임베디드 환경에서의 딥 러닝(Deep Learning) 기반 실시간 성별 인식

정현욱, 김대회, Wisam J, Baddar, 노용만
한국과학기술원 전기 및 전자공학과
e-mail : ymro@kaist.ac.kr

Real-time Gender Classification based on Deep Learning in Embedded System

Hyunwook Jeong, Dae Hoe Kim, Wisam J. Baddar, Yong Man Ro
School of Electrical Engineering, KAIST

요 약

사물 인터넷(IoT)의 확산에 따라 기계가 사용자의 정보를 인식하는 일이 매우 중요해졌다. 그 중에서도 성별은 사용자의 특징을 판단하는 결정적인 요소 중 하나이다. 하지만 아직 성별 인식에 관련된 연구는 여전히 도전적이며 향상시킬 부분이 많이 남아있다. 본 논문에서는 deep-convolutional neural network (DCNN)를 이용하여 높은 성능을 갖는 성별 인식 네트워크를 제안하며, 이를 모바일 GPU 보드에 임베디드 포팅(porting)하여 실시간 성별인식 시스템을 구성한 뒤, PC 환경과 모바일 GPU 환경에서 제안하는 시스템의 성능을 비교, 분석한다.

1. 서론

최근 사람과 사물, 공간을 연결해주는 사물 인터넷(IoT) 기술이 확산되고 있다. 사물 인터넷 네트워크를 구성하는 사물들이 실제 환경에 부합하는 서비스를 제공하기 위해서는 프로세스를 진행하기 전 인간에 대한 기본적인 정보(성별, 연령, 감정 등)를 자동으로 인지해야 한다.

그 중에서도 성별은 사회적으로, 생물학적으로 남, 여 각각에 대해 기대되는 분야가 다르기 때문에 사용자를 파악하는데 있어서 매우 중요한 정보이다. 하지만 현재 성별 인식 기술들이 실제 환경에 적용되기 위해서는 극복해야 할 부분들이 많이 있다. 최근까지 주로 사용된 성별 인식 방법으로는 support vector machine (SVM) 분류기를 이용한 성별 인식 방법 [1]과 local binary pattern (LBP) 특징을 사용한 성별 인식 방법[2] 등이 있다. 이러한 기존의 방법들은 촬영된 이미지의 품질이 낮거나 조명과 같은 외부 효과가 있을 경우 인식률이 급격히 떨어지는 문제가 있다[3]-[5]. 따라서 우수한 성별 인식 성능을 위해서는 사진의 품질에 상관 없이 개개의 이미지 데이터로부터 성별 인식에 가장 영향력 있는 특징을 찾아내는 방법이 요구된다. 여러 방법 가운데 딥 러닝은 이미지에서 물체의 특징을 결정하는 핵심적인 특징들을 추출할 수 있다[4]. 그래서 본 논문에서는 딥 러닝(Deep Learning)을 사용한 네트워크를 제안하여 향상된 성별인식 알고리즘을 제시한다.

실제 사물 인터넷 환경에서는 낮은 전력 소비를 위해 개별적인 컴퓨터 기반의 통제가 아닌 각각의 사물에 임베디드(embedded) 포팅(porting)된 집적회로 기반의 맞춤형 통제가 이루어진다. 본 논문에서는 실제 사물 인터넷 환경과 유사한 시스템을 구축하기 위하여 Jetson TK1 (NVIDIA)를 이용하여 모바일 GPU 보드에 딥 러닝 기반의 실시간 성별인식 시스템을 구성하였다. 더 나아가, PC 환경과 임베디드 환경에서의 성능 차이를 비교 및 분석 하였다.

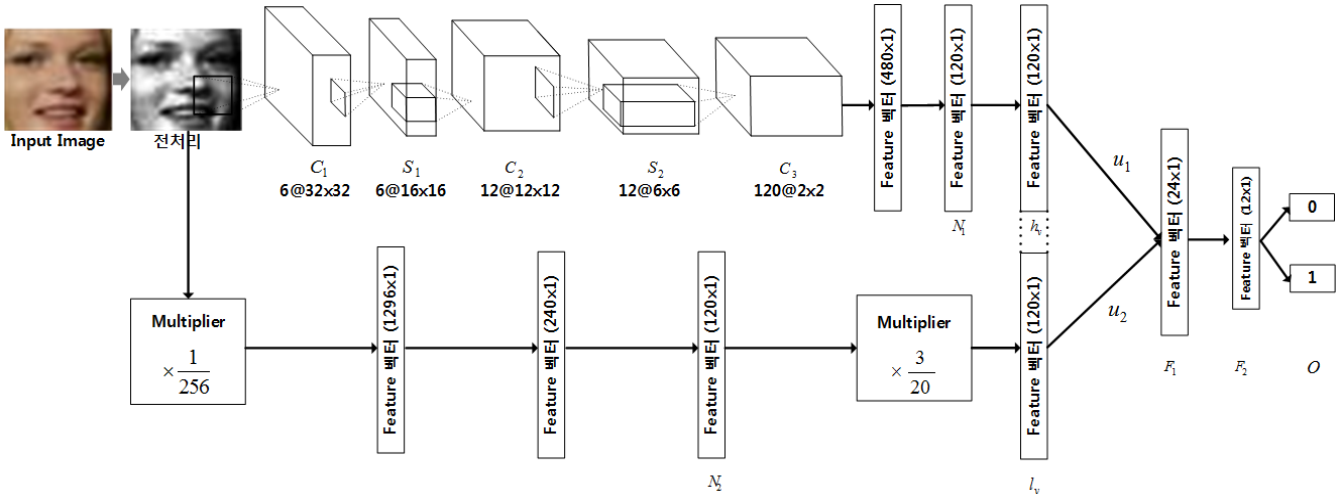
본 논문의 구성은 다음과 같다. 2 절에서는 제안하는 딥 러닝 네트워크 및 알고리즘을 설명한다. 3 절에서는 실험 환경 및 결과를 설명하고 마지막으로 4 절에서 결론을 맺는다.

2. 제안하는 방법

성별 인식 학습을 위해서 먼저, 얼굴 이미지들을 가지고 전처리(preprocessing) 과정을 수행한다. 그리고 (그림 1)의 제안하는 네트워크를 거치면서 계산된 비용함수를 최소화 하는 방향으로 학습한다. 자세한 내용은 다음 하위 항목에서 설명된다.

2.1 성별 인식을 위한 전처리

본 논문에서는 이미지 데이터를 네트워크에 입력하기 전 다음과 같은 전처리 과정을 거친다. 먼저, 얼굴 검출기 [6]를 이용하여 얼굴을 검출한 후, 검출된 영역을 잘라내어 크기를 모두 36 x 36 으로 변환시킨다. 사용한 datasets 에는 다양한 인종이 포함되어 있고, 다양한 촬영 환경으로 인한 다양한 밝기 효과가 반영되어 있기 때문에 이를 보정하기 위해 RGB 영상을 그레이 스케일로 변환하고, 히스토그램 평활



(그림 1) 본 논문에서 제안하는 딥 러닝 구조

화를 수행하여 다양한 밝기 변화를 보정해 주었다.

2.2 제안하는 네트워크 구조

전처리 된 이미지를 학습시키기 위해 기존 연구 [7]에서 더 개선된 네트워크를 제안한다. 먼저 얼굴의 윤곽 및 구조와 같은 상위 단계 정보를 얻기 위해 (그림 1)의 상위 네트워크와 같이 convolutional neural network(CNN) 구조를 설계하였다. CNN 구조는 3 개의 convolution layer 와 2 개의 subsampling layer 로 구성되어 있다. 각각의 convolution layer 직후에는 활성화 함수(activation function)를 거친다. 본 논문에서는 성별(남, 여) 즉, 2 개의 라벨로 분류하기 때문에 각 feature 의 점수가 0 이하, 1 이상과 같이 한쪽으로 크게 치우치는 것을 방지하기 위해 활성화 함수로 sigmoid 함수를 사용하였고 이는 식 (1)과 같이 표현된다.

$$\sigma(x) = sigm(x) = \frac{1}{1+e^{-x}}, \qquad (1)$$

식(1)에서 e 는 지수함수를 나타내고 x는 feature map 에서의 unit 값이다. 그리고 convolution layer 를 거쳤을 때, 각각의 convolution layer 에서 feature 들이 갖는 값은 다음과 같이 표현될 수 있다.

$$a_j^l = sigm(\sum_k w_{jk}^l * a_k^{l-1} + b_j^l), \qquad (2)$$

식 (2)에서 w_{jk}^l 는 (l -1)-번째 layer 의 k -번째 feature map 과 연산하여 l-번째 layer 에서 j-번째 feature map 을 생성하는 가중치 행렬, a_j^l 는 l-번째 layer 에서 j 번째 feature map 을 의미하고, b_j^l 는 l -번째 layer 에서 j-번째 feature map 의 bias 를 가리킨다.

또한 convolution layer 사이에 있는 sampling layer 의 경우 sigmoid 함수와 마찬가지로 각 feature 값을 0 부터 1 사이에서 동작영역 (active region)에 분포 시키기 위해 average pooling 방법을 사용한다 [8].

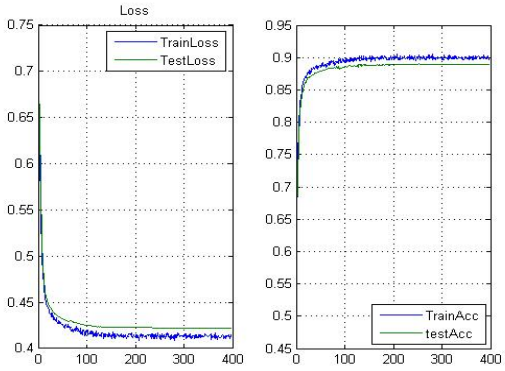
하지만 첫 번째 구조만을 통해 학습시킨 결과, 저조한 인식률을 보였고 이를 개선하기 위해 추가적인 구조를 삽입하였다. (그림 1)의 하위 네트워크는 얼굴 이미지에서 노이즈와 같이 불필요한 정보를 제거하기 위해 사용된 deep neural network (DNN) 구조이다. 36 x 36 크기의 입력 영상을 1296 x 1 벡터로 변환시킨 후, 벡터의 급격한 정보 손실을 줄이기 위해 2 번의 fully connected layer 를 거쳐 벡터의 차원을 240 차원, 120 차원 순으로 감소시켰다. 하위 네트워크에서 fully-connected layer 만을 사용할 경우 N_2 layer 의 feature 값이 N_1 layer 의 feature 값보다 너무 크게 나온다. N_2 layer 의 feature 값을 줄이기 위해 입력영상을 1296 x 1 로 벡터화(vectorization)하기 전에 multiplier 를 연결하였다. 그리고 네트워크가 학습 과정에서 상위 네트워크와 하위 네트워크의 가중치 조절을 하는 동안 최적의 비율을 찾지 못하고 local minima 에 걸릴 가능성이 높다 [9]. 이를 해결하기 위해 실험적인 방법을 통해 하위 네트워크의 가장 오른쪽에 추가적인 multiplier 를 삽입하였다.

마지막으로 (그림 1)에서 오른쪽 부분은 상위 네트워크와 하위 네트워크의 상호 작용을 위해 연쇄(concatenate)시키는 역할을 한다. 연쇄 처리를 위해 상위 네트워크의 출력물을 벡터화하고 두 번째 구조와 unit 개수를 맞추어주기 위해 fully connected layer 를 삽입하였다. 이후 첫 번째 구조와 두 번째 구조를 연쇄작용(concatenate)을 거쳐 240 차원 벡터를 생성했고 다음과 같이 표현된다.

$$F_1 = sigm(u_1^{\mathrm{T}} \cdot h_v + \beta u_2^{\mathrm{T}} \cdot l_v + b^l), \qquad (3)$$

식 (3)에서 u_1 은 상위 네트워크의 가중치를 u_2 는 하위 네트워크의 가중치를 나타내며 h_v 와 l_v 는 각각 연쇄시키는 과정에서 각각 상위 네트워크와 하위 네

트워크에서 온 벡터이다. 그리고 β는 상위 네트워크와 하위 네트워크의 가중치를 조절해 주기 위한 상수이다.



(그림 2) 반복 학습에 따른 loss 값과 성별 인식률

이전과 마찬가지로 계산 도중 정보의 급격한 손실을 방지하기 위해 두 개의 fully-connected layer (240→24→12→2)를 output layer 앞에 삽입하였다. F_1 layer 의 출력과 연산되는 가중치 행렬을 k_{F_1}, F_2 layer 의 출력과 연산되는 가중치 행렬을 k_{F_2} 라고 하면 output layer (O)는 아래 식 (4)와 같다.

$$o = k_{F_2}^{\mathrm{T}} \cdot k_{F_1}^{\mathrm{T}} \cdot F_1, \tag{4}$$

2.3 네트워크 학습 방법

본 논문이 제안하는 네트워크는 수 많은 이미지 데이터를 학습하기 위해 stochastic gradient descent (SGD) 방법을 사용한다. 상위 네트워크와 하위 네트워크의 비를 조절하기 위해 n번째 epoch 에서 가중치 조절 상수를 β_n , softmax loss function 을 $L(W,b,\beta)$ 라고 할 때, 비용함수를 최소로 하는 β 값을 찾는 방식으로 학습이 진행된다. 여기서, 비용함수를 직접 β로 미분하는 과정은 복잡하기 때문에 다음과 같이 chain rule 을 사용한다.

$$\frac{\partial L(W,b\beta)}{\partial \beta} = \frac{\partial L}{\partial o} \cdot \frac{\partial o}{\partial \beta}, \tag{5}$$

가장 최적화된 값을 구할 때까지 식 (6)과 같은 SGD 방식을 통해 가중치 조절 상수 β를 구한다.

$$\begin{cases} \beta_{n+1} = \beta_n - \alpha \cdot \dfrac{\partial L(W,b,\beta)}{\partial \beta} \\ 0 \le \beta_{n+1}, \beta_n \le 1 \end{cases}, \tag{6}$$

(그림 3) 잘못 분류된 예시.

3. 실험 결과 및 분석

3.1 사용된 데이터 및 성별 인식률

본 논문에서 제안하는 방법의 성능을 평가하기 위해 Caffe 라이브러리를 사용하였다 [10]. 그리고 LFW dataset 과 FERET dataset 을 혼합하여 training set 으로 남, 여 각각 7,500 장을 랜덤 추출하여 사용하였다. 이 때, 좌우 반전과 [-3°, 0°, 3°]로 회전시켜 dataset 을 6 배로 증가시켜 데이터 양을 늘렸다. 그리고 test set 으로는 training set 을 제외한 남, 여 각각 500 장을 랜덤 추출하여 사용하였다. 그 결과 <표 1>과 같이 89.8 %의 성별 인식률을 얻을 수 있었다. 동일 data set 으로 실험한 기존 연구 [7]와 비교하였을 때 성능이 약 26%정도 향상되었고 (그림 2)와 같이 수렴 속도도 증가하였다.

(그림 3)은 400 번의 epoch 이후 잘못 분류된 몇 가지 사례들이다. 성별이 잘못 인식된 사례들은 머리카락이나 선글라스 등으로 인한 이미지 가림 현상이 발생했거나 지나치게 측면 촬영이 이루어진 경우라고 볼 수 있다.

<표 1>제안하는 방법을 이용한 성별 인식률

	Yuxin Jiang 방법 [7]	제안한 네트워크
인식률	63%	89.8%

3.2 임베디드 환경 및 PC 실험과의 비교 분석

본 논문에서 제안하는 알고리즘을 구현한 임베디드 보드는 NVIDIA Jetson TK1 이고 Ubuntu 기반으로 동작을 한다. 그리고 CPU, GPU, 메모리를 하나로 패키징 한 Tegra Soc 가 Jetson TK1 에 장착되어 있기 때문에 별도의 PC 없이도 포팅(porting) 작업이 가능하다. 먼저 보드 세팅을 위해 보드에 NVIDIA installer 를설치 후 GPU 를 작동시키기 위한 Cuda SDK 를 설치한다. 그리고 Caffe, openCV 라이브러리를 이용하기 위한 패키지들을 설치한다 [11]. 이후 C++에서 구현한 프로그램을 보드에 포팅(porting)하고 컴파일 하여 실시간 성별인식 시스템을 구성하였다. (그림 4)는 제안하는 방법의 실시간 성별인식 시스템 동작 영상이다. 그리고 <

표 2>는 PC 환경에서의 동작 성능과 임베디드 환경에서의 실시간 동작 성능을 비교하여 나타낸 것이다.

(그림 4) 임베디드 환경에서의 실시간 얼굴인식.

<표 2> PC 사양과 임베디드 사양에서의 실시간 성별 인식 구현 시 필요한 소비전력 및 초당 프레임 수

	PC	임베디드 보드
소비전력	560 Watts[1)]	8 Watts[12]
초당 프레임 수	10~17 fps	4.8~6.2fps

4. 결론

본 논문에서는 성별 인식을 위한 새로운 딥 러닝 알고리즘을 제안하였다. 제안한 알고리즘에선 하위 네트워크가 가장 효과적으로 보조 역할을 할 수 있는 시스템을 찾는 방법과 2 개의 output 으로 feature 를 줄이는 동안 정보 손실량을 최대한으로 줄이기 위한 점이 고려되었고 이 결과 89.8%의 성능을 보여 이전 논문들에 비해 향상된 성능을 보여주었다. 그리고 이를 C++을 이용하여 실시간 프로그램으로 구현한 후 Jetson TK1 보드에 임베디드 포팅을 하고 PC 환경과 비교를 하였다. 초당 처리되는 프레임 수는 임베디드 보드가 PC 에 비해서 매우 적지만 소비전력 측면에선 PC 에 비해 큰 이점을 가져서 향후 IoT 가 적용된 가전제품에 사용자의 사전 정보로 많이 쓰일 것으로 보인다.

감사의 글

이 논문은 2015 년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2015R1A2A2A01005724)

참고문헌

[1] J. Zang and B. L. Lu, “A support vector machine classifier with automatic confidence and its application to gender classification,” *Neurocomputing* vol. 74, pp.1926-35, 2011

[2] E. Mäkinen and R. Raisamo, “Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces,” *IEEE Transactions on, Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 541-547, 2008.

[3] Jeong-Jik Seo, Hyung-Il Kim and Yong Man Ro, “Pose-robust and Discriminative Feature Representation by Multi-task Deep Learning for Multi-view Face Recognition,” in *IEEE ISM*, Dec. 2015

[4] Yeoreum Choi, Hyung-Il Kim and Yong Man Ro, “Two-step Learning of Deep Convolutional Neural Network for Discriminative Face Recognition under Varying Illumination,” in *IS&T International Symposium on Electronic Imaging*, Feb. 2016

[5] Heountaek Lim, Hak Gu Kim and Yong Man Ro, “Learning based hole filling method using deep convolutional neural networks for view synthesis,” in *IS&T International Symposium on Electronic Imaging*, Feb. 2016

[6] Bradski, Gary, "OpenCV," *Dr. Dobb's Journal of Software Tools*, 2000

[7] Jiang, Yuxin, et al., "Multi-feature deep learning for face gender recognition," in *IEEE ITAIC, 2014.*

[8] Y. Bengio, I. Goodfellow, and A. Courville: *Deep Learning*: MIT Press, 2015.

[9] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, “Why Does Unsupervised Pre-Training Help Deep Learning?,” *Journal of. Machine Learning Research*, vol. 11, pp. 625-660, 2010.

[10] Jia, Yangqing, et al., "Caffe: Convolutional architecture for fast feature embedding," *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014.

[11] Hurtado, Andres Felipe, et al., "Proposal of a Computer Vision System to Detect and Track Vehicles in Real Time Using an Embedded Platform Enabled with a Graphical Processing Unit," in *IEEE ICMEAE,* 2015.

[12] Alicea-Nieves, Christopher., "Caffe Framework on the Jetson TK1: Using Deep Learning for Real Time Object Detection,"

1) PC 소비전력은 저자 소유의 컴퓨터를 스펙에 맞추어 계산한 내용임.