非対称空間プーリングを用いた畳み込みニューラルネットワークによる高精度物体位置回帰

富樫陸 † 佐藤育郎 ‡ 中山英樹 †

† 東京大学大学院情報理工学研究科 ‡ デンソーアイティーラボラトリ

1 はじめに

従来の畳み込みニューラルネットワーク (CNN) において，プーリングが行う操作は，特徴マップの解像度を削減しつつタスクに必要な情報を伝搬させることである．画像の分類問題であれば，位置不変性を獲得できる上に計算コストを削減することができるが，画像内の物体位置を pixel wise で回帰するような問題においては，位置情報を落としていることが不利に働くと考えられる．位置情報を保存するために，プーリングを全て除いた CNN を考えることができるが，計算コストの面で非現実的である．そこで，本研究では，従来の画像分類の上で発展してきたディープニューラルネットワークのアーキテクチャを画像内位置回帰のために見直し，実行可能にするための次元削減を行いつつ，必要な情報の欠落を防ぐ手法を開発した．

2 提案手法

本研究では，CNN を二つのサブネットワークに分けて，それぞれ画像の垂直方向，水平方向に対してのみ偏向した非対称プーリングを行うモデルを提案する．図 1 のように，提案手法では一つの同じ画像を 2 つのニューラルネットワーク (CNN) に入力し，それぞれ異なる方向に対してのみプーリングを行っていく．最終的に全結合層と接続して，回帰を行う．本研究ではこのようなモデルと，従来の CNN との比較を行う．

2.1 非対称空間プーリング

従来の CNN におけるプーリングは，2×2 や 3×3 の正方形のプーリングウィンドウによって囲まれた領域に関して，平均や最大値をとるような集約処理を行うというものであるが，提案手法における非対称空間プーリングでは，2×1 や 1×2 のプーリングウィンドウによって集約を行う．

Convolutional neural networks with asymmetric spatial pooling for accurate object position regression
Riku Togashi†, Ikuro Sato‡and Hideki Nakayama†

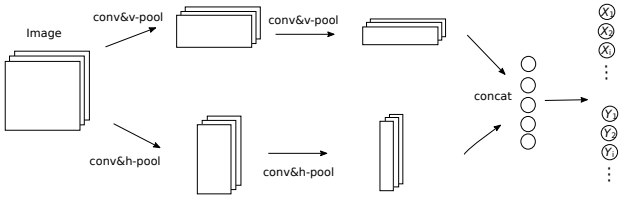

図 1: 非対称空間プーリングを用いた 2 ストリーム畳み込みニューラルネットワーク

2.2 2 ストリーム CNN

提案手法では 2 つの CNN に同時に同じ画像を入力する．図 1 に示したモデルでは，2 ストリーム CNN として全結合層の手前で各 CNN の出力を一つのベクトルとして concatenate する 2 ストリームネットワークとなっている．それに対して，実験的に，各ストリームを完全に独立なモデルとして，非対称プーリングによって保存される次元に関する目的変数だけをそれぞれ回帰する Independent モデルも本研究では比較を行った．

3 実験

3.1 データセット

本研究では，非対称空間プーリングを用いたモデルが従来の CNN に対して，画像内の物体位置の回帰に関する精度においてどのような違いがあるのかを比較する．

Leed sport pose dataset[1] は，Flickr の画像からいくつかのスポーツのタグがついた人間の写っている画像に対して，計 14 関節の位置がアノテーションされているようなデータセットである．このデータセットには 12,000 枚の画像が含まれている．本研究では ，train データ 10,000，test データ 2,000 で訓練，評価する．

3.2 比較するモデルの共通する部品

すべての CNN は，C(11)-P-LRN-C(7)-P-LRN-C(5)-P-LRN-F-F-F という構成である．ただし，C(K) は K×K の畳み込みであり，すべて Zero padding さ

図 2: LSP データセット:人間の 14 関節がアノテーションされている

れている．P は max-pooling， LRN は local response normalization[2]，F は fully-connected 層である．活性化関数はすべて LeakyReLU(α=0.333) を採用した．また，F にはすべて Dropout を 0.5 の割合でかけている．

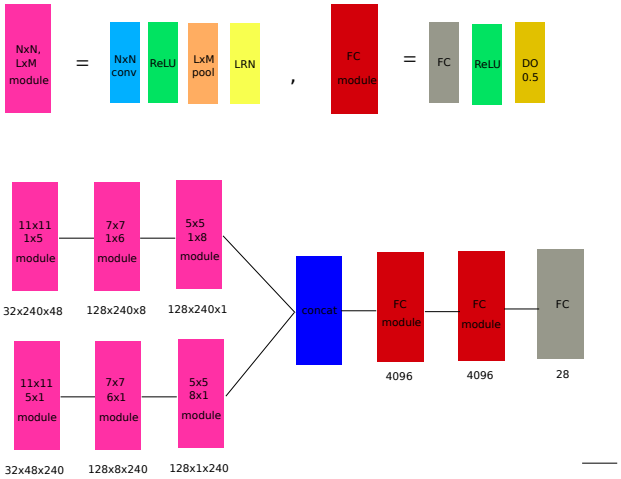

図 3: 2 ストリームモデルの詳細:いくつかの繰り替えされる部品 (module) を上で定義し，下のモデル記述で用いている．各層またはモジュールの下には出力テンソルのサイズを記述している．

3.3 最適化

最適化に関しては公平な比較のため，すべて共通するものを使用している．誤差関数は Mean Squared Error(MSE) で，最適化ソルバとしては Adam を用いる．学習の停止は validation による early stopping で行う．

3.4 実験結果

各モデルの test データに対する誤差関数の値をプロットしたものが図 4 である．表 1 に最終精度を示す．

表 1: 各モデルの比較

手法	誤差 (MSE)
baseline	92.69
Independent	128.12
2-stream	89.79

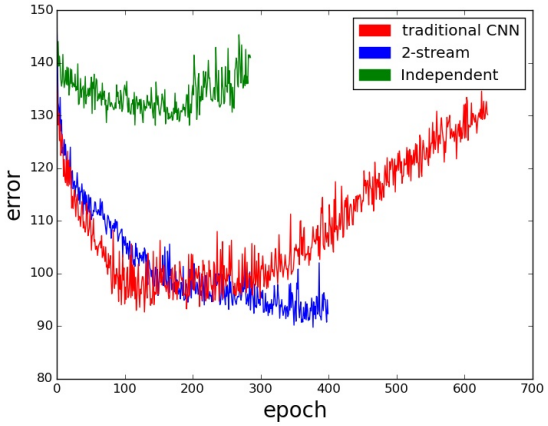

図 4: 各モデルの学習の進行と誤差関数の値の変化

4 考察

図 4 から，Independent は明らかに悪い性能になっている．これは，縦と横の方向に関する情報を共有しないことに起因したものであると思われる．人間の関節位置は関節同士で人間がとることのできる姿勢にいくらか二次元的に拘束されているが，各次元を分解して共有せず学習する Independent はそのような二次元的拘束を明らかに捉えることができない．2-stream は baseline をわずかに最終性能で勝っているが，特に注目したいのは，学習の収束が安定していることである．これは，非対称空間プーリングによる構造的な正則化が成功しているといえる結果である．

参考文献

[1] Johnson, Sam and Everingham, Mark, Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation, BMVC, 2010

[2] Alex Krizhevsky and Sutskever, Ilya and Geoffrey E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS, 2012