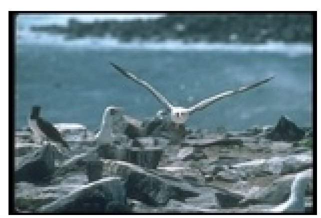
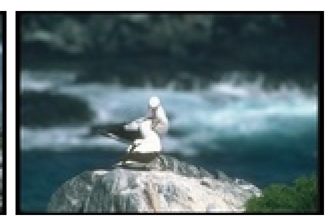
情報・システム工学概論
画像・映像認識のモデル化

機械情報工学科（機械B）
原田達也

実世界認識知能の構築





人と調和する情報機器の創出→
人の生活する実世界と情報世界の間に存在するギャップを埋めることが重要

画像アノテーション結果

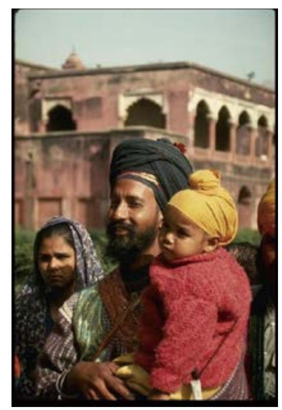
birds, booby, flight, rocks, water

buildings, ships, bridge, flag, sky

church, stone, buildings, chapel, people

sky, people, close-up, statue, clouds

buildings, water, city, light, night



people, woman, indian, pots, baby

cat, tiger, water, rocks, forest

一般的な視覚認識機能の困難さ

- 人の認識の曖昧性: weak labeling

Jet plane sky

cat tiger forest tree

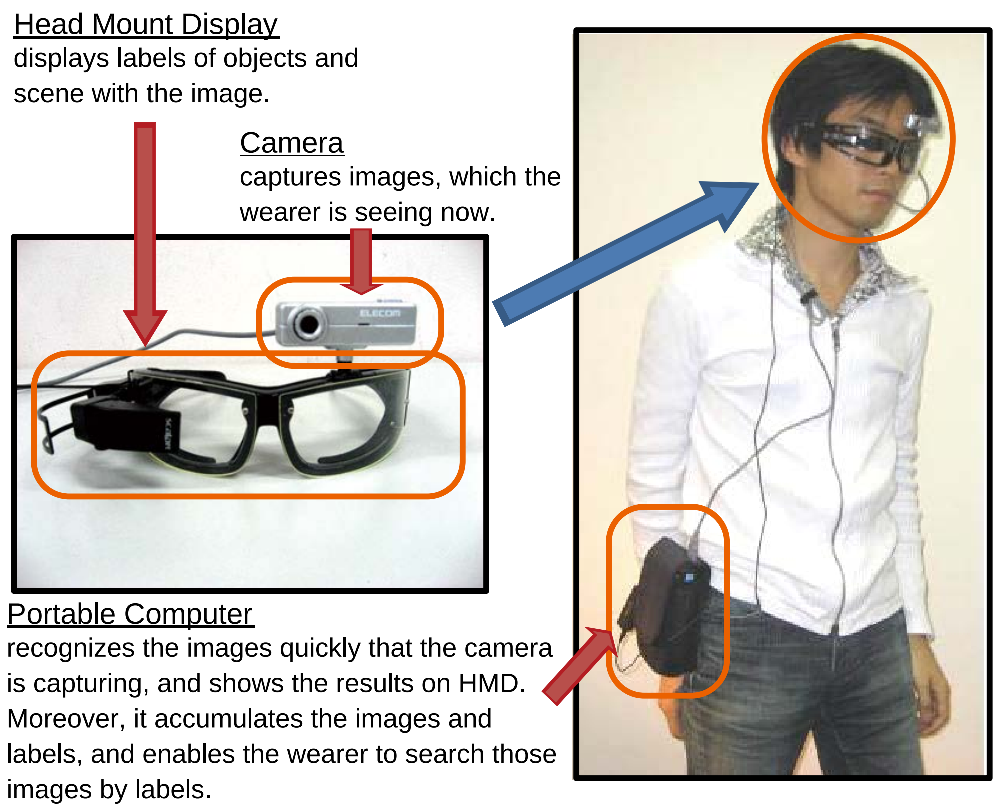
beach people water oahu

- 文脈の考慮

- Data drivenの特徴と意味との不一致: semantic gap
- 大量の学習データへのスケーラビリティ
- 多様な環境への対応：高速かつ安定な追加学習

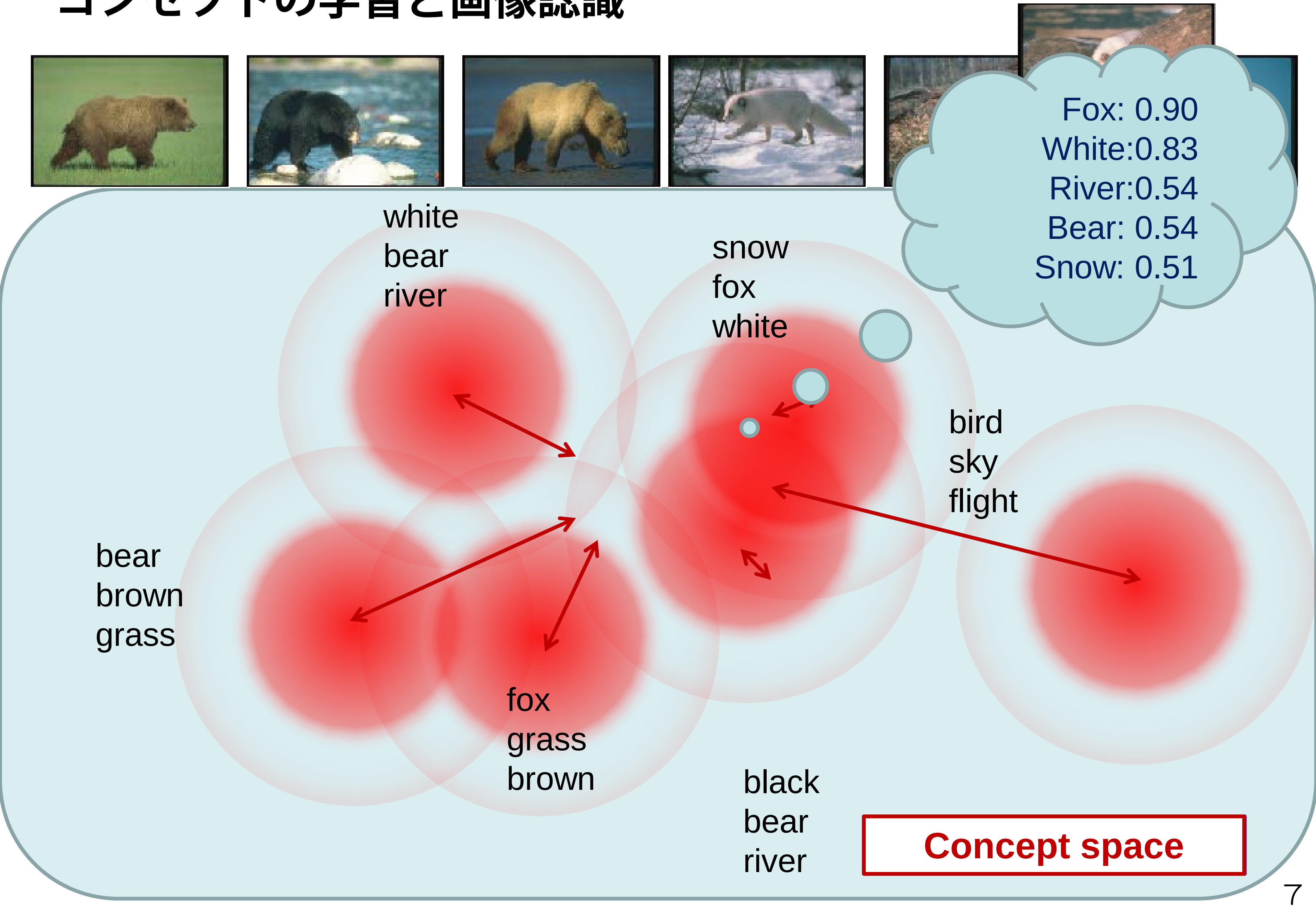
実世界応用１
人工知能ゴーグルの開発

- 提案手法の実世界応用：人工知能ゴーグル
  - 身の回りの物体の素早い認識・検索を実現
  - HMDによる情報提示，記憶支援（忘れ物検索）



AI Goggles
実世界におけるリアルタイムアノテーション

コンセプトの学習と画像認識



Large Scale Object Recognition

- ILSVRC (ImageNet Large Scale Visual Recognition Challenge)
  - Image recognition competition using large scale images
  - http://www.image-net.org/challenges/LSVRC/2012/index

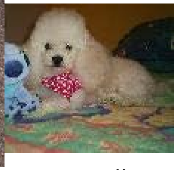
- Task 1: What's this image?
  - Learning 1.2 million images
  - Classifying 1000 object classes

- Task 2: Where's this object?
  - Detecting 1000 object classes in images

- Task 3: What kind of dog is this?
  - Fine-grained classification on 120 dog sub-classes
  - More difficult to classify objects than task 1

Shih-Tzu

Pomeranian



toy poodle

Deep CNN!

Task 1

Team	Flat Error
1) SuperVision Univ. of Toronto	0.153
2) ISI (ours) Ours Univ. of Tokyo	0.262
3) OXFORD_VGG Univ. of Oxford	0.270

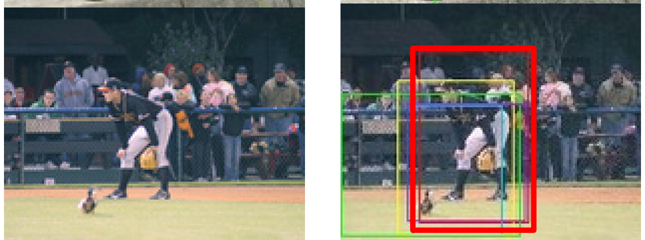
Task 3

Team	mAP
1) ISI (ours) Ours Univ. of Tokyo	0.323
2) XRCE/INRIA Xerox Research Centre Europe/INRIA	0.310
3) Uni Jena Univ. Jena	0.246

Results (2012)

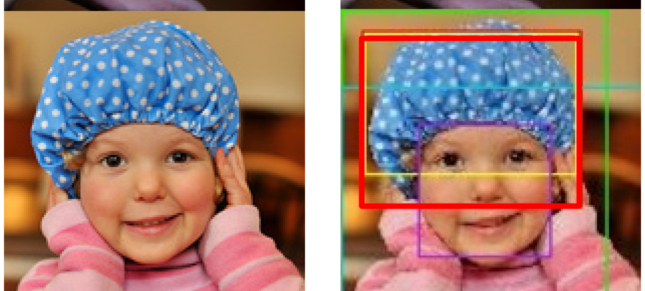
http://www.isi.imi.i.u-tokyo.ac.jp/pattern/ilsvrc2012/index.html

1. brown bear
2. Tibetan mastiff
3. sloth bear
4. American black bear
5. bison



1. baseball player
2. unicycle
3. racket
4. rugby ball
5. basketball

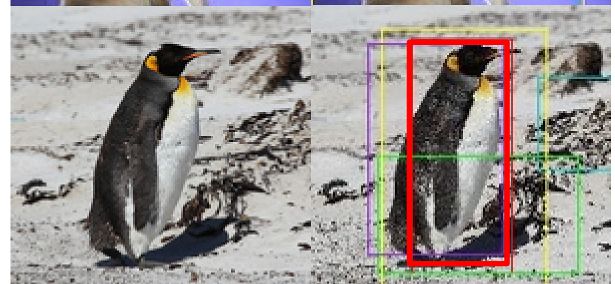
1. digital watch
2. Band Aid
3. syringe
4. slide rule
5. rubber eraser



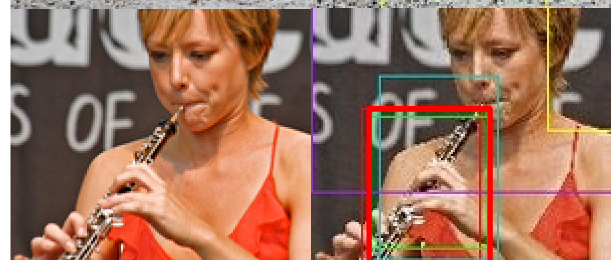
1. shower cap
2. bonnet
3. bath towel
4. bathing cap
5. ping-pong ball

1. diaper
2. swimming trunks
3. bikini
4. miniskirt
5. cello

1. Siamese cat
2. Egyptian cat
3. Ibizan hound
4. balance beam
5. basenji



1. king penguin
2. sea lion
3. drake
4. magpie
5. oystercatcher

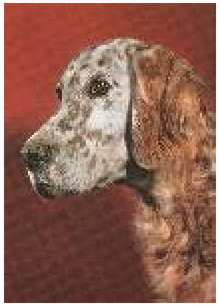


1. oboe
2. flute
3. ice lolly
4. bassoon
5. cello

1. beer bottle
2. pop bottle
3. wine bottle
4. Polaroid camera
5. microwave

1. butcher shop
2. swimming trunks
3. miniskirt
4. barbell
5. feather boa

Fine-grained object recognition results (2012)


English setter

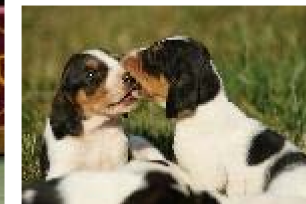

Siberian husky


Australian terrier


English springer


malamute


Great Dane


Walker hound

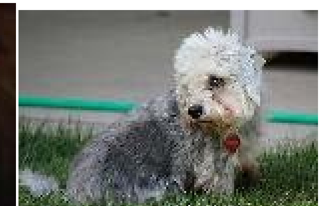

Welsh springer spaniel


whippet


Scottish deerhound


Weimaraner


soft-coated wheaten terrier


Dandie Dinmont


Old English sheepdog


otterhound


bloodhound


Airedale


giant schnauzer


black-and-tan coonhound


papillon


Staffordshire bullterrier


Mexican hairless


Bouvier des Flandres


miniature poodle


Cardigan


malinois

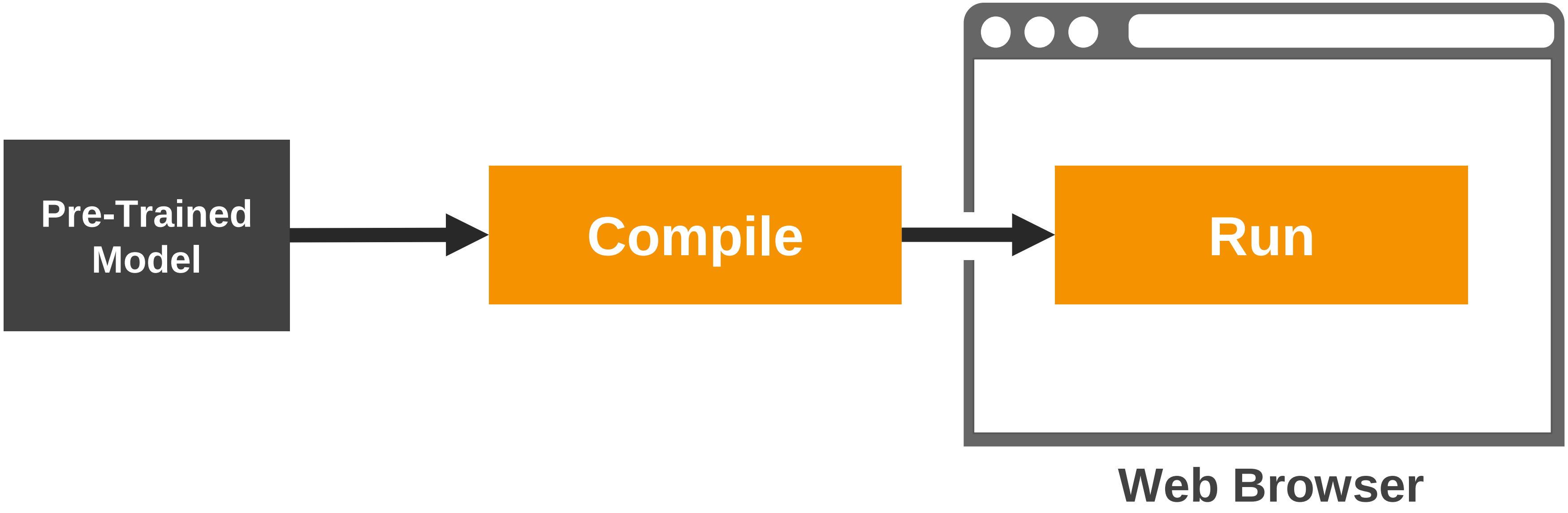
WebDNN:

Fastest DNN Framework on Web Browser

https://mil-tokyo.github.io/webdnn/

M. Hidaka, Y. Kikura, Y. Ushiku, T. Harada. WebDNN: Fastest DNN Execution Framework on Web Browser. ACM Multimedia Open Source Software Competition, 2017. Honorable Mention Open source software Award.

No need to install any applications and libraries in your smartphone and laptop

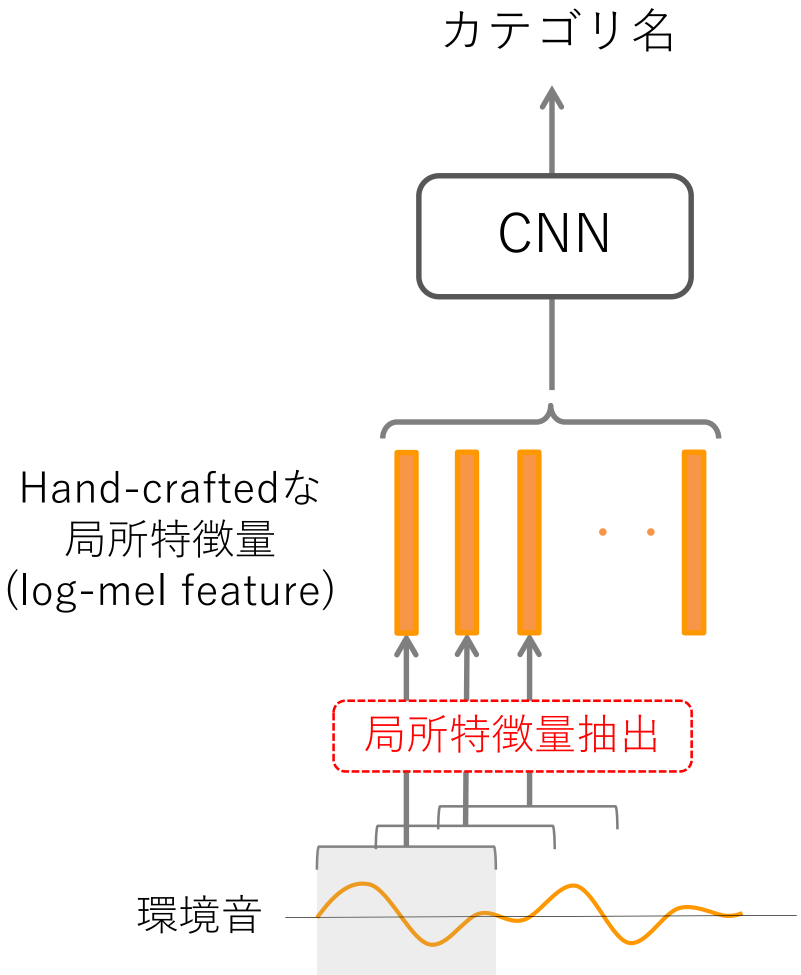


- WebDNN compile and optimize pretrained model to execute on web browser
- Tensorflow, Keras model, Caffe model, Chainer chain is supported
- Dynamic parameters (e.g. sequence length in RNN) is also supported

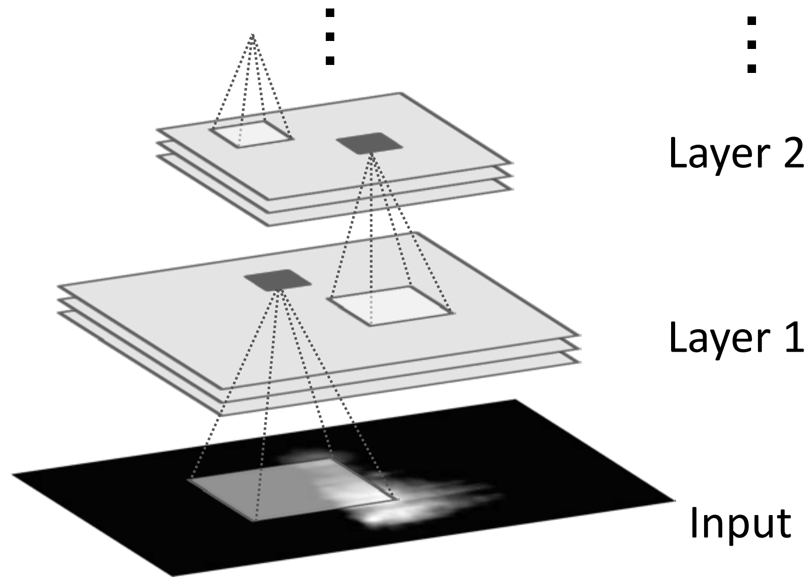
Sound Recognition

環境音識別手法

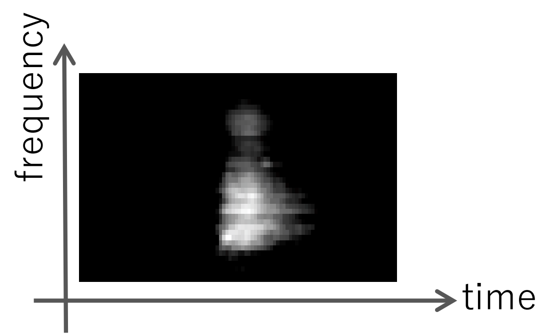
[Piczak, 2015]



画像化



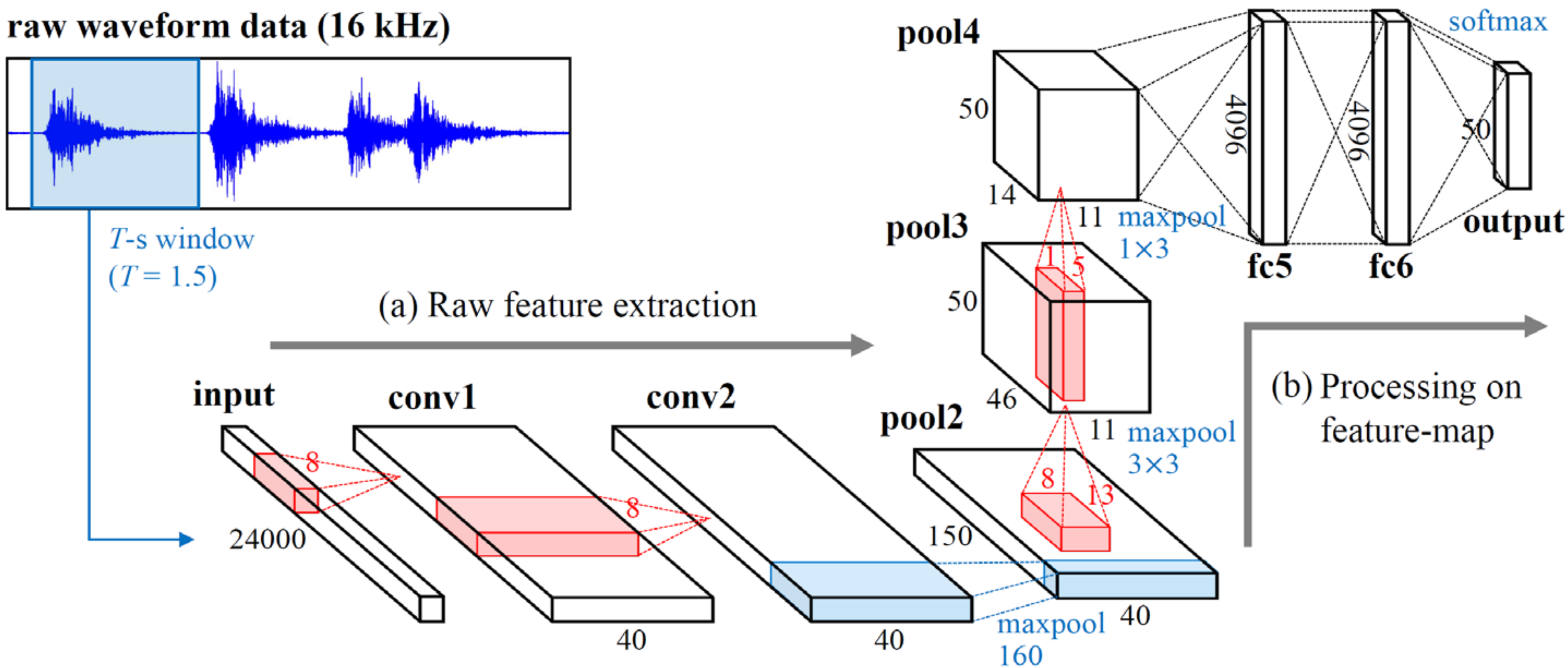
画像のように形状を持つので
CNNで識別できる



例: 犬の鳴き声

EnvNet

Yuji Tokozume and Tatsuya Harada. ICASSP, accepted, 2017



エンドツーエンドで学習可能な環境音モデル

実験結果

Yuji Tokozume and Tatsuya Harada. ICASSP, accepted, 2017

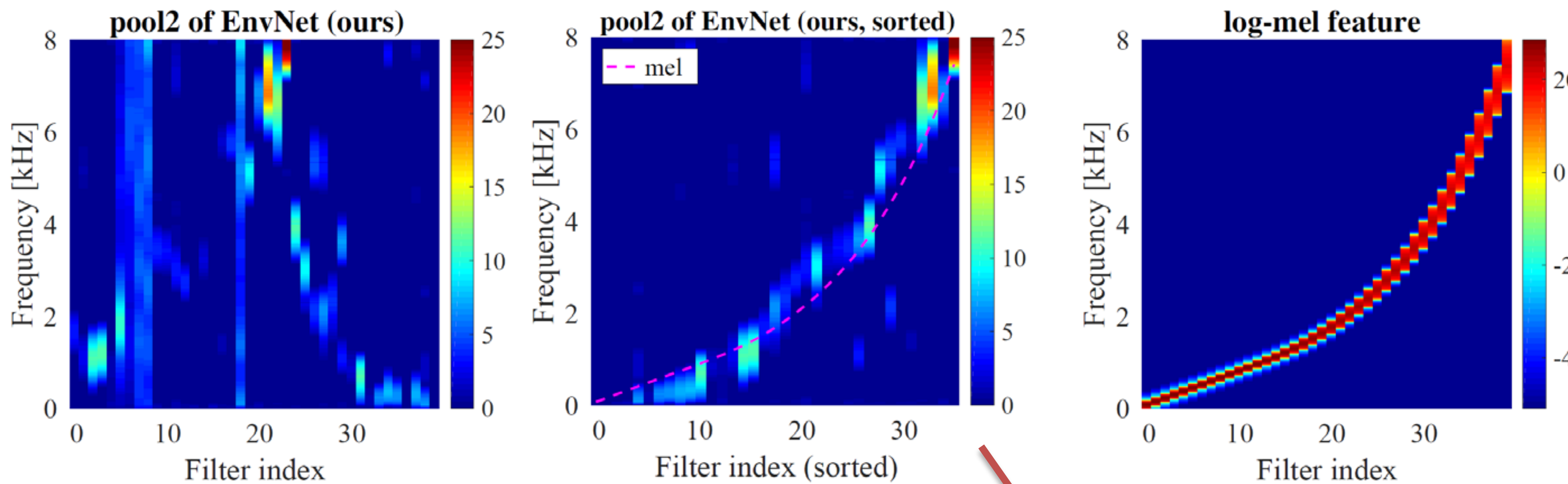
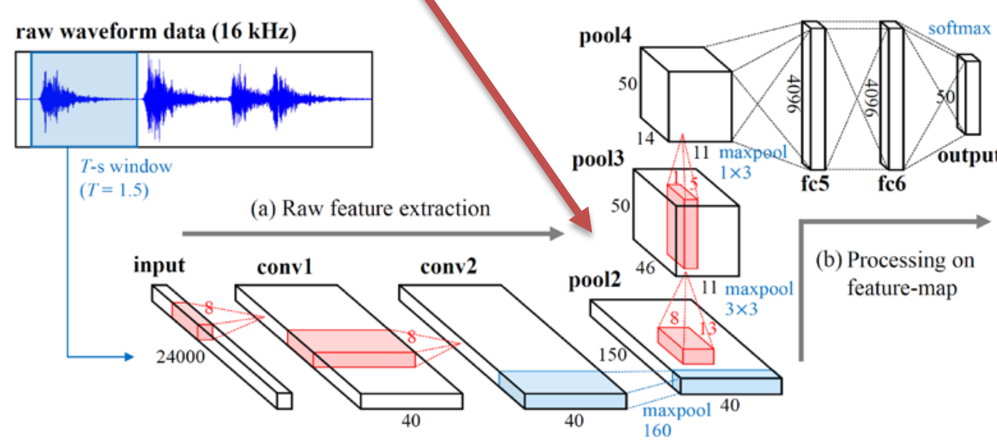
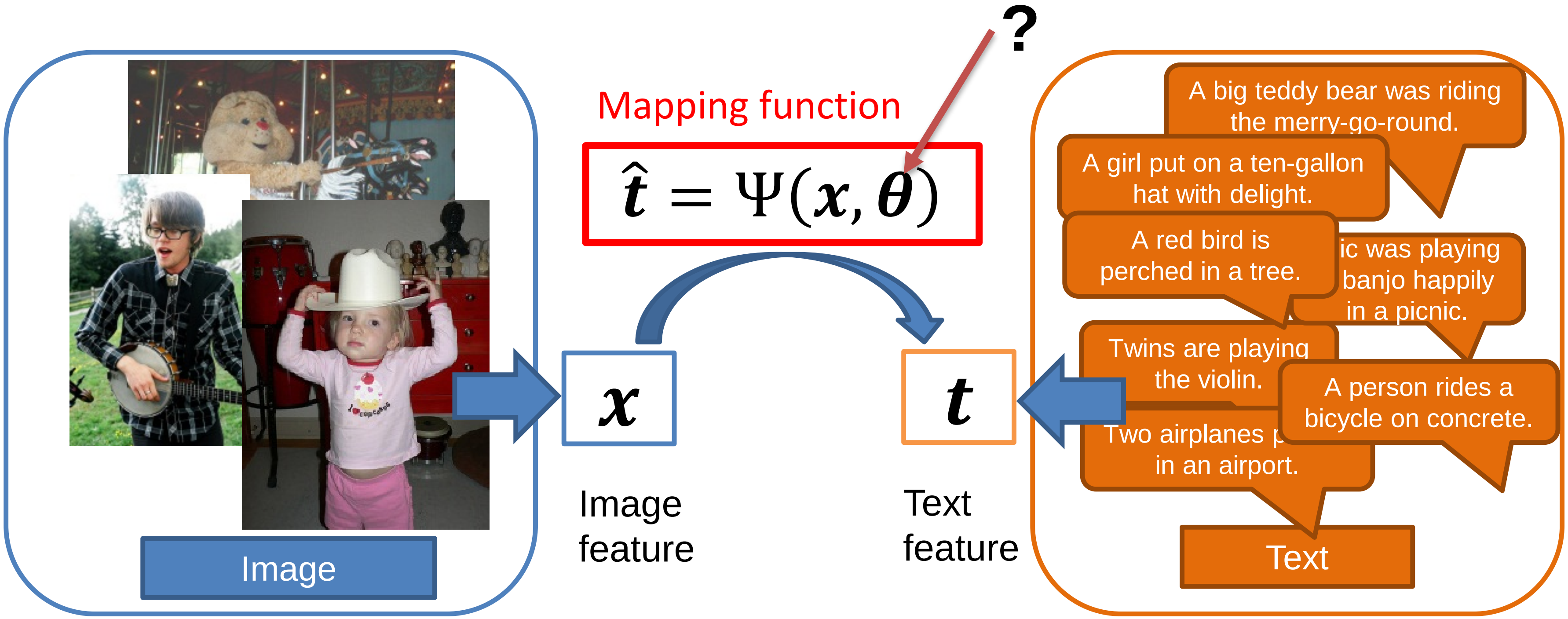


Table 1. Comparison and combination with logmel-CNN. The error in this table means the standard deviation among the accuracies for the five-fold cross-validation.

logmel-CNN			
static	delta	EnvNet (ours)	Accuracy [%]
✓			58.9 ± 2.6
✓	✓		66.5 ± 2.8
		✓	64.0 ± 2.4
✓		✓	69.3 ± 2.2
✓	✓	✓	71.0 ± 3.1
Piczak logmel-CNN [4]			64.5
Human [1]			81.3



Overview: Machine Learning for Visual Recognition



Learning the relationships between images and text

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \underbrace{r(\boldsymbol{\theta})}_{\text{Risk}}$$

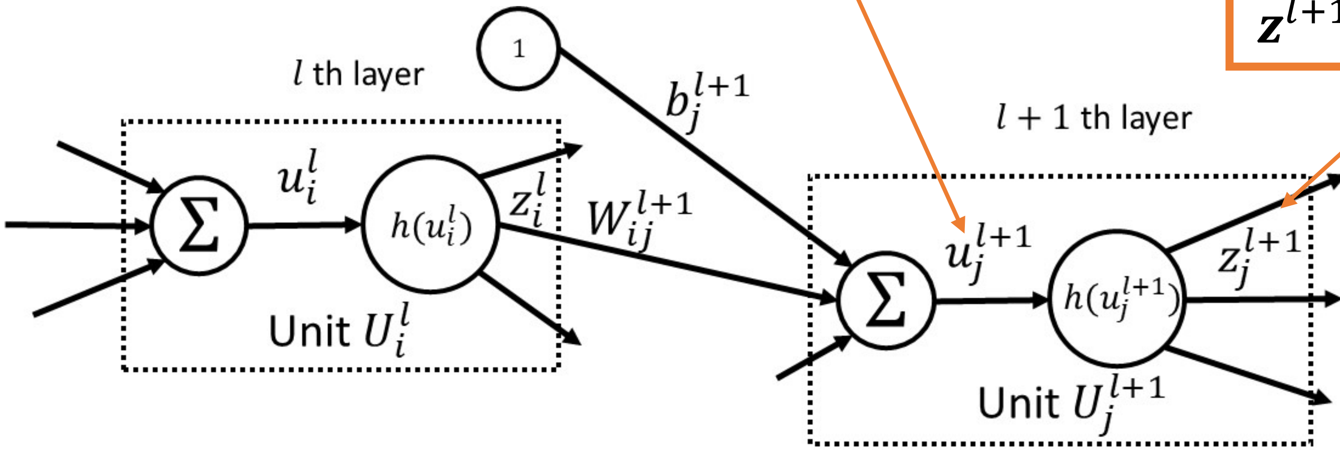
$$r(\boldsymbol{\theta}) = \mathbb{E}\left[\underbrace{l(\Psi(\boldsymbol{x}, \boldsymbol{\theta})|\boldsymbol{t})}_{\text{Loss function}}\right]$$

Deep Neural Networks

$\boldsymbol{u}^{l+1} \in \mathbb{R}^{|U^{l+1}|}, \boldsymbol{z}^{l} \in \mathbb{R}^{|U^{l}|}$

$\boxed{\boldsymbol{u}^{l+1} = (W^{l+1})^T \boldsymbol{z}^l + \boldsymbol{b}^{l+1}}$

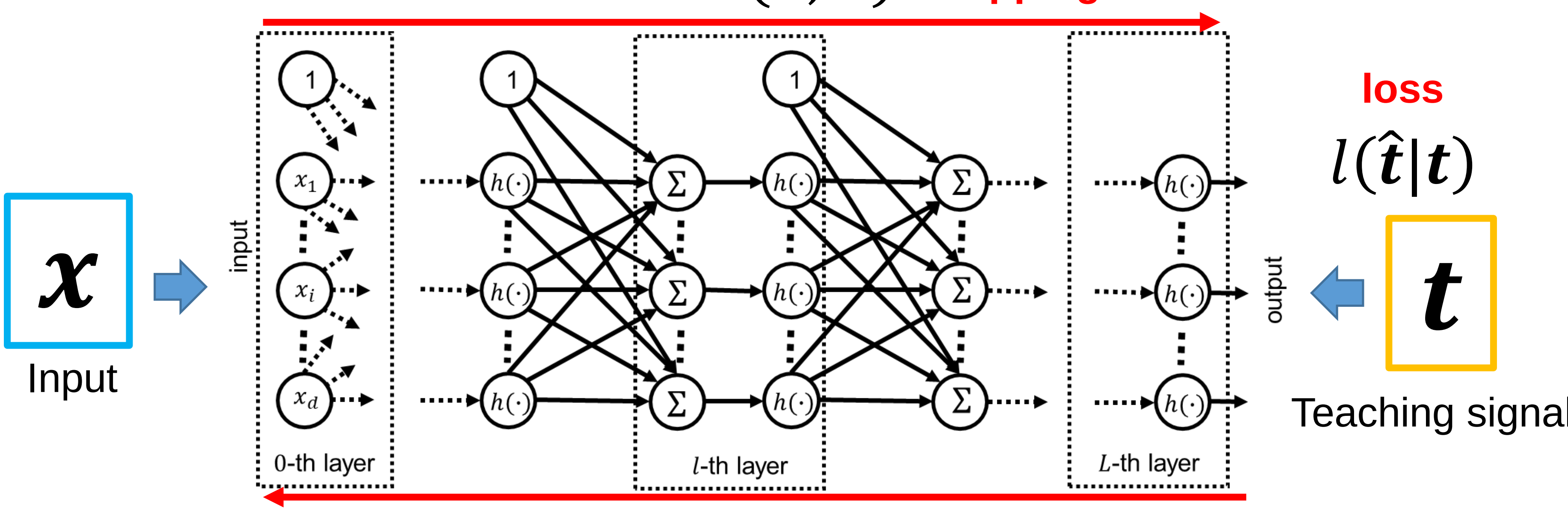
$\boxed{\boldsymbol{z}^{l+1} = h^l(\boldsymbol{u}^{l+1})}$



$\hat{\boldsymbol{t}} = \Psi(\boldsymbol{x}, \boldsymbol{\theta})$ **mapping**

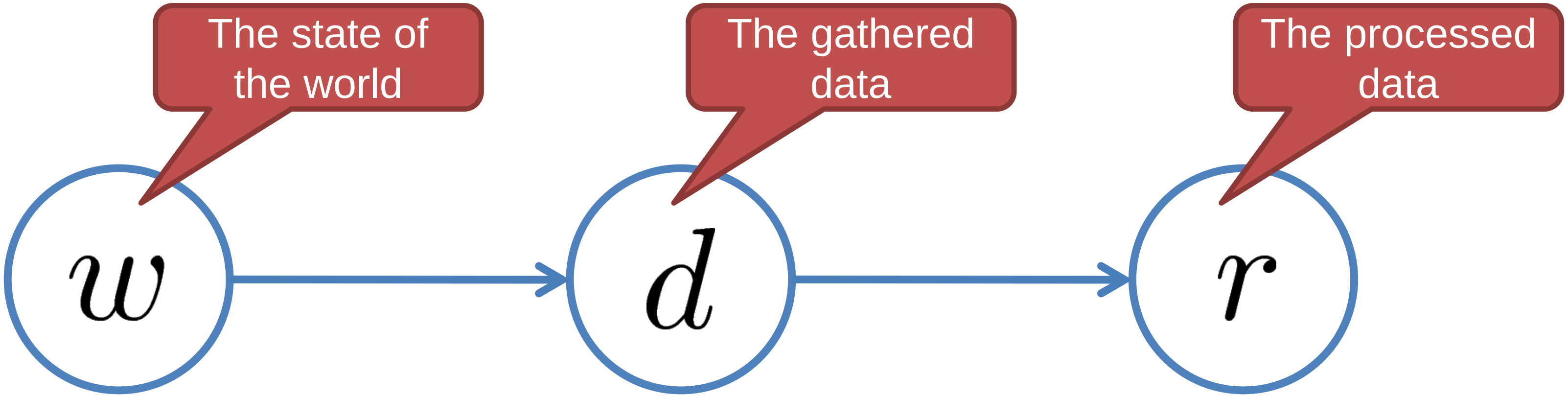
loss

$l(\hat{\boldsymbol{t}}|\boldsymbol{t})$



$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \epsilon C \nabla_w l(\Psi(\boldsymbol{x}, \boldsymbol{\theta}_t)|\boldsymbol{t})$ **back propagation**

The data processing theorem



Markov chain $P(w,d,r) = P(w)P(d\,|\,w)P(r\,|\,d)$

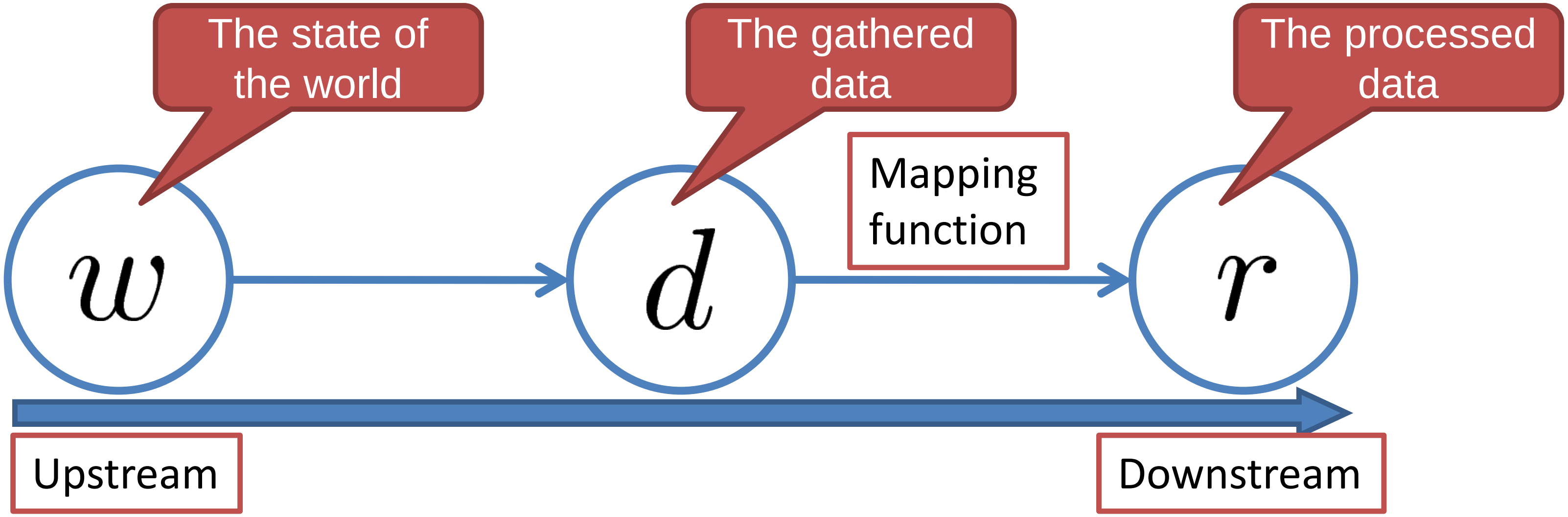
The average information

$$I(w;d) \geq I(w;r)$$

The data processing theorem states that data processing can only destroy information.

David J.C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press 2003.

The data processing theorem



Markov chain $P(w, d, r) = P(w)P(d \mid w)P(r \mid d)$

The average information

$$I(w; d) \geq I(w; r)$$

The data processing theorem states that data processing can only destroy information.

David J.C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press 2003.

Caltech-101

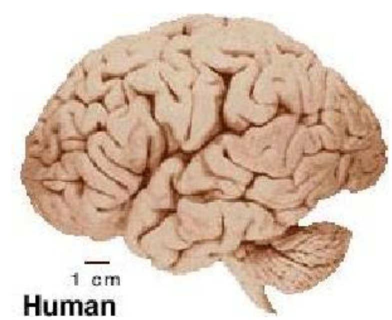
http://www.vision.caltech.edu/Image_Datasets/Caltech101/

- Pictures of objects belonging to 101 categories.
- About 40 to 800 images per category.
- Most categories have about 50 images. Collected in September 2003 by Fei-Fei Li, Marco Andreetto, and Marc 'Aurelio Ranzato.
- The size of each image is roughly 300 x 200 pixels.

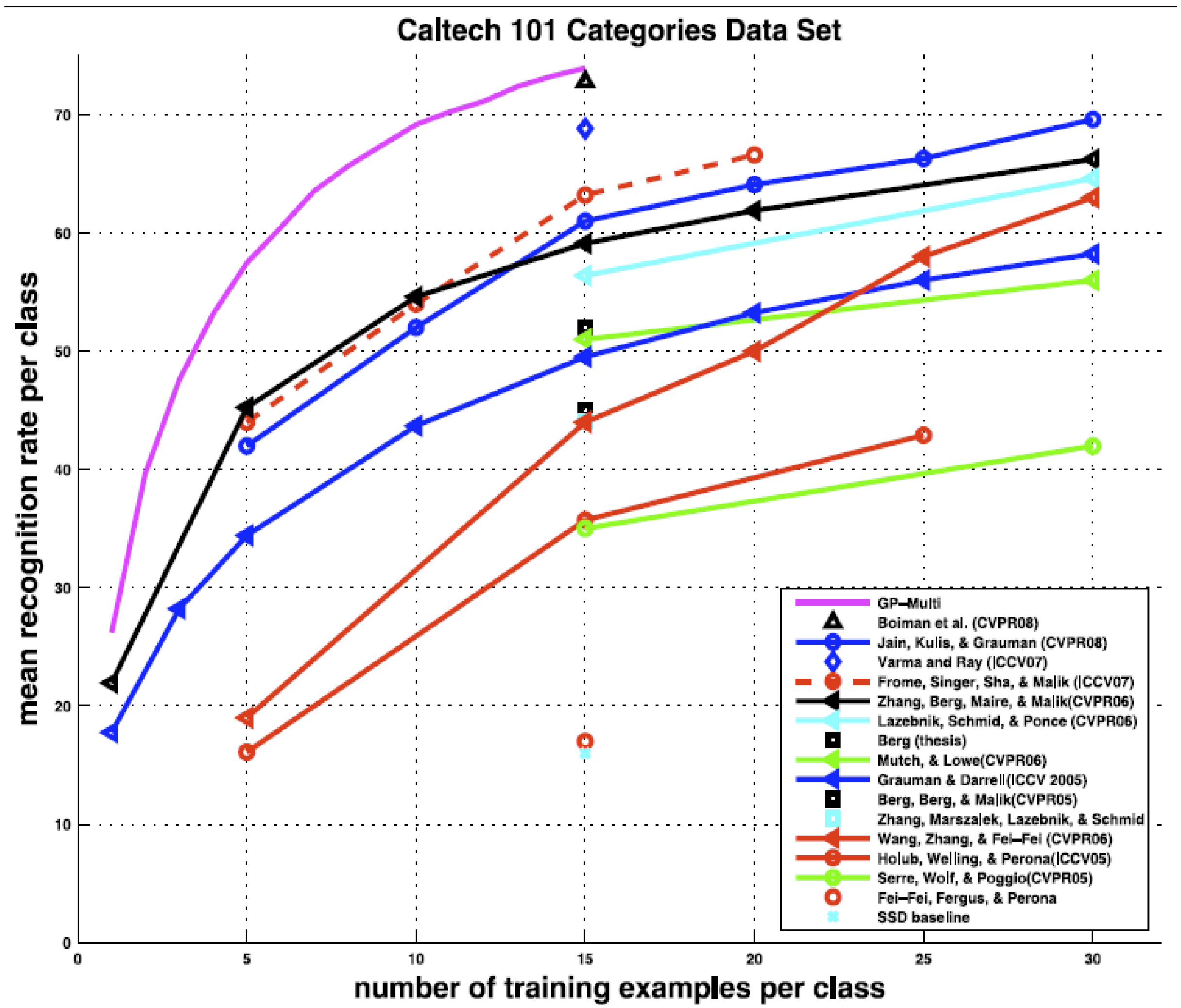






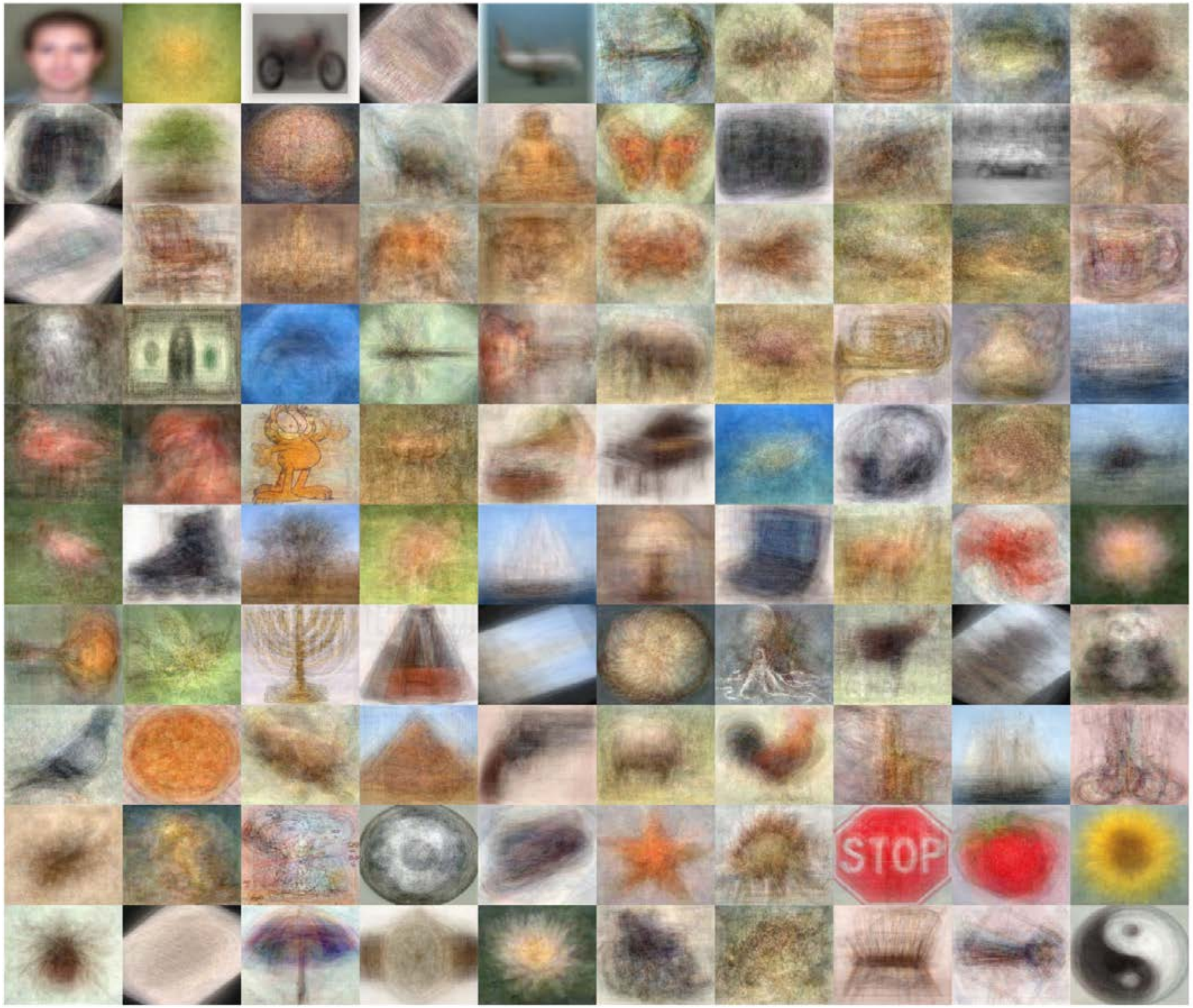


Recognition Rate on Caltech101 (2004-2008)



Gaussian Processes for Object Categorization. A. Kapoor, K. Grauman, R. Uratsun, and T. Darrell. In International Journal of Computer Vision (IJCV), Vol. 88, No. 2, 2010.

Dataset Bias



http://www.vision.caltech.edu/Image_Datasets/Caltech101/averages100objects.jpg

The rise of the modern dataset

Antonio Torralba, Alexei A. Efros. Unbiased Look at Dataset Bias. CVPR, 2011.

Development of dataset: a reaction against the biases and inadequacies of the previous datasets in explaining the visual world

- COIL-100 dataset
  - a reaction against model-based thinking of the time
  - an embrace of data-driven appearance models that could capture textured objects

- 15 Scenes dataset, Corel Stock Photo
  - a reaction against the simple COIL-like backgrounds
  - an embrace of visual complexity

- Caltech101
  - partially a reaction against the professionalism of Corel's photos
  - an embrace of the wilderness of the Internet

- MSRC, LabelMe
  - a reaction against the Caltech-like single-object-in-the-center mentality
  - the embrace of complex scenes with many objects

- PASCAL VOC
  - a reaction against the lax training and testing standards of previous datasets

- Tiny Images, ImageNet, SUN09
  - a reaction against the inadequacies of training and testing on datasets that are just too small for the complexity of the real world

TinyImages

- A. Torralba, R. Fergus, W. T. Freeman. 80 million tiny images: a large dataset for non-parametric object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.30(11), pp. 1958-1970, 2008.

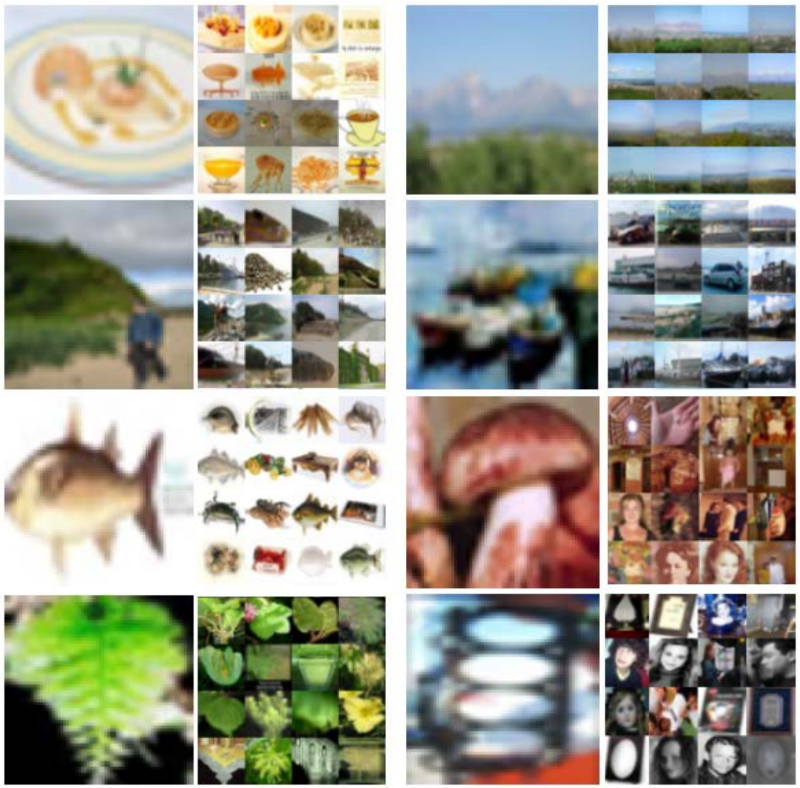
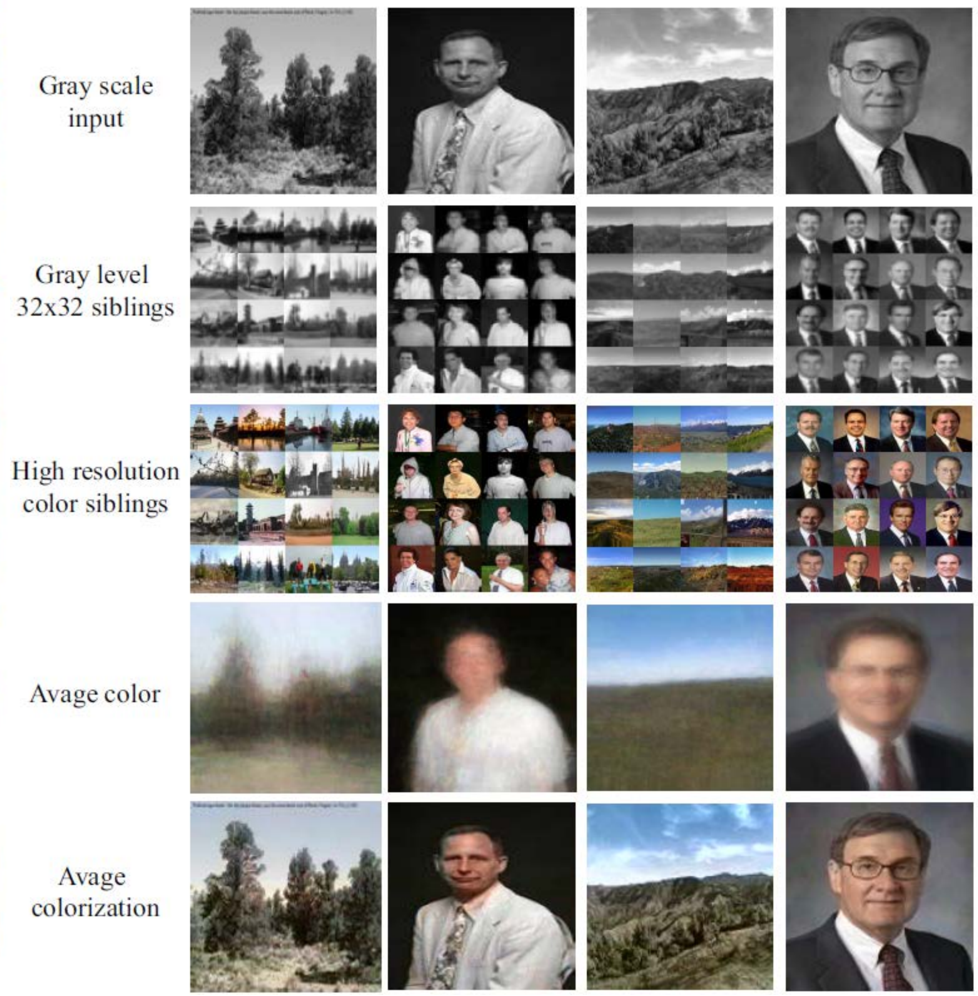


Fig. 1. 1st & 3rd columns: Eight 32×32 resolution color images. Despite their low resolution, it is still possible to recognize most of the objects and scenes. These are samples from a large dataset of 10^8 32×32 images we collected from the web which spans all visual object classes. 2nd & 4th columns: Collages showing the 16 nearest neighbors within the dataset to each image in the adjacent column. Note the consistency between the neighbors and the query image, having related objects in similar spatial arrangements. The power of the approach comes from the copious amount of data, rather than sophisticated matching methods.



ImageNet

- ImageNet
  - 12 million images, 15 thousand categories
  - Image found via web searches for WordNet noun synsets
  - Hand verified using Mechanical Turk
- WordNet
  - Source of fraction of English nouns
  - Also used the labels
  - Semantic hierarchy
  - Contains large o collect other datasets like tiny images (Torralba et al)
  - Note that categorization is not the end goal, but should provide information for other tasks, so idiosyncrasies of WordNet may be less critical

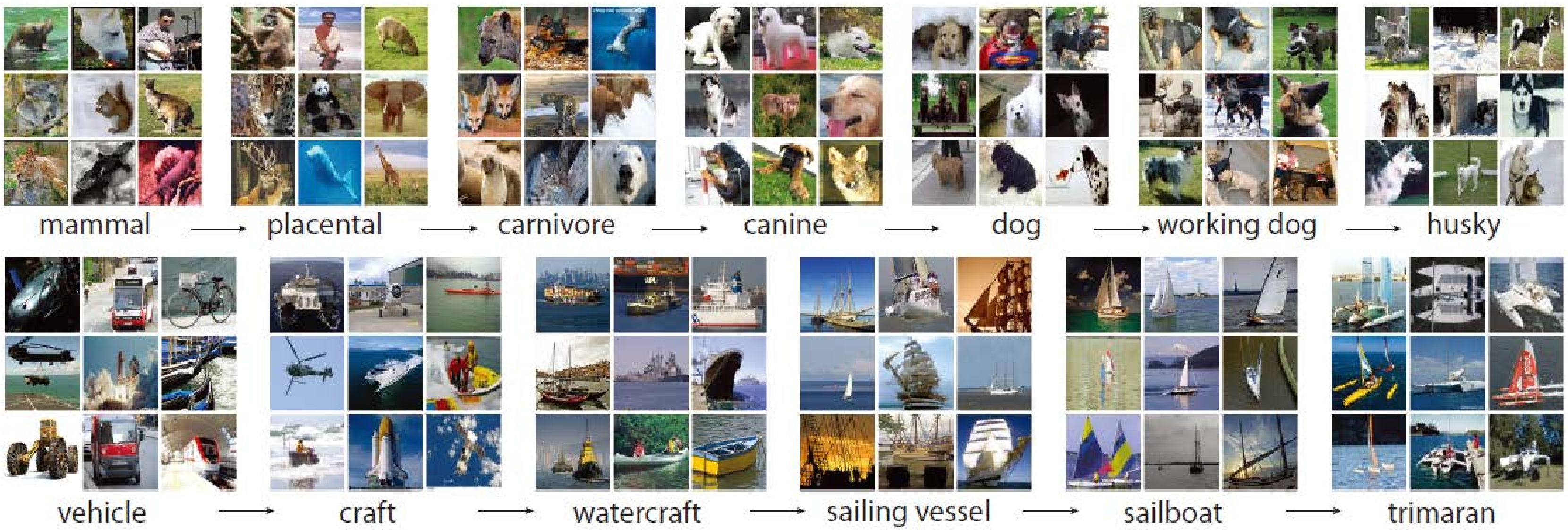
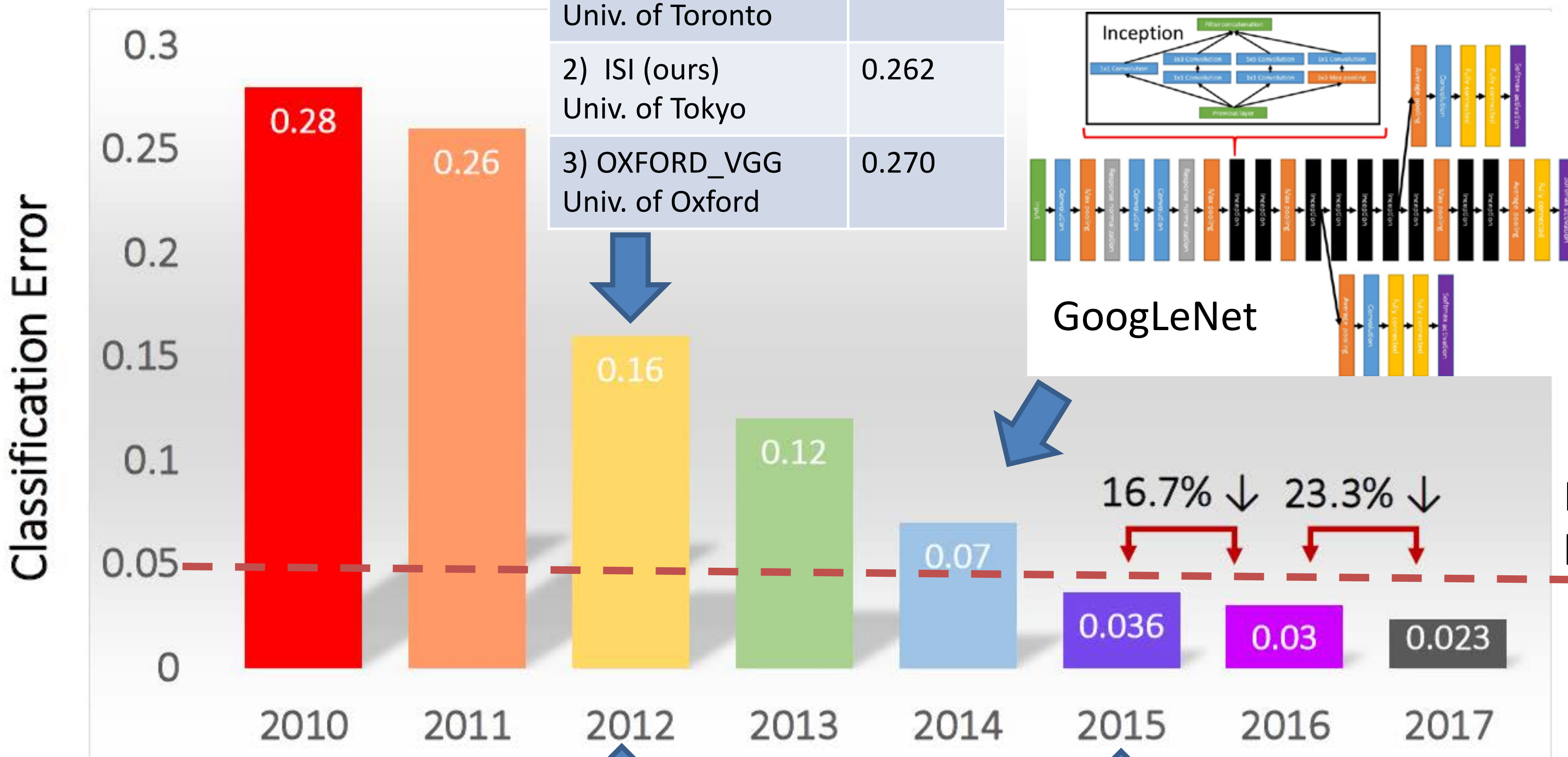


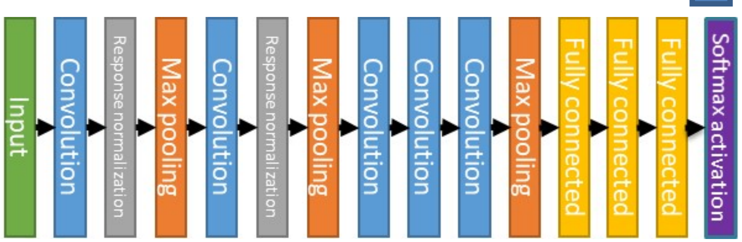
Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the **top** row is from the mammal subtree; the **bottom** row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.

Deng et al., CVPR2009

ILSVRC (Large Scale Visual Recognition Challenge)

Team	Flat Error
1) SuperVision Univ. of Toronto	0.153
2) ISI (ours) Univ. of Tokyo	0.262
3) OXFORD_VGG Univ. of Oxford	0.270

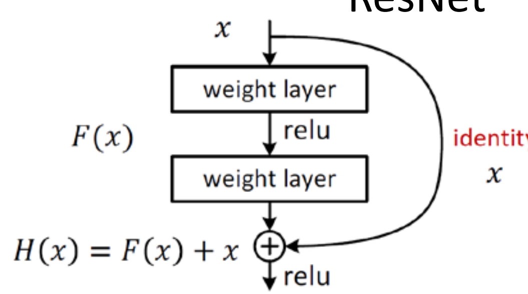




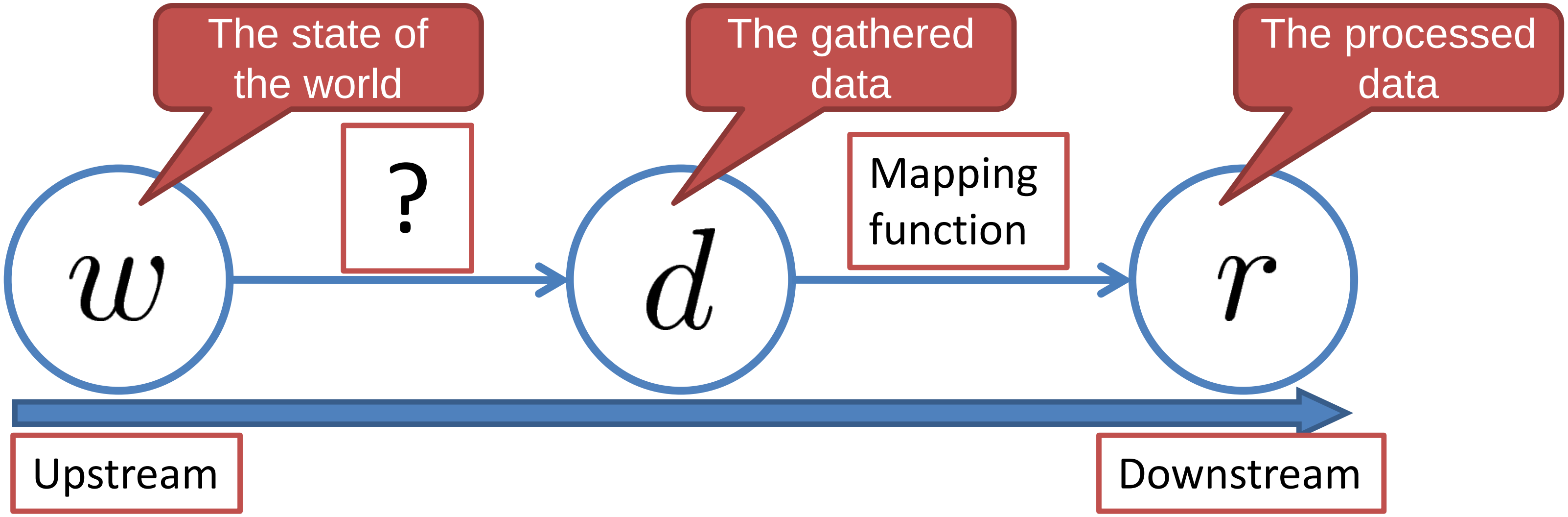
AlexNet

- Residual net

ResNet



The data processing theorem revisited



Markov chain $P(w, d, r) = P(w)P(d \mid w)P(r \mid d)$

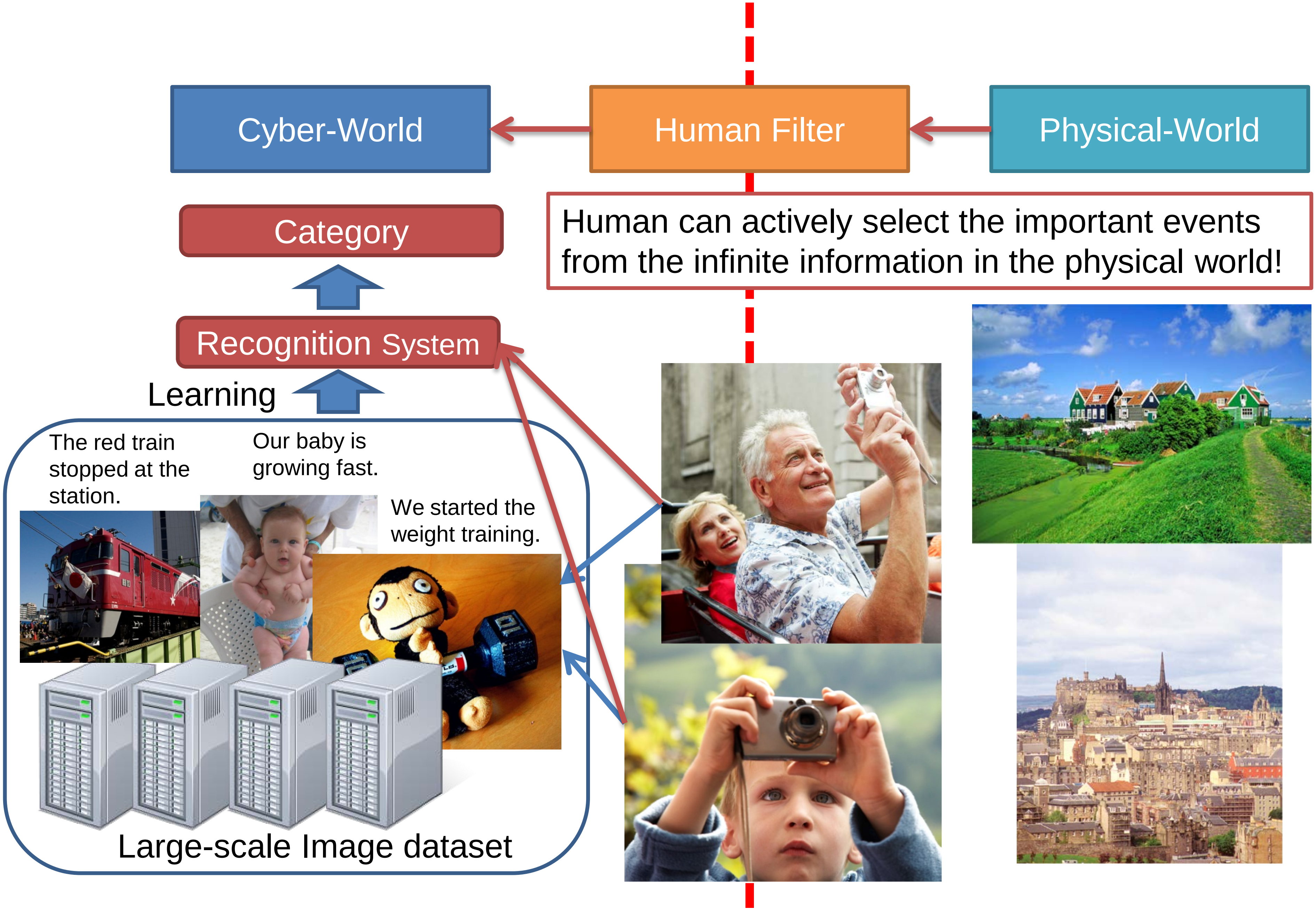
The average information

$$I(w; d) \geq I(w; r)$$

The data processing theorem states that data processing can only destroy information.

David J.C. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press 2003.

Framework of Recognition System



Journalist Robot

Since 2006

- Many interesting events in the physical-world are overlooked.
- Infinite information is embedded in the physical-world.
- What should we focus on in the physical-world?
- Journalist Robot
  - moves about in the physical-world, finds news-like events, recognizes scenes and objects, interviews with people, and finally generates the articles.
  - is a grand challenge of intelligent robot.

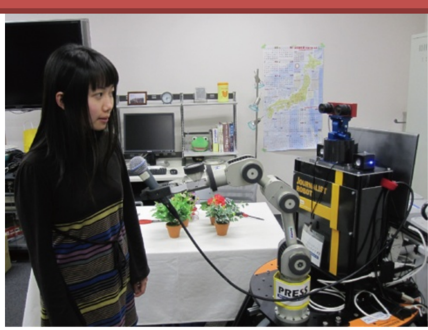
Image Recognition

meerkat
1. meerkat
2. snow leopard
3. Komodo dragon
4. raccoon
5. common iguana

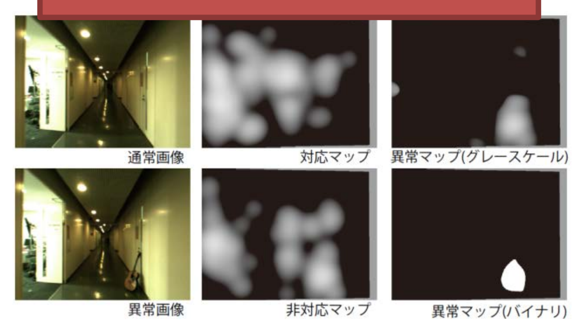
pumpkin seed
1. pumpkin seed
2. french fries
3. Dungeness crab
4. cashew nut
5. jigsaw puzzle

Article Generation

Interviewing



News Detection



Anomaly Detection

Automatic Article Generation in 2011

Results

News article generated (in Japanese)

What is this strange thing?

2011/02/12 23:34:11

Witness said, “Practicing poster session for coming conference. It is about a robot finding news”.

異常発見! これは一体!?

2011/02/12 23:34:11

付近の人によれば、「学会が近いので発表練習をしています。自分でニュースを探してくるロボットの研究です」らしい。

The picture taken by the system near the abnormal object.

journalistrobot I found: http://localhost/zoomed_news_image.png Witness said, “Practicing poster session for coming conference. It is about a robot finding news”.

about 19 minutes ago from api

In twitter client:

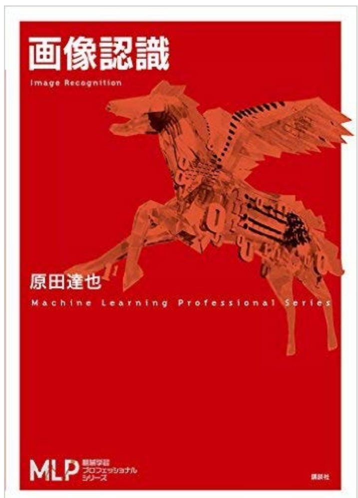
All Friends journalistrobot (localhost)

I found: http://localhost/zoomed_news_image.png Witness said, “Practicing poster session for coming conference. It is about a robot finding news”.

journalistrobot, [+]
Sat 12 Feb 23:34 via api

⇨ The followers of the system gets easy access to the news.

画像認識の教科書



画像認識 (機械学習プロフェッショナルシリーズ)
単行本 – 2017/5/25
原田 達也 (著)

￥ 3,240　　288ページ

■おもな内容
第1章 画像認識の概要
第2章 局所特徴
第3章 統計的特徴抽出
第4章 コーディングとプーリング
第5章 分類
第6章 畳み込みニューラルネットワーク
第7章 物体検出
第8章 インスタンス認識と画像検索
第9章 さらなる話題(セマンティックセグメンテーション/画像からのキャプション生成/画像生成と敵対的生成ネットワーク)