Efficient Equivariant Network Supplementary Materials

Anonymous Author(s)
Affiliation
Address
email

A Detail Analysis of Data Efficiency

As shown in Table 5 and Figure 3, the test error of all models increase as training data size decreases. Let's take a closer look at the gap between different models. First, the gap between R18 and $p4$-R18 is 2.17% at $5k$ and increases to 6.02% at $1k$, while the gap between $p4$-R18 and $p4m$-R18 is 1.7% at $5k$ and increases to 3.69% at $1k$. This indicates symmetry prior of model plays a more important role when data is less. Then, we compare our models with their corresponding counterparts. The gap between $p4$-R18($p4m$-R18) and $p4$-E^4R18($p4m$-E^4R18) is $1.11\%(0.87\%)$ at $5k$, and it increases to $5.3\%(5.22\%)$ at $1k$. Furthermore, in Figure 3, the red line($p4$-E^4R18) and blue dash line($p4m$-R18) intersect, which means that $p4$-E^4R18 is even more superior than $p4m$-R18 when data is less although the latter one has more capacity when we increase training data. The above results demonstrate greater data efficiency of E^4 model compared with G-CNNs, or it can better utilize the symmetry prior of data.

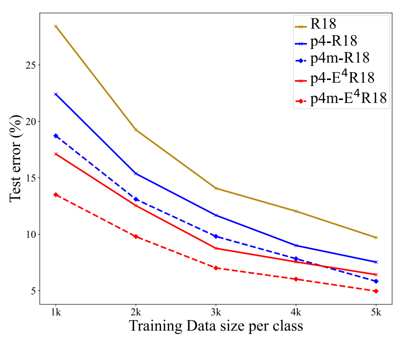

Figure 3: Trend of test error(%) on various training data sizes.

Table 5: Test error(%) of models trained on different training data sizes

Module/Dataset size	1k	2k	3k	4k	5k
R18	28.43	19.25	14.07	12.04	9.70
$p4$-R18	22.41	15.37	11.68	9.01	7.53
$p4$-E^4R18(Ours)	17.11	12.54	8.75	7.55	6.42
$p4m$-R18	18.72	13.10	9.81	7.68	5.83
$p4m$-E^4R18(Ours)	13.50	9.79	7.01	6.02	4.96

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

B MNIST-rot Model Architecture

Please refer to Table 6.

Table 6: Architecture of E^4-Net on Mnist-rot classification, p means dropout rate.

Layer	Kernel size	Output channels
Group convolution BatchNorm+ReLu	3×3	16
E^4-layer BatchNorm+ReLu Spatial-wise max pooling	5×5 2×2	16
E^4-layer BatchNorm+ReLu+Dropout(p=0.1)	5×5	16
E^4-layer BatchNorm+ReLu+Dropout(p=0.1) Spatial-wise max pooling	5×5 2×2	16
E^4-layer BatchNorm+ReLu+Dropout(p=0.1)	5×5	16
E^4-layer BatchNorm+ReLu+Dropout(p=0.1)	5×5	16
E^4-layer BatchNorm+ReLu+Dropout(p=0.1) Global max group pooling layer	5×5	16
Fully connected+Softmax		10

Here, Global max group pooling [1] is the operation act on each channel:

$$f_i = max_{g \in \mathcal{G}} f_i(g), \tag{13}$$

and i denotes the channel index. The hyperparameters we use in this architecture are kernel size $k = 5$, reduction ratio $r = 1$, and the number of slices $s = 2$. In the large model, we increase the channel dimension to 24, the number of slices to 3, and keep other hyperparameters the same.

C Details of CIFAR10 and CIFAR100 Experiments

We take standard ResNet-18 [2], which is composed of an initial convolution layer, followed by 4 stage Res-Blocks and one final classification layer. Each stage contain 2 Res-Blocks and each block contain 2 convolution layers. The channel dimensions of each stage of Res-Blocks are 64→64→128→256→512. For $p4$-R18 ($p4m$-R18), we replace all the conventional layers with $p4$ ($p4m$) convolutions layers, and modify the channel dimensions at each stage Res-Blocks as 32→32→64→128→256 (22→22→44→88→176) to keep the parameter almost invariant. Then, for building our model $p4$-E^4R18 ($p4m$-E^4R18), the second group convolution layer in each Res-Block of $p4$-R18 ($p4m$-R18) is replaced by our E^4-layer with $k = 3$, $r = 2$, and $s = C_l/2$. Here, C_l is the channel dimension at that layer.

References

[1] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, pages 2990–2999, 2016.

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