6R-03

空間的注意と位置符号化を用いたCNNの位置汎化能力の獲得

昼間 彪吾† 森 裕紀‡ 尾形 哲也†

†早稲田大学基幹理工学部表現工学科 ‡早稲田大学次世代ロボット研究機構

1 序論

画像認識の分野で標準的なアーキテクチャとなっている Convolutional Neural Network（CNN）は、一般物体認識に用いる場合、位置情報を消すために Pooling を行い位置普遍性を確保している。CNN の応用は認識にとどまらずロボットの End-toEnd 制御に用いることもあるが、この場合には物体などの位置情報が決定的に重要である。しかし、CNN は位置に関する汎化能力が低く、様々にして位置を変えたデータを用意する必要があり、データの収集コストが高くなってしまう。本研究では、CNN の位置汎化能力が何故低くなるのかを理論的に示し、改善手法として、CNN と空間的注意機構、位置符号化を組み合わせた手法を提案する。

CNN 層では、オブジェクトを表現する画像特徴量の位置に対応するニューロンが反応するが、隣り合うニューロン同士に連続的な関係性を示す情報を明示的に与えていない。そのため、学習データが疎であり、オブジェクトが空間的に連続に変化するデータが揃っていない場合、CNN の重み共有を行っていたとしても、その後の層で回帰問題を解く場合に学習データになかったオブジェクト位置のニューロンが反応しても対応する出力が不定と解釈されるため、汎化できず誤差が大きくなってしまう。

本研究では、自然言語処理で主に利用されている注意機構 [1] を画像処理に応用することで CNN の位置の汎化性能の獲得の効率化を目指す。画像処理に注意機構を導入した研究として Self-Attention GAN[2] や Non Local Neural Network[3] があるが、いずれも画像内の離れた位置の特徴量同士の関係のモデリングに用いており、位置の汎化性向上には用いられていない。

2 空間的注意機構

本論文の提案モデルである、空間的な注意機構（Spatial Attention、SA）は、ベースとしている注意機構同様、図 1 にあるように Key、Query、Value の役割を持つ表現持ち、それら 3 つを用いて「類似度推定」と「位置符号化」の 2 つの処理を適用する構造となっている。

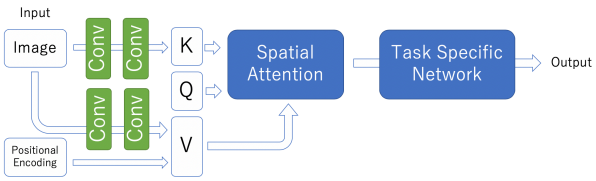


図 1: 空間的注意機構を導入したモデル構造

まず類似度推定では、Key と Query を用いて入力画像内の注意を向ける領域を示す Attention Map を生成する。ここで、Key は入力画像 I_{img} に畳み込み層 f_{key} で抽出した画像特徴量表現 K である。

$$K = f_{key}(I_{img}) \tag{1}$$

Key を Query とチャンネル方向に内積（$\odot$）を取り、Softmax 関数をかけて生成した A_{map} が Attention Map となる。ここで、Query は Key のチャンネル方向の次元数と同次元のベクトル q である。

$$A_{map} = softmax\left(K \odot q\right) \tag{2}$$

次に位置符号化では、求めた Attention Map と Value の積による重み付き平均により、画像内の注目点の位置情報を画像表現から XY 座標に変換した A_{out} を出力する。ここで、Value は幅・高さが Attention Map と等しい画像で、各ピクセルにその位置の XY 座標が配置されている行列 V とし、Positional Encoding と定義する。また、V に画像特徴量表現を連結することで注目位置の特徴量を同時に抽出することが可能である。

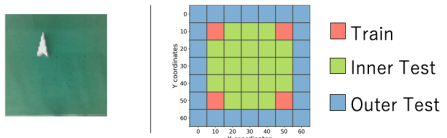
$$A_{out} = A_{map}^{\mathrm{T}} V \tag{3}$$

本手法では、位置符号化処理で Value で特徴量の位置を座標として抽出することで明示的に空間情報を考慮した構造とした。それにより教示位置に限らず、任意に配置された特徴量の位置情報が、線形補間により安定して抽出可能である。

3 実験

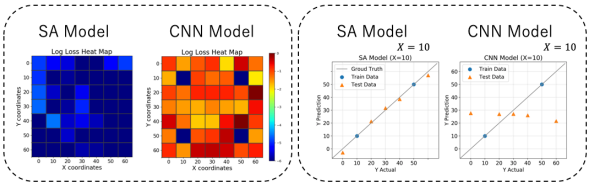
提案する空間的注意機構の有用性の検証のため、カメラ画像内の物体の位置と角度の同時推定タスクを行う。モデルを図 1 に示す。この実験は、例えばロボットアームによるピッキング動作を End-to-End で行う際に、手先の到達位置を決める物体位置（Extrinsic）情報と物体を把持する

Non Localized CNN With Spatial Attention Mechanism
†Hyogo Hiruma, Waseda Unversity The Department of Intermedia Art and Science
†Tetsuya Ogata, Waseda Unversity The Department of Intermedia Art and Science
‡Hiroki Mori, Waseda Unversity Future Robotics Organization



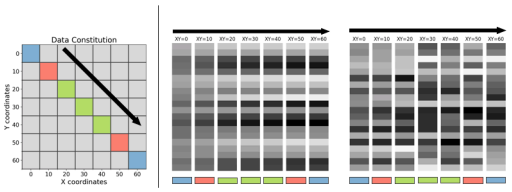
(a) Sample Image Data　(b) Data Structure

図 2: (a) 使用データのサンプル画像、(b) データの構成（グリッド区切りの各セルがデータセット内の座標に対応）



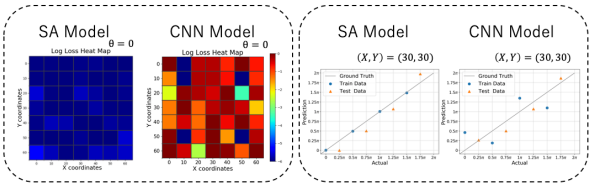
(a) Position Wise Prediction Error　(b) Prediction/Truth Comparison

図 3: 位置推定のモデルごとの比較 (a) 位置毎の位置推論誤差のヒートマップ、(b)X 座標固定時の、Y 座標の予測値と真値の比較



Data Transition Direction　SA Model Hidden Values　CNN Model Hidden Values

図 4: 位置の、空間的に連続なデータ推論時のネットワーク内部表現の推移の連続性の比較



(a) Position Wise Prediction Error　(b) Prediction/Truth Comparison

図 5: 角度推定のモデルごとの比較 (a) 教示方向 $\theta = 0$ を向いているときの各位置での平均角度推定誤差のヒートマップ、(b) 位置固定時の角度の予測度と真値の比較

向きを決める物体方向や物体形状といった（Intrinsic）情報を統合的に扱えるかどうかをテストしている。

実験設定として、緑のステージ上に白い矢印を、図 2 が示す 49 箇所に配置し、各位置で矢印を 45 ℃ずつ回転させた場合の俯瞰カメラの画像と実空間上の座標をデータとして用いる。ここで、学習データは図 2 に示す 4 箇所のみとし、8 方向中 4 方向を学習データに用いる。

ネットワーク構造は図 1 の構造を用い、SA の出力を実世界座標と角度情報へそれぞれ変換する全結合層を 3 層ずつを連結したモデルを使用する。また、畳み込み層 3 層、全結合層 3 層から成る一般的な CNN モデルを位置推定タスクの比較モデルとして用いる。

4　結果と考察

まず位置推定について、位置毎の誤差、予測値と真値の比較を図 3*a*、3*b* に示す。結果、SA モデルは学習データ、テストデータともに高精度で推論が可能であることが示され、位置の汎化が確認できる。

一方、図 3*a*、3*b* に示される比較モデルの一般的な CNN の結果は 4 つの教示位置に過学習し、未教示位置の推論で誤差が高くなるものだった。特に図 3*b* に示すように、テストデータのみ真値から離れた、平均値を出力する傾向にある。

以上の結果を解析する為、ネットワークの内部表現を図 4 に示す。この結果、一般的な CNN がオブジェクトの角度固定にもかかわらず不連続で、離散的な構造となっており、一方 SA モデルは位置の遷移に合わせて中間表現の値が滑らかに変化する、空間情報を考慮した連続的な構造となっていることがわかる。

次に角度推定について、位置毎の誤差、予測値と真値の比較を図 5*a*、図 5*b* に示す。結果、一般的な CNN が教示位置でのみ低い誤差で推論する中、SA は教示・未教示の位置での推論誤差が均等に低い為、任意の位置の特徴量の抽出が可能であることが確認できる。ただし、図 5*b* が示すように一般的な CNN が全データで誤差が大きくなったのに対して、SA は未教示方向のデータのみ誤差が大きくなったことが示される。原因は教示方向が 4 方向と疎であったことだと考えられ、画像特徴量そのものを扱う場合は従来の CNN の同様の傾向を示している。

5　まとめと今後の展望

本研究では、提案した空間的注意機構を用いたモデルが、一般的な CNN モデルと比較して高い位置汎化性能を持ち、実世界データに対しても疎かつ少数のデータで高精度の学習が可能であることが示された。位置の汎化に必要なデータ数の削減は、深層学習によるロボット制御などデータ収集コストが高い領域において特に有用となる。今後は本提案機構を発展させ、複数の注目点を扱う Multi Attention や、ロボット制御への応用の研究をしていく所存である。

参考文献

[1] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” 2014, cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation. [Online]. Available: http://arxiv.org/abs/1409.0473

[2] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” 2018.

[3] X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” 2017.