H-031

畳み込みニューラルネットワークを用いた顔表情分類の実験的評価

Experimental Evaluation of Facial Expression Classification Using Convolutional Neural Networks

高橋 佑典 * 松川 徹 † 鈴木 英之進 †
Yusuke Takahashi Tetsu Matsukawa Einoshin Suzuki

1. はじめに

人の顔表情は国や文化圏が異なっても共通であり，相手の顔を見ることでコミュニケーションをとることができる [1]．そこで本実験は顔の表情に注目し，ロボットが人の表情を認識できるように 100 人顔画像のデータベースを用いて学習を行う [2]．100 人顔画像データベースは，成人 100 人が 25 種類の表情をした顔画像から構成されている．

顔画像の学習方法には，深層学習 [3] を用いた．深層学習とは，多層のモデル構造を用いて，データ内にある特徴について多段の組み合わせを考慮することで，高次の概念を学習する機械学習の方法論である．深層学習を用いた画像認識の方法は，設計した特徴を用いず，多数の層の学習によってデータから本質的なデータ，すなわち特徴を抽出できる．深層学習のライブラリの中で最も処理が速く，活発に開発されている Caffe[4]‡ を用い，顔表情の学習に適したモデルを実験により調べる．

本論文の構成を述べる．第 2 節では，本研究の基盤となっている畳み込みニューラルネットワークについて説明すると共に，深層学習を行うツールである Caffe とそのパラメータの説明を行う．第 3 節と第 4 節では，Caffe に 100 人顔画像のデータベース [2] を用いた学習を行い，実験結果と考察を示す．

2. 予備知識

2.1. 畳み込みニューラルネットワーク

これより参考文献 [5] に従い，畳み込みニューラルネットワークを説明する．画素数 $n_x \times n_x$ の画像 $\mathbf{x}$ に対する，画素数 $n_w \times n_w$ のフィルタ $\mathbf{w}$ の畳み込みを考える．この畳み込みを $\mathbf{h} = \mathbf{x} * \mathbf{w}$ と書くと，出力 $\mathbf{h}$ は $n_h \equiv n_x - n_w + 1$ のサイズの画像になる．畳み込みニューラルネットワークでは通常，複数のフィルタ $\mathbf{w}^1, \ldots, \mathbf{w}^L$ から構成され，フィルタごとに異なる出力 $\mathbf{h}^1, \ldots, \mathbf{h}^L$ を得る．

畳み込みニューラルネットワークは，この入力 $\mathbf{x}$ から出力 $\mathbf{h}$ への計算に対応する配線を基本構造とする．具体的には，入力画像の画素とユニットが 1 対 1 で対応する入力層に対し，その $n_w \times n_w$ の部分集合とすぐ上の第1 層のユニット 1 つが結合され，その結合重みが畳み込むフィルタ $\mathbf{w}$ に相当する．畳み込みニューラルネットワークの特徴は，層間の結合が位置ごとに局所的になされ，その結合重みが上位層のユニット間で共有されていることである．

これに加えて畳み込みニューラルネットワークでは，その次の層でプーリングと呼ばれる処理が行われる．プーリングとは，畳み込みの出力の解像度を落とす処理であり，その際に位置情報を捨てることで微小な位置変化に対する不変性を実現する．プーリング方法にはいくつかのバリエーションがある．**平均プーリング** (average pooling) では，

$$h'_i = \frac{1}{|P_i|} \sum_{j \in P_i} h_j \tag{1}$$

のように $k-1$ 層での小領域 P_i での応答の平均を k 層のニューロン i の値とする．**マックスプーリング** (max pooling) では，

$$h'_i = \max_{j \in P_i} h_j \tag{2}$$

のように，$k-1$ 層での小領域 P_i 内の応答の最大値を k 層の値とする．

平均プーリングは，直下小領域 P_i の応答の総和に揺れなどがなければ，出力は特徴に関する情報を損なうことなく解像度を落とすことができ，マックスプーリングは，直下小領域 P_i 内の応答がスパースな特性を持っている場合，無駄な入力を高効率に分離できる．なおマックスプーリングの誤差逆転伝播計算では，順伝播時にプーリング領域で選択された最大値を出力したユニットを記憶しておき，逆伝播時に利用することで計算を行う．

畳み込みニューラルネットワークはこのフィルタの畳み込みとプーリングをこの順で何度か繰り返す下図 1 のような構造をとる．つまりフィルタ出力層，プーリング層の 2 層を，プーリング層を次の層の入力層とする形で積み重ねる．プーリングはユニット数を減少させるので，上に行くほど層のユニット数が小さくなる．最終層には，クラス数に応じた数のユニットを持つ出力層を置く．

畳み込みニューラルネットワークは，フィルタとプーリングをどのように行うかによってそのネットワーク構造が決まる．具体的にはまず，フィルタとプーリング層の数，各層でのフィルタのサイズおよびフィルタ出力数を決める必要がある．入力画像が大きい場合，ネットワークの自由度を低下させるために，フィルタを全画素で適用せず，間引いた位置で適用することもある．その場合にはその間隔 (stride) を決める必要がある．プーリング層では，下層のプーリングする範囲，プーリングの間隔 (stride) を定める．

2.2.Caffe

Caffe (Convolutional Architecture for Fast Feature Embedding) は畳み込みニューラルネットワークのライ

*九州大学 大学院システム生命科学府, Graduate school of System Life Sciences, Kyushu University
†九州大学 大学院システム情報科学研究院， Faculty of Information Science and Electrial Engineering, Kyushu University
‡http://caffe.berkeleyvision.org/

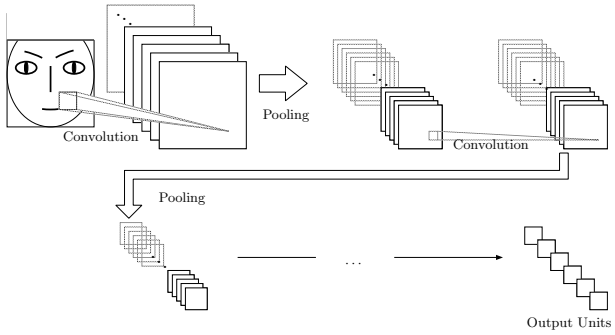

図 1: 畳み込みニューラルネットワーク

ブラリのひとつであり，画像認識において高い性能を示す．Yangqing Jia 氏が制作し， Berkeley Vision and Learning Center (BVLC) によって拡張されている [4]．

実装が難しい深層学習を，Caffe を用いることで比較的簡単に行うことができる．GPU と CPU の使用も簡単な設定で切り替えることが出来る．Caffe は NVIDIA K40 を用いた場合，1 日に 4000 万枚の画像を処理することができる．つまり画像 1 枚の学習に 5ms，画像 1 枚のテストに 2ms しか必要とせず，既存の深層学習のライブラリの中で，畳み込みニューラルネットの実行が最も速い．

2.3.100 人顔画像データベース

100 人顔画像データベースは，成人 100 人が図 3 のように，25 種類の表情をした顔画像から構成されている [2]．100 人中男性が 66 人，女性が 34 人という構成になっている．国籍は，中国人 65 人，インドネシア人 12 人，日本人 18 人，マレーシア人 1 人，シンガポール人 1 人，スウェーデン人 2 人とベトナム人 1 人となっている．

図 2: 100 人顔画像データベース:25 表情

データベースの各画像は表 (1) の表情クラスに分類され，ラベルが付けられている．

本実験においては，この 25 表情の中から，参考文献 [1] において，6 つ感情が全人類に普遍的であることから，幸福，悲しみ，驚き，恐怖，怒り，嫌悪の 6 表情を用いる．

表 1: 25 表情の表情名およびクラス番号

表情	クラス番号
Happy:幸福	1
Sad:悲しみ	2
Pleased:感謝	3
Angry:怒り	4
Confused:困惑	5
Tired:疲れた	6
Shocked/Surprised:驚いた	7
Irritated:いらいらした	8
WTF:理解不能	9
Triumph:得意な/勝ち誇った	10
Fear:恐怖	11
Bereft:希望を失った/悲しく孤独な	12
Flirty:チャラチャラした/気がある	13
Serious:真面目な	14
Silly:馬鹿な/愚かな	15
Hollow/Blank:うかつな	16
Incredulous:疑い深い	17
Confident:自信がある	18
Fierce:凶暴な	19
Despondent/Pouty:意気消沈した	20
Drunk:酔っ払った	21
Rage:激怒した	22
Sarcastic:皮肉な	23
Disgusted:嫌悪	24
Ill/Nauseous:気分が悪い	25

図 3: 100 人顔画像データベース:6 表情

3. 実験

100 人顔画像のデータベースを Caffe で学習できるようにするために，学習用データセット，テスト用データセットを作成した．テスト用データセットに 10 人の顔画像を格納し，学習用データセットに残りの 90 人の顔画像を格納した．1 人の 1 表情あたり，10 枚の顔画像を用意した．同時に，学習用とテスト用のラベルテキストを作成した．学習が終わった後，テスト用のディレクトリに格納する 10 人を変更し，同様の作業を行い 10 交差検定によって実験結果を出した．

データセットに格納する画像は，あらかじめ加工した画像を用いた．それぞれの画像を 80*80 にリサイズし，グレースケールにした後，ヒストグラムの平坦化を行った．さらに全ての画像を 5°，10° 回転させた画像と，それらを反転させた画像を加えた．

図 4 にある基本モデルは，参考文献の畳み込みニューラルネットワークによる表情認識のモデルを参考にし

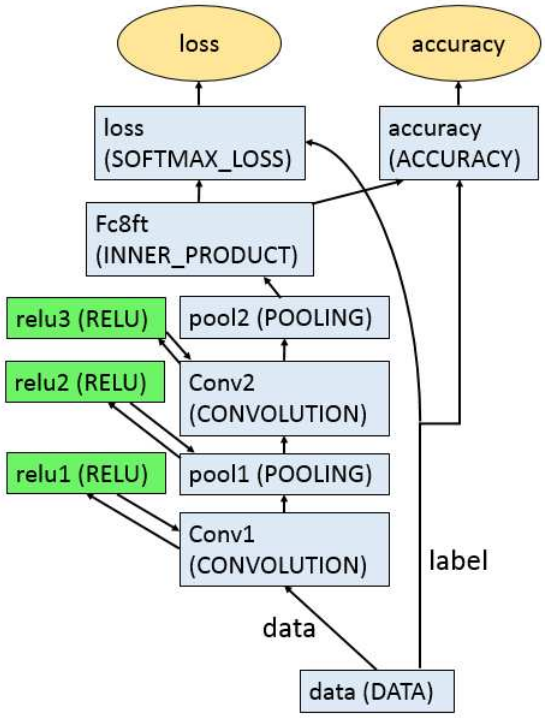

図 4: 実験に用いる基本モデル

て構成した [6].

3.1. 学習率の補正の有無

周期ごとに学習率に補正値ガンマ γ を乗ずることで，初期学習の早さを変更することなく，フィルタの過学習を防止する [6]．ガンマ $\gamma = 0.1$ とし，学習率に補正をかけなかった場合と，周期 step_size で学習率にガンマ γ を乗じ補正した場合の精度を測定した．

3.2. フィルタ枚数の変更

畳み込み層のフィルタ枚数を変更し，変更した場合の精度を測定した．フィルタの枚数を変更することで，学習する特徴の数を変更することができる．フィルタの枚数を増やすことで，より多くの特徴を学習できるが，パラメータ数が増加するので，学習の時間も増加する．実験により，顔表情の認識に有効なフィルタの枚数を調べる．前節の実験結果より，これより以下の実験においては周期 600 で，学習率にガンマ $\gamma = 0.1$ をかける．

3.3. フィルタサイズの変更

畳み込み層のフィルタサイズを変更し，変更した場合の精度を測定した．フィルタサイズを変更することで，学習する特徴の大きさを設定することができる．これによって顔表情の認識に有効な特徴を学習できるフィルタの大きさを調べた．顔表情フィルタ枚数は前節の実験結果より，畳み込み層 1 と畳み込み層 2 共に 8 枚とした．

3.4. 畳み込みニューラルネットワークと LBP の比較

本実験では，深層学習である畳み込みニューラルネットワークを用いて分類学習を行った．この手法を用いることで，設計した特徴を用いず，分類に有効な特徴を自動的に抽出できる．LBP(Local Binary Pattern)[7] によって特徴抽出を行い，LIBSVM[8] を学習装置として使用して表情分類を行った場合と，畳み込みニューラルネットワークで表情分類を行った場合で性能を比較した．

3.5. 実験条件

プーリング方法の 1 つであるマックスプーリングは，画像認識において高い性能を示す [9]．そこで本実験では，プーリングはマックスプーリングを用いる．

使用した GPU: GeForce GTX 980

4. 実験結果

4.1. 学習率の補正の有無

下図 5 は，学習率の補正周期を変更し，繰り返し学習させたときの精度を示している．補正周期が 10000 とは，学習率に補正をかけない場合の精度の変化を示している．

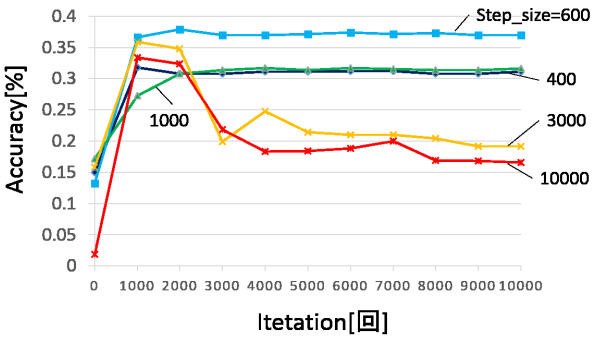

図 5: 学習率補正の周期を変更したときの精度

図 5 について，step_size=10000 のとき，つまり学習率に補正を加えなかった場合，精度は Iteration=1000 を境に減少し，約 17%まで減少している．クラス数が 6 個であるため，この精度は学習によってなにも成果が得られていないことを示している．step_size=3000 のとき，精度は補正を加えなかった場合とほとんど同じ変化だったが，学習率が下がった Iteration=3000 後にわずかに精度が上昇し，約 20%に収束した．step_size=1000 以下のとき，精度は上昇したのち，一定の値に収束した．

4.2. フィルタ枚数の変更

図 6 は，フィルタ枚数を変更し，繰り返し学習させたときの，精度を示している．モデル定義ファイルの num_output はフィルタの枚数を示しているので，畳み込み層 1 と畳み込み層 2 の num_output を変更した．図の num_output が 4*8 であれば，第 1 層のフィルタが 4 枚，第 2 層のフィルタを 8 枚に設定していることを表す．

下図 6 より，フィルタの枚数を多くしたとき，学習率が大きく下がった．原因として，フィルタの枚数を多くしたことでネットワークの自由度が増加したことによる過学習が考えられたので，学習率に補正をかけ

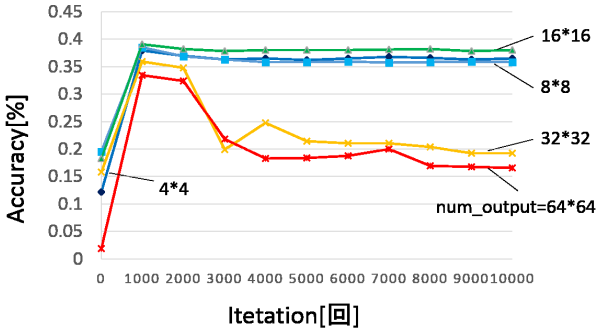

図 6: フィルタの枚数を変更したときの精度

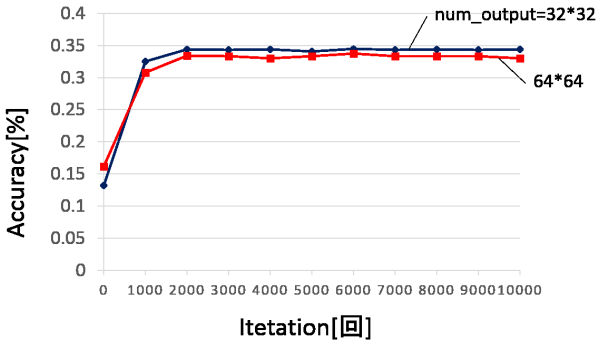

図 7: 学習率の step_size=400 のときの精度

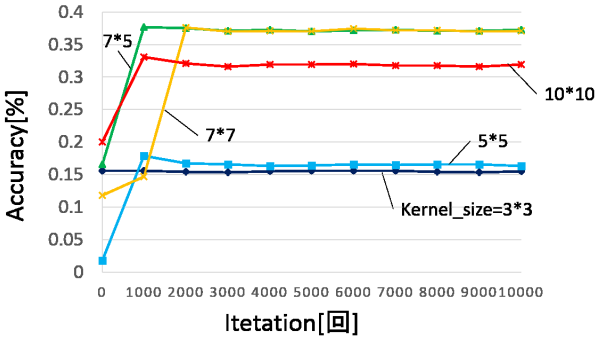

図 8: フィルタのサイズを変更したときの精度.1

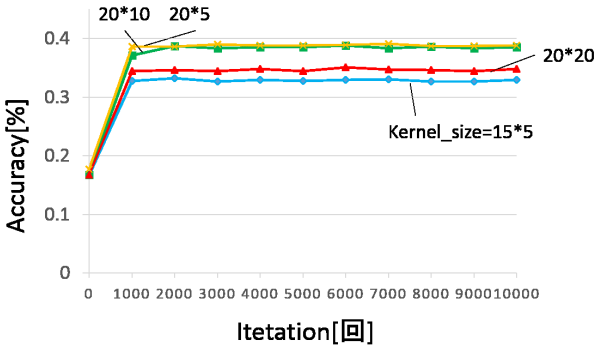

図 9: フィルタのサイズを変更したときの精度.2

る周期 step_size を 600 から 400 にしたところ，下図 7 のように精度が改善された．

4.3. フィルタサイズの変更

図 8 と図 9 は，フィルタのサイズを変更し，繰り返し学習させたときの精度を示している．モデル定義ファイルの kernel_size はフィルタのサイズを示しているので，畳み込み層 1 と畳み込み層 2 の kernel_size を変更した．図の kernel_size が 7*5 であれば，第 1 層のフィルタサイズを 7*7，第 2 層のフィルタサイズを 5*5 に設定しているということである．

図 8 のように，フィルタサイズを小さくすると，精度は約 16%になり，うまく学習がされなくなった．フィルタサイズが 3*3 であるとき，学習した畳み込み層 1 のフィルタは下図 10 のようになり，畳み込み層 1 の出力は図 12 となった．図 12 は入力画像 11 からほとんど変化しておらず，畳み込み層 1 が特徴を得られていないため精度が低くなったことがわかる．

4.4. 畳み込みニューラルネットワークと LBP の比較

表 2 は畳み込みニューラルネットワークと LBP の精度を示している．1 列目は，学習データセットとテストデータセット 画像を無作為に選出した場合，2 列目は学習データセットとテスト データセットに同一の人物を用いなかった場合の精度を表している．

畳み込みニューラルネットワークを用いた分類学習は，画像を無作為に選出した場合と，学習とテストに同一人物を使用しなかった場合のどちらにおいても，LBP よりも高い精度を示した．

表 2: 畳み込みニューラルネットワークと LBP の精度

手法	画像を無作為に選出した場合	学習とテストに同一人物を使用しなかった場合
深層学習	0.974	0.420
LBP	0.872	0.235

5. 結論

本研究では深層学習のライブラリである Caffe を用いて，人の 6 種類の表情を学習するのに適したモデルを調べた．実験の結果より，6 クラスの学習において，30～40%の精度を示した．学習画像を事前に加工することについては，反転画像の追加によって精度が向上した．LBP を用いた比較手法の精度が約 24%であったのに対して畳み込みニューラルネットワークは特徴を与えることなく高い精度差を示したといえる．

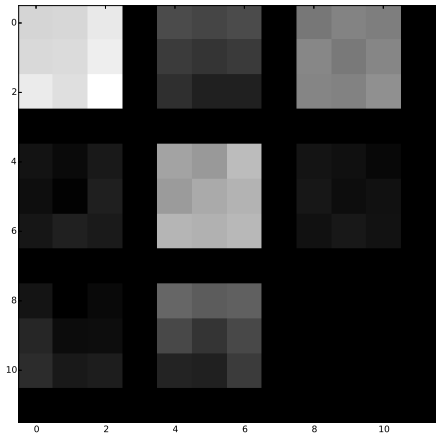

図 10: フィルタサイズが 3*3 のときの畳み込み層 1 のフィルタ

図 11: 学習したモデルに入れる画像

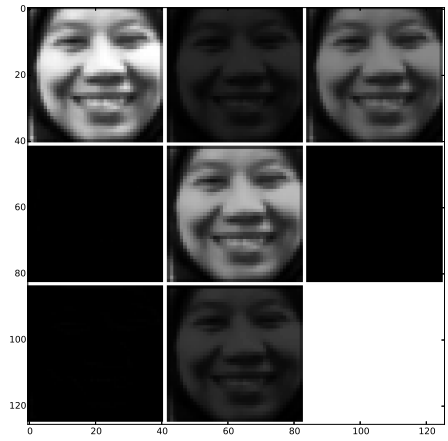

図 12: フィルタサイズが 3*3 のときの畳み込み層 1 の中間出力

参考文献

[1] W.V. フリーセン P. エクマン 著. 表情分析入門:表情に隠された意味をさぐる.

[2] A. Erna, L. Yu, K. Zhao, W. Chen, and E. Suzuki. Facial Expression Data Constructed with Kinect and their Clustering Stability. *Active Media Technology,Lecture Notes in Computer Science 8610 (AMT 2014)*, 2014.

[3] L. Deng and D. Yu. Deep Learning: Methods and Applications. *Signal Processing*, 7:3–4.

[4] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[5] 竹島 由里子 金谷 健一 日野 英逸 村田 昇 岡谷 貴之斎藤 真樹 藤代　一成, 高橋　成雄. コンピュータビジョン最先端ガイド.

[6] S. Lawrence, C.L. Giles, A.C. Tsoi, and A.D. Back. Face Recognition: A Convolutional Neural-Network Approach. *IEEE Transactions on Neural Networks*, 8(1):98–113, 1997.

[7] C. Shan, S. Gong, and P.W. McOwan. Facial Expression Recognition Based on Local Binary Patterns: A Comprehensive Study . *Image and Vision Computing*, 27(6):803–816, 2009.

[8] C.C Chang and C.J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), Article 27, 2011.

[9] A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *NIPS*, 2012.