MIL at ImageCLEF 2014: Scalable System for Image Annotation

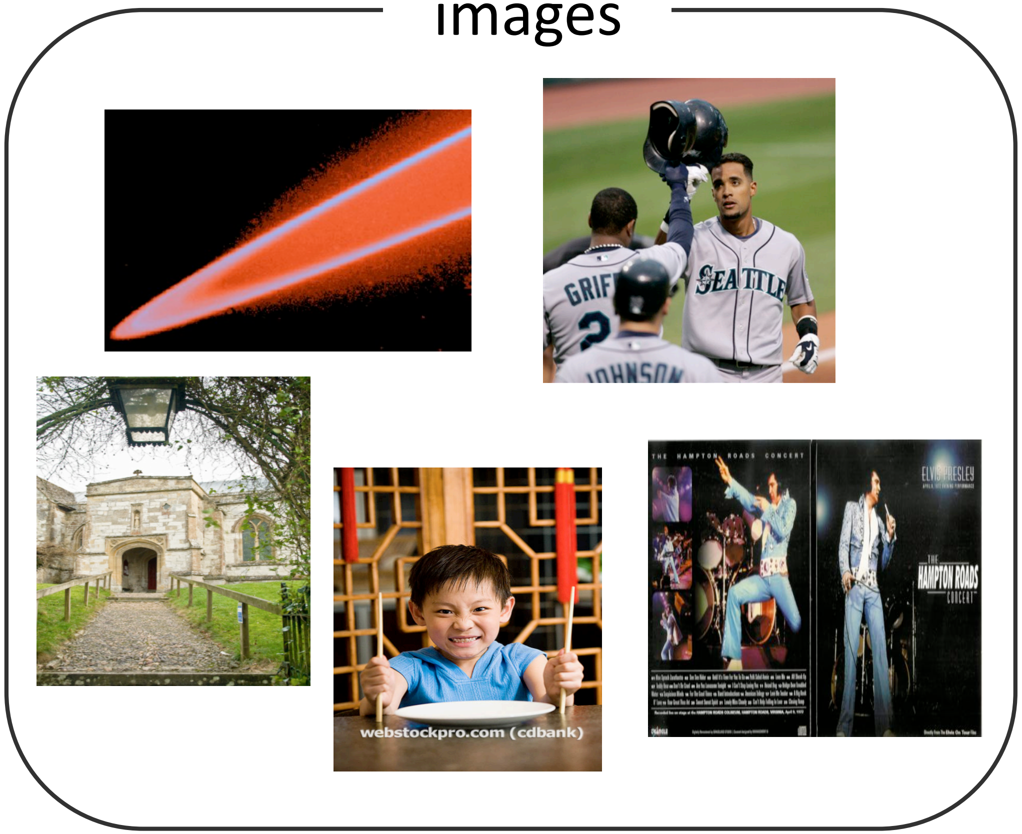
Machine Intelligence Laboratory, the University of Tokyo, Japan

Atsushi Kanehira, Masatoshi Hidaka, Yusuke Mukuta, Yuichiro Tsuchiya, Tetsuaki Mano, Tatsuya Harada

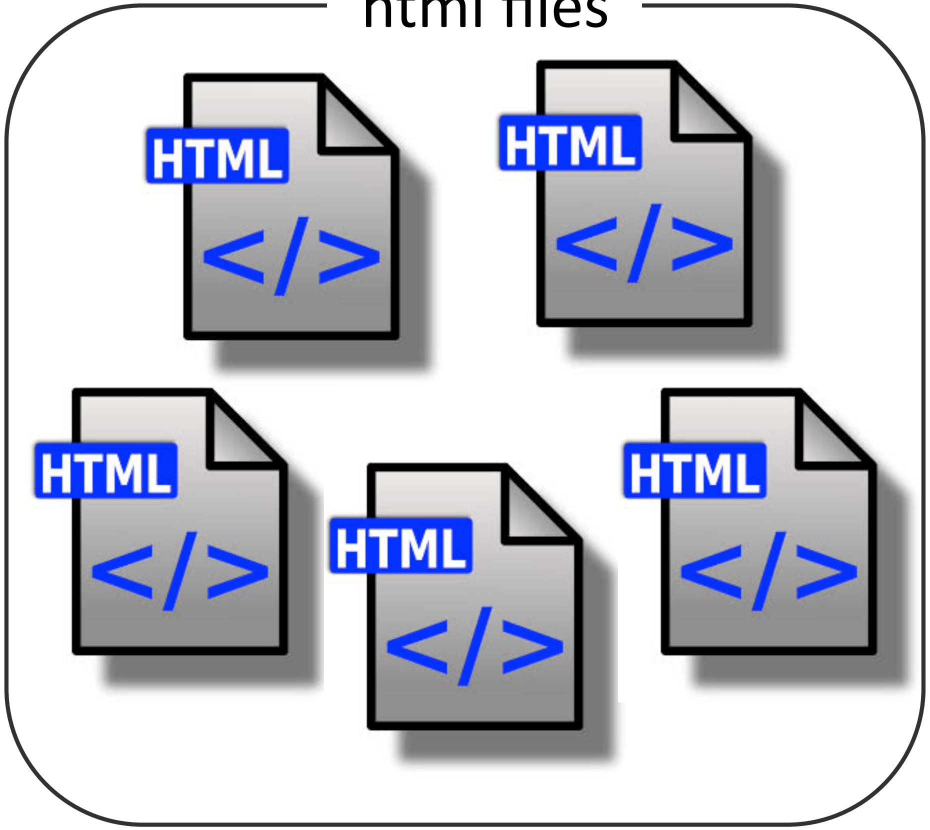
Task

- Construct image annotation system, which has scalability and high recognition performance
  - Given 500 thousands of images and webpages

images

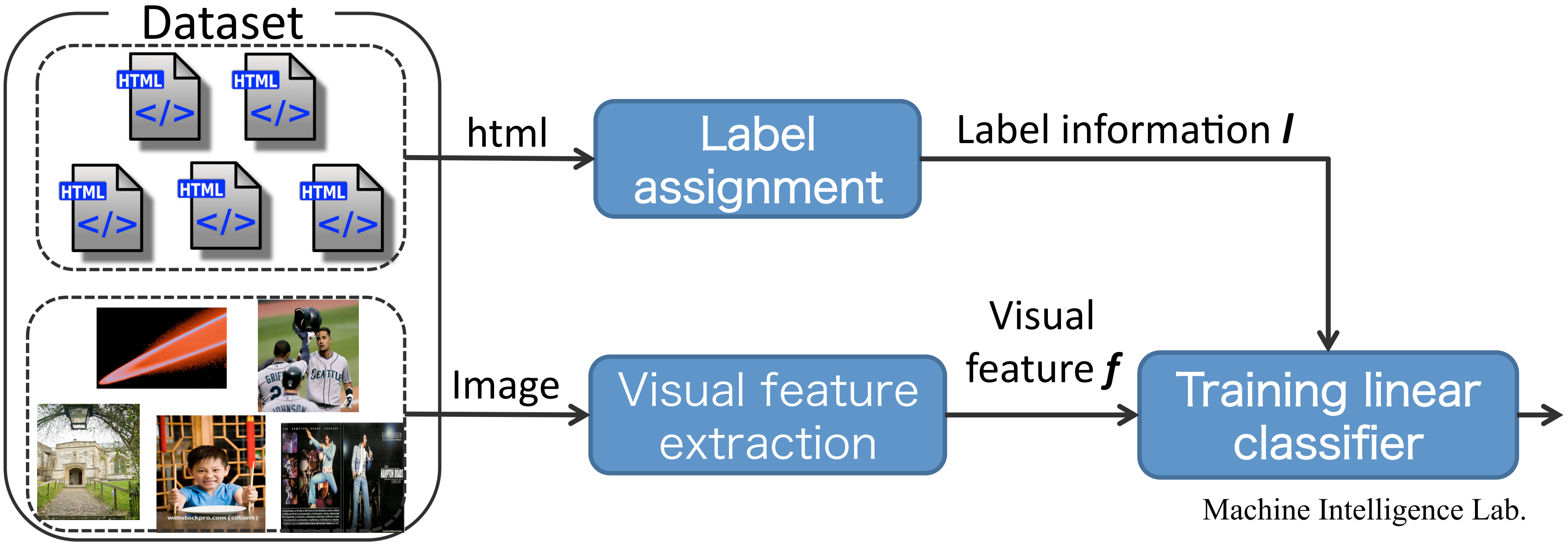


html files



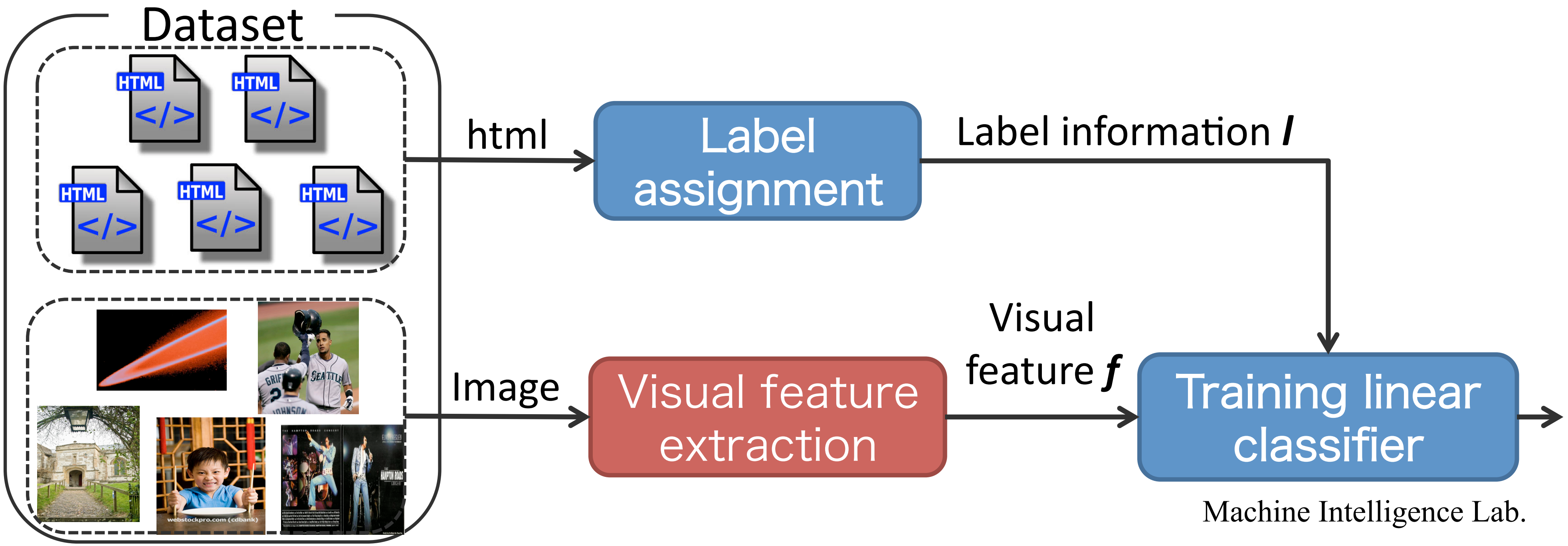
Methodology Overview

- Visual feature
  - Combination of Fisher Vector (FV) and deep convolutional neural network (CNN) based feature
- Label assignment
  - Page title and attributes of image tags
- Linear classifier
  - Passive Aggressive with Averaged Pairwise Loss (PAAPL)



Methodology Overview

- Visual feature
  - Combination of Fisher Vector (FV) and deep convolutional neural network (CNN) based feature
- Label assignment
  - Page title and attributes of image tags
- Linear classifier
  - Passive Aggressive with Averaged Pairwise Loss (PAAPL)



Visual Feature Extraction

- Combination of two types of visual features
  - Fisher Vector as generative feature
  - Deep CNN based feature as discriminative feature

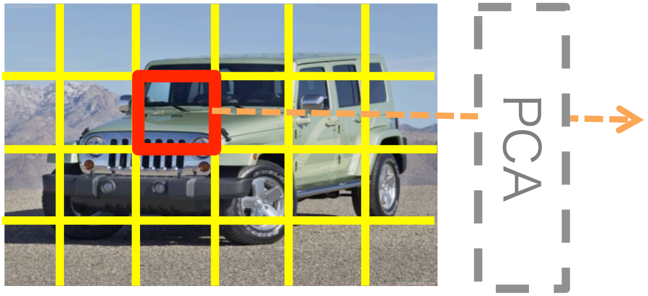
  - These can represent different kinds of information.

- Assuming that
  - These two features mutually compensate for representational ability.
  - Combining different type of features improves performance of annotation system.

Visual Feature Extraction (Fisher Vector)

- Improved Fisher Vector [F. Perronnin et al., 2010]
  - 4 local descriptors: SIFT, C-SIFT, GIST, LBP
  - Dimension of FV = 262,144 (64 × 256 × 2 × 8)
    - Dimension reduction of local feature with PCA : 64
    - Components of GMM : 256
    - Spatial pyramid : 1x1, 2x2, and 3x1 cells

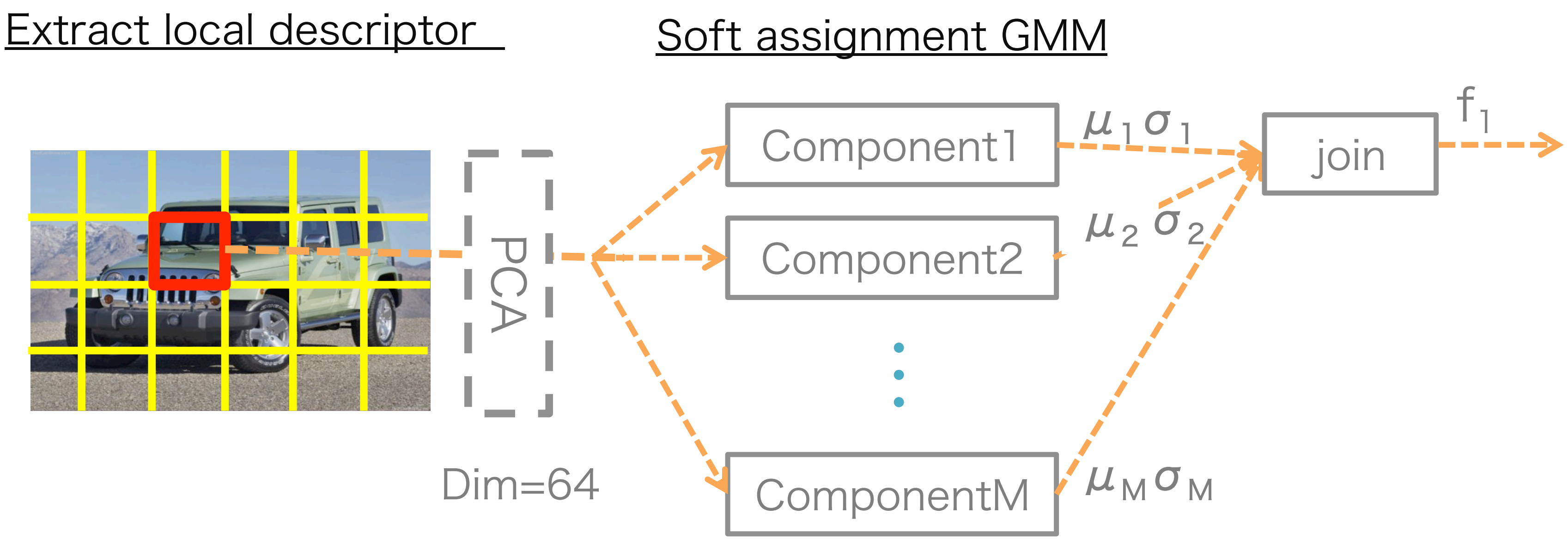
Extract local descriptor



Dim=64

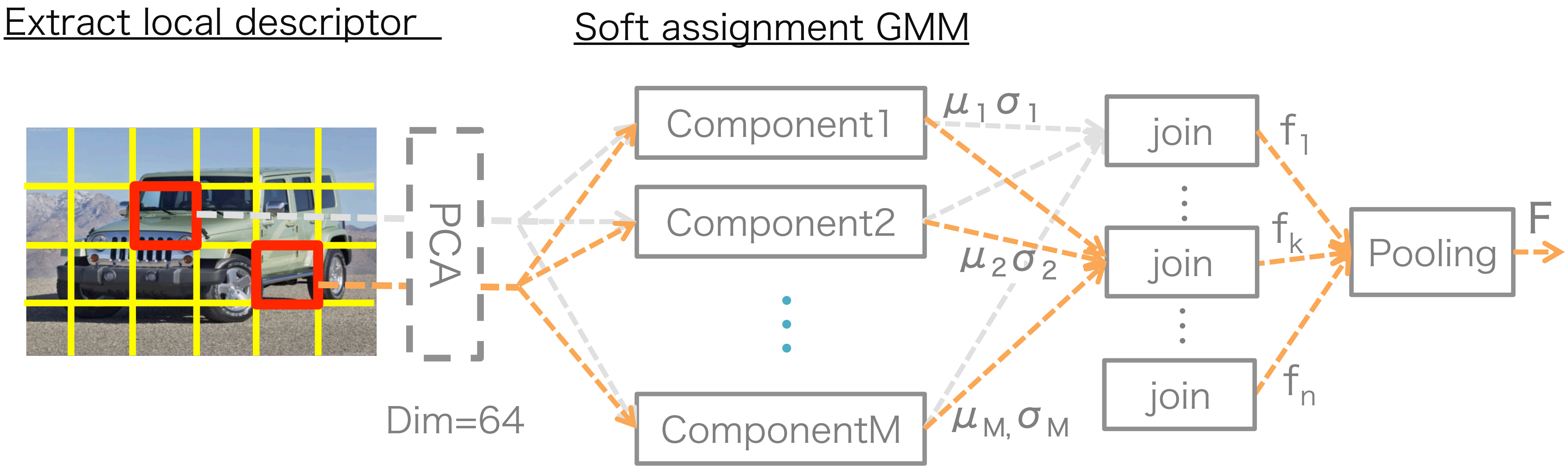
Visual Feature Extraction (Fisher Vector)

- Improved Fisher Vector [F. Perronnin et al., 2010]
  - 4 local descriptors: SIFT, C-SIFT, GIST, LBP
  - Dimension of FV = 262,144 (64 × 256 × 2 × 8)
    - Dimension reduction of local feature with PCA : 64
    - Components of GMM : 256
    - Spatial pyramid : 1x1, 2x2, and 3x1 cells



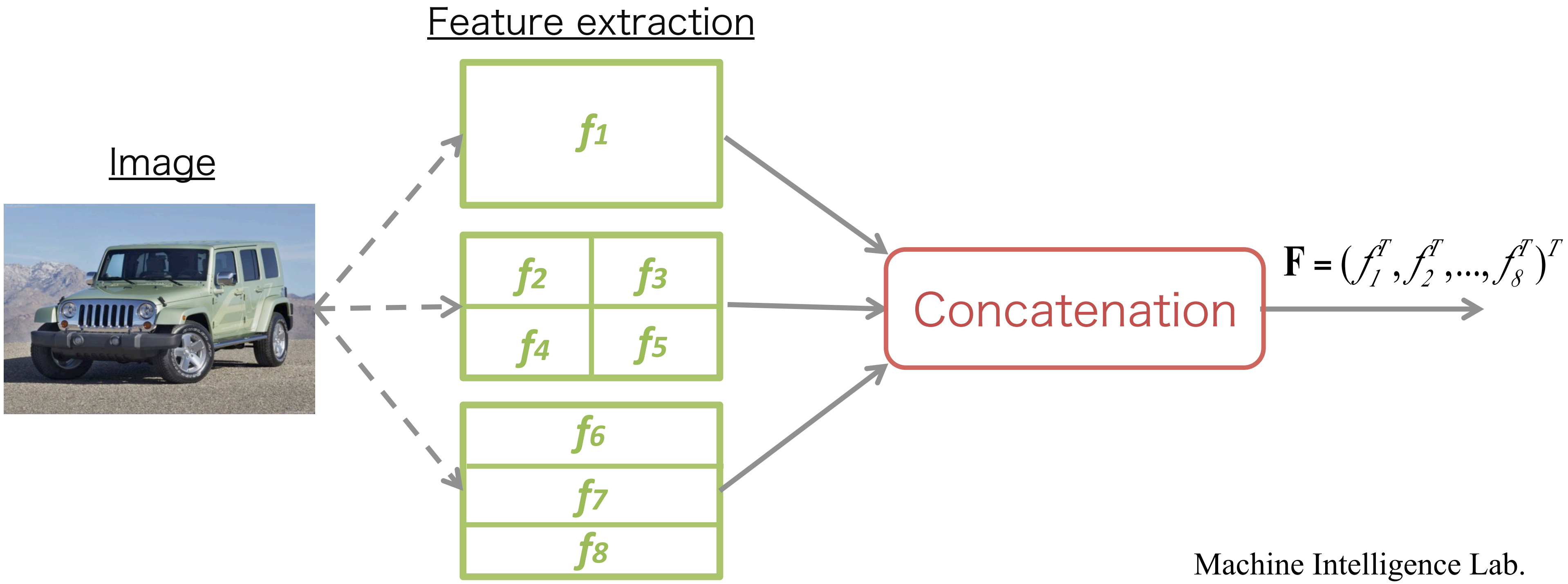
Visual Feature Extraction (Fisher Vector)

- Improved Fisher Vector [F. Perronnin et al., 2010]
  - 4 local descriptors: SIFT, C-SIFT, GIST, LBP
  - Dimension of FV = 262,144 (64 × 256 × 2 × 8)
    - Dimension reduction of local feature with PCA : 64
    - Components of GMM : 256
    - Spatial pyramid : 1x1, 2x2, and 3x1 cells



Visual Feature Extraction (Fisher Vector)

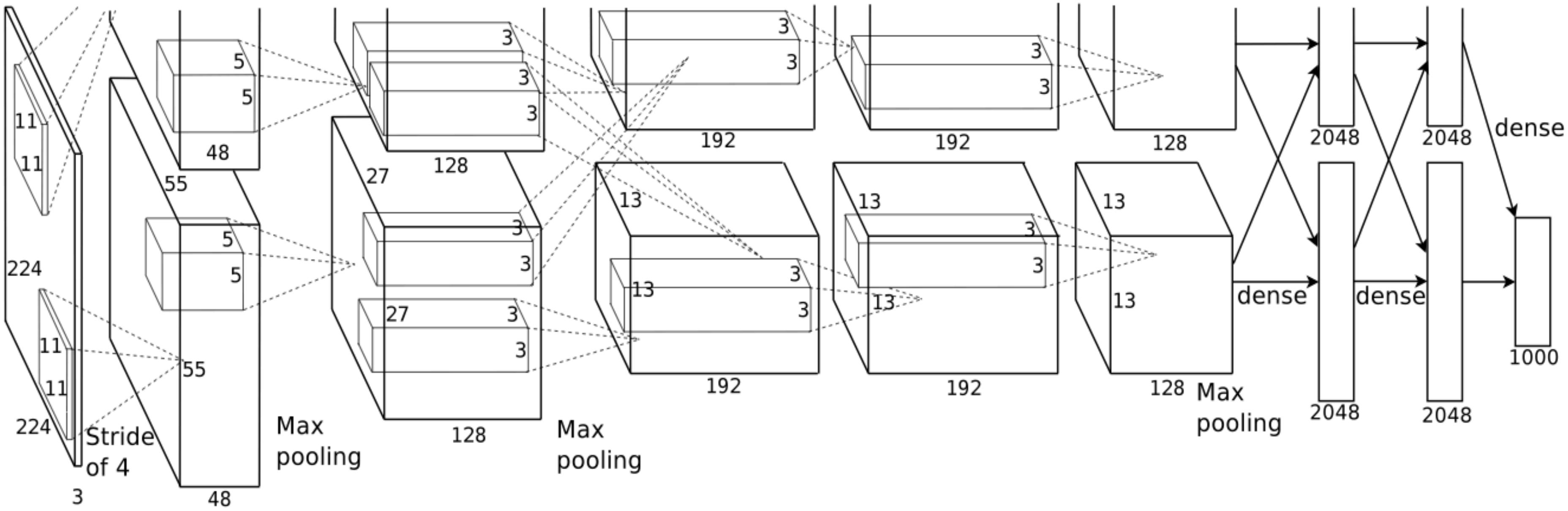
- Improved Fisher Vector [F. Perronnin et al., 2010]
  - 4 local descriptors: SIFT, C-SIFT, GIST, LBP
  - Dimension of FV = 262,144 (64 × 256 × 2 × 8)
    - Dimension reduction of local feature with PCA : 64
    - Components of GMM : 256
    - Spatial pyramid : 1x1, 2x2, and 3x1 cells



Visual Feature Extraction (deep CNN based feature)

- Deep convolutional neural network (CNN) based feature
  - Extracted from the activation of a pre-trained CNN model
  - Can be re-purposed to other tasks. [J. Donahue et al., 2014]

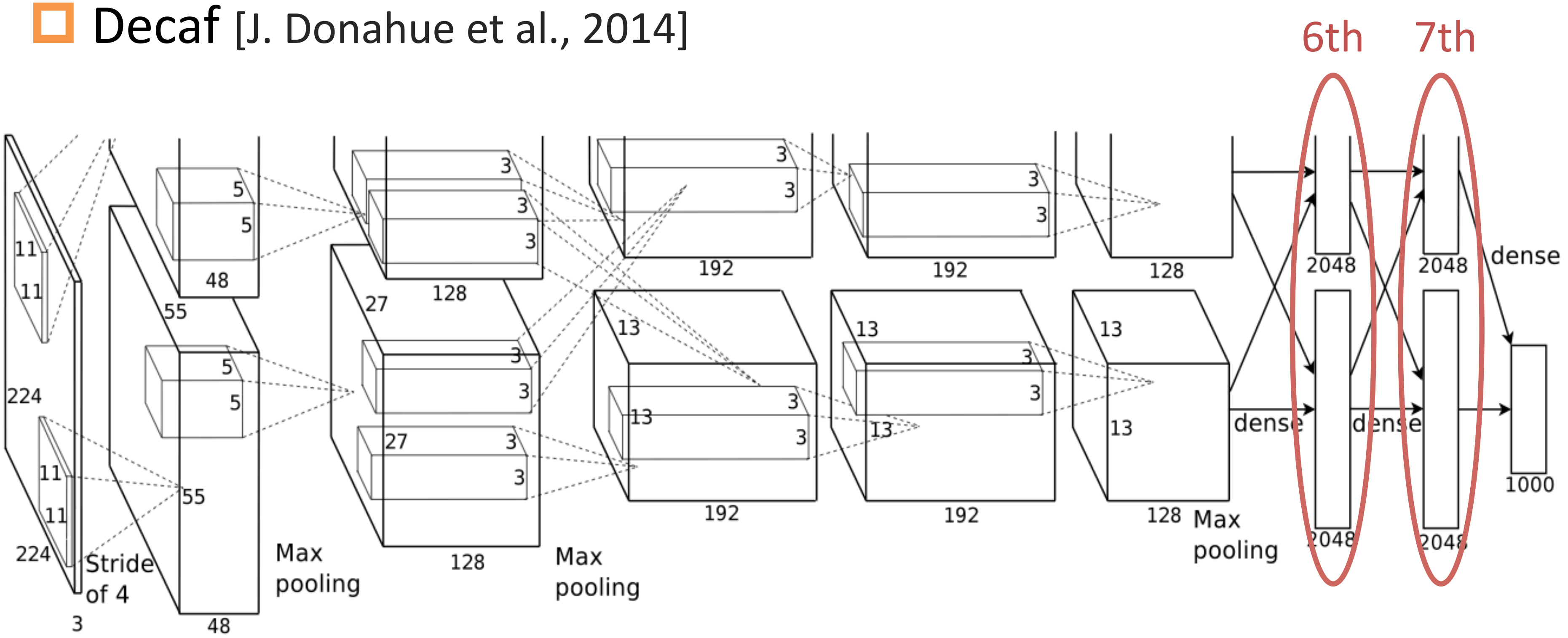
- CNN model includes five convolutional and three fully connected layers. [A. Krizhevsky et al., 2012]



ImageNet Classification with Deep Convolutional Neural Networks
In NIPS, Vol. 1, p. 4, A. Krizhevsky et al 2012

Visual Feature Extraction (deep CNN based feature)

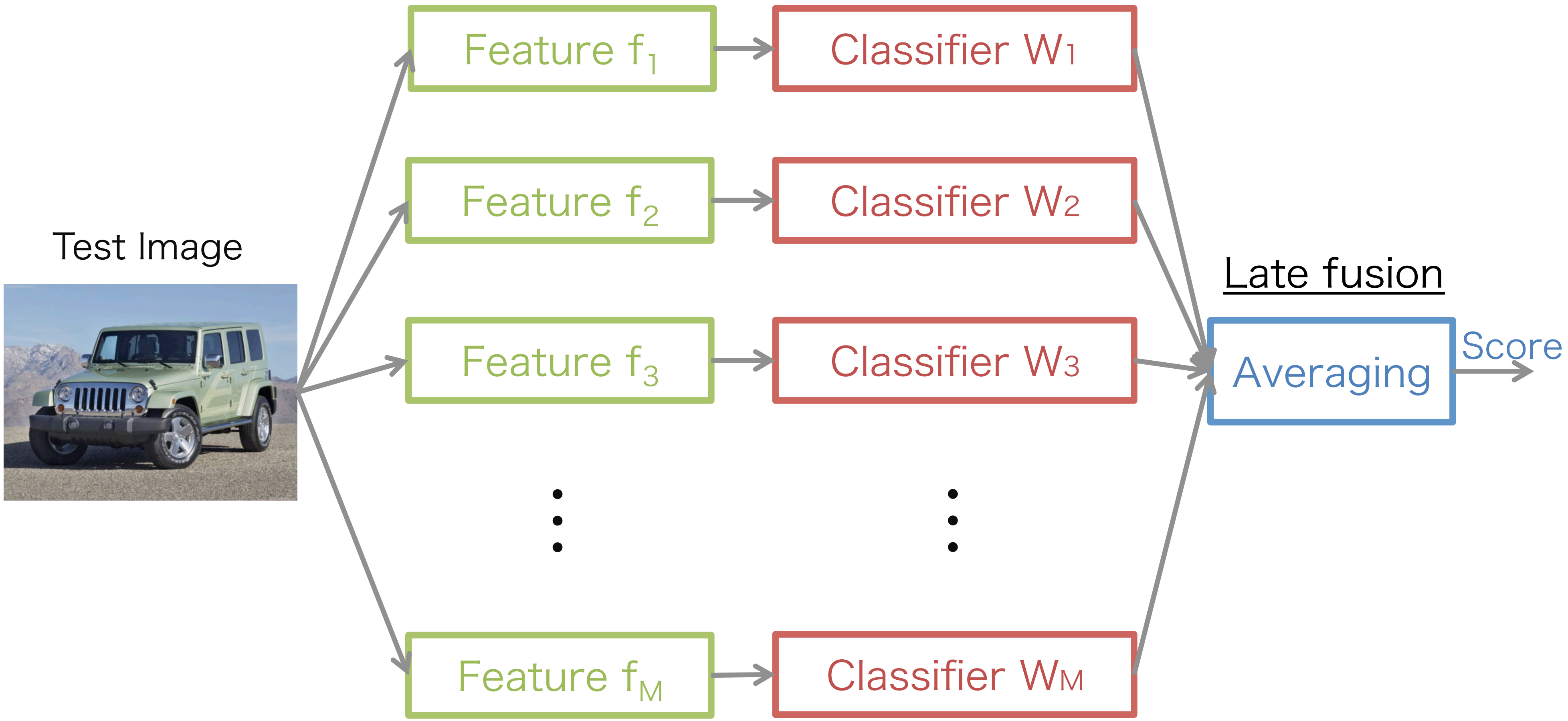
- 4 types of features
  - layer: 6th and 7th
  - activation function: linear and Rectified Linear Unit (ReLU)
    - **linear**: *f*=*x* , **ReLU**: *f*=max(*0*,*x*)
  - dimension: 4096
- Decaf [J. Donahue et al., 2014]



ImageNet Classification with Deep Convolutional Neural Networks
In NIPS, Vol. 1, p. 4, A. Krizhevsky et al 2012

Feature Combination

- Combination of Visual Features



Test Image

Feature f_1 → Classifier W_1

Feature f_2 → Classifier W_2

Feature f_3 → Classifier W_3

⋮

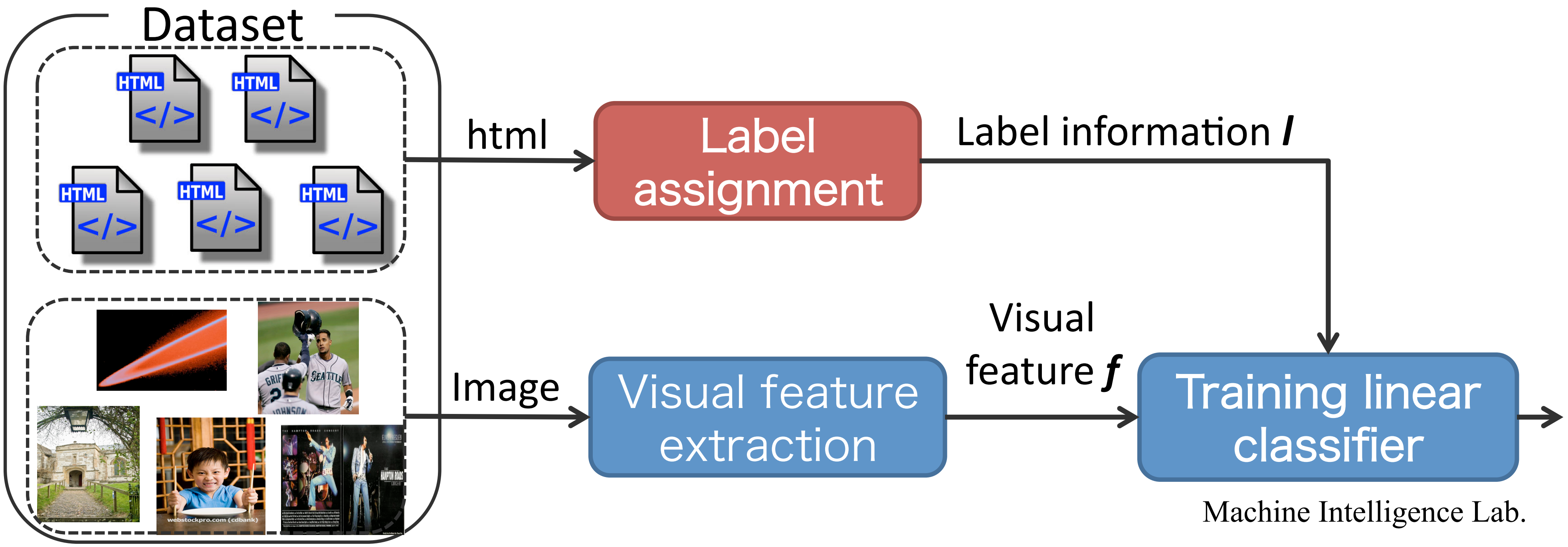
Feature f_M → Classifier W_M

Late fusion

Averaging → Score

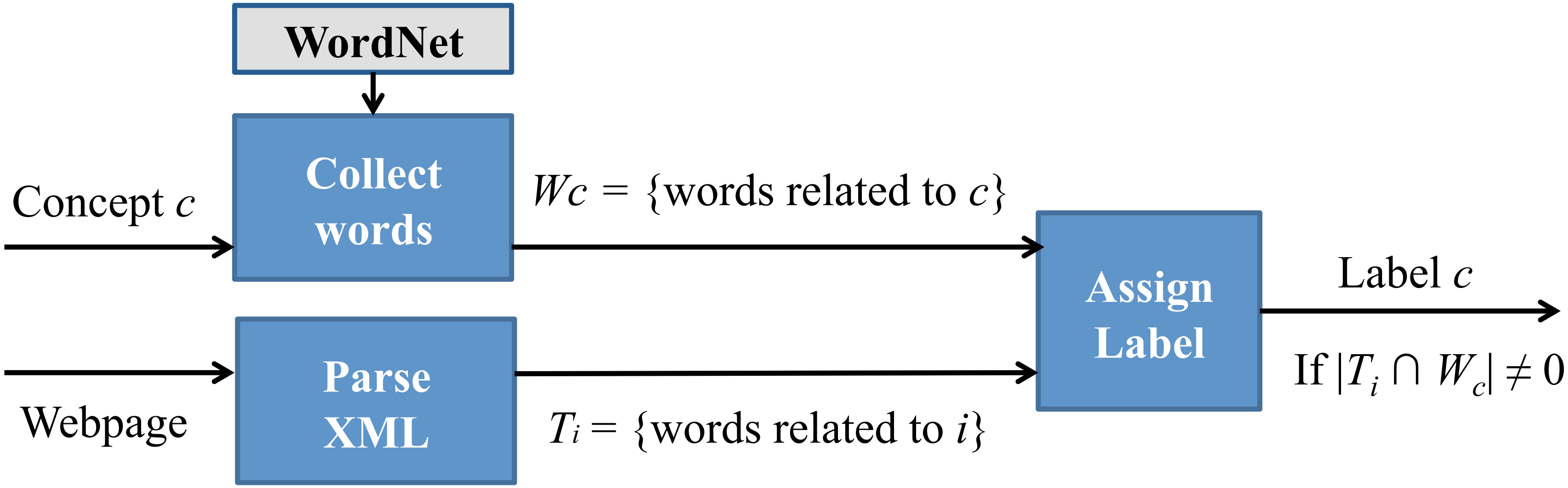
Methodology Overview

- Visual feature
  - Combination of Fisher Vector (FV) and deep convolutional neural network (CNN) based feature
- Label assignment
  - Page title and attributes of image tags
- Linear classifier
  - Passive Aggressive with Averaged Pairwise Loss (PAAPL)



Label assignment

- Extract words T_i related to the image *i*
  - Page title and src, title, alt attributes of image tag
- Extract words W_c related to the concept *c*
  - Synonyms and hyponyms of the concept *c* from WordNet
- If W_c and T_i have some common words, *i* is labeled as *c*.

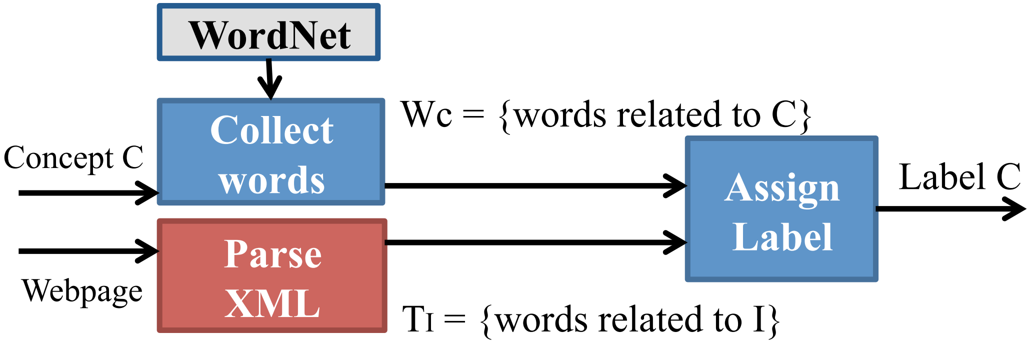

WordNet
Concept *c*
Collect words
Wc = {words related to *c*}
Webpage
Parse XML
T_i = {words related to *i*}
Assign Label
Label *c*
If $|T_i \cap W_c| \neq 0$

Label assignment (Example)

Image tag

<img src=corgi_2215843b.JPG
title = “All aboard“
alt="All aboard: the Queen and Prince
Edward with one of the family's corgi />

T = {Queen, Prince, corgi, family, abroad}



Label assignment (Example)

Image tag

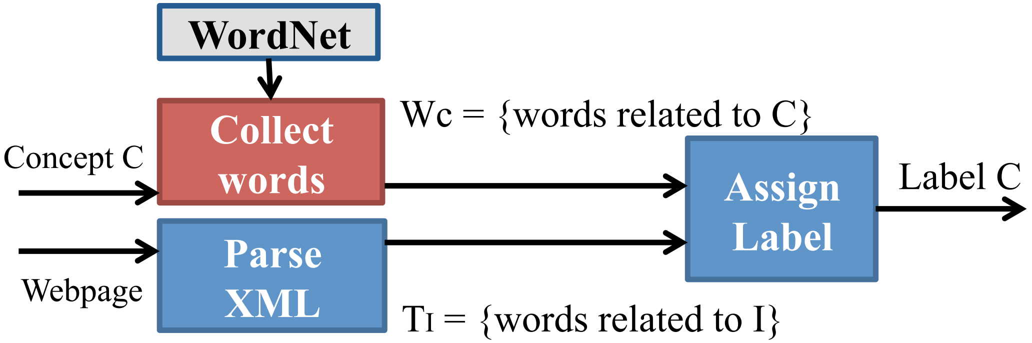
```
<img src=corgi_2215843b.JPG
        title = “All aboard“
        alt="All aboard: the Queen and Prince
        Edward with one of the family's corgi />
```

T = {Queen, Prince, corgi, family, abroad}

Dog

Get synonyms and hyponyms

W = {dog, puppy, corgi, …}



WordNet

Concept C

Collect words

W_C = {words related to C}

Webpage

Parse XML

T_I = {words related to I}

Assign Label

Label C

Label assignment (Example)

Image tag

<img src=corgi_2215843b.JPG
title = “All aboard“
alt="All aboard: the Queen and Prince Edward with one of the family's corgi />

T = {Queen, Prince, corgi, family, abroad}

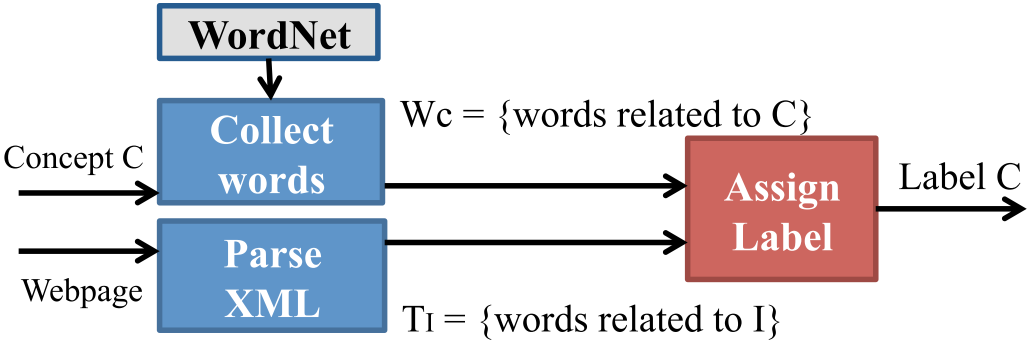
Judge having common words or not

Have common word: “corgi”

Dog

Get synonyms and hyponyms

W = {dog, puppy, corgi, …}



WordNet

Concept C

Collect words

Webpage

Parse XML

W_C = {words related to C}

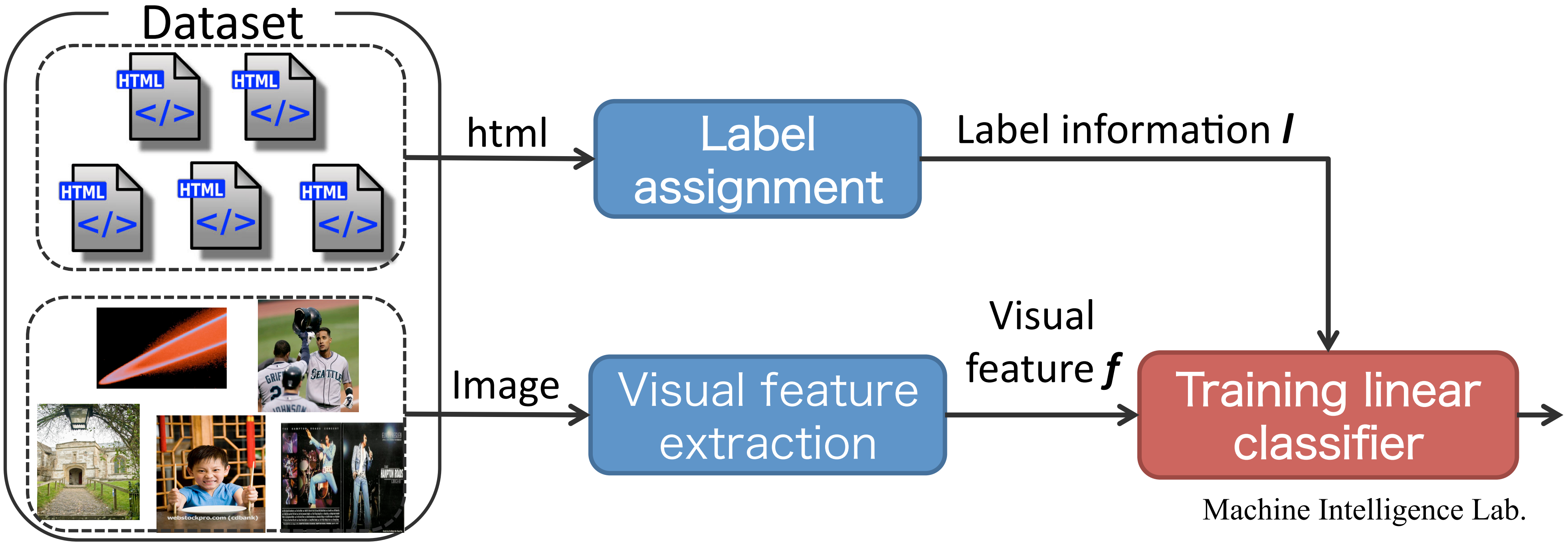
T_I = {words related to I}

Assign Label

Label C

Methodology Overview

- Visual feature
  - Combination of Fisher Vector (FV) and deep convolutional neural network (CNN) based feature
- Label assignment
  - Page title and attributes of image tags
- Linear classifier
  - Passive Aggressive with Averaged Pairwise Loss (PAAPL)



Training linear Classifier (PAAPL)

- Passive Aggressive with Averaged Pairwise Loss (PAAPL) [Y. Ushiku et al., 2012]
  - Extension of Passive Aggressive (PA) for multi-label tasks
    - Fast convergence : handle multiple pairs of concept for one sample
    - Scalability and robustness to outliers

Training linear Classifier (PAAPL)

Update rule of PAAPL

Training sample

- visual feature $\boldsymbol{x}_t$
- correct labels $\mathbf{y}$ = (bird, train)

1. Calculate scores of all concepts.

2. Pick [min/max]-score from [correct/incorrect] labels.

3. Update the model using hinge-loss.

4. For all correct labels, repeat 2,3.

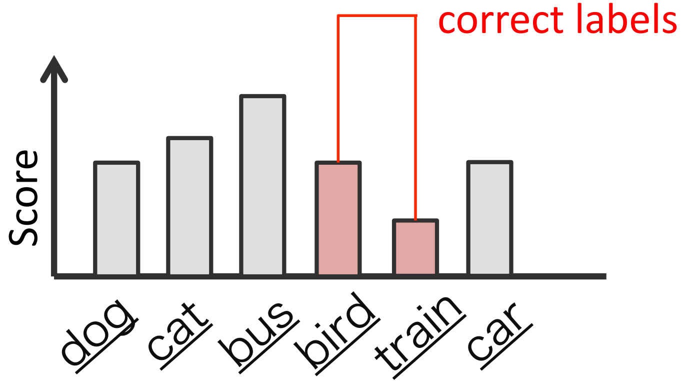
Training linear Classifier (PAAPL)

Update rule of PAAPL

Training sample

- visual feature $\boldsymbol{x}_t$
- correct labels $\mathbf{y}$ = (bird, train)

1. Calculate scores of all concepts.



2. Pick [min/max]-score from [correct/incorrect] labels.

3. Update the model using hinge-loss.

4. For all correct labels, repeat 2,3.

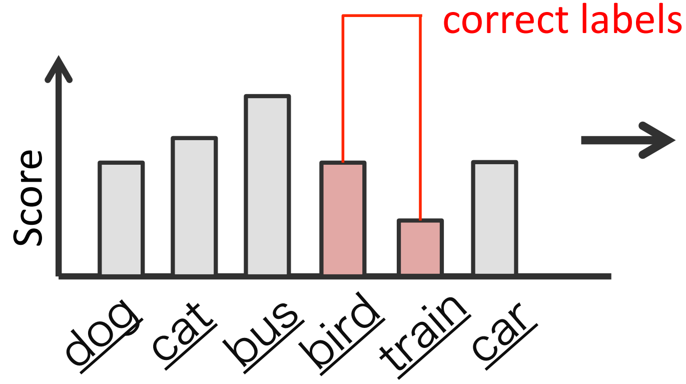
Training linear Classifier (PAAPL)

Update rule of PAAPL

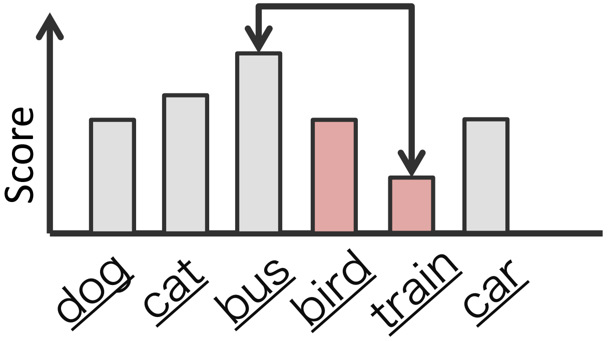
Training sample

- visual feature $\boldsymbol{x}_t$
- correct labels $\mathbf{y}$ = (bird, train)

1. Calculate scores of all concepts.



2. Pick [min/max]-score from [correct/incorrect] labels.



3. Update the model using hinge-loss.

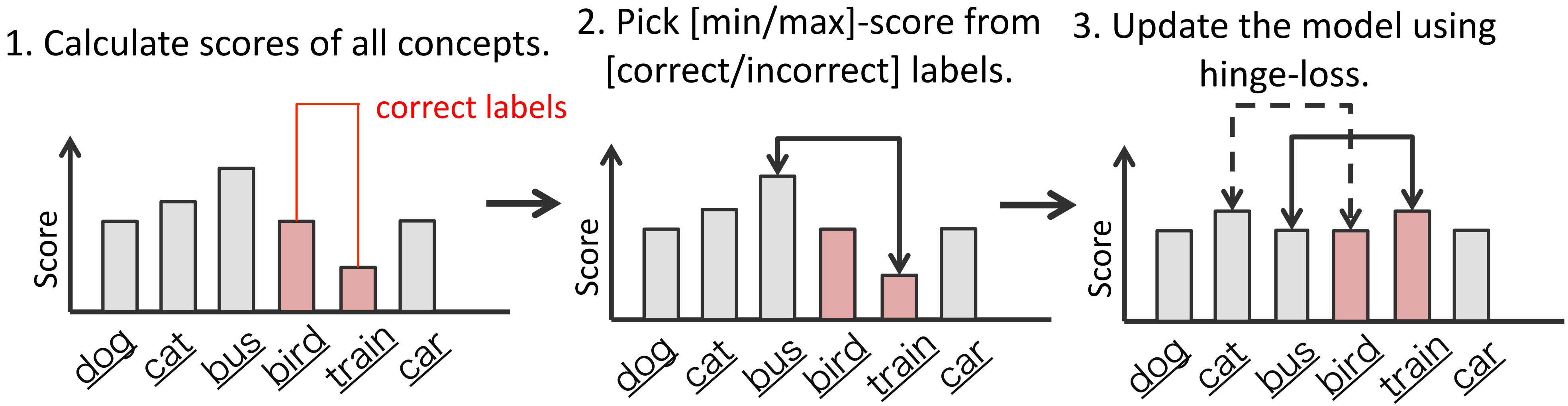
4. For all correct labels, repeat 2,3.

Training linear Classifier (PAAPL)

Update rule of PAAPL

Training sample

- visual feature $\boldsymbol{x}_t$
- correct labels $\mathbf{y}$ = (bird, train)



1. Calculate scores of all concepts.

2. Pick [min/max]-score from [correct/incorrect] labels.

3. Update the model using hinge-loss.

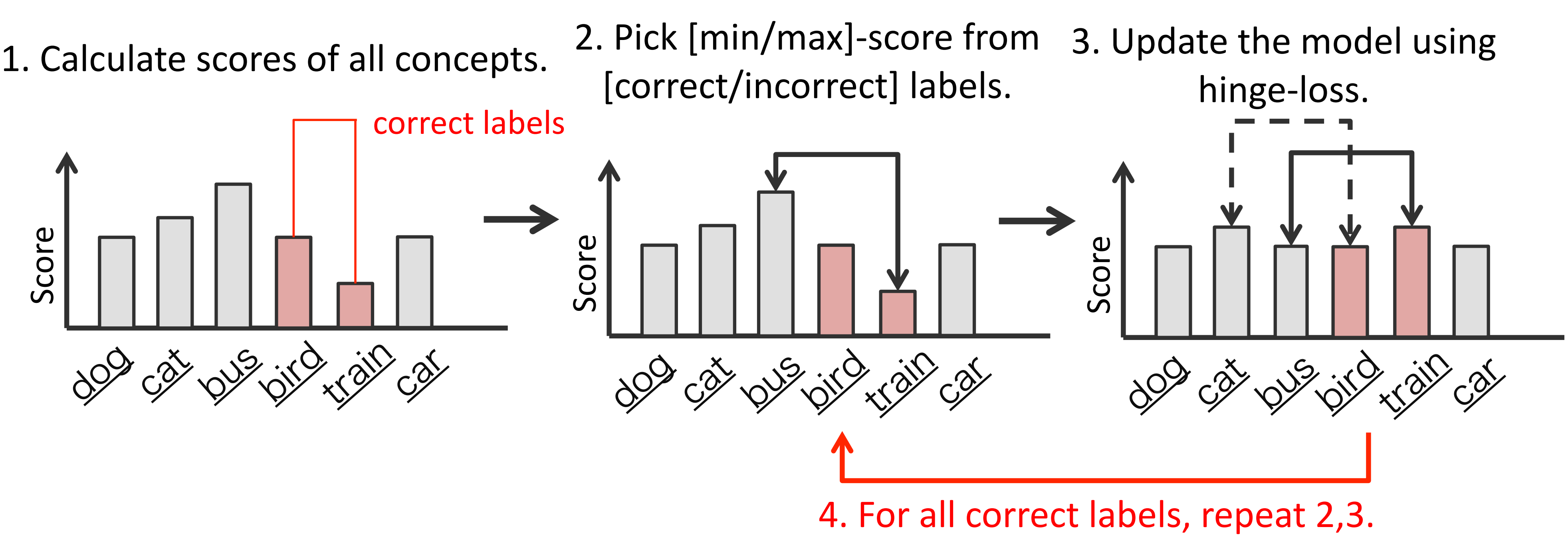
4. For all correct labels, repeat 2,3.

Training linear Classifier (PAAPL)

Update rule of PAAPL

Training sample

- visual feature $\boldsymbol{x}_t$
- correct labels $\mathbf{y}$ = (bird, train)



1. Calculate scores of all concepts.

2. Pick [min/max]-score from [correct/incorrect] labels.

3. Update the model using hinge-loss.

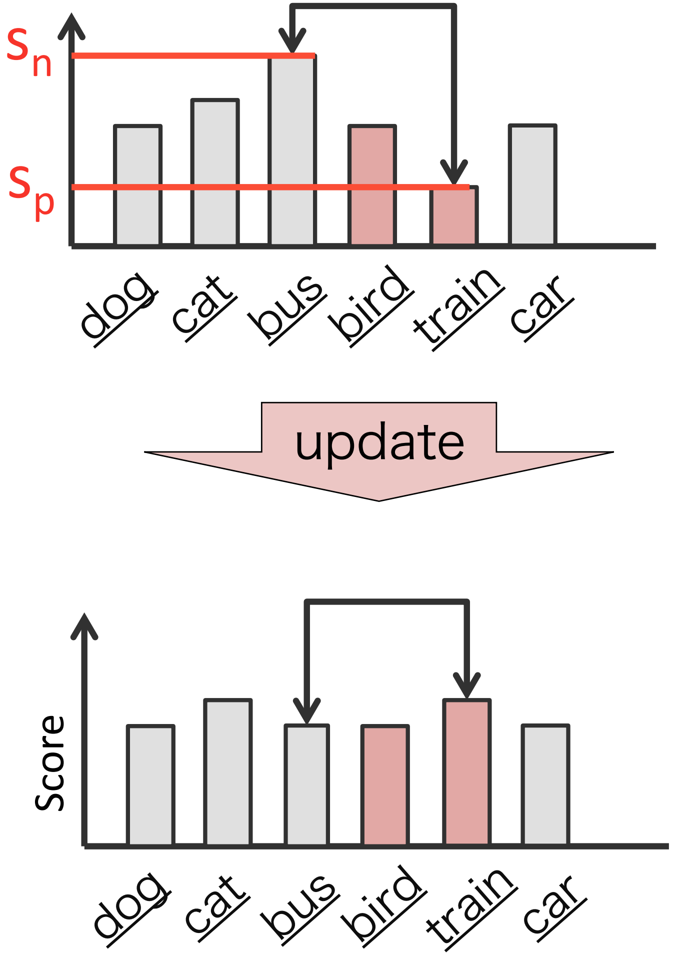
4. For all correct labels, repeat 2,3.

Training linear Classifier (PAAPL)

Update rule of PAAPL

Hinge-loss

$$l = \begin{cases} 0 & (\text{if } s_p - s_n > 1) \\ 1 - (s_p - s_n) & (\text{otherwise}) \end{cases}$$



Update model

$$\mathbf{w}_{t+1}^{p} = \mathbf{w}_{t}^{p} + \frac{l}{2|\mathbf{x}_t|^2 + 1/D}\mathbf{x}_t$$

$$\mathbf{w}_{t+1}^{n} = \mathbf{w}_{t}^{n} - \frac{l}{2|\mathbf{x}_t|^2 + 1/D}\mathbf{x}_t$$

Experiment

- The number of samples
  - Train : 500,000
    - 121,331 are labeled at validation.
    - 210,388 are labeled at test.
  - Development : 1,940
  - Test : 7,291
- Decide concepts with scores in the top 4% of all given concepts.
- 3 experiments
  1. To find the best combination of FVs
  2. To find the best combination of deep CNN features
  3. To try feature combination and compare with single features

Result (FV)

- Best combination of FVs
  - 4 features (4 local descriptors)
  - Combination of all features achieved the best performance.

result

C-SIFT	GIST	LBP	SIFT	MF-samples (devel)
✓				0.286
	✓			0.292
		✓		0.284
			✓	0.294
✓	✓	✓		0.347
✓	✓		✓	0.350
✓		✓	✓	0.348
	✓	✓	✓	0.344
✓	✓	✓	✓	**0.356**

Result (FV)

- Best combination of FVs
  - 4 features (4 local descriptors)
  - Combination of all features achieved the best performance.

result

C-SIFT	GIST	LBP	SIFT	MF-samples (devel)
✔				0.286
	✔			0.292
		✔		0.284
			✔	0.294
✔	✔	✔		0.347
✔	✔		✔	0.350
✔		✔	✔	0.348
	✔	✔	✔	0.344
✔	✔	✔	✔	**0.356**

Combination of more feature is better

Result (deep CNN based feature)

- Best combination of deep CNN based features
  - 4 features (layer and activation function)
  - Combination of all features achieved the best performance.

result

6th (ReLU)	6th	7th (ReLU)	7th	MF-samp (devel)
✔				0.325
	✔			0.348
		✔		0.346
			✔	0.360
✔		✔		0.358
	✔		✔	0.371
✔			✔	0.356
	✔	✔		0.366
✔	✔	✔	✔	**0.373**

Result (deep CNN based feature)

- Best combination of deep CNN based features
  - 4 features (layer and activation function)
  - Combination of all features achieved the best performance.

result

6th (ReLU)	6th	7th (ReLU)	7th	MF-samp (devel)
✔				0.325
	✔			0.348
		✔		0.346
			✔	0.360
✔		✔		0.358
	✔		✔	0.371
✔			✔	0.356
	✔	✔		0.366
✔	✔	✔	✔	**0.373**

Combination of more feature is better

Result (deep CNN based feature)

- Best combination of deep CNN based features
  - 4 features (layer and activation function)
  - Combination of all features achieved the best performance.

result

6th (ReLU)	6th	7th (ReLU)	7th	MF-samp (devel)
✔				0.325
	✔			0.348
		✔		0.346
			✔	0.360
✔		✔		0.358
	✔		✔	0.371
✔			✔	0.356
	✔	✔		0.366
✔	✔	✔	✔	**0.373**

Linear activation is better than ReLU

Discussion (experiment 1 and 2)

□ The more features combined, the better performance the system have.

□ ReLU reduces representational ability because it eliminates negative elements.

Result (feature combination)

- Compare performance
  - FVs and deep CNN based features and combination of them.

result

RUN	4 FVs	4 CNNs	MF-samples (devel)	MF-samples (test)
1	✔		0.356	0.240
2		✔	0.373	0.265
3	✔	✔	**0.394**	**0.275**

Increase
0.021 (devel)
0.010 (test)

- Combined feature is better than single one.

Result (feature combination)

□ Compare performance

- ➢ FVs and deep CNN based features and combination of them.

result

RUN	4 FVs	4 CNNs	MF-samples (devel)	MF-samples (test)
1	✔		0.356	0.240
2		✔	0.373	0.265
3	✔	✔	**0.394**	**0.275**

Increase
0.038 (devel)
0.035 (test)

➢ Combined feature is better than single one.

Conclusion

- Goal
  - Construction of image annotation system, which has scalability and high recognition performance
- Methodology
  - Visual feature : Combination of Fisher Vector and deep CNN based feature
  - Label assignment : Page title and attributes of image tag
  - Training classifier : Passive Aggressive with Pairwise Loss (PAAPL)
- Result
  - Combination of these features contributes to improvement of recognition performance.

Thank you for kind attention.

Experiment Results – Text Extraction

- Experiment of using text around image tag (imageCLEF 2013)

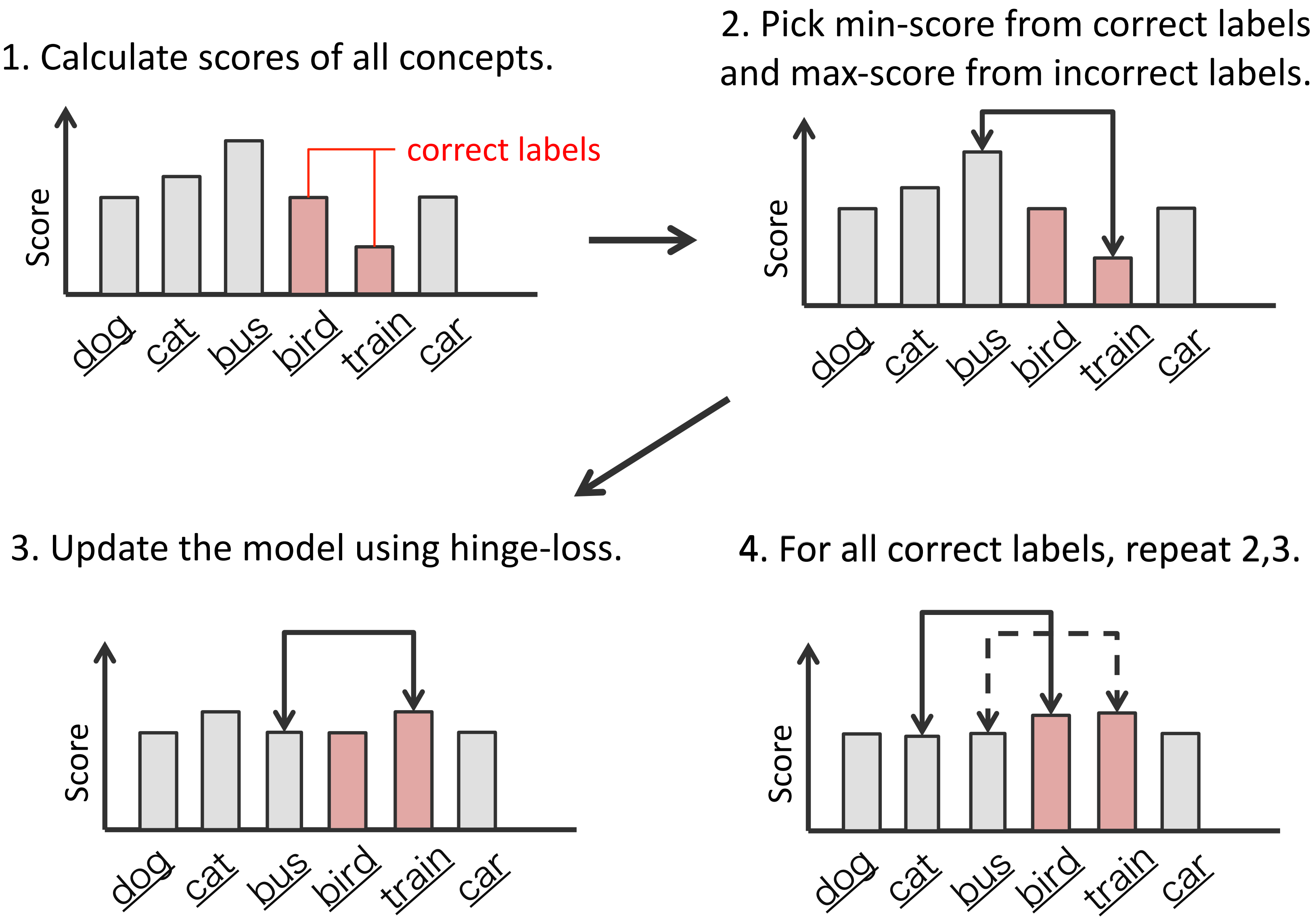
Text around image [max word distance]	MF-samples [%]	Number of images with label	Average number of labels
-	26.0	111247	0.6
10	26.1	140448	0.9
100	23.0	186394	2.6
1000	20.7	193971	5.3

- Experiment of using Synonym and hyponym (imageCLEF 2013)

Synonym	Hyponym	MF-samples [%]
		23.4
✔		23.2
	✔	26.1
✔	✔	**26.6**

Training linear Classifier (PAAPL)

Update rule of PAAPL



1. Calculate scores of all concepts.

2. Pick min-score from correct labels and max-score from incorrect labels.

3. Update the model using hinge-loss.

4. For all correct labels, repeat 2,3.