**Highlight**

(1) SENet has higher performance compared to conventional CNN.

(2) SE block consists of two parts: squeeze and excitation.

(3) SE blocks can be integrated into modern architectures.

(4) SENet focus on dynamic and non-linear dependencies between channels by using global information.

**what is the problem to be tackled?**

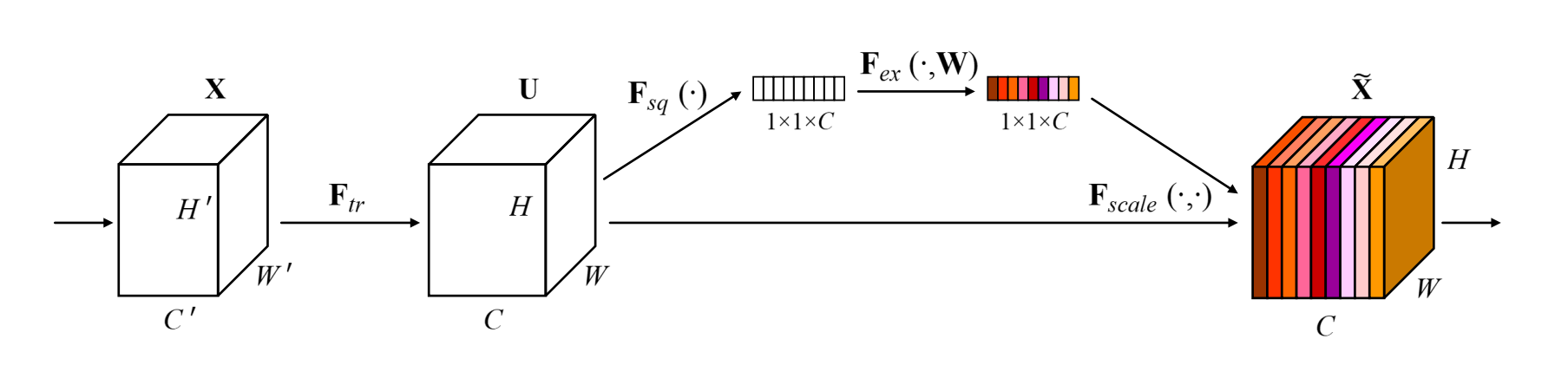
Convolution neural networks are useful models for tackling a wide range of visual tasks. At each convolutional layer in the network, a collection of filters expresses neighborhood spatial connectivity patterns along input channels—fusing spatial and channel-wise information together within local receptive fields. By interleaving a series of convolutional layers with non-linear activation functions and down-sampling operators, CNNs are able to produce robust representations that capture hierarchical patterns and attain global theoretical receptive fields. However, this kind of method is based on the assumption that channel relationships can be formulated as a composition of instance-agnostic functions with local receptive fields. In this paper, it is shown that dynamic and non-linear dependencies between channels using global information can ease the learning process, and significantly enhance the representational power of the network.

**what is the difficulty to solve the problem?**

(1) The first difficulty is to tackle the issue of exploiting channel dependencies and the signal to each channel in the output features should be considered. Each of the learned filters operates with a local receptive field and consequently each unit of the transformation output is unable to exploit contextual information outside of this region.

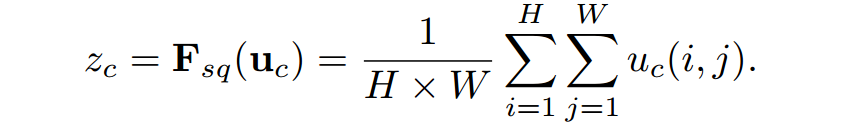
(2) The second one is to make use of the information aggregated in the squeeze operation, it ought to fully capture channel-wise dependencies. It means that the function must be flexible, which means it must be capable of learning a nonlinear interaction between channels. Besides, it must learn a non-mutually-exclusive relationship.

**how is the problem solved in the paper?**

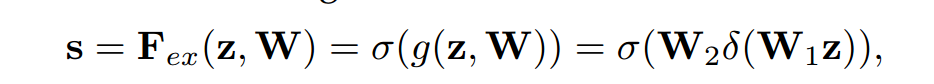


In this paper, Squeeze-and-Excitation block is put forward to solve this problem. It allows the network to perform feature recalibration, through which it can learn to use global information to selectively emphasize informative features and suppress less useful ones. As is shown in the picture above, the computational process is:

(1) The features U are first passed through a squeeze operation, which produces a channel descriptor by aggregating feature maps across their spatial dimensions. This is achieved by using global average pooling to generate channel-wise statistics. Formally, a statistic z 2 RC is generated by shrinking U through its spatial dimensions H × W , such that the c-th element of z is calculated by:



(2) The aggregation is followed by an 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. The computational process is as follows.



where δ refers to the ReLU [64] function, W1 2 R Cr ×C and W2 2 RC× Cr .

(3) These weights are applied to the feature maps U to generate the output of the SE block.

**what is the novelty in the paper, particularly in methodology?**

(1) In methodology, this model focus on the channel relationship rather than fusing spatial and channel-wise information together. It adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels;

(2) In model architecture, it didn't use deeper architectures to better performance. Instead, it provides a unit with gating mechanism to explicitly model dynamic, non-linear dependencies between channels using global information.

(3) In mechanism, it doesn't employ the popular attention mechanism which is proven to be powerful. In contrast, it comprises a lightweight gating mechanism which focuses on enhancing the representational power of the network by modelling channel-wise relationships in a computationally efficient manner.

(4) Experiments show that SE blocks can be integrated into many kinds of modern architecture easily without adding too much burden in the aspect of memory usage and time consumption. Meanwhile, it can achieve better performance compared to other state-of-the art architectures.

**what are the significant experimental results?**

(1) In Image Classification: experiments on the ImageNet 2012 dataset which comprises 1.28 million training images and 50K validation images from 1000 different classes.

SE-ResNet-50 achieves a single-crop top-5 validation error of 6.62%, exceeding ResNet-50 (7.48%) by 0.86% and approaching the performance achieved by ResNet-101 network (6.52% top-5 error). SE-ResNet-101 (6.07% top-5 error) outperforms the deeper ResNet-152 network (6.34% top-5 error) by 0.27%.

(2) In Scene Classification: experiments on the Places365-Challenge dataset which comprises 8 million training images and 36500 validation images across 365 categories.

SE-ResNet-152 (11.01% top-5 error) achieves a lower validation error than ResNet-152 (11.61%top-5 error) It also surpasses the previous state-of-the-art model Places-365-CNN which has a top-5 error of 11.48% on this task.

(3) In Object Detection: experiments on the COCO dataset which comprises 80k training images and 40k validation images.

SE-ResNet-50 outperforms ResNet-50 by 1.3% (a relative5.2% improvement) on COCO’s standard AP metric and by 1.6% on AP@IoU=0.5. ResNet-101 architecture achieving a 0.7%

improvement (or 2.6% relative improvement) on the AP metric.

(4) In ILSVRC 2017 Classification Competition

The winning entry comprised a small ensemble of SENets that employed a standard multi-scale and multi-crop fusion strategy to obtain a top-5 error of 2.251% on the test set.

**what are the advantages and disadvantages of the method/solution proposed in the paper?**

There are some significant advantages about SE blocks:

(1) SE blocks bring significant improvements in performance for existing state-of-the-art CNNs at minimal additional computational cost.

(2) The structure of the SE block is simple and can be used directly in existing state-of-the-art architectures by replacing components with their SE counterparts.

(3) SE blocks are also computationally lightweight and impose only a slight increase in model complexity and computational burden.

(4) SE blocks can be used as atomic building blocks for these search algorithms, and were demonstrated to be highly effective in this capacity in concurrent work.

And there is an inevitable disadvantage:

It introduces more parameters costing more time and memory. Take it as an example. SE-ResNet-50 introduces ∼2.5 million additional parameters beyond the ∼25 million parameters required by ResNet-50, corresponding to a ∼10% increase. Therefore, a single pass forwards and backwards through SE-ResNet-50 takes longer time than ResNet-50. When it is performed on GPUs, takes ResNet-50 takes 190 ms, compared to 209 ms for SE-ResNet-50 with a training mini-batch of 256 images; And ResNet-50 takes 164 ms in comparison to 167 ms for SE-ResNet-50 while on CPU. And SE-ResNet-50 requires ∼3.87 GFLOPs, corresponding to a 0.26% relative increase over the original ResNet-50.

**how can the proposed method be further developed?**

(1) This method can be used in more state-of-the-art model to better their performance in various fields which requiring strong discriminative features.

(2) The feature importance values produced by SE blocks may be of use for other tasks such as network pruning for model compression.

(3) Other design of SE blocks like changing the number of FC layers or changing the reduction ratio may be tested on various architecture to find the best design.

(4) This kind of methodology can be used to find some more effective models which focus on the relationship between different channels.