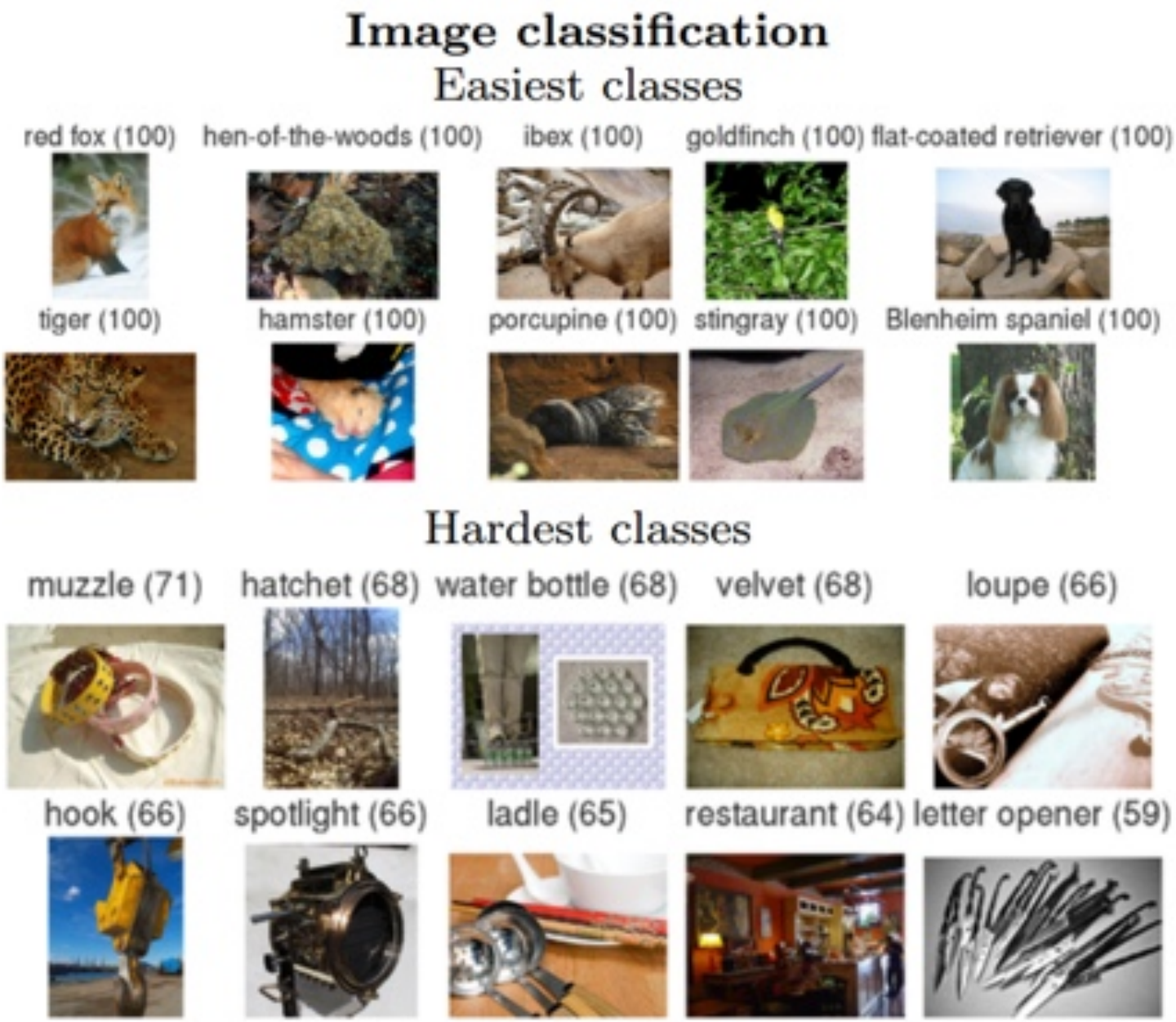
コンピュータビジョン
深層学習応用

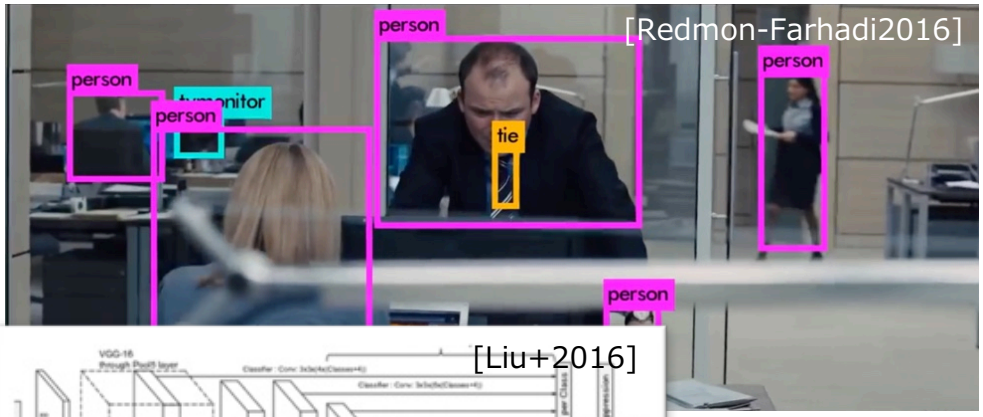
1000クラス物体認識



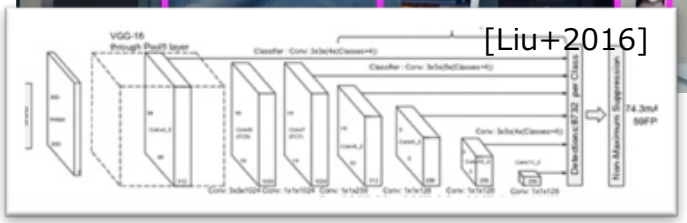
https://www.technologyreview.com/s/530561/the-revolutionary-technique-that-quietly-changed-machine-vision-forever/

認識

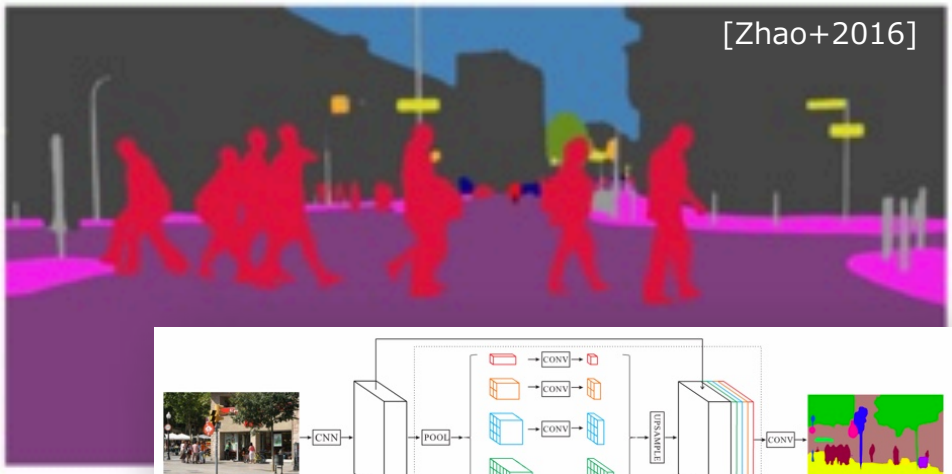
物体検出



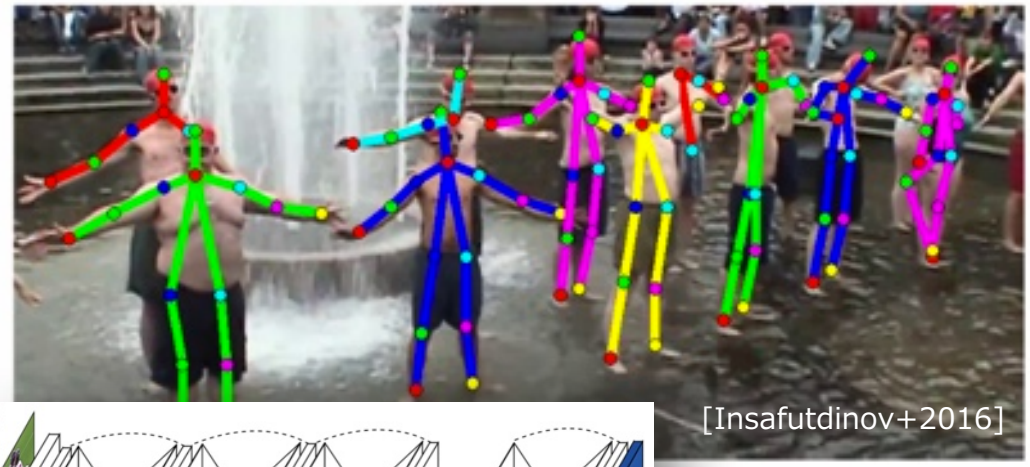
[Redmon-Farhadi2016]

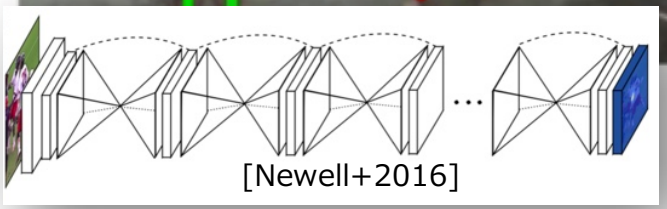

[Liu+2016]

画素レベルの認識

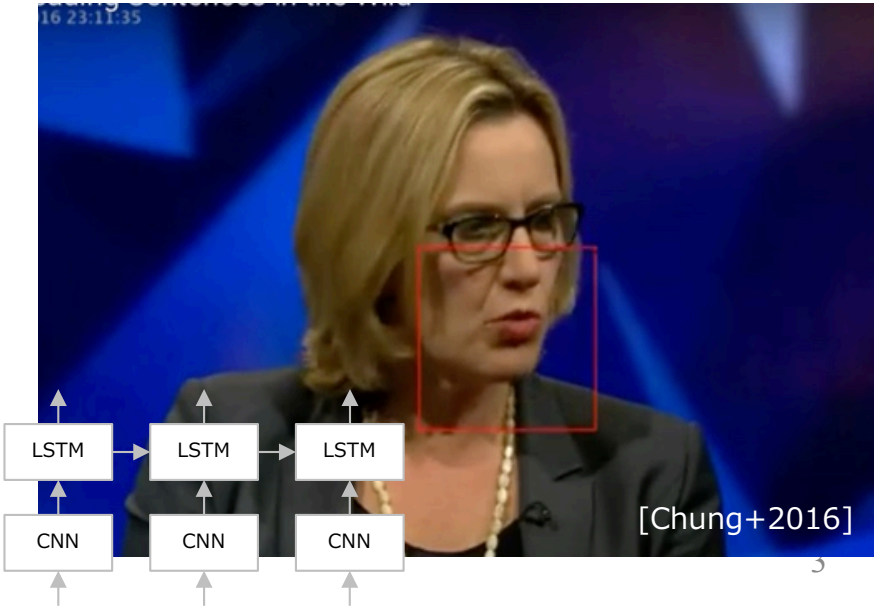

[Zhao+2016]

人体ポーズ


[Insafutdinov+2016]


[Newell+2016]

読唇

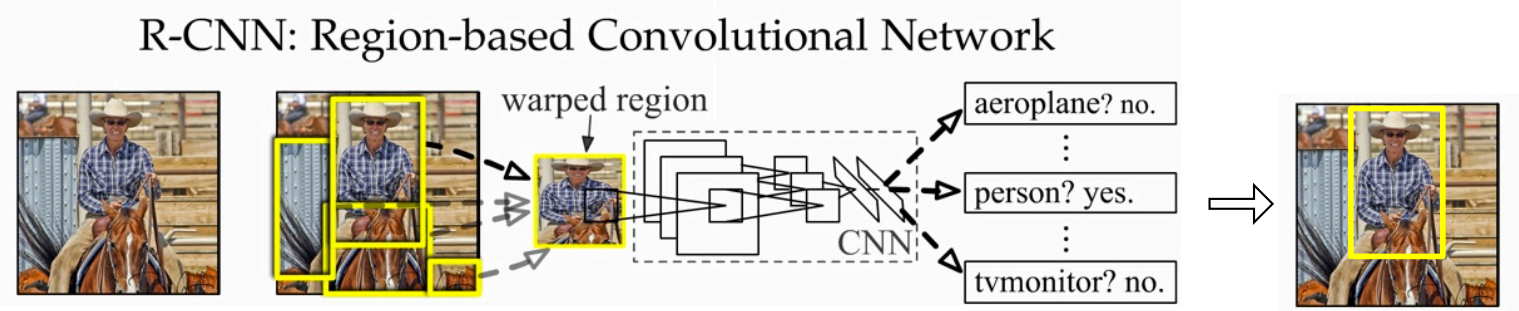


[Chung+2016]

物体検出

R-CNN/Faster-RCNN [Girshick+2015,Ren+2016]

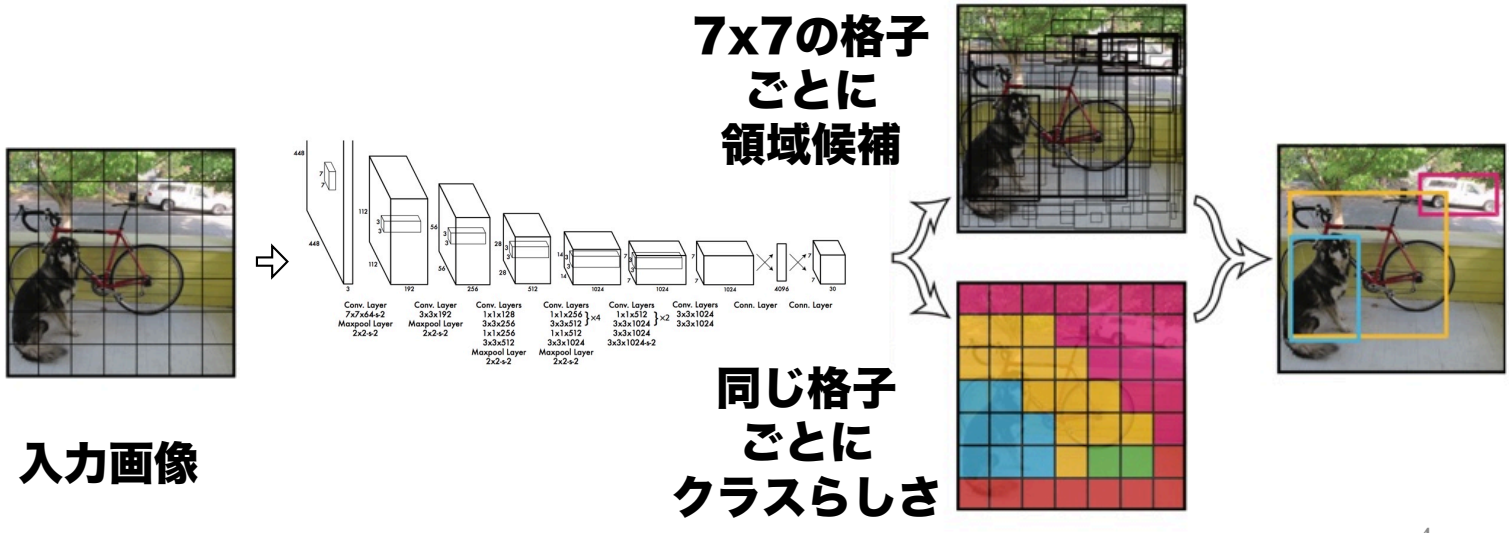
- 物体認識のクラス分類用CNNを流用
- 候補領域を何らかの方法で（大量に）作り検証



入力画像　物体領域候補　クラス分類　＋どれでもない

YOLO(You Only Look Once)/YOLOv2 [Redmon+2015,2016]

- 物体領域の(x,y)とサイズ(w,h)を直接出力(回帰)するCNNをデザイン
- 格子ごとに(x,y)&(w,h)とクラス尤度

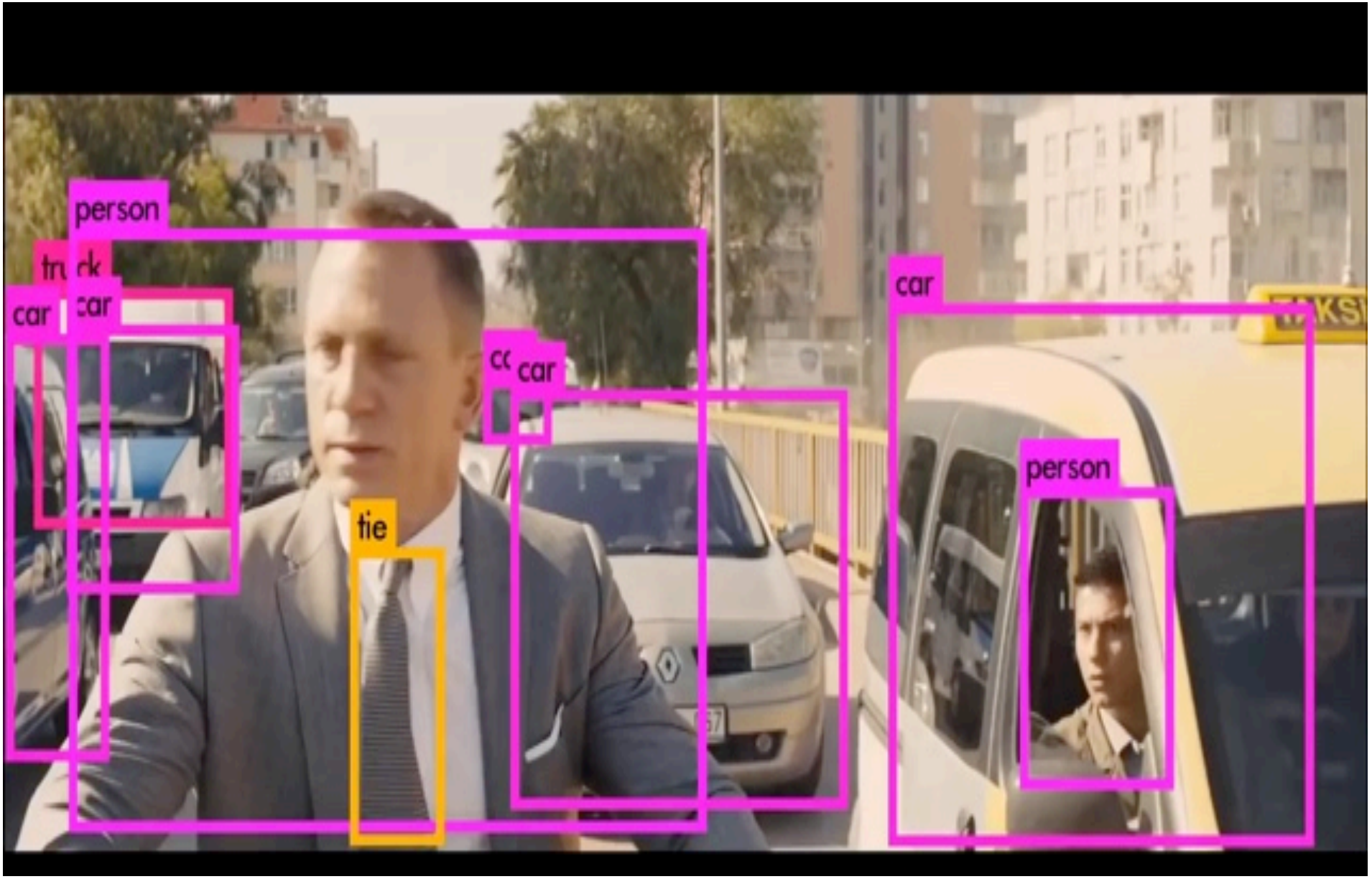


7x7の格子ごとに領域候補

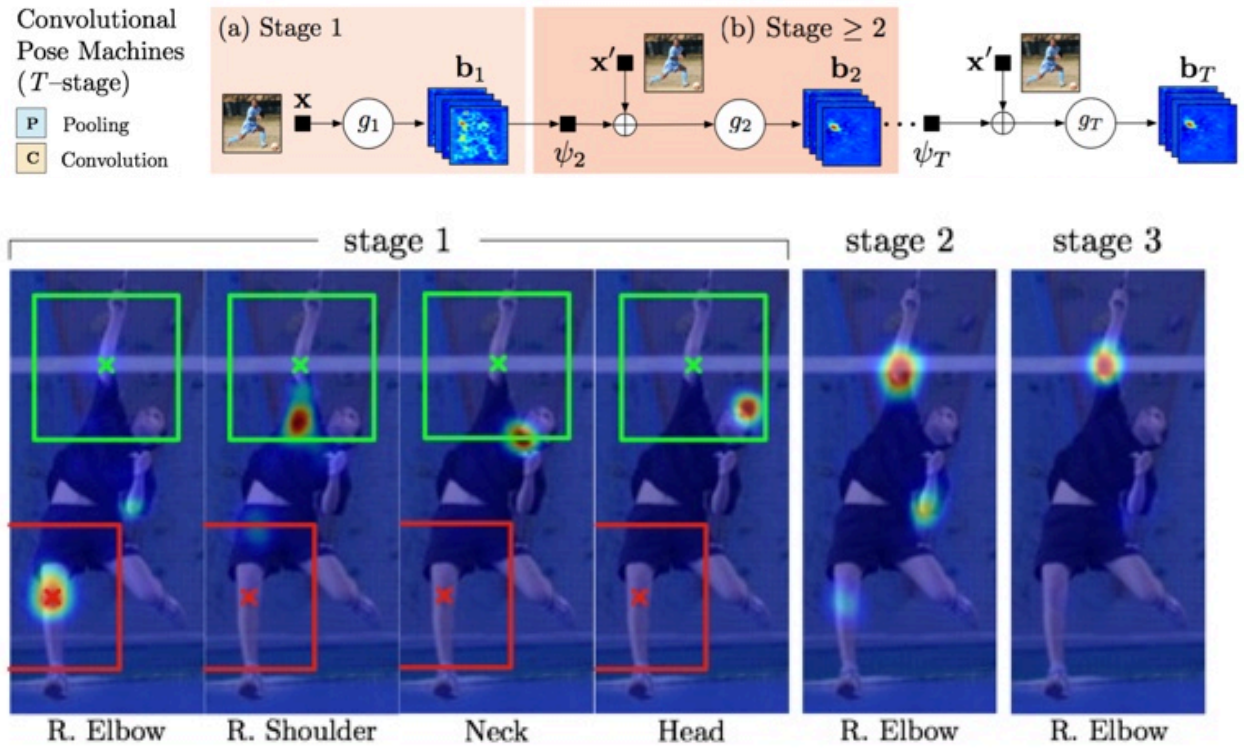
同じ格子ごとにクラスらしさ

入力画像

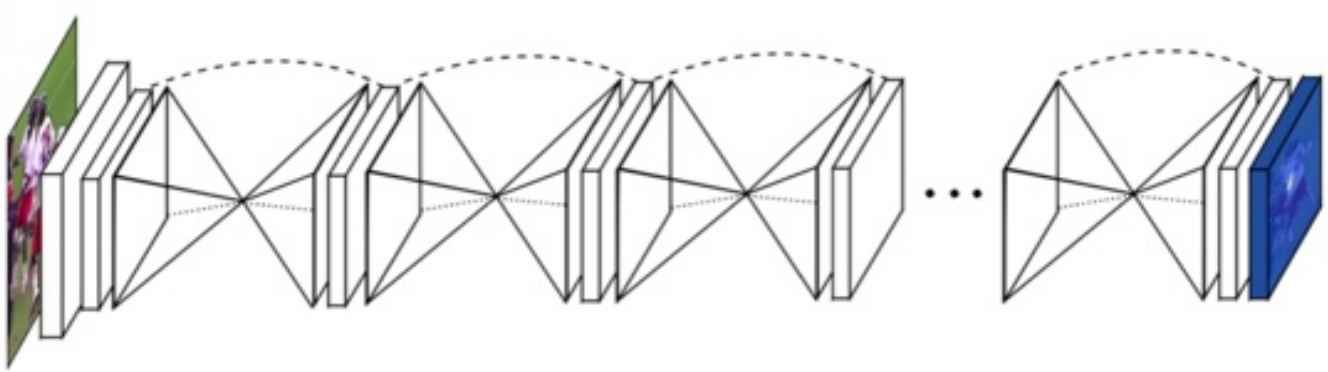
物体検出



人体姿勢推定



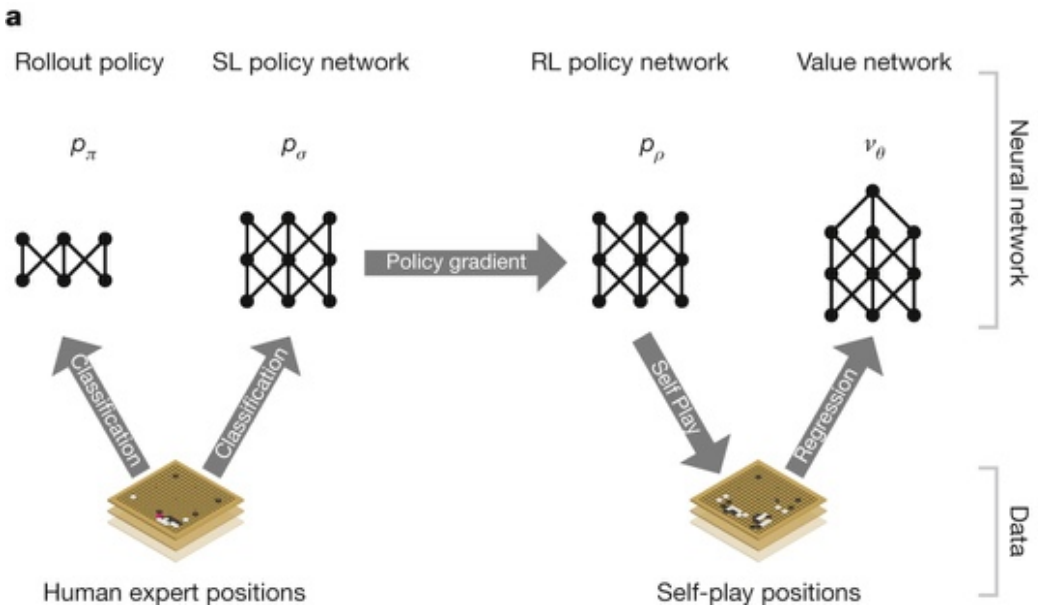
Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, Convolutional Pose Machines, 2016

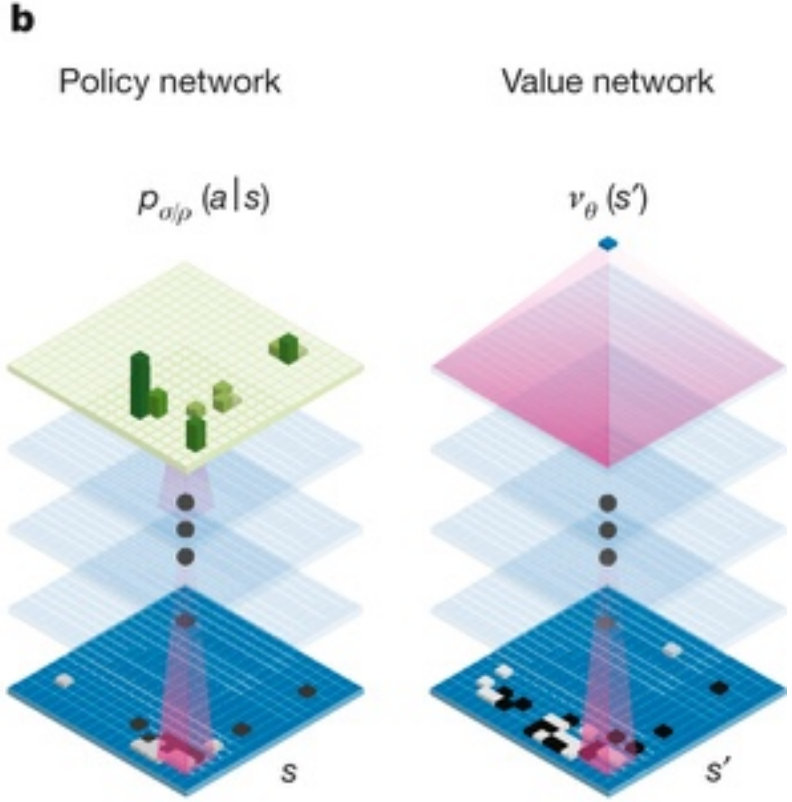


Newell+, Stacked Hourglass Networks, 2016

AlphaGo

D Silver *et al. Nature* **529,** 484–489 (2016) doi:10.1038/nature16961

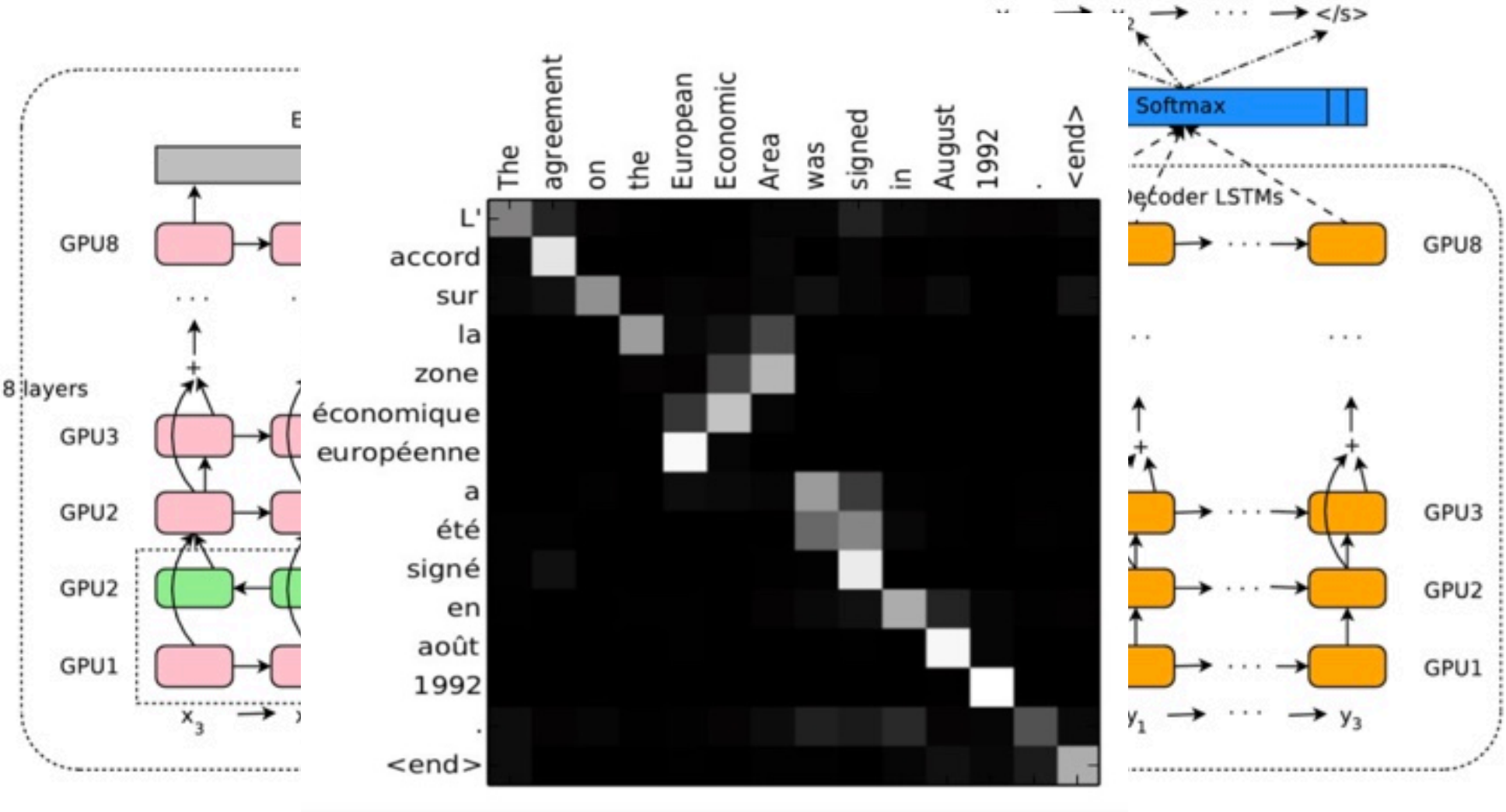




nature

機械翻訳（Google‘s NMT）

Wu+, Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation, arXiv, Oct. 2016

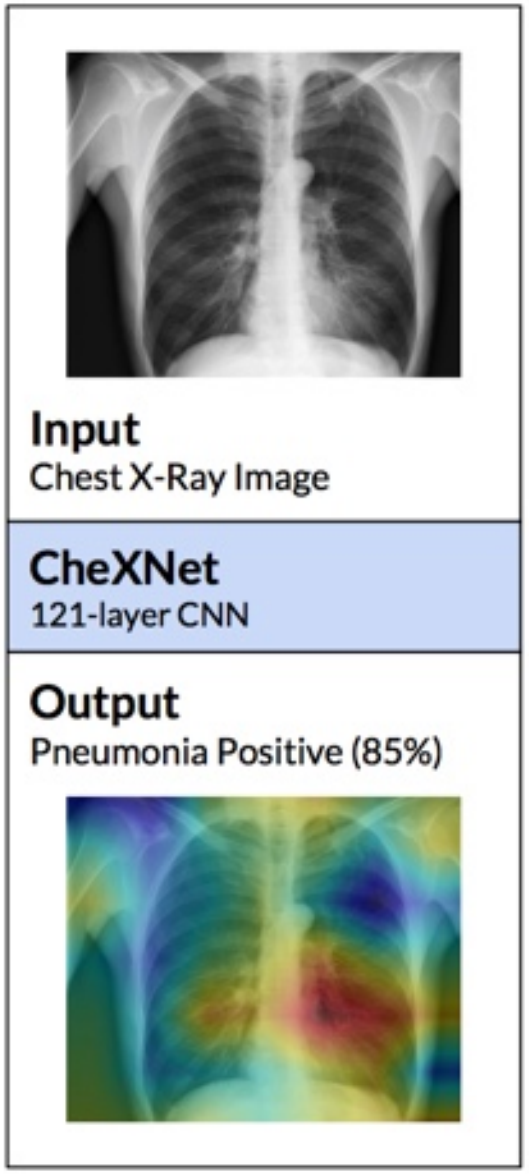


X線画像からの肺炎検査

Rajpurkar+, CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, arXiv2017

- X線画像から12種の病気の診断
- 112,120枚の画像を深層学習
- 放射線科医の平均を上回る診断精度

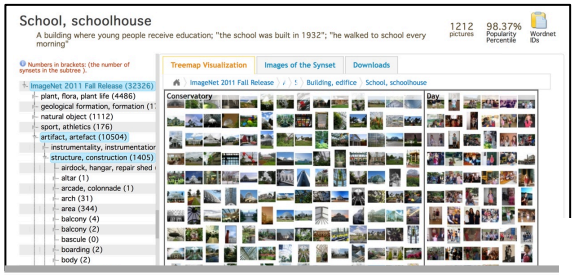
	F1 Score (95% CI)
Radiologist 1	0.383 (0.309, 0.453)
Radiologist 2	0.356 (0.282, 0.428)
Radiologist 3	0.365 (0.291, 0.435)
Radiologist 4	0.442 (0.390, 0.492)
Radiologist Avg.	0.387 (0.330, 0.442)
CheXNet	0.435 (0.387, 0.481)



タスクごとの学習データ

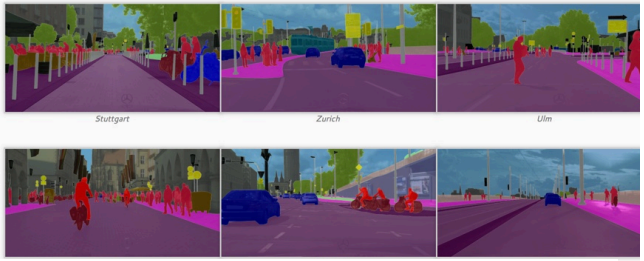
IMAGENET http://image-net.org/index

- 一般物のクラス認識
- 21841クラス・14,197,122枚
- スタンフォード大・プリンストン大他



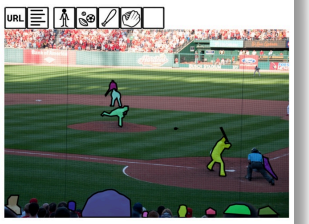
CITYSCAPES https://www.cityscapes-dataset.com/examples/

- 車載カメラの画素単位の物体クラス
- 30クラス・50都市・５,000枚/20,000枚
- ダイムラー・ダルムシュタット工科大他



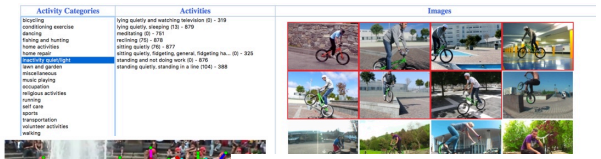
COCO http://cocodataset.org

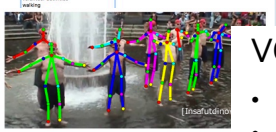
- 物体クラスとその画像領域
- 80物体+91物体以外クラス・330,000枚
- Cornel U, Microsoft他



MPII Human Pose Dataset http://human-pose.mpi-inf.mpg.de

- 人体ポーズ
- 前身関節位置・40,000人物・25,000枚
- Max Planck Institute Informatik





CelebA http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

- 顔画像
- 40属性・10,177人物・202,599枚
- 香港中文大学

ACTIVITYNET http://activity-net.org

- ビデオクリップ内人物行動
- 200クラス・20,000クリップ(648時間)
- サウジ王立科技大・ノルテ大（コロンビア）

VQA http://visualqa.org/index.html

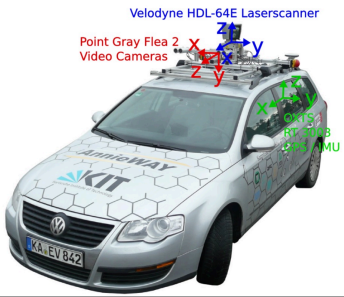
- 画像中のシーンに対する質問と答え
- 画像当たり平均5.4質問・10回答・265,016枚
- Virginia Tech., Georgia Tech.

Q: Where is the cat?
Q: Is the cat safe?
Q: What color is the car?

Q: What toy does the elephant have in his cage?
Q: Is this animal in an enclosure?
Q: Where is the woman?

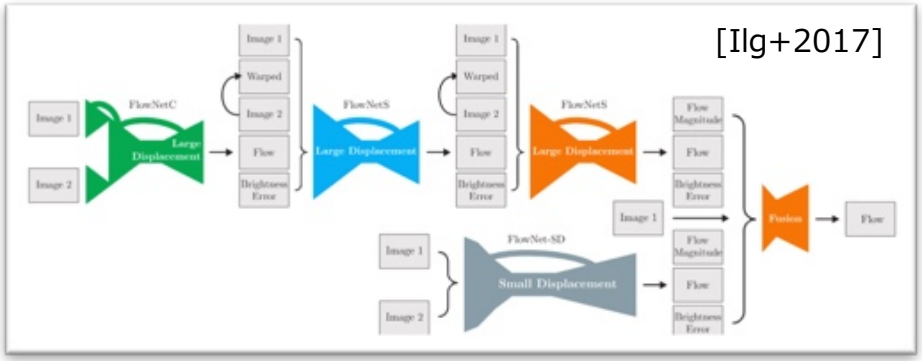
KITTI Vision http://www.cvlibs.net/datasets/kitti/index.php

- 車載カメラ映像からのステレオ視・オプティカルフロー他
- >100GB
- Karlsruhe Institute of Tech., TTI Chicago



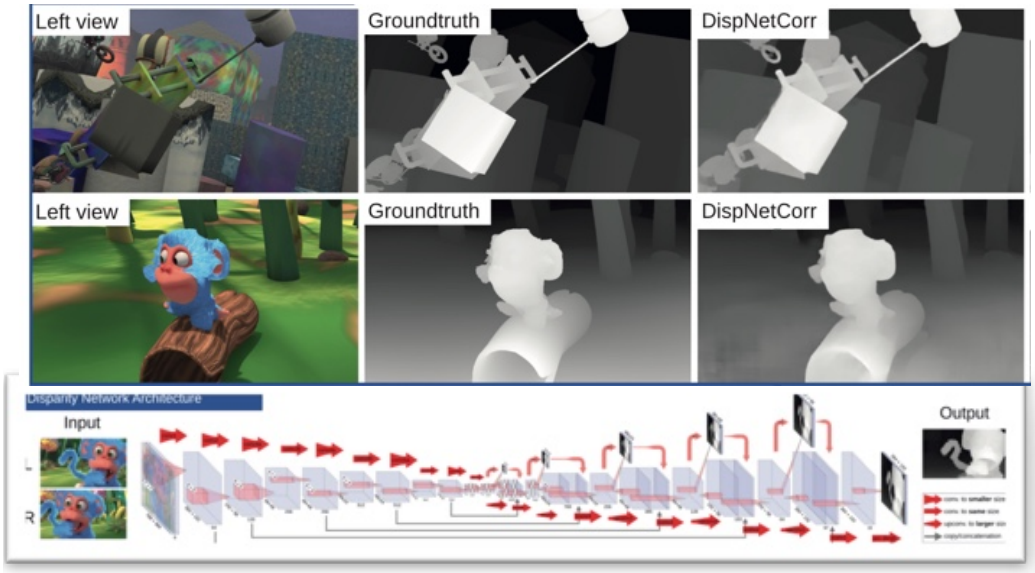
モーション・多視点幾何

オプティカルフロー



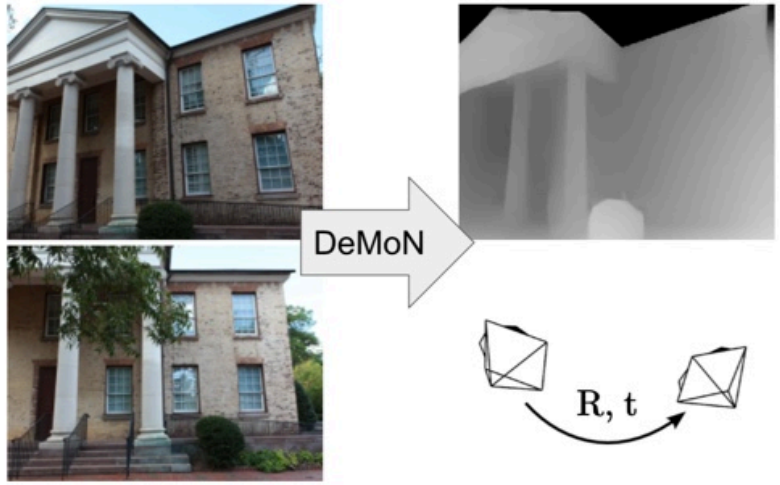
視差（ステレオ画像）

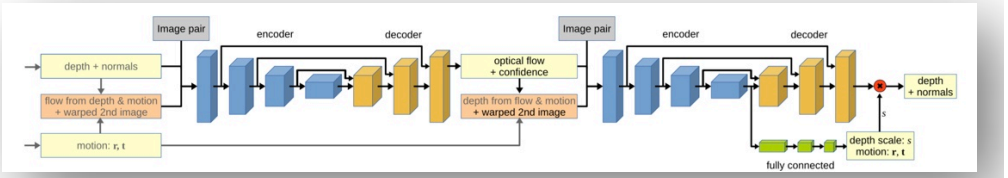
[Mayer+2017]



カメラ姿勢

[Ummenhofer+2017]

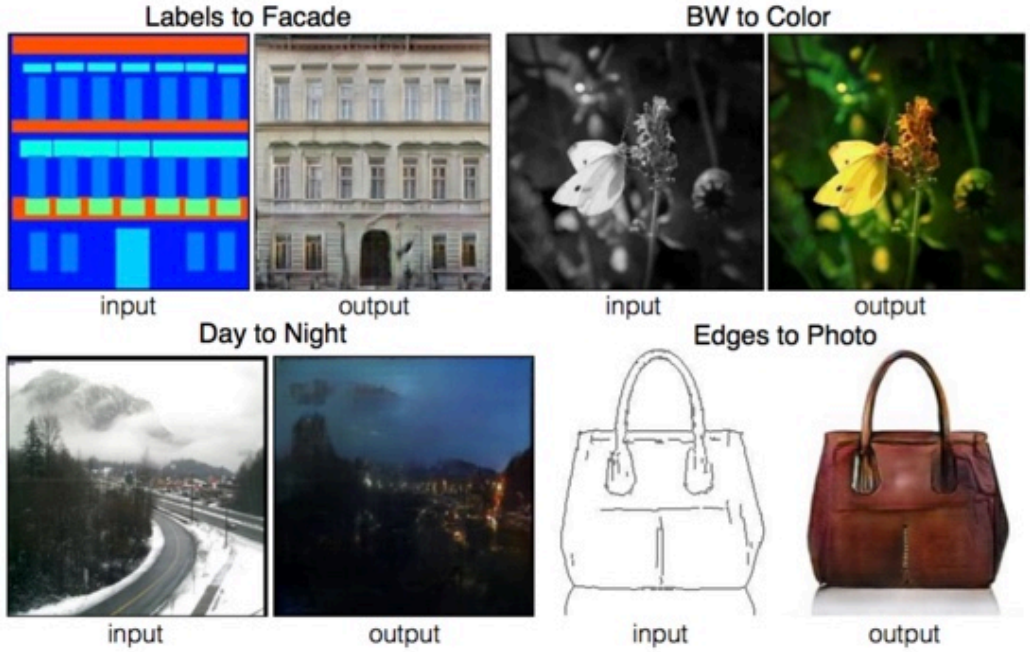




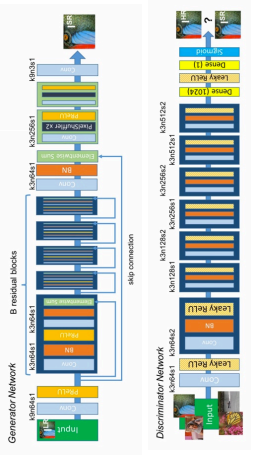
画像変換・合成

スタイル変換

色付け・昼夜変換・線画→実写

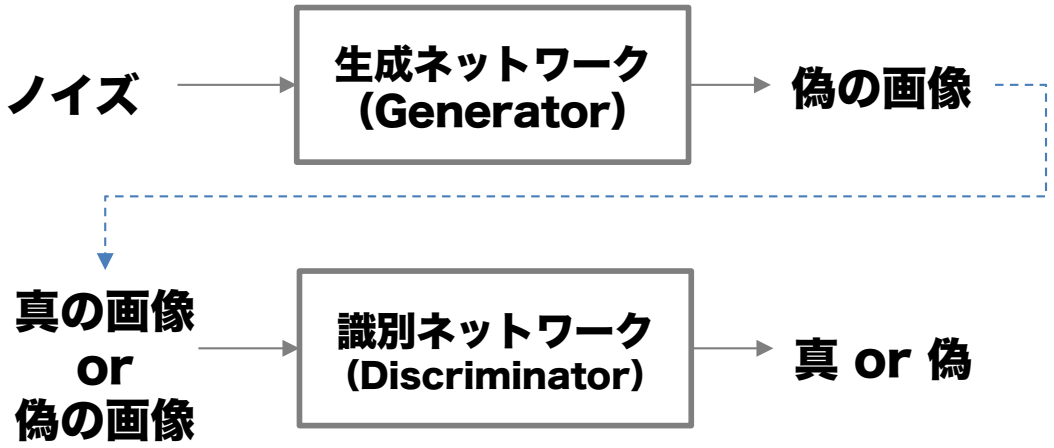


超解像



Generative Adversarial Network (GAN)

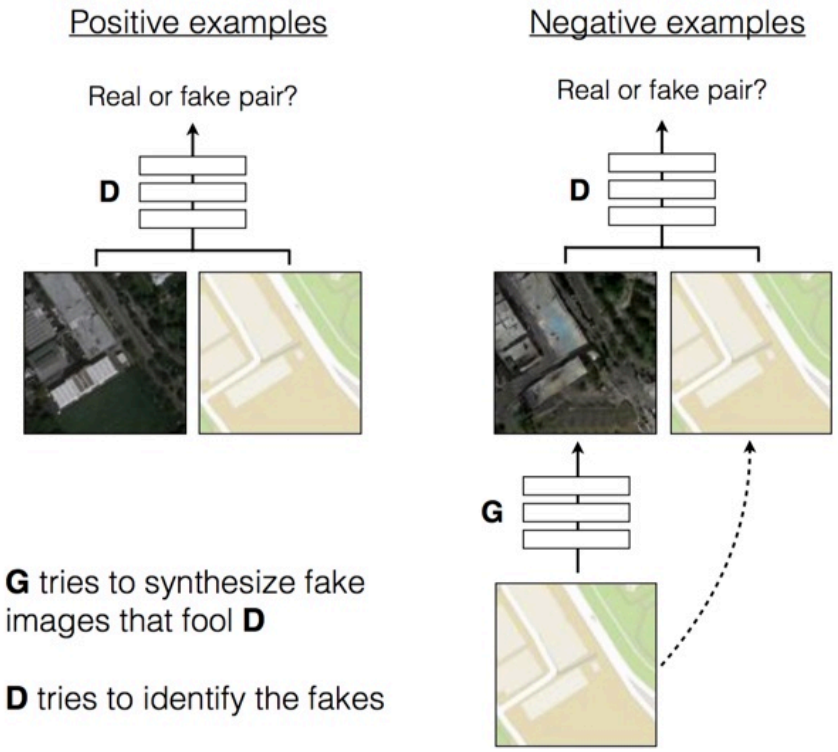
- 二つのネットワークを競争的(adversarial）に学習
  - Gはなるべく本物らしい画像を生成してDを騙すように，　Dは本物と偽物を正確に見極めてGに騙されないように



$$\min_G \max_D V(D,G)$$

$$V(D,G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z))]$$

GAN (unconditional)



**“pix2pix”
(Conditional GAN)**

Generative Adversarial Network (GAN)

- 二つのネットワークを競争的(adversarial）に学習
  - Gはなるべく本物らしい画像を生成してDを騙すように，　Dは本物と偽物を正確に見極めてGに騙されないように

Wasserstain GAN [Arjovsky+2017]

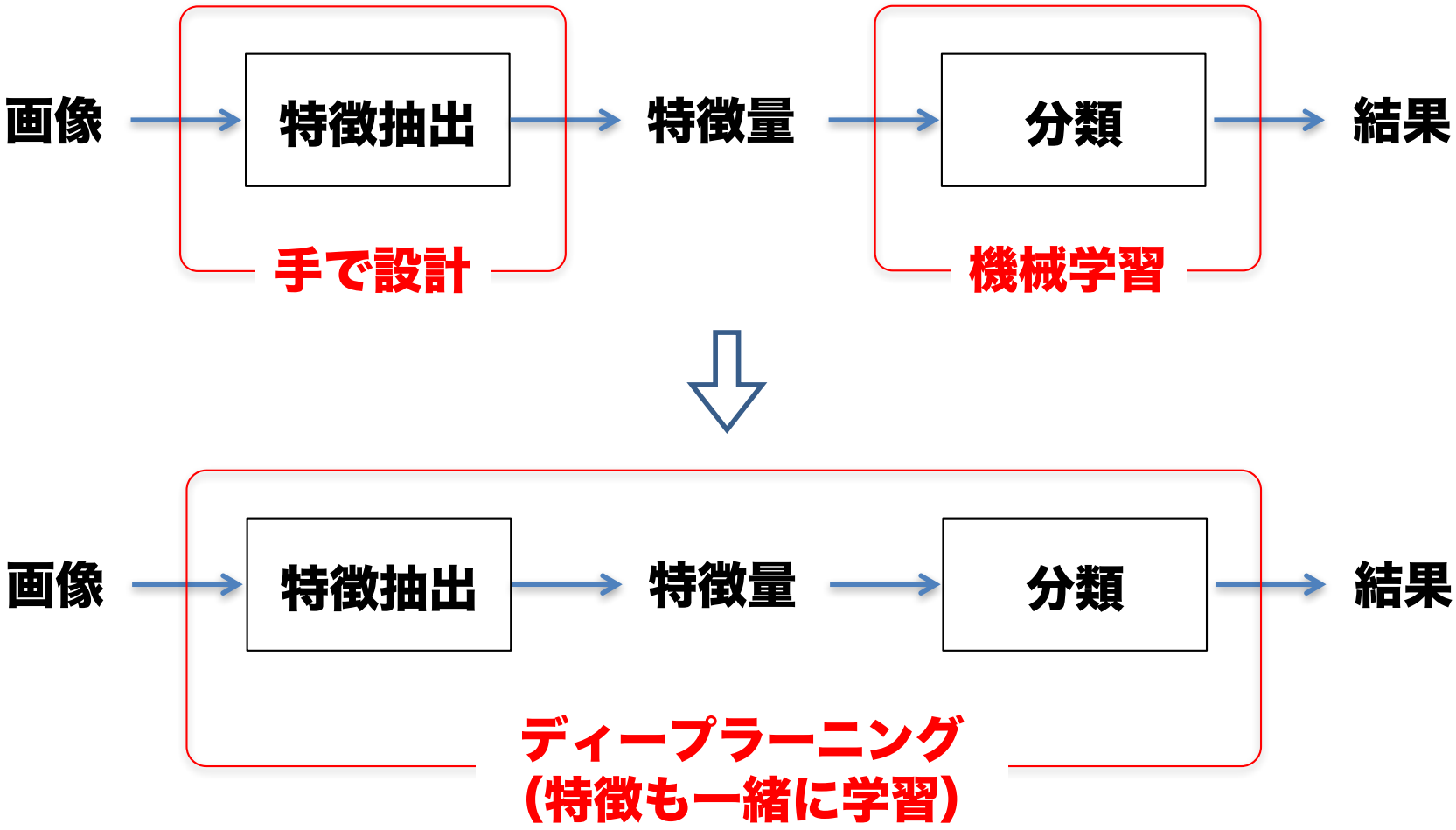
顔の動きや表情の転送

Kim+, Deep Video Portraits, SIGGRAPH2018

深層学習を手元の問題に適用するには？

- 類似のタスク（問題）を探す→なければ２へ
  - 物体カテゴリ認識・シーン認識，物体検出，セマンティック(インスタンス)セグメンテーション，…
- 「解きたい問題」をDLで扱える形に落とし込む
  - 入力は何か？
    - 画像(2D)，動画像(3D)，距離画像，点群，系列データ，グラフ構造etc.
  - 出力は？
  - 入力→出力のマッピングがDNNで表現可能か？それは学習可能か？
- モデルを考える
  - ネット構造：CNN, RNN, ハイブリッド，アテンション，Graph-NN(GNN/GCN)
  - ロス関数，前処理
- データを集める
  - たくさん集まらない場合は転移学習を考える
  - 場合によってはCGなどで合成・ドメイン適応なども
- 学習方法を考える
  - 教師あり（分類・回帰・距離軽量学習），半教師(semi-supervised)，競争的(adversarial)など

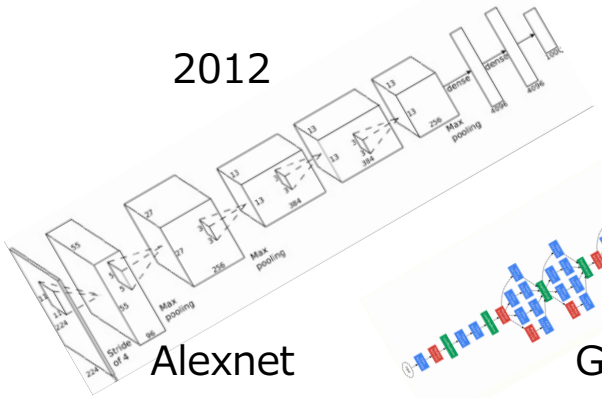
特徴の設計から特徴の学習へ



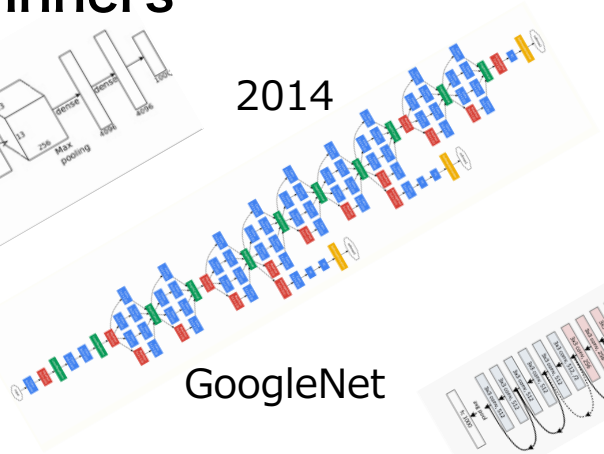
ネットワーク構造のデザイン

- **ILSVRC Winners**

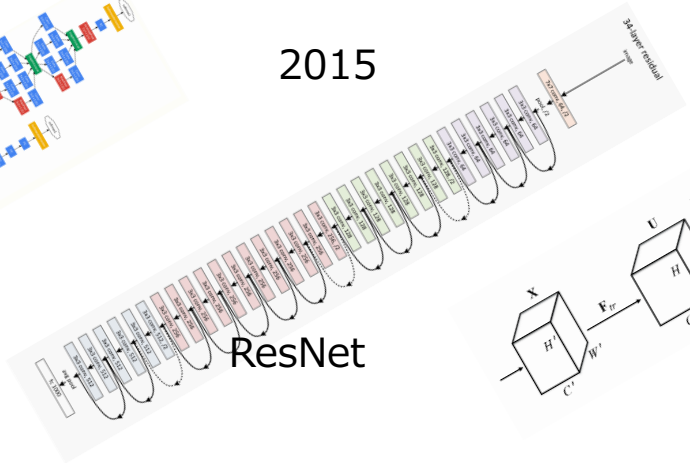
2012 Alexnet



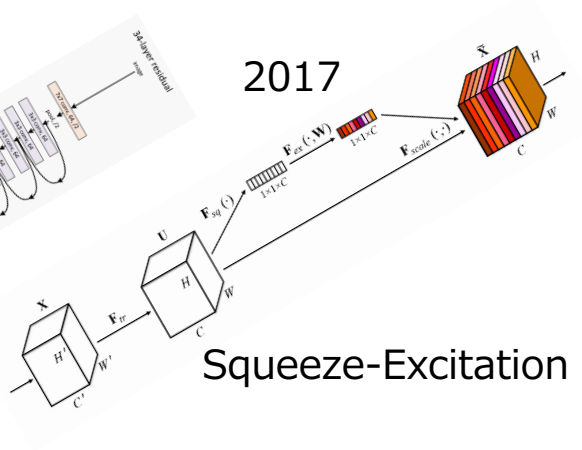
2014 GoogleNet



2015 ResNet

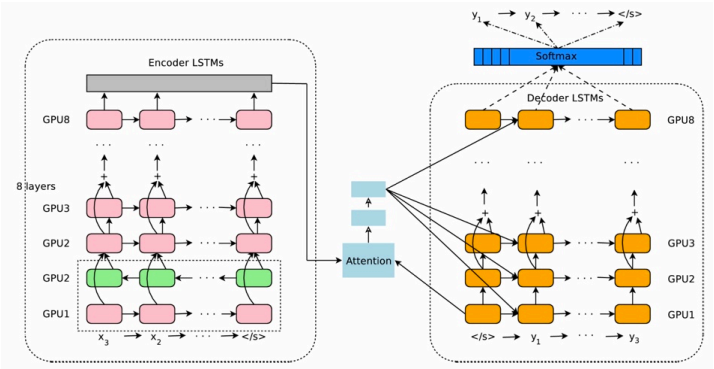


2017 Squeeze-Excitation



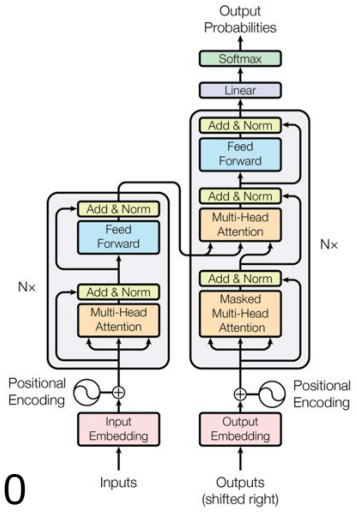
- **ニューラル機械翻訳の変遷**

Google's NMT
[Wu+16.09]



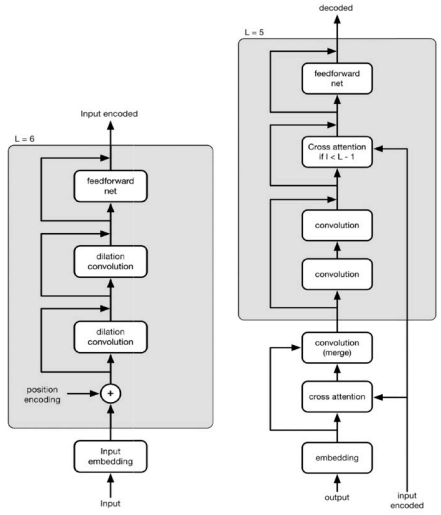
EN-FR BLUE: 39.92

Attention is all you need
[VasWani+17.06]



41.0

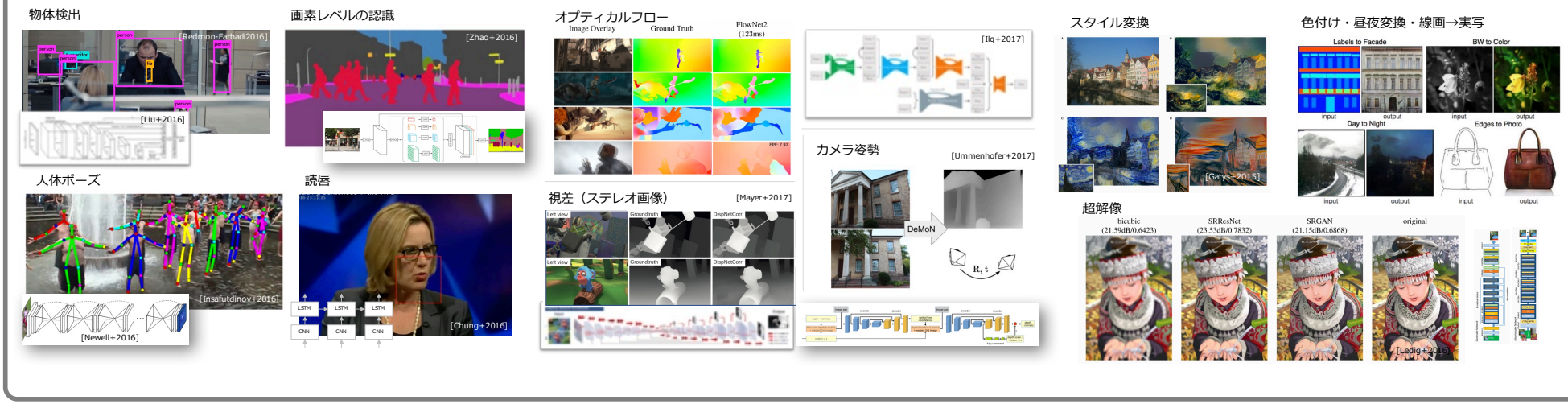
CNN is all you need
[Chen-Wu,17.12]



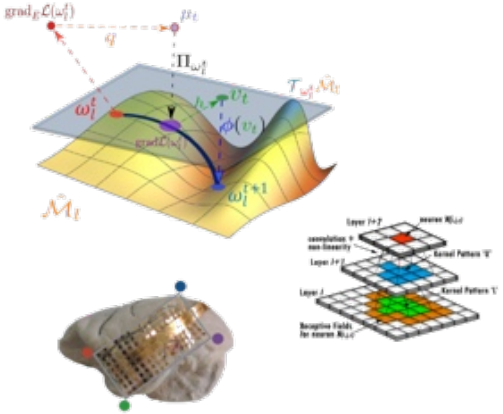
45.54

フロンティア（＝今はできないこと）

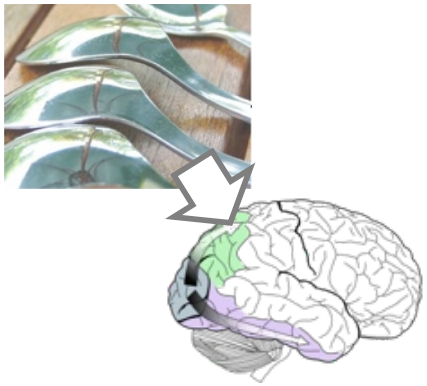
画像のあらゆる問題をCNNで解く時代



深層学習の理解

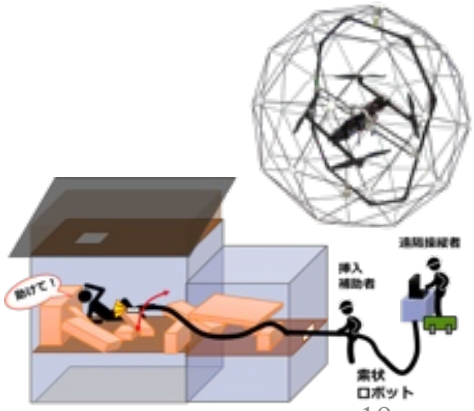


教師データがない？



画像の理解

ロボット



教師データが少ない場合の対策

- ディープラーニングはdata-hungry
  - 汎化能力は「期待するほど」高くない
- 対策
  - データ拡張(data augmentation)
  - 転移学習(transfer learning/domain transfer)
  - ドメイン適応(domain transfer)
    - 目的と同一タスクの学習データがあるが，入力データが少し違う
  - データの合成
    - 目的タスクに近い入力データを合成；ラベル付けはタダ
  - 自己教師(self-supervised/self-taught)
    - 目的と(全然)違うタスクながら，ラベル付けがタダ
  - 半教師学習(semi-supervised)
    - 少量のラベル付きデータ＋大量のラベルなしデータ
  - その他
    - 弱教師(weakly-supervised)，無教師(unsupervised)，マルチタスク学習(multi-task learning)

CGによる学習データ生成

Wood+, Learning an appearance-based gaze estimator from one million synthesized images, 2016

- 視線推定のための画像データを合成

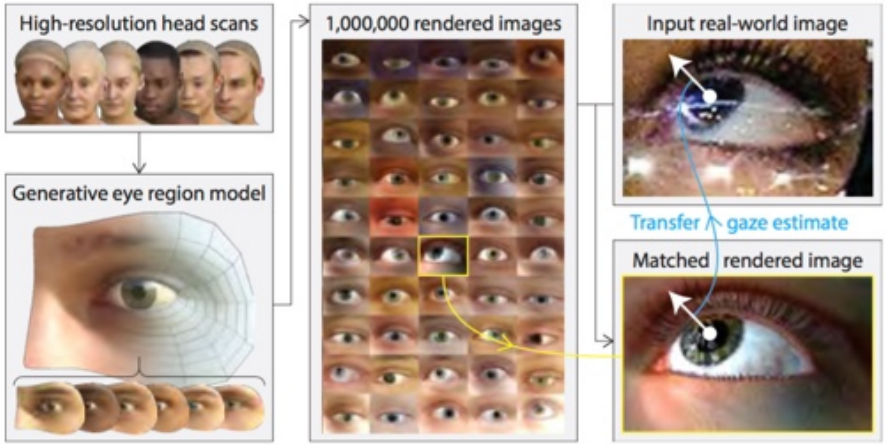
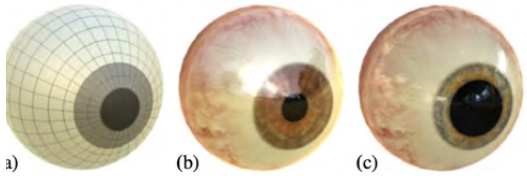


Figure 1: *We rendered one million realistic images of eyes using our generative eye region model. These are matched to an input image using a nearest-neighbor approach for gaze estimation. Our model manages to find good matches even with extreme gaze angles and glare from glasses.*

and glare from glasses.



igure 2: Our eyeball mesh (a) shown rendered with physically-ased materials and refraction effects. We model pupillary con-raction (b) and dilation (c) as part of the refraction shader.

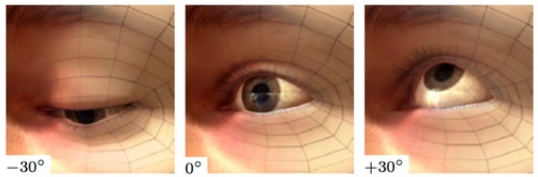


Figure 7: *We use anatomically inspired procedural geometric methods to animate eyelid, avoiding the need to manually rig the model. Shown are renderings for eyeball pitch at 0° and ±30°.*

Figure 13: *Nearest-neighbour pairs showing in-the-wild images (top) and our renders (bottom) along with estimated gaze (green). The top three rows show qualitatively good gaze estimates, even under difficult lighting, low resolution, and extreme gaze angles. The bottom row shows failure cases from unmodelled variation e.g. makeup and hair.*

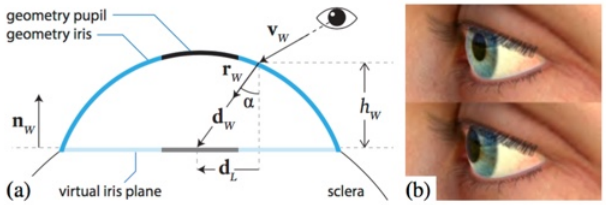
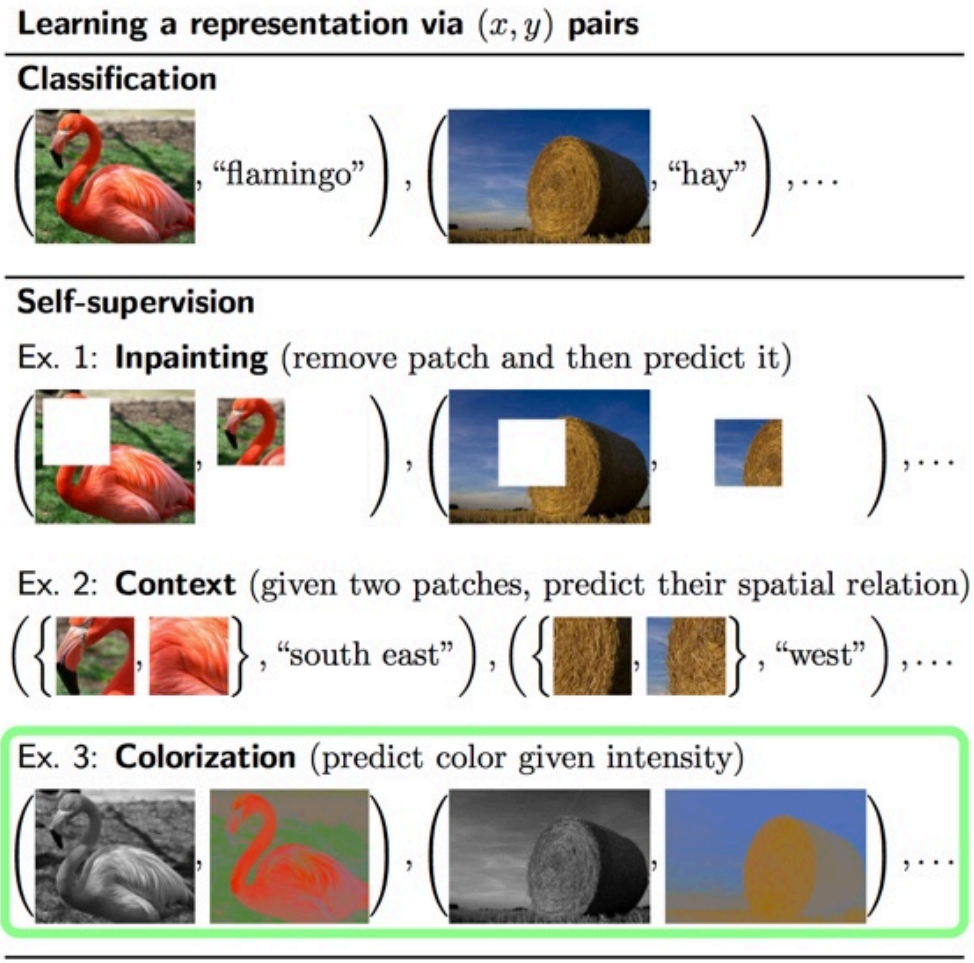


Figure 3: *We model iris refraction by altering texture look-ups. In (a), a viewed pixel is refracted correctly to show black (pupil) instead of blue (geometry surface). Example renders with (top) and without (bottom) refraction are shown in (b).*

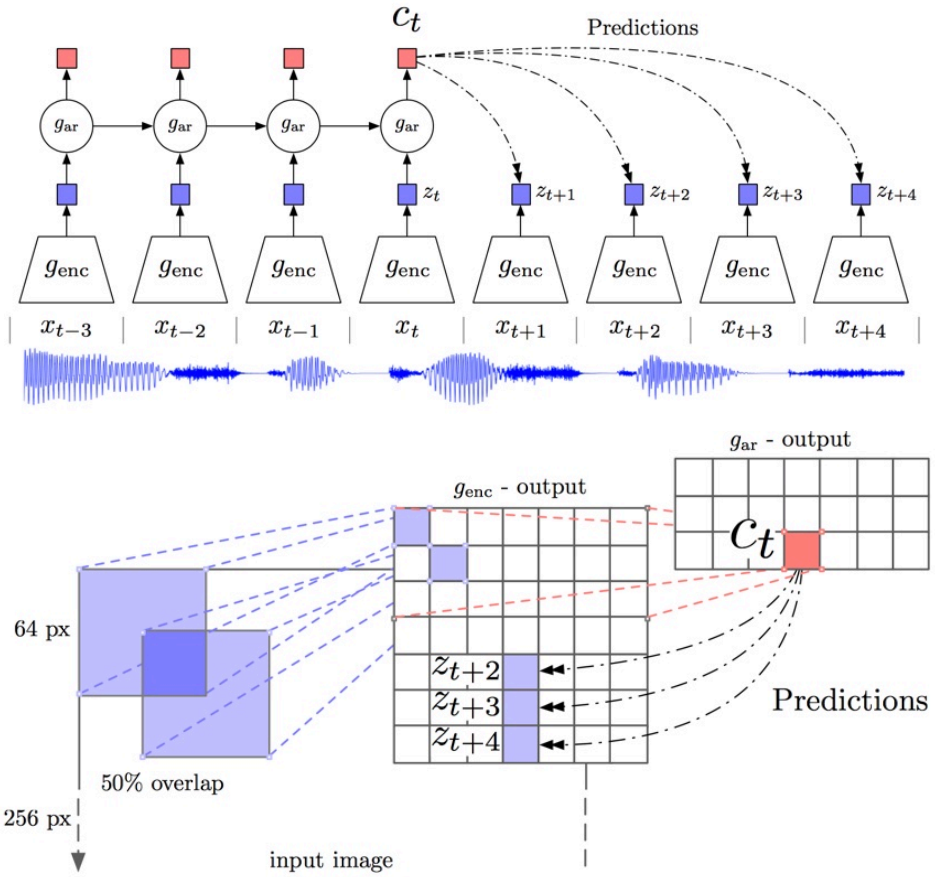
Figure 10: *We use HDR panoramic images for reflections and ambient light in the scene. Here you can see two example equi-rectangular panoramas, with example eye renderings.*

自己教師学習 (self-supervised learning or self-taught …)

- 目的タスクと違う”Proxy task”で”pre-train”



[Larsson-Maire-Shakhnarovich CVPR17]



未来や近接画素の特徴を予測するが，多次元特徴そのものではなく，予測対象とその元になるものの間の相互情報量をなるべく保つ潜在表現を学習

Contrastive Predictive Coding
[van den Oord-Li-Vinyals arXiv18]

推論の可視化

中間層出力の可視化 [Zou+CVPR16]

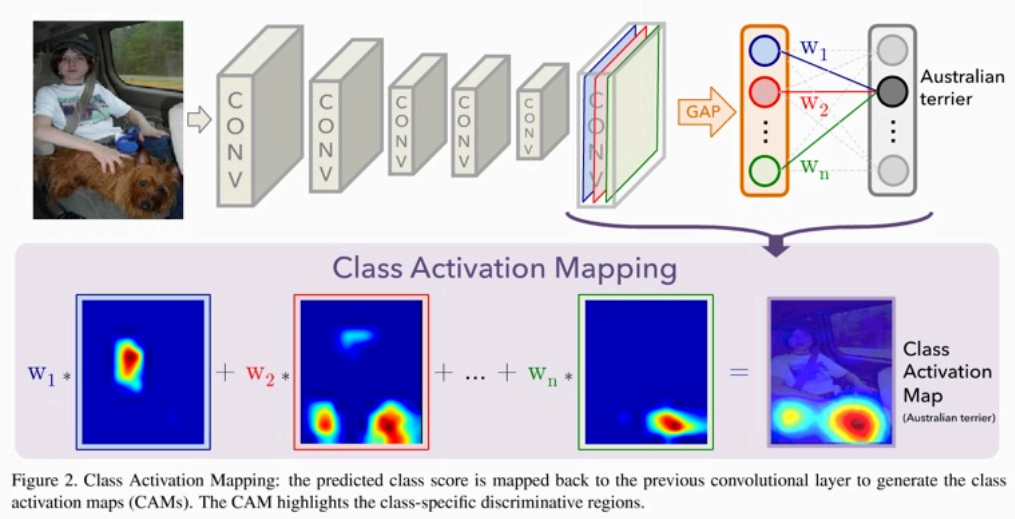
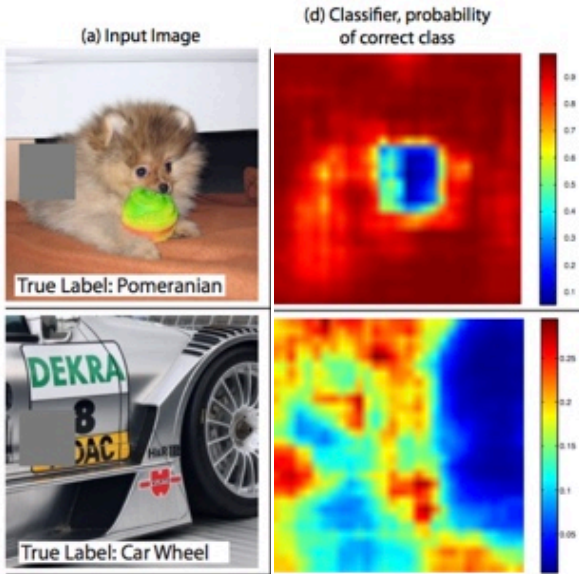


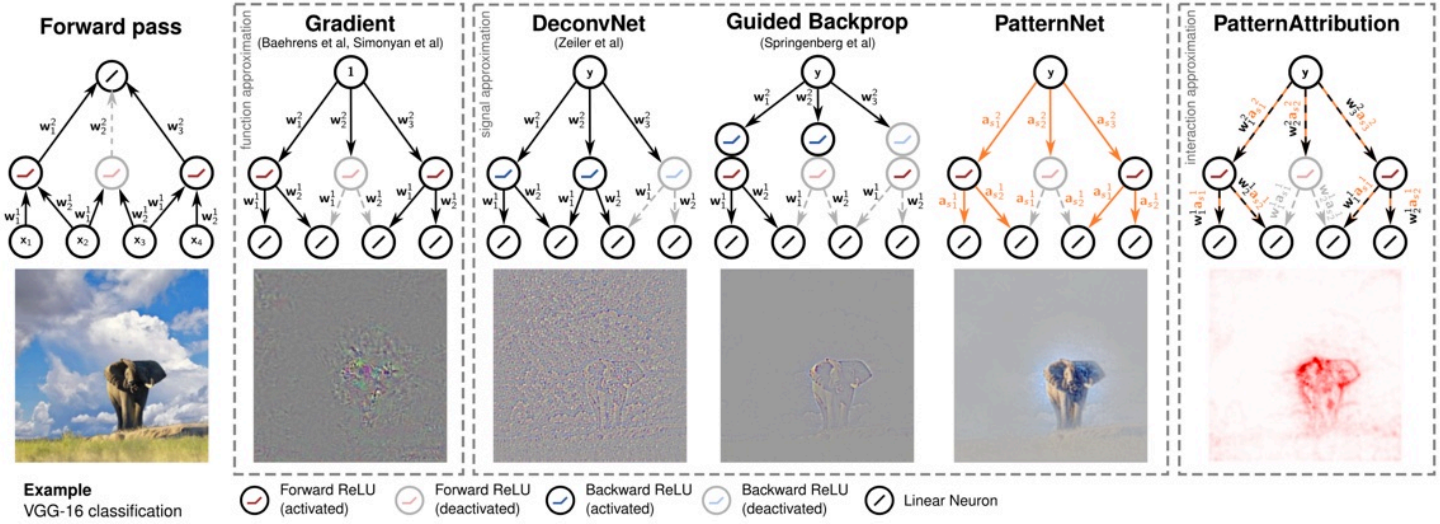
Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

マスキング（隠して影響を見る）



[Zeiler-Fergus14]

寄与度の逆算（出力側から逆伝播）



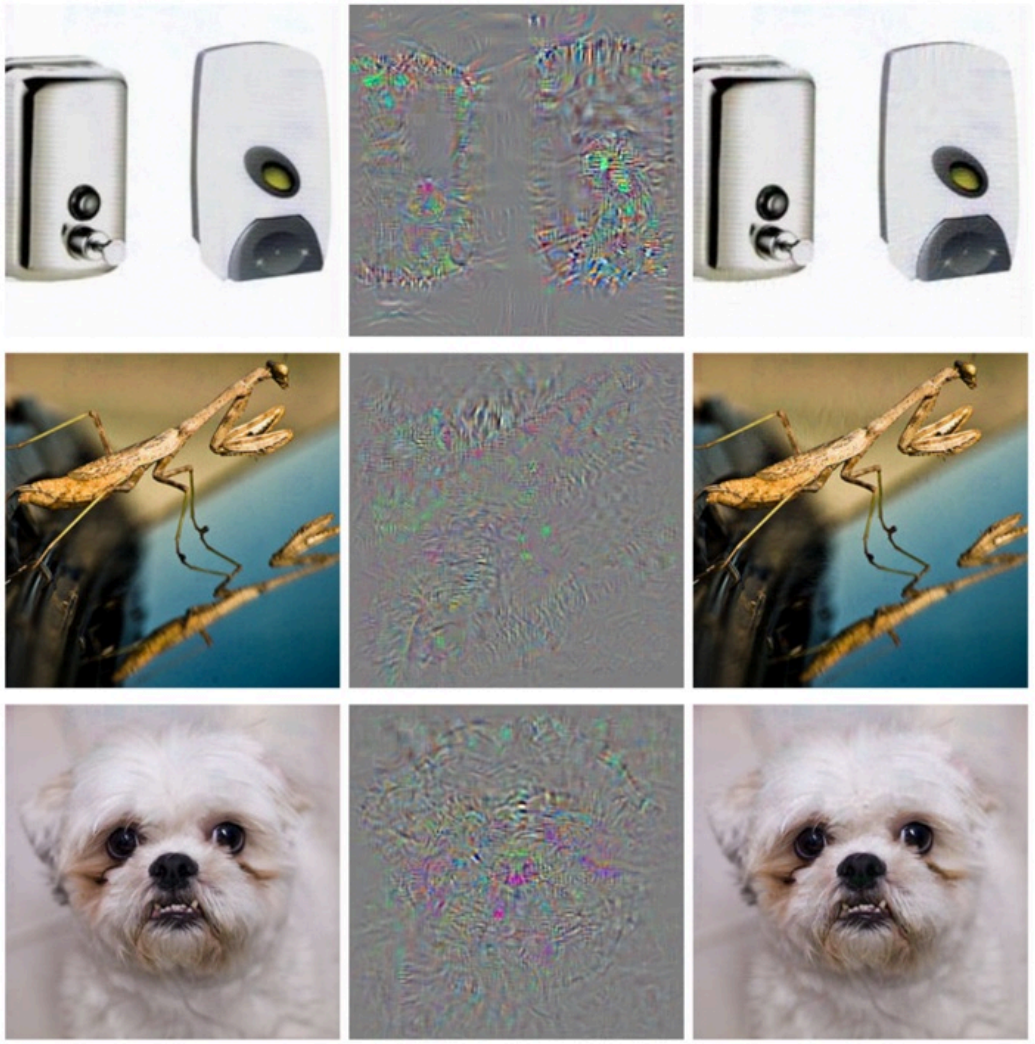
[Kindermans+2018]

説明のための構成（アテンション等）

CNNを騙す

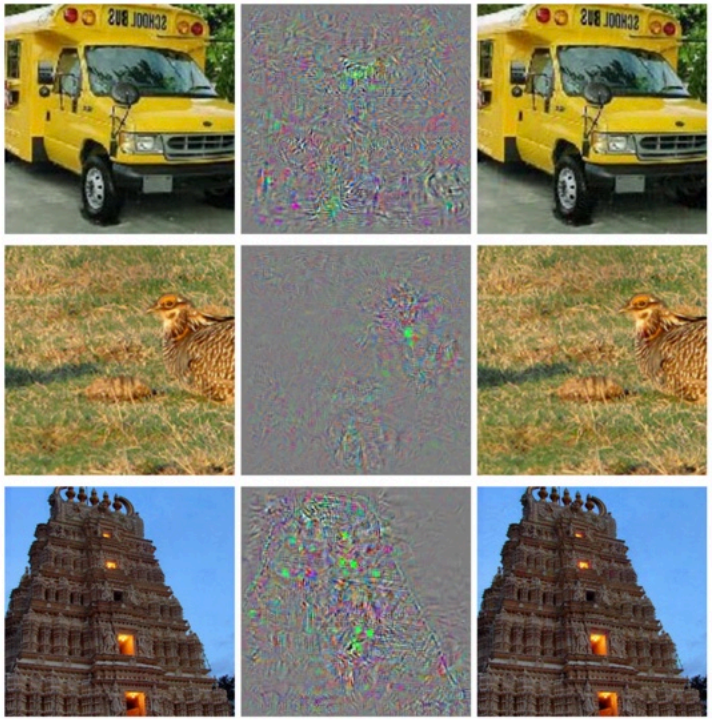
Szegedy+, Intriguing properties of neural networks, 2014

- 人には同じに見えるが，CNNには全然違って見える



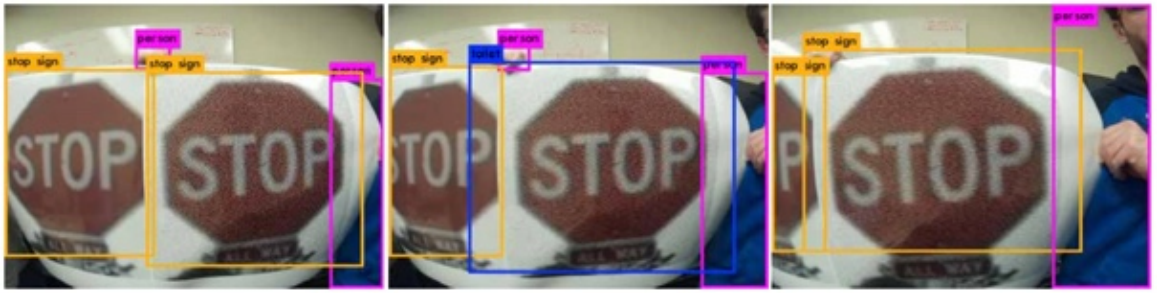
Correctly recognized | **Additive noise** | **Recognized as“Ostrich”**

- Minimize $\|r\|_2$ subject to:
  1. $f(x+r) = l$
  2. $x + r \in [0,1]^m$



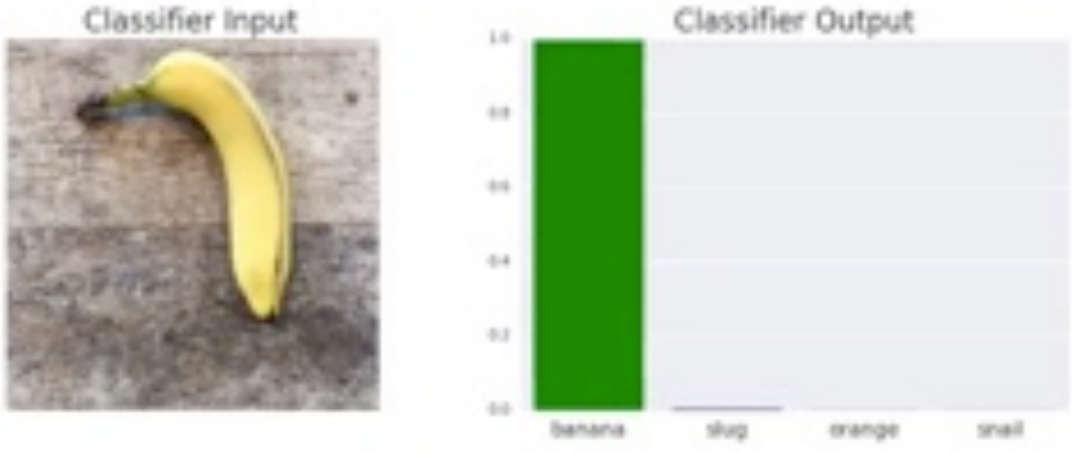
実世界での攻撃

Adversarial exampleは撮影条件に敏感だから心配しすぎでは？
[Lu+2017]



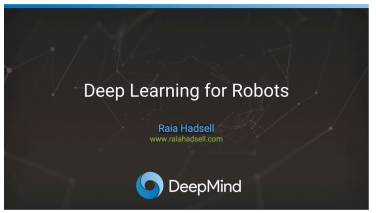
ロバストなadversarial example
[Athalye+ ICLR2018]
- 撮影条件によらず間違えるように作る

Adversarial patches
[Brown+arXiv Dec. 2017]
- 対象とするCNNに，シーンの他の物体を無視させ，指定の物体で置き換え



深層強化学習のロボット応用

Raia Hadsell@DeepMind，ERF2017でのキーノート講演



Could deep RL allow robots to learn end-to-end?

- Sensorimotor control ✓
- Exploration of complex spaces ✓
- Strategy and decision making ✓

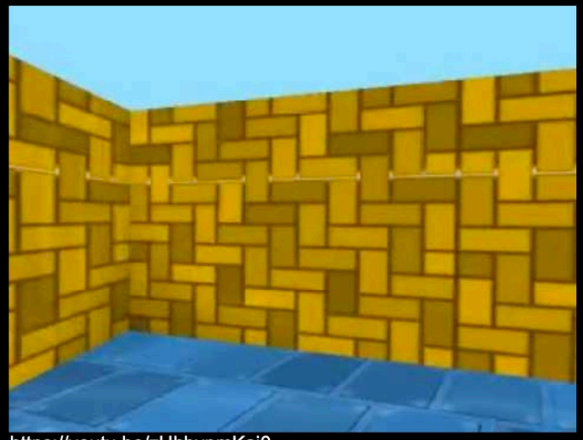
Google DeepMind General Artificial Intelligence

General Atari Player

https://www.youtube.com/watch?v=Erkt7HelEco

[Mnih et al, *Playing Atari with Deep Reinforcement Learning*, 2014]

Maze navigation



https://youtu.be/zHhbypmKaj0

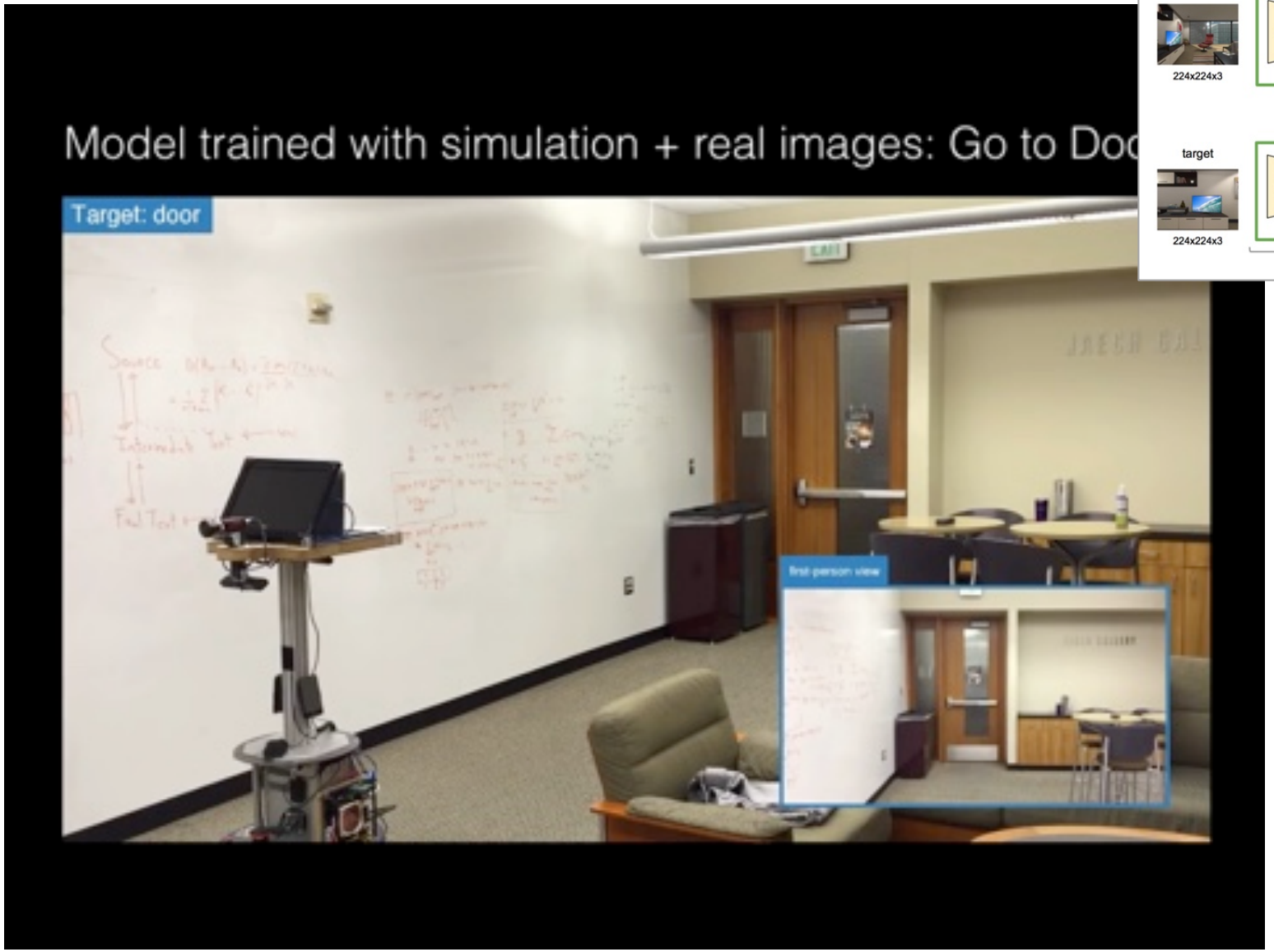
Lesson: use supervised learning when possible

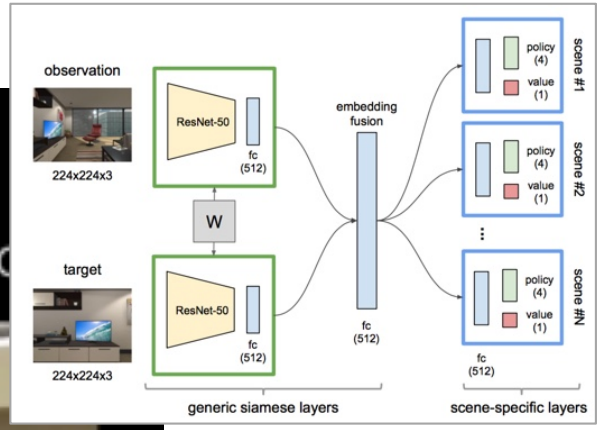
AlphaGo

Google DeepMind General Artificial Intelligence

強化学習による室内ナビゲーション

Zhu+, Target-driven Visual Navigation in Indoor Scenes using Deep Reinforcement Learning, ICRA17





ドローン自律飛行

Giusti+, A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots, 2016

パタン認識として解く視覚運動制御：CNNで自動運転

Bojarski, End to End Learning for Self-Driving Cars, arXiv, 2016

- 画像からステアリング操作出力をCNNが直接計算
  - パスは１次元的

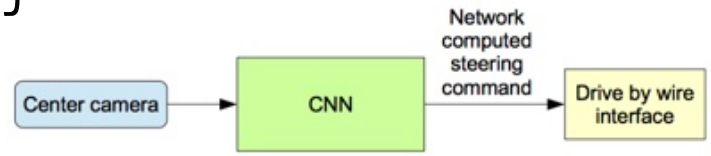


Figure 3: The trained network is used to generate steering commands from a single front-facing center camera.

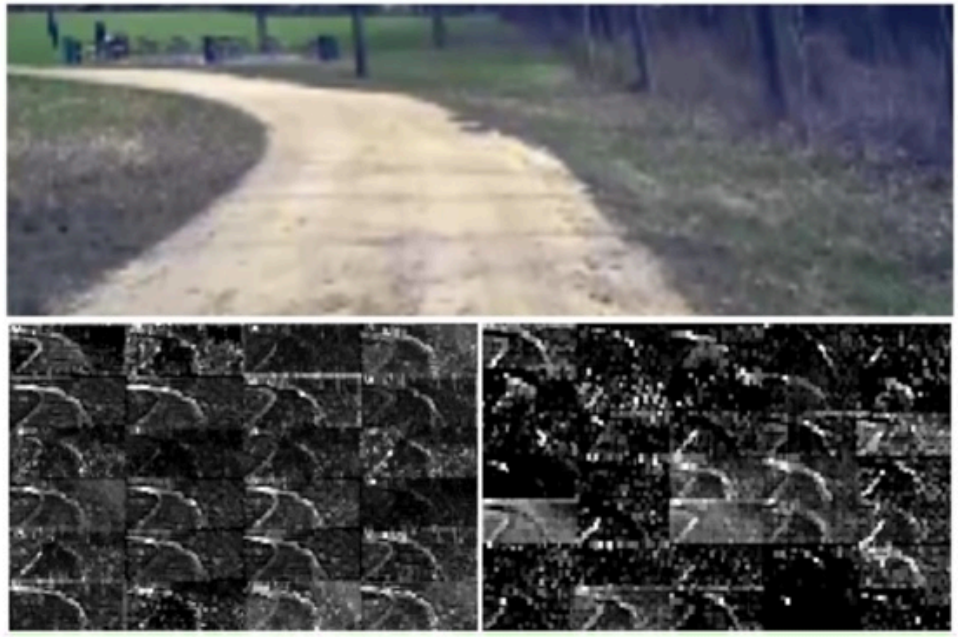


Figure 7: How the CNN "sees" an unpaved road. Top: subset of the camera image sent to the CNN. Bottom left: Activation of the first layer feature maps. Bottom right: Activation of the second layer feature maps. This demonstrates that the CNN learned to detect useful road features on its own, i. e., with only the human steering angle as training signal. We never explicitly trained it to detect the outlines of roads.

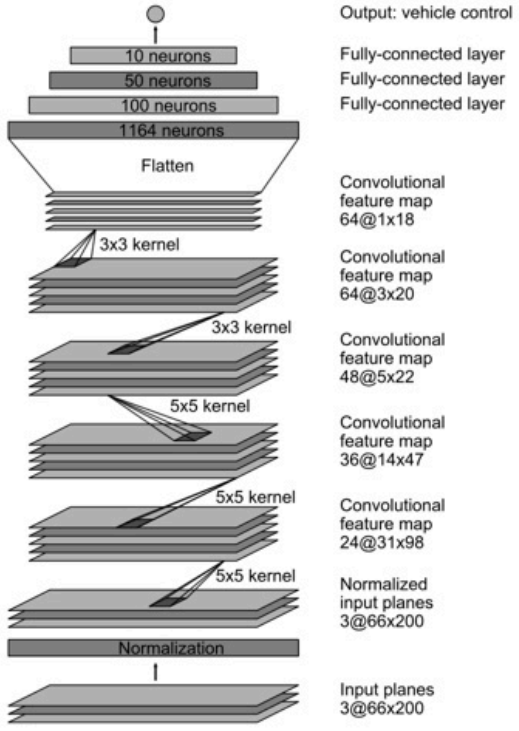
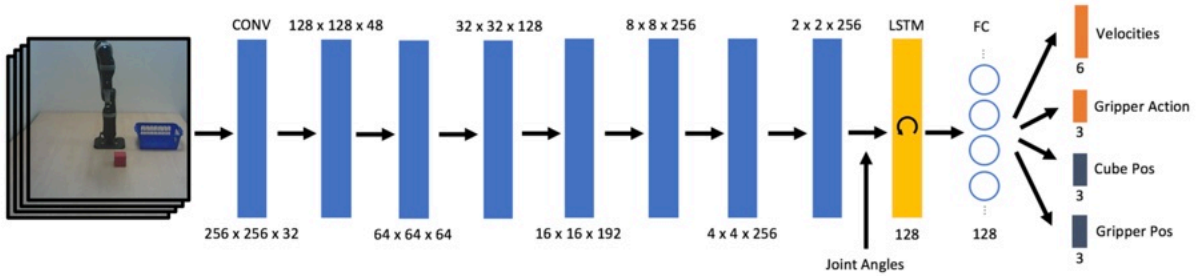


Figure 4: CNN architecture. The network has about 27 million connections and 250 thousand parameters.

パタン認識として解く視覚運動制御：マニピュレータのpick&place

James, Davison, Johns, Transferring End-to-End Visuomotor Control from Simulation to Real World for a Multi-Stage Task, arXiv17

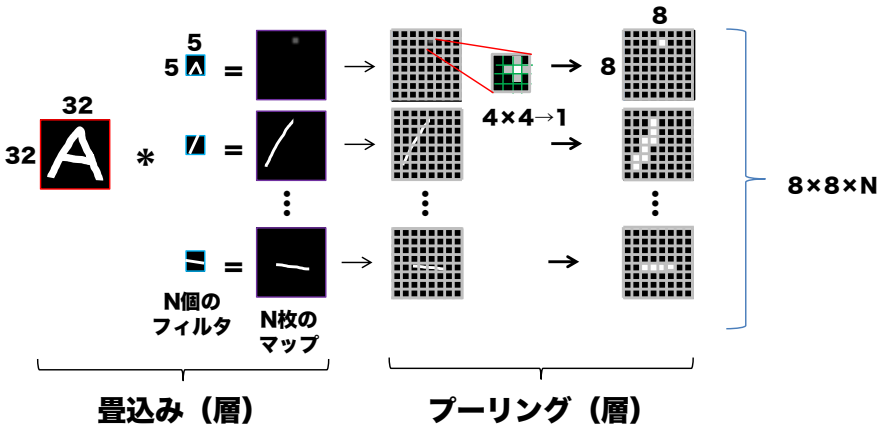
- CNN+LSTM：最新の４画像を入力，axillary出力を同時学習



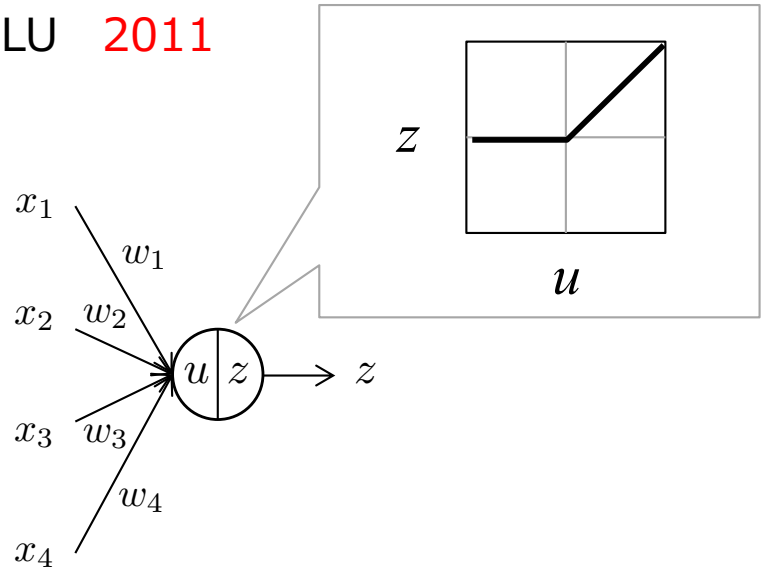
CG画像で学習(ランダム照明・背景)

深層学習の中核技術

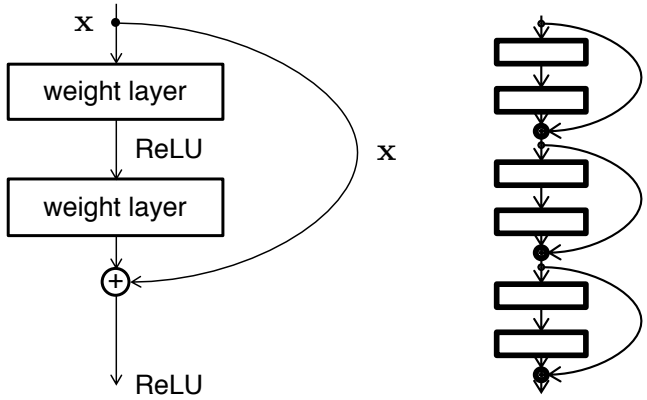
Convlution+Downsampling



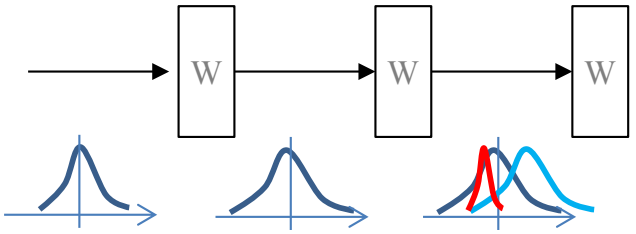
ReLU 2011



Skipconnection (ResNet) 2015

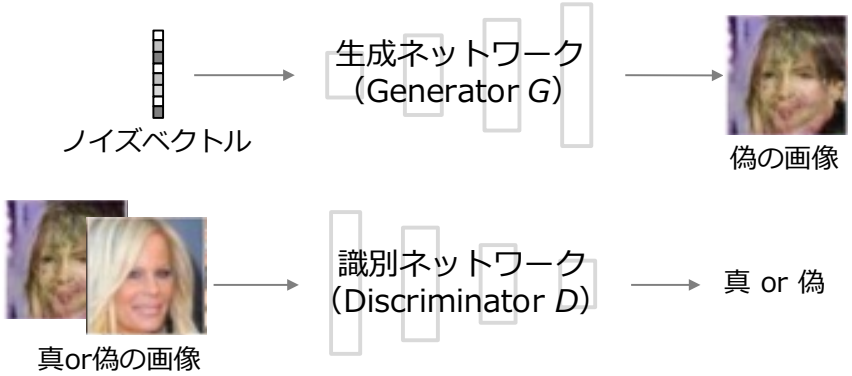


Batch Normalization 2015

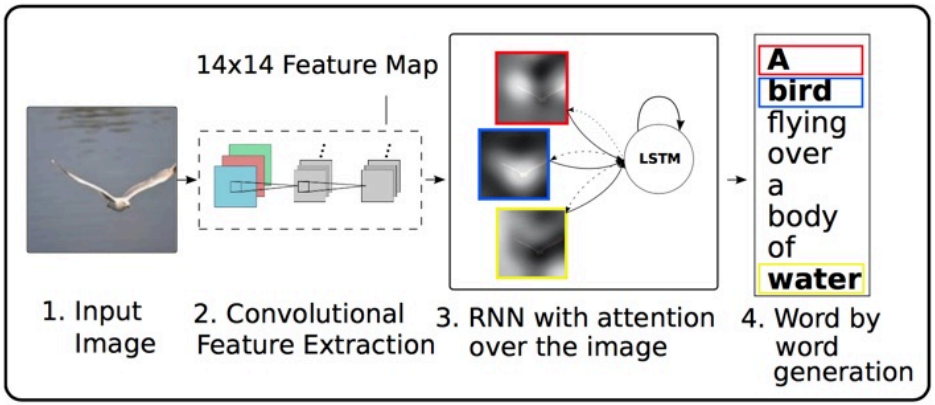


深層学習の中核技術

Adversarial training

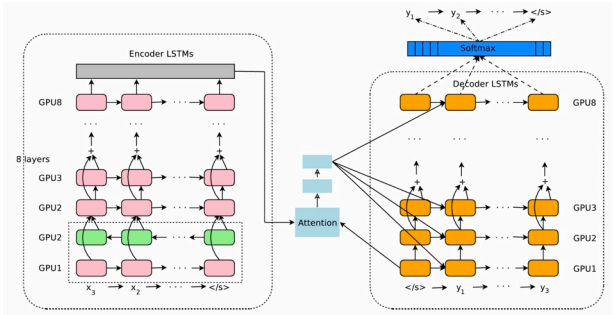


Attention

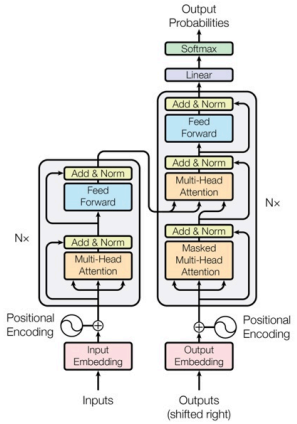


[Xu+2015]

Models for sequences

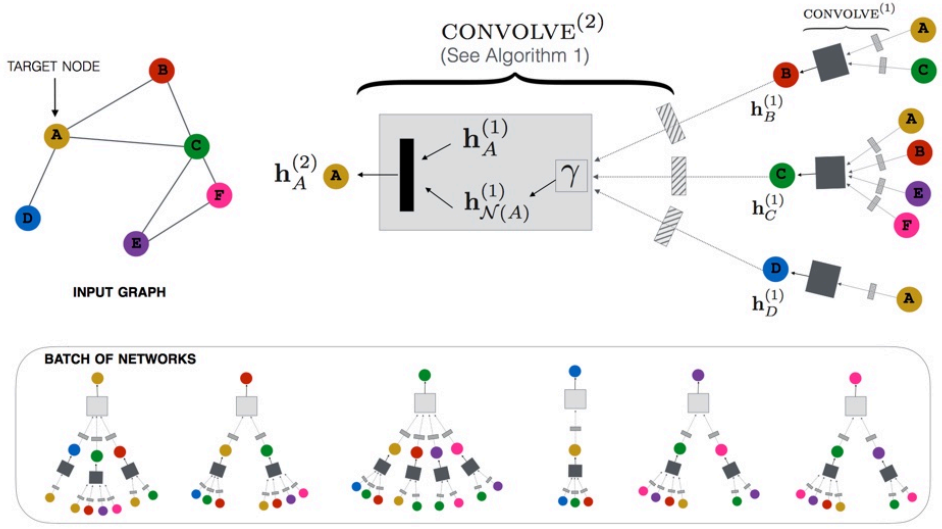


[Wu+16.09]



[VasWani+17.06]

Models for graphs



[Ying+2018]