深層畳み込みニューラルネットワークによる画像特徴抽出と転移学習

東京大学 大学院情報理工学系研究科
創造情報学専攻
中山 英樹

目次

- １．画像認識分野におけるdeep learningの歴史と発展
- ２．畳み込みニューラルネット (CNN)を用いた転移学習
- ３．実践方法

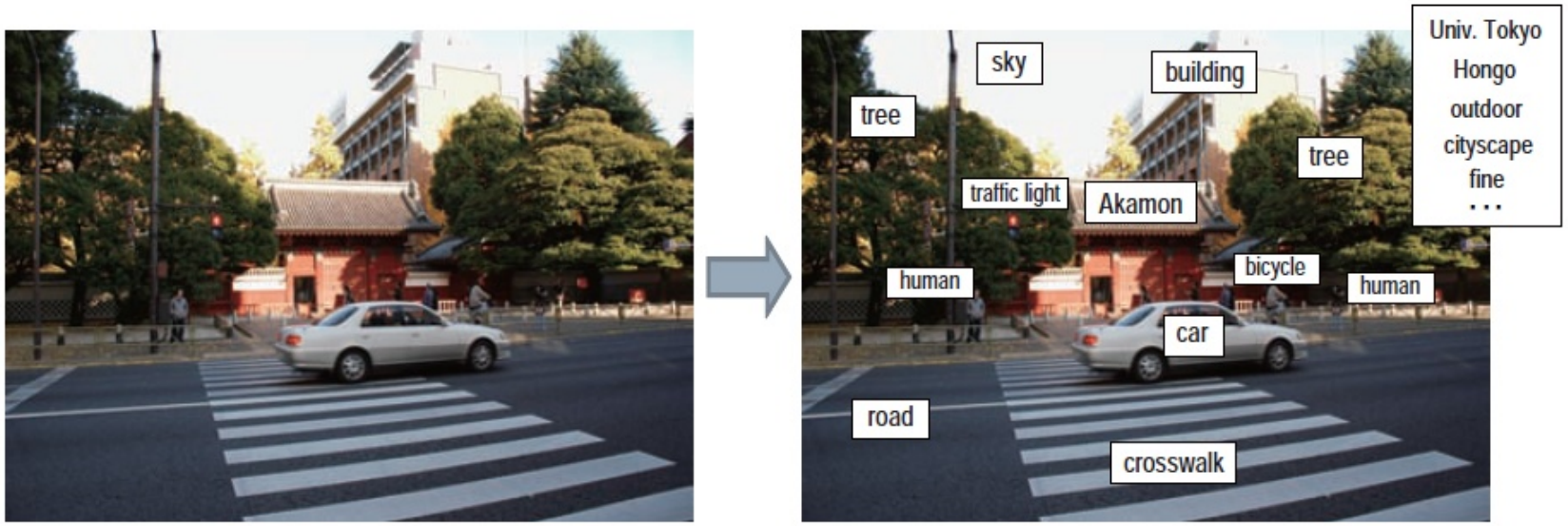
目次

- １．画像認識分野におけるdeep learningの歴史と発展
- ２．畳み込みニューラルネット (CNN)を用いた転移学習
- ３．実践方法

一般画像認識（一般物体認識）

- 制約をおかない実世界環境の画像を単語で記述
  - 一般的な物体やシーン、形容詞、印象語
  - 2000年代以降急速に発展（コンピュータビジョンの人気分野）
  - 幅広い応用先

    デジタルカメラ、ウェアラブルデバイス、画像検索、ロボット、…

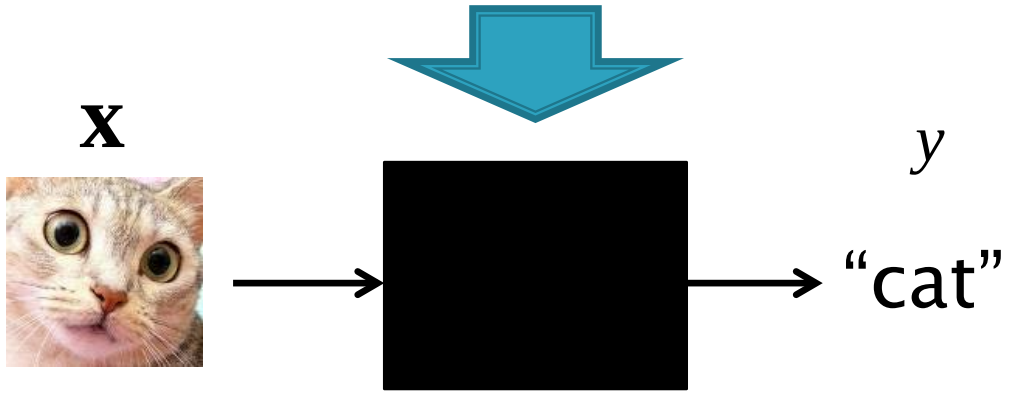


物体カテゴリ識別

- 機械学習（教師付）

$$\{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$$

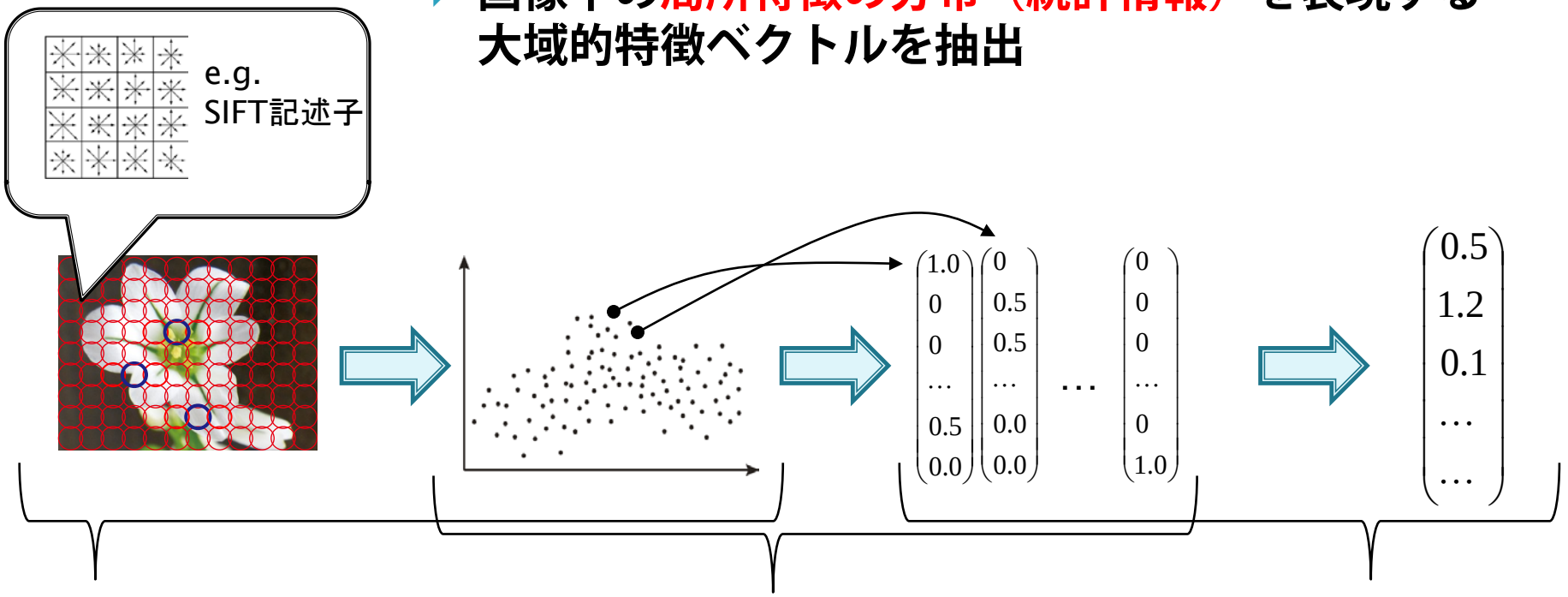
大量のラベル付き訓練データ
(x:画像，y:ラベル)



未知のデータ（学習データに含まれない）のカテゴリを正しく認識させることが目標

深層学習以前の画像特徴抽出の枠組

- **画像中の局所特徴の分布（統計情報）を表現する大域的特徴ベクトルを抽出**



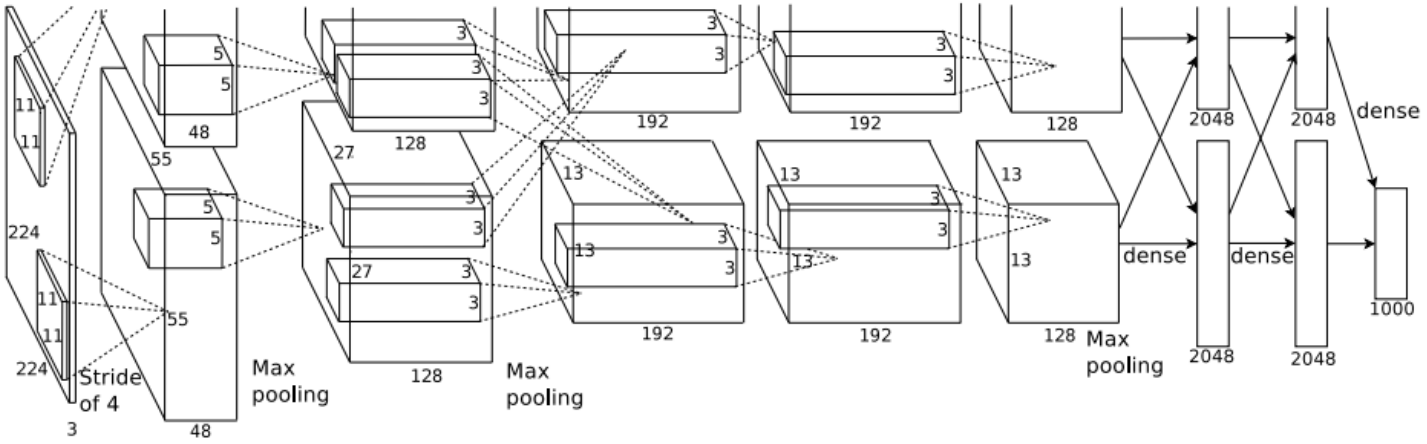
1. 局所特徴抽出
- SIFT, SURF, HOG, etc.
- Dense sampling
  （回転、スケールの
  正規化なし）

2. 量子化
- **K-means**
- **スパースコーディング**
- **Gaussian mixture model**

3. プーリング
- **最大値プーリング**
- **平均値プーリング**

Deep learning ブレイク後の画像認識

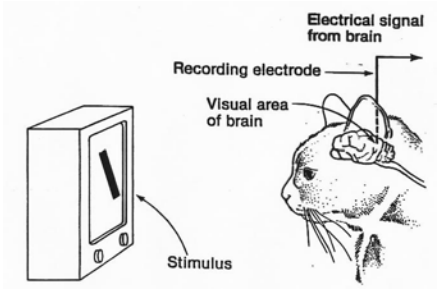
- 畳み込みニューラルネットワーク（CNN)
  - 脳の視覚野の構造を模倣した多層パーセプトロン
  - ニューロン間の結合を局所に限定（パラメータ数の大幅な削減）

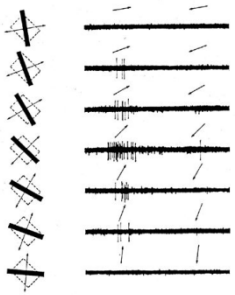


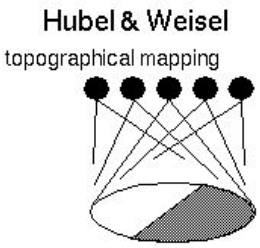
[A. Krizhevsky et al., NIPS'12]

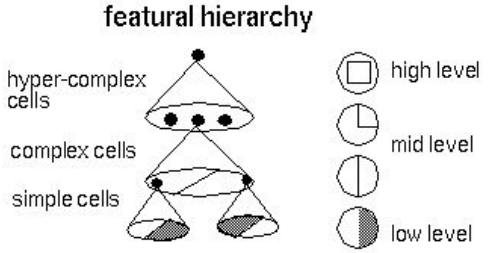
歴史

- Hubel and Wiesel (1950~60年代)
  - 一次視覚野(V1)の発見、分析
  - 単純細胞・複雑細胞・超複雑細胞

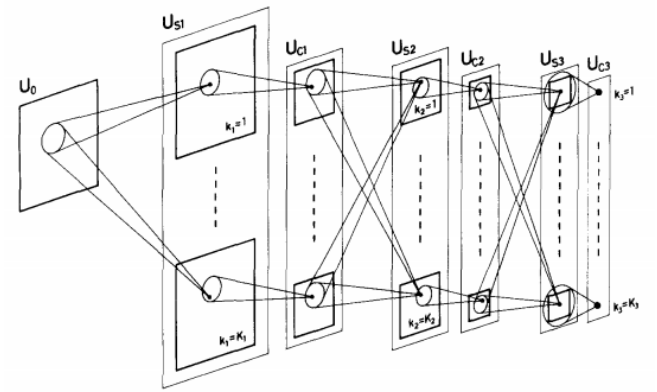








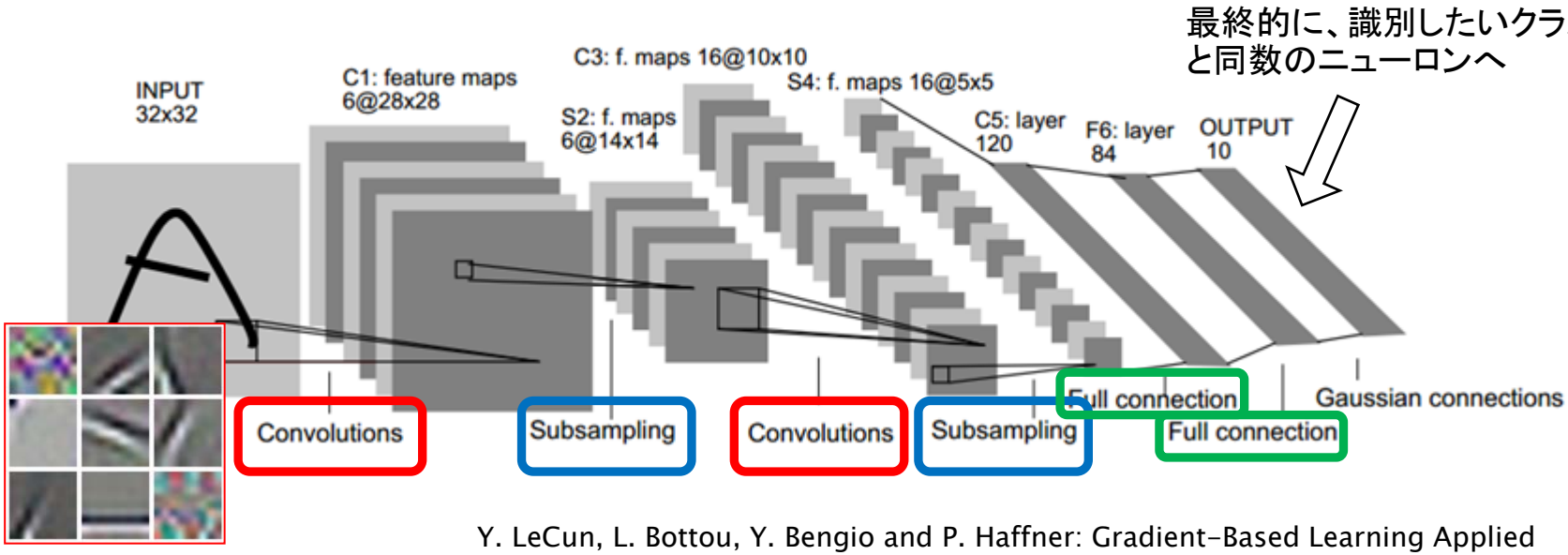
- ネオコグニトロン （福島邦彦先生、1980年)



Kunihiko Fukushima, “Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position“, Biological Cybernetics, 36(4): 93-202, 1980.

Convolutional neural network (CNN)

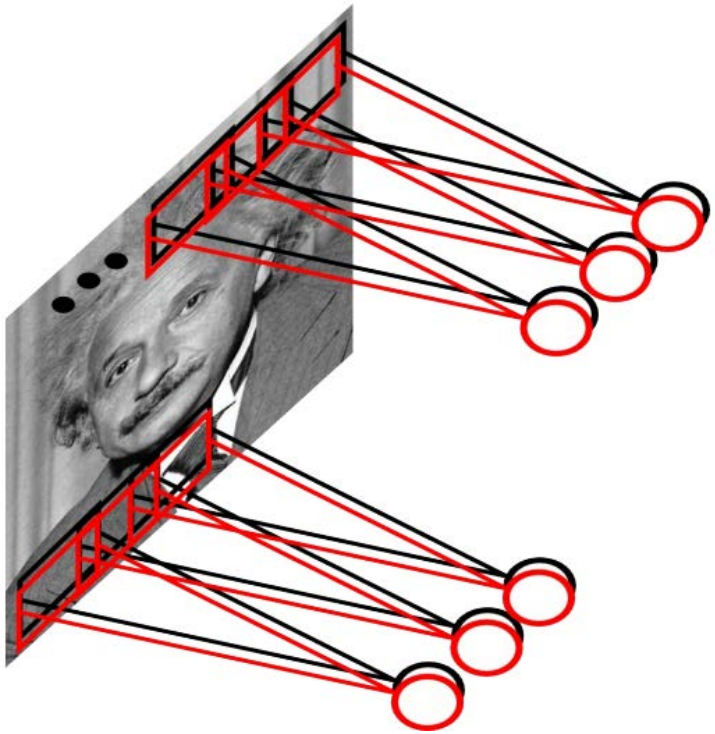
- 局所領域(受容野)の畳み込みとプーリングを繰り返す多層ネットワーク
  - 段階的に解像度を落としながら、局所的な相関パターンを抽出
  - 要するに、さまざまな解像度での特徴の共起をみている
  - 誤差逆伝播法による全体最適化



Y. LeCun, L. Bottou, Y. Bengio and P. Haffner: Gradient-Based Learning Applied to Document Recognition, Proceedings of the IEEE, 86(11):2278-2324, 1998.

畳み込み層

- 各フィルタのパラメータは全ての場所で共有
  - 色の違いは異なる畳み込みフィルタを示す



※もちろん入力は生画像のみ
とは限らない（中間層など）

非線形活性化関数（とても重要）

$$r = \phi(\mathbf{w} * \mathbf{h} - \theta)$$

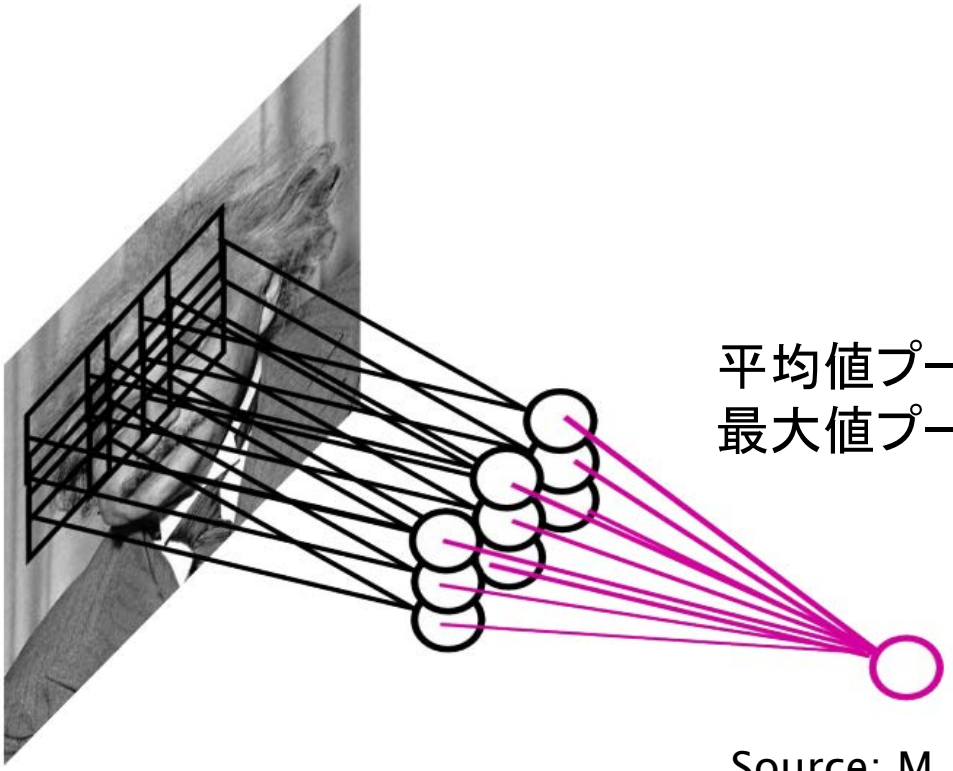
フィルタの係数　　入力　　バイアス

例えば、5x5の畳み込み、
10チャンネルの入力の場合、
5x5x10=250個

Source: M. Ranzato, CVPR'14 tutorial slides

プーリング層

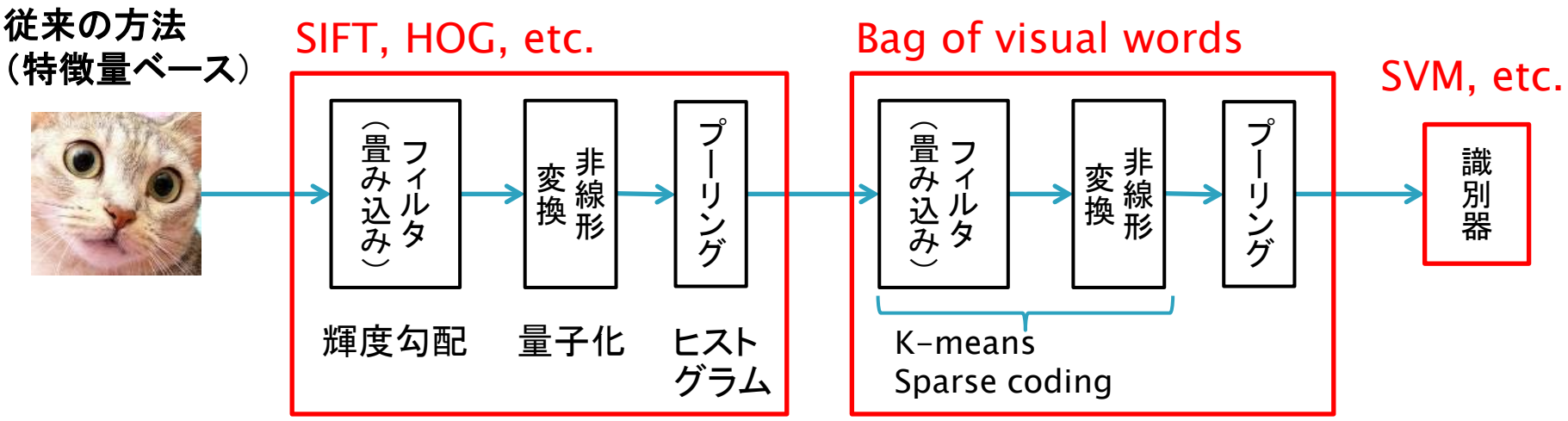
- 一定領域内の畳み込みフィルタの反応をまとめる
  - 領域内での平行移動不変性を獲得

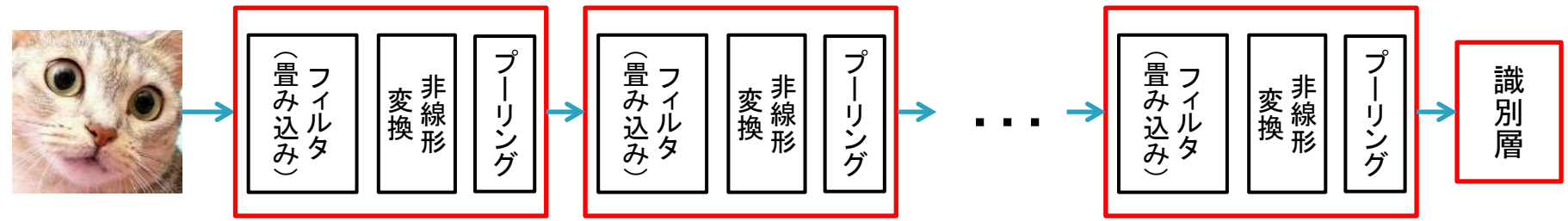


平均値プーリング、
最大値プーリングなど

Source: M. Ranzato, CVPR'14 tutorial slides

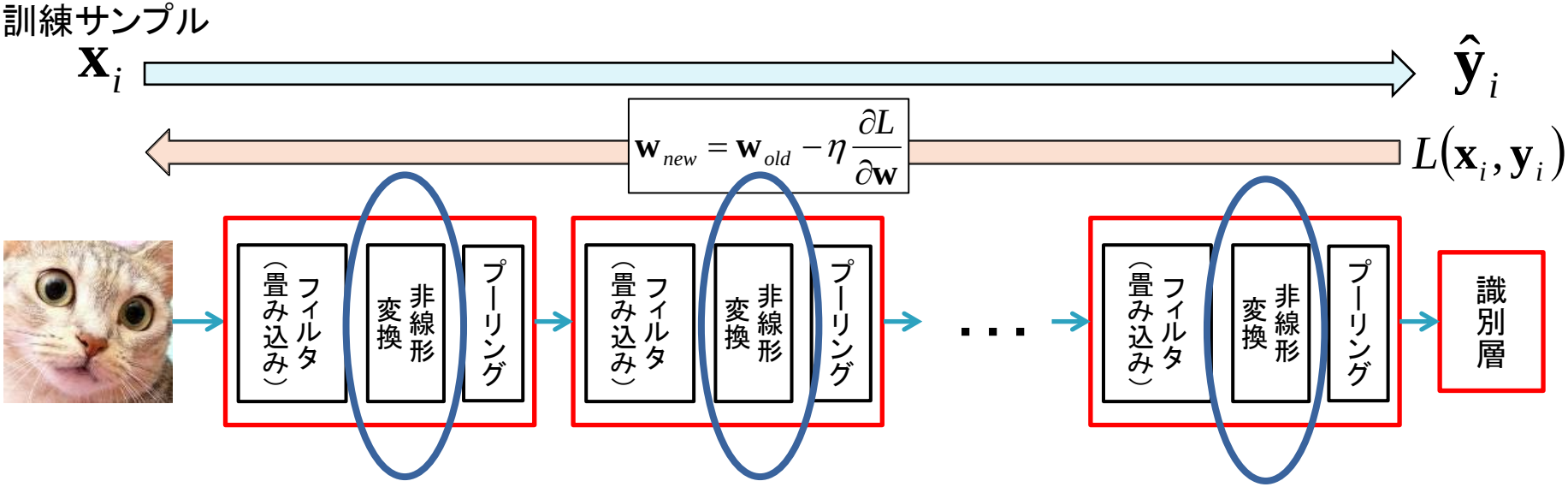
CNNと従来の特徴抽出法の関係





ニューラルネットは全てを学習で決める

- End-to-endでパラメータを最適化
- 誤差逆伝播法
- 非線形変換(活性化関数)の設計が重要
  - 少なくとも微分可能でないといけない



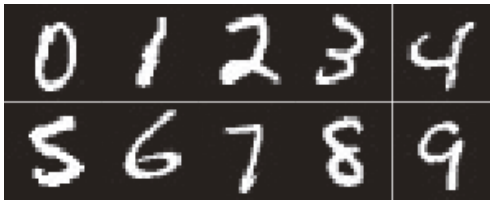
CNNの冬の時代：2010年頃まで

①計算機能力の不足

- 多数のフリーパラメータを扱うことは現実的に困難であった

- 小さな画像を用いた基礎研究が主流

  - MNISTデータセット [LeCun]
    - 文字認識、28 x 28ピクセル、6万枚



  - CIFAR-10/100 データセット [Krizhevsky]
    - 物体認識、32 x 32ピクセル、5万枚

CNNの冬の時代：2010年頃まで

②データの不足

- パラメータ最適化の効果がほとんど得られず

- Caltech-101 [Fei-Fei et al., 2004]
  - 一般的な解像度の画像データセット (9144枚、102クラス)
  - CNNを適用した場合、ランダム結合と最適化後でほとんど性能差なし
    [Jarrette, ICCV'09]





http://www.vision.caltech.edu/Image_Datasets/Caltech101/

CVPR 2012- One Author's Withdrawal Statement

"We are withdrawing it for three reasons: 1) the scores are so low, and the reviews so ridiculous, that I don't know how to begin writing a rebuttal without insulting the reviewers; 2) we prefer to submit the paper to ICML where it might be better received. （中略）

Getting papers about feature learning accepted at vision conference has always been a struggle, and I 've had more than my share of bad reviews over the years. Thankfully, quite a few of my papers were rescued by area chairs. （中略）

This time though, the reviewers were particularly clueless, or negatively biased, or both. （中略）

So, I 'm giving up on submitting to computer vision conferences altogether. CV reviewers are just too likely to be clueless or hostile towards our brand of methods. Submitting our papers is just a waste of everyone's time （中略）

Regardless, I actually have a keynote talk at [Machine Learning Conference], where I'll be talking about the results in this paper."

Figure from
[Ramanan et al, 2009]

Large-scale recognition

2010

カテゴリ数:10^3~10^4

サンプル数: 10^6~10^7

Figure from Russakovsky et al.

ImageNet Large-scale Visual Recognition Challenge (ILSVRC)

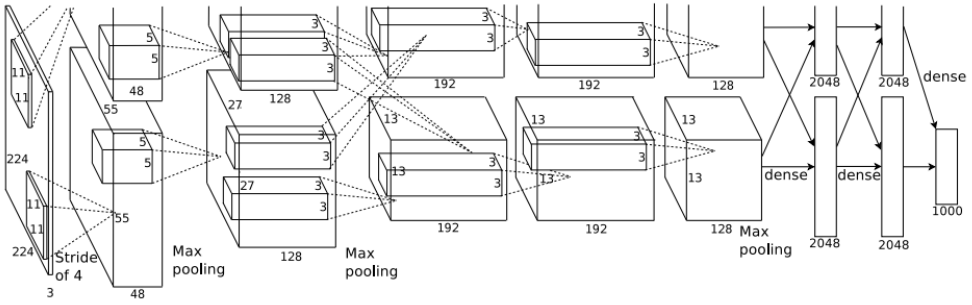
Russakovsky et al., "ImageNet Large Scale Visual Recognition Challenge", 2014.

- **ImageNetのデータの一部を用いたフラッグシップコンペティション (2010年より開催)**
  - **ImageNet [Deng et al., 2009]**
    - クラウドソーシングにより構築中の大規模画像データセット
    - 1400万枚、2万2千カテゴリ（WordNetに従って構築）

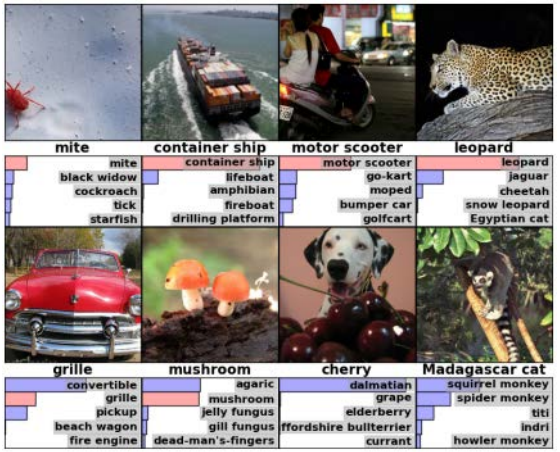
- **コンペでのタスク**
  - **1000クラスの物体カテゴリ分類**
    - 学習データ120万枚、検証用データ5万枚、テストデータ10万枚
  - **200クラスの物体検出**
    - 学習データ45万枚、検証用データ2万枚、テストデータ4万枚

ILSVRC 2012 の衝撃

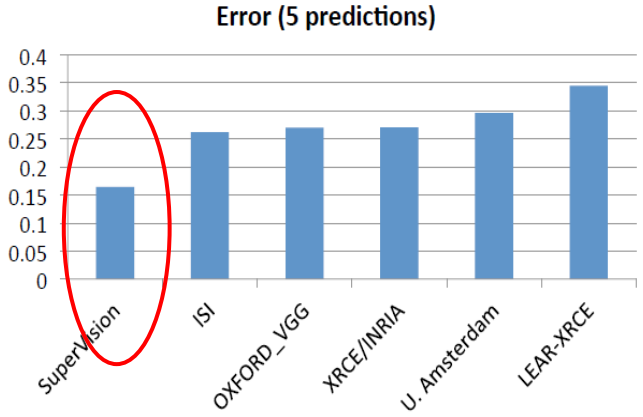
- 1000クラス識別タスクで、deep learning を用いたシステムが圧勝
  - トロント大学Hinton先生のチーム (AlexNet)



[A. Krizhevsky *et al.*, NIPS'12]

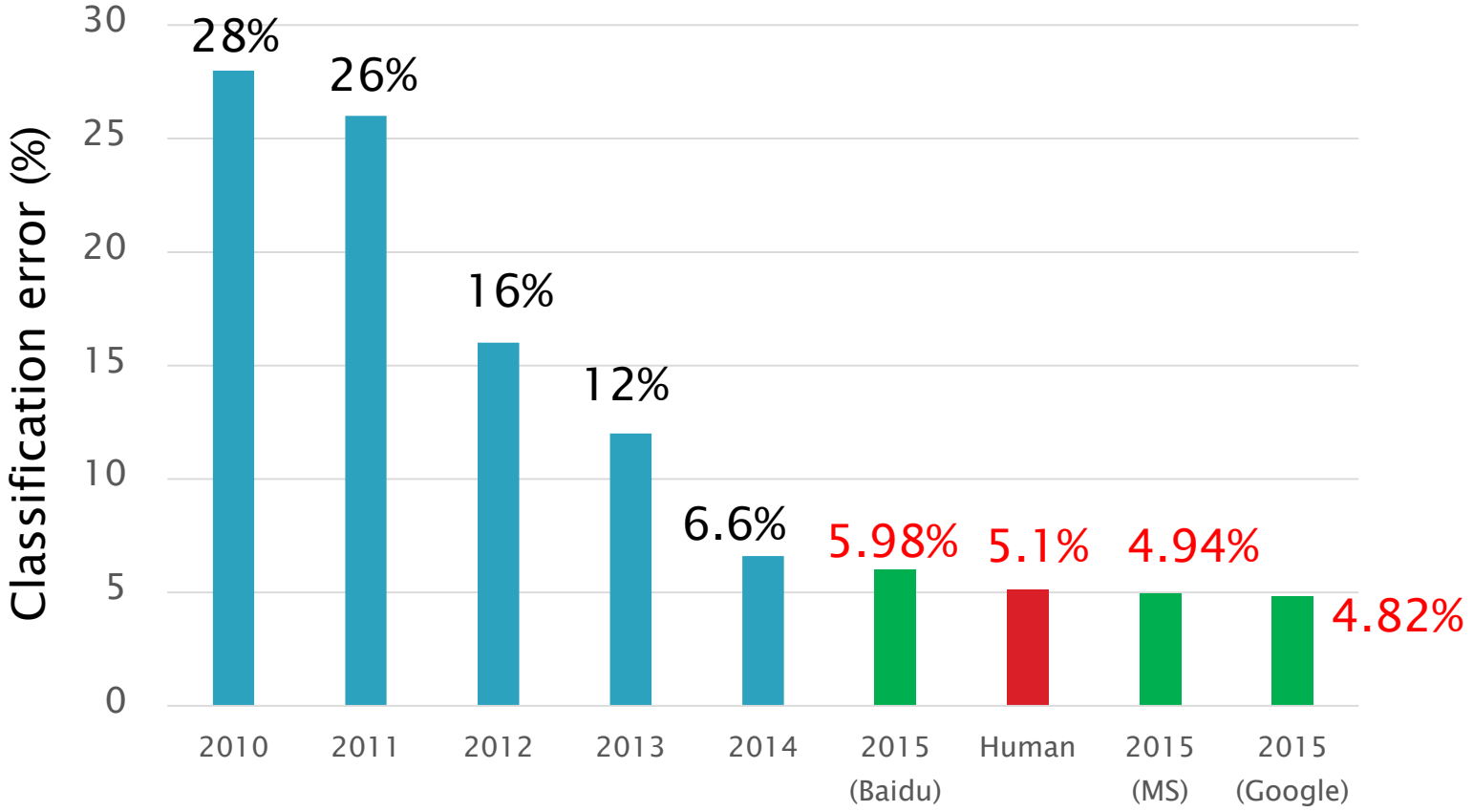


エラー率が一気に10%以上減少！
(※過去数年間での向上は1~2%)



その後も急激な性能向上が続く

- エラー率が 16% (2012) → 4.8% (2015)

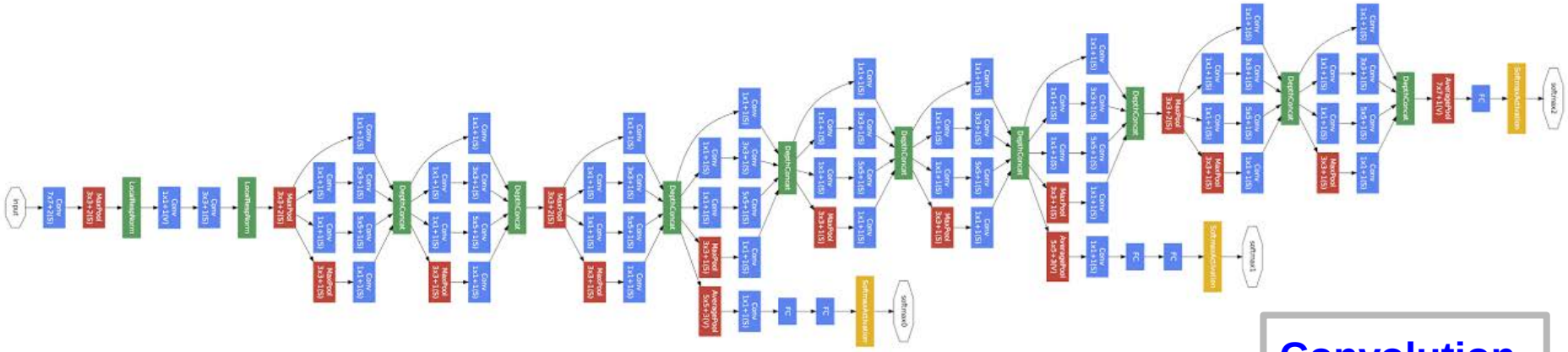


Wu et al., “Deep Image: Scaling up Image Recognition”, 2015.

He et al., “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”, 2015.

2014年のCNN

- GoogLeNet (22層) [Szegedy et al., 2014]
  - NINベース
  - ILSVRC 2014 で優勝
  - 独自の並列分散フレームワークで学習

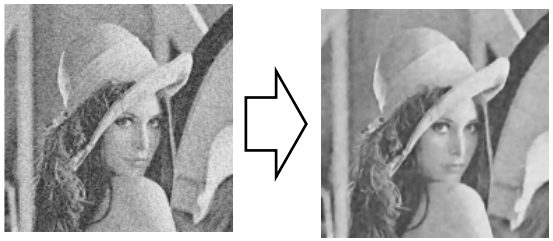


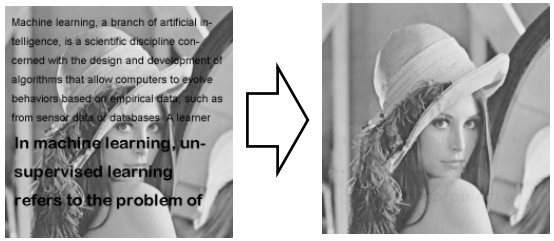
Convolution
Pooling
Softmax
Other

http://www.image-net.org/challenges/LSVRC/2014/slides/GoogLeNet.pptx

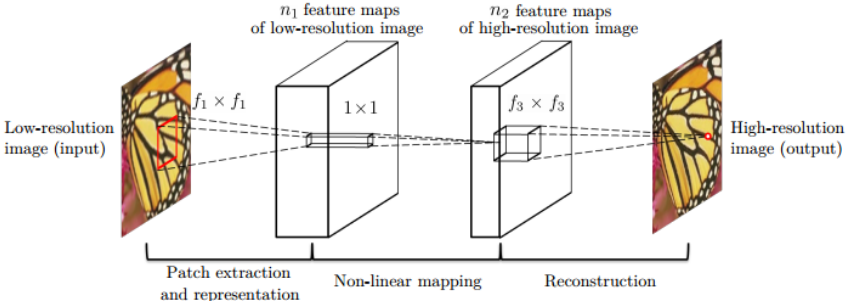
画像処理の諸分野でも広い応用

- デノイジング・インペインティング [Xie et al., NIPS'12]
  - 画像のノイズ除去
  - Stacked denoising auto-encoder





- 超解像 [Dong et al., ECCV'14]
  - 低解像度画像から
    高解像度画像を復元（推定）



- ボケ補正 [Xu et al., NIPS'14]

目次

- １．画像認識分野におけるdeep learningの歴史と発展
- ２．畳み込みニューラルネット(CNN)を用いた転移学習
- ３．実践方法

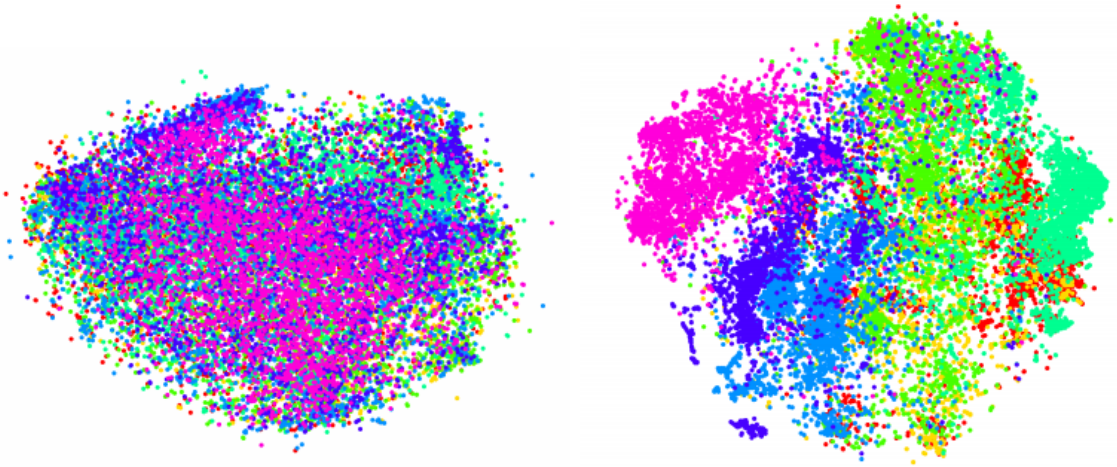
ILSVRC 2012 での議論

- ① ネットワーク内で何が起こっているのか？何を見ているのか？
- ② 学習で得られた知識、構造は他タスクへ一般化できるのか？

各層のニューロンの出力

- 層を上るにつれ、クラスの分離性能が上がる

ILSVRC'12 の
validation data
（色は各クラスを示す）



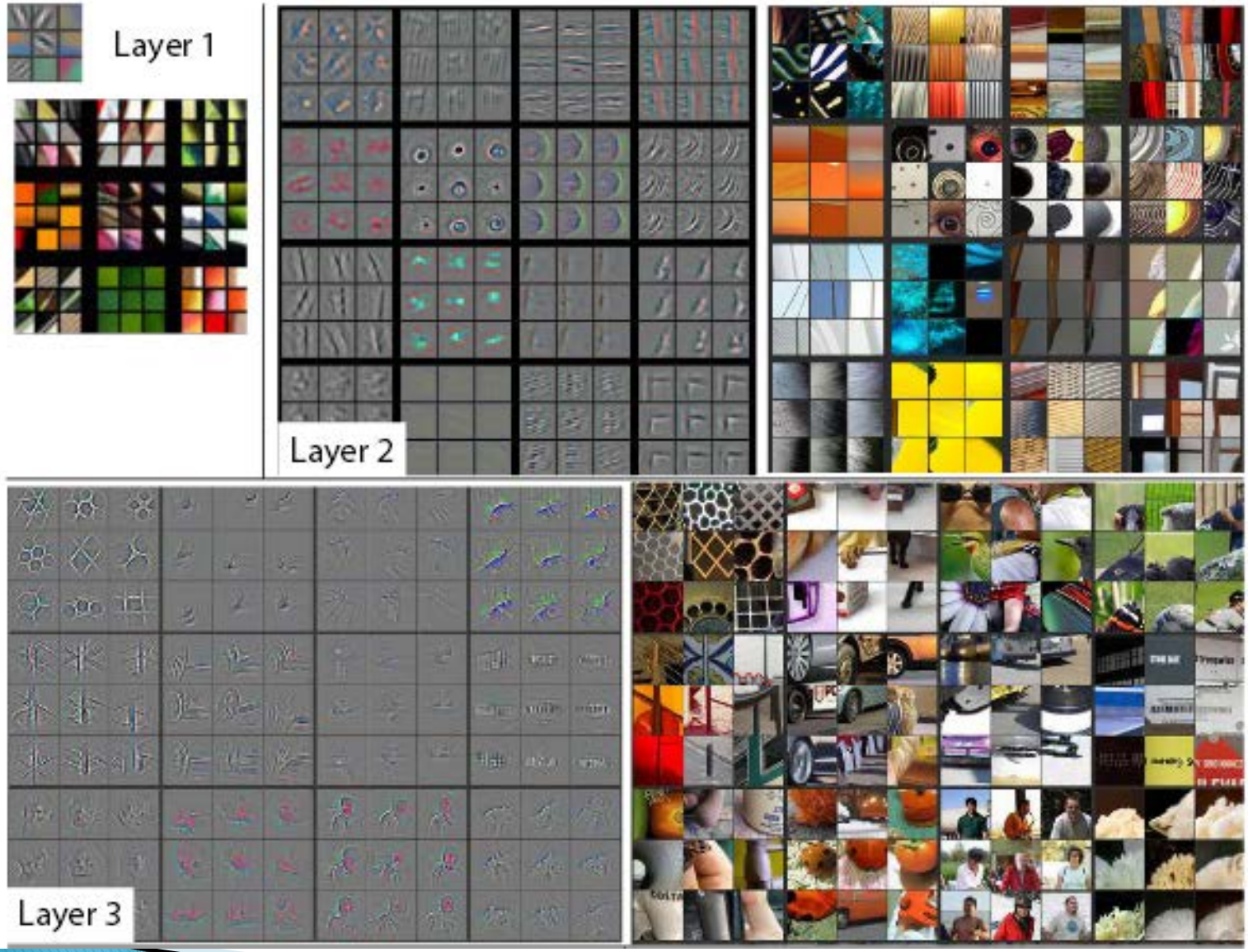
(c) DeCAF_1 第1層

(d) DeCAF_6 第6層

J. Donahue et al., "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", In Proc. ICML, 2014.

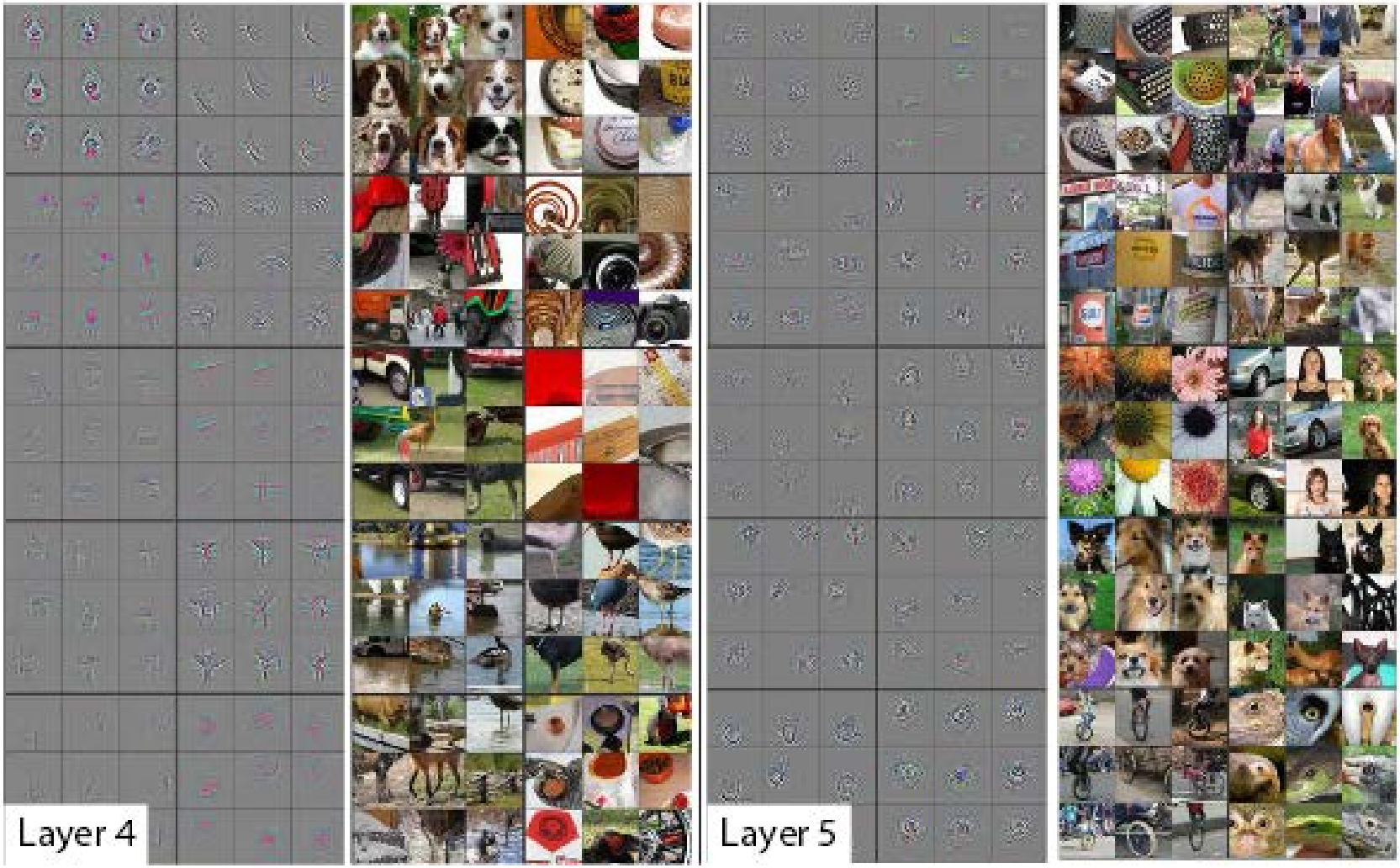
中間層の可視化

Matthew D. Zeiler and Rob Fergus, “Visualizing and Understanding Convolutional Networks”, In Proc. ECCV, 2014.



中間層の可視化

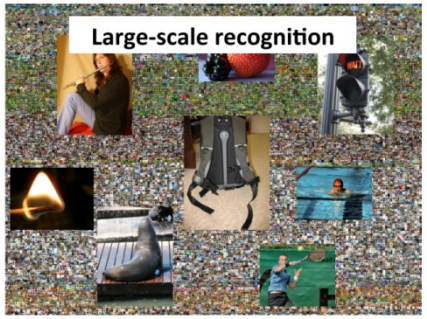
Matthew D. Zeiler and Rob Fergus, “Visualizing and Understanding Convolutional Networks”, In Proc. ECCV, 2014.



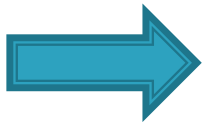
転移学習

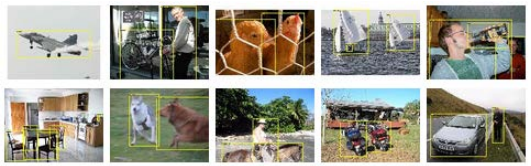
- 転移学習
  - 新規タスクの効果的な仮説を効率的に見つけ出すために，一つ以上の別のタスクで学習された知識を得て，それを適用する問題 [神嶌, 2010]

- 画像認識の場合
  - あるドメイン(データセット)で学習した識別器(特徴抽出器)を他ドメインでの識別器構築に役立てる



ImageNet ILSVRC'12
130万枚、1000クラス

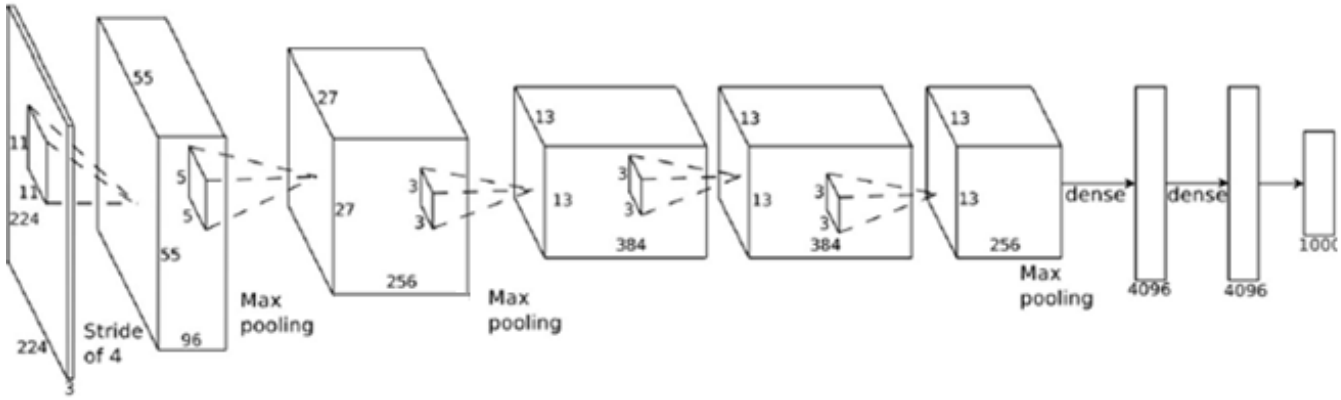




PASCAL VOC 2007
5千枚、20クラス

CNNを用いた転移学習

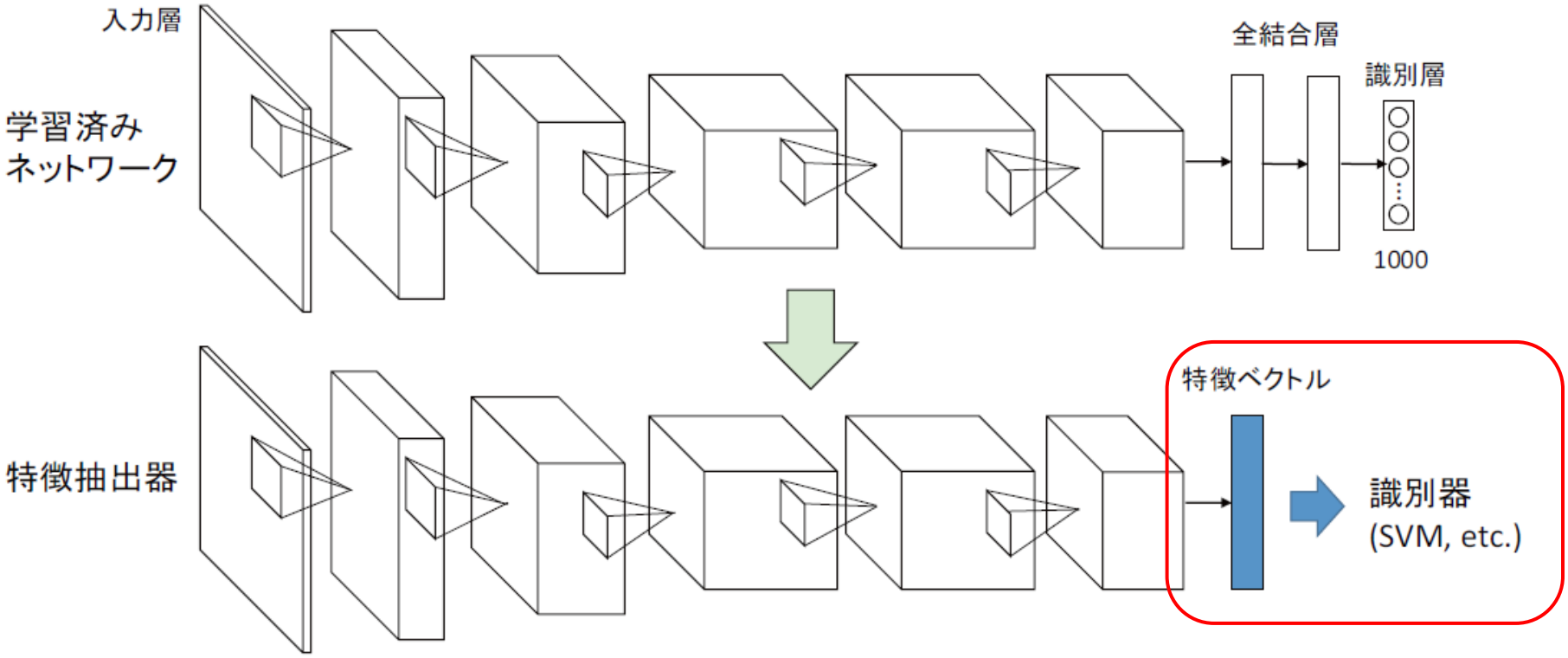
- 学習済ネットワークを転用
  - 転用先のタスクと何らかの関係がある(と期待できる)十分に大規模なデータセットで学習したネットワーク



- 大きく分けると二つのアプローチ
  - 特徴抽出器として利用 (Pre-trained feature)
  - Fine-tuning

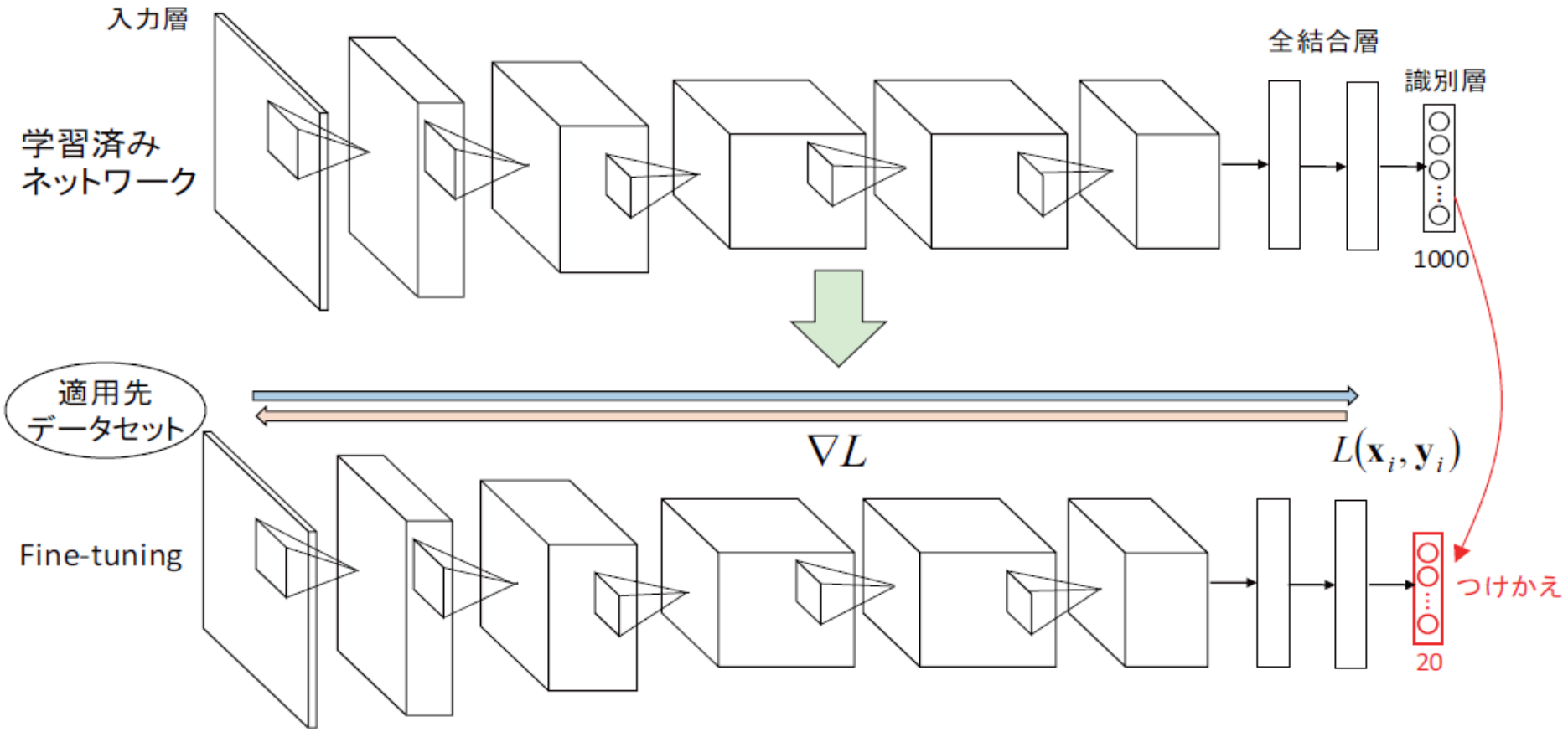
特徴抽出器として利用

- 学習済ネットワークを特徴抽出器として用いる
  - 中間層の出力を利用して識別器を構築
  - どの層を選ぶかは重要（タスク依存？）



Fine-tuning

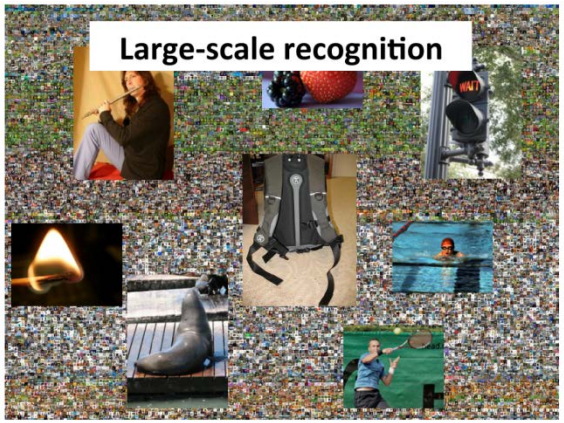
- 学習済ネットワークを初期値とし、適用先データセットで
  さらに学習を進める
- 教師なし事前学習とは異なる概念であることに注意



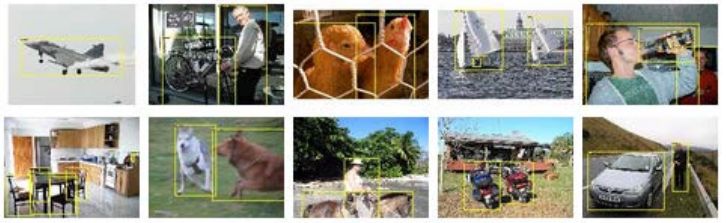
学習済みネットワークの効果

- ILSVRC 2012 → VOC 2007 の例 (検出成功率、mAP%)
  - フルスクラッチCNN:　40.7
  - Pre-trained feature:　45.5
  - Fine tuning:　54.1

Agrawal et al., “Analyzing the Performance of Multilayer Neural Networks for Object Recognition”, In Proc. ECCV, 2014.



ImageNet ILSVRC’12
130万枚、1000クラス

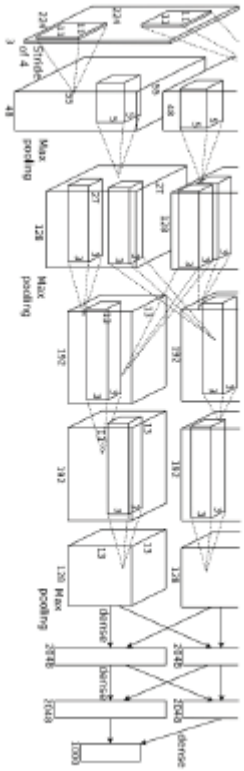


PASCAL VOC 2007
5千枚、20クラス

学習済みネットワーク自体の性能も重要

- 2012年以降も劇的な向上が続いている

AlexNet (8層)



VGG (19層)

GoogLeNet (22層)

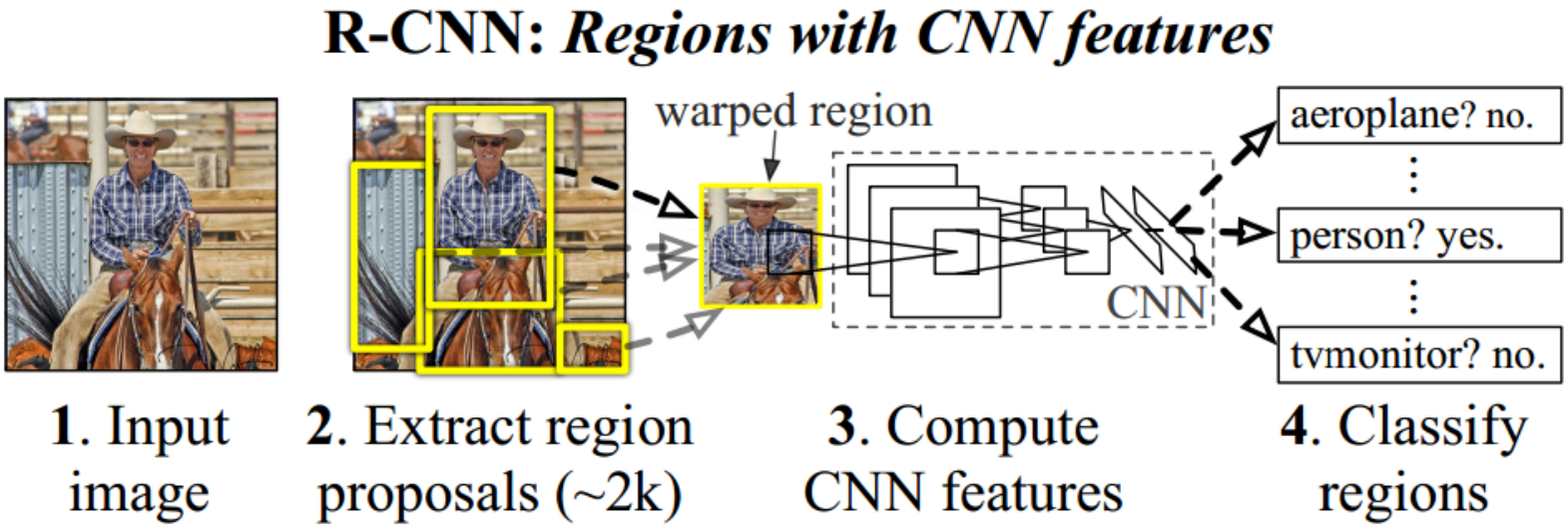
異なる学習済みネットワークの比較

- ILSVRC 2012 → VOC 2007 でfine-tuningをした場合の性能比較 (検出成功率、mAP%)
  - AlexNet : 58.5 ← ILSVRC'12 winner, エラー率16%
  - Small VGG: 60.2
  - VGG-16: 66.0 ← ILSVRC'14 二位, エラー率7.4%

Redmon et al., "You Only Look Once : Unied, Real-Time Object Detection", arXiv preprint, 2015.

物体検出(detection)への応用

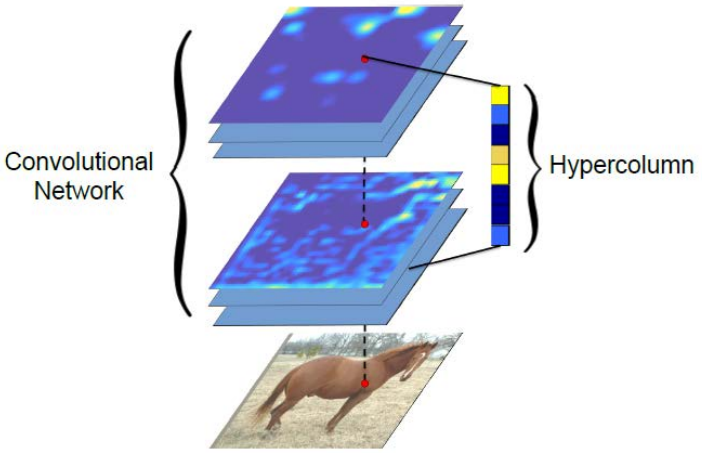
- R-CNN [Girshick et al., CVPR'2014]
  - 物体の領域候補を多数抽出（これ自体は別手法）
  - 領域を識別するためのCNNをfine-tuning

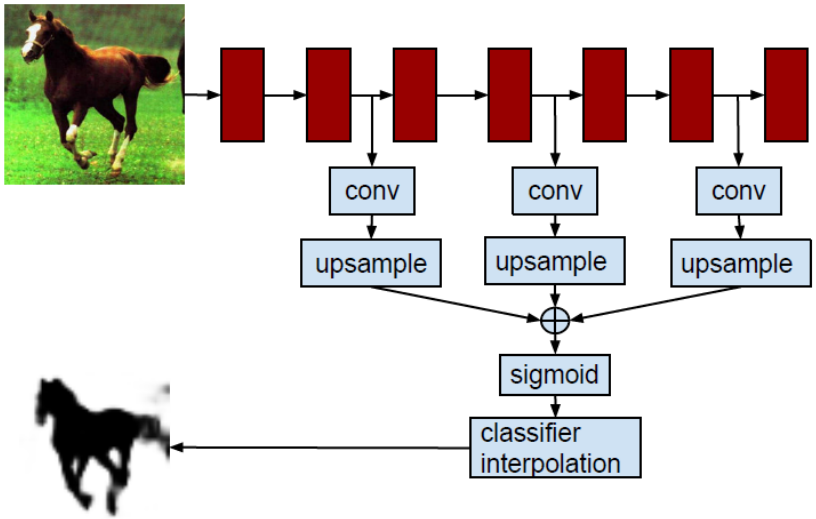


Semantic segmentation への応用

- ピクセルレベルでの物体領域認識

- Hypercolumns [Hariharan et al., 2014]
  - 各ピクセルについて、位置的に対応する中間層の反応をすべてとり特徴量として用い、識別器を学習





注意：何でもうまくいくわけではない

- Pre-trainingに用いる外部データセットが、所望のタスクを内包するものでなければ効果が薄い
  - ImageNetはあくまでも物体認識のデータセット

- 参考：Fine-grained competition 2013

https://sites.google.com/site/fgcomp2013/

Team	Aircraft	Birds	Cars	Dogs	Shoes	Overall
Inria-Xerox	**81.46**	71.69	**87.79**	52.90	**91.52**	**77.07**
CafeNet	78.85	**73.01**	79.58	57.53	90.12	75.82

Fisher vector

CNN (fine-tuning)

飛行機、車、靴データセットなど、ImageNet上にあまりデータが存在しないドメインに関してはターゲットの学習データのみ用いたFisher vector (BoVW) の方が良かった

他のデータセットの例

- MIT-Places [Zhou et al., NIPS'14]
  - 大規模シーンカテゴリデータセット (205カテゴリ、200万枚)

- 他のシーン認識タスクへの転移で高い性能

シーン認識データセット ⇨

	SUN397	MIT Indoor67	Scene15	SUN Attribute
Places-CNN feature	**54.32±0.14**	**68.24**	**90.19±0.34**	**91.29**
ImageNet-CNN feature	42.61±0.16	56.79	84.23±0.37	89.85

物体・行動認識データセット ⇨

	Caltech101	Caltech256	Action40	Event8
Places-CNN feature	65.18±0.88	45.59±0.31	42.86±0.25	94.12±0.99
ImageNet-CNN feature	**87.22±0.92**	**67.23±0.27**	**54.92±0.33**	**94.42±0.76**

目次

- １．画像認識分野におけるdeep learningの歴史と発展
- ２．畳み込みニューラルネット (CNN)を用いた転移学習
- ３．実践方法

計算環境

- ハードウェア
  - Fine-tuningにはGPU計算機が必要
  - ビデオメモリの容量がボトルネックになる場合が多い
    - メインメモリとの通信は遅い
    - ネットワークのパラメータはもちろん、できるだけ多くの学習サンプルをビデオメモリに積みたい

- Titan X (約15万円)
  - コストパフォーマンス的にお薦め

- Tesla K20 (約40万円), K40 (約80万円)
  - より信頼性が高い

オープンソースソフトウェア

- 2012年頃から、著名な研究チームによる主導権争い
  - Caffe/Decaf： UC Berkeley
  - Theano/Pylearn2： Univ. Montreal
  - Torch7： Univ. New York
  - Cuda-convnet2： Univ. Toronto (Alex Krizhevsky)
  - Chainer： PFI/PFN

Framework	License	Core language	Binding(s)	CPU	GPU	Open source	Training	Pretrained models	Development
Caffe	BSD	C++	Python, MATLAB	✓	✓	✓	✓	✓	distributed
cuda-convnet [7]	unspecified	C++	Python		✓	✓	✓		discontinued
Decaf [2]	BSD	Python		✓		✓	✓	✓	discontinued
OverFeat [9]	unspecified	Lua	C++,Python	✓				✓	centralized
Theano/Pylearn2 [4]	BSD	Python		✓	✓	✓	✓		distributed
Torch7 [1]	BSD	Lua		✓	✓	✓	✓		distributed

Y. Jia et al., “Caffe: Convolutional Architecture for Fast Feature Embedding”, ACM Multimedia Open Source Competition, 2014.

Caffe (BSD 2-Clause license)

- 画像系ではデファクトスタンダード
  - トップクラスに高速
  - オープンソースコミュニティとして確立

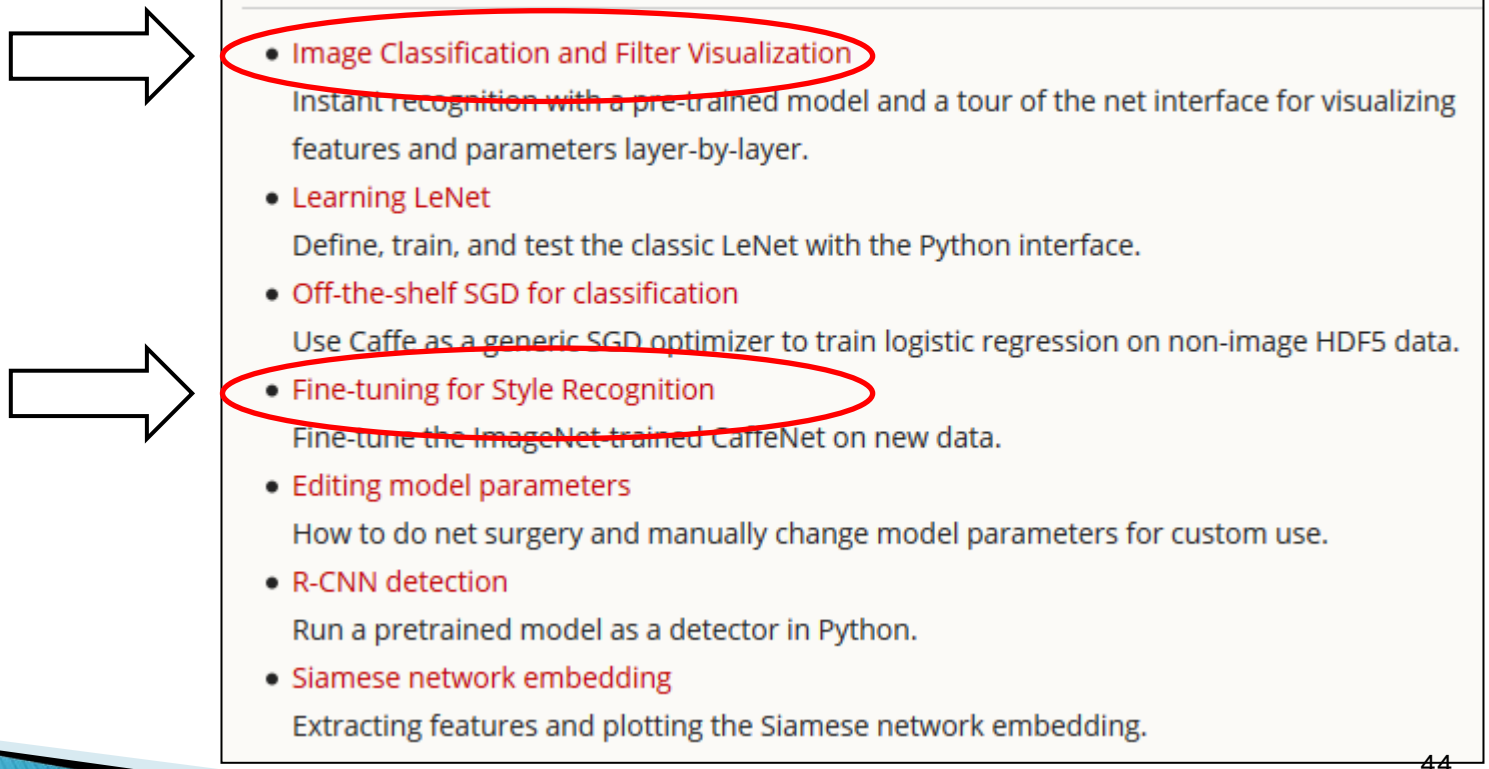
- Model Zoo
  - 各自が作った学習済みネットワークを共有する枠組み
  - AlexNetはもちろん、VGG、GoogLeNet なども
  - 最新の成果を極めて容易に試せる

Caffe (BSD 2-Clause license)

- Webドキュメントが充実　http://caffe.berkeleyvision.org/
  - ImageNet等の結果を再現可能
  - IPython notebookによるコード実例多数

Pre-trained feature

Fine tuning



学習済ネットワークの読み込み

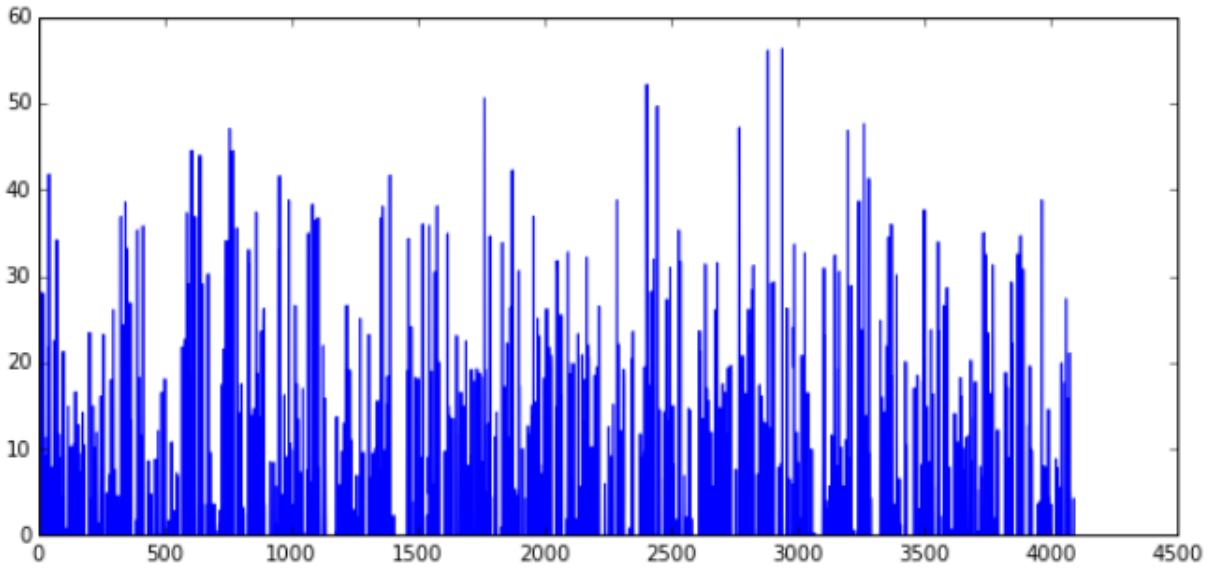
```
In [3]: caffe.set_mode_cpu()
net = caffe.Net(caffe_root + 'models/bvlc_reference_caffenet/deploy.prototxt',
                caffe_root + 'models/bvlc_reference_caffenet/bvlc_reference_caffenet.caffemodel',
                caffe.TEST)
```

```
In [6]: net.blobs['data'].data[...] = transformer.preprocess('data', caffe.io.load_image(caffe_root + 'examples/images/cat.jpg'))
out = net.forward()
```

画像のセット・フィードフォワード

```
In [36]: feat = net.blobs['fc6'].data[0]
plt.subplot(2, 1, 1)
plt.plot(feat.flat)
plt.subplot(2, 1, 2)
_ = plt.hist(feat.flat[feat.flat > 0], bins=100)
```

特徴量の取得



本日のまとめ

- 画像認識分野における深層学習の成功
  - 畳み込みネットワークの構造は重要
  - 大規模教師付データ(ImageNet)のもたらしたブレークスルー
  - 学習済みネットワークの転移は非常に有効

- 十分な教師付学習データの有無がボトルネック
  - 現在成功しているものはほぼ全てImageNet絡み
  - 静止画は比較的アノテーションが楽（クラウドソーシングがスケールしやすい）
  - 動画像認識等ではまだまだ