Squeeze-and-Excitation Networks

Abstract

“Squeeze-and-Excitation” (SE) block,
adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels
slight additional computational cost.

Introduction

CNN
fusing spatial and channel-wise information together within local receptive fields
the relationship between channels
feature recalibration
use global information to selectively emphasise informative features and suppress less useful ones

first passed through a squeeze operation, which produces a channel descriptor by aggregating feature maps across their spatial dimensions
excitation operation, which takes the form of a simple self-gating mechanism that takes the embedding as input and produces a collection of per-channel modulation weights

In earlier layers, it excites informative features in a class-agnostic manner, strengthening the shared low-level representations.
. In later layers, the SE blocks become increasingly specialised, and respond to different inputs in a highly class-specific manner
feature recalibration

can be used directly
computationally lightweight
impose only a slight increase in model complexity and computational burden

Related Work

Deeper Architectures
Much of this research has concentrated on the objective of reducing model and computational complexity, reflecting an assumption that channel relationships can be formulated as a composition of instance-agnostic
dynamic, non-linear dependencies between channels -> ease learning process / significantly enhance the representational power

Algorithmic Architecture Search
SE blocks can be used as atomic building blocks for these search algorithms

Attention and Gating Mechanisms
SE block - lightweight gating mechanism
modelling channel-wise relationships in a computationally efficient manner

Squeeze-and-Excitation Blocks

$$\mathbf{u}_c = \mathbf{v}_c * \mathbf{X} = \sum_{s=1}^{C'} \mathbf{v}_c^s * \mathbf{x}^s.$$

Squeeze: Global Information Embedding
unable to exploit contextual information outside of this region

-> squeeze global spatial information into a channel descriptor
 global average pooling

Excitation: Adaptive Recalibration
fully capture channel-wise dependencies.
-> Excitation

조건 두 가지
it must be flexible
must learn a non-mutually-exclusive relationship

$$\mathbf{s} = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z})),$$

Instantiations
The SE block can be integrated into standard architectures

MODEL AND COMPUTATIONAL COMPLEXITY
good trade-off between improved performance and increased model complexity
ResNet-50 and SE-ResNet-50
ResNet-50 ~3.86 GFLOPs
SE-ResNet-50 requires ~3.87 GFLOPs
0.26% relative increase
the accuracy of SE-ResNet-50 = ResNet-101 network requiring ~7.58 GFLOPs

	original		re-implementation			SENet		
	top-1 err.	top-5 err.	top-1 err.	top-5 err.	GFLOPs	top-1 err.	top-5 err.	GFLOPs
ResNet-50 [13]	24.7	7.8	24.80	7.48	3.86	$23.29_{(1.51)}$	$6.62_{(0.86)}$	3.87
ResNet-101 [13]	23.6	7.1	23.17	6.52	7.58	$22.38_{(0.79)}$	$6.07_{(0.45)}$	7.60
ResNet-152 [13]	23.0	6.7	22.42	6.34	11.30	$21.57_{(0.85)}$	$5.73_{(0.61)}$	11.32
ResNeXt-50 [19]	22.2	-	22.11	5.90	4.24	$21.10_{(1.01)}$	$5.49_{(0.41)}$	4.25
ResNeXt-101 [19]	21.2	5.6	21.18	5.57	7.99	$20.70_{(0.48)}$	$5.01_{(0.56)}$	8.00
VGG-16 [11]	-	-	27.02	8.81	15.47	$25.22_{(1.80)}$	$7.70_{(1.11)}$	15.48
BN-Inception [6]	25.2	7.82	25.38	7.89	2.03	$24.23_{(1.15)}$	$7.14_{(0.75)}$	2.04
Inception-ResNet-v2 [21]	$19.9^{\dagger}$	$4.9^{\dagger}$	20.37	5.21	11.75	$19.80_{(0.57)}$	$4.79_{(0.42)}$	11.76

SE-ResNet-50, it adds approximately 2.5 million additional parameters, which is around a 10% increase compared to the ~25 million parameters required by ResNet-50
Most of these parameters come from the final stage of the network.
removing final stage has a minimal impact on performance (4%)
그러니까 제거를 고려해볼 수 있다

Experiments (Classification)

	original	re-implementation	SENet

	top-1 err.	top-5 err.	top-1 err.	top-5 err.	GFLOPs	top-1 err.	top-5 err.	GFLOPs
ResNet-50 [13]	24.7	7.8	24.80	7.48	3.86	$23.29_{(1.51)}$	$6.62_{(0.86)}$	3.87
ResNet-101 [13]	23.6	7.1	23.17	6.52	7.58	$22.38_{(0.79)}$	$6.07_{(0.45)}$	7.60
ResNet-152 [13]	23.0	6.7	22.42	6.34	11.30	$21.57_{(0.85)}$	$5.73_{(0.61)}$	11.32
ResNeXt-50 [19]	22.2	-	22.11	5.90	4.24	$21.10_{(1.01)}$	$5.49_{(0.41)}$	4.25
ResNeXt-101 [19]	21.2	5.6	21.18	5.57	7.99	$20.70_{(0.48)}$	$5.01_{(0.56)}$	8.00
VGG-16 [11]	-	-	27.02	8.81	15.47	$25.22_{(1.80)}$	$7.70_{(1.11)}$	15.48
BN-Inception [6]	25.2	7.82	25.38	7.89	2.03	$24.23_{(1.15)}$	$7.14_{(0.75)}$	2.04
Inception-ResNet-v2 [21]	$19.9^{\dagger}$	$4.9^{\dagger}$	20.37	5.21	11.75	$19.80_{(0.57)}$	$4.79_{(0.42)}$	11.76

	original		re-implementation				SENet			
	top-1 err.	top-5 err.	top-1 err.	top-5 err.	MFLOPs	Params	top-1 err.	top-5 err.	MFLOPs	Params
MobileNet [64]	29.4	-	28.4	9.4	569	4.2M	$25.3_{(3.1)}$	$7.7_{(1.7)}$	572	4.7M
ShuffleNet [65]	32.6	-	32.6	12.5	140	1.8M	$31.0_{(1.6)}$	$11.1_{(1.4)}$	142	2.4M

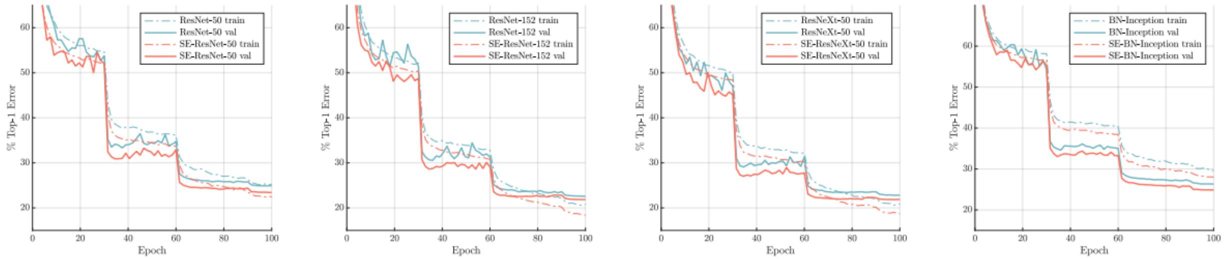


TABLE 4
Classification error (%) on CIFAR-10.

	original	SENet
ResNet-110 [14]	6.37	5.21
ResNet-164 [14]	5.46	4.39
WRN-16-8 [67]	4.27	3.88
Shake-Shake 26 2x96d [68] + Cutout [69]	2.56	2.12

TABLE 5
Classification error (%) on CIFAR-100.

	original	SENet
ResNet-110 [14]	26.88	23.85
ResNet-164 [14]	24.33	21.31
WRN-16-8 [67]	20.43	19.14

Shake-Even 29 2x4x64d [68] + Cutout [69]	15.85	15.41

ABLATION STUDY

ImageNet 8 GPUs ResNet-50

Removed biases from the FC layers

Label-smoothing regularization

Reduction Ratio

Ratio r	top-1 err.	top-5 err.	Params
2	22.29	6.00	45.7M
4	22.25	6.09	35.7M
8	22.26	5.99	30.7M
16	22.28	6.03	28.1M
32	22.72	6.20	26.9M
original	23.30	6.55	25.6M

Squeeze Operator

Squeeze	top-1 err.	top-5 err.
Max	22.57	6.09
Avg	**22.28**	**6.03**

Excitation Operator

Excitation	top-1 err.	top-5 err.
ReLU	23.47	6.98
Tanh	23.00	6.38
Sigmoid	**22.28**	**6.03**

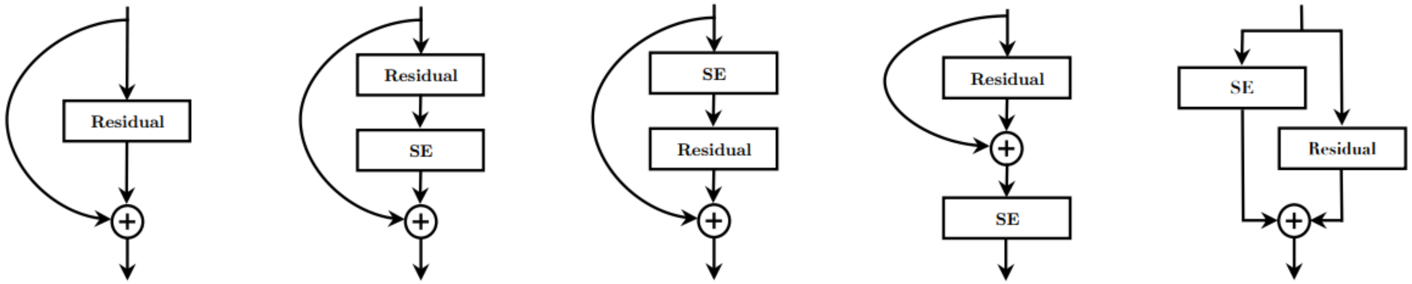
Different Stages

Stage	top-1 err.	top-5 err.	GFLOPs	Params
ResNet-50	23.30	6.55	3.86	25.6M
SE_Stage_2	23.03	6.48	3.86	25.6M

SE_Stage_3	23.04	6.32	3.86	25.7M
SE_Stage_4	22.68	6.22	3.86	26.4M
SE_All	22.28	6.03	3.87	28.1M

Integration strategy

Design	top-1 err.	top-5 err.
SE	22.28	6.03
SE-PRE	22.23	6.00
SE-POST	22.78	6.35
SE-Identity	22.20	6.15



(a) Residual block (b) Standard SE block (c) SE-PRE block (d) SE-POST block (e) SE-Identity block

Design	top-1 err.	top-5 err.	GFLOPs	Params
SE	22.28	6.03	3.87	28.1M
SE_3×3	22.48	6.02	3.86	25.8M

Role of SE Blocks

understand the relative importance of the squeeze operation and how the excitation mechanism operates in practice

empirical approach to examining the role played by the SE block

Effect of Squeeze

	top-1 err.	top-5 err.	GFLOPs	Params
ResNet-50	23.30	6.55	3.86	25.6M
NoSqueeze	22.93	6.39	4.27	28.1M

NoSqueeze	22.93	6.39	4.27	28.1M
SE	**22.28**	**6.03**	3.87	28.1M

Role of Excitation
how excitations vary across images of different classes, and across images within a class

earlier layers - the distribution across different classes is very similar
greater depth - the value of each channel becomes much more class-specific as different classes exhibit different preferences to the discriminative value of features,

last stage of the network - similar pattern emerges over different classes
feature recalibration

SE blocks produce instance-specific responses which nevertheless function to support the increasingly class-specific needs of the model at different layers in the architecture.

Conclusion

m dynamic channel-wise feature recalibration
inability of previous architectures to adequately model channel-wise feature dependencies