

GABOR FILTER INCORPORATED CNN FOR COMPRESSION

Akihiro Imamura
Embedded AI R&D department
DENSO CORPORATION
Shinagawa, Tokyo, Japan
akihiro.imamura.j5f@jp.denso.com

Nana Arizumi
Embedded AI R&D department
DENSO CORPORATION
Shinagawa, Tokyo, Japan
nana.arizumi.j8j@jp.denso.com

Abstract—Convolutional neural networks (CNNs) are remarkably successful in many computer vision tasks. However, the high cost of inference is problematic for embedded and real-time systems, so there are many studies on compressing the networks. On the other hand, recent advances in self-attention models showed that convolution filters are preferable to self-attention in the earlier layers, which indicates that stronger inductive biases are better in the earlier layers. As shown in convolutional filters, strong biases can train specific filters and construct unnecessarily filters to zero. This is analogous to classical image processing tasks, where choosing the suitable filters makes a compact dictionary to represent features. We follow this idea and incorporate Gabor filters in the earlier layers of CNNs for compression. The parameters of Gabor filters are learned through backpropagation, so the features are restricted to Gabor filters. We show that the first layer of VGG-16 for CIFAR-10 has 192 kernels/features, but learning Gabor filters requires an average of 29.4 kernels. Also, using Gabor filters, an average of 83% and 94% of kernels in the first and the second layer, respectively, can be removed on the altered ResNet-20, where the first five layers are exchanged with two layers of larger kernels for CIFAR-10.

Index Terms—compression, CNN, Gabor

I. INTRODUCTION

In recent years, convolutional neural networks (CNNs) have achieved remarkable success in computer vision tasks, e.g. image classification [1], object detection [2], semantic segmentation [3]. The networks have grown deeper [4], [5], and wider [6] to achieve higher accuracy in these tasks. Hence, the networks are increasing to have more parameters with a high cost of inference operations. It is a major problem for mobile devices and embedded systems applications, which has a limitation in memory, power, and computation speed. Also, it is an issue for real-time systems like image classification for web services, which also has a time budget to process inferences.

Hence, there has been a significant amount of works on model compression, and acceleration [7], [8]. Among which, network pruning from well-trained CNN has been widely studied [9]–[11]. There are different granularities of pruning. Fine-grained level (e.g., weight level) sparsity gives the highest flexibility and has high compression. However, it also gives unstructured sparsity and needs specialized hardware or/and sparse libraries to harvest the best efficiency from the com-

pression [11]–[13]. On the contrary, structured pruning does not require them. Hence, there are increasingly more works on channel-level pruning [14]–[16] and kernel-level pruning [17]. Note that the channel-level pruning in one layer also makes kernel-level pruning in the next layer. Layer-level pruning is also possible for deep CNNs [18], [19]. In this work, we also focus on pruning without introducing unstructured sparsity.

The reason why we can prune the weights is that after training, the filters become small enough to be ignored; however, we do not know which to be pruned without training the network [20]. Hence, we can assume that unnecessary filters become zero by training with weight decay. However, the filters might still have redundancy. The success of compressing the network using low-rank decomposition of the convolutional layers [21], [22] indicates that features required to represent the layer are relatively small.

The recent advances in self-attention models showed that convolution filters are preferable to self-attention filters in the earlier layers [23]–[25], which indicates that stronger inductive biases are better in the earlier layers. So, additional inductive biases might help choose the right filters in the earlier layers. Since it has been known that the networks are trained to become the Gabor filters in the earlier layers of CNNs [1], we can insert additional biases by enforcing filters to be the Gabor filters.

We observed that the first layer of VGG-16 on CIFAR-10 requires an average of $13.2 \times 7 \times 3$ channels of Gabor kernels without accuracy degradation. On the contrary, the original VGG-16 has $64 \times 3 \times 3$ channels in the first layer. Hence, in the first two layers, it makes 24% computation and memory reduction. Even more, an average of $162.6 \times 7 \times 7$ kernels out of 192 can be removed without accuracy degradation using kernel-level pruning. Similar behavior is observed for ResNet-20 on CIFAR-10. We extended this idea to the second layer and experimented with altered ResNet-20, where the first five layers are exchanged with two layers of 7×7 and 5×5 without shortcut connection. It shows that 83% and 94% of kernels can be removed in the first and second layers, respectively, without accuracy degradation.

Using well-trained CNNs, we can fit and retrain the features by restricting to Gabor filters in the earlier layers. This way, we can insert stronger inductive biases in the earlier layers and

help to choose more specific features. As with classical feature detection, clear feature prediction can significantly reduce the size of CNNs. Finding better features and reduce redundant channels is an excellent way to compress CNNs, because it does not create unstructured sparse pruning. Furthermore, any other compressing methods can be applied after our channel-level pruning since it does not change the structure of CNNs.

II. RELATED WORK

The Gabor filters are the classical method that can extract local features [26], and it is particularly popular because of invariance in scale, rotation, and translation. Also, the linear convolution of the Gabor filters is a good representation of the simple cells in the human visual cortex [27], [28], which is a good indication that the early stage of feature extraction of the human visual system also uses Gabor filters. Analogously, the earlier layers of CNNs are similar to Gabor filters [1], [29]. Hence, there are several studies on utilizing Gabor filters with CNNs.

Gabor filters can be used as pre-processing of the input data [30]–[32], or initialization [33]. [34] uses fixed Gabor filters for the first or the second layer to reduce training complexity. [35] incorporated selected Gabor orientation filters to each layer to improve the robustness of the networks. [36] uses Gabor filters for the interpretability of the networks in applying person re-identification. They train Gabor filters' parameters but do not learn the scale or the center of the filters since their interests are on the features' interpretability and not to remove the features. In our work, we also used backpropagation to learn the parameters of Gabor filters in the network to enforce the layer to be the suitable feature detection.

III. METHODOLOGY

In recent CNNs, kernel sizes are mostly 1×1 and 3×3 . It is mostly because of computational efficiency. However, the first layer can have larger filters because the input images are usually one or three channels, and the computational cost in channel size is much smaller. In fact, there are several popular networks with larger first filters [1], [5]. We could expand the kernels in the first layer for the other CNNs with a smaller kernel. Expanding kernels itself should not change the accuracy drastically. Even though the kernel size becomes larger, if the number of channels compressed, then the total computation cost can be smaller than the original CNNs. Also, a larger kernel size means it covers larger receptive fields, so if the first few layers are only to expand receptive fields, then the layers can also be pruned. It is more common to use larger kernels in earlier layers for larger images.

Once the first layer's kernel is expanded and trained, we can fit the Gabor filters to the well-trained network. Then, train Gabor filter parameters using backpropagation. This way, we can enforce learned features to be Gabor filters. After retraining, prune channels using the L1 norm criterion as in [37].

Fig. 1 illustrates the schematic of this method in yellow. We also show the other methods for comparison. Here, "retrain"

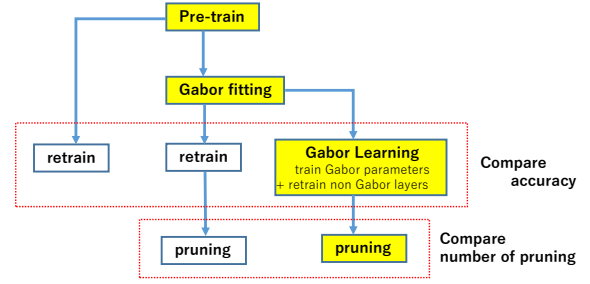


Fig. 1: Schematic of pruning methods

does not mean resuming the learning. We want to compare each method, so we retrain the network with the same hyper-parameters, including learning rates. Even though the network training is converged in the pretraining stage, it is known that the changing learning rate increases accuracy [38].

A. Gabor Fitting

The Gabor filter [39] is defined by a sinusoidal wave multiplied by the Gaussian function. We use the real part

$$G = a \exp\left(-\frac{\hat{x}^2 + \gamma \hat{y}^2}{2\sigma^2}\right) \cos\left(2\pi \frac{\hat{x}}{\lambda} + \psi\right), \quad (1)$$

where

$$\hat{x} = (x - x_0) \cos(\theta) + (y - y_0) \sin(\theta), \quad (2)$$

$$\hat{y} = -(x - x_0) \sin(\theta) + (y - y_0) \cos(\theta), \quad (3)$$

λ is the wavelength of the sinusoidal factor, θ is the orientation of the normal to the parallel stripes, ψ is the phase offset, σ is the standard deviation of the Gaussian envelope, γ is the spatial aspect ratio, a is the amplitude, and (x_0, y_0) is the center. Consider the kernels are located at the (x, y) grids from $(1, 1)$ to (k, k) , where k is the size of the kernel.

Construct Gabor filters using the following parameters and choose the closest to the original kernel using the L2 norm.

- amplitude a to be $-1, -0.5, 0, 0.5, 1$
- center (x_0, y_0) to be all the combination between 1 to size of kernel
- orientation θ and phase offset ψ to be $0, \pi/4, \pi/2, 3\pi/4$
- standard deviation of the Gaussian envelope σ and wavelength λ to be $1, 2, 3, 4, 5$
- aspect ratio γ to be $0.2, 0.4, 0.6, 0.8, 1$

B. Gabor Learning

Using backpropagation, we can train the parameters of Gabor filters, $\lambda, \theta, \psi, \sigma, \gamma, a, x_0$, and y_0 . The details of backpropagation are written in [36]. Contrary to their method, we did not set a range of values of parameters. Also, we add the amplitude and the center of the filters for learnable parameters. We intend to prune the filters, so amplitude should dissipate, and we can remove filters altogether. We did not use regularization loss of the Gabor parameter since we did not observe the difference. We used the Autograd package in PyTorch [40] to implement. For our experiments, the same learning rate is applied to all parameters of Gabor filters.

TABLE I: Accuracy of Gabor Learning

first layer	3×3	7×7
VGG-16		
pre-train	93.37 ± 0.09	93.23 ± 0.13
without fit + retrain	93.44 ± 0.18	93.30 ± 0.13
fit + retrain	-	93.39 ± 0.08
fit + Gabor learning	-	93.11 ± 0.08
ResNet-20		
pre-train	91.25 ± 0.14	90.80 ± 0.07
without fit + retrain	91.89 ± 0.09	91.48 ± 0.10
fit + retrain	-	91.41 ± 0.08
fit + Gabor learning	-	91.42 ± 0.10

IV. EXPERIMENTS

In CNNs, input feature maps X_i are convolved with channels C_i for n_{i+1} times and produce output feature maps X_{i+1} at the i th layer. The channels C_i has n_i kernels of the size $k_i \times k_i$. We use shorthand notation of $k_i \times k_i, n_i$ for the channels. In our experiments, all input images are 3 channels.

We used CIFAR-10 [41] image database, which has 32×32 natural images. Note, 7×7 kernels can cover a large part of the image. There are 5000 training images and 1000 test images. Standard data augmentation and normalization of input data are also applied.

We conducted experiments, as shown in Fig. 1 on two networks, VGG [42] and ResNet [5]. Both of which are widely popular networks and commonly used for pruning experiments. We followed VGG-16 from [37] and ResNet-20 from [43] for pre-processing, hyper-parameters, and learning schemes. It has been known that the network training is sensitive to initialization and order of training [44]. So, we ran all the experiments 5 times and took an average. We pruned channels and kernels that do not reduce accuracy by more than 0.2%. The channels and the kernels are chosen from the minimum according to the L1-norm. The testing for each case is operated with the same 79 batches with 128 images.

A. Gabor Fitting of first layer

For both VGG-16 and ResNet-20, we expanded 3×3 kernels of the first layer to 7×7 . As for ResNet-20, we remove two layers to fit the receptive field since it only has 0.5% degradation of accuracy. We did not do so for VGG-16, because it causes more than 2% of accuracy loss. First, we train VGG-16 for 350 Epochs twice with the same hyperparameters and ResNet-20 for 170 Epochs once, then fit the first layer to the Gabor filters. After retraining using conventional convolution and Gabor learning, we compare the accuracy as shown in Tab. I. Fitting does not degrade the accuracy.

The percentile of prunable kernels and channels are shown on Tab. II. For VGG-16, the percentile of channel pruning is the same as the results with 3×3 kernels without fitting presented in [37]. Note that for ResNet-20, two layers are also removed, but that is not shown on the table.

TABLE II: First layer pruning

	VGG-16		ResNet-20	
	channel	kernel	channel	kernel
retrain	49%	55%	14%	21%
Gabor learning	79%	91%	50%	75%

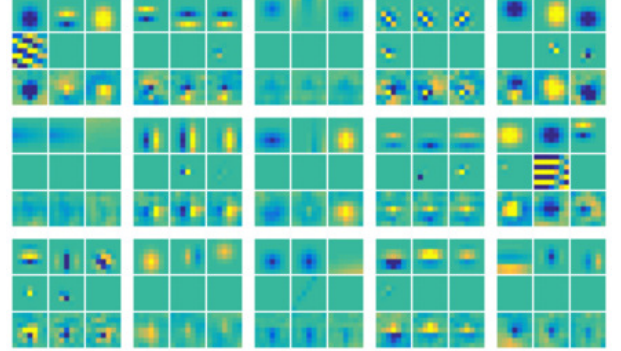


Fig. 2: Examples of kernels in the first layer channels of ResNet-20 for both conventional convolutional learning and Gabor learning after Gabor fit. Inside the 3×3 block, the first row shows after the Gabor fit, the second row shows the Gabor learning after Gabor fit, and the third row shows the conventional convolutional learning after Gabor fit. Each column shows filters for RGB images.

TABLE III: Accuracy of expanded ResNet-20 after second layer Gabor fitting

	7×7	5×5	accuracy
5×5 fitting	retrain	retrain	91.92 ± 0.13
	Gabor	retrain	91.88 ± 0.07
	retrain	Gabor	91.70 ± 0.06
	Gabor	Gabor	91.60 ± 0.42
no fitting	Gabor	retrain	91.96 ± 0.14

To show how the kernels were learned, Fig. 2 shows examples of kernels in the first layer channels of ResNet-20 for both conventional convolutional learning and Gabor learning after Gabor fit. Each kernel is normalized, and the colormap is shown in 6σ . For each block with 3×3 kernels, the first row shows after the Gabor fit, the second row shows the Gabor learning after Gabor fit, and the third row shows the conventional convolutional learning after Gabor fit. Both of the learnings are starting from the first row. The columns are corresponding to the 3 input channels of RGB images. It clearly shows that conventional convolutional learning does not change much from the Gabor fit. Still, many of the kernels after Gabor learning dissipates, and some form different structures. However, once the kernels are close to zero in initialization, it does not change to other structures. It seems that unnecessary filters dissipate by weight decay.

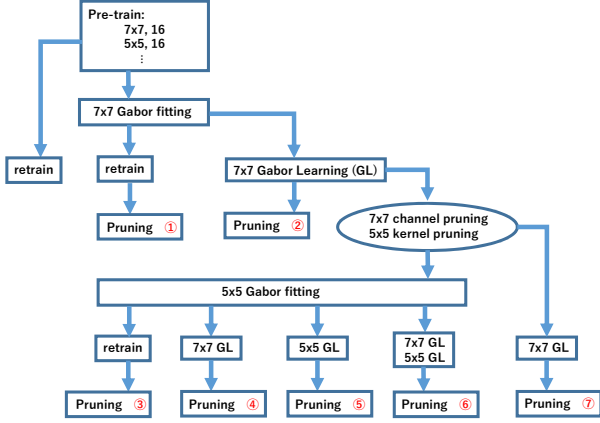


Fig. 3: Schematic of second layer Gabor learning

TABLE IV: Second layer prunings

		kernel	level	pruned
First run				
7×7 fitting	retrain ①	7×7	kernel	$10\% \pm 3\%$
			channel	$5\% \pm 5\%$
	GL ②	7×7	kernel	$63\% \pm 11\%$
			channel	$35\% \pm 3\%$
Second run				
5×5 fitting (7×7 results from GL)	7×7 retrain 5×5 retrain ③	7×7	kernel	$50\% \pm 2\%$
			channel	$49\% \pm 3\%$
		5×5	kernel	$61\% \pm 4\%$
			channel	$51\% \pm 11\%$
	7×7 GL 5×5 retrain ④	7×7	kernel	$78\% \pm 4\%$
			channel	$59\% \pm 3\%$
		5×5	kernel	$59\% \pm 2\%$
			channel	$41\% \pm 8\%$
	7×7 retrain 5×5 GL ⑤	7×7	kernel	$70\% \pm 5\%$
			channel	$66\% \pm 7\%$
		5×5	kernel	$94\% \pm 1\%$
			channel	$71\% \pm 13\%$
	7×7 GL 5×5 GL ⑥	7×7	kernel	$83\% \pm 3\%$
			channel	$60\% \pm 6\%$
		5×5	kernel	$94\% \pm 1\%$
			channel	$63\% \pm 6\%$
5×5 no fitting	7×7	kernel	$79\% \pm 5\%$	
		channel	$56\% \pm 6\%$	
	5×5	kernel	$61\% \pm 7\%$	
		channel	$41\% \pm 8\%$	

B. Gabor Fitting of first two layers

We have shown that learning the Gabor parameters in the first layer chooses fewer features. We expand this idea to the second layer. The schematic of the training and pruning is in Fig. 3. Each number in the figure indicates the experiments in the following table.

The first five layers of ResNet-20 are all 3×3 kernels with

16 output channels, which we denote $3 \times 3, 16$. We expanded the kernels and removed layers to make it to two layers of $7 \times 7, 16$, and $5 \times 5, 16$ without shortcut connection. We trained the network four times, with the same hyperparameters as the previous run, and used this as a reference. Note, the original ResNet-20 has an accuracy of 92.38 ± 0.18 after four runs, and it also shows convergence. The Gabor fitting is the same as in previous cases, except the scale a , which is scaled with each kernel's maximum absolute value.

We first prune unnecessary $7 \times 7 \times 3$ input channels from Gabor learnings. To do so, we also pruned kernels in the second layer. Then, we repeat the Gabor fitting for 5×5 kernels.

Table III shows the accuracy of all the combinations of learnings with 5×5 Gabor fitting after the first Gabor learning. All of them have similar results, i.e., less than 0.4% difference. Table IV shows the total number of average pruning after the second run. Their results include the pruning of $7 \times 7 \times 3$ channels in the first Gabor learnings, and it shows that further prunings of the first layer are possible. In the second layer, input features are no longer correlated to each other like in the case of RGB inputs; hence the standard deviation of prunable channels is large.

Using second Gabor learning in either the first or the second layer, ④, ⑤, and ⑥, makes significantly more prunable channels/kernels in the first layer. It might be because the first Gabor learning makes the first layer close enough to Gabor filters.

To check the influence on Gabor fitting in the second layer, we retrained the network without 5×5 Gabor fitting and compared it with the network with 5×5 Gabor fitting, i.e., comparing ④ and ⑦. It shows that 5×5 Gabor fitting does not make a clear difference in the number of prunable kernels without Gabor learning. It implies that the structure of parameter space, not just the initial values, is essential for learning.

Same as the first layer, 5×5 Gabor learning, ⑤, and ⑥, can make significantly more prunable channels/kernels. It could be possible to extend Gabor learning to further layers.

V. CONCLUSION AND DISCUSSION

In this paper, we propose to retrain networks by Gabor learning in the earlier layers of CNNs to decrease the number of features/channels. Using strong inductive biases by restricting features to Gabor filters help networks to choose specific filters. As with classical image processing, we showed choosing the right features for filters can make a smaller number of filters after learning.

Our method does not compete with the other compressing methods since it does not alter the structure of the networks. Hence, it can be used as a good preprocessing for network compression.

REFERENCES

- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

- [2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.
- [3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [4] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
- [5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [6] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
- [7] Y. Cheng, D. Wang, P. Zhou, and T. Zhang, "A survey of model compression and acceleration for deep neural networks," *arXiv preprint arXiv:1710.09282*, 2017.
- [8] J. Cheng, P.-s. Wang, G. Li, Q.-h. Hu, and H.-q. Lu, "Recent advances in efficient computation of deep convolutional neural networks," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 64–77, 2018.
- [9] Y. LeCun, J. S. Denker, and S. A. Solla, "Optimal brain damage," in *Advances in neural information processing systems*, 1990, pp. 598–605.
- [10] B. Hassibi and D. G. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," in *Advances in neural information processing systems*, 1993, pp. 164–171.
- [11] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.
- [12] S. Han, J. Pool, J. Tran, and W. Dally, "Learning both weights and connections for efficient neural network," in *Advances in neural information processing systems*, 2015, pp. 1135–1143.
- [13] Y. Guo, A. Yao, and Y. Chen, "Dynamic network surgery for efficient dnns," in *Advances in neural information processing systems*, 2016, pp. 1379–1387.
- [14] Y. He, X. Zhang, and J. Sun, "Channel pruning for accelerating very deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1389–1397.
- [15] Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2736–2744.
- [16] J.-H. Luo, J. Wu, and W. Lin, "Thinet: A filter level pruning method for deep neural network compression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5058–5066.
- [17] S. Anwar, K. Hwang, and W. Sung, "Structured pruning of deep convolutional neural networks," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 13, no. 3, pp. 1–18, 2017.
- [18] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *European conference on computer vision*. Springer, 2016, pp. 646–661.
- [19] W. Wen, C. Wu, Y. Wang, Y. Chen, and H. Li, "Learning structured sparsity in deep neural networks," in *Advances in neural information processing systems*, 2016, pp. 2074–2082.
- [20] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *arXiv preprint arXiv:1803.03635*, 2018.
- [21] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. De Freitas, "Predicting parameters in deep learning," in *Advances in neural information processing systems*, 2013, pp. 2148–2156.
- [22] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 1943–1955, 2015.
- [23] S. d'Ascoli, H. Touvron, M. Leavitt, A. Morcos, G. Biroli, and L. Sagun, "Convit: Improving vision transformers with soft convolutional inductive biases," *arXiv preprint arXiv:2103.10697*, 2021.
- [24] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *arXiv preprint arXiv:1906.05909*, 2019.
- [25] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16519–16529.
- [26] A. K. Jain, N. K. Ratha, and S. Lakshmanan, "Object detection using gabor filters," *Pattern recognition*, vol. 30, no. 2, pp. 295–309, 1997.
- [27] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *JOSA A*, vol. 2, no. 7, pp. 1160–1169, 1985.
- [28] J. P. Jones and L. A. Palmer, "An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex," *Journal of neurophysiology*, vol. 58, no. 6, pp. 1233–1258, 1987.
- [29] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 2011, pp. 215–223.
- [30] H. Yao, L. Chuyi, H. Dan, and Y. Weiyou, "Gabor feature based convolutional neural network for object recognition in natural scene," in *2016 3rd International Conference on Information Science and Control Engineering (ICISCE)*. IEEE, 2016, pp. 386–390.
- [31] A. Calderón, S. Roa, and J. Victorino, "Handwritten digit recognition using convolutional neural networks and gabor filters," *Proc. Int. Congr. Comput. Intell.*, 2003.
- [32] B. Kwolek, "Face detection using convolutional neural networks and gabor filters," in *International Conference on Artificial Neural Networks*. Springer, 2005, pp. 551–556.
- [33] S.-Y. Chang and N. Morgan, "Robust cnn-based speech recognition with gabor filter kernels," in *Fifteenth annual conference of the international speech communication association*, 2014.
- [34] S. S. Sarwar, P. Panda, and K. Roy, "Gabor filter assisted energy efficient fast learning convolutional neural networks," in *2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*. IEEE, 2017, pp. 1–6.
- [35] S. Luan, C. Chen, B. Zhang, J. Han, and J. Liu, "Gabor convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 9, pp. 4357–4366, 2018.
- [36] Y. Yuan, J. Zhang, and Q. Wang, "Deep gabor convolution network for person re-identification," *Neurocomputing*, vol. 378, pp. 387–398, 2020.
- [37] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.
- [38] L. N. Smith, "Cyclical learning rates for training neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 464–472.
- [39] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.
- [40] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in neural information processing systems*, 2019, pp. 8026–8037.
- [41] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [43] N. Zmora, G. Jacob, L. Zlotnik, B. Elharar, and G. Novik, "Neural network distiller," Jun. 2018. [Online]. Available: <https://doi.org/10.5281/zenodo.1297430>
- [44] P. Krähenbühl, C. Doersch, J. Donahue, and T. Darrell, "Data-dependent initializations of convolutional neural networks," *arXiv preprint arXiv:1511.06856*, 2015.