DATA-ORIENTED SCENE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Most deep learning backbones are evaluated on ImageNet. Using scenery images as an example, we conducted extensive experiments to demonstrate the widely accepted principles in network design may result in dramatic performance differences when the data is altered. Exploratory experiments are engaged to explain the underlining cause of the differences. Based on our observation, this paper presents a novel network design methodology: data-oriented network design. In other words, instead of designing universal backbones, the scheming of the networks should treat the characteristics of data as a crucial component. We further proposed a Deep-Narrow Network and Lossless Pooling module, which improved the scene recognition performance using less than half of the computational resources compared to the benchmark network architecture ResNets.

1 INTRODUCTION

Since the development of AlexNet (Krizhevsky et al., 2012), a number of variations of deep Convolutional Neural Networks (CNNs) emerged. Motivated by the success of studies (Simonyan & Zisserman, 2015; Szegedy et al., 2016; He et al., 2016a), the networks become deeper by adding more convolutional layers to improve performance for the targeted problems. On the other hand, arguments on the benefits of increasing network width gained support from researchers. Zagoruyko et al. (Zagoruyko & Komodakis, 2016) presented wider deep residual networks that significantly improve the performance over ResNet (He et al., 2016a). The wider deep network achieved state-of-the-art performance on ImageNet (Deng et al., 2009) and CIFAR (Krizhevsky et al., 2009), showing consistently better accuracy than its ResNet counterparts. Xie et al. (Xie et al., 2017) proposed a multi-branch architecture called ResNeXt by widening the residual blocks and cooperating group convolution. In contrast to ResNet, ResNeXt (Xie et al., 2017) demonstrated a performance boost on a larger ImageNet-5K set and the COCO object detection dataset (Lin et al., 2014). ResNeSt (Zhang et al., 2020) preserved the wider network layout and multi-branch strategy and introduced a modulated architecture to improve the feature learning process. The proposed ResNeSt networks further improved the performance on the ImageNet dataset.

Despite the benefits brought by increased depth and width (i.e., number of channels), side effects such as vanishing gradient and requirement of large number of training examples make it difficult to employ a deep network for learning from complex scenery images. Is a deeper network more suitable for extracting features from scenery images for better understanding complex views? Or shall we employ more channels to achieve improved performance? To answer these questions, we need to have a better understanding of the functions of network layers and channels. A number of studies have been conducted in the past years. Lu et al. (Lu et al., 2017) argued that an integration of both depth and width provides a better understanding of the expressive power of neural networks. Tan and Le (Tan & Le, 2019) showed that it is critical to balance the network depth and width by maintaining a constant depth/width ratio and demonstrated the effectiveness of their approach on ResNet and MobileNet. Besides manually designed networks, deep Neural Architecture Search were proposed to optimize the network depth and width (Zoph & Le, 2017; Guo et al., 2020). However, most methods, if not all, were developed and evaluated using ImageNet (Deng et al., 2009) and CIFAR (Krizhevsky et al., 2009). These data sets commonly depict an object near the center of the image, and the label tells what the object is. That is, such an image is mostly dominated by one object and, hence, is referred to as "object-centric". On the other hand, a scene image presents a complex view consists of multiple objects and background clutters. This inadvertent data bias could potentially lead to the ignorance of the characteristics of different data. The features that are crucial for object recognition dominantly affect the design of the CNNs.

Deep networks with more layers use a variety of receptive fields to extract distinctive scale features whereas networks with more channels capture fine-grained patterns (Tan & Le, 2019). However, in many applications, both types of information that are crucial for accurately recognizing an image, but the prominence difference between them is seldom explored.

This paper attempts to bring out a new perspective on neural network design: the data itself has the preference. Learning the overall spatial layout is crucial to recognize the entire scene, thus scene recognition favors the networks that can better learn the spatial information. For the object-centric images, typical examples only consist of one single object, the spatial layout does not contribute much to the semantic meaning of the image. As the differences between certain object categories are subtle, the detailed patterns and textures of objects are likely to be more representative. The networks that emphasize learning various features can better fit the requirement of object recognition tasks. By considering the distinct characteristics of scene images, our hypothesis is that for scene recognition, learning spatial-wise information improves the performance of CNNs in a more effective manner compared to learning channel-wise information. To evaluate this hypothesis, we conducted comprehensive experiments and our results show scene images gain clear benefits from deepening the network, and the performance change caused by altering the width is marginal. We further proposed a Deep-Narrow Network, which increases the depth of the network as well as decreased the width of the network. We design a Lossless Pooling component and use it in our Deep-Narrow Network to extract spatial features.

2 Related Work

Deep Networks Network depth has played an integral role in the success of CNNs. With the increase of depth, the network can better approximate the target function with richer feature hierarchies, which enables the boost of performance. The success of VGGNet (Simonyan & Zisserman, 2015) and Inception (Szegedy et al., 2015; 2016; 2017) on ILSVRC competition further reinforced the significance of the depth. ResNet (He et al., 2016b), which is a continuance work of deeper networks, revolutionized the possible depth of deep networks by introducing the concept of residual learning and identity mapping into CNNs. ResNet's effective methodology enables the network to be extremely deep and demonstrated improved performance in image recognition tasks.

Wide Networks Network width has also been suggested as an essential parameter in the design of deep networks. Wide ResNet (Zagoruyko & Komodakis, 2016) introduced an additional factor to control the width of the ResNet. The experimental results showed that the widening of the network might provide a more effective way to improve performance compared to making ResNet deeper. Xception (Chollet, 2017) can be considered as an extreme Inception architecture, which exploits the idea of depth-wise separable convolution. Xception modified the original inception block by making it wider. This wider structure has also demonstrated improved performance. ResNeXt (Xie et al., 2017) introduced a new term: cardinality to increase the width of ResNet and won the 2016 ILSVRC classification task. With the success of ResNeXt, it is widely accepted that widen the deep network is an effective way to boost model performance. A most recent wider network: ResNeSt (Zhang et al., 2020) preserved the wide architecture of ResNeXt and achieved superior performance on image and object recognition tasks.

Effects of Depth and Width Although depth and width are proven to be essential parameters in network architecture design, the effect of depth and width, i.e., what do deep and wide networks learn remains seldom explored. Most of the existing literature focus on the effect of width and depth separately or the trade-off between depth and width in the network design (Lu et al., 2017). Tan and Le (Tan & Le, 2019) claimed that deep networks can make use of a larger receptive field while wide networks can better capture fine-grained features. Nguyen et al. (Nguyen et al., 2021) explored the effects of width and depth and found a characteristic structure named block structure. They demonstrated that for different models, the block structure is unique, but the representations outside the block structure trends to be similar despite the setting of depth and width. In our paper, we analyze the effect of depth and width in CNNs from the perspective of image characteristics.

3 EXPERIMENTAL STUDY

3.1 DATA SETS AND EXPERIMENTAL SETTINGS

To understand the impact of network structure to the data set and ultimately the applications, we use ImageNet 2012 (Deng et al., 2009) and Places Standard dataset (Zhou et al., 2017) as our evaluation data sets. ImageNet 2012 is the benchmark object recognition data set that consists of 1,000 classes and 1.28 million training images. An image in ImageNet 2012 usually contains a single object that is highly distinctive from the background. Place365 Standard dataset is designed for scene recognition and contains 1.8 million training images of 365 classes. The images in Places365 Standard datasets present more complex scenery images.

We train deep network models and compute the single-crop (224×224 pixels) top-1 and top-5 accuracy based on the application of the models to the validation set. We train each model for 100 epochs on eight Tesla V100 GPUs with 32 images per GPU (the batch size is 256). All models are trained using synchronous SGD (Stochastic Gradient Descent) with a Nesterov momentum of 0.9 and a weight decay of 0.0001. The learning rate is set to 0.1 and is reduced by a factor of 10 in every 30 epochs. In the training of ResNet and its variants, we follow the settings in (He et al., 2016b).

3.2 COMPARISON ON DIFFERENT DATA SETS

We conducted our comparison study using ResNet and its variants on Places365 and ImageNet data sets. The results are reported in Tables 1 (varying network depth) and 2 (varying network width). By increasing the network depth from 50 to 101, i.e., ResNet-50 and ResNet-101, we obtained a performance improvement of 1.40% and 2.32% on Place365 and ImageNet data sets in terms of Top-1 accuracy. Theoretically, if the widening of the network is more effective to improve the performance as stated in the previous literature, we should expect more accuracy increase when we double the width of the networks. However, doubling the width leads to a top-1 accuracy increase of 3.28% on ImageNet, but only 0.94% on Place365. More surprisingly, for ResNeXt, which also doubled the network width, the relative performance increase on ImageNet is 2.34% in terms of top-1 accuracy, but the number is only 0.14% on Place365.

Table 1: Top-1 and top-5 accuracy (%) and complexity comparison by changing the network depth. The number of parameters is in million. Note that for the cases using Places365-Standard dataset we calculate the number of parameters base on 365-class models (Place365).

Data	Model	Top-1	Top-5	GFLOPs	# Parameters
Places365	ResNet-18	54.22	84.63	1.82	11.36
	ResNet-50	55.69	85.80	4.12	24.26
	ResNet-101	56.47	86.25	7.84	43.25
ImagineNet	ResNet-18	70.52	89.56	1.82	11.69
	ResNet-50	76.02	92.80	4.12	25.56
	ResNet-101	77.78	93.72	7.84	44.55

The trend that the model performance on ImageNet is more sensitive to width change compared with Place365 is also true when we decrease network depth or narrow down the width. When the network depth decreased from 50 to 18, ImageNet suffered a 7.23% relative top-1 accuracy decrease, and for Place365 it is 2.71%., that is, the top-1 performance drop on ImageNet is around 2.7 times of the top-1 performance drop on Places. But when we decreased the width of ResNet-50 to half of the original size, the number changed to 4.7. The statement that widening of the network might provide a more effective way to improve performance is biased towards ImageNet (object-centric data) and ignores the characteristics of scenery images.

4 ANALYSIS OF THE IMPACT OF DEPTH AND WIDTH OF DEEP NETWORKS

4.1 COMPLEXITY OF IMAGES

Our first hypothesis is the performance difference is caused by the complexity of scenery images. This hypothesis is aroused by the distinct complexity difference between scene images and object-

Table 2: Top-1 and top-5 accuracy (%) and complexity comparison by changing the network width. The number of parameters is in million. Note that for the cases using Places365-Standard dataset we calculate the number of parameters base on 365-class models (Place365). The numbers within parenthesis are the width scaling factors.

Data	Model	Top-1	Top-5	GFLOPs	# Parameters
Places365-Standard	ResNet-50 ($\times$ 1)	55.69	85.80	4.12	24.26
	ResNet-50 ($\times$ 2)	56.21	86.11	11.43	67.58
	ResNet-50 ($\times$ .5)	55.07	85.12	1.07	6.27
	ResNet-50 ($\times$ .25)	52.16	82.85	0.29	1.67
	ResNeXt-50	55.77	85.99	4.27	23.73
ImageNet	ResNet-50 ($\times$ 1)	76.02	92.80	4.12	25.56
	ResNet-50 ($\times$ 2)	78.51	94.09	11.43	68.88
	ResNet-50 ($\times$ .5)	72.08	90.78	1.07	6.92
	ResNet-50 ($\times$ .25)	64.04	85.76	0.29	1.99
	ResNeXt-50	77.80	94.30	4.27	25.03

centric images: object-centric images always only contain one major object which occupies a large portion of the view, scene images always consist of multiple objects and background clutters. Figure 1 shows two samples from benchmark object-centric data set (ImageNet) and benchmark scene data set (Place365), respectively. Figure 1 (a) is labeled as "bald eagle", in which the eagle stands in the center of the view and occupies a large portion of the entire image; figure 1 (b) is labeled as "forest-broadleaf", the entire view consists of not only a bird but also tree branches and leaves. As the correct recognition of scenery images relies on multiple components, a scene image is typically considered more complex than an object-centric image.

(a) ImageNet (Bald eagle) (b) Place365 (Forest-broadleaf) (c) CUB-200-2011 (Black-footed albatross) (d) CUB-200-2011 (Sooty albatross)

Figure 1: Example images from benchmark object recognition data set ImageNet (a), benchmark scene recognition data set Places365 (b), and fine-grained data set Caltech-UCSD Birds-200-2011 (c and d).

To evaluate this hypothesis, we introduced another data set that is also widely believed to be "complex": the fine-grained image classification data set. Fine-grained classification is considered to be a more complex task as the classes in the data set can only be discriminated by local and subtle differences. CUB-200-2011 is a data set consists of 200 different species of birds, which serves as a benchmark data set for fine-grained classification tasks. Figure 1 shows two samples in the CUB-200-2011 data set. In Figure 1, black-footed albatross (c) and sooty albatross (d) are considered to be two different categories in classification. The two albatrosses are similar in appearance, and the differentiating of them is challenging due to the subtle traits that characterize the different species are not straightforward.

We conducted comparison experiments on two "complex" data sets (Places365 and CUB-200-2011) and one "simple" data set. The results are shown in Table 3. Using the benchmark ResNet-50 as the backbones, we observed that on CUB-200-2011, the relative top -1 accuracy increased by 1.81% when doubling the width and dropped 4.28% when we narrowed the width to half of the original. This performance change caused by altering the width is much acute compared to the result on Place365 (0.94% and 1.11%, doubling and halving the width respectively) under the same settings, which demonstrated that a wider network is able to effectively enhance the recognition of "complex" fine-grained features. This result does not agree with our first hypothesis: if the performance difference is originated from the complexity of the data, we should observe a moderate

performance change on CUB-200-2011 data set along with the changing of network width. The results proved that the performance variance on different data set is not the consequence of the complexity (rich fine-grained details) of the data.

Table 3: Top-1 accuracy (%) of ResNet-50 on Places365, ImageNet, and Caltech-UCSD Birds-200-2011 by changing network width.

Width Scaling Factor	Places	ImageNet	CUB-200-2011
2	56.21	78.51	71.53
1	55.69	76.02	70.26
0.5	55.07	72.08	67.25
0.25	52.16	64.04	61.29

4.2 SPATIAL V.S. CHANNEL

Based on the observations, we came up with the second hypothesis: for the scene recognition task, instead of learning more fine-grained features, learning spatial information is more crucial. As defined in (Fan et al., 2020), spatial information refers to the spatial ordering on the feature map. Intuitively, for the images that only contain one object, the semantic meaning related to spatial layout is limited; for scene images, the spatial structures, namely, scene contextual information likely to contribute more to the understanding of the scene. Thus, for scene recognition tasks, learning spatial information is more crucial.

To verify this hypothesis, we conducted the experiments by gradually passing low and high-frequency information on Place365 and ImageNet data sets. Generally speaking, the high-frequency information in the image refers to the regions where the intensity of the image (brightness/gray-scale) changes drastically, which are often called the edges or contour; the low-frequency information in the image refers to the regions where the image intensity changes smoothly, such as large patches of color. As shown in Figure 2, the image filtered by low-pass filter tends to present proximate or blurred patterns of the original image, the images filtered by high-pass filter better preserved the spatial information.

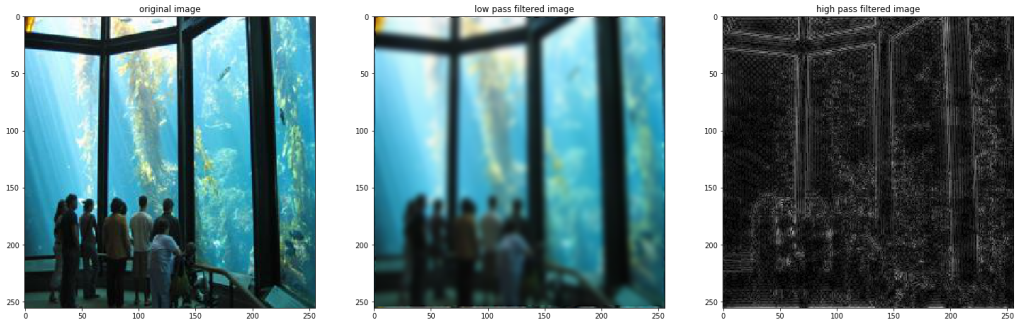
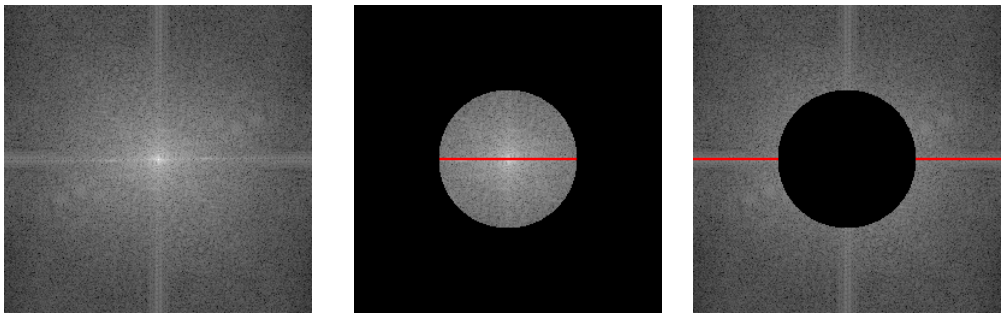


Figure 2: Results on scenery image using low and high pass filter.

To understand the importance of low and high frequency information in different data sets, we designed the low pass and high pass filter based on Fourier Transform. We transformed the testing images into spectrum domain using Fourier Transform and applied both low and high pass filters to test how low/high frequency information can affect the model performance on different data sets. Figure 3 shows the design of the filters: for low pass filters, we masked the high frequency components; and for high pass filters, we masked the low frequency components.

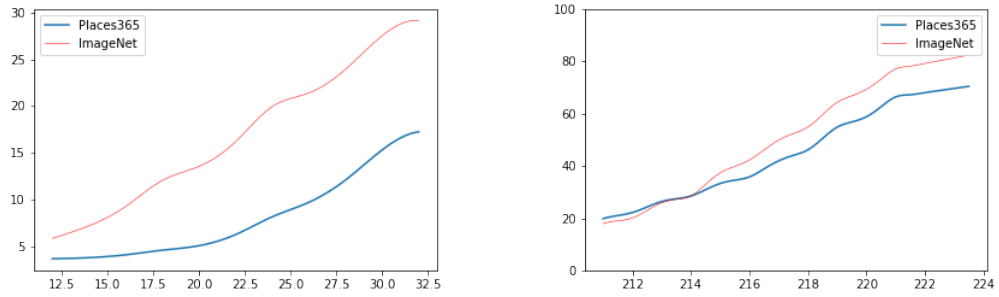
For a fair comparison, we randomly selected 100 classes from Place365 and ImageNet data sets to deploy the experiments. Note that as scene recognition is considered a harder task compared to object recognition, the classification accuracy on Place365 is lower despite the chance is the same. The results are shown in Figure 4. In both of the sub-figures, the x-axis denotes the size of the corresponding low/high pass filter in the spectrum domain (the maximum size is 224), and the y-axis denotes the top-1 accuracy. Through the comparison, we observed enlightening phenomena:



(a) Image in frequency domain (b) Low pass filter (c) High pass filter

Figure 3: Illustration of low and high pass filters. In figure (b) and (c), the mask (black region) denotes information removed by corresponding filters; the length of red line denotes the corresponding filter size.

when we gradually passing the low frequency information, the performance increase on ImageNet is steeper than on Place365 (Figure 4 (a)). Surprisingly, when we use the low pass filter of size 33, ImageNet can achieve a top-1 accuracy of nearly 30% while the chance is 1%. This denotes the correct recognition of object-centric images heavily relies on low frequency information. On the contrary, when we gradually passing the high frequency information, we can observe that the model trained on scenery data set is more sensitive to high-frequency information (Figure 4 (b)). Notably, when the size of the high-pass filter is around 210 to 214., the top-1 accuracy on Place365 even exceeds the top-1 accuracy on ImageNet. This observation demonstrated that recognition of scenery images is susceptible to high frequency information.



(a) (b)

Figure 4: Top-1 accuracy (%) on Place365 and ImageNet data sets using low pass filters (a) and high pass filters (b). The x-axis denotes the size of the corresponding filters.

Our observation perfectly fits the experimental results we discussed in Section 3. The high and low frequency information in images can approximately present the learned spatial and channel-wise information in deep networks. Wider networks have an expanded number of channels, which enables the network to learn more fine-grained features. Deep networks have an increasing number of layers and larger receptive fields, which enables the network to learn more spatial information. Thus for the scene recognition tasks, deepening the network is more efficient than increasing the width of the network. We proposed two directions to better enhance the scene recognition networks: deepening and narrowing the network, and decreasing the spatial information lost.

5 DEEP-NARROW NETWORK

5.1 DEEP-NARROW STRUCTURE

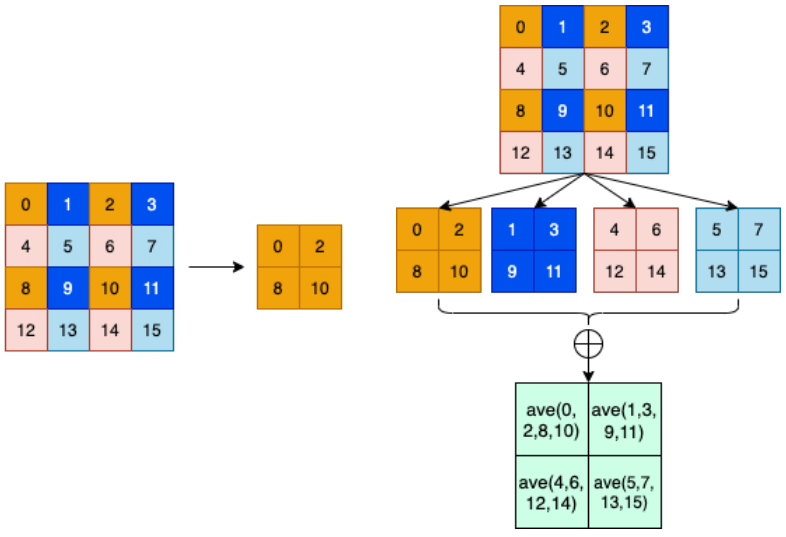
Based on our observation, we argued that as correctly recognizing scenery images is susceptible to the learning of spatial information, designing the networks with larger depth and smaller width can potentially be an effective and effective option. Based on this notion, we proposed a Deep-Narrow architecture that increases the number of layers in ResNet to 101 and decreases the width of the network to half of the size in benchmark ResNet.

Table 4 shows the performance comparisons among the benchmark ResNet-50, i.e., ResNet-50 ($\times$ 1), ResNet-50 with half the width, i.e., ResNet-50 ($\times$ .5)), and our Deep-Narrow Network on Place365 data set. Deep-Narrow Network achieves comparable evaluation scores with benchmark ResNet-50 using only less than half of the FLOPs and parameters: the relative top-1 actuary dropped 0.20%. On the ImageNet, the Deep-Narrow Architecture obtained a relative top-1 accuracy drop of 1.35%. The results re-certified that scene recognition is highly dependent on learning spatial information. Different from the assertions in the previous literature that widening of the network might provide a more effective way than by making ResNet deeper to improve the model performance, we demonstrated that data have their preference and the network design should rely on the characteristics of data.

Table 4: Top-1 accuracy (%) using ResNet with different depth and width. The number of parameters is in million. The numbers within parenthesis are the width scaling factors.

Data	Model	Top-1	Top-5	GFLOPs	Parameters
Place365	ResNet-50 ($\times$ 1)	55.69	85.80	4.12	24.26
	ResNet-50 ($\times$ .5)	55.07	85.12	1.07	6.27
	Deep-Narrow Network	55.58	85.80	2.00	11.03
ImageNet	ResNet-50 ($\times$ 1)	76.02	92.80	4.12	25.56
	ResNet-50 ($\times$ .5)	72.08	90.78	1.07	6.92
	Deep-Narrow Network	74.99	92.31	2.00	11.68

5.2 LOSSLESS POOLING



(a) (b)

Figure 5: The schema of down-sample component in ResNet and Lossless Pooling module from the view of spatial dimension. (a) Down-sample component in ResNet from the view of spatial dimension. During the down-sampling process, 3/4 of the spatial information are discarded and the number of channels is doubled. (b) Lossless Pooling module from the view of spatial dimension.

Table 5: Top-1 accuracy (%) using Deep-Narrow Network and Lossless Pooling (LP) module. The number of parameters is in million.

Data	Model	Top-1	Top-5	GFLOPs	Parameters
Place365	ResNet-50 ($\times$ 1)	55.69	85.80	4.12	24.26
	Deep-Narrow Networ	55.58	85.80	2.00	11.03
	Deep-Narrow Network (LP)	55.91	86.12	2.00	11.03
ImageNet	ResNet-50 ($\times$ 1)	76.02	92.80	4.12	25.56
	Deep-Narrow Network	74.99	92.31	2.00	11.68
	Deep-Narrow Network (LP)	74.63	92.13	2.00	11.68

Besides deepening the network to better process spatial information, we also designed a Lossless Pooling module to better preserve the spatial information. As shown in Figure 5(a), in ResNet, the down-sampling process is conducted by doubling the width of the network and discard 3/4 of the features along the spatial dimension. This design is not suitable for scene recognition as discarding spatial information will likely damage the performance. We designed a Lossless Pooling module to preserve the spatial information (Figure 5(b)). Instead of directly discarding 3/4 of the spatial information, we separate the features maps into four sections along the spatial dimension. Then we conduct convolution on the four feature maps and merged the result together via averaging operation. By leveraging Lossless Pooling, we are able to make use of all the spatial information without increasing the Flops and number of parameters.

As shown in Table 5, by integrating Lossless Pooling module with Deep-Narrow Network, our design outperforms benchmark ResNet using less than half of the FLOPs (computational resources) and the number of parameters on the benchmark Place365 data set. Specifically, adding Lossless Pooling to our Deep-Narrow Network brings a relative top-1 accuracy increase of 0.59% (Table 5), and leads to 0.48% relative top-1 accuracy drop on ImageNet. which demonstrated the effectiveness and efficiency of data-oriented network design.

To demonstrate the superiority of the proposed method, we compare with several state-of-the-art approaches that also tried to minimize the information lost caused by shrinking the spatial resolution, including ResNet-D (He et al., 2019) and Antialiased-CNN (Zhang, 2019). ResNet-D used average pooling and convolution operation to maintain more spatial information; Antialiased-CNN leveraged convolution-based pooling strategies to enhance the conventional pooling. We also implemented a baseline strategy named ResNet-Ave, which simply using averaging instead of discarding 3/4 information. We can observe that Deep-Narrow Network with Lossless Pooling achieves higher accuracy than the comparison methods using less computational resources.

Table 6: Top-1 accuracy (%) on Place365 using different backbones. The number of parameters is in million. Note that we calculate the number of parameters base on 365-class models (Place365).

Model	Top-1	Top-5	GFLOPs	# Parameters
ResNet-50 ($\times$ 1)	55.69	85.80	4.12	24.26
Deep-Narrow Network (LP)	55.91	86.12	2.00	11.03
ResNet-Ave	55.60	85.58	2.26	11.03
ResNet-D	55.79	85.93	2.26	11.03
Antialiased-CNN	55.85	85.93	2.26	11.03

6 CONCLUSION

This paper studies the impacts of scenery images in the network architecture design using ImageNet (object recognition data set) and Places365 (scene recognition data set) as examples. Carefully designed experiments showed that the characteristics of data sets affect the performance of the models: wider networks often achieve better performance of recognition of images with a prominent object and have less impact on the recognition of scenery images. We further evaluated our hypothesis by conducting comparison experiments and demonstrated that learning spatial-wise information is more substantial in scene recognition tasks compared to object classification tasks. This phenomenon explained why widening the networks is less effective than deepening the networks for scene recognition as deeper networks can better learn the spatial information in the training exam-

ples. Thus deploying the networks that have larger depth and smaller width, and emphasize the spatial information learning likely to benefit scene recognition backbone designs. Our proposed Deep-Narrow Network and Lossless Pooling module re-certified the effectiveness and efficiency of taking advantage of data properly. Our design achieved a better accuracy compared to benchmark ResNet-50 using less than half of the computation resources.

ACKNOWLEDGMENTS

The authors would like to thank Google for providing the academic research grant and computation resources.

REFERENCES

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, pp. 1251–1258, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Yue Fan, Yongqin Xian, Max Maria Losch, and Bernt Schiele. Analyzing the dependency of convnets on spatial information. In *DAGM German Conference on Pattern Recognition*, pp. 101–115. Springer, 2020.

Shaopeng Guo, Yujie Wang, Quanquan Li, and Junjie Yan. Dmcp: Differentiable markov channel pruning for neural networks. In *CVPR*, pp. 1539–1547, 2020.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, Los Alamitos, CA, USA, Jun 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016b.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NIPs*, 25:1097–1105, 2012.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755. Springer, 2014.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: a view from the width. In *Nips*, pp. 6232–6240, 2017.

Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do wide and deep networks learn the same things? Uncovering how neural network representations vary with width and depth. *ICLR*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pp. 1–9, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pp. 2818–2826, 2016.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, pp. 4278–4284, 2017.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, pp. 6105–6114. PMLR, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pp. 1492–1500, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*. British Machine Vision Association, 2016.

Hang Zhang, Chongruo Wu, Zhongyue Zhang, Yi Zhu, Haibin Lin, Zhi Zhang, Yue Sun, Tong He, Jonas Mueller, R Manmatha, et al. Resnest: Split-attention networks. *arXiv preprint arXiv:2004.08955*, 2020.

Richard Zhang. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*, pp. 7324–7334. PMLR, 2019.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *ICLR*, 2017.