심층 신경망의 지역 기술자를 이용한 피셔 커널 기반 영상 표현

유동근*, 박승균**, 이준영*, 권인소*
*KAIST 전기 및 전자 공학과
** KAIST 산업 및 시스템 공학과
e-mail : dgyoo@rcv.kist.re.kr

Fisher Kernel for Deep Neural Activations

Donggeun Yoo*, Sunggyun Park**, Joon-Young Lee*, In So Kweon*
*Dept of Electrical Engineering, KAIST
**Dept of Industrial & Systems Engineering, KAIST

요 약

In this paper, we briefly review the methodology of the suitable combination of Fisher Kernel and deep neural activations, proposed by Yoo *et al*. [1]. To integrate both representations, a fair amount of neural activations are extracted and aggregated by Fisher kernel framework, which has been modified with a simple scale-wise normalization essential to make it suit-able for CNN activations. The representation demonstrates new state-of-the-art performances on two public datasets.

1. 서론

Image representation is one of the most important factors for visual recognition tasks. With a success of local descriptors [2], many researches have devoted in studying Fisher kernel [3] that aggregates local statistics to globally represent an image. A major benefit of this representation is their invariance property to scale changes, location changes, occlusions and background clutters.

In recent, drastic advances of visual recognition are achieved by deep convolutional neural networks (CNNs) [4], which jointly learn the whole feature hierarchies starting from image pixels to the final class posterior with stacked non-linear processing layers. If sufficient training data to train a CNN is provided, intermediate activations from a CNN pre-trained on independent large data have been successfully applied as a generic image representation. [5,6] Most of previous approaches extract multiple activations from a fully connected layer by feeding multiply jitterred images and average them to get a final image representation. However, this representation are sensitive to the geometric variance such as scale/location changes and occlusions.

In this paper, we briefly review an image representation proposed by Yoo *et al*. [1], which combine the traditional Fisher kernel and deep neural activations to achieve geometric invariance. In this representation, so called multi-scale pyramid pooling (MPP), abundant amount of multi-scale local activations are extracted and aggregated by Fisher kernel [3] with a simple but important scale-wise normalization. MPP demonstrates substantial improvements on both scene and object classification tasks compared to the previous CNN-based representations.

2. Multi-scale Pyramid Pooling

To obtain a fair amount of multi-scale activations from a CNN, the fully connected layers are replaced with equivalent convolution layers. In this setting, images of multiple scales can be fed to a CNN and result in multiple activation vectors where each vector is CNN activations from the corresponding local patch. With this technique, thousands of dense local activations (4,410 per image) from multiple scale levels are extracted in a reasonable extraction time (0.46 seconds per image on a server with a CPU of 2.6GHz Intel Xeon and a GPU of GTX TITAN Black).

For representing an image from the multi-scale local activations, all the activations are merged into a single vector by Fisher kernel. To adopt the Fisher kernel suitable to CNN activation characteristics, the traditional Fisher kernel is modified as follows. Given multi-scale local activations, the dimensionality is reduced by PCA and a Gaussian mixture model (GMM) is trained as a visual dictionary. After that, the local

activations of each scale are aggregated to a scale-wise Fisher vector. The scale-wise Fisher vector are then merged into one global vector by average pooling after L_2-normalization. Average pooling is a natural pooling scheme for Fisher kernel rather than vector concatenation. Following the Improved Fisher Kernel framework [3], the power normalization followed by L_2-normalization is applied to the final vector to tackle burstiness [7] phenomena.

5. 검증

MPP is evaluated over two datasets for image classification, including scene MIT Indoor 67 and PASCAL VOC 2007. A CNN used in MPP is Alex [4] network, and pre-trained on ILSVRC'12 dataset. Henceforth, we denote this model by "Alex". For MPP, seven scales are used since the seven scales can cover large enough scale variations. Each scaled image in the pyramid has twice resolution than the previous scale starting from the standard size defined in each CNN (e.g. 227*227). Each scale image is fed to the CNN and 4,410 activation vectors of 4,096 dimension are obtained from the seventh convolution layer. Each activation vector is then reduced to 128 dimension by PCA, and a visual vocabulary (GMM of 256 Gaussian distributions) is also trained. Consequently, one 65,536 dimensional Fisher vector is computed, and further power- and L_2-normalized. One-versus-rest linear SVMs with a quadratic regularizer and a hinge loss are trained finally.

Table 1. Performances of baselines and MPP.

Method	Top-1 Acc. (Indoor 67)	mAP(%) (VOC 2007)
Alex-FC7	57.91	72.36
AP(Alex-FC7)	60.90	73.75
NFK(Alex-FC7)	71.49	74.96
MPP(Alex-FC7)	**75.67**	**79.54**

Table 1 shows the experimental results. Alex-FC7 is a standard activation vector from FC7 of Alex network with a center-crop from an input image. AP(CNN-FC7) denotes the average pooling of activations from 10 jittered images. NFK(CNN-FC7) pools the multi-scale local activations into a Fisher vector by the traditional Fisher kernel without scale-wise vector normalization. MPP(Alex-FC7) denotes the MPP method.

We compare MPP with several baseline methods. As expected, the most basic representation, Alex-FC7, performs the worst for all datasets. The average pooling improves the performance, however, the improvement is marginal. Another baseline method NFK exploiting multi-scale CNN activations shows better results than the other baselines. It shows that image representation based on CNN activations can be enriched by utilizing multi-scale local activations. Compared to NFK, MPP achieves an extra but significant performance gain of +4.18% and +4.58% for each dataset.

6. 결론

We have reviewed the multi-scale pyramid pooling for better use of neural activations from a pre-trained CNN. Through this study, we discover that it is important to take scale characteristic of neural activations into consideration for the successful combination of a Fisher kernel and a CNN. Multi-scale local activations contribute to better scale invariance when they meet the simple scale-wise normalization. The comprehensive test with various pooling strategies so far shows that MPP can be used as a primary image representation in wide visual recognition tasks for better performance.

7. Acknowledgement

This work was supported by the Technology Innovation Program (No. 10048320), funded by the Ministry of Trade, Industry & Energy (MI, Korea).

참고문헌

[1] D. Yoo, S. Park, J. Lee, and I. S. Kweon, "Multi-scale Pyramid Pooling for Deep Convolutional Representation," In CVPR Workshop (Deep Vision), 2015.
[2] D. G. Lowe, "Distinctive Image Features from Scale-invariant Keypoints." International Journal on Computer Vision (IJCV), 60(2):91- 110, 2004.
[3] F. Perronnin, J. S´anchez, and T. Mensink, "Improving the Fisher Kernel for Large-scale Image Classification," ECCV, 2010.
[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," In NIPS, 2013.
[5] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the Devil in the Details: Delving Deep into Convolutional Nets," In BMVC, 2014.
[6] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN Features Off-the-shelf: An Astounding Baseline for Recognition," In CVPR Workshop (Deep Vision), 2014.
[7] H. Jegou, M. Douze, and C. Schmid, "On the Burstiness of Visual Elements," In CVPR, 2009.