3Q1-12in1

深層距離学習における contrastive loss の分析と高速化

A Speed-up Method of the Contrastive Loss for Deep Metric Learning

櫻井 隆平 *1　松見 匠規 *2　李 周浩 *1
Ryuhei Sakurai　Shoki Matsumi　Joo-Ho Lee

*1立命館大学 情報理工学部
College of Information Science & Engineering, Ritsumeikan University

*2立命館大学大学院 情報理工学研究科
Graduate School of Information Science and Engineering, Ritsumeikan University

The contrastive loss is a loss function for training neural networks to learn appropriate distance function between two points of input. We propose the soft contrastive loss that speeds up the learning with the contrastive loss. Through experiments of writer identification task on offline handwriting images, we confirmed that the training with our proposed method was much faster than the training with the ordinary contrastive loss and reached comparable accuracy.

1. はじめに

画像認識における顔認識，person re-identification，オフライン手書き文書の書き手識別のような人物識別タスクでは，機械学習により識別器を構築する際に，構築時に利用できる訓練データセットと評価時に用いるテストデータセットとの間で，人物クラス集合が異なっているという問題設定を扱うことがある．このような問題設定では，あらかじめ定義されたクラス集合を対象とする多クラス分類器を構築するという，非常に確立された手法を適用することができないため，最近傍法にもとづく検索問題に落とし込む手法が主流である．最近傍法では画像同士の良い距離をいかにして定義するかに焦点があてられる．

特に近年では，畳み込みニューラルネットワーク（CNN）を用いて距離関数を学習させるという手法（深層距離学習）により，一定の成功が得られている．深層距離学習において提案されている損失関数の中でも代表的なものとして contrastive loss (CL) [1] がある．これは古くから知られた手法であり，また現代でも用いられているが，学習が遅く訓練に長時間を要するという難点がある．

そこで本研究では，CL による訓練を高速化するための soft contrastive loss (SCL) という手法を提案する．また，提案手法をオフライン手書き文書画像の書き手識別タスクに適用し，CL により訓練した場合とほぼ同等の精度が得られる CNN を短時間で訓練できることを示す．

2. 関連研究

深層距離学習では contrastive loss (CL) [1] と triplet loss (TL) [2] が代表的な損失関数であり，家具の検索 [3] や顔認識 [4, 5] など応用事例も多い．

近年ではペアや 3 つ組ではなく，ミニバッチ内の全ての点間の距離や分布を考慮した損失関数が相次いで提案されてきており，非常に効率的な訓練が可能となっている [6, 7, 8]．しかし本研究の焦点は CL の高速化であるため，それらとの関連については議論しない．

連絡先: 櫻井隆平，立命館大学情報理工学部，滋賀県草津市野路東 1-1-1，077-561-5238，rsakurai@fc.ritsumei.ac.jp

3. 提案手法

初めにニューラルネットワークを用いた距離学習の概要を述べる．まず 2 枚の画像間の適切な相違度（逆類似度）を表現する距離関数を獲得する．そのうえで，その距離にもとづく最近傍法により分類や検索を行う．しかし，距離関数を直接的に定義することは，評価時に非効率的である（あるクエリに対して全ての DB 中のインスタンスとのペアについて関数を適用する必要があるため）．そのため通常は，NN を用いて非線形な座標変換を学習させ，入力 2 点の変換先の空間におけるユークリッド距離が，元の空間での適切な距離になっているようにする．この変換先のベクトル表現を埋め込みベクトルと表記する．そのような NN を獲得するために，クラスラベルの付与された画像からなるデータセットを用いて，次のような損失関数を最小化する．すなわち，2 枚の画像のペアが同じクラスに属しているがそれらの埋め込みベクトルの距離が大きいこと，あるいは異なるクラスの画像の埋め込みベクトル同士の距離が小さいことを損失とする．

$X := [0,1]^{W\times W}$ をサイズが $W\times W$ の画像の集合，$Y := \mathbb{R}^D$ を D 次元ベクトル，$f_\theta : X \mapsto Y$ をパラメタ θ を持つ畳み込みニューラルネットワークとする．$x, x' \in X$ を 2 枚の画像，$y, y' = f_\theta(x), f_\theta(x')$ をそれらの埋め込みベクトルとし，x と x' の距離を $D(x, x') := \|y - y'\|_2$ と定義する．ここで，$\|\cdot\|_2$ は L2 ノルムである．訓練データセット $\{(x_1, c_1), \ldots, (x_N, c_N)\}$ を用いて NN を訓練する．ここで N は全事例数，$c_i \in \{1, \ldots, C\}$ は $x_i \in X$ のクラスラベルであり，クラス数は C とする．以下に具体的な損失関数を記述する．

3.1 Contrastive Loss

データセット中の 2 つの事例 (x, c) と (x', c') について，もし x と x' が同じクラス（すなわち $c = c'$）であるときこれを正ペアと呼び，そうでない（すなわち $c \neq c'$）なら負ペアと呼ぶ．contrastive loss (CL) [1] は埋め込みベクトルのペアと，そのペアの正負を表す指示変数に対して，次のように定義される:

$$L(d, t) := td^2 + (1 - t)\max(0, margin - d)^2 \qquad (1)$$

ここで，$d := \|y - y'\|_2$，$t := \delta(c, c') \in \{0, 1\}$ であり，$\delta(\cdot, \cdot)$ はクロネッカーのデルタである．また正の定数 $margin$ は超

パラメタである．訓練においては，データセットからいくつかの事例ペアをランダムに抽出し，上記損失関数をパラメタ θ について最小化する．

3.2 Soft Contrastive Loss による CL の高速化

実験で示されるように，CL を用いた訓練は精度の良い解を与えるものの，学習速度が遅い．そこで，学習を高速化させるための手法として，*soft contrastive loss* (SCL) による訓練を提案する．SCL を説明する前に，正ペア増大 (positive pair augmentation, *P-aug*) を導入する．

3.2.1 positive pair augmentation

データセットから構成できる全ての可能なペアにおける，正ペアと負ペアの比率を考える．各クラスに属する事例数は，クラス数に対して相対的に小さいため，負ペアの数は正ペアよりもおよそ C 倍多い．そのため，全てのペアから一様な確率でペアを抽出すると，正ペアの出現する確率は $1/C$ 程度となる．この不均衡のために，正ペア同士を近づけるような更新の頻度が低いことが，CL による訓練が遅い原因となる．そこで，不均衡の補正を目的として，正ペアと負ペアの出現頻度を同程度にするために，正ペアを仮想的に C 倍に複製する．この補正を P-aug と呼ぶ．P-aug により学習が高速化されるはするが，しかし訓練データへの過学習が引き起こされる．

3.2.2 soft contrastive loss

Soft contrastive loss (SCL) は，高速化のための P-aug が引き起こす過学習を抑制するための損失関数である．CL 関数 (1) は，0 又は 1（すなわち負ペア又は正ペア）の値をとる指示変数 t に対して定義されているが，$t \in [0,1]$ の実数に緩和しても有効である．これは，正ペアであるとことの程度を連続的な度合いにより表現するソフト指示変数とみなすことができ，一方の CL における指示変数は，ハードな割り当てを表現している．

ここで，類似度関数 $S : X \times X \mapsto [0,1]$ を定義し，$t = S(x,x')$ により事例ペアのソフト指示変数を与える．具体的には，事前訓練した多クラス分類器の出力をソフト目的変数 [9] として，その内積を類似度関数として用いる．入力 x に対して，ソフトマックス関数を適用する前のロジットを出力とする分類器を $g_\phi : X \mapsto \mathbb{R}^C$ とする．この分類機は距離学習に用いるものと同じデータセットを使って事前に構築しておく．また，ソフト目的変数を $z = \mathrm{softmax}(g_\phi(x)/T)$ により定義する．ここで，T は温度超パラメタであり，z の平滑化の度合いを表す．最後に，事例ペア x と x' についてソフト指示変数を $S(x,x') := \langle z, z' \rangle$ により定義する．ここで，$\langle \cdot, \cdot \rangle$ は内積である．

SCL は，CL を用いた訓練に加えて正則化の効果がある．前述の観点からみると，CL における 0 または 1 の値をとる指示変数は，ハード目的変数 z としてクラス c への所属を表す one-hot ベクトルを用いていることに相当する．この場合は，正ペアについての損失は 0 に向かおうとする．つまり各クラスについて正ペアの埋め込みベクトルは 1 点に集中するようになる．これが原因で，P-aug を用いたときに過学習が起きる．

一方，ソフト指示変数の使用は，そのような極端な解が回避されるように働く．図 1 は，埋め込みベクトル 2 点間の距離 d についての CL 関数とその導関数を，指示変数 t が 0，1 とその中間それぞれの値についてプロットしたものである．なお *margin* は 0.2 である．青色と緑色のプロットがそれぞれ $t = 0$ と $t = 1$ に対応しており，その中間としてソフト指示変数により対応しているものがその他のプロットである．ソフト指示変数が比較的大きい値（例えば紫色の $t = 0.75$）の場合

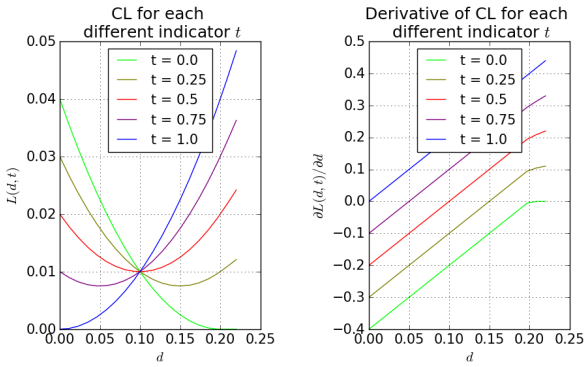

図 1: ソフト指示変数 t が 0 から 1 の間のいくつかの値をとったときの（左）CL 関数と（右）その導関数．与えられた画像ペアの類似度 (すなわち t) に応じて，それらの埋め込みベクトル間の距離の収束先が決定される．

は，SCL 関数は小さい d に最小値をとる（例えば $d = 0.05$）．これは，SCL 関数は正ペアの埋め込みベクトル間の距離を近づけるが，その極小点を超えるほど近づけはしないことを意味する．そのため，P-aug を用いたときにも過学習を生じさせない．

4. 実験

まず実験の目的と結果を概説する．

距離学習の未知クラスへの汎化能力: 距離学習により得られる埋め込みベクトルが，最近傍法による検索性能の観点で優れた特徴表現となっていることを確かめる．結果として，得られた特徴表現は，訓練データ中に出現していないクラスにおいても良い性能を発揮した．

SCLP の有効性の検証: 4 種類の訓練方法について精度と学習速度を比較する．contrastive loss (CL), CL に P-aug を適用したもの (CLP), 提案手法である soft contrastive loss に P-aug を適用したもの (SCLP) と，比較手法としての Triplet loss (TL) である．結果として，CL が最も良い精度であったが，SCLP はほとんど精度を悪化させずに顕著な高速化を達成した．

4.1 手書き文書画像データセットと性能評価指標

オフライン手描き文書画像データセットの CVL データベース [10] を対象とした，書き手識別タスクのための距離学習を行う．CVL データベースは 311 人の書き手による 1604 枚の手書き文書画像により構成される．訓練セットとテストセットがあり，訓練セットは 27 人の書き手の各々7 枚（すなわち全部で 189 枚）の手書き文書画像からなり，テストセットは 283 人の書き手の各々5 枚（すなわち全部で 1415 枚）の手書き文書画像からなる．書き手にはいくつかの制約が課されている．7 種類の規定のテキストがあり，罫線付きの下敷きを用いて書かれた文書はベースラインが直線的に揃っている．それ以外の要素，例えば文字の大きさ，改行の位置，行間の幅，余白の大きさなどは書き手の自由である．図 2 に画像の実例を示す．

訓練時には未知のクラスからなるテストデータセットに対して，最近傍法を用いた分類・検索精度により性能を評価する．具体的には，[10] に定義されている評価指標として，ソフト評価，ハード評価，検索評価を用いる．まず，訓練済み CNN に

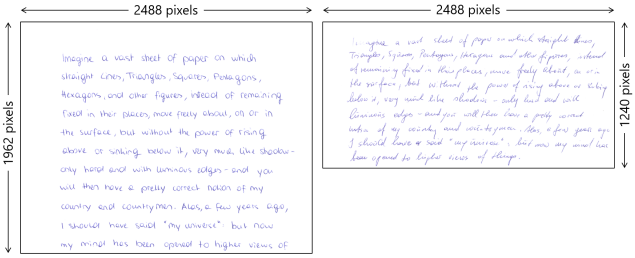

図 2: CVL データベースの 2 枚の手書き文書画像．テキスト内容は同じだが，異なる書き手によって書かれているため，文字や行間の幅などのスタイルが異なる．

より全てのテスト用画像を埋め込みベクトルに変換する．次に，全ての画像ペアの距離を，埋め込みベクトルのユークリッド距離により算出する．各画像ごとに，その他の画像を距離の小さい順にソートすることで，類似度順を得る．その上で，各評価指標を以下のように定義する．

4.1.1 ソフト評価指標

あるクエリ画像に対し，類似度の上位 N 以内にクエリ画像と同クラスの画像が少なくとも 1 つあればソフト一致とし，全テスト画像のソフト一致率をソフト評価指標とする

4.1.2 ハード評価指標

あるクエリ画像に対し，類似度の上位 N 以内の全ての画像がクエリ画像と同クラスであればハード一致とし，全テスト画像のハード一致率をハード評価指標とする

4.1.3 検索評価指標

あるクエリ画像に対し，類似度の上位 N 画像のうちクエリ画像と同クラスであるものの割合を再現率とし，全テスト画像の再現率の平均を検索評価指標とする

4.2 実験設定の詳細

4.2.1 訓練セットとテストセットの交換

CVL データベースは公式的に訓練セットとテストセットを分割して提供している．しかし訓練セットの画像枚数が少なく，効果的な訓練が難しいため，本研究ではテストセットで訓練し，訓練セットを評価用として用いた．

4.2.2 画像の前処理

実験では，オリジナル画像とともに配布されている，上下余白削除済み画像を用いた．さらに左右の余白を自動的に除去し，画像サイズを縦横半分に縮小し，グレースケール化したのち白黒反転した（すなわち黒の背景に白のストローク）．このように前処理した文書画像から，正方形に切り出した部分画像を CNN への入力とする．訓練時にはランダム位置から切り出すことでデータ増大する．

4.2.3 ネットワークの構成と最適化

実験では，全ての訓練手法で同一のネットワーク構成（表 1）を用いた．特に，入力は 198×198 サイズのグレースケール画像で，出力は 50 次元の L2 正規化ベクトルである．全ての畳み込み変換と全結合変換（表 1 の Conv と FC）の後にバッチ正規化 [11] と ReLU を適用した（ただし最後の FC を除く）．

ミニバッチ SGD と *Adam* [12] によりネットワークのパラメタを最適化した．すべての訓練手法でミニバッチ数は 20，$margin$ は 0.2，学習率は 10^{-5} とし，Adam の超パラメタはデフォルト値を用いた (すなわち $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$)．

表 1: CNN の構成

Type	Kernel size, stride	Output channels, size
Input	-	1, 198
Conv	3, 1	50, 196
Max Pool	2, 2	50, 98
Conv	3, 1	100, 96
Conv	1, 1	100, 96
Max Pool	2, 2	100, 48
Conv	3, 1	200, 46
Conv	3, 1	200, 44
Max Pool	2, 2	200, 22
Conv	3, 1	400, 20
Conv	1, 1	400, 20
Max Pool	2, 2	400, 10
Conv	1, 1	400, 10
Conv	1, 1	400, 10
Average Pool	10, 1	400, 1
FC	-	400
FC	-	50
L2 Normalization	-	50

表 2: テストセットにおけるソフト評価指標 (%)

Method	Top 1	Top 2	Top 5
Baseline	93.2	97.5	99.4
TL	98.8	99.4	**100**
CL	**100**	**100**	**100**
SCLP	99.4	**100**	**100**

表 3: テストセットにおけるハード評価指標 (%)

Method	Top 2	Top 3	Top 4
Baseline	89.5	80.2	**71.0**
TL	92.0	82.1	70.4
CL	**95.7**	**87.7**	69.1
SCLP	93.2	83.3	67.3

表 4: テストセットにおける検索評価指標 (%)

Method	Top 2	Top 3	Top 4
Baseline	93.5	90.5	87.5
TL	95.7	92.2	89.4
CL	**97.8**	**95.1**	**90.1**
SCLP	96.6	93.2	88.0

4.3 実験結果

4.3.1 各訓練手法の性能評価

表 2，3，4 は CL，TL とベースライン手法のテストセットに対する性能評価である．ここで，ベースライン手法として，分類器の最後の隠れ層を特徴ベクトルに用いた．分類器は距離学習と同じ訓練セットにより訓練され，ネットワーク構成も出力層を訓練セットに含まれるクラス数である 282 次元としたことを除いて同一である．なお，CL と TL では，訓練中にテストセットにおける最高のソフト評価指標を記録した際のパラメタにより，その他の評価指標も評価した．すなわち，ホールドアウト検証などは行っていない．

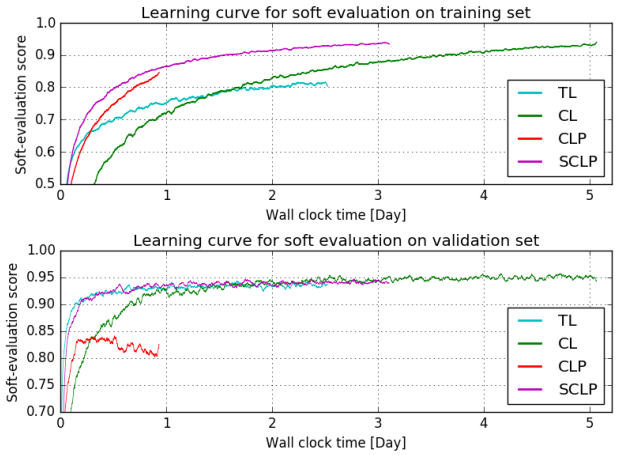

図 3: （上）訓練セットと（下）テストセットにおける学習曲線の比較．横軸は実時間で縦軸はソフト評価指標（大きいほど良い）．triplet loss (TL，水色), contrastive loss (CL，緑), CL with P-aug (CLP，赤) と提案手法の SCL with P-aug (SCLP，紫)．見やすさのために移動平均により平滑化している．

実験の結果，CL，TL，SCLP の全てで一貫してベースライン手法の性能を上回った（ただしハード Top4 評価指標を除く）．この結果から，距離学習に基づく手法が，訓練時には未知のクラスに対しても良い特徴表現を与えることが確認できる．事前に定義されたクラス集合に対して識別的に訓練された分類器の中間表現よりも，距離学習により得られた特徴表現のほうが，最近傍法ベースの分類に適用する際に適していることがわかる．また，CL のほうが一貫して TL，SCLP よりも良い性能であった．さらに，提案手法の SCLP は TL を僅差であるが上回った．ただし，TL では hard negative mining を行っているため，実装の簡便さの点でも SCLP に利点がある．

4.3.2 各訓練手法における学習曲線の挙動

訓練中の学習の進み具合をみるために，図 3 に異なる訓練手法同士の学習曲線を比較する．訓練セット（上）とテストセット（下）の両方で，序盤において TL は急速に学習が進行しており CL は低速であるが，最終的には CL が TL を上回る様子が確認できる．CLP は，TL と同様に序盤において CL よりも急速に学習が進行するが，過学習によりテストセット上のスコアがすぐに悪化しており，性能は良くない．提案手法の SCLP は，CLP と同様に急速な立ち上がりであるが，過学習が回避されている．

4.4 議論

実験結果から，以下のような観察が得られた．

正ペアの出現確率を上げると序盤の学習速度が高速化する: 訓練時に正ペアと負ペアが等確率で出現する全ての手法に比べて，CL が低速であった結果からこのことが確認できる．

正ペアを過剰に近づけようとすると過学習が生じる: SCP では過学習が生じ，SCLP ではそれが生じなかった結果からこのことが確認できる，

5. まとめ

本研究では，深層距離学習における contrastive loss 関数を用いた訓練を高速化するための soft contrastive loss 関数を提案した．実験を通して，提案手法による顕著な高速化が確認された．十分な訓練時間を経た最終的な精度では CL が SCL を上回るが，しかし概ね匹敵する精度であった．

参考文献

[1] Chopra, Sumit, Raia Hadsell, and Yann LeCun. ”Learning a similarity metric discriminatively, with application to face verification.”, CVPR 2005.

[2] Schroff, Florian, Dmitry Kalenichenko, and James Philbin. ”Facenet: A unified embedding for face recognition and clustering.”, CVPR. 2015.

[3] Bell, Sean, and Kavita Bala. ”Learning visual similarity for product design with convolutional neural networks.”, SIGGRAPH 2015.

[4] Hu, Junlin, Jiwen Lu, and Yap-Peng Tan. ”Discriminative deep metric learning for face verification in the wild.”, CVPR 2014.

[5] Parkhi, Omkar M., Andrea Vedaldi, and Andrew Zisserman. ”Deep face recognition.” British Machine Vision Conference. Vol. 1. No. 3. 2015.

[6] Oh Song, Hyun, et al. ”Deep metric learning via lifted structured feature embedding.”, CVPR 2016.

[7] Sohn, Kihyuk. ”Improved deep metric learning with multi-class n-pair loss objective.”, NIPS 2016.

[8] Song, Hyun Oh, et al. ”Learnable Structured Clustering Framework for Deep Metric Learning.” arXiv preprint arXiv:1612.01213 (2016).

[9] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. ”Distilling the knowledge in a neural network.” arXiv preprint arXiv:1503.02531 (2015).

[10] Florian Kleber, Stefan Fiel, Markus Diem and Robert Sablatnig, CVL-Database: An Off-line Database for Writer Retrieval, Writer Identification and Word Spotting, In Proc. of the 12th Int. Conference on Document Analysis and Recognition (ICDAR) 2013, pp. 560-564, 2013.

[11] Ioffe, Sergey, and Christian Szegedy. ”Batch normalization: Accelerating deep network training by reducing internal covariate shift.” arXiv preprint arXiv:1502.03167 (2015).

[12] Kingma, Diederik, and Jimmy Ba. ”Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980v8 (2014).