Weakly Supervised Learning

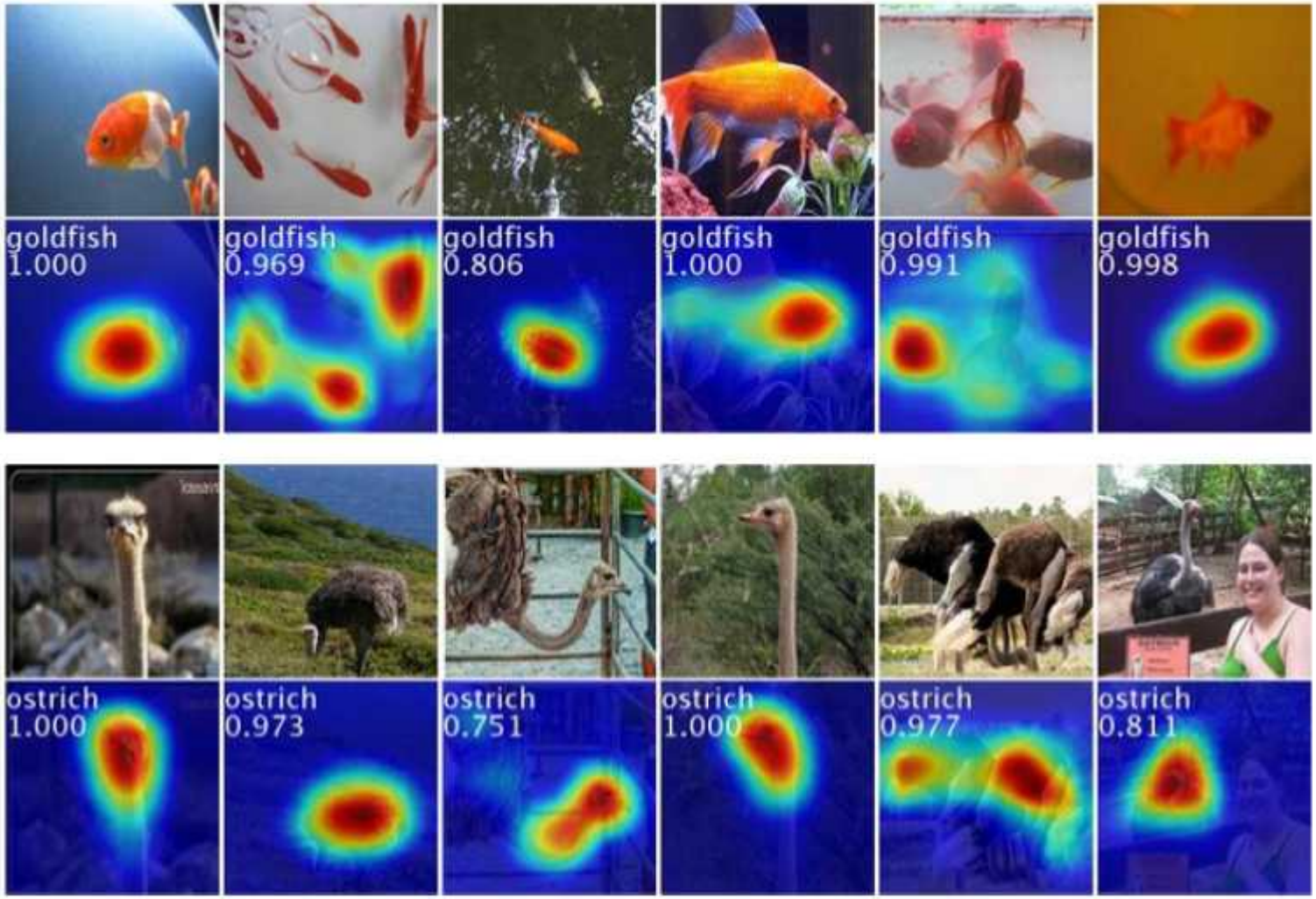
Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT

Based on Junho Cho's slide (@PIL SNU)

Weakly Supervised?

Introduction



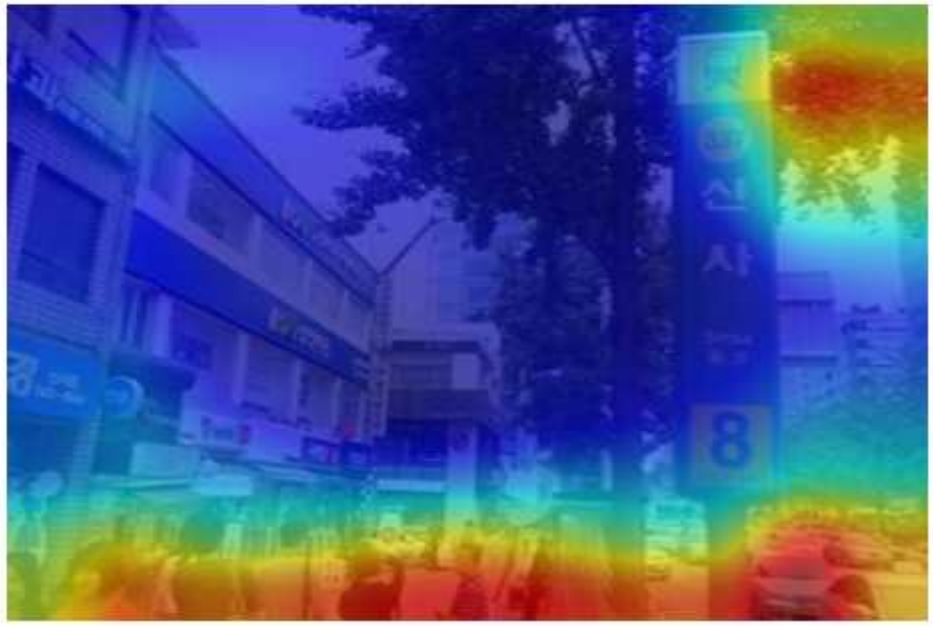
Introduction

Introduction

You can try it yourself on http://places.csail.mit.edu/demo.html

Predictions:

- **Type of environment:** outdoor
- **Semantic categories:** crosswalk:0.35, plaza:0.10, hospital:0.07, office_building:0.06, motel:0.05
- **SUN scene attributes:** man-made, naturallight, nohorizon, mostlyverticalcomponents, leaves, foliage, trees, openarea, glass, shopping
- **Informative region for the category *crosswalk* is:**

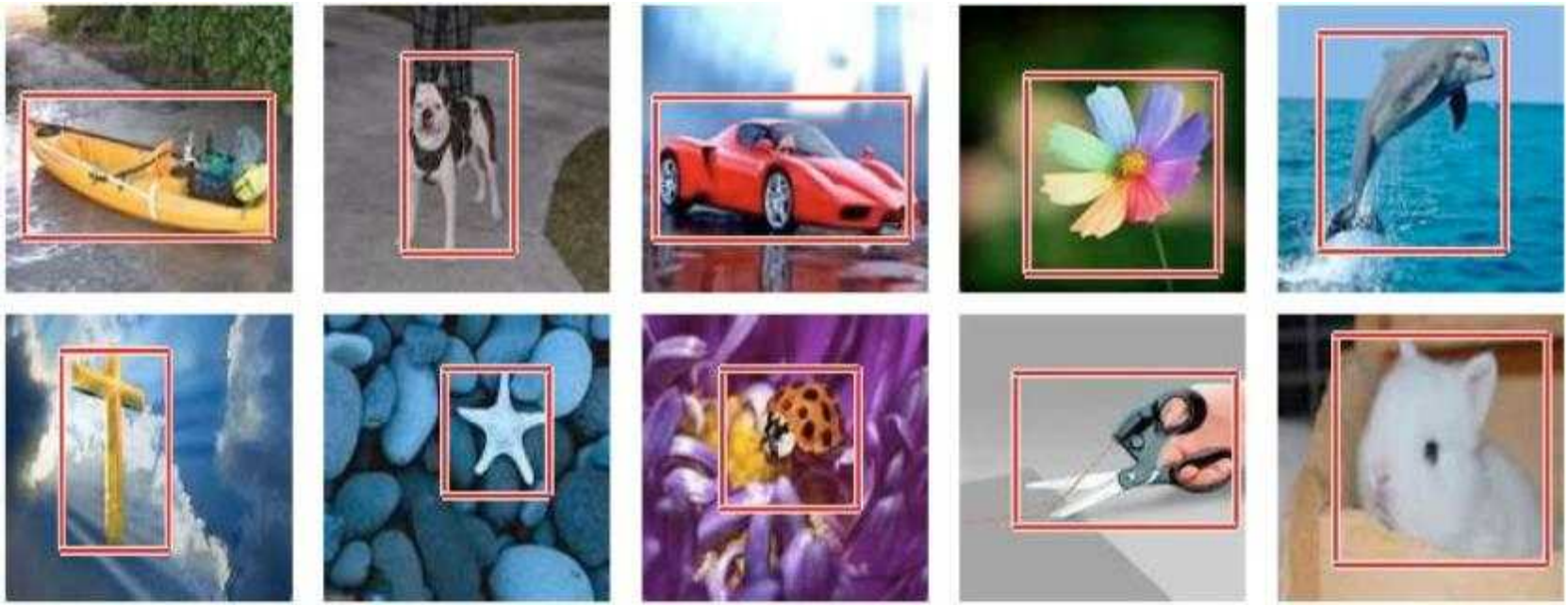


Introduction

Learning Deep Features for Discriminative Localization – CVPR2016

◆ **Weakly supervised object localization**

- Only trained with **class label** on image
- Yet able to localize object very well
- Close to fully supervised learned AlexNet

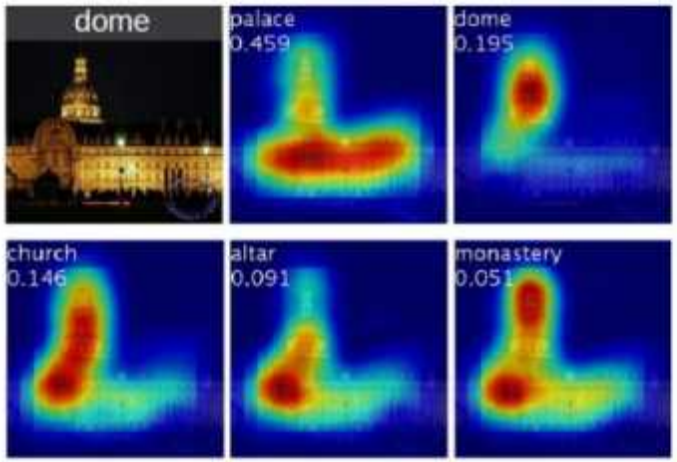


Introduction

Learning Deep Features for Discriminative Localization – CVPR2016

◆ **Visualizing the internal representation of CNNs.**

- Localization by **Class Activation Map (CAM)**
- Units activated by some visual pattern within its receptive field
- Visualize what activates for the output
- in One CNN forward pass.



Class activation maps of top 5 predictions

Class activation maps for one object class

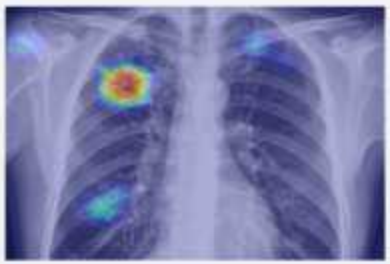
Lunit

https://vimeo.com/161429649

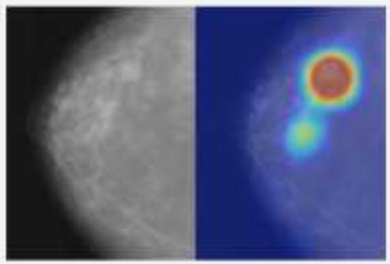
Lunit

Start-Up named after "**Learning unit**"

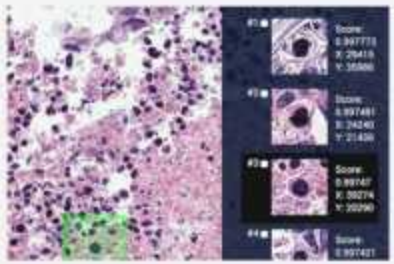
- Data-driven medical image startup
- But Medical images lack annotations!
- Hard to get from doctors.
- Annotation effort is too big.



Digital Chest X-ray



Digital Mammography



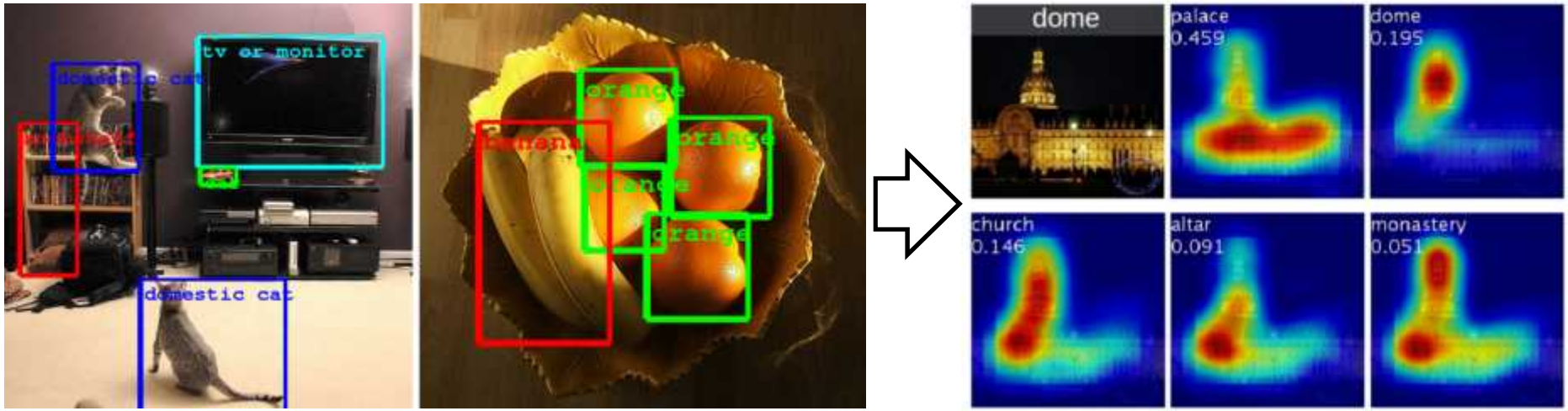
Digital Pathology

Fundus Photography (TBD)

Weakly Supervised Object Localization

Usually **supervised learning** of localization is **annotated with bounding box**

What if **localization is possible with image label** without bounding box annotations?

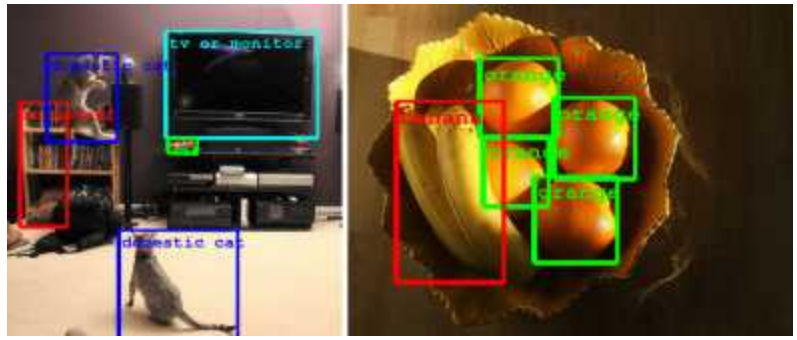


Today's seminar: Learning Deep Features for Discriminative Localization

1512.04150v1 Zhou *et al.* 2015 CVPR2016

Localization task (ILSVRC)

- Classification and **localize** its position
- ILSVRC LOC: 1000 classes and each object annotated with bounding box
  - Predict 5 class labels and 5 bounding boxes for each class label.



Supervised localization (VGGnet, Deep Residual Net)

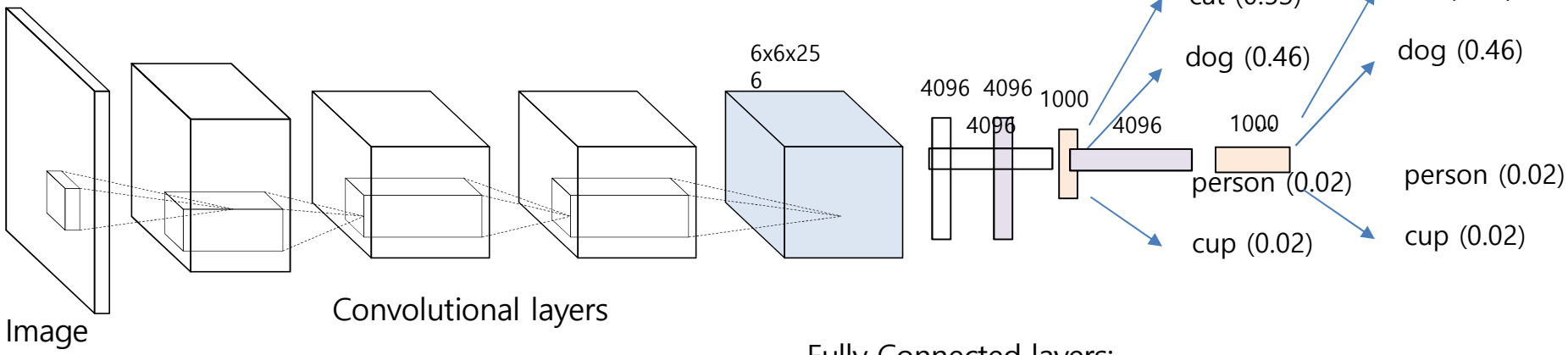
- **Per-class regression for each class.**
  - learn a bounding box regressor (4D vector: x, y, w, h) for each class
- **Train image-level classifier** to predict class labels of an image.
  - **then localization** by predicting bounding boxes based on the predicted classes.
- Pre-train networks for ImageNet classification and then fine-tune for localization.

Easy approach: localization by classification

- Localization by **classification model** (AlexNet)
- Classification map with **Fully Convolutional Network** instead single classification.

< Output (classes)>

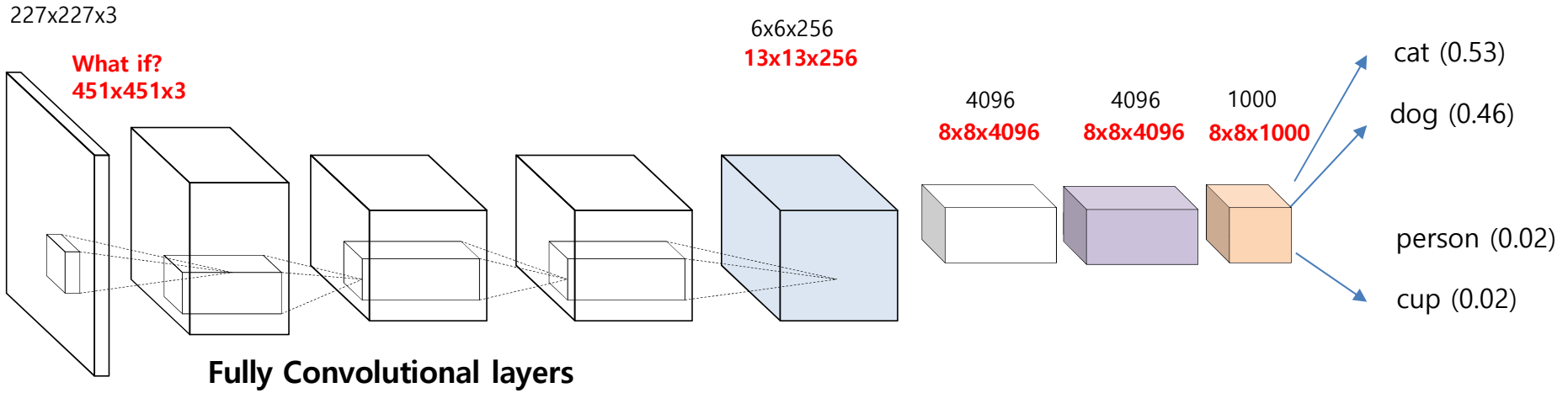
< AlexNet trained on ImageNet >



Fully Connected layers:
Can be considered as 1x1 convolution

Easy approach: localization by classification

< AlexNet trained on ImageNet >

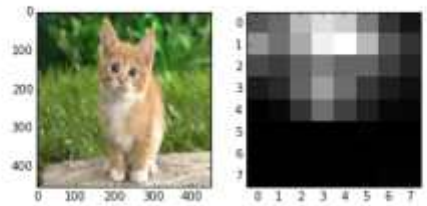


Blob(Tensor) sizes

```
In [5]: [(k, v.data.shape) for k, v in net_full_conv.blobs.items()]
Out[5]: [('data', (1, 3, 451, 451)),
         ('conv1', (1, 96, 111, 111)),
         ('pool1', (1, 96, 55, 55)),
         ('norm1', (1, 96, 55, 55)),
         ('conv2', (1, 256, 55, 55)),
         ('pool2', (1, 256, 27, 27)),
         ('norm2', (1, 256, 27, 27)),
         ('conv3', (1, 384, 27, 27)),
         ('conv4', (1, 384, 27, 27)),
         ('conv5', (1, 256, 27, 27)),
         ('pool5', (1, 256, 13, 13)),
         ('fc6-conv', (1, 4096, 8, 8)),
         ('fc7-conv', (1, 4096, 8, 8)),
         ('fc8-conv', (1, 1000, 8, 8)),
         ('prob', (1, 1000, 8, 8))]
```

```
out = net_full_conv.forward_all(data=np.asarray([transformer.preprocess('data', im)]))
print out['prob'][0].argmax(axis=0)
# show net input and confidence map (probability of the top prediction at each location)
plt.subplot(1, 2, 1)
plt.imshow(transformer.deprocess('data', net_full_conv.blobs['data'].data[0]))
plt.subplot(1, 2, 2)
plt.imshow(out['prob'][0,281])
```

```
[[282 282 281 281 281 281 277 282]
 [281 283 283 281 281 281 281 282]
 [283 283 283 283 283 283 287 282]
 [283 283 283 281 283 283 283 259]
 [283 283 283 283 283 283 283 259]
 [283 283 283 283 283 283 259 259]
 [283 283 283 283 259 259 259 277]
 [335 335 283 259 263 263 263 277]]
Out[11]: <matplotlib.image.AxesImage at 0x12379a690>
```

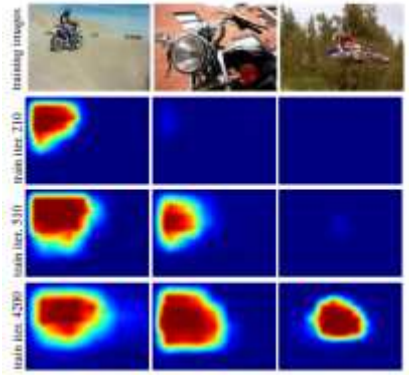


281 tiger cat
282 tabby
283 persian

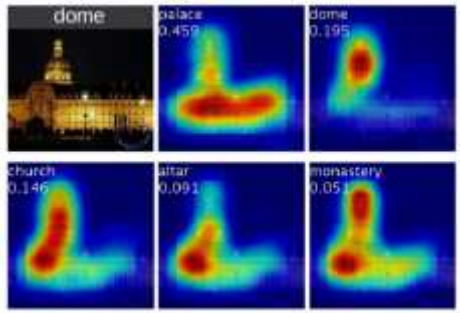
Thus classification CNN already able to localize

Related Works

- Convolutional layers can localize objects.
- The ability is lost when **fully-connected(fc) layers** used for classification.
- Network in Network and GoogLeNet use fully convolutional network
  - avoid use of fc layers (except last layer)
- Global pooling layer used as structural regularizer, preventing overfitting.
- [Oquab et al 2015] use **global max pooling** to localize
  - Limited to point lying in the boundary of object rather tan full extent of object.



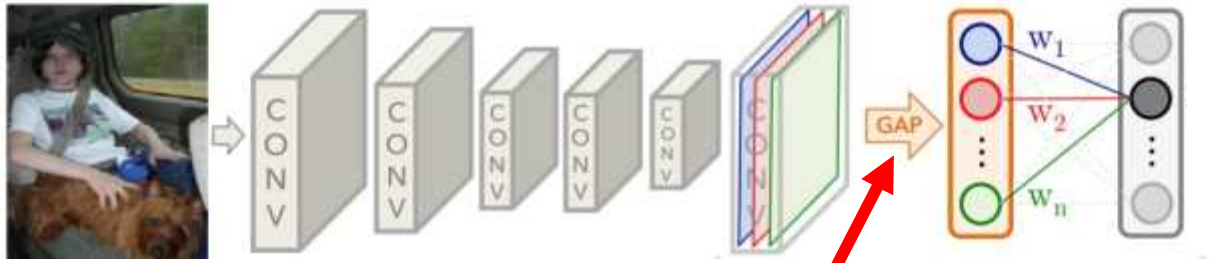
Is object localization for free? – Weakly-supervised learning with convolutional neural networks
arxiv Oquab *et al.* CVPR2015



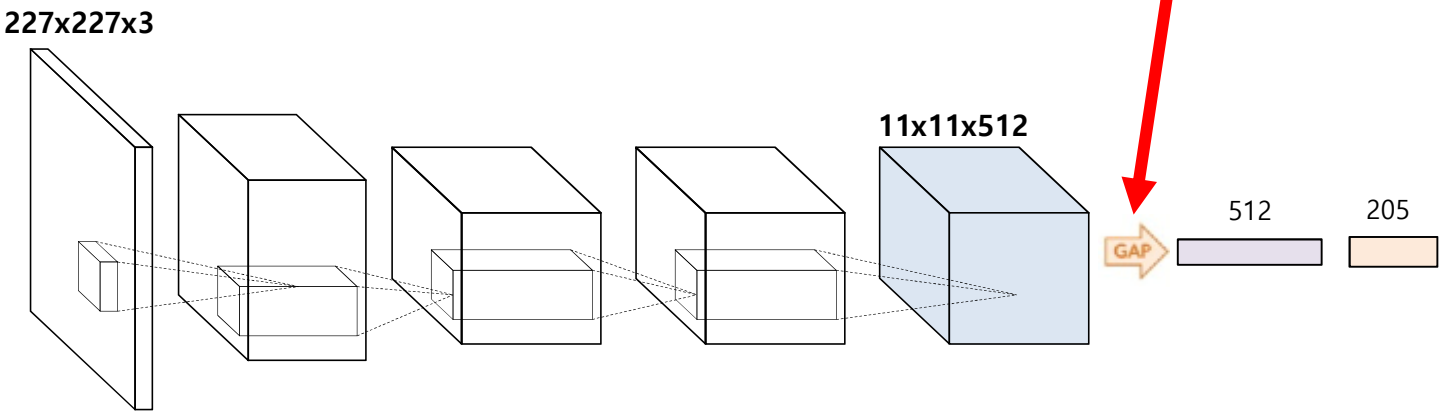
Today's seminar: Learning Deep Features for Discriminative Localization
1512.04150v1 Zhou *et al.* 2015 CVPR2016

Architecture

AlexNet+GAP+places205



11x11 Avg Pooling: Global Average Pooling (GAP)




```
('data', (10, 3, 227, 227))
('conv1', (10, 96, 55, 55))
('pool1', (10, 96, 27, 27))
('norm1', (10, 96, 27, 27))
('conv2', (10, 256, 27, 27))
('pool2', (10, 256, 13, 13))
('norm2', (10, 256, 13, 13))
('conv3', (10, 384, 13, 13))
('conv4', (10, 384, 13, 13))
('conv5', (10, 384, 13, 13))
('pool5', (10, 384, 11, 11))
('conv6', (10, 512, 11, 11))
('conv7', (10, 512, 11, 11))
('pool8_global', (10, 512, 1, 1))
('fc9', (10, 205))
('prob', (10, 205))
→ alexnetplusCAM_places205
```

Class activation map (CAM)

- **Identify important image regions** by projecting back the weights of output layer to convolutional feature maps.
- CAMs can be generated for each class in single image.
- Regions for each categories are different in given image.
  - palace, dome, church …

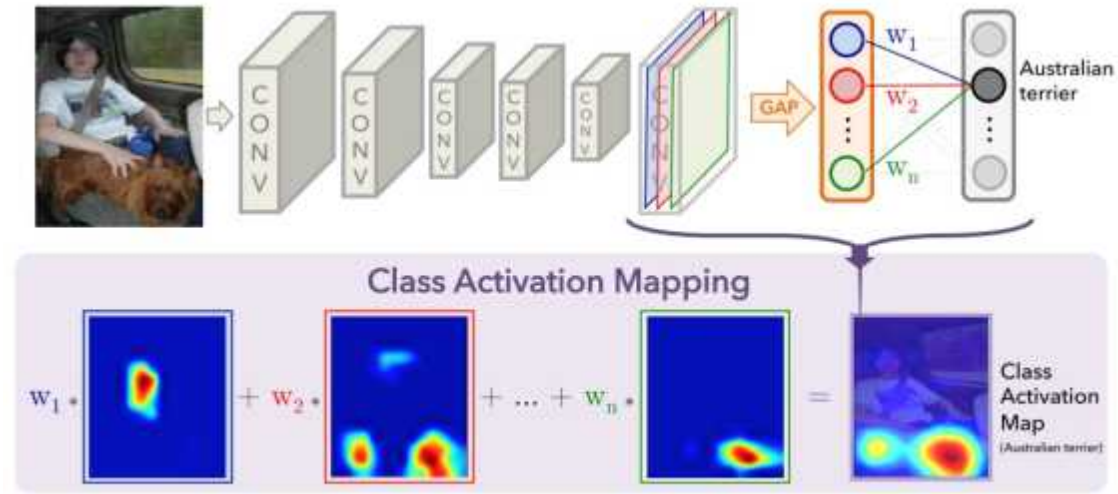


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Results

- CAM on top 5 predictions on an image
- CAM for one object class in images

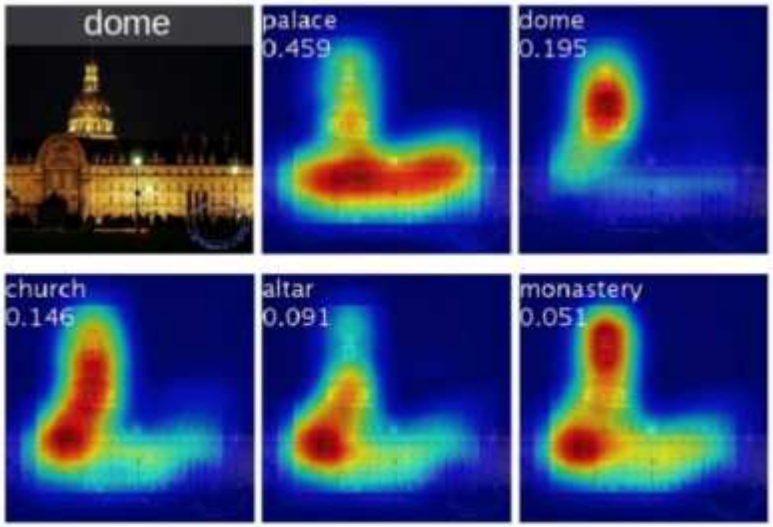


Figure 4. Examples of the CAMs generated from the top 5 predicted categories for the given image with ground-truth as dome. The predicted class and its score are shown above each class activation map. We observe that the highlighted regions vary across predicted classes e.g., *dome* activates the upper round part while *palace* activates the lower flat part of the compound.

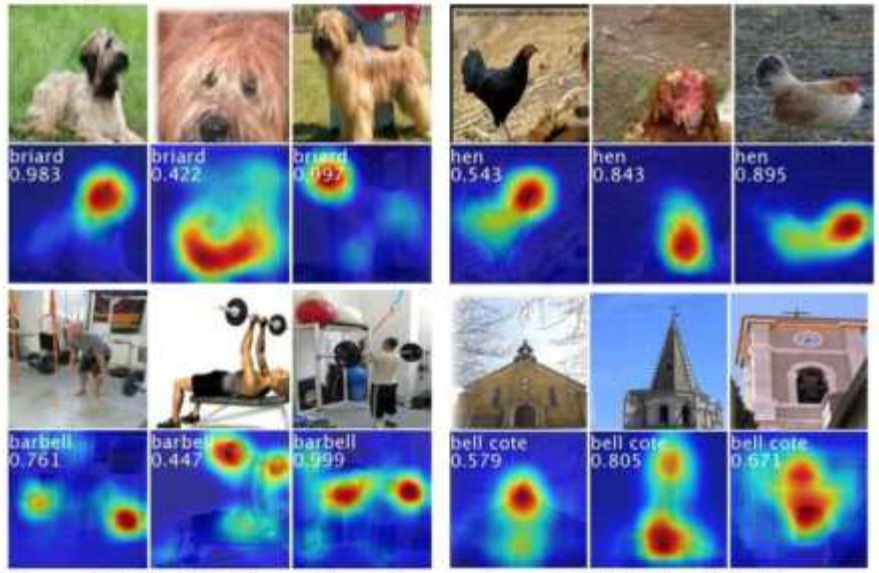
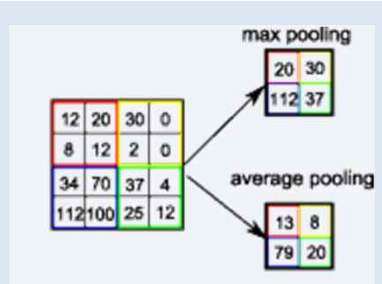


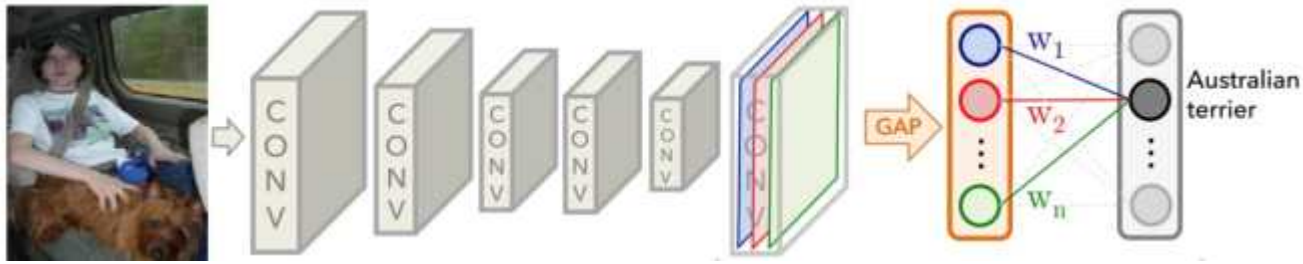
Figure 3. The CAMs of four classes from ILSVRC [20]. The maps highlight the discriminative image regions used for image classification e.g., the head of the animal for *briard* and *hen*, the plates in *barbell*, and the bell in *bell cote*.

GAP vs. GMP

- Oquab et al. CVPR2015
  Is object localization for free? weakly-supervised learning with convolutional neural networks.
  - Use global max pooling(GMP)
- Intuitive difference between GMP and GAP?
  - GAP loss encourages identification on the extent of an object.
  - GMP loss encourages it to identify **just one** discriminative part.
  - GAP, average of a map maximized by finding **all discriminative** parts of object
    - if activations is all low, output of particular map reduces.
  - GMP, low scores for all image regions except the most discriminative part
    - do not impact the score when perform MAX



pooling



GAP & GMP

- GAP (upper) vs GMP (lower)
- GAP outperforms GMP
- GAP highlights more **complete** object regions and less background noise.
- Loss for average pooling benefits when the network identifies ***all* discriminative** regions of an object

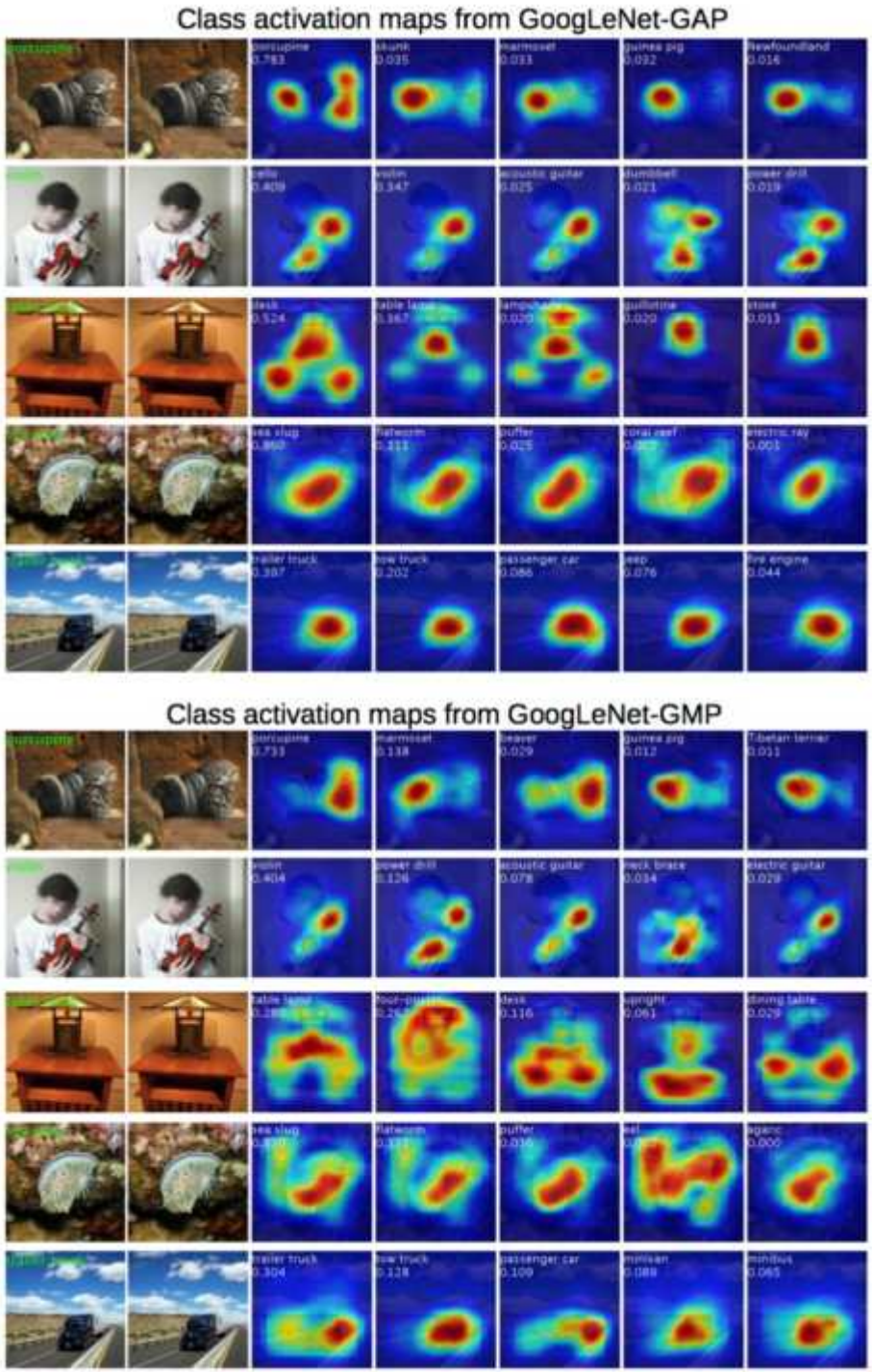


Table 1. Classification error on the ILSVRC validation

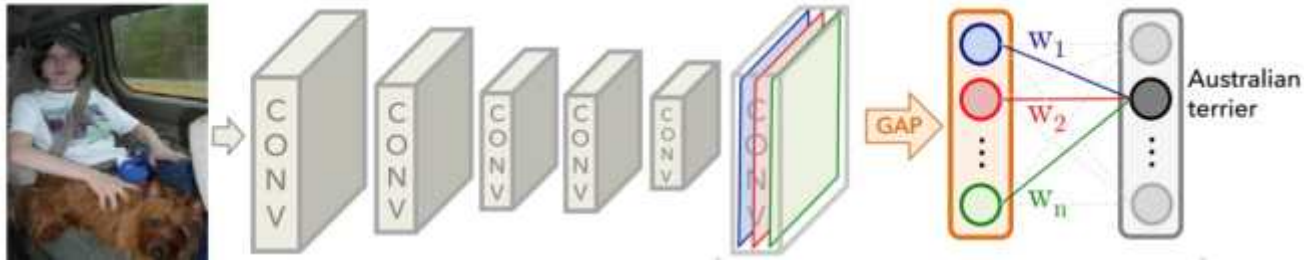
Networks	top-1 val. error	top-5 val. error
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
AlexNet*-GAP	44.9	20.9
AlexNet-GAP	51.1	26.3
GoogLeNet	31.9	11.3
VGGnet	31.2	11.4
AlexNet	42.6	19.5
NIN	41.9	19.6
GoogLeNet-GMP	35.6	13.9

Table 2. Localization error on the ILSVRC validation set. *Back prop* refers to using [22] for localization instead of CAM.

Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

Experiments!

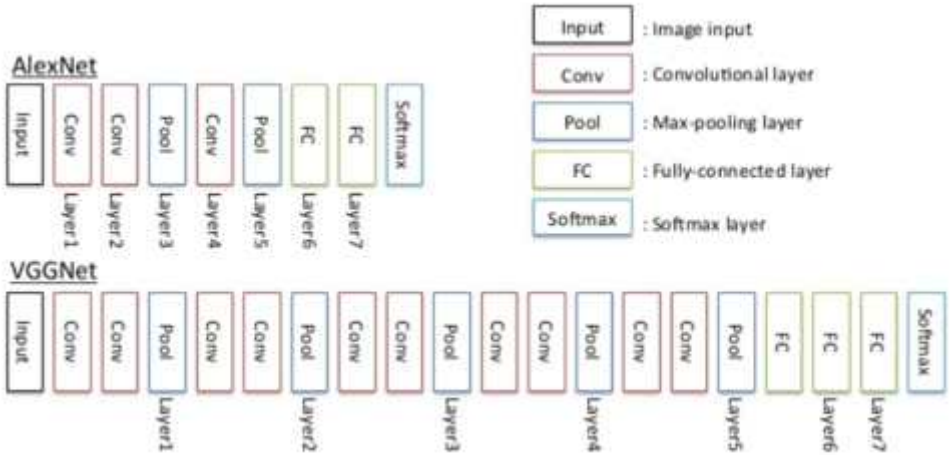
- **AlexNet, VGGnet, GoogeLeNet**
  : Trained with ImageNet from the scratch
  - All remove Fully Connected(FC) layers
  - Replace with GAP & FC softmax layer

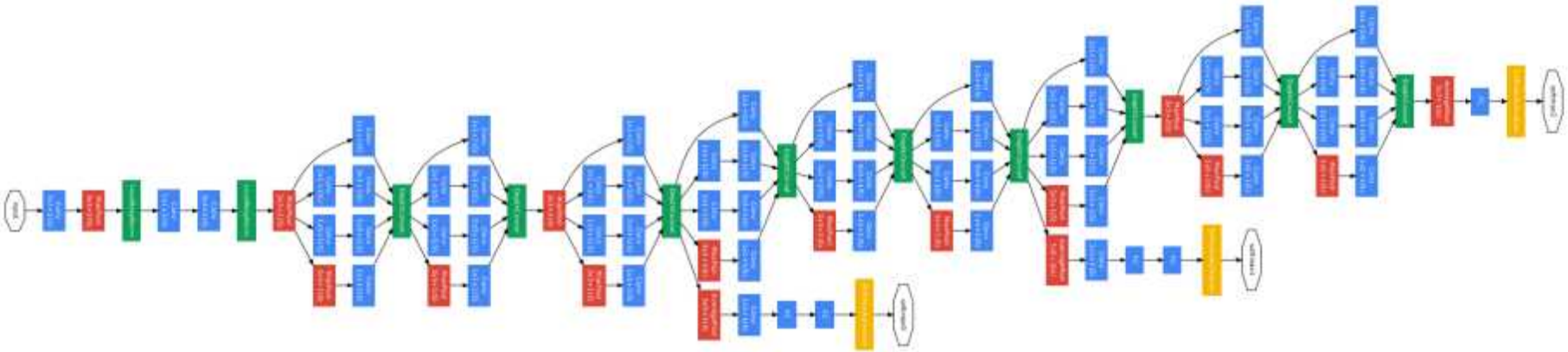


- **Better if Conv layer before GAP has higher spatial resolution**
  - *mapping resolution*
  - Removed last conv layer and add higher spatial resolution
  - Ex) VGGnet: Conv5-3 3x3x512 → 3x3x1024

AlexNet, VGGnet, GoogLeNet

- Reminder on popular nets.
- GoogeLeNet is most convolutional among these.
- Performs the best





Classification

Table 1. Classification error on the ILSVRC validation set.

Networks	top-1 val. error	top-5 val. error
VGGnet-GAP	33.4	12.2
GoogLeNet-GAP	35.0	13.2
AlexNet*-GAP	44.9	20.9
AlexNet-GAP	51.1	26.3
GoogLeNet	31.9	11.3
VGGnet	31.2	11.4
AlexNet	42.6	19.5
NIN	41.9	19.6
GoogLeNet-GMP	35.6	13.9

Classification error:

- 1~2% performance drop on Top-5
- AlexNet-GAP drops with removal of FCs
- GooLeNet-GMP & GooLeNet-GAP is similar
- High classification performance is important to achieve good localization
  - as it involves identifying both object category and bounding box location

Localization

- Generation of Bounding Box from CAM.
  - thresholding to segment the heatmap
  - 20% of the max value of CAM, cover with bounding box.
    - on Top 5 predicted classes.

Table 2. Localization error on the ILSVRC validation set. *Back-prop* refers to using [22] for localization instead of CAM.

Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

GT bounding box

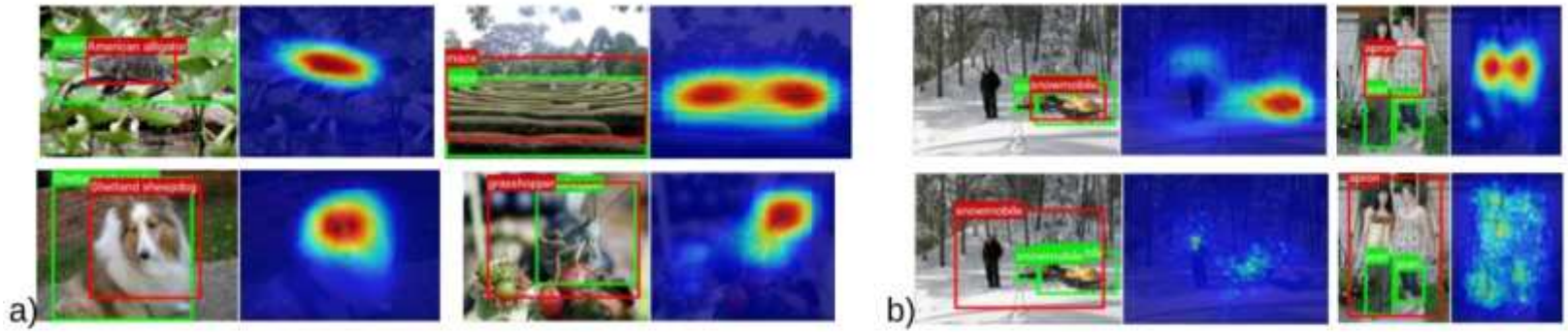


Figure 6. a) Examples of localization from GoogleNet-GAP. b) Comparison of the localization from GooleNet-GAP (upper two) and the backpropagation using AlexNet (lower two). The ground-truth boxes are in green and the predicted bounding boxes from the class activation map are in red.

Localization

Weakly vs. weakly

- **GoogLeNet-GAP** beats **GoogLeNet** and **GoogLeNet-GMP**
  - **Weakly-supervised!!**
  - Not trained on single annotated box.

Weakly vs. fully-supervised

- (heuristic)
  - loose bbox from top 3rd predicted class.
- **GoogLeNet-GAP is closed to AlexNet that learned supervised method.**
- Impressive but still behind fully-supervised.

Table 2. Localization error on the ILSVRC validation set. *Backprop* refers to using [22] for localization instead of CAM.

Method	top-1 val.error	top-5 val. error
GoogLeNet-GAP	**56.40**	**43.00**
VGGnet-GAP	57.20	45.14
GoogLeNet	60.09	49.34
AlexNet*-GAP	63.75	49.53
AlexNet-GAP	67.19	52.16
NIN	65.47	54.19
Backprop on GoogLeNet	61.31	50.55
Backprop on VGGnet	61.12	51.46
Backprop on AlexNet	65.17	52.64
GoogLeNet-GMP	57.78	45.26

Table 3. Localization error on the ILSVRC test set for various weakly- and fully- supervised methods.

Method	supervision	top-5 test error
GoogLeNet-GAP (heuristics)	weakly	**37.1**
GoogLeNet-GAP	weakly	42.9
Backprop [22]	weakly	46.4
GoogLeNet [24]	full	26.7
OverFeat [21]	full	29.9
AlexNet [24]	full	34.2

Deep Features for Classification

- Usage CNN features (fc6, fc7 of AlexNet) as **a generic feature**
  → Used as very powerful feature without training on the task.
- Use the features learned by GAP CNN. **Not trained for particular task!**
- Obtain weights by training a linear SVM on the output of GAP layer
  - similar to mapping weight to softmax layer
- Compare performance.

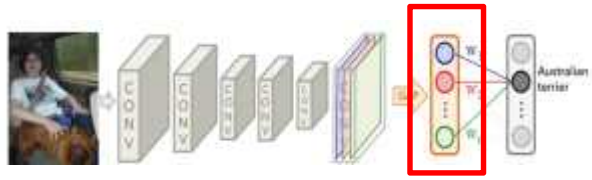


Table 5. Classification accuracy on representative scene and object datasets for different deep features.

	SUN397	MIT Indoor67	Scene15	SUN Attribute	Caltech101	Caltech256	Action40	Event8
`fc7` from AlexNet	42.61	56.79	84.23	84.23	87.22	67.23	54.92	94.42
`ave pool` from GoogLeNet	51.68	66.63	88.02	92.85	92.05	78.99	72.03	95.42
`gap` from GoogLeNet-GAP	51.31	66.61	88.30	92.21	91.98	78.07	70.62	95.00

- GoogLeNet-GAP outperforms AlexNet.
- GoogLeNet-GAP is competitive with the state-of-the-art as generic visual features.

Deep Features for Localization

- Also localizable with CAM
  - Very effective for generating localizable deep features for generic tasks.

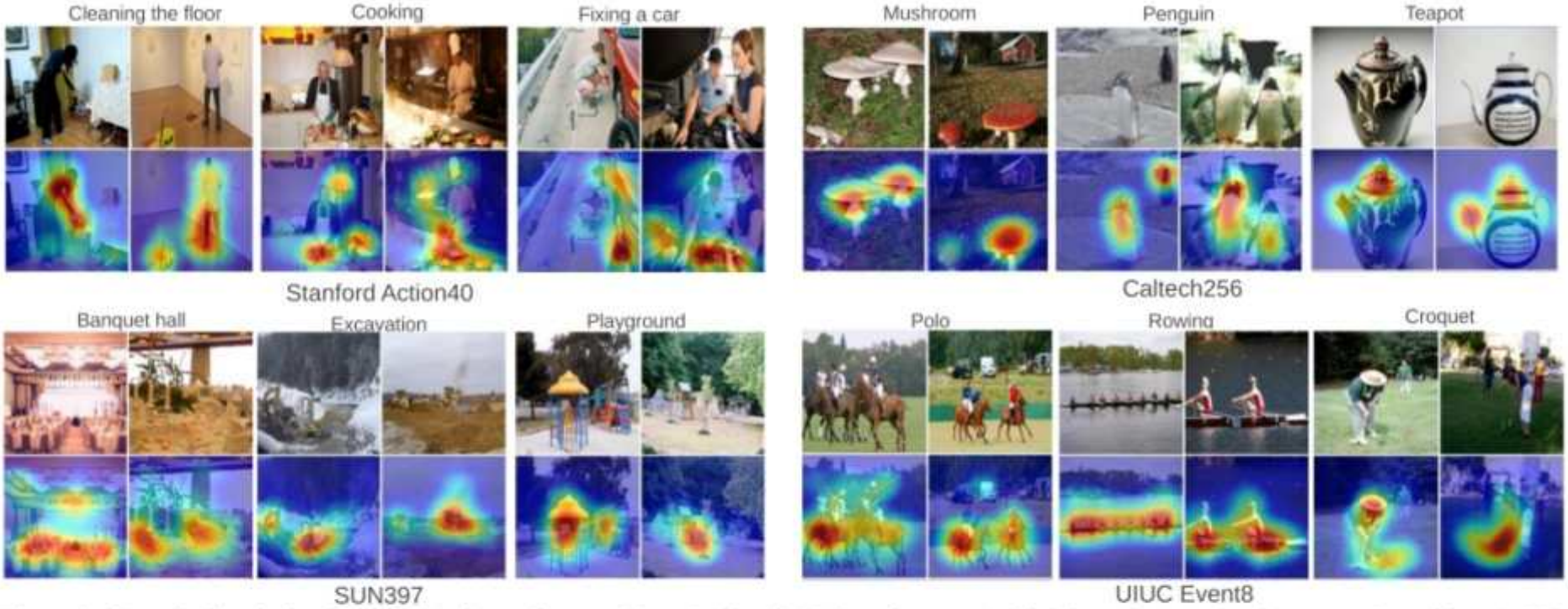


Figure 8. Generic discriminative localization using our GoogLeNet-GAP deep features (which have been trained to recognize objects). We show 2 images each from 3 classes for 4 datasets, and their class activation maps below them. We observe that the discriminative regions of the images are often highlighted e.g., in Stanford Action40, the mop is localized for *cleaning the floor*, while for *cooking* the pan and bowl are localized and similar observations can be made in other datasets. This demonstrates the generic localization ability of our deep features.

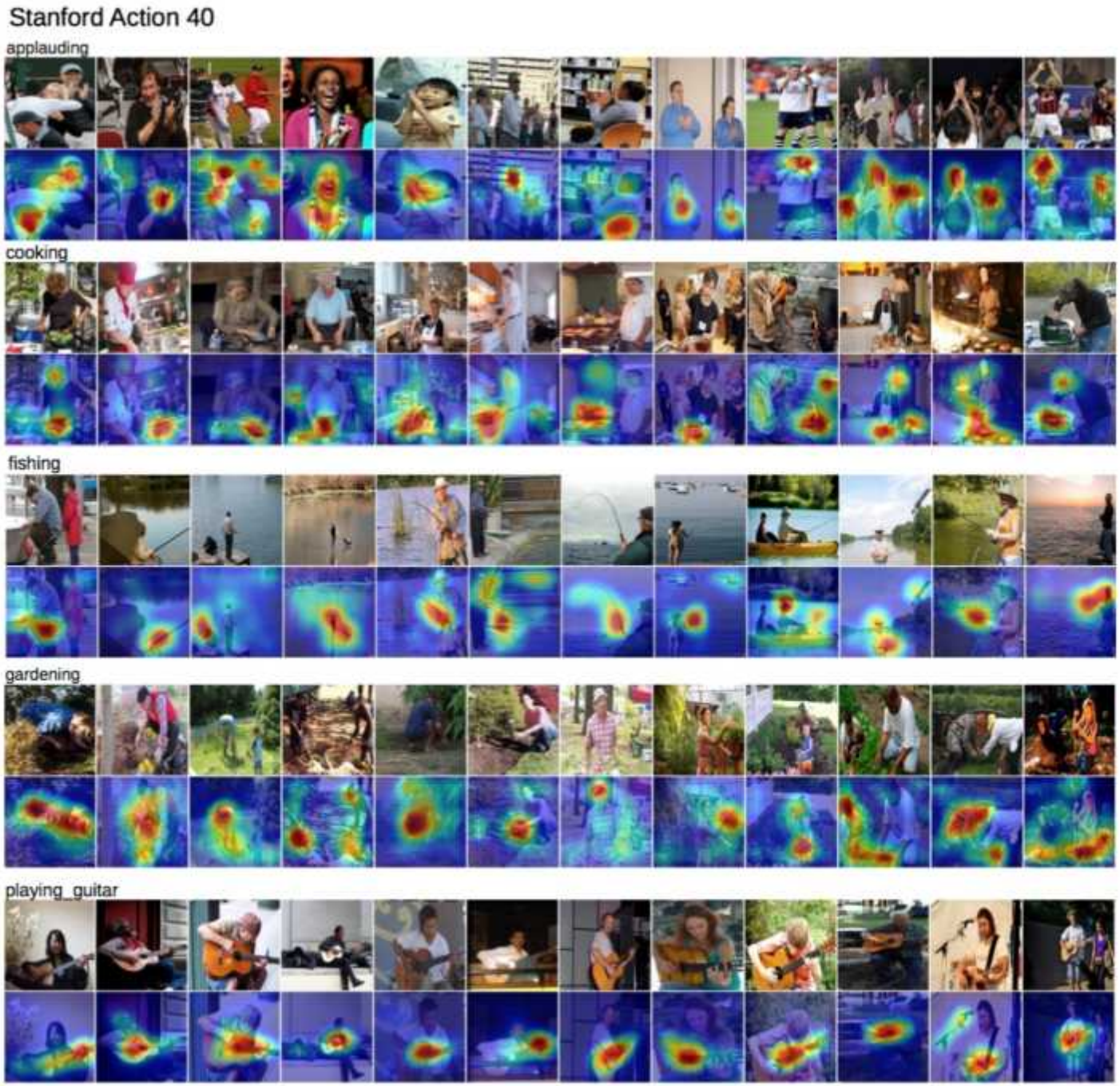


Figure 6. Examples of the class activation maps for 5 action classes from Stanford Action 40 dataset.

Scene Recognition + Localization

- 10 scene categories from SUN dataset. 4675 fully annotated images.
- one-vs-all linear SVM for each scene category.
- Compute the **CAMs** using the weights of the linear SVM.
- **Plot CAM for predicted scene category**
  - list top 6 objects that most frequently overlap
- High activations regions correspond to objects indicative of the particular scene category.

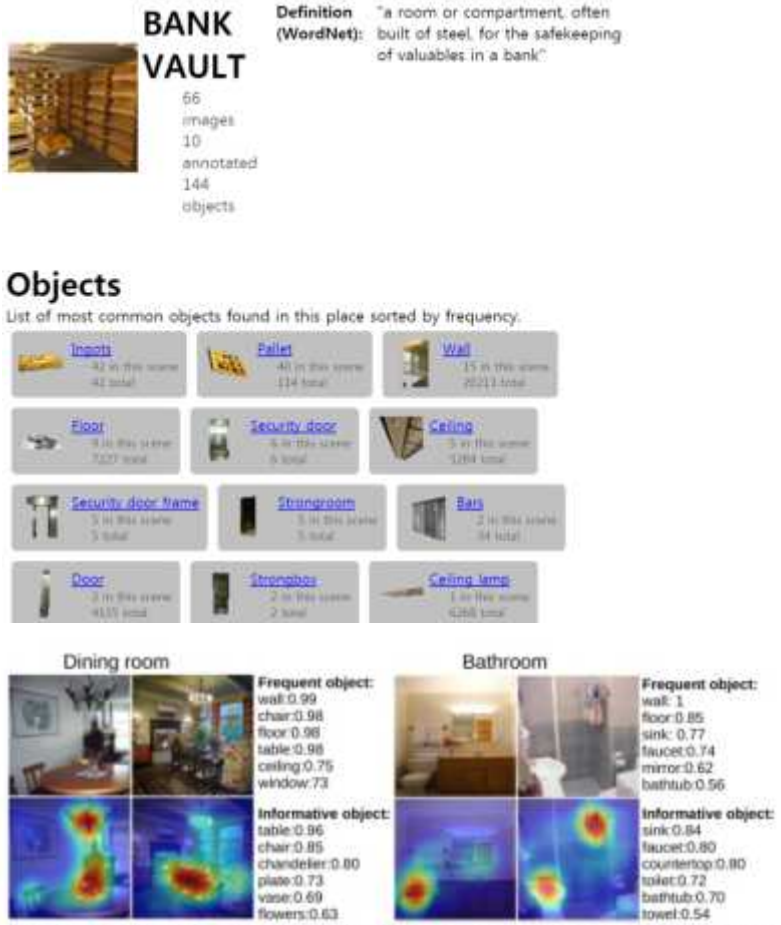


Figure 9. Informative objects for two scene categories. For the dining room and bathroom categories, we show examples of original images (top), and list of the 6 most frequent objects in that scene category with the corresponding frequency of appearance. At the bottom: the CAMs and a list of the 6 objects that most frequently overlap with the high activation regions.

Concept localization

Concept localization in weakly labeled images

- Positive set: short phrase in text caption
- Negative set: randomly selected images
- Model catch the concept, phrases are much more abstract than object name.

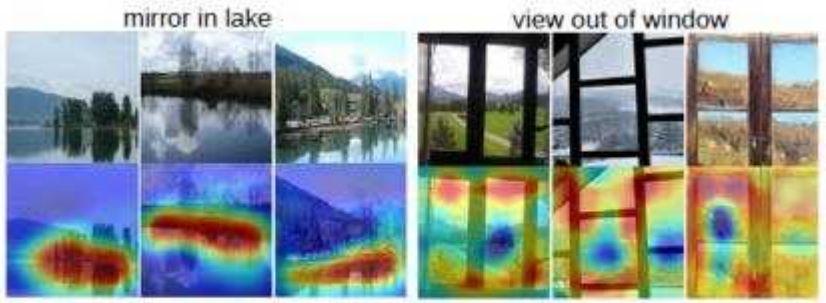


Figure 10. Informative regions for the concept learned from weakly labeled images. Despite being fairly abstract, the concepts are adequately localized by our GoogLeNet-GAP network.

Weakly supervised text detector

- Positive set: 350 Google Street View images that contain text.
- Negative set: outdoor scene images in SUN dataset
- Text highlighted without bounding box annotations.

Figure 11. Learning a weakly supervised text detector. The text is accurately detected on the image even though our network is not trained with text or any bounding box annotations.

Conclusion

- General technique called **Class Activation Mapping (CAM)** with Global Average Pooling has been proposed
- Classification-trained CNNs can learn object localization.
  - **without bounding box annotations.**
- CAM is useful to **visualize** the predicted class scores.
  - Highlights discriminative object parts detected by CNN.
- Weakly-supervised object localization on ILSVRC benchmark, and GAP-CNN can perform accurate object localization.
- CAM localization technique generalized to other visual recognition tasks.
  - Generic localizable deep features applicable to various tasks.
    - Import object in scene, concept localization, text detection.