Learning Deep Features for Discriminative Localization

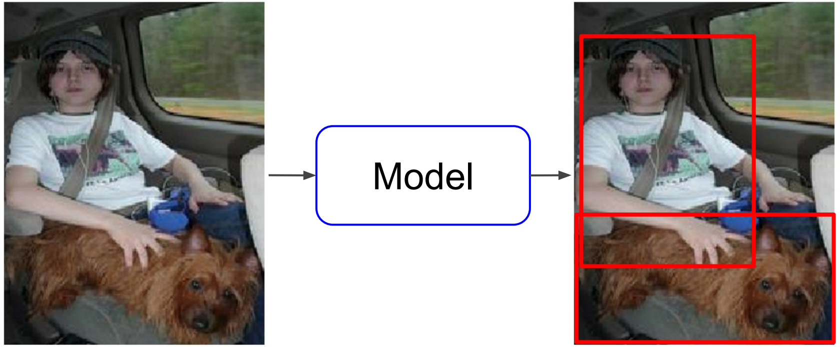
Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba

Presented by Marc Bolaños

Motivation

1) Weakly-Supervised Localization

- Need for an end-to-end model
- Fast inference at test time



2) CNN Visualization Techniques

- Easy to implement and understand CNN visualization

Cleaning the floor

Related Work

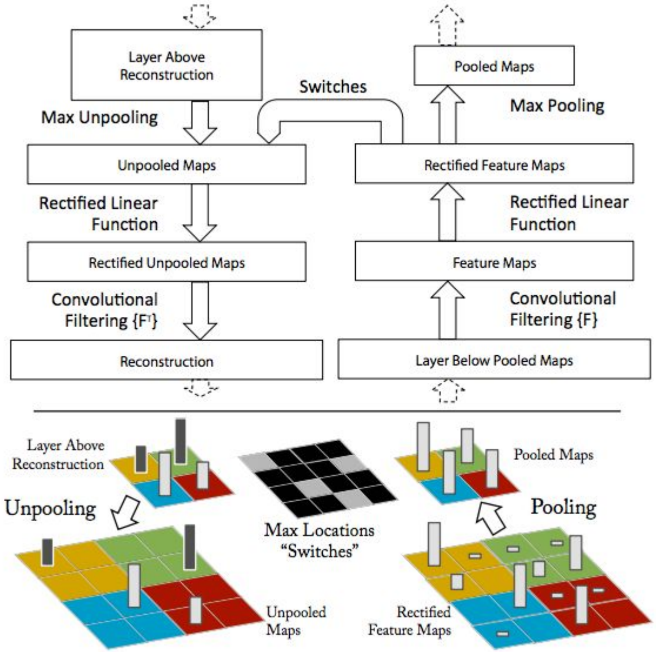
1) Weakly-Supervised Localization

Re-localization and refinement

Cinbis, R.G., Verbeek, J. and Schmid, C., 2015. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning. *arXiv preprint arXiv:1503.00949*.

Related Work

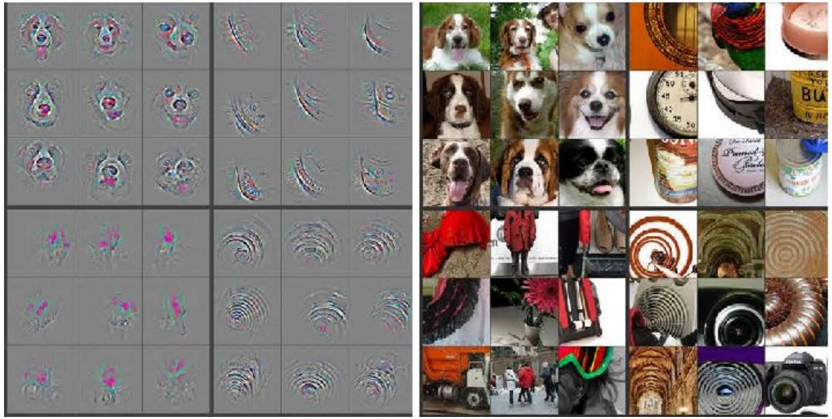
2) CNN Visualization Techniques

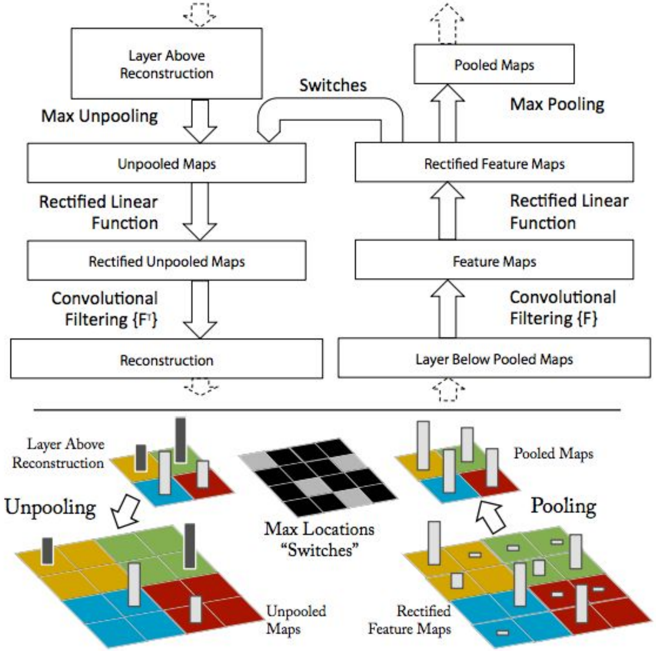


Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *Computer vision–ECCV 2014*. Springer International Publishing, 2014. 818-833.

Related Work

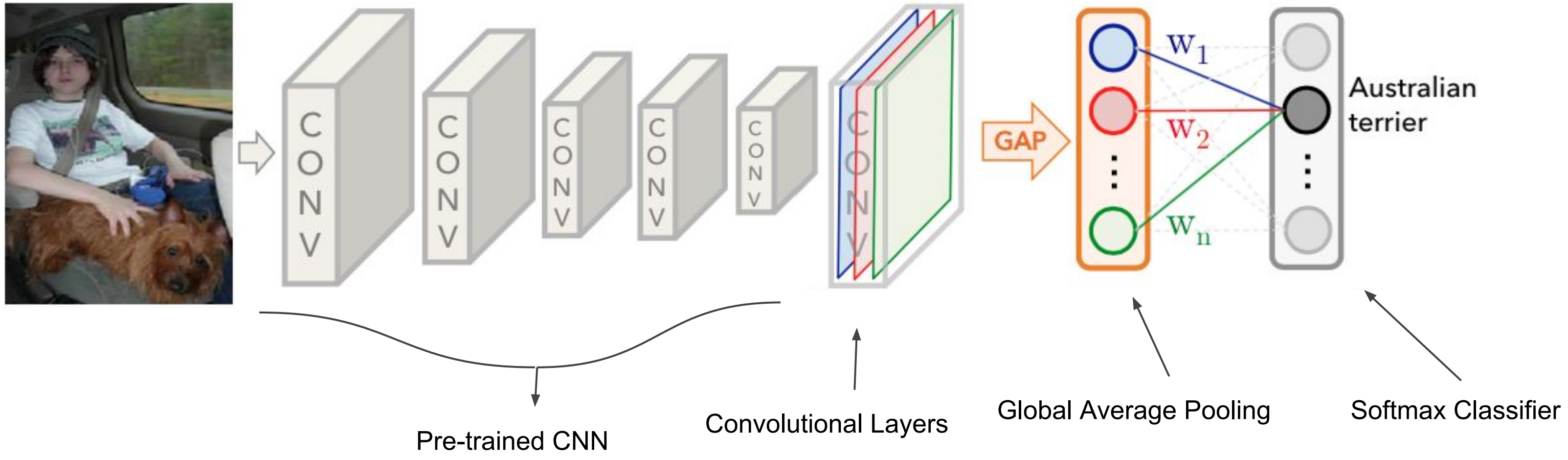
2) CNN Visualization Techniques



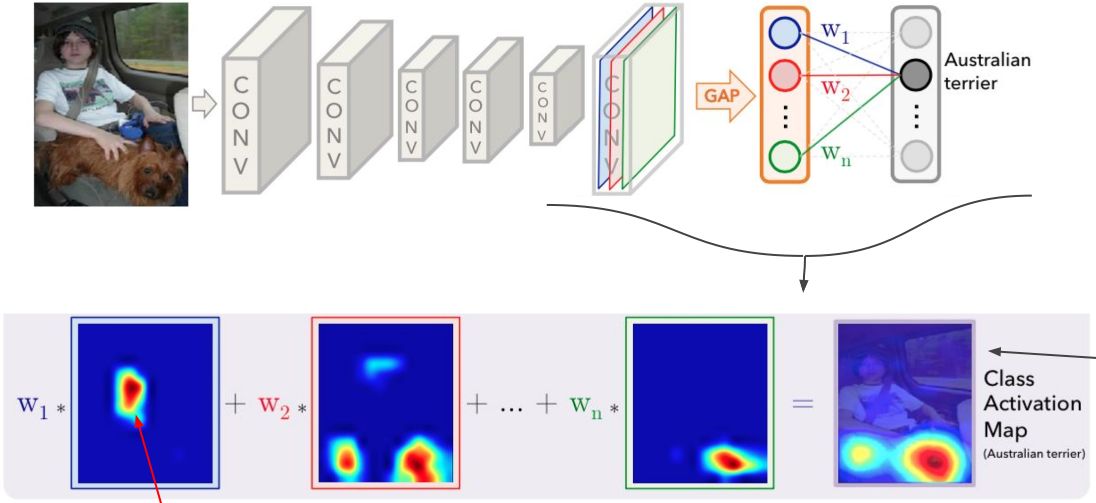


Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *Computer vision–ECCV 2014*. Springer International Publishing, 2014. 818-833.

Methodology - Global Average Pooling Model



Methodology - Class Activation Mapping



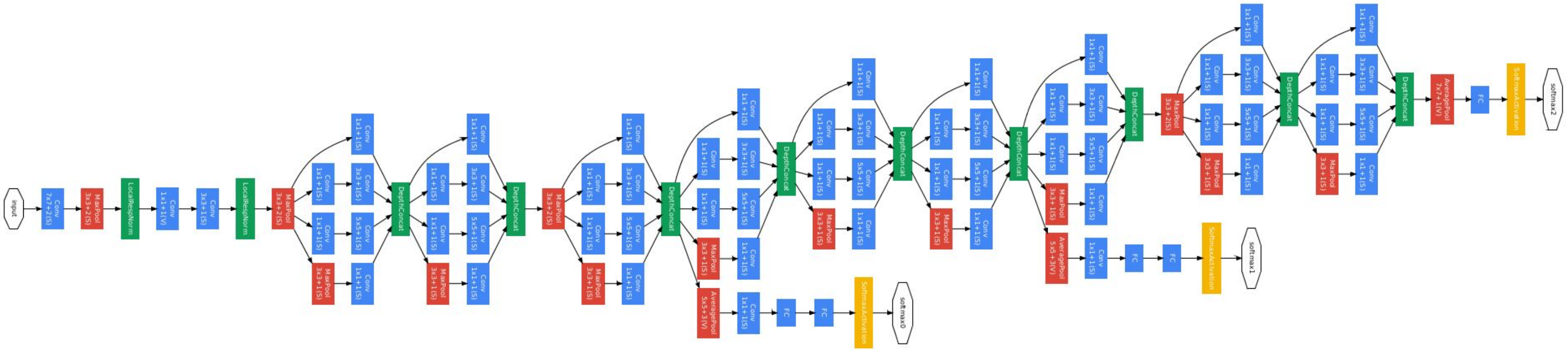
$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

CAM for class c

$f_k(x,y)$

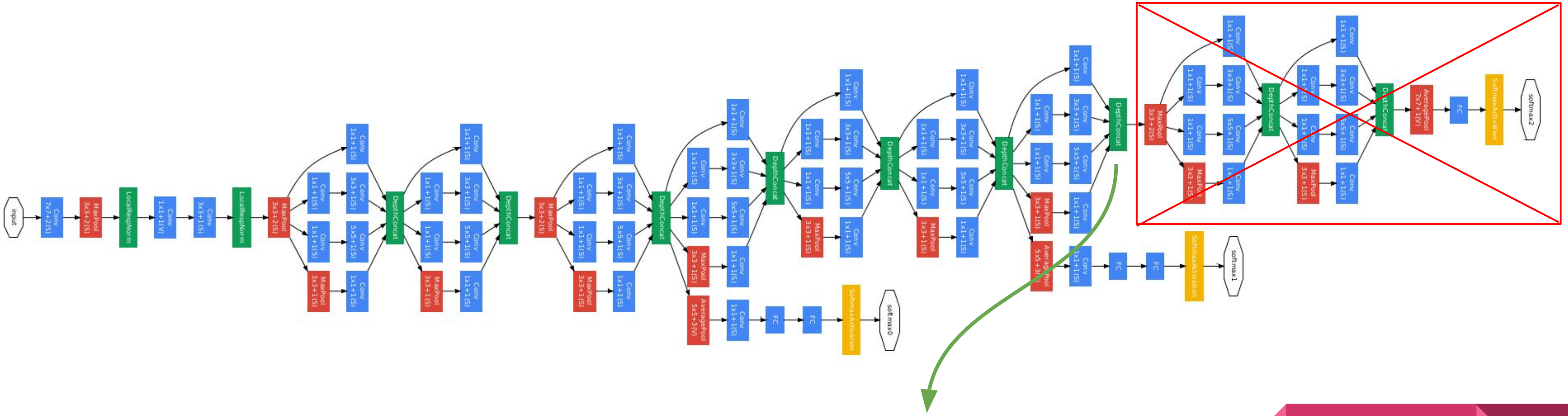
Methodology - Weakly-supervised Object Localization

GoogleNet Network Setup



Methodology - Weakly-supervised Object Localization

GoogleNet Network Setup

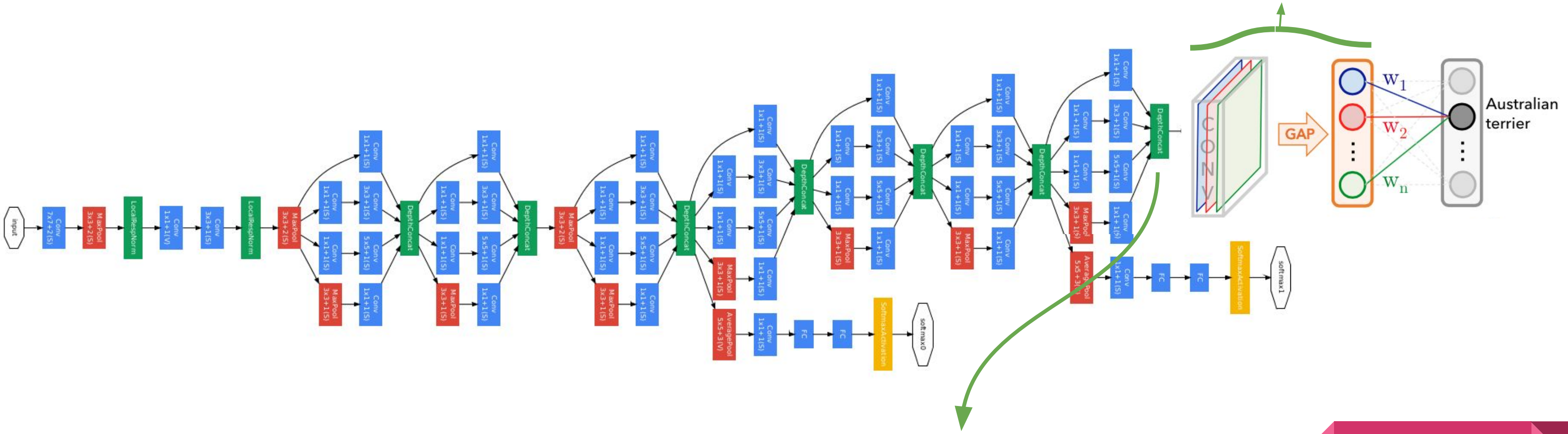


14x14 resolution

Methodology - Weakly-supervised Object Localization

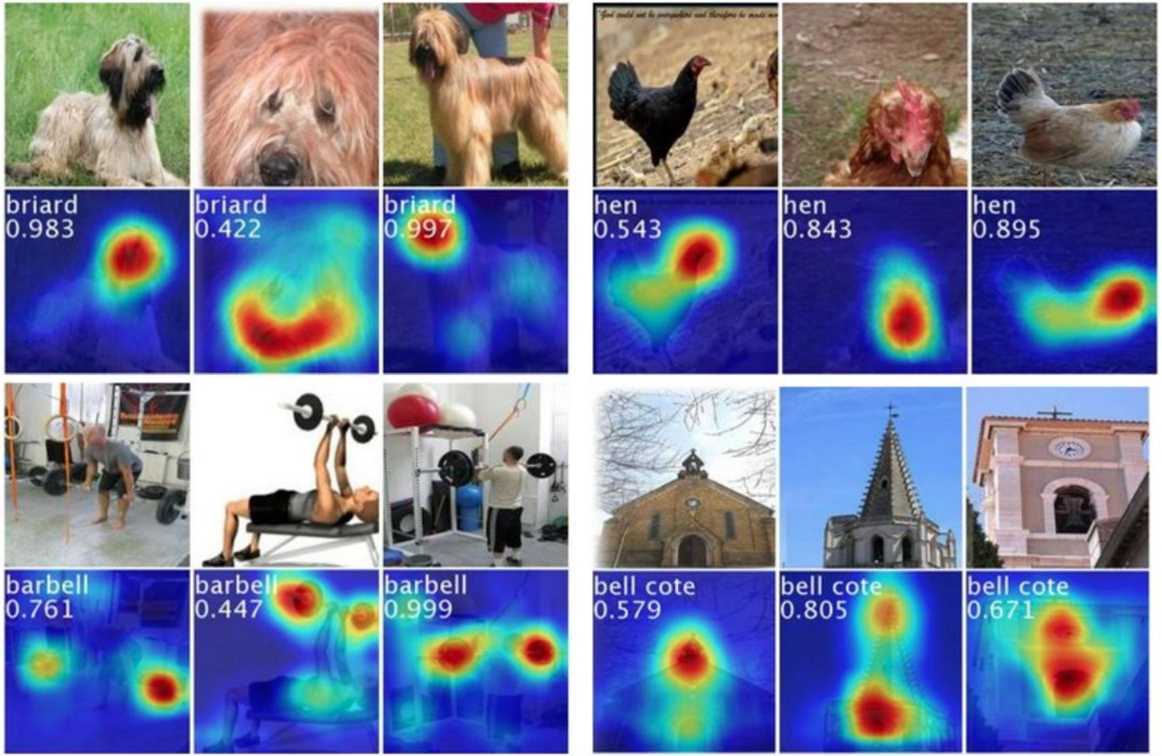
GoogleNet Network Setup

1024 filters/kernels

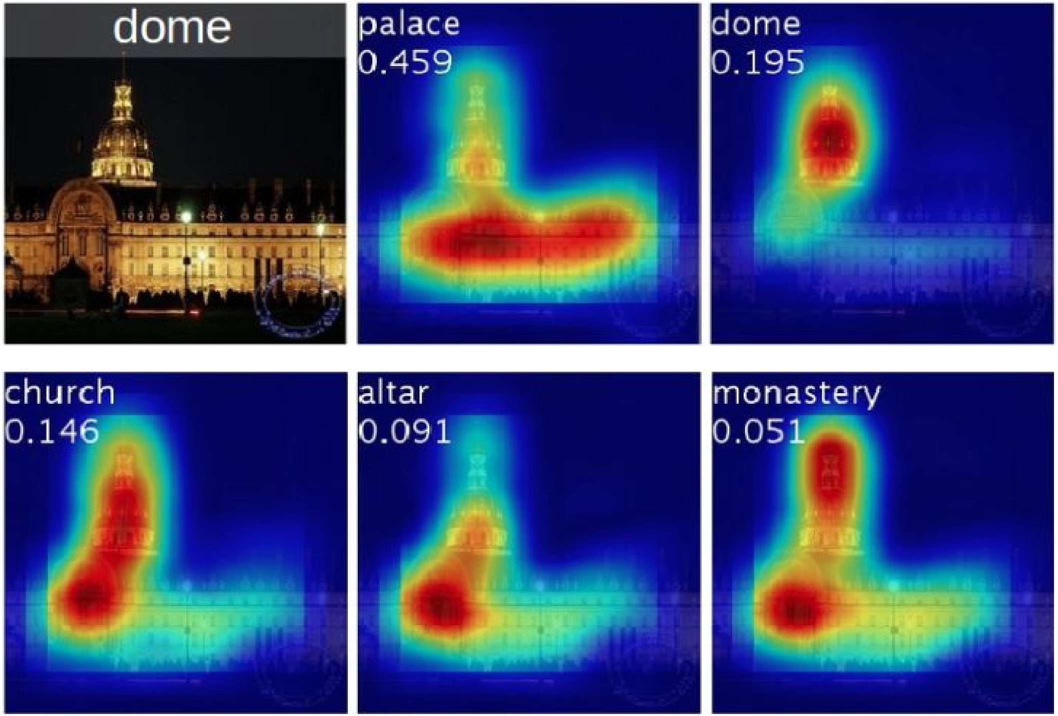


14x14 resolution

Results - Classification Examples



Results - Classification Examples



Results - Classification

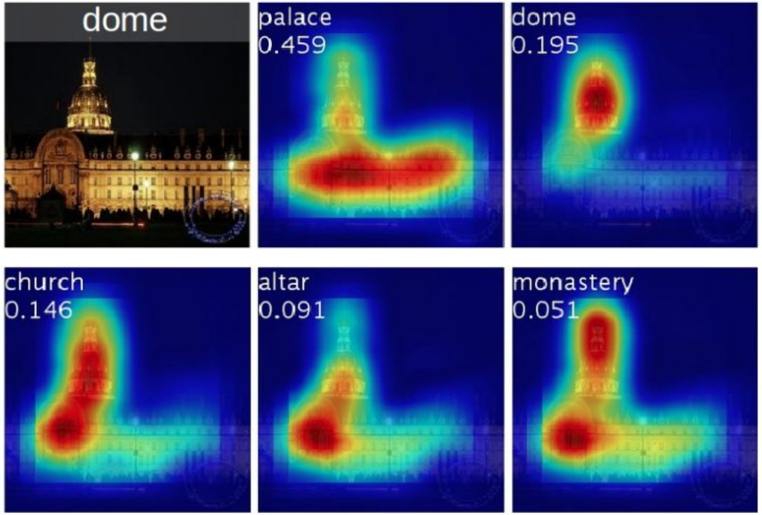
Table 1. Classification error on the ILSVRC validation set.

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

- Small performance drop during classification.
- GAP works slightly better than GMP during classification.

Methodology - Bounding Box Generation

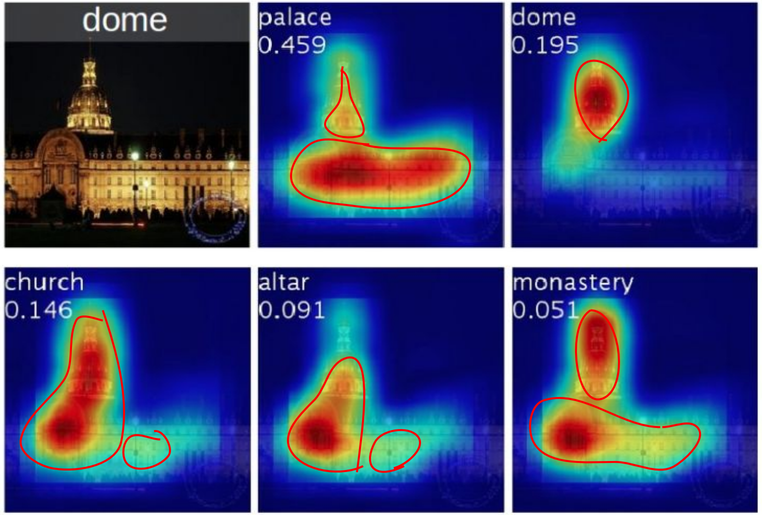
For each CAM on the top-5 predicted categories.



Methodology - Bounding Box Generation

For each CAM on the top-5 predicted categories.

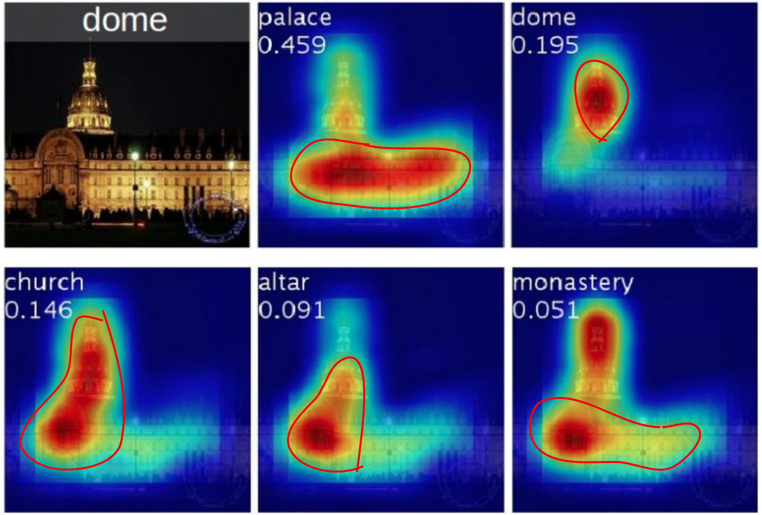
1) Threshold the heatmap. Select regions with values > 20% of max heat.



Methodology - Bounding Box Generation

For each CAM on the top-5 predicted categories.

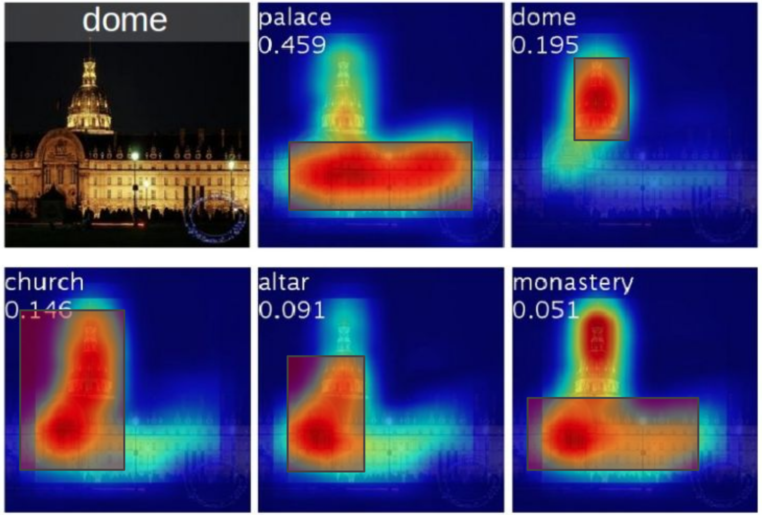
1) Threshold the heatmap. Select regions with values > 20% of max heat.
2) Get biggest connected component.



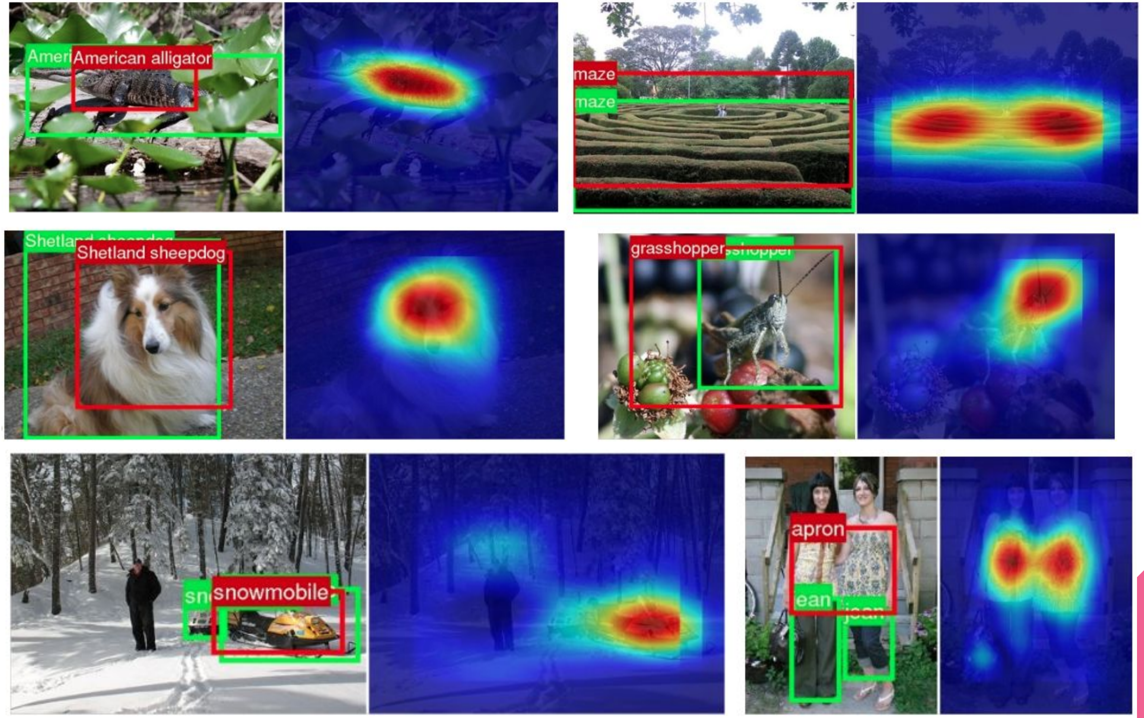
Methodology - Bounding Box Generation

For each CAM on the top-5 predicted categories.

1) Threshold the heatmap. Select regions with values > 20% of max heat.
2) Get biggest connected component.
3) Generate BBox covering the component



Results - Bounding Box Generation Examples



Results - Bounding Box Generation

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

- GAP works better than [22]
- GAP works slightly better than GMP during localization

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

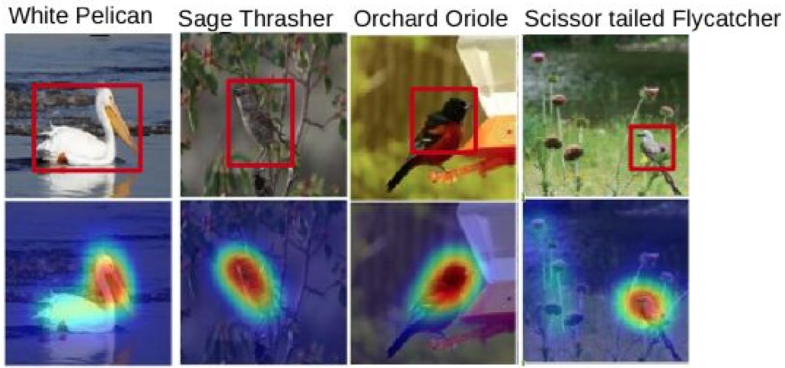
- GAP, although is weakly trained, gets closer to the full supervision methods

Heuristic: select two bounding boxes (one tight and one loose) from the top 1st and 2nd predicted classes and one loose from the top 3rd predicted class.

[22] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *International Conference on Learning Representations Workshop*, 2014.

Results - Fine-grained Recognition

Use a linear SVM on the extracted features



CUB200 Dataset, 200 bird classes

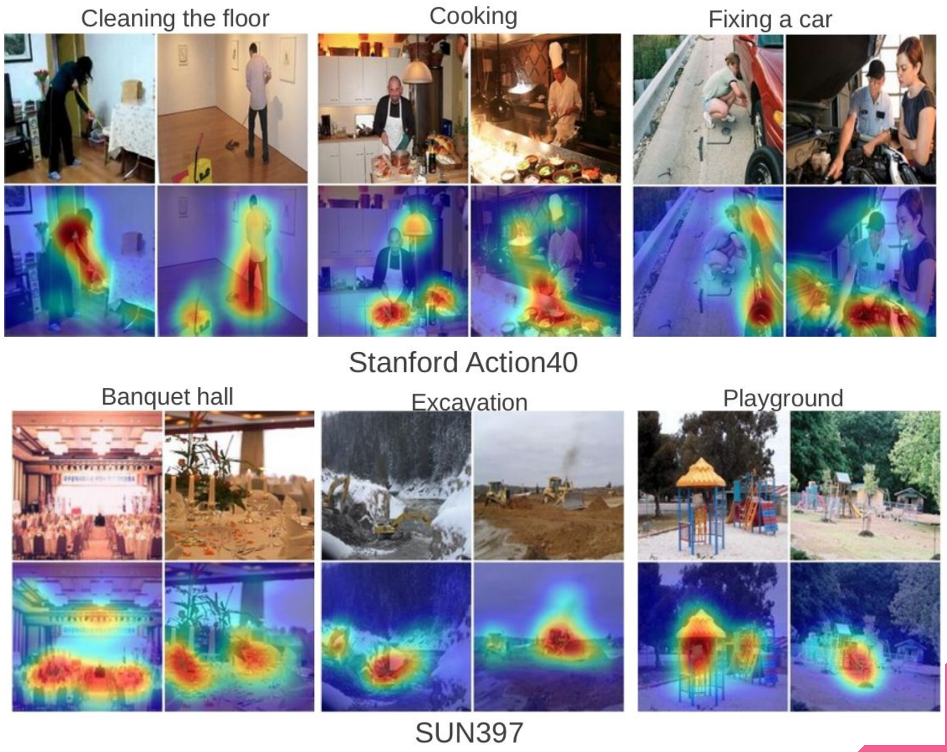
Table 4. Fine-grained classification performance on CUB200 dataset. GoogLeNet-GAP can successfully localize important image crops, boosting classification performance.

Methods	Train/Test Anno.	Accuracy
GoogLeNet-GAP on full image	n/a	63.0%
GoogLeNet-GAP on crop	n/a	67.8%
GoogLeNet-GAP on BBox	BBox	70.5%
Alignments [7]	n/a	53.6%
Alignments [7]	BBox	67.0%
DPD [31]	BBox+Parts	51.0%
DeCAF+DPD [3]	BBox+Parts	65.0%
PANDA R-CNN [30]	BBox+Parts	76.4%

- GAP weakly supervised gets comparable results (63.0%) to other methods
- Improves more (67.8%) if the localized crop is used again for classification
- Close to R-CNN (fully supervised) when using BBoxes for training (also fully supervised).

Results - Pattern Discovery

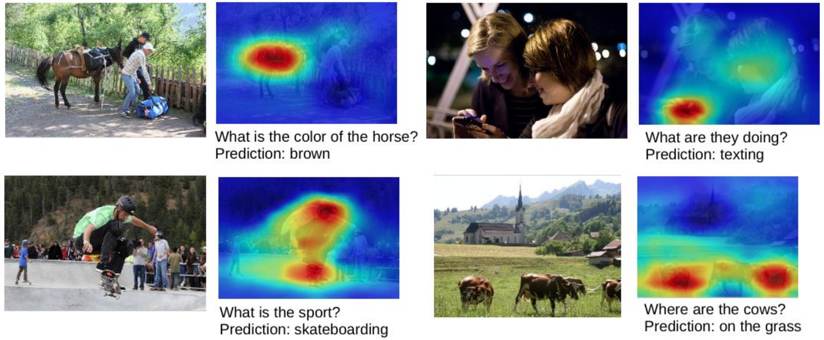
Use a linear SVM on the extracted features



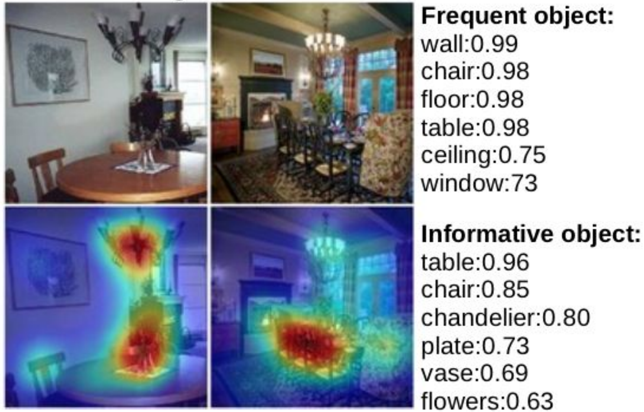
Results - Pattern Discovery

Text Detector

Question answering



Concept localization



Conclusions

- They propose a **simple yet effective weakly-supervised** object localization method (CAM).
- Easy to interpret **visualization technique**.
- Easy to use on the **top of a pre-trained CNN**.
- With potential uses on **several problems**:
    - Classification
    - Localization
    - Concept detection
    - Activity recognition
    - Text detection
    - Q&A