概説

《学修項目》

- 深層学習の基礎と事例
- 画像認識技術と適用事例

《キーワード》

> 機械学習、深層学習、ニューラルネットワーク、Deep Neural Network、パラメータ学習、U-Net、DenseNet、Attention/Transformer、CNN、RNN、GAN、ImageNet、MNIST、CIFAR-10

《参考文献，参考書籍》

- [1] 東京大学MIセンター公開教材 「AI基礎：3-4 深層学習の基礎と展望」 《利用条件CC BY-NC-SA》
- [2] 東京大学MIセンター公開教材 「AI基礎：3-5 認識」 《利用条件CC BY-NC-SA》

1. はじめに

近年、人工知能（AI）の中で、最も注目を集めているのが機械学習の一種である深層学習（Deep Learning）の技術であろう。ここで、機械学習とは、一般に、大量のデータを学習機に与えて学習させることにより、目的の出力を得る技術である。

例えば、犬と猫の写真をどちらかに識別する問題の場合、学習機に犬と猫の写真画像を大量に与え、それと同時にその写真が犬か猫かの正解データを与えることにより（このような学習を教師つき学習と呼ぶ）、学習機に犬か猫かを識別する学習をさせる。

このようにして構築したモデルに、犬か猫の写真を与えると、そのどちらかに識別してくれるようになる。深層学習は、このような機械学習の1つであり、学習機は多層構造のニューラルネットワーク（人間の神経回路網を模倣したネットワーク）で構成されている。また、深層学習は、現在、医療、自動運転、自然言語処理、音声認識、データ解析など、様々な分野に適用されており、その有効性が示されている。

2. 教材の構成

そこで、本教材では、ニューラルネットワークを画像認識の分野に適用する下記事例について取り上げ、その手法や性能について言及する。

- ① 文字認識における曲線の屈曲度の判定
- ② 楽譜の音符検出における記号判定
- ③ MLPを用いた数字画像の識別
- ④ CNNを用いた数字画像・写真画像の識別

なお、従来のニューラルネットワークの学習では、画像から識別に有効であると考えられる特徴をあらかじめ抽出し、その特徴量をネットワークの入力に与えて学習する手法が用いられてきた。例えば、犬と猫の識別の場合、目の色が識別に大きく関わっていると開発者が判断した場合、目の色の特徴量を抽出して、それを学習に使用する、といった方式である。本教材では、①と②の事例がこの手法に該当する（ここでは、3層型の単純なニューラルネットワークを用いる）。

また、③の事例では、機械学習プラットフォームであるTensorFlowを用いた数字画像識別の例を示す（ここでは、簡単のため、特別な特徴量抽出をせず、画像の輝度値そのものを入力として、数字画像の識別を行う例を挙げる）。

一方、④の事例では、現在よく使われている深層学習の1つであるCNN（Convolutional Neural Network：畳み込みニューラルネットワーク）を用いた事例を紹介する。ここでは、特徴抽出機能を包含するネットワークが構築できることを示す（例えば、犬と猫の識別の場合、開発者が目の色のような特徴をあらかじめ抽出して与えることなしに、その写真画像そのものをネットワークに与えて学習することができる）。ここで、現在の

最新技術だけを紹介せず、旧来の手法①、②、③も紹介するのは、そこで使われているニューラルネットワークの学習技術が、最新技術の手法④にも使われているからである。

3. 深層学習の基礎と事例 [1]

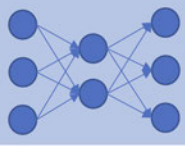
3.1 深層学習の基礎、Neural Network、Deep Neural Network

深層学習

深層学習とはニューラルネットワークをモデルとする機械学習のことで以下のような関係性があります.

機械学習（Machine Learning, ML）
有限の観測データから背後に潜む規則を獲得.
教師あり学習・教師なし学習・強化学習という枠組みがある.

深層学習（Deep Learning, DL）
ニューラルネットワークを用いた機械学習.
AIにパラダイムシフトを起こした.



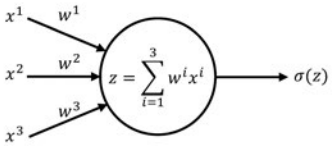
ニューラルネットワーク：
人の脳を模した学習機械.
画像・音声などの認識，自然言語処理，
深層強化学習などで用いられる.

ニューラルネットワークの原理

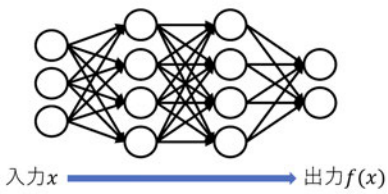
ニューラルネットワークは脳を模した機械学習モデルです.

ニューロン（ノード）が重みつきの枝を通じて接続し合いネットワークを構成します.

ニューロンは接続関係にあるニューロンから枝の重みに従い信号を受け取ったのち，活性化関数と呼ばれる適当な変換をかけ信号を出力します.

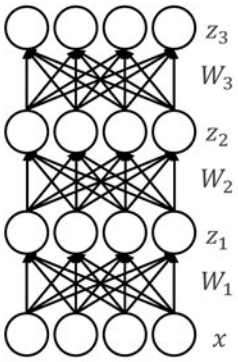


この信号の変換を多数のニューロンを用いて層状に積み重ね，入力データの複雑な変換を実現したモデルがディープニューラルネットワーク（DNN）です.

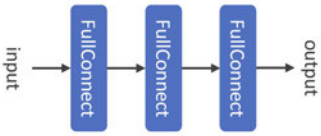


枝の重みをパラメータとして経験損失最小化を行いニューラルネットワークの学習を行います.

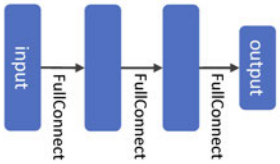
ディープニューラルネットワークの構造



層を重ねディープニューラルネットワークを構成.
左図：全結合層から成る３層ニューラルネットワーク.
簡易的に下図のように表現する場合もあります.



変換部分に注目した表現.



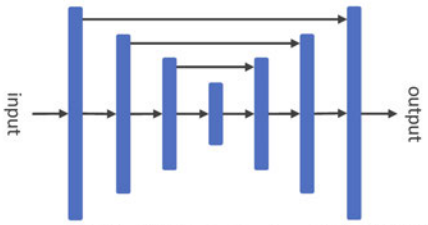
ニューロンの集まりに
注目した表現.

層は特定の用途・目的に応じデザインされ，種々のものがあります.
全結合層，ソフトマックス層，活性化層，畳み込み層，プーリング層，
LSTM層，バッチ正規化層，ドロップアウト層.

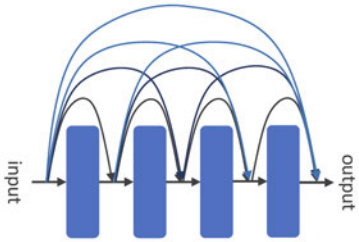
3.2 深層学習器のモデリング、応用タスク、認識精度の向上

柔軟なモデリング

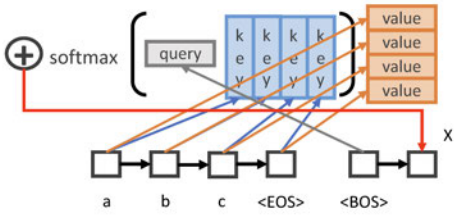
層をブロックとして組み合わせ自由自在なモデリングが可能です.
タスクに応じて直感的な依存関係を柔軟に記述することできます.



U-Net: 画像セグメンテーションで有効なモデル



DenseNet: 高精度な認識モデル



Attention: 入力(query)に対応する値(value)を返す辞書(keyとvalueのペア).

TransformerはAttention機構を組み込んだ機械翻訳モデル.

実世界で進む深層学習の応用と革新

タスク依存の直感的なモデリングはニューラルネットワークの精度を向上させる上での大きな利点です.

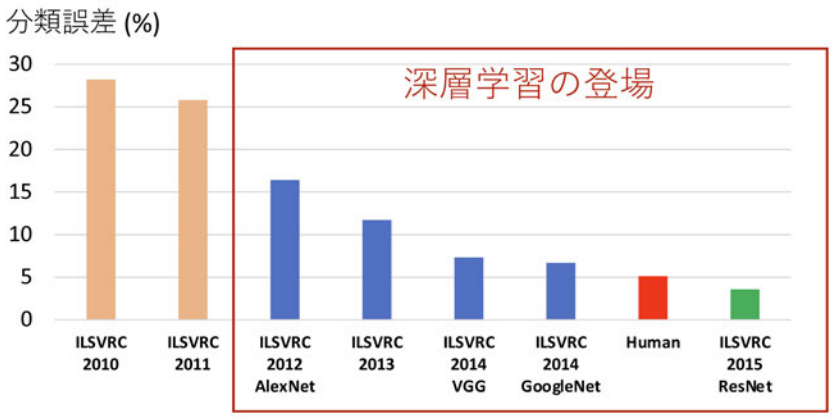
実際，タスクに応じてさまざまなニューラルネットワークモデルが提案され学習技法の発展とあわせて劇的な精度向上がもたらされました.

例えば以下のタスクで非常に優れたモデルや学習法が開発されました.

- **画像認識**
  畳み込みニューラルネットワーク（CNN）
- **自然言語処理**
  再帰型ニューラルネットワーク（RNN），Transformer
- **画像生成**
  敵対的生成ネットワーク（GAN）
- **音声生成**
  WaveNet

画像認識精度の向上

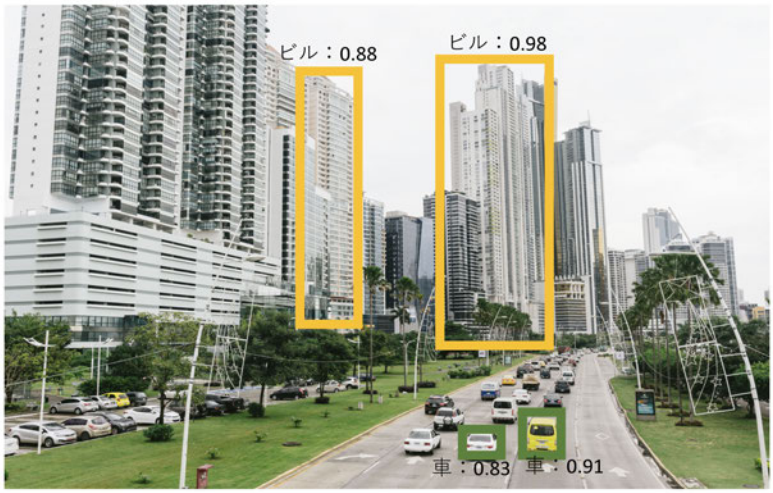
ImageNet（画像認識の大規模データセット）のコンペティションでの画像認識精度は深層学習時代のCNN登場以降，飛躍的に向上.



3.3 物体検出、マスキング、画像生成の例

事例：物体検出

画像認識より一段むずかしい問題設定.
どこに何が写っているかを検出します.



「どこに」＋「なにが」

事例：マスキング

物体検出より詳しくピクセル単位で認識する.
ピクセルが何を構成しているか予測します.

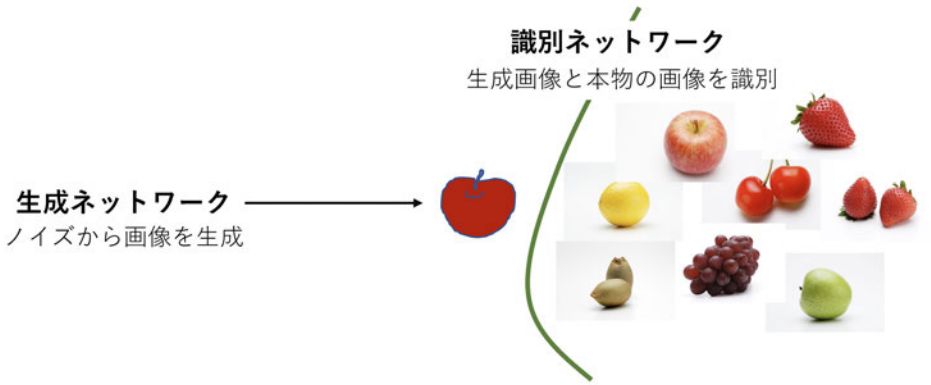
事例：画像生成

敵対的生成ネットワーク(GAN)により高精細な画像の生成も可能です.

GANでは識別ネットワークと生成ネットワークという二つのネットワークが以下の指針で敵対的に交互に学習されます.

1. **識別ネットワーク**：本物の画像か生成ネットワークが生成した画像かを識別.
2. **生成ネットワーク**：識別ネットワークを騙すような画像を生成.

この敵対的学習法により次第に高精細画像が生成されるようになります.

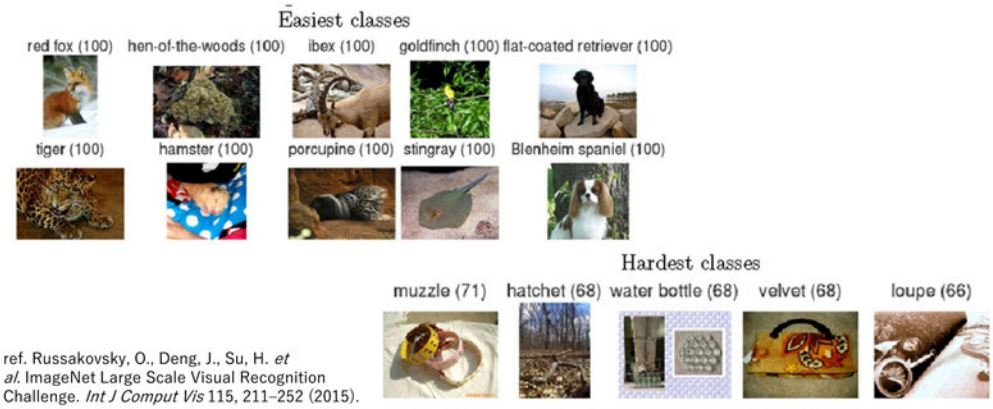


4. 大規模データセットを用いた深層学習器、適用事例 [2]

4.1 大規模データセットを用いた学習、認識精度向上

深層学習による画像認識の幕開け

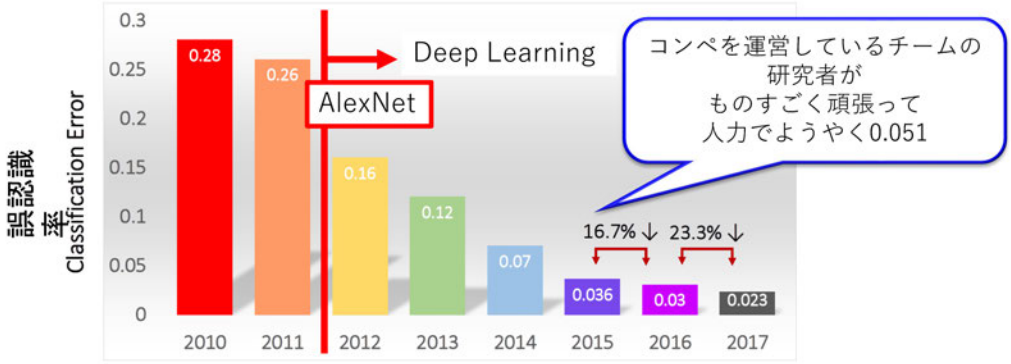
- クラウドソーシングによる大規模データセット：ImageNet
- ImageNet Large-Scale Visual Recognition Challenge (ILSVRC)
  - 1000クラス、学習120万枚、検証5万枚、テスト10万枚
  - マルチラベル：1枚の画像に複数ラベル＋信頼度
- 2011年のILSVRCで優勝したモデルのエラー率は26%
  → 2012年にDeep Learningを使ったモデルが登場→15%に激減！



ref. Russakovsky, O., Deng, J., Su, H. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* 115, 211–252 (2015).

物体認識精度は人間を超えた!?

- ILSVRC2012で、トロント大学Geoffrey Hinton教授率いるグループが初めて多層ニューラルネットワークによる画像認識モデルを提案
- 主著者の名前をもじってAlexNetと呼ばれる
- 翌年から上位チームはすべてDeep Learningを採用
- 2015年についに人間の精度を超えた



Excerpted from ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2017 Overview
Andrej Karpathy, "What I learned from competing against a ConvNet on ImageNet", Sep 2, 2014, http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/

深層学習はそれまでの画像処理と何が違ったのか？

- AlexNet (2012)登場より前の画像認識では、CNNにおける第1~2層の出力に相当する情報を使って認識していた
  → Deep Networkと対比してShallow networkと呼ばれる
- 2000年前後にいくつかの技術革新
  - 数学的解法：勾配消失問題（層を深くすると学習が進まなくなる現象）に対する効率的な解決法の提案（1990年代後半）
  - GPGPUの発展：コンピュータグラフィックスの描画に用いられていたGPUをベクトル計算機とみなして気象や地震シミュレーション等、数値計算に利用（2006年NVIDIAがCUDA提供開始）
  - Big Data時代の到来：
    Webで画像やテキストなどが大量に収集できるようになり、モデルの学習に使えるデータが爆発的に増加
- 様々な画像処理タスクに対する学習データセットが公開
  - 学習データを使わない自己教師あり学習（self-supervised learning)の研究も進められている

4.2 画像認識の活用事例

顔領域検出

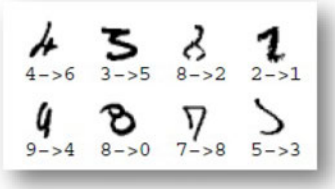
- 写真の中で「顔」らしい領域を検出する
- カメラが自動的に顔にフォーカスを合わせる際にも用いられる

ref. (2020/4/6): Wikimedia commons: File:Kasahara Saitama Kasahara Jinjo Elementary School 1920 1.jpg パブリックドメイン
https://ja.wikipedia.org/wiki/%E3%83%95%E3%82%A1%E3%82%A4%E3%83%AB:Kasahara_Saitama_Kasahara_Jinjo_Elementary_School_1920_1.jpg

- ref. (2020/4/6): Wikimedia commons: File:Kasahara Saitama Kasahara Jinjo Elementary School 1920 1.jpg パブリックドメイン

手書き文字認識

- 画像認識初期のタスク
- よく用いられるデータセット：MNIST (Mixed National Institute of Standards and Technology database)
  - 「0~9」の10種類の数字の認識
    ＝10クラス分類タスク
  - 各画像に数字が1つ記入
  - 解像度は28x28 pixel
  - 訓練データ：60,000枚
  - 評価データ：10,000枚
- 間違いやすいサンプルを含む



ref. (2020/4/6): THE MNIST DATABASE of handwritten digits http://yann.lecun.com/exdb/mnist/
Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998.

- ref. (2020/4/6): THE MNIST DATABASE of handwritten digits
- Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998

- ref. (2020/04/06): The CIFAR-10 dataset

- ref. (2020/4/6) Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields", CVPR2017,

memo