

Squeeze-and-Excitation Networks

The ETRI logo is displayed in a stylized font, with the letters 'E', 'T', and 'R' in blue and 'I' in orange. It is centered within a white rectangular box that is placed over the background image.

UST-ETRI

석사2학기 김형민

21.05.14

Squeeze-and-Excitation Networks

Jie Hu^{1*}

hujie@momenta.ai

¹ Momenta



현재 bytedance AI 소속

Li Shen^{2*}

lishen@robots.ox.ac.uk

² Department of Engineering Science, University of Oxford

ETRI



현재 텐센트 AI 소속

Gang Sun¹

sungang@momenta.ai



Momenta 소속

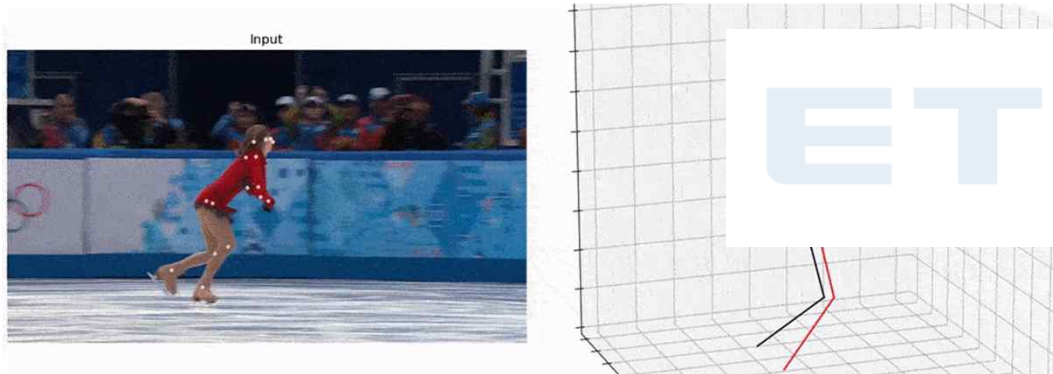
기관 : Momenta, Oxford 대학교
인용 : 6436회



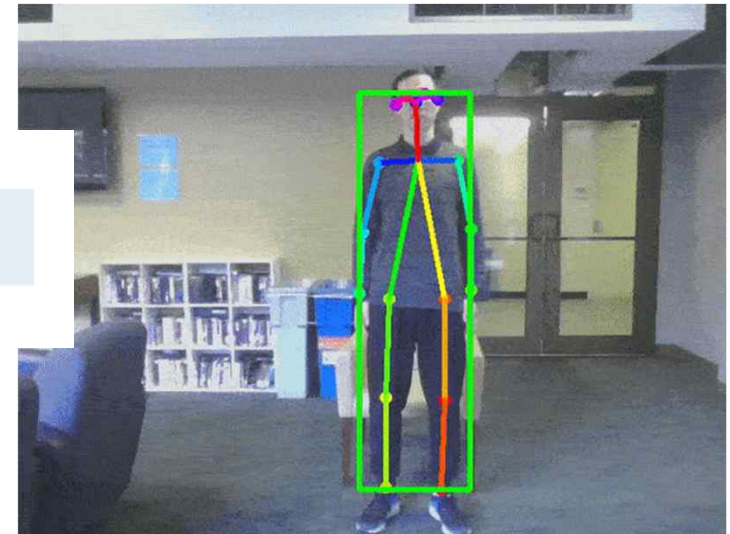
ETRI Motivation

CNN

최근 CNN 구조가 여러 Vision task를 효과적으로 풀게 되었다.



Pose estimation



Action recognition and detection

From : <https://blog.usejournal.com/3d-human-pose-estimation-ce1259979306>

대부분 응용에서 Visual feature extraction을 위해

Image Net에서 잘 학습된 feature로 사용 했다.



이미지넷 성능을 끌어 올리는 것이 주요 이슈.

때문에 Image Classification Task는 Computer Vision 분야의 core task이다.

깊게 해보자.

Figure 2. Residual learning: a building block.



Gradient vanishing 문제 해결위해
Skip-Connection 도입.
덧셈 항으로 gradients가 0이 되는 문제 해결.
동시에 Bottleneck 구조와 3x3 conv 적용

Residual Connection 등장.

[10] K. He et al. "Deep Residual Learning for Image Recognition", CVPR, 2015

Deeper makes better

method			error (%)
Maxout [10]			9.38
NIN [25]			8.81
DSN [24]			8.22
	# layers	# params	
FitNet			
Highway			7.72±0.16
Highway			
ResNet			
ResNet	32	0.40M	7.51
ResNet	44	0.66M	7.17
ResNet	56	0.85M	6.97
ResNet	110	1.7M	6.43 (6.61±0.16)
ResNet	1202	19.4M	7.93

Table 6. Classification error on the **CIFAR-10** test set. All methods are with data augmentation. For ResNet-110, we run it 5 times and show “best (mean±std)” as in [43].

More layers consume more cost

method			error (%)
Maxout [10]			9.38
NIN [25]			8.81
DSN [24]			8.22
	# layers	# params	
FitNet [35]	19	2.5M	8.39
Highway [42,			.16)
Highway [42,			—
ResNet			
ResNet			
ResNet			
ResNet	56	0.85M	6.97
ResNet	110	1.7M	6.43 (6.61±0.16)
ResNet	1202	19.4M	7.93

Table 6. Classification error on the **CIFAR-10** test set. All methods are with data augmentation. For ResNet-110, we run it 5 times and show “best (mean±std)” as in [43].

또한, 많은 데이터를 필요로 한다.
하지만 질 좋은 데이터는 제작이 힘들다.

층을 많이 쌓는 것은 한계다.

동일한 수준에.  이 이루어져야.

또한 Domain specific한 supervision 없는
self-learning mechanism이어야 한다.

Capture the more spatial correlation

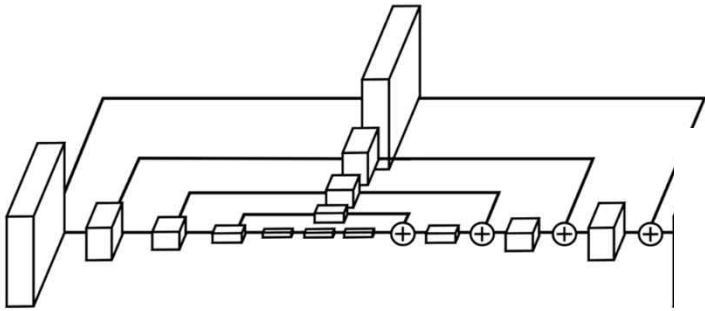
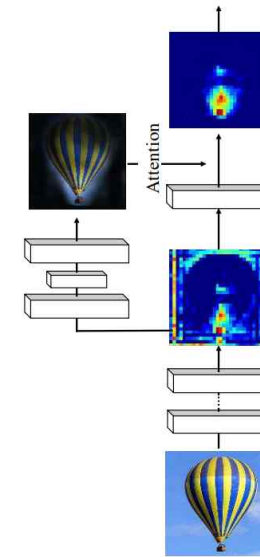


Fig. 3. An illustration of a single “hourglass” module. Each box in the figure corresponds to a residual module as seen in Figure 4. The number of features is consistent across the whole hourglass.

[31] stacked hourglass
multi-scale manner

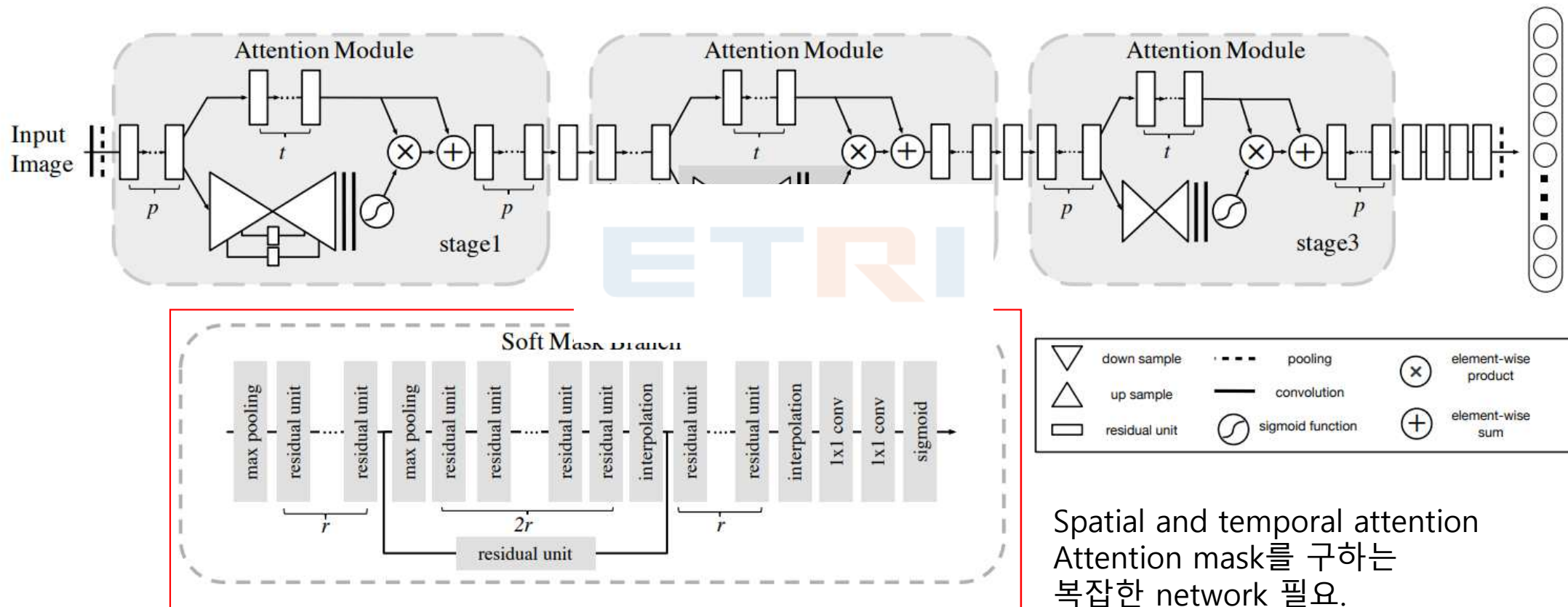


Attention mechanism

Residual attention (2017.04)

[31] A. Newell, et al. “Stacked hourglass networks for human pose estimation, ECCV, 2016
F. Wang et al. “Residual Attention Network for Image Classification”, CVPR, 2017

Residual Attention Network (17.04)

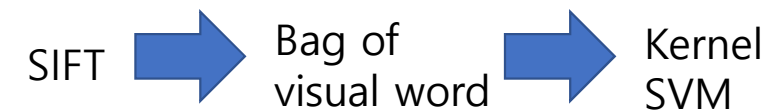
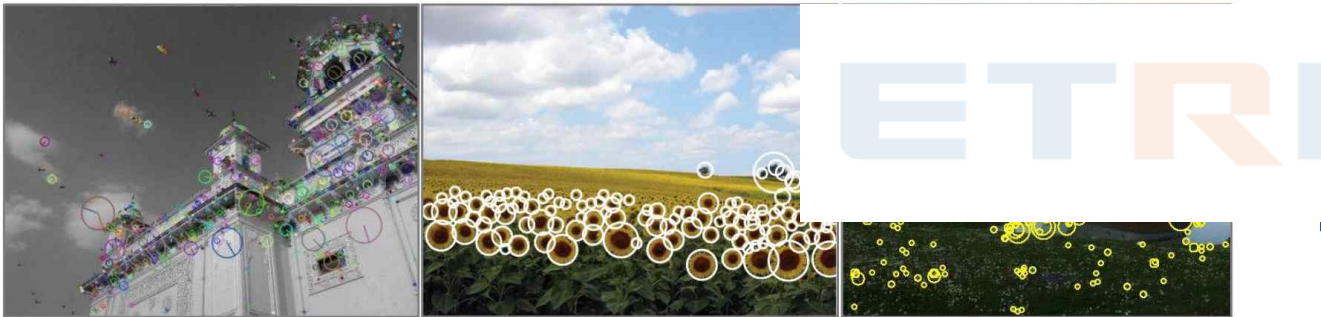


ETRI :-?

Filter in Computer Vision

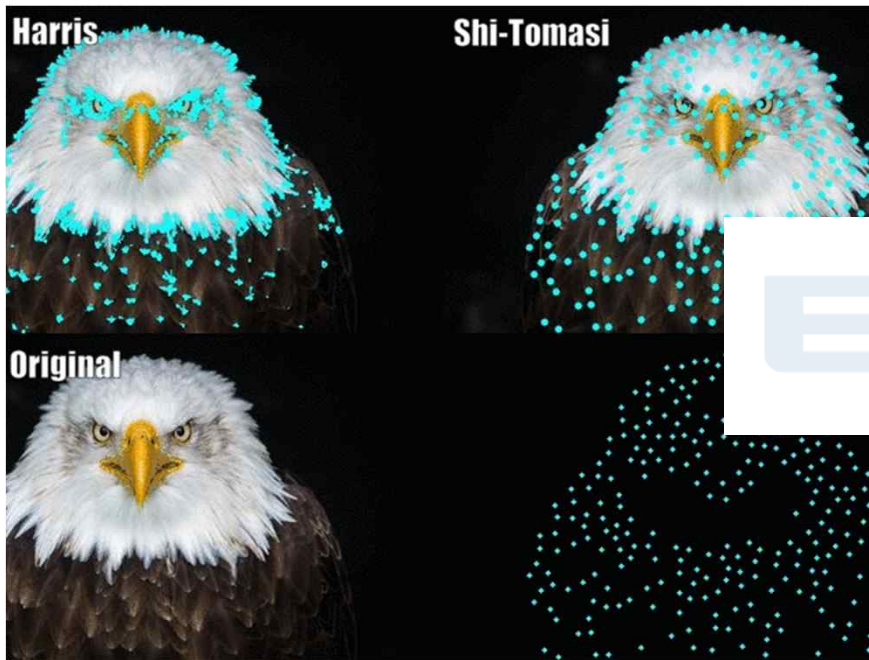
Local image feature

- An interesting part of an image

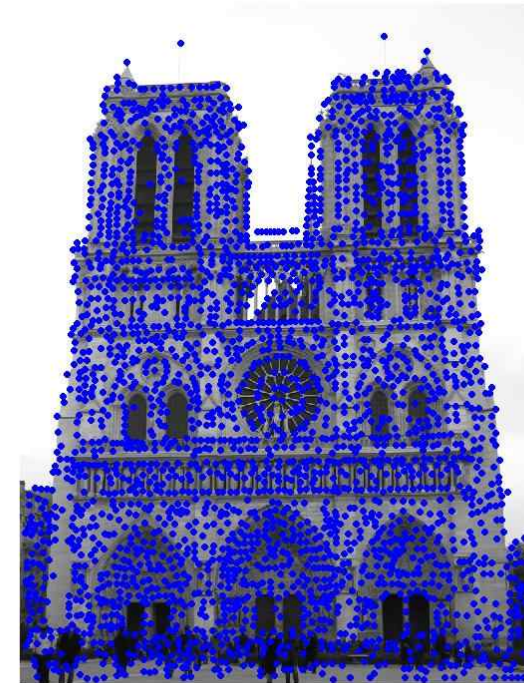


CNN 이전에는 unique하게 검출될 수 있는 부분에 대해서 feature extraction을 진행.

Interest point detection using filters



Corner detection kernel



Blob detection kernel

From : <https://medium.com/pixel-wise/detect-those-corners-aba0f034078b>

Convolutions can learn the filters

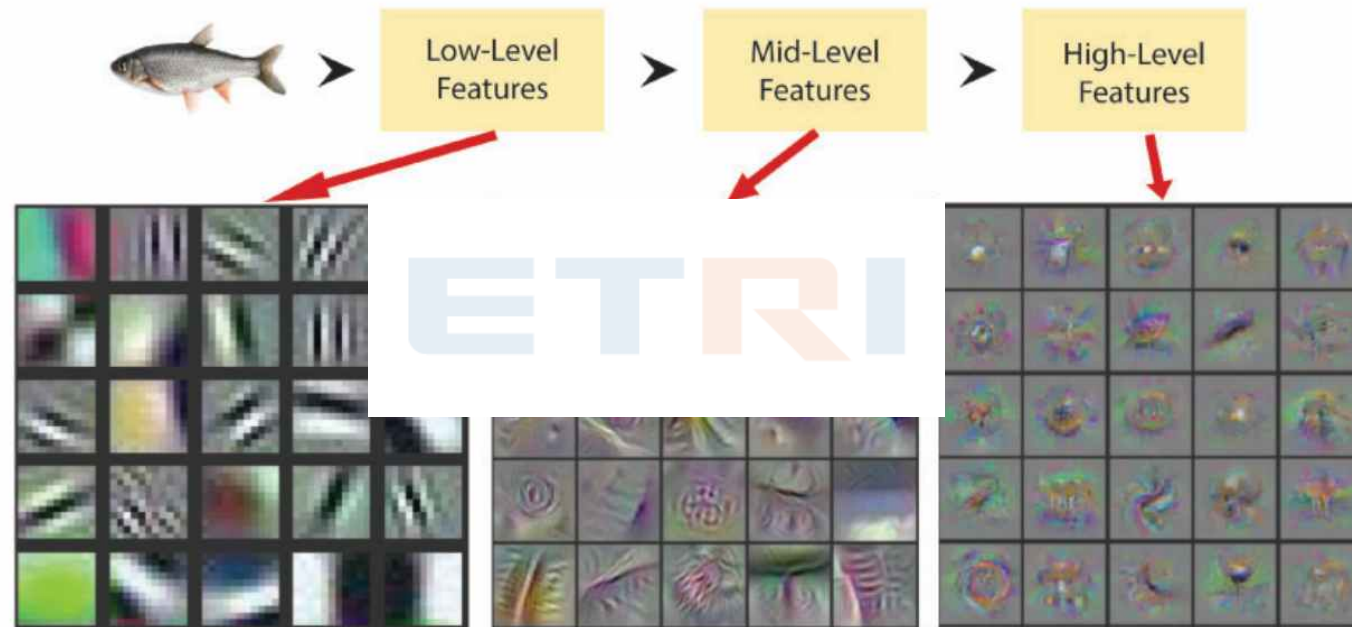


Figure 4. Hierarchical representation learning by a CNN where the initial layer detects simple patterns like edges and gradients while higher layers detect more abstract features (Yosinski *et al.*, 2015).

Block 1 in VGG



From : <https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/>

Block 2 in VGG



From : <https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/>

Block 3 in VGG



From : <https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/>

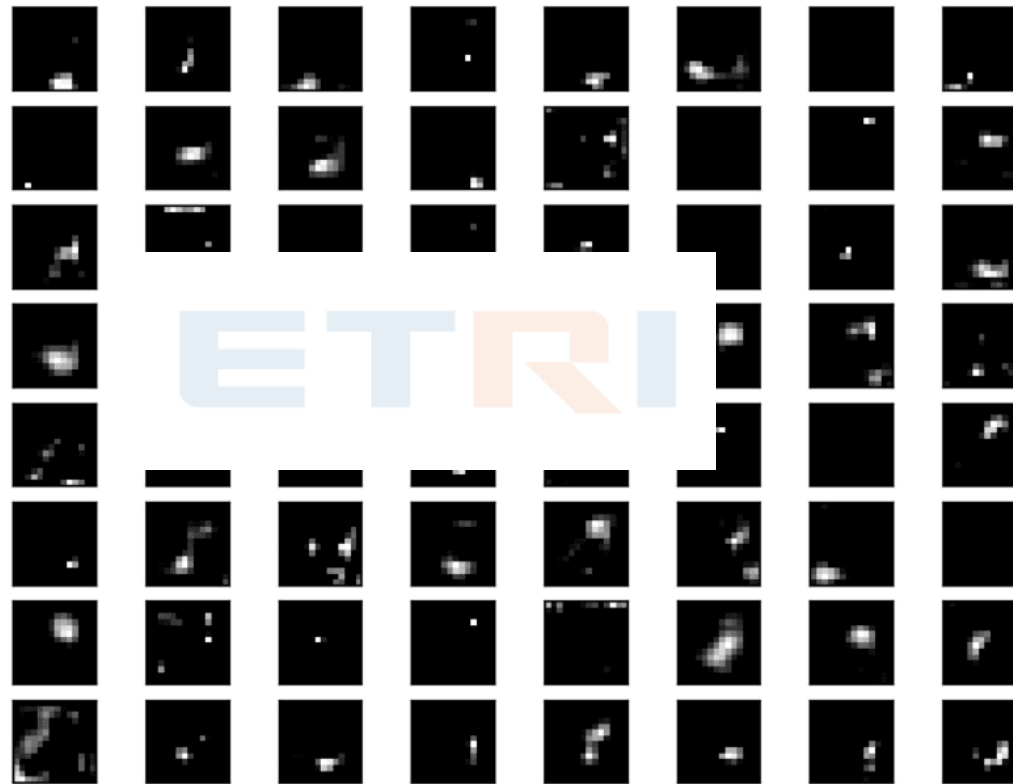


Block 4 in VGG



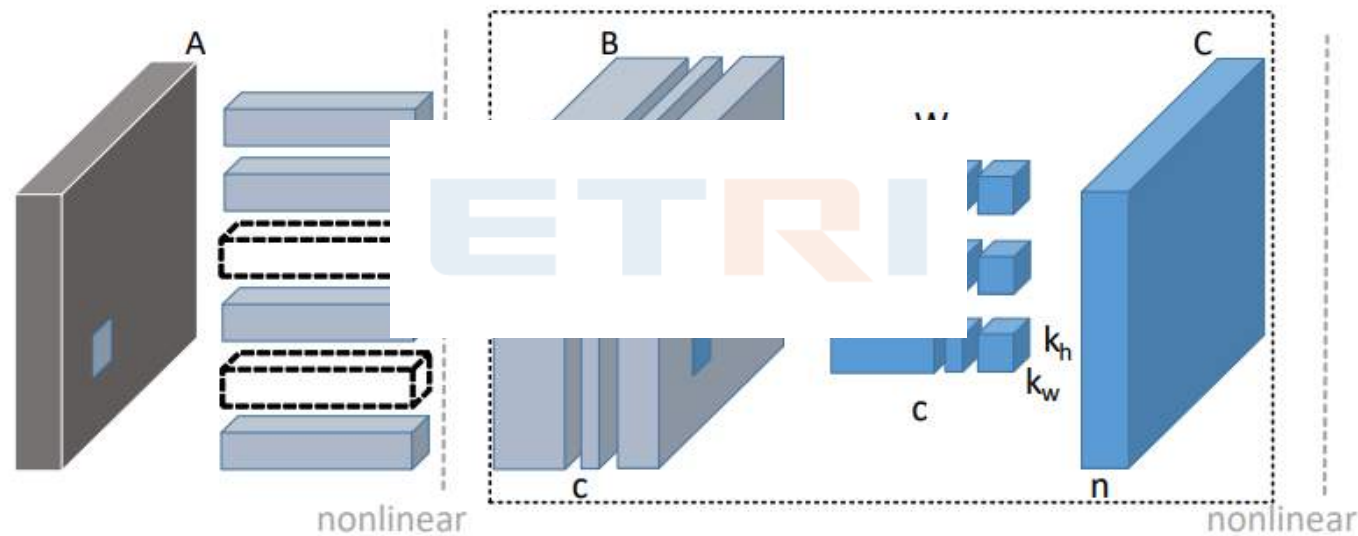
From : <https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/>

Block 1 in VGG



From : <https://machinelearningmastery.com/how-to-visualize-filters-and-feature-maps-in-convolutional-neural-networks/>

Channel pruning



“사실 일부 커널만이 의사결정에 중요한 영향을 미친다.”

동적으로 ETRI 을 준다면?



Squeeze and Excitation Networks

Squeeze and Excitation Network (17.09)

기존의 대부분의 접근법들은 Spatial Attention 기법.
채널간 correlation만 고려하면 어떨까?

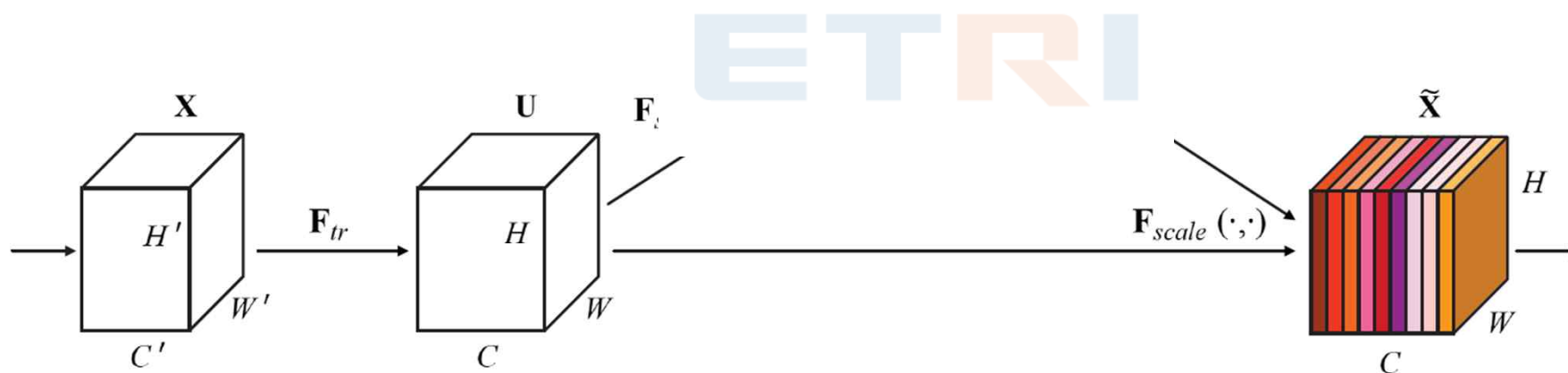


Figure 1: A Squeeze-and-Excitation block.

The Squeeze Operation

Global Average Pooling을 통한, channel-wise feature vector 생성

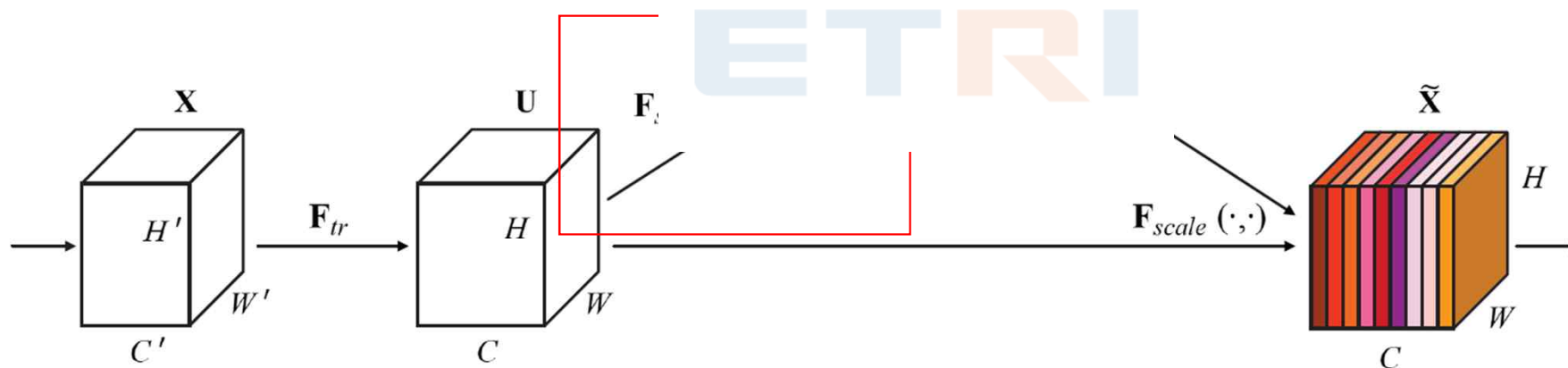


Figure 1: A Squeeze-and-Excitation block.

Global average pooling

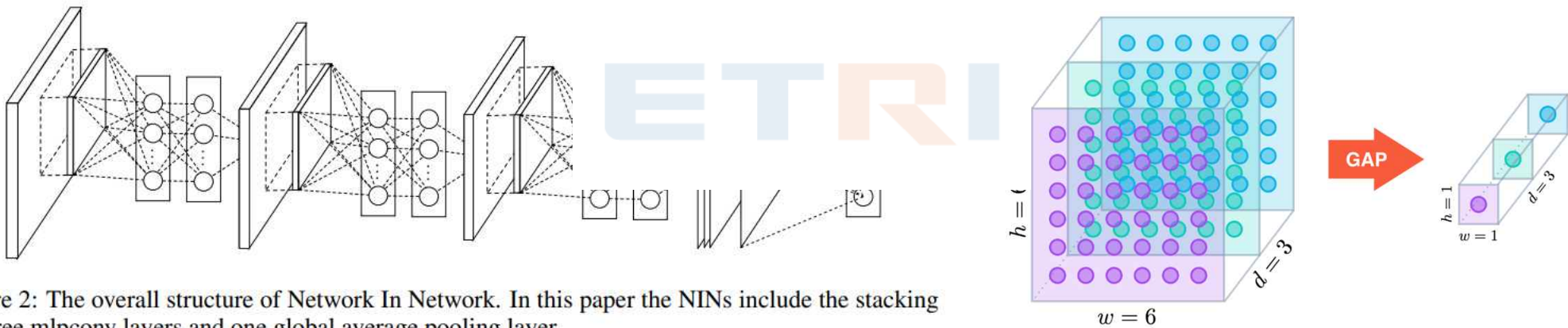


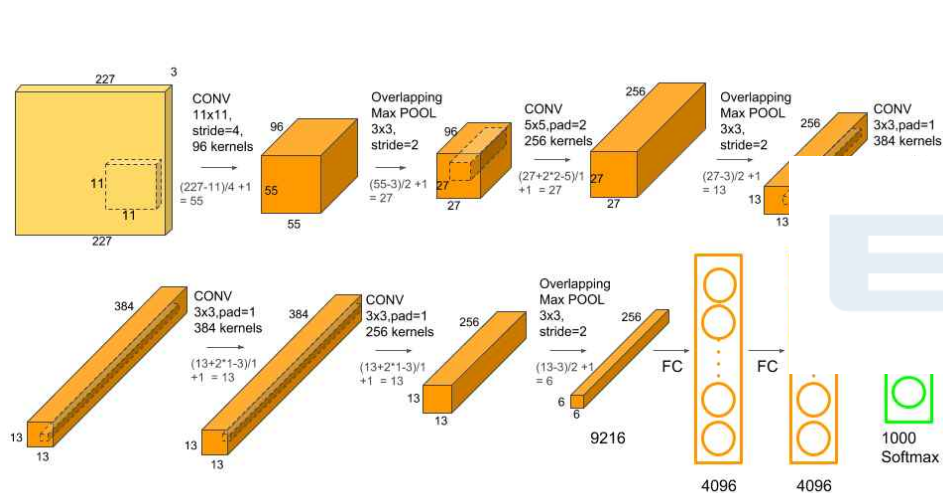
Figure 2: The overall structure of Network In Network. In this paper the NINs include the stacking of three mlpconv layers and one global average pooling layer.

Global average pooling

However, the fully connected layers are prone to overfitting, thus hampering the generalization ability of the overall network. Dropout is proposed by Hinton et al. [5] as a regularizer which randomly sets half of the activations to the fully connected layers to zero during training. It has improved the generalization ability and largely prevents overfitting [4].

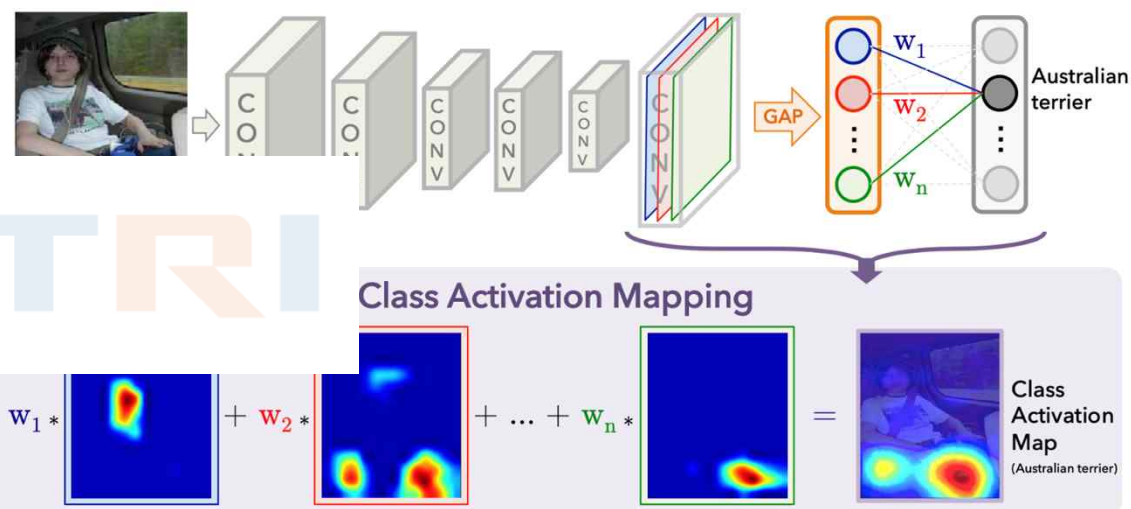
In this paper, we propose another strategy to replace the fully connected layers in CNN. The feature map for each corresponding category of the classification task in the convolutional layer is fed directly into the softmax layer. One advantage of global average pooling over the fully connected layers is that it is more native to the convolution structure by enforcing correspondences between feature maps and categories. Thus the feature maps can be easily interpreted as categories confidence maps. Another advantage is that there is no parameter to optimize in the global average pooling thus overfitting is avoided at this layer. Furthermore, global average pooling sums out the spatial information, thus it is more robust to spatial translations of the input.

Global average pooling



[AlexNet]

Flatten -> conv 채널과 예측 class와의
관계 해석 불가능



[CAM]

GAP 적용시 어느 피쳐맵이
의사결정에 얼마나 영향을 미치는지 유추 가능.

The Squeeze Operation

Global Average Pooling을 통한, channel-wise feature vector 생성

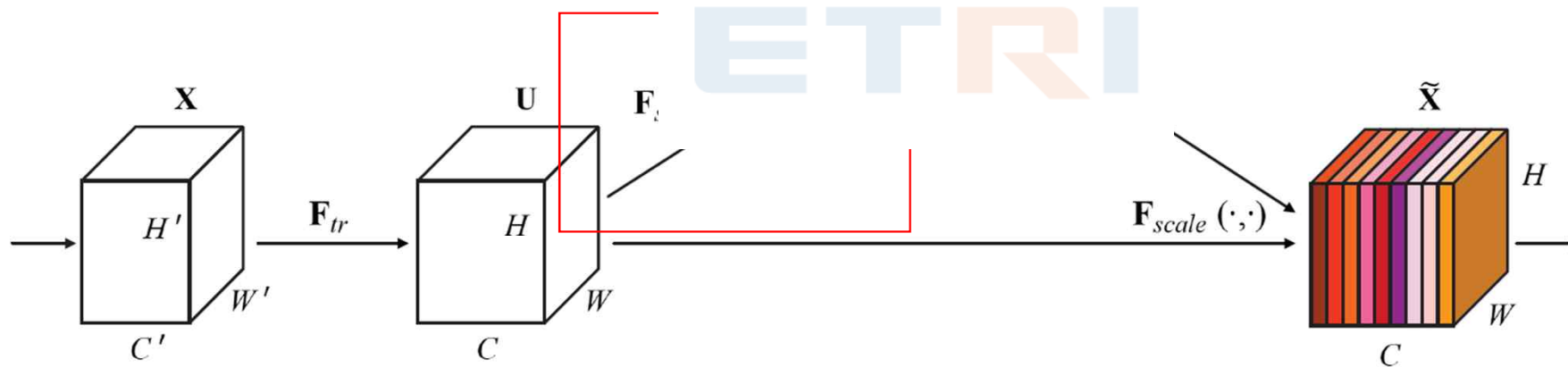


Figure 1: A Squeeze-and-Excitation block.

The Squeeze Operation

Channel-wise importance vector 생성.

Weights를 곱하여 중요도 모델링
Activation으로 sigmoid 사용해 0~1 사이 중요도 값 출력

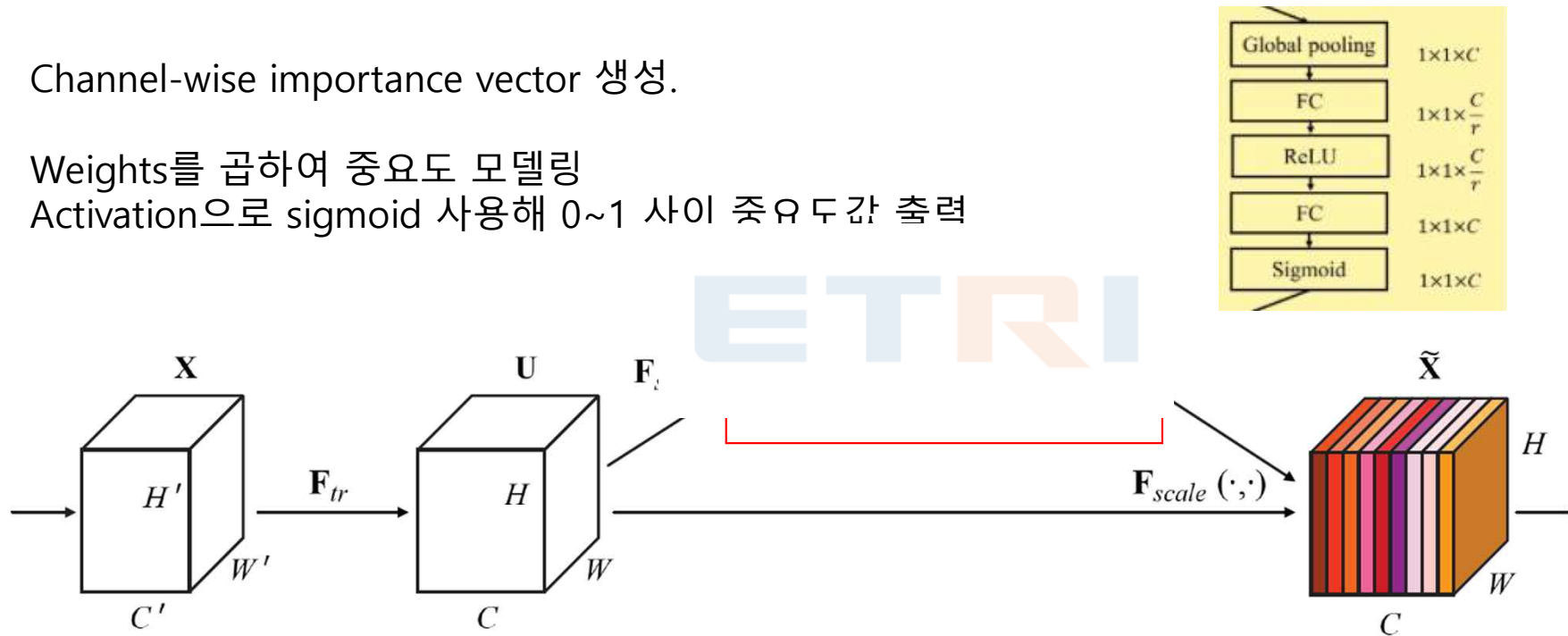
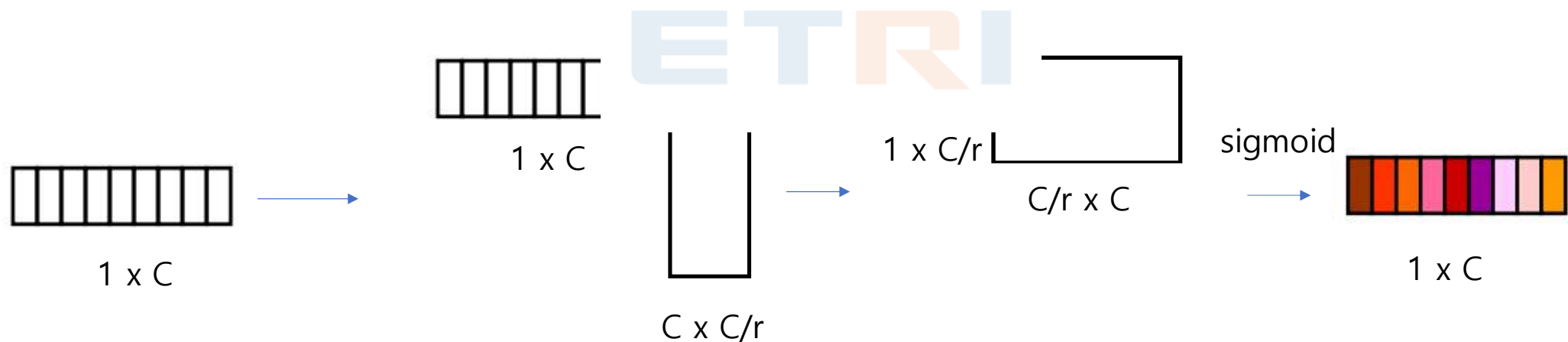
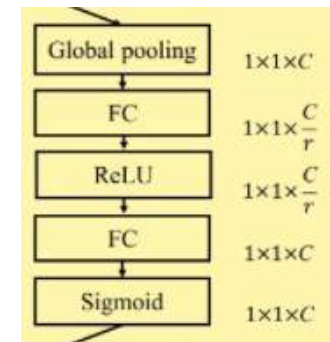


Figure 1: A Squeeze-and-Excitation block.

The Squeeze Operation

Bottleneck in squeeze operation

Efficiency를 위해 중간 weights matrix의 크기를 $1/r$ 로 감소



The Excitation Operation

Channel-wise Attention

앞서 모델링된 채널 중요도와 입력 피쳐맵 간 Hadamard product
입력 피쳐맵이 recalibration 되었다.

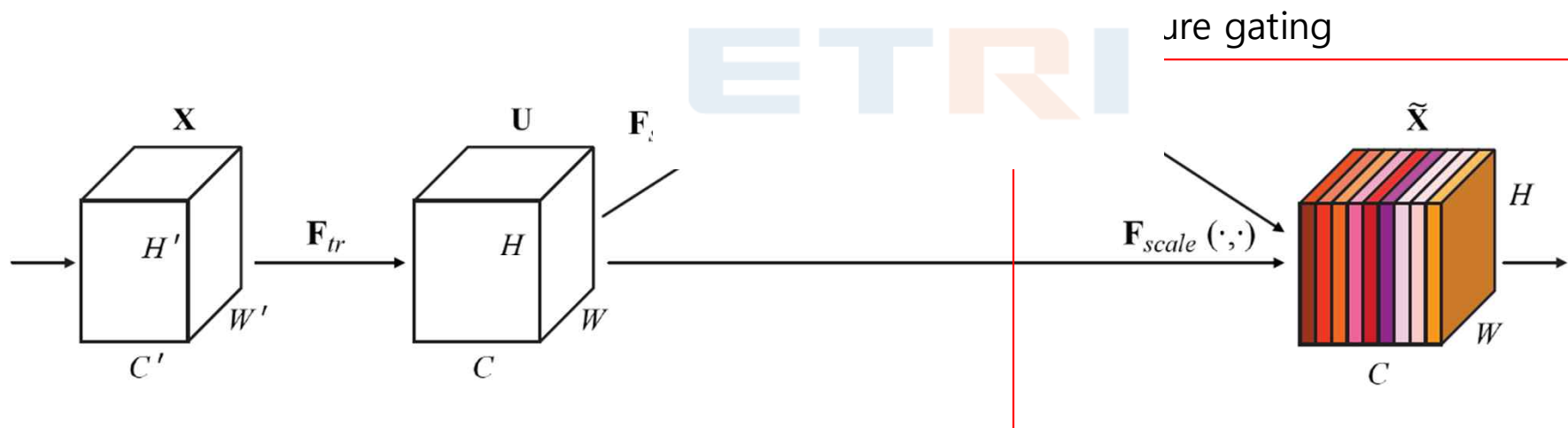



Figure 1: A Squeeze-and-Excitation block.

Contributions

1. 새로운 CNN backbone network를 설계 하는 것은 어렵다. (hyper parameters searching 등)
 - > SE Block은 간단한 구조로 설계되어 **구현이 쉽고** 현재 SOTA 모델에 **바로 적용 가능**하다.
2. SE Block은 적은 cost로 동작한다.
 - > 추가적인 parameter 연산이 원본 **때문에 efficient** 하다.
3. 성능이 좋다.
 - > 간단하여 이해하기 쉽고, 구현도 쉬우며, 적용성이 넓고, **높은 성능 향상효과**가 있다.
 - > ILSVRC 2017 classification competition에서 top5 error 2.251%를 기록하며 당해 1위 등극.

The ETRI logo consists of the letters 'ETRI' in a stylized font. The 'E' and 'I' are light blue, while the 'T' and 'R' are light orange. The 'R' has a unique shape with a diagonal cut. The logo is centered horizontally and is flanked by two vertical orange lines, one above and one below it.

ETRI

Experiments

Experiment : 다른 모델에 적용

	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 [10]	24.7	7.8	24.80	7.48	3.86	23.29 _(1.51)	6.62 _(0.86)	3.87
ResNet-101 [10]	23.6	7.1	23.17	6.52	7.58	22.38 _(0.79)	6.07 _(0.45)	7.60
ResNet-152 [10]	23.0	6.7	22.42	6.34	11.30	21.57 _(0.85)	5.73 _(0.61)	11.32
ResNeXt-50 [47]	22.2	-	22.11	5.90	4.94	21.10 _(1.01)	5.49 _(0.41)	4.25
ResNeXt-101 [47]	21.2	5				20.70 _(0.48)	5.01 _(0.56)	8.00
VGG-16 [39]	-					25.22 _(1.80)	7.70 _(1.11)	15.48
BN-Inception [16]	25.2	7.				24.23 _(1.15)	7.14 _(0.75)	2.04
Inception-ResNet-v2 [42]	19.9 [†]	4.				19.80 _(0.57)	4.79 _(0.42)	11.76

Table 2: Single-crop error rates (%) on the ImageNet validation set and complexity comparisons. The *original* column refers to the results reported in the original papers. To enable a fair comparison, we re-train the baseline models and report the scores in the *re-implementation* column. The *SENet* column refers to the corresponding architectures in which SE blocks have been added. The numbers in brackets denote the performance improvement over the re-implemented baselines. † indicates that the model has been evaluated on the non-blacklisted subset of the validation set (this is discussed in more detail in [42]), which may slightly improve results. VGG-16 and SE-VGG-16 are trained with batch normalization.

SE block 적용시 모든 modern architecture에서 error 감소 효과 보임
또한 ResNet구조로 살펴볼 때 모든 depth에서 동일하게 성능 향상 효과를 보임

Experiment : 다른 모델에 적용

	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 [10]	24.7	7.8	24.80	7.48	3.86	23.29 _(1.51)	6.62 _(0.86)	3.87
ResNet-101 [10]	23.6	7.1	23.17	6.52	7.58	22.38 _(0.79)	6.07 _(0.45)	7.60
ResNet-152 [10]	23.0	6.7	22.42	6.34	11.30	21.57 _(0.85)	5.73 _(0.61)	11.32
ResNeXt-50 [47]	22.2	-	22.11	5.90	4.24	21.10 _(1.01)	5.49 _(0.41)	4.25
ResNeXt-101 [47]	21.2	5				20.70 _(0.48)	5.01 _(0.56)	8.00
VGG-16 [39]	-					25.22 _(1.80)	7.70 _(1.11)	15.48
BN-Inception [16]	25.2	7.				24.23 _(1.15)	7.14 _(0.75)	2.04
Inception-ResNet-v2 [42]	19.9 [†]	4.				19.80 _(0.57)	4.79 _(0.42)	11.76

Table 2: Single-crop error rates (%) on the ImageNet validation set and complexity comparisons. The *original* column refers to the results reported in the original papers. To enable a fair comparison, we re-train the baseline models and report the scores in the *re-implementation* column. The *SENet* column refers to the corresponding architectures in which SE blocks have been added. The numbers in brackets denote the performance improvement over the re-implemented baselines. † indicates that the model has been evaluated on the non-blacklisted subset of the validation set (this is discussed in more detail in [42]), which may slightly improve results. VGG-16 and SE-VGG-16 are trained with batch normalization.

거의 없다시피한 GFLOPs 상승만으로 엄청난 성능 향상 달성.

Experiment : 다른 모델에 적용

	original		re-implementation		SENet			
	top-1 err.	top-5 err.	top-1 err.	top-5 err.	top-1 err.	top-5 err.	MFLOPs	Million Parameters
MobileNet [13]	29.4	-	29.4	-	25.3 _(3.8)	7.9 _(2.2)	572	4.7
ShuffleNet [52]	34.1	-	33.3	-	31.7 _(2.2)	11.7 _(1.9)	142	2.4

Table 3: Single-crop error rates (%) on the ImageNet validation set and complexity comparisons. Here, MobileNet refers to “1.0 MobileNet-224” in [13] and ShuffleNet refers to “ShuffleNet $1 \times (g = 3)$ ” in [52].

경량 구조에서도 상당한 error 감소 효과를 보임.

ImageNet2017 Competition

	224 × 224		320 × 320 / 299 × 299	
	top-1 err.	top-5 err.	top-1 err.	top-5 err.
ResNet-152 [10]	23.0	6.7	21.3	5.5
ResNet-200 [11]	21.7	5.8	20.1	4.8
Inception-v3 [44]	-	-	21.2	5.6
Inception-v4 [42]	-	-	20.0	5.0
Inception-ResNet-v2 [42]	-	-	19.9	4.9
ResNeXt-101 (64 × 4d) [47]	20.4	5.3	19.1	4.4
DenseNet-264 [14]	22.15	6.12	-	-
Attention-92 [46]	-	-	19.5	4.8
Very Deep PolyNet [51] †	-	-	18.71	4.25
PyramidNet-200 [8]	20.1	5.4	19.2	4.7
DPN-131 [5]	19.93	5.12	18.55	4.16
SENet-154	18.68	4.47	17.28	3.79
NASNet-A (6@4032) [55] †	-	-	17.3 [‡]	3.8 [‡]
SENet-154 (post-challenge)	-	-	16.88[‡]	3.58[‡]

Multi-scale/multi-crop fusion strategy(ensemble)

On Validation set

ETRI

challenge에서 우승.

Table 4: Single-crop error rates of state-of-the-art CNNs on ImageNet validation set. The size of test crop is 224 × 224 and 320 × 320 / 299 × 299 as in [11]. † denotes the model with a larger crop 331 × 331. ‡ denotes the post-challenge result. SENet-154 (post-challenge) is trained with a larger input size 320 × 320 compared to the original one with the input size 224 × 224.

Experiment : 다양한 task에서도 잘 동작



ETRI



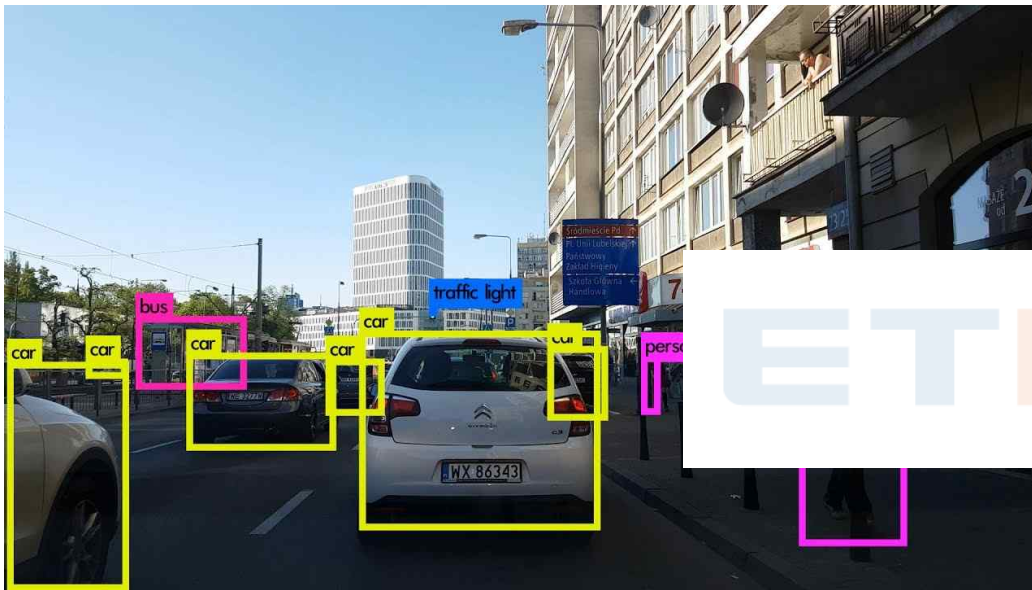
Scene Recognition

SE block 적용하는 것으로 SOTA 달성.

	top-1 err.	top-5 err.
Places-365-CNN [37]	41.07	11.48
ResNet-152 (ours)	41.15	11.61
SE-ResNet-152	40.37	11.01

Table 5: Single-crop error rates (%) on Places365 validation set.

Experiment : 다양한 task에서도 잘 동작



Object Detection (COCO 2014)

SE block 을 trunk network에 적용 하면 Object detection task 에서도 성능향상됨을 보임.

	AP@IoU=0.5	AP
ResNet-50	45.2	25.1
esNet-50	46.8	26.4
Net-101	48.4	27.2
esNet-101	49.2	27.9

Table 1. Object detection results on the COCO 40k validation set by using the basic Faster R-CNN.

<https://www.youtube.com/watch?v=EhcpGpFHCrw>

Ablation Study – bottleneck reduction ratio

Ratio r	top-1 err.	top-5 err.	Million Parameters
4	23.21	6.63	35.7
8	23.19	6.64	30.7
16	23.29	6.62	28.1
32			.9
original			.6

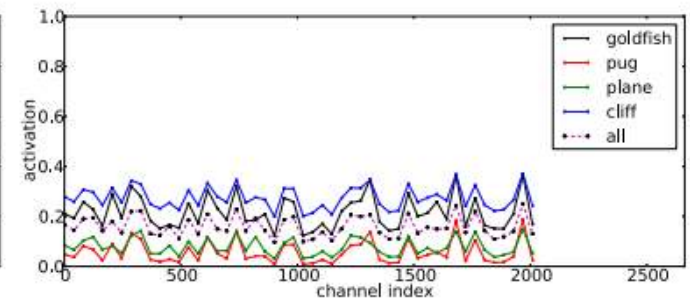
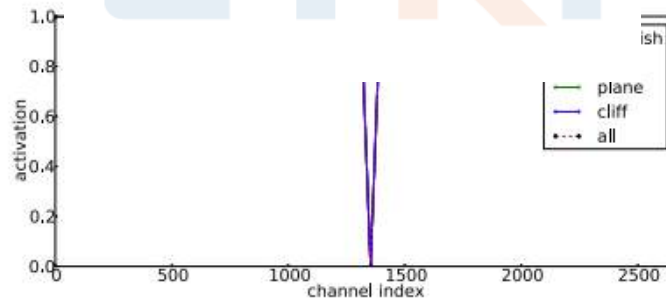
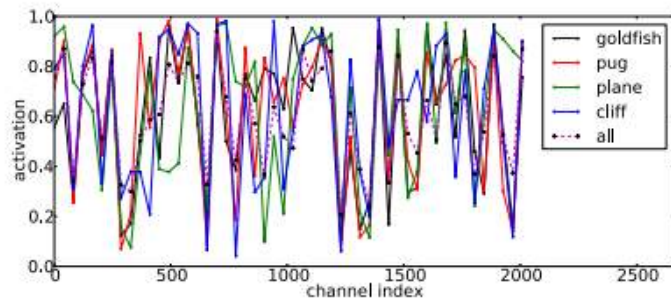
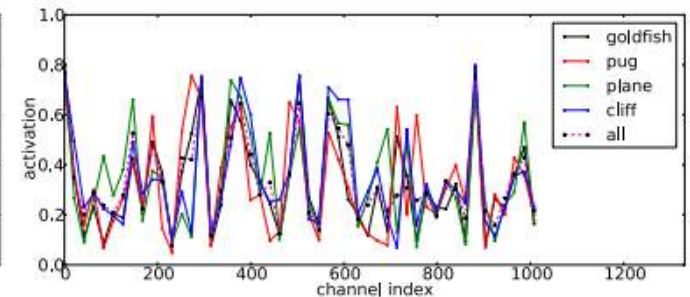
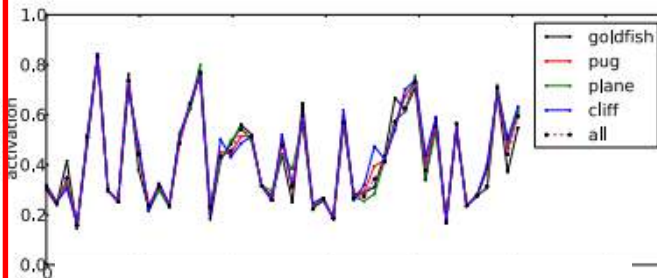
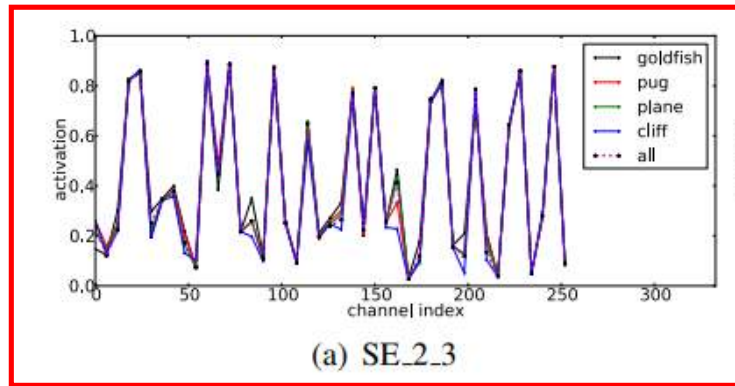
Table 7: Single-class validation set and parameter sizes for SE-ResNet-50 at different reduction ratios r . Here *original* refers to ResNet-50.

많이 reduction할 경우 파라미터는 감소하나 에러율 증가.
trade-off를 찾아 사용하면 된다.

논문에서는 complexity와 performance간 trade-off로 16선택

초기단계 : distribution이 거의 동일 (아직까지는 매우 저수준의 feature extraction)

Analysis : self-gating이 어떻게 동작?



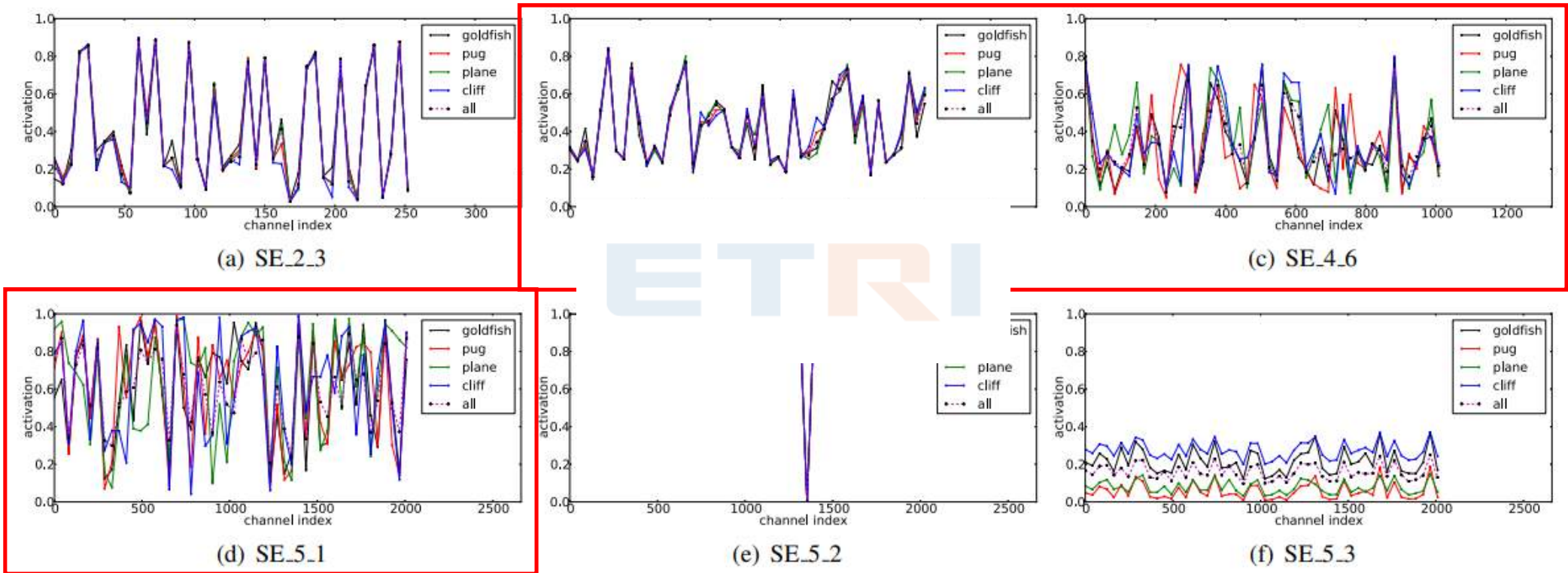
(a) goldfish (b) pug (c) plane (d) cliff

activation in the different modules of SE-ResNet-50 on ImageNet. The module is named as

Fig. 8. Sample images from the four classes of ImageNet used in the experiments described in Sec. 7.2.

중간단계 : class wise distribution의 차이가 점점 심해짐 (클래스별로 최적화된 channel selection 과정)

Analysis : self-gating이 어떻게 동작?



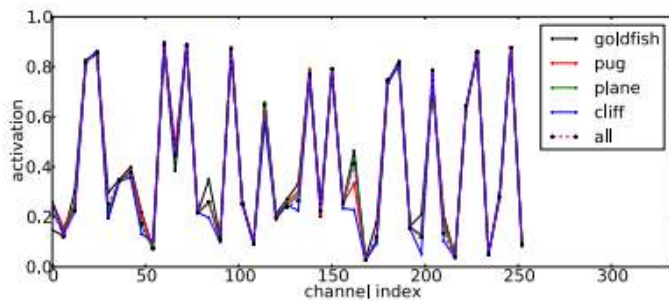
(a) goldfish (b) pug (c) plane (d) cliff

Fig. 8. Sample images from the four classes of ImageNet used in the experiments described in Sec. 7.2.

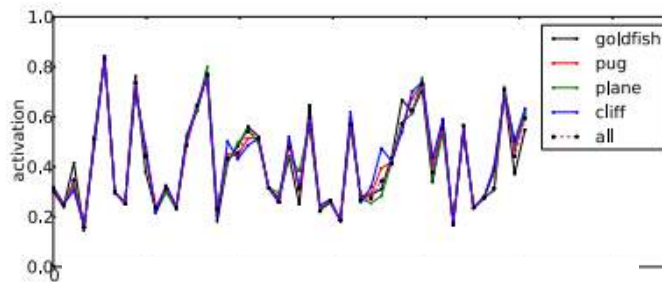
activation in the different modules of SE-ResNet-50 on ImageNet. The module is named as

SE5_2 : 거의 모든 activation이 1로 수렴 -> 해당 단계에서는 gating 안하는 것이 최적화된 모델

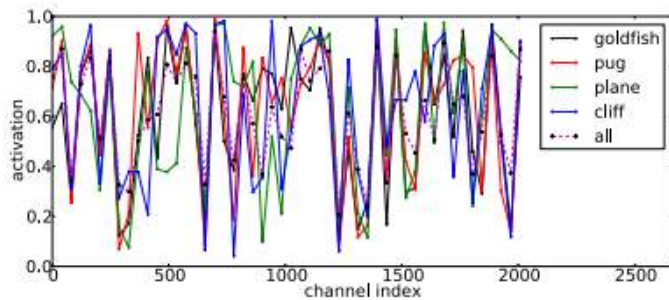
Analysis : self-gating이 어떻게 동작?



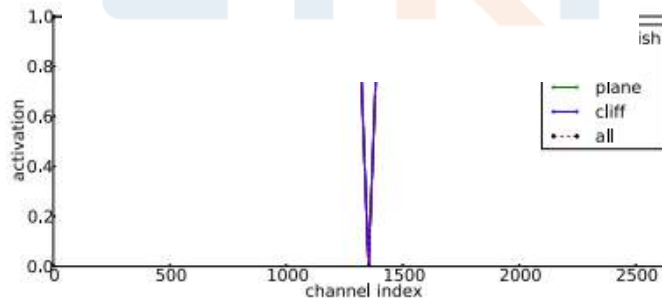
(a) SE_2_3



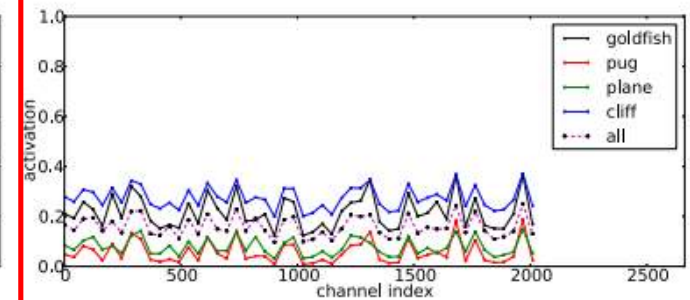
(c) SE_4_6



(d) SE_5_1



(e) SE_5_2



(f) SE_5_3



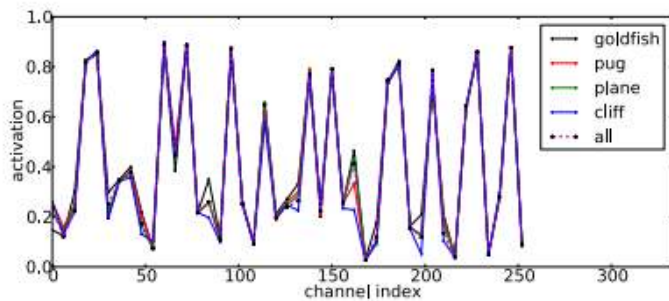
(a) goldfish (b) pug (c) plane (d) cliff

activation in the different modules of SE-ResNet-50 on ImageNet. The module is named as

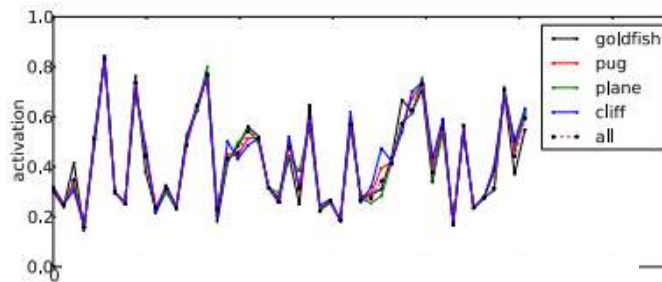
Fig. 8. Sample images from the four classes of ImageNet used in the experiments described in Sec. 7.2.

SE5_3 : 마지막 block (scale만 다르고 모두 동일한 distribution) -> 역시 gating 안 해도 무방

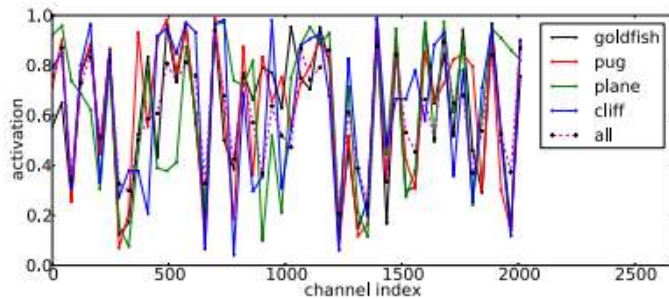
Analysis : self-gating이 어떻게 동작?



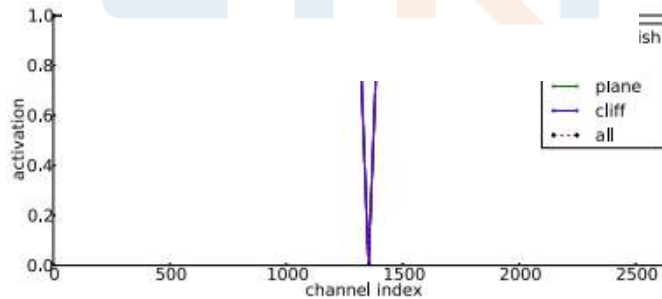
(a) SE_2_3



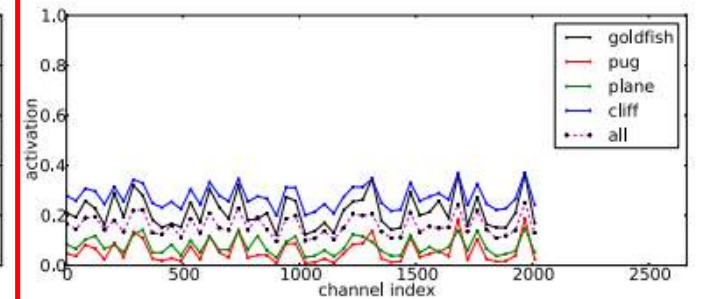
(c) SE_4_6



(d) SE_5_1



(e) SE_5_2



(f) SE_5_3

ETRI




(a) goldfish (b) pug (c) plane (d) cliff

activation in the different modules of SE-ResNet-50 on ImageNet. The module is named as

Fig. 8. Sample images from the four classes of ImageNet used in the experiments described in Sec. 7.2.

Conclusion

- Automatic Neural Architecture Search를 사람의 인사이트로 능가할 수 있다.
- 그동안 다루어지지 않았던, 피쳐  | 대해서 알 수 있었다.
- ImageNet 챌린지에서 1위 등극.
- 다양한 테스트에 적용 가능할 것으로 기대됨.

ETRI.