SUMMARY

No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects

Many state-of-the-art computer vision models nowadays are based on Convolutional Neural Networks (CNNs). They have astounding performances in many computer vision tasks such as image classification or object detection. However, their performance drops when dealing with low resolution images or small object detection. Therefore, the paper suggests a defect in the design of these CNN architectures which causes the drop in the model's performance and proposes a fix for it.

According to the paper, popular CNN models such as: AlexNet, VGGNet, Resnet, R-CNN series, YOLO series, SSD, EfficientDet… excel in image classification and object detection tasks, as long as they have fine images, medium to large objects as inputs in training and interference. However, all of their performance degrades when the images' resolution is reduced. This is caused by the use of strided convolution and/or pooling layers in CNNs architecture. Usually, with high quality inputs there is plenty of redundant pixel information that strided convolution and pooling can conveniently skip and the model can still learn features quite well. However, in tougher tasks when images are blurry or objects are small, the assumption of redundant information no longer applies and the model starts to suffer from loss of fine-grained information and poorly learned features.

In order to fix this problem, the article proposes a SPD-Conv layer. This layer consists of a Space-to-depth (SPD) layer followed by a non-strided convolution layer. The SPD layer will slice out the original feature map (X) of size $S \times S \times C_1$ into a sequence of sub feature maps, each sub feature map downsample X by a factor of $scale$, and then concatenate these sub feature maps along the channel dimension to create a new feature map (X'), which has a reduced spatial dimension by a factor of scale and an increased channel dimension by a factor of $scale^2$. After that, the non-strided convolution layer with C_2 filters, where $C_2 < scale^2 C_1$, and further transforms feature map X'($\frac{S}{scale}$,$\frac{S}{scale}$,$scale^2 C_1$) to X”($\frac{S}{scale}$,$\frac{S}{scale}$,C_1). This non-strided convolution layer is used to retain the discriminative feature information as much as possible.

SPD-Conv for Low-Resolution Images and Small Objects 5

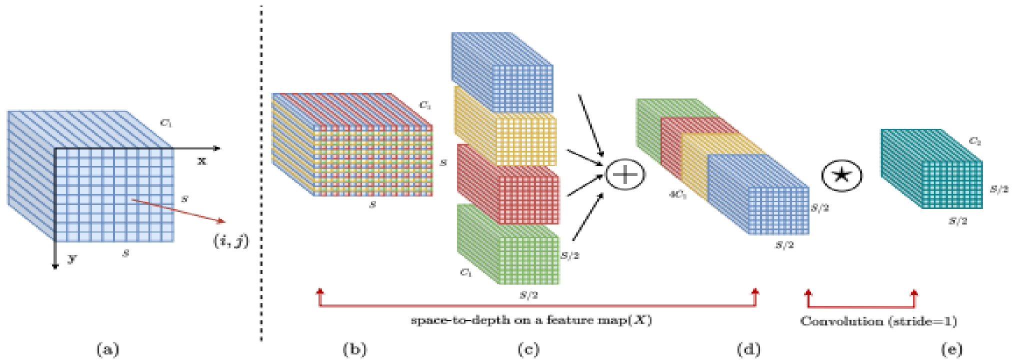
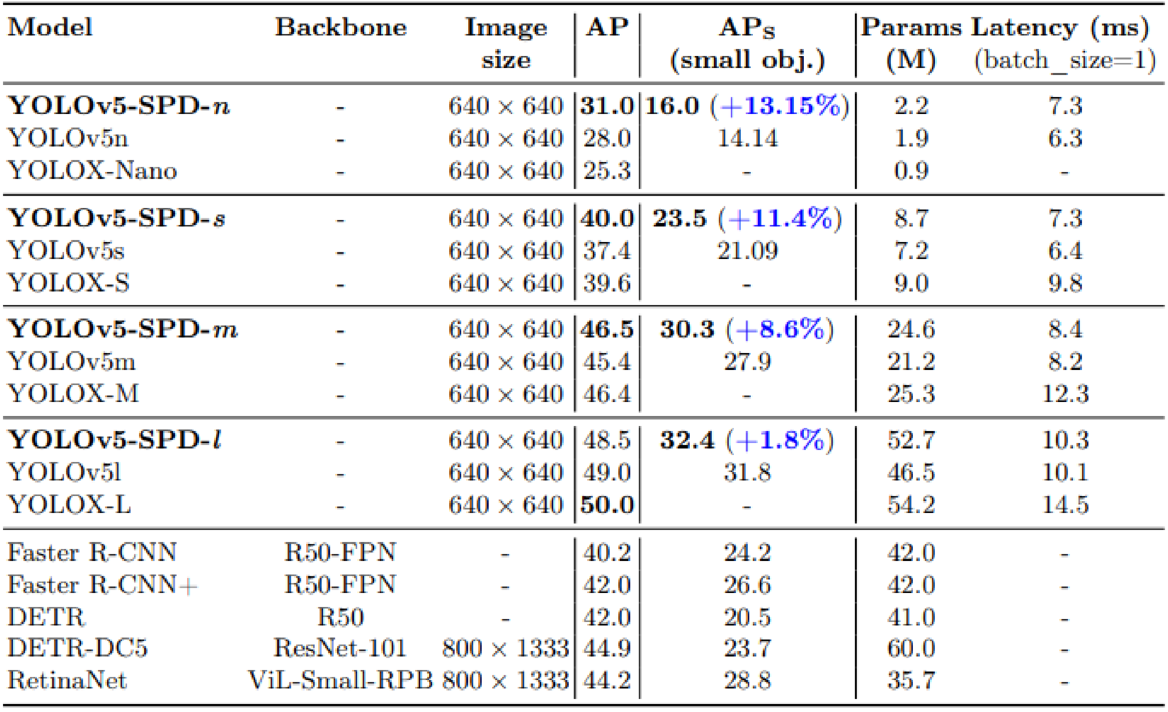


Fig. 3: Illustration of SPD-Conv when $scale = 2$ (see text for details).

After replacing stride-2 convolutions layers with the SPD-Conv layers in the YOLOv5 model, here is the result after training the new YOLOv5 model from scratch with the COCO-2017 dataset for object detection task:

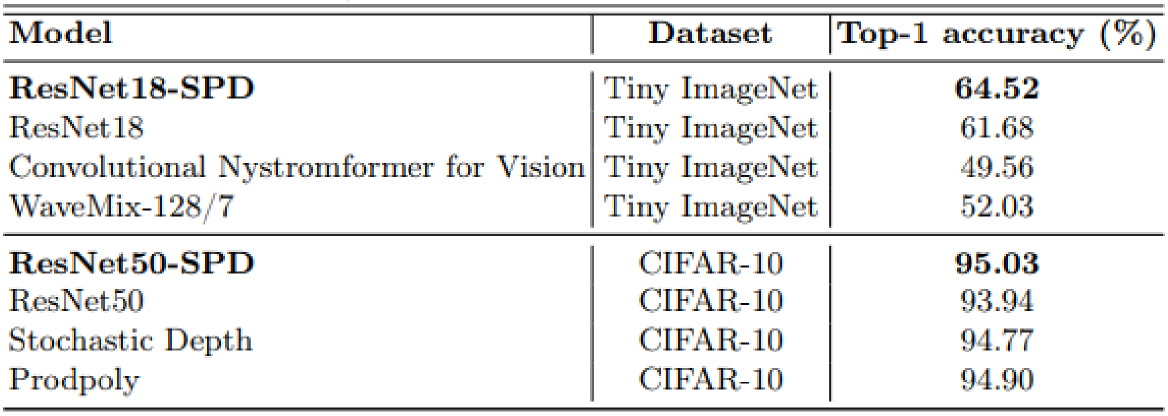
Table 4: Comparison on MS-COCO validation dataset (val2017).



Model	Backbone	Image size	AP	AP_S (small obj.)	Params (M)	Latency (ms) (batch_size=1)
YOLOv5-SPD-*n*	-	640 × 640	**31.0**	**16.0** (+**13.15%**)	2.2	7.3
YOLOv5n	-	640 × 640	28.0	14.14	1.9	6.3
YOLOX-Nano	-	640 × 640	25.3	-	0.9	-
YOLOv5-SPD-*s*	-	640 × 640	**40.0**	**23.5** (+**11.4%**)	8.7	7.3
YOLOv5s	-	640 × 640	37.4	21.09	7.2	6.4
YOLOX-S	-	640 × 640	39.6	-	9.0	9.8
YOLOv5-SPD-*m*	-	640 × 640	**46.5**	**30.3** (+**8.6%**)	24.6	8.4
YOLOv5m	-	640 × 640	45.4	27.9	21.2	8.2
YOLOX-M	-	640 × 640	46.4	-	25.3	12.3
YOLOv5-SPD-*l*	-	640 × 640	48.5	**32.4** (+**1.8%**)	52.7	10.3
YOLOv5l	-	640 × 640	49.0	31.8	46.5	10.1
YOLOX-L	-	640 × 640	**50.0**	-	54.2	14.5
Faster R-CNN	R50-FPN	-	40.2	24.2	42.0	-
Faster R-CNN+	R50-FPN	-	42.0	26.6	42.0	-
DETR	R50	-	42.0	20.5	41.0	-
DETR-DC5	ResNet-101	800 × 1333	44.9	23.7	60.0	-
RetinaNet	ViL-Small-RPB	800 × 1333	44.2	28.8	35.7	-

The same method is applied to the ResNet18 and ResNet50 models. After replacing the strided convolution layers and the max pooling layers with SPD-Conv layers, then run new model trainings with Tiny ImageNet and CIFAR-10 datasets for image classification task, the models achieved these results:

Table 6: Image classification performance comparison.



Model	Dataset	Top-1 accuracy (%)
ResNet18-SPD	Tiny ImageNet	**64.52**
ResNet18	Tiny ImageNet	61.68
Convolutional Nystromformer for Vision	Tiny ImageNet	49.56
WaveMix-128/7	Tiny ImageNet	52.03
ResNet50-SPD	CIFAR-10	**95.03**
ResNet50	CIFAR-10	93.94
Stochastic Depth	CIFAR-10	94.77
Prodpoly	CIFAR-10	94.90

As we can see from the result, by simply replacing the strided convolution and pooling layers with the SPD-Conv layer, the models have significantly improved its accuracy, while maintaining the same level of parameter size. This new design has proved to be effective in downsampling feature maps while retaining the discriminative feature information. Moreover, this new design can easily be applied to any CNN architecture.

Comment

I stumbled upon this paper, and even though it's been around since August 7, 2022, I only got around to reading it recently. It's exactly the kind of paper that grabs my attention when I look into publications. I'm really into staying updated on all the progress happening in deep learning and machine learning fields, especially when they're talking about finding defects and making improvements on existing architectures. Many companies nowadays are using these types of models for their machine learning solutions, therefore publications like this can be a great way for them to further improve their models performance.