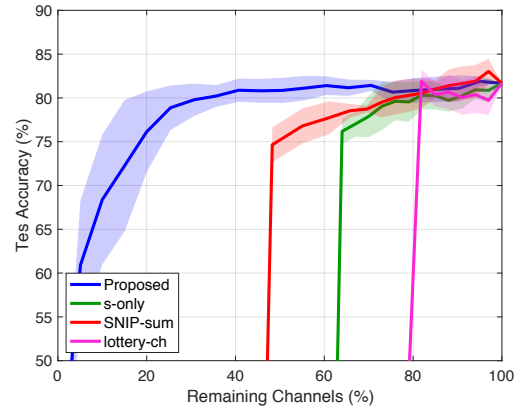


Figure 8: Test accuracy with respect to the percentage of remaining channels over pruning methods on WRN-18 for Caltech101 dataset