ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 23-03-16
j.inst.Korean.electr.electron.eng.Vol.27,No.3,313~318,September 2023
https : //doi.org/10.7471/ikeee.2023.27.3.313
101

# Vision Transformer를 활용한 비디오 분류 성능 향상을 위한 Fine-tuning 신경망

# Fine-tuning Neural Network for Improving Video Classification Performance Using Vision Transformer

이 광 엽*, 이 지 원*, 박 태 룡*★

Kwang-Yeob Lee*, Ji-Won Lee*, Tae-Ryong Park*★

## Abstract

This paper proposes a neural network applying fine-tuning as a way to improve the performance of Video Classification based on Vision Transformer. Recently, the need for real-time video image analysis based on deep learning has emerged. Due to the characteristics of the existing CNN model used in Image Classification, it is difficult to analyze the association of consecutive frames. We want to find and solve the optimal model by comparing and analyzing the Vision Transformer and Non-local neural network models with the Attention mechanism. In addition, we propose an optimal fine-tuning neural network model by applying various methods of fine-tuning as a transfer learning method. The experiment trained the model with the UCF101 dataset and then verified the performance of the model by applying a transfer learning method to the UTA-RLDD dataset.

## 요 약

본 논문은 Vision Transformer를 기반으로 하는 Video Classification의 성능을 개선하는 방법으로 fine-tuning를 적용한 신경망을 제안한다. 최근 딥러닝 기반 실시간 비디오 영상 분석의 필요성이 대두되고 있다. Image Classification에 사용되는 기존 CNN 모델의 특징상 연속된 Frame에 대한 연관성을 분석하기 어렵다는 단점이 있다. 이와 같은 문제를 Attention 메커니즘이 적용된 Vistion Transformer와 Non-local 신경망 모델을 비교 분석하여 최적의 모델을 찾아 해결하고자 한다. 또한, 전이 학습 방법으로 fine-tuning의 다양한 방법을 적용하여 최적의 fine-tuning 신경망 모델을 제안한다. 실험은 UCF101 데이터셋으로 모델을 학습시킨 후, UTA-RLDD 데이터셋에 전이 학습 방법을 적용하여 모델의 성능을 검증하였다.

*Key words : Fine-tuning, Transfer Learning, Video Classification, Vision Transformer, Non-Local, Attention*

* Dept. of Computer Eng., Seokyeong University
★ Corresponding author
E-mail : kylee@skuniv.ac.kr, Tel: +82-2-940-7745
※Acknowledgment
This work was supported by Seokyeong University in 2022 and by Seokyeong University in 2023.
Manuscript received Sep. 18, 2023; revised Sep. 24, 2023; accepted Sep. 26, 2023.


## Ⅰ. 서론

최근 높은 정확도와 빠른 추론 성능을 동시에 필요로 하는 Video Classification에 관한 연구의 중요성이 대두되고 있다. 딥러닝 기반의 CCTV영상에서 이상 행동 감지와 같은 Human Action Recognition, YOLO 모델[1]을 활용한 실시간 Video Object Detection 등이 그 예다.

Image Classification 연구는 과거부터 연구가 활발히 진행되어온 분야다. 2012년 Alexnet [2]이 ILSVRC (ImageNet Large Scale Visual Recognition Challenge)

대회에서 Top 5 error를 15.3%로 압도적인 성능을 보이면서 CNN(Convolutional Neural Network)에 많은 연구가 이루어져 왔으며 그 결과 Image Classification에서는 전이 학습[3] 방법 통해 생성된 pre-trained model을 활용하는 것이 효과적이다. 그러나 비디오 영상의 분류에는 비디오의 매 프레임을 CNN 모델의 입력으로 사용하기 때문에 높은 정확도와 빠른 예측 성능을 동시에 기대하기 어렵다는 단점이 존재한다. 또한, 연속된 프레임 간의 연관성을 고려하지 못하기 때문에 독립된 연산에 의한 속도 저하 문제가 존재한다. 다른 연구들에서는 이를 CNN 모델에서 추출한 feature maps를 RNN, LSTM 같은 시계열 모델의 입력으로 사용하여 해결하고자 했다. 또 다른 접근 방식으로, 기존 pre-trained된 2차원 필터를 3차원으로 팽창한 필터를 사용하는 3D CNN 구조를 활용하는 연구도 진행되고 있다.

본 논문에서는 비디오 영상 분류의 정확도를 높이고 신경망 개발을 용이하게 하기 위해 전이 학습 방법으로 fine-tuning을 제안한다. 특히, 자연어 처리와 같은 seq2seq 모델에서 높은 정확도를 가지는 Attention 메커니즘을 적용한 Vision Transformer[4]를 기반으로 fine-tuning을 진행하였다. Attention 메커니즘에 높은 성능을 보이는 Video Vision Transformer[5](이하 ViViT)와 Non-Local Neural Networks[6]를 활용한 Video Classification 성능을 비교 분석한다. 이후 더 높은 성능의 attention 메커니즘에 fine-tuning을 적용하여 전이 학습을 진행하면서 최적의 fine-tuning 구조를 제안한다. 실험은 두 모델을 UCF101 데이터셋[7]으로 학습시킨 후, UTA-RLDD 데이터셋[8]에 전이 학습 방법을 적용하여 모델의 성능을 검증하였다.

## Ⅱ. 본론

본론 1장에서 활용한 두 모델을 이해하는데 필요한 관련 연구를 소개한다. 2장에서는 모델의 입력으로 사용한 데이터셋, 3장에서는 학습 환경 및 실험 방법, 4장에서는 실험 결과, 5장에서는 모델의 성능을 검증한 결과를 제시한다.

### 1. 관련 연구

가. Video Vision Transformer

Vision Transformer(이하 ViT)[9]는 transformer 구조를 활용하여 최소한의 연산으로 2D 이미지를 처리한다는 특징을 가진다. ViT는 N개의 이미지 패치를 추출하여 linear projection을 수행한 후 1차원의 z 토큰들로 rasterization 한다. 인코더에 입력되는 토큰의 순서는 식 1과 같다.

$$z = [z_{cls}, E_{x1}, E_{x2}, \dots, E_{xn}] + p \tag{1}$$

z는 선택적으로 학습된 token이며, classification의 layer에서 마지막 representation으로 사용된다. transformer의 self-attention 연산은 입력의 순서가 바뀌어도 같은 결과를 도출한다는 특징이 있다. 이러한 문제점을 보완하기 위해 위치 임베딩 p를 추가한다.

이후 토큰들은 L개의 트랜스포머 레이어로 이루어진 인코더를 통과하게 된다. 각각의 레이어는 Multi-Headed Self-Attention(MSA), Layer Normalization(LN)[10], 그리고 MLP block으로 이루어져 있으며 계산식은 식 2, 3과 같다.

$$Y^l = MSA(LN(z^l)) + z^l \tag{2}$$

$$z^{l+1} = MLP(LN(y^l)) + y^l \tag{3}$$

마지막으로, 인코딩된 입력을 선형 분류기를 사용하여 분류한다. 이러한 구조 덕에 ViT는 크기가 다른 입력 토큰들에 대해서도 유동적인 연산이 가능하다는 장점을 가진다.

T x H x W x C 크기를 가지는 비디오를 Nt x Nh x Nw x d 크기의 z 토큰의 시퀀스로 매핑하는 방법엔 2가지가 있다. 첫 번째는 분리한 token을 한 번에 transformer 인코더의 입력으로 사용하는 Unifrom frame sampling이다. 두 번째는 비디오를 프레임 단위로 묶어 패치 단위로 나눠서 결과적으로 spatio-temporal 정보를 혼합하여 사용하는 Tubelet Embedding이다.

인코더에 사용하는 구조엔 Spatio-temporal attention, Factorised encoder, Factorised self-attention, Factorised dot-product attention이 있다. 단순히 spatio-temporal 토큰을 인코더에 넣은 Spatio-temporal attention은 마지막까지 두 정보를 독립적으로 처리한 후 최종 단계에서 융합하는 “late fusion” 기법이다. 나머지 세 모델은 Spatial과 Temporal encoder를 분리하여 더 적은 연산량으로 앞의 모델과 유사한 성능을 보이는 특징을 가진다.

나. Non-Local Neural Networks

Non-local은 이미지 내의 모든 픽셀의 가중 평균을

계산하는 필터링 알고리즘이다[11]. long-range dependency를 캡처하는 것이 dnn에서 중요한 문제이다. Non-Local은 비디오에서 픽셀과 거리 차이에 상관없이 long-range dependency를 capture할 수 있다는 장점이 있다. 또한, 비교적 적은 레이어 개수로 인한 연산 효율성과 동시에 좋은 성능을 가진다. 이와 더불어 다양한 크기의 입력 크기를 받을 수 있으며, 다른 연산과 쉽게 결합할 수 있다는 이점이 있다. 이러한 특징은 픽셀과 시간 거리에 대한 상호작용이 존재하는 Video Classification에 유리하게 작용한다.

연구에서의 Non-Local 연산은 식 4와 같다.

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j) \tag{4}$$

i는 출력 포지션, j는 모든 입력 포지션을 의미한다. x는 이미지, 시퀀스, 비디오와 같은 입력 신호를, y는 x와 같은 사이즈를 가지는 출력 신호를 의미한다. 함수 f는 i와 모든 j에 대한 관계를 계산하고, 함수 g는 j에 대한 입력 신호에 대한 표현을 계산한다. 최종적으로, 연산의 결과는 C(x)에 의해 정규화된다.

논문에서는 단순성을 위해 함수 g를 식 5와 같이 정의하였다.

$$g(x_j) = W_g x_j \tag{5}$$

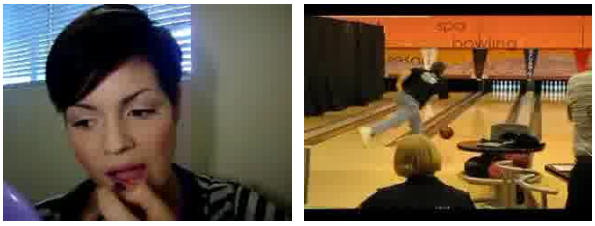
함수 f에 대하여 Gaussian, Embedded Gausian, Dot product, Concatenation 수식으로 성능을 검증해 봤지만, top-1과 top-5 정확도에서 모두 큰 차이가 없는 연구 결과를 관찰할 수 있다.

$$z_i = W_z y_i + x_i \tag{6}$$

최종적으로 Non-local Block 수식에 대한 정의는 식 6과 같으며, $x_i$는 residual connection을 의미한다. 이를 Backbone이 RestNet인 2D ConvNet과 Inflated 3D ConvNet에 삽입한 형태가 바로 Non-local Neural Networks이다.

## 2. 데이터셋

본 논문에서는 두 모델을 학습시키기 위해 UCF101 Action Recognition 데이터셋을 사용하였다. 원본 데이터셋은 101개 클래스를 가지는 13,320개의 비디오로 이루어져 있다. 영상들은 다양한 카메라 움직임, 객체의 외형 및 동작, 각도, 배경 및 조명 조건 아래서 촬영되었다. 클래스의 예로 립스틱 바르기, 골프, 자전거 타기, 볼링 등이 있으며, 비디오의 크기는 320×240이다. 반면, 비디오의 길이는 고정되어 있지 않기 때문에 150 Frame 이상의 비디오를 선별하여 모델의 입력으로 사용하였다. 최종적으로 99개의 클래스를 갖는 7,575개의 비디오를 학습에 사용하였으며, 이를 4:1 비율로 나눠 각각 6,060 / 1,515개의 데이터를 train과 validation에 사용하였다.



(a) ApplyLipstick (b) Bowling

Fig. 1. sample images of UCF101 Dataset.
그림 1. UCF101 데이터셋의 샘플 이미지

또한, 성능 검증을 위해 졸음 정도에 따라 분류한 UTA-RLDD 데이터셋을 사용하였다. 원본 데이터셋은 18세 이상의 60명 참가자가 그들의 휴대폰 또는 웹캠을 이용하여 직접 녹화한 영상으로 이루어져 있다. 모든 동영상은 두 눈이 보일 수 있는 각도로 촬영되었으며, 실제 차량 내의 휴대폰 거치대와 유사한 거리에 촬영 기기를 두도록 설계되었다. 데이터셋은 Alert(일반 상태), Drowsy(졸음 상태), 그리고 Low Vigilant(중간 상태) 총 3개의 클래스를 가진다. 본 논문에서는 "졸음" 클래스 영상을

Fig. 2. sample images of UTA-RLDD Dataset.
그림 2. UTA-RLDD 데이터셋의 샘플 이미지

필요로 하기 때문에 Drowsy 상태 비디오만을 선별하였다. 또한, UCF101 데이터셋과 같은 크기의 비디오를 만들기 위해 가로 비율이 더 큰 영상들을 골라 동일하게 320×240 크기로 resize한 후 사용하였다.

### 3. 학습 환경 및 실험 방법

본 실험은 Quadro RTX 8000 환경에서 Python 3.7.4, PyTorch 1.13.1, CUDA 11.7 버전을 사용하였다.

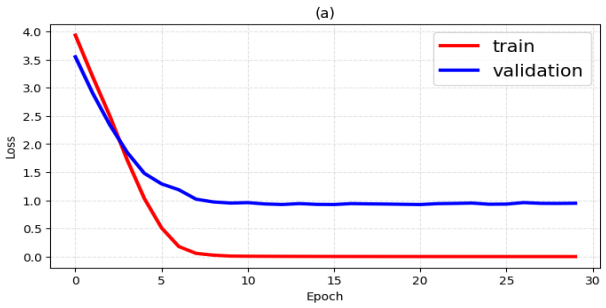
ViViT의 image path size는 40, frame patch size는 10, spatial depth는 6, temporal depth는 6, head 개수는 8, dimension은 1024, mlp dimension은 2048로 설정하였다.

Non-Local Neural Networks의 Backbone으로 ResNet-50을, pairwise function은 Embedded Gaussian으로 설정하여 학습을 진행하였다.

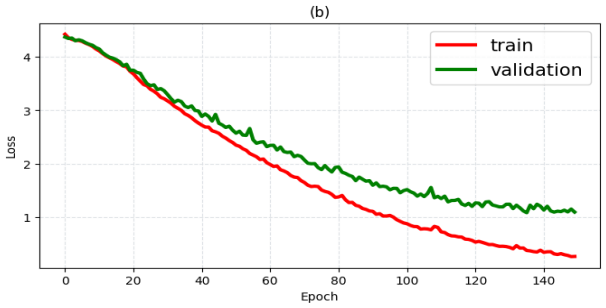
Table 1. Comparison of learning process.
표 1. 학습 과정 비교

| | ViViT | Non-Local |
|---|---|---|
| epoch | **30** | 150 |
| 1epoch(min) | **21~23** | 12~13 |
| parameters | **125,689,699** | 25,812,643 |
| validation accuracy | **77.81%** | 71.48% |

### 4. 실험 결과




(a) ViViT




(b) Non-Local Neural Networks

Fig. 3. Train / Validation Loss Graphs for Two Models.
그림 3. 두 모델의 Train / Validation Loss 그래프

두 모델의 loss 그래프는 그림 3과 같다. ViViT의 경우 1 epoch 당 소요 시간이 긴 대신, 적은 epoch로도 높은 정확도를 갖는다는 사실을 확인할 수 있다.

## Ⅲ. Fine-tuning Video Classification

높은 성능의 attention 메커니즘을 비교한 앞선 실험 결과 ViViT가 video classification에서 좋은 성능이 나왔기 때문에 본 논문에서는 ViViT를 대상으로 다양한 구조의 Fine-tuning 신경망을 만들고 전이 학습을 통하여 가장 좋은 fine-tuning을 결정한다.

### 1. Fine-tuning 구조 설계

Fine-tuning은 일반적으로 pre-trained network의 마지막 분류 단계를 대상으로 진행된다. 주로 FC(Fully Connected) Layer를 제거하고 새로운 분류가 추가된 FC로 수정된 network가 만들어 진다. 즉, target training set에서 적절하게 조정된 gradient descent가 추가되어 fine-tuning 된다.

본 논문의 다음 실험에서도 알 수 있듯이 FC층만 교체하여 새로운 분류를 만드는 것은 좋은 성능을 갖지 못하기 때문에 Fine-tuning의 몇 가지 구조를 비교 연구하였다. Fine-tuning은 이미 학습된 안정적인 가중치를 크게 변경하면 기존 네트워크에서 얻어진 학습 손실이 발생하기 때문에 미세 조정으로 접근하게 된다.

Fine-tuning방법은 기존의 학습 데이터와 새롭게 학습하려는 데이터의 크기와 유사성에 따라 다음과 같이 나누어 볼 수 있다[12].

case ① 데이터셋의 크기가 크고 유사성이 작을 때는 사전 학습모델의 구조만 사용하고 전체 새로 학습한다.

case ② 데이터셋의 크기가 크고 유사성이 높을 때는 convolutional layer 일부분을 고정하고 나머지 계층과 classifier를 새로 학습한다.

case ③ 데이터셋의 크기가 작고 유사성이 작을 때는 convolutional layer의 층수와 freezing layer의 층수를 적절하게 조절하는 것이 어렵다. 새롭게 많은 계층을 학습하면 과적합이, 적은 계층인 경우 학습이 어렵게 된다.

case ④ 데이터셋의 크기가 작고 유사성이 높을 때는 classifier 부분만을 변형하여 학습한다.

본 논문에서는 기존 학습 데이터셋으로 UCF101 Action Recognition를 사용하였으며 새로운 학습 데이터셋은 UTA-RLDD 이다. UTA-RLDD 는 비교적 데이터셋이

적고 유사성도 작은 경우로 case ③에 해당하여 새롭게 학습할 부분을 정하기에 어려운 경우가 된다.

### 2. Transformer 특징에 기반한 fine-tuning

본 논문에서는 위의 case ③에 해당하는 fine-tuning을 위한 전이 학습 문제를 해결하기 위해 전이 학습 타겟 모델이 되는 ViViT의 transformer 구조의 특징에 착안하였다. Transformer는 비디오 프레임 이미지에서 CNN 계층을 이용한 객체 특징 추출이 진행된 후 특징 맵은 토큰화되고 시계열에 해당하는 계층에서 어텐션을 거쳐 MLP 층에서 분류가 된다.

위 과정을 크게 구분하여 보면 특징 추출을 위한 spatial layer들과 시계열의 temporal layer로 신경망이 구성되는 것을 알 수 있으며 이 경계에서 학습의 변화가 크다. Non-local에서는 temporal layer와 spatial layer를 확연한 특징으로 구분할 수 없으나 ViViT는 spatial layer 30개, temporal layer 30개로 확연한 특징을 보인다. 본 논문에서는 ViViT의 spatial layer까지 학습 가중치를 사용하고 temporal layer부터 새로운 학습을 통한 fine-tuning을 진행하였다.

Table 2. Comparison of Fine-tuning Performance.
표 2. Fine-tuning 성능 비교

| case | val_acc (%) | Drowsy acc (%) | Fine-tuning Learning Method |
|---|---|---|---|
| **case1** | 94.91 | 98.6 | All Layers |
| **case2** | 94.84 | 98.6 | spatial_transformer + temporal_transformer + FC) |
| **case3** | **95.43** | **100*** | **temporal_transformer + FC** |
| **case4** | 94.58 | 92.0 | FC |

* Because the average number of data per lable in the ucf101 dataset is 76.5, we also experimented with 75 drawsy data, resulting in 100%.

### 3. 제안하는 Fine-tuning 성능 분석

ViViT에 기반한 Fine-tuning 학습 방법은 위에서 설명한 네 가지 경우로 나누었다. 표 2에서 나타내었듯이 첫 번째는 ViViT를 UTA-RLDD 데이터셋으로 전체 학습하였고 두 번째는 특징 추출 CNN층 일부를 포함한 나머지 모두를 학습하였다. 세 번째는 본 논문에서 제안하는 방법으로 spatial 층과 temporal 층을 분리한 방법으로 학습 검증 정확도(val_acc)와 새로운 분류인 drowsy 분류 정확도(Drowsy acc)가 가장 높았다. 네 번째는 FC만 drowsy 분류가 포함되도록 교체한 것으로 가장 낮은 정확도를 보였다.

## Ⅳ. 결론

본 논문에서는 기존 Video Classification의 문제점을 해결하고 높은 정확도와 빠른 예측 속도를 가지는 모델을 찾기 위해 Video Vision Transformer와 Non-Local Neural Networks의 성능을 비교 분석하였다. 그 결과 정확도가 높으면서 Transformer의 구조에서 statial layer와 temporal layer를 분리하여 fine-tuning이 가능한 ViViT를 대상으로 전이 학습을 위한 fine-tuning 구조를 제안하였다. 제안된 구조는 ViViT에 drowsy 데이터셋으로 전이 학습하고 검증하였을 때 기존보다 높은 정확도를 보였다. 향후 fine-tuning 파라메터의 다양한 분석을 통하여 정확도 향상을 위한 추가 연구가 수행될 예정이다.

## References

[1] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” *IEEE 2016,* pp. 779-788, 2016. DOI: 10.48550/arXiv.1506.02640

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” *Communications of the ACM,* Vol.60, No.6, pp.84-90, 2017.

[3] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, “A Comprehensive Survey on Transfer Learning,” *Proceedings of the IEEE,* Vol.109, No.1, pp.43-76, 2021. DOI: 10.48550/arXiv.1911.02685

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, “Attention Is All You Need,” *NIPS 2017,* pp.6000-6010, 2017. DOI: 10.48550/arXiv.1706.03762

[5] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, Cordelia Schmid, “ViViT: A Video Vision Transformer,” *ICCV 2021*, pp.6836-6846, 2021. DOI: 10.48550/arXiv.2103.15691

[6] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He, “Non-local Neural Networks,” *IEEE,*

pp.7794-7803, 2017.
[7] Khurram Soomro, Amir Roshan Zamir and Mubarak Shah, "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild," CRCV-TR-12-01, 2012. DOI: 10.48550/arXiv.1212.0402
[8] Reza Ghoddoosian, Marnim Galib, Vassilis Athitsos, "A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection," *CVPRW 2019*, pp.178-187, 2019.
DOI: 10.48550/arXiv.1904.07312
[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," https://arxiv.org/abs/2010.11929
DOI: 10.48550/arXiv.2010.11929
[10] Jimmy Lei Ba, Jamie Ryan Kiros, Geoffrey E. Hinton, "Layer Normalization," https://arxiv.org/abs/1607.06450
[11] A. Buades, B. Coll and J.-M. Morel, "A non-local algorithm for image denoising," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),* Vol.2, pp.60-65, 2005. DOI: 10.1109/CVPR.2005.38.
[12] https://newindow.tistory.com/254

## BIOGRAPHY

**Kwang-Yeob Lee** (Life Member)

1985 : BS degree in Electronics Engineering, Sogang University
1987 : MS degree in Electronics Engineering, Yonsei University.
1994 : PhD degree in Electronics Engineering, Yonsei University.
1989~1995.2 : Senior Researcher, Hyundai Electronics Inc.
1995.3~present : Professor, Dept. of Computer Engineering, Seokyeong University

**Ji-Won Lee** (Member)

2024 : BS degree candidate in Computer Engineering, Seokyeong University.

**Tae-Ryong Park** (Member)

1985 : Hangyang University, Dept. of Mathmatics(BS)
1987 : Hangyang University, Dept. of Mathmatics(MS)
1995 : Hangyang University, Dept. of Mathmatics(Ph.D)
1994~ : Seokyeong Univeristy, Dept. of Computer Engineering, Professor
〈Research interests〉 Crypto Algorithm, Computer Security, Computer Arithmetic, Recongnition Algorithm, Machine Learning