Journal of Korea Multimedia Society Vol. 24, No. 10, October 2021(pp. 1380-1390)
https://doi.org/10.9717/kmms.2021.24.10.1380

비디오 얼굴인식을 위한 다중 손실 함수 기반 어텐션 심층신경망 학습 제안

김경태[†], 유원상[††], 최재영[†††]

Attention Deep Neural Networks Learning based on Multiple Loss functions for Video Face Recognition

Kyeong Tae Kim[†], Wonsang You[††], Jae Young Choi[†††]

ABSTRACT

The video face recognition (FR) is one of the most popular researches in the field of computer vision due to a variety of applications. In particular, research using the attention mechanism is being actively conducted. In video face recognition, attention represents where to focus on by using the input value of the whole or a specific region, or which frame to focus on when there are many frames. In this paper, we propose a novel attention based deep learning method. Main novelties of our method are (1) the use of combining two loss functions, namely weighted Softmax loss function and a Triplet loss function and (2) the feasibility of end-to-end learning which includes the feature embedding network and attention weight computation. The feature embedding network has a positive effect on the attention weight computation by using combined loss function and end-to-end learning. To demonstrate the effectiveness of our proposed method, extensive and comparative experiments have been carried out to evaluate our method on IJB-A dataset with their standard evaluation protocols. Our proposed method represented better or comparable recognition rate compared to other state-of-the-art video FR methods.

Key words: Attention Mechanism, Video Face Recognition, Deep Convolutional Neural Networks, Combined loss, End-to-end learning

1. 서 론

얼굴 인식은 어떤 인물인지 판별하는 기술로 패턴 인식 및 컴퓨터 비전 분야에서 오랫동안 연구되어 온 분야이며 최근 몇 년 동안 점점 더 많은 관심을 받고 있다. 특히 얼굴 검증(Face Verification) 및 얼굴 식별(Face Identification)은 금융 신원 인증 및 비디오 감시와 같은 컴퓨터 비전 응용에 널리 사용되고

※ Corresponding Author : Jae Young Choi, Address: (17035) 81 Oedae-ro, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, Korea, TEL : +82-31-330-4906, FAX : +82-31-330-4906, E-mail : jychoi@hufs.ac.kr
Receipt date : Aug. 12, 2021, Revision date : Sep. 30, 2021
Approval date : Oct. 12, 2021
[†] Division of Computer and Electronic Systems Engineering, Hankuk University of Foreign Studies (E-mail : rudxo9251@gmail.com)
[††] Dept. of Information and Communications Engineering, Sun Moon University (E-mail : wyou@sunmoon.ac.kr)
[†††] Division of Computer and Electronic Systems Engineering, Hankuk University of Foreign Studies

※ This research was supported by Hankuk University of Foreign Studies Research Fund.
※ This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2018R1D1A1A09082615.
※ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1A2C1092322).
※ This research was supported by the "Development of technology using analysis of ocean satellite images (20210046)" funded by the Korea Institute of Marine Science & Technology Promotion(KIMST)

있으며 현재까지 지속적으로 연구되고 있다. 대용량 훈련 데이터셋, 심층합성곱 신경망 및 최적화된 손실 함수를 기반으로 구축된 최신 방법은 높은 성능을 달성하지만 큰 포즈 변화, 중첩 또는 조명으로 인한 얼굴의 모양 변화로 인해 올바른 인식이 어렵다. 또한 실제 응용에서는 단일 얼굴 영상이 아닌 비디오를 처리해야 된다. 단일 영상 기반 얼굴인식에 비해 비디오(시퀀스) 기반 얼굴인식에서 인물에 대한 더 많은 정보를 활용하기 때문에 서로 보완적일 수 있고 서로 다른 포즈의 이미지와 같이 단일 이미지보다 더 많은 정보를 제공한다. 그렇기 때문에 다양한 포즈, 중첩, 조명 등 얼굴인식 저해 요소들로부터 벗어나 얼굴 시퀀스를 자연스럽게 통합할 수 있다. 비디오 얼굴 인식의 핵심 문제는 비디오 얼굴의 적절한 표현을 구축하여 서로 다른 프레임의 정보를 효과적으로 통합하여 불필요한 정보는 버리면서 유익한 정보는 유지할 수 있도록 하는 것이다.

하나의 단순한 접근 방식은 얼굴 시퀀스를 프레임(Frame) 단위로 심층 합성곱 신경망에 입력하여 추출한 특징을 집합으로 비디오 얼굴을 표현한다[1,2,3,4,5,6]. 이러한 표현은 모든 프레임에 걸쳐 정보를 포괄적으로 유지한다. 그러나 두 얼굴 시퀀스를 비교하려면 두 얼굴 시퀀스 사이의 모든 프레임 쌍에서 비교를 해야 된다. n을 비디오 프레임의 평균 수 라고 가정하면 계산 복잡도는 프레임 쌍 매칭 작업 당 $O(n^2)$이기 때문에 대용량 인식에 바람직하지 않다. 게다가, 이러한 집합 기반 표현은 비디오 얼굴 시퀀스 당 $O(n)$의 공간 복잡성을 발생시켜 많은 메모리 용량을 요구한다. 또 다른 접근 방식은 얼굴 시퀀스를 프레임 단위로 심층 합성곱 신경망에 입력하여 추출한 모든 특징을 최대/평균 풀링(Pooling)을 하여 특징을 통합하여 표현한다. 그러나 최대/평균 풀링의 문제점은 시퀀스의 일부 영상이 인식에 적합하지 않을 수 있다는 것이다. 비디오에서 얼굴을 샘플링하다보면 영상 흔들림 혹은 흐림으로 인해 노이즈가 발생된 영상이 시퀀스를 구성할 수 있다. 만약 노이즈가 발생된 영상이 시퀀스에 포함되어 있어도 동일하게 처리가 되고, 최대/평균 풀링이 모든 영상의 특징을 통합하는 데 사용되는 경우 노이즈가 심한 이미지는 올바르지 못한 최종 표현을 만들어내 인식 성능을 저하시킨다.

최근 비디오 얼굴인식 방법에서 대표적인 기술인 어텐션 기법(Attention mechanism)을 활용한 연구가 활발히 진행되고 있다. 영상 인식에서 어텐션 기법은 전체 또는 특정 영역의 입력 값을 반영해서 그 중 어떤 부분을 집중해야 하는지를 나타내준다. 또한 많은 프레임이 존재할 때 어느 프레임에 집중해야 되는지를 나타내준다. 어텐션 기법을 적용한 최신 기술 중 하나인 Neural Aggregation Network(NAN)[7]은 여러 개의 프레임으로부터 단일 특징을 추출하는 방법을 제안하였다. 제안된 NAN는 다음과 같은 모듈로 구성된다. 모든 프레임을 심층 합성곱 신경망(Deep Convolutional Neural Network)을 거쳐 각각의 특징을 추출한 하는 특징 임베딩(Feature Embedding)모듈과 모든 특징들을 어텐션 블록을 이용하여 하나의 특징 벡터를 생성하는 통합(Aggregation)모듈로 이루어져 있다. 여기서 통합 모듈의 어텐션 블록은 각 프레임의 특징 벡터에 가중치를 적용하여 하나의 특징 백터를 생성한다. 이때 상대적으로 좋은 품질의 프레임은 높은 가중치를 생성하고, 반대로 품질이 안 좋은 프레임은 낮은 가중치를 생성하는 것을 보여주었다. 하지만 NAN은 End-to-end 방식으로 훈련되는 것이 아닌 각각의 모듈을 따로 훈련을 해야 되는 번거로움이 있으며 별로로 훈련시키기 때문에 특징 임베딩 모듈과 통합 모듈이 서로 보완적으로 훈련되지 않는다.

각 시퀀스의 프레임 혹은 세트의 이미지는 동일한 인식대상자에 속한다. 시퀀스의 프레임 혹은 세트의 이미지는 상호 보완적일 수 있으며, 더 높은 정확도로 이어진다. 그러나 각 샘플의 품질을 보장 할 수 없으며, 품질이 좋지 않은 샘플은 성능저하를 일으킨다. 이러한 문제를 해결하기 위해 QAN(quality aware network)[8]를 제안하였다. 네트워크에는 2개의 모듈이 있으며, 첫 번째 모듈은 각 샘플에 대한 외형 특징 임베딩(Appearance feature embedding)을 추출하며, 두 번째 분기(Branch)는 각 샘플의 품질 점수를 예측한다. 이 후 세트에 잇는 모든 샘플의 특징 및 품질 점수를 통합하여 최종 특징 임베딩을 생성한다. 상대적으로 좋은 품질을 갖는 영상은 높은 품질 점수를 갖게 되고, 화질, 중첩, 조명, 포즈 등 다양한 변화에 의해 열화된 영상에 대해서는 낮은 품질 점수가 할당되는 것을 실험적으로 보였다. 하지만 QAN는 영상의 품질 우수한지 저조한지만 고려하며 얼굴 영상의 세밀한 부분까지는 고려되지 않는다.

본 논문에서는 비디오 얼굴 인식을 위한 어텐션 기법 중 하나인 NAN을 활용하여 가중치가 적용된 Softmax 손실 함수와 Triplet 손실 함수를 결합한 종단학습(End-to-end learning) 방법을 제안한다. 제안 방법은 기존 관련 방법들과 비교하여 다음과 같은 독창성을 갖는다.

- 기존 NAN은 각 프레임에 대해 공통된 심층 신경망 모델을 이용하여 특징을 추출한 후 다중의 특징 벡터를 구성한 뒤, 이 특징들을 어텐션(Attention) 블록을 이용하여 하나의 융합된 특징 벡터를 생성하였다. [7]의 저자는 훈련 시 어텐션 블록에 훈련이 더 집중 될 수 있도록 End-to-end 학습이 아닌 별도의 모델로 설계하였다. 제안한 방법은 End-to-end 학습으로 모델을 설계하였으며, 이는 결합된 손실 함수를 활용하여 학습하기 때문에 특징 추출 네트워크가 NAN에 영향을 오히려 긍정적인 영향을 미치게 된다.

제안한 방법은 단일 얼굴 및 비디오 얼굴 인식 데이터셋인 IJB-A[9] 데이터셋에 대해 기존 전통적인 결합방법과 최신 연구들 간의 성능 비교를 통해 실험을 진행하였다. 전통적인 결합방법으로는 대표적으로 각 프레임의 특징을 추출하여 특징들을 평균(Average) 혹은 특징 연결(Concatenation), 점수 융합(Score fusion)이 있다. 특징 평균은 CNN(Convolutional Neural Network)를 통해 추출된 특징들을 평균하여 동일한 크기의 특징을 다시 완전 연결 계층을 통해 분류하는 방법이며, 특징 연결은 CNN를 통해 추출된 특징들을 1차원 벡터로 연결한 특징을 다시 완전 연결 계층을 통해 분류하는 방법이다. 점수 융합(Score fusion)은 CNN를 통해 나온 최종 분류 점수를 평균하여 분류하는 방법이다. 실험 결과 제안한 방법이 전통적인 결합방법보다 우수한 성능을 달성하였다. 또한 다른 최신 연구[7,14-22]와 비교하였을 때 비슷하거나 우수한 성능을 달성하였다.

2. 제안한 방법

Fig. 1은 제안한 방법의 전체도(Overall framework)이다. 제안한 방법은 입력 얼굴 비디오 시퀀스를 완전계층연결 계층이 수정된 VGG Face 신경망[2]에 입력하여 심층 특징 정보를 추출한다. 이후 각 프레임에 대한 심층 특징 정보를 모두 연결한 뒤 어텐션 신경망인 Neural Aggregation Network (NAN)

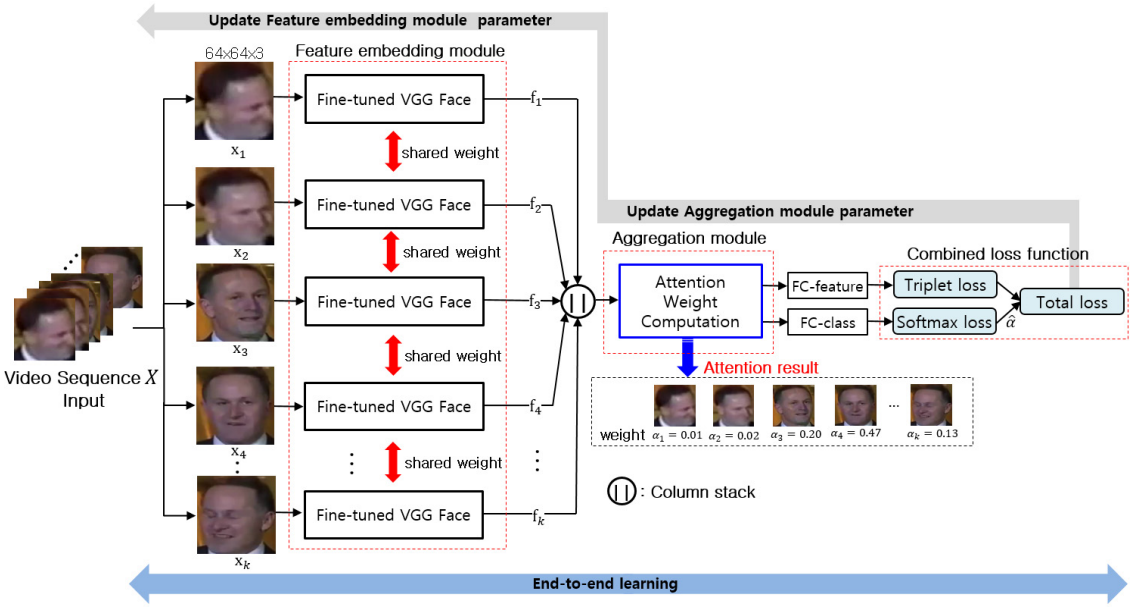

Fig. 1. Overall framework for face recognition with Attention Network using combined triplet and softmax losses. Each frame extracts each feature through a feature embedding module. The extracted features are stacked in the column direction and input to the aggregation module. The aggregation module calculates a weight for each feature. Each calculated weight is combined with each feature to extract a single feature. This feature is learned end-to-end with a combined loss function.

에 입력되어 각 프레임에 중요도를 나타내는 가중치 및 가중치가 적용되어 융합된 심층 특징을 출력한다. 이 특징들은 Triplet 손실 함수와 가중치가 적용된 Softmax 손실 함수에 의해 제안된 신경망이 End-to-end 방식으로 훈련된다.

2.1 특징 추출 모듈(Feature embedding module)

특징 추출 신경망은 사전 학습된 신경망인 VGG-Face를 사용하여 파인 튜닝(Fine-tuning)을 통해 특징 정보를 추출하였다. VGG-Face에서는 인터넷 검색을 통해 직접 만든 대용량의 얼굴인식을 위한 데이터셋인 VGG 얼굴 데이터셋을 공개하고, 이 데이터를 이용하여 Table 1과 같이 15개의 합성곱 계층(Convolution Layer)으로 구성된 딥 네트워크 구조를 학습시켰다. VGG에서는 VGG-Face 학습 모델을 제공할 뿐만 아니라, ImageNet 영상인식 챌린지(Challenge)에서의 VGG 구조와 마찬가지로 상대적으로 간단한 3×3 컨볼루션 필터를 이용하여 학습시킴으로써 VGG-Face는 LFW 데이터셋에 대해 Deep Face보다 약 1% 정도 개선된 98.95% 성능을 달성하였다.

본 연구에서는 VGG-Face의 합성곱 계층은 그대로 사용하였으며, 완전 연결 계층의 차원을 수정한 상태에서 심층 특징을 추출하여 어텐션 신경망인 Neural Aggressive Network(NAN)에 입력으로 활용한다. 연결된 소프트맥스 계층은 이전 계층인 완전 연결 계층 노드 (fully-connected layer node)의 특징 맵 h와 해당 가중치 W를 활용하여 비디오 얼굴 시퀀스의 특징 $\mathbf{f}^{d}$는 다음과 같이 표현된다.

$$\mathbf{f}^{d}=\sum_{j=1}^{H} h^{(j)} W^{(j,d)}+b \tag{1}$$

여기서 H는 완전 연결 계층의 노드의 개수이고, d는 특징정보의 차원이며, b는 바이어스이다. 더 나은 가독성을 위해 수식 (1)에서 인덱스 d는 생략한다.

2.2 집계 모듈(Aggregation module)

얼굴 비디오 데이터 $(X^{i},y_{i})_{i=1}^{n}$라고 할 때 n은 총 비디오 얼굴 시퀀스이며 X^{i}은 얼굴 비디오 시퀀스를 나타낸다. $X^{i}=\mathrm{x}_1^{\mathrm{i}},\mathrm{x}_2^{\mathrm{i}},\mathrm{x}_3^{\mathrm{i}},...,\mathrm{x}_{\mathrm{k}}^{\mathrm{i}}$에서 $\mathrm{x}_{\mathrm{k}}^{\mathrm{i}},\mathrm{k}=1,...,\mathrm{K}$는 얼굴 비디오의 k-번째 프레임(frame)을 나타내며, y_i는 해당 인식대상자의 ID를 의미한다. 각 얼굴 프레임 $\mathrm{x}_{\mathrm{k}}^{\mathrm{i}}$은 특징 추출 모듈에 의해 특징 $\mathbf{f}_k^{i}$으로 추출된다. 더 나은 가독성을 위해 인덱스 i는 생략한다. NAN는 하나의 인식 대상자에 대한 많은 프레임들에 대한 특징 정보에 대해 가중치 $\{a_k\}_{k=1}^{K}$를 구하여 특징 정보와 다음과 같이 결합한다.

$$\mathbf{r}=\sum_{k} a_k \mathbf{f}_{\mathrm{k}} \tag{2}$$

이때 가중치는 특징 추출 모듈에서 추출한 단일 얼굴 이미지 특징과 동일한 크기를 갖는다.

2.3 어텐션 블록(Attention blocks)

Fig. 2는 제안된 방법의 어텐션 블록이며 집계 모

Table 1. VGG-Face Architecture. It consists of a total of 15 convolution layers and 3 fully connected layers. In this experiment, 15 convolutional layers were used. And only the fully connected layer was modified and used for feature extraction.

layer	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
type	input	conv	relu	conv	relu	mpool	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv
name	-	conv1_1	relu1_1	conv1_2	relu1_2	pool1	conv2_1	relu2_1	conv2_2	relu2_2	pool2	conv3_1	relu3_1	conv3_2	relu3_2	conv3_3	relu3_3	pool3	conv4_1
support	-	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3
filt dim	-	3	-	64	-	-	64	-	128	-	-	128	-	256	-	256	-	-	256
num filts	-	64	-	64	-	-	128	-	128	-	-	256	-	256	-	256	-	-	512
stride	-	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1
pad	-	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1

layer	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
type	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	relu	mpool	conv	relu	conv	relu	conv	softmax
name	relu4_1	conv4_2	relu4_2	conv4_3	relu4_3	pool4	conv5_1	relu5_1	conv5_2	relu5_2	conv5_3	relu5_3	pool5	fc6	relu6	fc7	relu7	fc8	porb.
support	1	3	1	3	1	2	3	1	3	1	3	1	2	7	1	1	1	1	1
filt dim	-	512	-	512	-	-	512	-	512	-	512	-	-	512	-	4096	-	4096	-
num filts	-	512	-	512	-	-	512	-	512	-	512	-	-	4096	-	4096	-	2622	-
stride	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	1	1	1
pad	0	1	0	1	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0

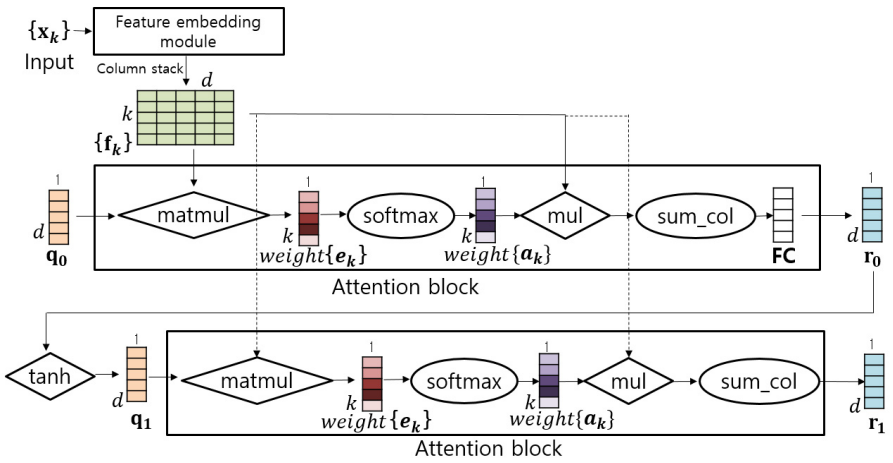

Fig. 2. All input face images $\{\mathbf{x}_k\}$ are processed by a feature embedding module with a deep CNN, yielding a set of feature vectors $\{\mathbf{f}_k\}$. These features are passed to the aggregation module, producing a single 512-dimensional vector $\mathbf{r}_1$ to represent the input faces images. This compact representation is used for recognition.

듈(Aggregation module)을 통해 나온 특징 정보들이 어텐션 블록에 입력된다. 어텐션 블록은 특징 임베딩 모듈에서 모든 특징 정보를 읽고 선형 가중치를 생성한다. 특징 정보 $\mathbf{f}_k$는 어텐션 블록이 커널 $\mathbf{q}$와 내적(dot product)을 통해 해당하는 의미 집합 $\{e_k\}$을 생성한다. 그런 다음 $\sum_k a_k = 1$인 양의 가중치 $\{a_k\}$를 생성하기 위해 softmax 연산자로 전달된다. 이 두 연산은 다음과 같은 수식으로 설명할 수 있다.

$$e_k = \mathbf{q}^{\mathrm{T}} \mathbf{f}_{\mathrm{k}} \tag{3}$$

$$a_k = \frac{\exp(e_k)}{\sum_j \exp(e_j)} \tag{4}$$

위의 수식을 보면 알 수 있듯이 어텐션 블록에서 모든 특징 정보는 각 특징정보의 해당하는 하나의 가중치를 선택한다는 것을 알 수 있다. 따라서 입력 $\{\mathbf{f}_k\}$의 수는 특징 정보 $\mathbf{f}_k$와 동일한 차원인 어텐션 블록의 출력인 집계(Aggregation) $\mathbf{r}$의 크기에 영향을 주지 않는다. 게다가, 집계 결과는 식에 따라 $\mathbf{f}_k$의 입력 순서에 불변하기 때문에 입력순서를 바꾸어도 집계 표현 $\mathbf{r}$에는 영향을 미치지 않는다. 더욱이, 어텐션 블록은 필터 커널 $\mathbf{q}$에 의해 변조되며, 이는 역전파 및 경사하강법을 통해 학습할 수 있다.

2.4 Triplet과 Softmax 손실 함수를 결합한 다중 손실 함수

Triplet 손실 함수는 기준 데이터 (Query)로부터 기준 데이터와 동일한 신원인 데이터(Positive example)는 인접하게, 기준 데이터와 동일한 신원이지 않은 데이터(Negative example)는 먼 공간에 떨어트려 적절한 분리 공간에 배치한다. 하지만 인식 대상자 간의 세밀한 차이를 학습하기 힘들다. 즉 유사하게 생긴 인식 대상자들 간의 분류 능력이 저하된다. 따라서 본 논문은 [10]를 참고하여 가중치가 적용된 softmax 손실 함수를 추가하여 ID간의 세밀한 차이를 학습할 수 있도록 Triplet 손실 함수와 softmax 손실 함수 결합한 손실 함수로 학습하며 다음과 같다.

$$L_{tri}(X^i) = \sum_{i=1}^{P} \sum_{(+)=1}^{K} \ln\left(1 + \exp\left(\max_{(+)=1\sim K} D(\mathbf{f}(x_a^i), \mathbf{f}(x_{(+)}^j)) - \min_{\substack{(-)=1\sim P,\\ j\neq i}} \sum_{(-)=1\sim K} D(\mathbf{f}(x_a^i), \mathbf{f}(x_{(-)}^j))\right)\right) \tag{5}$$

$$L_s(X) = -\frac{1}{M}\sum_{i=1}^{M} \log \frac{\exp(W_{y_i}^T \mathbf{f}(\mathrm{x}^{\mathrm{i}}) + \mathrm{b}_{\mathrm{y_i}})}{\sum_{j=1}^{C} \exp(W_j^T \mathbf{f}(\mathrm{x}^{\mathrm{i}}) + \mathrm{b}_{\mathrm{j}})} \tag{6}$$

수식 (5)은 Triplet 손실 함수로 여기서 $\sum_{i=1}^{P}\sum_{(+)=1}^{K}$는 모든 앵커(anchor)를 나타내며, 앵커는 배치(batch) 데이터 선택 시 기준이 되는 데이터를 의미한다. 그리고 $\max_{(+)=1\sim K} D(\mathbf{f}(\mathrm{x}_{\mathrm{a}}^{\mathrm{i}}), \mathbf{f}(x_{(+)}^j))$는 앵커를 기준으로 동일한 인식대상자중 식별이 어려운 샘플(hardest positive), $\min_{\substack{(-)=1\sim P,\\ j\neq i}} \sum_{(-)=1\sim K} D(\mathbf{f}(x_a^i), \mathbf{f}(x_{(-)}^j))$는 앵커를 기준으로 동일하지 않

은 인식대상자중 식별이 어려운 샘플(hardest negative)를 나타낸다. 여기서 a,(+),(−)는 각각 앵커(anchor), 앵커와 동일한 인식대상자의 샘플(positive), 앵커와 동일하지 않은 인식대상자의 샘플(negative)를 나타내며, P는 인식대상자의 수이고, 각각의 인식대상자 당 K개의 영상을 의미한다. 그리고 D는 두 특징 정보의 유클리드 거리를 나타내며, softplus 함수인 $\ln(1+\exp(\bullet))$을 사용하였다. 수식 (6)은 Softmax 손실 함수로 $y_i \in \{1,2,...,M\}$는 영상 x^i의 부류(class)를 나타내며, M은 총 부류의 개수이다. 그리고 W, b는 완전 연결 계층의 파라미터이다.

수식 (5)와 (6)을 결합을 하되, 부류를 판별할 수 있는 Softmax 손실 함수에 너무 편향되면 오히려 재인식 성능이 저하될 수 있기 때문에 Softmax 손실 함수에 가중치를 적용하여 결합하였고, 최종 결합 손실 함수는 수식 (7)과 같다.

$$L_{total}(X) = L_{tri}(X) + \hat{a} \times L_s(X) \tag{7}$$

여기서 $\hat{a}$는 결합 손실 함수의 가중치이며, 한쪽 손실 함수에 편향되지 않으면서 인식대상자 간의 세밀한 차이를 학습하도록 [10]를 참고하여 가중치를 결정하였다.

3. 실험 결과 및 고찰

3.1 실험 데이터셋

본 연구에서 제안한 방법은 얼굴 인식 실험 평가로 많이 사용되는 IJB-A 데이터셋[9]를 사용하여 평가하였다. IJB-A 데이터셋은 총 500명의 인식대상자에 대해 25,812장의 영상(프레임)으로 구성되어 있으며, 메타 데이터도 함께 제공한다. 메타 데이터는 얼굴의 위치 및 각도, 눈, 코 등의 위치와 실내/외, 나이, 피부 톤, 인식 대상자의 ID등 다양한 정보가 포함되어 있다. 본 연구에서는 얼굴의 위치 정보를 활용하여 얼굴 부위만 잘라내어 실험을 진행하였다. Fig. 3은 훈련 및 테스트에 사용한 IJB-A 데이터셋의 샘플 중 일부를 나타낸다. 재인식 평가방법은 IJB-A 표준 프로토콜[9]를 준수하여 실험을 진행하였다. IJB-A 표준 프로토콜은 1:1의 얼굴 검증(Face verification)과 1:N의 얼굴 식별(Face identification)을 구하게 되어져 있다. 1:1의 얼굴 검증은 입력으로 들어오는 두 개의 얼굴 영상이 동일 인물인지 판단하게 되며, 1:N의 얼굴 식별은 입력으로 들어오는 하나의 얼굴 영상이 사전에 등록된 N명의 인물 중 어떤 인물에 해당하는지를 판단하게 된다. 각 프로토콜마다 10개의 집합으로 구성되어있으며, 10개의 집합에 대해 평균과 표준편차를 구하여 성능을 도출하게 된다.

3.2 실험 및 구현환경

본 실험에서 모든 입력 이미지는 64×64로 크기를 조정하였고, 사전 훈련 모델인 VGG-Face[2]을 사용하였다. 그리고 미니 배치(mini-batch) 생성 시 K = 6, P = 5~10으로 샘플링(Sampling)하여 30~60개의 영상으로 구성하였다. 그리고 [11]를 참고하여 학습 속도 미세조절(learning rate decay) 방식을 채택하였으며 다음과 같다.

$$\epsilon(t) = \begin{cases} \epsilon_0 & \text{if } t \le t_0 \\ \epsilon_0 0.001^{\frac{t-t_0}{t_1-t_0}} & \text{if } t_0 \le t \le t_1 \end{cases} \tag{8}$$

여기서 초기 학습 속도(learning rate)는 $\epsilon_0 = 3e-4$이고 훈련 반복(iteration)은 $t_0 = 25000$, $t_1 = 50000$이며, t_1에 도달했을 때 훈련은 멈추게 된다. 그리고 훈련에 사용된 Adam optimizer [12]은 $\epsilon = 10^{-3}, \beta_1 = 0.9, \beta_2 = 0.999$로 설정하였다. 또한 결합된 손실 함수에서 가중치 $\hat{a}$는 [10]에 따라 0.5로 설정하였다. 실험 시 하드웨어는



Fig. 3. Example images used in our experiments on the preprocessed IJB-A dataset [9].

ASUS x299칩셋(메인보드), i9-10900X(CPU), 128GB RAM(메모리), RTX 2080ti(GPU)를 사용하였다.

3.3 실험 결과 및 전통적인 결합 방법들과의 성능비교

본 실험에서는 어텐션(Attention) 기법의 우수성을 보이기 위해 제안한 방법이 기존의 전통적인 결합 방법들과 비교하였다. 전통적인 결합 방법이란 각 프레임의 특징을 추출하여 특징들을 평균(Average) 혹은 특징 연결(Concatenation), 점수 융합(Score fusion) 하는 것을 나타낸다. 특징 평균은 CNN(VGG-Face)를 통해 추출된 특징들을 평균하여 동일한 크기의 특징을 다시 완전 연결 계층을 통해 분류하는 방법이며, 특징 연결은 CNN(VGG-Face)를 통해 추출된 특징들을 1차원 백터로 연결한 특징을 다시 완전 연결 계층을 통해 분류하는 방법이다. 단, 이 방법은 최대 5개의 프레임만 적용되도록 하였다. 점수 융합(Score fusion)은 CNN(VGG-Face)를 통해 나온 최종 분류 점수를 평균하여 분류하는 방법이다. Table 2는 앞서 언급한 전통적인 결합 방법과 제안한 방법의 인식 성능을 나타낸다. 실험 결과 전통적인 방법들은 제안한 방법보다 성능이 낮았으며, 다음과 같은 단점을 가지고 있다.

- CNN+Average 방법은 오 분류된 프레임의 특징이 올바르게 분류된 프레임의 특징에 악영향을 미치게 되어 Rank-1 기준 0.882±0.015으로 저조한 성능을 보였다.
- CNN+Concatenation 방법은 프레임의 수가 고정되어 있기 때문에 프레임이 많을 경우 모든 프레임을 사용하기 못하며 프레임 수가 적을 경우 성능에 악영향을 미친다. 또한 입력 순서에 따라 특징이 연결되는 순서에 민감하다. 실험결과를 보면 Rank-1 기준 0.758±0.049으로 전통적인 방법 중 가장 낮은 성능을 보였다.
- CNN+Score Fusion 방법은 오 분류된 프레임의 분류 점수가 올바르게 분류된 프레임의 분류 점수에 악영향을 미치게 되지만 CNN+Average 방법보다는 덜 민감하며 Rank-1 기준 0.884±0.018으로 다른 전통적인 결합방법보다 우수하였다.

따라서 전통적인 방법들은 1)입력 프레임 수가 고정되어야 하며, 2)입력 순서에 인식 성능이 민감하고, 3)오 분류된 영상 혹은 특징들에 의해 오류 전파가 심하다는 단점을 가지고 있다. 앞서 언급한 단점들은 제안한 어텐션 기법을 통해 어느정도 보완해서 성능을 개선시킨다. 제안한 어텐션 기법은 입력 개수에 상관없으며, 입력 순서에도 영향을 받지 않는다. 또한 인식 저해요소(포즈 변화, 중첩, 저해상도 등)가 있는 영상에 대해 어텐션 계산을 통해 가중치가 낮게 결정하여 무시되기 때문에 강인하다.

3.4 실험 결과 및 최신 방법들과의 성능비교

본 실험에서는 제안한 방법의 우수성을 평가하기 위하여 최신 방법들과의 인식 성능을 비교하였다. Table 3는 최신 방법과 제안한 방법에서 Softmax 손실 함수만을 사용하였을 때의 성능, 가중치가 적용된 Softmax 손실 함수와 Triplet 손실 함수가 결합된 손실 함수를 사용하였을 때의 성능을 비교하였다. 실험 결과, 결합된 손실 함수를 활용하여 훈련한 방법이 Softmax 손실 함수만 사용했을 때 보다 TAR@FAR=0.1 기준 0.006±0.105 향상하였으며 Rank-1 기준 0.020±0.003 향상하였다. 또한 기존 NAN[7]와 비교했을 때 TAR@FAR=0.1 기준 0.008±0.003 향상하였으며 Rank-1 기준 0.004±0.001 향상하였다. 이러한 성능 향상은 제안한 방법이 End-to-end 방식으로 모델을 설계하였으며, 이는 결합된 손실 함수를

Table 2. Performance evaluation about traditional method vs. Proposed method on the IJB-A dataset. For verification, the true accept rates (TAR) vs. false positive rates (FAR) are reported. For identification, the true positive identification rate (TPIR) vs. false positive identification rate (TPIR) and the Rank-N accuracies are presented.

Method	1:1 Verification TAR			1:N Identification TPIR				
	FAR=0.001	FAR=0.01	FAR=0.1	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5	Rank-10
CNN+Average	0.581±0.039	0.678±0.031	0.706±0.031	0.472±0.028	0.591±0.012	0.822±0.015	0.893±1.0.018	0.905±0.016
CNN+Concatenation	0.615±0.091	0.686±0.079	0.767±0.059	0.426±0.053	0.557±0.033	0.758±0.049	0.874±0.031	0.885±0.028
CNN+Score Fusion	0.645±0.047	0.743±0.029	0.886±0.027	0.632±0.019	0.740±0.014	0.884±0.018	0.917±0.017	0.922±0.014
Proposed method (softmax+triplet loss)	0.753±0.037	0.891±0.011	0.965±0.006	0.798±0.015	0.912±0.013	0.940±0.017	0.970±0.012	0.975±0.001

Table 3. Performance evaluation about State-of-the-Arts method vs. Proposed method on the IJB-A dataset. For verification, the true accept rates (TAR) vs. false positive rates (FAR) are reported. For identification, the true positive identification rate (TPIR) vs. false positive identification rate (TPIR) and the Rank-N accuracies are presented.

Method	1:1 Verification TAR			1:N Identification TPIR				
	FAR=0.001	FAR=0.01	FAR=0.1	FPIR=0.01	FPIR=0.1	Rank-1	Rank-5	Rank-10
B-CNN[13]	-	-	-	0.143±0.027	0.341±0.032	0.588±0.020	0.796±0.017	-
LSFS[14]	0.514±0.060	0.733±0.034	0.895±0.013	0.383±0.063	0.613±0.032	0.820±0.024	0.929±0.013	-
DCNN*manual*+metric[15]	-	0.787±0.043	0.947±0.011	-	-	0.852±0.018	0.937±0.010	0.954±0.007
Triplet Similarity[16]	0.590±0.050	0.790±0.030	0.945±0.002	0.556±0.065	0.754±0.014	0.880±0.015	0.95±0.007	0.974±0.005
Pose-Aware Models[17]	0.652±0.037	0.826±0.018	-	-	-	0.840±0.012	0.925±0.008	0.946±0.007
Deep Multi-pose[18]	-	0.876	0.954	0.52	0.75	0.846	0.927	0.947
FPN[19]	0.775	0.852	0.901	-	-	0.914	0.930	0.948
Template Adaptation[20]	0.836±0.027	0.939±0.013	0.979±0.004	0.774±0.049	0.882±0.016	0.928±0.010	0.977±0.004	-
DR-GAN[21]	0.539±0.043	0.774±0.027	-	-	-	0.855±0.015	94.7±0.011	
NAN[7]*	0.745±0.038	0.876±0.008	0.960±0.005	0.779±0.019	0.905±0.008	0.936±0.019	0.968±0.009	0.972±0.009
Proposed method (Only softmax loss)	0.721±0.043	0.865±0.009	0.959±0.007	0.757±0.016	0.894±0.009	0.920±0.014	0.947±0.011	0.956±0.009
Proposed method (softmax+triplet loss)	0.753±0.037	0.891±0.011	0.965±0.006	0.798±0.015	0.912±0.013	0.940±0.017	0.970±0.012	0.975±0.001

: Results obtained through direct implementation indicates that results are obtained through provided by the following online link: https://github.com/YirongMao/NAN

활용하여 학습하기 때문에 특징 추출 네트워크가 어텐션 신경망인 NAN에 긍정적인 영향을 미치게 된다. Fig. 4는 IJB-A[9] 데이터셋 중 일부를 제안한 방법에 실행했을 때 입력 시퀀스에 대한 각각의 프레임의 가중치 a_k를 나타낸 것으로, 얼굴이 정면에 가까우며, 영상 품질이 우수하고, 중첩이 없는 영상일수록 높은 가중치를 매기는 것을 확인 할 수 있다. 이는 비디오 얼굴인식을 위한 다중 손실 함수 기반 어텐션 신경망 학습 방법의 유용성을 증명한다. Fig. 5는 오인식된 일부 샘플을 나타낸 것으로 포즈 변화

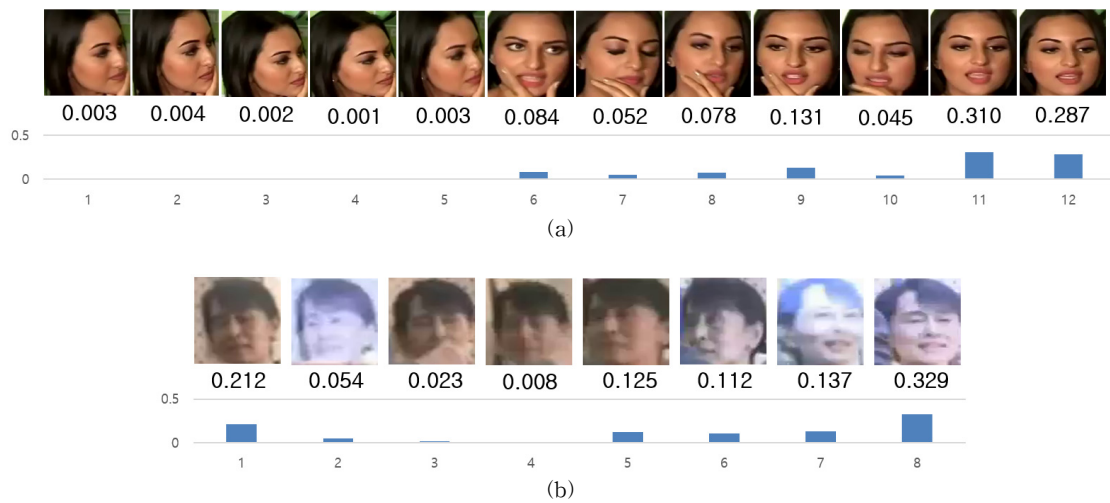

Fig. 4. Weight a_k extraction results of the proposed method for video face images. (a) The weight is set low when the face angle or occlusion occurs, and the weight is set high for the frame that is close to the front without occlusion. (b) The weight is set low for the frame with relatively poor face quality, and the weight is set high for the frame that is close to the front and shows the eyes, nose and mouth.



Fig. 5. Example face sequences incorrectly recognized by our proposed method. Most of the incorrectly recognized images have large pose variation, occlusion and low-resolution.

가 크거나 중첩이 심하고 해상도가 낮은 영상이 대체적으로 오인식되고 있다. Fig. 5에서 얼굴 영상 아래의 수치와 그래프는 수식 (4)의 어텐션 가중치 a_k를 나타낸다. 제안된 방법은 오인식된 일부 샘플들이 있더라도 포즈변화가 적고 중첩이 심하며 해상도가 낮은 영상이 존재한다면 가중치를 작게 결정하여 무시가 가능하다. 하지만 Fig. 5처럼 얼굴 영상 시퀀스가 낮은 품질(포즈 변화, 중첩, 저해상도 등)의 얼굴 영상들로 구성될 때 낮은 품질의 영상에 가중치를 높게 결정되어 잘못 인식되는 경우가 발생한다.

5. 결 론

본 논문에서는 비디오 얼굴 인식을 위한 어텐션 기법 중 하나인 NAN을 활용하여 가중치가 적용된 Softmax 손실 함수와 Triplet 손실 함수를 결합한 종단학습(End-to-end learning) 방법을 제안한다. Softmax 손실 함수는 인식 대상자를 분류할 때 사용되는 손실 함수이며, Triplet 손실 함수는 같은 인식 대상자간에는 가깝게 다른 인식 대상자간에는 멀게 만들어 성능을 향상시키는 손실 함수이다. 이 손실 함수를 가중치를 적용하여 최적으로 결합한 손실함수를 사용하여 성능을 향상하였다. 또한 End-to-end 학습으로 모델을 설계하였으며, 이는 결합된 손실 함수를 활용하여 학습하기 때문에 특징 추출 네트워크가 어텐션 신경망인 NAN에 영향을 오히려 긍정적인 영향을 미치게 된다. 실험은 전통적인 결합방식과 최신 연구결과들과 성능을 비교를 하였다. 전통적인 결합방법으로는 대표적으로 각 프레임의 특징을 추출하여 특징들을 평균(Average) 혹은 특징 연결(Concatenation), 점수 융합(Score fusion)이 있다. 비교실험 결과, 제안한 방법이 전통적인 결합방법보다 우수한 성능을 달성하였다. 또한, 결합된 손실 함수를 활용하여 훈련한 방법이 Softmax 손실 함수만 사용했을 때 보다 TAR@FAR=0.1 기준 0.006±0.105 향상하였으며 Rank-1 기준 0.020±0.327 향상하였다. 최신연구들과 비교해본 결과, 기존 NAN와 비교했을 때 TAR@FAR=0.1 기준 0.008±0.335 향상하였으며 Rank-1 기준 0.004±0.141 향상하였다. 이러한 성능 향상은 제안한 방법이 End-to-end 학습으로 모델을 설계하였으며, 이는 결합된 손실 함수를 활용하여 학습하기 때문에 특징 추출 네트워크가 어텐션 신경망에 긍정적인 영향을 미치게 된다.

REFERENCE

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the Gap to Human-level Performance in Face Verification,” *IEEE Conference on Computer Vision and Pattern*

Recognition (CVPR), pp. 1701-1708, 2014.

[2] O.M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," *Proceedings of the British Machine Vision Conference (BMVC)*, Vol. 1, pp. 6, 2015.

[3] Y. Sun, X. Wang, and X. Tang, "Deeply Learned Face Representations are Sparse, Selective, and Robust," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2892-2900, 2015.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815-823, 2015.

[5] I. Masi, A.T. a. Tran, J. Toy Leksut, T. Hassner, and G. Medioni, "Do We Really Need to Collect Millions of Faces for Effective Face Recognition?," *European Conference on Computer Vision (ECCV)*, 2016.

[6] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," *European Conference on Computer Vision (ECCV)*, pp. 499-515, 2016.

[7] J. Yang, P. Ren, D. Zhang, D. Chen, F. Wen, H. Li, and G. Hua, "Neural Aggregation Network for Video Face Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4362-4371, 2017.

[8] Y. Liu, J. Yan, and W. Ouyang, "Quality Aware Network for Set to Set Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5790-5799, 2017.

[9] B.F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, and K. Allen, et al. "Pushing the Frontiers of Unconstrained Face Detection and Recognition: IARPA Janus Benchmark A," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931-1939, 2015.

[10] K.T. Kim and J.Y. Choi, "Development of Combined Architecture of Multiple Deep Convolutional Neural Networks for Improving Video Face Identification," *Journal of Korea Multimedia Society*, Vol. 22, No. 6, pp. 655-664, 2019.

[11] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," *arXiv preprint*, arXiv:1703.07737, 2017.

[12] D.P. Kingma and J. Ba. Adam, "A Method for Stochastic Optimization," *arXiv preprint*, arXiv: 1412.6980, 2014.

[13] A.R. Chowdhury, T.Y. Lin, S. Maji, and E. Learned-Miller, "One-to-many Face Recognition with Bilinear CNNs," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1-9, 2016.

[14] D. Wang, C. Otto, and A.K. Jain, "Face Search at Scale: 80 Million Gallery," *arXiv preprint*, arXiv:1507.07242, 2015.

[15] J.C. Chen, R. Ranjan, A. Kumar, C.H. Chen, V. Patel, and R. Chellappa, "An End-to-end System for Unconstrained Face Verification with Deep Convolutional Neural Networks," *IEEE International Conference on Computer Vision Workshops*, pp. 118-126, 2015.

[16] S. Sankaranarayanan, A. Alavi, C. Castillo, and R. Chellappa, "Triplet Probabilistic Embedding for Face Verification and Clustering," *arXiv preprint*, arXiv:1604.05417, 2016.

[17] I. Masi, S. Rawls, G. Medioni, and P. Natarajan, "Pose-aware Face Recognition in the Wild," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4838-4846, 2016.

[18] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, and I. Masi, et al. "Face Recognition using Deep Multi-pose Representations," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1-9, 2016.

[19] F.J. Chang, A. Tuan Tran, T. Hassner, I. Masi,

R. Nevatia, and G. Medioni, “Faceposenet: Making a Case for Landmark-free Face Alignment,” *IEEE International Conference on Computer Vision Workshops*, pp. 1599-1608, 2017.

[20] N. Crosswhite, J. Byrne, C. Stauffer, O. Parkhi, Q. Cao, and A. Zisserman, “Template Adaptation for Face Verification and Identification,” *Image and Vision Computing*, pp. 35-48, 2018.

[21] L. Tran, X. Yin, and X. Liu, “Disentangled Representation Learning GAN for Pose-Invariant Face Recognition,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1415-1424, 2017.



김 경 태

2016년 중원대학교 의료공학과 학사
2017년～2019년: 한국외국어대학교 컴퓨터.전자공학부 석사
2019년～현재: 한국외국어대학교 컴퓨터.전자공학부 박사
관심분야 : 머신러닝, 패턴인식, 영상처리, 인공지능



유 원 상

2008년 KAIST 정보통신공학 석사
2013년 Otto-von-Guericke University Magdeburg 전기공학 박사
2019년～2012년 독일 Leibniz Institute for Neurobiology 연구원
2013년～2019년 미국 Children’s National Hospital 선임연구원
2020년～현재선문대학교 정보통신공학과 조교수
관심분야 : 영상처리, 인공지능, 의료영상, 뇌공학



최 재 영

2011년 KAIST 전기 및 전자공학과 박사
2008년～2012년 토론토대학 연구원
2012년～2013년 펜실베니아대학 연구원
2013년～2014년 삼성전자 책임연구원
2014년～2016년 중원대학교 의료공학과 조교수
2016년～현재 한국외국어대학교 컴퓨터.전자공학부 부교수
관심분야 : 딥러닝, 머신러닝, 패턴인식, 영상처리