중심확장 알고리즘이 보강된 식별적 특징학습을 통한 얼굴인식 향상기법

강명균*, 이상철**, 이인호***
*국민대학교 소프트웨어학과
**대구경북과학기술원 융합연구원
***㈜넥시스
e-mail : kmk0202@gmail.com

Improving Discriminative Feature Learning for Face Recognition utilizing a Center Expansion Algorithm

Myeong-Kyun Kang*, Sang C. Lee**, In-Ho Lee***
*Dept. of Computer Science, Kookmin University
**Convergence Research Institute, DGIST
***Nexys Crop.

요 약

좋은 특징을 도출할 수 있는 신경망은 곧 대상을 잘 이해하고 있는 신경망을 의미한다. 그러나 얼굴과 같이 유사한 이미지를 분류하기 위해서는 신경망이 좀 더 구분되는 특징을 도출해야한다. 본 논문에서는 얼굴과 같이 유사도한 이미지를 분류하기 위해 오차함수에 중심확장(Center Expansion)이라는 오차를 추가한다. 중심확장은 도출된 특징이 밀집되면 클래스를 분류하는 매니폴드를 구하기 어려워져 분류 성능이 하락되는 문제를 해결하기 위해 제안한 것으로 특징이 밀집될 가능성이 높은 부분에 특징이 도출되지 않도록 강제하는 방식이다. 학습 시 활용하는 오차는 일반적으로 분류 문제를 위해 사용되는 softmax cross-entropy 오차와 각 클래스의 분산을 줄이는 오차 그리고 제안한 중심확장 오차를 조합해 구할 것이다. 본 논문에서는 제안한 중심확장 오차를 조합한 모델과 조합하지 않은 모델이 결과적으로 특징 도출과 분류에 어떠한 영향을 주었는지 알아볼 것이다. 중심확장을 조합해 학습한 모델이 어떤 영향을 주었는지 알기 위해 본 논문에서는 Labeled Faces in the Wild 를 활용해 분류 실험을 진행할 것이다. Labeled Faces in the Wild 을 활용해 실험한 결과 중심확장을 활용한 모델과 활용하지 않은 모델간의 성능을 차이를 확인할 수 있었다.

1. 서론

합성곱 신경망(Convolutional Neural Network)은 최근 이미지 인식 성능을 크게 향상 시켰다. 딥러닝 기술의 발전과 ImageNet(Deng et al. 2009)과 같은 큰 이미지 제공, 그리고 GPU 와 같은 높은 수준의 컴퓨터성능 향상으로 합성곱 신경망은 컴퓨터 비전에서 매우 중요하게 되었다. ImageNet Large-ScaleVisual Recognition Challenge (ILSVRC)을 통해 많은 이미지 분류를 학습하면서 더 이상 얕은 깊이의 신경망으로 학습 하는 것이 아닌 더 깊은 깊이의 신경망으로 학습하는 것이 가능해졌다. [1], [2] 이전과 비교도 되지 않을 정도로 깊은 깊이의 신경망을 활용하면서 신경망은 고차원의 특징까지도 도출하게 되었고, 이는 곧 인식 성능향상으로 이어지게 되었다. [3] 더 깊은 신경망을 설계해 보다 더 좋은 특징을 도출하기 위해 이미 다양한 방식이 제안되었다. 그 중에는 레이어를 수정해 잔여를 학습하게 하는 방식이 제안되기도 했다. [4] 더 나아가 층을 통째로 건너뛰는 방식이 제안되기도 했다. [5] 물론 깊은 신경망을 사용하는 데에는 깊은 합성곱 신경망이 좋은 특징 표현을 가진다는 가정이 있어, 더 깊은 신경망을 설계하기 위해 다양한 기법들이 사용되었지만, 분류하고 싶은 대상이 일반적인 이미지와 다른 성격을 갖고 있으면, 신경망의 깊이와 함께 추가적으로 고려해야 할 사항이 있다. [6], [7]

본 논문에서는 합성곱 신경망이 더 구분되는 특징을 도출할 수 있도록 강제해 얼굴과 같이 유사한 이미지도 분류할 수 있도록 한다. 도출된 특징이 각 클래스의 중심과의 거리를 줄임과 동시에 모든 특징의 중심으로부터 팽창시키는 모델을 설계해 유사한 이미지도 분류할 수 있도록 한다. 클래스 간에 거리를 팽창시키면서 학습하는 방식을 본 논문에서는 중심확장(Center Expansion)이라고 부른다. 중심확장을 이용한 모델은 이미지를 분류하는 방법에도 차이가 있다. 중심확장을 이용한 분류방법은 일반적으로 모델에 활용

하는 softmax 를 활용해 분류하지 않고, 클래스의 중심과 도출된 특징의 거리를 비교해 가장 가까운 중심을 같은 클래스로 선택해 분류한다. 실험에 사용한 데이터는 Labeled Faces in the Wild(LFW)[8]를 사용하였고, 실험 결과를 평가는 LFW 이미지를 짝지어 동일 클래스의 이미지 인지 아닌지를 확인하며 진행하였다.

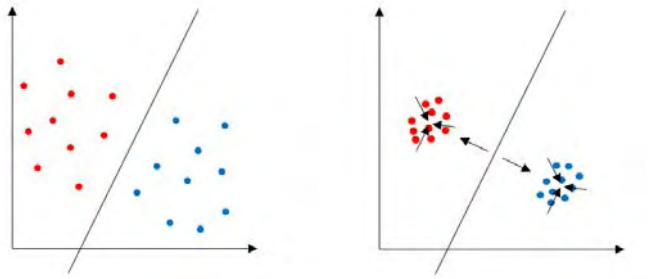
논문의 구성은 2 절에서 식별적 특징을 도출하기 위해 학습한 모델과 관련 신경망이 얼굴인식에 어떤 영향을 주었는지를 논의한다. 추가로 관련 식별적 특징을 활용해 학습한 모델이 어떻게 해서 얼굴인식을 성능을 증가시키려 했던 것인지 살펴본다. 3 절에서는 식별적 특징을 도출하는 오차함수를 정의하며, 본 논문에서 제안하는 중심확장(Center Expansion) 기법을 설명한다. 4 절에서는 실험을 통해 중심확장을 사용한 오차함수와 사용하지 않은 모델간의 차이를 확인하고 5 절에서는 관련 결과를 갖고 토의를 한다. 6 절에서는 본 논문의 결론을 맺는다.

2. 관련연구

일반적으로 합성곱 신경망은 softmax 한 값과 정답 레이블의 cross-entropy 오차를 최소화하는 방식을 활용해 학습한다. 하지만 얼굴인식 문제에서는 softmax 오차 만을 이용해 학습한 모델이 어려움을 겪고 있다. 이는 모델이 도출한 특징이 서로 밀집되어 있어 각 클래스간의 구분이 어려워지기 때문이다. 얼굴과 같이 유사도가 높은 이미지 분류를 위해서는 학습되는 특징이 서로 분류(Classification)되어야 할 뿐만 아니라, 특징이 더 많이 구분되고 차별적 구분 즉, 식별(Discrimination)되는 모델이 필요하다.

최근 이러한 문제를 해결하기 위해 도출한 특징이 더 구분되도록 강제하는 시도가 있었다. contrastive loss 는 자신과 유사한 분류와 자신과 다른 분류를 레이블해 자신과 다른 분류의 거리는 마진보다 커지도록 해 식별적 특징을 학습시켰다. [9] Triplet loss 는 자신과 같은 분류의 거리는 최소화하고 자신과 다른 분류의 거리는 최대화, 추가로 다른 분류와 같은 분류의 거리가 마진보다 커지도록 해 식별적 특징을 학습시켰다. [10] 중심점 오차(Center loss)를 활용해 자신이 속한 클래스의 중심을 구한 후 자신과 해당 중심과의 거리를 최소화해 식별적 특징을 학습시켰다. [11]

이처럼 모델이 식별적 특징을 갖도록 강제하면서 학습시킨 모델이 최근 얼굴인식 분야에서 높은 수준의 인식률을 보여주고 있다. 추가로 식별적 특징을 갖도록 강제한 모델이 분류 시 몇 가지 장점을 갖게 된다. 모델로 도출한 특징은 이전의 softmax 에 비해 더 구분되기에 특징 간의 유클리드 거리를 측정해 분류가 가능하다. 도출된 특징과 각 클래스의 중심을 비교해 가장 가까운 중심을 찾아 분류를 한다거나 더 나아가 여러 이미지를 활용해 특징을 도출하고 각 특징의 유클리드 거리를 비교해 자신과 가장 유사한 이미지 순으로 분류할 수 있다. 특히 클래스 결과를 받지 않고 특징을 도출해 분류를 하는 것은 실용적으로 매우 장점을 갖는다. 모델이 새로운 분류의 이미지가 추가될 때마다 다시 학습을 시켜 결과를 도출하지 않고 새로 추가된 분류의 이미지의 특징과 비교해 분류하는 점이다.



(그림 1) 심화 학습된 특징(deeply learned features) 분포도 비교: (좌)구별적 특징(separable features) 분포와 (우) 식별적 특징(discriminative features) 분포

본 논문에서는 각 클래스의 중심과 거리를 최소화하는 오차함수와 모든 특징의 중심과 거리를 팽창시키는 중심확장을 이용해 모델이 더 좋은 식별적 특징을 갖도록 강제한다. 더 좋은 식별적 특징을 갖도록 학습을 강제한 모델은 각 특징 간의 거리를 늘려 최종 분류 성능을 높이려고 한다.

3. 중심확장(Center Expansion) 기법

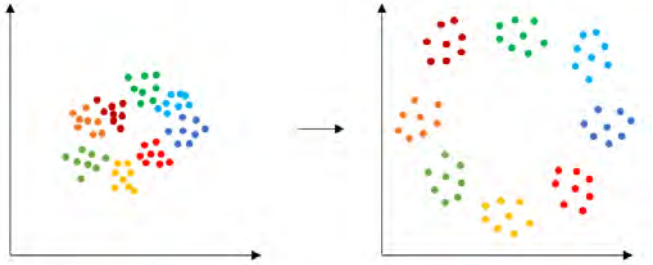
모델이 식별적 특징을 도출하기 위해서는 식별적 특징을 도출되도록 강제하는 오차함수가 있어야 한다. 가장 쉽게 식별적 특징을 도출하게 하는 것은 각 클래스 내 분산을 작아지게 학습하는 것이다. 도출된 특징이 클래스 내 분산이 작아지면 결과적으로 다른 클래스와의 겹칠 수 있는 가능성이 줄어들어 결과적으로 특징이 식별되게 된다. 클래스내 분산을 최소화하기 위해 본 논문에서는 각 클래스의 중심과 소속된 클래스의 특징 간의 거리를 최소화하려 한다. 클래스 내 분산을 최소화하는 오차함수는 식(1)와 같다.

$$L_c = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i - c_{yi}\right\|_2^2 \qquad (1)$$

식(1)의 c_{yi}는 y 클래스 특징의 중심을 의미하며, L_c는 m 개의 도출된 이미지의 특징을 비교해 오차를 구하는 함수이다. c_{yi}는 매시간 모든 학습데이터와 비교해 도출하는 것이 가장 이상적이지만 성능상 효율이 없으므로 mini-batch 내에서 각각의 클래스의 중심을 구한다.

식별적 특징을 갖도록 학습하는 이유는 각 특징이 너무 밀집되지 않도록 하는 것이다. 특징이 밀집되면 모든 값을 만족시키기 위해 함수가 고차가 되거나 모든 값을 만족하지 못해 저차가 되어, 각각 오버피팅이과 언더피팅이 된다. 이는 결과적으로 모델의 분류 성능의 하락으로 이어지게 된다. 지도 학습을 통해 모델을 학습 시키는 것은 결국 각 클래스의 매니폴드를 구하는 것과 같은데 밀집 문제와 더불어 만약에 분류해야 할 클래스의 개수가 많아지면 적합한 매니폴드를 구하는 것이 상대적으로 적은 클래스에 비해

힘들어 진다. 본 논문에서는 밀집될 가능성이 높고 매니폴드를 구하기 어려운 공간일수록 특징이 도출되지 않도록 강제해 적합한 매니폴드가 가능하게 하려 한다. 일반적으로 가장 밀집될 가능성이 높은 공간은 모든 특징의 중심이므로 모든 특징의 중심에 특징이 가까워 질수록 높은 오차를 주려고 한다.



(그림 2) 중심확장(center expansion) 기법을 통해 식별적 심화학습(discriminative deep learning)된 특징 분포도 변화

도출된 특징이 중심으로 밀집되지 않도록 강제하는 오차함수는 식(2)와 같다.

$$L_g = \frac{1}{2}\sum_{i=1}^{m}\left[\arg max \left\|x_i - c_g\right\|_2^2 - \left\|x_i - c_g\right\|_2^2 + \gamma\right] \quad (2)$$

식(2)의 c_g는 모든 특징의 중심을 의미한다. 물론 c_g 또한 매시간 모든 학습데이터와 비교해 도출하는 것이 가장 이상적이지만, 이 또한 성능상의 이유로 mini-batch 내에서 각 클래스의 중심c_j을 사용해 구한다. 식(2)의 γ는 하이퍼파라미터로 도출한 특징과 c_g 간의 거리가 최소한 γ보다 커지도록 강제한다. 가장 멀리 있는 특징인 식 (2) 의 $\arg max \left\|x_i - c_g\right\|_2^2$ 은 mini-batch 내에서 x_i와 c_g차가 가장 큰 것을 찾아 도출하며, 식(2)의 L_g는 mini-batch 내에서 c_g와 가장 멀리 있는 특징과 자신을 각각 각각의 차를 제곱한 것의 차가 γ보다 커지도록 오차를 주어 강제한다. 이 값을 최소화 하면 특징이 γ보다 커지도록 강제하면서 도출된 특징이 중심으로부터 멀어지게 학습할 수 있다. 다음은 x_i에 의한 L_c와 L_g 식(3)와 식(4) 그리고 gradient c_j와 c_g식(5)와 식(6)이다.

$$\frac{\partial L_c}{\partial x_i} = x_i - c_{yi} \quad (3)$$

$$\frac{\partial L_g}{\partial x_i} = \arg max \left(x_i - c_g\right) - \left(x_i - c_g\right) \quad (4)$$

$$\Delta c_j = \frac{\sum_{i=1}^{m}\delta(y_i = j)\cdot(c_j - x_i)}{1 + \sum_{i=1}^{m}\delta(y_i = j)} \quad (5)$$

$$\Delta c_g = \frac{\sum_{i=1}^{m}\delta(y_i = j)\cdot(c_g - c_j)}{1 + \sum_{i=1}^{m}\delta(y_i = j)} \quad (6)$$

식(5)의 δ는 내부 조건을 만족하면 1 이고 내부 조건을 만족하지 못하면 0 이다. 이는 자신이 속한 클래스의 중심이 오로지 자신 클래스의 특징만이 영향을 미칠 수 있도록 하기 위함이다. 식(5)의 Δc_j는 도출된 특징과 자신이 속한 클래스의 중심의 차를 사용해 각 클래스에 해당하는 c_j를 수정한다. 실제로 Δc_j를 활용해 c_j를 갱신할 때는 하이퍼파라미터 α를 활용해 균형을 맞춘다. 식(6)의 Δc_g는 학습 시 mini-batch 내에 있는 각 클래스의 c_j를 활용해 갱신하며 하이퍼파라미터 β를 활용해 균형을 맞춘다. 하지만 L_g와 L_c만을 이용해 모델을 학습할 수는 없다.

최종 오차함수는 제안한 L_g 와 L_c 그리고 softmax cross-entropy 오차를 합한 식(7)와 같다.

$$L_t = L_s + \lambda_c L_c + \lambda_g L_g \quad (7)$$

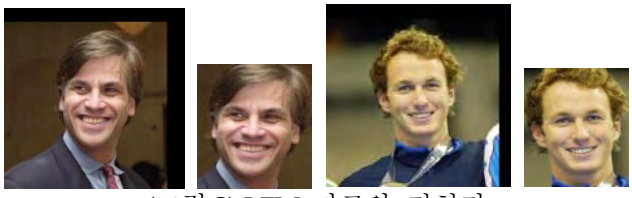
$$L_t = -\sum_{i=1}^{m}\log\frac{e^{W_{yi}^T x_i + b_{yi}}}{\sum_{j=1}^{n} e^{W_j^T s_i + b_j}} + \frac{\lambda_c}{2}\sum_{i=1}^{m}\left\|x_i - c_{yi}\right\|_2^2 + \frac{\lambda_g}{2}\sum_{i=1}^{m}\left[\arg max \left\|x_i - c_g\right\|_2^2 - \left\|x_i - c_g\right\|_2^2 + \gamma\right] \quad (8)$$

식(7)은 일반적으로 학습시킬 때 활용되는 경사 하강 기법을 사용할 수 있다. L_c에 의한 오차는 식(8)의 c_{yi}와 가까워 질수록 지수적으로 작아지고 L_g에 의한 오차는 식(8)의 c_g와 멀어 질수록 지수적으로 작아진다. 따라서 모델의 특징이 각 특징의 중심과 너무 가까워 지거나 혹은 특징이 비정상적으로 특징의 중심과 멀어지는 것을 완화시킨다. 이는 학습될 특징이 각 클래스의 중심과 멀어질수록 오차를 키워 식별적 특징이 되도록 함과 더불어 접점이 생길 가능성이 높은 부분인 식(8)의 c_g 주변의 특징은 오차를 키워 다른 클래스간의 밀집을 피하도록 학습을 강제한다. 식(7)의 λ_c와 λ_g는 스칼라 값으로 두 개의 오차함수의 영향을 조정하는데 활용한다. 식(7)의 λ_c 와 λ_g 값이 커질 수록 특징의 분포는 자신의 클래스의 중심에 가까워지며, 0 일 경우에는 softmax 만 활용해 학습한 모델과 결과가 같게 된다. 식(8)의 c_g와 특징의 차 그리고 c_{yi}와 특징의 차 두 개를 오차함수로 추가함으로 각 특징은 자신의 클래스 내의 분산을 줄임과 동시에 아닌 모든 특징의 중심과 멀어지게 학습한다.

하이퍼파라미터α와 β는 클래스의 중심이 학습데이터에 의해 너무 쉽게 변경되지 않도록 설정한 것으로 학습 시 생기는 불균형을 제어하기 위해 추가하였다.

4. 알고리즘 구현과 실험

실험은 Inception-ResNet 기반으로 softmax cross-entropy 오차와 중심점 오차(center loss)을 사용해 학습한 모델과 softmax cross-entropy 오차와 중심점 오차(center loss), 중심확장(center expansion)을 사용해 학습한 모델에 대해 수행하였다. [12] 학습에 사용된 이미지는 Labeled Faces in the Wild (LFW)으로 총 5,749 명의 사람과 13,233 장의 이미지를 사용했다. 학습에 사용할 이미지는 그림 3 과 같이 모두 얼굴 부분만 추출해 잘랐으며, 모든 잘린 얼굴 이미지는 양 눈과 코를 같은 위치로 정렬하는 전처리 과정을 거쳤다.



(그림 3) LFW 자료와 전처리

알고리즘
학습 데이터 $\{x_i\}$.
파라미터 θ_c, W, $\{c_j | j = 1, 2, \ldots, n\}$, c_g.
Learning rate μ.
하이퍼파라미터 α, β, λ_c, λ_g.

while not 수렴 do
$t \leftarrow t + 1$
오차계산 $L^t = L_s^t + L_c^t + L_g^t$
역전파계산 $\frac{\partial L^t}{\partial x_i^t}$ **for each i** $\frac{\partial L^t}{\partial x_i^t} = \frac{\partial L_s^t}{\partial x_i^t} + \lambda_c \cdot \frac{\partial L_c^t}{\partial x_i^t} + \lambda_g \cdot \frac{\partial L_g^t}{\partial x_i^t}$
갱신 W. $W^{t+1} = w^t - \mu^t \cdot \frac{\partial L^t}{\partial W^t} = W^t - \mu^t \cdot \frac{\partial L_s^t}{\partial W^t}$
갱신 c_j. j 마다 $c_j^{t+1} = c_j^t - \alpha \cdot \Delta c_j^y$
갱신 c_g. j 마다 $c_g^{t+1} = c_g^t - \beta \cdot \frac{1}{n} \sum_{j=1}^{n} c_j$
갱신 θ_c. $\theta_c^{t+1} = \theta_c^t - \mu^t \sum_i^m \frac{\partial L^t}{\partial x_i^t} \cdot \frac{\partial x_i^t}{\partial \theta_c^t}$
end while

학습은 mini-batch 를 사용했으며 mini-batch 크기는 90 으로 하였다. 별도로 0.8 의 확률로 Drop-out 기법을 사용하였으며, RMSProp 를 통하여 경사 하강을 수행하였다. RMSProp 의 learning rat 는 0.1 을 사용하였고, weight-decay 는 0.0005 로 하였다. 하이퍼파라미터 λ_c 와 λ_g, γ는 0.1 와 0.1, 0 으로 설정했고 하이퍼파라미터 α와 β는 0.05 와 0.05 로 설정했다.

실험결과 평가는 Labeled Faces in the Wild(LFW) 이미지를 무작위로 짝지어 동일한 사람일 경우와 같지 않은 경우를 평가하였다. 정확도 평가함수는 식 (11) 와 같다.

$$\text{pss} = \text{predict same as same count} \quad (9)$$
$$\text{pdd} = \text{predict different as different count} \quad (10)$$
$$\text{accuracy} = (\text{pss} + \text{pdd}) / \text{testcount} \quad (11)$$

실험결과 60 번 epoch 동안 cross-entropy 오차와 중심점 오차 만을 사용해 학습한 모델은 97.21%의 성능을 보였고, cross-entropy 오차와 중심점 오차, 중심확장을 사용해 학습한 모델은 98.01%의 성능을 보였다.

중심확장 (center expansion)을 조합한 모델은 조합하지 않은 모델에 비해 효율적으로 특징을 학습할 수 있다. 난수로 초기화된 파라미터로 도출한 특징은 학습 초기 무작위로 분포되어 각각 클래스의 중심이 모든 특징의 중심과 같아지게 된다. 그리고 이는 곧 각 클래스의 분산을 줄이는 오차 때문에 모든 특징이 한 점으로 학습 초반에 모이게 된다는 것인데, 추후 식별적 특징을 도출하도록 강제한다고 하더라도 이미 특징이 밀집된 공간 내에서 분포해야 하므로 효율적으로 학습하기 상대적으로 어려워진다. 하지만 본 논문에서 제안한 식(7) 중심확장을 활용하면 모든 특징의 중심으로부터 γ거리 내에는 특징이 도출되지 않도록 강제하기도 하며 또한 특징들이 모든 특징의 중심으로부터 점진적으로 멀어지게 해 밀집 문제를 완화한다. 본 논문에서 얻게 된 분류 성능 향상은 도출된 특징이 충분한 공간 내에서 분류했기에 얻을 수 있었다.

5. 결론

본 논문에서는 모델이 더 식별적 특징을 도출하도록 강제한 모델을 학습시키고, 그에 따른 효과를 실험하였다. 실험 결과 본 논문에서 제안한 중심확장 기법을 추가로 조합한 모델이 기존 대비하여 오차율을 28.67% 더 줄일 수 있었다. 추후 연구로는 더 많은 사람과 이미지를 포함하고 있는 Youtube faces Database 등을 활용해서 본 논문의 성능을 검증할 것이다.

Acknowledgement
본 연구는 산업통상자원부/한국산업기술진흥원의 "지역특화산업육성사업"에 의해 수행되었습니다.

참고문헌

[1] Florent Perronnin, Jorge S´anchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," ECCV2010
[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet classification with deep convolutional neural networks," NIPS2012
[3] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," ICLR2015
[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. "Deep residual learning for image recognition," CVPR2016
[5] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q. Weinberger, "Deep Networks with Stochastic Depth," ECCV2016
[6] Geoffrey E. Hinton, "Learning multiple layers of representation," Trends in cognitive sciences 11(10), Nov. 2007
[7] Yoshua Bengio, "Learning deep architectures for AI," Foundations and trends in machine learning 2(1), 2009
[8] Gary B. Huang, Marwan Ramesh, Tamara Berg, Eric Learned-Miller, "Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments," Workshop on faces in real-life images, Oct. 2008
[9] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," CVPR2006
[10] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet: A unified embedding for face recognition and clustering," CVPR2015
[11] Yandong Wen, Kaipeng Zhang, Zhifeng Li and Yu Qiao, "A discriminative feature learning approach for deep face recognition," ECCV2016
[12] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, Alex Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," CoRR2016