深層学習の画像認識・処理への応用の現在と今後

岡谷貴之
東北大学大学院情報科学研究科
理化学研究所革新知能統合研究センター

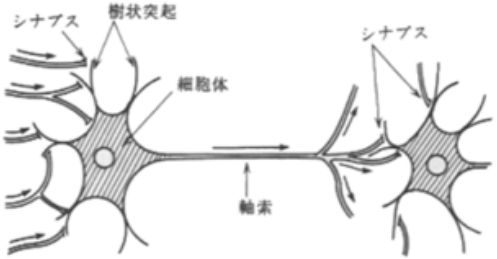
ディープラーニングの汎用性と特殊性
（深層学習）

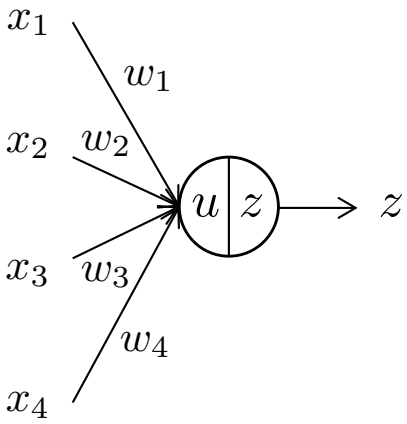
- 機械学習全体が緩やかに進展する中，ディープラーニングのみ速いペースで発展（≒応用分野を拡大）
- 画像・音声のような「生データ」を扱う問題で圧倒的

…

- 数多くの「ディープラーニングを要しない問題」が機械学習の適用を待っている

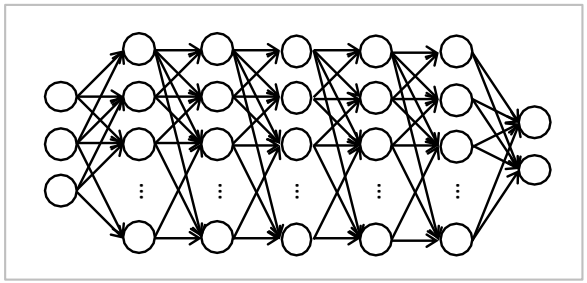
ディープラーニング

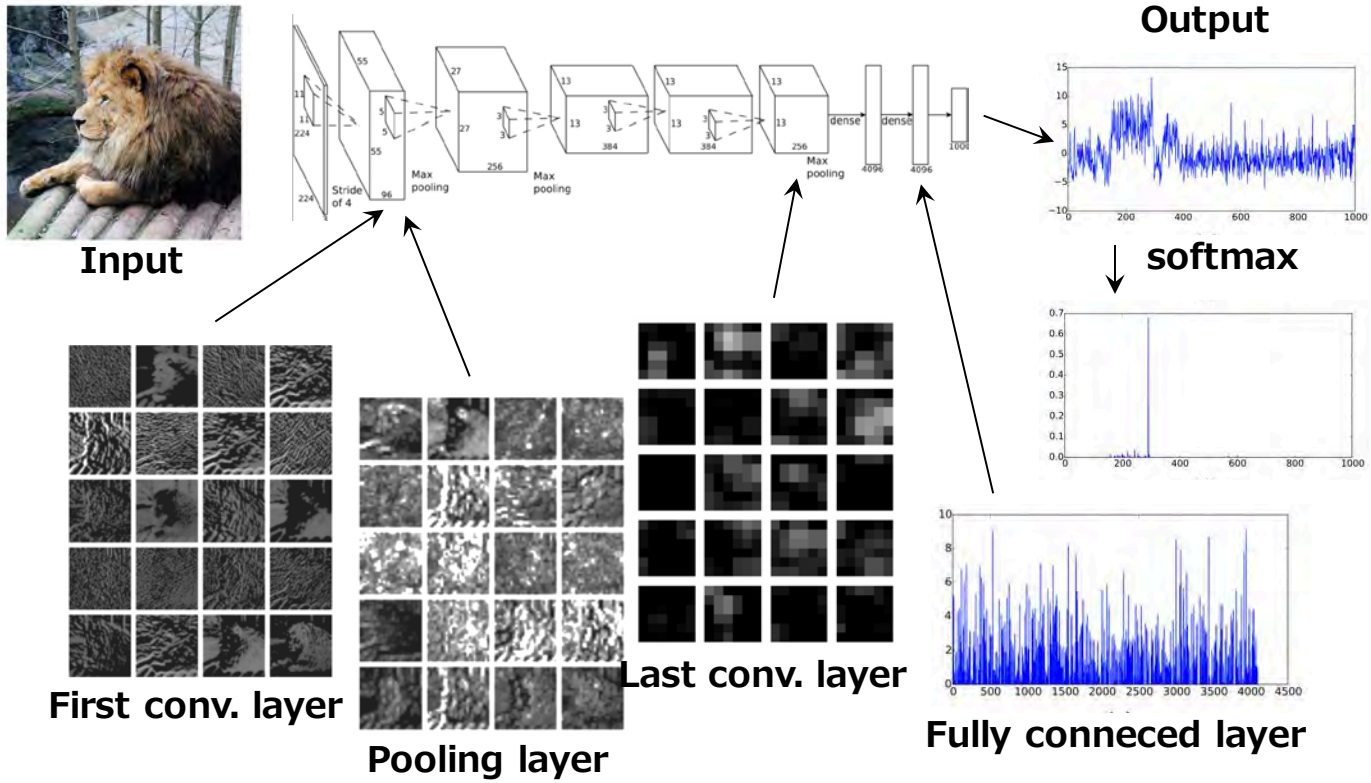




$$u = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b$$

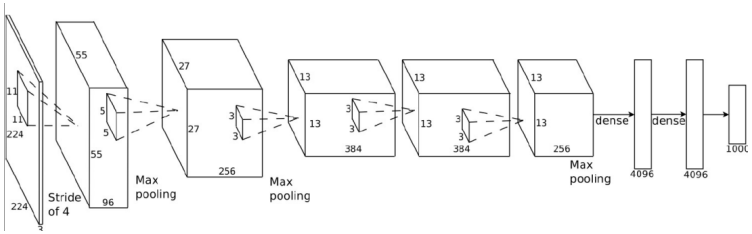
$$z = f(u)$$

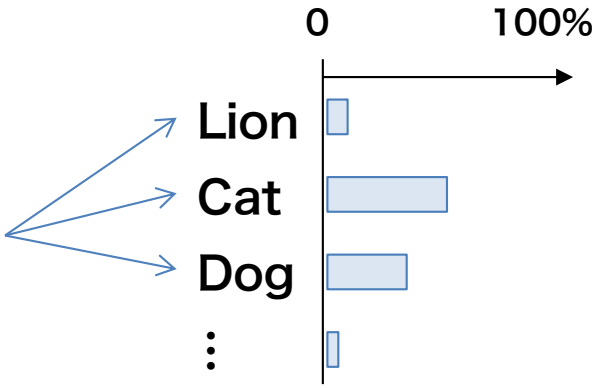




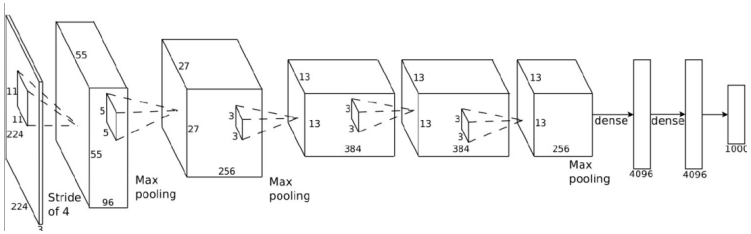
End-to-Endの教師あり学習

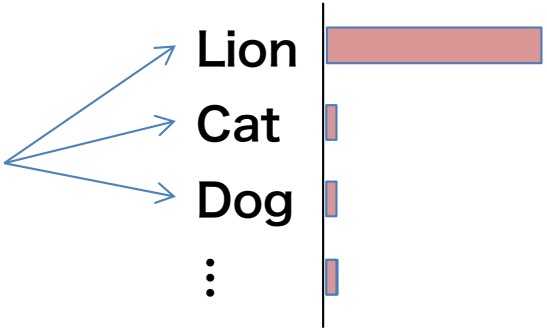
1. サンプルを入力し，結果を出力させる



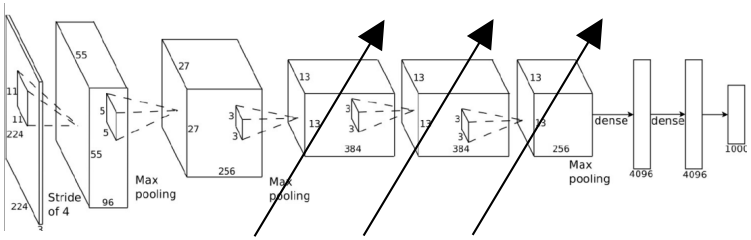


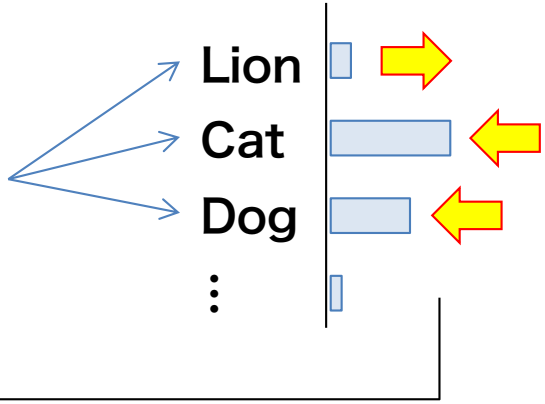
2. 正誤を評価（誤差を計算）





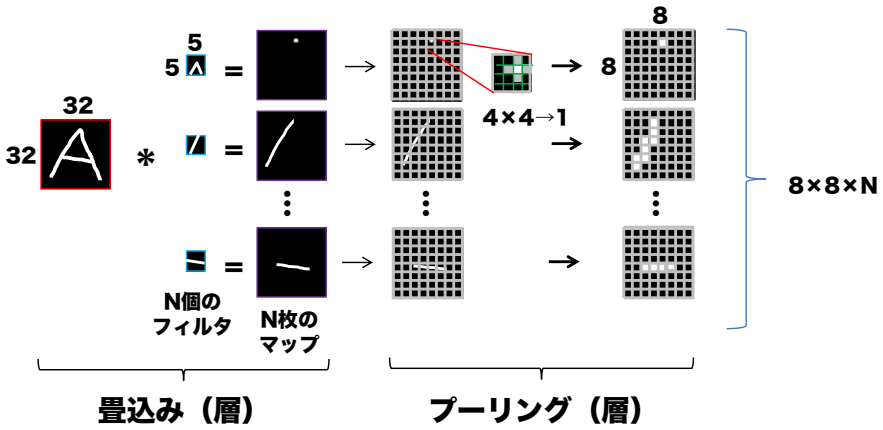
3. 重みを調節（誤差逆伝播法）



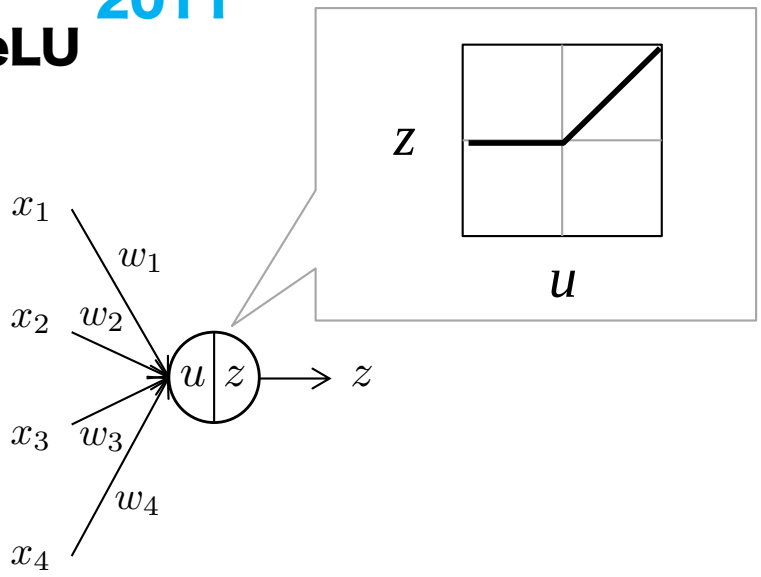


深層学習の中核技術

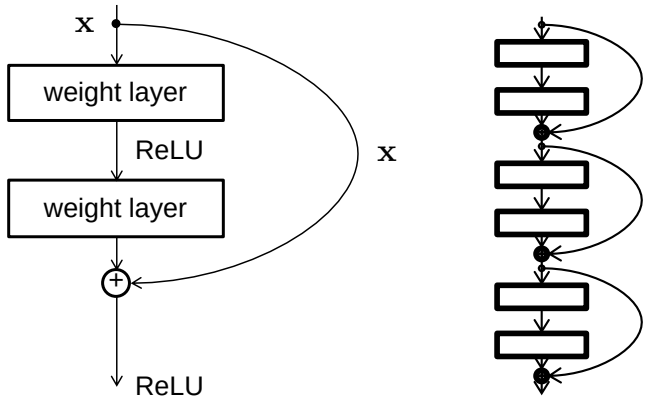
CNN：畳込みニューラルネット 1980-90



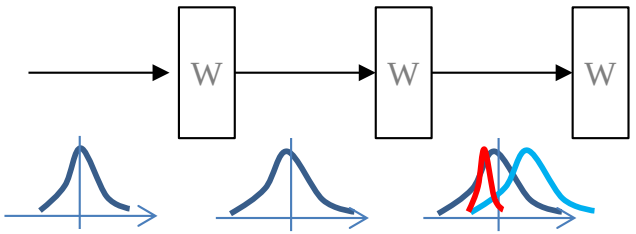
ReLU 2011



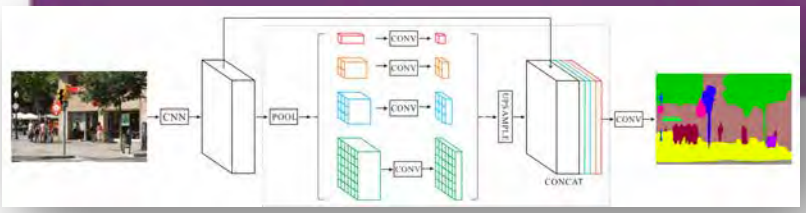
Skipconnection (ResNet) 2015



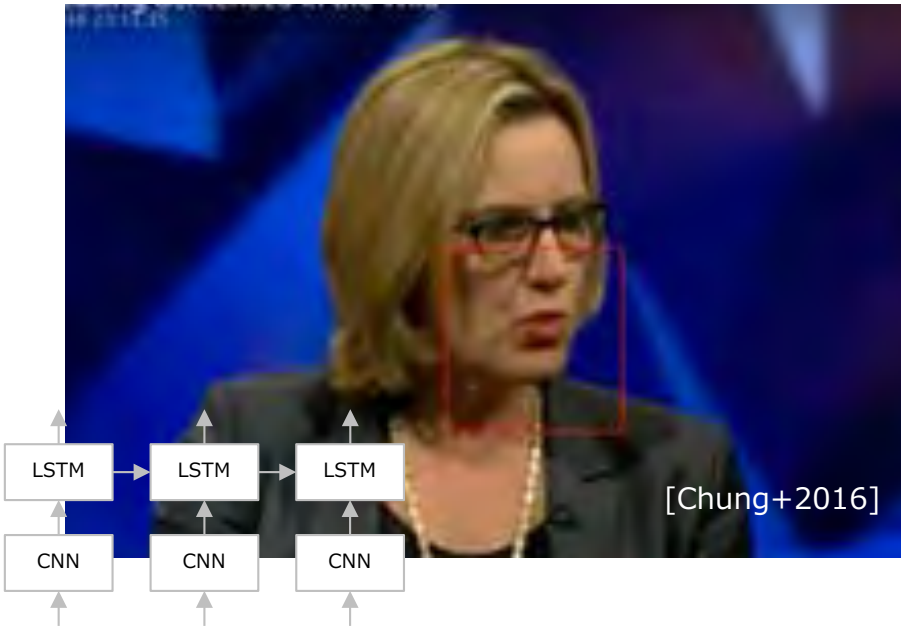
Batch Normalization 2015



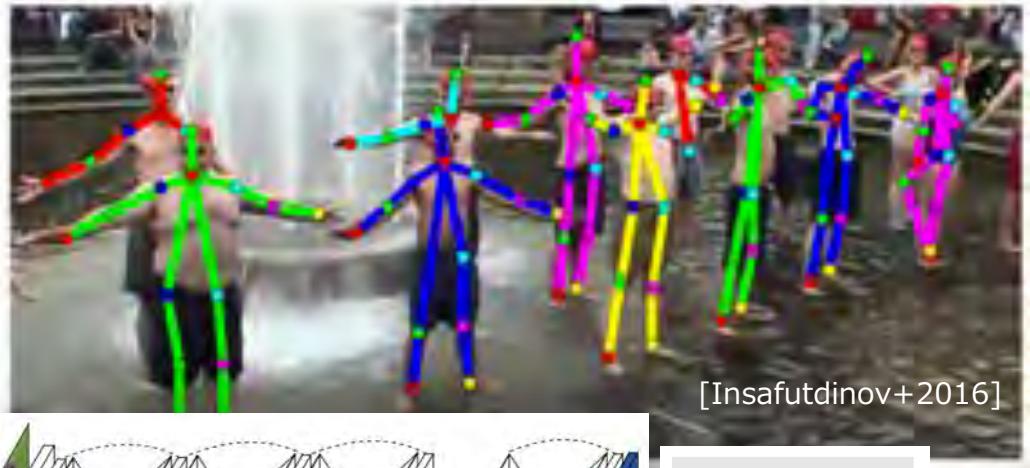
画素レベルの認識

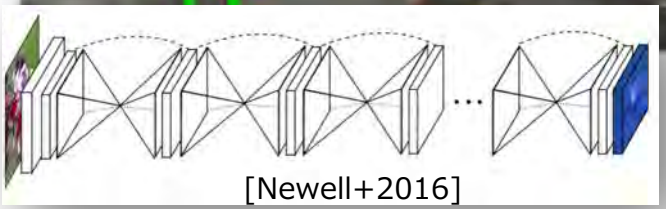


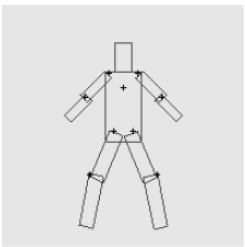
読唇



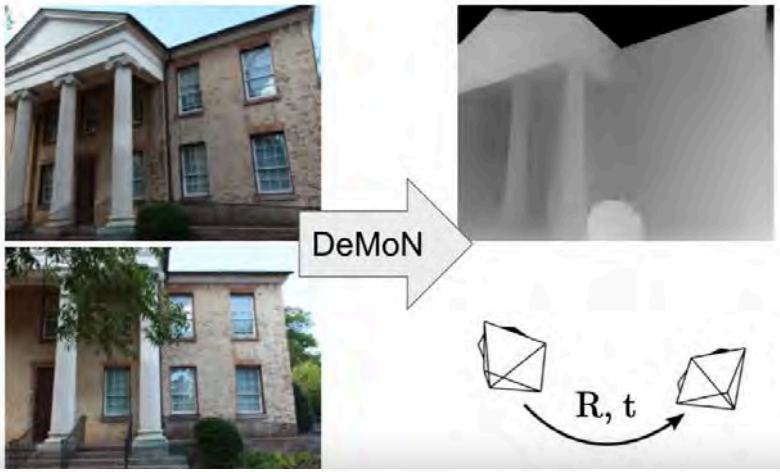
人体ポーズ

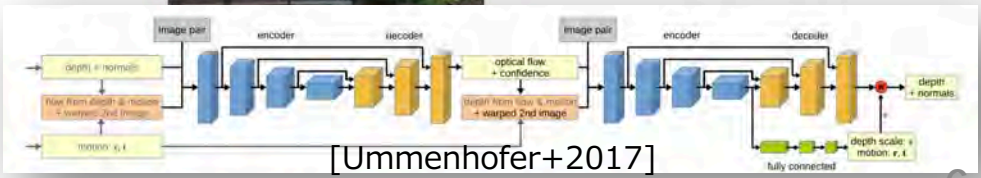




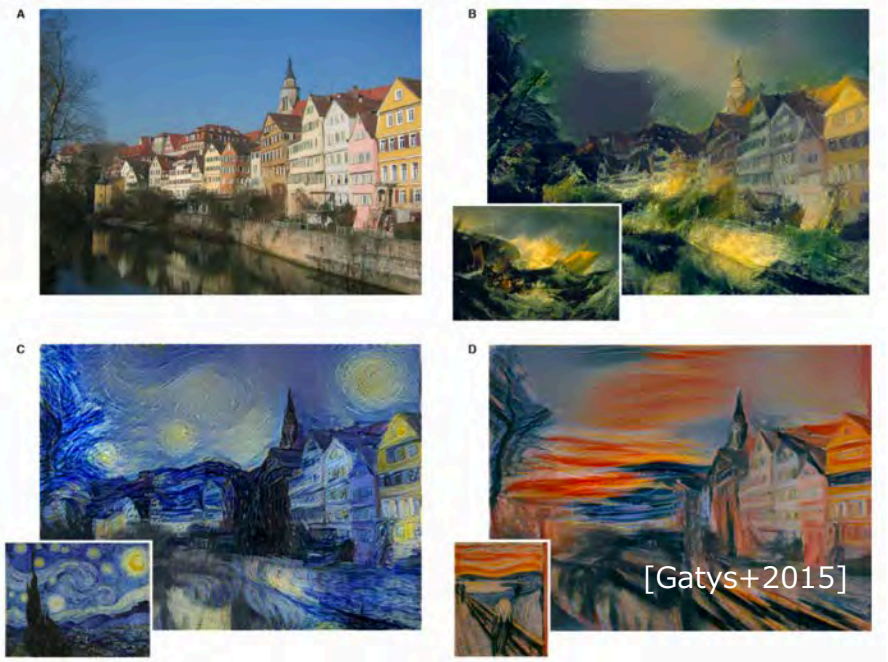


カメラ姿勢と奥行き

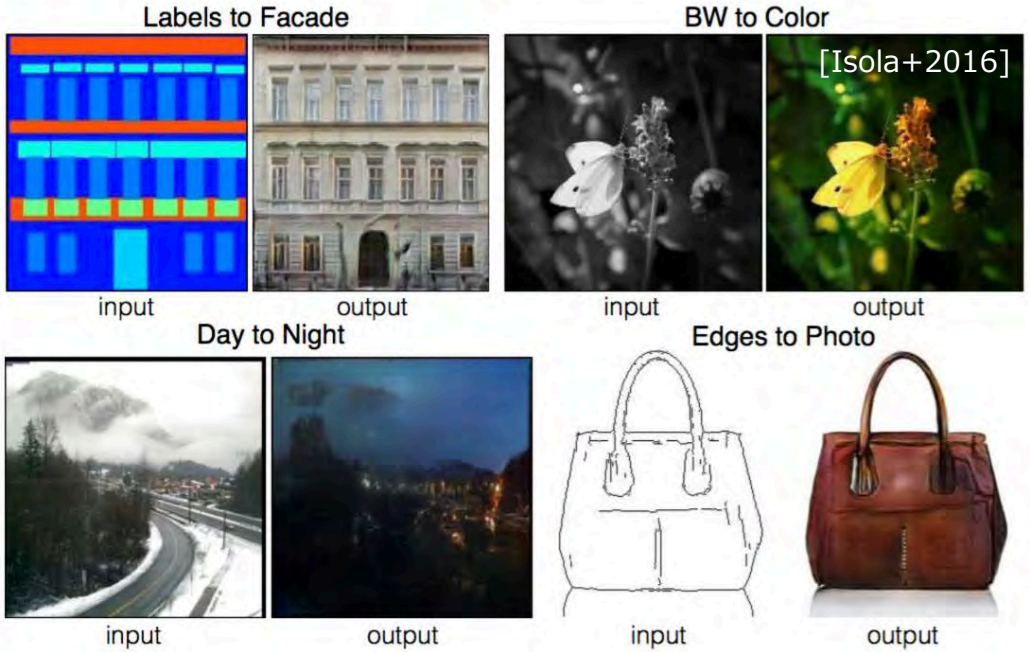




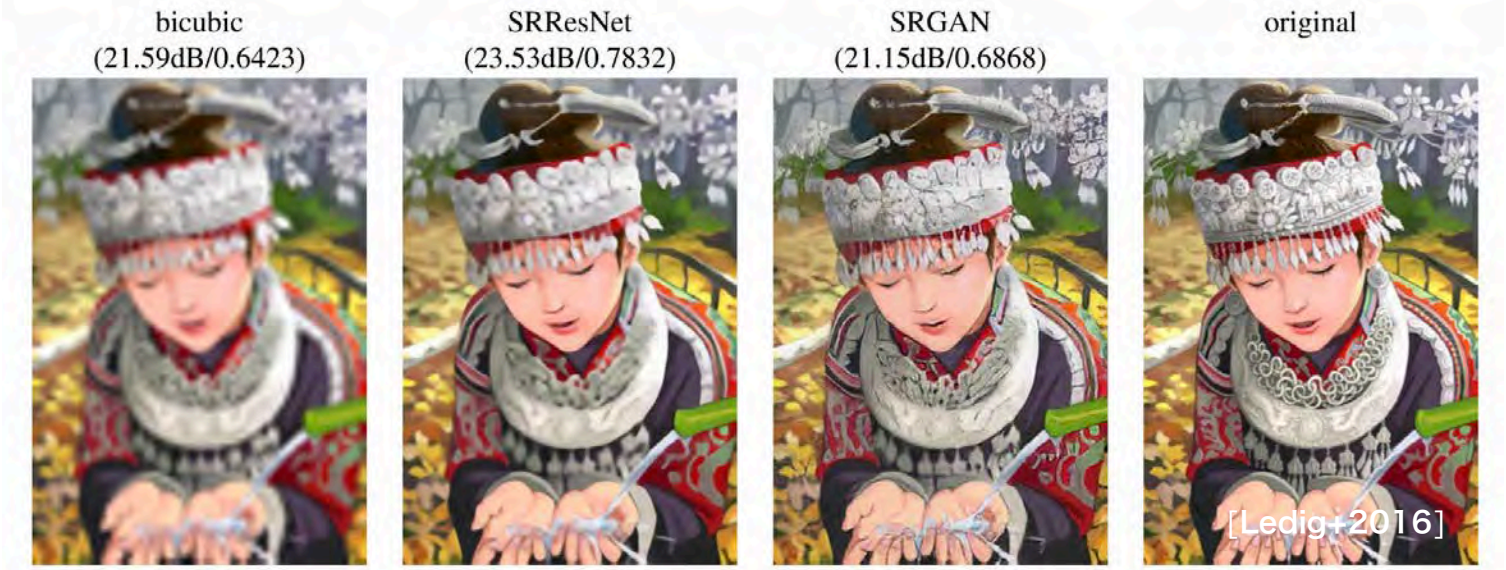
スタイル変換

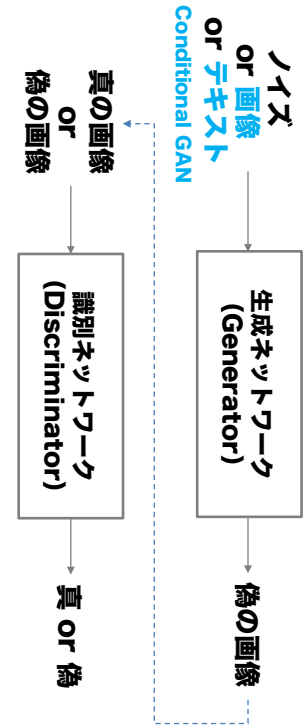


画像変換

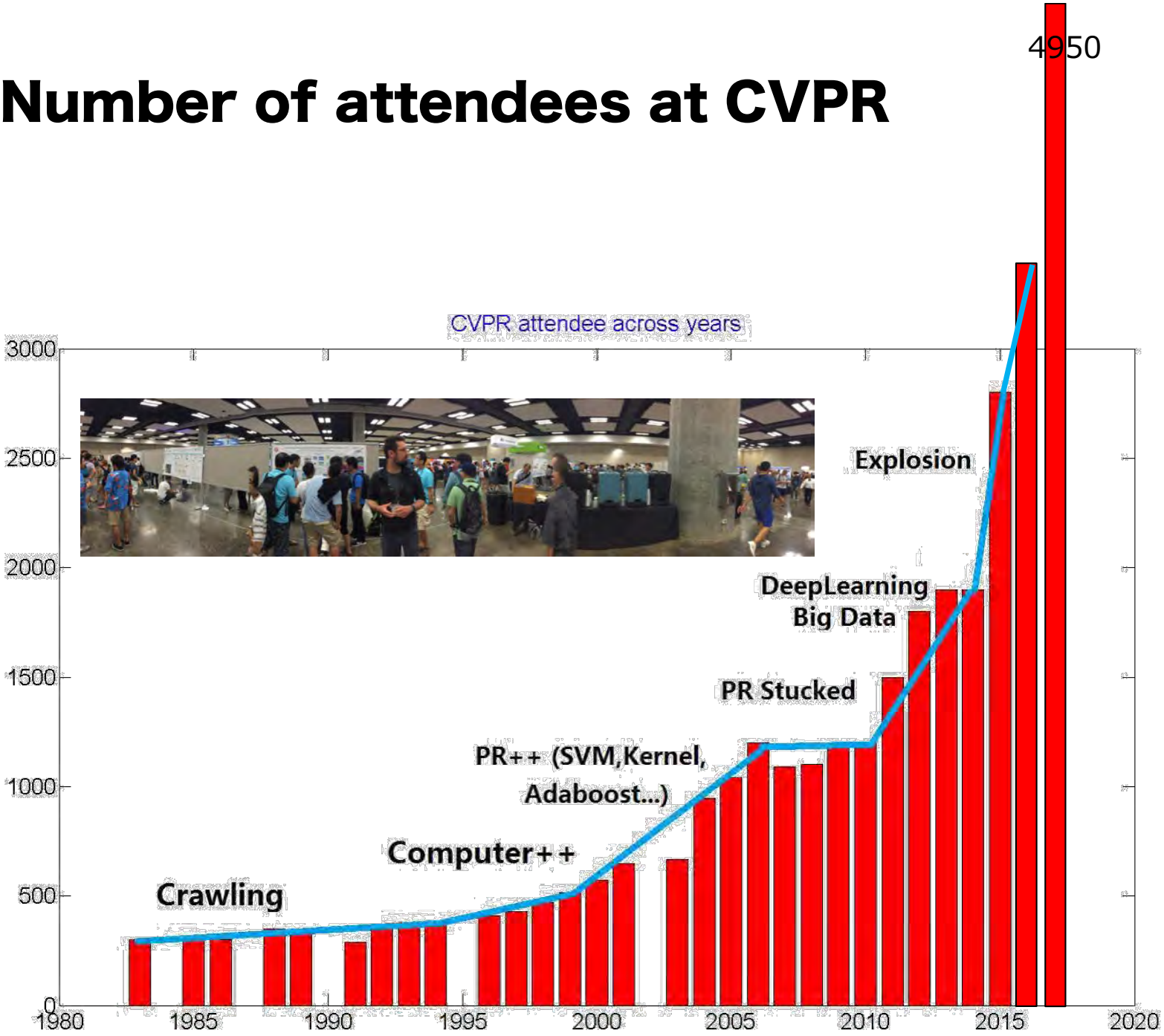


超解像





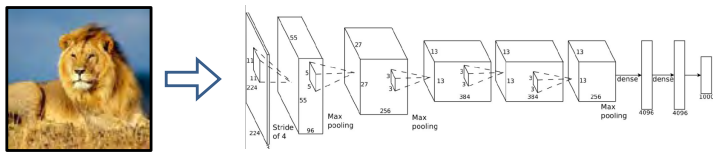
Number of attendees at CVPR



From a blogpost of Yann LeCun Created by Changbo Hu

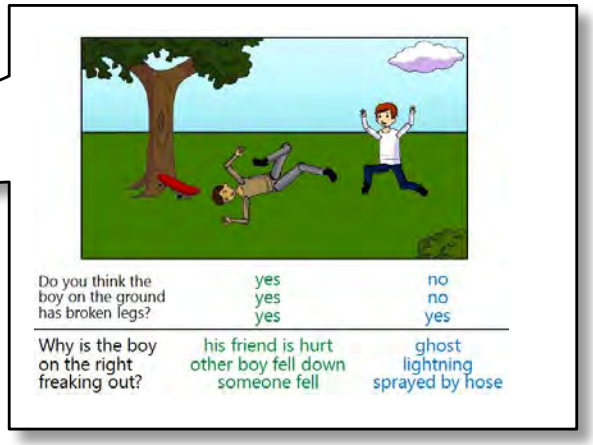
画像分野の研究の現状

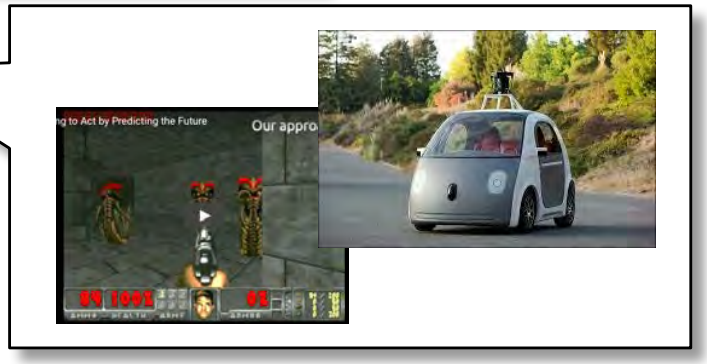
- ディープラーニングにより多くの問題が解決へ
  - 大量の入出力ペアを用いたend-to-end学習



- 残された問題

1. 真の画像理解
2. 出力がはっきりしない
3. データが集まらない
4. サイバースペースから実世界へ
5. DNN(CNN)の理解

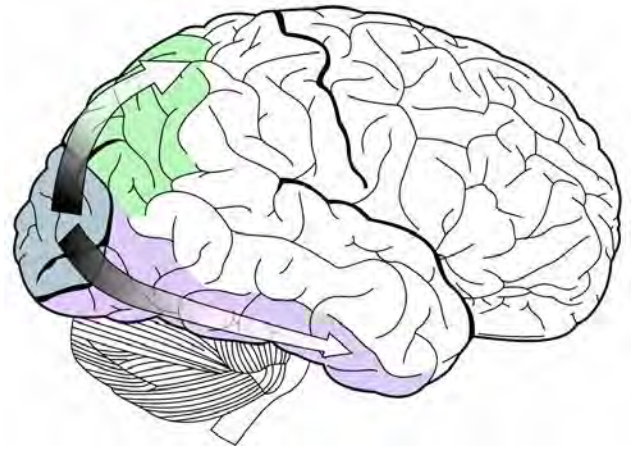


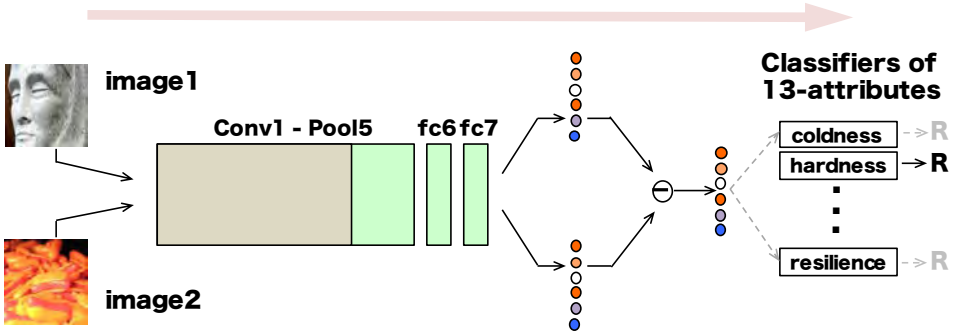


質感の画像認識

Liu+, Understanding Deep Representations Learned in CNNs for Material Recognition, VSS2016

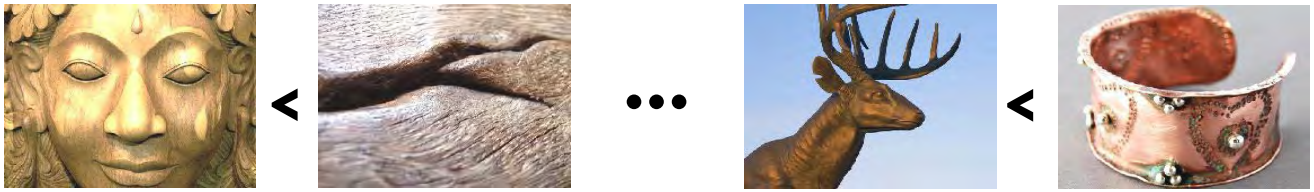
- 光沢感，透明感，柔らかさ，滑らかさ，…



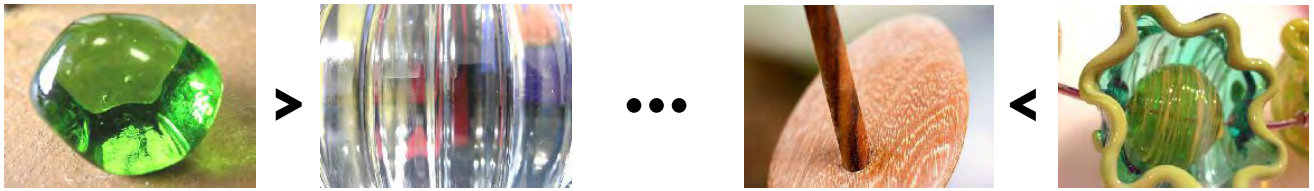


学習データ＝質感形容詞の比較データ

13 Attributes

Aged: 

Hard: 

Transparent: 

.
.
.

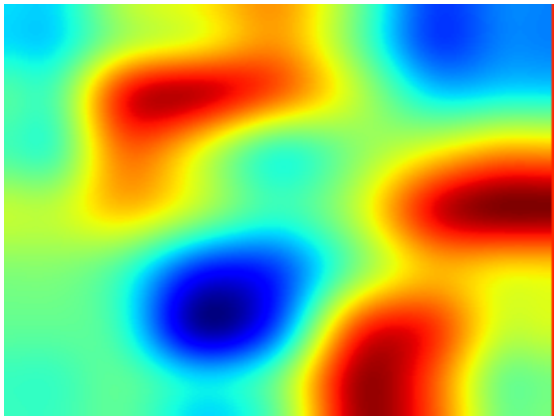
Wet: 

質感の画像認識

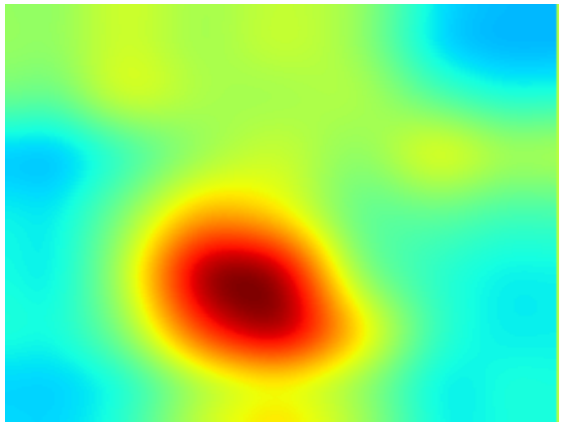
Liu+, Understanding Deep Representations Learned in CNNs for Material Recognition, VSS2016

INPUT

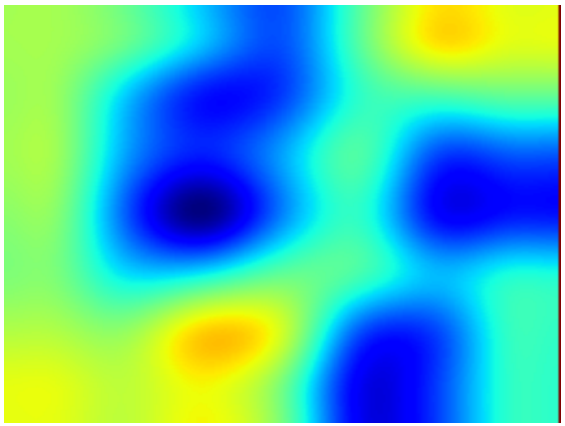
aged



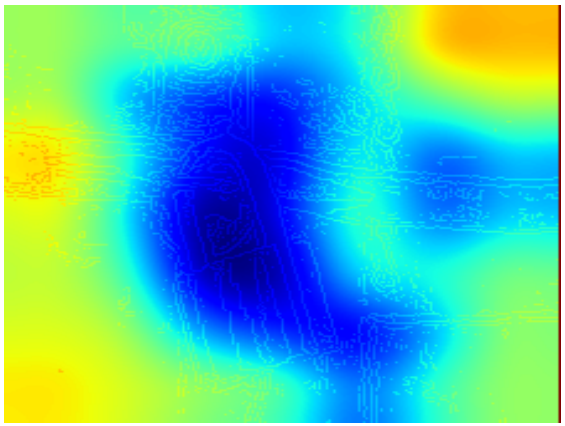
cold



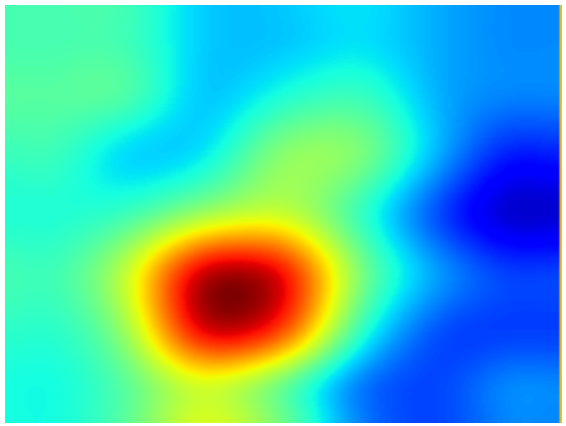
clean



fragile

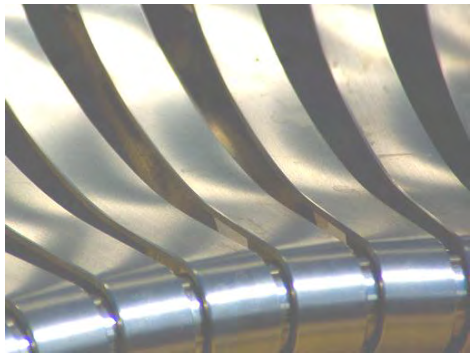


gloss

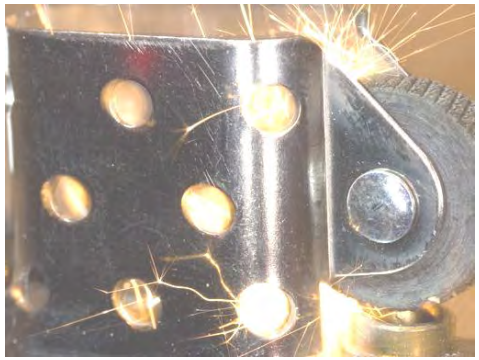


曖昧な「正解」

どちらがより「冷たく」感じますか？
（5人中）

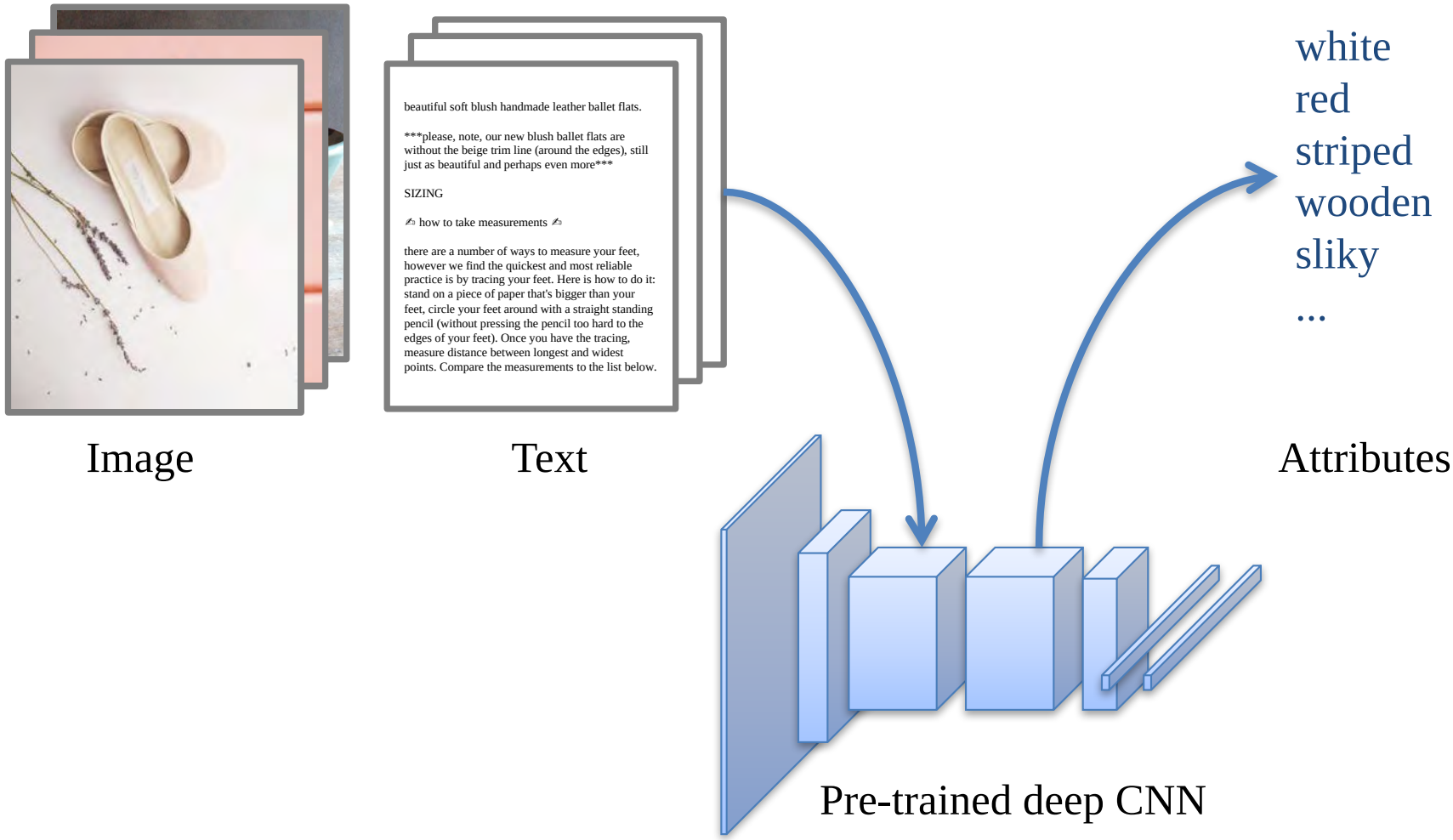
5  vs  0

4  vs  1

3  vs  2

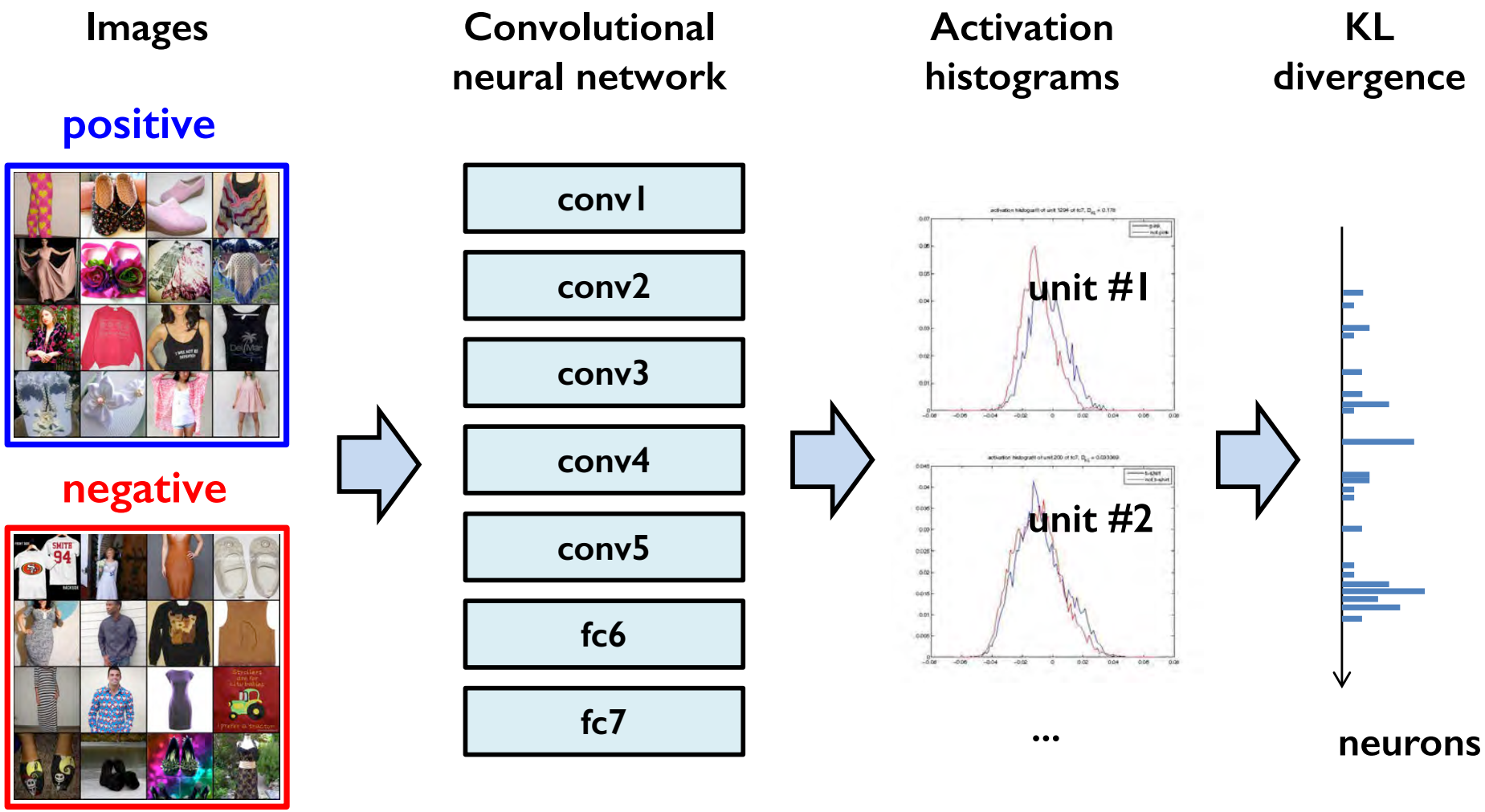
質感を表現する語彙の発見

Vittayakorn+, Automatic Attribute Discovery with Neural Activations, ECCV2016



質感を表現する語彙の発見

Vittayakorn+, Automatic Attribute Discovery with Neural Activations, ECCV2016



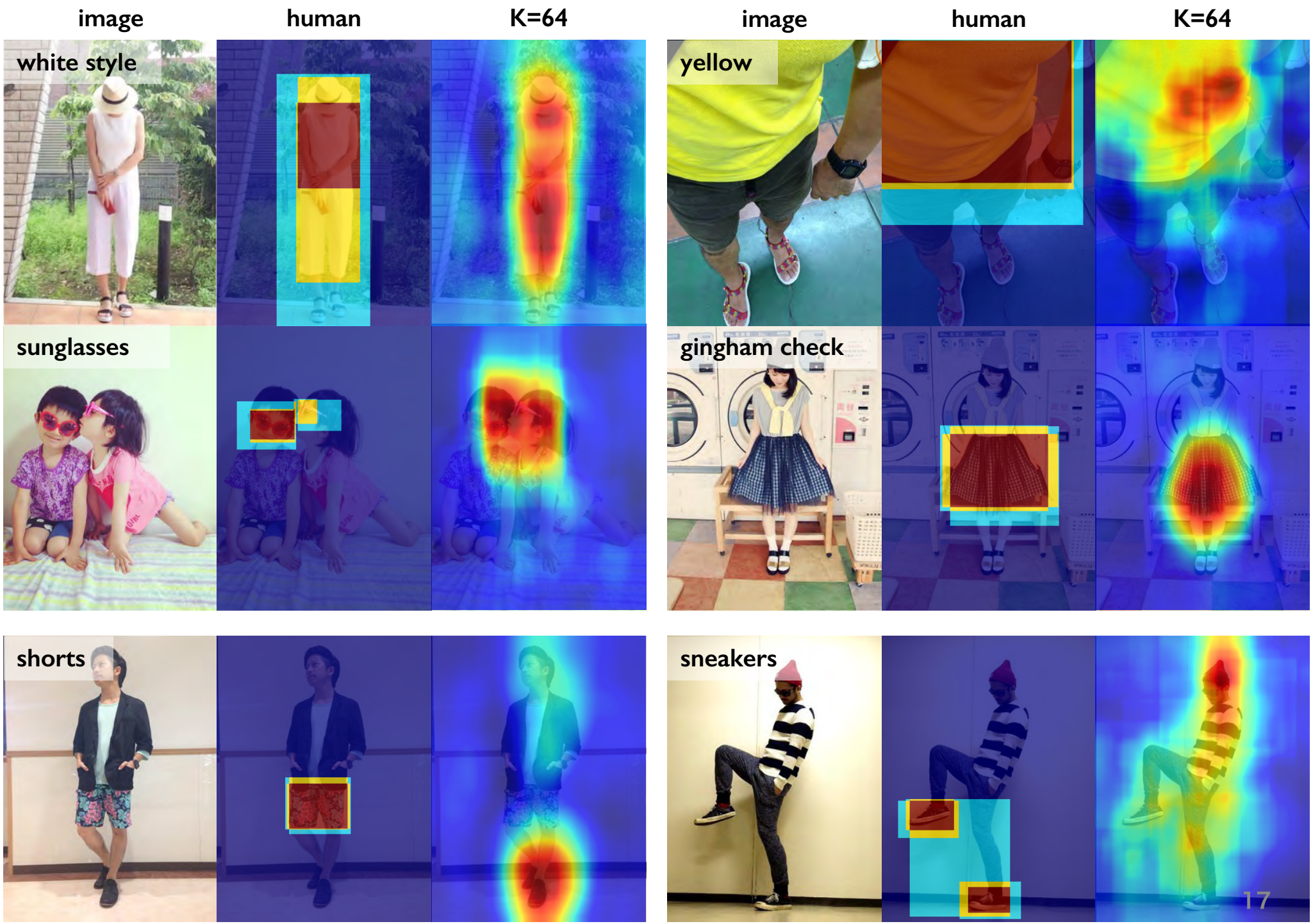
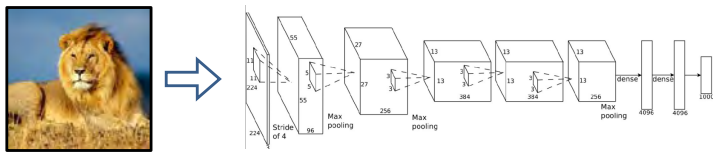
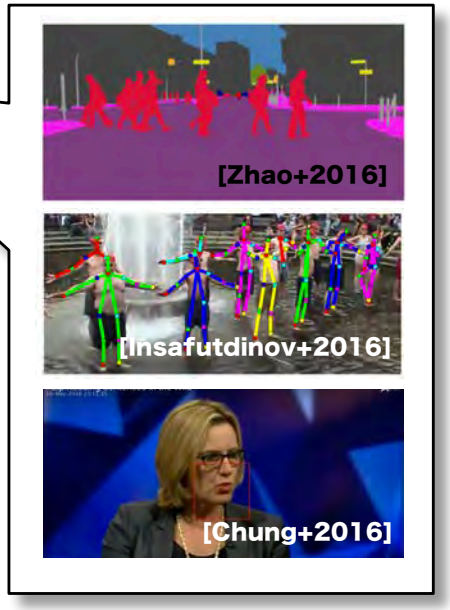


image
human
K=64
image
human
K=64
white style
yellow
sunglasses
gingham check
shorts
sneakers

画像分野の研究の現状

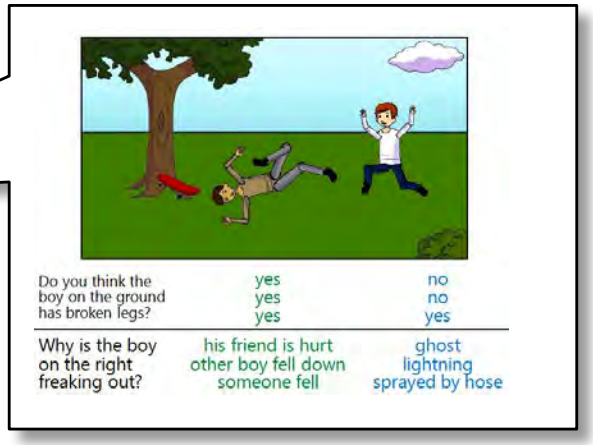
- ディープラーニングにより多くの問題が解決へ
  - 大量の入出力ペアを用いたend-to-end学習

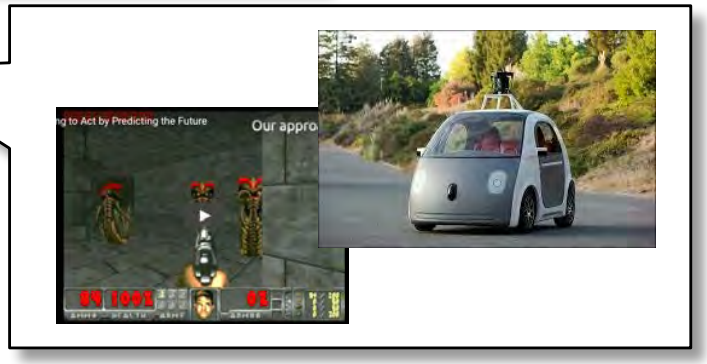




- 残された問題

1. 真の画像理解
2. 出力がはっきりしない
3. データが集まらない
4. サイバースペースから実世界へ
5. DNN(CNN)の理解

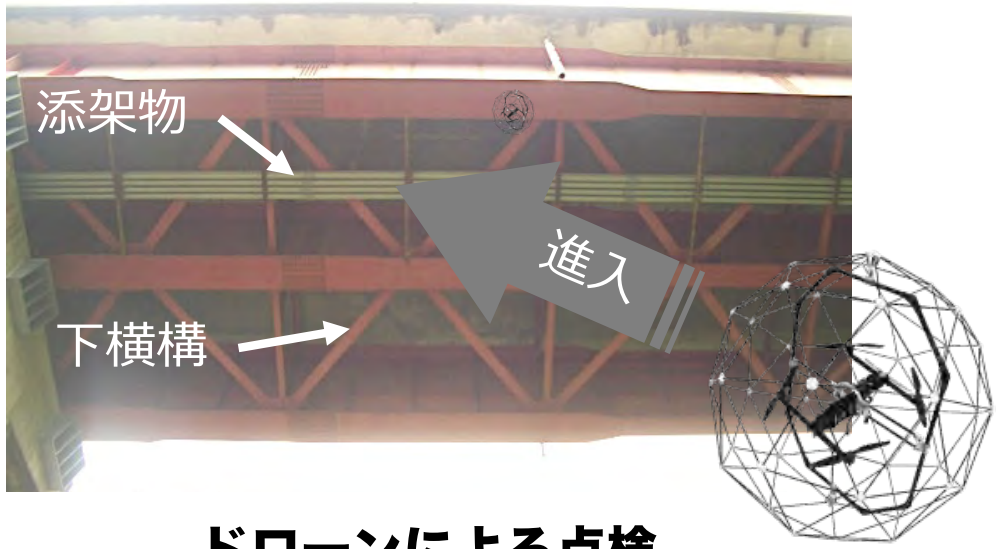




ドローンを用いた橋梁点検

- 橋梁点検の効率化が急務
  - 老朽化しつつある全国70万橋・うち４割が40年超
- なぜ橋梁？
  - トンネルや道路・滑走路等と比較して格段に難しい
  - 形状・構造・周囲環境の多様性 ⇒ 専用計測装置開発が困難
- 内閣府SIP→理研AIP
  - 内閣府SIP「インフラ維持管理・更新・マネジメント技術」
  - 東北大コンソ「橋梁の打音検査ならびに近接目視を代替する飛行ロボットシステムの研究開発」

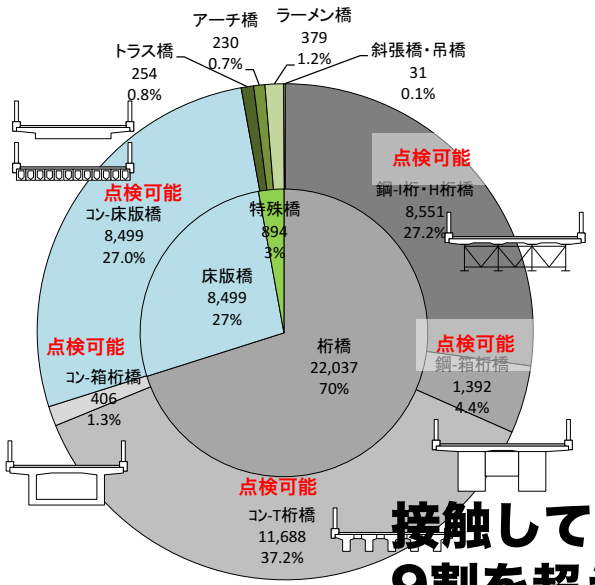
従来：橋梁点検車の点検



ドローンによる点検

球殻ドローン [SIP東北大コンソーシアム(代表：大野和則@東北大)]

- ぶつかりながら飛行し、近接撮影が可能な受動回転球殻ヘリ

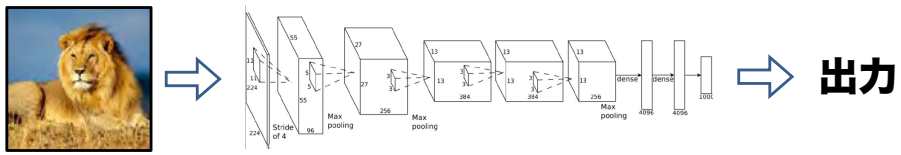


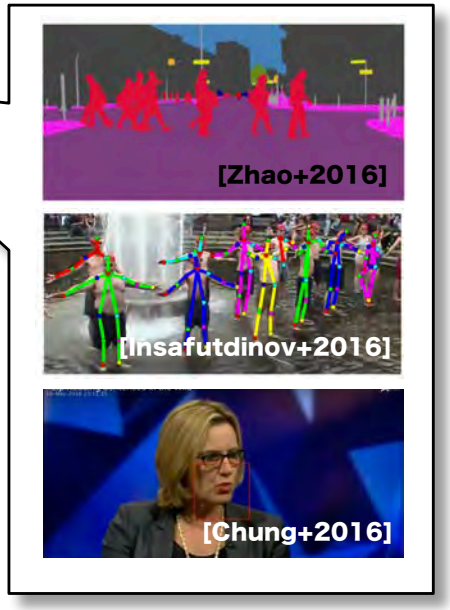
遠隔操縦で２０m上空にある橋梁の横構や添架物を接触しながらすり抜け、床版などを近接撮影

接触しても構わないため
9割を超える橋梁の点検が可能に！

画像分野の研究の現状

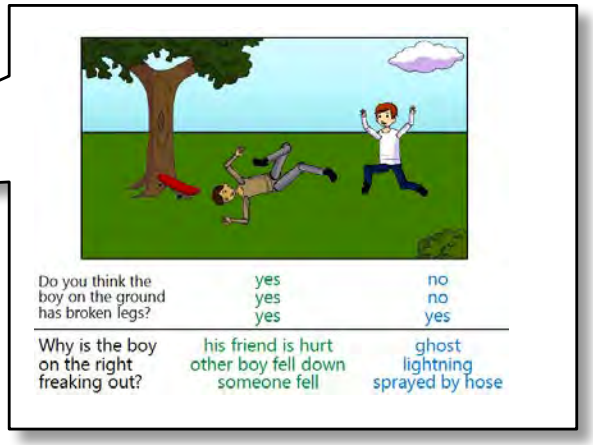
- ディープラーニングにより多くの問題が解決へ
  - 大量の入出力ペアを用いたend-to-end学習

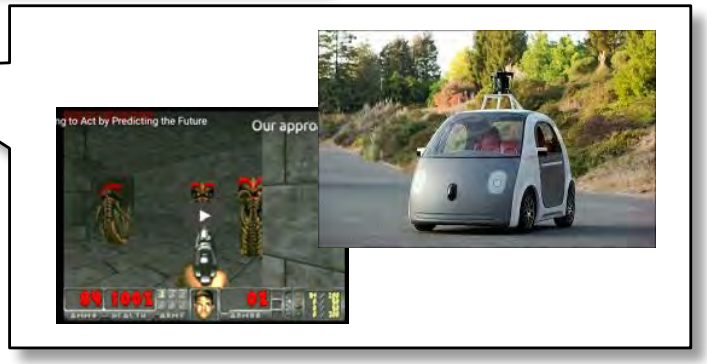




- 残された問題

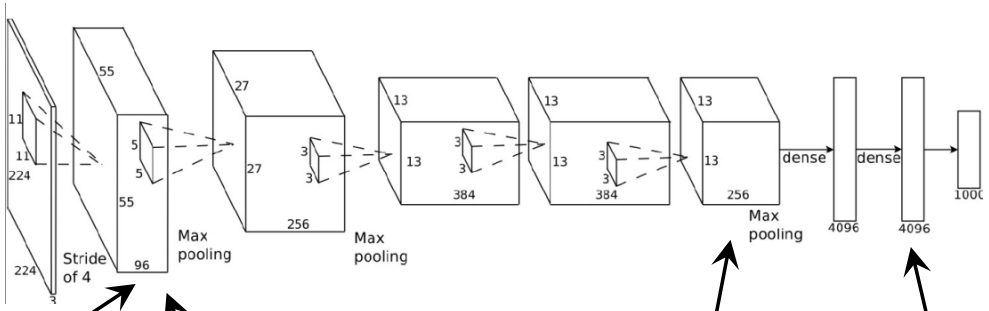
1. 真の画像理解
2. 出力がはっきりしない
3. データが集まらない
4. サイバースペースから実世界へ
5. DNN(CNN)の理解



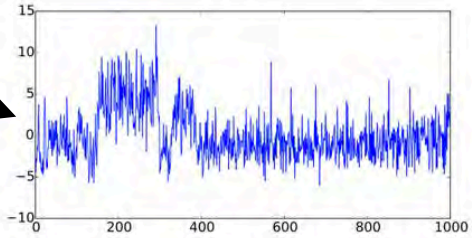


入力画像に対する中間層出力

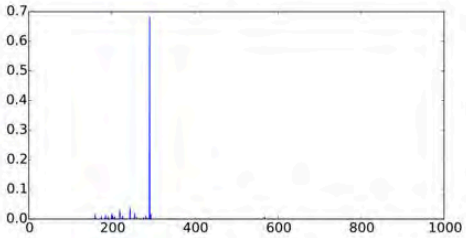
入力

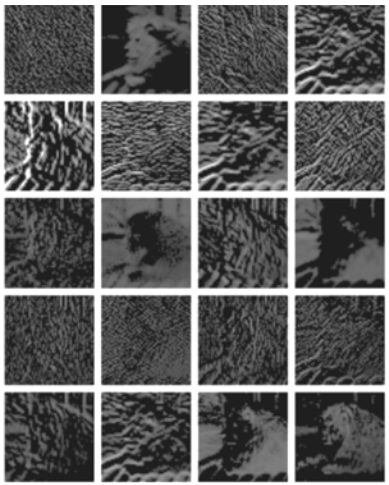


出力層

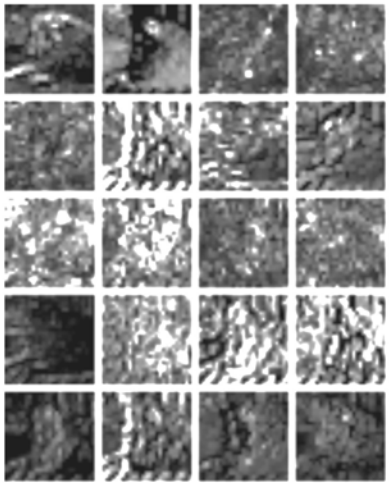


↓ softmax

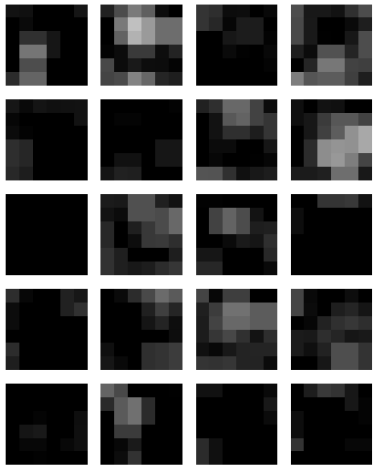




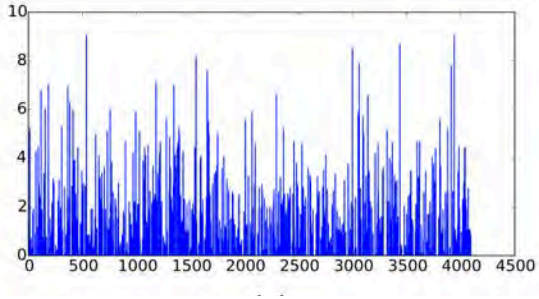
初期畳込み層



プーリング層

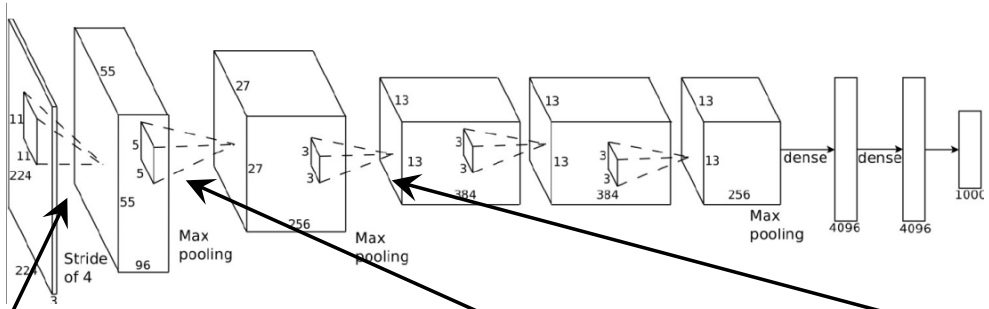


最終畳込み層

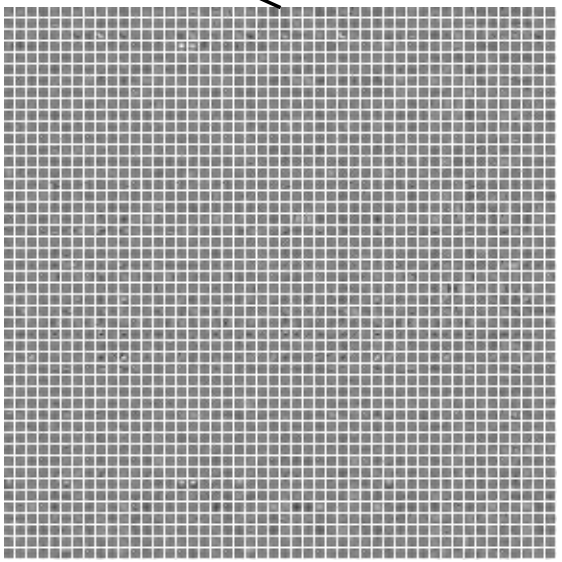


全結合層

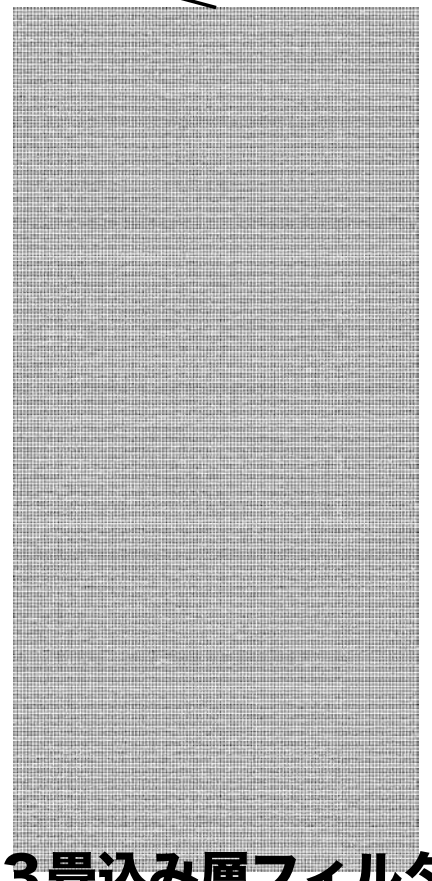
学習した重み（畳込みフィルタ）



第１畳込み層フィルタ



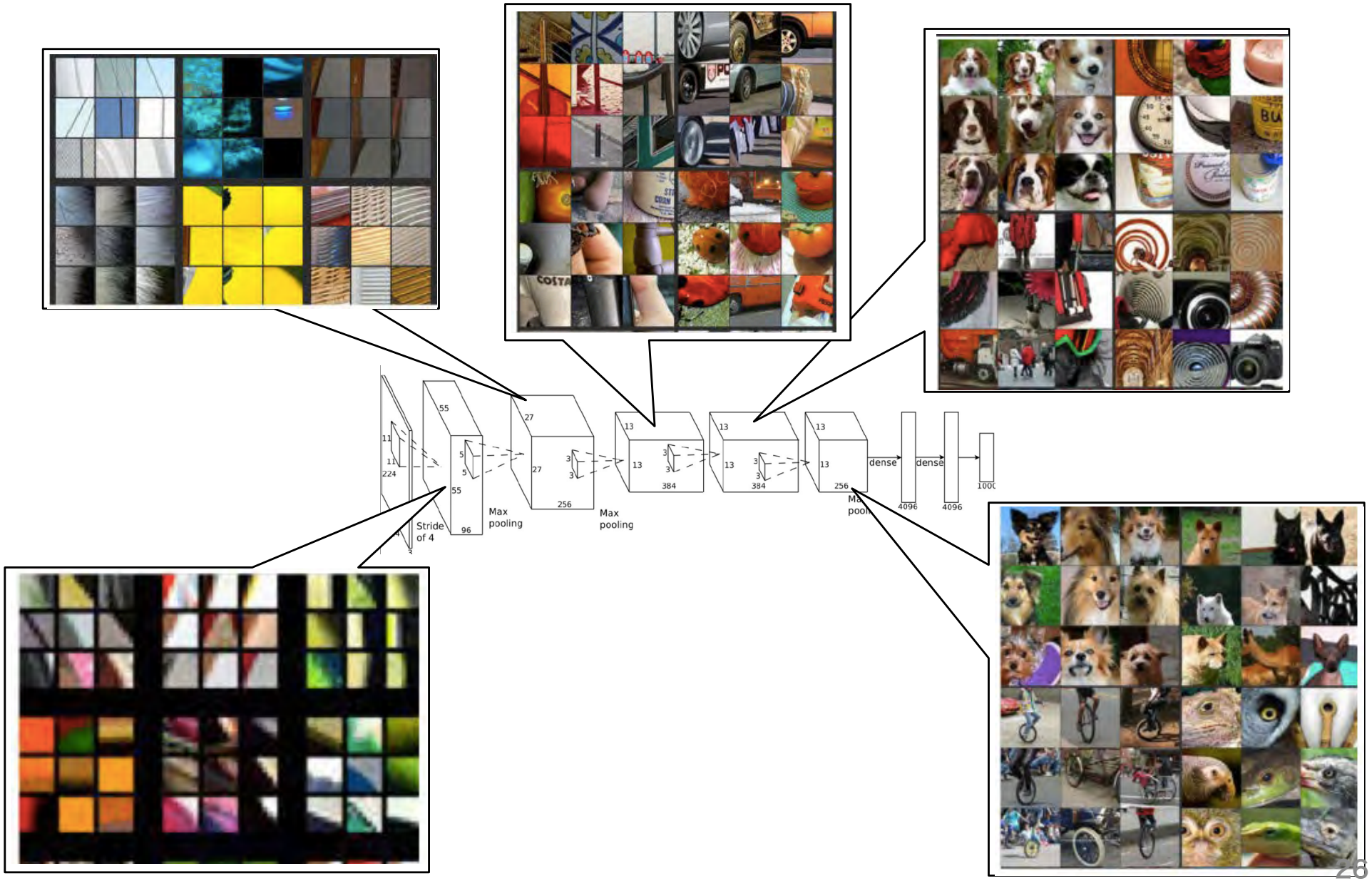
第２畳込み層フィルタ



第３畳込み層フィルタ

各ユニットを最も活性化させる入力

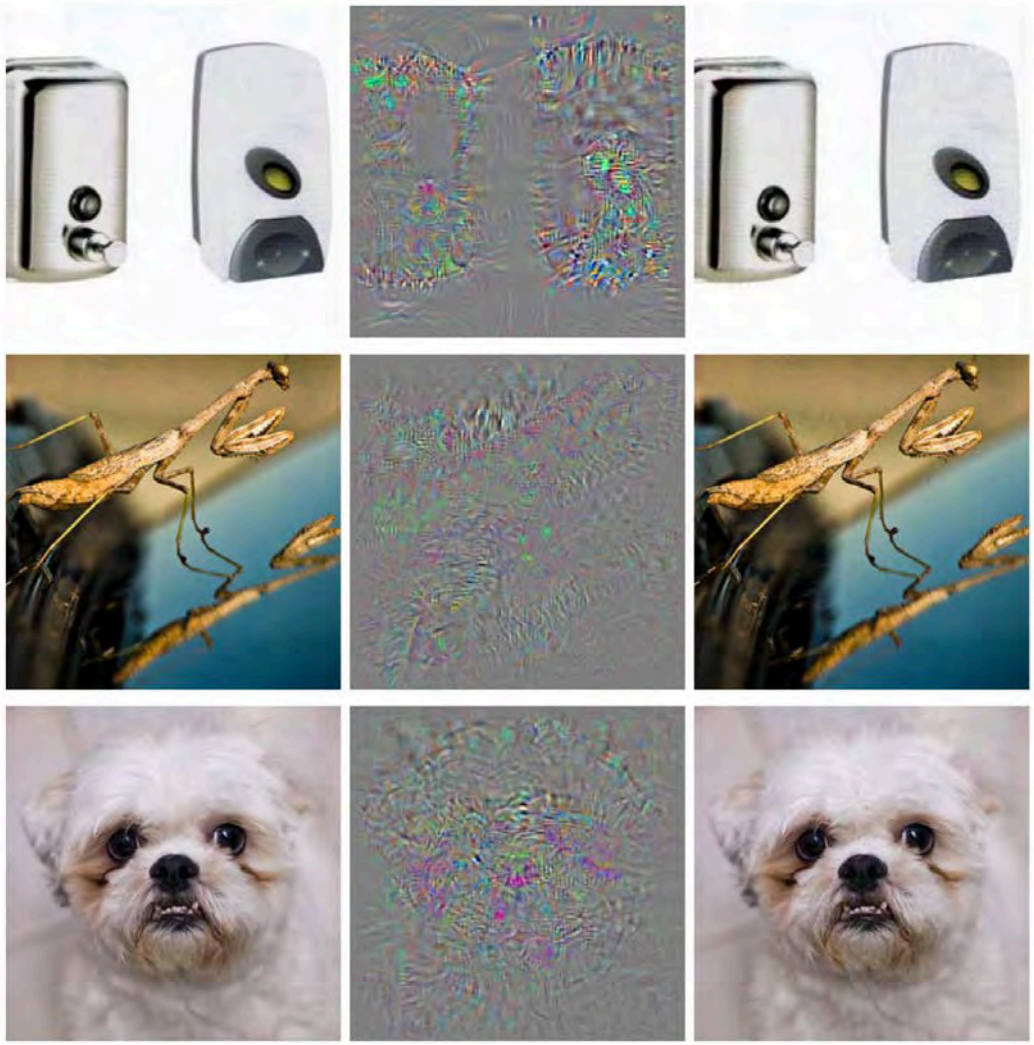
Zeiler-Fergus, Visualizing and understanding convolutional networks, 2014



CNNを騙す

Szegedy+, Intriguing properties of neural networks, 2014

目立たないわずかな改変 ⇒ 誤認識

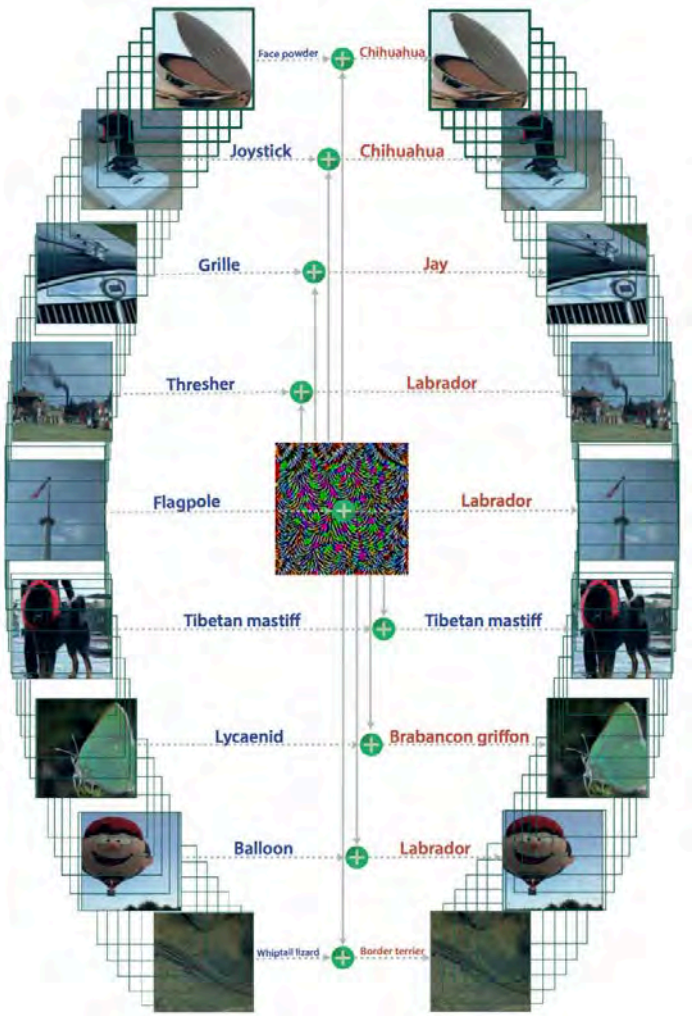


Correctly recognized

Additive noise

Recognized as“Ostrich”

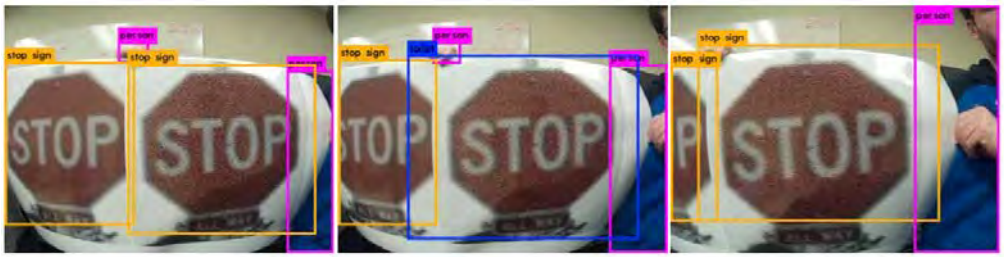
ユニバーサルな改変：画像・ネットワークによらず誤認識

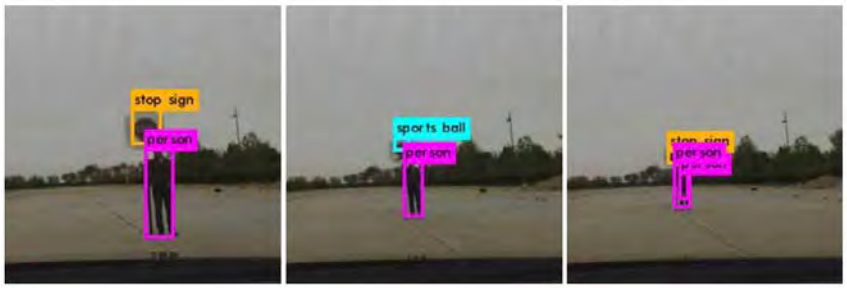


[Moosavi-Dezfooli+2017]

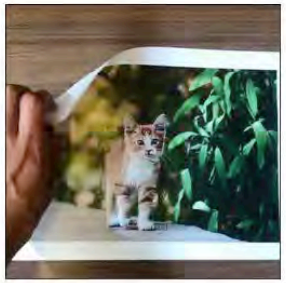
CNNを騙す

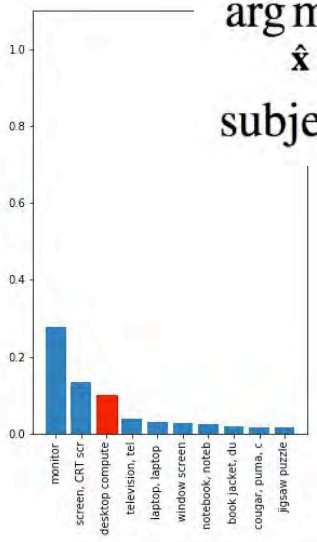
- 現実には懸念不要では？
  - NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles [Lu+2017]





- 「ロバストな」改変は可能
  - Synthesizing Robust Adversarial Examples [Athalye-Sutskever2017]





$$\arg\max_{\hat{\mathbf{x}}} \quad \mathop{\mathbb{E}}_{t\sim T} \log P\left(\hat{y} \mid t(\hat{\mathbf{x}})\right)$$

$$\text{subject to} \quad \left\|\mathbf{x}-\hat{\mathbf{x}}\right\|_{\infty} \leq \varepsilon$$

情報理論からのアプローチ

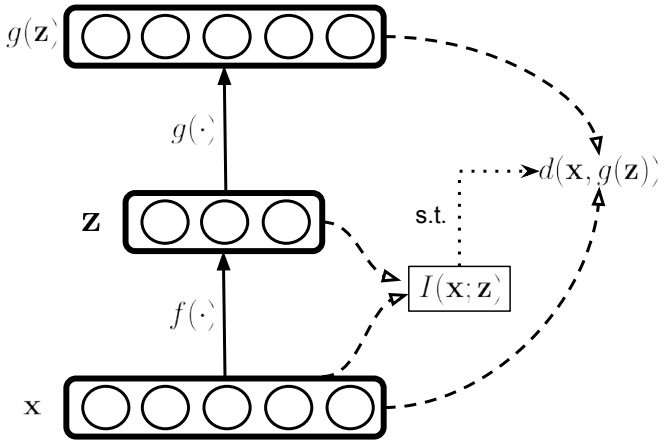
Zhang-Ozay-Sun-Okatani, Information Potential Autoencoeders, arXiv 2017

- **データとその符号の相互情報量を最小化**

$$\min_{f,g} I(\mathbf{x};\mathbf{z})$$

$$\text{s.t.} \quad \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[d(\mathbf{x}, g(\mathbf{z}))] \leq D$$

確率的符号化： $\mathbf{z} = \mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \boldsymbol{\epsilon},$
[Kingma-Welling2014]



変分自己符号化器(VAE) [Kingma-Welling2014]

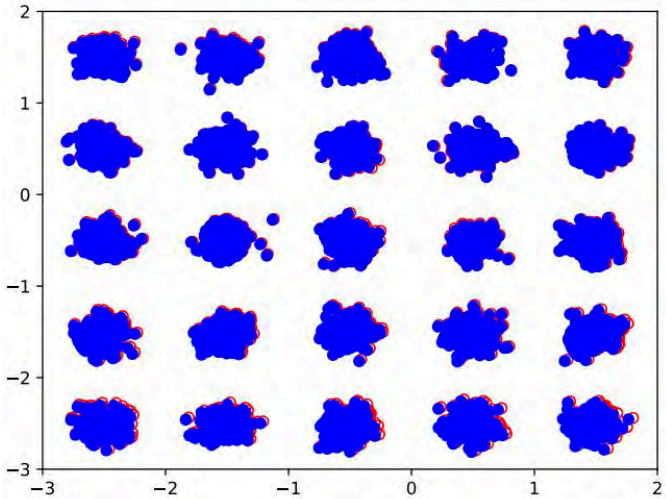
$$I(\mathbf{x};\mathbf{z}) \leq \frac{1}{2N}\sum_{i=1}^{N}\left(\left|\left|\mu(x_i)\right|\right|_2^2 + \left|\left|\sigma^2(x_i)\right|\right|_1 - \log\left|\text{diag}(\sigma^2(x_i))\right| - 1\right)$$

情報ポテンシャル自己符号化器(IPAE)

$$I(\mathbf{x};\mathbf{z}) \leq \frac{1}{2KN^2}\sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{j=1}^{N}\left(\left(\mu(x_j) - \mu(x_i) - \sigma(x_i)\odot\epsilon_k\right)^2 \oslash \sigma^2(x_j) - 1\right)$$

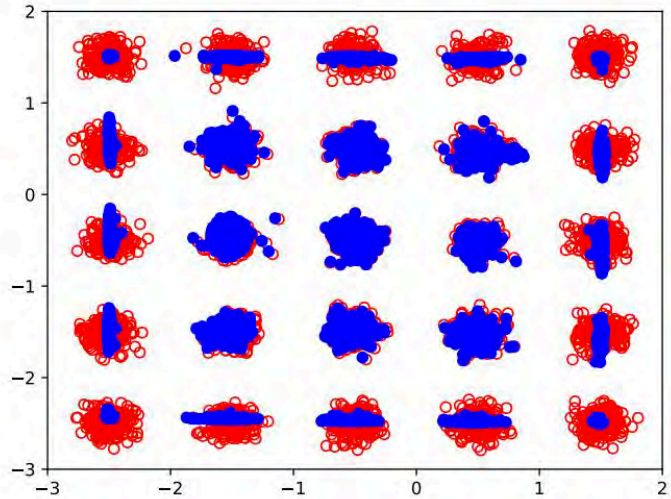
情報理論からのアプローチ

Zhang-Ozay-Sun-Okatani, Information Potential Autoencoeders, arXiv 2017

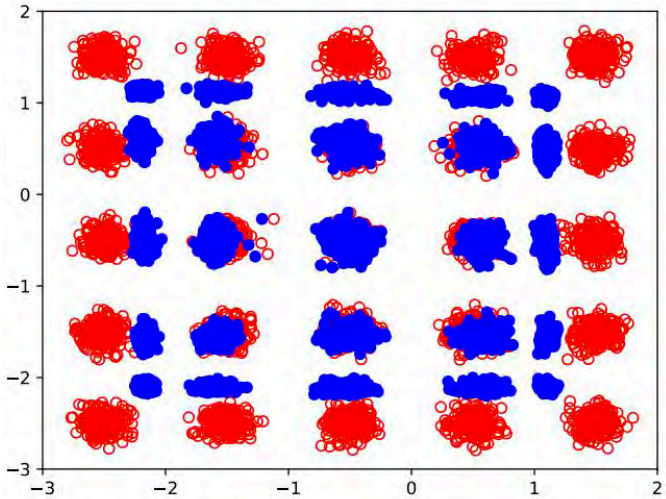


(a) VAE, $\beta = 0.0001$, $\mathcal{E} = 0.00972$.

VAE（従来）

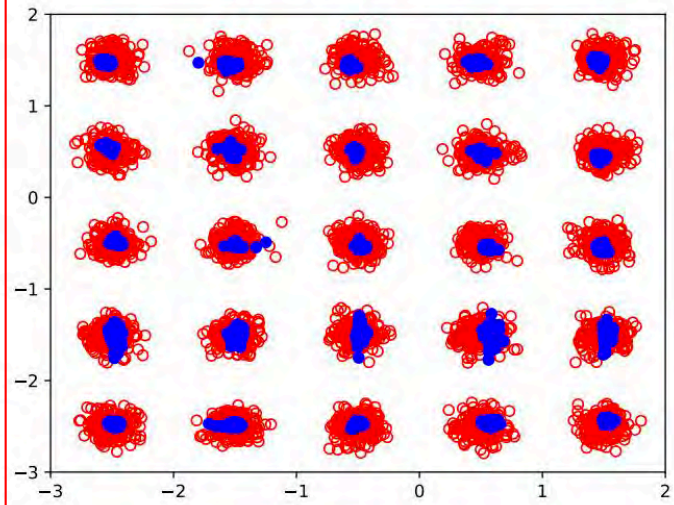


(b) VAE, $\beta = 0.1$, $\mathcal{E} = 0.00728$.



(c) VAE, $\beta = 0.5$, $\mathcal{E} = 0.0653$.

IPAE（提案）



(d) IPAE, $\beta = 0.001$, $\mathcal{E} = 0.00204$.

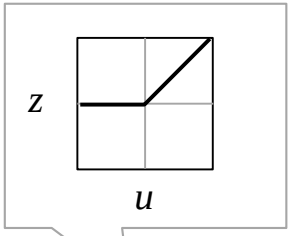
Red circles denote input data and the blue dots denote reconstructed data. $\mathcal{E}$ is the average euclidean distance of all reconstructed samples to the means of Gaussian from which they were drawn.

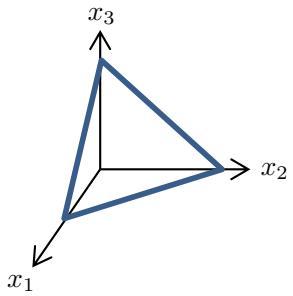
小さくて高性能なネット

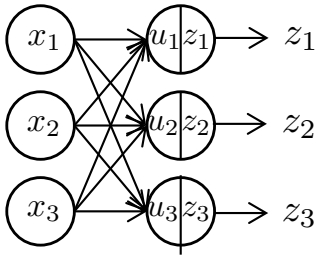
Zhang+, Truncating Wide Networks using Binary Tree Architectures, arXiv 2017

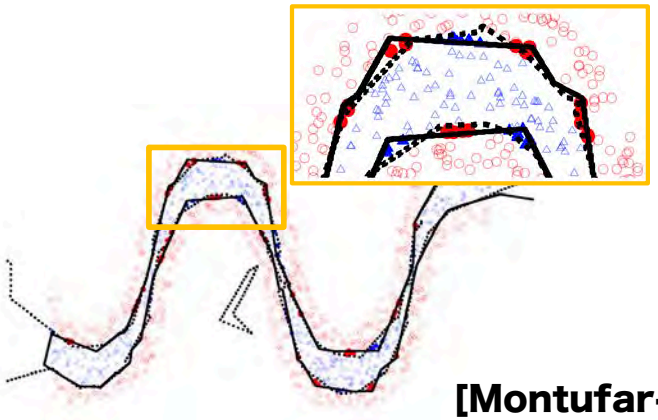
多層ネットの表現力 ＝入力空間の線形領域数：

$$\left(\prod_{i=1}^{L-1}\left\lfloor\frac{n_i}{n_0}\right\rfloor^{n_0}\right)\sum_{j=0}^{n_0}\binom{n_L}{j}$$

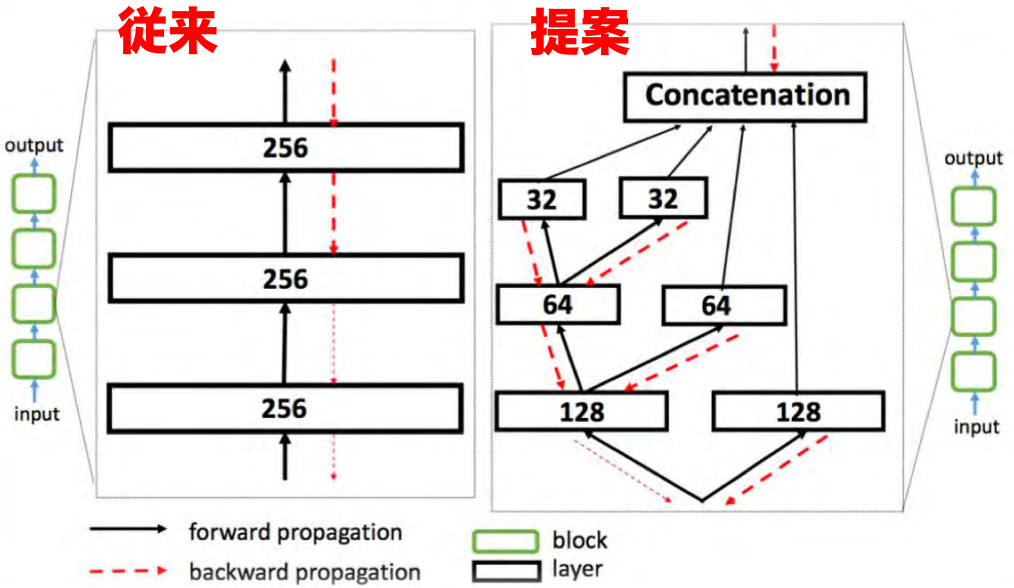








[Montufar+2014]

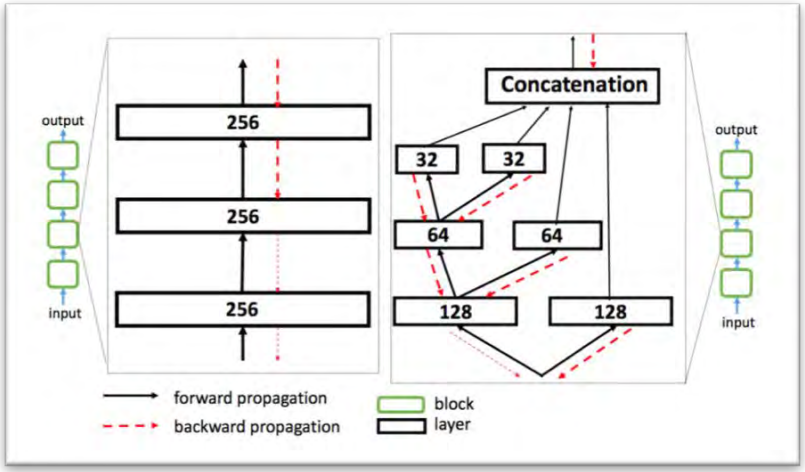


Proposition 3.3. *Suppose that we are given a BitNet stacked by L fully connected BitBlocks each having width D and depth K. The parameter size of the given BitNet is $\mathcal{O}\left(\frac{4}{3}L\left(1-\frac{1}{4^K}\right)D^2\right)$. The maximal number of linear regions of functions that can be computed by the given BitNet in an n-dimensional ($D \geq 2^K n$) input space is lower bounded by $\mathcal{O}\left(\left(\frac{D}{2^K n}\right)^{nKL}\right)$.* ■

小さくて高性能なネット

Zhang+, Truncating Wide Networks using Binary Tree Architectures, arXiv 2017

- １/3〜1/4のサイズで同等の性能
- 少ないパラメータで大きな表現力を持つ構造



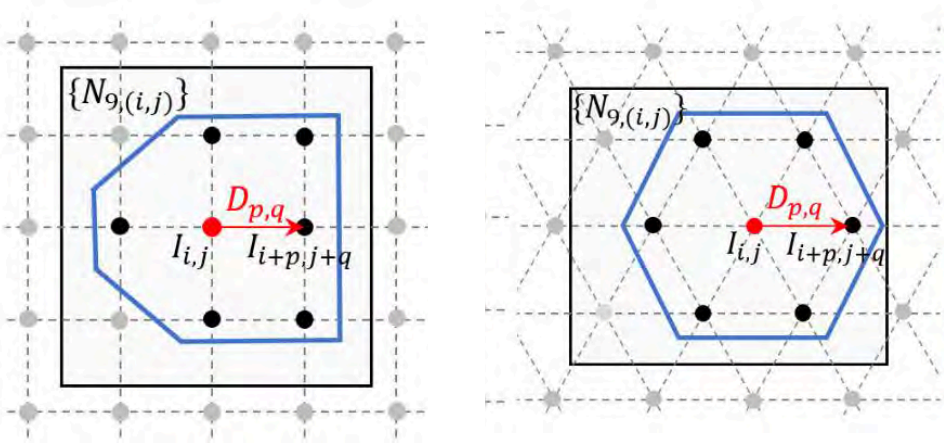
Model	Depth	Param	FLOP	Cifar10	Cifar100
NIN [17]	-	-	-	8.81	35.67
ELU [3]	-	-	-	6.55	24.28
DSN [16]	-	-	-	7.97	34.57
AllCNN [23]	-	-	-	7.25	33.71
ResNet [6]	1202	10.2M	-	4.91	–
preact-ResNet [7]	1001	10.2M	-	4.62	22.71
Stochastic Depth ResNet [9]	110	1.7M	-	5.25	24.98
FractalNet [15]	40	22.9M	-	5.24	22.49
Wide ResNet (d=4,k=2,n=6) [28]	38	8.9M	1.34×10^9	4.97	22.89
BitNet (d=4,k=3,n=4)	38	3.7M	0.53×10^9	4.82	22.19
BitNet (d=4,k=4,n=3)	38	2.7M	0.39×10^9	4.65	22.60
BitNet (d=4,k=2,n=6)	38	5.4M	0.78×10^9	5.31	23.22
BitNet (d=4,k=6,n=2)	38	1.7M	0.24×10^9	4.77	23.87
Wide ResNet (d=10,k=2,n=2) [28]	14	17.1M	2.64×10^9	4.56	21.59
BitNet (d=10,k=2,n=2)	14	9.6M	1.32×10^9	4.17	20.48
BitNet (d=10,k=4,n=1)	14	3.9M	0.49×10^9	4.97	23.88
Wide ResNet (d=10,k=2,n=3) [28]	20	26.8M	4.06×10^9	4.44	20.75
BitNet (d=10,k=2,n=3)	20	15.6M	2.21×10^9	**3.78**	19.29
BitNet (d=10,k=3,n=2)	20	10.2M	1.41×10^9	3.81	19.37
Wide ResNet (d=12,k=2,n=4) [28]	26	52.5M	7.87×10^9	4.33	20.43
BitNet (d=12,k=2,n=4)	26	31.2M	4.45×10^9	4.07	**19.06**
BitNet (d=12,k=4,n=2)	26	14.9M	2.06×10^9	4.11	19.22

26.8M	4.06×10^9	4.44	20.75
15.6M	2.21×10^9	**3.78**	19.29
10.2M	1.41×10^9	3.81	19.37
52.5M	7.87×10^9	4.33	20.43
31.2M	4.45×10^9	4.07	**19.06**
14.9M	2.06×10^9	4.11	19.22

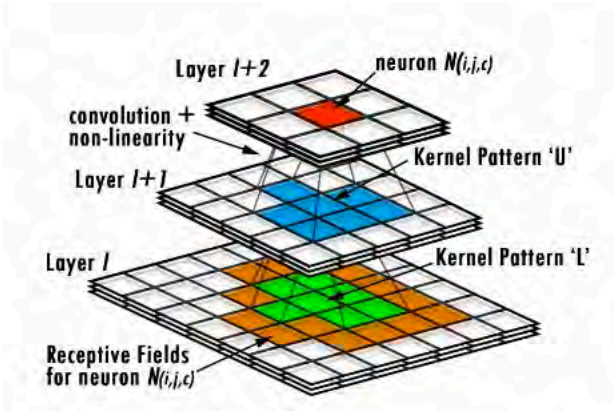
Quasi-hexagonal kernel

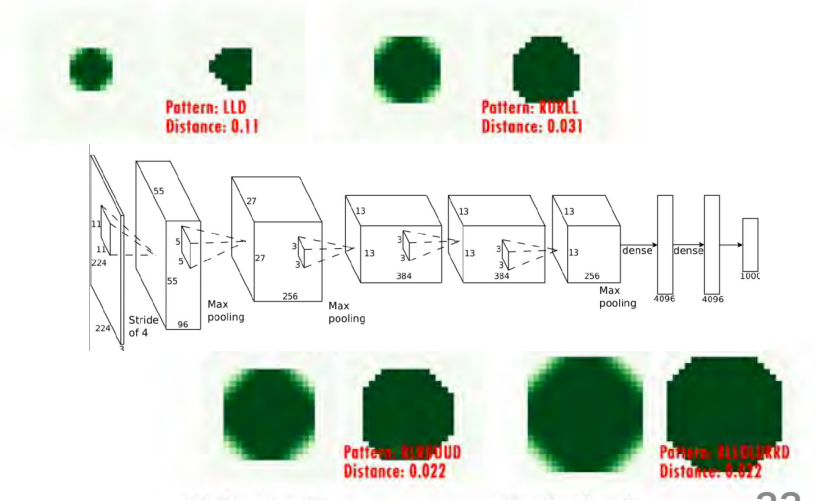
Sun+, Design of Kernels in Convolutional Neural Networks for Image Classification, ECCV2016

- **Use of 凸-shape filters (quasi-hexagonal kernels)**



- **Different orientation for each layer**
  - **More flexible receptive field**



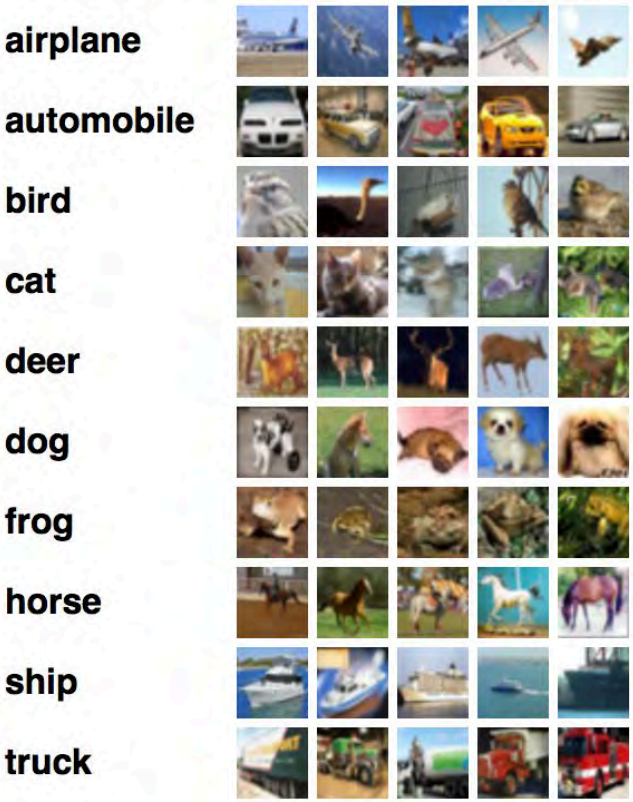


Quasi-hexagonal kernel

Sun+, Design of Kernels in Convolutional Neural Networks for Image Classification, ECCV2016

- **Change of filter shapes alone improves accuracy for CIFAR-10/100**

Model	Testing Error (%) CIFAR-10	CIFAR-100	# of Params.
NIN [17]	10.41	35.68	$\approx 1M$
DSN [15]	9.69	34.57	$\approx 1M$
ALL-CNN [30]	9.08	33.71	$\approx 1.4M$
RCNN [16]	8.69	31.75	$\approx 1.9M$
Spectral pool [23]	8.6	31.6	–
FMP [4]	–	31.2	$\approx 12M$
BASE-A-AD	8.71	31.2	$\approx 1.4M$
QH-B-AD	8.54	30.54	$\approx 1.4M$
QH-C-AD	**8.42**	**29.77**	$\approx 2.4M$

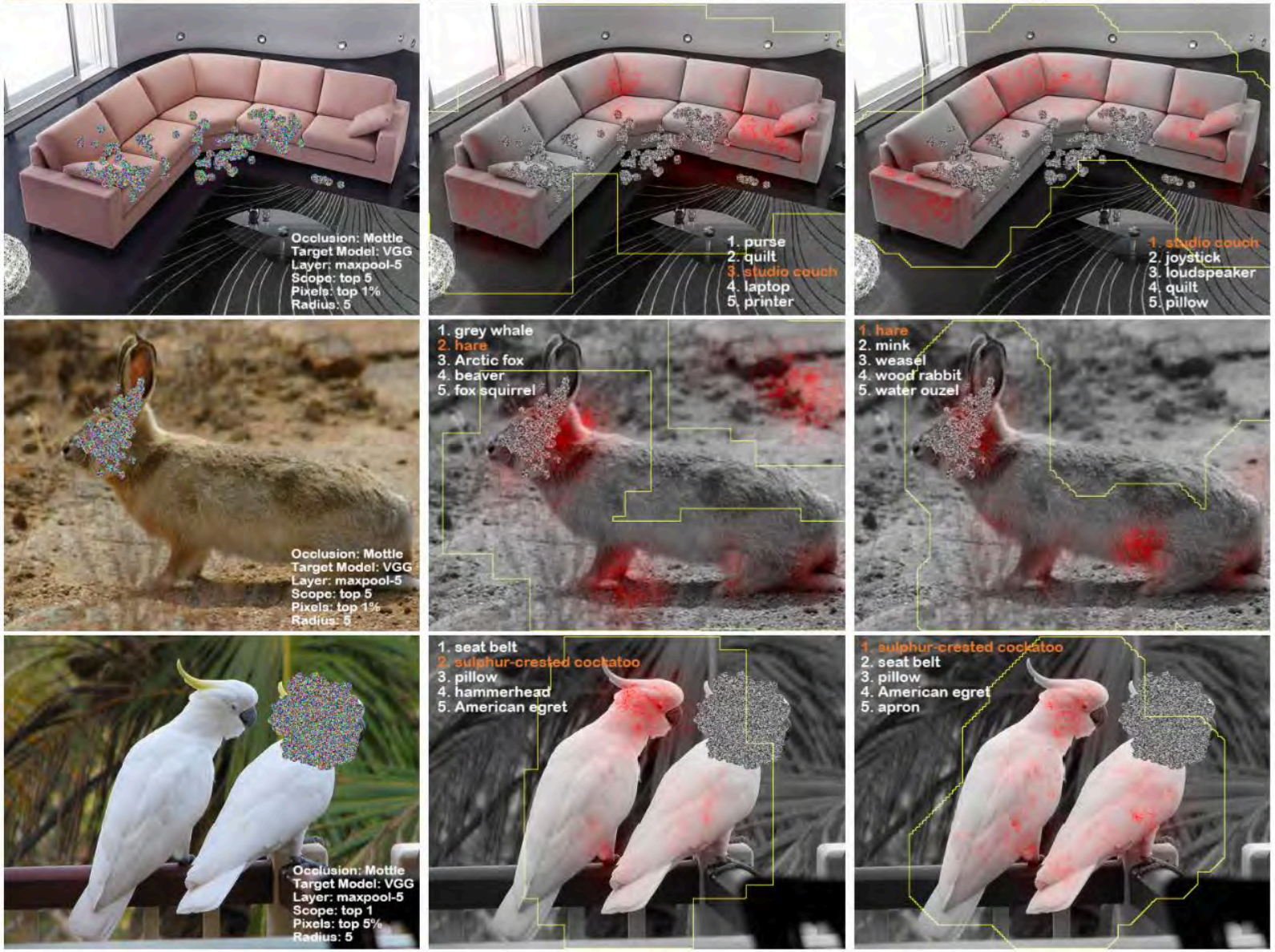


ILSVRC2012 (large scale object recognition)

Model	BASE	QH-BASE	REF-A-BASE	REF-B-BASE
top-1/5 val.error (%)	31.2/12.3	**29.2/11.1**	31.4/12.4	31.2/12.2

Quasi-hexagonal kernel

Sun+, Design of Kernels in Convolutional Neural Networks for Image Classification, ECCV2016



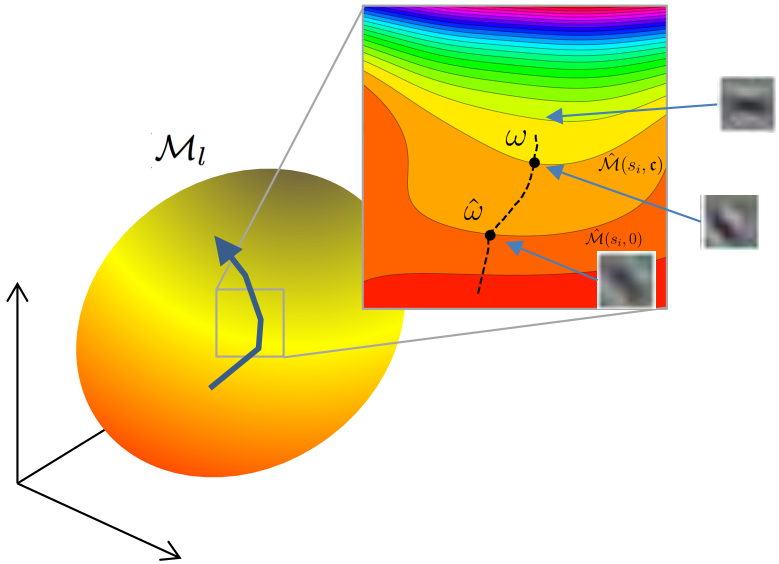
Occluded Image **BASE** **QH-BASE**

畳込みフィルタの多様体制約付き学習

Ozay-Okatani, arXiv2016 / Ozay-Hatsutani-Okatani, arXiv2017

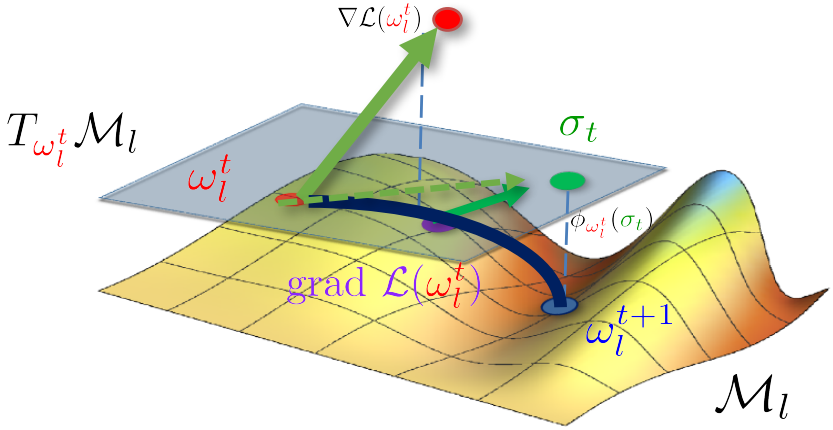
- **フィルタの制約＝行列多様体**

Sphere:	$\|W\|_F^2 = 1$
Oblique:	$\mathrm{ddiag}(W^\top W) = I$
Stiefel: (Orthogonal)	$W^\top W = I$



- **同一CNN（ResNet-110）での結果**

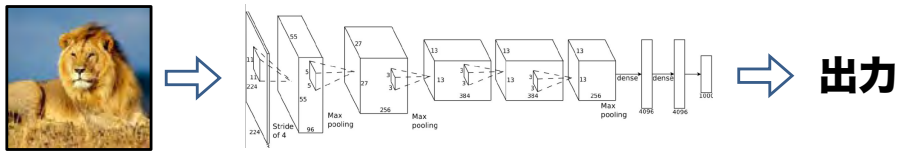
	Cifar10	Cifar100
RCD [Huang+2016]	6.41	27.22
Ours (One manifold)	5.91	25.39
Ours (Product of four manifolds)	5.17	23.79
RSD [Huang+2016]	5.23	24.58
Ours (One manifold)	4.73	23.09
Ours (Product of four manifolds)	4.31	22.03



Weights (filter kernels) are updated on a manifold.

まとめ（画像分野の研究の現状）

- ディープラーニングにより多くの問題が解決へ
  - 大量の入出力ペアを用いたend-to-end学習



- 残された問題

1. 真の画像理解
2. 出力がはっきりしない
3. データが集まらない
4. サイバースペースから実世界へ
5. DNN(CNN)の理解