形状認識と機械学習

2018年8月31日

神奈川工科大学

春日 秀雄

研究分野としての形状認識

- 「形状認識」とは対象の形を識別すること
  - 図形や文字といった2次元のデータの識別も形状識別の一種と言えるが、今回は主に扱う形状を3次元形状とする

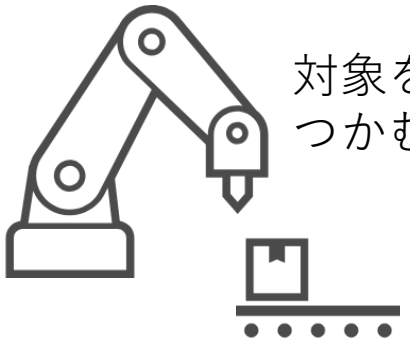
- 対象となるのが実空間の『物』であることが多いため、「物体認識」という言葉も同じ意味で使われることが多い
- 「物体認識」という研究分野には大別して２つの問題領域がある
  - 特定物体認識
  - 一般物体認識

特定物体認識

- 同一の物体（形状）であるか識別する
  - 文字認識をはじめ、２次元の対象を識別する手法は多数考案されているため、ここでは３次元物体の形状を認識する手法に限定して述べる

２次元のデータの識別例

- ３次元の物体認識はロボットビジョンの中核となる技術



対象を発見し、ロボットアームで正確に対象をつかむには３次元物体認識が欠かせない

３次元の特定物体認識

- 特定の対象がどの位置にどのような姿勢であるか認識する
  - 空間的な情報を認識

- ３次元物体の形状認識手法を大別すると以下の２種類
  - アピアランスベースの物体認識
    - 外観（画像）から判定する
  - モデルベースの物体認識
    - ３Ｄモデル（形状データ）を利用して判定する

アピアランスベース物体認識

- 入力画像と３次元物体の多視点画像群との照合を行う
  - ２Dデータと２Dデータの照合
    - 単なる画像のパターンマッチング問題とすることができる

- メリット
  - 一般的なカメラを入力デバイスとして利用できるので、安価にシステムを構築できる
- デメリット
  - 多視点画像を用いるため、辞書データとして多くの画像を用意しなければならない

モデルベース物体認識

- 距離画像などの３次元的なデータと３次元モデルを照合
  - ３Dデータと３Dデータの照合
    - ３次元データのマッチング問題は２次元データである画像のマッチングより複雑

- メリット
  - 多数の画像ではなく１つの形状データ（３次元CADモデルなど）を用意すればよい
- デメリット
  - 入力には距離情報の得られる３次元センサーが必要

３次元データのマッチング問題

- ２次元のパターンマッチングと同じく、３次元データにおける局所特徴量を抽出し、その特徴量の比較を行いマッチングする
- 対象のすべての場所の３次元データから特徴量を求めるのではなく、キーポイントと呼ばれる特定の場所の周りの特徴量を求めて照合する手法が一般的
  - SHOT特徴量、PFH特徴量、FPFH特徴量など

一般物体認識

- 物体（必ずしも同一の形状ではない）のカテゴリ（一般的な名称等）を識別する
  - 入力画像に自動的にラベルを付ける問題等

例）

右の画像すべてに猫というラベルを付けるような問題

※　対象がすべて同じ外見をしているとは限らない

→ 猫

一般物体認識のためのアプローチの変遷

- 古典的な手法としては、幾何学的なルールや３次元モデルに基づく手法がある
- その後、多数のデータを集めてその統計量をもとに評価を行う統計的パターン認識が発展する
- さらに、その統計量（≒特徴量）に基づく判定に機械学習を用いた手法が登場する
- そして、ディープラーニングを用いた手法の登場によって劇的に精度が進化
  - 2012年のImageNet large scale visual recognition challenge (ILSVRC) に登場したAlexNet以降、画像認識においてはディープラーニングを用いた手法が一般化する

- 十数年前には「画像認識の研究において最も困難な課題の１つ」とまで言われていたが目覚ましい進歩を遂げた

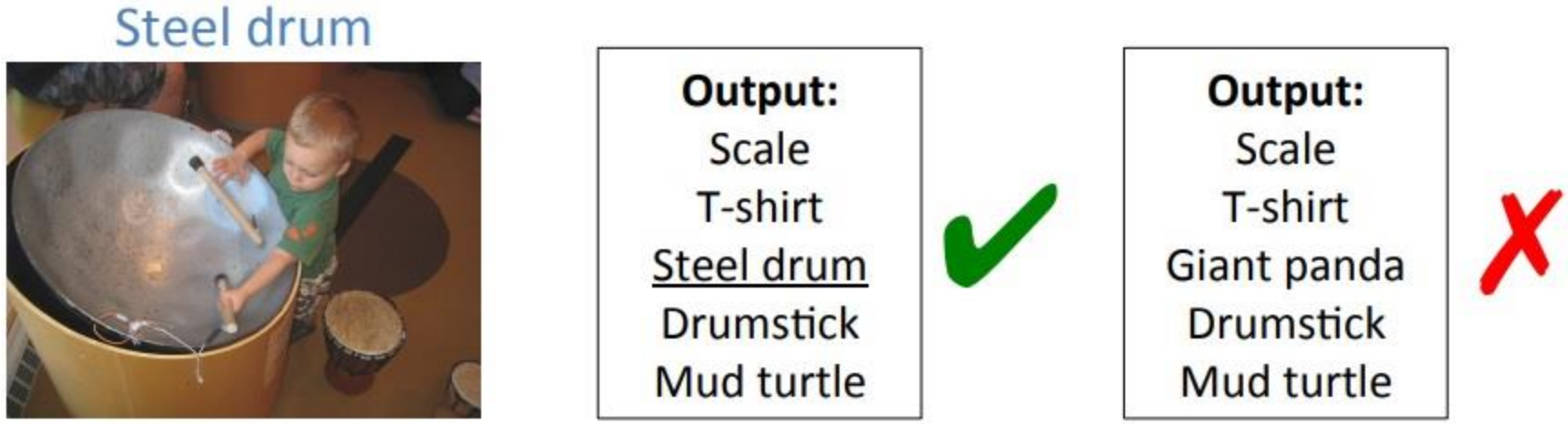
ImageNet Large Scale Visual Recognition Challenge (ILSVRC)

- 2010年から始まった大規模画像認識のコンペティション
- ImageNetは大量のラベル付けされた画像のデータセットを持つ
  - 2018年の時点で21,841クラス、1400万以上の画像がある

- 2017年のILSVRCの内容は以下の３つ
  - Object localization for 1000 categories.
    - 画像内の物体が1000カテゴリの何でどこにあるか判定する
  - Object detection for 200 fully labeled categories.
    - 画像内で指定された200カテゴリの物体を検出する
  - Object detection from video for 30 fully labeled categories.
    - 動画内で指定された30カテゴリの物体を検出する

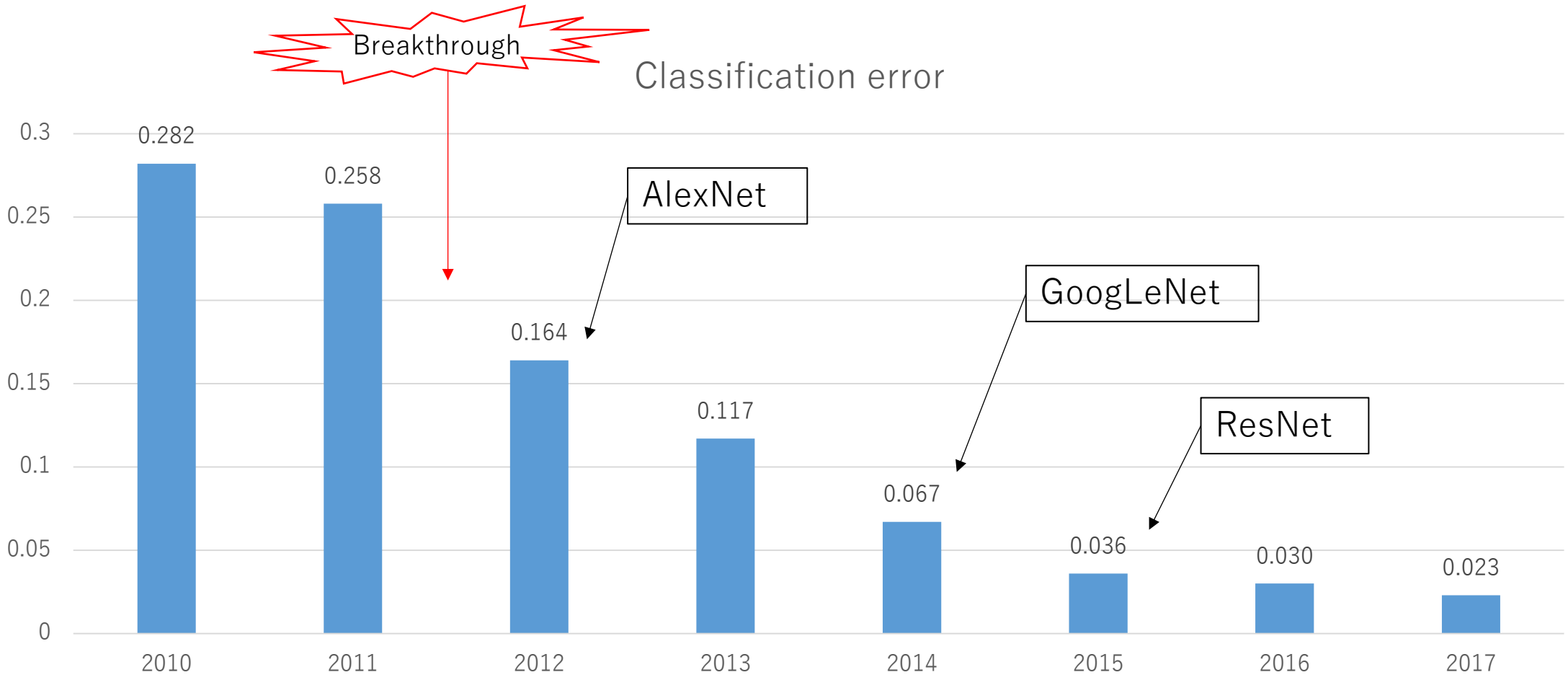
ILSVRCの画像分類

- 競技内容は年によって変わるが、“画像内の物体のカテゴリ分類の精度を競う”という競技は2010年から続いている
  - カテゴリの数は1000
  - 画像内の物体に５つまでラベルを付けて、その中に正解が含まれていればよい



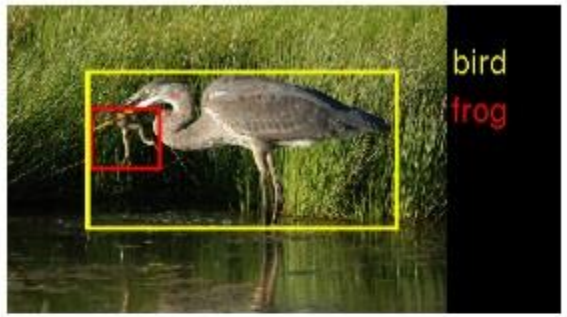
http://www.image-net.org/challenges/LSVRC/2013/slides/ILSVRC2013_12_7_13_clsloc.pdf

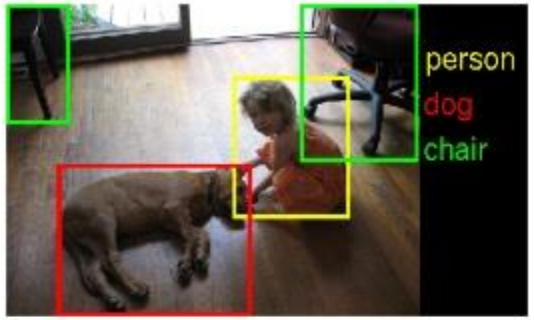
ILSVRC優勝チームの認識精度の推移

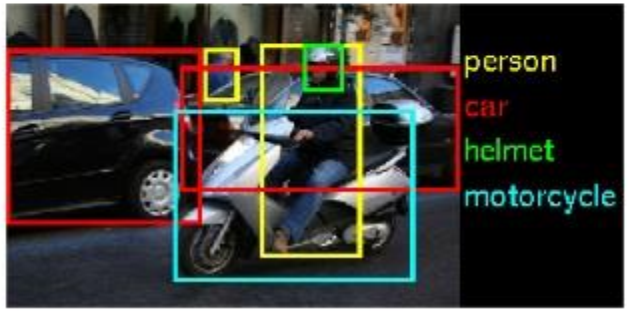


ILSVRCの物体検出

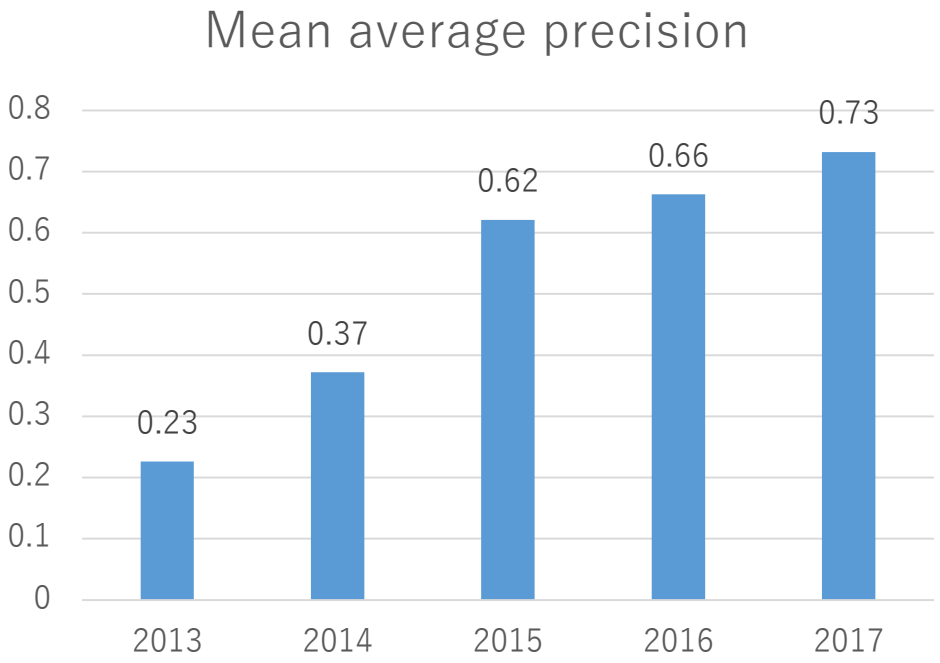
- 物体検出は2013年から行われている
  - カテゴリ数は200
  - 画像内に何が存在するか判別する
  - どれだけ多くのカテゴリを検出できるか競う







http://image-net.org/challenges/LSVRC/2013/index



ほぼ限界まで達したカテゴリ分類の精度

- 2015年にはILSVRCの画像分類の誤認識率が３％台となり、人間の識別能力を超えたと言われるようになった
- 2017年には誤認識率が2.3％となっているが、もはやこれ以上の精度向上を競う意味はほぼなくなっている
  - ImageNetのデータセットは人間が画像を分類した場合でも４～５％は間違うと言われている

- 物体検出（画像内で指定されたカテゴリの物体を検出する）でも2017年には73％の精度となっている

ディープラーニング

- ディープラーニングとは層を深くしたニューラルネットワーク

- では、なぜ層を深くするのか？
  - 層を深くすることによって認識性能が向上
    - 実際に AlexNet（8層）～ GoogLeNet（22層）～ ResNet（152層）のように層が深くなるにつれて性能が向上している
  - ネットワークのパラメータを少なくすることができる
    - 層を深くしなかった場合に比べて、少ないパラメータで同等の空間領域をカバーできる
  - 学習効率が上がる
    - ニューラルネットワークでは層ごとに特徴を学習し、それを重ねることで複雑なパターンを識別するので、層を深くすることで各層の学習すべき特徴をよりシンプルにすることができる

畳み込みニューラルネットワーク

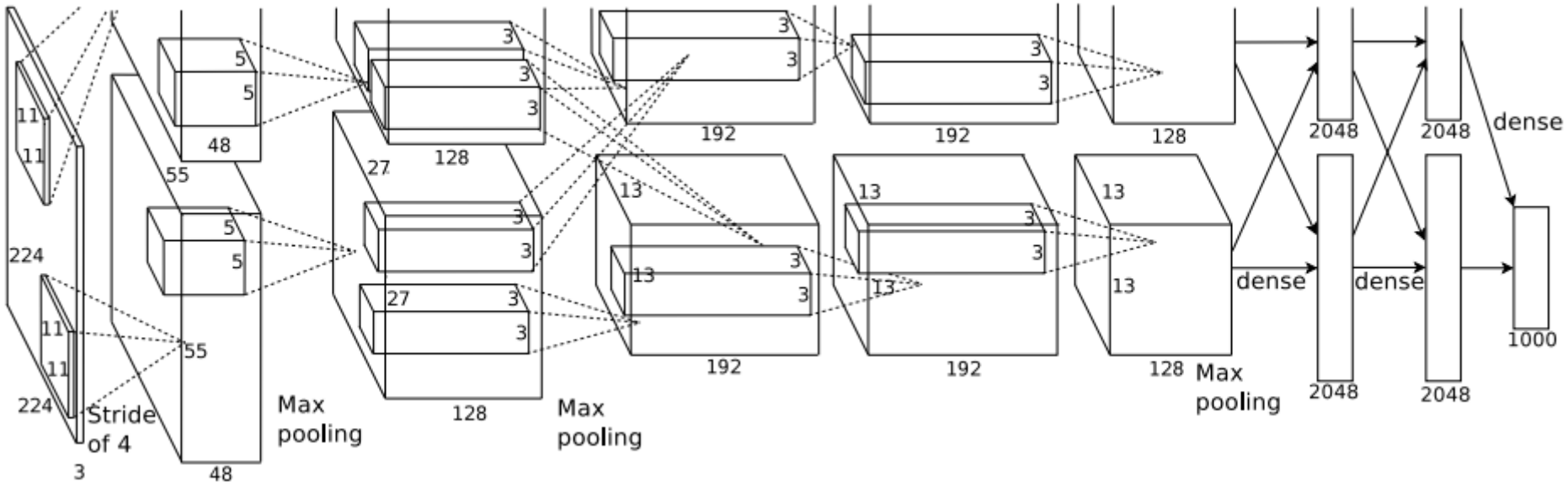
- Convolutional Neural Network略してCNNと呼ばれる
- 画像認識の分野で優れた性能を発揮するニューラルネットワーク
  - 人間の脳の視覚野の仕組みに似た機能を持つと言われている

- その特徴は畳み込み層とプーリング層を持つこと

- 畳み込み層では全結合ではなく一部の領域のみを結合する
  - 局所特徴を抽出するような役割を持つようになる
- プーリング層では領域の最大値または平均値を求めて空間を縮小する
  - 位置変化に対してロバストになる

AlexNet

- 2012年のILSVRCで優勝したモデル
- 畳み込み層５つ、全結合層３つのネットワーク
- 活性化関数にReLUを使用
- 過学習を抑制するDropoutを適用



ImageNet Classification with Deep Convolutional Neural Networks(Alex Krizhevsky et al.)より引用

VGG

- 2014年のILSVRCで準優勝したモデル
- 構造はシンプルだが最大19層まで層を増やしたネットワーク
  - 層の深さに応じて「VGG16」、「VGG19」と呼ばれることもある
- 3×3の小さなフィルターで畳み込みを連続して行っている

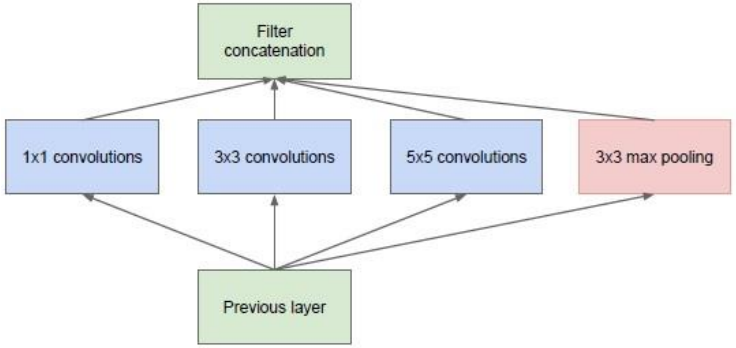
19層のモデル

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

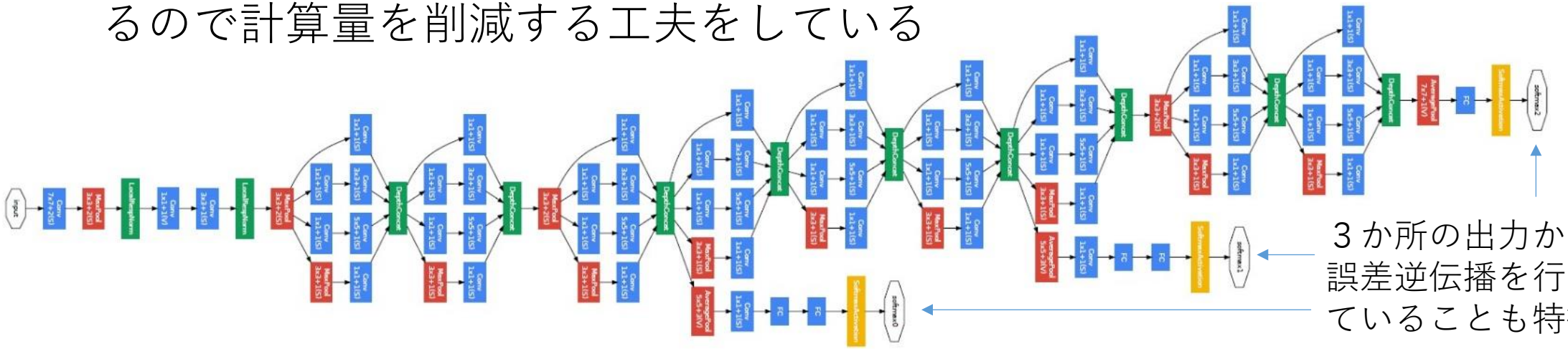
https://arxiv.org/pdf/1409.1556.pdfより引用

GoogLeNet

- 2014年のILSVRCで優勝したモデル
- インセプションモジュールという独自の構造を重ねてネットワークを構築している
  - 異なるフィルターサイズの畳み込み結果を結合して次の層に渡している
  - 多様な特徴を取得できるが計算コストが高くなるので計算量を削減する工夫をしている



インセプションモジュール

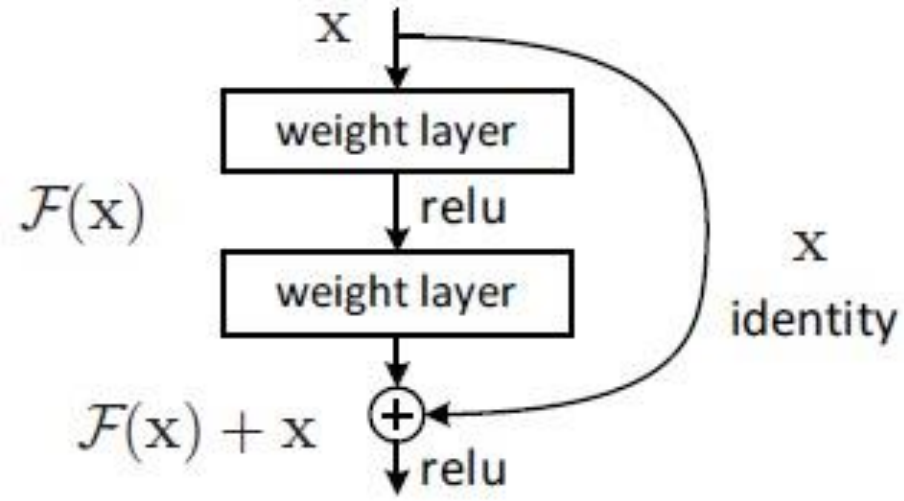


3か所の出力から誤差逆伝播を行っていることも特徴

https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/43022.pdfより引用

ResNet

- 2015年のILSVRCで優勝したモデル
- 152層という非常に深い構造のネットワーク
- 右図のような層をまたいで接続する経路を持つユニットを重ねて構築されている
  - 勾配の消失によって学習が進まなくなるという問題を解決し、非常に深いネットワークを実現した



https://arxiv.org/pdf/1512.03385.pdfより引用

- 他にも以下の特徴がある
  - プーリング層の代わりに畳み込みのストライドを2にしてサイズを小さくしている
  - 全結合層がなく畳み込み層でも小さなフィルターを使っておりパラメータが少ない（VGG19の18%）
  - 過学習を抑制するDropoutを使用していない

ResNet

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

https://arxiv.org/pdf/1512.03385.pdfより引用

３次元の一般物体認識

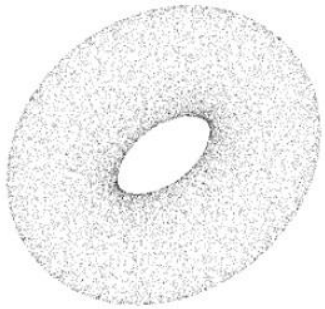
- ３次元データから物体のカテゴリを識別する

- 画像による一般物体認識の場合と同じく、必ずしも同じ形状ではない物体をカテゴリ分類する

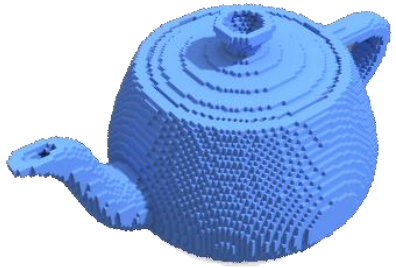
- より高度な作業を行うロボットでは特定物体認識だけでなく、様々な物体を識別する機能が求められている
  - 例えば、荷物の配送を行うシーンでは次々と新しい物が登場するので、特定物体認識の技術だけでは不十分

３次元データをどのように扱うか？

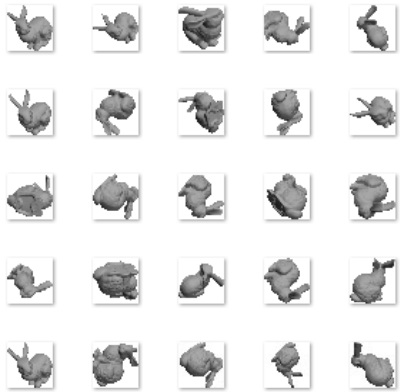
- ３次元的な特徴量を持った点群として扱う



- ボクセルデータとして扱う



- 多視点画像として扱う



これらデータを使って機械学習を行う

3D Shape Retrieval Contest (SHREC)

- 形状検索のアルゴリズムの性能を競うコンテスト
  - 2006年から開催され、ＣＡＤモデル、顔、分子モデル、ハンドジェスチャーなど、毎年様々なトラックが開催される

- 大規模３Ｄ形状検索のトラックでは、３次元の一般物体認識技術が活用されている

大規模３D形状検索（SHREC'16, SHREC'17）

- クエリのＣＡＤモデルに対して検索結果に同じカテゴリの物体がどれだけ含まれるか競う

- 学習および評価に用いられるデータセットはShapeNet Core55のデータセット
  - 50,000以上のモデルが55のカテゴリに分類されている
  - 55のカテゴリの中に204のサブカテゴリがある

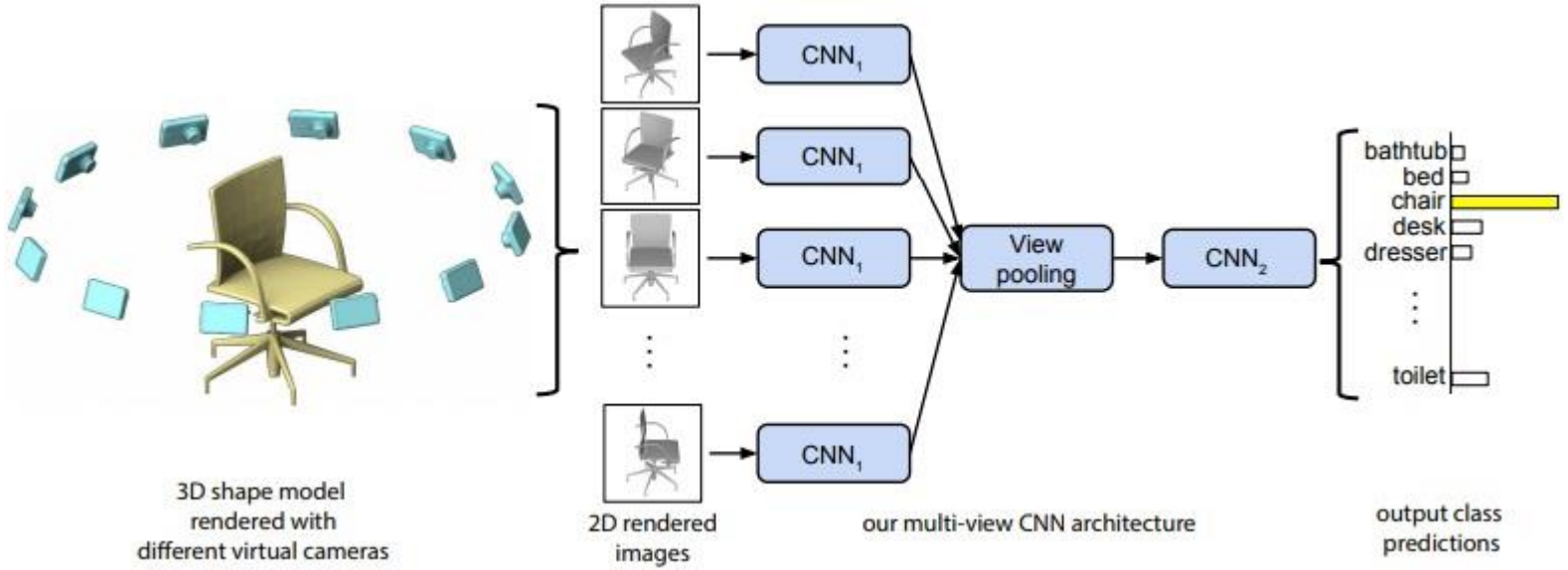
- 2016年に5チーム、2017年に8チームが参加し、そのほとんどがディープラーニングを用いていた

優勝したチームの手法

- 姿勢をそろえたモデルを対象としたタスク
  - 2016年　MVCNN: Multi-view Convolutional Neural Network (H. Su et al)
  - 2017年　RotationNet (A. Kanezaki)

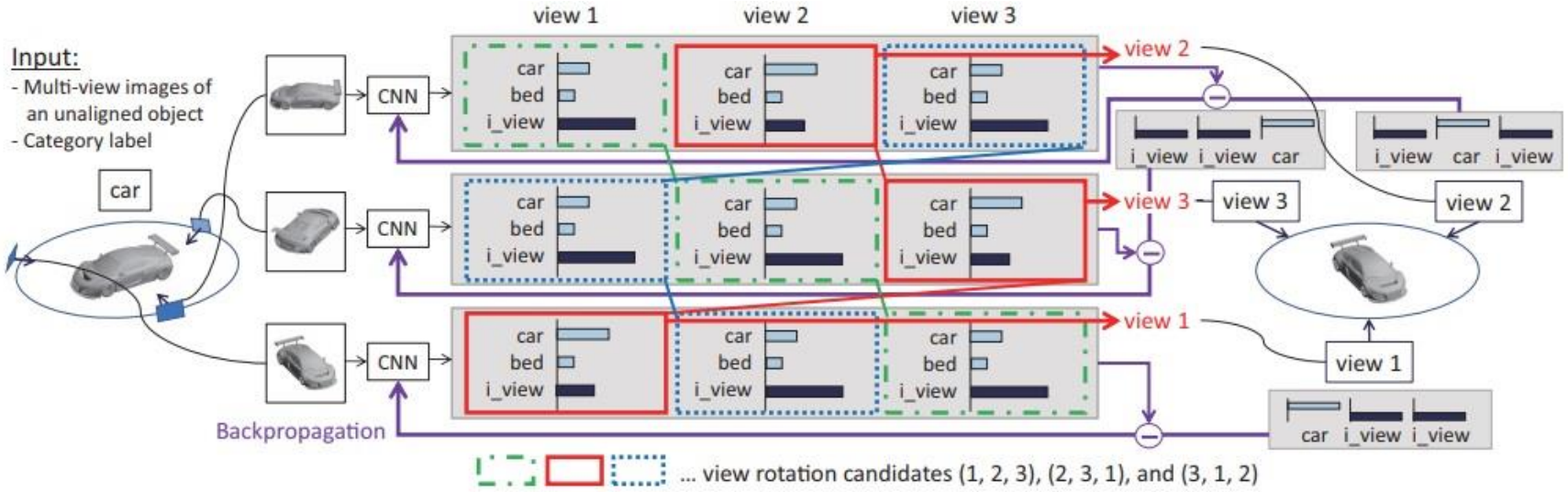
- どちらも多視点画像をCNNで学習する手法
  - レンダリングされた画像を入力に用いている
    - ２D画像のディープラーニングの手法を活用

MVCNN



Multi-view Convolutional Neural Networks for 3D Shape Recognition
H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. IEEE ICCV, 2015. より引用

RotationNet



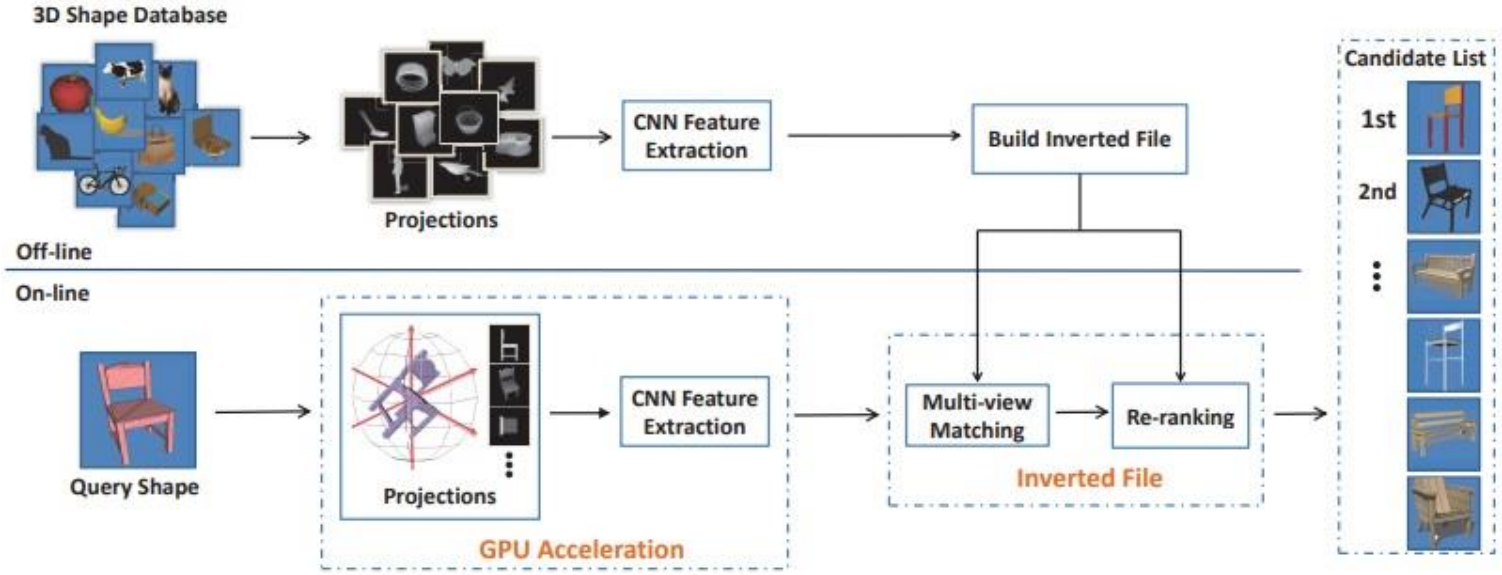
RotationNet: Joint Object Categorization and Pose Estimation Using Multiviews from Unsupervised Viewpoints
A. Kanezaki, Y. Matsushita, and Y. Nishida (https://arxiv.org/pdf/1603.06208.pdf) より引用

優勝したチーム

- 姿勢がそろっていないモデルを対象としたタスク
  - 2016年　GIFT: A Real-time and Scalable 3D Shape Search Engine (S. Bai et al)
  - 2017年　DLAN: Deep aggregation of local 3D geometric features (T. Furuya et al)

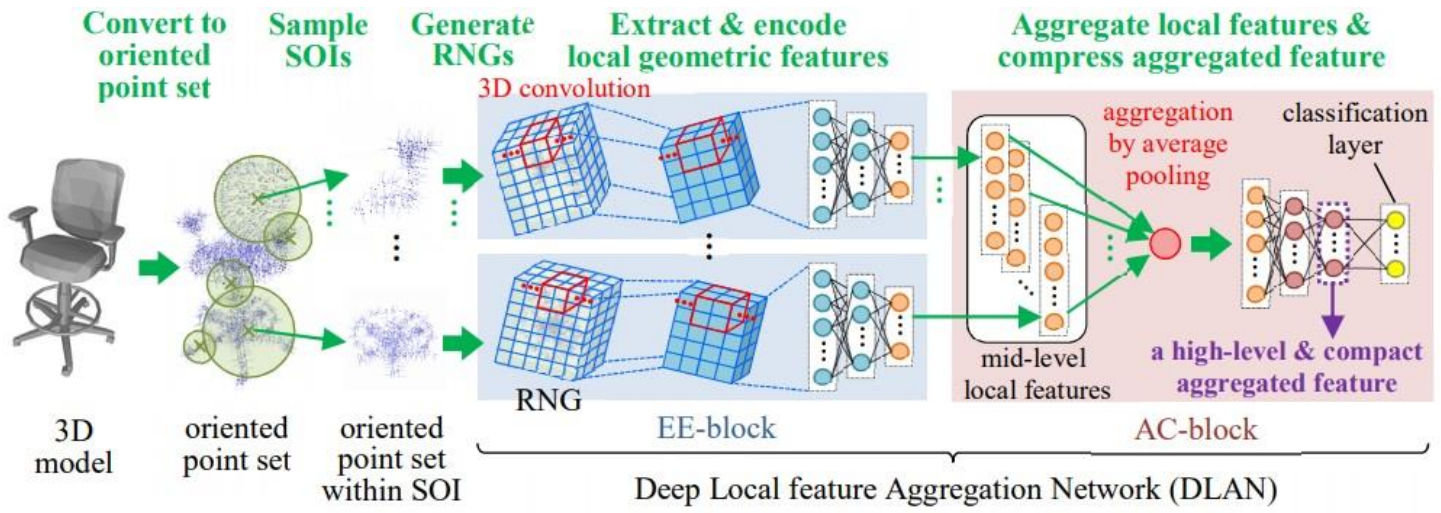
- GIFTは多視点画像を用いる手法
- DLANは点群を用いる手法

GIFT



GIFT: A Real-time and Scalable 3D Shape Search Engine
S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. J. Latecki. CVPR 2016 より引用

DLAN



Deep Aggregation of Local 3D Geometric Features for 3D Model Retrieval
T. Furuya, R. Ohbuchi. BMVC 2016 より引用

今後の予想（希望的観測も含めて）

- 特徴量を抽出してからニューラルネットワークに入力するのではなく、end-to-endの機械学習が増えてくる可能性がある