INTERPRETING CNN COMPRESSION USING INFORMATION BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we first investigate the representation learned in convolutional neural networks at the filter-wise granularity by computing the mutual information between channels of higher conv-layers and input or output variables. Then we identify the approximate minimal sufficient statistics of learned representation based on the information bottleneck principle and propose a novel approach to automatically compress a neural network. This approach prunes a large trained network structurally and automatically by extracting relevant information backpropagately layer by layer in the post-training phase. Our experimental results match the two fundamental data processing inequalities, and prove that mutual information is a fundamental element for examining the efficiency of the internal representations at the filter-wise granularity. In addition, using the information bottleneck principle to interpret structure compression is an efficient method to get closer to the information theoretic limit of compression/prediction problem. Finally, from the observed results, we argue that compression is causally linked to the improved generalization performance.

1 INTRODUCTION

Despite massive model compression approaches are proposed, there is still lacking comprehensive theoretical understanding of the compression methods that developed with iteratively retraining strategy. Previous work Tishby & Zaslavsky (Apr. 2015) proposed to analyze and understand Deep Neural Networks (DNNs) with the theoretical framework of IB principle Tishby et al. (Sep. 1999), which showed that any DNN can be quantified in the Information Plane: the Mutual Information values that each layer preserved on the input and output variables. They argued that both the optimal architecture, number of layers and features/connections are related with the information bottleneck trade-off between compression and prediction for each layer. Following Tishby & Zaslavsky (Apr. 2015), the authors of Shwartz-Ziv & Tishby (2017) give a new view that separating the training optimization process of SGD into two phases: the fitting phase and the compression phase, they also observed that compression cannot be linked to architectural simplicity, whether these claims all hold true is the subject of an ongoing debate Saxe et al. (May 2018). Subsequently, several works (see (Kolchinsky et al., Sep.2018), (Alemi et al., Apr. 2017)) proposed different methods for performing IB for discrete or continuous, possibly non-Gaussian features by claiming that computing mutual information in the IB functional Tishby & Zaslavsky (Apr. 2015) is, in general, computationally challenging. Achille & Soatto. (2017) and Achille & Soatto (2018) further analyze and control the characteristics of representations with IB-based principle. To bring the practice in line with the theory, they proposed to replace mutual information terms with tractable bounds in order to obtain objective functions that can be computed and optimized using neural-network-based methods. Further, Amjad & Geiger (2019) argued that the IB functional does not capture desirable properties of intermediate representations, such as allowing robustness to noise and deploying such architecturally/computationally economical CNNs on embedded/edge devices with limited computational resources and real-time processing constraints. They also showed that the IB functional leads to an ill-posed optimization problem in deterministic DNNs by either being infinite for almost all parameter settings or by being a piecewise constant function of parameters, hence it is not fit for training DNNs, these problems might be solved by replacing the IB functional with a more well-behaved cost function as a remedy.

1.1 CONTRIBUTIONS

Rather than adding to the debate or providing new tractable bounds to replace IB terms for obtaining feasible SGD-optimized cost functions, inspired by Amjad & Geiger (2019), Yu et al. (2018) and HH & J. (1999), in this paper we review the IB functional from the perspective of considering the desired property of representation simplicity for architecturally/computationally economical benefits. We first analyze the interpretability of learned representation at the filter-wise granularity for the CNNs, then we simply replace the IB functional compression term with a filter selection function to select the most informative filters for discriminating objects and use the highly correlated AUC value to replace the original IB prediction term. More precisely, (a) we analyze how much information each filter captured from both the information-theoretic and human visualization semantics' view on any trained CNNs, which is qualified collectively with two quantities: the virtualization of information plane, and the overlapped heatmaps synthesized from selected high-informative and low-informative filters respectively. (b) we propose the post-training filter selection approach based on the IB principle. Given large trained CNNs, we evaluate the final AUC results while incrementally pruning/selecting filters ranked by informative scores preserved on the input and output labels, which is equivalent to minimize the IB functional. (c) Finally, we turn to the general problem of model compression against proposing a automatic framework to make each latent representation layer as simple as possible without changing the original network structure and the object functions. The framework works backpropagately by first applying the filter selection approach to the top intermediate representation layer, then refining relevant information preserved on class labels in the subsequent layers, all of the low-informative and redundant filters are pruned off during the post-training phase without retraining. Our experimental results admit the statement of significant pruning after training without performance degradation and complement the general research area of lacking interpretability regards model compression. Moreover, our compression results demonstrate that structure compression is in some extent causally linked to the improved generalization performance.

To the end, our work fits the thinking of compressing CNNs structurely using a modified IB functional to instill the property of representation simplicity with explicitly clear interpretability in the post-training phase. Hence it is very flexible, theoretically guided and practically interpretable for accelerating structure compression.

2 RELATED WORKS

The authors of paper Tishby & Zaslavsky (Apr. 2015) started to introduce the new idea of using the IB principle to analyze and understand the inner workings of DNNs, and formulated the goal of deep learning as an information theoretic trade-off between compression and prediction. They suggested that DNNs should learn to extract the most efficient informative features, or approximate minimal sufficient statistics, with the most compact architecture (i.e. minimal number of layers, with minimal number of units within each layer). They argued that compression is thus necessary for generalization, the hidden layers must compress the input in order to reach the optimal point on the information curve where it gives the trade-off between the complexity and the accuracy of the representation. They also suggested the IB functional as an optimization criterion for training DNNs. Several works have been published as a follow up of Tishby & Zaslavsky (Apr. 2015) either from experimental and theoretical IB-based DNN analysis or IB-based DNN training and optimization. On the one hand, in paper Shwartz-Ziv & Tishby (2017) they suggested to open the black box of deep neural network with information. Whereas the authors of Saxe et al. (May 2018) disagreed with them analytically and empirically to present none of those claims hold true. On the other hand, Amjad & Geiger (2019) stated that the successes reported in IB-based training methods such as Kolchinsky et al. (Sep.2018) and Alemi et al. (Apr. 2017) in terms of goals of generalization, adversarial robustness cannot be attributed to the use of IB functional but should be considered as an outcome of making DNNs stochastic.

Further, Amjad & Geiger (2019) defined a list of properties of an intermediate representation desirable for the classification task, such as sufficient, maximally compressed, admitting a simple decision function and robust, which presented that the goal formulated in Tishby & Zaslavsky (Apr. 2015) is not sufficient for the classification tasks. In addition, the IB functional would not produce architecturally economical intermediate representations without possible remedies.

Meanwhile, despite there is already an ample amount of research on compressing neural network ((LeCun et al., 1990), (Hassibi & Stork., 1993), (Han et al., 2015), (Li et al., 2017)), we will discuss those which are highly related to our work. Recently proposed Dai et al. (2018) compressed DNNs using the information bottleneck principle instantiated via a tractable variational bound, in which network are trained to regularize most of neurons approximately uninformative. Similarly, the entropy-based framework proposed in work Luo & Wu. (2017) pruned several unimportant filters, whereas we go deeper to prune those both highly informative and redundant filters but helpless for discriminating class labels. In addition, instead of pruning and retraining iteratively with a compression rate as a hyperparameter for deciding the pruning boundary, we compress the network backpropagately to optimize an expected setting of AUC result, and this structure compression can be applied in any layer for any flexible AUC result.

To the end, seldom works investigated this topic from the view of both information theory and interpretability but Li et al. (2019), we address it by connecting the IB principle with architecturally/computationally economical benefits to open the black box of model compression .

3 Post-Training Filter Selection using Information Bottleneck

Suppose that Y is a binary class variable, X are features at the input of the CNNs, and $\hat{X}$ is a latent representation for the input X, the IB compression term $I(X;\hat{X})$ can be seen as a data-dependent regularization term that depends on the representation $\hat{X}$ rather than the parameters of the CNN (see (Amjad & Geiger, 2019)), which means it is not necessary to be optimized in the parameter training process. Instead we compress $I(X;\hat{X})$ during the post-training phase to utilize the facility of transferring learning. By this way a large state-of-the-art CNN model can be compressed to fit a small dataset. Suppose there are high dimensional filters $T \in R^C$ in one intermediate convolutional layer that each characterizing some properties of the input, in other words, the amount of information that the convolutional layer gained from input X is preserved in C joint information sources $\{T_1, T_2, \ ... \ , T_C\}$, as suggested in work Yu et al. (2018), the IB functional is formulated as

$$\begin{aligned} R_{IB} &= I\left(\hat{X};Y\right) - \beta I\left(X;\hat{X}\right) \\ &= I\left(\{T_1, T_2, \ ... \ , T_C\};Y\right) \ - \beta I\left(X;\{T_1, T_2, \ ... \ , T_C\}\right) \end{aligned} \tag{1}$$

For some trade-off parameter $\beta \ > \ 0$. The positive Lagrange multiplier operates as a trade-off parameter between the complexity of the representation $I(X;\hat{X})$ and the amount of preserved relevant information $I(\hat{X};Y)$.

As mentioned earlier in work HH & J. (1999), it is easy to see that the joint mutual information has the property

$$I\left(X;T_1,T_2, \ ... \ ,T_C\right) \ - I\left(X;T_1,T_2, \ ... \ ,T_{C-1}\right) = \ I\left(T_C;X|T_1,...,T_{C-1}\right) \geqslant 0 \tag{2}$$

Which means that discarding variable T_C will always decrease the mutual information. Therefore, we reformulate the original IB problem of finding the optimum minimal representation $\hat{X}$ as to construct a simple and elegant algorithm to select the minimal subset $k << C$ filters, $T_S \in R^k$, that perform optimally based on the joint mutual information with respect to the input and output variables. In general, we would like to maximize the mutual information between the subset of selected filters T_S and the target variable Y while using the minimal size of k filters in the same layer.

$$\tilde{S} = \arg\max_S I(T_S;Y), \qquad s.t.|S| = k, \tag{3}$$

where k is the number of filters we want to select.The original trade-off parameter β is implicitly becoming into the cardinality of subset T_S dividing the cardinality of T filters: $\beta = \dfrac{|T_S|}{|T|}$.

But minimizing this subset is an NP-hard optimisation problem, because the set of possible combinations of filters grows exponentially. We solve it with the simplest greedy step-wise pruning algorithm conversely. Filters are pruned incrementally, one or a group of filters at a time, greedy step-wise filter pruning thus selects the filters that at each step results in almost no degradation on the IB prediction term. Such filter selection method can be guided with both saliency and selection

criterion, the selection criterion is based on the generalization error on a test data set which can be approximated by the AUC results R_{AUC}. While the saliency criterion is implemented by ranking the filters according to its relevance to the input and output variables, which is interpretable as well to let one understand and know how far it can reach the information theoretic limit.

Let $\overline{S}^{t-1} = \{c_1, ..., c_{t-1}\}$ be the set of selected pruned filters before time step t and $T_S \cup T_{\overline{S}} = T$, then our greedy step-wise method selects the next pruned filter set c_t such that

$$c_t = \underset{\Delta S \notin \overline{S}^{t-1}}{\arg\max} I(T_{\overline{S}^{t-1} \cup \Delta S}; y), \quad \Delta S \in S^{t-1}, \tag{4}$$

where ΔS is determined by the criterion

$$I(T_i; X) \geqslant ... \geqslant I(T_j; X) \tag{5}$$

bidirectionally and iteratively. The pruned filters are selected bidirectionally due to the observations that filters can be quantified as high-informative and low-informative ones as shown in 5.3 and 5.4, those high-informative filters in the subsequent layers might learn environmental information for upper layers to discriminate objects from around, so we first maximally prune the low-informative filters, then prune the high-informative filters incrementally in the precise of not hurting expected performance. At last considering the redundant problem mentioned in Abbasi-Asl & Yu. (2017), the filters with intermediate information scores are iteratively pruned off with greedy small step-wise selection method.

For example, given a trained large CNN that achieved state-of-the-art prediction precision $I(T; Y)$ on the supervised classification task, we assume $I(T; Y)$ is directly bounded by some constant AUC result R_{AUC}. Thus we wish to find such a T_S for any arbitrary small $0 < \delta < 1$ that can produce a $\hat{R}_{AUC}$ allow $|R_{AUC} - \hat{R}_{AUC}| < \delta$. More precisely, we select such a subset T_S by progressively pruning candidate filters in terms of the mutual information ranking list bidirectionally and iteratively sample on the rest of list at last. The candidates are those either low-informative on the input variable and output variable or those capturing massive redundant information on the environment that are not helpful for discriminating the object.

In particularly, let the binary class variable is $Y^k, k \in \{0, 1\}$, for the binary class variable Y^k, we may get three ranking subset $S_{k \in \{0,1,01\}}$, corresponding to different MI scores list between each filter and the input and output variable assembles $\{Y^0,\ Y^1,\ Y^0 \cup Y^1\}$ respectively. It is intriguing to find out that different combinations of this saliency criterion reveal non-trivial roles of filters playing on the classification task, (as seen in 5.5).

4 Automatic Backpropagation Information Compression

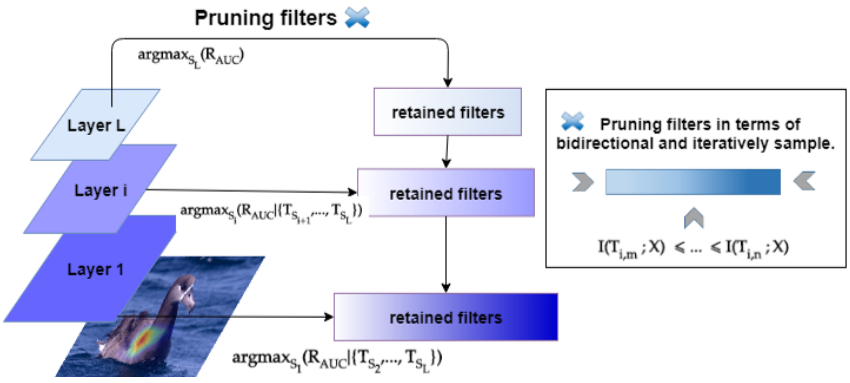

Figure 1: Automatic Backpropagation Information Compression Flow

We assume that for guaranteeing the expected prediction precision, the upper layers need to be compressed first due to their close spatial association with the output layer, so the filter selection

approach is applied layer by layer backpropagately to extract relevant information preserved on class labels. Specifically, the automatic backpropagation information compression flow of Figure 1 shows that the most relevant filters on the top conv-layer are retained, then we freeze this conv-layer and select the informative filters in the subsequent layers to hold the AUC performance, this is achieved one layer at a time.

$$\begin{aligned}\min_{\forall s\in S}\tilde{S}_i &= \min_{\forall s\in S}\gamma.\arg\max_{S_i} I(T_{S_i};Y|\{T_{S_{i+1}},...,T_{S_L}\}), \qquad s.t.|S_i|=k_i \\ &\propto \min_{\forall s\in S}\gamma.\arg\max_{S_i}(R_{AUC}|\{T_{S_{i+1}},...,T_{S_L}\})\end{aligned} \tag{6}$$

Where k_i is the number of filters we want to retain at each layer $i \in [L, 1]$, γ is used to flexibly adjust the expected R_{AUC}. To this end, we select the most informative filters contribute to state-of-the-art prediction performance in a layer-wise way, by pruning low-informative or massively redundant filters which plays non-trivial roles on the prediction of class labels.

5 Numerical Experiments and Results

5.1 Experimental Setup

For the numerical studies in this paper we explored state-of-the-art classification convolutional neural networks VGG-16 and ResNet-50, with standard settings and no other architecture constraints. Using the pre-trained ImageNet weights, the networks were fine-tuned with classical SGD algorithms on the CUB200-2011 Wah et al. (2011) and the Pascal VOC Part Chen et al. (2014) to demonstrate the interpretability and effectiveness of our approach.

5.2 Mutual Information Calculation

Mutual Information is a Shannon entropy-based measure of dependence between random variables, it is increasingly popular used to evaluate the quality of representation learned for DNN. However, MI is notoriously difficult to compute, particularly in continuous and high-dimensional settings. Therefore several methods are developed to apply them to deep neural networks which are computationally more efficient. The binning-based approach yielding discrete latent representation is attractive because of its computational efficiency when the numbers of bins is not too large, which is illustrated by the empirical mutual information plots from Shwartz-Ziv & Tishby (2017) and Saxe et al. (May 2018). The kernel density approach of Kolchinsky et.al.Kolchinsky & Tracey (19, 2017) consists in fitting a mixture of Gaussians (MoG) to samples of the variable of interest and subsequently compute an upper bound on the entropy of the MoG [48]. The method of Kraskov et al.Kraskov et al. (2004) used nearest neighbor distances between samples to directly build an estimate of the entropy. Recently, Belghazi et al. (2018) proposed a new non-parametric estimator for mutual information which involves the optimization of a neural network to tighten a bound. It is unfortunately computationally hard to test how these estimators behave in high dimension, as even for a known distribution the computation of the entropy is intractable in most cases. In our work, we apply the binning approach to obtain a finite MI value for analyzing how much information each filter captured, the neuron's relu output are binned into 100 equal intervals between the lowest and highest activation values. Though this binning approach may not exactly estimate the real mutual information, it is highly correlated with the mutual information as argued in work Goldfeld et al. (Nov. 2018).

5.3 Visualization OF Information Plane

As an extension work of Shwartz-Ziv & Tishby (2017), we visualize the information paths of CNNs in the information plane at the filter-wise granularity and analyze architectures of VGG-16 and ResNet-50 in terms of their efficiency in preserving the relevant information in each channel. The three colors of dots in Figure 2 (a) and (b) represent information preserved on filters for layers $block5\_conv3$, $block5\_conv2$, $block4\_conv3$ of VGG-16 and $activation\_49$, $activation\_40$, $activation\_22$ of ResNet-50 respectively, our experimental results are consistent with the successive Markov chain and DPI explanation of a K-layers CNN for both VGG-16 and ResNet-50, where as layers go deeper, the information preserved on labels is getting lost. (a) and (b) also show that filters

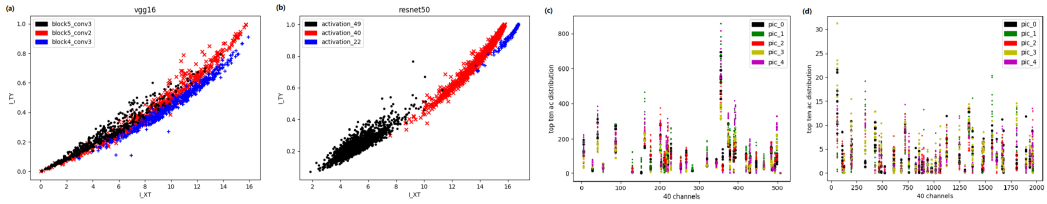

Figure 2: (a) and (b) Information Plane of filters (c) and (d) High activations distribution of high-informative filters

can be quantified with mutual information, and suggest that those low-informative filters could be discarded without significant performance degradation. More interestingly, we make an hypothesis from results of (a) and (b) that ResNet-50 might extract more disentangled features and the informative filters are more decentralized than VGG-16. VGG-16 is more apt to optimize a fraction of filters which dominate most of the patterns of target objects. (c) and (d) represent the top ten activations produced by top 40 high-informative filters on 5 random sampled pictures for VGG-16 and ResNet-50 respectively, which further verified our previous hypothesis that VGG-16 has centralized filters almost learn everything need to predict the labels than ResNet-50. We can intuitively understand these conclusions in next section.

5.4 VISUALIZATION OF OVERLAPPED HEATMAPS TO THE ORIGINAL IMAGES

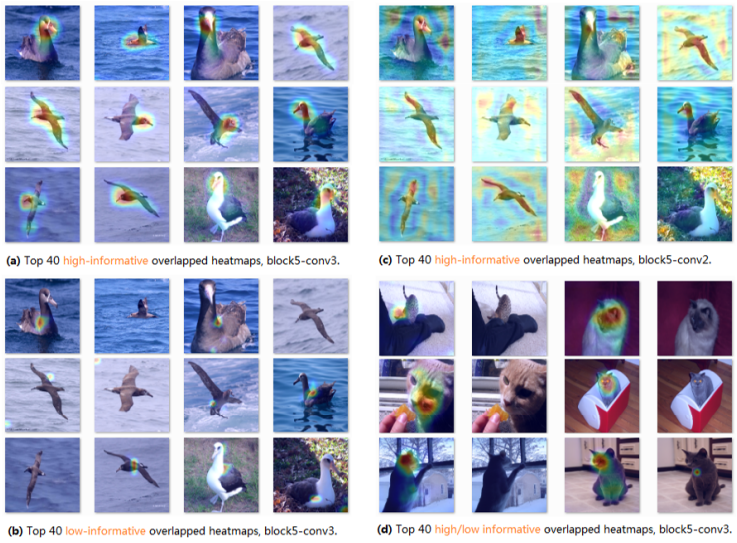

Figure 3: Overlapped heatmaps to the original images.

We explore the semantic interpretability of learned representations in CNNs on the CUB200-2011 and Pascal VOC Part dataset by visualizing how many information these filters captured. The CUB200-2011 dataset contains 200 species of birds with $N = 1960$ samples, and the Pascal VOC Part dataset select the cat class with $N = 441$ pictures, each of them are trained as the positive samples of a binary class separately, and the negative samples are selected from Places dataset Zhou et al. (2014) with $N = 2000$ and $N = 500$ samples respectively. Figure 3(a) and (b) are the overlapped heatmaps to the CUB200-2011 dataset that selected respectively from top 40 high-informative and low-informative filters of $block5\_conv3$ layer in VGG-16. Obviously, (a) demonstrates that the top 40 high-informative filters perfectly highlight object part patterns such as the head part to describe its characteristics. Oppositely, (b) shows that the top low-informative filters almost learned nothing useful knowledge. Based on these observations, pruning the low-informative filters is plausible to achieve the architecturally/computationally economical benefits. Similarly, Figure 3(d) jointly ver-

ifies above results on the Pascal VOC Part dataset. Figure 3(c) are the overlapped heatmaps that selected from top 40 high-informative filters at layer $block5\_conv2$ in VGG-16, which represent that these top 40 informative filters actually learned the environments information around the target object. However, they are probably unimportant for discriminating the target class, and can be pruned off, which is further demonstrated in the 5.5 section. In addition, observations tell that some filters with intermediate information scores might learn redundant information, these numbers can be pruned off without performance degradation.

To the end, experimental results show that our method has captured the most important information encoded in the filters of the mid-level layers in CNNs, the meaningful and useful characteristics of learned representations can be visualized and explained from the information theoretic perspective.

5.5 Compression Result

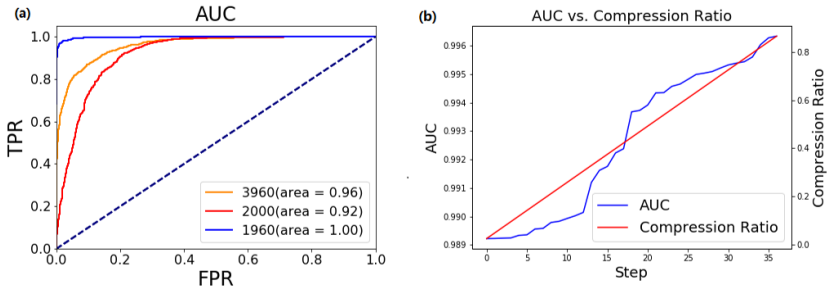

Figure 4: AUC results with respect to different numbers of pruned filters

In Figure 4(a), the three curves are generated by pruning low-informative filters with respect to different ranking list mentioned in section 3. More interestingly, it shows that the R_{AUC} still holds 1 after pruning low-informative 506 filters of total 512 in the $block5\_conv3$ layer with the criterion Y^1, we attribute this phenomenon to the success of finding the correct order of informative filters, which learned sufficient representations on the positive samples. This result further evaluates our hypothesis that VGG-16 has seldom key filters playing non-trivial roles on the classification tasks. Whereas the performance degraded largely on the criterion Y^0, since filters are ranked by the relevance with negative samples, intuitively speaking, these high-informative filters selected by Y^0 probably learned the representations irrelevant with the positive samples. So pruning these high-informative filters affects to predict the objects correctly. Regarding the criterion $Y^0 \cup Y^1$, the result is a little degraded, this is because this ranking list is a mixture order of Y^0 and Y^1. Based on these observations, our automatic backpropagation information compression method ranks the filters with the criterion Y^0 for state-of-the-art performance.

Figure 4(b) shows R_{AUC} vs compressing ratio results in layers $block5\_conv3$, $block5\_conv2$, $block4\_conv3$ of VGG-16 using our automatic backpropagation information compression method on the cat category of Pascal VOC Part dataset. As the compression ratio increasingly grows up, the generalization performance is gradually improved as well. This phenomenon represents that compression in some extent causes improved generalization performance.

Below results in Table 1 show that to guarantee an expected AUC result 0.997, great compression ratio is achieved with our automatic backpropagation information compression method, which also reveal that the learned representation behaves like a bottleneck to distort relevant information, and this distortion rate is getting smaller as layer goes deeper.

6 Discussion

Motivated by the work in Amjad & Geiger (2019), we propose a novel post-training filter selection approach to achieve structure compression using the information bottleneck principle. We explore the desired properties of representation simplicity by replacing the IB functional compression term with a filter selection function, and correlating the IB prediction term with AUC results. Our experimental results give an insight on interpreting the inner working mechanism of model compression

Table 1: Overall performance of our approach to prune filters on the VGG-16 model with prior fixed AUC threshold 0.997.

Layer	# Number of Filters		# Parameters		
	Original	Pruned	Original	Pruned	Percentage
Conv1-1	64	3	1.75K	84	5%
Conv1-2	64	3	36.06K	1.69K	5%
Conv2-1	128	24	72.13K	13.52K	19%
Conv2-2	128	6	144.13K	6.76K	5%
Conv3-1	256	63	288.25K	70.94K	25%
Conv3-2	256	72	576.25K	162.07K	28%
Conv3-3	256	55	576.25K	123.80K	21%
Conv4-1	512	234	1.15M	526.73K	46%
Conv4-2	512	350	2.30M	1.58M	69%
Conv4-3	512	270	2.30M	1.22M	53%
Conv5-1	512	336	2.30M	1.51M	66%
Conv5-2	512	450	2.30M	2.03M	88%
Conv5-3	512	506	2.30M	2.28M	99%
Dense1	-	-	6.42M	6.20M	97%
Dense2	-	-	257	0	0%
Total	-	-	21.14M	15.72M	74%

on supervised CNN classification tasks, which also demonstrate that the quality of filters can be evaluated from the information theoretic view. Precisely, the questions of how much relevant information each filter captured on class labels, how many filters are minimally sufficient to do the classification tasks in each hidden layer, are partly answered by visualizing the information plane and overlapped heatmaps at the filter-wise granularity. Additionally, the automatic backpropagation information compression method proposed in this work admitted the statement of significant pruning after training without performance degradation, with which one may step further to guide model architecture design on the direction of deepness or wideness. More importantly, on the case where a large trained model should be applied to a small dataset, the experimental results demonstrate that our method can compress the model while improve the generalization performance.

REFERENCES

Reza Abbasi-Asl and Bin Yu. Interpreting convolutional neural networks through compression. *arXiv preprint arXiv:1711.02329*, 2017.

A. Achille and S. Soatto. Information dropout: Learning optimal representations through noisy computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Alessandro Achille and Stefano Soatto. On the emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*, 2017.

A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy. Deep variational information bottleneck. *Proc. International Conference on Learning Representations (ICLR), Toulon*, Apr. 2017.

R.A. Amjad and B.C. Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence; preprint available: arXiv:1802.09766 [cs.LG]*, 2019.

Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062,ICML'2018*, 2018.

X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille. Detect what you can: Detecting and rep- resenting objects using holistic models and body parts. CVPR, 2014.

Bin Dai, Chen Zhu, and David Wipf. Compressing neural networks using the variational information bottleneck. *arXiv preprint arXiv:1802.10399*, 2018.

Z. Goldfeld, E. van den Berg, K. H. Greenewald, I. Melnyk, N. Nguyen, B. Kingsbury, and Y. Polyanskiy. Estimating information flow in neural networks. *arXiv:1810.05728v3 [cs.LG]*, Nov. 2018.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *NIPS*, 2015.

Babak Hassibi and David G Stork. Second order derivatives for network pruning: Optimal brain surgeon. *NIPS*, 1993.

Yang HH and Moody J. Feature selection based on joint mutual information. *Advances in intelligent data analysis, proceedings of international ICSC symposium*, pp. 22–25, 1999.

A. Kolchinsky and B.D. Tracey. Estimating mixture entropy with pairwise distances. entropy. *arXiv:1705.02436v7 [cs.IT]*, 19, 2017.

A. Kolchinsky, B. D. Tracey, and D. H. Wolpert. Nonlinear information bottleneck. *arXiv:1705.02436v7 [cs.IT]*, Sep.2018.

A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical Review E,69:066138*, 2004.

Yann LeCun, John S Denker, and Sara A Solla. Optimal brain damage. *NIPS*, 1990.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *ICLR*, 2017.

Yuchao Li, Shaohui Lin, Baochang Zhang, Jianzhuang Liu, David Doermann, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Exploiting kernel sparsity and entropy for interpretable cnn compression. *arXiv:1812.04368*, 2019.

Jian-Hao Luo and Jianxin Wu. An entropy-based pruning method for cnn compression. *arXiv preprint arXiv:1706.05791*, 2017.

A. M. Saxe, Y. Bansal, J. Dapello, M. Advani, A. Kolchinsky, B. D. Tracey, and D. D. Cox. On the information bottleneck theory of deep learning. *Proc. International Conference on Learning Representations (ICLR), Vancouver*, May 2018.

R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. *Proc. IEEE Information Theory Workshop (ITW), Jerusalem*, pp. 1–5, Apr. 2015.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *Proc. Allerton Conf. on Communication, Control, and Computing, Monticello, IL*, pp. 368–377, Sep. 1999.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. technical report, in california institute of technology., 2011.

Shujian Yu, Kristoffer Wickstrøm, Robert Jenssen, and Jose C Principe. Understanding convolutional neural networks with information theory: An initial exploration. *arXiv preprint arXiv:1804.06537*, 2018.

Bolei Zhou, Agata Lapedriza1, Jianxiong Xiao2, Antonio Torralba1, and Aude Oliva1. Learning deep features for scene recognition using places database. NIPS, 2014.

A Appendix

You may include other additional sections here.