FGVC: Fine Grained Visual Classification

Jinna Cui

Vision@OUC

May 25, 2017

Content

1. Image Classification
   - 1.1 Dataset
   - 1.2 Methods
2. FGVC
   - 1.1 Dataset
   - 1.2 Methods

Image Classification Introduction

What is image classification?

Image classification: an image processing method that distinguishes different kinds of targets according to different features reflected in the image information. （Baidu Encyclopedia）

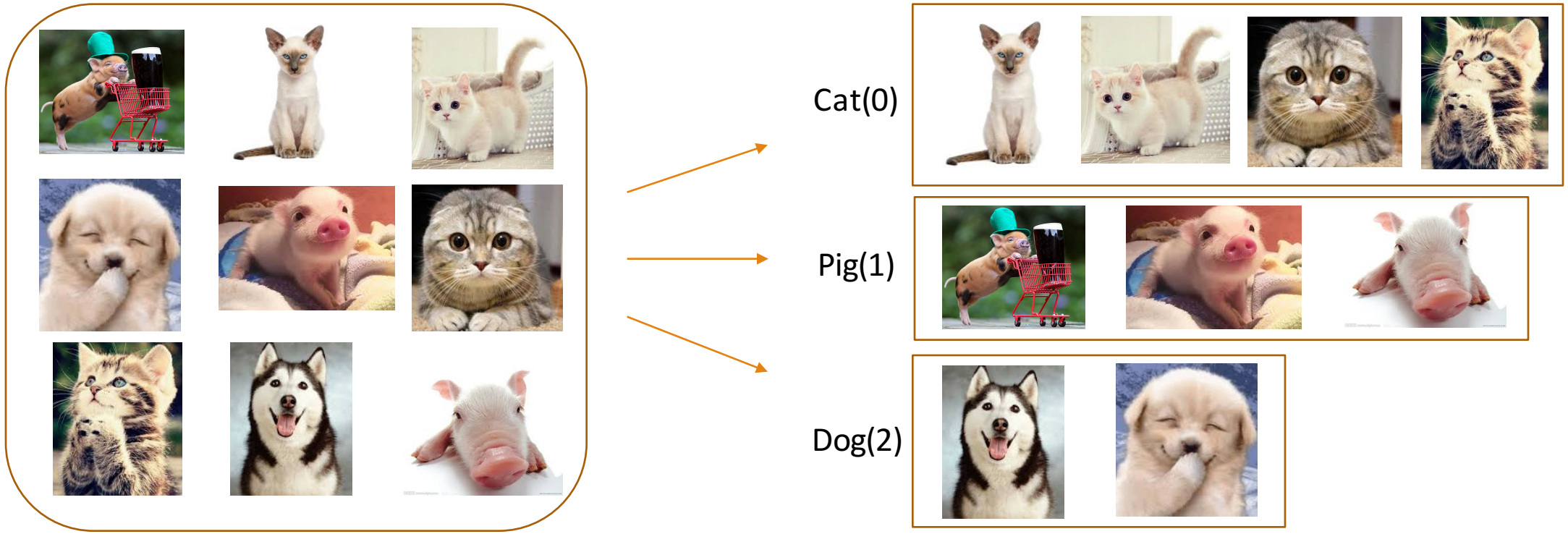


Image Classification

Application:

1. Public security system: Fingerprint identification & Face recognition
2. Medical system: Assisted medical care
3. Commercial: Trademark management, Online shopping, Searching…
4. Geography: Remote sensing image classification
5. Ecology: Species classification

…

Image Classification

What we see

(194, 153, 151)

194	230	250	254	232	206	211	232	229	179	189	186	158	132	132	135	144	153	147	144	162	86	49	65	30	31	41	26
213	243	235	215	189	188	227	238	234	233	181	135	128	112	129	138	136	127	139	139	116	103	60	57	58	68	55	39
237	243	241	209	217	247	226	242	217	153	132	146	120	128	151	165	153	152	154	127	108	120	134	39	57	34	18	55
253	245	238	203	244	222	214	226	168	122	128	121	102	100	149	187	177	167	182	123	109	115	117	89	42	43	49	33
230	233	205	236	235	196	206	203	121	147	112	90	94	111	152	188	186	191	202	118	100	89	104	129	75	54	35	44
220	212	202	241	227	183	192	166	138	150	154	75	53	119	176	181	174	179	195	84	86	93	95	150	131	58	65	36
182	184	196	242	201	169	187	132	152	175	176	88	29	150	190	193	163	177	150	53	102	154	126	144	142	90	39	34
187	158	211	235	193	168	225	135	118	175	161	172	143	170	192	219	208	186	181	124	157	148	188	186	167	87	37	27
168	167	195	221	181	152	182	158	75	131	163	172	187	177	173	177	179	179	194	179	183	189	190	191	158	114	66	58
161	157	154	179	157	165	171	175	114	80	110	128	126	137	157	169	160	167	153	149	161	174	180	148	97	146	74	58
148	168	142	145	148	202	186	161	158	105	94	100	101	92	102	132	127	130	106	121	127	104	96	87	132	173	84	58
154	180	167	158	158	178	173	145	136	127	93	97	85	55	83	75	64	81	68	76	87	74	67	103	236	178	110	61
128	151	168	169	161	141	124	104	101	97	106	88	74	82	93	84	113	124	100	66	53	63	107	200	236	175	86	48
92	91	114	127	107	84	89	156	182	183	175	153	145	148	160	153	150	142	132	136	126	103	122	175	233	184	109	70

What the computer sees

Image Classification

92% Pig

4.8% Cat

3.2% Dog

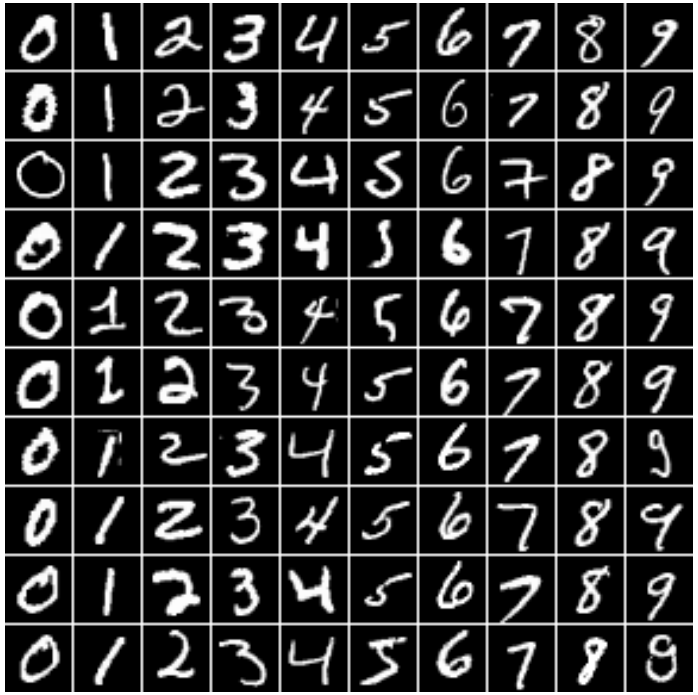
What the computer outputs

Image Classification Dataset

Datasets

MNIST(1998)：

Number of training images：60000
Number of testing images：10000
Number of categories: 10

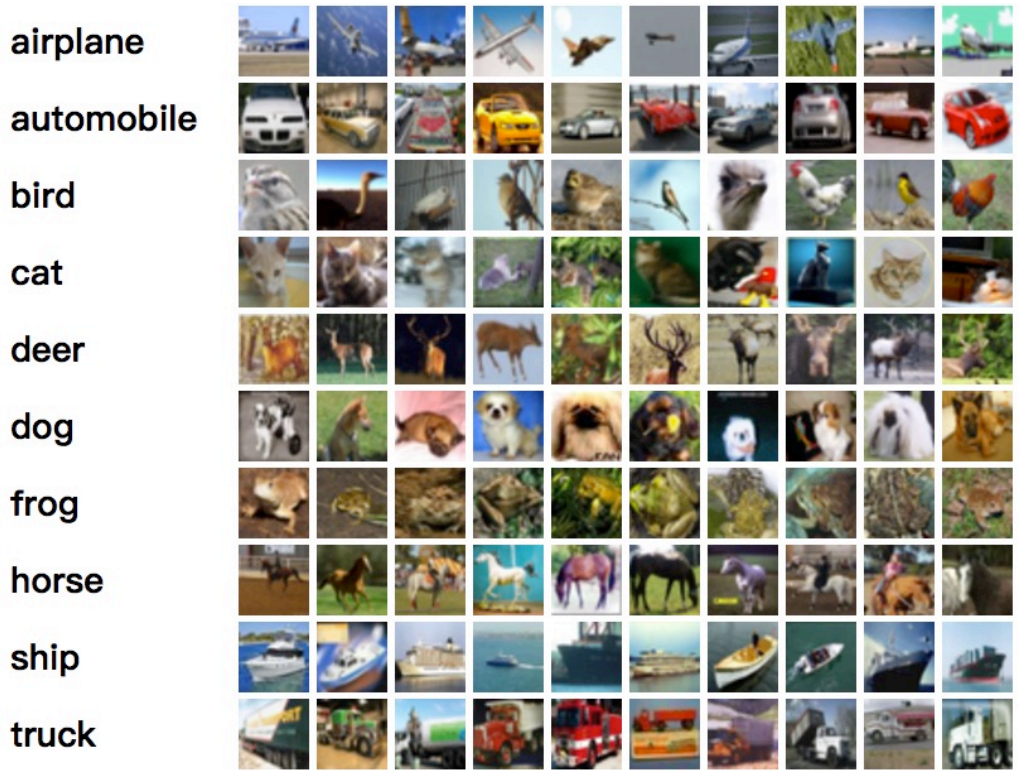


Datasets

CIFAR-10：

Number of training images： 50000
Number of testing images： 10000
Number of categories: 10

CIFAR-100:
Number of training images： 50000
Number of testing images： 10000
Number of categories: 100



https://www.cs.toronto.edu/~kriz/cifar.html

Datasets

Caltech-101:

Number of images：9144
Number of categories: 102

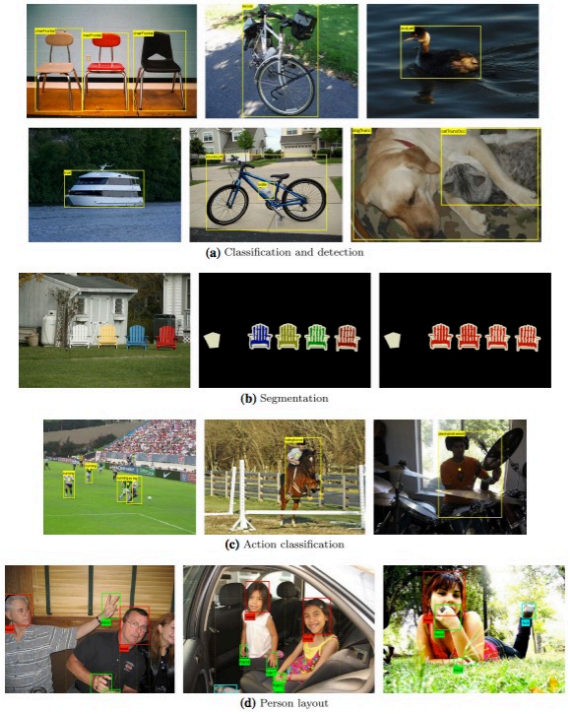
Caltech-256:

Number of images：30607
Number of categories: 257

Datasets

PASCAL VOC(2005-2012):

Number of images：11000
Number of categories: 20
Number of Object instances: 27000
Number of segmentation: 7000


(a) Classification and detection

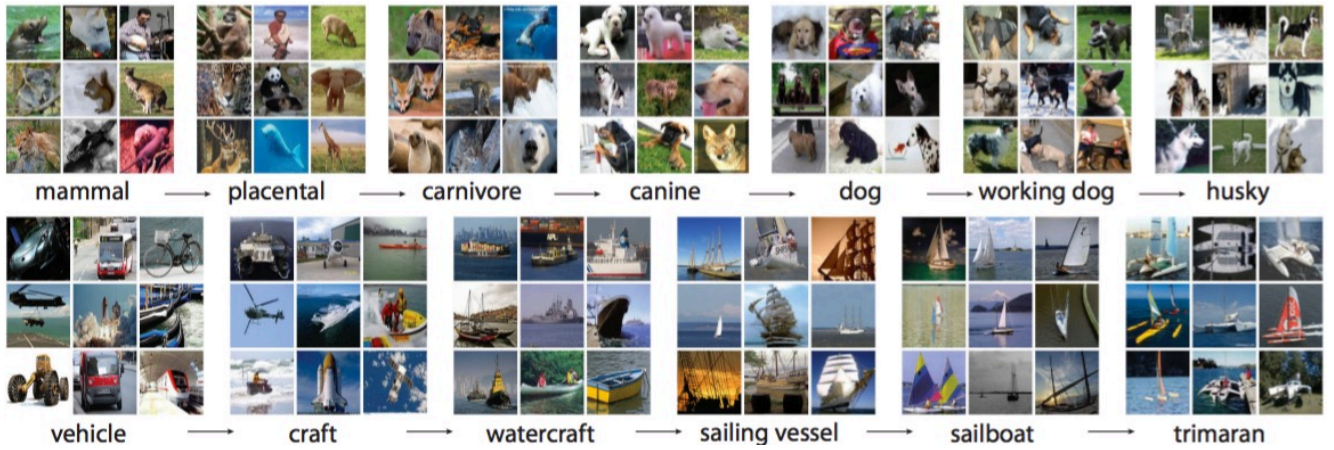
(b) Segmentation

(c) Action classification

(d) Person layout

Datasets

ImageNet-1000:



Number of training images： 1.3M
Number of testing images： 100K
Number of categories: 1000

http://www.image-net.org/about-stats

Datasets

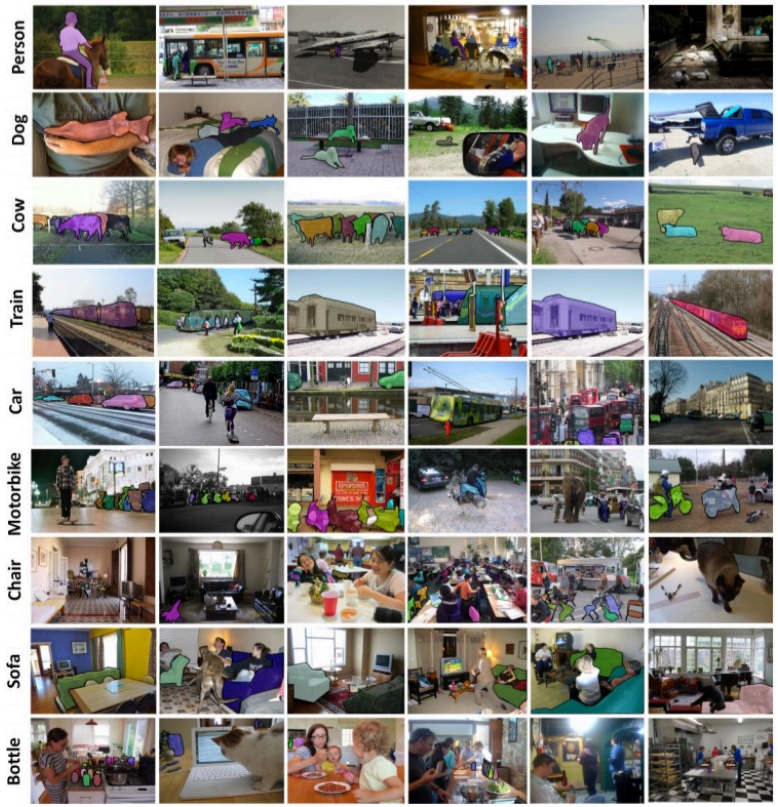
MS COCO:

Number of training images: 165482

Number of validation images: 81208

Number of testing images: 81434

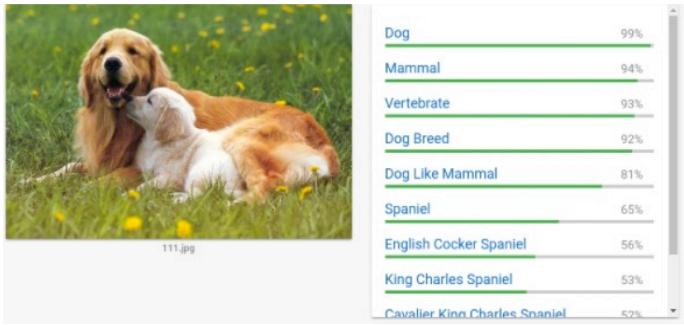
Number of categories: 102

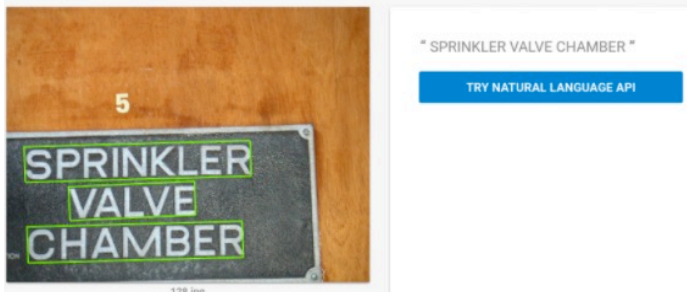


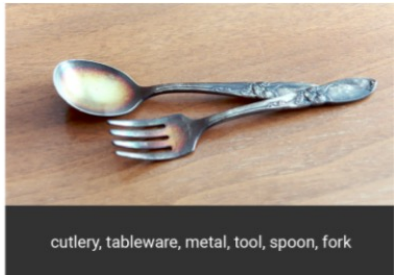
Datasets

Google Open Image dataset:

9 million images
Over 6000 categories







https://github.com/openimages/dataset

vision@OUC

Evaluation Indexes

Evaluation Indexes

- Overall accuracy
- top-1 error
- top-5 error
- Confusion matrix
  - Precision
  - Recall
  - F-measure

Evaluation Indexes

top-5 error rate & top-1 error rate：

top-n error rate： the fraction of test images for which the correct label is not among the five labels considered most probable by the model.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Evaluation Indexes

Confusion matrix：

Confusion matrix is a kind of Visualization tools to show whether different categories images are confused.

Predict

Actual		Girl	Boy
	Girl	18	2
	Boy	9	71

Evaluation Indexes

TP：标签为正类，并且被分到正类的样本数

FP：标签为负类，但是被分到正类的样本数

FN：标签为正类，但是被分到负类的样本数

TN：标签为负类，并且被分到负类的样本数

Overall Accuracy（准确率）:准确率 = $\frac{TP+TN}{TP+FP+FN+TN}$

Precision（精确率）: P = $\frac{TP}{TP+FP}$

Recall（召回率）: R = $\frac{TP}{TP+FN}$

		Predict: Girl	Predict: Boy
Actual	Girl（Positive）	18（TP）	2（FN）
	Boy（Negative）	9（FP）	71（TN）

$$overall\ auucray = \frac{18+71}{18+2+9+71} = 0.89$$

$$P = \frac{18}{18+9} = 0.6667$$

$$R = \frac{18}{18+2} = 0.9$$

Evaluation Indexes

F-measure：

F_1-measure： harm-mean of precision & recall

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R} \longrightarrow F_1 = \frac{2PR}{P+R} = \frac{2TP}{2TP + FP + FN}$$

$$\downarrow$$

$$F_a = \frac{(a^2+1)PR}{a^2(P+R)}$$

Classification Methods

Classification Process

Image Classification

- **Image processing：** Image enhancement, Image restoration, Image segmentation
- **Feature extraction：** SIFT, SURF, HOG, LBP, FAST, LoG, DoG
- **Feature representation：** Fisher Vector, Vector quantization, Soft quantification, FMM, LCC
- **Feature selection：** Search measurement based, Evaluation criteria based
- **Dimensionality reduction：** PCA, LDA, LLE, Feature hashing
- **Classifier：** SVM, K-nearest, Random forest, adaboost

Shortcomings

- Hand-designed features
  - High costs
  - Subjective
  - Poor effectiveness
- Artificial integrated system

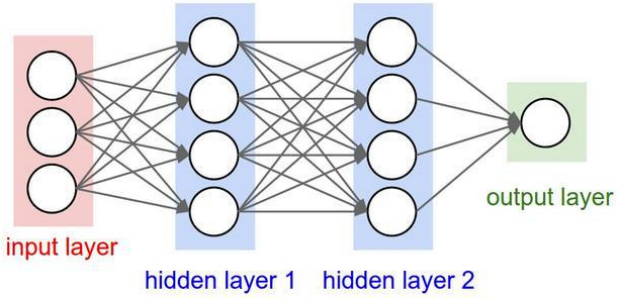
Deep Learning

Deep Learning

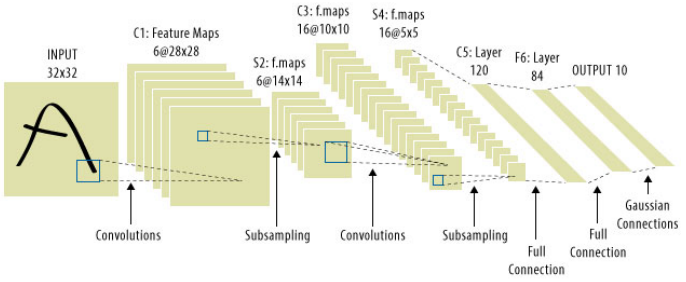
What is Deep Learning?

Deep learning is a class of machine learning algorithms that use a cascade of many layers of nonlinear processing units for feature extraction and transformation.

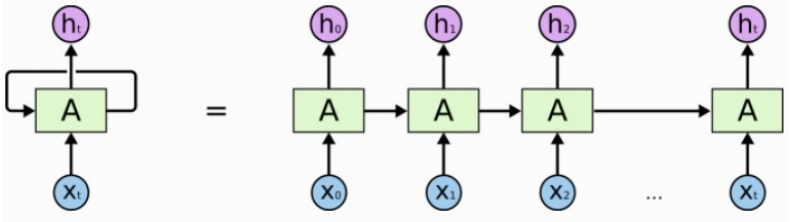
MLP



CNN



RNN



Deep Learning

Deep Learning: Revolution

ILSVRC：

- Good classification effect
- Transportability
- End to End
- Hierarchical features

Year	Top-5 error	Methods
2010 Champion	28.20%	HOG + LBP + SVM
2011 Champion	25.70%	FV + SVM
2012 Champion	16.42%	DCNN: AlexNet (8 layers)
2013 Champion	11.74%	DCNN: Network visualization (AlexNet based)
2014 Champion	6.66%	DCNN: GoogLeNet (22 layers + Inception)
2014 Second	7.32%	DCNN: VGGNet (19 layers)
2015 Champion	3.6 %	DCNN: ResNet (152 layers)

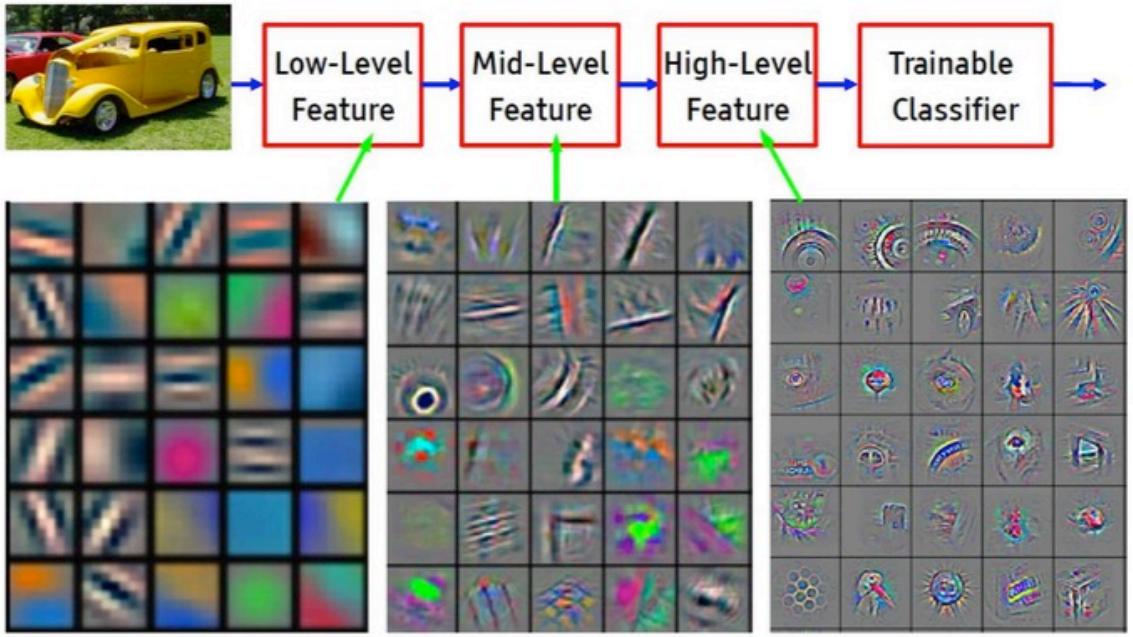
Deep Learning

End - to - End:

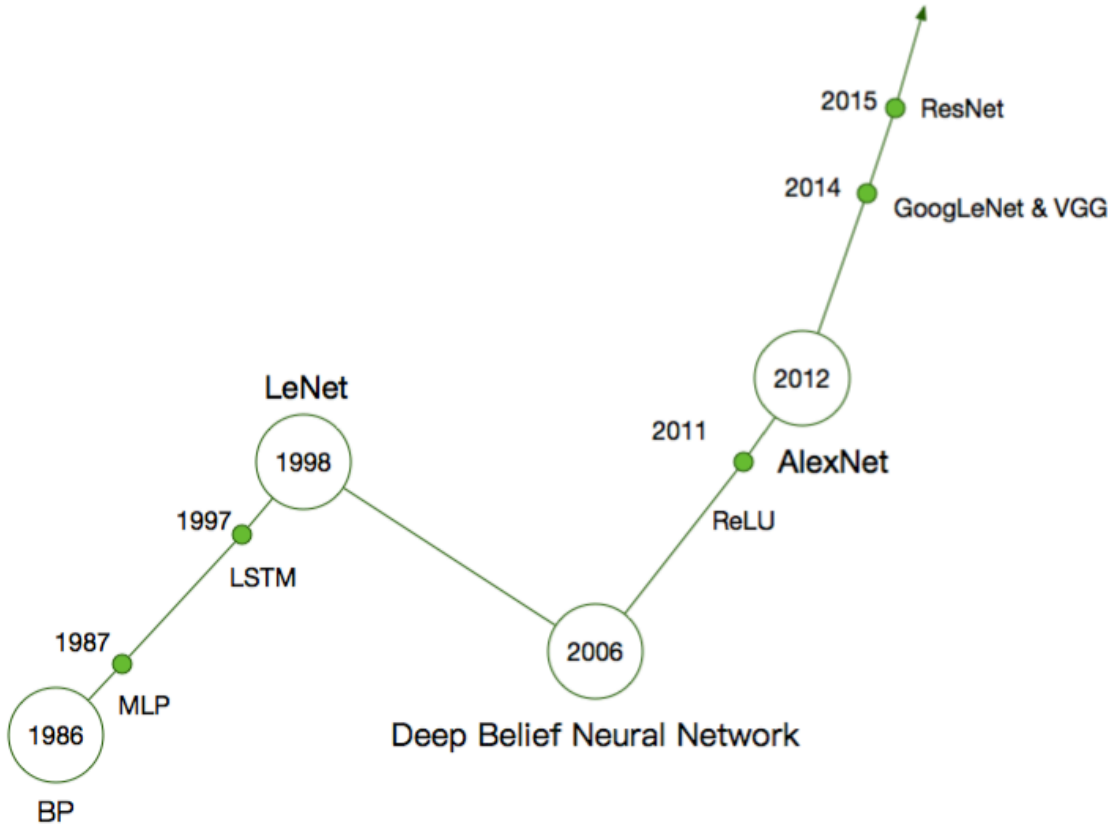
Deep Learning

Hierarchical features :

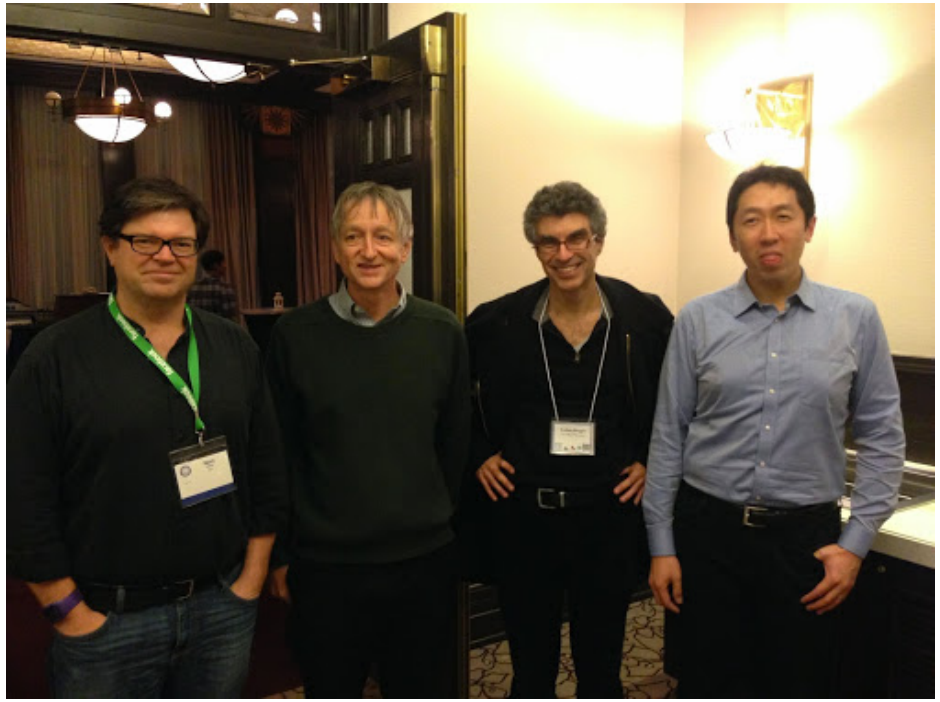
Pixel => Edge => Texton => Motif => Part => Object



Deep Learning



Deep Learning

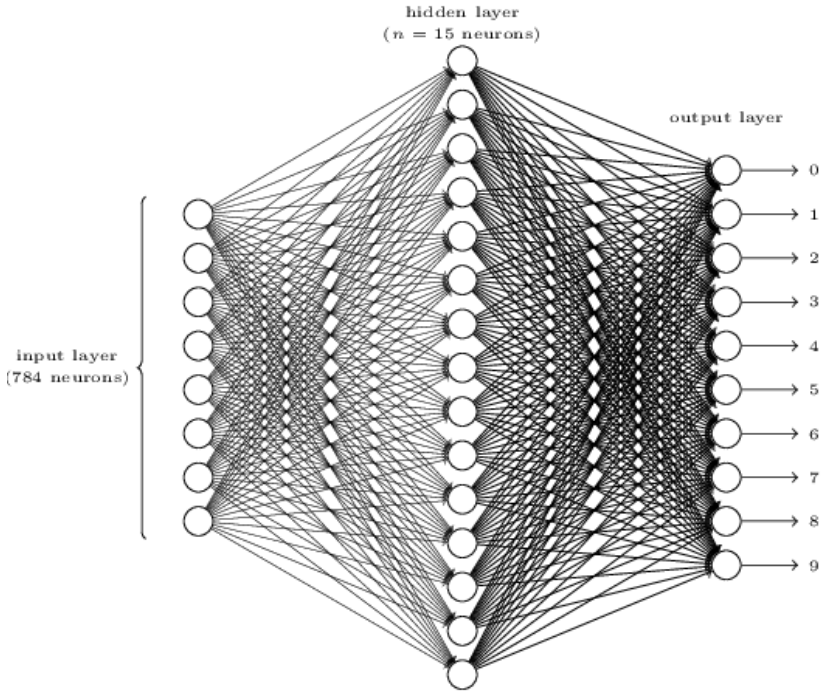


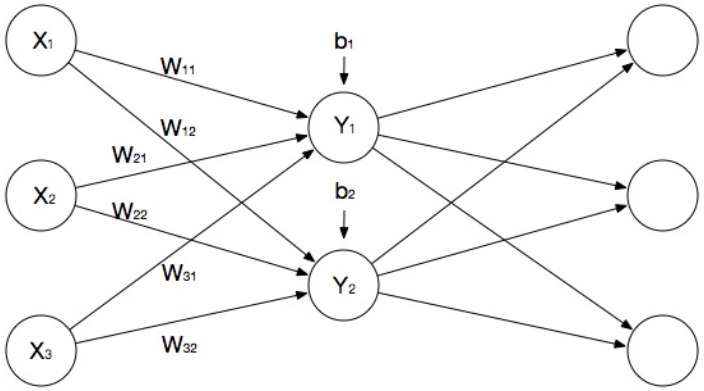
From the left：

LeCun
Hinton
Bengio
NG

MLP

Multilayer Perceptron：

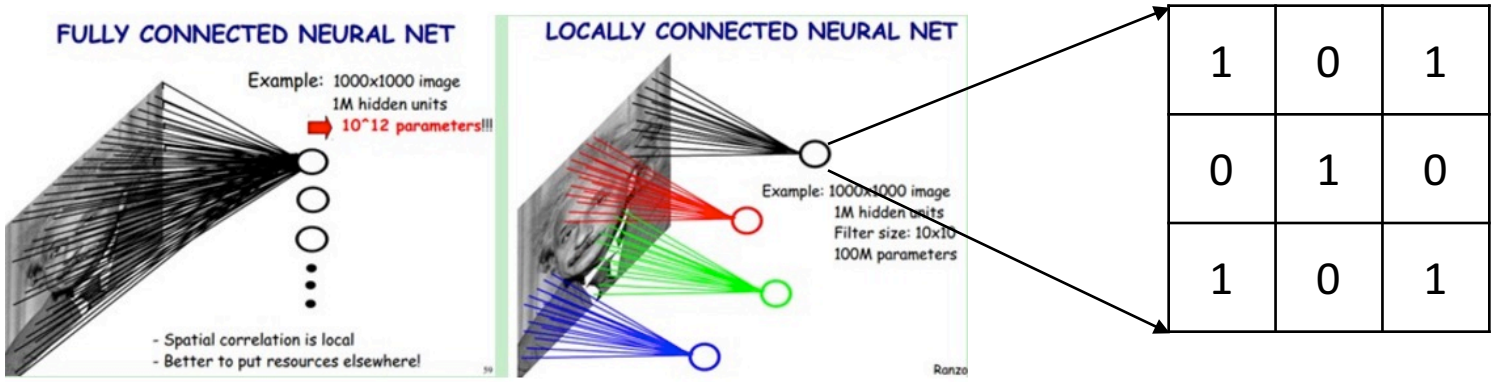




$$Forward:\ y = W^T x + b,\quad y \in R^{m\times 1}, x \in R^{n\times 1}, W \in R^{m\times n}$$

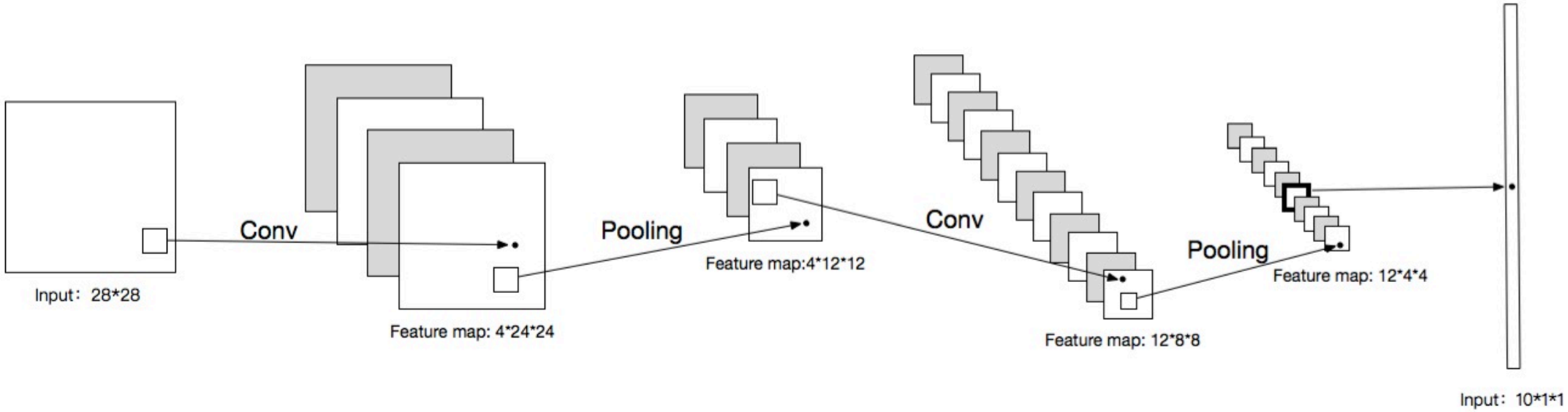
$$Backward:\ \frac{\partial L}{\partial x} = W \times \frac{\partial L}{\partial y}, \frac{\partial L}{\partial y} = x \times (\frac{\partial L}{\partial y})^T$$

CNN



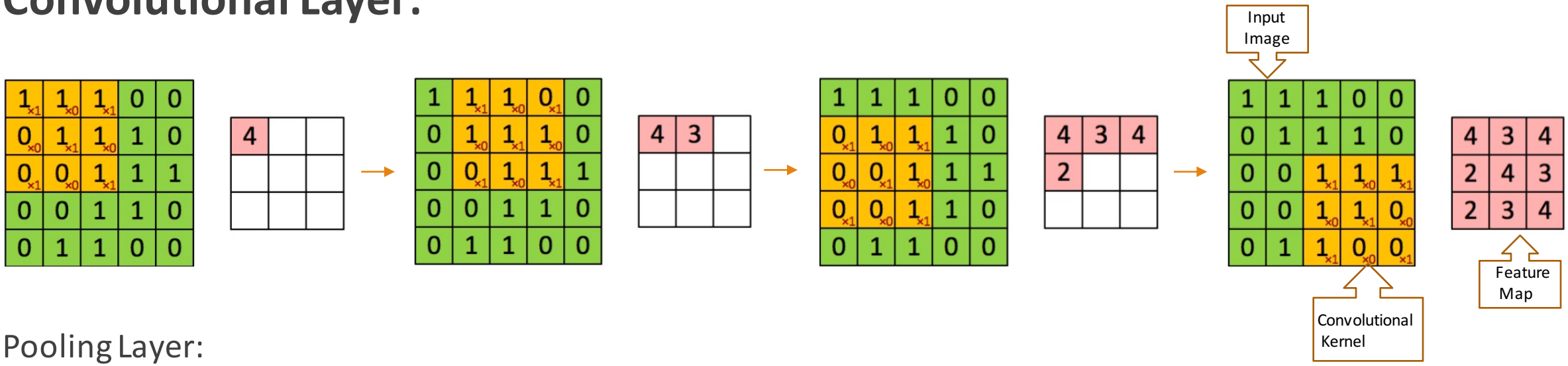
- Sparse Connectivity
- Shared Weights

CNN


Conv
Pooling
Conv
Pooling
Input：28*28
Feature map: 4*24*24
Feature map:4*12*12
Feature map: 12*8*8
Feature map: 12*4*4
Input：10*1*1

CNN

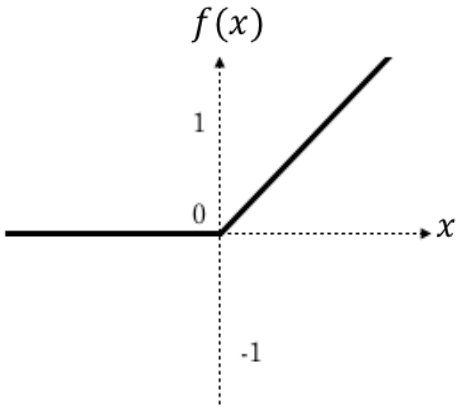
Convolutional Layer:



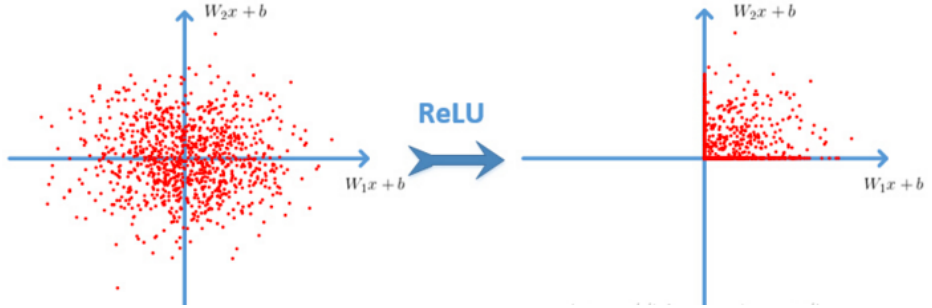
Pooling Layer:

CNN

ReLU:



$$f(x) = max(0, x)$$



Glorot X, Bordes A, Bengio Y. Deep Sparse Rectifier Neural Networks[C]//Aistats. 2011, 15(106): 275.

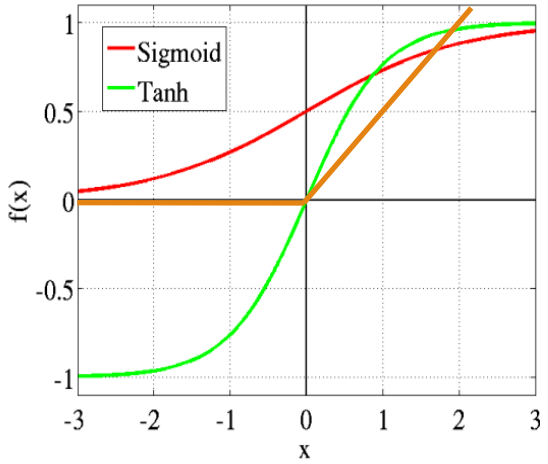
CNN

Advantages of ReLU：

1、Faster

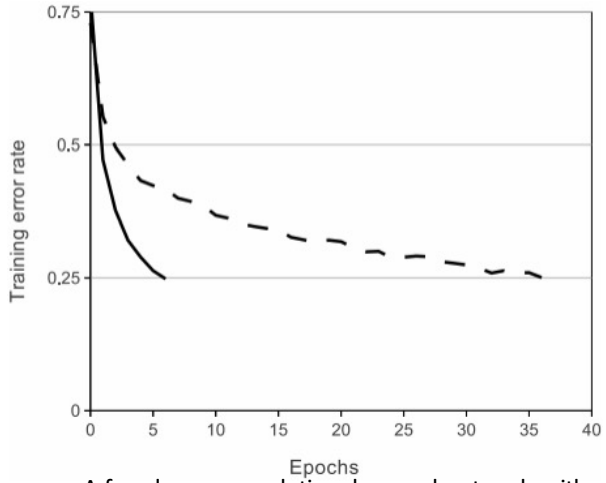
2、Simple Gradient

3、Sparsity



Sigmoid: $f(x) = \frac{1}{1+e^{-x}}$

Tanh: $f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$

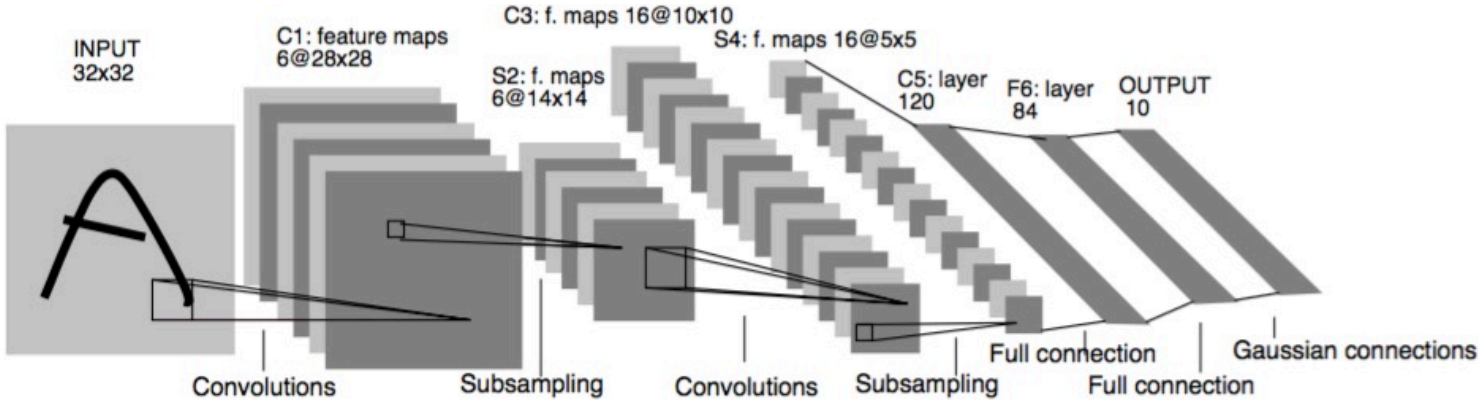


A four-layer convolutional neural network with ReLUs (solid line) reaches a 25% training error rate on CIFAR-10 six times faster than an equivalent network with tanh neurons (dashed line).

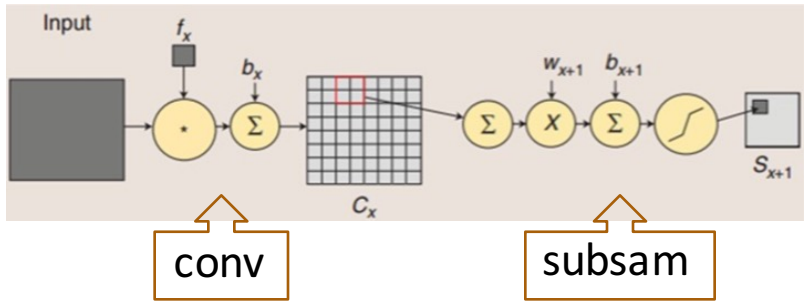
CNN: LeNet-5

LeNet-5(1998):

Network Architecture:



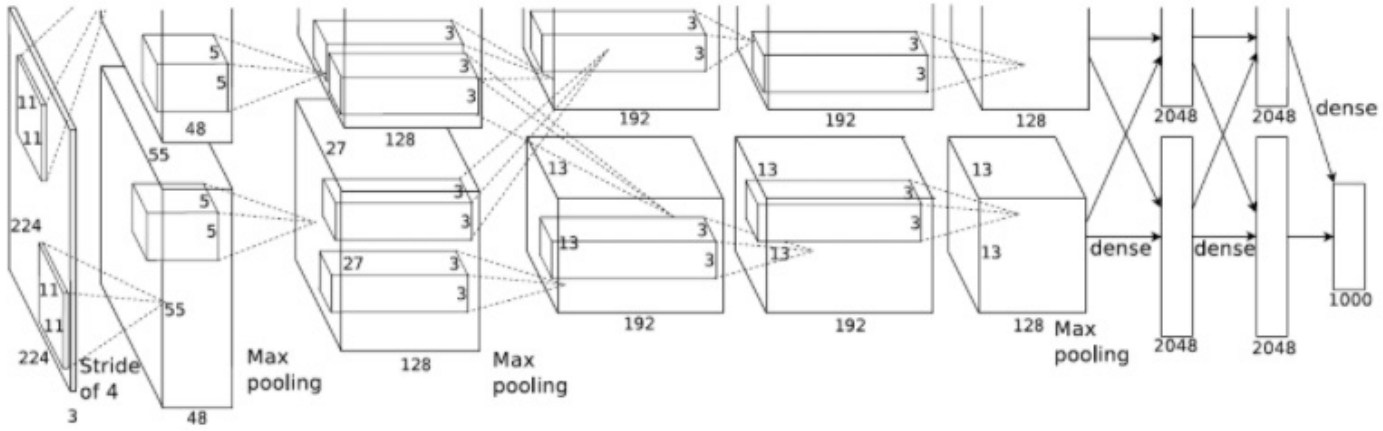
Process of Convolutions & Subsumpling:



LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

CNN: AlexNet

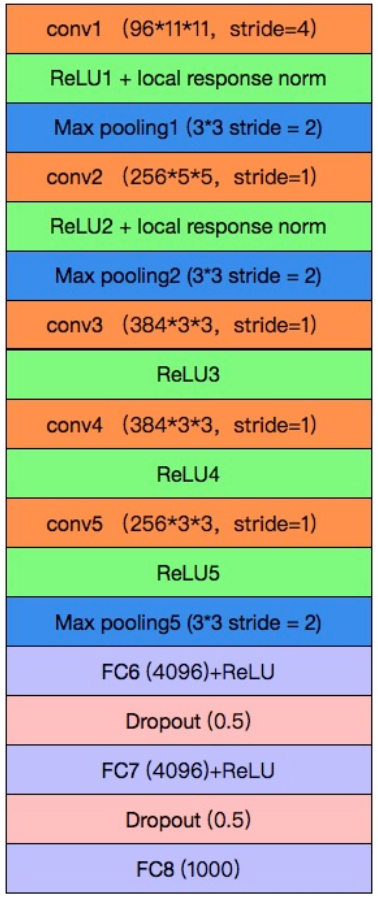
AlexNet(2012):



NVIDIA GTX 580 3GB GPU * 2

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

CNN: AlexNet



Novel:

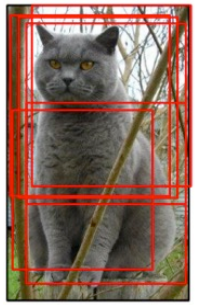
1. Data Augmentation
2. ReLU Nonlinearity
3. Local Response Normalization
4. Overlapping Pooling
5. Dropout
6. Training on Multiple GPUs

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

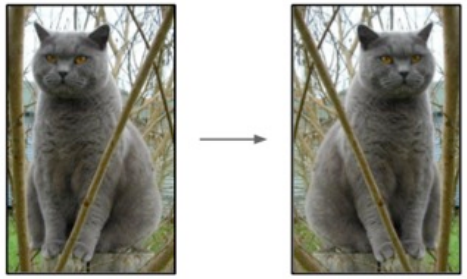
CNN: AlexNet

Data Augmentation:

1. Image crop and horizontal reflections.
2. Alter the intensities of the RGB channels in training images.

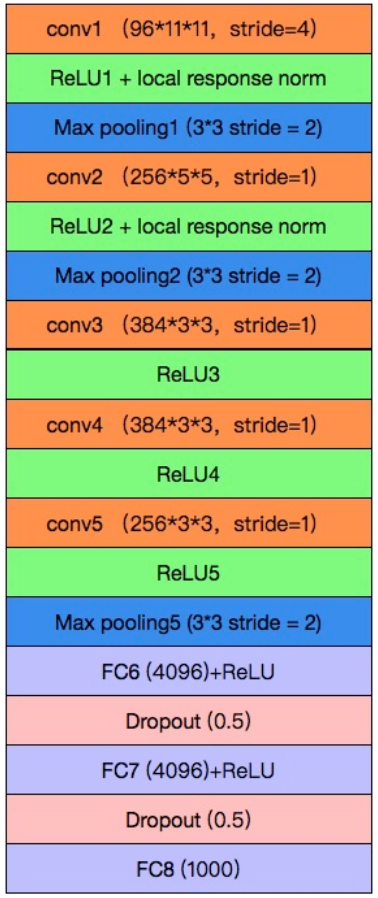


crop



reflection

CNN: AlexNet

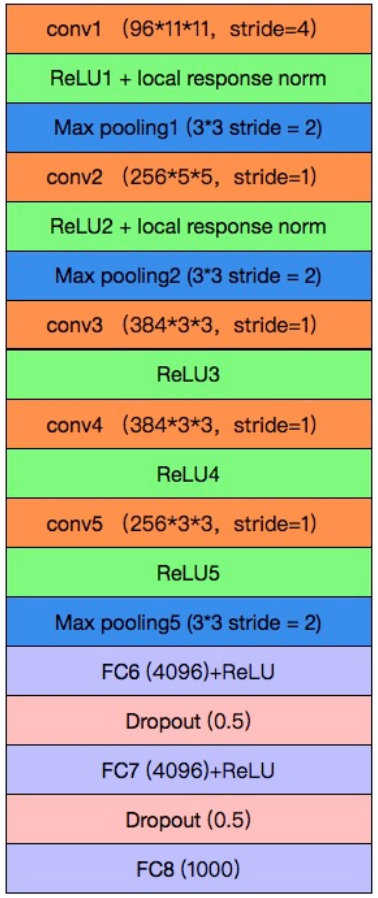


Novel:

1. Data Augmentation
2. ReLU Nonlinearity
3. Local Response Normalization
4. Overlapping Pooling
5. Dropout
6. Training on Multiple GPUs

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

CNN: AlexNet



Novel:

1. Data Augmentation
2. ReLU Nonlinearity
3. Local Response Normalization
4. Overlapping Pooling
5. Dropout
6. Training on Multiple GPUs

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

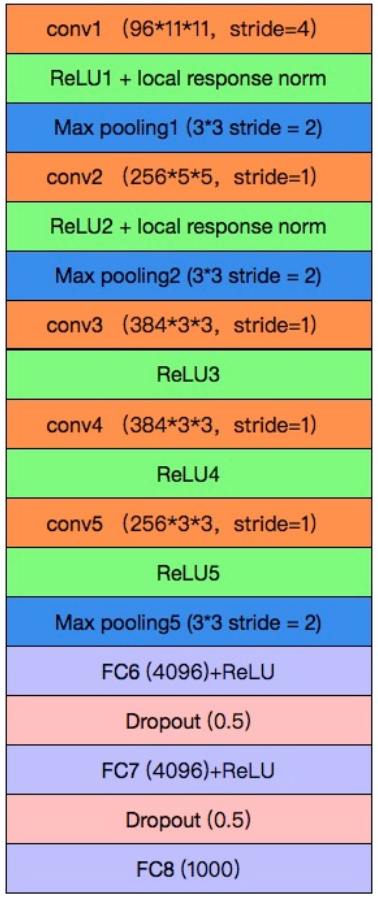
CNN: AlexNet

Local Response Normalization：

$$b_{x,y}^{i} = a_{x,y}^{i} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_{x,y}^{j})^2 \right)^{\beta}$$

Effect: Reduces top-1 error and top-5 error by 1.4% and 1.2%

CNN: AlexNet

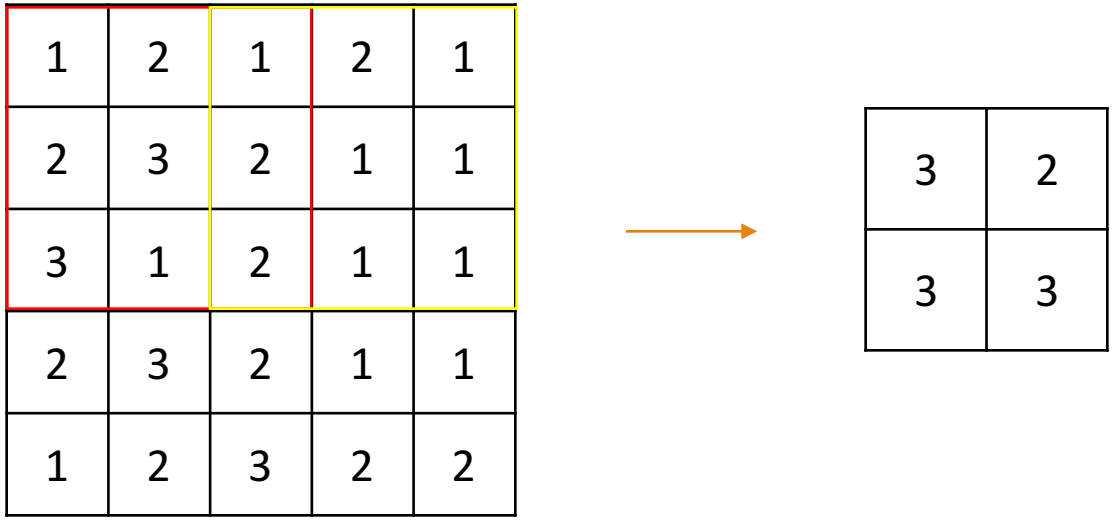


Novel:

1. Data Augmentation
2. ReLU Nonlinearity
3. Local Response Normalization
4. Overlapping Pooling
5. Dropout
6. Training on Multiple GPUs

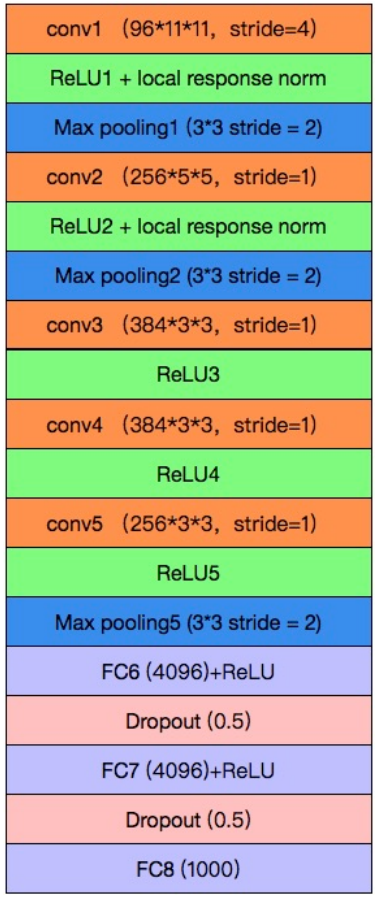
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

CNN: AlexNet



Effect: Reduces top-1 error and top-5 error by 0.4% and 0.3%

CNN: AlexNet



Novel:

1. Data Augmentation
2. ReLU Nonlinearity
3. Local Response Normalization
4. Overlapping Pooling
5. Dropout
6. Training on Multiple GPUs

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

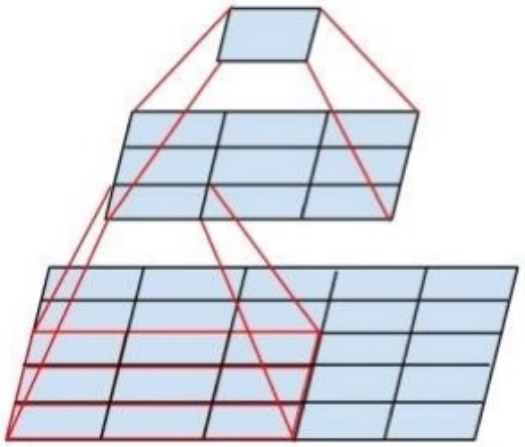
CNN: VGGNet(2014)

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

- More layers
- Non-overlapping pooling
- No LRN
- Smaller kernel size and stride

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

CNN: VGGNet



	5*5	3*3 + 3*3
Parameter Amount	5*5+1=26	2*(3*3+1)=20

Advantages：

1. Less parameters
2. Multiple 3*3 kernels can replace much bigger kernel
3. More nonlinear

Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

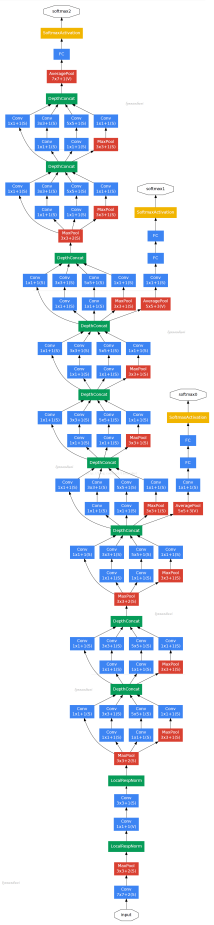
CNN

1. Deeper network is more prone to over-fitting

2. Dramatically increase use of computational resources

3. Too much parameters of full connect layers.

CNN: GoogLeNet(2014)



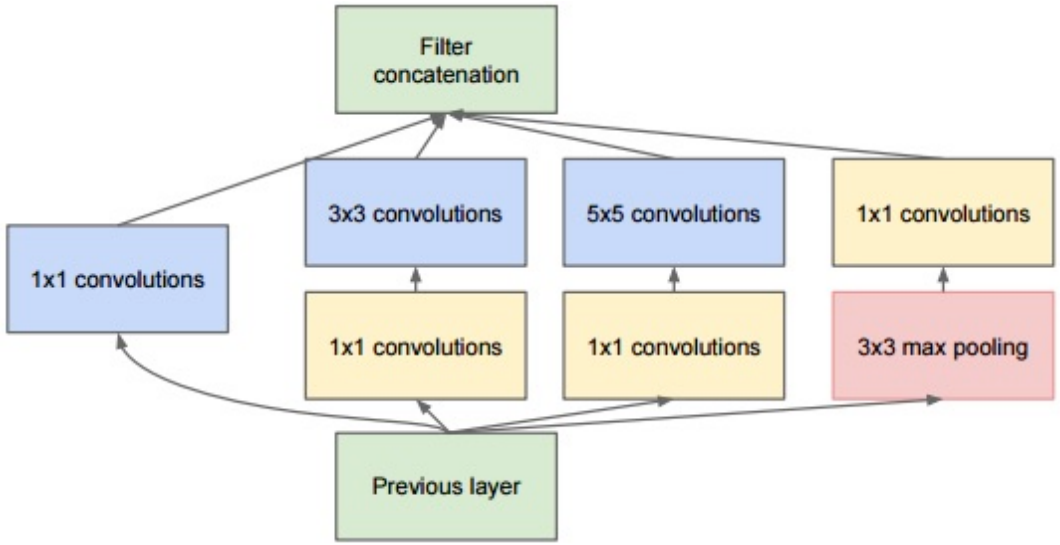
type	patch size/ stride	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj	params	ops
convolution	7×7/2	112×112×64	1							2.7K	34M
max pool	3×3/2	56×56×64	0								
convolution	3×3/1	56×56×192	2		64	192				112K	360M
max pool	3×3/2	28×28×192	0								
inception (3a)		28×28×256	2	64	96	128	16	32	32	159K	128M
inception (3b)		28×28×480	2	128	128	192	32	96	64	380K	304M
max pool	3×3/2	14×14×480	0								
inception (4a)		14×14×512	2	192	96	208	16	48	64	364K	73M
inception (4b)		14×14×512	2	160	112	224	24	64	64	437K	88M
inception (4c)		14×14×512	2	128	128	256	24	64	64	463K	100M
inception (4d)		14×14×528	2	112	144	288	32	64	64	580K	119M
inception (4e)		14×14×832	2	256	160	320	32	128	128	840K	170M
max pool	3×3/2	7×7×832	0								
inception (5a)		7×7×832	2	256	160	320	32	128	128	1072K	54M
inception (5b)		7×7×1024	2	384	192	384	48	128	128	1388K	71M
avg pool	7×7/1	1×1×1024	0								
dropout (40%)		1×1×1024	0								
linear		1×1×1000	1							1000K	1M
softmax		1×1×1000	0								

- Delete full connect layer
- Inception architecture
- Auxiliary classifiers

Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

CNN: GoogLeNet

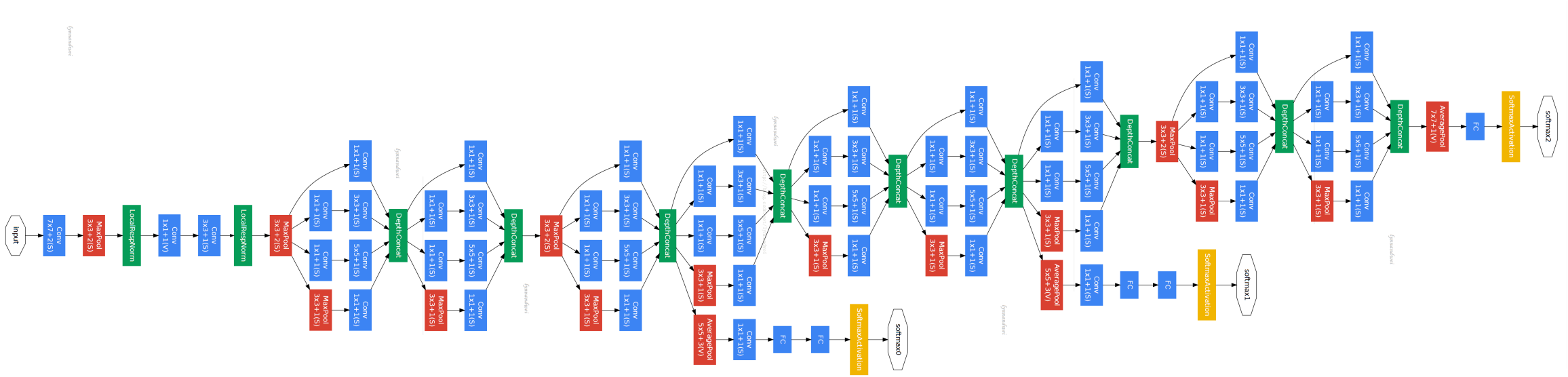
Inception Architecture：



Main idea: consider how an optimal local sparse structure of a convolutional vision network can be approximated and covered by readily available dense components

Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

CNN: GoogLeNet



Fine-grained classification

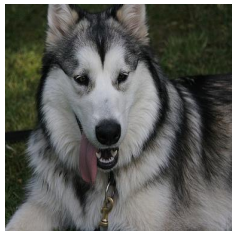
Fine-Grained Classification

aims to distinguish subordinate-level categories which are very similar.

Husky(0)



Malamute(1)




Fine-Grained Classification

Applications:

- Species identification
  - Flowers
  - Fishes
  - Birds
  - Dogs

…

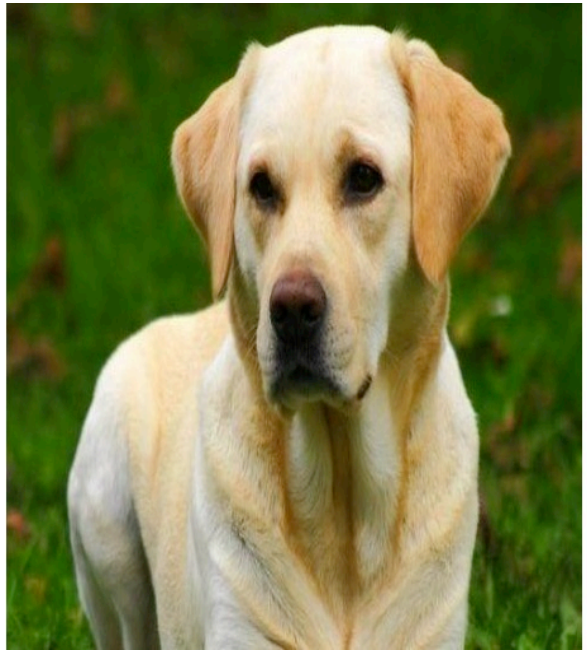
Challenges:

1. Limited amount of database
2. low inter-breed varition

Norfolk Terrier or Cairn Terrier?

Challenges:

3. High intra-breed variation



Challenges：

4. Innumerable Poses：

Challenges:

5. Complex background:

Fine-grained Classification

Part！ Part！ ！ Part！ ！ ！

Dataset: CUB-200-2011

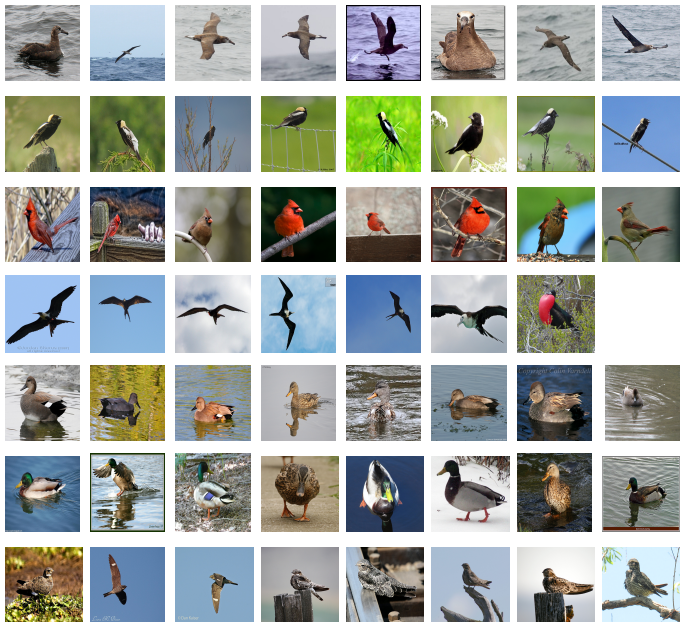
Number of categories:200

Number of images: 11788

Annotations per image:

15 Part Locations,

1 Bounding Box



Dataset: CUB-200-2011

Dataset: Stanford Dogs

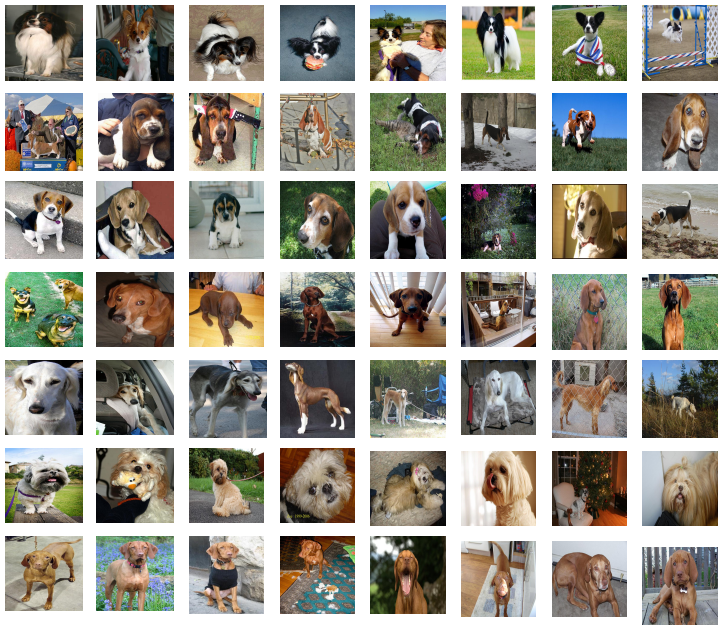
Number of categories: 120

Number of images: 20580

Annotations per image:

Class Labels

Bounding Boxes



Dataset: UCEFOOD256

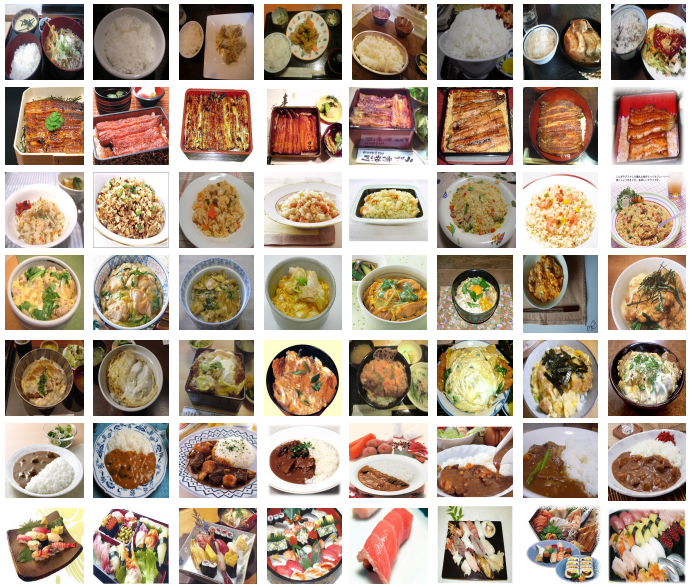
Number of categories: 256

Number of images: 31910

Annotations per image:

Class Labels

Bounding Boxes



Dataset: FGVC-Aircraft

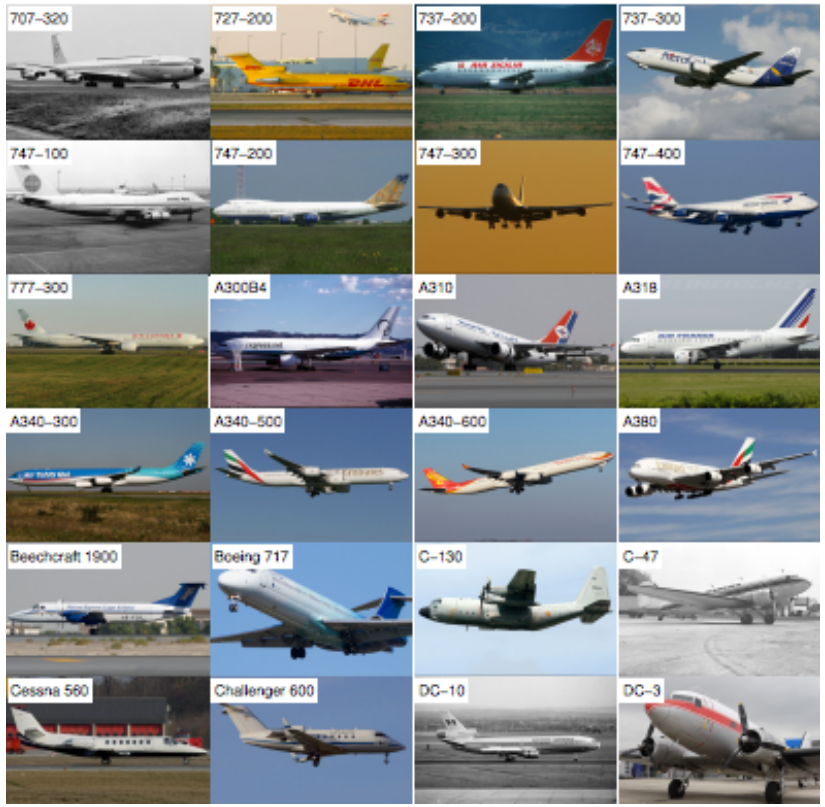
Number of categories: 102

Number of images: 10200

Annotations per image:

Class Labels

Bounding Boxes



Dataset: Cars

Number of categories: 196

Number of images: 16185

Annotations per image:

Class Labels

Bounding Boxes

http://ai.stanford.edu/~jkrause/cars/car_dataset.html

Dataset: Oxford Flowers

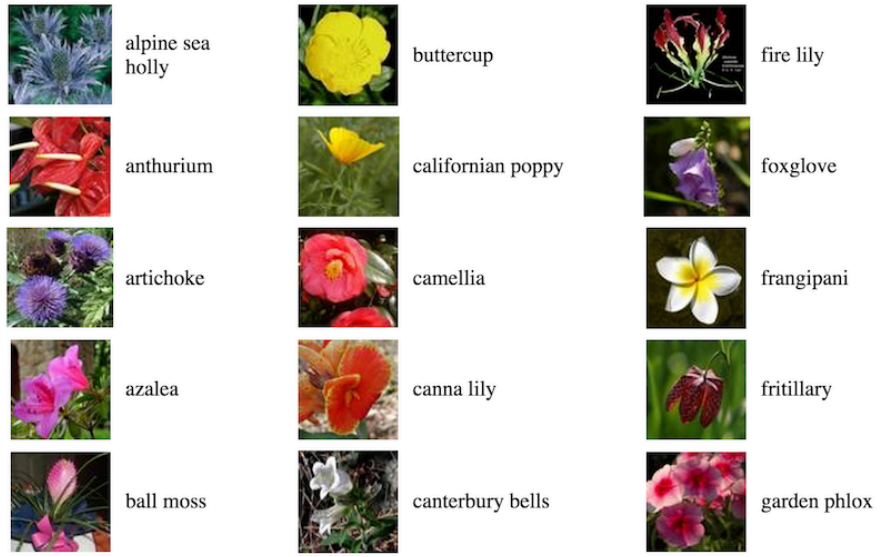
Number of categories: 102

Number of images: 8189

Annotations per image:

Class Labels

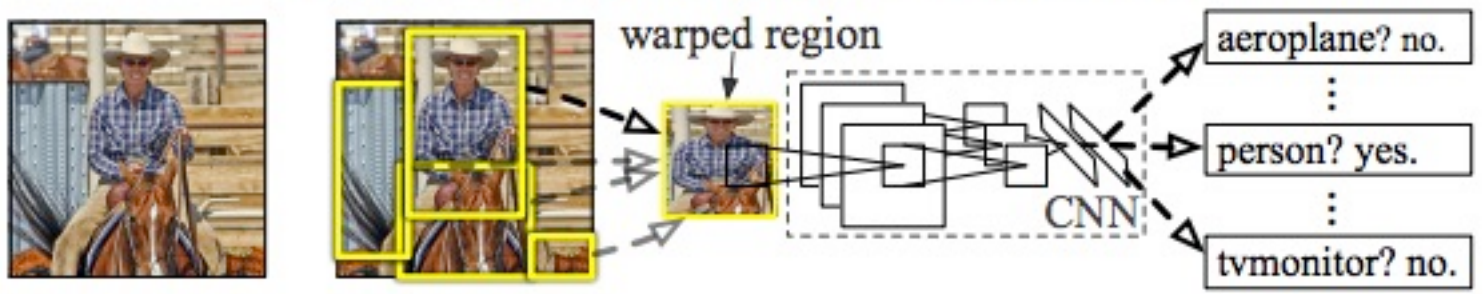
Segmentation



http://www.robots.ox.ac.uk/~vgg/data/flowers/

Fine-grained Classification

Locate Object & Part → Extract Object & Part Features → Classification



Fine-grained Classification

1. Part location + RGB color histogram + SIFT + BoW + SVM + boundingbox and part: 17.3%

2. AlexNet + Bounding Box: 61%

Fine-grained Classification

1. Image Enhancement

- Online images
- Rotation

2. Transfer Learning

- Pre-Trained Model
- Fine-Tune Model

Fine-grained Classification

Strongly-supervised vs Weakly-supervised

Strongly-supervised	Weakly-supervised
Labels, Bounding boxes, part locations…	Labels

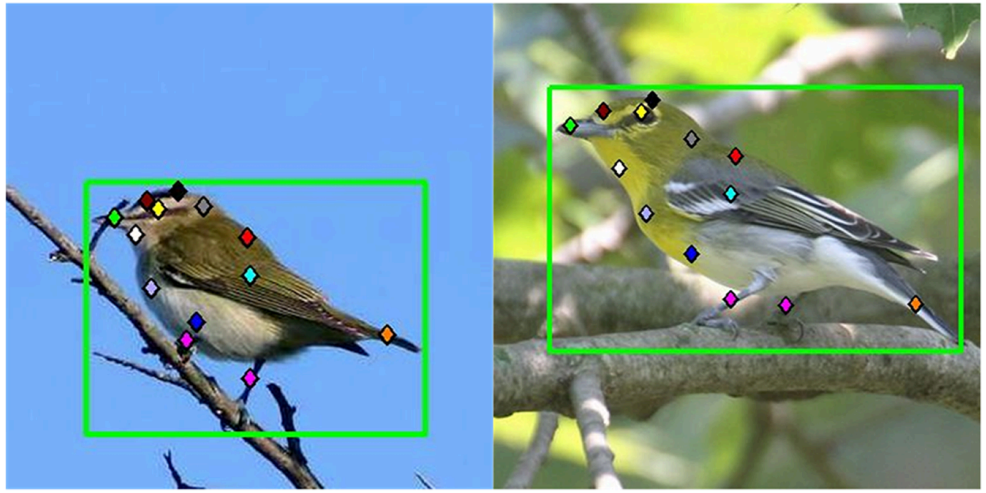
Strongly-supervised

Advantages:

Precise part annotation

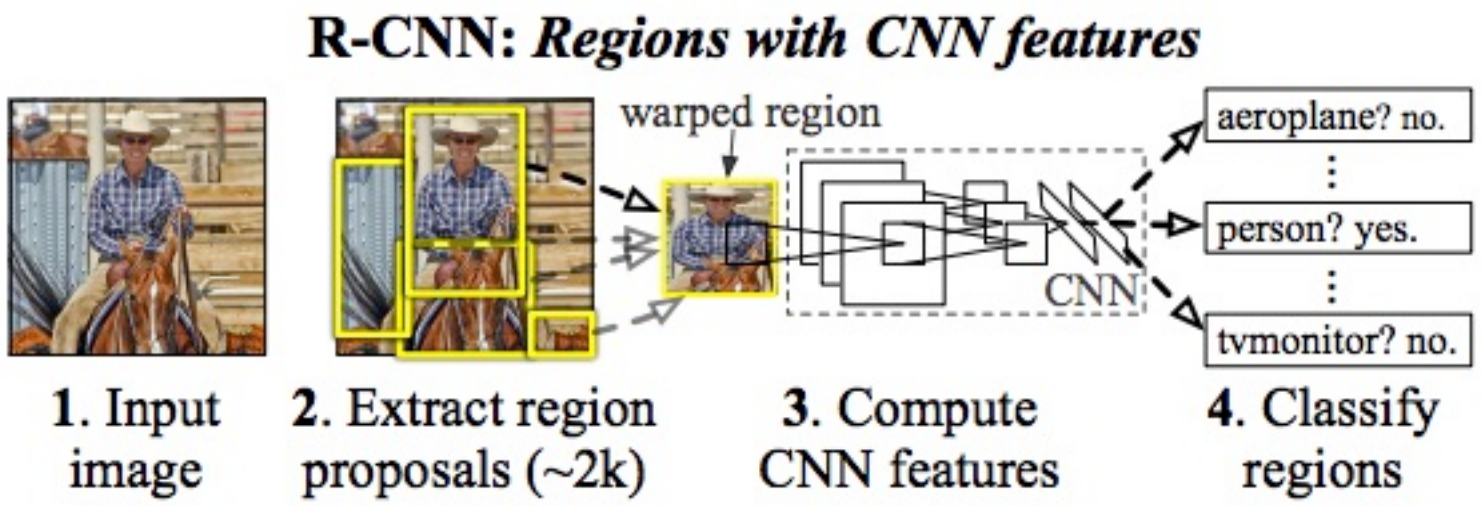
Shortages:

High costs



Strongly-supervised

R-CNN(Regions with CNN features)



Localization + classification

Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.

Strongly-supervised: R-CNN

Localization：

Selective search: obtain region proposals

Warped region:

Warp all regions into normal size

Feature extraction:

AlexNet: extract a 4096-dimensional feature vector

Classification:

SVM: optimize one linear SVM per class

Strongly-supervised: R-CNN

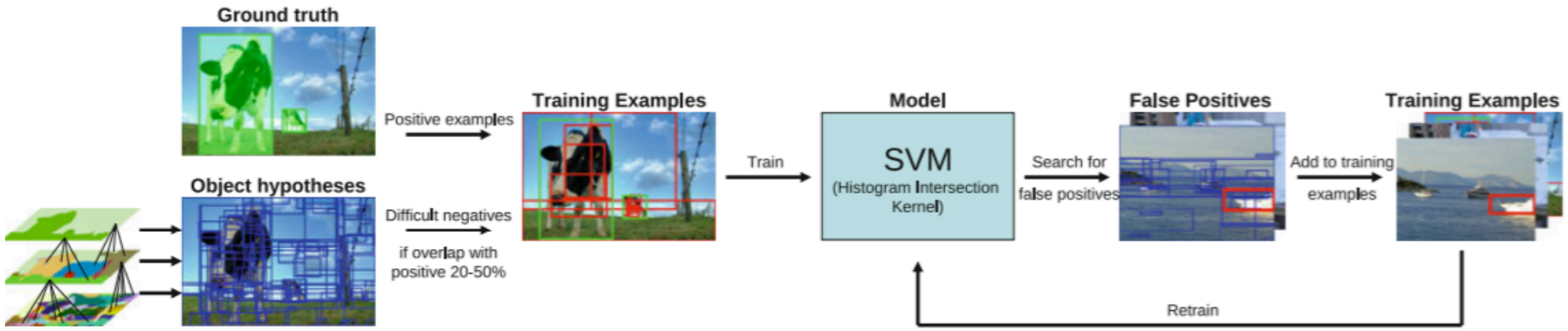
Selective search:

1. Efficient GraphBased Image Segmentation to get regions
2. calculate the comparability of each regions
3. Emerge the most similar two regions

Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

Strongly-supervised: R-CNN

Selective search:



4. Select negative examples

color, texture, size, fit

5. SVM

6. Test

Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

Strongly-supervised: R-CNN

Localization：

Selective search: obtain region proposals

Warped region:

Warp all regions into normal size → 224*224

Feature extraction:

AlexNet: extract a 4096-dimensional feature vector

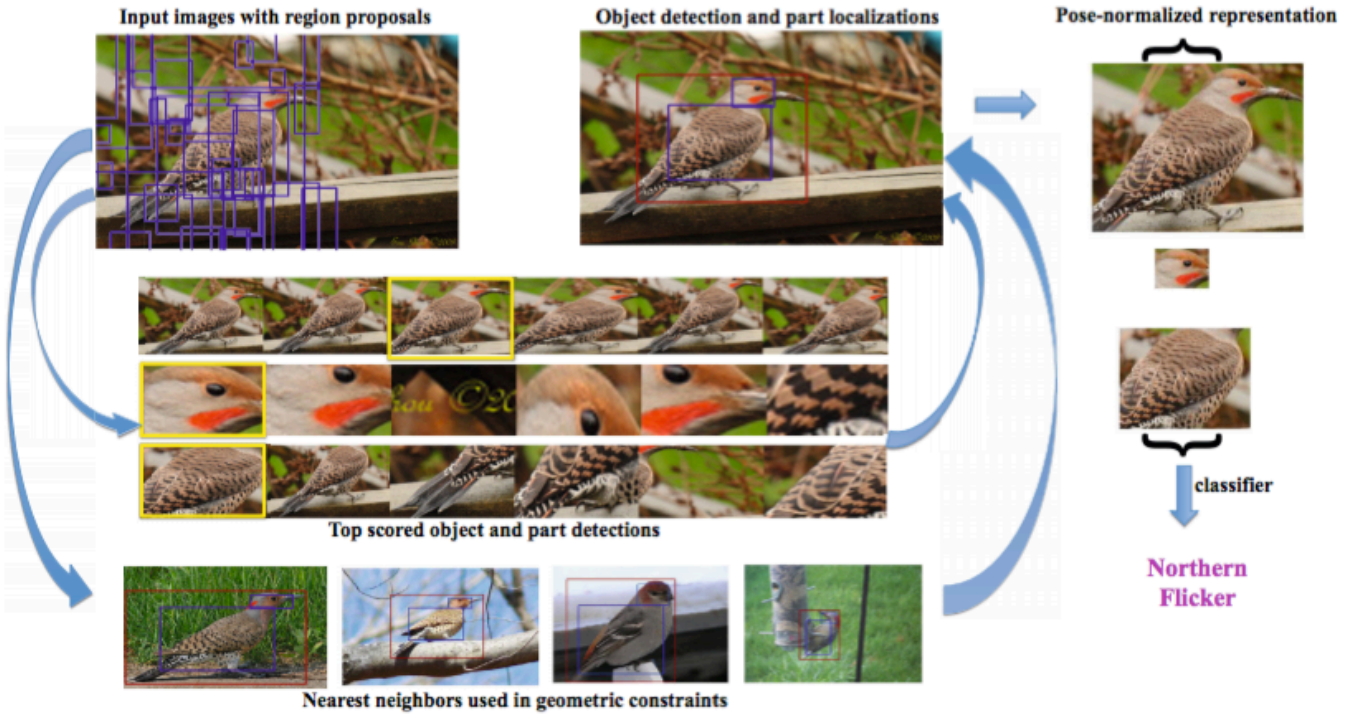
Classification:

SVM: optimize one linear SVM per class

Strongly-supervised: Part R-CNN

Part R-CNN:

R-CNN: Object detection and part (head & body) Localization.

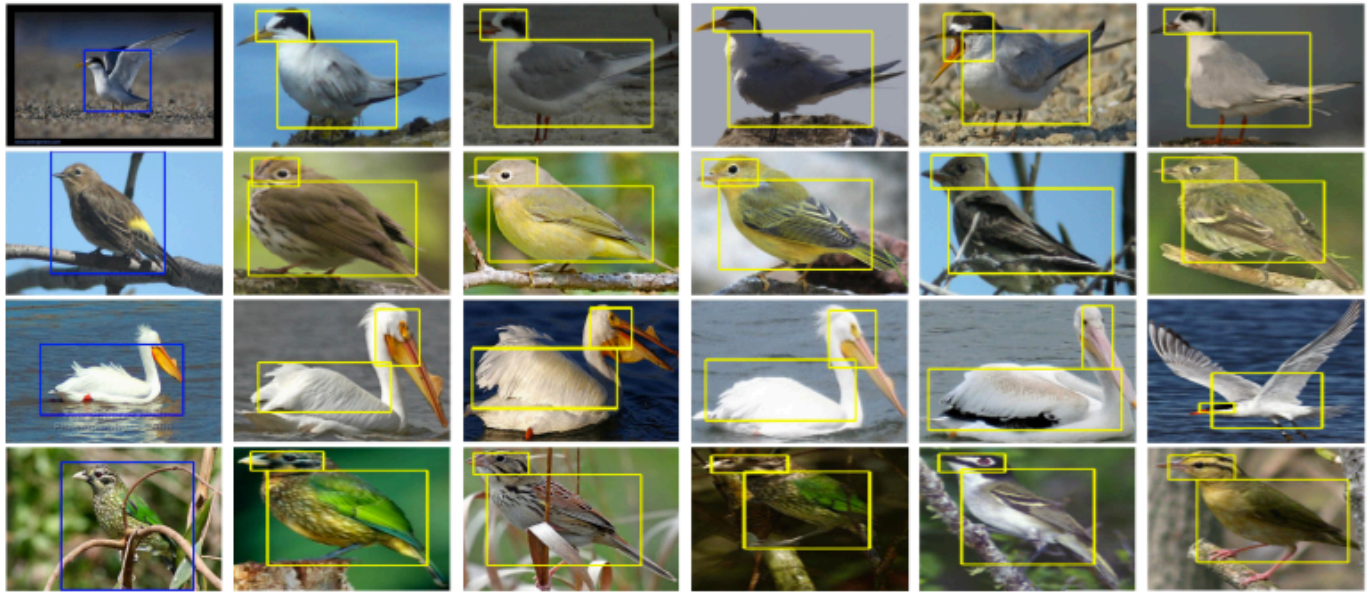


Constraints:

- All part windows are inside object windows
- constraints over the layout of the parts relative to the object

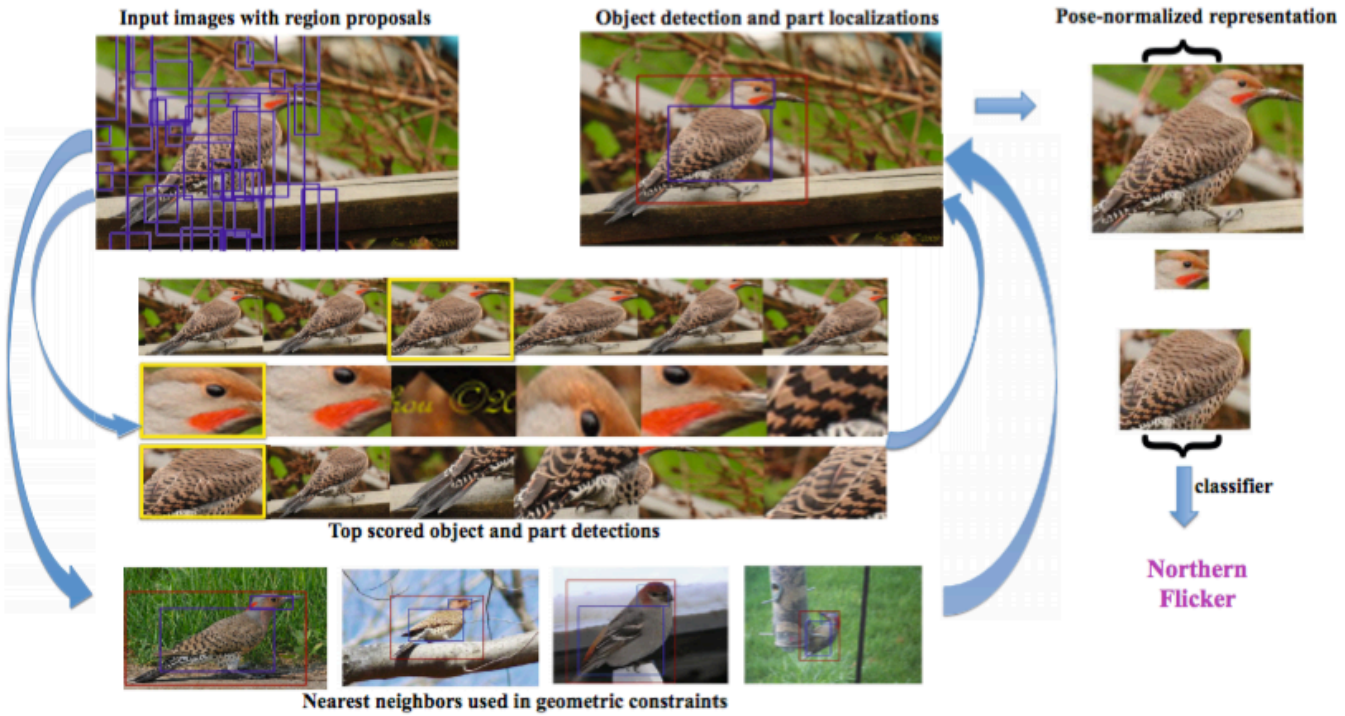
Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]//European conference on computer vision. Springer International Publishing, 2014: 834-849.

Strongly-supervised: Part R-CNN



Strongly-supervised: Part R-CNN

Part R-CNN:
R-CNN: Object detection and part (head & body) Localization.



Constraints:

- All part windows are inside object windows
- constraints over the layout of the parts relative to the object

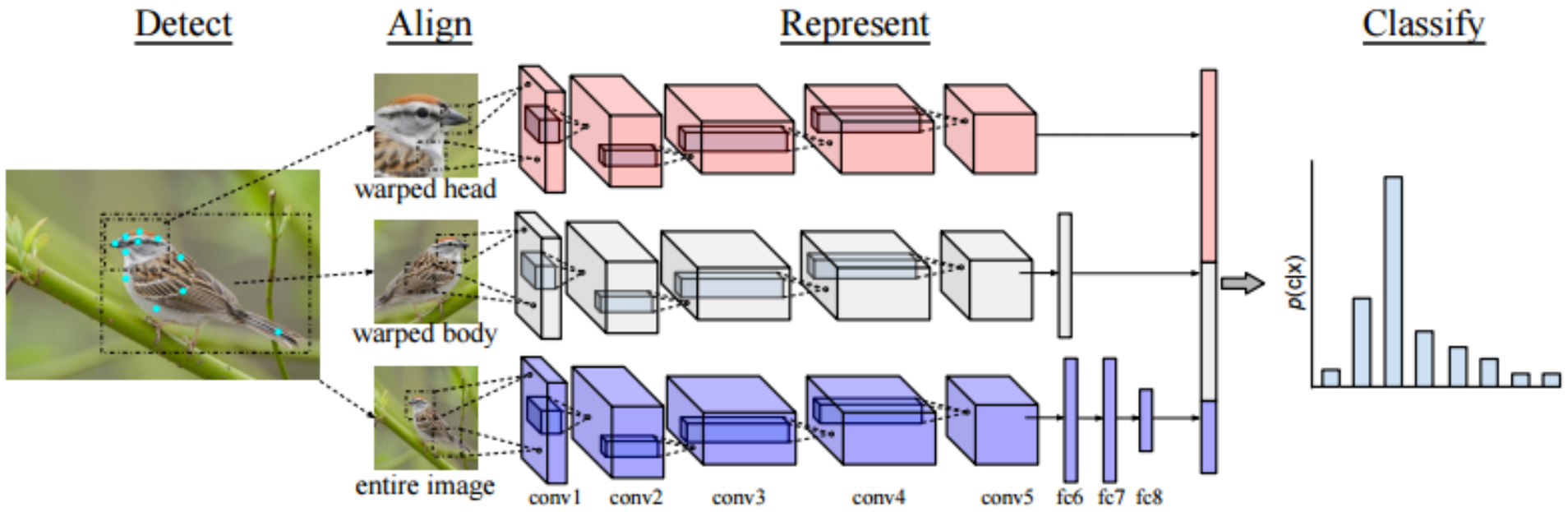
Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]//European conference on computer vision. Springer International Publishing, 2014: 834-849.

Strongly-supervised: Part R-CNN

Bounding Box Given	
DPD [47]	50.98%
DPD+DeCAF feature [14]	64.96%
POOF [7]	56.78%
Symbiotic Segmentation [10]	59.40%
Alignment [20]	62.70%
Oracle	72.83%
Oracle-ft	82.02%
Ours (Δ_{box})	67.55%
Ours ($\Delta_{\text{geometric}}$ with δ^{MG})	67.98%
Ours ($\Delta_{\text{geometric}}$ with δ^{NP})	68.07%
Ours-ft (Δ_{box})	75.34%
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{MG})	**76.37%**
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{NP})	76.34%
Bounding Box Unknown	
DPD+DeCAF [14] with no bounding box	44.94%
Ours (Δ_{null})	64.57%
Ours (Δ_{box})	65.22%
Ours ($\Delta_{\text{geometric}}$ with δ^{MG})	65.98%
Ours ($\Delta_{\text{geometric}}$ with δ^{NP})	65.96%
Ours-ft (Δ_{box})	72.73%
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{MG})	72.95%
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{NP})	**73.89%**

Zhang N, Donahue J, Girshick R, et al. Part-based R-CNNs for fine-grained category detection[C]//European conference on computer vision. Springer International Publishing, 2014: 834-849.

Pose Normalized CNN



Branson S, Van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets[J]. arXiv preprint arXiv:1406.2952, 2014.

Pose Normalized CNN

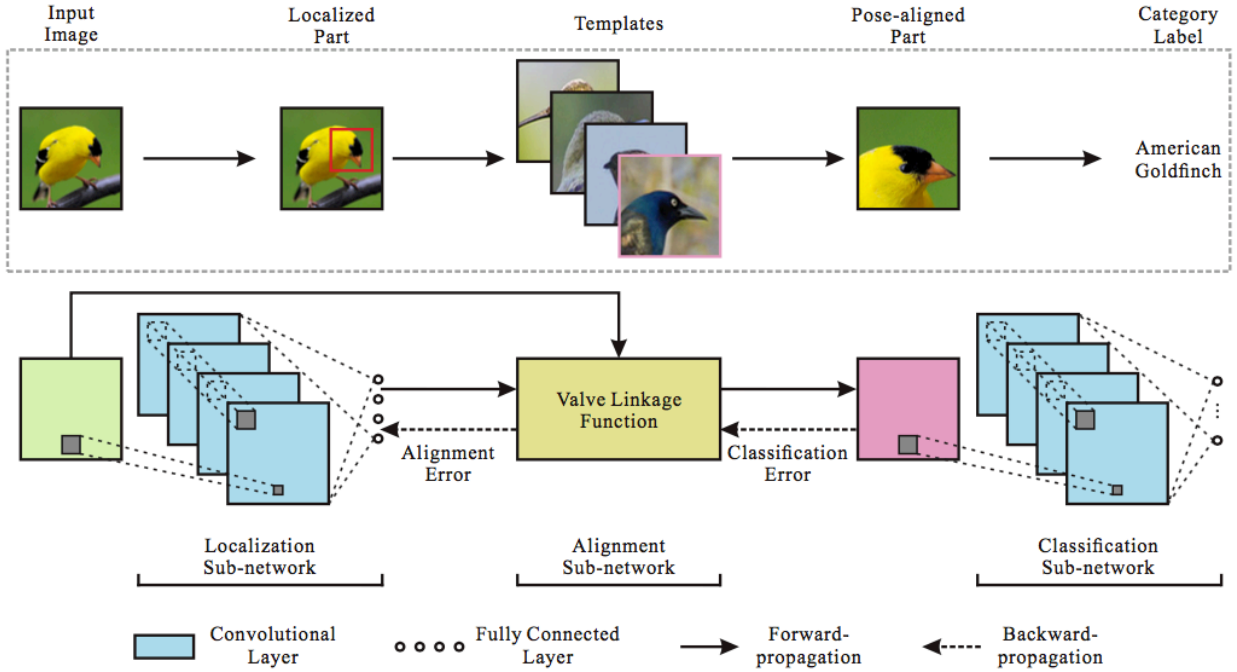
Method	Oracle Parts	Oracle BBox	Part Scheme	Features	Learning	% Acc
POOF [8]		✓	Sim-2-131	POOF	SVM	56.8
Alignments [21]		✓	Trans-X-4	Fisher	SVM	62.7
Symbiotic [13]		✓	Trans-1-1	Fisher	SVM	61.0
DPD [51]		✓	Trans-1-8	KDES	SVM	51.0
Decaf [11]		✓	Trans-1-8	CNN	Logistic Regr.	65.0
CUB [43]			Trans-1-15	BoW	SVM	10.3
Visipedia [33]			Trans-1-13	Fisher	SVM	56.5
Ours			Sim-5-6	CNN	SVM+CNN-FT	**75.7**
CUB Loc. [43]	✓	✓	Trans-1-15	BoW	SVM	17.3
POOF Loc. [8]	✓	✓	Sim-2-131	POOF	SVM	73.3
Ours Loc.	✓	✓	Sim-5-6	CNN	SVM+CNN-FT	**85.4**

Branson S, Van Horn G, Belongie S, et al. Bird species categorization using pose normalized deep convolutional nets[J]. arXiv preprint arXiv:1406.2952, 2014.

Strongly-supervised

Deep LAC:

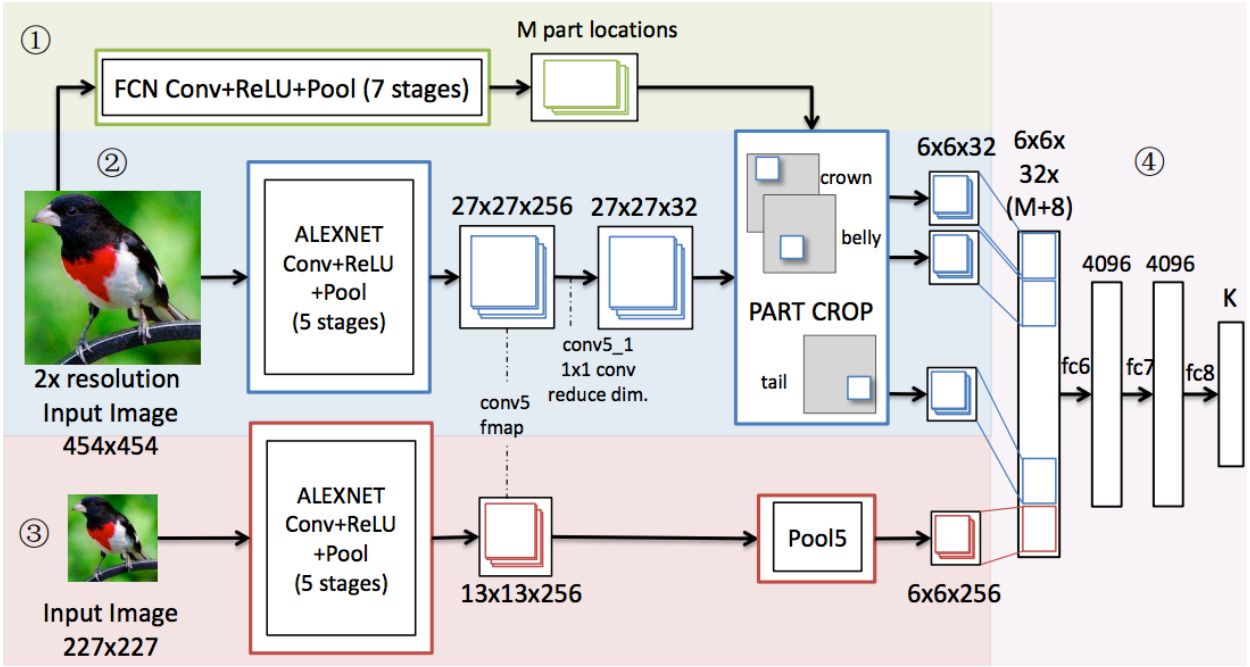
Localization + Alignment + Classification



Lin D, Shen X, Lu C, et al. Deep lac: Deep localization, alignment and classification for fine-grained recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1666-1674.

Strongly-supervised

Part-stack CNN:

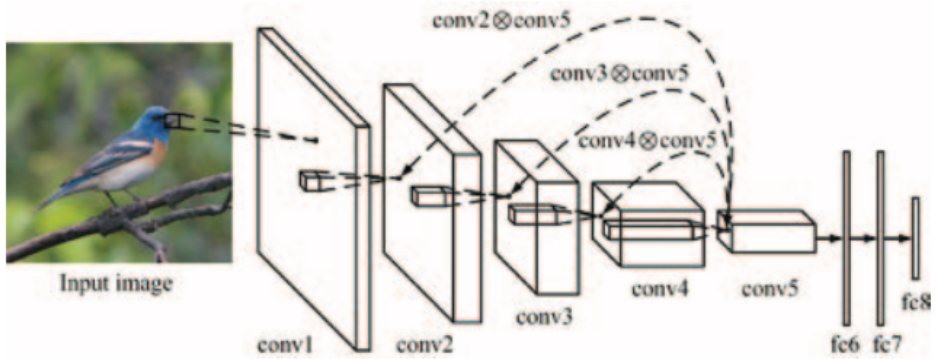


Two-stream neural network for object & part
FCN: Fully Convolutional Network

Huang S, Xu Z, Tao D, et al. Part-stacked cnn for fine-grained visual categorization[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1173-1182.

Strongly-supervised

Coarse-to-fine:

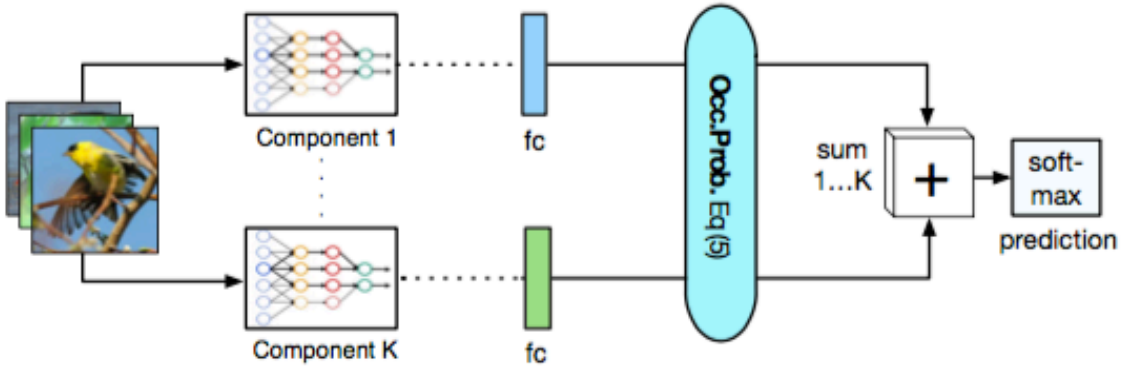


Dataset	Annotation	Accuracy (%)
CUB-2011	BBox	83.7
Aircraft-100	Bbox & Part	87.6
Cars-196	BBox	91.5

Wang Y, Zhang X Y, Zhang Y, et al. Exploiting coarse-to-fine mechanism for fine-grained recognition[C]//Image Processing (ICIP), 2016 IEEE International Conference on. IEEE, 2016: 649-653.

Strongly-supervised

Mix DCNN: Mixture of Deep Convolutional Neural Networks



K Components

occupation : $\alpha_k = \frac{e^{C_k}}{\sum_{c=1}^{K} e^{C_c}}$

Ge Z Y, Bewley A, McCool C, et al. Fine-grained classification via mixture of deep convolutional neural networks[C]//Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on. IEEE, 2016: 1-6.

Strongly-supervised

Strongly-supervised fine-grained classification methods' classification results on CUB-2011 dataset:

Method	Architecture	Bbox (training)	Bbox (testing)	Parts (training)	Parts (testing)	Accuracy (%)
Part R-CNN	AlexNet	✓	✓			73.9
Pose Normalized CNN	AlexNet	✓	✓			75.7
Pose Normalized CNN	AlexNet	✓	✓	✓	✓	85.4
Deep LAC	AlexNet	✓	✓			80.3
Part-stack CNN	AlexNet	✓	✓	✓		76.2
Coarse-to-fine	AlexNet	✓	✓			83.7
Mix DCNN	AlexNet	✓	✓			74.1

Strongly-supervised

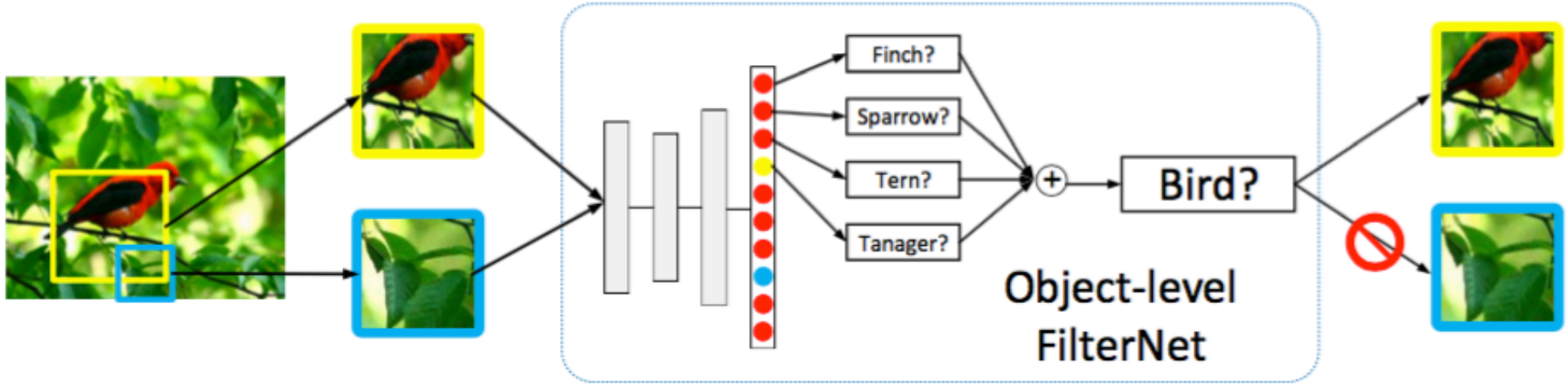
- Too much annotation costs a lot
- Without End - to - End training

Weakly-Supervised

Only Label!!!

Weakly-Supervised： Two-level Attention

Two-level Attention: Object-Level & Part-Level

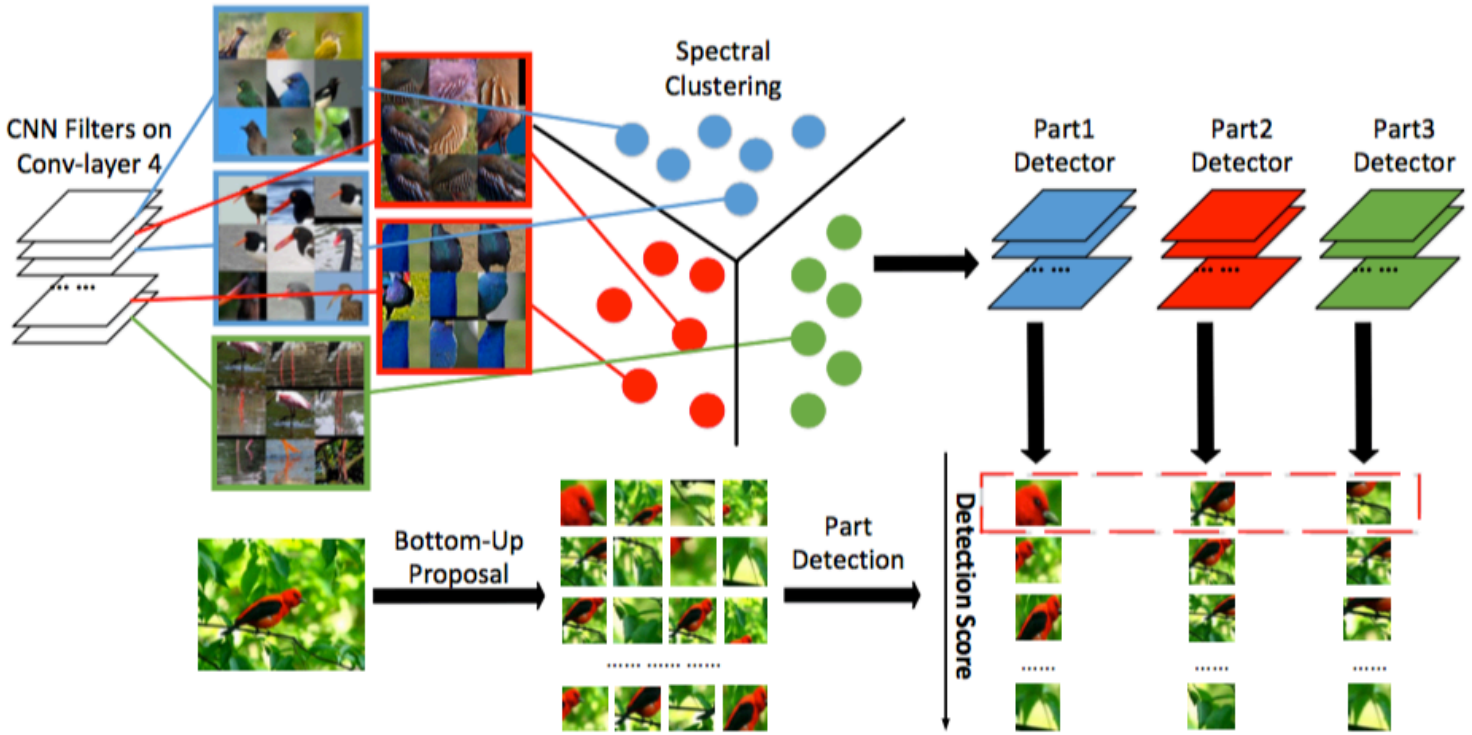


1. Selective search: regions
2. CNN: regions selection

Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 842-850.

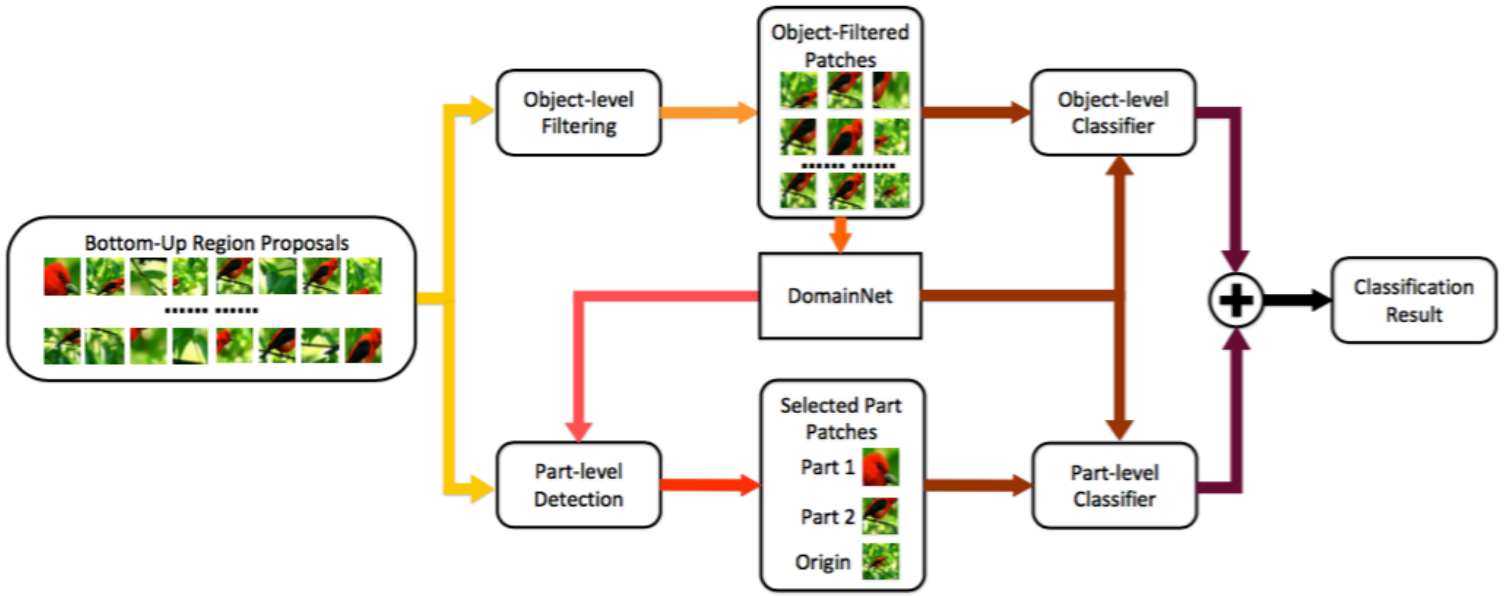
Weakly-Supervised： Two-level Attention

Spectral Clustering:



Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 842-850.

Weakly-Supervised： Two-level Attention



We turn a Convolutional Neural Net (CNN) pre-trained on ILSVRC2012 1K category into a *FilterNet*. *FilterNet* selects patches relevant to the basic-level category, thus processes the object-level attention. The selected patches drive the training of another CNN into a domain classifier, called *DomainNet*.

Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 842-850.

Weakly-Supervised：Two-level Attention

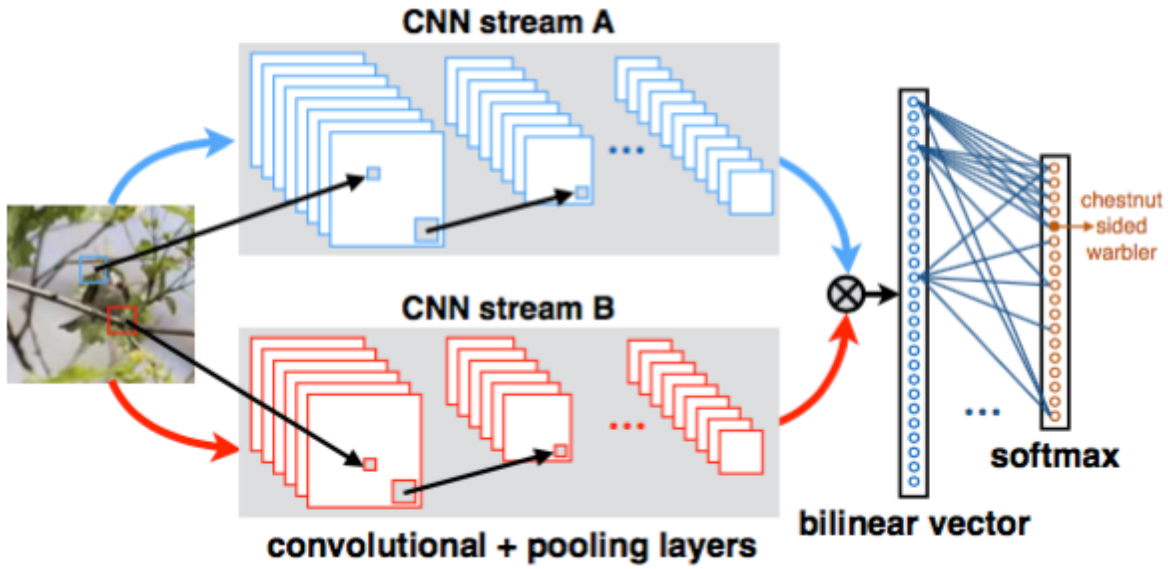
	Training phase		Testing phase		
Method	BBox Info	Part Info	BBox Info	Part Info	Accuracy (%)
Object-level attention					67.6
Part-level attention					64.9
Two-level attention					**69.7**

CNN	AlexNet	VGGNet
Accuracy (%)	69.7	77.9

Xiao T, Xu Y, Yang K, et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 842-850.

Weakly-Supervised： B-CNN

Bilinear CNN Model :



$$B = (f_A, f_B, P, C)$$

f_A, f_B: Feature Extraction;

P: Pooling;

C: Classifier.

Lin T Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1449-1457.

Weakly-Supervised：B-CNN

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

CNN-F	64x11x11 st. 4, pad 0 LRN, x2 pool	256x5x5 st. 1, pad 2 LRN, x2 pool	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 -	256x3x3 st. 1, pad 1 x2 pool	4096 drop-out	4096 drop-out	1000 soft-max
CNN-M	96x7x7 st. 2, pad 0 LRN, x2 pool	256x5x5 st. 2, pad 1 LRN, x2 pool	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 x2 pool	4096 drop-out	4096 drop-out	1000 soft-max
CNN-S	96x7x7 st. 2, pad 0 LRN, x3 pool	256x5x5 st. 1, pad 1 x2 pool	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 -	512x3x3 st. 1, pad 1 x3 pool	4096 drop-out	4096 drop-out	1000 soft-max

Chatfield K, Simonyan K, Vedaldi A, et al. Return of the devil in the details: Delving deep into convolutional nets[J]. arXiv preprint arXiv:1405.3531, 2014.
Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

Weakly-Supervised： B-CNN

Bilinear CNN Model :

1. CNNs are pre-trained on the ImageNet dataset.
2. Only use the convolutional layers
3. Sum-pooling to aggregate the bilinear features
4. classification function: logistic regression or linear SVM.
5. End-to-end training: BP

Lin T Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1449-1457.

Weakly-Supervised： B-CNN

Lin T Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1449-1457.

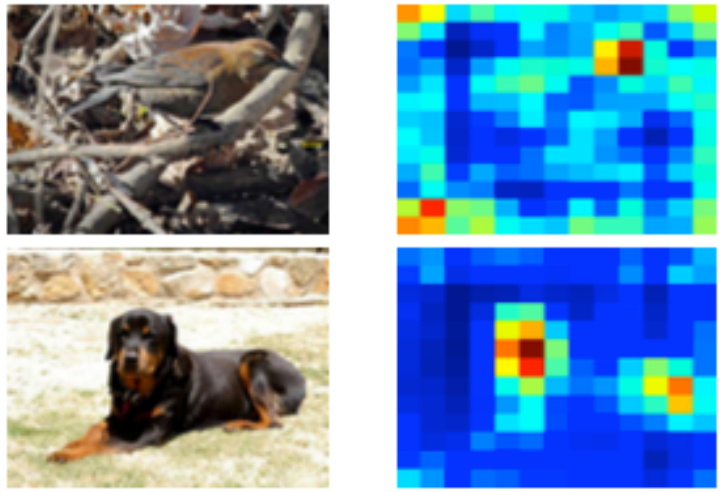
Weakly-Supervised：B-CNN

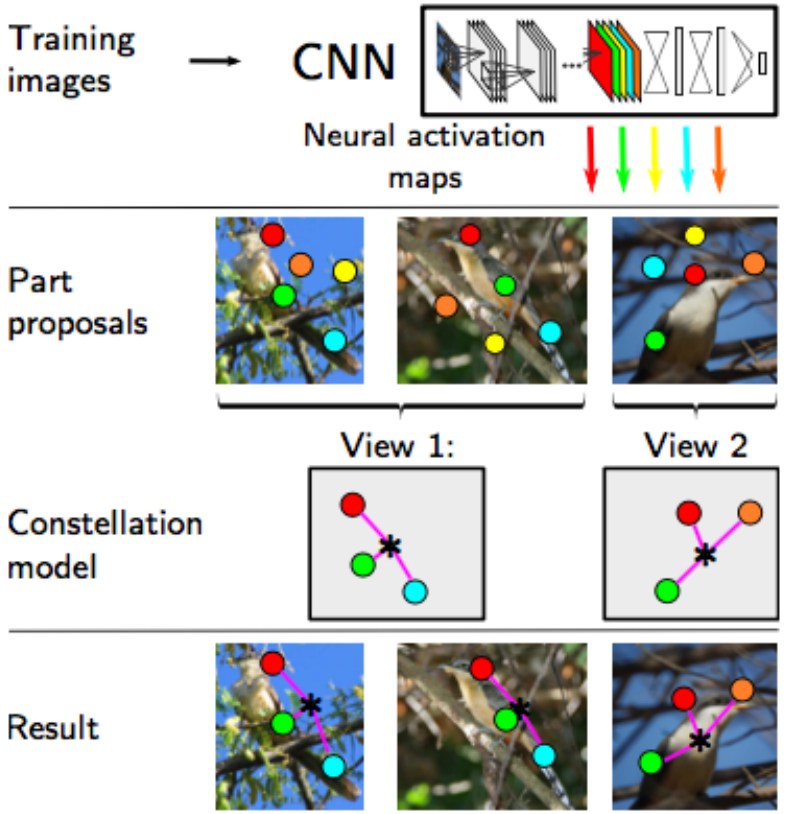
method	birds		birds + box		aircrafts		cars		FPS
	w/o ft	w/ ft	w/o ft	w/ ft	w/o ft	w/ ft	w/o ft	w/ ft	
FV-SIFT	18.8	-	22.4	-	61.0	-	59.2	-	10†
FC-CNN [M]	52.7	58.8	58.0	65.7	44.4	57.3	37.3	58.6	124
FC-CNN [D]	61.0	70.4	65.3	76.4	45.0	74.1	36.5	79.8	43
FV-CNN [M]	61.1	64.1	67.2	69.6	64.3	70.1	70.8	77.2	23
FV-CNN [D]	71.3	74.7	74.4	77.5	70.4	77.6	75.2	85.7	8
B-CNN [M,M]	72.0	78.1	74.2	80.4	72.7	77.9	77.8	86.5	87
B-CNN [D,M]	80.1	**84.1**	81.3	**85.1**	78.4	83.9	83.9	**91.3**	8
B-CNN [D,D]	80.1	84.0	80.1	84.8	76.8	**84.1**	82.9	90.6	10
Previous work	**84.1** [19], 82.0 [21] 73.9 [38], 75.7 [2]		82.8 [21], 73.5 [24] 73.0 [7], 76.4 [38]		72.5 [4], 80.7 [16]		**92.6** [21], 82.7 [16] 78.0 [4]		†on a cpu

Lin T Y, RoyChowdhury A, Maji S. Bilinear cnn models for fine-grained visual recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1449-1457.

Weakly-Supervised

Constellations:

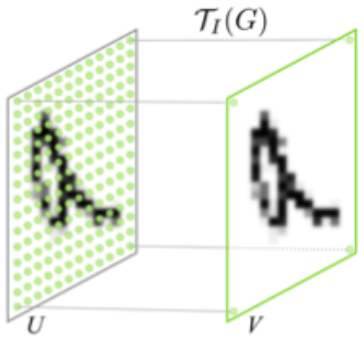


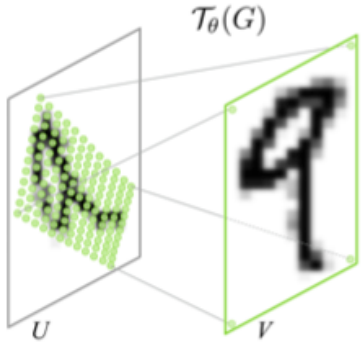


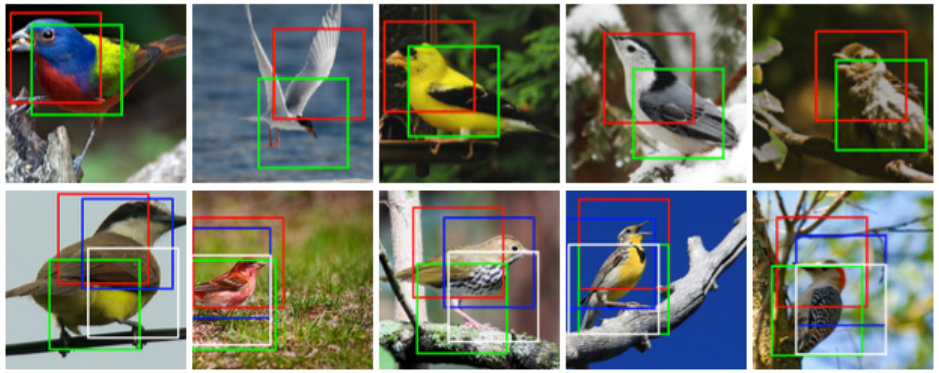
Simon M, Rodner E. Neural activation constellations: Unsupervised part model discovery with convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1143-1151.

Weakly-Supervised

Spatial Transformer Networks:





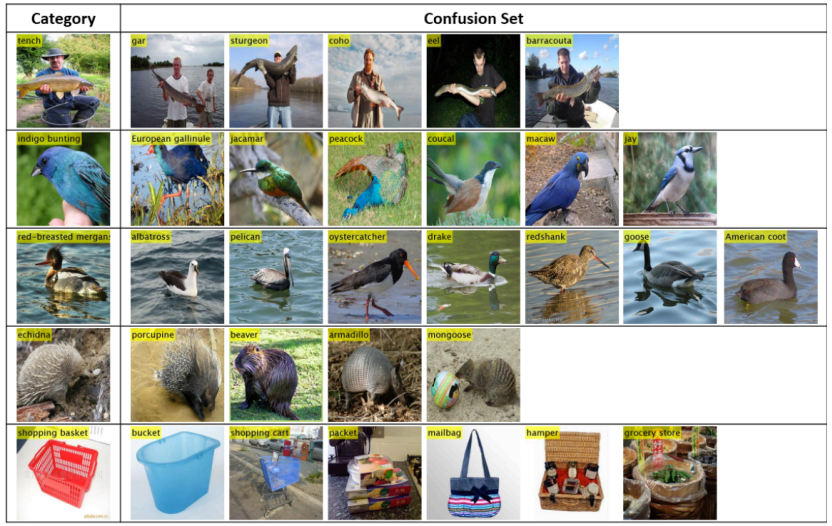


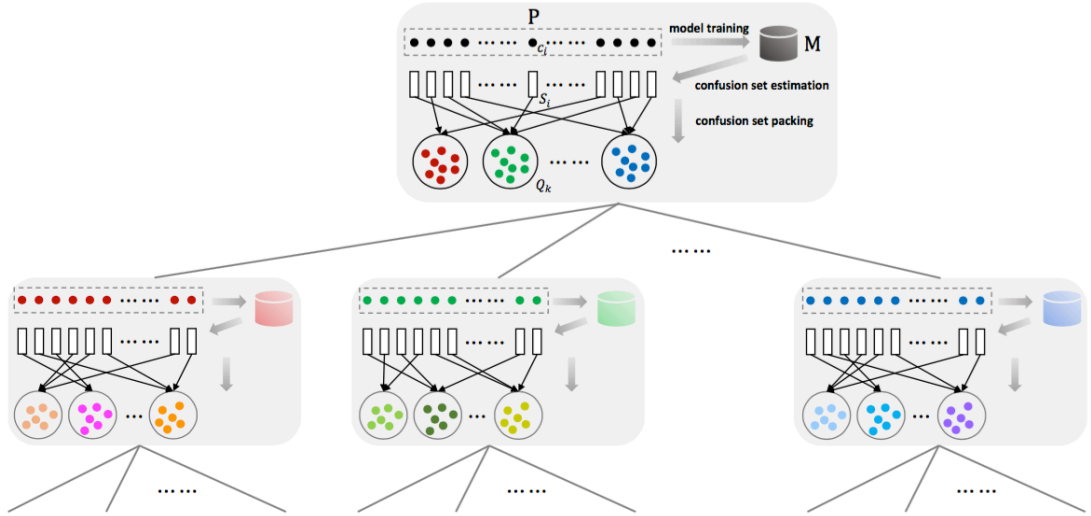
2ST: 83.9%
4ST: 84.1%

Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in Neural Information Processing Systems. 2015: 2017-2025.

Weakly-Supervised

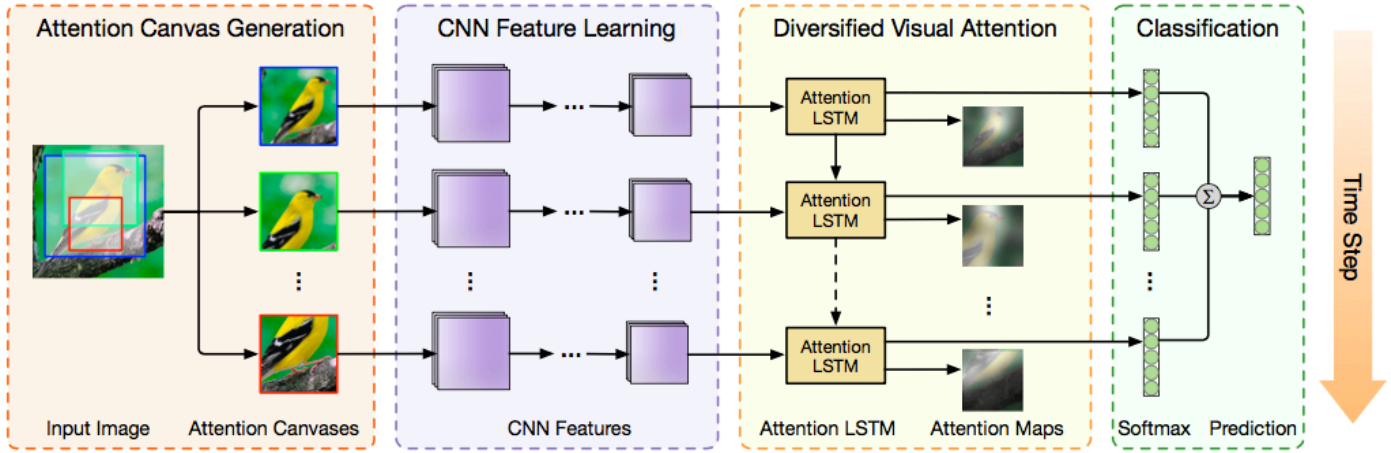
CNN Tree：





Weakly-Supervised

DVAN: Diversified Visual Attention Networks



Zhao B, Wu X, Feng J, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017.

vision@OUC

Weakly-Supervised

Weakly-supervised fine-grained classification methods' classification results on CUB-2011 dataset:

Method	Architecture	Accuracy (%)
Two-level attention	AlexNet	69.7
Two-level attention	VGGNet	77.9
Bilinear CNN	VGGNet	84.1
Constellations	VGGNet	81
ST Net	Inception	84.1
DVAN	VGGNet	79

Fine-grained Classification

Low classification accuracy:

- Part location
- feature extraction
- End - to - End
- Dataset

Thanks for watching!