The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

Convolutional Neural Networkによる画像認識と視覚的説明

Convolutional Neural Network for Image Recognition and Visual Explanation

山下隆義 *1
Takayoshi Yamashita

*1中部大学
Chubu university

Deep learning technologies in the field of computer vision are gradually introducing into our daily life. These technologies have been achieved by introducing the methods that improve recognition performance, such as deeper models, a method of stability training of the deeper model. Moreover, to achieve the productization, the visual explanation that explains the decision making of deep learning to a user has been proposed. In this paper, we present a trend of deep learning technologies, which have been used on image recognition methods such as image classification, object detection, and visual explanation.

1. はじめに

画像中に存在する物体のカテゴリや位置等を認識する画像認識技術は，Advanced Driver Assistance System (ADAS) や防犯システム，SNS やスマートフォンのアプリケーション等で幅広く応用されている．ADAS では自動ブレーキシステム等の歩行者や自動車を検出する処理で用いられ，SNS やアプリケーション等のレジャーな分野では顔の形状を画像等から推定して自動で編集するアプリケーション等へ応用されている．

このような，画像認識技術が身近に取り入れられるようになった要因の一つとして，深層学習の発展により認識性能が向上したことが深く関係している．画像認識分野における深層学習では，Deep Convolutional Neural Network (CNN) [Alex 12] をベースにした手法が一般的に用いられている．CNN は，複数のカーネルで構築される畳み込み層を主体に構築されたニューラルネットワークであり，学習によりカーネルを更新することで画像認識に有効な特徴量を獲得する．CNN が画像分類で高い性能を発揮した後，物体検出やセマンティックセグメンテーション等の様々な画像認識タスクへ応用されるようになった．これは，CNN のモデルの発展や深いモデルを安定して学習できる手法が数多く提案されたことが大きな貢献となっている．

一方で，CNN による性能向上のみでなく，推論時における CNN の判断根拠をユーザへ伝える技術も提案されている．CNN を導入した製品が誤認識によりユーザへ何かしらの危害を与えたとき，ユーザへなぜこのような行動を起こしたのかを説明する必要がある．そのため，これらの判断根拠を解析する手法は CNN を用いた製品を商品化する際に重要な技術となる．CNN の判断根拠を解析する研究は活発に取り組まれており，様々なアプローチが提案されている．特に，画像認識分野では CNN が推論時に注視した領域をマップで表現した Attention map を用いることで，判断根拠を解析する視覚的説明が用いられる．視覚的説明は，CNN が画像認識する際に注視した領域を可視化することができるため，直感的に CNN の判断根拠を知ることができる．

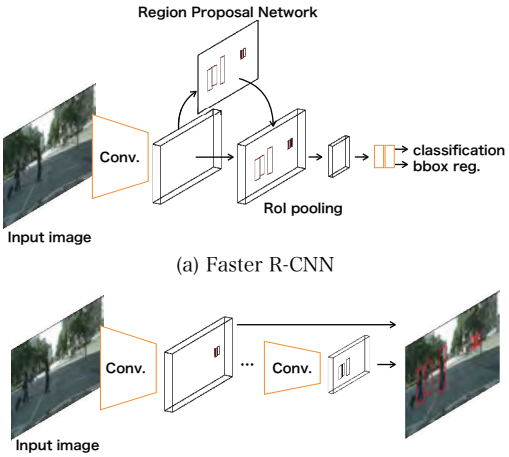
本稿では，画像認識分野で用いられる CNN の最近の動向について述べる．まず，画像分類において用いられる CNN のモデルについて述べた後，物体検出のモデルについて述べる．そして，CNN の学習時に用いられる正規化や学習方法について述べる．最後に，CNN の判断根拠を解析する視覚的説明の研究事例について述べる．

連絡先: 山下隆義，中部大学，愛知県春日井市松本町 1200，0568-51-1111，0568-51-1111，takayoshi@isc.chubu.ac.jp

2. 画像分類における CNN のモデル

物体検出法やセグメンテーション，属性認識等で用いられる CNN ベースのモデルは，画像分類で提案されたモデルをベースにネットワークを構築している例が多い [Ren 15, Liu 16, Badrinarayanan 15, He 17]．画像分類における CNN の初代の代表的なモデルとして，AlexNet [Alex 12] がある．AlexNet は，大規模な画像認識コンペティションである ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [Russakovsky 15] でトップとなったモデルである．AlexNet の構造は，5 層の畳み込み層と 2 層の全結合層から構築された CNN であり，Local Response Normalization (LRN) や Dropout，Rectified Linear Units (ReLU) 等の大量のデータで深いネットワークを学習するためのテクニックが導入されている．AlexNet の提案後，より深いネットワークを構築することで，高精度な画像分類を実現するアプローチが取られた．深いネットワークは，単純な特徴と複雑な特徴を同時に学習できるため，より高い精度で認識が可能である．初期におけるこのアプローチの代表的なモデルとして，VGGNet [Karen 15] や GoogLeNet [Szegedy 15] が挙げられる．VGGNet は最大で 19 層の CNN を構築し，GoogLeNet は Inception module を導入することで 22 層の CNN を構築している．その後，Batch Nornalization [Sergey 15] や He の正規化 [He 15] 等，深いネットワークの学習を安定させるためのアプローチが提案された．

2016 年には，He らが Residual Learning を導入した ResNet を提案し，100 層以上の CNN を構築した [He 16]．ResNet は複数の層を繋げる際にバイパス構造を取り入れており，Residual Learning を取り入れた Residual unit を構築している．Residual Learning の導入により，大量の層で構築されたネットワークを安定して学習できるようになり，CNN ベースの画像分類の精度がさらに向上した．ResNet が提案された後は，ResNet をベースにした様々なモデルが提案されている [Zagoruyko 16, Huang 17, Xie 17].



(a) Faster R-CNN

(b) Single Shot Multi-box Detector

図 1: 物体検出の代表的なモデル

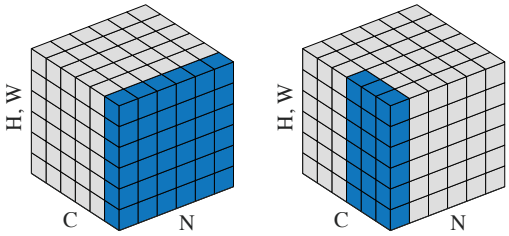
これらの画像分類モデルの進化は，物体検出や属性認識，セグメンテーションにも大きな影響を与えている．物体検出やセグメンテーションで用いられるネットワークは画像分類のネットワークモデルをベースに構築するため，画像分類の性能向上に伴い他の認識タスクの性能も向上できる．そのため，画像分類のモデルは他の画像認識タスクの観点からも，重要な立ち位置に属している．

3. 物体検出

CNN ベースの画像分類法の発展に伴い，CNN をベースとした物体検出法も大きく発展している．物体検出は，図 **??**(b) のように画像中の物体の位置とそのカテゴリを推定する技術である．CNN ベースの物体検出法は，R-CNN [Girshick 14] と Fast R-CNN [Girshick 15] をベースに進化を遂げている．R-CNN ベースの手法では，物体の候補領域を検出し，検出した候補領域を CNN でカテゴリ分類とバウンディングボックスの修正を行う，2 段階の検出構造を採用している．しかし，R-CNN と Fast R-CNN は計算コストが高く，リアルタイムな物体検出が困難である．リアルタイムで物体検出が可能な手法として，Faster R-CNN [Ren 15] がある．Faster R-CNN は，Region Proposal Network (RPN) を図 1(a) のように導入しており，物体候補領域の検出と，カテゴリ分類とバウンディングボックスの修正を 1 つのネットワークで End-to-End に処理できる．Faster R-CNN は，インスタンスセグメンテーションへも応用されている．Faster R-CNN をインスタンスセグメンテーションへ応用した Mask R-CNN [He 17] は，検出した物体領域のマスクを出力するブランチを追加することで，高精度なインスタンスセグメンテーションを実現している．また，R-CNN のような 2 段階の検出構造を用いずに，図 1(b) のようにネットワークの応答値から直接検出スコアを出力できる Single Shot Multi-box Detector (SSD) [Liu 16] も提案されている．SSD は，物体候補領域の検出を必要としないため，Faster R-CNN より高速に物体を検出できる．

4. CNN の汎化性能を向上させる手法

大量のデータを用いて深いネットワークを安定して学習するために，様々なテクニックが導入されている．ネットワークの学習時に正規化，最適化方法を導入することで，学習の収束を早めたり，学習時における勾配の発散等を防ぐことができる．



(a) Batch Normalization (b) Group Normalization

図 2: Batch Normalization の発展系の手法 (文献 [Yuxin 18] から引用)

4.1 正規化

CNN はカーネルと入力画像または特徴マップの局所領域から内積値を求めるため，入力値にノイズが発生した場合等に応答値のばらつきが発生し，認識性能を低下させる原因となる．そのため，一般的には畳み込み層や全結合層の応答値を正規化し，精度低下を抑制する．AlexNet では LRN が用いられていたが，他の特定のネットワークに対しては精度が向上しなかったり，Batch Normalization [Sergey 15] が提案されたことで，現在は一般的に用いられていない．

Batch Normalization は，特定のチャンネルをミニバッチ単位で正規化し，平均を 0 と分散を 1 にする．図 2(a) のようにミニバッチ単位で特定のチャンネルを正規化することで，内部共変量シフトが大幅に変動するのを防いでいる．しかし，Batch Normalization は内部共変量シフトを獲得するために，ミニバッチのサイズを 16 以上にする必要がある．ミニバッチのサイズは大きいほど計算コストが増加するため，物体検出等の使用メモリ量が膨大なモデルで十分な性能を発揮できない．この問題を解決する手法として，Group Normalization がある．Group Normalization は，図 2(b) のように数枚のチャンネルのみ用いて正規化する．これにより，Group Normalization は少量のミニバッチサイズでも従来の Batch Normalization と同等の精度を得ることができる．

4.2 学習方法

確率的勾配降下法 (stochastic gradient descent; SGD) によりネットワークを学習する場合，学習率 η を学習の過程で変化させる Learning rate schedule が一般的に導入されている．ネットワークの学習における学習率は，パラメータの更新量を制御する係数である．学習率を大きく設定した場合は学習の収束が早くなるが，勾配が発散しやすい．一方で，学習率を小さく設定した場合は最適解を獲得しやすいが，学習の収束が遅くなる．この問題を解決するために，学習の過程で学習率を変更することで，最適解を獲得しやすくしている．Learning rate schedule は，一般的には学習率を減衰させる Learning rate drop が一般的に用いられる．SGD で Learning rate drop を用いる場合，指定した更新回数に達した際に学習率を減衰する．学習率を特定の更新回数で下げることで，算出される学習誤差をより下げることができ，認識率を向上できる．

一方で，学習率を下げるタイミングを手動で決定するのではなく，学習の過程で自動に決定する方法も提案されている [Zeiler 12, Tieleman 12]. Adaptive Gradient (AdaGrad) [Duchi 11] は，ネットワークの各パラメータに対し

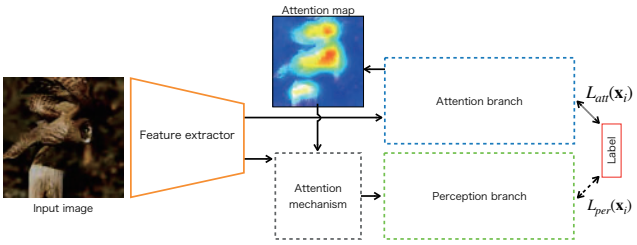


図 3: Attention Branch Network の構造

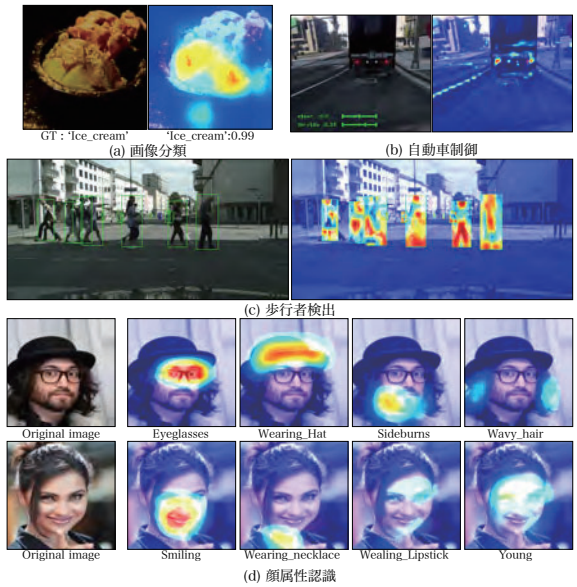


図 4: ABN が獲得した Attention map の例

て学習率を設計し，自動で調整しながら学習できる．AdaGrad をベースとした学習法は，数多く提案される [Zeiler 12, Kingma 14]．

5. 視覚的説明

深層学習における画像認識分野では，推論時に注視した領域をマップで表現した Attention map から判断根拠を解析する視覚的説明の研究が取り組まれている [Zeiler 14, Zhou 16, Ramprasaath 17]．Attention map の獲得には，勾配を用いた Bottom-up の手法とネットワークの応用値を用いる Top-down の手法の 2 種類がある．Bottom-up の手法の例として，Guided Backpropagation と Gradient-weighted Class Activation Mapping (Grad-CAM) [Daniel 17] がある．Guided Backpropagation と Grad-CAM は，逆伝播の特定のクラスにおける正値の勾配のみ用いることで，Attention map を獲得する．Guided Backpropagation と Grad-CAM は特定のクラスにおける Attention map を様々なプレトレーニングモデルから獲得できるため，CNN の解析手法として広く一般的に用いられている．Grad-CAM は，特定のクラスの出力層のユニットから勾配を発生させ，特徴マップを獲得する．そして，順伝播時における最後の畳み込み層の特徴マップに対して，GAP を施す．GAP により獲得した特徴ベクトルは重みとして使用し，勾配ベースの特徴マップに対して重み付き和を求める．この重み付き和で求めた特徴マップに ReLU を施すことで，特

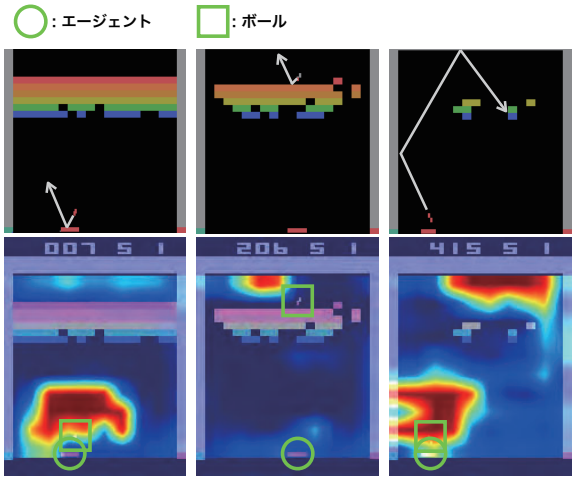


図 5: Breakout における Attention map の例

定のクラスに対する注視領域を獲得することができる．

視覚的説明における Top-down の手法は，ネットワークが出力した応答値を用いることで Attention map を獲得できる．Top-down の手法は Attention map を獲得するためにネットワークを再構築して再学習する必要があるが，順伝播の過程で各クラスにおける注視領域を獲得できる．Top-down の手法の代表的な手法である CAM [Zhou 16] は，畳み込み層の応答値と全結合層の結合重みを用いることで，各クラスにおける Attention map を獲得できる．CAM は全結合層を畳み込み層に入れ替える等の処理が必要なため，画像分類においては性能低下を引き起こしやすい．

この問題を解決した Top-down な方法として，Attention Branch Network (ABN) [Fukui 18] がある．ABN は，図 3 のように Feature extractor と Attention branch，Perception branch の 3 つのモジュールから構成されている．Feature extractor は，入力画像から特徴マップを抽出するモジュールである．Attention branch は，畳み込み層をベースに構築されたブランチであり，Attention map を出力する．Feature extractor は，入力画像から特徴マップを獲得するモジュールである．抽出した特徴マップは Attention branch へ入力され，Attention map を出力する．ABN は，画像分類をはじめとした歩行者検出，マルチタスク学習等の様々な画像認識タスクから，図 4 のように Attention map を獲得できる．また，ABN はエージェントの制御で用いられる深層強化学習へも応用できる．ABN を深層強化学習へ応用した際に獲得した Attention map を，図 5 に示す．図 5 の例は，Atari ゲームの一つである Breakout を深層強化学習で操作している例である．図 5 の結果から，ボールを跳ね返す，ブロックの奥でボールが跳ね返る等のゲームスコアを獲得する直前のシーンにおいて，Attention map が強く反応していることがわかる．

6. おわりに

本稿では，画像認識における深層学習の動向についてまとめ，紹介した．画像認識における深層学習の発展により，画像分類をはじめとした物体検出等の性能を大幅に向上させた．また，性能向上だけでなく，深層学習の推論結果に対する判断根拠を解析する研究も活発に取り組まれている．

参考文献

[Alex 12] Alex, K., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in *Neural Information Processing Systems*, pp. 1097–1105 (2012)

[Badrinarayanan 15] Badrinarayanan, V., Handa, A., and Cipolla, R.: SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling, *arXiv preprint arXiv:1505.07293* (2015)

[Daniel 17] Daniel, S., Nikhil, T., Been, K., Fernanda, B. V., and Martin, W.: SmoothGrad: removing noise by adding noise (2017)

[Duchi 11] Duchi, J., Hazan, E., and Singer, Y.: Adaptive Subgradient Methods for Online Learning and Stochastic Optimization, *Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159 (2011)

[Fukui 18] Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H.: Attention Branch Network: Learning of Attention Mechanism for Visual Explanation, *arXiv preprint arXiv:1812.10025* (2018)

[Girshick 14] Girshick, R., Donahue, J., Darrell, T., and Malik, J.: Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, in *Computer Vision and Pattern Recognition*, pp. 580–587 (2014)

[Girshick 15] Girshick, R.: Fast R-CNN, in *International Conference on Computer Vision* (2015)

[He 15] He, K., Zhang, X., Ren, S., and Sun, J.: Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification, in *International Conference on Computer Vision*, pp. 1026–1034 (2015)

[He 16] He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, *Computer Vision and Pattern Recognition*, pp. 770–778 (2016)

[He 17] He, K., Gkioxari, G., Dollár, P., and Girshick, R.: Mask R-CNN, in *International Conference on Computer Vision* (2017)

[Huang 17] Huang, G., Liu, Z., Maaten, van der L., and Weinberger, K. Q.: Densely connected convolutional networks, in *Conference on Computer Vision and Pattern Recognition* (2017)

[Karen 15] Karen, S. and Andrew, Z.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *International Conference on Learning Representations* (2015)

[Kingma 14] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *1412.6980* (2014)

[Liu 16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-y., and Berg, A. C.: SSD : Single Shot MultiBox Detector, in *European Conference on Computer Vision*, pp. 1–15 (2016)

[Ramprasaath 17] Ramprasaath, S., R., Michael, C., Abhishek, D., Ramakrishna, V., Devi, P., and Dhruv, B.: Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, in *International Conference on Computer Vision*, pp. 618–626 (2017)

[Ren 15] Ren, S., He, K., Girshick, R., and Sun, J.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in *Neural Information Processing Systems*, pp. 91–99 (2015)

[Russakovsky 15] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge, *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211–252 (2015)

[Sergey 15] Sergey, I. and Christian, S.: Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, in *International Conference on Machine Learning*, pp. 448–456 (2015)

[Szegedy 15] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A.: Going deeper with convolutions, in *Computer Vision and Pattern Recognition*, pp. 1–9 (2015)

[Tieleman 12] Tieleman, T. and Hinton, G.: Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude, COURSERA: Neural Networks for Machine Learning (2012)

[Xie 17] Xie, S., Girshick, R. B., Dollár, P., Tu, Z., and He, K.: Aggregated Residual Transformations for Deep Neural Networks, *Computer Vision and Pattern Recognition*, pp. 5987–5995 (2017)

[Yuxin 18] Yuxin, W. and Kaiming, H.: Group Normalization, in *European Conference on Computer Vision*, pp. 3–19 (2018)

[Zagoruyko 16] Zagoruyko, S. and Komodakis, N.: Wide Residual Networks, in *British Machine Vision Conference* (2016)

[Zeiler 12] Zeiler, M. D.: ADADELTA: An Adaptive Learning Rate Method, *1212.5701* (2012)

[Zeiler 14] Zeiler, M. D. and Fergus, R.: Visualizing and Understanding Convolutional Networks, in *European Conference on Computer Vision*, pp. 818–833 (2014)

[Zhou 16] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A.: Learning Deep Features for Discriminative Localization, *Computer Vision and Pattern Recognition* (2016)