

深層学習による動画像からの表情認識手法の開発

10 月 9 日 (水)

小松 大起

1 動画像による表情認識を行う深層学習モデル

1.1 CNN

CNN は、畳み込み層とプーリング層を 1 つのペアーとし、それらが複数回重ね合わせて構成される順方向性ニューラルネットワークである。ここで、対象とする画像を $X \times Y$ pixels の RGB の階調値とし、 k 番目の階調の素子 (i, j) の画素値を $I_{ij}^{(k)}$ とする。ただし、 $k = 1$ が R, $k = 2$ が G, $k = 3$ が B とする。最初の層の畳み込み層の a 番目のフィルターの (i, j) 番目の素子の内部状態を $y_{ij}^{(1)(a)}$, その出力を $\hat{y}_{ij}^{(1)(a)}$, プーリング層の出力を $z_{ij}^{(1)(a)}$ とすると、各々以下のように与えられる。

$$y_{ij}^{(1)(a)} = \sum_{k=1}^3 \left(\sum_{x \in W} \sum_{y \in W} w_{ij}^{(1)(a)(k)} I_{i+x, j+y}^{(k)} + b_{ij}^{(1)(a)(k)} \right) \quad (1)$$

$$\hat{y}_{ij}^{(1)(a)} = \max(y_{ij}^{(1)(a)}, 0) \quad (2)$$

$$z_{ij}^{(1)(a)} = \max_{x \in W, y \in W} \hat{y}_{i+x, j+y}^{(1)(a)} \quad (3)$$

ここで、 $w_{ij}^{(1)(a)(k)}$ は入力層と畳み込み層間のシナプス結合加重、 W は各素子が入力を受ける範囲を与える配置集合（受容野）、 $b_{ij}^{(1)(a)(k)}$ は閾値である。

ℓ 番目の層の畳み込み層の出力 $\hat{y}_{ij}^{(\ell)(a)}$ 及びプーリング層の出力 $z_{ij}^{(\ell)(a)}$ は式 (??) 及び (??) と同じであるが、 ℓ 番目の層の畳み込み層の内部状態 $y_{ij}^{(\ell)(a)}$ は異なり、以下の式で与えられる。

$$y_{ij}^{(\ell)(a)} = \sum_{\alpha=1}^{N(\ell-1)} \sum_{x \in W} \sum_{y \in W} w_{ij}^{(\ell)(a, \alpha)} z_{i+x, j+y}^{(\ell-1)(\alpha)} + b_{ij}^{(\ell)(a)} \quad (4)$$

最終層 (L) の内部状態 $y_k^{(L)}$ は、前層のプーリング層の出力 $z_{ij}^{(L-1)(a)}$ との全結合として、以下のように与えられる。

$$y_k^{(L)} = \sum_{\alpha=1}^{N(L-1)} \sum_i \sum_j w_{kij}^{(L)(\alpha)} z_{ij}^{(L-1)(\alpha)} + b^{(L)k} \quad (5)$$

そして、その出力は、ソフトマックス関数により、以下のように与えられる。

$$\hat{y}_k^{(L)} = \frac{y_k^{(L)}}{\sum_i y_i^{(L)}} \quad (6)$$

1.2 Long Short-Term Memory(LSTM)

LSTM は RNN を拡張したものであり、畳み込み層の内部状態を、以下のように変更した。

$$y_{ij}^{(\ell)(a)}(t) = y_{ij}^{(\ell)(a)} + \sum_{\tau=1}^T \sum_{\alpha=1}^{N(\ell)} \sum_{x \in W} \sum_{y \in W} v_{ij}^{(\ell)(a, \alpha)} y_{i+x, j+y}^{(\ell)(\alpha)}(t-\tau) \quad (7)$$

ここで、 $y_{ij}^{(\ell)(a)}$ は式 (??) であり、 $y_{ij}^{(\ell)(a)}(t)$ は t 回目の学習時の畳み込み層の内部状態の値である。RNN は、LSTM の $T = 1$ に相当する。数値実験では、 $T = 10$ としている。

2 先週までの作業

- 院試の勉強及び、研究についての勉強

3 今週の作業

- 院試の勉強。
- 研究に関する勉強。keras の cifar-10 というデータセットを用いて CNN を行った。

4 来週以降の作業

- 院試の勉強。
- 実験を行う。