IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received November 28, 2021, accepted December 12, 2021, date of publication December 14, 2021,
date of current version December 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3135576

Kernel Product Neural Networks

HAO XU[1], SHUYUE ZHOU[1], YANG SHEN[1], KENAN LOU[2], RUIHUA ZHANG[1], ZHEN YE[1], XIAOBO LI[3], AND SHUAI WANG[4]

[1]College of Engineering, Lishui University, Lishui 323000, China
[2]College of Business, Lishui University, Lishui 323000, China
[3]College of Mathematics and Computer Science, Zhejiang Normal University, Jinhua 321004, China
[4]LPL, CNRS, Aix-Marseille University, 13100 Aix-en-Provence, France

Corresponding authors: Kenan Lou (nanekuol@126.com) and Yang Shen (tlsheny@163.com)

This work was supported in part by the Baishanzu National Park Scientific Research Project under Grant 2021KFLY07, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21F020004, in part by the Key Research and Development Projects of Science and Technology Bureau in Lishui City, China, under Grant 2019ZDYF04, in part by the National Natural Science Foundation of China under Grant 61373057, in part by the Science and Technology Planning Project of Lishui City under Grant 2019RC05, and in part by the Science and Technology Public Welfare Project of Lishui City under Grant 2021SJZC005.

ABSTRACT Attention is an important field to explore the importance of each convolutional kernel channel/weight. The existing attention methods mostly use the Squeeze-and-Excitation (SE) technology to extract the global nonlinear feature vectors as the weights of corresponding feature maps. However, the pooling operators and fully-connected layers used in SE technology extract global features at the cost of much valuable information loss and the parameter amount increase. Actually, the feature map containing full information is a ready-made and better attention for other feature maps in the same layer. Simultaneously the products of feature maps will bring powerful non-linearity. Seeing this, Kernel Product (KP) technology is proposed to simply get useful nonlinear attention. To verify the effectiveness of KP, the proposed KP module is employed on Selective Kernel Networks (SKNets) to take the place of the original SE technology. The variety of SKNets is called Kernel Product Networks (KPNets) in this paper. In addition, identity mapping is used to solve the non-convergence problem in very deep neural networks. The KPNets are evaluated on ImageNet-1k, CIFAR-10, and CIFAR-100. The experiment results show that KPNets outperform many state-of-the-art methods and get a similar but more efficient performance than its SKNets with counterpart.

INDEX TERMS Attention, non-linearity, kernel product.

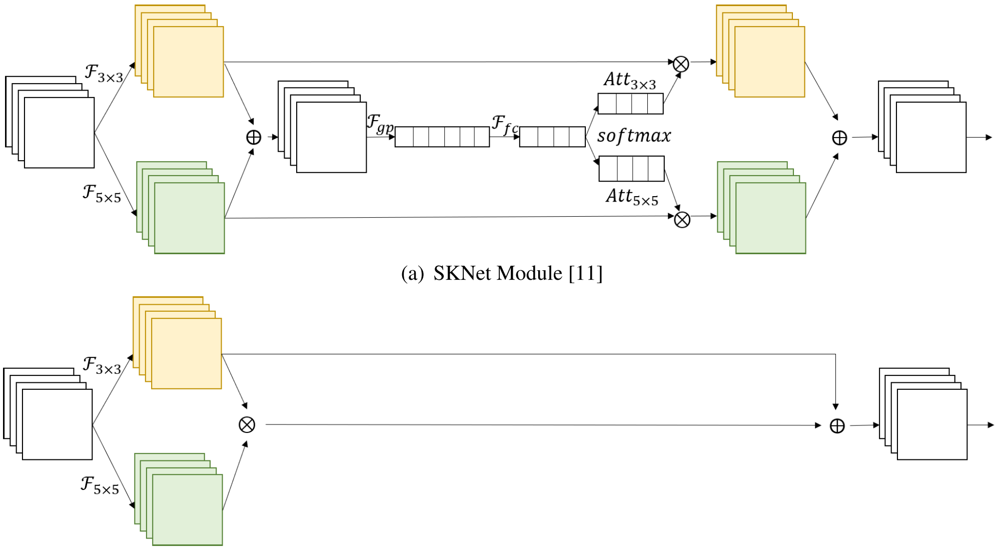
I. INTRODUCTION

Attention mechanism has been studied for many years in a variety of CV fields. It is motivated by the human vision mechanism in which visual neurons are partly activated by the salient objects in the seen scenes. Squeeze-and-Excitation technology proposed by [1] has been widely used to extract weights of corresponding feature map channels/weights [1], [2]. As reported in [1], the test classification accuracy of ImageNets has been improved substantially compared with other methods at that time. Furthermore, besides classification tasks, attention mechanism has been also used in many other tasks such as object detection [3], [4], semantic segmentation [5], [6], super resolution [7], [8], action recognition [9], [10], etc. As the most popular attention method, SE technology used pooling operators to achieve the invariant feature of each channel, bringing nonlinearity at the same time. Then the linear layers are used to exchange information across channels. Its nature is an explicit representation of the importance of convolutional kernel channels/weights. SE technology has achieved great success but some questions have still been proposed.

A natural question is yielded: why the attention learned by SE technology has not been automatically learned in the vanilla convolutional kernels? A plausible explanation is that the extra global information and non-linearity brought by SE technology make the pattern of fully trained SE kernels more robust to unseen data and noise. The Convolutional Neural Networks is based on an important assumption, that is the learned good weight values in one local region are also fit in other regions of the whole image [12]. However, not all extracted local features can represent the global features. Much global information can not be fused in the convolution feature representation because of the intrinsic local characteristics of convolutional kernels.

SE technology actually alleviates this local problem of convolutional kernels. SE technology extracts more useful global information through global pooling operators together with

The associate editor coordinating the review of this manuscript and approving it for publication was Joewono Widjaja.



(a) SKNet Module [11]

(b) KPNet Module

FIGURE 1. (a) SKNets fuses two SE modules to get different attention for feature maps convoluted by different kernel sizes [11]. (b) KPNets drops the middle operators and uses the product of feature maps convoluted by different kernel sizes. The activation functions used in the module are not given in the figure.

training, and then the extracted global information is used to make up the local characteristics of vanilla convolutional kernels by multiplying the extracted weights and the corresponding feature maps together. Moreover, the non-linearity brought by the global pooling operators and multiplication makes the SE modules more robust to unseen data. The non-linearity and fused global information are the keys that make Squeeze-and-Excitation technology work.

However, the SE technology extracts the global information of output feature maps at the cost of much useful information loss caused by the global pooling operators [12]. To maintain the valuable global information, the layer-wise output feature maps containing abundant global information are taken into consideration. This information has not been used amply and properly by existing attention methods. SE technology uses the multiply operators to impose the yielded weights on the corresponding feature maps. Similar to SE technology, the output feature maps in the same layer are multiplied together to retain these advantages of SE, formally $f(x) = \langle f_l^1(x), f_l^2(x)\rangle$.

Regarding the non-linearity, except global maximum pooling operators, many other technologies also can bring CNNs with nonlinearity. Nonlinear functions, such as ReLU, sigmoid, etc., are also used to introduce non-linearity into convolutional kernels and fully-connected layers. Moreover, except point-wise nonlinear functions, patch-wise nonlinear functions have been proposed to impose nonlinear kernels, such as polynomial kernels, Gaussian kernels, *etc.*, on inner product operators [13].

For instance, in [13], $f(x)\cdot f(x)$ is used to yield non-linearity in the polynomial kernels. The above multiplication, i.e., $f(x) = \langle f_l^1(x), f_l^2(x)\rangle$, can also bring non-linearity. Additionally, to fuse the extracted information in the corresponding feature maps, fully-connected layers are used to be multiplied by the original feature maps along with the channels. The fully-connected layers are also used to get better weights of the corresponding feature maps by training.

Having shown these, the product of feature maps in the same layer could be used directly to make full use of the information contained in output feature maps by considering the feature maps in the same layer as the attention of other feature maps. The product of different feature maps is able to bring enough non-linearity to the whole model.

To verify the effectiveness of the proposed Kernel Product (KP) technology, KP is embedded in the Selective Kernel module to replace the original SE technology. As shown in Fig. 1, the difference between SK module and KP module explicitly displays that many middle operators of SK module are dropped in KP module.

Our main contributions are listed as follows.

1) A new attention mechanism, which is called Kernel Product (KP) technology, is proposed to replace the original SE module. KP drops the pooling operators and linear layers of SE technology. Instead, the product of feature maps in the same layer is used to yield non-linearity and fuse the global information. KP modules contain fewer parameters and are simpler than SE modules.
2) The gradient and loss landscape analyses in Section III-B show that KP modules show have more powerful optimizable ability than SK modules. Also, Meanwhile KP modules keep share the similar representation ability as SK modules.
3) The feasibility of kernel products is discussed by demonstrating that KP can introduce enough non-linearity without dropping much information in

original feature maps. Also, the convergence problem brought by the multiplication is addressed by the identity transformation, which makes KP modules avoid gradient propagation problems occurring in the deep networks.

II. RELATED WORKS

A. ATTENTION MECHANISM

Attention mechanism in CNNs originates from Squeeze-and-Excitation Networks [1]. It is a lightweight module that can be embedded in the vanilla convolutional layers. The module imposes the extracted channel weights on the corresponding feature maps. It is inspired by originates from the neuroscientific study on the visual cortex [14], and motivates numerous studies in deep learning. Details of the basic SE technology are shown in Section III-A1.

Woo *et al.* proposed CBAM [2] where spatial attention and channel attention can be jointly used to get more robust models. In [2], series connections of spatial and channel attention get better performance in contrast to parallel connections. BAM [15] is similar to CBAM. In this paper, channel attention and spatial attention are extracted simultaneously and then are combined to be 3-D attention maps with the same shape as the original feature maps before the attention maps are imposed on the corresponding feature maps.

SKNets [11] used SE technology to extract different attentions followed by a softmax function from a fused feature map. The fused feature map consists of feature maps convoluted with different size kernels. Then the extracted channel attentions are imposed on the corresponding feature maps. As discussed in [16], the softmax function, which is used to map weights into probability form, might be an important point to improve the generalization of models.

In ECANet [17], Wang *et al.* argued that the reduction part used in SE module would bring a side effect with reference to the performance of SENet. To address this problem, parameter shared 1-D convolutional kernels are used to replace the original fully-connected layers, which is motivated by group convolution.

Switchable Attention [18] used a SN-like [19] view to fuse three existing attention mechanisms, including local spatial attention, global spatial attention, and channel attention. This method selects an appropriate attention mechanism for each layer by adaptive learning.

B. IDENTITY MAPPING

As convolutional neural networks go deeper, vanishing and exploding gradient problems will block the convergence of networks. ResNets [20] introduces identity mapping to CNNs, which makes networks go to 1000+ layers and optimizes the models better. Reference [21] analyzed the importance of identity mappings theoretically and experimentally. In this paper, He *et al.* [21] proved that signals could be directly propagated from one block to any block in this fully block stacked architecture during forward and backward processes.

Gao *et al.* proposed DenseNets [22] that all levels features should be connected by identity mappings. Identity mappings not only can address gradient propagation problems, but also fuse low-level and high-level feature maps. Different level features fusion brings very good generalization. As reported in [22], DenseNet-100 with only 0.8M parameters outperforms ResNet-1001 with 10.2M parameters.

According to these two advantages, identity mappings are used in KP module to alleviate the training divergence problem, which is brought by a lot of multiply operators in KP modules. Furthermore, identity mappings could bring better performance for KPNets without extra parameters.

C. GROUP CONVOLUTION

Group convolution was introduced in AlexNets [23] to deep learning to reduce the parameter amount of CNNs while maintaining the performance of the vanilla CNNs. ResNexts [24] and MobileNets [25] verified the effectiveness of group convolution with residual style networks and developed it into depth-wise style.

Similar to SK modules, group convolution is used in KP module to reduce the parameters and improve the representation ability by using larger width convolutional kernels and sharing more parameters at the same time.

III. METHOD

In this section, SE and KP modules are formulated firstly. Then, to verify the optimization and representation ability of KP modules, the visualization methods proposed in [26] are used. Lastly, the analysis experiments are conducted on feature maps and gradients to verify that KP modules get almost the same representation ability as SK modules.

A. FORMULATION

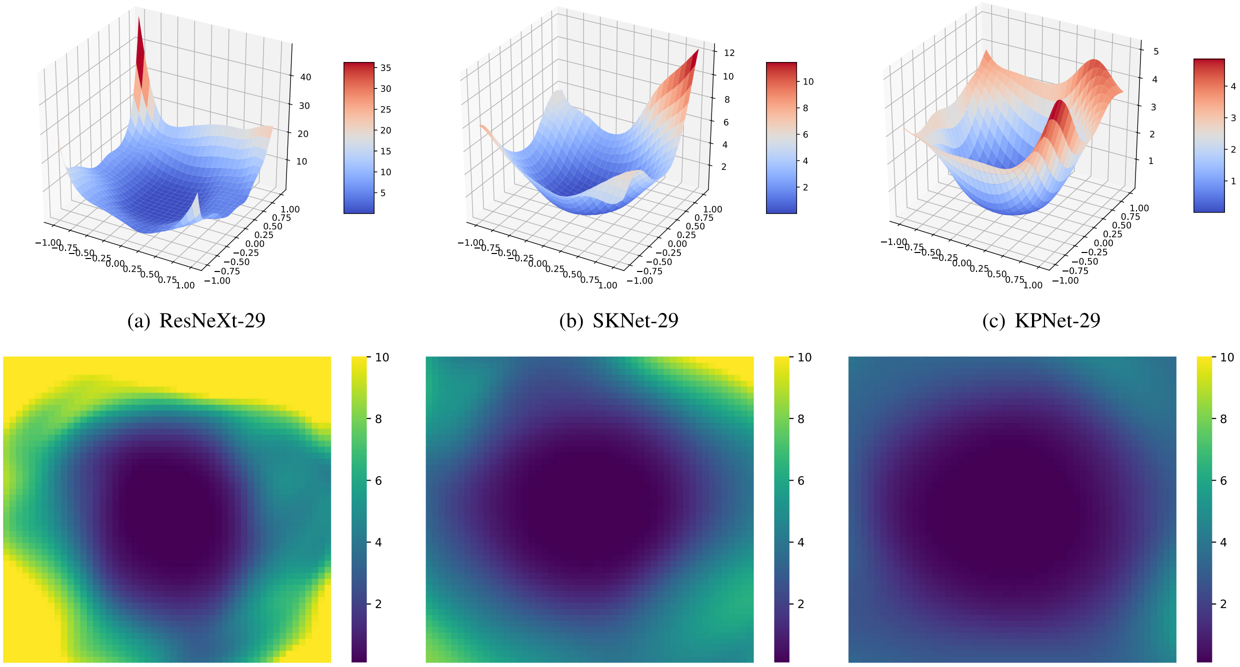
1) SE TECHNOLOGY FORMULATION

Firstly, the original SE module is formulated as follows. $X \in \mathbb{R}^{N\times C\times W\times H}$ denotes input feature maps and $Y \in \mathbb{R}^{N\times C\times W\times H}$ denotes the output feature maps weighted with attention. The attention yielded by SE modules is denoted by $Att \in \mathbb{R}^{N\times C}$. The linear layers used as dimensionality reduction and excitation is represented by a linear mapping, formally $Linear(X) = W \times X + B$, where $W, B \in \mathbb{R}^{C\times\frac{C}{4}\times C}$ represent weights and biases of the linear transformations respectively. The attention generation process is formulated as

$$\begin{aligned} Att &= W \times \mathcal{F}_{\mathcal{GP}}(X, k, s) + B, \\ Y &= Att \times X, \end{aligned}$$

where $\mathcal{F}_{\mathcal{GP}}(\cdot, \cdot, \cdot)$ denotes the global pooling operator, while k and s represent pooling sizes and stride sizes respectively.

As is shown in Fig. 1(a), SKNets, which is a variety of SENets. SKNets use different-size convolutional kernels to yield different feature maps with different receptive fields. Then the stack or sum of the yielded feature maps is used to extract the corresponding attention with the softmax function



(a) ResNeXt-29 (b) SKNet-29 (c) KPNet-29

(d) ResNeXt-29 (e) SKNet-29 (f) KPNet-29

FIGURE 2. The 3D loss landscape [26] visualization of (a) ResNeXt-29, (b) SKNet-29, and (c) KPNet-29 on CIFAR-10. For each point in the loss surface plots [26] of (d) ResNeXt-29, (e) SKNet-29, and (f) KPNet-29, we calculate the maximum and minimum eigenvalue of the Hessian, and map the ratio of these two.

for weighting the feature maps by the SE technology. Finally, the sum of the weighted feature maps is used as the output of SK module, which can select an appropriate size adaptively [11]. The added Softmax function convert makes the attention into probability which is more suitable for neural networks [11]. SK modules are formulated as

$$\begin{aligned} F_1 &= \mathcal{F}_{k_1\times k_1}(X), \quad F_2 = \mathcal{F}_{k_2\times k_2}(X),\\ F_X &= \mathcal{F}_{concat}(F_1, F_2),\\ [Y_1, Y_2] &= F_X \times Softmax\left(W \times \mathcal{F}_{\mathcal{GP}}(F_X, s_1, s_2) + B\right),\\ Y &= Y_1 + Y_2, \end{aligned} \tag{1}$$

where F_1, F_2 are feature maps convoluted by kernels with k_1, k_2 sizes. F_X is the stack or sum of F_1 and F_2. Both concatenation and added operators are linear and are only used to fuse point-wise information in both feature maps. Finally Squeeze-and-Excitation technology is used to get attentions for these two different kernel-size feature maps, i.e., adaptively selecting the suitable kernel size.

2) KERNEL PRODUCT FORMULATION

As shown Equation 1, information should be appropriately fused in a point-wise and global approach. Then the fused feature map is used to extract attentions for F_1 and F_2. The extracted attentions are viewed as importance weights to be multiplied by F_X. The shape of output feature maps is the same as X.

Actually, global information fusion, attention extraction, and weights multiplying can be replaced by the product of feature maps. The Kernel Product module is formulated as

$$Y = \prod_{i=0}^{n} \mathcal{F}_{k_i}(X) + \mathcal{F}_{3\times 3}(X), \tag{2}$$

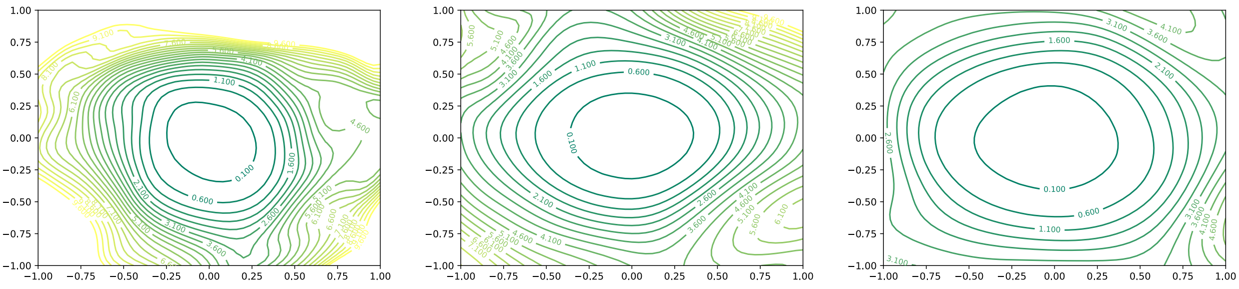
where $\mathcal{F}_{k_i}(\cdot)$ denotes different feature maps convoluted by kernels with size k_i. The product of different feature maps brings enough non-linearity and robustness.

In SKNets, different feature maps are yielded by different size convolutional kernels in the same layer. The different receptive field feature maps can be directly used, as shown in Equation 2. Full information in the feature maps can be maintained by dropping pooling operators, while parameters can be reduced by dropping fully-connected layers. The multi-size local and global information are completely fused in the final output, and the good knowledge is learned adaptively by an optimization algorithm.

B. REPRESENTATION ANALYSIS OF KP

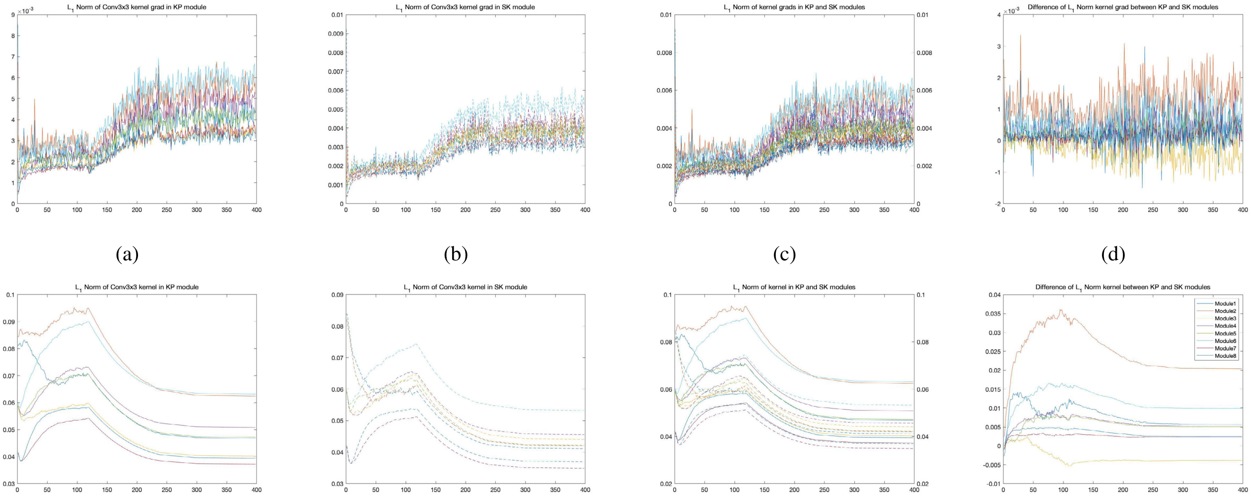
Firstly, non-linearity brought by KP module is shown by gradients of feature maps. We assume that only two different-size kernels are used, i.e. $n = 2$, formally

$$\begin{aligned} Y &= \mathcal{F}_{3\times 3}(X) \times \mathcal{F}_{5\times 5}(X) + \mathcal{F}_{3\times 3}(X)\\ \frac{\partial Y}{\partial \mathcal{F}_{3\times 3}} &= \mathcal{F}_{5\times 5}(X) \end{aligned} \tag{3}$$



(a) ResNeXt-29 (b) SKNet-29 (c) KPNet-29

FIGURE 3. **The 3D loss contour [26] visualization of ResNeXt-29, KPNet-29, and SKNet-29 on CIFAR-10. Similar to Fig. 2, obviously KPNet-29 performs best among these three models.**



(a) (b) (c) (d)

(e) (f) (g) (h)

FIGURE 4. **To Show KP module can be comparable to SK module with many operators reduction, the training trends of 3 × 3 kernels gradient of KP modules and SK modules are shown in (a) and (b) respectively measured by L_1 norm. In (c), both trends are shown together. And (d) gives the trend of difference between KP and SK. Similarly, (e), (f), (g), and (h) show the training trend and comparison of 3 × 3 kernels in both KP and SK modules by L_1 norm. Obviously, KP module is similar to SK.**

As shown in Equation 3, the partial derivative of output Y referring to convolutional kernel $\mathcal{F}_{3\times3}$ is $\mathcal{F}_{5\times5}(X)$, which is varied by the input X. Referring to SE technology, the partial derivative is $\frac{\partial Y}{\partial \mathcal{F}_{3\times3}} = Att_{3\times3}$. As shown in Fig. 1, $Att_{3\times3}$ is the attention of feature maps of convolutional kernel with size 3. The non-linearity brought by SE technology is also induced by the partial derivative which is varied by input X.

From the view of partial derivative, we speculate that the nonlinearity brought by attention is actually caused by the product of different information. If attention information is added to the target feature maps, i.e. $Y_{3\times3} = \mathcal{F}_{3\times3} + Att_{3\times3}$, the non-linearity is only brought by global pooling and activation functions in the $Att_{3\times3}$ part.

For verifying the optimization property and generalization of the proposed KPNets, loss landscape visualization method is used. Li *et al.* proposed filter normalization [26] to visualize the loss landscape of the target models. As argued in [26], the better minimizers can be simply obtained if the training loss landscape is flatter. These better minimizers tend to lead to better generalization of the fully trained models. As shown in Fig. 3, obviously the contours of KPNet-29 loss landscape are flat, i.e., they can easily get good minimizers compared to SKNet-29 and ResNeXt-29. Fig. 2 gives the 3D surface of loss landscape and the loss surface plots of these three methods. If the landscape is flat, it means that the model is easy-to-optimizable. Also the generalization of the fully trained models is partly related to the optimized situations of models. In the loss surface plots, Blue color indicates a more convex region (near-zero negative eigenvalues relative to the positive eigenvalues), while yellow indicates significant levels of negative curvature. Combined with the scale of legends, obviously KPNet-29 performs best among these three models.

Moreover, to verify the viewpoint of non-linearity and the representation ability of the proposed KP module, we investigate the weight L_1 norm and gradient L_1 norm of convolutional kernel with size 3 in each KP module in 400 training epochs respectively. As shown in the top row of Fig. 4, the training trends of kernel weight norm and kernel gradient norm in KPNet-29, which are shown in Fig. 4(e) and 4(a), are very similar to the trends of SKNet-29, which are shown in Fig. 4(f) and 4(b). The difference between these two weight norm trends is shown in the top-right subfigure of Fig. 4, in which only the orange line (i.e., the first KP module) generates high deviation. As demonstrated in the Fig. 4, we verify that the representation ability of KP module is equivalent to SK module. In addition, the difference between modules in the deep layers, which contribute to the final decisions, is consistently observed around zeros. This evidence supports that KPNets achieve similar results as SKNets at the training stage.

IV. EXPERIMENTS

A. BASIC SETTINGS

To show the performance of our proposed KP module effectively, many classification experiments are conducted on the public datasets, including ImageNet-1k, CIFAR-10, and CIFAR100. Moreover, the training strategies and used hyperparameters are set to be the same as SKNets for experimental impartiality.

B. ImageNet CLASSIFICATION

For ImageNet-1k, which comprises 12.8 million training images and 50K validation images from $1,000$ classes, only Top-1 errors on validation set are reported and only training images with size 224×224 are used because of our computation resource limit. On data augmentation, to align with most state-of-the-art attention methods, we follow the standard practice and perform the random-size cropping to 224×224 and random horizontal flipping [11]. The practical mean channel subtraction is adopted to normalize the input images for both training and testing. Label-smoothing regularization [27] is used during training. In ImageNet-1k experiments, synchronous SGD with momentum 0.9 to training KPNet. Batch size is set to 256 and a weight-decay is set to 1e-4 for the network robustness. The initial learning rate is set to 0.1 and decreases by 10 every 30 epochs. All models are trained for 100 epochs from scratch, using the weight initialization strategy in [28]. Same as SKNet-50, the initial kernel channel number are set to 128 and group number is set to 32, which represents for group convolution. The results reported on ImageNet-1k are the averages of 3 runs by default.

From the classification accuracy aspect, as shown in Table 1, KPNets achieves better results than the compared SOTA methods. For 50 convolutional layers networks, KPNet-50 outperforms ResNeXt-50, SENet-50 by 1.41% and 0.3% respectively. For 101 layers networks, KPNet-101 outperforms ResNeXt-101, SENet-101 by 0.86% and 0.33% respectively. Compared with SKNets, the performance of KPNets is similar to SKNets while KPNets keep fewer parameters.

TABLE 1. Comparison results of KP and many state-of-the-art methods on ImageNet-1k.

Methods	Top-1 err	#P	GFLOPs
ResNeXt-50 [24]	22.23	25.0M	4.24
SENet-50	21.12	27.7M	4.25
SKNet-50	20.79	27.5M	4.47
KPNet-50	20.82	25.2M	4.51
ResNeXt-101	21.11	44.3M	7.99
SENet-101	20.58	49.2M	8.0
SKNet-101	20.19	48.9M	8.46
KPNet-101	20.25	46.4M	8.52

TABLE 2. Comparison results of KP and other state-of-the-art methods on CIFAR-10 and 100.

Methods	CIFAR-10	CIFAR-100	# P
ResNeXt-29	3.87	18.56	25.2M
SENet-29	3.68	17.78	35.0M
SKNet-29	3.47	17.33	27.7M
KPNet-29	3.45	17.37	25.5M

TABLE 3. Inference performance of different methods requirements, latency is measured with batch size 1 on a single core Intel CPU 10920X.

Methods-CIFAR100	Latency	Memory Footprint
SKNet29-CIFAR00	35.15ms	435M
KPNet29-CIFAR100	11.92ms	396M

C. CIFAR CLASSIFICATION

To evaluate the performance of KPNets on smaller datasets, more experiments are conducted on CIFAR-10 and 100 [29]. The training and test sets consist of 50k and 10k images respectively, and each image contains 32×32 pixels. To align with SKNets, all CIFAR experiments are conducted on 2 GPUs with batch size 128 for 300 epochs. The learning rate is initialized to 0.1 for CIFAR-10 and 0.05 for CIFAR-100, and is divided by 10 at 150 epoch and 200 epoch. Following [20], the weight decay is set to 5e-4 and the momentum is set to 0.9. In order to prevent overfitting on these small datasets, 5×5 kernels are replaced by 1×1 [11].

The results in Table 2 indicate that KPNets and SKNets outperform ResNeXts and SENets by a large margin. Compared with SKNets, KPNets perform best on CIFAR-10 and have the similar performance to SKNets on CIFAR-100 with fewer parameters.

In addition, to further validate the efficiency of KPNets, the inference latency and memory footprint are computed for both SKNets and KPNets which are comparable due to the similar classification performance. As shown in Table 3, KPNets run 2.9× faster than the SKNets on CIFAR-100. Furthermore, KPNets have fewer memory footprint than SKNets.

TABLE 4. Comparison results of the different kernel size selection, 1 × 1 and 5 × 5.

Datasets	1 × 1	5 × 5
CIFAR-10	3.45	4.25
CIFAR-100	17.37	18.24

These results indicate that KPNets are really fast on the real hardware.

For verifying the argument that the kernel size 1 × 1 is more suitable for CIFAR datasets, the results of ablation experiments given in Table 4 show that the kernel sizes of the convolutional layers should be set to 3 × 3 and 1 × 1. This set outperforms 3 × 3 and 5 × 5 by 0.8% and 0.87% on CIFAR-10 and CIFAR-100, respectively.

V. CONCLUSION

We proposed an alternative attention mechanism called Kernel Product to replace Selective Kernel. This mechanism drops global pooling operators and linear layers, with the product of feature maps in the same layer being attentions of other feature maps. The product of feature maps can yield nonlinearity and global information fusion, which are two relatively critical aspects. According to the loss landscapes of KPNets, the proposed KP modules can get minimizers more easily than SK modules with fewer parameters. Additionally, in terms of the inference latency and memory footprint, KPNets are more efficient than SKNets.

REFERENCES

[1] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[2] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[3] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7289–7298.

[4] Y. Zhang, P. Zhao, D. Li, and K. Konstantin, "Spatial attention based real-time object detection network for Internet of Things devices," *IEEE Access*, vol. 8, pp. 165863–165871, 2020.

[5] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, "Context encoding for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7151–7160.

[6] Y. S. Hariyani, H. Eom, and C. Park, "DA-CapNet: Dual attention deep learning based on U-Net for nailfold capillary segmentation," *IEEE Access*, vol. 8, pp. 10543–10553, 2020.

[7] M. Suganuma, X. Liu, and T. Okatani, "Attention-based adaptive selection of operations for image restoration in the presence of unknown combined distortions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9039–9048.

[8] A. Muqeet, M. T. B. Iqbal, and S.-H. Bae, "Hran: Hybrid residual attention network for single image super-resolution," *IEEE Access*, vol. 7, pp. 137020–137029, 2019.

[9] H. Sang, Z. Zhao, and D. He, "Two-level attention model based video action recognition network," *IEEE Access*, vol. 7, pp. 118388–118401, 2019.

[10] Z. Shi, L. Cao, C. Guan, H. Zheng, Z. Gu, Z. Yu, and B. Zheng, "Learning attention-enhanced spatiotemporal representation for action recognition," *IEEE Access*, vol. 8, pp. 16785–16794, 2020.

[11] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 510–519.

[12] S. Sabour, N. Frosst, and G. Hinton, "Dynamic routing between capsules," in *Proc. Conf. Workshop Neural Inf. Process. Syst.*, Oct. 2017, pp. 1–11.

[13] C. Wang, J. Yang, L. Xie, and J. Yuan, "Kervolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2019, pp. 31–40.

[14] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.*, vol. 160, no. 1, pp. 106–154, Jan. 1962.

[15] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.

[16] Y. Chen, X. Dai, M. Liu, D. Chen, and Z. Liu, "Dynamic convolution: Attention over convolution kernels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11030–11039.

[17] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1–3.

[18] Q. Cheng, H. Li, Q. Wu, F. Meng, L. Xu, and K. N. Ngan, "Learn to pay attention via switchable attention for image recognition," in *Proc. IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, Aug. 2020, pp. 291–296.

[19] P. Luo, J. Ren, Z. Peng, R. Zhang, and J. Li, "Differentiable learning-to-normalize via switchable normalization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019, pp. 1–19.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," 2016, *arXiv:1603.05027*.

[22] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImagNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[24] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1492–1500.

[25] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2017, pp. 1–9.

[26] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *Proc. Neural Inf. Process. Syst.*, 2018, pp. 1–18.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[29] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep, Jan. 2009, pp. 1–58.

HAO XU received the B.S. and M.S. degrees in information and communication engineering and the Ph.D. degree in computer engineering from Honam University, South Korea, in 2012, 2014, and 2017, respectively. He is currently a Lecturer with the College of Engineering, Lishui University, Zhejiang, China. His current research interests include deep learning and the IoT.

SHUYUE ZHOU received the B.S. degree in software engineering from Heilongjiang University and the M.S. degree from the College of Information Science and Technology, Ningbo University. He is currently a Lecturer with the College of Engineering, Lishui University, Zhejiang, China. His current research interests include multi-label classification imbalanced data and bioinformatics.

YANG SHEN received the Ph.D. degree in computer science from Shanghai Jiao Tong University, in 2011. He is currently an Associate Professor with the College of Engineering, Lishui University, Zhejiang, China. His current research interests include image processing and machine learning.

KENAN LOU received the M.S. degree in tourism studies and the Ph.D. degree in hotel tour from Honam University, South Korea, in 2013 and 2018, respectively. She is currently a Lecturer with the College of Business, Lishui University, Zhejiang, China. Her current research interests include wisdom tourism and machine learning.

RUIHUA ZHANG received the Ph.D. degree majoring in mechatronics engineering from the Harbin Institute of Technology, in 2006. She is currently a Professor with the School of Engineering, Lishui University. Her research interest includes intelligent manufacturing equipment and automation.

ZHEN YE received the B.S. and Ph.D. degrees in computer science and technology from Zhejiang University, China, in 2007 and 2013, respectively. He is currently a Lecturer with the College of Engineering, Lishui University, Zhejiang, China. His current research interest includes machine learning.

XIAOBO LI received the B.S. degree in microelectronics from Nankai University, China, in 1990, the Master of Engineering (Research) degree from The University of Sydney, Australia, in 2004, and the Ph.D. degree in pathology and pathophysiology from Zhejiang University, China, in 2012.

He is currently a full-time Professor and a Ph.D. Supervisor with the Department of Computer Science and Technology, College of Mathematics and Computer Science, Zhejiang Normal University, China. He has authored more than 40 articles. His current research interests include bioinformatics, machine learning, data mining, and tumor pathology.

Dr. Li is a member of the Bioinformatics Committee of China Computer Federation and the Bioinformatics and Artificial Life Committee of Chinese Association for Artificial Intelligence. He was a recipient of the Wu Wen Jun AI Science and Technology Award, in 2016, and the Science and Technology Award of China General Chamber of Commerce, in 2017.

SHUAI WANG received the B.S. degree in biological engineering from Shandong Agricultural University, in 2011, and the Ph.D. degree in cognitive neuroscience from East China Normal University, in 2019. He currently works at Aix-Marseille University as a Postdoctoral Researcher. His research interests include neuroimaging and machine learning.