MIL at ImageCLEF 2013: Scalable System for Image Annotation

Machine Intelligence Laboratory, the University of Tokyo, Japan

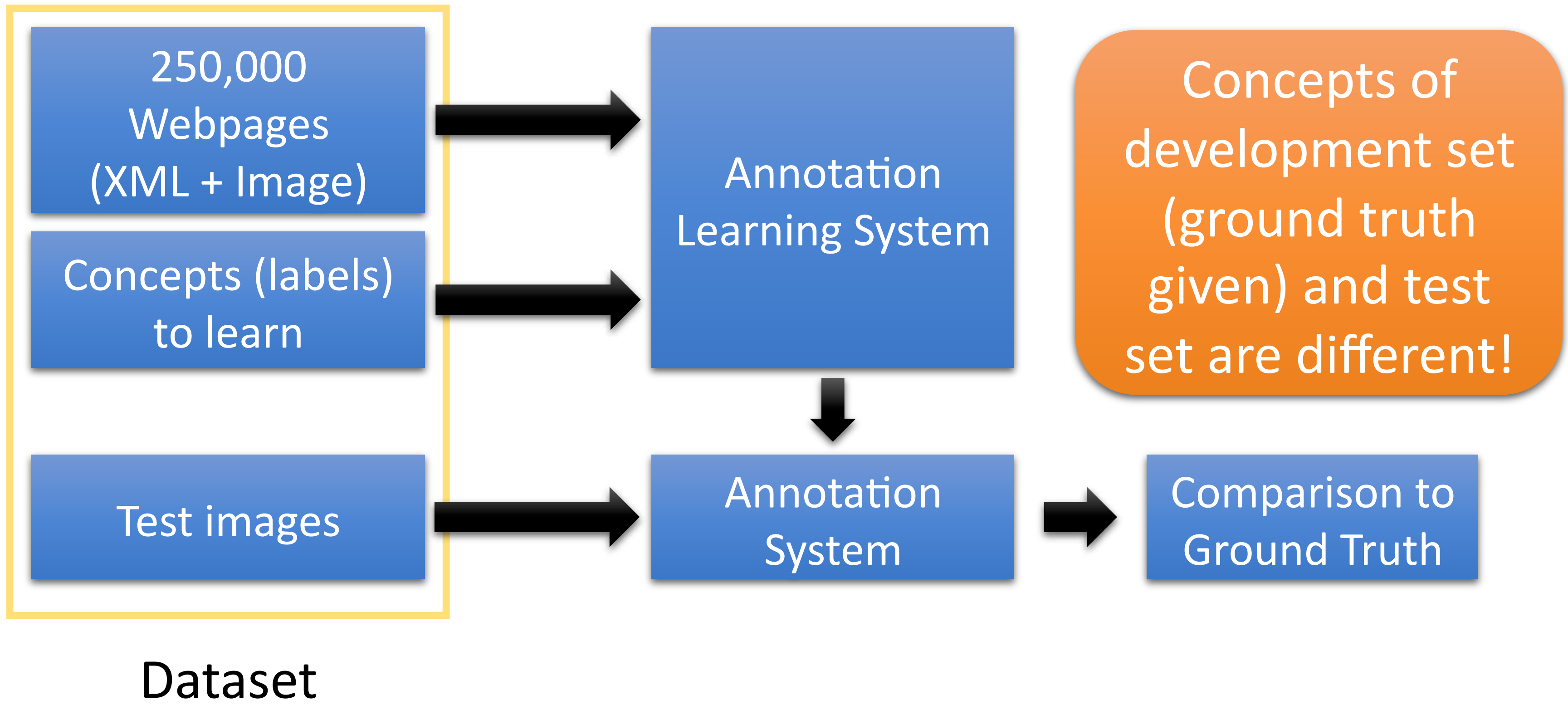
Masatoshi Hidaka

Naoyuki Gunji

Tatsuya Harada

Scalable Concept Image Annotation Task

- To make image annotation system from wild web data



Contents

- Scalable Concept Image Annotation Task
  - Image Feature; Fisher Vector, state-of-the-art
  - Textual Feature; our original method which **supports concept set change**
  - Multilabel Annotation Learning; PAAPL, scalable to the dataset size

Learning Pipeline

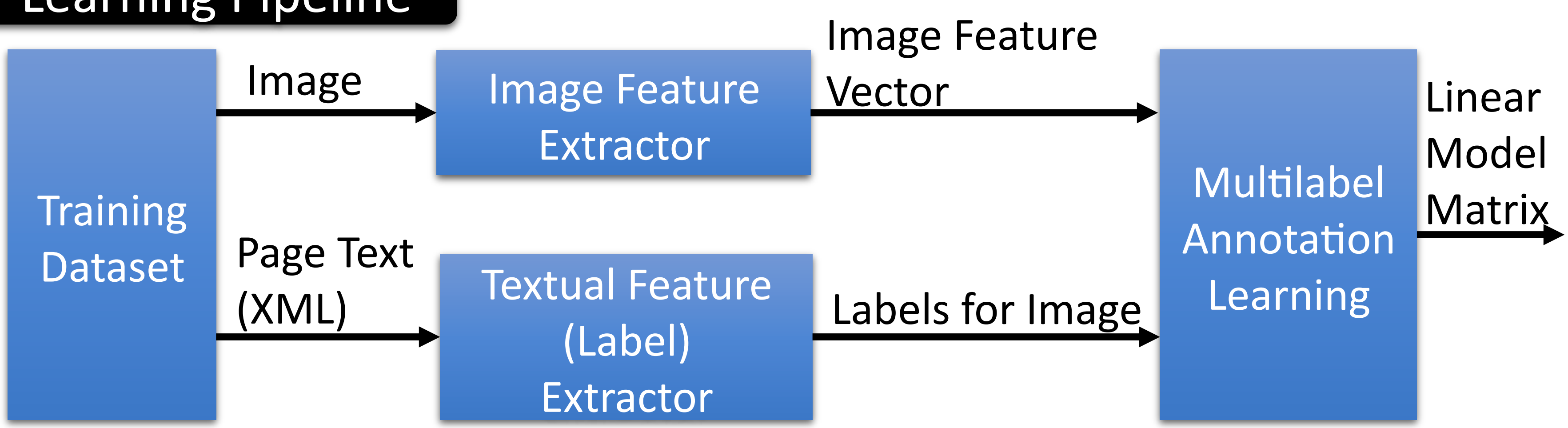


Image Feature – Fisher Vector [Perronnin et al., 2010]

- Local descriptor
  - SIFT, C-SIFT, GIST, LBP are used separately
  - Using GIST not for global image feature, but for local descriptor
- Statistic calculation
  - Calculate local descriptors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ statistic using Gaussian Mixture Model $w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ calculated by random sample in dataset beforehand

$$\boldsymbol{u}_i = \frac{1}{N\sqrt{w_i}} \sum_{n=1}^{N} \gamma_n(i) \boldsymbol{\Sigma}_i^{-\frac{1}{2}} (\boldsymbol{x}_n - \boldsymbol{\mu}_i)$$

Average

$$\boldsymbol{v}_i = \frac{1}{N\sqrt{2w_i}} \sum_{n=1}^{N} \gamma_n(i) [\boldsymbol{\Sigma}_i^{-1} \mathrm{diag}\left((\boldsymbol{x}_n - \boldsymbol{\mu}_i)(\boldsymbol{x}_n - \boldsymbol{\mu}_i)^T\right) - \mathbf{1}]$$

Variance

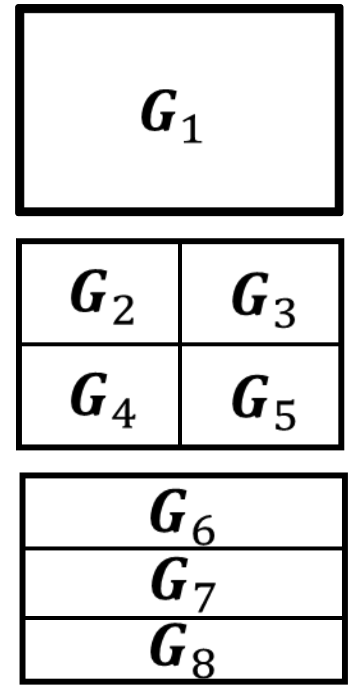
Image → Local Descriptor Extraction → Statistic Calculation → Norma-lization → Spatial Info → Image Feature Vector

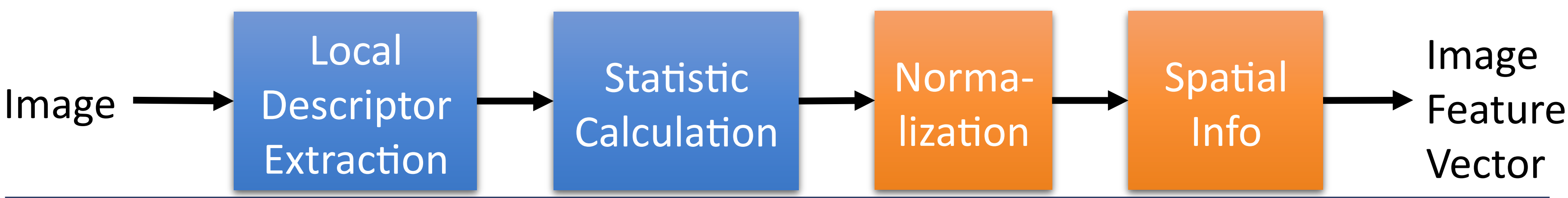
Image Feature – Fisher Vector [Perronnin et al., 2010]

- Normalization
  - FV representation: $\boldsymbol{G} = [\boldsymbol{u}_1^T, \boldsymbol{v}_1^T, \ldots, \boldsymbol{u}_K^T, \boldsymbol{v}_K^T]^T$
  - Power normalization: $\mathrm{sign}(\boldsymbol{G})|\boldsymbol{G}|^{1/2}$
- Spatial Information
  - Calculate FVs for divided 8 areas and concatenate them

$$\boldsymbol{G} = [\boldsymbol{G}_1^T, \boldsymbol{G}_2^T, \ldots, \boldsymbol{G}_8^T]^T$$

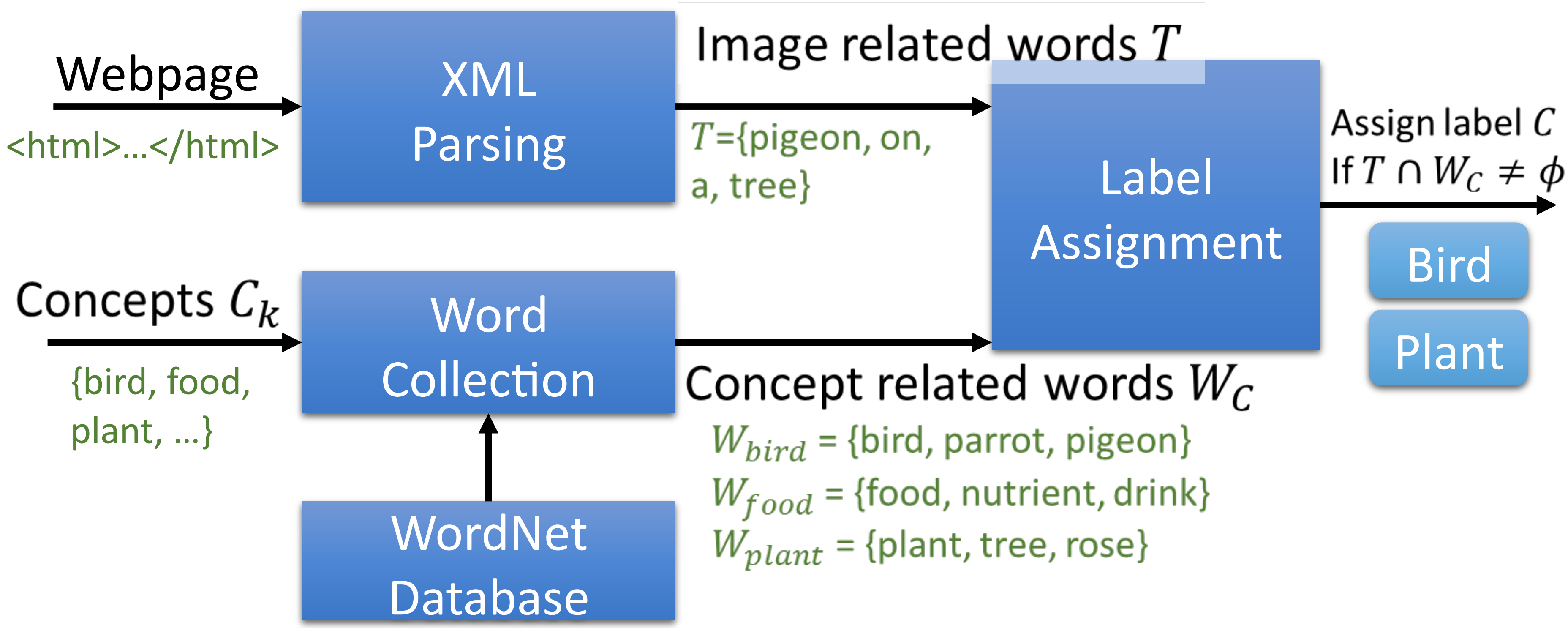
- The dimension of our FV is 262144



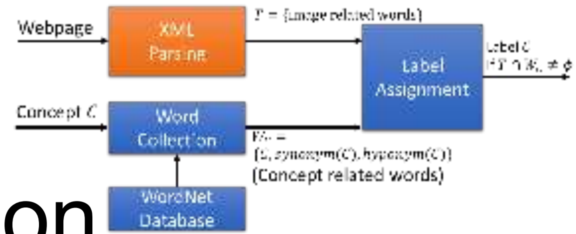


Textual Feature – Pipeline

- Supporting concepts of both development and test set is required
- Use WordNet [Fellbaum, 1998] as an external source
- Fast and significantly improves performance



Textual Feature – Text Extraction



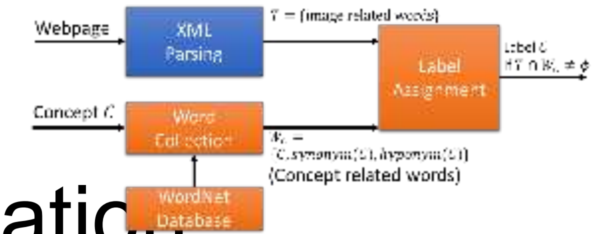
- Webpage is NOT concentrating on one image
  - Range of text corresponding to the image is limited
- Parse XML and extract elements
  - Page Title
  - Img tag attributes (filename, alternative text, title)
  - Text displayed near the image
- Select text closely related to the image
- Regard text as a set of words T
  - Not considered about grammar

<h1>Swim With Dolphins Bahamas</h1>
<img src=
"bahamas-dolphin-encounters.jpg"
alt="Swim With Dolphins Bahamas">
</div>
The popular Bahamas Dolphin Encounters specializes in creating opportunities for humans to interact safely with dolphins.

T = {swim, with, dolphin, bahama, encounter, …}

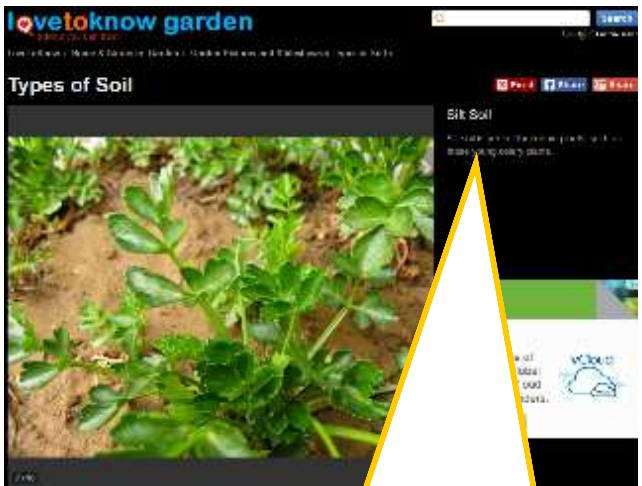
Image related words

Text far from image is less related to the image
=> Consider only certain words distant from image



Textual Feature – Label Estimation

- Simplest method (used in ImageCLEF 2012) [Ushiku et al., 2012]



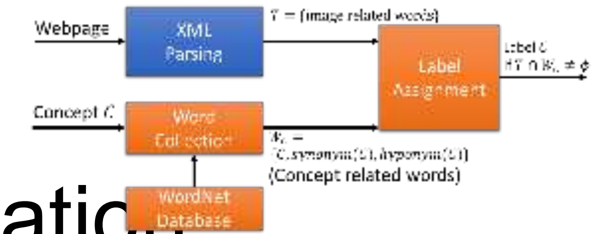
http://garden.lovetoknow.com/wiki/Slideshow:Types_of_Soil#7

Silt soil is perfect for certain plants, such as these young celery plants

Word match with label

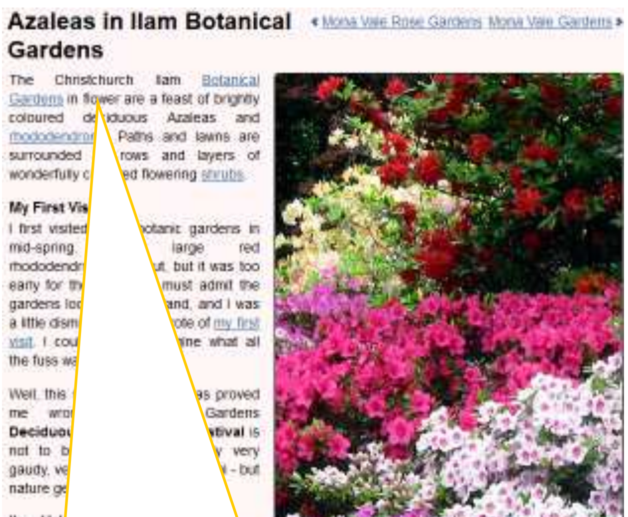
Image + Estimated Label

Label set	Car	Fish	Plant	Sea	…



Textual Feature – Label Estimation

- Problem: related word cannot be used



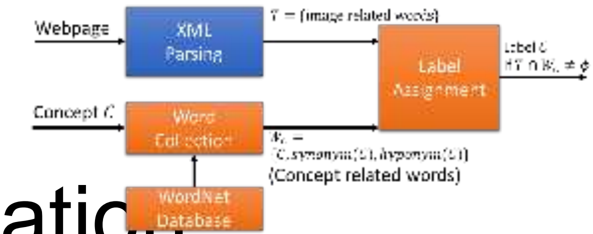
http://www.mooseyscountrygarden.com/botanical-gardens/ilam-azaleas.html

Azaleas in Ilam Botanical Gardens The Christchurch…

Azalea is a plant

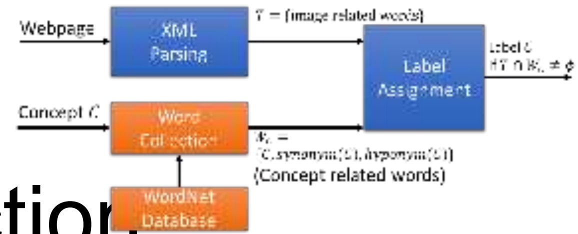
Image + Estimated Label

Label set	Car	Fish	Plant	Sea	…



Textual Feature – Label Estimation

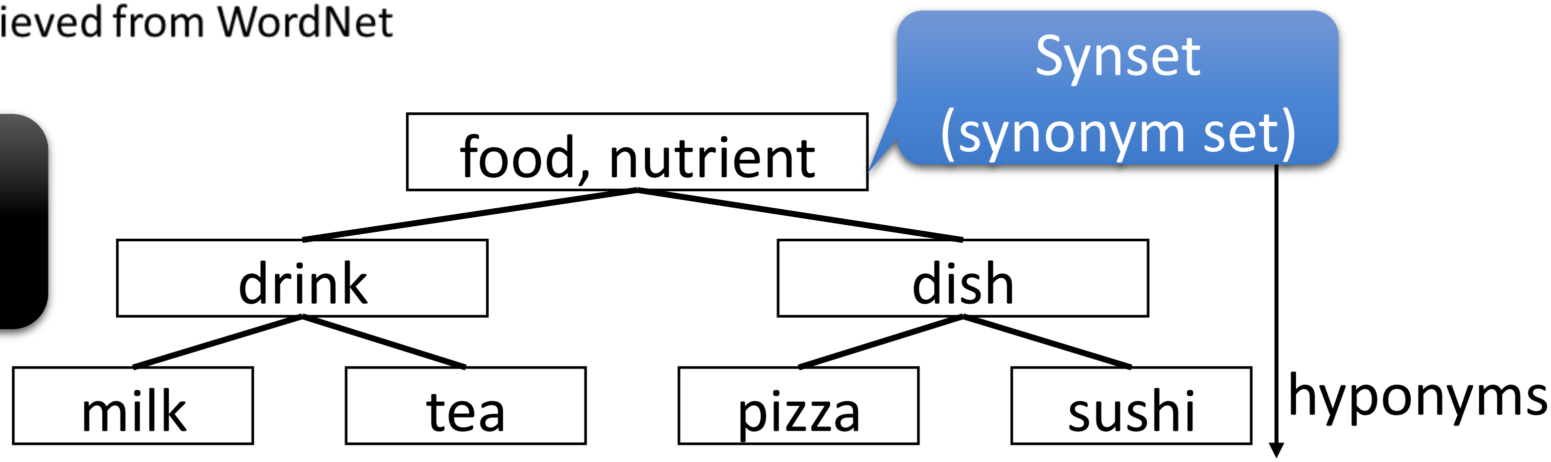
- Using related words are important
- [Jin et al., 2005] used semantic distance from WordNet to remove irrelevant keywords from annotation
- [Villegas et al., 2012] used words from definition of concept in English dictionary and constructed probabilistic model
- We try to collect more concept related words simply



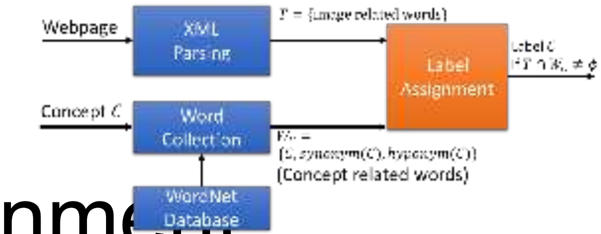
Textual Feature – Word Collection

- Collect words W_C related to each concept C
- Use synonyms and hyponyms of the concept word
  - Quite simpler than other methods (e.g. Google Distance)
  - Retrieved from WordNet

WordNet Hierarchy (Simplified)



$C = \text{food}$ → $W_{food} = \{\text{food, nutrient, drink, milk, ...}\}$



Textual Feature – Label Assignment

- A label is assigned to the image if image related words contains any of concept related words

From webpage

T = {pigeon, on, a, tree} (image related words)

W_{bird} = {bird, parrot, $\boxed{\text{pigeon}}$}

W_{food} = {food, nutrient, drink}

W_{plant} = {plant, $\boxed{\text{tree}}$, rose}

From WordNet

Bird

Plant

Estimated labels

Online Multilabel Annotation Learning

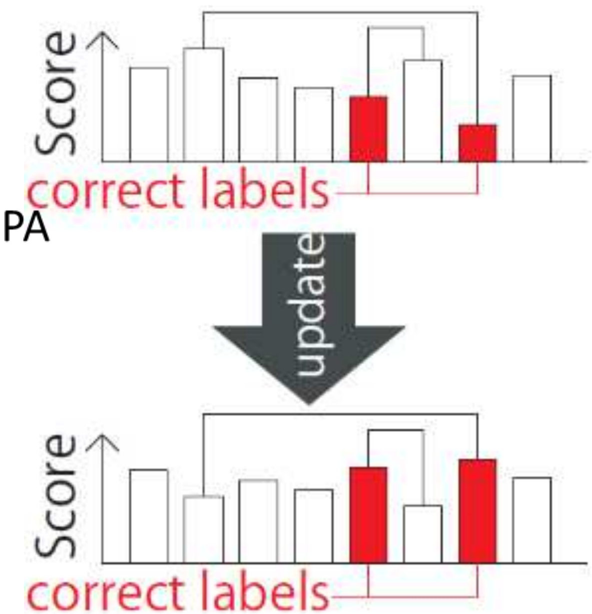
- To make system scalable, linear model based approach is adopted
  - K-NN based approach: complexity of recognizing is $O(N)$ (N is dataset size)
  - Kernel based approach: complexity of learning is $O(N^2)$
- PAAPL: Passive Aggressive with Averaged Pairwise Loss [Ushiku et al., 2012]
- Passive Aggressive [Crammer et al., 2006] based method
  - Online; requires less RAM
  - Robust to noise of label data
- Converges faster than original PA in multilabel learning

PAAPL – Learning Flow (summarized)

- Update models $\boldsymbol{\mu}^C$ sequentially for each training sample by following
  - Fetch training sample; image feature $\boldsymbol{f}$, assigned labels Y, not assigned labels $\bar{Y}$
  - Find a label r in Y, a label s in $\bar{Y}$ by follows

Mistakenly low scored label

$$r = \underset{r \in Y}{\operatorname{argmin}} \, \boldsymbol{\mu}^r \cdot \boldsymbol{f}$$
$$s = \underset{s \in \bar{Y}}{\operatorname{argmax}} \, \underbrace{\boldsymbol{\mu}^s \cdot \boldsymbol{f}}_{\text{Score}}$$



  - Calculate hinge-loss l and update models according to PA

$$l = \max(1 - (\boldsymbol{\mu}^r \cdot \boldsymbol{f} - \boldsymbol{\mu}^s \cdot \boldsymbol{f}), 0)$$
$$\boldsymbol{\mu}^r_{new} = \boldsymbol{\mu}^r + l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$
$$\boldsymbol{\mu}^s_{new} = \boldsymbol{\mu}^s - l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$

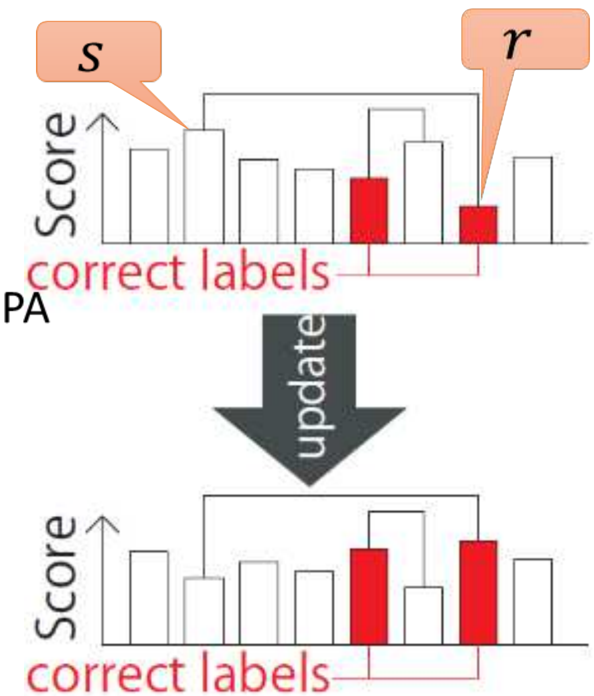
  - **Repeat above for previously not selected labels**
    - This procedure is not in original PA

PAAPL – Learning Flow (summarized)

- Update models $\boldsymbol{\mu}^C$ sequentially for each training sample by following
  - Fetch training sample; image feature $\boldsymbol{f}$, assigned labels Y, not assigned labels $\bar{Y}$
  - Find a label r in Y, a label s in $\bar{Y}$ by follows

Mistakenly low scored label

$$r = \underset{r \in Y}{\operatorname{argmin}} \, \boldsymbol{\mu}^r \cdot \boldsymbol{f}$$
$$s = \underset{s \in \bar{Y}}{\operatorname{argmax}} \, \underbrace{\boldsymbol{\mu}^s \cdot \boldsymbol{f}}_{\text{Score}}$$



  - Calculate hinge-loss l and update models according to PA

$$l = \max(1 - (\boldsymbol{\mu}^r \cdot \boldsymbol{f} - \boldsymbol{\mu}^s \cdot \boldsymbol{f}), 0)$$
$$\boldsymbol{\mu}^r_{new} = \boldsymbol{\mu}^r + l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$
$$\boldsymbol{\mu}^s_{new} = \boldsymbol{\mu}^s - l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$

  - **Repeat above for previously not selected labels**
    - This procedure is not in original PA

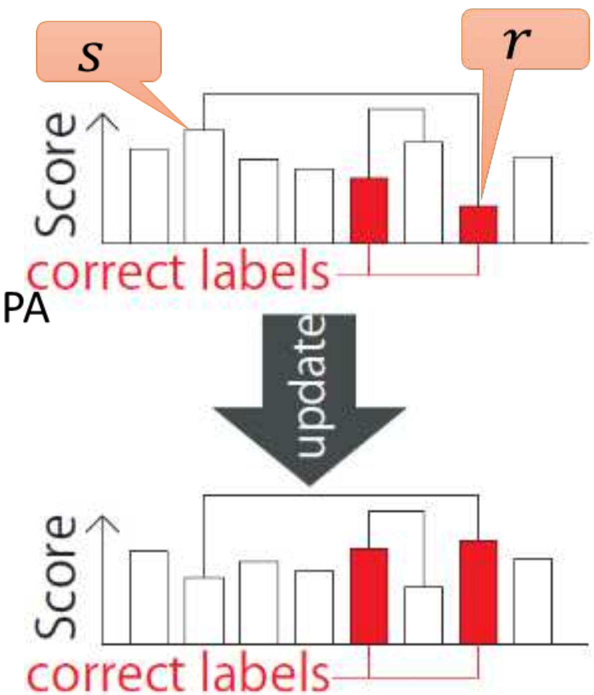
PAAPL – Learning Flow (summarized)

- Update models $\boldsymbol{\mu}^C$ sequentially for each training sample by following
  - Fetch training sample; image feature $\boldsymbol{f}$, assigned labels Y, not assigned labels $\bar{Y}$
  - Find a label r in Y, a label s in $\bar{Y}$ by follows

Mistakenly low scored label

$$r = \underset{r \in Y}{\operatorname{argmin}}\, \boldsymbol{\mu}^r \cdot \boldsymbol{f}$$

$$s = \underset{s \in \bar{Y}}{\operatorname{argmax}}\, \underbrace{\boldsymbol{\mu}^s \cdot \boldsymbol{f}}_{\text{Score}}$$



  - Calculate hinge-loss l and update models according to PA

$$l = \max(1 - (\boldsymbol{\mu}^r \cdot \boldsymbol{f} - \boldsymbol{\mu}^s \cdot \boldsymbol{f}), 0)$$

$$\boldsymbol{\mu}^r_{new} = \boldsymbol{\mu}^r + l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$
$$\boldsymbol{\mu}^s_{new} = \boldsymbol{\mu}^s - l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$

  - **Repeat above for previously not selected labels**
    - This procedure is not in original PA

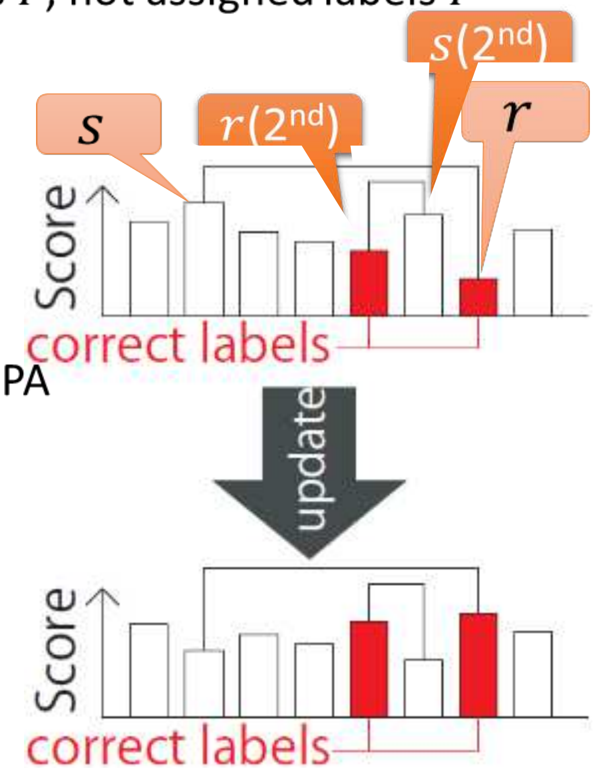
PAAPL – Learning Flow (summarized)

- Update models $\boldsymbol{\mu}^C$ sequentially for each training sample by following
  - Fetch training sample; image feature $\boldsymbol{f}$, assigned labels Y, not assigned labels $\bar{Y}$
  - Find a label r in Y, a label s in $\bar{Y}$ by follows

Mistakenly low scored label

$$r = \underset{r \in Y}{\operatorname{argmin}}\, \boldsymbol{\mu}^r \cdot \boldsymbol{f}$$

$$s = \underset{s \in \bar{Y}}{\operatorname{argmax}}\, \underbrace{\boldsymbol{\mu}^s \cdot \boldsymbol{f}}_{\text{Score}}$$



  - Calculate hinge-loss l and update models according to PA

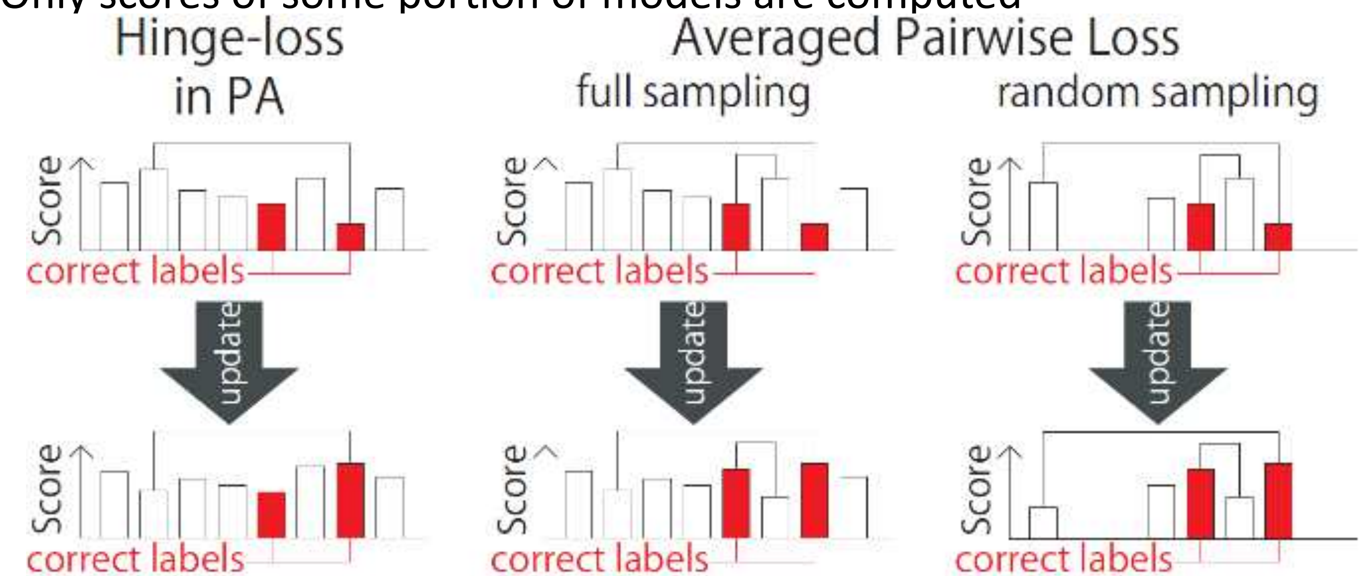
$$l = \max(1 - (\boldsymbol{\mu}^r \cdot \boldsymbol{f} - \boldsymbol{\mu}^s \cdot \boldsymbol{f}), 0)$$

$$\boldsymbol{\mu}^r_{new} = \boldsymbol{\mu}^r + l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$
$$\boldsymbol{\mu}^s_{new} = \boldsymbol{\mu}^s - l/(2\,|\boldsymbol{f}|^2 + 1/D) \cdot \boldsymbol{f}$$

- **Repeat above for previously not selected labels**
  - This procedure is not in original PA

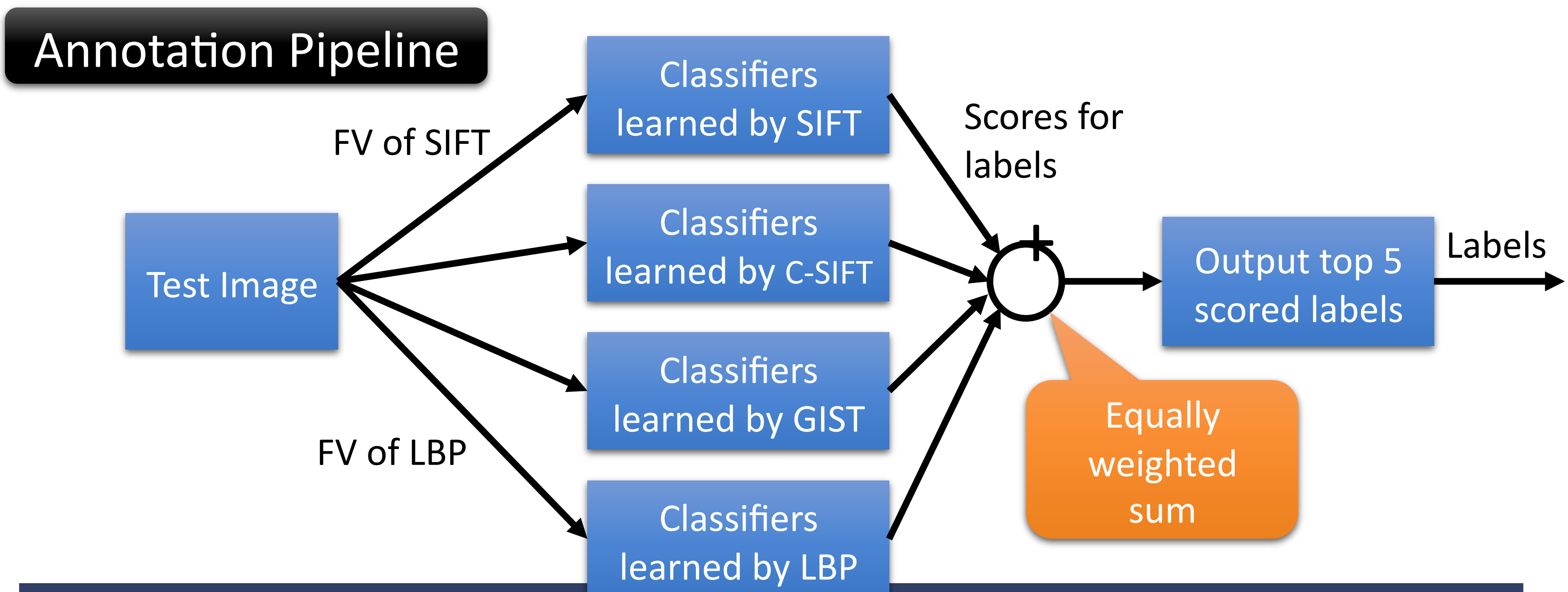
PAAPL – Advantages

- Score computation process is heavy part of PA
  - PAAPL updates all pairs of models by one score computation
  - It makes convergence faster
- To make faster, random sampling is adopted
  - Only scores of some portion of models are computed



Multiple Feature Score Combination

- Scores of models which were learned by different image features are summed in annotation step
  - Which combination is best is evaluated by experiment



Experiment Condition

- We applied these methods to ImageCLEF 2013 dataset
- Experiment order
  1. Label estimation condition (whether to use synonyms and hyponyms)
  2. Text extraction condition (whether to use page title etc.)
  3. Comparison of image local descriptors and their score combination
- Image feature for first two experiment is provided C-SIFT + BoVW
- Evaluation was done by F-measure for development set
- Submitted runs are computed with best parameters for development set

Experiment Results – Label Estimation

- Whether we should use synonyms and hyponyms

$C = \text{food}$ → $W_{food} = \{\text{food}, \underbrace{\text{nutrient},}_{\text{Synonym}} \underbrace{\text{drink, milk, ...}}_{\text{Hyponym}}\}$

Webpage's text contains W_{food} => label "food" assigned

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

+ 3pts

Using both synonyms and hyponyms is the best

Experiment Results – Text Extraction

- What elements of webpages we should use (best 3 & baseline shown)

Text around image [max word distance]	Img tag attributes	Page title	MF-samples [%]	Number of images with label
-	✔		**27.6**	80009 **[lowest]**
10	✔		26.6	129050
10	✔	✔	26.1	140448
1000	✔	✔	20.7	193971

Baseline

People use filename to manage photos

+ 7pts

<h1>Swim With Dolphins Bahamas</h1>
<img src=
"bahamas-dolphin-encounters.jpg"
alt="Swim With Dolphins Bahamas">
</div>
The popular Bahamas Dolphin Encounters specializes in creating opportunities for humans to interact safely with dolphins.

10 words after image

Text around image (max distance 10 words)

Img tag attributes

Experiment Results – Image Local Descriptor

- Best 5 combinations and 4 single features (Fisher Vector applied)

C-SIFT	GIST	LBP	SIFT	MF-samples [%]	Test set MF-samples
✔	✔		✔	**34.6** [ISI-1]	**33.2**
✔	✔	✔	✔	34.3 [ISI-2]	32.7
✔	✔	✔		34.2 [ISI-3]	31.8
	✔		✔	34.0 [ISI-4]	32.4
	✔	✔	✔	33.9 [ISI-5]	31.7
✔				31.2	
	✔			**32.4**	
		✔		27.9	
			✔	31.1	
Provided C-SIFT + BoVW				27.6	

+ 7pts

Submitted runs

GIST is the best among single descriptor

Conclusion

- Visual Feature
  - Fisher Vector with four local descriptors was used and the combination of C-SIFT, GIST and SIFT showed superior performance than provided C-SIFT + BoVW
- Textual Feature
  - Using synonyms and hyponyms for label estimation improved performance
  - Selecting text related to image also highly improved performance
    - Img tag attributes were the most important
  - Worked well in concepts of both development set and test set
- Learning
  - The method which is scalable to the size of dataset was adopted

Experiment Results – Text Extraction (All)

Text around image [max word distance]	Img tag attributes	Page title	MF-samples [%]	Number of images with label	Average number of labels
10			25.4	113802	0.7
100			23.1	183545	2.5
1000			20.2	192210	5.2
-	✔		**27.6**	80009	0.4
10	✔		26.6	129050	0.8
100	✔		23.8	185471	2.5
1000	✔		21.3	193170	5.3
-		✔	24.6	92254	0.5
10		✔	25.5	134318	0.9
100		✔	22.9	185471	2.5
1000		✔	20.5	193497	5.3
-	✔	✔	26.0	111247	0.6
10	✔	✔	26.1	140448	0.9
100	✔	✔	23.0	186394	2.6
1000	✔	✔	20.7	193971	5.3

Textual Feature – Implementation Detail

- Text Extraction
  - Words are singularized by ActiveSupport library
- Word Collection
  - Used synset of synset id specified in the concept list
  - Ambiguous words (words of multiple meaning) are not used as related words
    - The word which appears in multiple synset in WordNet is judged to be ambiguous
  - Hyponyms are gathered from all depths from the synset of concept