ISSN 1598-0170 (Print)
ISSN 2287-1136 (Online)
http://www.jksii.or.kr

딥러닝을 이용한 인스타그램 이미지 분류☆

Instagram image classification with Deep Learning

정 노 권1　　조 수 선1*
Nokwon Jeong　　Soosun Cho

요　약

본 논문에서는 딥러닝의 회선신경망을 이용한 실제 소셜 네트워크 상의 이미지 분류가 얼마나 효과적인지 알아보기 위한 실험을 수행하고, 그 결과와 그를 통해 알게 된 교훈에 대해 소개한다. 이를 위해 ImageNet Large Scale Visual Recognition Challenge(ILSVRC)의 2012년 대회와 2015년 대회에서 각각 우승을 차지한 AlexNet 모델과 ResNet 모델을 이용하였다. 평가를 위한 테스트 셋으로 인스타그램에서 수집한 이미지를 사용하였으며, 12개의 카테고리, 총 240개의 이미지로 구성되어 있다. 또한, Inception V3모델을 이용하여 fine-tuning을 실시하고, 그 결과를 비교하였다. AlexNet과 ResNet, Inception V3, fine-tuned Inception V3 이 네 가지 모델에 대한 Top-1 error rate들은 각각 49.58%, 40.42%, 30.42% 그리고 5.00%로 나타났으며, Top-5 error rate들은 각각 35.42%, 25.00%, 20.83% 그리고 0.00%로 나타났다.

☞ 주제어 : 이미지 분류, 회선 신경망, 인스타그램 이미지

ABSTRACT

In this paper we introduce two experimental results from classification of Instagram images and some valuable lessons from them. We have tried some experiments for evaluating the competitive power of Convolutional Neural Network(CNN) in classification of real social network images such as Instagram images. We used AlexNet and ResNet, which showed the most outstanding capabilities in ImageNet Large Scale Visual Recognition Challenge(ILSVRC) 2012 and 2015, respectively. And we used 240 Instagram images and 12 pre-defined categories for classifying social network images. Also, we performed fine-tuning using Inception V3 model, and compared those results. In the results of four cases of AlexNet, ResNet, Inception V3 and fine-tuned Inception V3, the Top-1 error rates were 49.58%, 40.42%, 30.42%, and 5.00%. And the Top-5 error rates were 35.42%, 25.00%, 20.83%, and 0.00% respectively.

☞ keyword : Image Classification, Convolutional Neural Network, Instagram Images

1. Introduction

Recently, many applications using Convolutional Neural Network have been appearing in the field of image processing[1, 2, 3]. In 2012, the SuperVision team led by professor Geoffrey Hinton of the University of Toronto won the image recognition competition named ILSVRC(ImageNet Large Scale Visual Recognition Challenge). In several years, beginning with their model named AlexNet[4], lots of network structures were studied. In every year ILSVRC, the performance of top class neural network models has been increased[5-10].

However, these results based on the same data set, ImageNet data set. Our study starts with some doubts about the data. The ImageNet data set consists of well refined images, so it is different from unrefined images created by users in social media. Our question is how much it would be effective the neural networks trained on the ImageNet, on those images created by actual users.

In this situation, there was a research for classifying and tagging actual social networks images using neural networks trained by ImageNet[11]. According to the researchers, tagging using the 2012 Toronto model is more accurate than the tagging by Instagram users.

1 Department of Computer Science and Information Engineering, Korea National University of Transportation, Uiwang Gyeonggi, 16106, Korea.
* Corresponding author (sscho@ut.ac.kr)
[Received 5 June 2017, Reviewed 12 June 2017(R2 26 July 2017), Accepted 27 July 2017]
☆ This work (Grants No. C0395945) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2016.
☆ A preliminary version of this paper was presented at ICONI 2016 and was selected as an outstanding paper.

http://dx.doi.org/10.7472/jksii.2017.18.5.61

We developed this research further. We evaluate not only the AlexNet[4] model used in previous research and won the championships in the 2012 ILSVRC, but also the ResNet[8] model won the 2015 ILSVRC . For evaluation, we used Instagram images. Instagram is one of the most popular image-based social network service. We examined how much the performances of the networks were improved.

In addition, we performed fine-tuning to classify some untrained categories, and to increase the performance of other categories. We used another CNN model called Inception V3 model[9] which is a variant of GoogLeNet using Inception module. It is known that this model has higher performance of classification than that of ResNet, and is frequently introduced on tutorials of fine-tuning. And then we compared those results.

Finally, to apply those networks to actual image classification, we will introduce several difficulties and some useful lessons found from our study.

2. Related Works

Image classification, tagging, and retrieval are getting more and more important because of explosively increasing images on social network services. So there are many methods and researches to classify the images such as social tag method, local information method, and content based method[12-14].

Content based image classification uses several features such as shape, color, texture and etc. Thus, content based image classification using those various features has been performed steadily for a long time[13, 14].

Recently, many researches and applications in the field of computer science pay attention to Neural networks, and one of these cases is research of G. Kim et al.[15]. Especially, in the field image processing, there are several neural network models called Convolutional Neural Network to extract feathers and classify images.

Convolutional Neural Network is the Neural Network using Convolution processing. Convolutional Neural Network consists of stacks of three type of layers: Convolutional Layer, Pooling Layer, and Fully-connected Layer. Convolutional layer performs Convolution processing that is the element-wise multiplication process of the input and the weights matrix called the kernel or filter. and its result is regarded as feather. Deeper layer makes higher level feather. In Convolutional Neural Network, Learning means adjusting those weights. To learn the proper weights, a process called Backpropagation is needed. Forwarding means the process of passing the input through the layers, multiplying the weights, and performing the activation function that reacts if it is above a certain threshold and ignores the below. After forwarding to last layer, Backpropagation is performed. Backpropagation means the update of those weights to decrease the error, the difference between correct and output. To update the weights, backward sequentially, calculate the gradient of the error function for each layer weight and add this variation in the negative direction to the original weight. The Pooling Layer performs down-sampling to reduce the amount of computation. The Pooling Layer performs pooling processing that is a computation for extracting a representative value from the n * n receptive field, and generally uses maxpooling to extract the maximum value. The Pooling, however, is not used well recently. because there is too much information loss. Fully-Connected Layer is connected to all activations of the previous layer and calculates the score of the classes to perform classification.

These Convolutional Neural Network has been studied extensively in recent years. In several years, beginning with their model named AlexNet[4], lots of network structures were studied. In every year ILSVRC, the performance of top class neural network models has been increased[5-10].

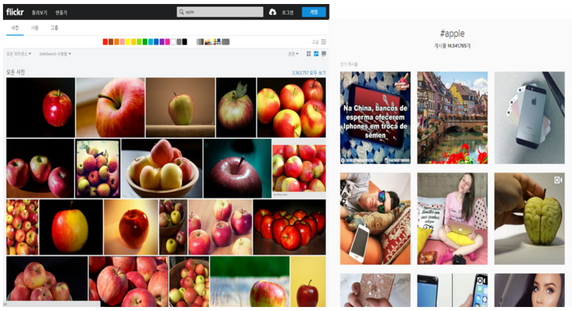
In this situation, there was a research for classifying and tagging actual social networks images using neural networks trained by ImageNet[11]. According to H. Jang et al.[11], tagging using the 2012 Toronto model is more accurate than the tagging by Instagram users.

3. Experiments and Evaluations

3.1. AlexNet v.s. ResNet

In this experiment, we used the ImageNet 2012 dataset[16] as the training set. It has been used in the various field of image processing for competitions in the ILSVRC. The data set consists of 1.2 million images devided into 1,000 categories.

And we used images from Instagram as a test set have been selected. Recently, the most famous image-based social network services are Flickr and Instagram. Figure 1. shows the images searched with 'apple' tag on Flickr and Instagram.



(Figure 1) Flickr(left) and Instagram(right) images retrieved using ‘apple’ tag.

As you can see from the photos, Flickr has refined images with relatively higher resolution and refined tags compared to Instagram. It is mainly used by expert users with DSLR cameras. On the contrary, Instagram images created by smartphone has more generality. As a result, its images and tags relatively less refined compared to Flickr. In this study, Instagram images were collected to use it as a test set for evaluation because we judged that not refined, actual images could be appropriate for our research goal, social image classification.

For classification, we set 12 categories. And 20 images per category have been collected to be used in the experiment. We excluded some collected images that are inappropriate images for the category, for example, it contains adult content. The 12 categoies was selected on those themes that general social network user may be post, such as 'dog', 'cat', and 'car.'

These category labels don't always coincide with 1,000 category labels of the ImageNet 2012 dataset. Besides, the 1,000 category labels of ImageNet are so concrete and deep compared to tags generally in use that it is different from generally used words. For example, ImageNet 2012 dataset is using many labels in the such as 'Egyptian cat', 'Persian cat.' and so on. They corresponds to the category label 'cat' used in the experiment which is more generally used. In such cases, we judged the case that the network generated labels such as 'Egyptian cat', 'Persian cat', as being classified into the category 'cat'. The 12 categories used in the experiment and the corresponding ImageNet labels are shown in the Table 1.

(Table 1) 12 categories corresponding to ImageNet labels

Category	ImageNet labels
bike	'mountain bike, all-terrain bike, off-roader'/'bicycle-built-for-two, tandem bicycle, tandem'/'tricycle, trike, velocipede'/'unicycle, monocycle'//'moped'/'motor scooter, scooter'
bird	'ptarmigan'/'jay'/'brambling, Fringilla montifringilla'/'hornbill'/'bald eagle, American eagle, Haliaeetus leucocephalus'/'drake'/'bulbul'/'goose'/etc...
butterfly	'cabbage butterfly'/'sulphur butterfly, sulfur butterfly'/'lycaenid, lycaenid butterfly'/'ringlet, ringlet butterfly'/'monarch, monarch butterfly, milkweed butterfly, Danaus plexippus'/
car	'sports car, sport car'/'racer, race car, racing car'/'minivan'/'tow truck, tow car, wrecker'/'beach wagon, station wagon, wagon, estate car, beach waggon, station waggon, waggon'/'limousine, limo'/'convertible'/'cab, hack, taxi, taxicab'/'police van, police wagon, paddy wagon, patrol wagon, wagon, black Maria'/etc...
cat	'Egyptian cat'/'tabby, tabby cat'/'Persian cat'/'tiger cat'
chair	'barber chair'/'rocking chair, rocker'/'folding chair'/'studio couch, day bed'/'park bench'
dog	'English foxhound'/'golden retriever'/'Chihuahua'/'Pembroke, Pembroke Welsh corgi'/'chow, chow chow'/'Labrador retriever'/'Blenheim spaniel'/ etc...
flower	'vase'/'daisy'/'pot, flowerpot'/
night sky	
rainbow	
strawberry	'strawberry'
waterfall	'valley, vale'/'cliff, drop, drop-off'/'dam, dike, dyke'

Even though most CNN networks require fixed image sizes but our test images from Instagram have various sizes. So almost all images from Instagram should be cropped or resized.

In this study, we used the ImageNet 2012 dataset as a training set. And as classification engines, we used AlexNet[4] model and the ResNet-50[8] model. Even though the system is equipped with the latest GPU (Graphics Processing Unit), usually too much training time is required for training of these neural networks. Therefore the pre-trained networks have been frequently used to shorten the training time as existing studies[17, 18].

Table 2 is a summary for the experimental results. The classification results with AlexNet for the 12 categories showed 49.58% as top-1 error rate and 35.42% as top-5 error rate and those with ResNet were 40.42% and 25.00%, respectively. The number of layers of AlexNet is 8. Compared to AlexNet, the number of layers of ResNet is 50. We recognized that as a network became deeper better results appeared.

The number of correct labels in case of Categories 'night sky' and 'rainbow' was 0. That is because there was no label that corresponds to the 1,000 ImageNet categories as in Table 1. In other words, these categories were not learned by the neural networks. In general, the results are not bad excepting such categories that were not learned by the neural networks. Excluding categories 'night sky' and 'rainbow', the results for 10 categories showed 39.50% top-1 error rate and 22.50% top-5 error rate in the case of AlexNet, and 28.50% and 10.00% in the case of ResNet respectively. It indicates that quite good results are shown for categories that have been learned. However, the classification of deeper categories showed relatively poor results. For example, classification results for varieties of flowers under 'flower' category was almost 'vase', 'daisy', 'pot, flowerpot' even though the flower was 'rose'.

Categories which showed relatively poor results include the images with high aspect ratios that cause much distortion in the process of image resizing. And the images with mixtures of multiple objects as shown in Figure. 2 produced poor results too.

For instance, in the case of images with category 'butterfly', it is highly likely that mixtures of plant objects such as flowers and trees. So not only one flower but also many flowers would appear around the butterfly due to the characteristics of those images. In the case of the left image of Figure 2, rather than main object 'bird', sub-objects or scene objects can be mainly appeared with labels such as 'television, television system' or 'microwave, microwave oven'. Actually those labels were included in top-5 labels. In the case of the right part of Figure 2, instead of 'butterfly', wrong labels were appeared such as 'shower curtain', 'honeycomb', 'mask', 'ocarina', and 'sweet potato' appeared as top-5 labels. We found that simple image classification performance was lower on such images with multiple objects.

〈Table 2〉 The classification results from AlexNet and ResNet-50

	AlexNet		ResNet-50	
Category	Top-1 error(%)	Top-5 error(%)	Top-1 error(%)	Top-5 error(%)
bike	10 (50.00)	2 (10.00)	5 (25.00)	1 (5.00)
bird	11 (55.00)	8 (40.00)	4 (20.00)	2 (10.00)
butterfly	12 (60.00)	9 (45.00)	9 (45.00)	3 (15.00)
car	6 (30.00)	3 (15.00)	4 (20.00)	0 (0.00)
cat	13 (65.00)	7 (35.00)	8 (40.00)	5 (25.00)
chair	7 (35.00)	3 (15.00)	10 (50.00)	5 (25.00)
dog	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
flower	9 (45.00)	7 (35.00)	9 (45.00)	2 (10.00)
*nightsky	20 (100.00)	20 (100.00)	20 (100.00)	20 (100.00)
*rainbow	20 (100.00)	20 (100.00)	20 (100.00)	20 (100.00)
strawberry	4 (20.00)	2 (10.00)	3 (15.00)	1 (5.00)
waterfall	7 (35.00)	4 (20.00)	5 (25.00)	1 (5.00)
total	119 (49.58)	85 (35.42)	97 (40.42)	60 (25.00)
* excepted total	79 (39.50)	45 (22.50)	57 (28.50)	20 (10.00)

(Figure 2) Two images in which multiple objects are mixed up

3.2. Inception V3 with fine-tuning

In this experiment, we used another CNN model called Inception V3 model[9] which is a variant of GoogLeNet using Inception module. It is known that this model has higher performance of classification than that of ResNet, and is frequently introduced on tutorials of fine-tuning. Actually we have tried Inception V3 with fine-tuning, a transfer learning method. It means selective re-training on a pre-trained network. So, it shorten the learning time by using the pre-trained network's weight to initialize a network. For fine-tuning the model Inception V3, 9,975 Flickr images were used as a training set and we performed evaluation with 1,200 Flickr images as a validation set. Test set is same as the previous experiment which consists of 240 Instagram images.

We have fine-tuned pre-trained Inception V3 model as follows; only re-train last fully connected layer that serves an output classifier, 'batch_size=32', 'max_number_of_steps =3000'. The 'batch_size' means the number of images per one training step and 'max_number_of_steps' means the number of repetitions of training steps with the images of batch_size. The training results from our fine-tuned Inception V3 were 99.33% top-1 accuracy and 100.00% top-5 accuracy on the 1,200 evaluation set.

Table 3 is a summary for the experimental results. The classification results with Inception V3 were more accurate than ResNet-50, the winner in previous experiment. The results from Inception V3 model for the 12 categories showed 30.42% top-1 error rate and 20.83% top-5 error rate. The results for 10 categories excluding categories 'nightsky' and 'rainbow' showed 16.50% top-1 error rate and 5.00% top-5 error rate.

(Table 3) The classification results from Inception V3

Category	ResNet-50	Inception V3		fine-tuned Inception V3
	Top-1 error(%)	Top-1 error(%)	Top-5 error(%)	Top-1 error(%)
bike	5 (25.00)	1 (5.00)	0 (0.00)	0 (0.00)
bird	4 (20.00)	1 (5.00)	1 (5.00)	0 (0.00)
butterfly	9 (45.00)	4 (20.00)	2 (10.00)	2 (10.00)
car	4 (20.00)	2 (10.00)	0 (0.00)	0 (0.00)
cat	8 (40.00)	6 (30.00)	2 (10.00)	0 (0.00)
chair	10 (50.00)	4 (20.00)	2 (10.00)	0 (0.00)
dog	0 (0.00)	0 (0.00)	0 (0.00)	1 (5.00)
flower	9 (45.00)	10 (50.00)	2 (10.00)	1 (5.00)
*nightsky	20 (100.00)	20 (100.00)	20 (100.00)	0 (0.00)
*rainbow	20 (100.00)	20 (100.00)	20 (100.00)	0 (0.00)
strawberry	3 (15.00)	0 (0.00)	0 (0.00)	0 (0.00)
waterfall	5 (25.00)	5 (25.00)	1 (5.00)	1 (5.00)
total	97 (40.42)	73 (30.42)	50 (20.83)	5 (2.08)
* excepted total	57 (28.50)	33 (16.50)	10 (5.00)	5 (2.50)

On the other hands, the results from fine-tuned Inception V3 was more accurate than non fine-tuned one. The results from fine-tuned Inception V3 showed only 5.00% top-1 error rate and 0.00% top-5 error rate. Especially, compared to non-fine-tuned one's, the results with category 'nightsky' and 'rainbow' was amazing. Because our fine-tuned Inception V3 engine has been trained and evaluated with Flickr images including 'nightsky' and 'rainbow' categories these wonderful results could be possible.

4. Lessons

In this study, we examined the enhancement of Deep learning technology, Convolution Neural Network, for image classification. To check the possibility to utilize

CNN as a actual social image classifier, we used Instagram images as a test set. The lessons obtained through this study are as follows;

1) In AlexNet and ResNet, the classification of social images is quite efficient even when these networks trained with ImageNet data set.

2) Better results could be obtained by using the deeper network structures in the case of social image classification.

3) Additional training, fine-tuning with additional training data is necessary for some untrained categories. Because social images under some categories are not classified so well with only ImageNet training data.

4) Object detection is required because image cropping/resizing processes affect image classification through networks.

5) It may be proper using Flickr to train networks for classification of social images, using Flickr images, because they are relatively well tagged and have high quality.

6) It could be obtained amazing classification results just by fine-tuning the last fully connected layer on the pre-trained network.

5. Conclusions

In this study, images from Instagram, which is one of the most representative social networks, were applied to actual image classification to check the possibility to utilize CNN as a social image classifier. In addition, various useful lessons obtained from experiments were introduced. Through experiments, we can find that social image classification using CNN was very effective, particularly with fine-tuned networks such as fine-tuned Inception V3 model. So, our study can be enhanced by re-training and fine-tuning the networks using images from actual users in social networks in addition to ImageNet data.

More over we can recognize that many interesting research subjects can be studied hereafter through various lessons. Recently, since studies on object detection using neural networks have been actively presented in addition to the field of image classification[19, 20, 21], if relevant technologies are applied, the results in this paper could be enhanced.

참고문헌(Reference)

[1] L.A. Gatys, A.S. Ecker, M. Bethge,"A Neural Algorithm of Artistic Style", arXiv:1508.06576, 2015. https://doi.org/10.1167/16.12.326

[2] H. Kagaya, K. Aizawa, M. Ogawa "Food Detection and Recognition Using Convolutional Neural Network", MM ’14 Proceedings of the 22nd ACM international conference on Multimedia, Pages 1085-1088, 2014. https://doi.org/10.1145/2647868.2654970

[3] F. N. Iandola, A. Shen, P. Gao, K. Keutzer, "DeepLogo: Hitting Logo Recognition with the Deep Neural Network Hammer", arXiv:1510.02131, 2015. https://arxiv.org/abs/1510.02131

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolution Neural Networks", In NIPS,, 2012. https://doi.org/10.1145/3065386

[5] M. Lin, Q. Chen, S. Yan," Network In Network", arXiv:1312.4400, 2013. https://arxiv.org/abs/1312.4400

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, "Going Deeper with Convolutions", arXiv:1409.4842, 2014. https://doi.org/10.1109/cvpr.2015.7298594

[7] K. Simonyan, A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", arXiv:1409.1556, 2014. https://arxiv.org/abs/1409.1556

[8] K. He, X. Zhang, S. Ren, J. Sun, "Deep Residual Learning for Image Recognition", arXiv:1512.03385, 2015. https://doi.org/10.1109/cvpr.2016.90

[9] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", arXiv:1512.00567, 2015. https://doi.org/10.1109/cvpr.2016.308

[10] K. He, X. Zhang, S. Ren, J. Sun, "Identity Mappings in Deep Residual Networks", arXiv:1603.05027, 2016. https://doi.org/10.1007/978-3-319-46493-0_38

[11] H. Jang, S. Cho, "Automatic Tagging for Social Images using Convolution Neural Networks", Journal

of KIISE, Vol 43,No. 1, pp. 47-53, 2016. https://doi.org/10.5626/jok.2016.43.1.47

[12] K. Tang, M. Paluri, L. Fei-Fei, R. Fergus, L. Bourdev, "Improving Image Classification with Location Context", arXiv:1505.03873v1, 2015. https://doi.org/10.1109/iccv.2015.121

[13] S. Cho, "Web Image Classification using Semantically Related Tags and Image Content", Journal of Internet Computing and Services, v.11, no.3, pp.15-24, 2010. http://ksci.kisti.re.kr/search/article/articleView.ksci?articleBean.atclMgntNo=OTJBCD_2010_v11n3_15

[14] Y. Mussarat, S. Muhammad, M. Sajjad and I. Isma, "Content Based Image Retrieval Using Combined Features of Shape, Color and Relevance Feedback" , KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 7, NO. 12, pp.3149, 2013. https://doi.org/10.3837/tiis.2013.12.011

[15] G. Kim, T. An, M. Kim, "Estimation of Crowd Density in Public Areas Based on Neural Network", KSII TRANSACTIONS ON INTERNET AND INFORMATION SYSTEMS VOL. 6, NO. 9, pp.2170, 2012. https://doi.org/10.3837/tiis.2012.09.011

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." IJCV, 2015. https://doi.org/10.1007/s11263-015-0816-y

[17] A. Vedaldi and K. Lenc, "MatConvNet - Convolutional Neural Networks for MATLAB", Proc. of the ACM Int. Conf. on Multimedia, 2015. https://doi.org/10.1145/2733373.2807412

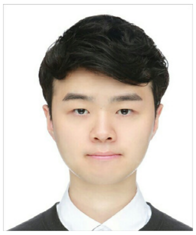
[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, "TensorFlow : Large-Scale Machine Learning on Heterogeneous Distributed Systems", arXiv:1603.04467, 2015. https://arxiv.org/abs/1603.04467

[19] R. Girshick, "Fast R-CNN", arXiv:1504.08083, 2015. https://doi.org/10.1109/iccv.2015.169

[20] K. He, X. Zhang, S. Ren, J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition", arXiv:1406.4729, 2014. https://doi.org/10.1007/978-3-319-10578-9_23

[21] S. Ren, K. He, R. Girshick, J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv:1506.01497, 2015. https://doi.org/10.1109/tpami.2016.2577031

◐ Authors ◑



정 노 권 (Nokwon Jeong)

Nokwon Jeong is a graduate student in computer science and information engineering at Korea National University of Transportation, Uiwang, Korea. His research interests include deep learning and image processing.

조 수 선 (Soosun Cho)

Soosun Cho has been a professor in computer science and information engineering at Korea National University of Transportation, Korea from 2004. She received the B.S. and the M.S. degrees in Computer Science and Statistics from Seoul National University, Seoul, Korea. And she received the Ph.D. degree from Chungnam National University, Daejeon, Korea. She worked as a senior research engineer at ETRI-Computer & Software Technology Lab from 1994 to 2004. She was a visiting scholar at University of Michigan-Ann Arbor from 2006 to 2007. Her current research interests include data mining, machine learning, and so on.