パターン認識 最終レポート
FaceNet: A Unified Embedding for Face Recognition and Clustering

37-196360　森田涼介

2019 年 7 月 30 日

1　論文の概要

FaceNet[1] についての概要を記す。

1.1　サマリー

- Bibliography
  - Paper: FaceNet: A unified embedding for face recognition and clustering
  - Date: 2015-06-12
  - Authors: Florian Schroff, Dmitry Kalenichenko, James Philbin
  - Published in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
- Keywords
  - face verification
  - face recognition
  - face clustering
  - deep convolutional network
  - triplet loss
- Model
  - Name: FaceNet
  - Input: Face Image (e.g. $224 \times 224 \times 3$)
  - Output: Embedding (128-D)
  - Network: CNN (e.g. Inception)
  - Loss: Triplet Loss
- Dataset
  - Labelled Faces in the Wild (LWF)
  - YouTube Faces
- Points of view
  - DCNN で顔画像を 128 次元の embeddings に変換

- embedding 間の Euclid 距離が，顔の類似度に対応
  - 距離が近いほど似ている
  - 顔認証，顔認識，顔画像のクラスタリングにそのまま使える
- Triplet Loss を用い，同一人物の顔画像の embeddings は近く，異なる人の embeddings は遠くなるように訓練
- Triplet の作り方を工夫し，loss の収束に効くような組み合わせを選ぶ
- 2 画像が同一人物かどうかの二値分類の正解率は，LWF が 99.63%，YouTube Faces が 95.12% で SOTA を達成

1.2 理論

画像 x の embedding を $f(x)$ $(\in \mathbb{R}^d)$ と表す。f は x を d 次元の Euclid 空間に写像する。また，embedding を d 次元の超球上に制限する（$||f(x)||_2 = 1$）。ある人の顔画像 x_i^{a} (anchor) とその人の他の画像 x_i^{p} (positive) との距離が，他の人の画像 x_i^{n} (negative) との距離より小さくなるように学習したい。式で表すと，

$$||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{p}})||_2^2 + \alpha < ||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{n}})||_2^2 \tag{1}$$

$$\forall \big(f(x_i^{\mathrm{a}}),\, f(x_i^{\mathrm{p}}),\, f(x_i^{\mathrm{n}})\big) \in \mathscr{T} \tag{2}$$

ここで，α はマージンで，$\mathscr{T}$ は訓練データにおける triplet の組み合わせである（$||\mathscr{T}|| = N$）。損失関数は，

$$\mathscr{L} = \sum_i^N \left[||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{p}})||_2^2 - ||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{n}})||_2^2 + \alpha\right]_+ \tag{3}$$

となる。

Triplet の選択について議論する。考えうる全ての triplet の組み合わせを生成しても，式 (1) が満たされているような組み合わせは学習に貢献しないため，ただ学習時間を増加させるだけになってしまう。そこで，式 (1) を満たさないような trpilet を選択することを考える。つまり，ある anchor 画像 x_i^{a} に対して，次のような x_i^{p}, x_i^{n} を選ぶ。

$$x_i^{\mathrm{p}} = \arg\max_{x_i^{\mathrm{p}}} ||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{p}})||_2^2$$

$$x_i^{\mathrm{n}} = \arg\min_{x_i^{\mathrm{n}}} ||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{n}})||_2^2$$

しかし，$\arg\max$, $\arg\min$ を全訓練データに対して計算するのは，計算時間やデータの質（ラベルの誤りや画像の質）を考えれば妥当ではない。そこで，ミニバッチの中で $\arg\max$, $\arg\min$ を考える。このとき，ミニバッチの中には，1 人につき 2 枚以上の画像が必要である。ただし，anchor-positive に関しては，$\arg\max$ でなく全組み合わせを考えてもよい。また，$\arg\min$ となるような x_i^{n} を選ぶと学習初期に悪い局所最適解，特に $f(x) = 0$ に収束してしまうことがある。そのため，

$$||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{p}})||_2^2 < ||f(x_i^{\mathrm{a}}) - f(x_i^{\mathrm{n}})||_2^2 \tag{4}$$

を満たすような x_i^{n} を選ぶとよい。

図 1: モデルの構造

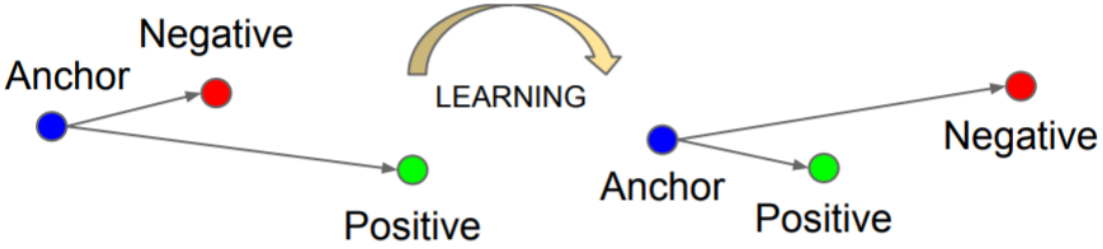

図 2: Triplet Loss

2　扱う問題と評価指標

顔認識・顔認証・顔画像のクラスタリングといった問題を扱う。顔認識・顔認証は，セキュリティや犯罪捜査など様々な場面で重要となる技術である。本論文では，顔画像を embedding に変換し，それらの距離を測ることでこれらの問題を解決している。すなわち，embedding 同士が近ければ似ており，遠ければ似ていないというようにして解いている。学習には Tripet Loss を用いていて，同一人物の顔画像の embedding を近く，異なる人物の顔画像の embedding を遠くするように最適化する。

評価方法について考える。x_i と x_j の間の距離を $D(x_i, x_j)$，同一人物かどうかの閾値を d とし，同一人物の画像のペアを $\mathscr{P}_{\text{same}}$，異なる人物の画像のペアを $\mathscr{P}_{\text{diff}}$ とする。同一人物のペアのうち，正しく同一人物であると判定された，true accepts のペアの集合を次のように表す。

$$\mathrm{TA}(d) = \left\{(i,\ j) \in \mathscr{P}_{\text{same}},\ D(x_i,\ x_j) \leq d\right\} \tag{5}$$

また，異なる人物のペアのうち同一人物であると判定された，false accepts のペアの集合を次のように表す。

$$\mathrm{FA}(d) = \left\{(i,\ j) \in \mathscr{P}_{\text{diff}},\ D(x_i,\ x_j) \leq d\right\} \tag{6}$$

これらを用いて，次の 2 つの評価指標を導入する。

$$\mathrm{VAL}(d) = \frac{|\mathrm{TA}(d)|}{|\mathscr{P}_{\text{same}}|}, \qquad \mathrm{FAR}(d) = \frac{|\mathrm{FA}(d)|}{|\mathscr{P}_{\text{diff}}|} \tag{7}$$

$\mathrm{VAL}(d)$ が高く，$\mathrm{FAR}(d)$ が低いほど性能が良い。

3　論文の選定理由

画像の embedding を得るためには大規模データセットで分類問題を解き，必要に応じて転移学習をしたり最後の層を除いたりといったことをすることが多い。それに対しこの論文では，Euclid 距離がそのまま画像の類似度に対応するような embedding を直接得る手法を提案している点が新しく，また応用の幅も広く面白い。特に，顔画像のようにクラス数がかなり多いときは，分類問題として解くよりもこのような学習方法の方が学習時のパラメータが少なくて済む点も，軽量化などの点で有用である。また，1 クラスに対するデータ数が ImageNet などの分類問題に比べれば少なくて済み，クラス数が一定である必要もないという特長もある。

このような手法は深層計量学習として語られるが，顔認証・顔認識以外の分野では例えばファッションアイテムの embedding を得る部分に用いられている。これらのように，クラス数が一定でなく，また細かい特徴を捉える必要のあるような分野に対しては，深層計量学習は特に有用性が高いといえる。

深層計量学習の先行研究としては Siamese Net (Contrastive Loss) がある。しかし，Triplet Loss はそれよりも性能が安定して高くなりやすく，深層計量学習におけるブレイクスルーになったとも言える点が，この論文を選んだ理由の大きなところである。

4　講義への意見・コメント

内容はわかりやすく，宿題の難易度も少し重めで勉強になりました。また，テーマも広く扱っていて，どんな手法があるのかを抑えることができました。ただ，資料の数式のフォーマットを整えて頂けるともっと見やすくていいと思いました。

参考文献

[1] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.