Mixconv-EfficientNeSt-Ghost (ME-NeSt-Ghost): Using a combination of mixed depthwise convolutions, Split-Attention units and Ghost convolutions for improving EfficientNet-V2 performance: Final Report

G024 (s1893731)

Abstract

This report introduces Mixconv-EfficientNeSt-Ghost, a novel CNN model architecture family that is built to improve upon the EfficientNet-V2 family of convolutional neural networks with the help of a combination of mixed-depthwise convolutions, split attention modules and ghost convolutions. The Mixconv-EfficientNeSt-Ghost family of models is more parameter efficient than the EfficientNet-V2 while at the same time providing comparable or even better results in image classification tasks. The novel model architecture was built upon the EfficientNet-V2 architecture with the Fused-MBConv block of the EffNet-V2 being replaced by the MixConv-Split-Attention Block (MS Block) and the MBConv block being replaced by the MixConv-SE-GhostConv Block (MS-Ghost Block). The performance of both models on the CIFAR-100 dataset was analysed.

1. Introduction

The recent trend in ConvNets design is to improve accuracy and parameter efficiency simultaneously. The EfficientNet (Tan & Le, 2019a) and EfficientNet-V2 (Tan & Le, 2021) family have stressed the importance of parameter efficiency, faster training and compound scaling in Convolutional Neural Networks. Depthwise convolutions are becoming increasingly popular and are an essential part of the EfficientNet Mobile-Conv (MB-Conv) building blocks. Depthwise convolutional kernels are applied to each individual channel separately, resulting in a reduction of the computational cost by a factor of C, where C is the number of channels. However, a disadvantage of these is that regular depthwise convolutions like the ones used in the EfficientNet (Tan & Le, 2019a), EfficientNet-V2 (Tan & Le, 2021), MobileNet and ShuffleNet (Zhang et al., 2017) overlook the factor of kernel size, having a fixed kernel size of 3x3, while research has shown that larger kernel sizes of 5x5 and 7x7 are useful for improving accuracy and efficiency. For this purpose Mingxing Tan and Quoc V. Le came up with mixed depthwise convolutional kernels (MixConv) (Tan & Le, 2019b) which naturally mixes up multiple kernel sizes in a single convlution resulting in an increase in both accuracy and parameter efficiency. The MixConv-EfficientNeSt-Ghost replaces the depthwise convolutions of the EfficientNet model with mixed-depthwise convolutions.

Another important module used for building this novel model architecture is the Split Attention module inspired by the paper ResNeSt:Split Attention Networks (Zhang et al., 2020) which proposes a channel-wise attention strategy applied with a multi-path network layout. This module can provide a generalization for both SK-Unit and Squeeze-Excitation module channel attention mechanisms and is utilised in the MixConv-Split Attention Block (MS-Block) which is a building block of this novel architecture.

Yet another interesting area of CNN research utilised in this novel model architecture is the Ghost module inspired from the paper GhostNet: More features from cheap operations (Han et al., 2019), which proposes a module that generates multiple ghost feature maps on applying a series of cheap linear transformations which when concatenated together reveal underlying information regarding intrinsic features. The ghost modules were utilised in the MixConv-EfficientNeSt-Ghost to replace the 1x1 convolutions carried out in the MB-Conv block of the EfficientNet models in order to increase accuracy and parameter efficiency simultaneously.

Overall, the ME-NeSt-Ghost family combines the advantages of large kernel sizes (with the help of mixed depthwise convolutions), along with the advantages of channel-wise attention (with the help of Split-Attention modules and Squeeze Excitation layers), along with the advantages of the cheap operations of Ghost convolutions to create a CNN family that is more parameter efficient than the EfficientNet-V2 while following identical compound scaling guidelines.

2. Data set and task

A resized CIFAR-100 dataset (cif) with 128x128 images was utilised for training the models for an image classification task. The 32x32 CIFAR-100 images were super resolved using the CAI Neural API (cai). Both the EfficientNet-V2 family and the ME-NeSt-Ghost family were trained on the dataset and the Top-1 accuracy of the image classification task on the test set was analysed. The training sets of the dataset consists of 50,000 images and the test set consists of 10,000 images. The training set was split into a ratio of 95:5 with the validation set having 2500 images and the training set 47,500 images.

Block Type	Layers	Channels	Stride
Conv3x3	1	24	2
MixConv-Split-Attention Block	2	24	1
MixConv-Split-Attention Block	4	48	2
MixConv-Split-Attention Block	4	64	2
MixConv-SE-Ghost Block	6	128	2
MixConv-SE-Ghost Block	9	160	1
MixConv-SE-Ghost Block	15	256	2
Conv1x1 + Pooling + FC	1	1792	-

Table 1. ME-NeSt-Ghost-S architecture

Block Type	Layers	Channels	Stride
Conv3x3	1	24	2
MixConv-Split-Attention Block	3	24	1
MixConv-Split-Attention Block	5	48	2
MixConv-Split-Attention Block	5	80	2
MixConv-SE-Ghost Block	7	160	2
MixConv-SE-Ghost Block	14	176	1
MixConv-SE-Ghost Block	18	304	2
MixConv-SE-Ghost Block	5	512	1
Conv1x1 + Pooling + FC	1	1792	-

Table 2. ME-NeSt-Ghost-M architecture

3. Methodology

The novel ME-NeSt-Ghost achitecture follows the same guidelines of compound scaling that were followed in the EfficientNet-V2 (Tan & Le, 2021) with identical width scaling, depth scaling and resolution scaling, with the ME-NeSt-Ghost S, ME-NeSt-Ghost M, ME-NeSt-Ghost L, ME-NeSt-Ghost XL being the corresponding versions of the EfficientNet-V2 S, EfficientNet-V2 M, EfficientNet-V2 L and EfficientNet-V2 XL respectively. The main difference being that the novel model architecture consists of 2 building blocks - MixConv-Split-Attention Block and the Mixconv-Squeeze-Excitation-Ghost Block which differ considerably from their corresponding Fused-MB-Conv Block and MB-Conv Block in the EfficientNet-V2 model architecture respectively.

Block Type	Layers	Channels	Stride
Conv3x3	1	24	2
MixConv-Split-Attention Block	4	32	1
MixConv-Split-Attention Block	7	64	2
MixConv-Split-Attention Block	7	96	2
MixConv-SE-Ghost Block	10	192	2
MixConv-SE-Ghost Block	19	224	1
MixConv-SE-Ghost Block	25	384	2
MixConv-SE-Ghost Block	7	640	1
Conv1x1 + Pooling + FC	1	1792	-

Table 3. ME-NeSt-Ghost-L architecture

Block Type	Layers	Channels	Stride
Conv3x3	1	24	2
MixConv-Split-Attention Block	4	32	1
MixConv-Split-Attention Block	8	64	2
MixConv-Split-Attention Block	8	96	2
MixConv-SE-Ghost Block	16	192	2
MixConv-SE-Ghost Block	24	256	1
MixConv-SE-Ghost Block	32	512	2
MixConv-SE-Ghost Block	8	640	1
Conv1x1 + Pooling + FC	1	1792	-

Table 4. ME-NeSt-Ghost-XL architecture

3.1. MixConv-Split-Attention Block

This block consists of a mixed depthwise convolutional module with kernel sizes 3x3, 5x5, 7x7, 9x9, followed by a split attention module with radix = 2 acting as an SK-Unit module (Li et al., 2019) and applying channel-wise attention on 2 branches. Research on attention methods indicates that attention mechanisms are more effective when applied to the initial layers of a convolutional neural network, since they provide a global context of a feature map which is only captured by the deeper layers of a regular convolutional neural network. Due to this reason and for the sake of parameter efficiency, the MixConv-Split-Attention Block constitutes the first 10 blocks of the ME-NeSt-Ghost-S, the first 13 blocks of ME-NeSt-Ghost-M, the first 18 blocks of ME-NeSt-Ghost-L and the first 20 blocks of ME-NeSt-Ghost-XL respectively while the rest of the network is built using the MixConv-Squeeze-Excitation-Ghost Block.

3.1.1. Mixed-Depthwise Convolutions: Details

Mixed Depthwise Convolutions (MixConv) (Tan & Le, 2019b) combine the computational advantages of depthwise convolutions with the advantages of large kernel sizes in convolutional neural networks. Kernel sizes are naturally mixed up in a single convolution by partitioning the channels into different groups and applying depthwise convolutions of different kernel sizes to each group. The channel size per group can be chosen in 2 ways -

1. Equal Partition - Each group has the same number of filters/channels.

2. Exponential Partition - The i^{th} group is provided 2^{-i} portion of the channels.

For example, given a 4-group MixConv with total channel size 32, the equal partition will divide the channels into (8, 8, 8, 8), while the exponential partition will divide the channels into (16, 8, 4, 4).

For simplicity, equal partition was chosen for the ME-NeSt-Ghost model architecture. After the convolution, Batch Normalization and activation function were utilised.

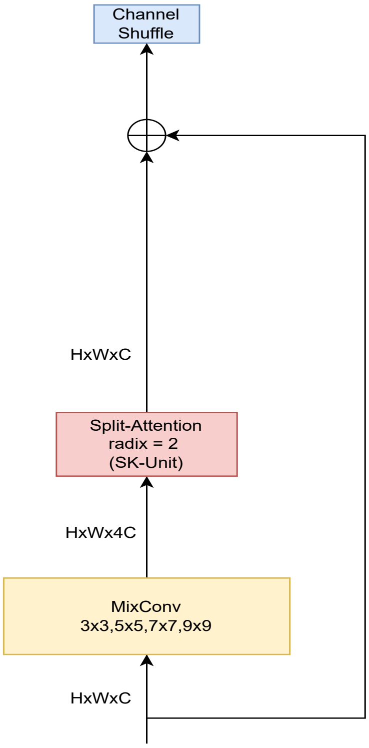

Figure 1. MixConv-Split-Attention Block consisting of a mixed depthwise convolutional module with kernel sizes - 3x3, 5x5, 7x7, 9x9. Followed by a Split Attention module with radix = 2 which acts like an SK-Unit (Li et al., 2019) carrying out channel-wise attention on 2 branches. Finally channel shuffle is carried out after the skip connection in order to aid information flow across feature channels, taking inspiration from ShuffleNet (Zhang et al., 2017)

3.1.2. Split Attention Units: Details

Split Attention units play the useful role of generalizing featuremap attention within a cardinal group setting in a computationally efficient manner. A split attention unit with cardinality k = 1 and radix = 1 applies "squeeze and excitation" an SE like attention mechanism to apply a global context to predict channel-wise attention factors.

In this particular implementation of the Split-Attention unit, input featuremap is first divided into r = 2 groups in which each group has a cardinality k = 1. A global pooling layer then aggregates over the spatial dimensions while keeping channel dimensions separated. After which 2 1x1 convolutions with number of groups = cardinality = 1 are added after the pooling layer to predict the attention weights for each split.

3.2. MixConv-Squeeze-Excitation-Ghost Block

The MixConv-SE-Ghost Blocks are utilised in the latter half of the model architecture and account for majority of the layers in the architecture. The MixConv-SE-Ghost Block corresponds to the MBConv block of the EfficientNet and EfficientNet-V2 model architectures. In order to make the MBConv Block more parameter efficient, 1x1 convolutions have been replaced by GhostConv operations and 3x3 depthwise convolutions have been replaced with a mixed depthwise convolution with kernel sizes 3x3,5x5,7x7,9x9.

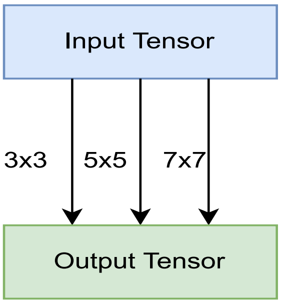

Figure 2. Example of a Mixed Depthwise Convolution where channels are divided into groups and a depthwise convolution of kernel size 3x3 is applied on the first group, a depthwise convolution of kernel size 5x5 applied to the second group and a depthwise convolution of kernel size 7x7 is applied to the third group

3.2.1. GhostConv: Details

GhostConv operations are parameter efficient operations described in (Han et al., 2019) which reduce the number of FLOPS required in the computation of the convolution. First a convolution operation is carried out on an input featuremap of HxWxC and a featuremap of HxWxC/2 is generated, on this featuremap a cheap depthwise convolutional operation of either kernel size 3x3 or 5x5 is applied to generate another HxWxC/2 featuremap, finally these two featuremaps are then concatenated together to generate a HxWxC featuremap.

A novel Mix-GhostConv module has been utilised for the ME-NeSt-Ghost model architecture. In this module the depthwise convolution of the original GhostConv operation is replaced with a mixed depthwise convolution with kernel sizes 3x3,5x5,7x7 and 9x9.

3.2.2. Squeeze Excitation Layer

The squeeze excitation layer (Hu et al., 2017) squeezes global spatial information into a channel descriptor by using Global Average Pooling/ Adaptive Average Pooling to generate channel-wise statistics. This aggregates global information about the whole image. In order to make use of this information and to capture channelwise dependencies, an "excitation" operation is carried out with the help of two Fully Connected layers in order to learn the non-linear interactions between channels.

4. Experiments

All experiments were carried out on Kaggle Notebooks with one Tesla P-100 GPU with 16GB RAM.

The aim of the experiments was to train the EffNet-V2-

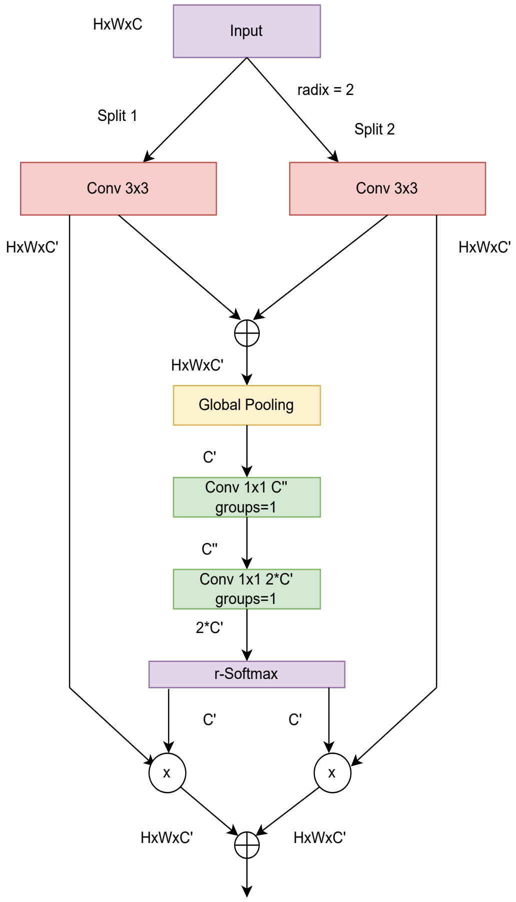

Figure 3. Split Attention Unit from (Zhang et al., 2020) with radix=2 and cardinality=1 has been utilised in the MixConv-Split-Attention block which applies SK-Unit (Li et al., 2019) like featuremap attention to 2 network streams.

S,M,L,XL and the ME-NeSt-Ghost-S,M,L,XL on the resized CIFAR-100 dataset with 128x128 images and compare the Top-1 Accuracy of the model on the test set. The training set was divided into a 95:5 ratio to produce a training set and a validation set.

In order to carry out the experiments, the images were resized to 256x256 for training all the models.

RandAugment (Cubuk et al., 2019) was utilized for carrying out image augmentations. It is a per-image augmentation strategy with adjustable magnitude ϵ which was chosen to be 10 for these experiments.

Each of the models was trained for 100 epochs using a **Label Smoothing Cross Entropy Loss** and the **AdamW optimizer** with a learning rate of $1e-4$ and weight decay of $1e-6$. **Cosine Annealing Learning Rate Scheduler** was utilised with a minimum learning rate of $1e-6$.

5. Related work

This novel architecture builds upon the EfficientNet-V2 (Tan & Le, 2021) by utilising the advantages of large kernel sized utilised in a parameter efficient manner in (Tan & Le, 2019b) along with the split attention units (Zhang et al., 2020) and parameter efficient GhostConv (Han et al., 2019) operations.

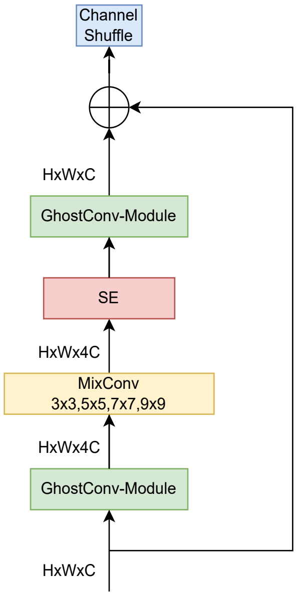

Figure 4. MixConv-SE-Ghost Block consisting of 2 GhostConv (Han et al., 2019) operations replacing the 1x1 convolutions of the MB-Conv Block of the EfficientNet-V2 (Tan & Le, 2021) in addition to a MixConv with kernel sizes 3x3,5x5,7x7,9x9 along with a Squeeze Excitation (SE) Layer (Hu et al., 2017) and finally a channel shuffle operation carried out after the skip connection.

Model	Number of Parameters	Top-1 Accuracy
ME-NeSt-Ghost-S	21M	**80.3%**
EffNet-V2-S	**20M**	79.6%
ME-NeSt-Ghost-M	**51M**	**80.8%**
EffNet-V2-M	53M	80.4%
ME-NeSt-Ghost-L	**102M**	**81.2%**
EffNet-V2-L	117M	80.7%
ME-NeSt-Ghost-XL	**172M**	**81.7%**
EffNet-V2-XL	207M	81.4%

Table 5. Model Performance (Top-1 Accuracy) on CIFAR-100 Test set

5.1. Comparison with EffNetV2 -

As discussed above, the novel model architecture follows the same compound scaling guidelines as the EffNet-V2 with the main difference being the building blocks of the 2 convolutional neural networks. With the Fused MB-Conv (Figure 8) being replaced by the MixConv-Split-Attention Block (Figure 1) and the MB-Conv (Figure 9)block being replaced by the MixConv-SE-Ghost Block (Figure 4).

Additionally, unlike the EfficientNet-V2 which uses the Swish/SiLU activation function, the Mish activation (Misra,

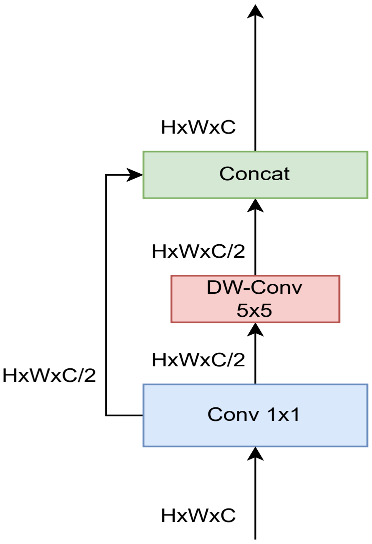

Figure 5. A standard GhostConv (Han et al., 2019) Module a HxWxC input feature map is first mapped to a HxWxC/2 feature map, a depthwise convolution of kernel size 5x5 is then applied to this feature map and the resulting feature map is then concatenated with the original HxWxC/2 feature map generated by the 1x1 Convolution. This increases accuracy and parameter efficiency especially for deeper layers with larger number of filters.

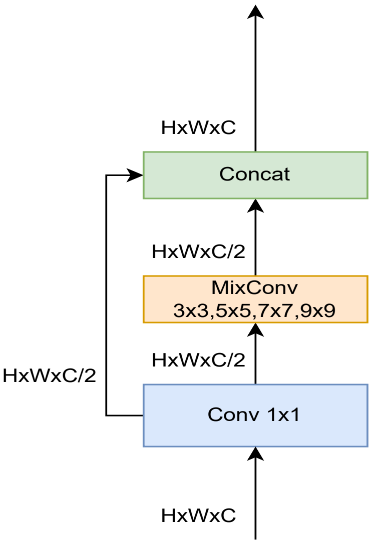

Figure 6. The novel Mix-GhostConv module used in the ME-NeSt-Ghost model architecture where instead of applying a depthwise convolution on the HxWxC/2 feature map, a mixed depthwise convolution with kernel sizes 3x3,5x5,7x7,9x9 is applied with the intention of introducing larger kernel sizes into the GhostConv module in a parameter efficient manner.

2019) was utilised for training the ME-NeSt-Ghost model architectures which is a regularized non-monotonic activation function with properties similar to the Swish activation function.

Channel Shuffle operations carried out after each skip connection were a cherry on top, inspired by (Zhang et al., 2017), it is an operation that helps information flow across feature channels in convolutional neural networks. Additionally, channel shuffle plays a useful role in acting as a regularizer (Kumawat et al., 2021).

6. Conclusions

From the experiments conducted it can be concluded that the ME-NeSt-Ghost family of model architectures manages to slightly outperform the EfficientNet-V2 while being more parameter efficient especially for the larger models of the family like the ME-NeSt-Ghost-L and ME-NeSt-Ghost-XL which had 102 million and 172 million parameters respectively compared to the 117 million parameters and 207 million parameters of the EfficientNet-V2-L and EfficientNet-V2-XL. The ME-NeSt-Ghost architectures trade-off model inference speed for model accuracy and parameter efficiency, while it is more parameter efficient and more accurate than the EfficientNet-V2, its inference time is approximately 4-6x higher than the EfficientNet-V2 which makes model training much slower. Though the ME-NeSt-Ghost is yet to be trained on large scale datasets like ImageNet, it has the potential to achieve SOTA results. It would be interesting to compare the model's performance with that of the modern CNNs that utilise Transformer Self Attention along with convolutions, for example the ConvNeXt (Liu et al., 2022) and CoAtNet (Dai et al., 2021). It would be interesting to analyse if split attention units which carry out featuremap attention on multiple branches can compete with the sophisticated transformer self attention mechanisms. The future of convnets today lies in a combination of self attention mechanisms and convolutions, resulting in more and more hybrd models achieving SOTA results, and it would be interesting to analyse the performance of the ME-NeSt-Ghost on replacing split attention units with transformer self attention. Yet another interesting experiment would be to build the ME-NeSt-Ghost in a ResNet-like fashion with the MixConv-Split-Attention Block and the MixConv-SE-Ghost Block being utilised in the 4 stages of a ResNet model architecture, which in effect would question the effectiveness of the compound scaling techniques of the EfficientNet-V2 which were established with the help of Neural Architecture Search (NAS).

References

Cai neural api. URL https://github.com/joaopauloschuler/neural-api.

Cifar-100 resized using cai super-resolution. URL https://www.kaggle.com/datasets/joaopauloschuler/cifar100-128x128-resized-via-cai-super-resolution.

Cubuk, Ekin D., Zoph, Barret, Shlens, Jonathon, and Le, Quoc V. Randaugment: Practical data augmentation with

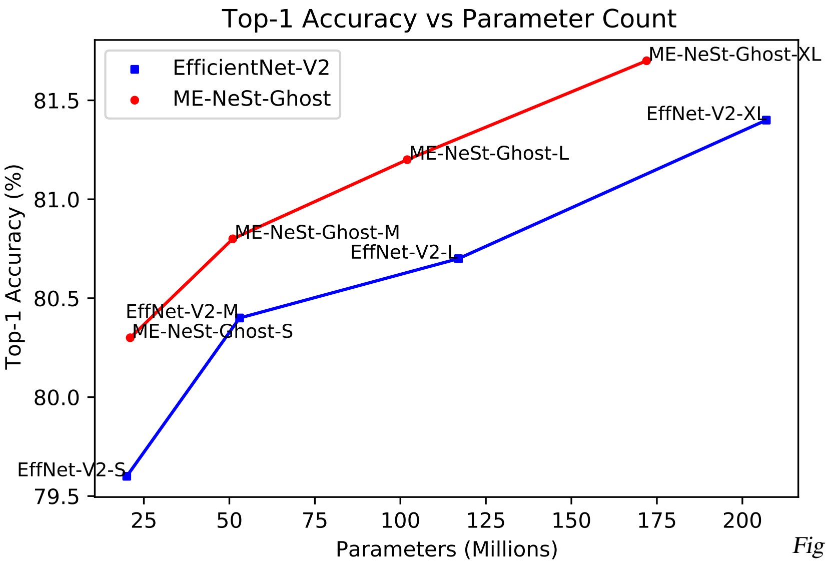

Figure 7. Top-1 Accuracy vs Parameter Count comparison for EfficientNet-V2 and ME-NeSt-Ghost on the CIFAR-100 Test Set

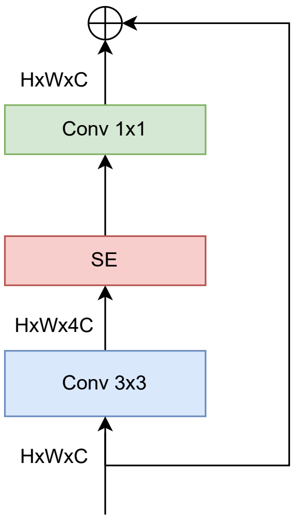

Figure 8. The Fused MB-Conv Block of the EfficientNet-V2

no separate search. *CoRR*, abs/1909.13719, 2019. URL http://arxiv.org/abs/1909.13719.

Dai, Zihang, Liu, Hanxiao, Le, Quoc V., and Tan, Mingxing. Coatnet: Marrying convolution and attention for all data sizes. *CoRR*, abs/2106.04803, 2021. URL https://arxiv.org/abs/2106.04803.

Han, Kai, Wang, Yunhe, Tian, Qi, Guo, Jianyuan, Xu, Chunjing, and Xu, Chang. Ghostnet: More features from cheap operations. *CoRR*, abs/1911.11907, 2019. URL http://arxiv.org/abs/1911.11907.

Hu, Jie, Shen, Li, and Sun, Gang. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017. URL http://arxiv.org/abs/1709.01507.

Kumawat, Sudhakar, Kanojia, Gagan, and Raman, Shanmuganathan. Shuffleblock: Shuffle to regularize deep convolutional neural networks. *CoRR*, abs/2106.09358, 2021. URL https://arxiv.org/abs/2106.09358.

Li, Xiang, Wang, Wenhai, Hu, Xiaolin, and Yang, Jian. Selective kernel networks, 2019.

Liu, Zhuang, Mao, Hanzi, Wu, Chao-Yuan, Feichtenhofer, Christoph, Darrell, Trevor, and Xie, Saining. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. URL https://arxiv.org/abs/2201.03545.

Misra, Diganta. Mish: A self regularized non-monotonic neural activation function. *CoRR*, abs/1908.08681, 2019. URL http://arxiv.org/abs/1908.08681.

Tan, Mingxing and Le, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019a. URL http://arxiv.org/abs/1905.11946.

Tan, Mingxing and Le, Quoc V. Mixconv: Mixed depthwise convolutional kernels. *CoRR*, abs/1907.09595, 2019b. URL http://arxiv.org/abs/1907.09595.

Tan, Mingxing and Le, Quoc V. Efficientnetv2: Smaller models and faster training. *CoRR*, abs/2104.00298, 2021. URL https://arxiv.org/abs/2104.00298.

Zhang, Hang, Wu, Chongruo, Zhang, Zhongyue, Zhu, Yi, Zhang, Zhi, Lin, Haibin, Sun, Yue, He, Tong, Mueller, Jonas, Manmatha, R., Li, Mu, and Smola, Alexander J. Resnest: Split-attention networks. *CoRR*, abs/2004.08955, 2020. URL https://arxiv.org/abs/2004.08955.

Zhang, Xiangyu, Zhou, Xinyu, Lin, Mengxiao, and Sun, Jian. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *CoRR*, abs/1707.01083, 2017. URL http://arxiv.org/abs/1707.01083.

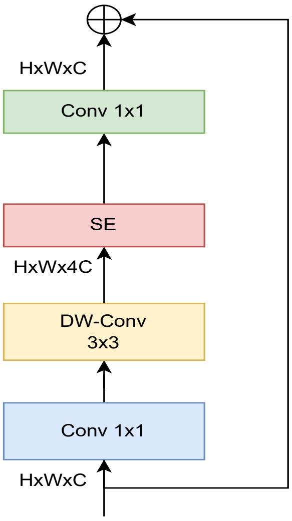

Figure 9. The MB-Conv Block of the EfficientNet-V2