시선 추론을 위한 단-대-단 경량화 딥러닝 프레임워크

오준오, 이정호*, 최상일
단국대학교 컴퓨터학과, *단국대학교 응용컴퓨터공학과

72200119@dankook.ac.kr, *wowo933@naver.com, choisi@dankook.ac.kr

End-to-End Lightweight Deep Learning Framework for Gaze Estimation

Oh Jun O, Lee Jung Ho*, Choi Sang Il
Department of Computer Science and Engineering, Dankook Univ.,
*Department of Applied Computer Science and Engineering, Dankook Univ.,

요 약

본 논문은 얼굴 이미지에서 시선을 추론하고, 지식 증류를 통해 추가적인 비용의 증가 없이 정확도를 향상시키는 학습 프레임워크를 제안한다. 실험 결과, 파라미터 감소에 따른 정확도 하락을 효과적으로 억제하고 기존 모델과 비교하여 유의미한 경량화와 지연 속도 하락을 보여주었다.

Ⅰ. 서 론

시선은 사람의 주의력을 나타내며, 의도 및 사회적 관계를 파악할 수 있는 대표적인 행동 중 하나로 다양한 어플리케이션에서 중요한 기능 요소로 사용된다. 시선 추론은 컴퓨터 비전 분야에서 활성화된 연구 주제 중 하나로, 모양 기반 혹은 외관 기반 등 다양한 접근 방식으로 연구가 이루어지고 있다. 컴퓨터 비전에서 합성곱 신경망의 성공 이후 시선 추론에도 적용되기 시작했으며 이에 따라 눈의 이미지를 사용하는 외관 기반 접근 방식이 급격한 성장을 이루었다. 본 실험 역시 외관 기반의 접근 방식으로 이해할 수 있다.

합성곱 신경망의 성공 이후, 컴퓨터 비전 분야는 여러 과제에서 전례 없는 성장세를 보였다. 그러나 아직도 큰 도전과제들이 남아있는데, 대표적인 문제점으로 자율주행, 의료 등 오차에 민감한 분야에서 활용 가능할 만큼 정확도를 보이지 못하고 있으며, 또 다른 문제점으로는 높은 연산량으로 인해 자원에 제약이 있는 상황에서 사용이 어렵다는 점이다. 이 문제를 해결하기 위해 다양한 방법이 제시되었고 그 예로 경량화된 네트워크 구조 설계[1], pruning[2], 양자화[3] 등이 제시되었으며 지식 증류 역시 이러한 접근 방식 중 하나로 이해할 수 있다. 지식 증류는 정확도를 향상시킬 뿐만 아니라 모델 압축에도 사용할 수 있어 경량화 문제에서 활성화된 연구 주제 중 하나이다.

본 논문에서는 시선 추론이라는 주어진 문제를 기존 연구와 달리 세부 문제로 나누지 않고 단-대-단으로 학습하면서도, 얼굴 이미지에서 눈 이미지를 찾는 등 세부 문제를 푸는 방법을 지식 증류를 통해 네트워크에 전달하여 정확도를 유지하고, 문제를 단-대-단 학습이 가능한 작은 네트워크로 해결하는 학습 프레임워크를 제안한다.

Ⅱ. 본 론

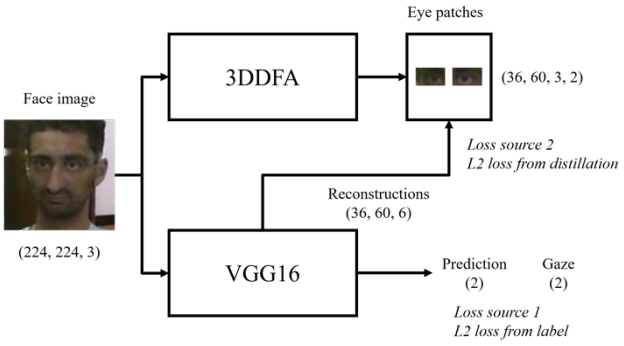

그림 1

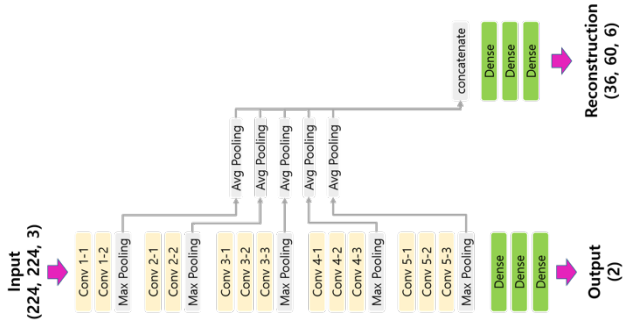

그림 2

(1) 지식 증류를 이용한 학습 전략

본 문단에서는 [그림 1]에 묘사된 바와 같이 지식 증류를 활용한 문제 해결 전략을 설명한다. 첫번째로 목표로 설정한 합성곱 신경망을 원래 구조와 깊이에 따라 여러 부분으로 나누는 것으로 시작한다. 예를 들어, 본 실험에서는 VGG16[4]을 Max Pooling 층을 기준으로 다섯 부분으로 나누었다. 두번째로, [그림 2]와 같이, VGG16 의 합성곱 층 중 Max Pooling 층을 통과한

후의 특징 층들의 해상도를 통일하고 전부 더한 뒤 복원기에 넣어 눈 이미지를 복원한다. 이 복원기는 학습 중에만 활성화되며 추론 과정에서는 제거될 수 있다. 이 구조에서 중요한 점은, 눈 이미지를 복원하는 과정을 통해 합성곱 층들이 사진에서 시선을 추론하기 위한 특징들을 추출할 뿐만 아니라, 해당 문제를 해결하기 위한 가장 중요한 정보인 눈 부위를 학습하도록 영향을 미치는 것이다. 학습 과정에서 다른 네트워크를 통해 추출된 눈 이미지를 복원하는 과정을 통해 그 네트워크의 지식을 증류할 수 있으며, 따라서 이 네트워크는 개념적으로 교사로 이해할 수 있다. 그러나 이 프레임워크에서 교사는 세부 문제에 대한 답을 제시하며 통상적인 지식 증류에 비해 더 간접적인 방식이다.

본 실험에서는 학생 네트워크의 성능을 향상시키고 문제를 해결하기 위해 학습 과정 중에 다음과 같은 두 개의 손실 함수를 사용한다. 첫번째 손실 함수는 모델의 예측값과 해당 이미지의 라벨, 즉 시선 벡터를 이용한 l2 손실 함수이며 다음과 같이 나타낼 수 있다.

$$\|Prediction - Label\|_2^2 \quad (1)$$

또한 두번째 손실 함수는 모델의 특정 특징 맵으로부터 복원된 이미지와 미리 학습된 네트워크에서 추출된 눈 이미지 간의 l2 손실 함수이며 수식을 통해 다음과 같이 나타낸다.

$$\alpha \cdot \|Image_{reconsturction} - Image_{teacher}\|_2^2 \quad (2)$$

해당 수식에서 α 는 0 이 아닌 양의 실수로 라벨에서의 loss 에 비해 복원 이미지의 loss 를 어느 정도 비율로 반영할지에 대한 hyper-parameter 로 실험에서는 대략 0.0005 의 값을 사용하였다.

그러므로 전체 손실 함수는 두 손실 함수를 더하여 다음과 같이 표현할 수 있다.

$$loss = \|Prediction - Label\|_2^2 + \alpha \cdot \|Image_{reconsturction} - Image_{teacher}\|_2^2 \quad (3)$$

(2) 실험 구성 및 결과

제안된 학습 프레임워크를 검증하기 위해 시선 추론 관련 데이터셋 중 MPII Gaze 데이터셋[5]과 RT-GENE 데이터셋[6]을 사용하여 결과를 평가하였다.

실험은 3 개 모델에 대해서 진행되었으며 RT-GENE 모델은 먼저 얼굴 랜드마크 추출 네트워크를 통해 얼굴에서 눈 이미지를 추출한 후, 해당 이미지에서 시선을 추론하는 방식을 취하고 있으며, VGG16 은 제안된 학습 프레임워크와 비교하기 위해 얼굴 이미지와 시선 라벨을 통해서만 학습하였다. 제안된 프레임워크를 통해 학습한 VGG16+KD 모델은 학습 중에만 눈 이미지 복원을 위해 추가적인 파라미터를 필요로 하며, 추론 시에는 이 모듈을 비활성화하기 때문에 VGG16 모델과 동일하다.

정확도 평가는 예측 벡터와 실제 시선 벡터의 cosine similarity 를 구한 후, 두 벡터가 이루는 각도 Θ를 radian 에서 degree 로 변환해 쉽게 파악할 수 있도록 했다. 따라서 결과값이 낮을수록 더 정확한 추론이다. 실험은 기존 논문과 비교하기 위해 각 논문에 기재된 실험 방식을 따랐으며, RT-GENE 데이터셋은 3 번의 교차 검증(3 fold cross validation)을 거친 후, 결과를 평균한 것이다.

	Number of Params	Latency (ms)	MPII Gaze Dataset	RT-GENE Dataset
RT-GENE	2000M	39.52	4.3±0.9	7.7±0.3
VGG16 (Baseline)	114M	7.50	14.09±6.8	23.9±8.7
VGG16+KD (Proposed)	114M	7.50	8.4±3.7	15.7±7.6

표 1

[표 1]은 실험에 대한 결과를 나타내고 있다. VGG 기반 모델들은 RT-GENE 에 비해 더 높은 오차를 기록하고 있다. 그러나 두 단계로 진행되는 RT-GENE 에 비해 단-대-단으로 동작하는 VGG 기반 모델들이 더 낮은 지연 시간을 보여주며, 사용하는 파라미터의 개수 역시 더 적다. 또 제안된 프레임워크를 이용하여 학습한 VGG16 모델의 경우 RT-GENE 과 비교할 때 두 검증 데이터셋에서 각각 4.1 degree, 8.0 degree 로 오차가 증가한 대신 전체 파라미터 개수와 지연 시간이 크게 감소했다. 이는 시선 추론 네트워크가 적용될 가능성이 높은 자원이 제약된 환경, 즉 모바일 또는 임베디드 시스템에서 활용될 여지를 높인다.

Ⅲ. 결 론

본 논문에서는 시선 추론을 위한 단-대-단 딥러닝 프레임워크를 제안한다. 또한 파라미터가 적어지면서 발생하는 정확도 하락 문제를 지식 증류를 통해 효과적으로 억제할 수 있음을 확인하였다. 그러나 [표 1]에서 보이듯 결과의 편차가 커 네트워크가 불안정한 문제는 개선이 필요하다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW 중심대학 사업(2017-0-00091)과 글로벌 핵심인재 양성지원사업의 연구 결과로 수행되었음(No.2020-0-01463)

참 고 문 헌

[1] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. “MobileNets: Efficient convolutional neural networks for mobile vision applications”, CVPR, 2017.

[2] S. Han, H. Mao, and W. J. Dally. “Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding”. 2016.

[3] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi. “Xnornet: Imagenet classification using binary convolutional neural networks”, ECCV, 2016.

[4] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”, ICLR, 2015

[5] X. Zhang, Y. Sugano, M. Fritz and A. Bulling, “MPIIGaze: Real-world dataset and deep appearance-based gaze estimation”. 2017

[6] T. Fischer, H. J. Chang and Y. Demiris “RT-GENE: Real-time eye gaze estimation in natural environments” ECCV, 2018