FaceNet:
A Unified Embedding for Face Recognition and Clustering

박태우

Introduction

- 얼굴을 비교하는데 있어 3가지 기준이 있음.

    1) Verification (같은 사람인가?)

    2) Recognition (이 사람이 누구인가?)

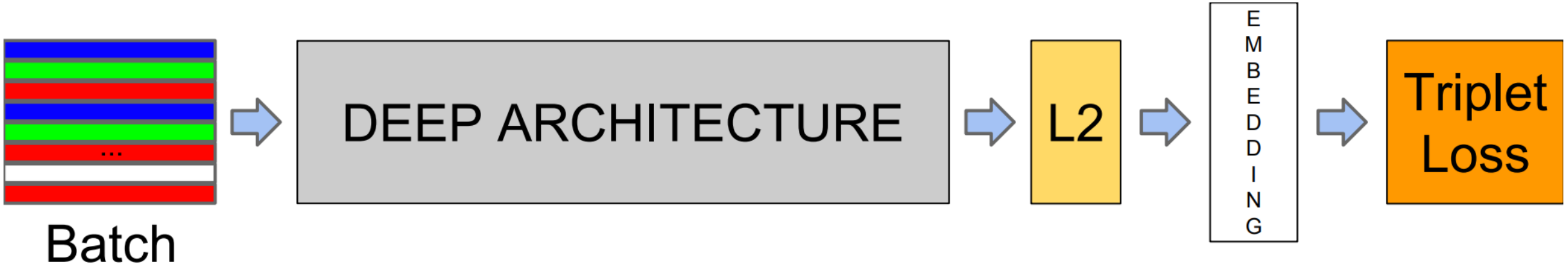
    3) Clustering (유사한 얼굴 찾기)

- Euclidean embedding으로 이미지들을 학습시킴.

- FaceNet은 학습을 기반으로 대상의 feature vector를 이용해 위 3가지를 모두 다룰 수 있도록 함.

- 학습을 위한 방법으로 Triplet loss를 제안.

Architecture



- Batch에 포함된 모든 이미지들을 NN에 통과시켜 d-dimension feature vector를 얻음.

- Feature vector들의 Euclidean distance를 구한 뒤 triplet loss를 계산 및 학습.

- 최종 결과로 서로 유사한 이미지는 거리가 가깝게, 다른 이미지는 거리가 멀게 함.

Triplet Loss

$f(x)$: 이미지 x를 d차원 vector로 표현.

이 때 $\|f(x)\|_2 = 1$을 만족. 즉 모든 embedding들이 원점으로부터 거리가 같도록 함.

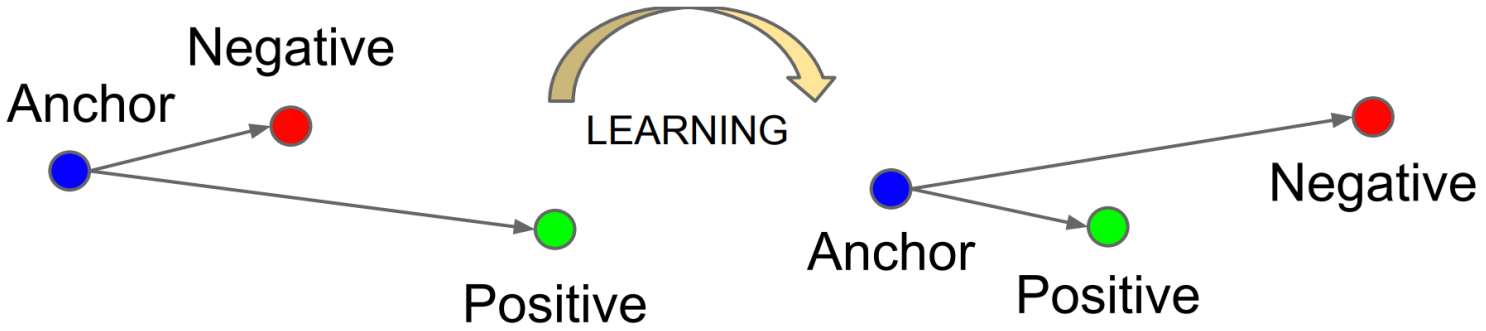
학습을 위해서 3개의 이미지를 한 세트로 묶게 되는데 다음과 같이 구성된다.

Anchor : 기준 이미지.

Positive : Anchor와 같은 인물의 이미지.

Negative : Anchor와 다른 인물의 이미지.

같은 인물은 가깝게, 다른 인물은 멀게 만들도록 학습시킴.



Triplet Loss

앞의 그림은 다음의 공식으로 표현됨.

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$$

여기서 α 는 학습의 효과를 극대화 하기 위한 margin 값.

최종적으로 Triplet loss는 다음과 같이 표현되고 이를 최소화하도록 한다.

$$L = \sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha\right]_+$$

Triplet Selection

본 논문에서는 triplet loss 학습을 위한 triplet 선별 방법을 중요하게 다룸.

학습 하는데 있어 앞장의 공식을 만족하는 triplet을 구성하면 학습의 효과가 없음.

그래서 공식을 만족하지 않는 triplet을 구성해 학습의 효과를 주려고 함.

즉 anchor <-> positive의 distance가 anchor <-> negative의 distance보다 먼 경우를 찾는다.

그리고 그 중에서 가장 차이가 큰 경우를 찾기 위해 다음을 만족하도록 한다.

$$\text{argmax}_{x_i^p} \| f(x_i^a) - f(x_i^p) \|_2^2 \qquad \text{argmin}_{x_i^n} \| f(x_i^a) - f(x_i^n) \|_2^2$$

Triplet Selection

하지만 수십, 수백만장 이미지에서 argmax, argmin을 만족하는 positive, negative를 찾기엔 사실상 불가능함.

그래서 mini-batch 내에서 hard point를 찾는 방법을 통해 계산량, 오버피팅 이슈를 해결함.

Mini-batch를 구성할 때는 hard positive를 뽑기보단, 한 사람당 40개 이미지를 넣고 $|f(x_i^a) - f(x_i^p)|_2^2 < |f(x_i^a) - f(x_i^n)|_2^2$ 를 만족하는 semi-hard를 이용해 학습 초기의 local minima에 빠지지 않도록 함.

Experiments

#dims	VAL
64	$86.8\% \pm 1.7$
128	$87.9\% \pm 1.9$
256	$87.7\% \pm 1.9$
512	$85.6\% \pm 2.0$

Feature vector의 dimension을 높인다고 결과가 좋진 않았음.

#training images	VAL
2,600,000	76.3%
26,000,000	85.1%
52,000,000	85.1%
260,000,000	86.2%

반면, training data의 수가 증가할 수록 정확도는 높아졌음.

참고 링크

https://kangbk0120.github.io/articles/2018-01/face-net

https://www.slideshare.net/ssuser1e0c53/facenet-a-unified-embedding-for-face-recognition-and-clustering

https://wwiiiii.tistory.com/entry/Pairwise-Triplet-Loss