# 딥 러닝을 이용한 영상 인식 기술 동향 -ILSVRC 사례를 중심으로☆

지 명 근*, 전 준 철*

◈ 목 차 ◈



## 1. 서 론

최근 영상 인식 분야에서는 SIFT[1], HOG[2]등과 같이 사람이 직접 영상에서 특징을 추출하여 영상을 인식하는 방법 대신 인공지능 기술인 딥 러닝을 사용하여 사람이 미처 인지하지 못하는 특징을 이용하는 방법이 각광받고 있다. 딥 러닝(Deep-Learning)이란 인간의 뇌에 있는 뉴런의 동작을 모방한 인공 신경망을 여러 층 쌓은 심층 신경망을 이용하는 방법이다[3].

딥 러닝에 의한 영상인식은 2012년 ILSVRC(Large Scale Visual Recognition Challenge)에서 딥 러닝에 기반한 AlexNet(SuperVision)[4]이 기존의 SIFT나 SVM을 이용한 ISI[5], VGG에 비하여 오차율이 약 10% 낮은 16%로 우수한 분류 결과를 보인 이후, 2013년 ILSVRC에서도 딥 러닝 방법을 쓴 ZF Net[7]가 12%의 정확도를 보이며 우승을 차지했으며, 2014년의 ILSVRC도 GoogLeNet[6]이 7%의 정확도로 우승을, 2015년 ILSVRC도 ResNet[9]이 4%의 정확도로 우승을 차지함으로써 성능을 입증하게 되었다[10].

본 논문에서는 이러한 딥 러닝을 이용한 영상 인식 방법의 기술 동향을 소개한다. 2장에서는 딥 러닝 기술의 바탕이 되는 인공 신경망에 대해 설명한다. 3장에서는 인공 신경망에서도 영상 인식에 최적화되어있는 합성곱 신경망(Convolutional Neural Network)에 대해서 설명한다. 4장에서 이를 이용한 영상 인식 연구에 대해 소개하고. 마지막 5장에서 결론을 맺는다.

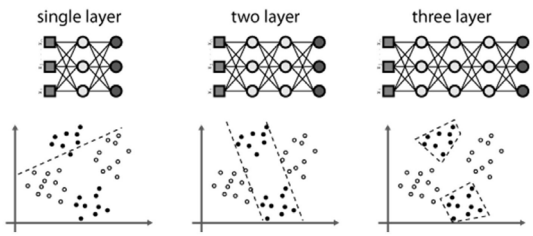
* Department of Computer Science, Kyonggi University, Gyeonggi-do, 443-760, Korea.

☆ 본 연구는 경기도의 경기도지역협력연구센터사업의 일환으로 수행하였음. [GRRC경기2017-B04:영상기반 지능정보 제조 서비스 연구]

## 2. 인공 신경망

인공 신경망이란 인간의 뇌에 있는 뉴런의 동작을 모방하기 위해 탄생한 방법이다. 가장 간단한 인공 신경망 구조로는 입력 레이어와 출력 레이어의 층이 하나인 단층 퍼셉트론이 있다[11]. 하지만 단층 퍼셉트론으로는 간단한 XOR 문제조차 해결하기 어려운 단점이 있었으며, 이를 해결하기 위하여 다층 퍼셉트론이 제안되었다[12].

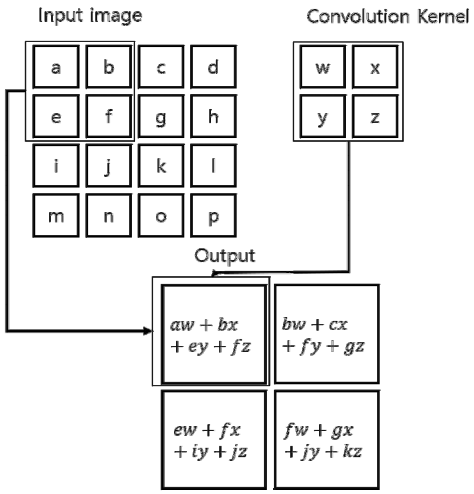
다층 퍼셉트론이란 입력 레이어와 출력 레이어 사이에 두 층 이상의 계층을 추가한 구조를 나타낸다. 아래 (그림 1)에서는 다층 퍼셉트론을 이용하여 XOR 문제를 해결하는 사례를 보여준다.




(그림 1) 단층 / 다층 퍼셉트론의 XOR문제 해결[13]

## 3. 합성곱 신경망(CNN)

합성곱 신경망(CNN:Convolutional Neural Network )란 앞서 소개된 인공 신경망의 층이 깊어지면 깊어질수록 더욱 많은 파라미터가 필요해지고, 과적합이 생기는 문제가 발생하고, 이를 해결하기 위해 도입된 방법이다[14]. CNN의 경우 여러 층이 있다는 것은 심층 신경망과 동일하지만, 층의 일부에 합성곱(컨볼루션) 레이어를 도입함으로써 문제를 해결하였다. 도입한 컨볼루션 레이어는 (그림 2)와 같다.



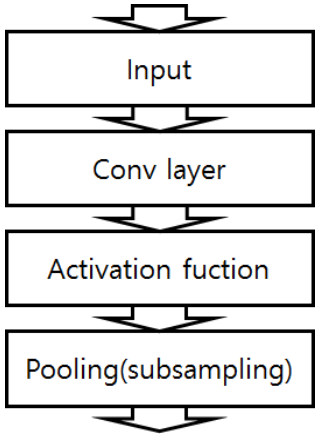

(그림 2) 컨볼루션 레이어의 작동 원리

컨볼루션 레이어는 이전 레이어와 현재 레이어의 노드가 전부 서로 연결되는 것이 아니라 컨볼루션 커널을 통하여 연결됨으로써 파라미터의 개수를 레이어 크기의 곱이 아닌 컨볼루션 커널의 크기로 축소하였다.

입력영상 $I$에 컨볼루션 커널 $F$ 가 적용된 출력 영상 $G$는 다음의 식 (1)과 같이 정의된다.

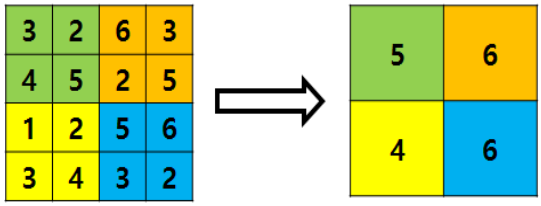
$$G[i,j] = \sum_{u=-k}^{k} \sum_{v=-k}^{k} I[u,v]\,F[i-u, j-v] \quad (1)$$

CNN은 기본적으로 (그림 3)과 같이 입력, 컨볼루션 레이어, 활성화 함수, 풀링의 레이어를 여러 층 가지고 있게 된다.




(그림 3) CNN의 기본 구조

활성화 함수는 컨볼루션 레이어를 통해 계산된 값을 다음 레이어로 넘기기 위해 사용하는 함수로, sigmoid[17], ReLu[16] 등이 있다. 풀링 레이어는 레이어의 출력을 줄여 네트워크의 사이즈를 줄이고 중요한 특징만을 얻기 위해 사용된다. 풀링 레이어의 종류에는 Max pooling과 Average pooling등이 있다.




(그림 4) Max pooling의 동작 방식

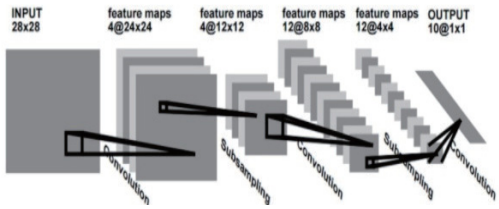
Max pooling은 (그림 4)와 같이 일정 범위 내의 최댓값 하나를 출력하는 레이어이다[4]. Average pooling 도 마찬가지로 일정 범위 내에서 평균을 다음 레이어로 넘기는 레이어이다.

## 4. CNN을 이용한 영상 인식 연구

### 4.1 LeNet[10]

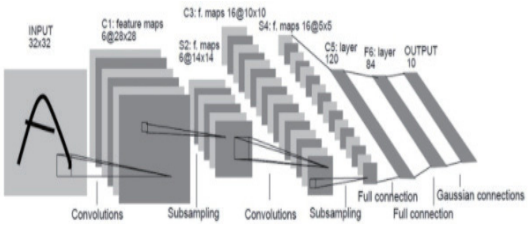
LeNet은 1990년 Yann LeCun이 우편번호와 수표의 필기체를 인식하기 위해서 연구개발 되었으며, 이는 최

초의 CNN이 되었다. 아래 (그림 5)는 1990년 발표된 LeNet-1의 구조를 보인다.




(그림 5) LeNet-1의 구조[15]

LeNet-1은 28x28 크기의 작은 영상을 사용했는데, 그 이유는 과거의 컴퓨터 사양 문제 때문이었다[15]. 그 후 입력 영상의 크기를 32x32로 확대하고 컨볼루션 커널의 크기를 늘리려 네트워크를 깊게 만든 구조인 LeNet-5를 발표하였다[10]. 아래 (그림 6)은 LeNet-5 구조를 보인다.

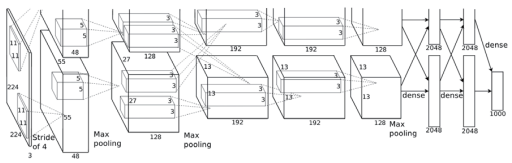



(그림 6) LeNet-5의 구조[10]

LeNet-5의 구조를 살펴 보면 3개의 컨볼루션 레이어와 2개의 서브샘플링 레이어, 1개의 전체 연결 레이어로 이루어져있는데, 컨볼루션 레이어와 서브샘플링 레이어로  영상의 크기를 1/4로 줄임으로써 기존의 심층 신경망, 즉 전체가 전체 연결 레이어로 이루어진 구조보다 영상의 크기, 회전, 위치 변화 등에 대해서 더 강건한 결과를 얻을 수 있었다[18].

## 4.2 AlexNet[4]

AlexNet은 2012년 ILSVRC에 참가하여 기존의 방법을 사용하였던 다른 팀들을 오차율이 10% 낮은 압도적인 성능으로 우승을 차지하여 딥 러닝이 영상 인식 분야에서 대표적인 방법이 되게 하였다. AlexNet의 구조는 다음 (그림 7)과 같다.

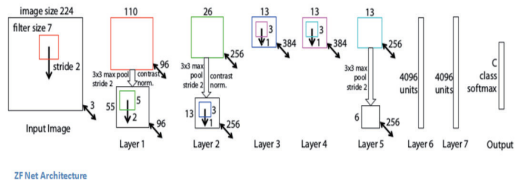



(그림 7) AlexNet의 구조[4]

AlexNet은 ILSVRC의 데이터셋인 ImageNet의 영상크기인 100x100x3 영상를 입력으로 받아 5개의 컨볼루션 레이어, 3개의 맥스 풀링 레이어, 2개의 전체 연결 레이어를 사용하였다. AlexNet에서는 크기를 줄이는 역할을 하는 레이어를 모든 컨볼루션 레이어 뒤에 쓰지 않고, 컨볼루션 연산 시 일정 범위를 뛰어넘는 스트라이드를 사용함으로써 레이어의 크기를 줄였다. 또한 활성화 함수를 학습속도가 빠른 ReLu를 사용하였다[16]. 마지막 레이어에 softmax 함수를 사용하여 1000개의 출력을 얻는다. AlexNet은 파라미터의 개수가 6000만개 이상으로 매우 많기 때문에 학습 데이터에 네트워크가 과적화되는 오버피팅 현상이 발생할 수 있기 때문에 Dropout 방법이 적용되었다[18].

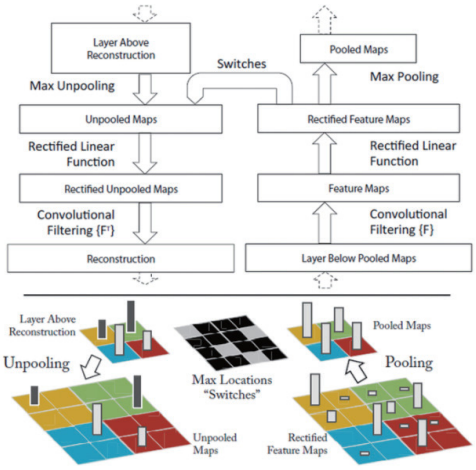
## 4.3 ZFNet[7]

ZFNet은 2013년 ILSVRC에서 오차율 12%로 우승을 하였다. ZFNet은 (그림 8)과 같이 AlexNet의 구조와 동일하며, 이를 시각화 기법을 통해 하이퍼 파라미터를 최적화시킴으로써 성능을 증가시켰다.



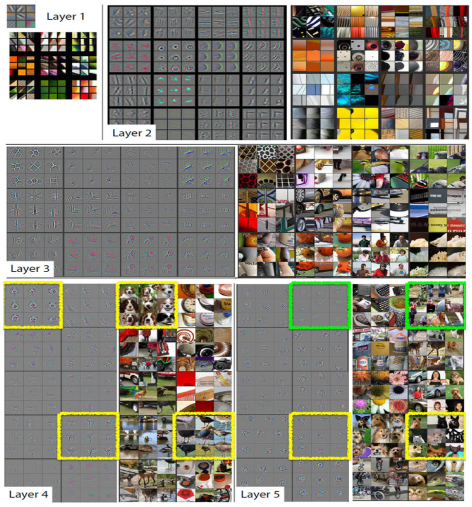

(그림 8) ZFNet의 구조[7]

시각화 기법이란 CNN의 내부 파라미터 변화를 관찰하기 위해 제안된 방법으로, 풀링 레이어를 통과하면서 사라진 값들을 복원하여 보여주는 방법이다.




(그림 9) 시각화 과정[7]

ZFNet의 시각화 과정은 (그림 9)와 같이 Switch를 도입하여 활성화 함수를 통과한 값들의 위치를 저장해두고, 복원 시 사용하게 된다. 이 Switch를 바탕으로 높은 값들의 위치를 표시하는 방식으로 동작하게 된다.

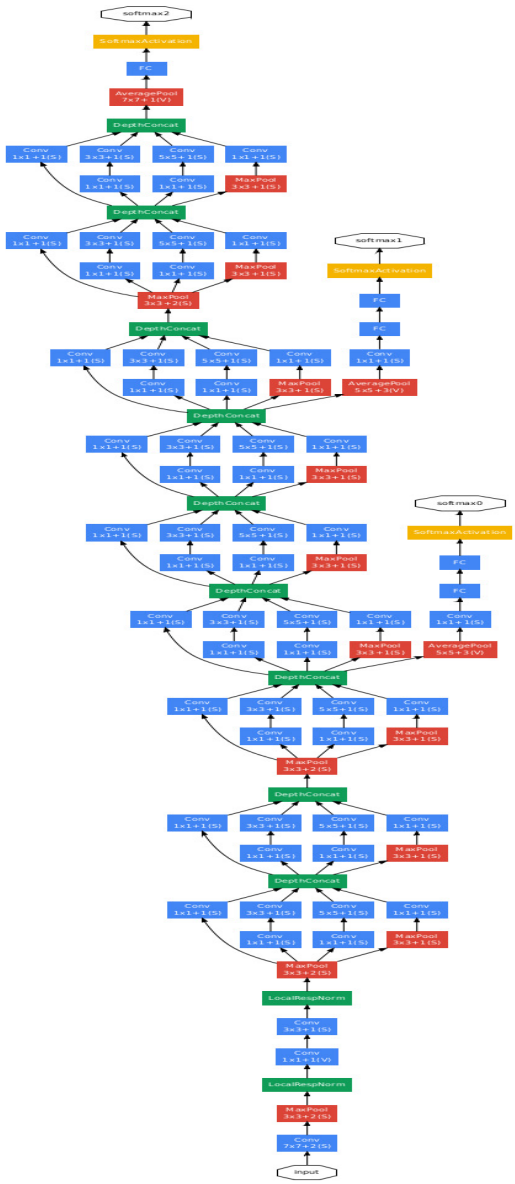



(그림 10) 시각화 결과[7]

(그림 10)의 결과를 보면 레이어1과 2에서는 영상의 코너, 에지, 색과 같은 낮은 레벨의 특징을 추출하는 것을 알 수 있고, 레이어가 점점 깊어질수록 구체적인 특징을 추출하는 것을 볼 수 있다.

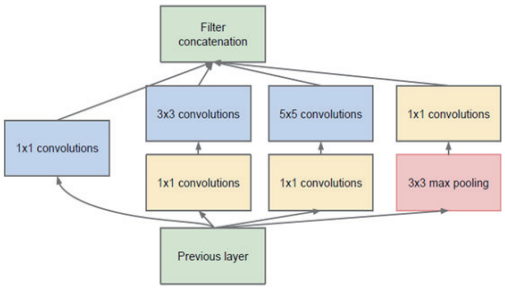
## 4.4 GoogLeNet[6]

GoogLeNet는 2014년 ILSVRC에서 7%의 오차율로 우승을 하였다. 이전까지의 구조와 달리 (그림 11)과 같이 22층의 레이어로 구성된 깊은 구조이다.



(그림 11)GoogLeNet 구조[6]

기존의 신경망과 달리 GoogLeNet은 (그림 12)와 같이 같은 층의 레이어에서 여러 사이즈의 컨볼루션 커널을 사용하는 inception module을 제안하였다.

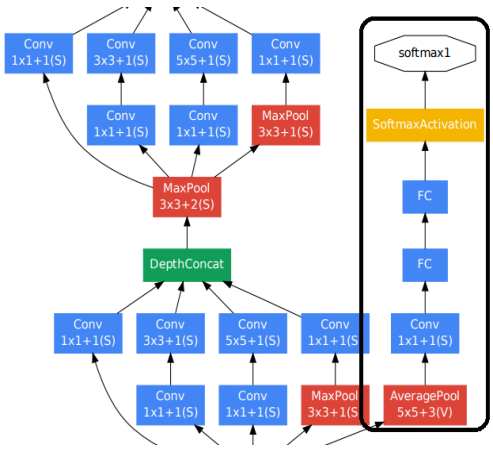



(그림 12) GoogLeNet의 inception module[6]

Inception module은 여러 종류의 컨볼루션 레이어가 사용되는 만큼 늘어난 파라미터로 인해 연산 시간이 증가할 수 있는데, 이를 1x1 컨볼루션 레이어를 이용하여 컨볼루션 커널의 개수를 줄임으로써 해결하였다.

GoogLeNet은 9개의 inception module을 포함한 총 22개의 레이어가 존재하며, 이전까지 레이어 마지막에 존재하던 전체 연결 레이어가 존재하지 않는다. 대신 global average pooling 레이어를 사용하고 있는데, 이를 통해 AlexNet등의 구조와 비교했을 때에도 깊이는 더 깊지만 파라미터의 수는 더 적게 될 수 있었다[6].

또 다른 GoogLeNet의 특징은 Auxiliary classifier이다.

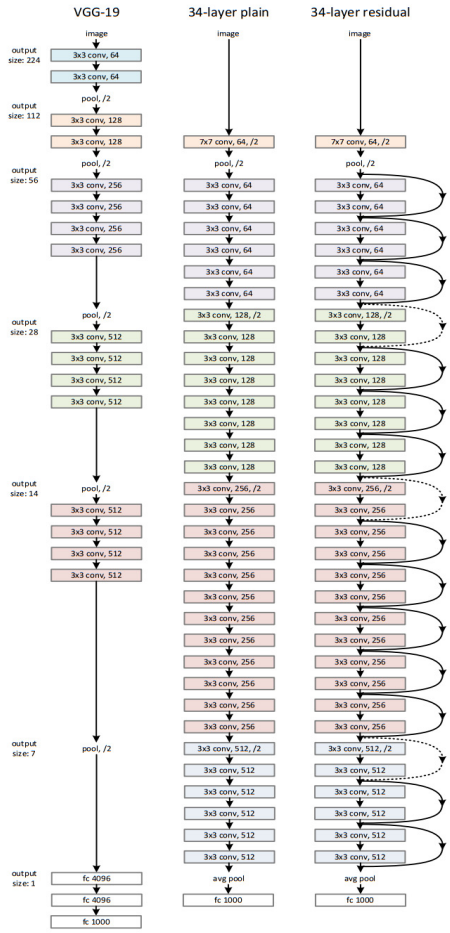



(그림 13) GoogLeNet의 Auxiliary classifier[6]

이는 최종 출력 레이어에서 오차를 계산하여 역전파시키는 역전파학습의 특성 상 레이어가 깊어질수록 학습이 잘 되지 않는 기울기 소실 문제가 생기는데, 이를 레이어 중간에 학습을 시킴으로써 해결하는 방법이다[19]. 이 방법은 학습 시에만 작동하며, 테스트 시에는 작동하지 않게 된다.

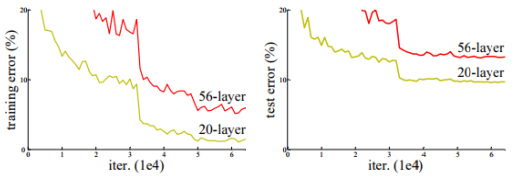
## 4.5 ResNet[9]

ResNet은 2015년 ILSVRC에서 오차율 4%로 우승을 하였다. 이는 사람이 분류한 오차율인 5%보다 높은 첫 번째 네트워크이다.



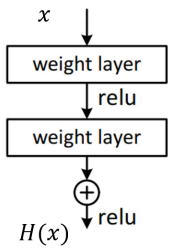

(그림 14) VGG-19모델(좌), 34층의 일반 CNN모델(중), 34층의 residual 네트워크(우)[6]

이 네트워크는 총 152층으로 이전의 다른 네트워크들보다 훨씬 깊은 구조를 가지고 있다. 이전의 네트워크들이 점점 깊어지는 것에서 알 수 있듯 깊은 네트워크는 성능이 좋지만, 파라미터가 많아져 학습이 느려지고, 기울기가 소실되는 등의 문제가 발생하게 된다. 아래 (그림 15)는 네트워크가 깊어졌을 때 학습이 제대로 되지 않는 현상을 실험한 결과이다[9].



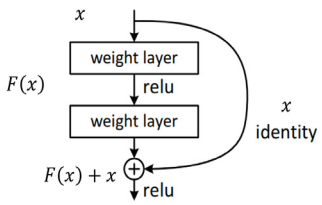

(그림 15) 20층과 50층 레이어를 가진 네트워크의 CIFAL-10 데이터셋 학습 결과[9]

이를 해결하기 위하여 ResNet에서는 Residual Learning이란 방법을 도입하였다.




(그림 16)일반적인 네트워크

일반적인 네트워크는 (그림 16)과 같이 $H(x)$를 얻을 수 있다면, 이는 $H(x)-x$를 얻을 수 있다는 것과 동일하다[9]. 이 때 $F(x)=H(x)-x$이면, $H(x)=F(x)+x$가 된다.




(그림 17) ResNet의 Residual Learning[9]

이때 $x$를 입력 레이어에서 가져와 $F(x)$에 더해주게 되면, 이 네트워크는 $H(x)-x$를 얻기 위한 학습을 하게 되며, 최적의 경우 $F(x)$가 0이 되어야 한다. 이렇게 0이 되는 방향으로 학습을 하게 되면 입력의 작은 움직임, 즉 나머지(residual)을 학습하게 되어 이를 Residual Learning이라고 부른다[9]

이 구조를 적용하게 되면 $x$가 출력층에 더해지는 것에 대한 연산량 증가를 빼면 파라미터도 증가하지 않고, 레이어를 건너뛰며 연결이 되기 때문에 학습이 간단해지게 된다. 따라서 깊은 네트워크를 좀 더 최적화 하여 깊은 네트워크의 장점인 정확도를 얻을 수 있다.

## 5. 결 론

본 논문에서는 딥 러닝을 이용한 영상 인식 시스템의 발전사를 2012년에서 2015년까지 ILSVRC에서 우승한 방법들의 특징들을 중심으로 간략하게 설명하였다. 본문에서 보이듯 딥 러닝을 이용한 영상 인식 방법은 기존의 방법들과 비교하여 높은 정확도를 나타내며, 매년 새로운 방법이 소개되어 다양한 분야에서 널리 쓰이고 있다. 영상 인식 분야에서는 향후 딥 러닝을 이용한 연구가 계속 될 것으로 예상된다.

## 참 고 문 헌

[1] David G Lowe. Distinctive image features from scaleinvariant keypoints. International journal of computer vision, Vol. 60, No. 2, pp. 91-110, 2004.

[2] Navneet Dalal and Bill Triggs. “Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition”, 2005. CVPR 2005. Computer Society Conference, Vol. 1, pp.886-893. 2005.

[3] Collobert R., and Weston J. “A unified architecture for natural language processing: Deep neural networks with multitask learning,” In Proceedings of the 25th international conference, pp. 160-167, 2008.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural

networks," Communications of the ACM, vol. 60, no. 6, pp. 84-90, May 2017.

[5] Harada, Tatsuya, and Yasuo Kuniyoshi. "Graphical Gaussian vector for image categorization." Advances in Neural Information Processing Systems, pp. 1547-1555, 2012.

[6] C. Szegedy et al., "Going deeper with convolutions," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[7] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in Computer Vision-ECCV 2014, Springer pp. 818-833, 2014.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] O. Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge," International Journal of Computer Vision, Vol. 115, No. 3, pp. 211-252, Apr. 2015.

[11] LeCun Y., Bottou L., Bengio Y., and Haffner P. "Gradient-based learning applied to document recognition," Proceedings of the IEEE, pp. 2278-2324, 1998.

[12] Raudys Š. "Evolution and generalization of a single neurone: I. single-layer perceptron as seven statistical classifiers," Neural Networks, pp. 283-296, 1998.

[13] Zhang, Zhengyou, et al. "Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron," Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on. IEEE, pp. 454-459, 1998.

[14] Maltarollo V. G., Honório K. M., and da Silva, A. B. F. "Applications of artificial neural networks in chemical problems," In Artificial neural networks architectures and applications, 2013.

[15] LeCun Y., and Bengio Y. "Convolutional networks for images, speech, and time series," The handbook of brain theory and neural networks, 1998.

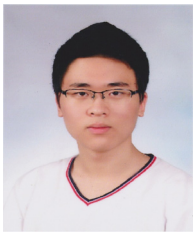
[16] Nair V., and Hinton G. E. "Rectified linear units improve restricted boltzmann machines," In Proceedings of the 27th international conference on machine learning, pp. 807-814, 2010.

[17] HAN, Jun; MORAGA, Claudio. "The influence of the sigmoid function parameters on the speed of backpropagation learning," From Natural to Artificial Neural Computation, pp. 195-201, 1995.

[18] SRIVASTAVA, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting." Journal of machine learning research, vol.15, pp. 1929-1958, Jun 2014.

[19] Hochreiter S. "The vanishing gradient problem during learning recurrent neural nets and problem solutions," International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, pp. 107-116, 1998.

## ◐ 저 자 소 개 ◑

**지 명 근(Myunggeun Ji)**

2017 B.S. in Computer Science, Kyonggi University, Suwon, Korea
2017~Present : M.S. Student in Computer Science, Kyonggi University, Suwon, Korea
Research Interests : Computer Vision, Augmented Reality
E-mail : jmg2968@gmail.com

**전 준 철(Junchul Chun)**

1984 B.S. in Computer Science, Chung-Ang University, Seoul, Korea
1986 M.S. in Computer Science(Software Engineering), Chung-Ang University, Seoul, Korea
1992 M.S. in Computer Science and Engineering (Computer Graphics), The Univ. of Connecticut, USA
1995 Ph.D. in Computer Science and Engineering (Computer Graphics), The Univ. of Connecticut, USA
2001.02～2002.02 Visiting Scholar, Michigan State Univ. Pattern Recognition and Image Processing Lab.
2009.02～2010.02 Visiting Scholar, Univ. of Colorado, Wellness Innovation and Interaction Lab.
Research Interests : Augmented Reality, Computer Vision, Human Computer Interaction
E-mail : jcchun@kgu.ac.kr