H-038

複数の代表点を用いた SDMM に基づく半教師あり行動認識手法
Semi-supervised Action Recognition with Multiple Centers based on SDMM

寺尾 颯人† 野口 渉‡ 飯塚 博幸‡* 山本 雅人‡*
Hayato Terao Wataru Noguchi Hiroyuki Iizuka Masahito Yamamoto

1. はじめに

行動認識とは，動画に含まれる一連の動作から被写体がどのような行動をしているかを推測する問題である．近年，ニューラルネットワークの発展に伴って行動認識の精度は大きく向上してきている．

しかし，これまで提案されてきた行動認識手法の多くは教師あり学習であり，学習に多くの教師ありデータを必要とする．一般に，データのラベル付けは人手でおこなわれるため，大量の教師ありデータを作成するためには時間がかかる．この問題を解決できる手法として，半教師あり学習がある．半教師あり学習は教師ラベルが付与された教師ありデータに加えて，教師ラベルが付与されていない教師なしデータを学習に利用することで，学習に必要な教師ありデータの数を減らすことができる．

ニューラルネットワークを用いた半教師あり学習の一つとして，教師なしデータに学習途中のネットワークの予測に基づいた疑似ラベルを付与し，それを教師として扱うという方法がある．例えば，半教師あり画像分類で提案された Mean Teacher [1]は学習中のパラメータについて指数平滑移動平均をとり，そのパラメータから予測される出力を疑似ラベルとして用いる．

また，ニューラルネットワークを用いない半教師あり行動認識手法として，Zengman らが提案した，Semisupervised Discriminant Multimanifold Analysis (SDMM) がある [2]．SDMM による学習を図 1(a)に示す．SDMM は通常の分類学習とともに，近傍でありながらも異なる行動クラスに属する動画同士の出力は離れるように，近傍かつ同じ行動クラスに属する動画同士の出力は近づくように学習した関数を用いて判別分析をおこなう．

本研究では Mean Teacher と SDMM の学習を組み合わせることで，ニューラルネットワークを用いた半教師あり行動認識手法を提案する．SDMM の学習は教師ありデータが特徴量空間で明瞭に分離するよう促すため，より正確な疑似ラベルを得られる可能性がある．提案手法では，SDMM と同様の学習をおこなえる SoftTriple Loss[3]を用いる．SoftTriple Loss を図 1(b)に示す．代表点は各行動クラスの動画を代表する点である．ネットワークは入力された動画から得られた特徴量を近傍の動画ではなく，代表点と比較して遠ざけたり近づけたりする．また，動画では同じ行動クラスであっても個人差や環境などが原因で動きが大きく変化することがある．このとき，特徴量も変化してしまうため，一つの点がすべての行動を代表してしまうと学習が難しくなる．この問題に対処するために代表点を複数用意し，特徴量が大きく変化してしまう場合でもうまく学習できるようにする．

† 北海道大学大学院情報科学院
Graduate School of Information Science and Technology, Hokkaido University
‡ 北海道大学人間知・脳・AI 研究教育センター
Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University
＊北海道大学大学院情報科学研究院
Faculty of Information science and Technology, Hokkaido University

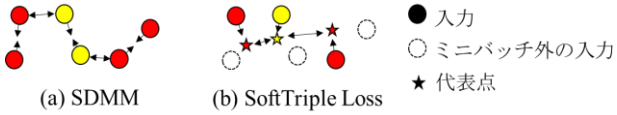

図 1 SDMM と SoftTriple Loss による学習方法の比較

2. 提案手法

本研究では，SoftTriple Loss と Mean Teacher を組み合わせた新しい学習手法を提案する．SoftTriple Loss を用いることで，先述したように SDMM と同様の学習をおこない，異なる行動クラスの教師ありデータを特徴量空間上でより明瞭に分離することができる．それによって，Mean Teacher で用いる疑似ラベルをより正確なものにできる．

2.1 SoftTriple Loss

教師ありデータとして，入力と教師ラベルのペア$(\boldsymbol{x}, y)$を考える．また，行動クラスの数をCとおく．まず，ネットワークを用いて入力$\boldsymbol{x}$を$\boldsymbol{h} \in \mathbb{R}^{C\times K}$に変換する．

$$\boldsymbol{h} = f(\boldsymbol{x}, \theta, \eta). \tag{1}$$

ただし，θはネットワークの重み，ηは訓練時に加えるランダムなノイズを表す．このとき，$\boldsymbol{h}$の要素$h_{c,k}$を$\boldsymbol{x}$とc番目の行動クラスに属するk番目の代表点との類似度とみなす．次に，行動クラスcの中で$\boldsymbol{x}$との類似度が最大の代表点を抽出する．抽出された代表点と$\boldsymbol{x}$の類似度S_cは次のように求められる．

$$S_c = \max_k h_{c,k}. \tag{2}$$

この類似度が$c = y$の場合はより大きく，$c \neq y$の場合は小さくなるように学習する．これによって，間接的に$\boldsymbol{x}$と同じ行動クラスに属する代表点は$\boldsymbol{x}$に近づき，異なる行動クラスに属するにも関わらず$\boldsymbol{x}$と近い代表点は$\boldsymbol{x}$から離れる．このような学習のために，スケールファクターλとマージンδを用いた以下の損失関数を定義する．

$$\ell_{SoftTriple} = H\left(\frac{exp(\lambda(\boldsymbol{S} - \delta\boldsymbol{p}))}{\sum exp(\lambda(\boldsymbol{S} - \delta\boldsymbol{p}))}, \boldsymbol{p}\right). \tag{3}$$

ただし，$\boldsymbol{S}$はS_cをc番目の要素として持つベクトル，$\boldsymbol{p}$はyを one-hot 表現したベクトル，$H(\boldsymbol{v}, \boldsymbol{w})$は確率分布$\boldsymbol{v}$と$\boldsymbol{w}$のクロスエントロピー誤差を表す．しかし，式(2)は不連続な max を用いているため学習が難しい．そのため，SoftTriple Loss は式(2)の代わりに，連続な以下の式を用いる．

$$S'_c = \sum_k \frac{exp(h_{c,k}/\gamma)}{\sum_k exp(h_{c,k}/\gamma)} h_{c,k}. \tag{4}$$

ただし，γはS_cとS'_cの値を近づける度合いを制御するハイパーパラメータである．

表 1　正答率の比較 [%]

%Label	10	20	30	40	50
提案手法 (*K*=3)	27.76±1.88	35.80±1.07	41.61±1.19	**45.59±0.82**	**48.66±0.92**
提案手法 (*K*=5)	27.82±2.11	**36.21±0.87**	41.39±0.82	45.54±0.68	48.56±0.82
提案手法 (*K*=10)	**27.99±2.09**	**36.21±0.98**	**41.71±1.30**	**45.59±0.71**	48.39±0.66
Mean Teacher	26.52±2.03	34.91±1.58	40.47±1.87	44.85±0.77	47.91±0.85
教師あり学習	16.60±2.10	24.55±1.61	31.02±1.57	35.56±1.18	38.44±1.16

一方，行動クラスごとに最適な代表点の数Kは異なる．しかし，半教師あり学習のもとで最適な代表点の数を決定することは困難である．そこで，Kの値を大きくとった上で以下の値を最小化するような学習も同時におこなう．

$$\ell_{regularize} = \frac{\sum_c^C \sum_i^K \sum_j^K \left\| \boldsymbol{w}_{c,i} - \boldsymbol{w}_{c,j} \right\|_2}{CK(K-1)}. \quad (5)$$

ただし，$\boldsymbol{w}_{c,k}$はネットワークの出力層において$h_{c,k}$を出力するニューロンの重みである．もし$\boldsymbol{w}_{c,i}$と$\boldsymbol{w}_{c,j}$が十分近くなれば，$h_{c,i}$と$h_{c,j}$は常に等しくなるため二つの代表点を一つの代表点にマージしたとみなせる．

2.2 Mean Teacher

Mean Teacher は半教師あり画像分類で提案された学習方法である．教師なしデータとして入力$\boldsymbol{u}$を考える．次に，2.1 節と同様の計算をおこない，類似度S'_cを得る．このとき，入力$\boldsymbol{u}$に対してネットワークが予測する行動クラスcの確信度q_cは次のようになる．

$$q_c = \frac{exp(\lambda S'_c)}{\sum_j exp(\lambda S'_j)}. \quad (6)$$

ここまでの計算を$\boldsymbol{q} = [q_1, \dots, q_C] = g(\boldsymbol{u}, \theta, \eta)$と記述したとき，教師なしデータに対する損失関数$\ell_{MeanTeacher}$を次のように定義する．

$$\ell_{MeanTeacher} = \| g(\boldsymbol{u}, \theta, \eta') - g(\boldsymbol{u}, \theta', \eta'') \|_2. \quad (7)$$

ただし，θ'は以下の式によってニューラルネットワークの学習がおこなわれるごとに更新される．

$$\theta' \leftarrow \alpha\theta' + (1-\alpha)\theta, \quad (8)$$

最終的に，提案手法の損失関数ℓは次のようになる．

$$\ell = \ell_{SoftTriple} + \beta_1 \ell_{MeanTeacher} + \beta_2 \ell_{regularize}. \quad (9)$$

ただし，β_1, β_2はハイパーパラメータである．

3. 実験

提案手法の有効性を評価するために提案手法と SoftMax Loss を用いた一般的な Mean Teacher および教師あり学習との比較実験をおこなった．また，提案手法において代表点の数が正答率に与える影響を明らかにするために，代表点の数Kが 3,5,10 と変化した場合についても比較した．

3.1 データセット

本研究では，データセットとして HMDB51 を用いた．HMDB51 は 51 種類の行動のいずれかに属する 6,766 個の動画で構成されている．著者らによってこれらの動画を訓練データとテストデータに分割する 3 種類の split が提供されているが，本研究では split1 を用いた．これらの動画はネットワークの入力として，[2]と同様に improved dense trajectories (IDTs)に基づく hand-crafted features に変換した．

3.2 実験設定

本研究で使用したネットワークは 4 層ニューラルネットワークで，中間層は入力層から順番に 512 次元，256 次元の値をそれぞれ出力する．活性化関数は tanh で，中間層で 50%の dropout を適用した．また，データ水増しとして入力に正規分布$N(0, 0.003)$からサンプルされたノイズを加えた．バッチサイズは教師ありデータ，教師なしデータ共に 32 とし，オプティマイザは Adam を用いた．学習回数は教師なしデータを基準として，300 エポックだけ学習をおこなった．また，HMDB51 はすべてのデータに対してラベルが付与されているが，半教師あり学習をおこなうために訓練データを一定の割合で教師ありデータと教師なしデータに分割し，教師ありデータの数だけラベルを用いた．訓練データに対する教師ありデータの割合は 10%, 20%, 30%, 40%, 50%とし，それぞれの場合で実験した．その他のハイパーパラメータは，$\lambda = 20, \gamma = 0.1, \delta = 0.01, \alpha = 0.999, \beta_1 = 1, \beta_2 = 0.2$とした．学習は 5 回ずつおこない，正答率の平均と分散を算出した．

3.3 実験結果

結果を表 1 に示す．まず，提案手法と教師あり学習を比較すると正答率が大幅に向上している．教師あり学習は教師なしデータを使わずに学習していることから，提案手法は教師なしデータを有効に活用できていることがわかる．また，提案手法は Mean Teacher と比較した場合にも正答率が向上している．このことから，SoftTriple Loss による学習が半教師あり学習に対して有効に働いたといえる．加えて，代表点が増えるにつれて正答率が向上する傾向もみられた．しかし，その傾向はわずかであり，代表点の数が結果に大きく影響することはなかった．これは，用いたデータセットでは個人差や環境による動きの変化が少なく，最適な代表点の数が小さかったためだと考えられる．

一方，SDMM は訓練データの約 20%に相当する 765 個のデータを教師ありデータとして学習し，44.07%の正答率を達成している．そのため，より教師なしデータから行動認識に役立つ情報を抽出できるような改良が必要である．

4. おわりに

本研究では，SoftTriple Loss と Mean Teacher を組み合わせた半教師あり行動認識手法を提案した．この手法は SoftTriple Loss を使わない通常の Mean Teacher と比較して高い正答率を出せることを実験によって示した．

謝辞

本研究の成果の一部は，独立行政法人情報通信研究機構(NICT)の委託研究 「データ連携・利活用による地域課題解決のための実証型研究開発(第 2 回)エッジコンピューティング環境を利用した動物のリアルタイム自動行動分類システムの開発」により得られたものです．

参考文献

[1] Tarvainen Antti, Harri Valpola “Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results”, Advances in Neural Information Processing Systems 30 (2017).

[2] Xu Zengmin, Ruimin Hu, Jun Chen, Chen Chen, Junjun Jiang, Jiaofen Li, Hongyang Li, “Semisupervised Discriminant Multimanifold Analysis for Action Recognition”, IEEE Transactions on Neural Networks and Learning Systems, Vol.30, Issue.10 (2019).

[3] Qian Qi, Lei Shang, Baigui Sun, Juhua Hu, Hao Li, Rong Jin, “SoftTriple Loss: Deep Metric Learning Without Triplet Sampling”, Proceedings of the IEEE International Conference on Computer Vision (2019).