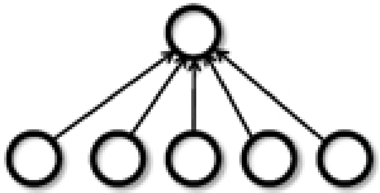
Neural Networks-2

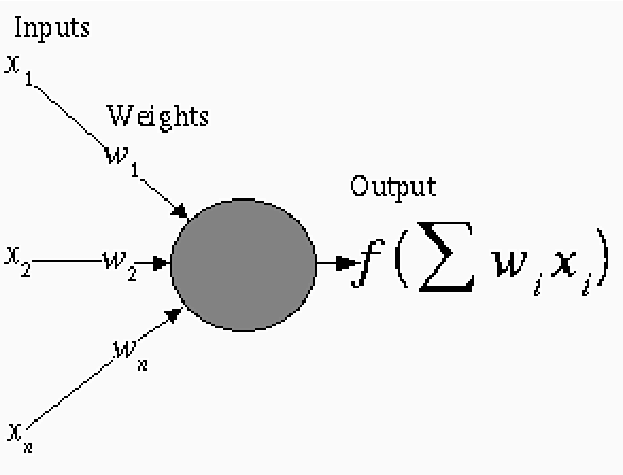
Changshui Zhang
Department of Automation，Tsinghua University
zcs@mail.Tsinghua.edu.cn

July, 2018

Neural Networks in 60’s

- Perceptron





1968, Perceptron, Frank Rosenblat

“Winter of Artificial Intelligence” Since 70’s

- 1969, *Perceptrons: An Introduction to Computational Geometry*, Marvin Minsky and Seymour Papert

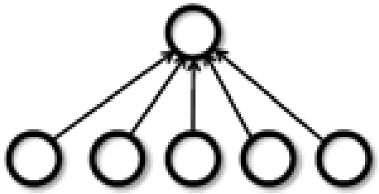
  Basic perceptrons were incapable of processing the exclusive-or circuit

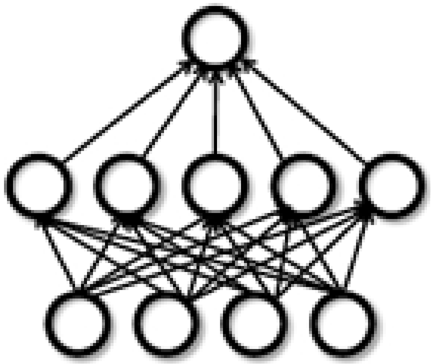
  Computers did not have enough processing power

- Cold winter

Neural Networks in 80’ s

- 1986, Backpropagation algorithm, Rumelhart



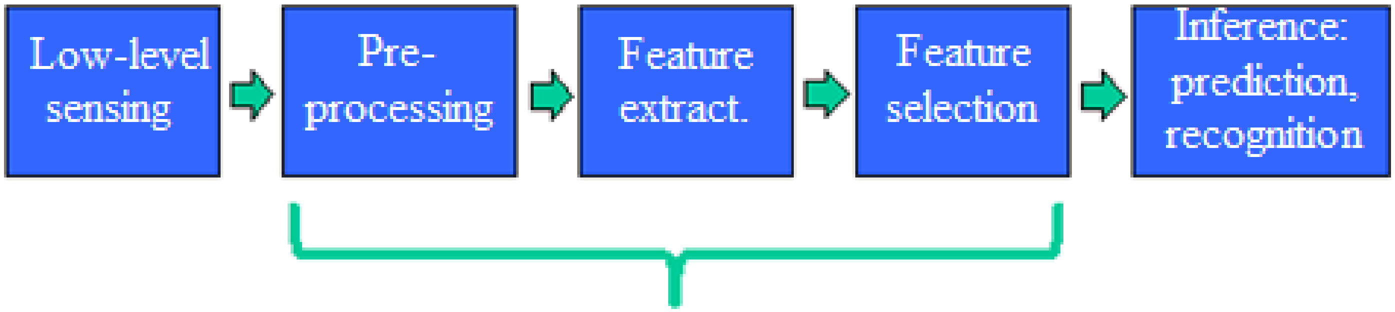


“Winter of Neural Networks” Since 90’s

- Non-convex
- Need a lot of tricks to play with
- Hard to do theoretical analysis
- Overfitting

The pipeline of machine visual perception

Most Efforts in Machine Learning



Low-level sensing

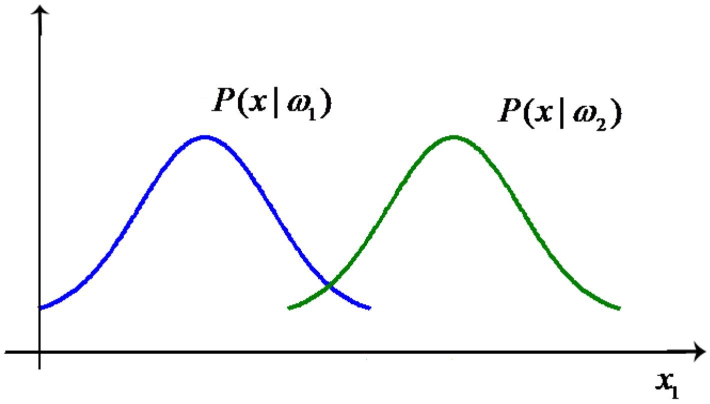
Pre-processing

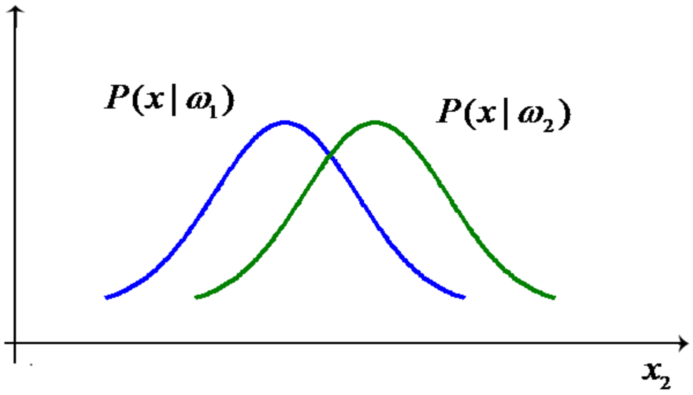
Feature extract.

Feature selection

Inference: prediction, recognition

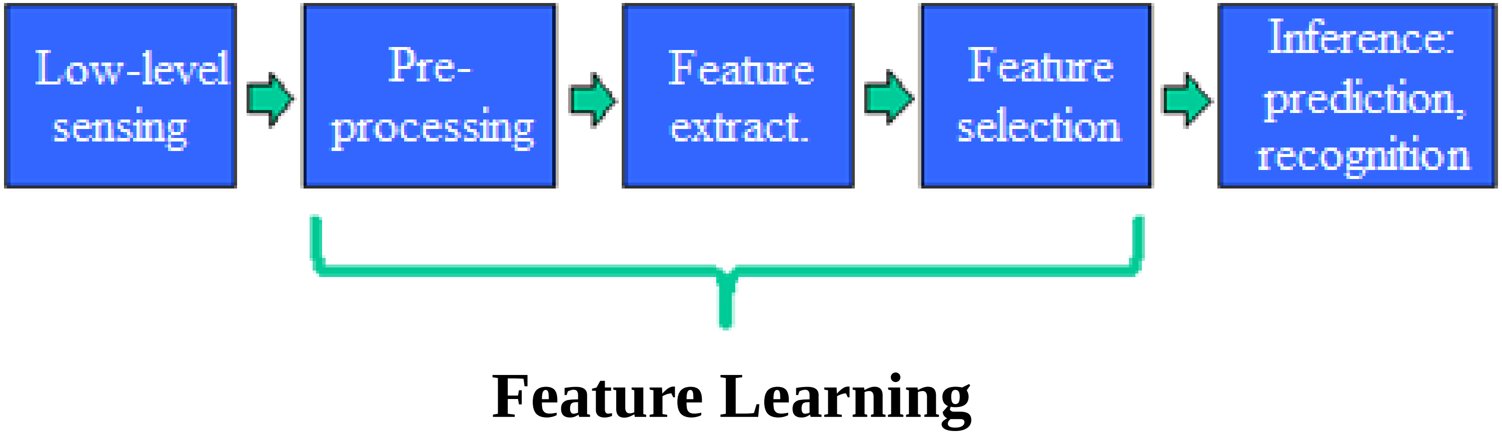
- **Most critical for accuracy**
- **Most time-consuming in development cycle**
- **Often hand-craft in practice**





Deep Learning: learning features from data

Machine Learning



Low-level sensing

Pre-processing

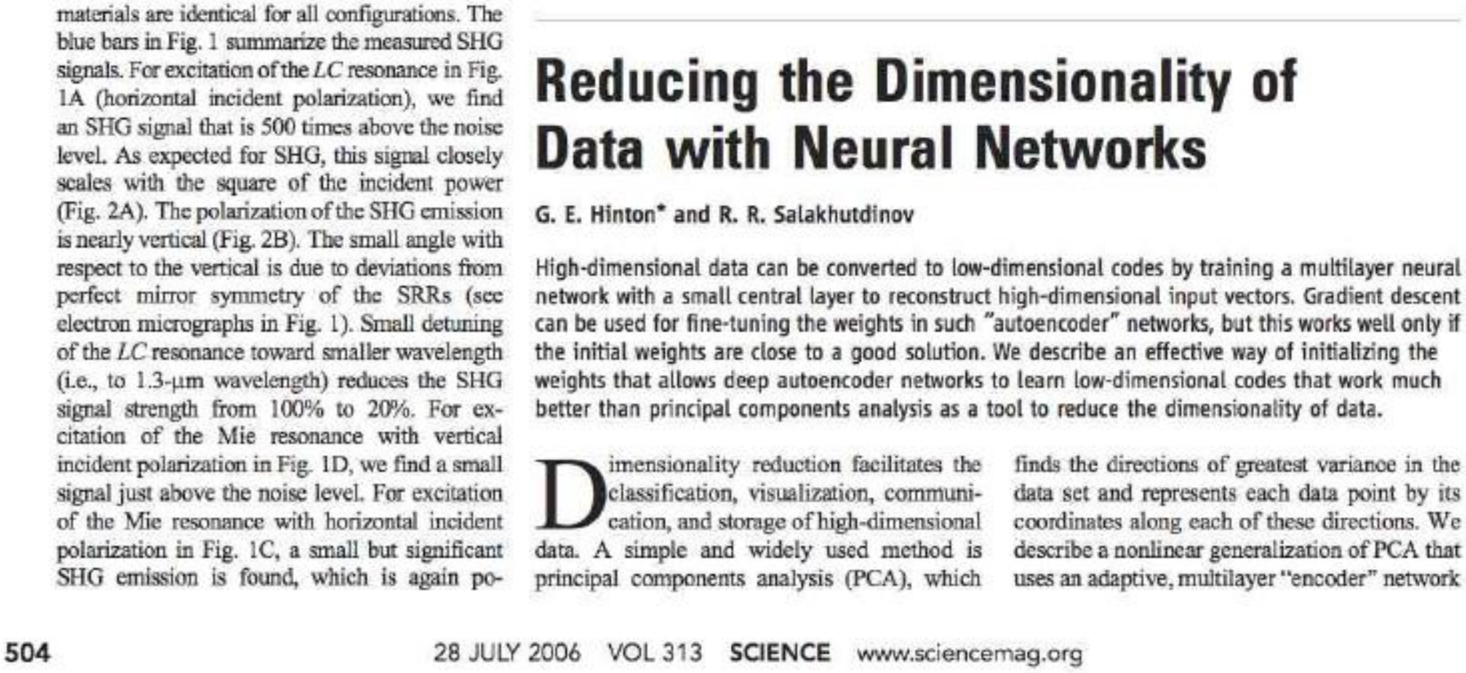
Feature extract.

Feature selection

Inference: prediction, recognition

Feature Learning

Deep Learning Since 2006



materials are identical for all configurations. The blue bars in Fig. 1 summarize the measured SHG signals. For excitation of the *LC* resonance in Fig. 1A (horizontal incident polarization), we find an SHG signal that is 500 times above the noise level. As expected for SHG, this signal closely scales with the square of the incident power (Fig. 2A). The polarization of the SHG emission is nearly vertical (Fig. 2B). The small angle with respect to the vertical is due to deviations from perfect mirror symmetry of the SRRs (see electron micrographs in Fig. 1). Small detuning of the *LC* resonance toward smaller wavelength (i.e., to 1.3-μm wavelength) reduces the SHG signal strength from 100% to 20%. For excitation of the Mie resonance with vertical incident polarization in Fig. 1D, we find a small signal just above the noise level. For excitation of the Mie resonance with horizontal incident polarization in Fig. 1C, a small but significant SHG emission is found, which is again po-

Reducing the Dimensionality of Data with Neural Networks

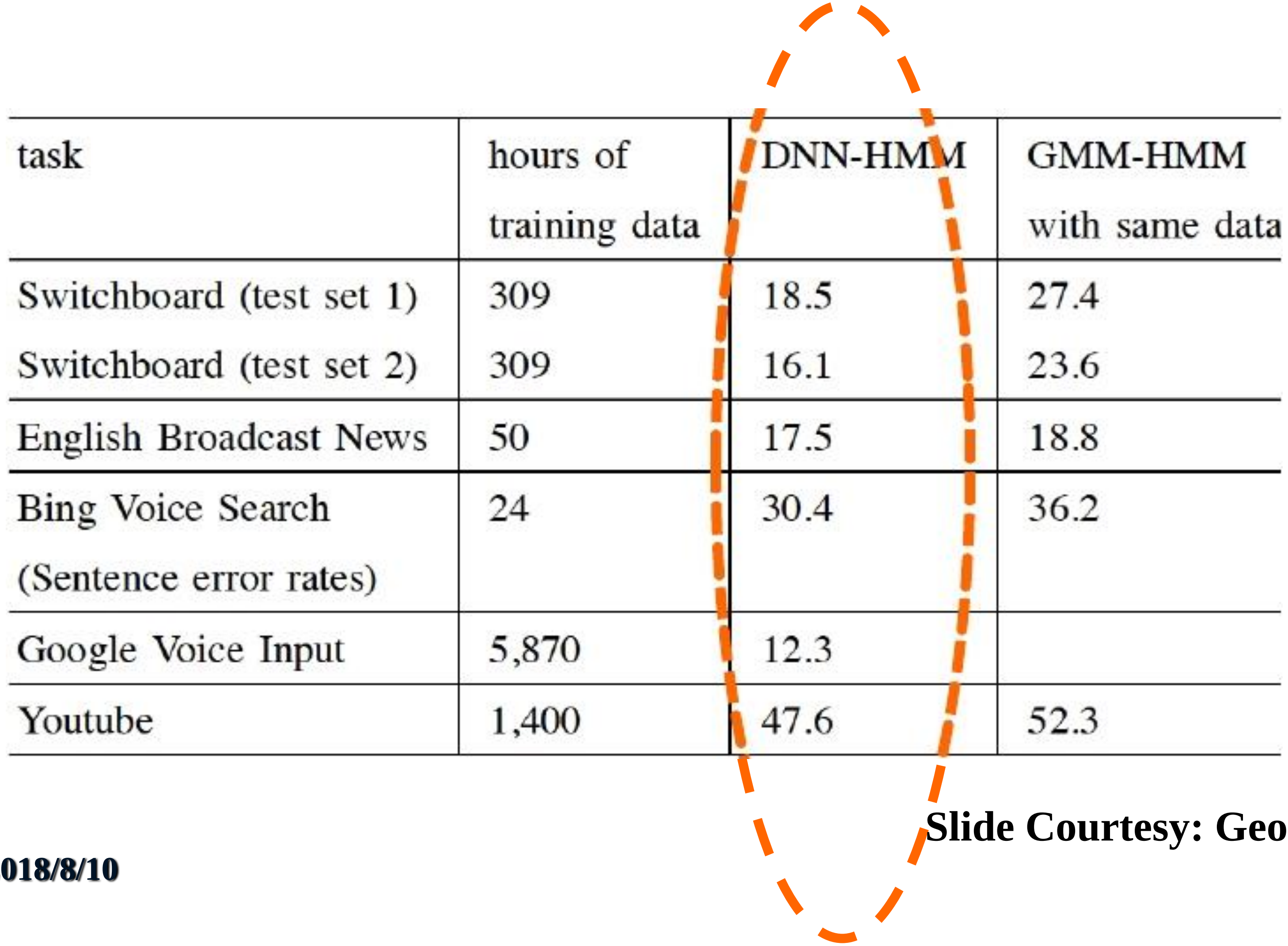
G. E. Hinton* and R. R. Salakhutdinov

High-dimensional data can be converted to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors. Gradient descent can be used for fine-tuning the weights in such "autoencoder" networks, but this works well only if the initial weights are close to a good solution. We describe an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis as a tool to reduce the dimensionality of data.

Dimensionality reduction facilitates the classification, visualization, communication, and storage of high-dimensional data. A simple and widely used method is principal components analysis (PCA), which finds the directions of greatest variance in the data set and represents each data point by its coordinates along each of these directions. We describe a nonlinear generalization of PCA that uses an adaptive, multilayer "encoder" network

Neural networks 第三次浪潮

Revolution on Speech



task	hours of training data	DNN-HMM	GMM-HMM with same data
Switchboard (test set 1)	309	18.5	27.4
Switchboard (test set 2)	309	16.1	23.6
English Broadcast News	50	17.5	18.8
Bing Voice Search (Sentence error rates)	24	30.4	36.2
Google Voice Input	5,870	12.3	
Youtube	1,400	47.6	52.3

Slide Courtesy: Geoff Hinton

Race on Image Net (Top 5 Hit Rate)

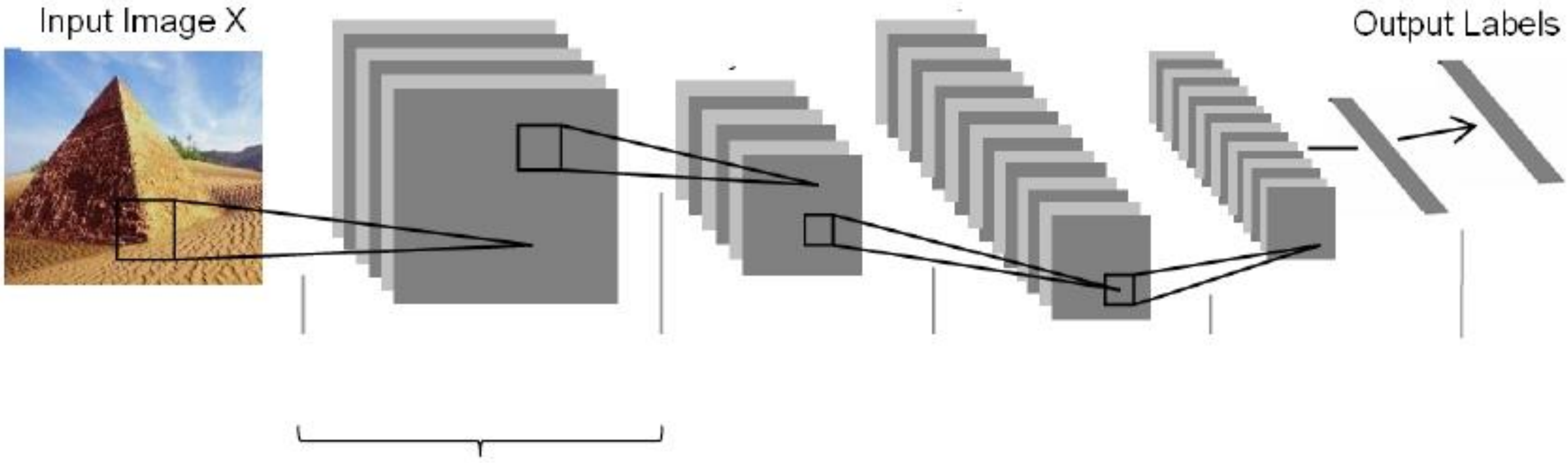
1000 object classes that we recognize

Images courtesy of ImageNet (http://www.image-net.org/challenges/LSVRC/2010/index)

72%, 2010

74%, 2011

The Best system on ImageNet (by 2012.10)



- This is a moderate deep model
- The first two layers are hand-designed

Challenge to Deep Learners

Key questions:

- **What if no hand-craft features at all?**
- **What if use much deeper neural networks?**

Our chief critic, Jitendra Malik, has said that this competition is a good test of whether deep neural networks really do work well for object recognition.

-- By Geoff Hinton

Answer from Geoff Hinton， 2012.10

1000 object classes that we recognize
Images courtesy of ImageNet (http://www.image-net.org/challenges/LSVRC/2010/index)

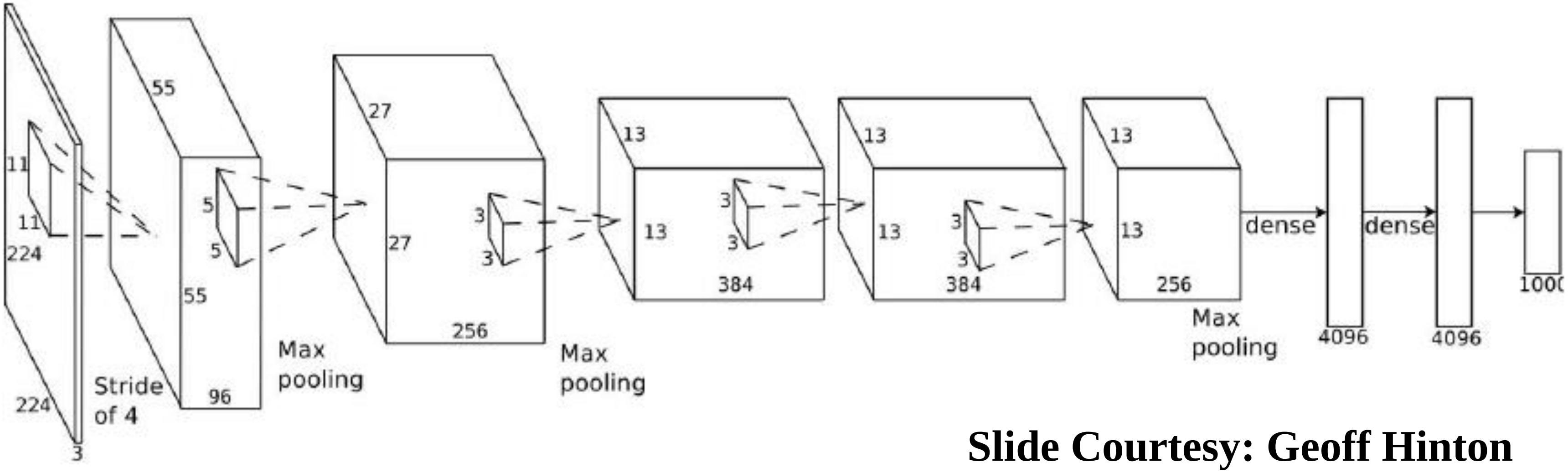
72%, 2010

74%, 2011

85%, 2012

The Architecture

- Max-pooling layers follow first, second, and fifth convolutional layers
- The number of neurons in each layer is given by 253440, 186624, 64896, 64896, 43264, 4096, 4096, 1000



Slide Courtesy: Geoff Hinton

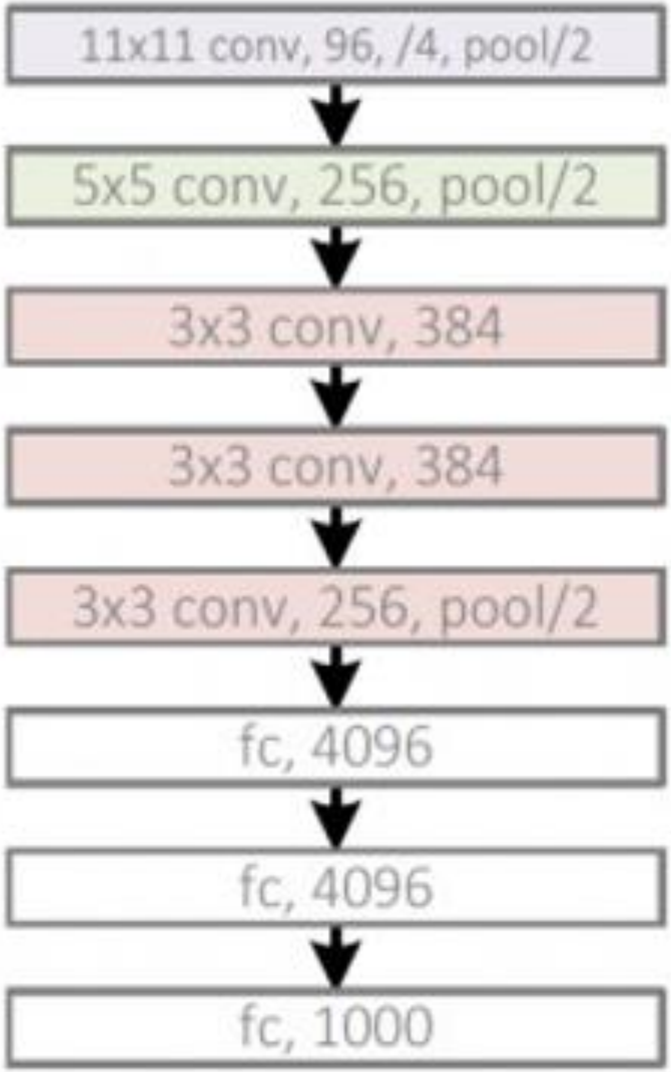
The Architecture

- – 7 hidden layers not counting max pooling.
- – Early layers are conv., last two layers globally connected.
- – Uses rectified **linear units** every layer.
- – Uses competitive normalization to suppress hidden activities.

Slide Courtesy: Geoff Hinton

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)



11x11 conv, 96, /4, pool/2

5x5 conv, 256, pool/2

3x3 conv, 384

3x3 conv, 384

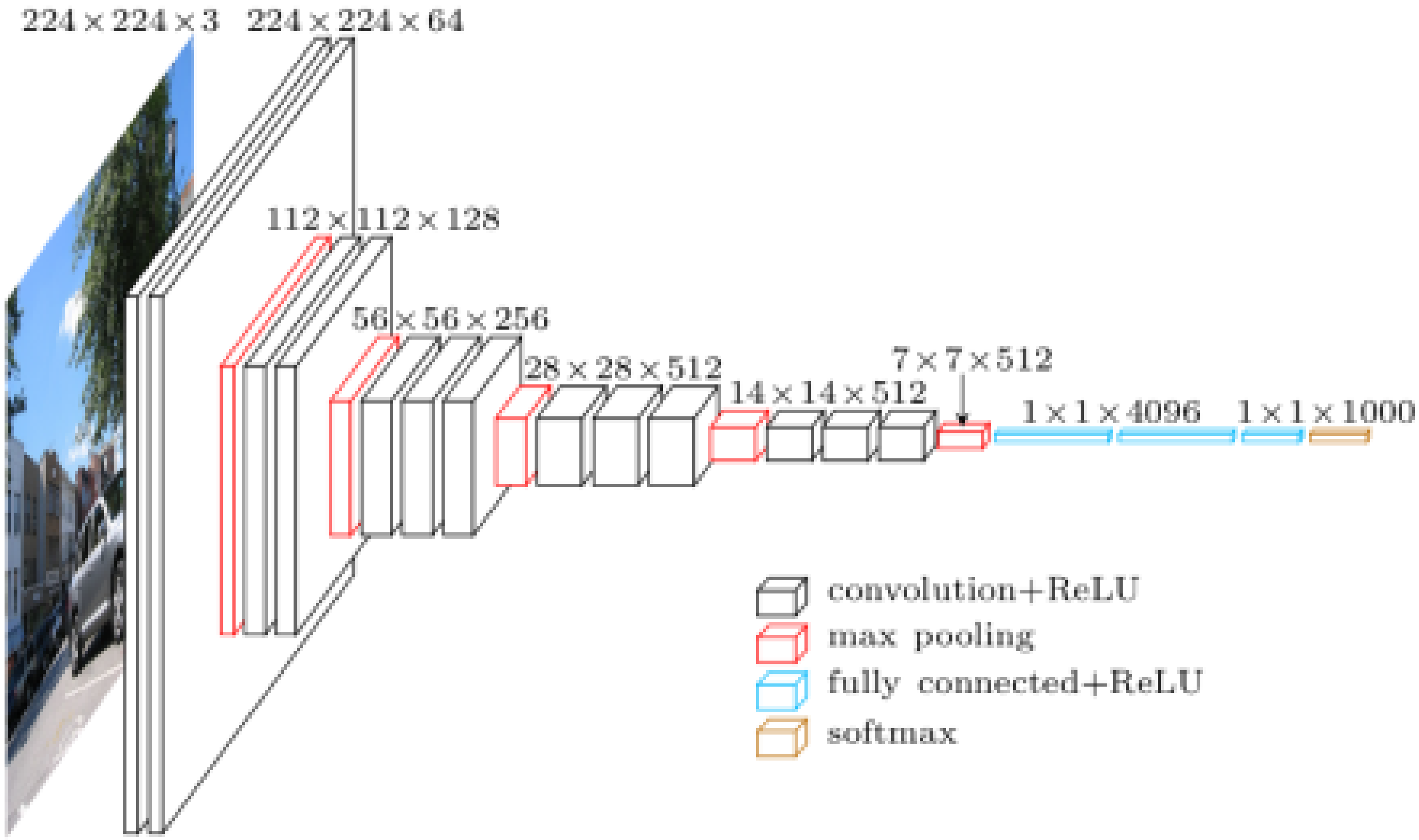
3x3 conv, 256, pool/2

fc, 4096

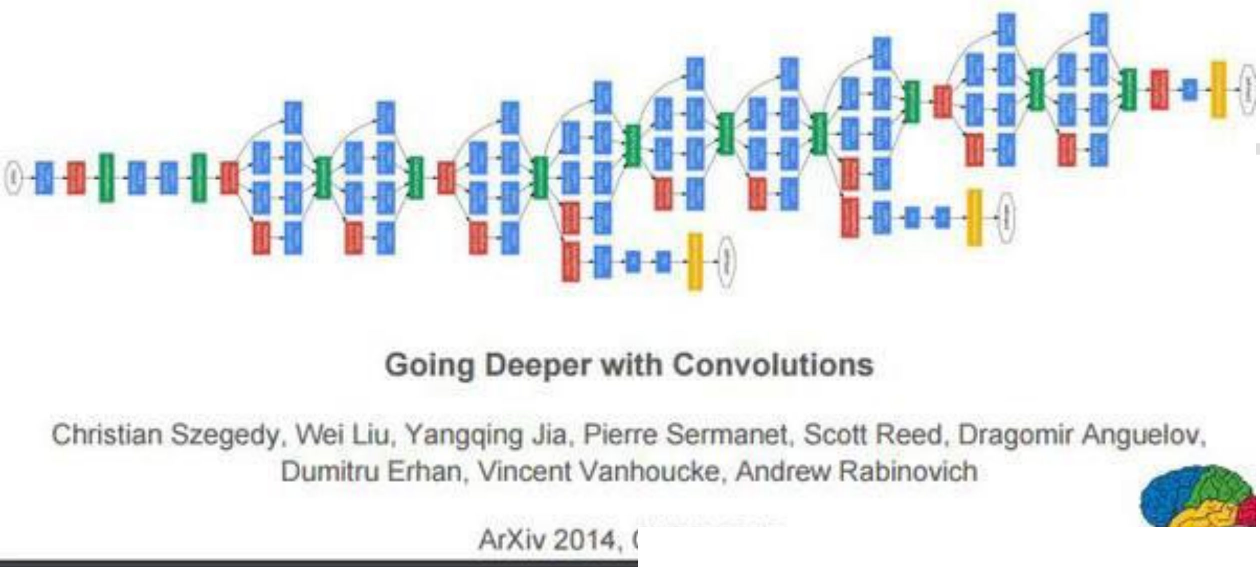
fc, 4096

fc, 1000

VGG ILSVRC 2014



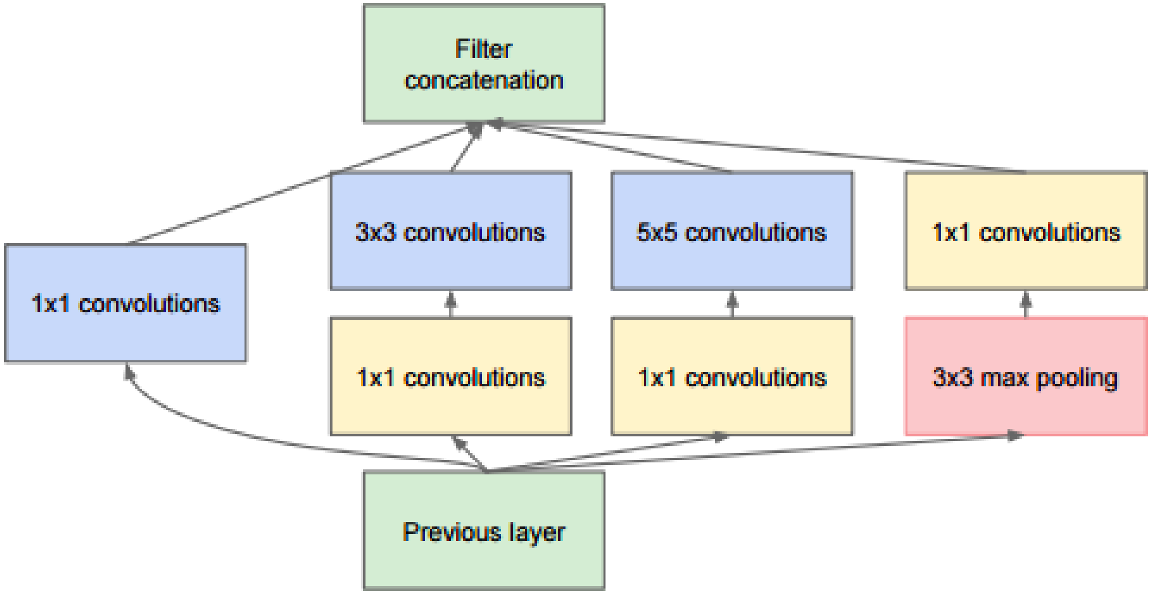
The Inception Architecture (GoogLeNet, 2014)



Going Deeper with Convolutions

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov,
Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich

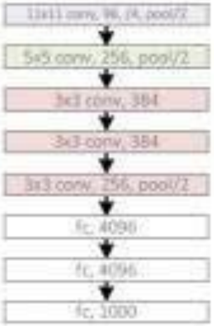
ArXiv 2014, (



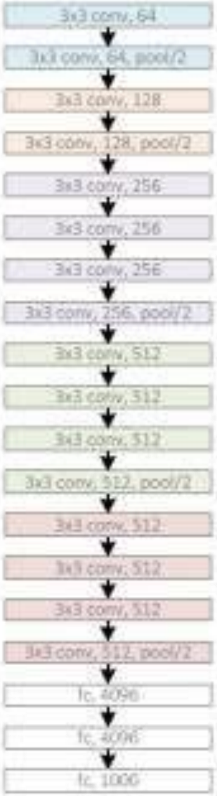
(b) Inception module with dimension reductions

Revolution of Depth

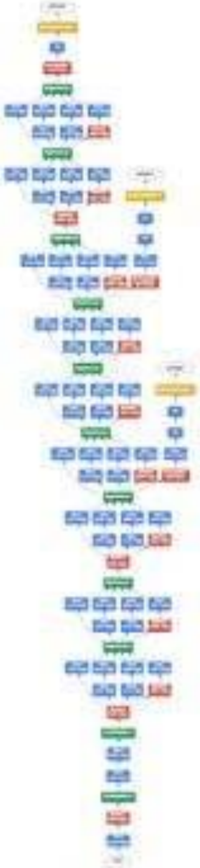
AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



GoogleNet, 22 layers
(ILSVRC 2014)

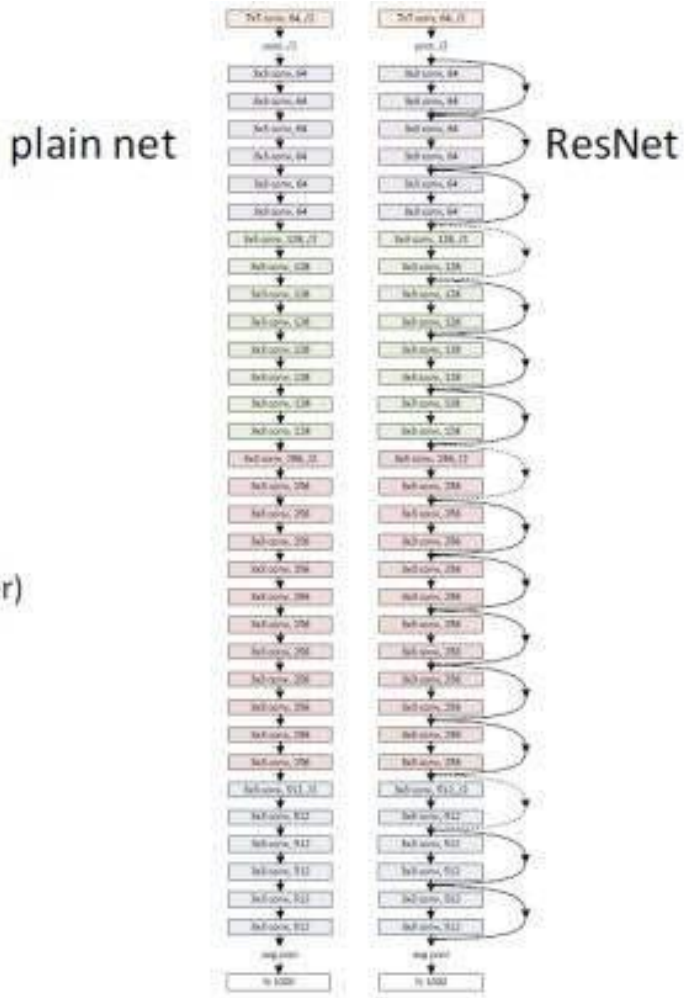


Network "Design"

- Keep it simple

- Our basic design (VGG-style)
  - all 3x3 conv (almost)
  - spatial size /2 => # filters x2 (~same complexity per layer)
  - Simple design; just deep!

- Other remarks:
  - no hidden fc
  - no dropout



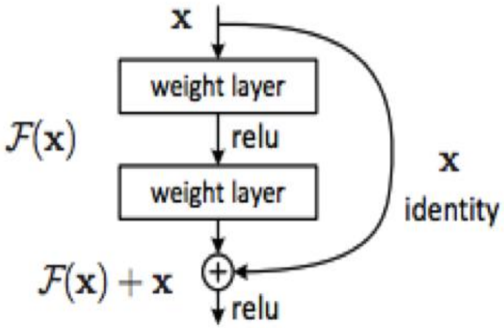
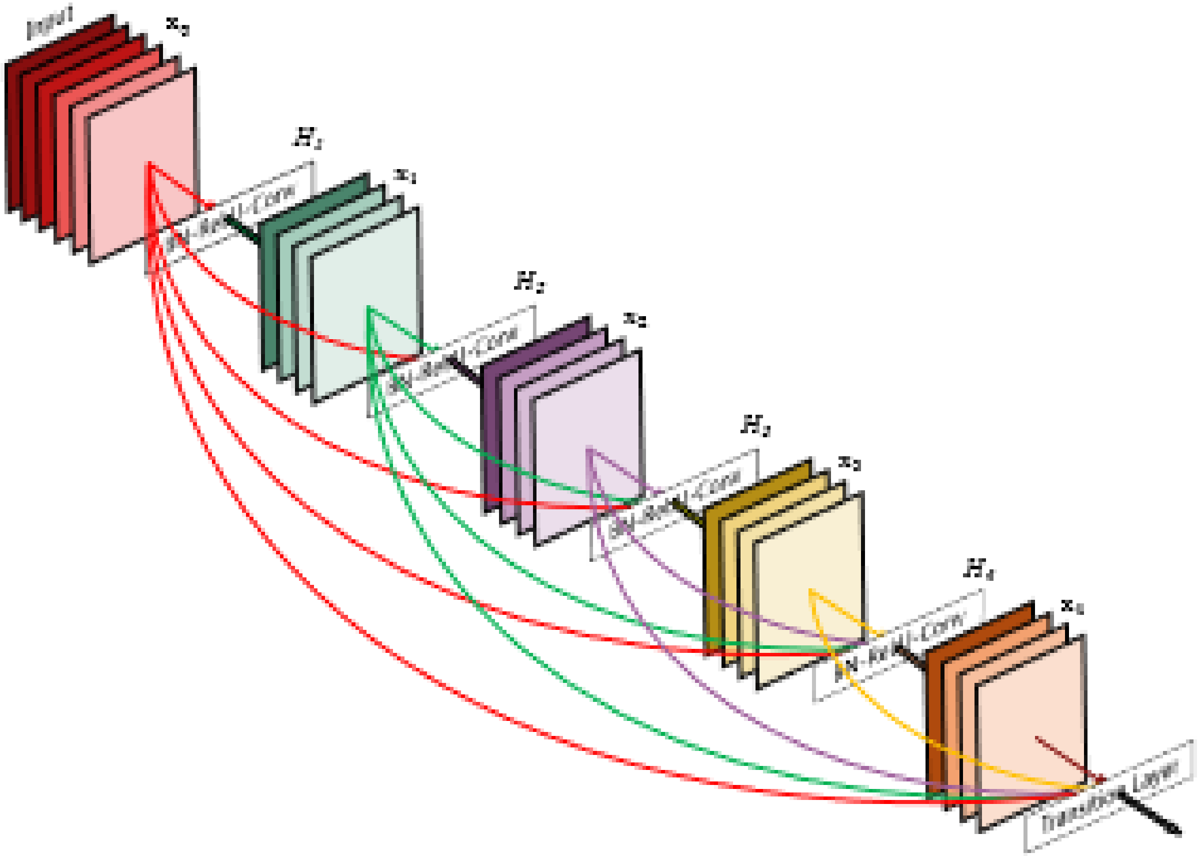

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2. Residual learning: a building block.

DenseNet



- Gao Huang , Zhuang Liu, etc., Densely Connected Convolutional Networks, CVPR 2017

Applications

- Digital recognition: MNIST
- Machine translation
- Image caption

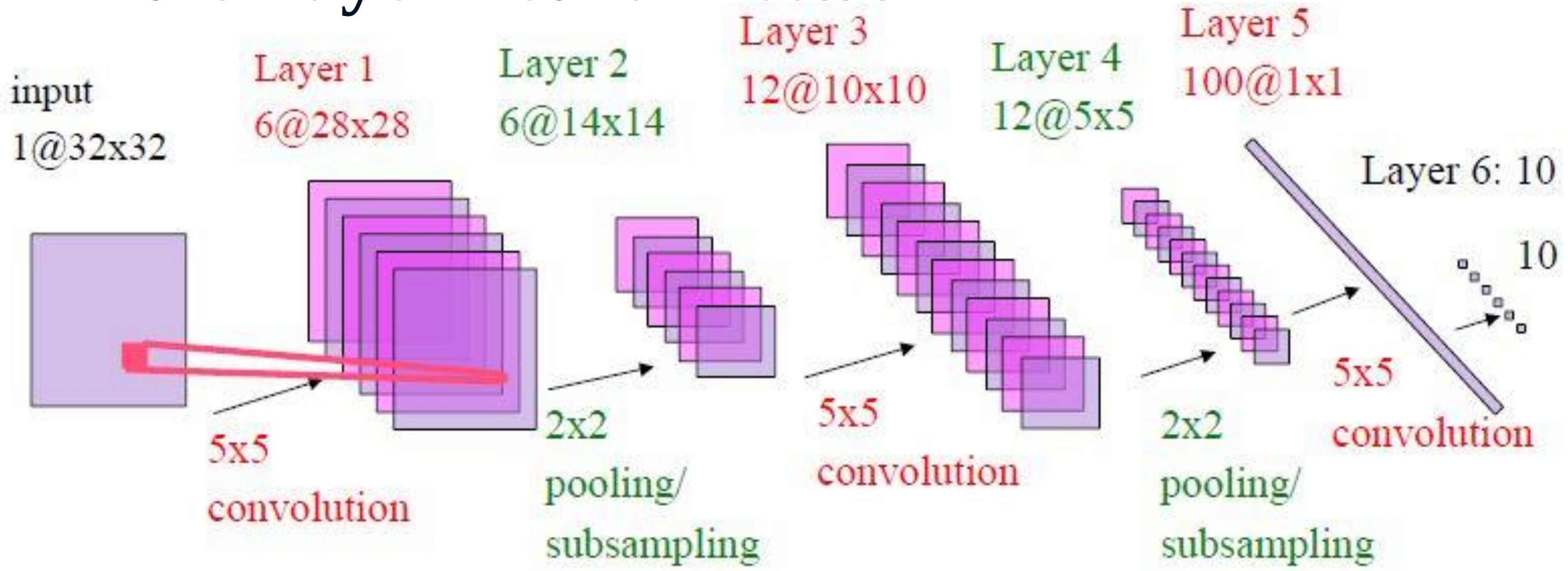
LeNet5

Convolution operator

Pooling operator

Kernel size

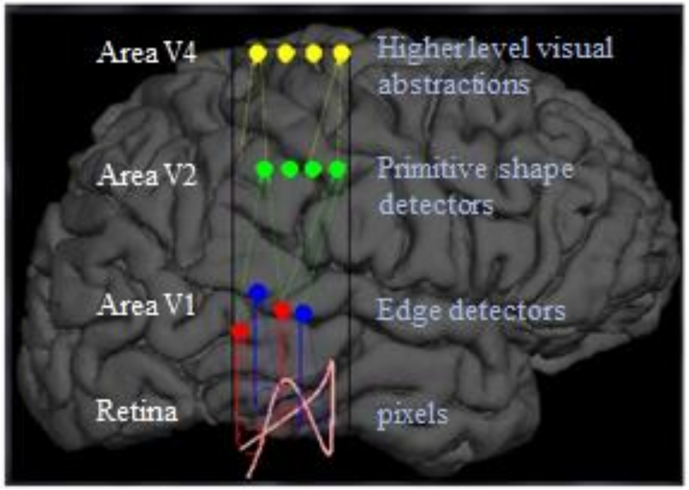
A multi-layer neural network

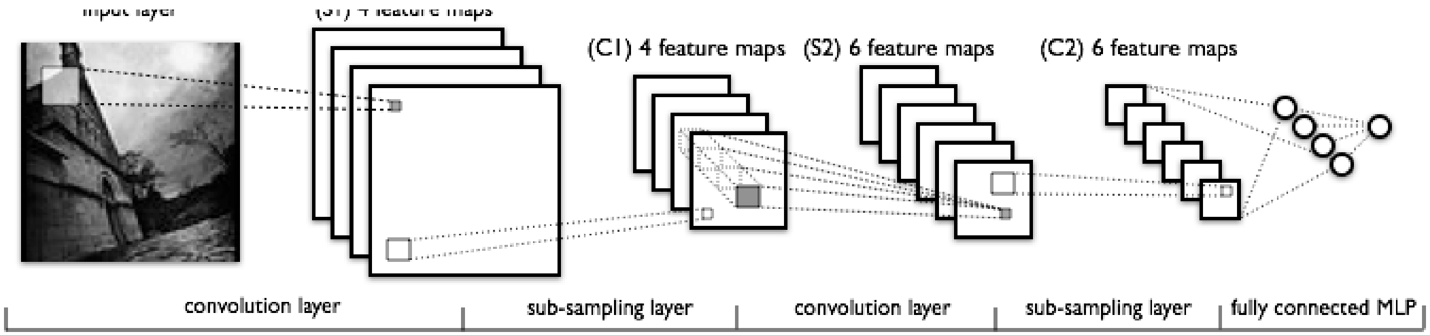


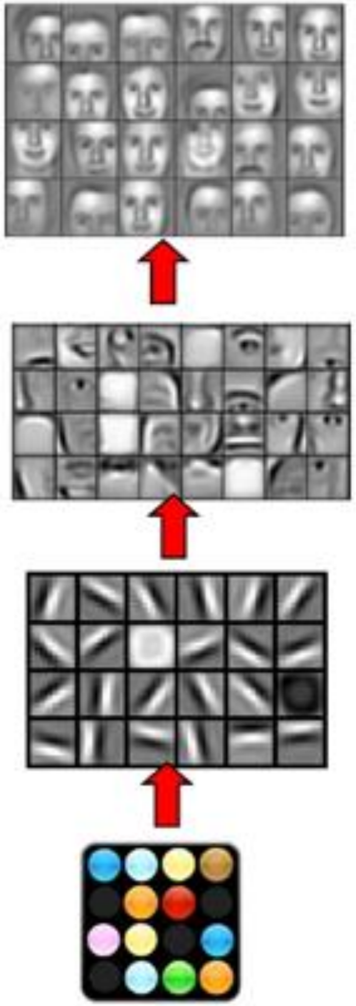
Simple ConvNet for MNIST [LeCun 1998]， 1993,1994,1995

CNN and Feature Learning

- **From low level feature to high level feature**

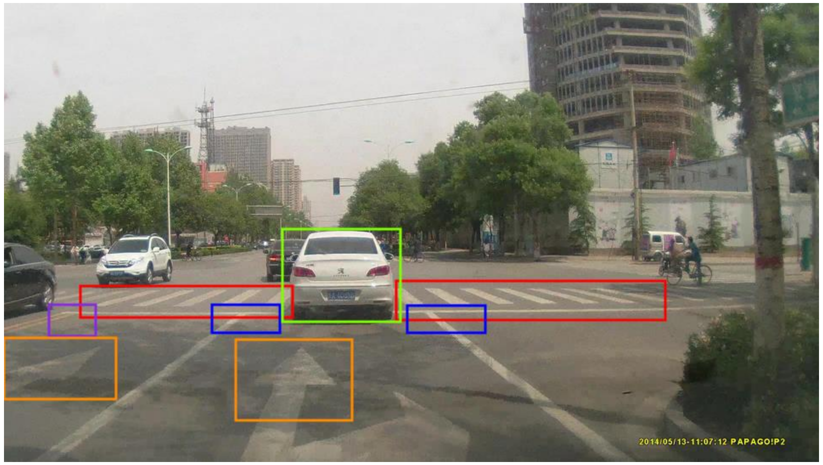


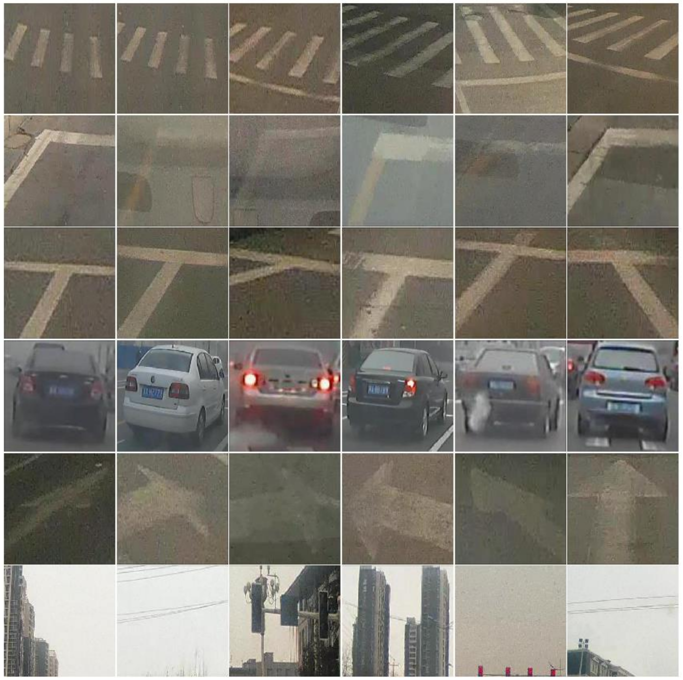




路上标识识别

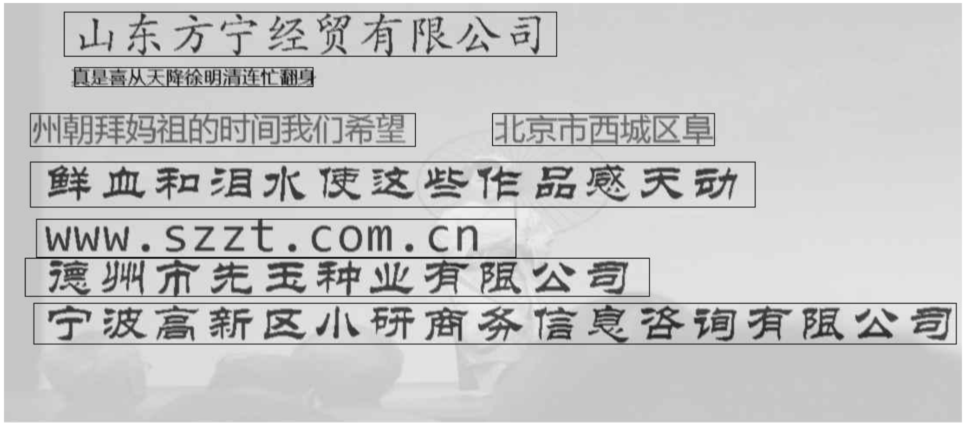
- 箭头、斑马线、停车线、车道线





文本行定位与OCR

- 自动确定每个文本行的坐标
- 单字识别
- 整行识别

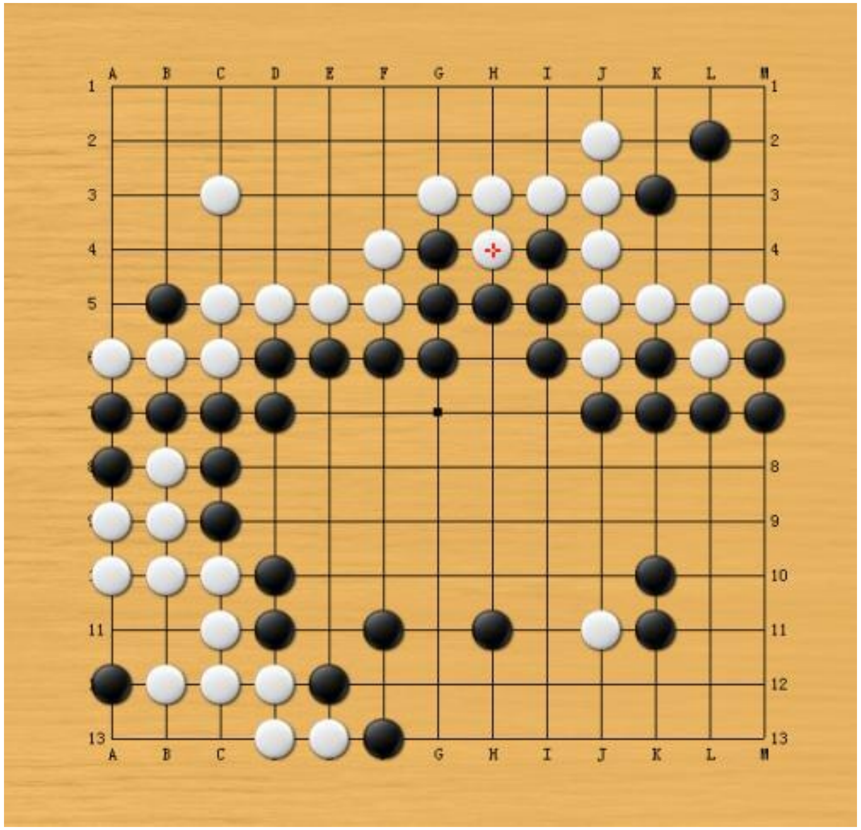


其它应用

- 超分辨率
- 图像分割
- 物体检测
- …

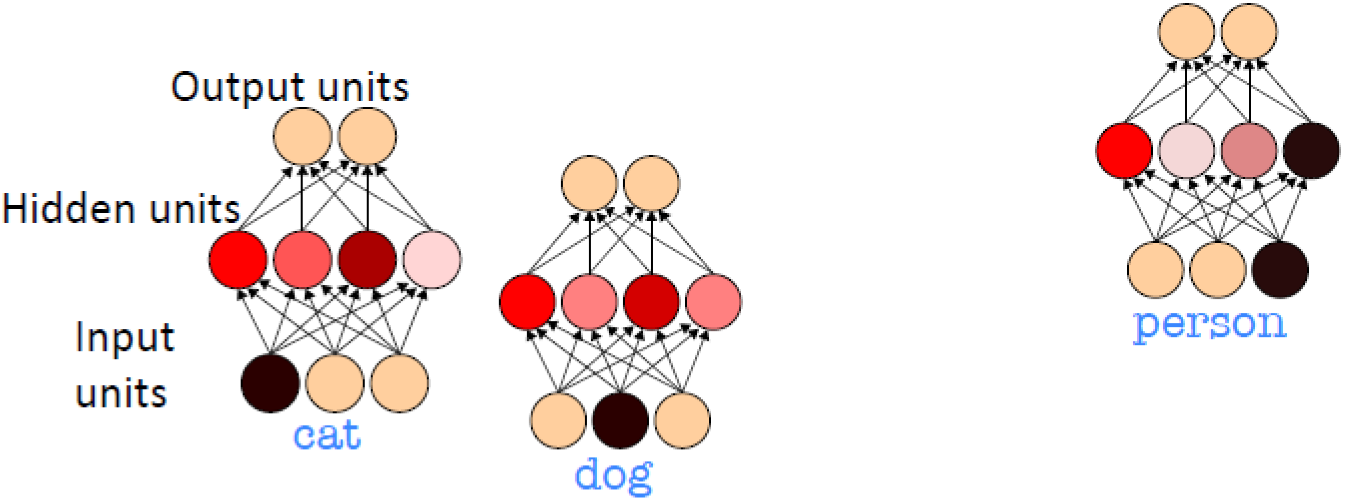
AlphaGo

- 输入：围棋图像
- 输出：动作



Comments

- **Concepts are not represented by symbols in our brain, but by patterns of activation (Connectionism, 1980’s)**
- **From “Deep Learning” NIPS’2015 Tutorial, Geoff Hinton, Yoshua Bengio & Yann LeCun**



Sequence to Sequence Learning with Neural Networks

- 实现英语句子到法语句子的翻译
- Encoder：一个多层LSTM将英语句子映射为一个固定长度的特征表达
- Decoder：另一个多层LSTM将特征表达解码为目标法语句子

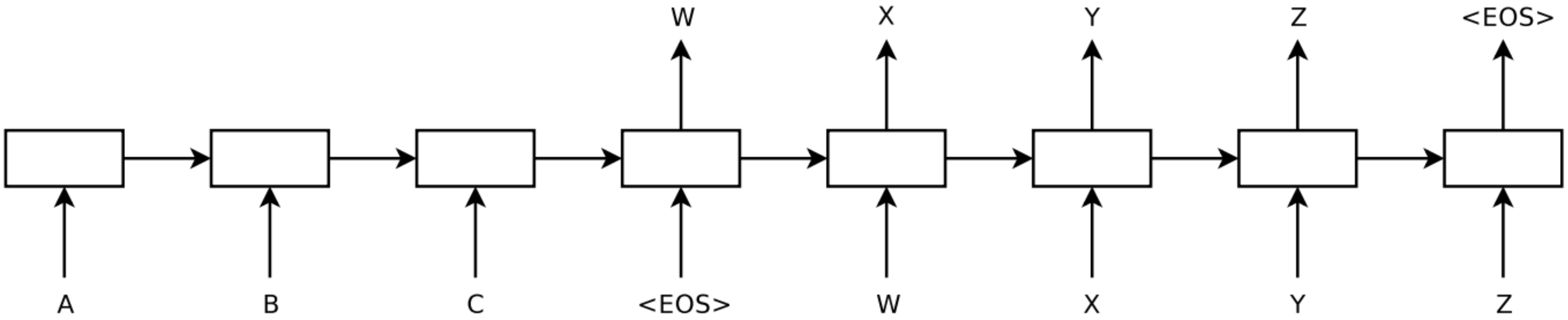


Figure from: I. Sutskever, O. Vinyals, Q. V. Le, Sequence to sequence learning with neural networks, NIPS, 2014

实现细节

- 结构采用Softmax输出作为目标词汇的概率表示

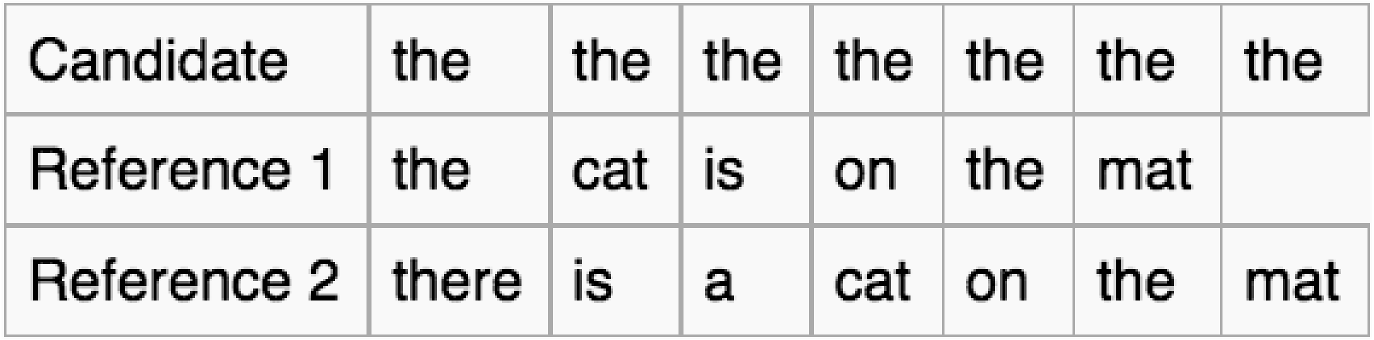
$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

- Encoder和Decoder是两个不同的LSTM
- 深层LSTM性能要好于浅层LSTM，文中采用了四层LSTM结构
- 目标函数

$$p(y_1, \cdots, y_{T'} | x_1, \cdots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \cdots, y_{t-1})$$

实现细节

- 作者发现将英语句子倒序输入可显著提升性能（BLEU: 25.9 → 30.6）
- 这种训练方式可能使得模型更容易建立起输入和输出的对应关系
- BLEU:

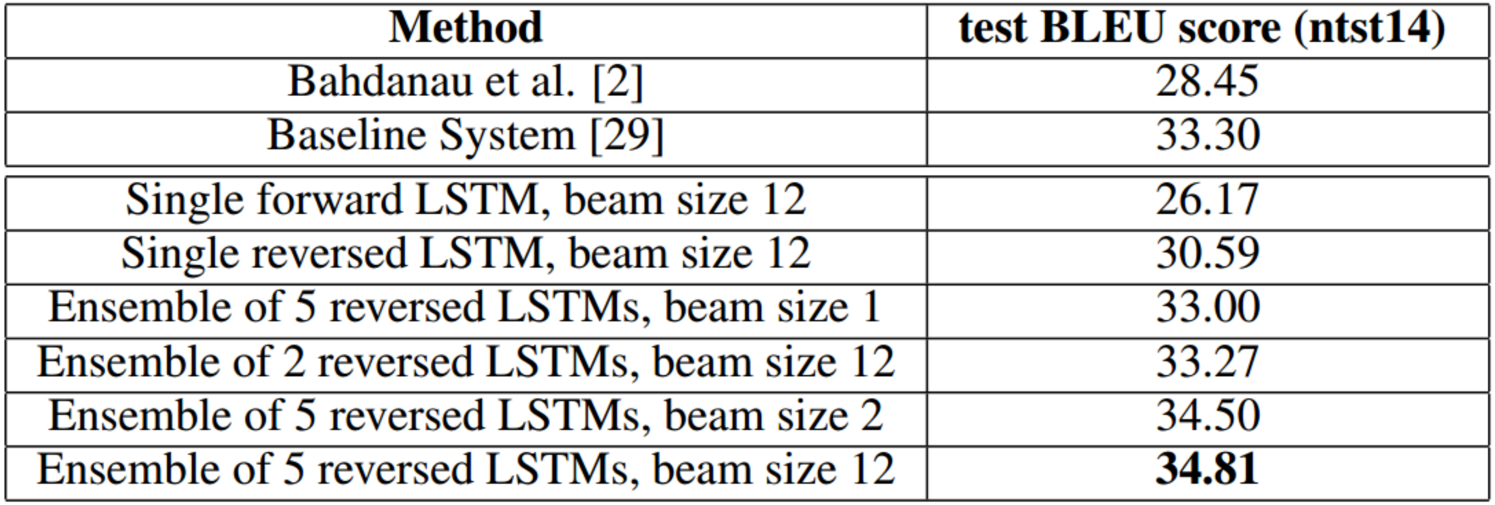


Candidate	the	the	the	the	the	the	the
Reference 1	the	cat	is	on	the	mat	
Reference 2	there	is	a	cat	on	the	mat

$$P = \frac{m}{w_t} = \frac{7}{7} = 1$$

$$P = \frac{2}{7}$$

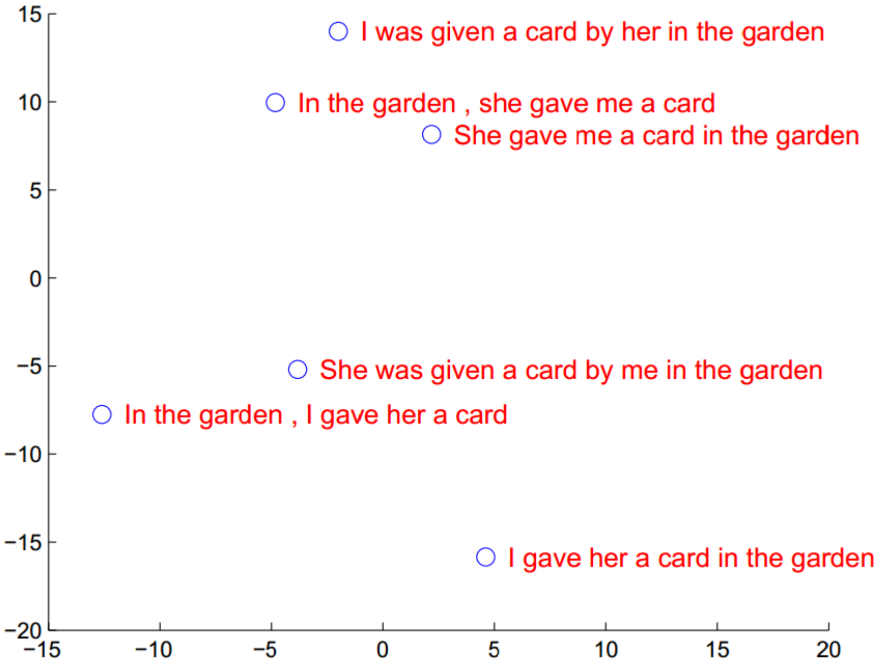
实验结果分析

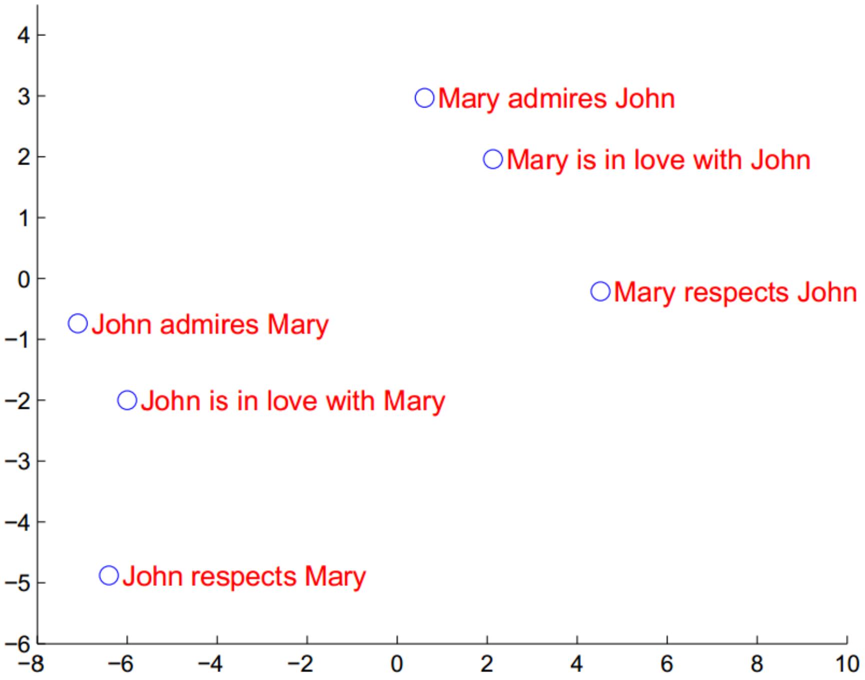


Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

实验结果分析

- 模型在较长的句子上仍然取得了不错的效果
- Encoder产生的特征向量具有一定的语义信息



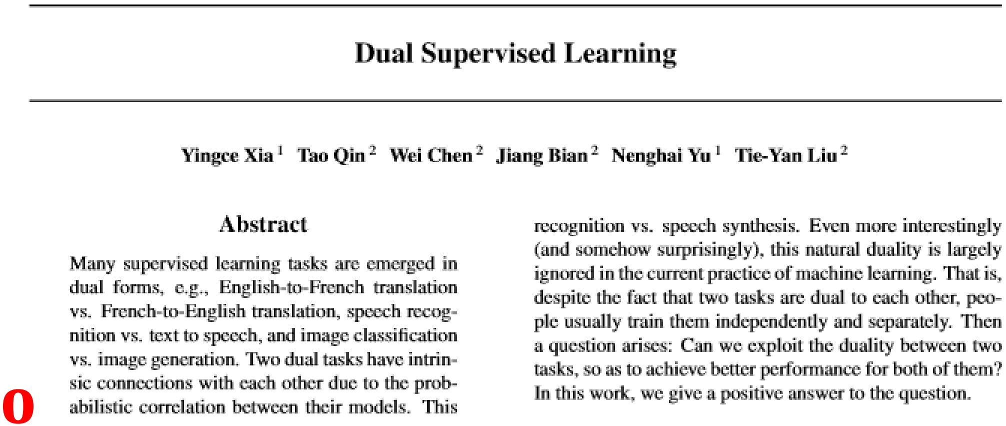


Interesting paper

Many supervised learning tasks are emerged in dual forms, e.g., English-to-French translation vs. French-to-English translation.

Training the models of two dual tasks simultaneously

Dual supervised learning can improve the practical performances of both tasks



Dual Supervised Learning

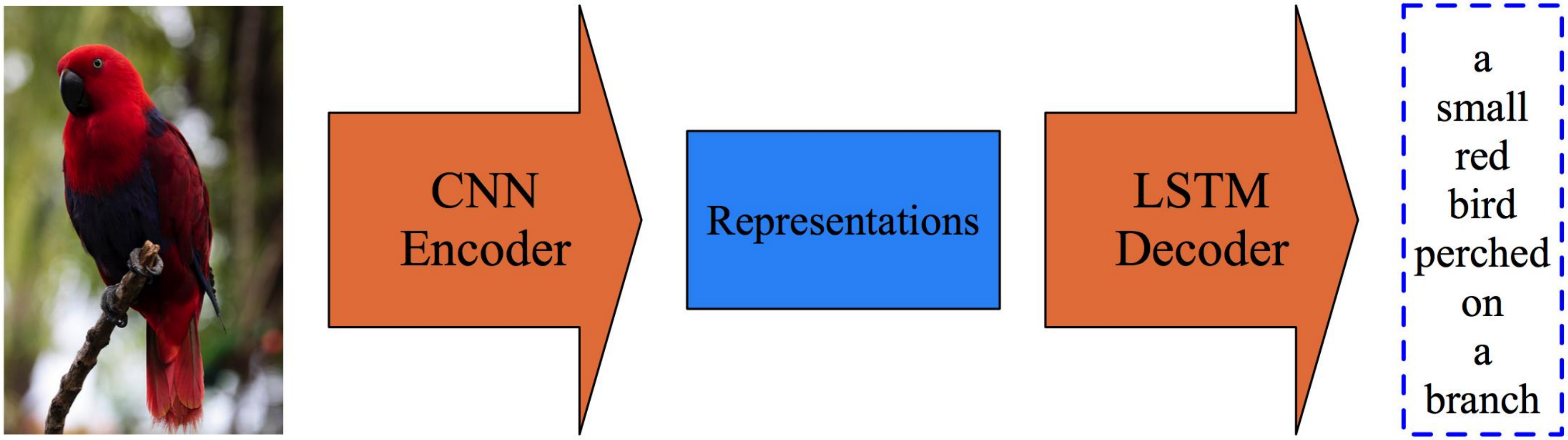
Yingce Xia [1] Tao Qin [2] Wei Chen [2] Jiang Bian [2] Nenghai Yu [1] Tie-Yan Liu [2]

Abstract

Many supervised learning tasks are emerged in dual forms, e.g., English-to-French translation vs. French-to-English translation, speech recognition vs. text to speech, and image classification vs. image generation. Two dual tasks have intrinsic connections with each other due to the probabilistic correlation between their models. This recognition vs. speech synthesis. Even more interestingly (and somehow surprisingly), this natural duality is largely ignored in the current practice of machine learning. That is, despite the fact that two tasks are dual to each other, people usually train them independently and separately. Then a question arises: Can we exploit the duality between two tasks, so as to achieve better performance for both of them? In this work, we give a positive answer to the question.

Image Caption

- Generate natural language sentence to describe image, involving:
  - Image understanding
  - language modeling



CNN Encoder

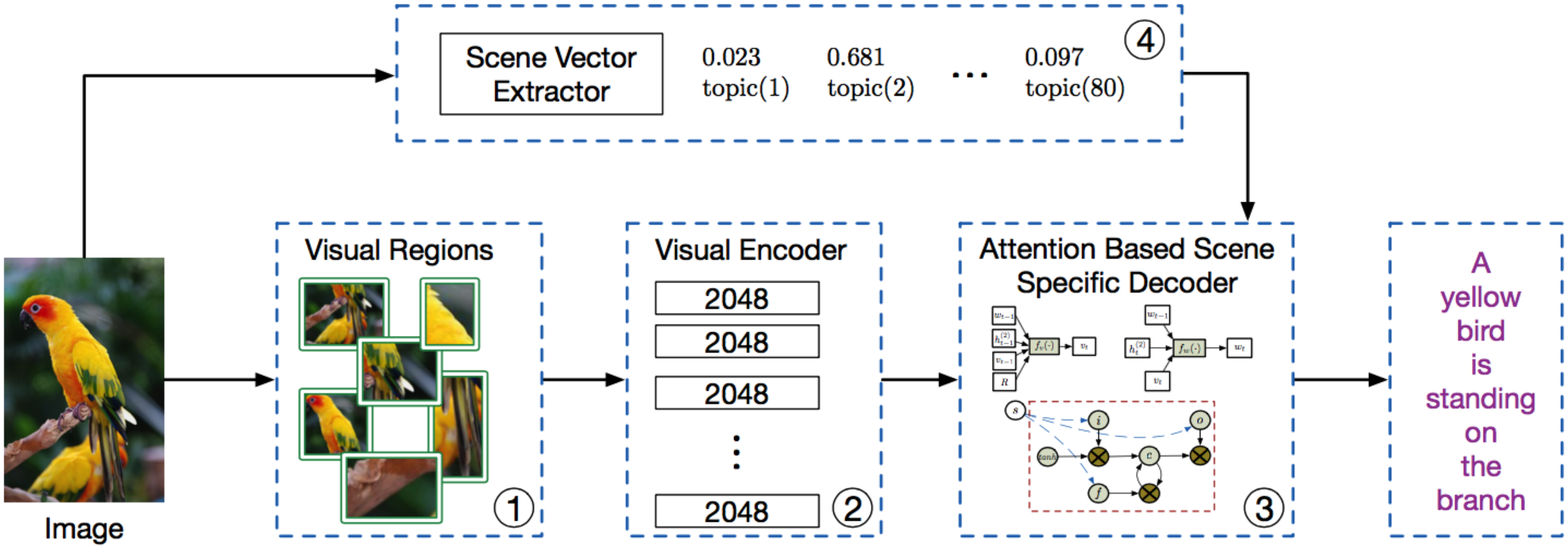
Representations

LSTM Decoder

a small red bird perched on a branch

Review

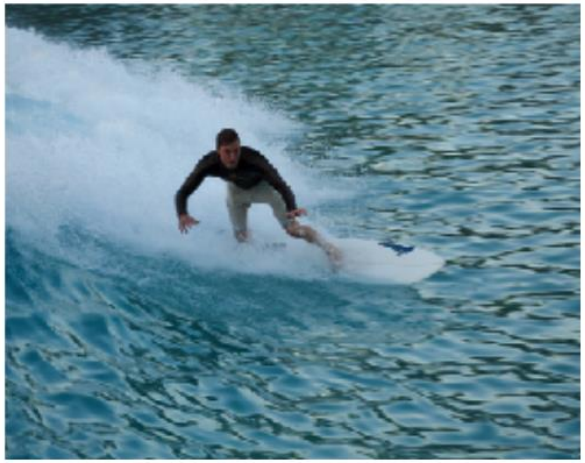
- Let’s review the model before entering experiments
  - CNN encoder + LSTM decoder
  - (1)Region-based attention + (2)scene-specific contexts



Experiment Setup

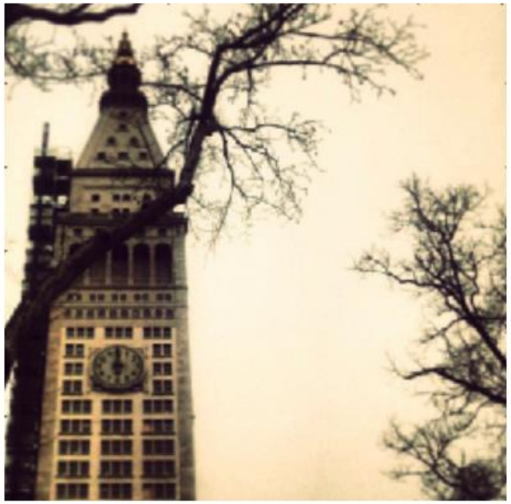
- Datasets:
  - MSCOCO: 82783, 40504, 40775
  - Flickr30K: 29000, 1000, 1000
  - Flickr8K: 6000, 1000, 1000
- Each image with 5 human labeled captions

Example Captions



[RA+SS]
a man riding a wave on top of a surfboard

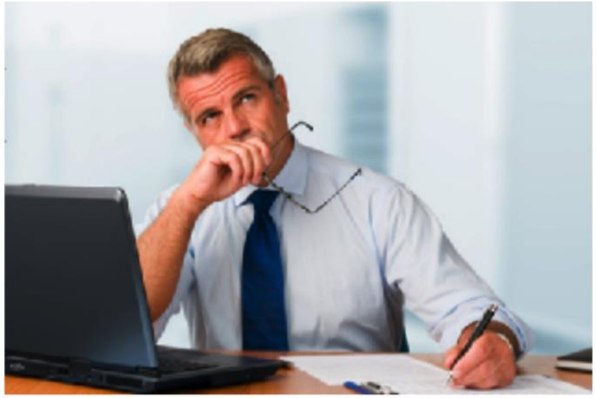
human:
1) a man riding a board on top of a wave in the ocean
2) a guy in a black and white outfit is surfing
3) a man in a wet suit riding a surfboard on a wave
4) a male surfing a large ocean wave on a white surfboard
5) a man is riding a surboard on a wave



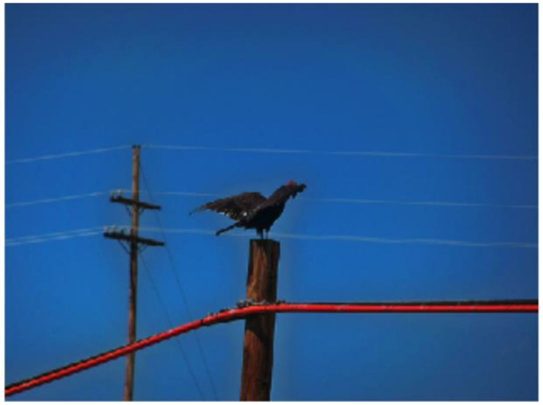
[RA+SS]
a tall tower with a clock on it

human:
1) a tall tower with a clock stands above a winter sky
2) there is a tree next to the clock tower
3) a large clock tower on a cloudy winter day
4) there is a clock in the center of a tower
5) a tower that will have a large clock at the top

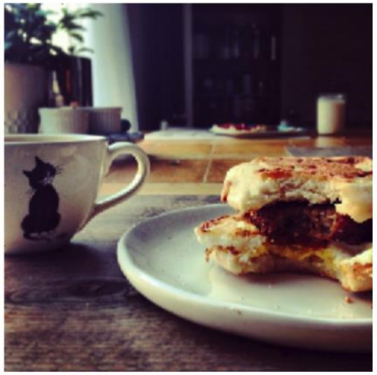
Example Captions



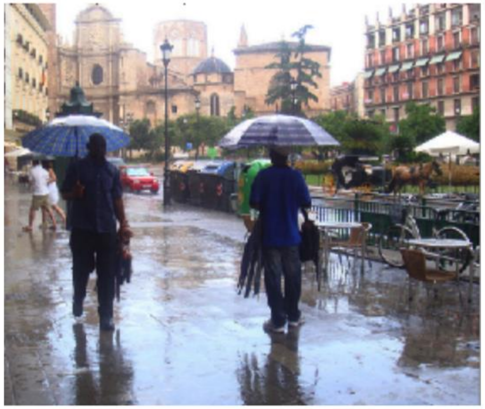
a man sitting at a table using a laptop



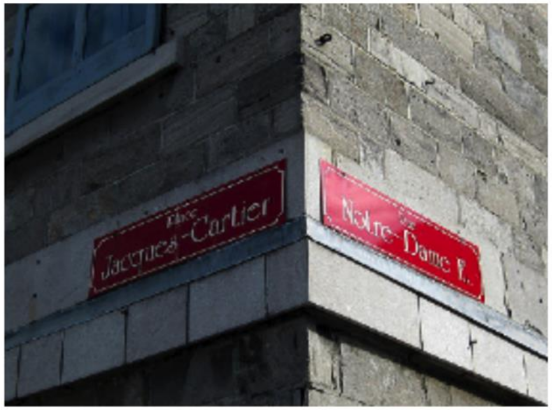
a bird perched on top of a power pole



a white plate topped with a half eaten sandwich



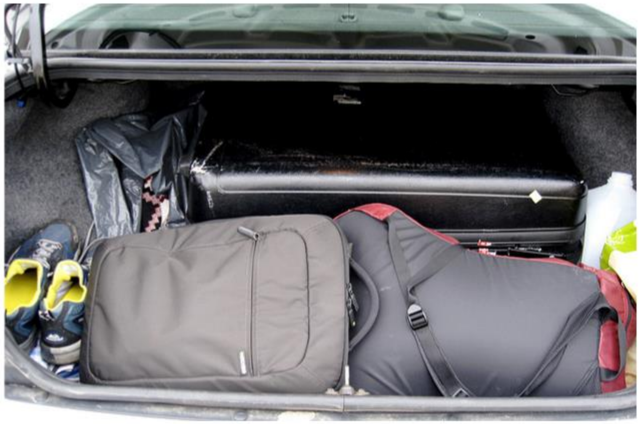
a group of people walking in the rain with umbrellas



a street sign on the corner of a building

Example Captions

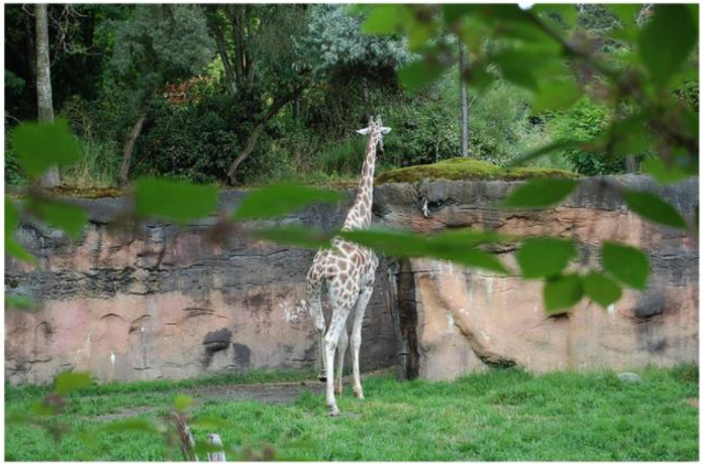
- SS helps to give high level scene information



[SS] a pile of luggage sitting in the back of a car ✓

[RA] a black suitcase with a bunch of luggage

[RA+SS] a luggage bag filled with lots of luggage



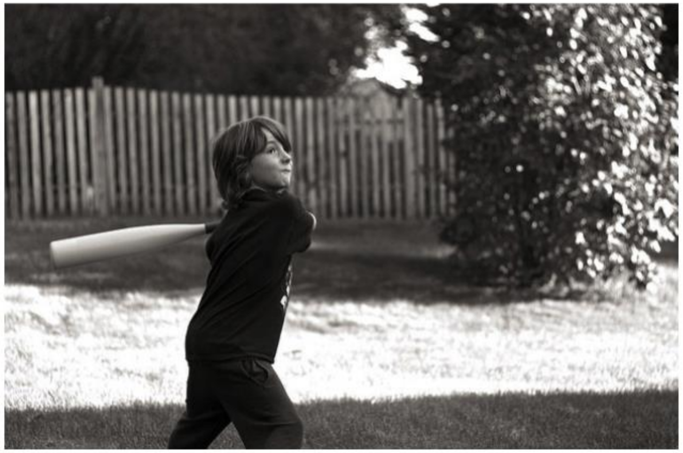
[SS] a giraffe standing in the middle of a field ✓

[RA] a giraffe standing next to a tall tree

[RA+SS] a giraffe standing in the grass near a tree

Example Captions

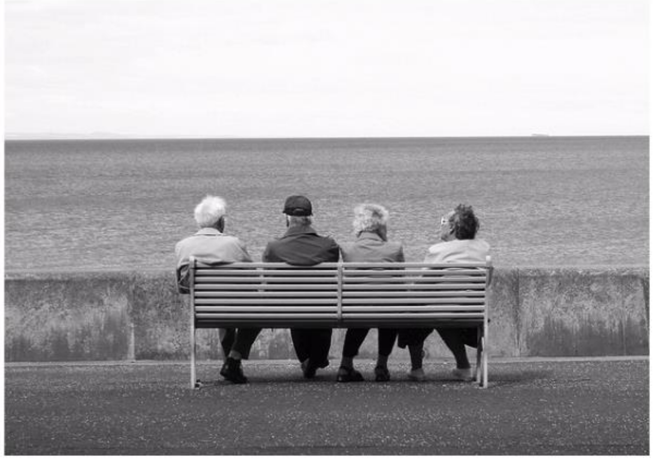
- RA better captures details



[SS] a young man is throwing a frisbee in a park

[RA] a young boy playing baseball on a field ✅

[RA+SS] a young boy is throwing a baseball bat



[SS] a person sitting on a bench near a body of water

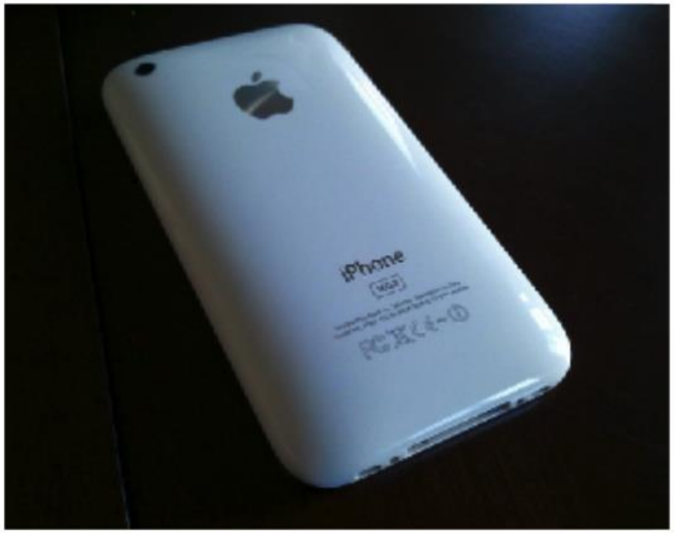
[RA] a group of people standing next to a bench

[RA+SS] a group of people sitting on top of a bench ✅

Example Captions

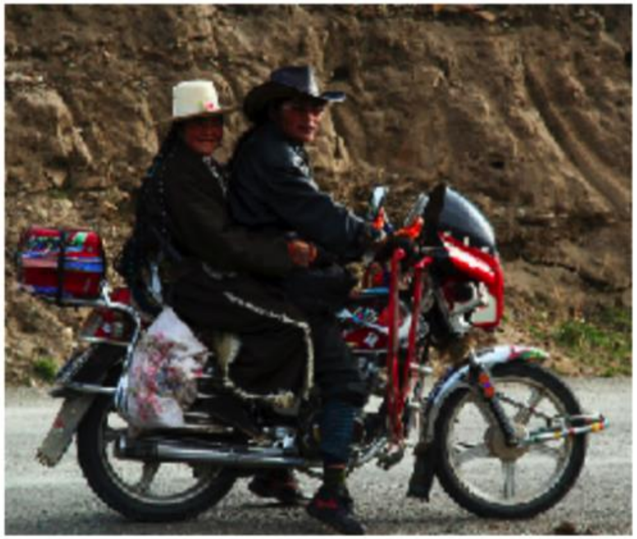
- Bad cases

wrong object



a wii controller sitting on top of a table

wrong counting

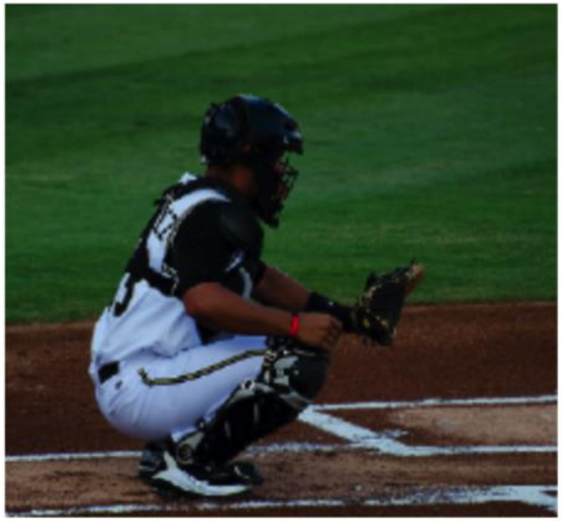


a man riding a motorcycle on a road

Example Captions

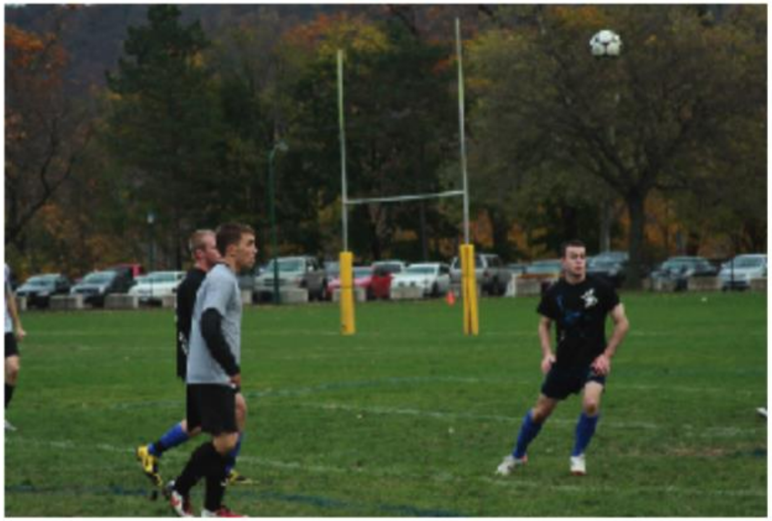
- Bad cases

wrong action



a baseball player holding a bat on a field

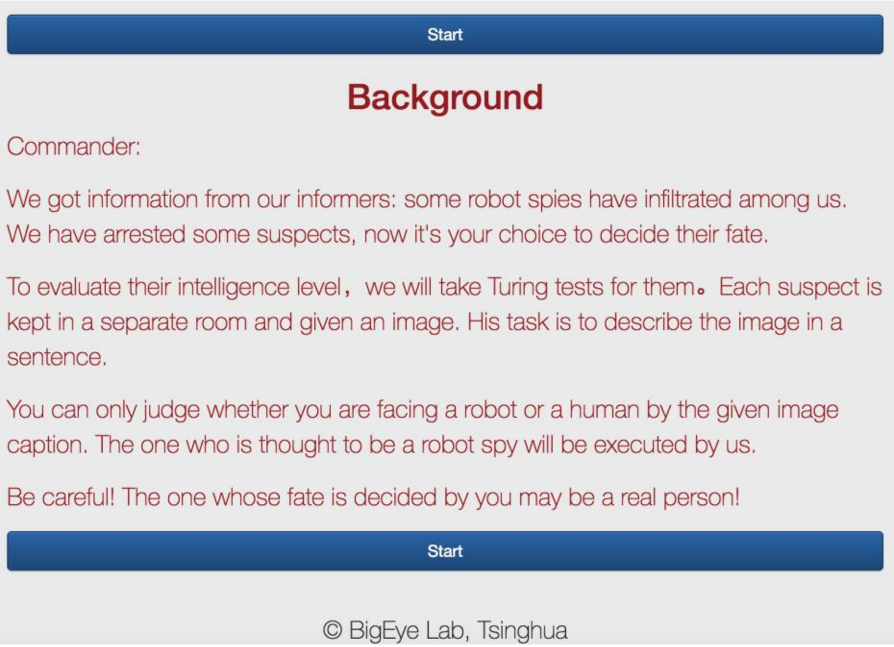
wrong scene



a group of young men playing a game of frisbee

Evaluation — Turing Test

- Given an image and a sentence, can you tell whether the sentence is generated by machine?



Background

Commander:

We got information from our informers: some robot spies have infiltrated among us. We have arrested some suspects, now it's your choice to decide their fate.

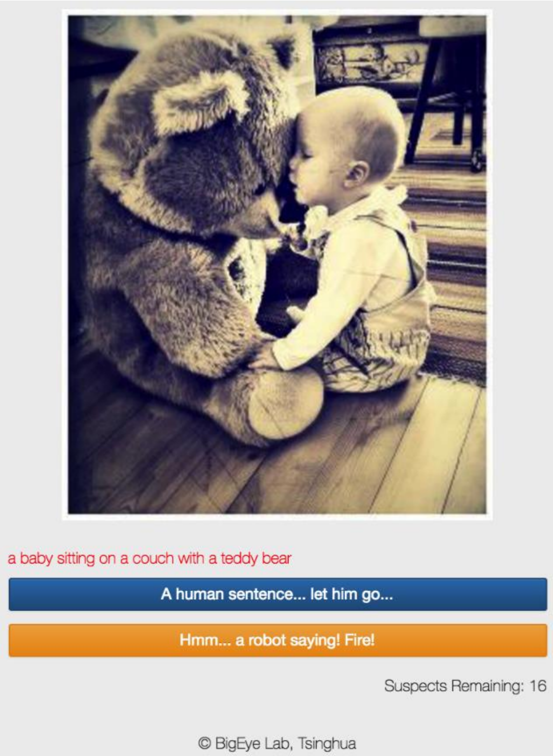
To evaluate their intelligence level， we will take Turing tests for them。 Each suspect is kept in a separate room and given an image. His task is to describe the image in a sentence.

You can only judge whether you are facing a robot or a human by the given image caption. The one who is thought to be a robot spy will be executed by us.

Be careful! The one whose fate is decided by you may be a real person!

© BigEye Lab, Tsinghua

u.c



a baby sitting on a couch with a teddy bear

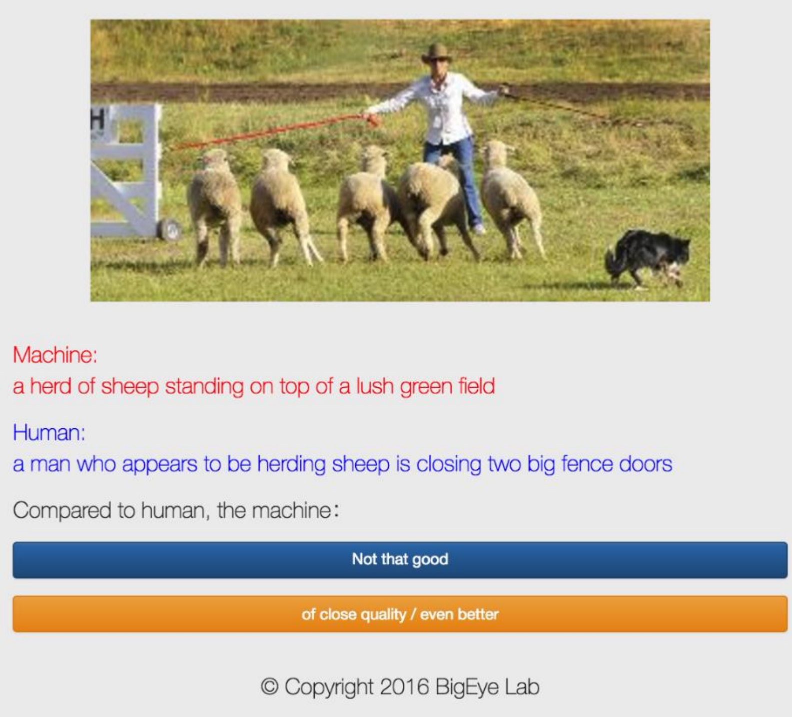
A human sentence... let him go...
Hmm... a robot saying! Fire!

Suspects Remaining: 16

© BigEye Lab, Tsinghua

Evaluation — Human Rating

- What do you think of machine sentence, compared to human sentence?
  - Try: http://bigeye.au.tsinghua.edu.cn:20829/



Machine:
a herd of sheep standing on top of a lush green field

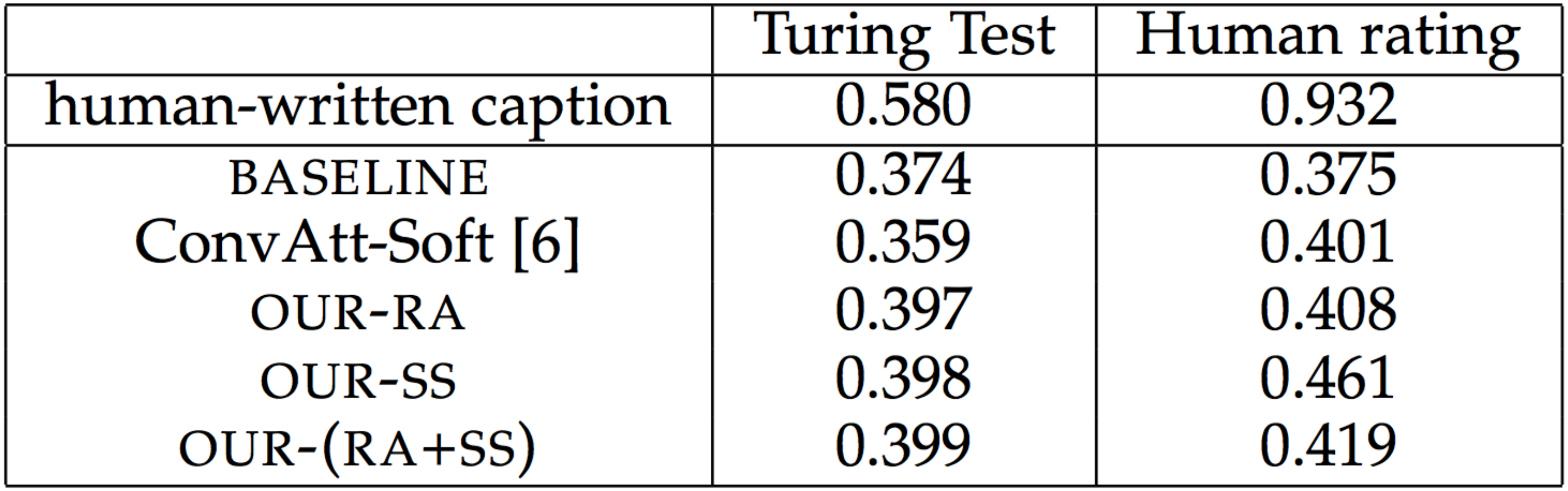
Human:
a man who appears to be herding sheep is closing two big fence doors

Compared to human, the machine:

- 0: machine not good
- 1: equal / machine bett

Results — Turing Test and Human Rating

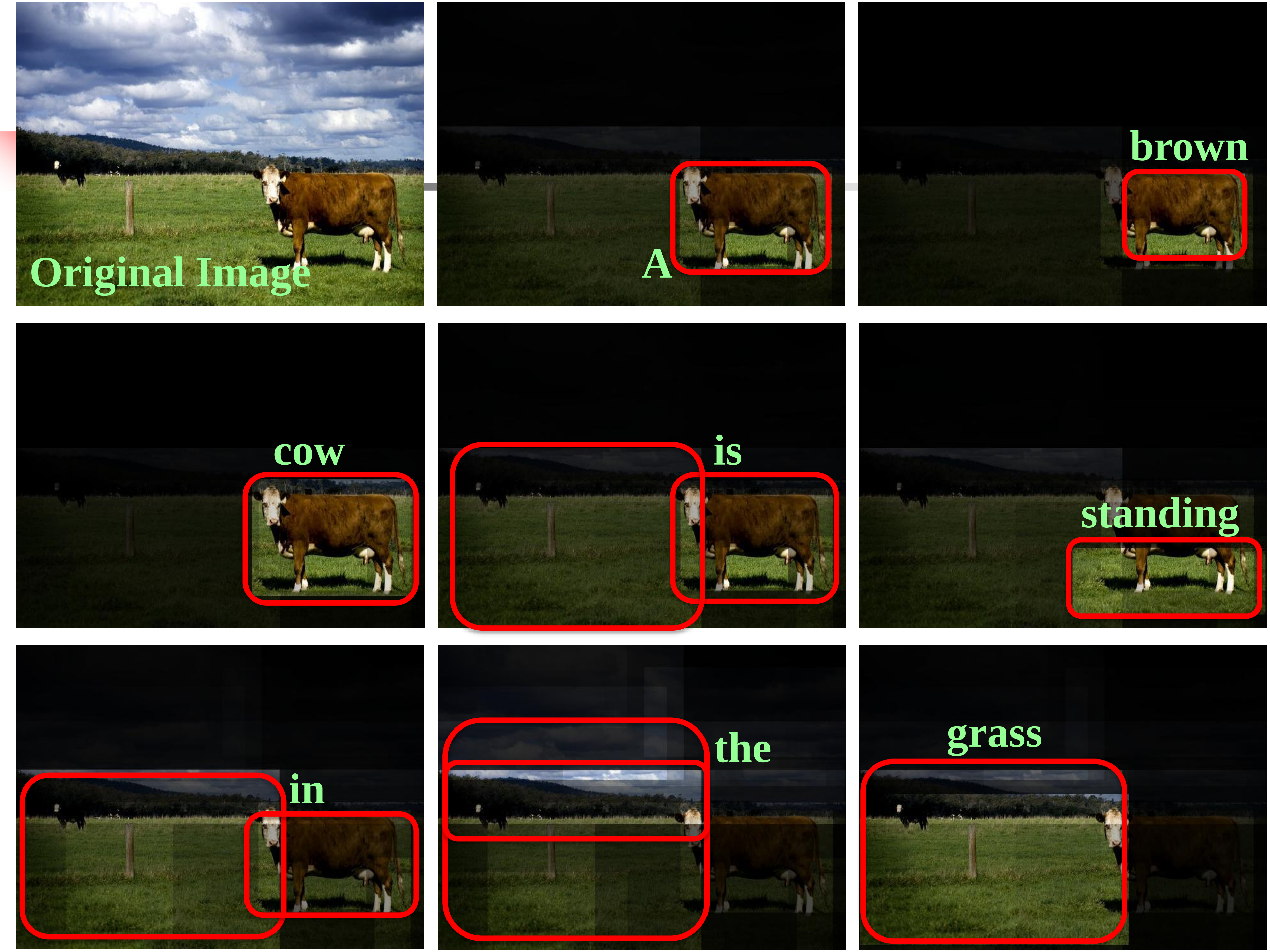
- Turing Test: collected 45380 judgements from 1900 IP addresses
- Human rating: collected 6000 judgements



	Turing Test	Human rating
human-written caption	0.580	0.932
BASELINE	0.374	0.375
ConvAtt-Soft [6]	0.359	0.401
OUR-RA	0.397	0.408
OUR-SS	0.398	0.461
OUR-(RA+SS)	0.399	0.419

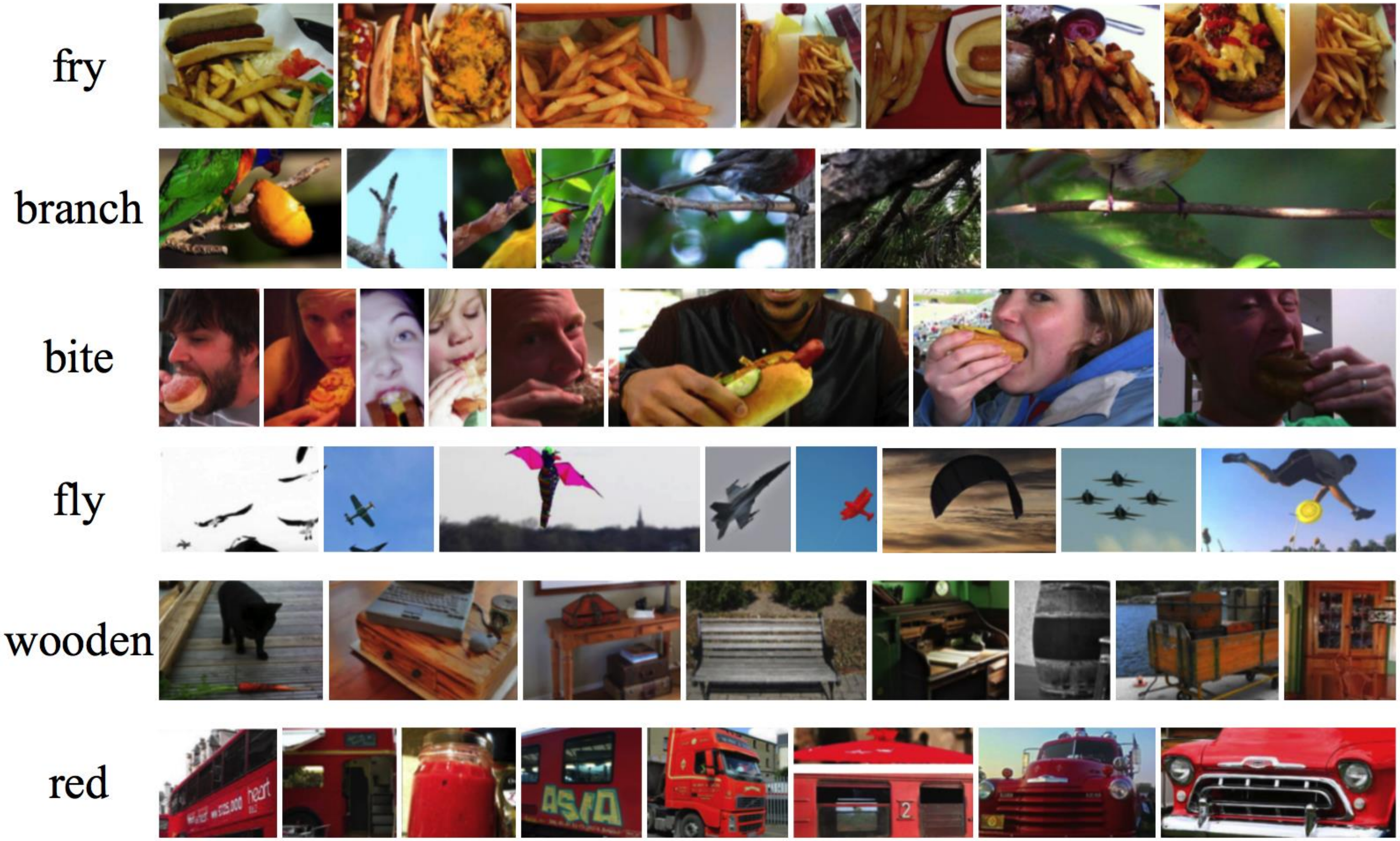
Attention Visualization

- Brighter parts correspond to larger attention weights
- Let’ s


Original Image
A
brown
cow
is
standing
in
the
grass

Region-word Matching

- Match a word to the region with maximum



Our paper

- http://arxiv.org/abs/1506.06272
- Kun Fu, Junqi Jin, Runpeng Cui, Fei Sha, Changshui Zhang. Aligning where to see and what to tell: image captioning with region-based attention and scene-specific contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2017

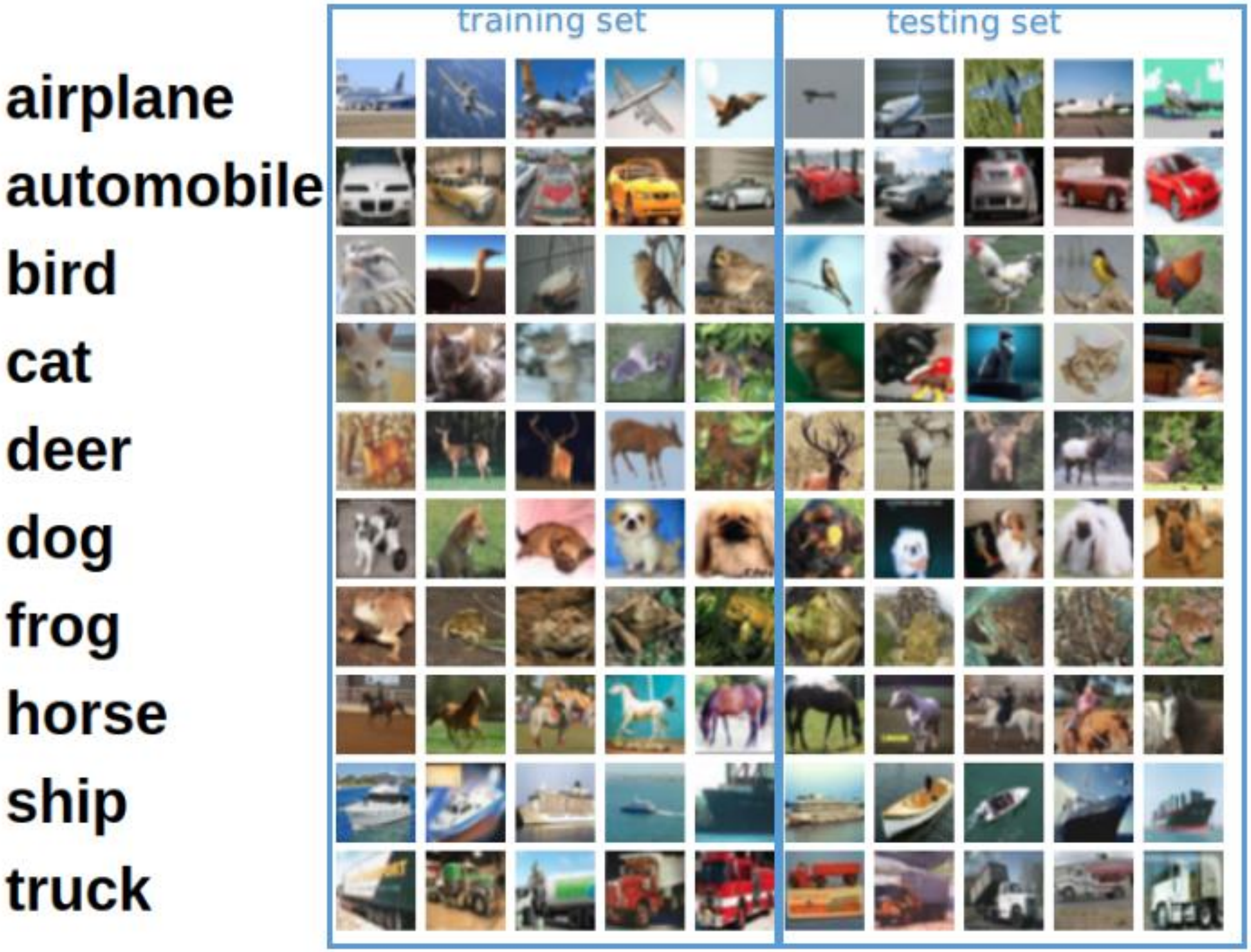
Soft Platform

- Caffee: Yangqing Jia
- TensorFlow: Google
- Caffee2, Torch: FaceBook
- Paddle: Baidu
- CNTK: MicroSoft

Challenges?

图像识别

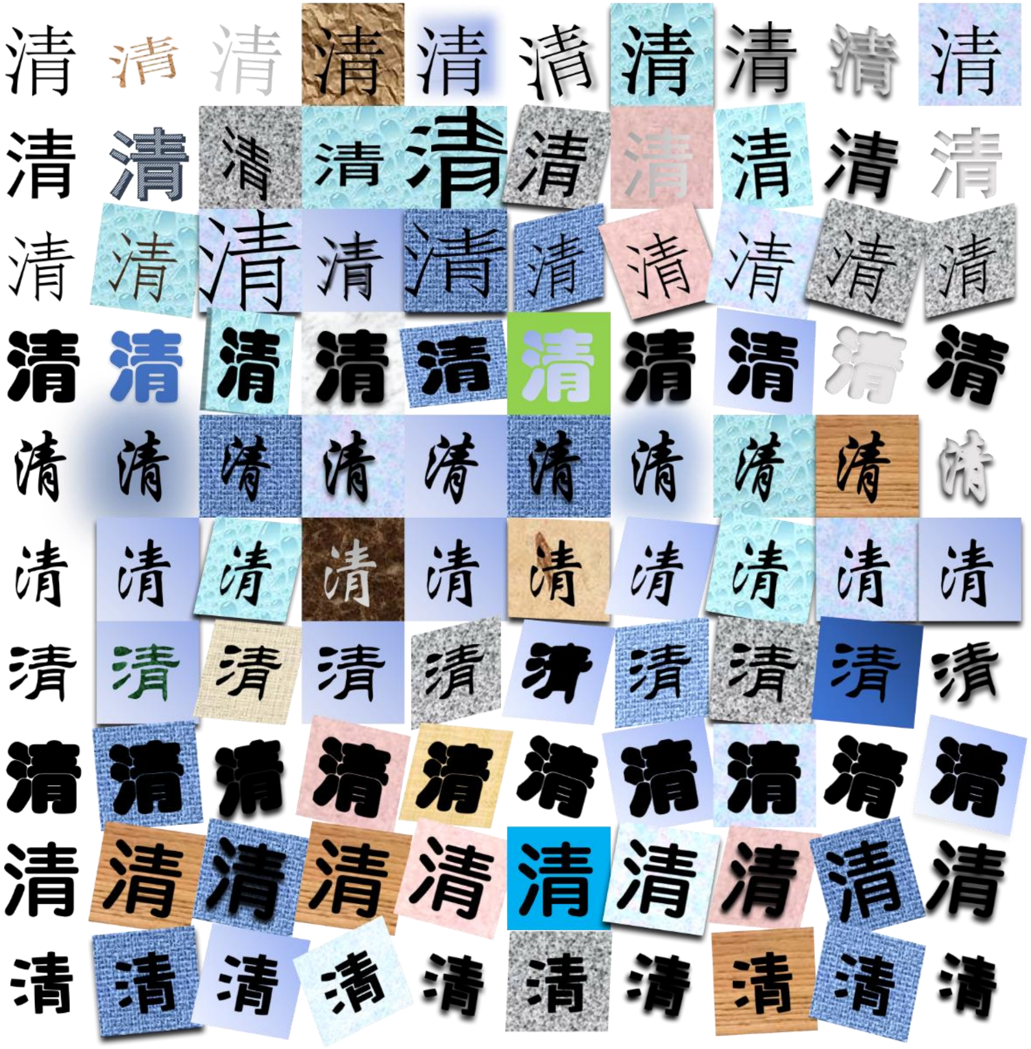
- 大量数据



大训练样本数据量下的图像分类任务
（图片引自于CIFAR-10数据集）

图像识别

- 大量的样本

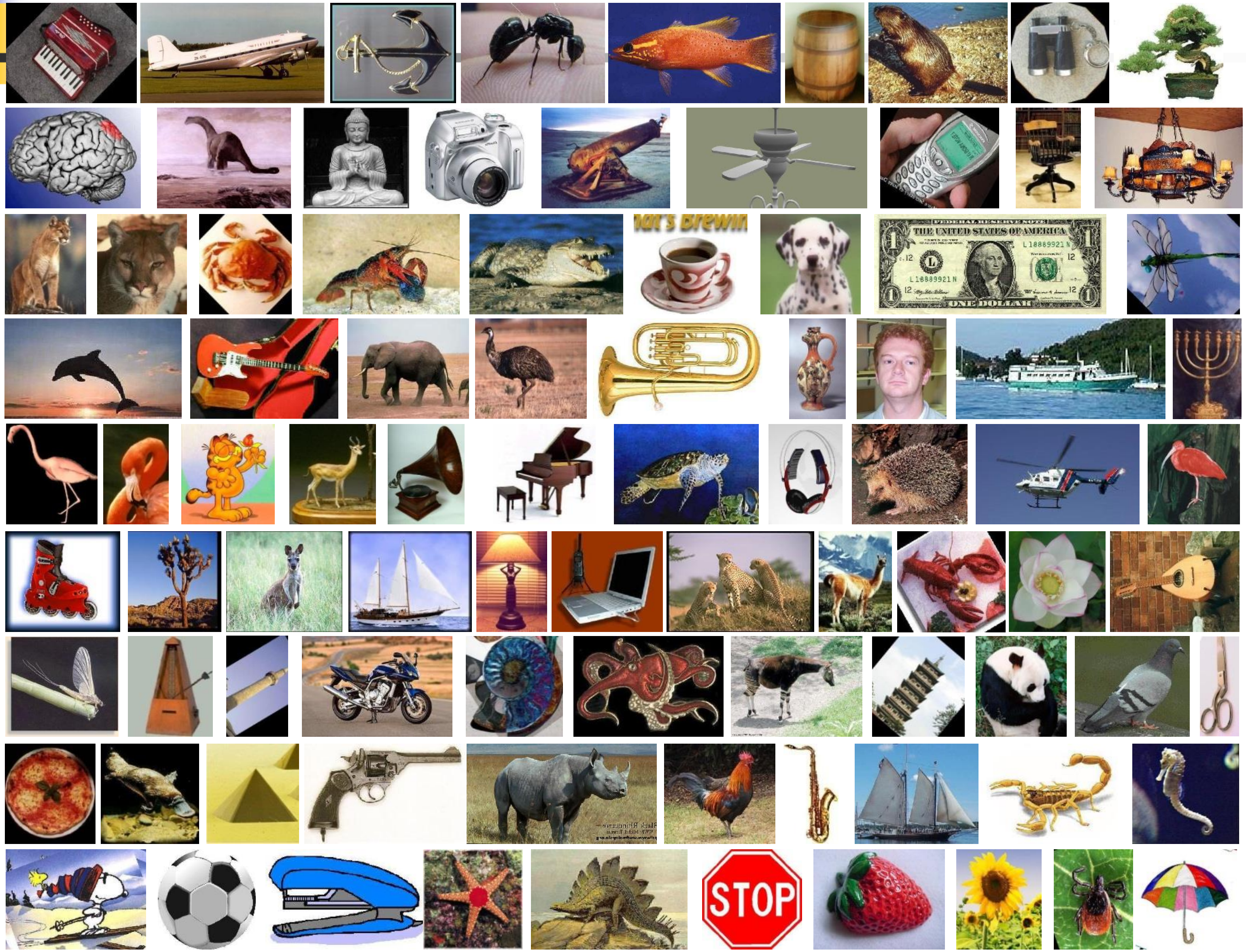


图像识别

困难

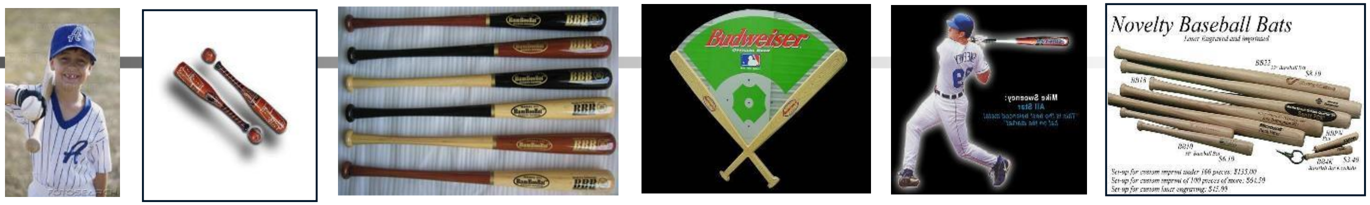
- 样本的获取
- 样本的标注
- 大数据量的训练
- 容易被攻击
- …

Caltech 101 images

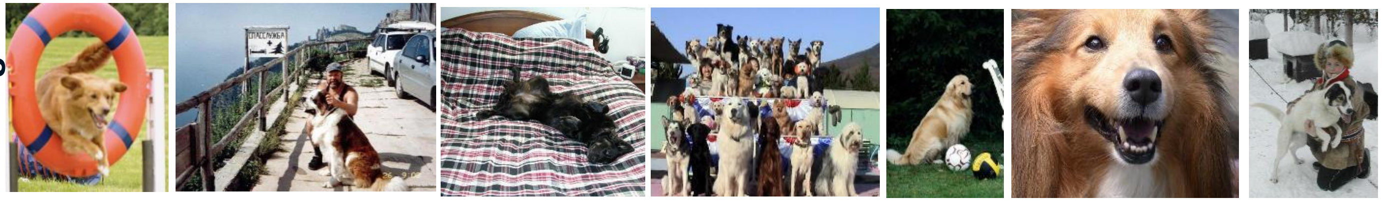


Caltech 256 images

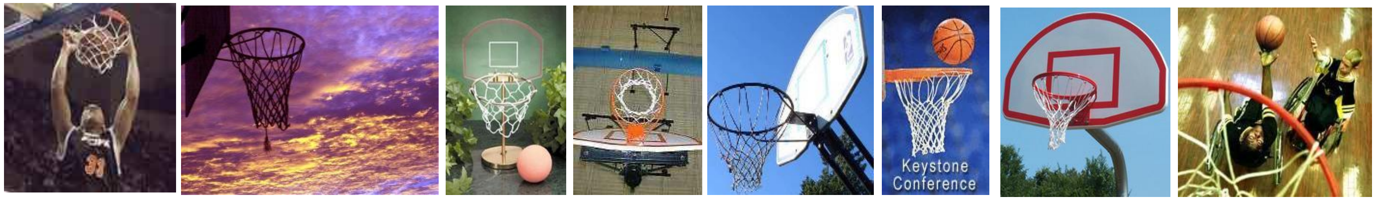
baseball-bat



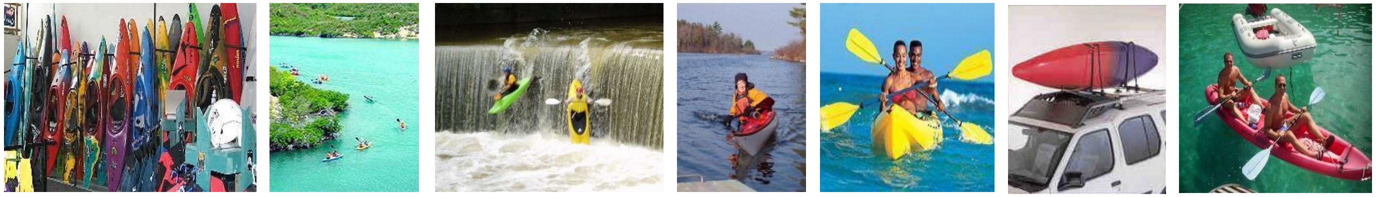
dog



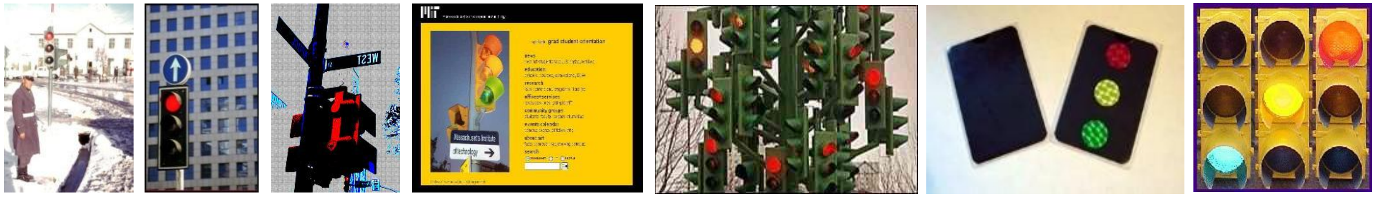
basketball-hoop



kayac



traffic light



Li FeiFei CVPR 2009

- 80,000 classes, 500-1000 images/class

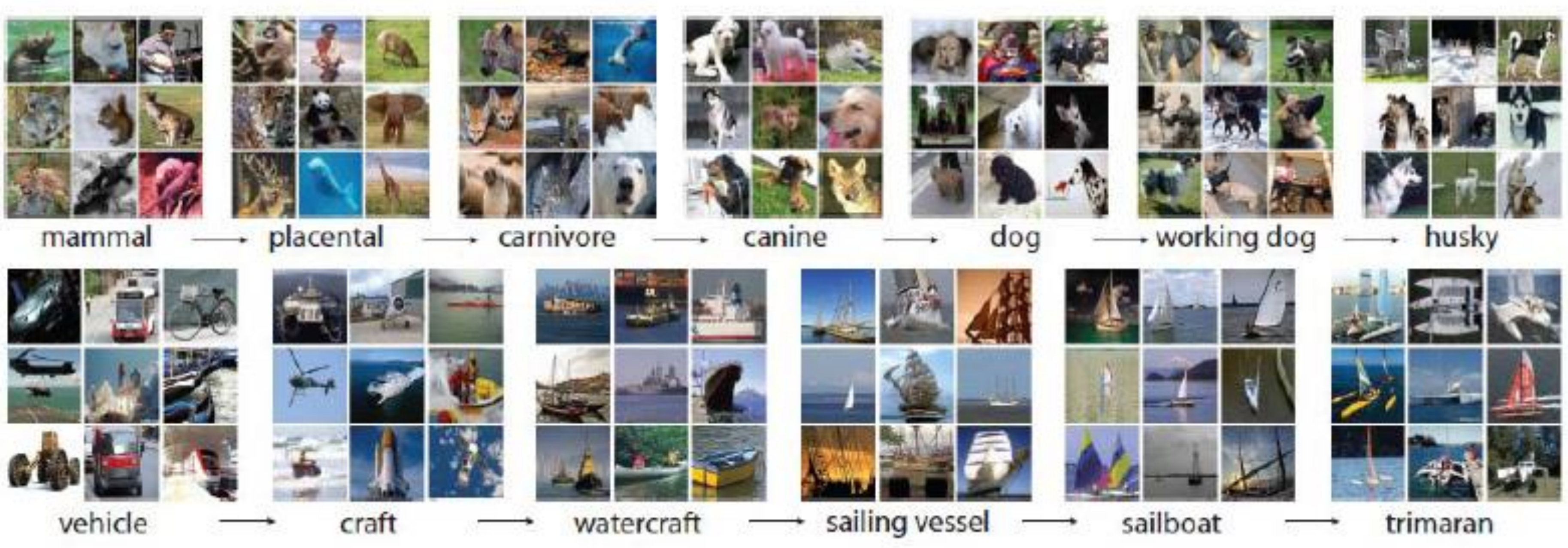

mammal → placental → carnivore → canine → dog → working dog → husky
vehicle → craft → watercraft → sailing vessel → sailboat → trimaran

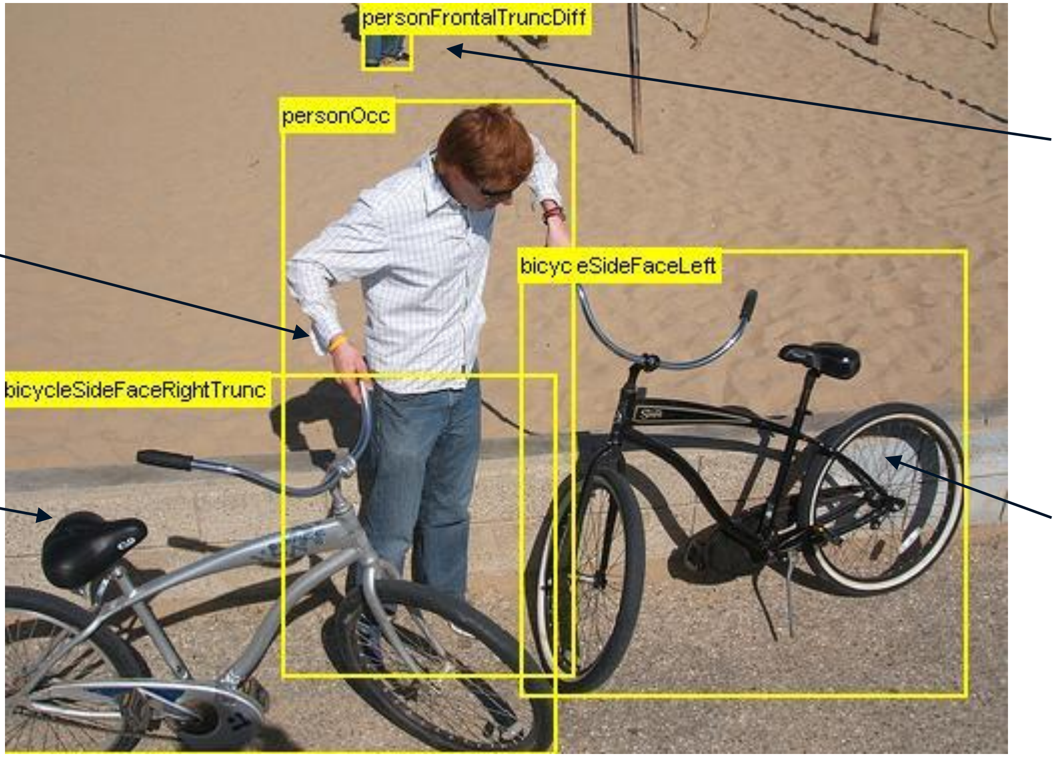
Figure 1: A snapshot of two root-to-leaf branches of ImageNet: the top row is from the mammal subtree; the bottom row is from the vehicle subtree. For each synset, 9 randomly sampled images are presented.

图像识别

图像获取

PASCAL Data Set

- Complete annotation of all objects
- Annotated in one session with written guidelines



Occluded
Object is significantly occluded within BB

Difficult
Not scored in evaluation

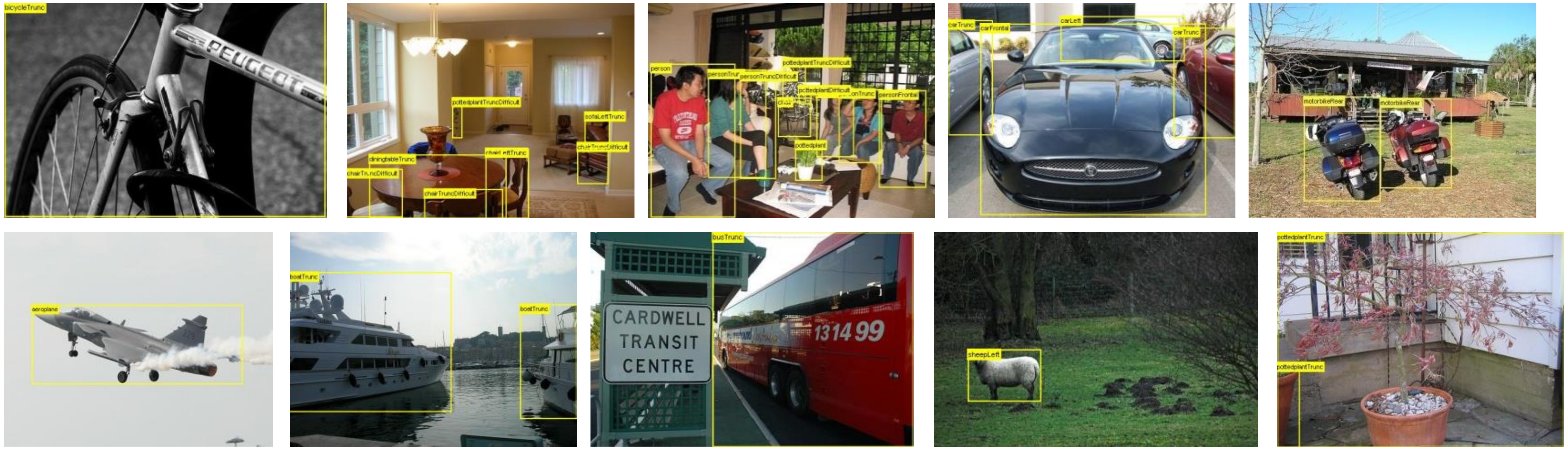
Truncated
Object extends beyond BB

Pose
Facing left

Dataset Content

- 20 classes: aeroplane, bicycle, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, train, TV

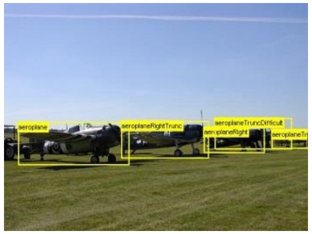
- Real images downloaded from flickr, not filtered for "quality"



- **Complex scenes, scale, pose, lighting, occlusion, ...**

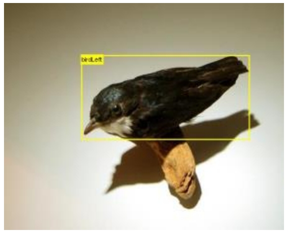
Examples

Aeroplane

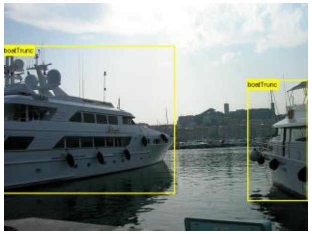


Bicycle

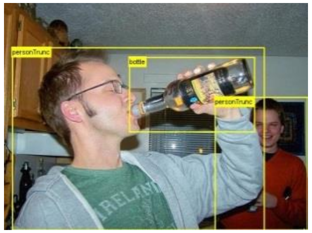
Bird

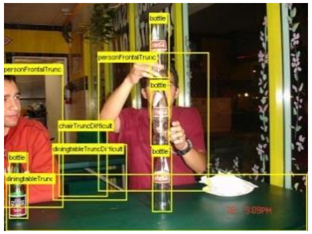


Boat

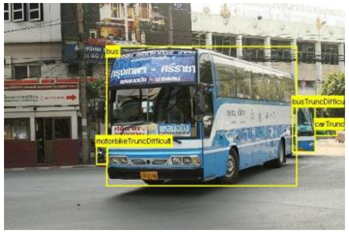


Bottle



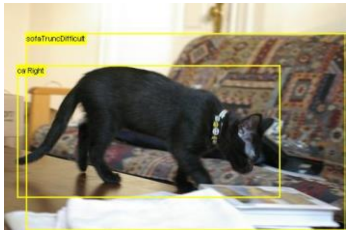


Bus

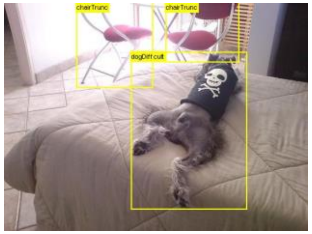


Car

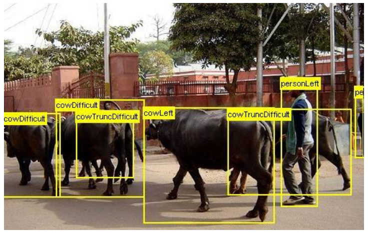
Cat



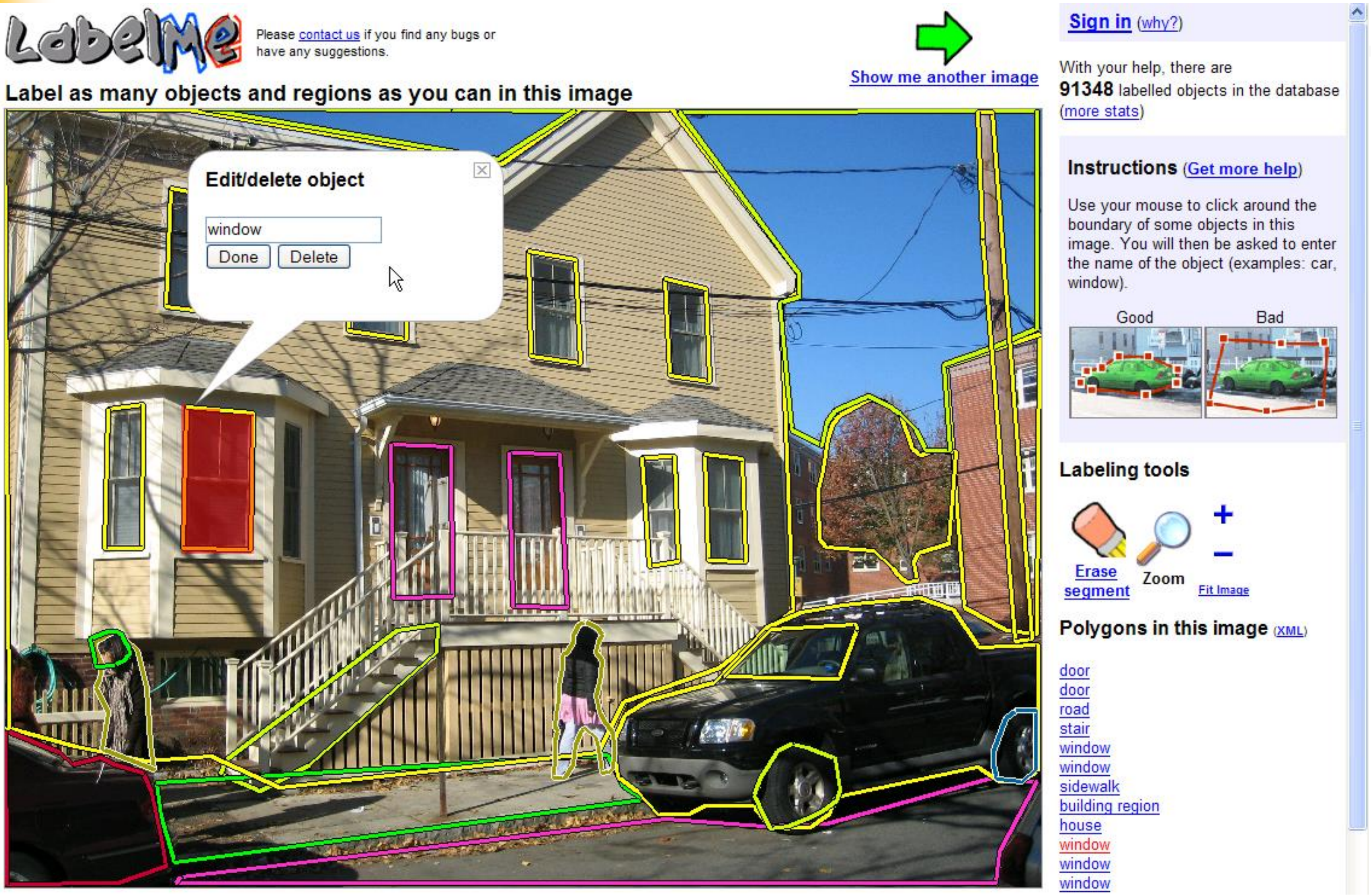
Chair



Cow



LabelMe



Please contact us if you find any bugs or have any suggestions.

Show me another image

Label as many objects and regions as you can in this image

Edit/delete object

window

Done Delete

Sign in (why?)

With your help, there are **91348** labelled objects in the database (more stats)

Instructions (Get more help)

Use your mouse to click around the boundary of some objects in this image. You will then be asked to enter the name of the object (examples: car, window).

Good Bad

Labeling tools

Erase segment Zoom + − Fit Image

Polygons in this image (XML)

- door
- door
- road
- stair
- window
- window
- sidewalk
- building region
- house
- window
- window
- window

Russell, Torralba, Freman, 2005

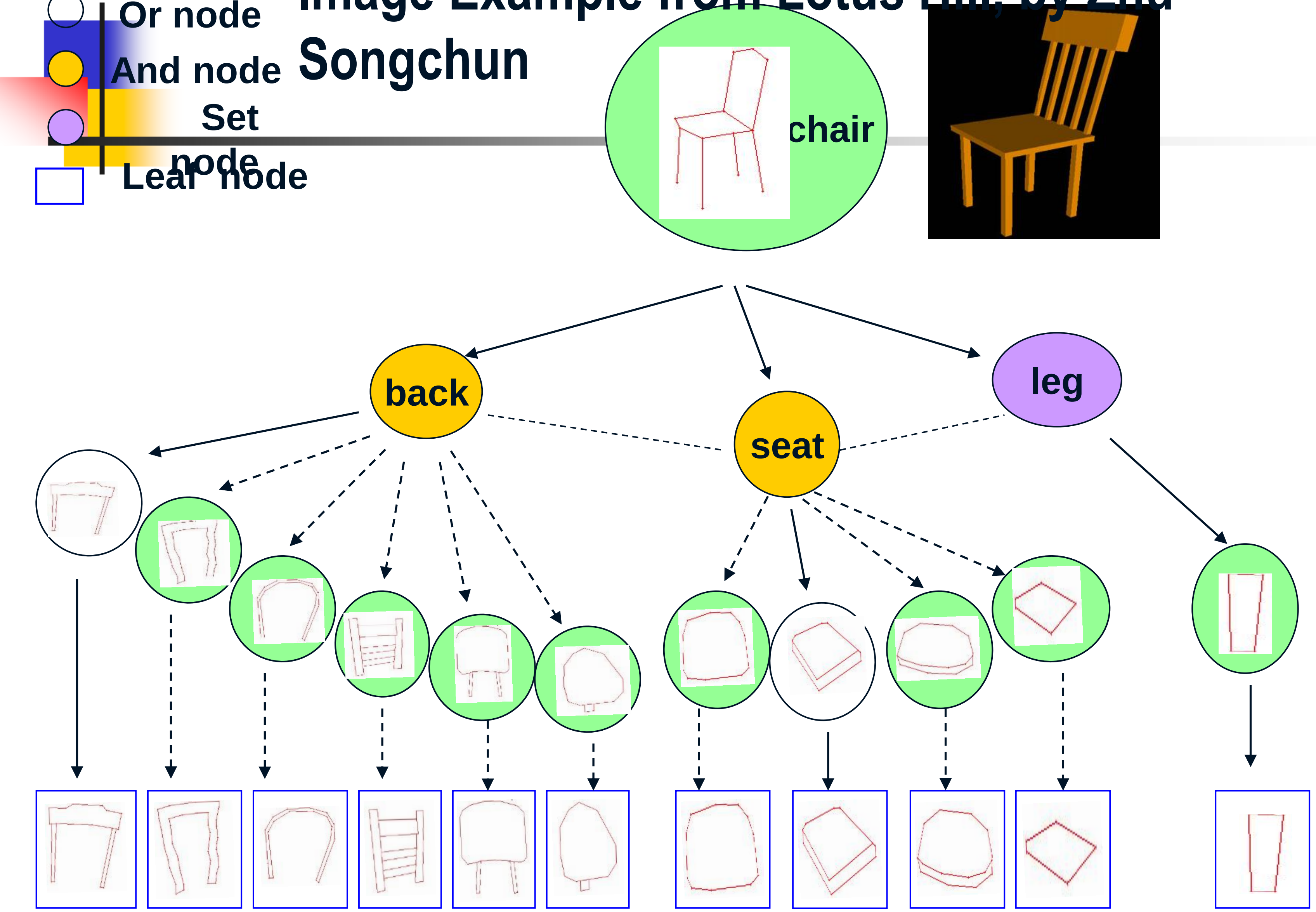
Image Example from Lotus Hill, by Zhu Songchun

○ Or node

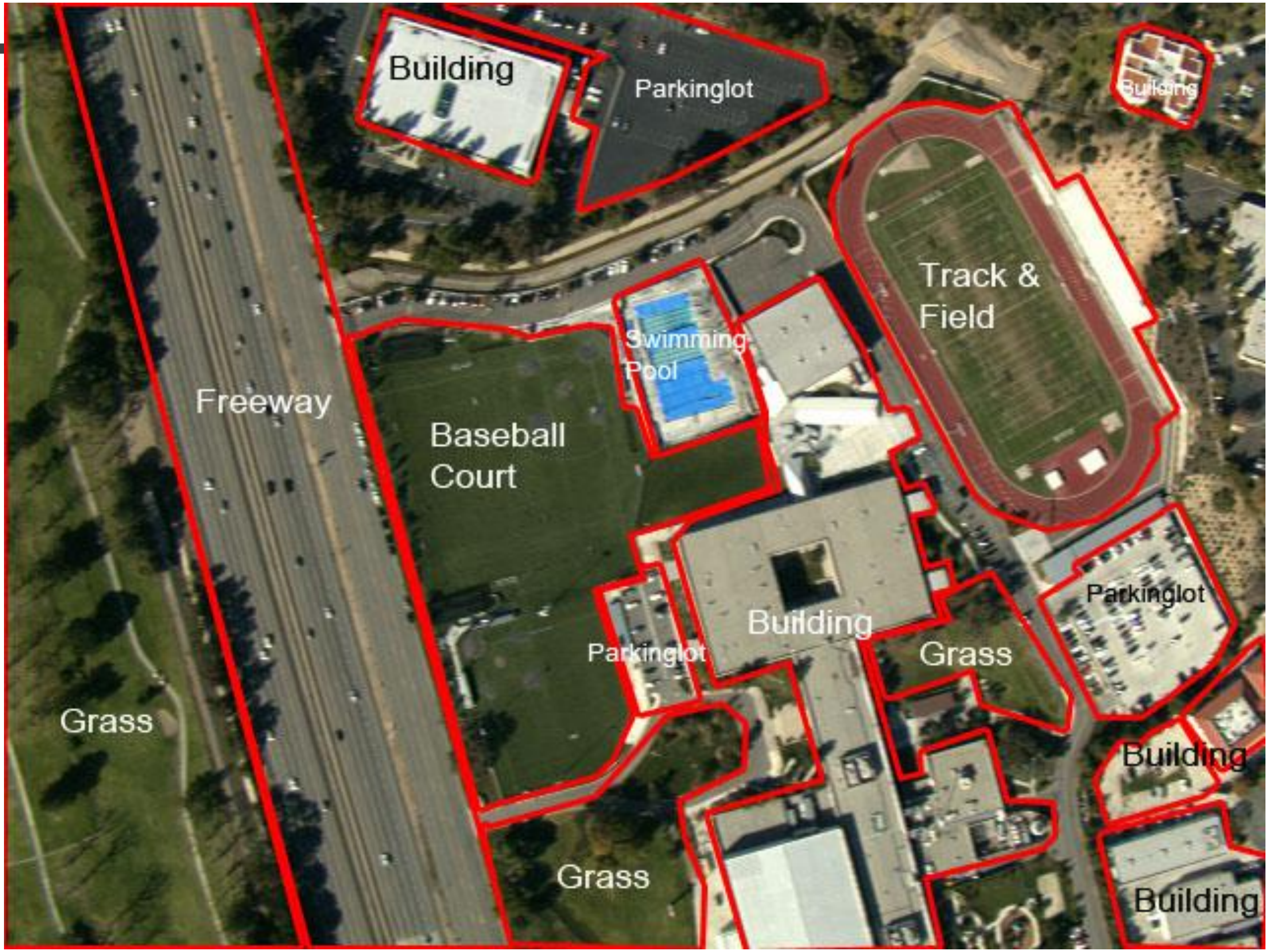
● And node

● Set node

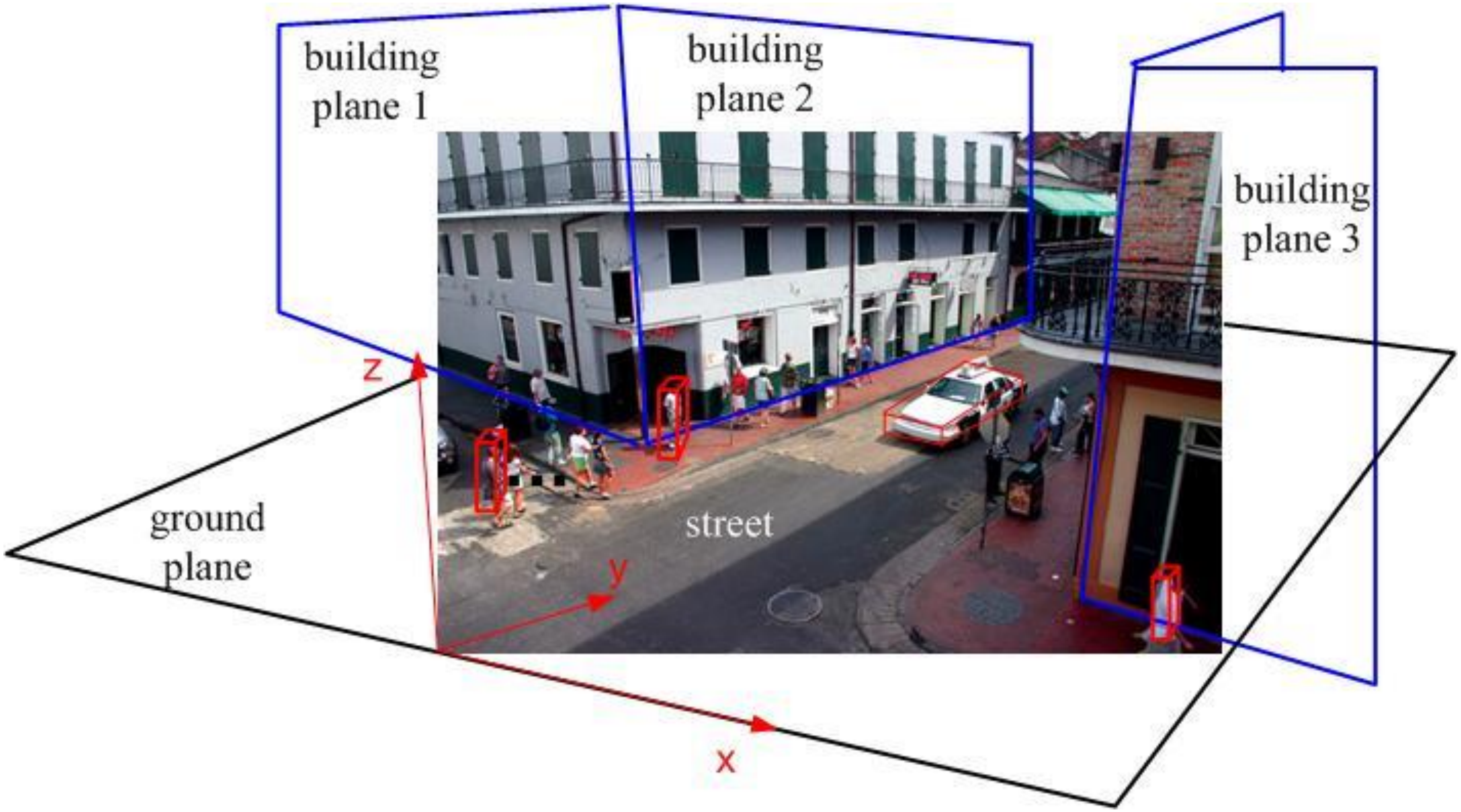
□ Leaf node



Scene example from Lotus Hill



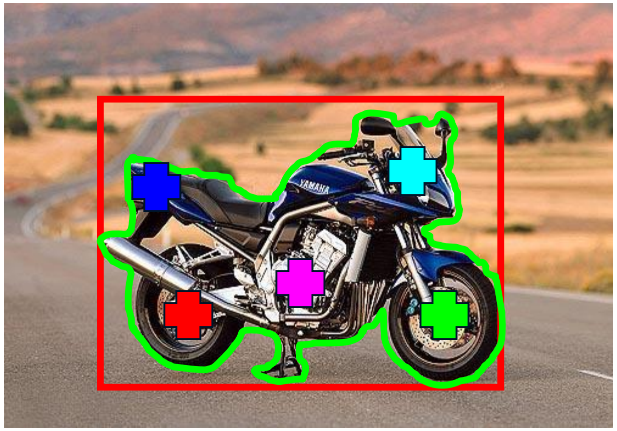
Scene example from Lotus Hill



Labeling

- Level of supervision
  - Manual segmentation; bounding box; image labels; noisy labels

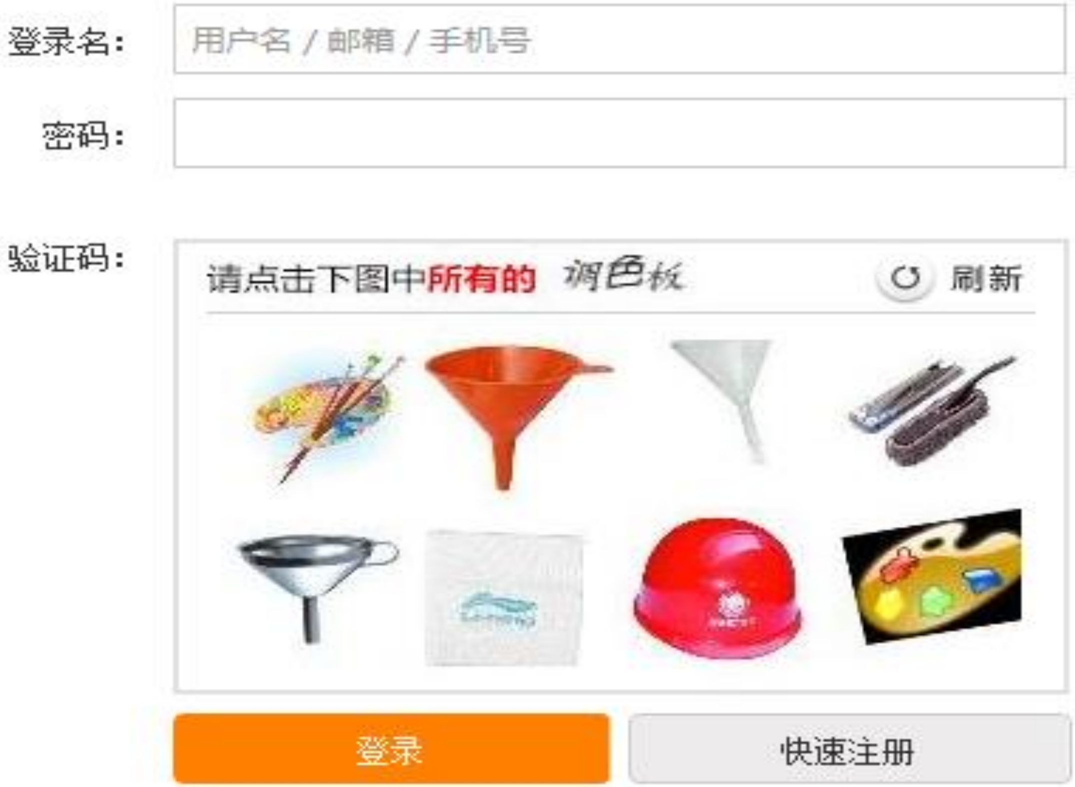
Contains a motorbike



图像识别

样本的标注

- 众包（Crowd sourcing）



登录名： 用户名 / 邮箱 / 手机号

密码：

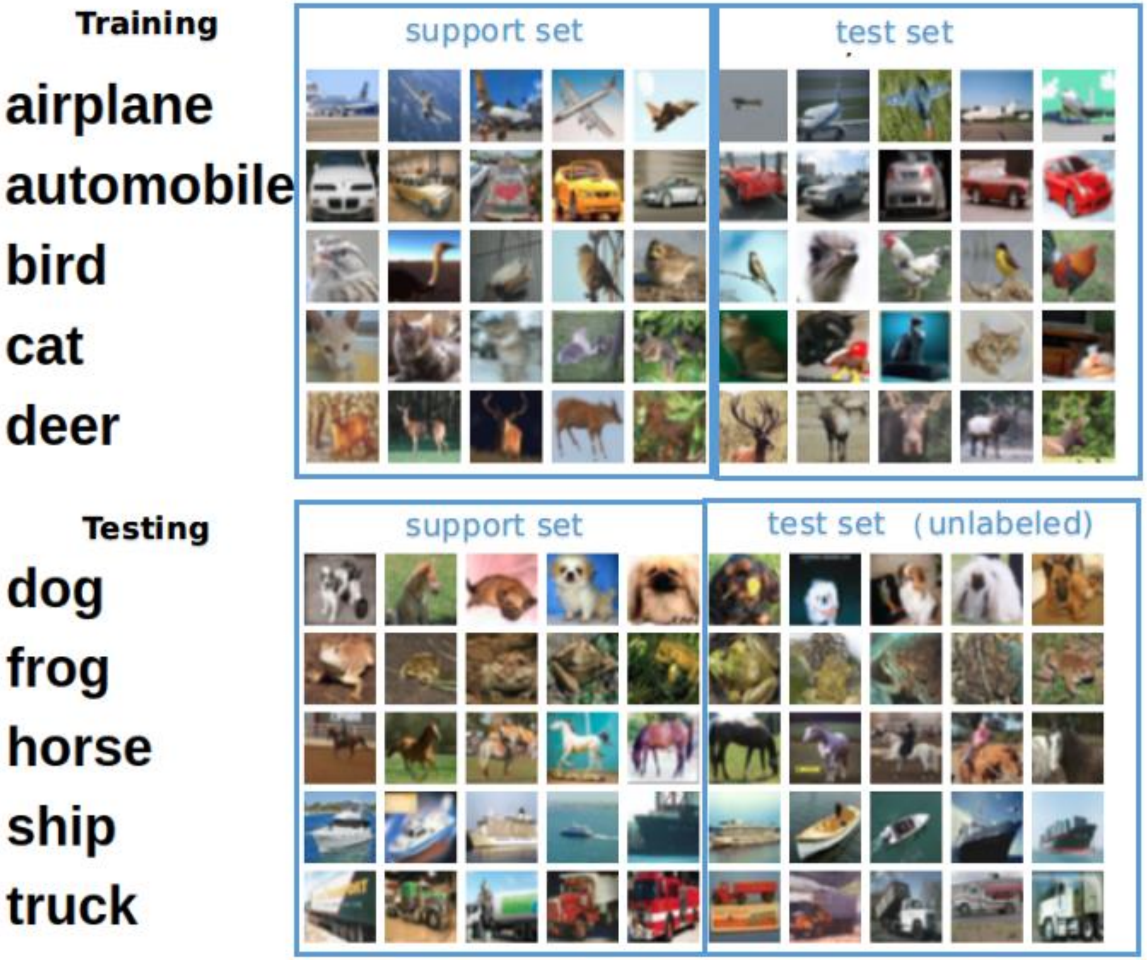
验证码：

请点击下图中**所有的** 调色板

图像识别

样本的标注

- A few shot learning
- One shot learning
- Zero shot learning



小样本数据量下的图像分类任务
（图片引自于CIFAR-10数据集合）

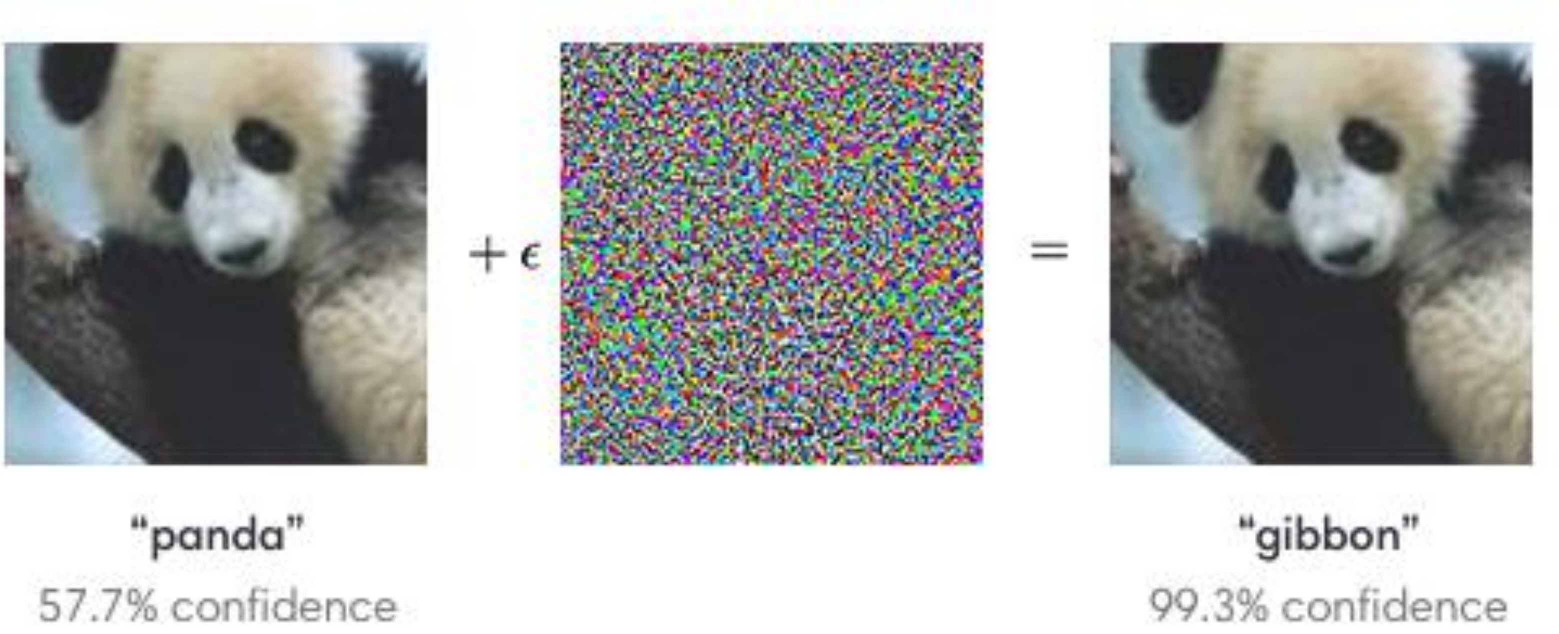
图像识别

样本的标注

- A few shot learning
- One shot learning
- Zero shot learning

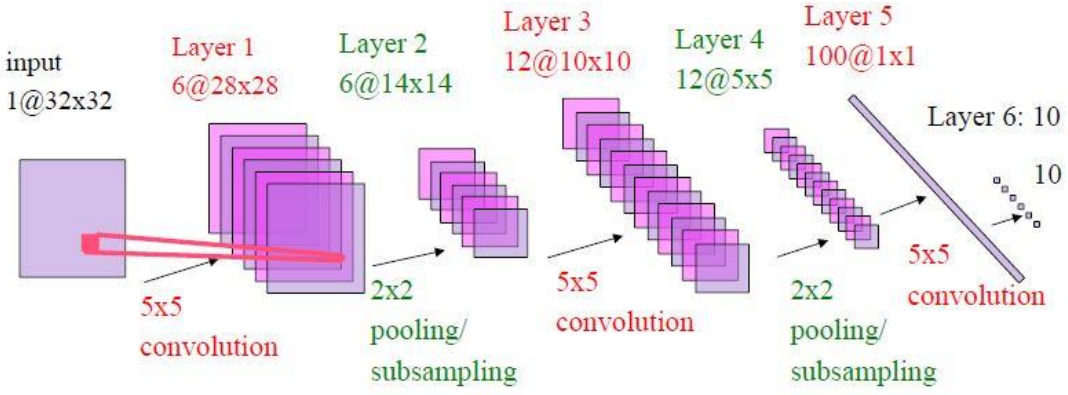
图像识别

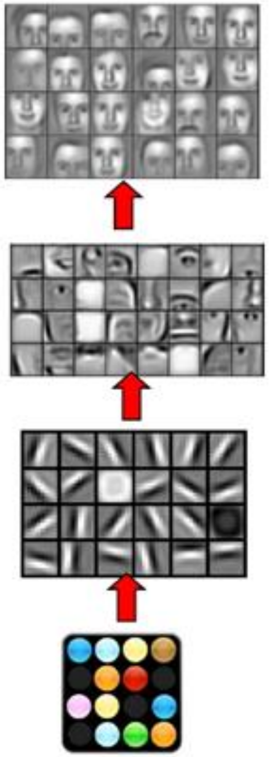
- 容易被攻击
- 机器学习的泛化问题



Interpretable Machine Learning

Neural Networks





Simple ConvNet for MNIST [LeCun 1998]，1993,1994,1995

谢　谢