Bluefield (KDE TUT) at LifeCLEF 2016 Plant Identification Task

Siang Thye Hang
Atsushi Tatsuma
Masaki Aono

TOYOHASHI
UNIVERSITY OF TECHNOLOGY

- Introduction and related work
- Modifying the VGG-16 layers CNN model
  - Spatial Pyramid Pooling (SPP)
  - Parametric Rectified Linear Unit (PReLU)
- Post processing
  - Unseen Category Sample Detection algorithm
  - Observation based identification
- Evaluation
  - Data augmentation and model training
  - Results and analysis
- Conclusion

Introduction (1/2)

- Accurate knowledge of plants is essential in agricultural development and biodiversity conservation
- However, identifying a species can be difficult even for professionals

query image

Daphne cneorum?

Rhododendron ferrugineum?

Oxalis articulata?

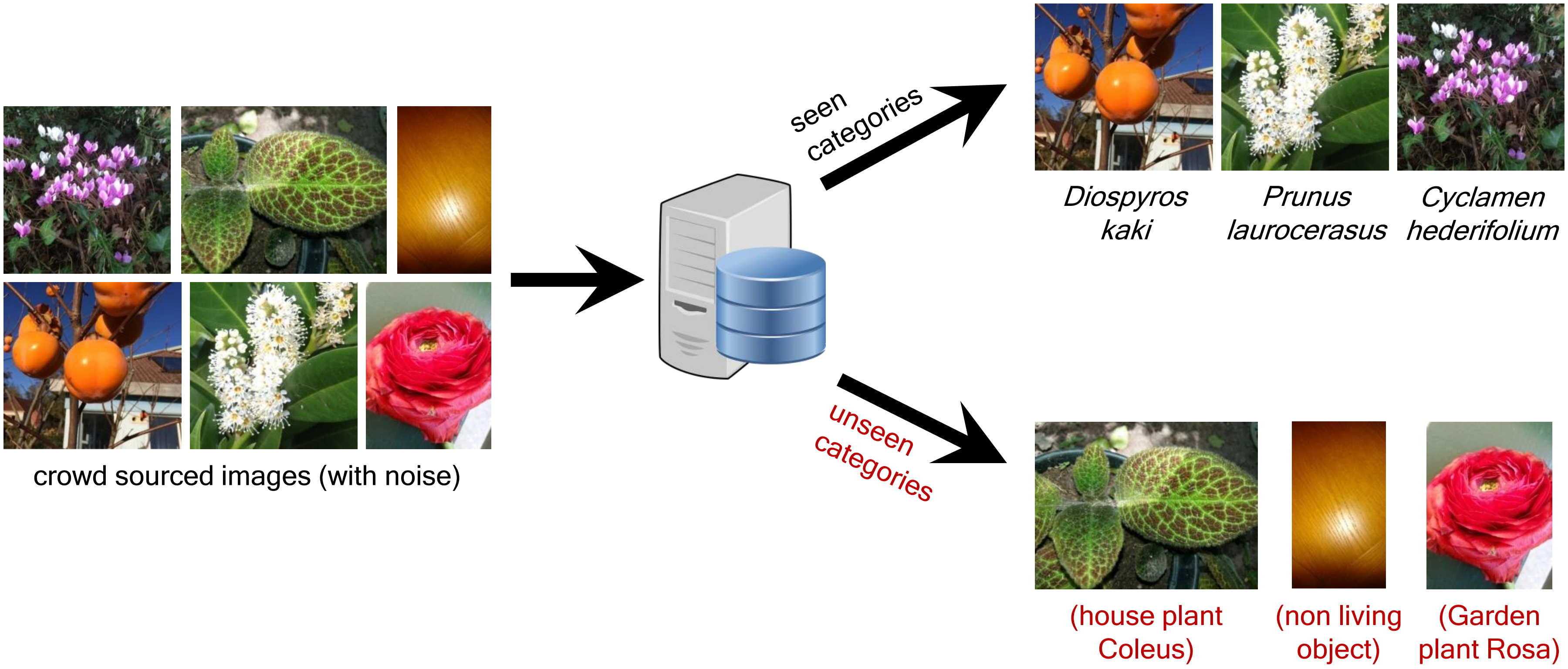
Silene acaulis?

. . .

- This leads to the consideration of using image retrieval technologies
  - Convolutional Neural Network (CNN) is widely used in various image retrieval tasks

Introduction (2/2)

- Crowd sourced data is a cost efficient method to provide data for image retrieval systems
- However, they usually contain higher level of noise and unknown objects
- In real world, instead of classification to a fixed number of categories, a good system is expected to be robust to unseen categories



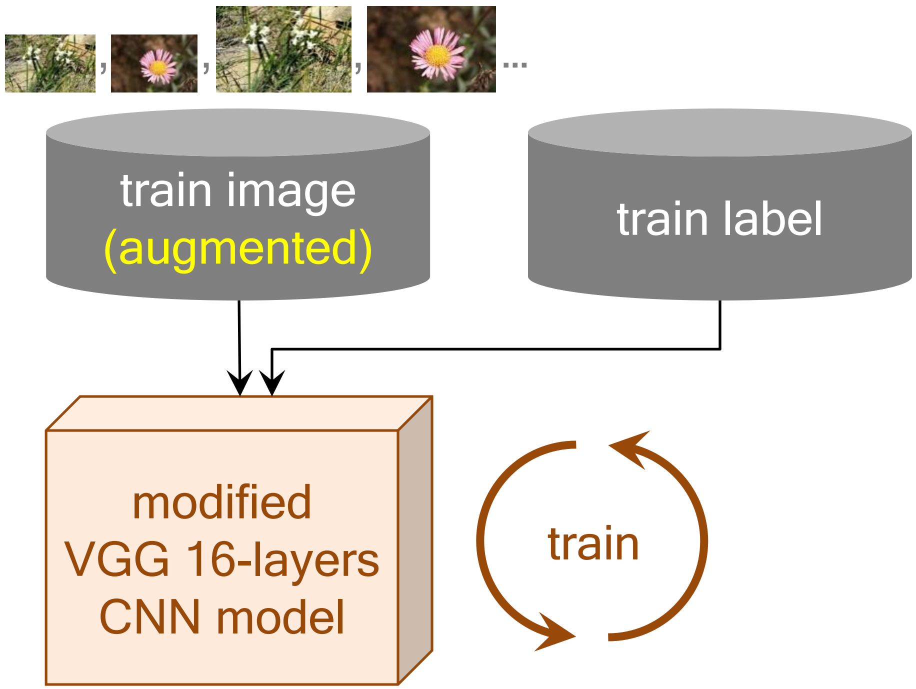
Related work

- PlantCLEF 2014's top performer: IBM Research Australia (IBM AU)
  - trained AlexNet CNN model
  - Fisher Vector encoded dense SIFT and Color Moments
  - Region of Interest based cropping and background removal

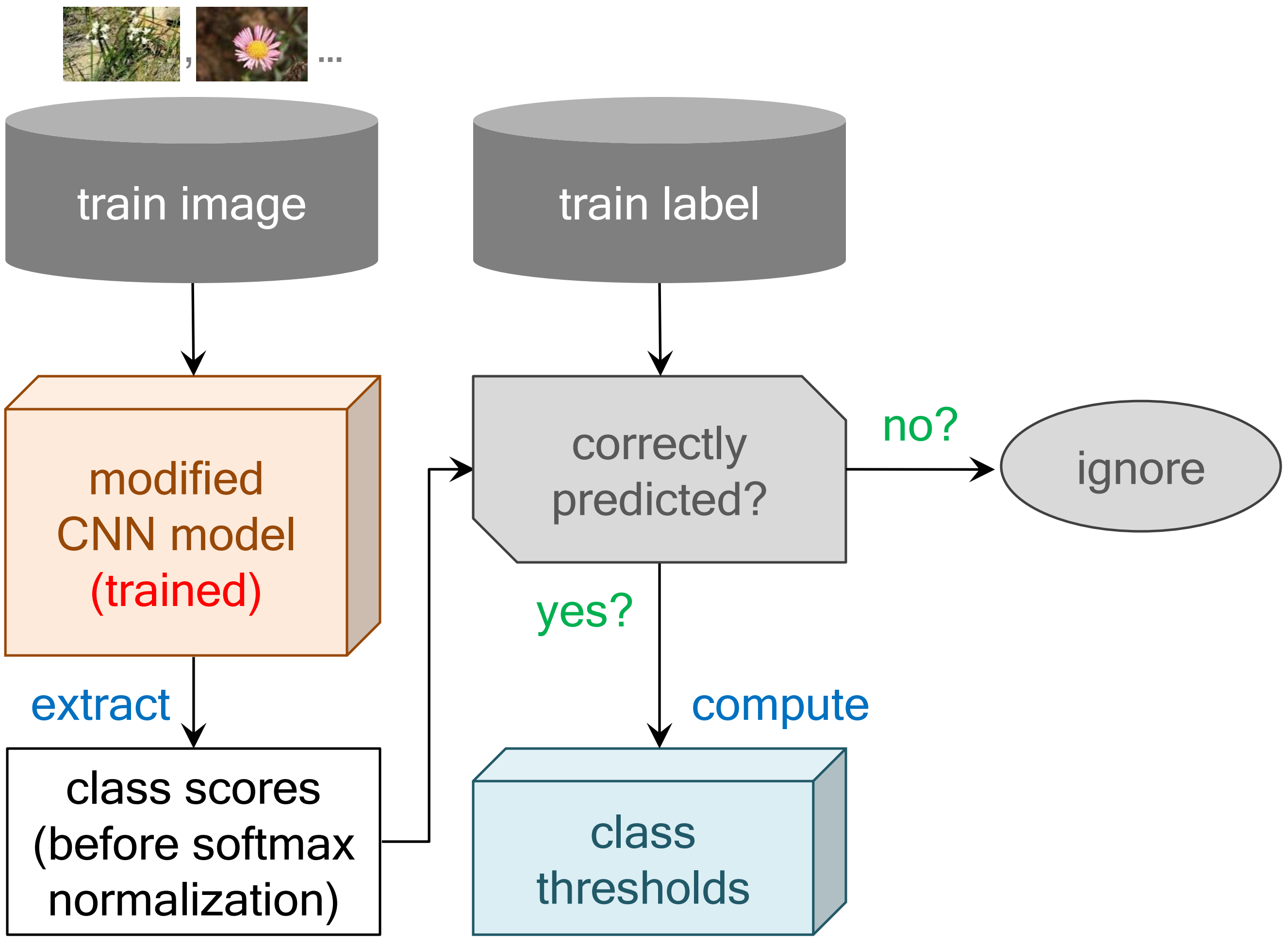
- PlantCLEF 2015's top performer: Seoul National University Medinfo (SNUMED INFO)
  - Transfer learning on 5 GoogLeNet CNN Models, outputs combined

- Our objective is to design a system that is robust to unseen categories
  - Automatic, minimizing the need of human intervention
  - Without relying on transfer learning
  - Without external dataset

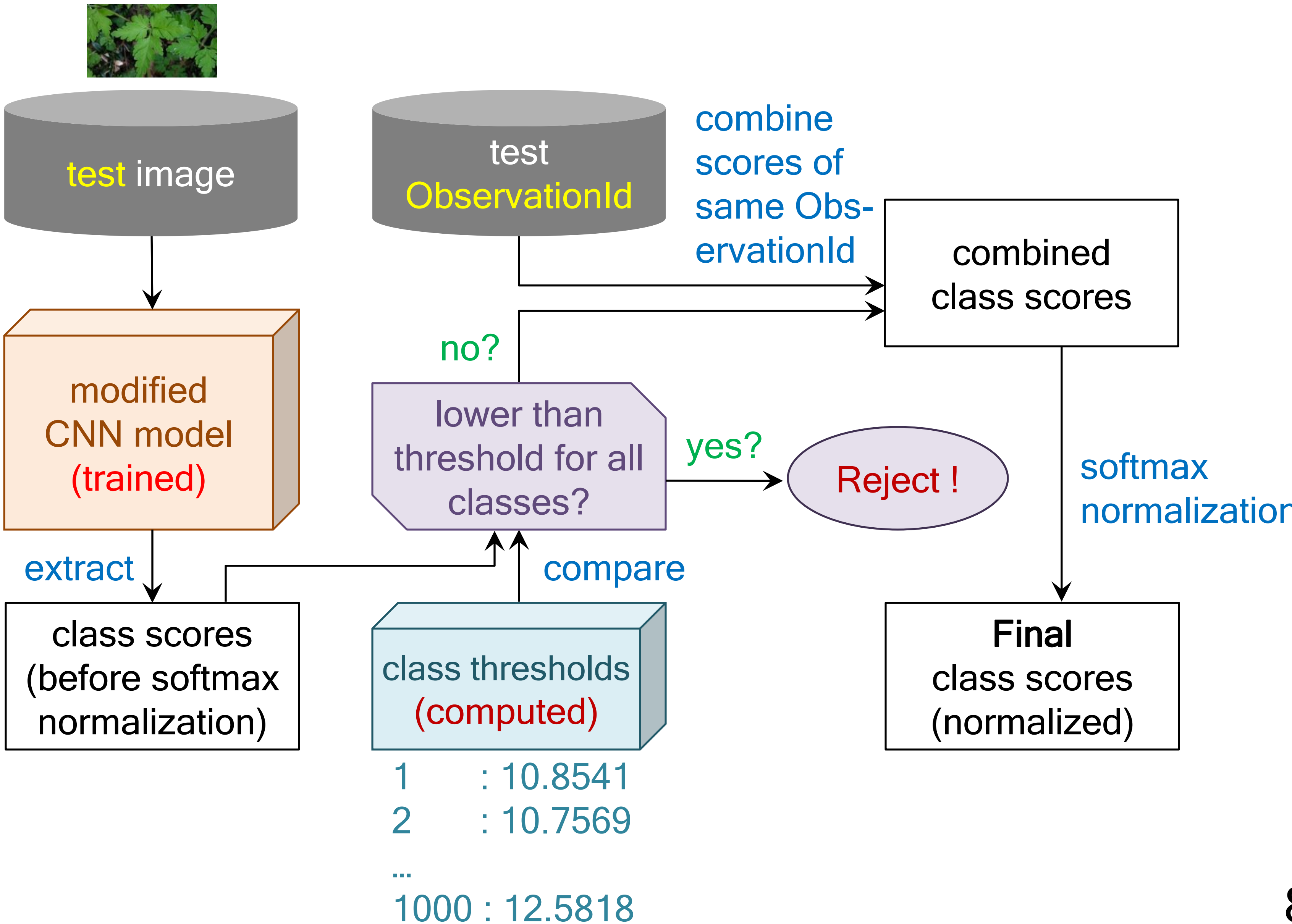
Proposed framework (train)



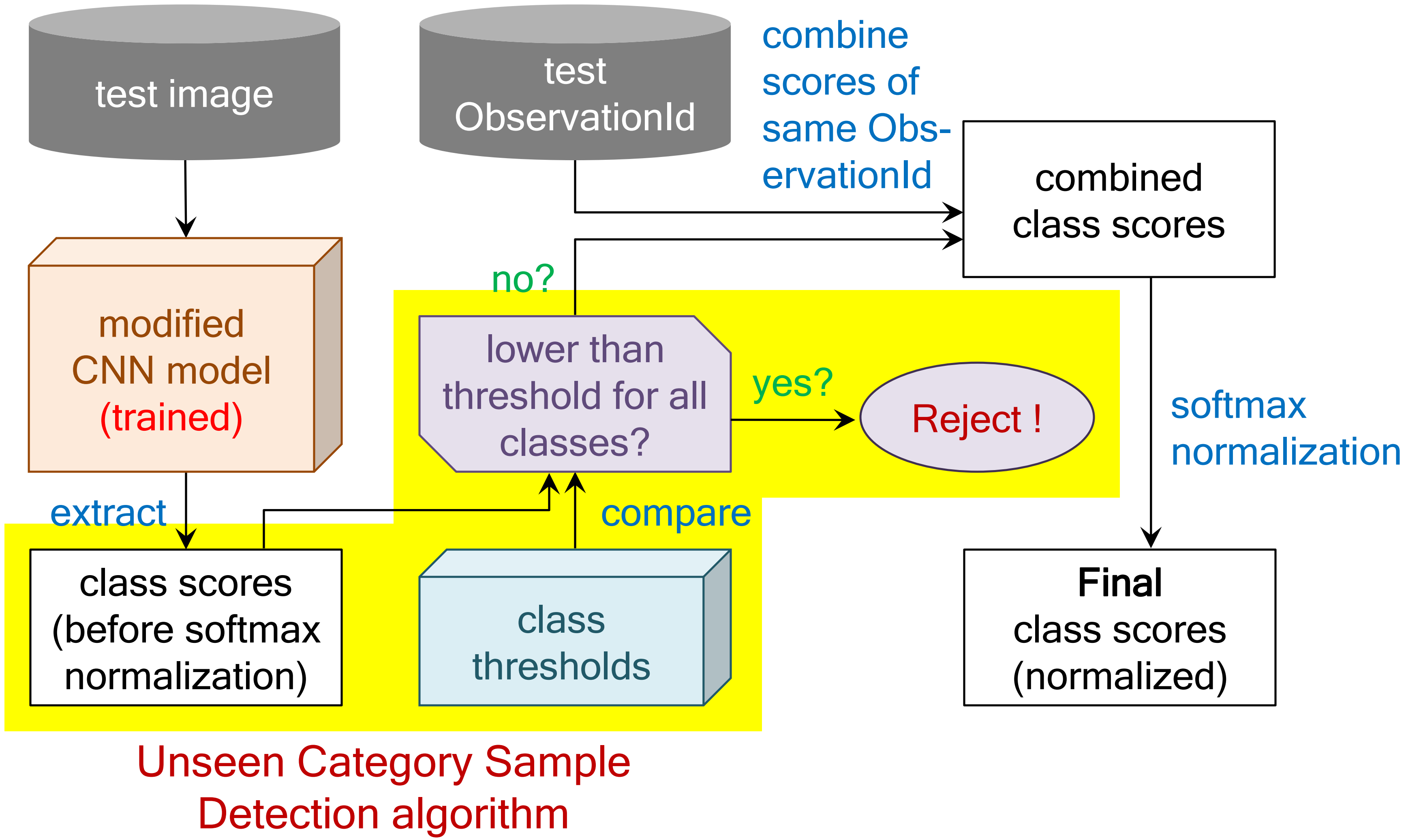
Proposed framework (train)



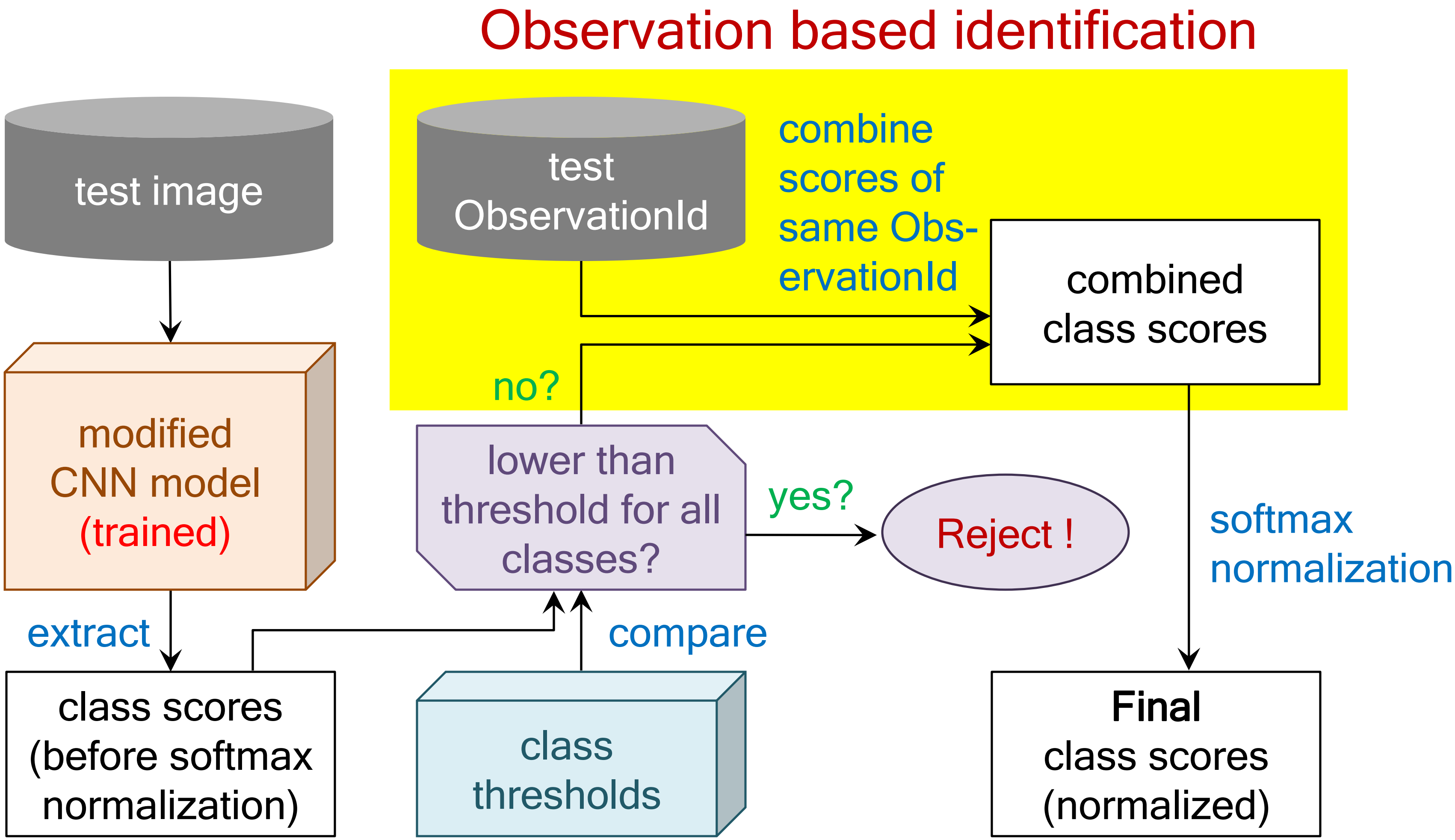
Proposed framework (test)



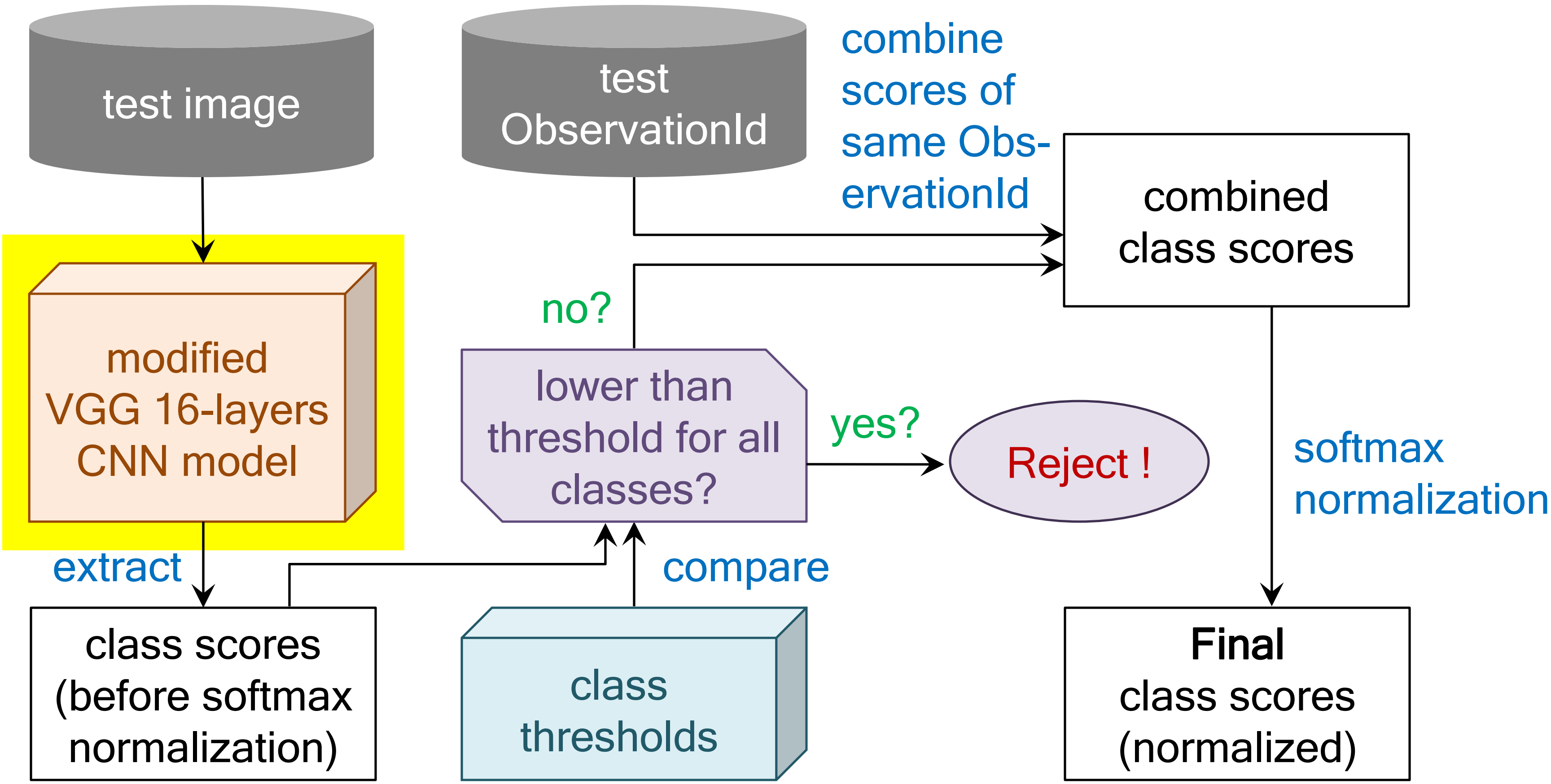
Proposed framework



Proposed framework



Proposed framework



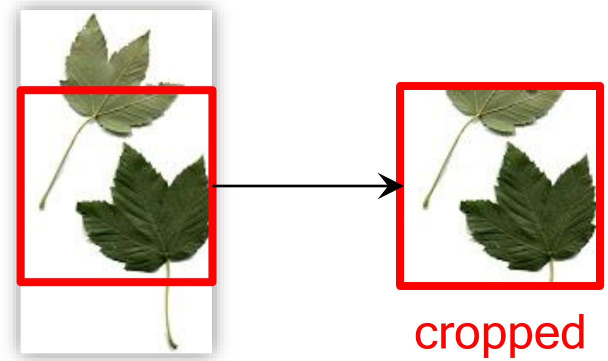
The VGG-16 layers CNN model*

input (224×224 RGB image)
conv3-64, conv3-64
maxpool
conv3-128, conv3-128
maxpool
conv3-256, conv3-256, conv3-256
maxpool
conv3-512, conv3-512, conv3-512
maxpool
conv3-512, conv3-512, conv3-512
maxpool
FC-4096, FC-4096
FC1000, softmax

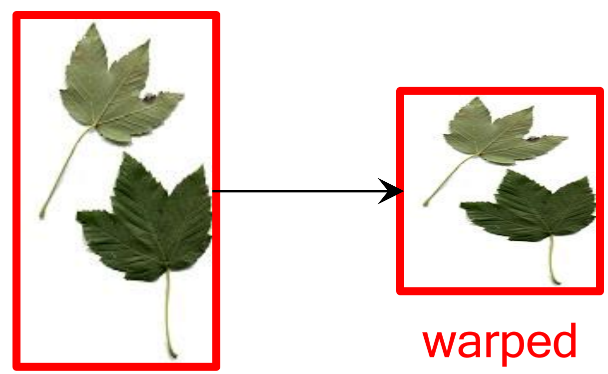
* K. Simonyan and A. Zisserman, "**Very Deep Convolutional Networks for Large Scale Image Recognition**," CVPR 2014

The VGG-16 layers CNN model

input (224×224 RGB image)
conv3-64, conv3-64
maxpool
conv3-128, conv3-128
maxpool
conv3-256, conv3-256, conv3-256
maxpool
conv3-512, conv3-512, conv3-512
maxpool
conv3-512, conv3-512, conv3-512
maxpool
FC-4096, FC-4096
FC1000, softmax



cropped



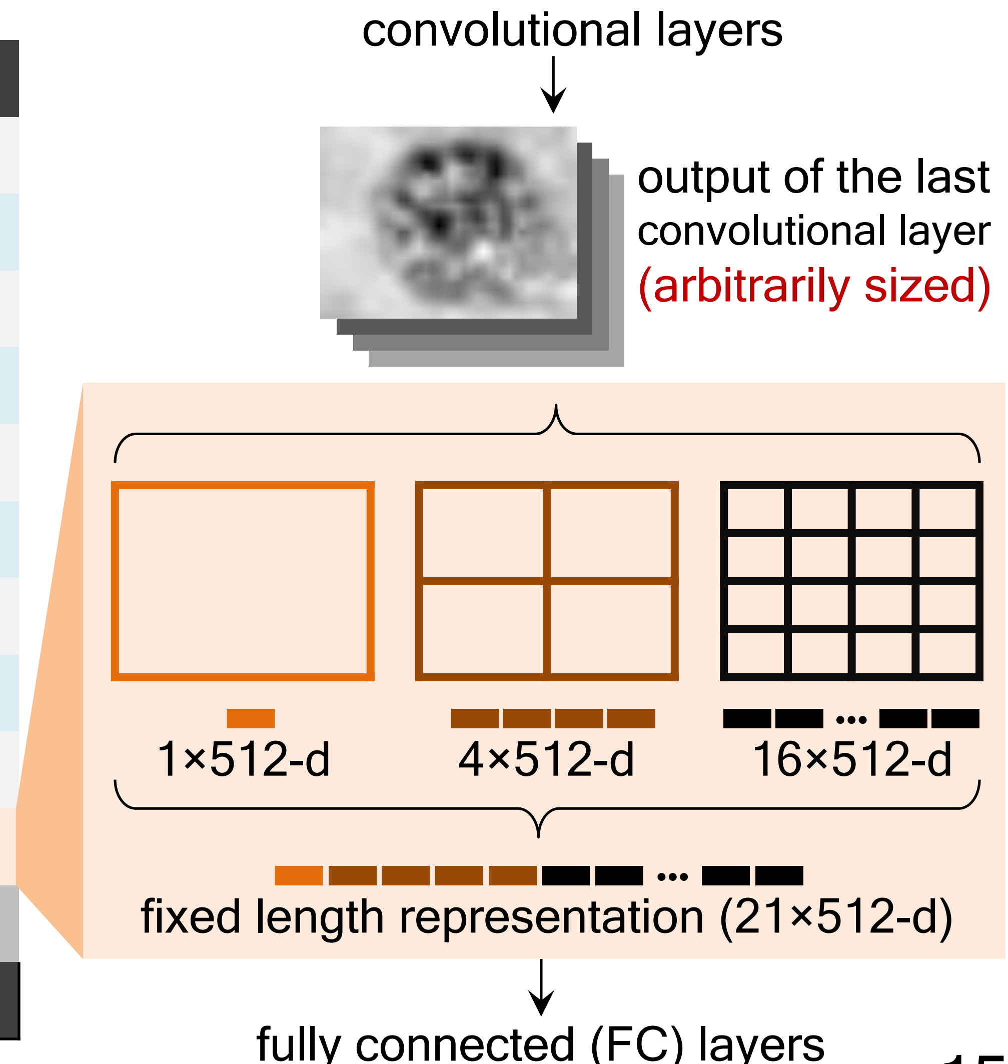
warped

Modifying VGG 16-layers model

input (224×224 RGB image)
conv3-64, conv3-64
maxpool
conv3-128, conv3-128
maxpool
conv3-256, conv3-256, conv3-256
maxpool
conv3-512, conv3-512, conv3-512
maxpool
conv3-512, conv3-512, conv3-512
~~maxpool~~
FC-4096, FC-4096
FC1000, softmax

Modifying VGG 16-layers model

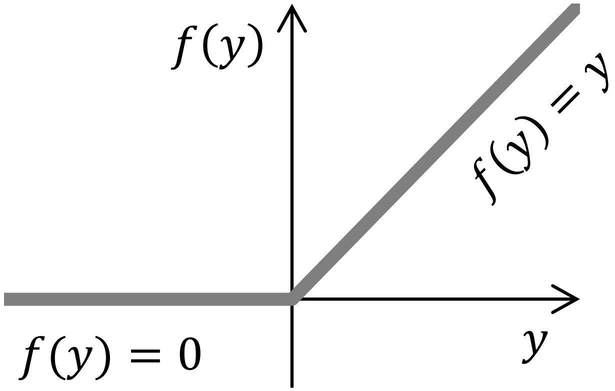
input (arbitrarily sized RGB image)
conv3-64, conv3-64
maxpool
conv3-128, conv3-128
maxpool
conv3-256, conv3-256, conv3-256
maxpool
conv3-512, conv3-512, conv3-512
maxpool
conv3-512, conv3-512, conv3-512
Spatial Pyramid Pooling*
FC-4096, FC-4096
FC1000, softmax



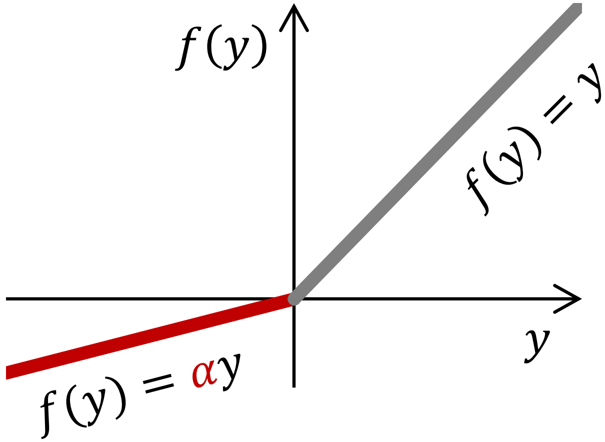
* K. He, X. Zhang, S. Ren and J. Sun, "**Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition**," TPAMI2015

Modifying VGG 16-layers model

Rectified Linear Unit (ReLU)



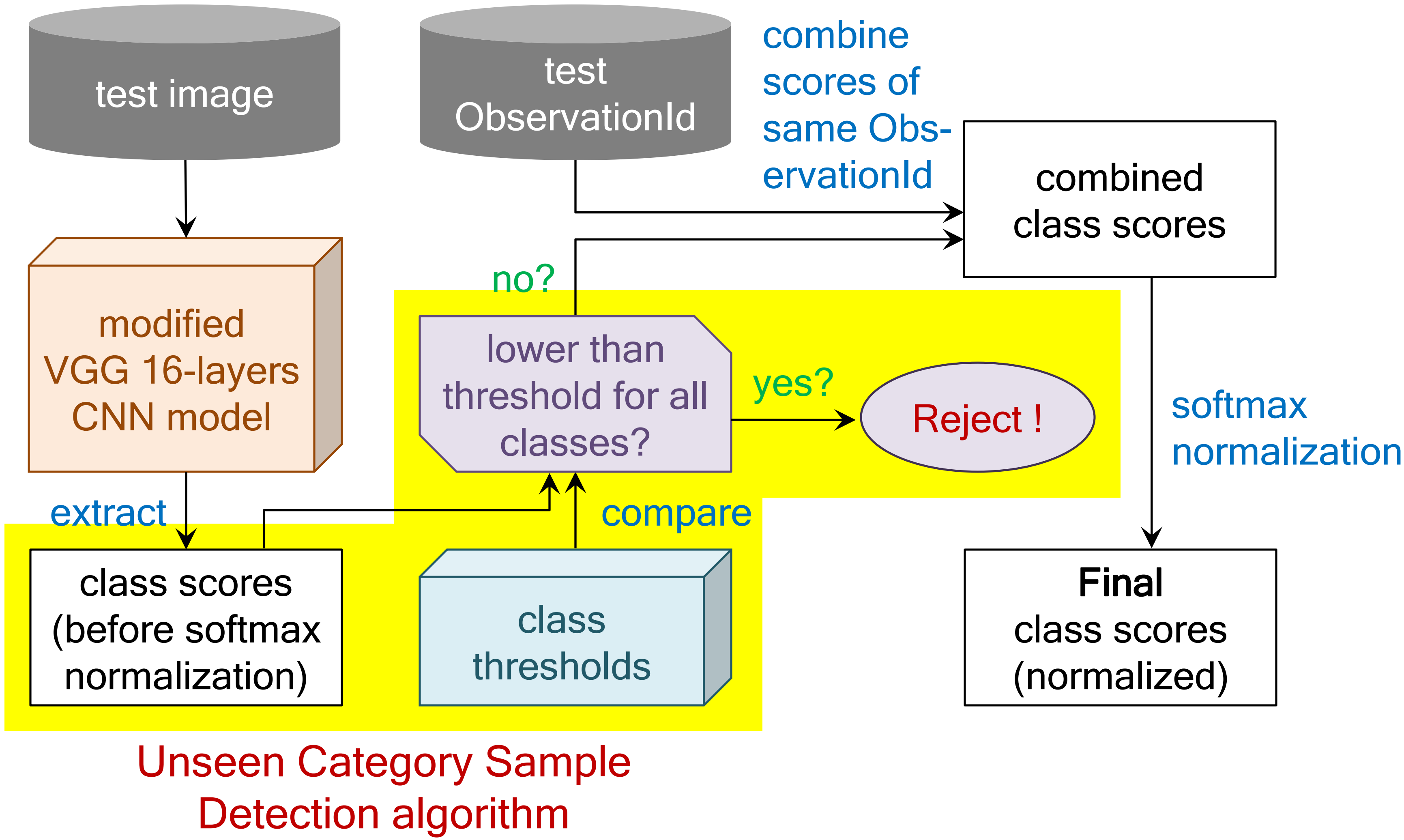
Parametric ReLU*



coefficient α for the negative part is not constant and is adaptively learned

* K. He, X. Zhang et. al., "**Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification**," CVPR2015

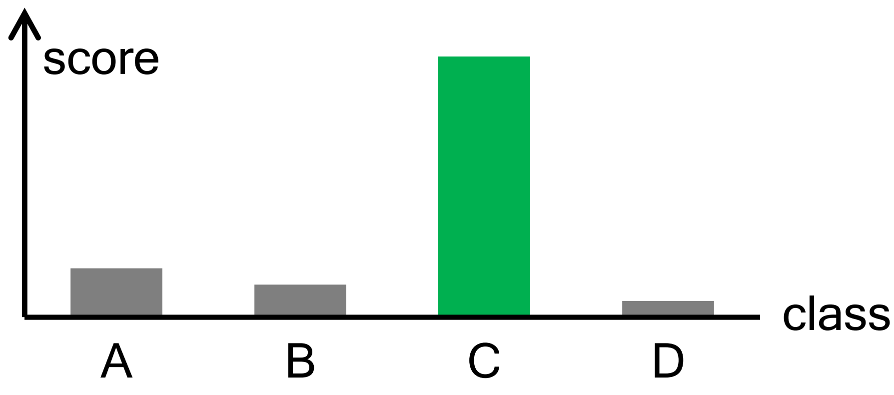
Proposed framework



Unseen Category Sample Detection algorithm

- The output of the last fully connected layer (FC-1000) can be assumed as output score of 1000 binary classifiers

- A sample will be classified to the class with highest score (One-versus-Rest)

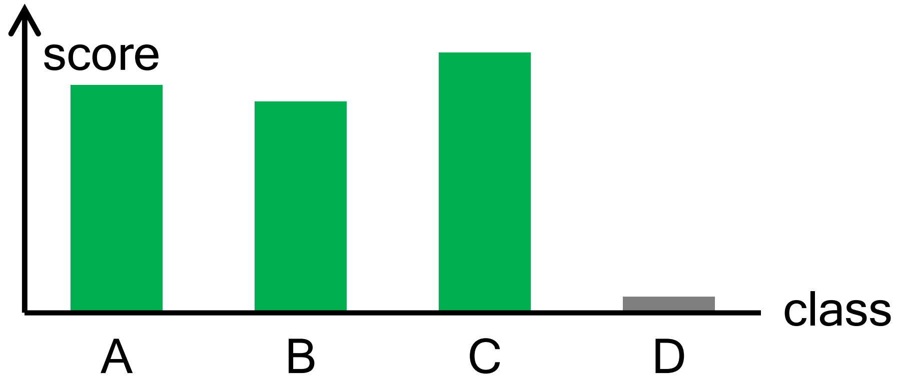


we may safely infer that
this sample belongs to class C

Unseen Category Sample Detection algorithm

- The output of the last fully connected layer (FC-1000) can be assumed as output score of 1000 binary classifiers

- A sample will be classified to the class with highest score (One-versus-Rest)

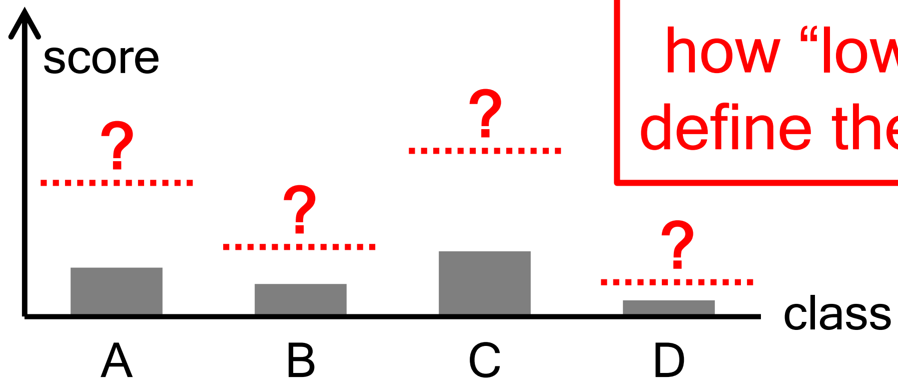


we may infer that this sample belongs to either class A or B or C

Unseen Category Sample Detection algorithm

- The output of the last fully connected layer (FC-1000) can be assumed as output score of 1000 binary classifiers

- A sample will be classified to the class with highest score (One-versus-Rest)



how “low” should we define the thresholds?

we may infer that this sample does not belong to any class (of unseen category)

Unseen Category Sample Detection algorithm

We propose to define the thresholds (for each class) as
"minimum score of correctly predicted training samples"

1. Assume scores $\boldsymbol{S}_{M \times N}$ of N classes and M training samples are extracted from the last fully connected layer (FC-1000)
   - Incorrectly predicted samples are omitted, leaving $M'(\leq M)$ samples
   - Only correctly predicted scores $\boldsymbol{S}_{M' \times N}$ are used

2. Thresholds $t_1, \ldots, t_N$ are computed with class-wise minima

$$t_i = \min_{k \in M'} S_{k,i}$$

3. During testing, any sample with scores lower than the computed thresholds for all of the classes will be rejected

Unseen Category Sample Detection algorithm

- When training a CNN model, a subset of samples are separated (from the training samples) for validation purpose

- So far, thresholds $t_1, \dots, t_N$ are computed based on training samples only
  - How about validation samples?

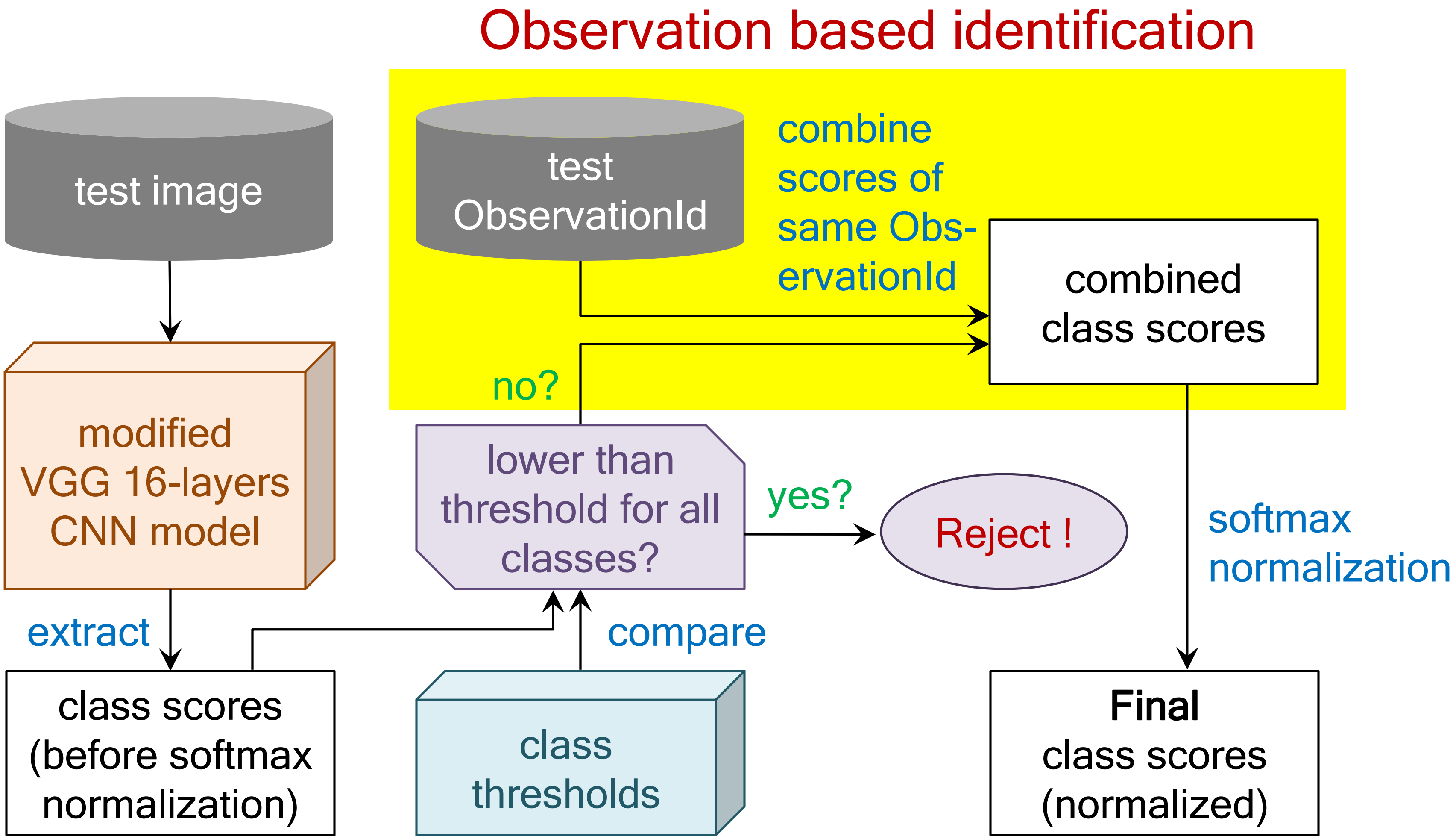
- Similarly, thresholds based on validation samples $v_1, \dots, v_N$ are computed

- Let Q be average of v_i/t_i for $i = 1, \dots, N$, we update the thresholds by multiplying $t_1, \dots, t_N$ with Q

$$t'_i = Qt_i$$

- As number of validation samples is usually less than training samples, we should expect that majority of v_i will be higher than t_i
  - Only v_i that are lower than t_i are considered into computation of Q

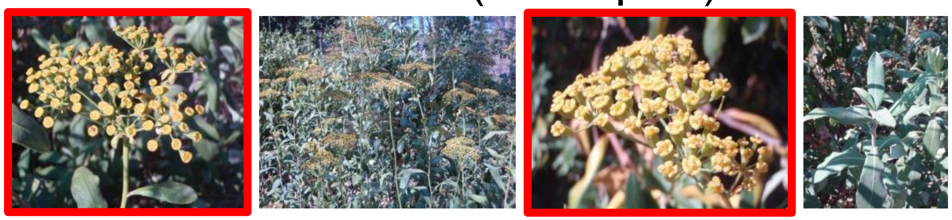
Proposed framework



Observation based identification

- Scores of samples with the same ObservationId are summed
  - Some samples are easily distinguishable while some are not
  - Samples with same ObservationId will have the same scores for all classes

ObservationId 16237 (5 samples)



•••

ObservationId 21782 (1 sample)

ObservationId 37494 (35 samples)

•••

- Introduction and related work

- Modifying the VGG-16 layers CNN model
  - Spatial Pyramid Pooling (SPP)
  - Parametric Rectified Linear Unit (PReLU)

- Post processing
  - Unseen Category Sample Detection algorithm
  - Observation based identification

- Evaluation
  - Data augmentation and model training
  - Results and analysis

- Conclusion

The PlantCLEF 2016 dataset

- 113204 labelled images
  - 111156 images to train CNN model
  - 2048 images to validate CNN model (every half epoch)

- 8000 unlabeled test images

Data Augmentation

- Train images
  - Resize such that shorter side becomes 224 and 336 (two scales) while preserving aspect ratio
  - Random cropping to 224×224
  - Random horizontal flipping

shorter side

longer side

- Validation & test images
  - Resize such that shorter side becomes 224 while preserving aspect ratio
  - No cropping
  - No flipping

Training the modified CNN model

- Model is initialized with Xavier's method*
- Learning rate 0.01 to 0.0001, batch size 50
  - 0.01 for 30 epochs
  - 0.001 for 15 epochs
  - 0.0001 for 8 epochs

- Final validation accuracy 0.626

* X. Glorot and Y. Bengio, "**Understanding the difficulty of training deep feedforward neural networks**," AISTATS 2015

Evaluation metric

- Mean Average Precision
  - including unknown classes and queries (official MAP)
  - restricted to a black list of potentially invasive species
  - ignoring unknown classes and queries

Submitted runs

Run1: Threshold t based on train images

Run2: Threshold t' based on train & validation images

Run3: Threshold t based on train images, use ObservationId

Run4: Threshold t' based on train & validation images, use ObservationId

Run	Official score MAP	MAP restricted to potentially invasive species	MAP ignoring unknown classes and queries
1	0.611	0.600	0.692
2	0.611	0.600	0.693
3	0.736	0.718	0.820
4	0.742	0.717	0.827

Official Evaluation Results

Run	Official score MAP	MAP restricted to potentially invasive species	MAP ignoring unknown classes and queries	
Bluefield Run 4	0.742	0.717	0.827	(with ObservationId)
SabanciUGebzeTU Run 1	0.738	0.704	0.806	
SabanciUGebzeTU Run 3	0.737	0.703	0.807	
Bluefield Run 3	0.736	0.718	0.820	(with ObservationId)
SabanciUGebzeTU Run 2	0.736	0.683	0.807	
SabanciUGebzeTU Run 4	0.735	0.695	0.802	
CMP Run 1	0.710	0.653	0.790	
LIIR KUL Run 3	0.703	0.674	0.761	
LIIR KUL Run 2	0.692	0.667	0.744	
LIIR KUL Run 1	0.669	0.652	0.708	
UM Run 4	0.669	0.598	0.742	
CMP Run 2	0.644	0.564	0.729	
CMP Run 3	0.639	0.590	0.723	
QUT Run 3	0.629	0.610	0.696	
Floristic Run 3	0.627	0.533	0.693	
UM Run 1	0.627	0.537	0.700	
Floristic Run 1	0.619	0.541	0.694	
Bluefield Run 1	0.611	0.600	0.692	(without ObservationId)
Bluefield Run 2	0.611	0.600	0.693	(without ObservationId)
BME TMIT Run 4	0.174	0.144	0.213	

Seen categories: correctly predicted

Seen categories: incorrectly predicted (1/2)

Query image

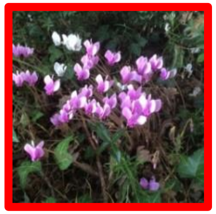
Shares ObservationId with

(100%) *Ailanthus altissima*

(0%) *Ruscus aculeatus* (ground truth)

Seen categories: incorrectly predicted (2/2)

Query image

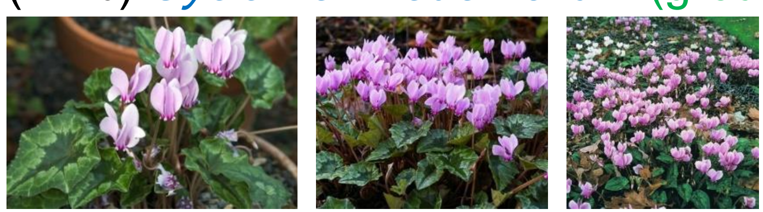


Shares ObservationId with

(43%) *Diplotaxis erucoides*

(41%) *Hesperis matronalis*

(11%) *Cyclamen hederifolium* (ground truth)

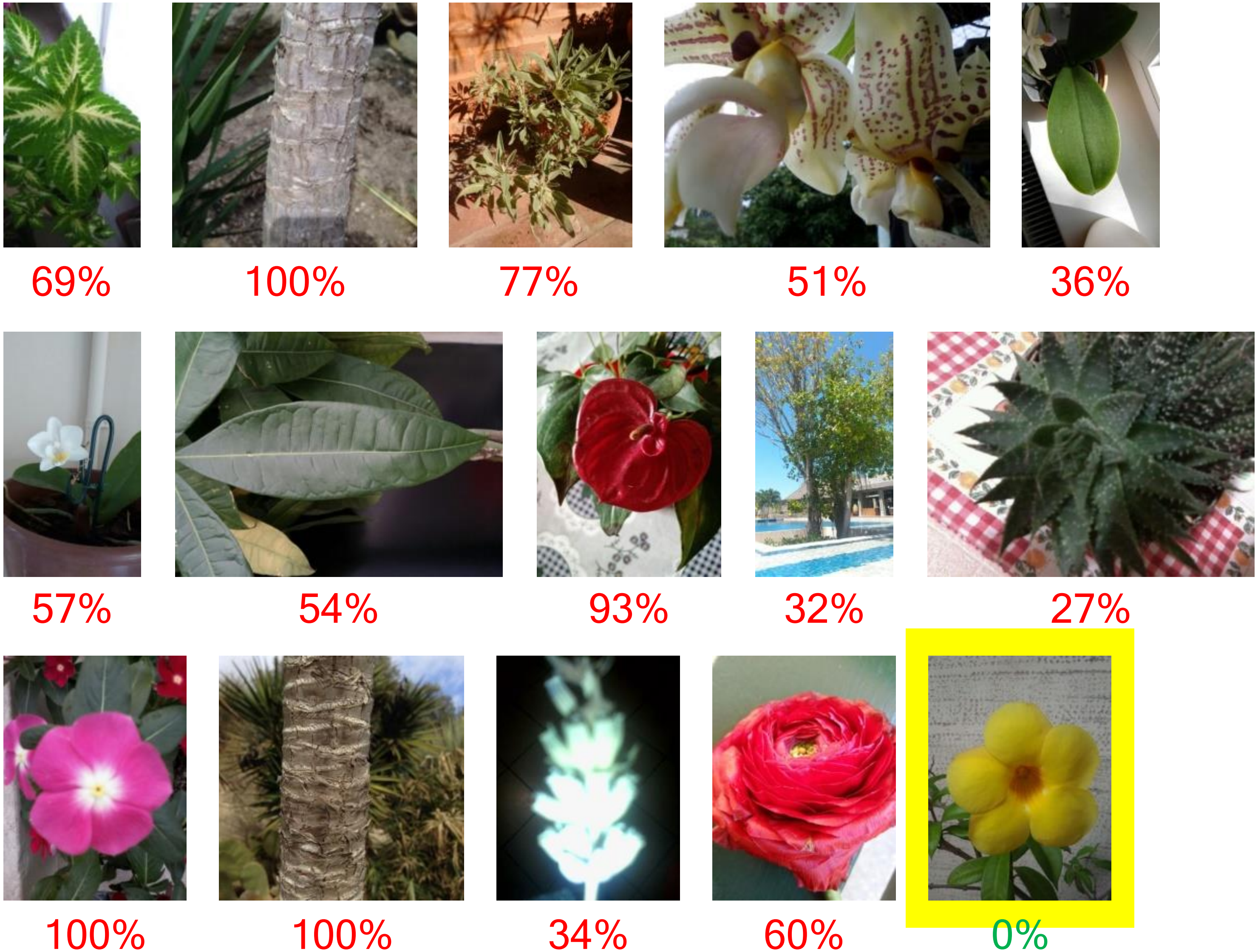


Comparing rejection algorithm

- By computing the **rate of change** from “MAP ignoring unknown classes and queries” to “Official score MAP (including unknown classes and queries)”, we may find out contribution of rejection algorithm
  - Less rate of change may imply better rejection algorithm
  - Apparently our rejection algorithm is generally weaker than other teams

Run	MAP ignoring unknown classes and queries (A)	Official score MAP (B)	Rate of change from A to B
Bluefield Run 4	0.827	0.742	-10.28%
SabanciUGebzeTU Run 1	0.806	0.738	-8.44%
CMP Run 1	0.790	0.710	-10.13%
LIIR KUL Run 3	0.761	0.703	-7.62%
UM Run 4	0.742	0.669	-9.84%
QUT Run 3	0.696	0.629	-9.63%
Floristic Run 3	0.693	0.627	-9.52%
BME TMIT Run 4	0.213	0.174	-18.31%

Result analysis: unseen categories



* Percentage are confidence (softmax normalized score) of first prediction

Result analysis: unseen categories

around half of 8000 test samples
are of unseen categories

Threshold	Number of samples rejected	Correctly rejected	Incorrectly rejected
$\boldsymbol{t}$ (Run 1, 3)	195	139	56
$\boldsymbol{t}'$ (Run 2, 4)	69	56	13

Conclusion

- We attempted to design a plant image identification system that is robust to unseen categories

- Factors that contribute the most to our system
  - Observation based identification (exploitation of ObservationId)
  - Multiscale (two scales) input images to train the CNN model
  - Arbitrarily sized input image during testing (Spatial Pyramid Pooling)

- To further improve our system
  - We should train the CNN model longer (> 100 epochs)
  - Randomly rotate input images
  - Randomly scale input images (instead of just two scales)
  - Exploit other metadata (GPS, Genus, Family, ...)