Hybrid Learning Framework for Large-Scale Web Image Annotation and Localization

Yong Li*, Jing Liu, Yuhang Wang, Bingyuan Liu, Jun Fu, Yunze Gao,
Hui Wu, Hang Song, Peng Ying, Hanqing lu

Image & Video Analysis Group, Institute of Automation, Chinese Academy of Sciences

September 9, 2015

Outline

1. Introduction to The Image Annotation Task
2. The Proposed Hybrid Learning Framework
3. Experiments and Results
4. Conclusions

Introduction to The Image Annotation Task

Two subtasks in Scalable Concept Image Annotation challenge 2015.

- Image concept detection and localisation: it is to annotate and localize 500,000 web images with 251 concepts.

- Generation of textual descriptions of images: it is to describe an image with a textual description of the visual content depicted in the image.

We focus on the subtask 1.

Introduction to The Image Annotation Task

Two subtasks in Scalable Concept Image Annotation challenge 2015.

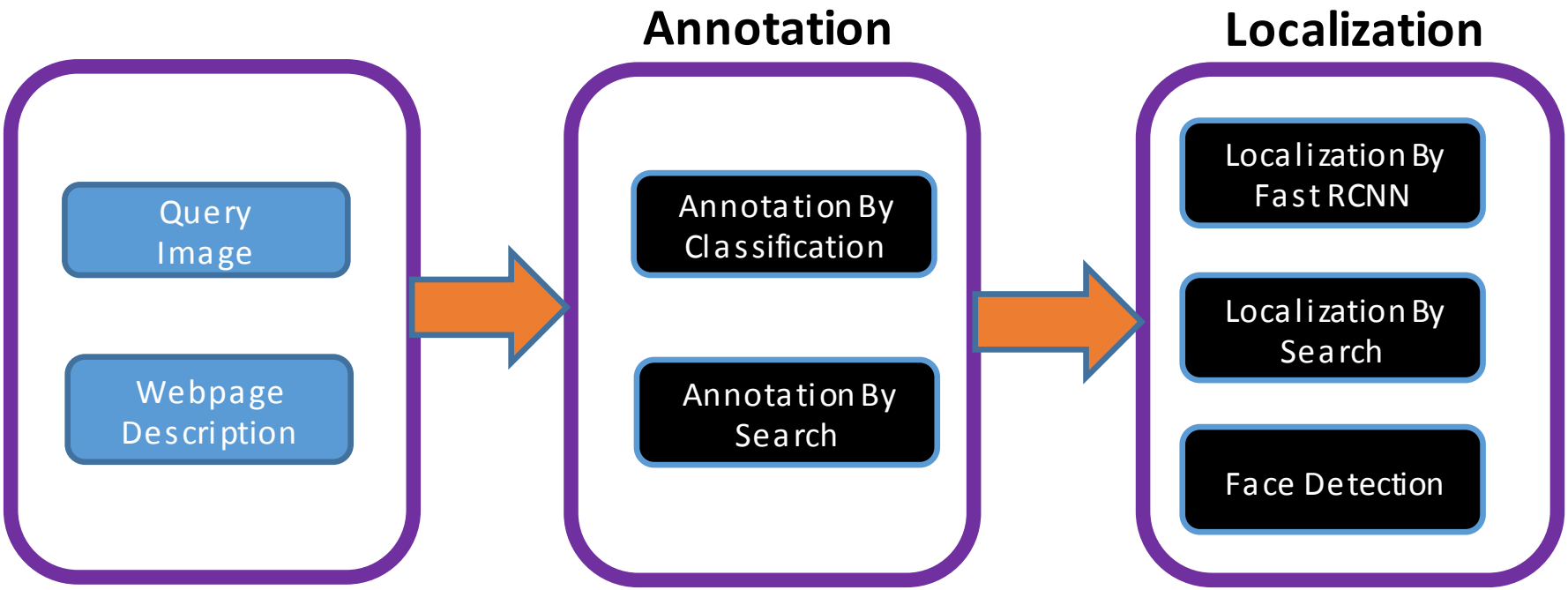
- Image concept detection and localisation: it is to annotate and localize 500,000 web images with 251 concepts.

- Generation of textual descriptions of images: it is to describe an image with a textual description of the visual content depicted in the image.

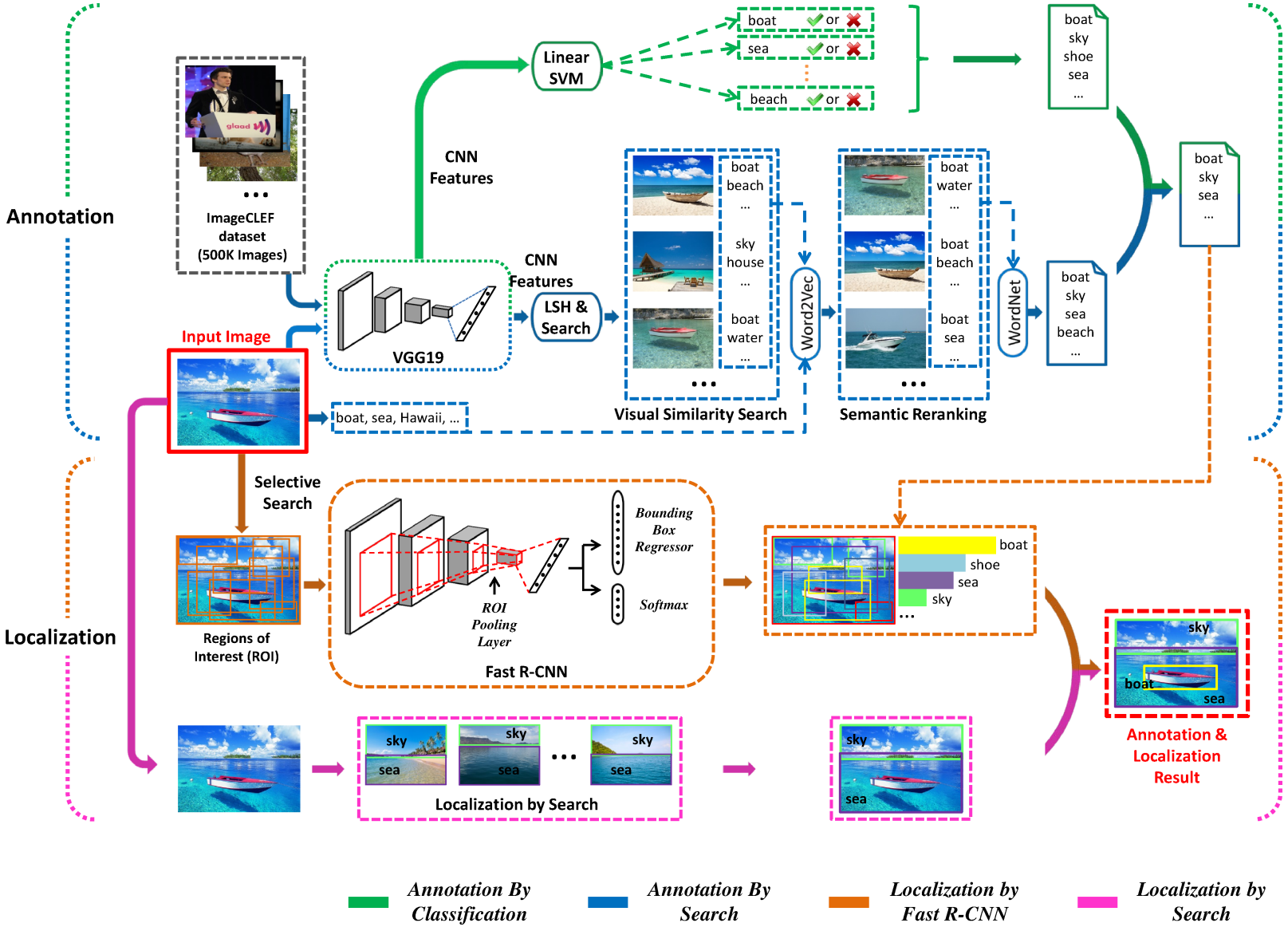
We focus on the subtask 1.

The Proposed Hybrid Learning Framework

We propose a two-stage hybrid learning framework as follows.



Flowchart of The Proposed Approach



Data Preparation

- We prefer multiple online resources to perform such task, including the ImageNet database, the Sun database, the WordNet, and the online image sharing website Flicker.
- There are 175 concepts concurrent in the ImageNet dataset and the ImageCLEF task simultaneously. Meanwhile, there are 217 concepts concurrent in the Sun dataset and the ImageCLEF task.
- For the concepts not in the ImageNet and Sun database or with very few annotated images, we crawled images from the website Flicker and filtered it with 50 images left for each concept.

The Proposed Hybrid Learning Framework

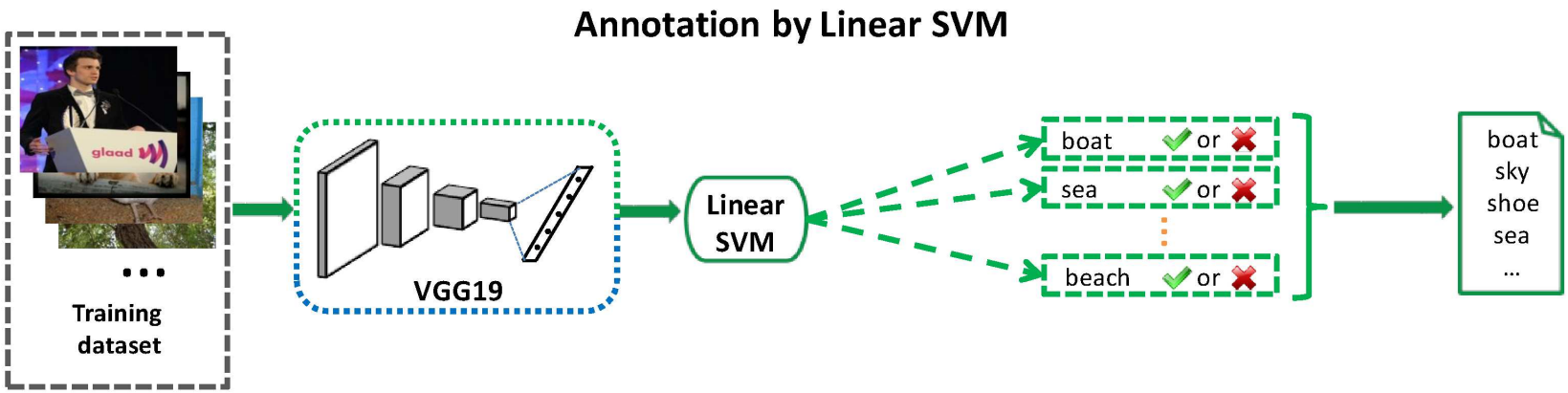
Image Annotation

- **Annotation by classification.**
- Annotation by search.

Concept Localization

- Localization By Fast RCNN.
- Localization By Search.
- Localization of Face Related Concepts.

Annotation by classification



- We train a linear SVM for each concept with one-vs-rest strategy.
- The visual feature is extracted with the released VGG19 model with dimension 4096.

Annotation by classification

- The ratio between the positive samples and negative samples is between 1:10 to 1:5.
- Due to images usually being labelled with multiple concepts in training data, the negative samples for a given concept classifier are selected as the ones whose all labels do not include the concept.
- The final output of SVM is normalized with the Logistic function $f(x) = \frac{1}{1+exp(-w^T x)}$ to make the output value between [0, 1].

The Proposed Hybrid Learning Framework

Image Annotation

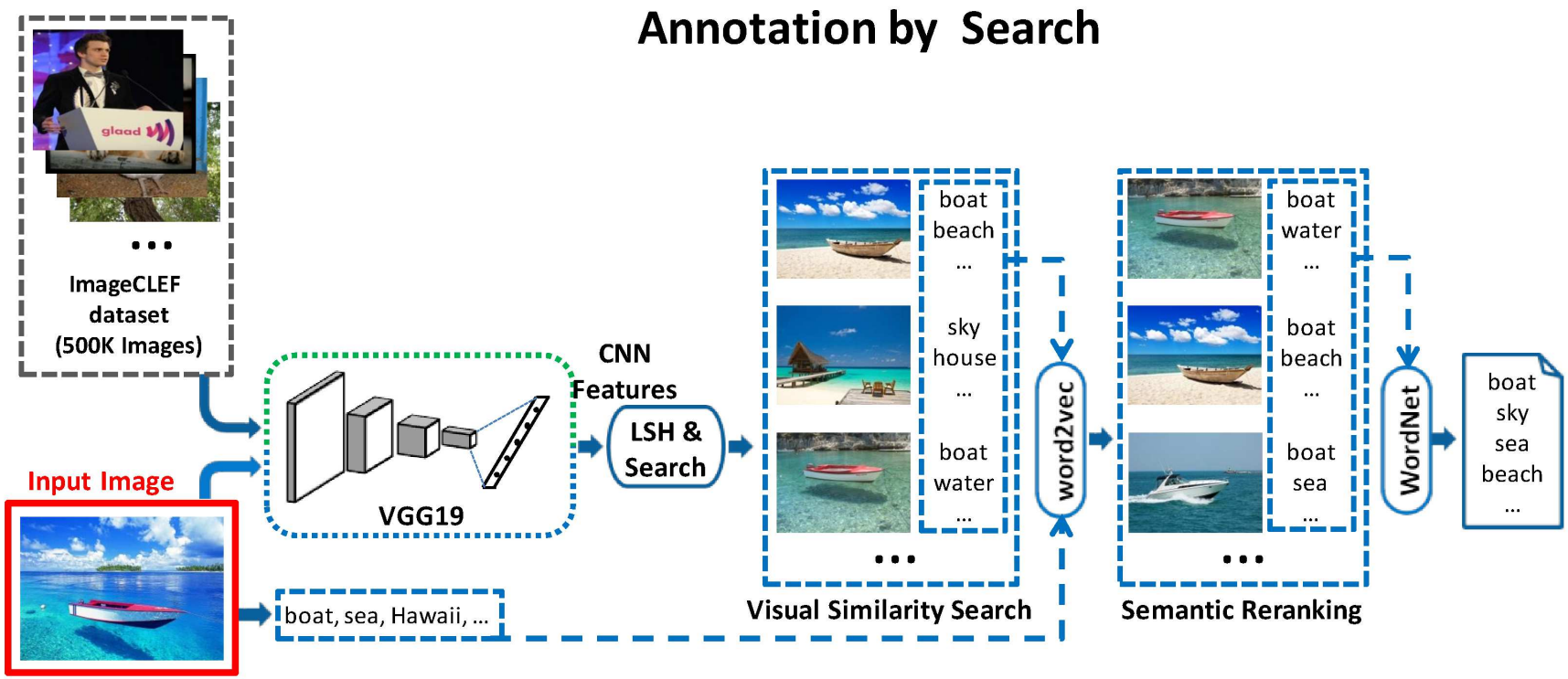
- Annotation by classification.
- **Annotation by search.**

Concept Localization

- Localization By Fast RCNN.
- Localization By Search.
- Localization of Face Related Concepts.

Annotation by Search

The search-based approach for image annotation works on the assumption that visual similar images should reflect similar semantical concepts, and most textual information of web images is relevant to their visual content.



Annotation by Search

Annotation by search mainly consists of the following three steps,

- Search for similar images: The extracted 4096-dimensional deep features are mapped to 32768-dimensional binary hash codes leveraging the random projection algorithm to save memory space and reduce retrieval time.
- Semantic Reranking: To further improve the results of visual similarity search, we explore the textual information of the given image, and perform the semantic similarity search on the top-N_A visually similar images to rerank the similar image set.
- Relevant concept selection: We employ a WordNet-based approach, which is similar to the solution in [1].

[1]: Budikova, P., Botorek, J., Batko, M., Zezula, P.: DISA at imageclef 2014: The search-based solution for scalable image annotation.

Annotation by Search

Annotation by search mainly consists of the following three steps,

- Search for similar images: The extracted 4096-dimensional deep features are mapped to 32768-dimensional binary hash codes leveraging the random projection algorithm to save memory space and reduce retrieval time.
- Semantic Reranking: To further improve the results of visual similarity search, we explore the textual information of the given image, and perform the semantic similarity search on the top-N_A visually similar images to rerank the similar image set.
- Relevant concept selection: We employ a WordNet-based approach, which is similar to the solution in [1].

[1]: Budikova, P., Botorek, J., Batko, M., Zezula, P.: DISA at imageclef 2014: The search-based solution for scalable image annotation.

Annotation by Search

Annotation by search mainly consists of the following three steps,

- Search for similar images: The extracted 4096-dimensional deep features are mapped to 32768-dimensional binary hash codes leveraging the random projection algorithm to save memory space and reduce retrieval time.
- Semantic Reranking: To further improve the results of visual similarity search, we explore the textual information of the given image, and perform the semantic similarity search on the top-N_A visually similar images to rerank the similar image set.
- Relevant concept selection: We employ a WordNet-based approach, which is similar to the solution in [1].

[1]: Budikova, P., Botorek, J., Batko, M., Zezula, P.: DISA at imageclef 2014: The search-based solution for scalable image annotation.

The Proposed Hybrid Learning Framework

Image Annotation

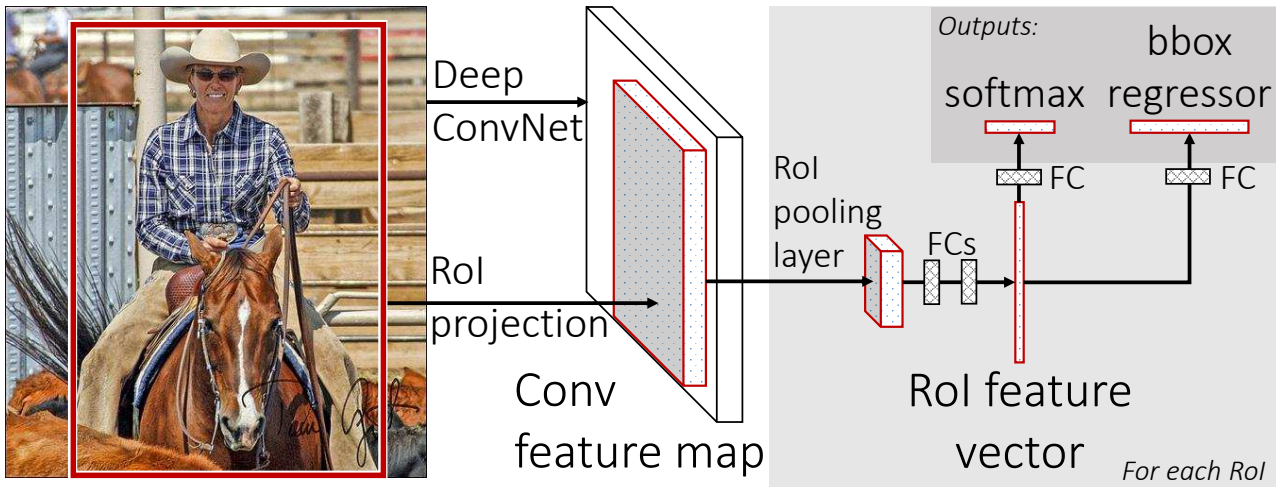
- Annotation by classification.
- Annotation by search.

Concept Localization

- **Localization By Fast RCNN.**
- Localization By Search.
- Localization of Face Related Concepts.

Localization By Fast RCNN

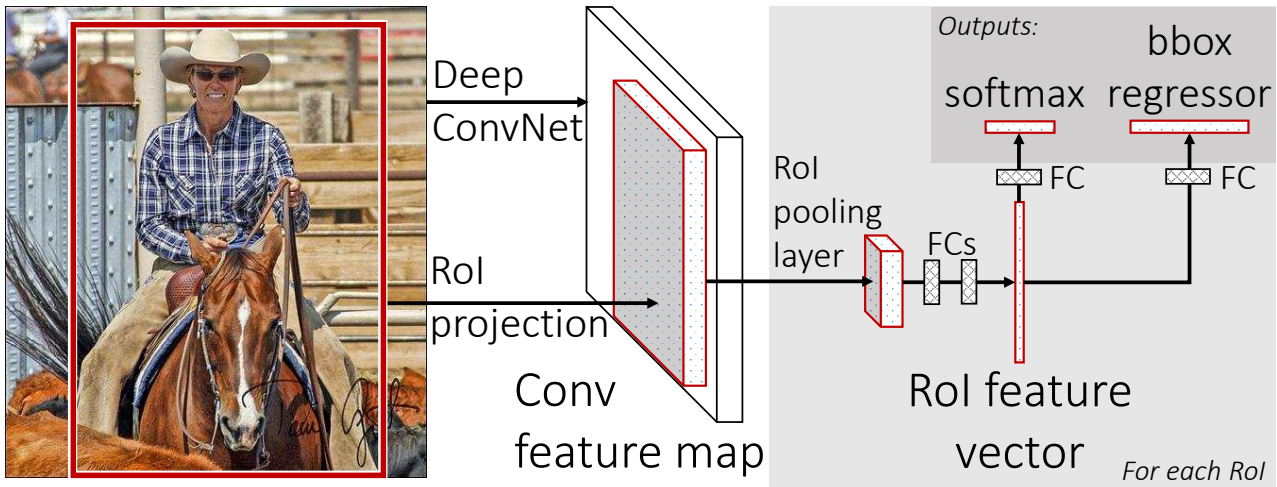
Fast RCNN originates from the method RCNN, which provides classification result and regressed location simultaneously.



Comparisons of such two methods will be given. We have used both methods during the competition.

Localization By Fast RCNN

Fast RCNN originates from the method RCNN, which provides classification result and regressed location simultaneously.



Comparisons of such two methods will be given. We have used both methods during the competition.

Localization By Fast RCNN

RCNN has demonstrated its advantage in object detection. However, it has notable drawbacks.

- Training is a multi-stage pipeline, which is expensive in space and time. One first finetunes a ConvNet for detection using cross-entropy loss. Then linear SVMs are fit to ConvNet features computed on warped object proposals. Finally, bounding-box regressors are learned.
- Test-time detection is slow. At test time, features are extracted from each warped object proposal in each test image. Detection with VGG16 takes 47s per image.

Localization By Fast RCNN

Compared with RCNN, Fast RCNN has the following advantages.

- It achieves higher detection quality(mAP).
- Training is single-stage, using a multi-task loss. It provides classification result and regressed location simultaneously for each candidate object proposal.
- No disk storage is required for feature caching.

For us, we preferred the RCNN method when decided to take part in the task. However, it is really tough to deal with 500,000 images with RCNN. We met fast RCNN two weeks before the deadline. It helped us a lot.

Localization By Fast RCNN

Compared with RCNN, Fast RCNN has the following advantages.

- It achieves higher detection quality(mAP).
- Training is single-stage, using a multi-task loss. It provides classification result and regressed location simultaneously for each candidate object proposal.
- No disk storage is required for feature caching.

For us, we preferred the RCNN method when decided to take part in the task. However, it is really tough to deal with 500,000 images with RCNN. We met fast RCNN two weeks before the deadline. It helped us a lot.

The Proposed Hybrid Learning Framework

Image Annotation

- Annotation by classification.
- Annotation by search.

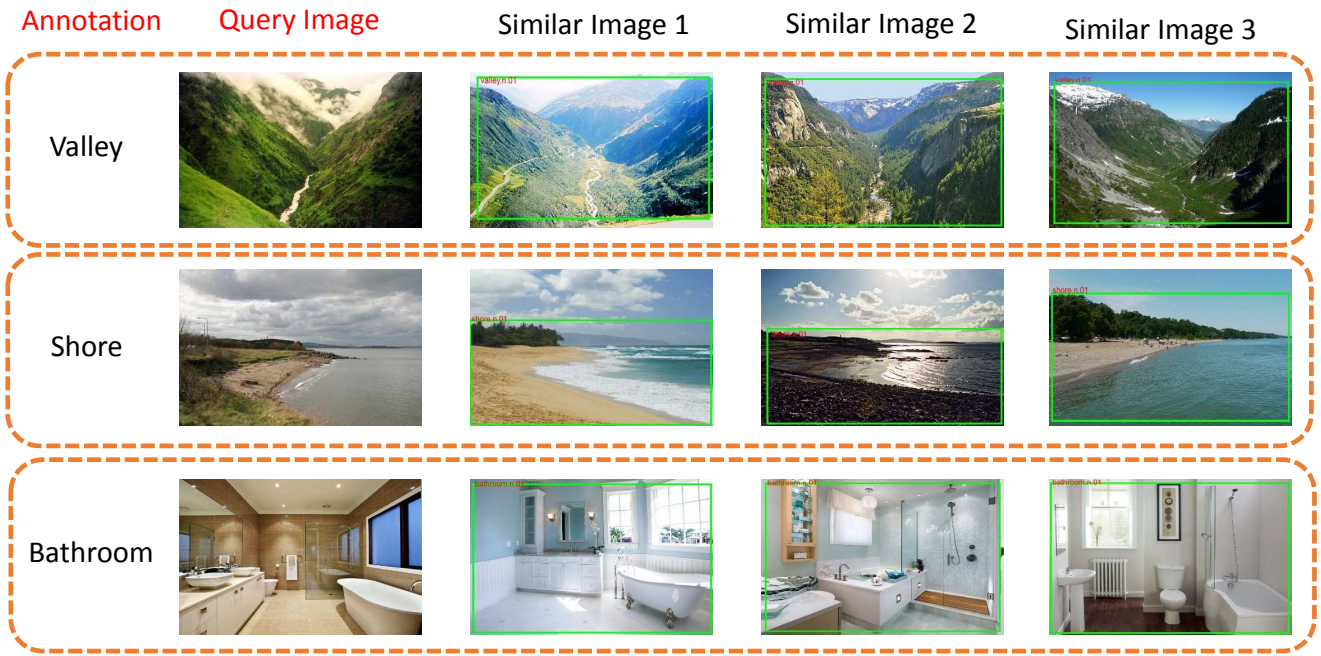
Concept Localization

- Localization By Fast RCNN.
- **Localization By Search.**
- Localization of Face Related Concepts.

Localization By Search

We give a special consideration to the 25 scene related concepts for concept localization (e.g., “beach”, “sea”, “river” and “valley”).

- We first find its top-N_L visually similar neighbors with the same concept in the localization training data.
- We use their merged bounding box as the location of the scenery concept.



The Proposed Hybrid Learning Framework

Image Annotation

- Annotation by classification.
- Annotation by search.

Concept Localization

- Localization By Fast RCNN.
- Localization By Search.
- **Localization of Face Related Concepts.**

Localization of Face Related Concepts

We give a special consideration to the person and face related concepts. Face detection and facial point localization have been actively studied over the past years and achieve satisfactory performance.

- Face
- Head
- Mouth
- Eye
- Nose
- Beard
- Hair
- Tongue
- Ear
- Neck
- Hand
- Leg
- Foot
- Arm

Linear classifiers are trained with the SIFT features extracted on the facial points to determine the following concepts.

- Man
- Woman
- Male child
- Female child

Localization of Face Related Concepts

We give a special consideration to the person and face related concepts. Face detection and facial point localization have been actively studied over the past years and achieve satisfactory performance.

- Face
- Head
- Mouth
- Eye
- Nose
- Beard
- Hair
- Tongue
- Ear
- Neck
- Hand
- Leg
- Foot
- Arm

Linear classifiers are trained with the SIFT features extracted on the facial points to determine the following concepts.

- Man
- Woman
- Male child
- Female child

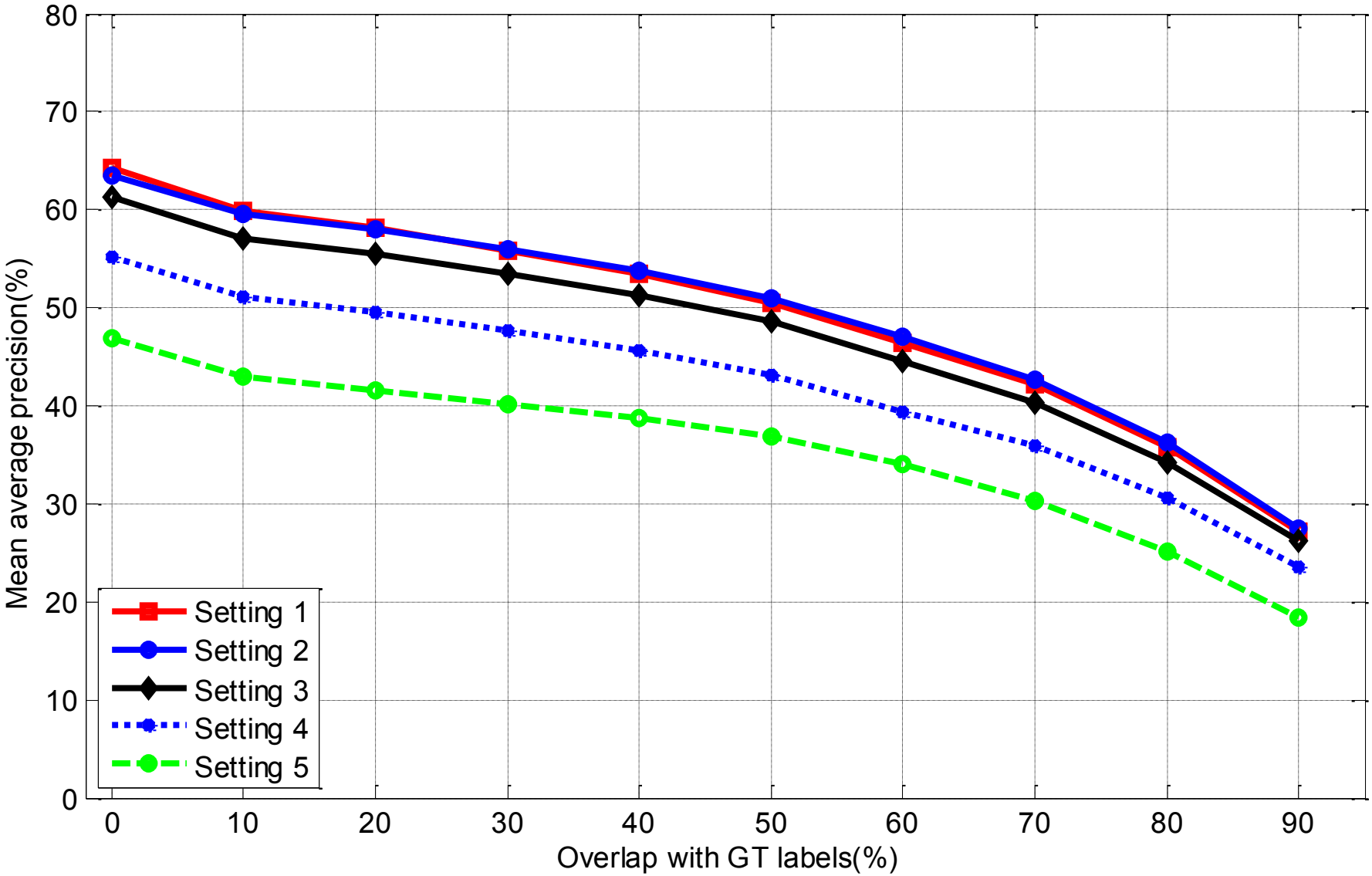
Experiments and Results

We have submitted 8 runs with 5 different settings of combinations with the above model modules, including Annotation By Classification (ABC), Annotation By Search (ABS), localization by Fast R-CNN (FRCN), Localization By Search (LBS) and Concept Extension (CE).

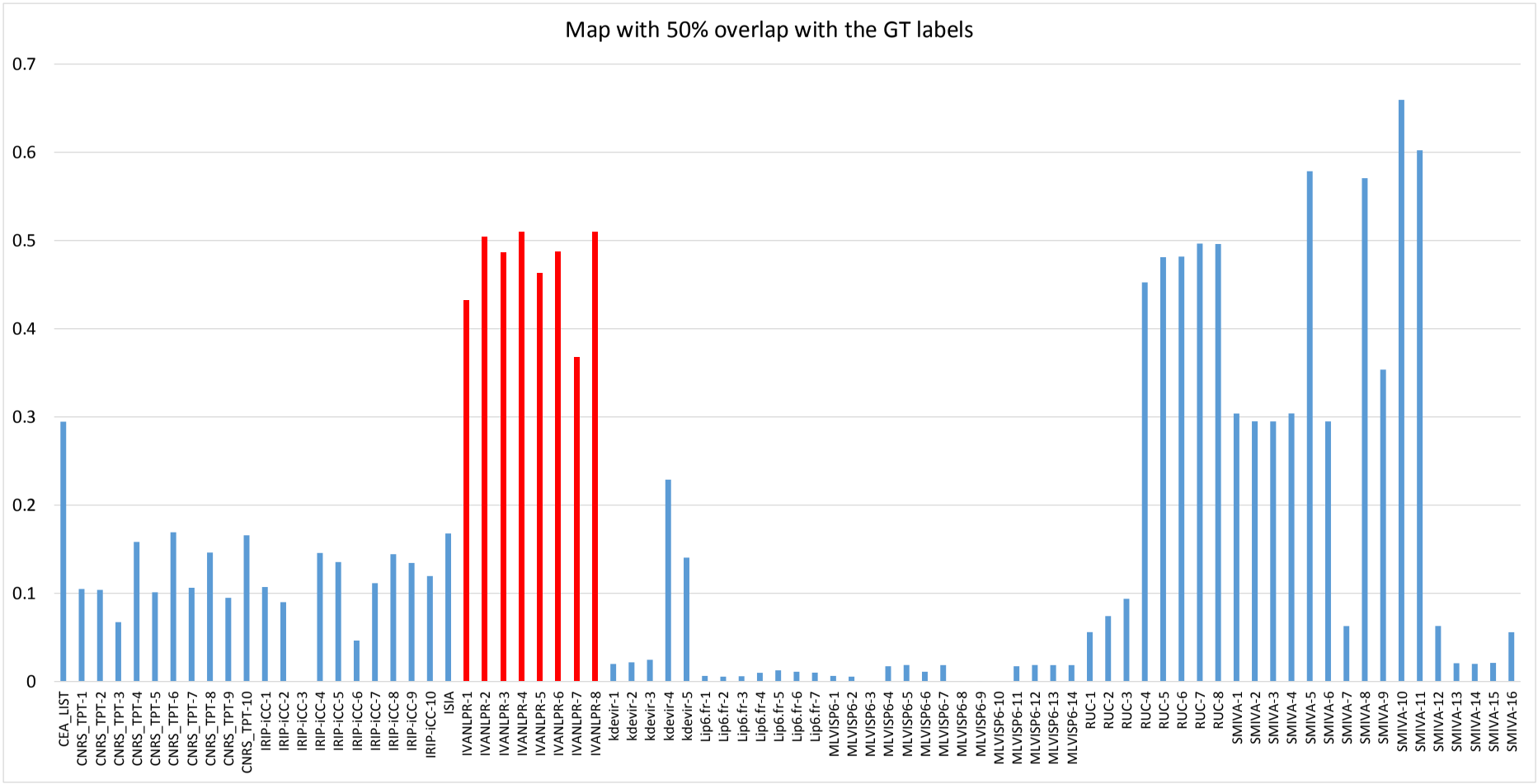
Method	ABC	CE	ABS	LBS	FRCN	SVM_Threshold	Overlap 0.5	Overlap 0
Setting 1	yes	no	yes	yes	yes	0.5	0.510	0.642
Setting 2	yes	yes	yes	yes	yes	0.4	0.510	0.635
Setting 3	yes	yes	no	yes	yes	0.4	0.486	0.613
Setting 4	yes	no	no	yes	yes	0.4	0.432	0.552
Setting 5	no	no	no	no	yes	0.4	0.368	0.469

- Effectiveness ABS is verified by comparing the result of setting 2 and setting 3 with 2 percent improvement.
- Effectiveness of FRCN is verified with setting 5, which is unsatisfactory. The proposed two stage process is more suitable to deal with such task.

MAP with Different Recall Rates



Submission Results of Different Teams



Conclusions

- We propose a two-stage hybrid learning framework for Large-Scale Web Image Annotation and Localization.
- It is important to analyse the concepts carefully and dealing with different kinds of concepts with different modules is necessary.
- For the proposed method, the performance is limited by the number of training samples with full annotations.

Thanks for your attention!

More info: http://www.foreverlee.net