SPANET: SPATIAL PYRAMID ATTENTION NETWORK FOR ENHANCED IMAGE RECOGNITION

Jingda Guo[1*], *Xu Ma*[1*], *Andrew Sansom*[1], *Mara McGuire*[2], *Andrew Kalaani*[3],
Qi Chen[1], Sihai Tang[1], Qing Yang[1], Song Fu[1]

[1]University of North Texas, Denton, TX 76203
[2]Texas A&M University - Corpus Christi, Corpus Christi, TX 78412
[3]Georgia Southern University, Statesboro, GA 30458

ABSTRACT

Attention mechanism has shown great success in computer vision. In this paper, we introduce Spatial Pyramid Attention Network (SPANet) to investigate the role of attention block for image recognition. Our SPANet is conceptually simple but practically powerful. It enhances the base network by adding Spatial Pyramid Attention (SPA) Blocks laterally. In contrast to other attention based networks that leverage global average pooling, our proposed SPANet considers both structural regularization and structural information. Furthermore, we investigate the topology structure of attention path connection and present three SPANet structures. SPA block is flexible to be deployed to various convolutional neural network (CNN) architectures. The experimental results show that our SPANet significantly improves the recognition accuracy without introducing much computation overhead compared with other CNN models. Codes are made publicly available [1].

1. INTRODUCTION

Convolutional neural networks have shown profound influence on a variety of visual processing applications. Hence, there are ever-increasing interests in CNN improvements. To enhance the performance of CNNs, recent works add more and more convolutional layers to the CNN architecture. For example, from 8-layer AlexNet [1] to 1000-layer ResNet [2,3], they aim to improve the accuracy of image recognition. Inevitably, more learnable layers introduce more parameters and prolong inference time.

In addition to making neural networks deeper, other efforts focus on investigating attention mechanisms [4] in CNNs. By informing a CNN network where to look and what to pay attention to, attention networks achieve a better performance with fewer layers. As an example, SENet [5] introduces Squeeze-and-Excitation (SE) blocks to study the channel dependencies in a CNN architecture. Although aforementioned CNN architectures achieve better performance for image recognition, the use of global average pooling (GAP) layers that aggregate a 3D feature map to a 1D attention map, would certainly cause loss of structural information in intermediate feature maps. To mitigate this problem, Convolutional block attention module (CBAM) [6] considers both channel-wise attention and spatial attention, which focuses on channel dependencies and structural information respectively.

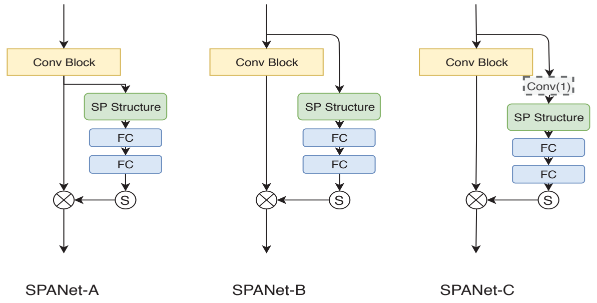


Fig. 1. Architecture of our SPANet. We design a Spatial Pyramid Structure to replace the traditional global average pooling. SPANet-A learns attention from current feature maps. SPANet-B learns from previous feature maps. SPANet-C adds an optional point-wise convolution to the attention path.

In this paper, we innovatively incorporate structural information to channel-wise attention blocks. We argue that the limitation originating from the global average pooling makes the shallow layers (which output big-size feature maps) unable to fully leverage the advantages of attention mechanism [7]. Following this argument, we present Spatial Pyramid Attention (SPA), which introduces a spatial pyramid structure to encode the intermediate features instead of using the simple global average pooling. Our proposed SPA is composed of two parts: one is a spatial pyramid structure which aggregates a 3D feature map into a 1D attention map, and the other is a combination of two fully-connected layers and a sigmoid-based activation layer, which sequentially encodes and decodes attention weights. These two parts are light-weight.

In terms of pooling schema, our spatial pyramid struc-

*Equal contribution

[1] https://github.com/13952522076/SPANet

978-1-7281-1331-9/20/$31.00 ©2020 IEEE

ture could be considered similar to SPPNet [8] and Region of Interesting Pooling [9]. In contrast, our spatial pyramid structure encodes a feature map with more structural information while SPPNet and Region of Interesting Pooling aim to obtain a fixed-length feature vector. In addition to being capable of retaining the spatial information in each channel, a major advantage of the proposed spatial pyramid structure is that it does not introduce any additional parameter. All layers in the spatial pyramid structure are not learnable, which is nearly cost-free. Compared to SENet [5], our structure only modifies the first fully-connected layer to tackle the large input size. The small computation overhead contributes to its enhanced performance.

Inspired by self-attention, we explore three topology structures of the Spatial Pyramid Attention module in our proposed SPANet, referred to as SPANet-A, SPANet-B and SPANet-C. SPANet-A learns attention from current feature maps, which follows a traditional self-attention path connection schema. SPANet-B learns from previous feature maps. SPANet-C adds an optional point-wise convolution to the attention path. Fig. 1 depicts the schemas of SPANet.

We comprehensively evaluate the performance of SPANet using CIFAR-100 and a down-sampled ImageNet dataset. Without bells and whistles, SPANet outperforms related state-of-art work [2,5,10,11]. Experimental results show that structural information in the attention mechanism, which we focus on, is a crucial factor for model performance. Compared to SENet that only considers the structural regularization in attention mechanism, our SPANet obtains 1.88% accuracy improvement on downsampled ImageNet [12].

2. RELATED WORK

Multi-Path Connection. Multi-path connection in deep learning was first used in Highway Networks [13, 14]. By allowing an unimpeded information flowed across several layers, a Highway Network is capable of reusing the information from previous layers, which facilities the training of deep networks. Moreover, gating units are employed to regulate the information flow. Subsequently, He *et al.* proposed Residual Networks (ResNet) [2, 3], which learn the residual functions by adding skip-connections. The ResNet shows that an identity mapping shortcut is crucial to ease the optimization [2,3]. Hence, ResNet discards the gating units used in Highway Networks and keeps the information passed though shortcuts. The better performance achieved by ResNet has made shortcut connections attractive. As a more dense reformulation, the work in [10] connects every convolutional layer in a deep convolutional network. Without introducing more parameters, it effectively alleviates the vanishing gradient problem and improves feature reuse.

In addition to shortcut connections, there are works studying the internal multi-path connections in convolutional blocks [15]. The InceptionV4 Network [15] is one of this kind. Besides a shortcut connection, each inception block in InceptionV4 contains 3-6 carefully designed paths. All these paths are integrated together using filter concatenation as input to the next block. More recently, attention based networks such as SENet [5] and CBAM [6] provide an independent attention path to learn the weight of each channel and achieve state-of-the-art performance.

Attention Mechanism. Attention Mechanism [4] has been prevailed in computer vision for years [16]. By adopting a gating function such as soft-max and Sigmoid, attention mechanism is able to selectively emphasize salient features as well as suppress insignificant features. Thus, visual features could be better captured and exploited. In [5], a Squeeze-and-Extraction block was proposed to learn the channel-wise attention for each convolutional layer, which provides an end-to-end training paradigm for attention learning. Inspired by SENet, Competitive-SENet [17] studies attention from both the residual path and the shortcut path. Although Competitive-SENet achieves promising performance, it is tailored particularly for Residual Networks [2], which limits its generalization to other models. Without being limited to channel-wise attention, Sanghyun Woo *et al.* [6] exploited the relation between channel-wise attention and spatial attention and proposed a Convolutional Block Attention Module (CBAM). CBAM is composed of two parts, i.e., a channel-wise attention part and a spatial attention part. The two attention parts in CBAM is able to tell what (channel) to look and where (spatial) to focus on. Unlike CBAM that learns channel-wise attention and spatial attention separately, our proposed SPANet learns channel-wise and spatial attention in an integrated fashion.

3. SPATIAL PYRAMID ATTENTION MODULE

Convolutional neural networks achieve great success in computer vision. Meanwhile, it ignores the weights of channels, which affects CNN's ability of discrimination. Attention mechanism, on the other hand, is capable of capturing channels' dependency, but ignores the structural information of channels. To enhance the representation power of CNNs, we introduce a spatial pyramid attention module. The proposed module considers the spatial pyramid structure which integrates average pooling of different sizes and explores the connection schema of attention paths.

3.1. Design Overview

Fig. 2 depicts the paradigm of our spatial pyramid attention module. The module learns a 1D attention map in an attention path, which is laterally connected to the original convolutional flow. The learned attention map is fed to each convolutional block in the original path. Such a design makes it possible to apply SPA module to various base models easily.

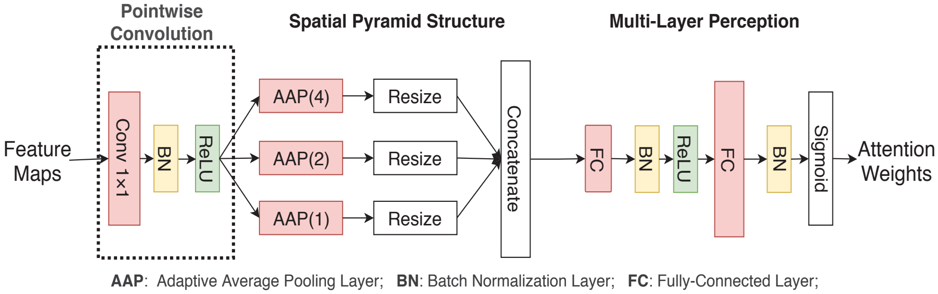


Fig. 2. Architecture of the spatial pyramid attention module. It is composed of three components, i.e., point-wise convolution, spatial pyramid structure, and multi-layer perception. Point-wise convolution is particularly designed for SPANet-C to match the channel number and integrate channel information. Spatial pyramid structure includes adaptive average pooling of three different sizes to integrate structural regularization and structural information in an attention path. Multi-layer perception learns an attention map from the output of the spatial pyramid structure. In SPANet-A, the input feature map is the current output of a block. In SPANet-B and SPANet-C, the input feature map is the previous output of a block with SPANet-C performing an optional point-wise convolution.

Suppose a CNN is composed of L layers, each of which outputs a feature map. We use x_l to denote the output of the l-th layer, where $l \in [1, L]$ is the index of a layer. We denoted adaptive average pooling and fully-connected layer as $P(\cdot, \cdot)$, $F_{fc}(\cdot)$ respectively. $C(\cdot)$ represents a concatenation operation, $\sigma(\cdot)$ is a Sigmoid activation function, and $R(\cdot)$ is referred to as re-sizing a tensor to a vector.

Given an intermediate feature map $x_l \in \mathbb{R}^{C \times W \times H}$, an attention mechanism based CNN model learns attention weights from the input x_l and multiplies each channel in x_l by learnable weights to produce an output. The output of Spatial Pyramid Structure $S(x_l)$ can be presented as:

$$S(x_l) = C\left(R\left(P\left(x_l, 4\right)\right),\ R\left(P\left(x_l, 2\right)\right),\ R\left(P\left(x_l, 1\right)\right)\right). \quad (1)$$

Omitting the batch normalization and activation layer for clarity, SPA module performs transformation $\mathfrak{F}$ as

$$\mathfrak{F}(x_l) = \sigma\left(F_{fc}\left(F_{fc}\left(S(x_l)\right)\right)\right). \quad (2)$$

Equation (2) presents the essence of transformation by the SPA module. The batch normalization layer, activation layers and point-wise convolution omitted in Equation (2) are included in the implementation and performance evaluation of the SPA module (Section 4). Following [8], we propose 3-level pyramid average pooling: 4×4, 2×2 and 1×1.

3.2. Attention Path Connection

Most of the existing self-attention based networks follow a path design pattern: they learn an attention map from a feature map and then apply the learned attention map to the original feature map [5, 18] . However, being confined to aforementioned schema compromises the exploration of attention path connections. For SPANet, we study the topology of attention path connections and explore three variations: SPANet-A, SPANet-B, and SPANet-C, as shown in Fig. 1.

SPANet-A feeds the current feature map x_l to the attention path to generate a 1D attention map. Accordingly, the output of a block in SPANet-A can be expressed as

$$x_l = \mathfrak{F}(x_l) \otimes x_l, \quad (3)$$

where $\otimes$ denotes element-wise multiplication. SPANet-A uses a similar schema as traditional self-attention path connections.

SPANet-B learns an attention map directly from x_{l-1} (where $x_{l-1} \in \mathbb{R}^{C' \times W' \times H'}$) instead of the processed x_l. The output of a block in SPANet-B is

$$x = \mathfrak{F}(x_{l-1}) \otimes x_l. \quad (4)$$

This design in SPANet-B is to assure that the attention path is independent of the original convolutional block path, enabling the attention path to learn more generalized weights. Note that although the two paths are independent of each other, they are not completely irrelevant because the attention path and the convolutional block path are trained jointly.

SPANet-C. Considering the channel number in x_{l-1} may not be equal to the channel number in x_l, the attention path might not produce the most accurate weights for x_l. Thus, we modify SPANet-B by adding a point-wise convolutional layer [19] at the beginning of the attention path if $C' \neq C$. We compute the output of SPANet-C as follows.

$$x = \mathfrak{F}(\mathfrak{C}(x_{l-1})) \otimes x_l, \quad (5)$$

where $\mathfrak{C}(\cdot)$ denotes the point-wise convolution operation. The point-wise convolution, which consists of a convolutional layer with a 1×1 filter and a batch normalization layer, aims to integrate channel information and match channel numbers of output feature maps. It makes the attention path further independent of x.

We focus on the topology structure of attention path connections in the preceding discussion. The implementation details of the attention mechanism are provided in Section 4. All of the three SPANets can be integrated with other CNN architectures. In the following discussion, SPANet refers to SPANet-C unless otherwise specified.

3.3. Spatial Pyramid Attention

Many existing attention based networks [5, 6, 6, 17] aggregate input feature maps into a 1D vector using global average pooling. They achieve structural regularization [20], but miss the structural information. In contrast, the spatial pyramid structure in our proposed attention module utilizes average pooling of three different sizes to both achieve structural regularization and explore structural information (as shown in Fig. 2).

3.3.1. *Spatial Pyramid Structure*

Global average pooling (GAP), which aggregates the global information in each channel, was introduced in [21] to replace the conventional fully-connected layers in CNNs. Since then, it has prevailed in computer vision for recognition [2], detection [22], segmentation [23], and more.

We note that existing work on global average pooling used the last feature map which is small in size (7 × 7 for example). However, attention based CNNs (e.g., [5], [6], [7], etc.) apply global average pooling on each feature map. As presented in [20], GAP behaves similarly to a structural regularizer and is capable of preventing over-fitting. However, applying GAP to every feature map overemphasizes the effect of regularization and misses the original feature representation and structural information, especially when a feature map is large. For example, aggregating a 112 × 112 feature map to a mean value causes significant loss of a features' representation capability, which affects feature learning.

To address this problem, we propose a *spatial pyramid structure* used in attention blocks. The spatial pyramid structure adaptively and averagely pools an input feature map to three scales: 4×4, 2×2, and 1×1. The spatial pyramid structure provides a combination of three regularization terms, i.e., the 4×4 average pooling captures more feature representation and structural information, the 1×1 average pooling is the traditional GAP with a strong structural regularization, and the 2×2 average pooling aims at a trade-off between structural information and structural regularization. Then we re-size the three outputs to three 1D vectors and combine all together to generate a 1D attention map. Our spatial pyramid structure is capable of both preserving the feature representation and inheriting the advantages of the global average pooling.

3.3.2. *Fully-Connected Layers*

The 1D attention map v extracted from the spatial pyramid structure is a concatenation of the outputs from three pooling layers. However, it cannot be used to learn channel dependency and its non-linear expression affects the effectiveness of the attention mechanism. To address this problem, we leverage the excitation block [5] to encode v and generate a 1D attention map $\tilde{v}$. The excitation block employs two fully-connected layers. Then a sigmoid layer is employed to normalize the output to a range of $(0, 1)$.

We use W_1 and W_2 to denote the first and second fully-connected layers respectively, where we set the reduction rate to r. Thus, the generated attention map is

$$\tilde{v} = sig\left(W_2 \rho\left(W_1 v\right)\right), \tag{6}$$

where ρ is a rectified linear unit (ReLU) function and sig denotes the sigmoid function. Like in SENet [5], we set r to 16.

3.3.3. *Point-wise Convolution*

The attention block in our proposed spatial pyramid attention module produces attention maps to analyze channel dependencies. In SPANet-B, the attention path learns a vector converted from a feature map with C' channels to multiply a feature map with C channels. However, the dis-match of channels may cause discrepancy in attention learning and decrease the performance of SPANet.

SPANet-C addresses this issue by adding a point-wise convolutional layer when $C' \neq C$. Specifically, the point-wise convolution is a convolutional layer with a filter in size of 1×1. By setting the input channel as C' and the output channel as C in the point-wise convolutional layer, we are able to match the number of channels and integrate channel information.

	Base	SENet	SPA-A	SPA-B	SPA-C
MobileNetV2	75.18	75.75	**75.81**	75.44	75.75
DenseNet	74.51	74.77	75.01	**75.22**	75.13
ResNeXt	77.93	**78.96**	78.76	78.56	78.63
VGG16	72.92	**73.0**	72.68	-	-

Table 1. Performance on CIFAR-100. SPANet and SENet achieve the best accuracy on four backbone CNN models. SPANet outperforms three backbone models.

4. PERFORMANCE EVALUATION

We comprehensively evaluate our SPANet on CIFAR-100 [24] and ImageNet [12]. Due to a lack of sufficient computing resources, we experiment on a downsampled ImageNet with 32 × 32 images. We compare ResNet + SPANet with SENet and ResNet. We also apply SPANet and SENet to several other base CNN architectures, including VGG [25], MobileNetV2 [11], DenseNet [10], and ResNext [26], to study the generalizability of SPANet.

4.1. Experiment Settings and Datasets

We implement SPANet using Pytorch. We train all models using a stochastic gradient descent method, with a 0.9 Nesterov momentum and a $5e^{-4}$ weight decay. The batch size is 512 and the learning rate is initialized as 0.1. We experiment on two common datasets: CIFAR-100 [24] and Downsampled ImageNet [12] (a downsampled version of the original ImageNet dataset). For training, we adopt a data augmentation scheme used in [2, 3]. We pad an original image by 4 pixels with value zero on each side and then randomly crop the padded image back to a size of 32 × 32 pixels. In addition, we horizontally flip 50% of images in random. To facilitate model training, we normalize the image data by using channels' means and standard deviations. On CIFAR-100, the epoch size is set to 300 and the learning rate is decreased by a factor of 10 every 70 epochs. On ImageNet, we set the epoch size to 100 and divide it by 10 at the 30th, 60th, and 90th epochs. All experiments are conducted on a server with 4 TESLA K80 GPUs.

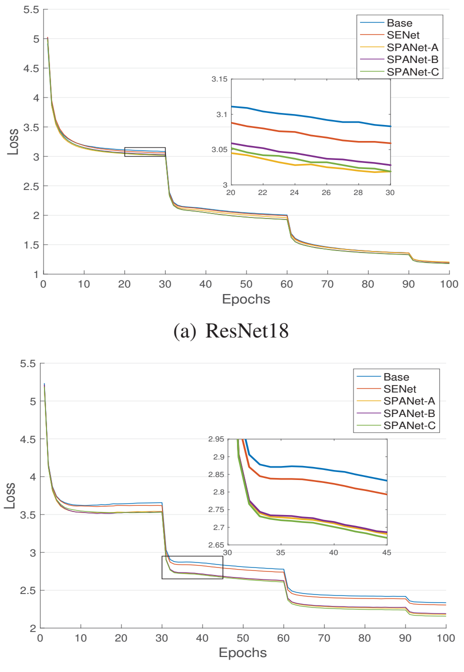
4.2. Experimental Results

We compare the performance of our SPANet with SENet and the base networks. We employ four base networks, i.e., light-weight model MobileNetV2 [11], heavy-weight model DenseNet [10], ResNeXt [26], and VGG16 [25].

Recognition accuracy. Table 1 shows the results on CIFAR-100. From the table, we can see that SPANet achieves the best performance in several scenarios but not all, while SENet outperforms

	Base		SENet		SPANet-A		SPANet-B		SPANet-C	
	Top1 acc.	Top5 acc.	Top1 acc.	Top5 acc.	Top1 acc.	Top5 acc.	Top1 acc.	Top5 acc.	Top1 acc.	Top5 acc.
MobileNetV2 [11]	44.306	69.496	44.556	70.264	46.370	71.678	46.242	71.298	**46.596**	71.812
DenseNet [10]	49.190	74.454	50.198	75.078	50.692	75.476	**51.886**	75.714	50.856	75.872
ResNeXt [26]	59.346	81.984	59.784	82.534	**61.210**	83.550	60.130	82.842	60.106	82.914
VGG16 [25]	49.214	73.698	49.612	73.512	**50.094**	74.144	-	-	-	-

Table 2. Performance on downsampled ImageNet. SPANet outperforms all four backbone CNN models and SENet.



(a) ResNet18

(b) MobileNetV2

Fig. 3. Training loss on the downsampled ImageNet (Best viewed in color). The three SPANets consistently produce less loss than SENet and the base networks. Similar results are also found on CIFAR-100.

the base networks in all cases. This is different from our intuition. As aforementioned, the global average pooling is for structural regularization, which mitigates over-fitting. Our spatial pyramid structure, on the other hand, uses both structural regularization and structural information to achieve a better learning capability. However, it may cause over-fitting on small datasets. The performance of SPANet on CIFAR-100 becomes stable, i.e., the training loss approaches zero. This indicates a larger training dataset can contribute to a better performance.

We further test SPANet on the downsampled ImageNet dataset. The results are presented in Table 2. Our major findings are 1) SPANet achieves the best performance over the base models. SPANet surpasses the base models, i.e., MobileNetV2, DenseNet, ResNeXt, and VGG16, by 2.290%, 2.696%, 1.864%, and 0.88% respectively in the Top-1 accuracy. Moreover, all three types of SPANet outperform the base models and SENet. These results show the effectiveness of SPANet and prove that using both structural regularization and structural information is imperative for the attention mechanism. 2) The best performance is not achieved by one particular type of SPANet. For example, SPANet-A has two greatest accuracy improvements, and SPANet-B and SPANet-C each has one. They deliver the best performance on different backbones. This result verifies it is necessary to investigate the topology structure of attention path connections. 3) The performance enhancement varies among the backbone models. SPANet achieves an improvement of 2.696% (2.040% over SENet) on DenseNet, but a small improvement of 0.88% (0.482% over SENet) on VGG16. This indicates the architecture of a base network may affect the effectiveness of SPANet. Note that the four networks represent different network architectures. MobileNetV2 is typically designed for lightweight models like [19, 27, 28]. DenseNet includes shortcut connections. ResNeXt is the first one that exposes ”cardinality” dimension. VGG is a popular plane-structure network. Our experimental results demonstrate that SPANet performs well on different types of CNN architectures.

Depth	SE	SE+	SE++	SPA-A	SPA-B	SPA-C
18	75.19	74.97	75.25	75.41	75.01	**75.56**
50	77.91	77.45	77.43	**78.21**	78.11	77.95
101	78.03	77.88	77.61	78.11	78.35	**79.17**

Table 3. Ablation results on CIFAR-100 based on ResNet. ’+’ means an SE block is connected to a previous feature map. ’++’ means an additional point-wise convolution is added to SENet+.

Training Loss. Next, we plot the training loss as shown in Fig. 3. Due to space limitations, we present the results on ResNet18 and MobileNetV2. In the figure, we can see SPANet achieves the least loss compared with other models. Among the three types of SPANet, SPANet-A, SPANet-B, and SPANet-C perform the best on ResNet18, DenseNet, and MobileNetV2 respectively. They employ different connection schemas, indicating that nuances in the different topology structures of attention path connections influence the performance of SPANet.

4.3. Ablation Analysis

In this set of experiments, we run a number of ablations to analyze SPANet. Tables 3 and 4 present the results.

Attention Connection. Unlike SPANet-A, SPANet-B uses an attention connection schema that learns attention from a previous feature map. We compare SENet with SENet+ and SPANet-A with SPANet-B because each pair only differs in the topology structure of their attention path connections. In the tables, we can see that SENet always performs better than SENet+ while the results of SPANet-A are mixed compared with SPANet-B. We observe similar results in Tables 1 and 2 where more backbone models are tested. Results on the downsampled ImageNet (shown in Table 4) indicate that the two attention path connection schemas on SENet and SPANet achieve comparable performance. Thus, we can conclude that the topology structure of an attention path connection should not be fixed to a

	Base		SENet		SENet+		SENet++		SPANet-A		SPANet-B		SPANet-C	
	Top1	Top5	Top1	Top5 .	Top1	Top5 .	Top1	Top5	Top1	Top5	Top1	Top5	Top1	Top5
ResNet18	53.632	77.200	53.526	77.424	53.754	77.412	53.668	77.610	54.502	78.184	54.236	78.026	**54.644**	78.430
ResNet50	60.434	82.476	59.414	81.716	59.316	81.652	59.132	81.648	61.304	83.396	61.364	83.514	**61.476**	83.502
ResNet101	61.860	83.522	60.928	82.862	60.922	82.774	60.826	82.628	62.672	84.122	**62.808**	84.382	61.020	83.184

Table 4. Ablation results on downsampled ImageNet. '+' and '++' have the same meanings as in the preceding table.

certain schema and further exploration is needed.

Point-wise Convolution. We also evaluate the impact of point-wise convolution on recognition accuracy. We compare SENet+ with SENet++ and SPANet-B with SPANet-C on both CIFAR-100 and downsampled ImageNet datasets. Experimental results show SENet+ consistently outperforms SENet++, while SPANet-C achieves a better performance than SPANet-B in four of six cases. This indicates that point-wise convolution improves the performance of SPANet-B, but not always among the three types of SPANet.

Spatial Pyramid Structure. We evaluate the spatial pyramid structure which is a substitute for global average pooling. We compare the performance of the base networks with that of SPANet-A based models. Tables 1, 2, 3, and 4 present the results. From the tables, we can see SPANet consistently outperforms the base and SENet based networks. Specifically, SPANet-MobileNetV2 surpasses MobileNetV2 by 2.064% and SENet based network by 1.814% on the ImageNet dataset. Compared to a 0.25% improvement made by SENet, SPANet-A significantly enhances the accuracy. These results show the importance of combining structural information and structural regularization in attention paths as discussed in Section 3.3.

5. CONCLUSIONS

We present the Spatial Pyramid Attention Network (SPANet), a new design to enhance the performance of CNN. SPANet introduces the spatial pyramid structure to the attention path, which integrates the structural information and structural regularization. We explore the topology structure of attention path connections and develop three types of SPANet using different connection schemes. Experimental results on two datasets demonstrate both the efficiency and effectiveness of SPANet.

ACKNOWLEDGMENT

The work is supported by National Science Foundation (NSF) grants NSF CNS-1761641 and CNS-1852134.

6. REFERENCES

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *ECCV*. Springer, 2016, pp. 630–645.

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *NeurIPS*, 2017, pp. 5998–6008.

[5] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[6] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[7] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang, "Residual attention network for image classification," in *CVPR*, 2017, pp. 3156–3164.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[9] Ross Girshick, "Fast r-cnn," in *ICCV*, 2015, pp. 1440–1448.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017, pp. 4700–4708.

[11] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *CVPR*, 2018, pp. 4510–4520.

[12] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter, "A downsampled variant of imagenet as an alternative to the cifar datasets," *arXiv preprint arXiv:1707.08819*, 2017.

[13] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Highway networks," *arXiv preprint arXiv:1505.00387*, 2015.

[14] Rupesh K Srivastava, Klaus Greff, and Jürgen Schmidhuber, "Training very deep networks," in *NeurIPS*, 2015, pp. 2377–2385.

[15] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.

[16] Tam V Nguyen, Qi Zhao, and Shuicheng Yan, "Attentive systems: A survey," *IJCV*, vol. 126, no. 1, pp. 86–110, 2018.

[17] Yang Hu, Guihua Wen, Mingnan Luo, and Dan Dai, "Competitive inner-imaging squeeze and excitation for residual network," *arXiv preprint arXiv:1807.08920*, 2018.

[18] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le, "Attention augmented convolutional networks," *arXiv preprint arXiv:1904.09925*, 2019.

[19] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[20] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.

[21] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.

[22] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *CVPR*, 2017, pp. 7263–7271.

[23] Wei Liu, Andrew Rabinovich, and Alexander C Berg, "Parsenet: Looking wider to see better," *arXiv preprint arXiv:1506.04579*, 2015.

[24] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," Tech. Rep., Citeseer, 2009.

[25] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[27] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *CVPR*, 2018, pp. 6848–6856.

[28] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *ECCV*, 2018.