UF Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

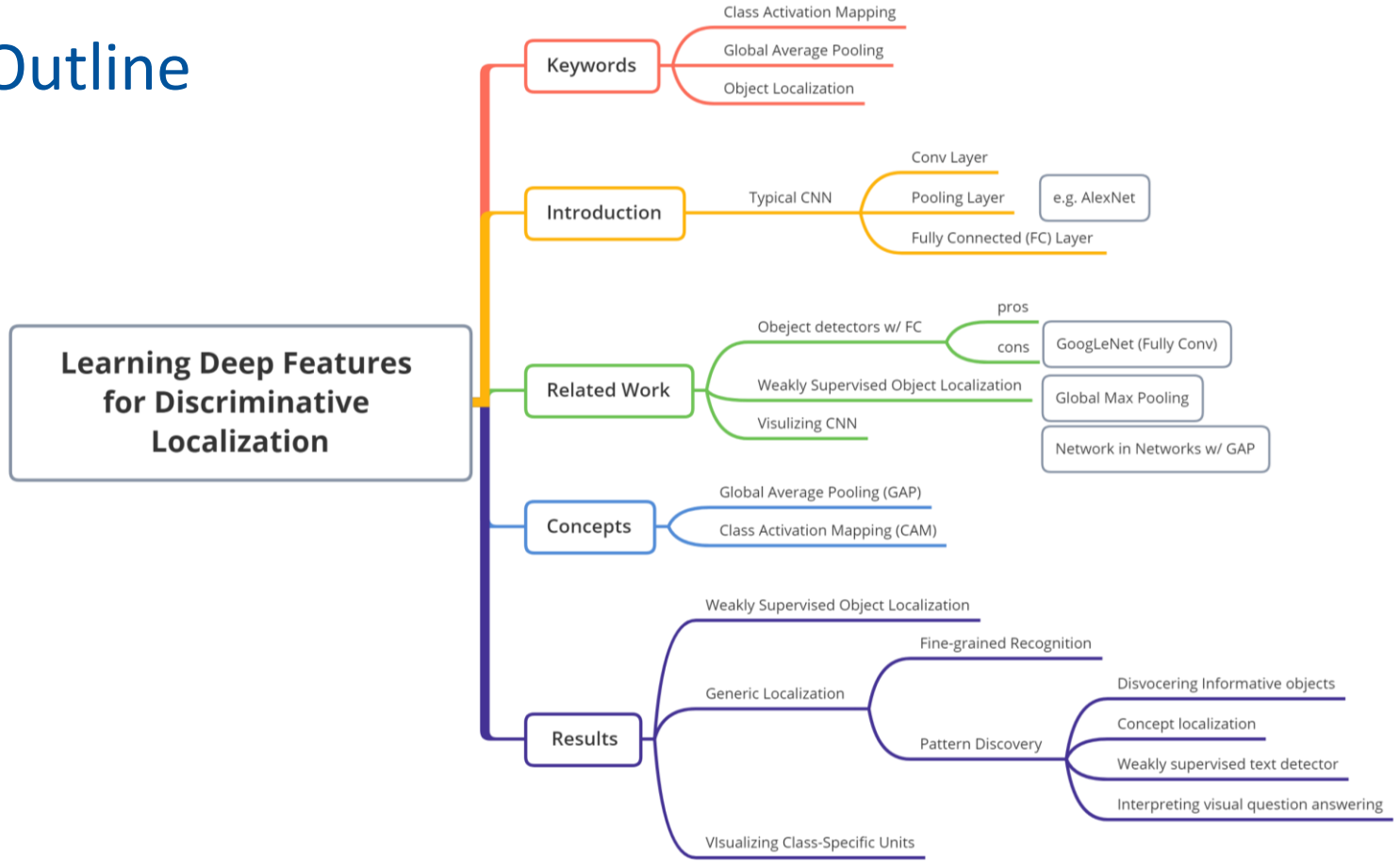
Learning Deep Features for Discriminative Localization

CVPR 2016

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

Presentor: Zhiyun Ling

Outline

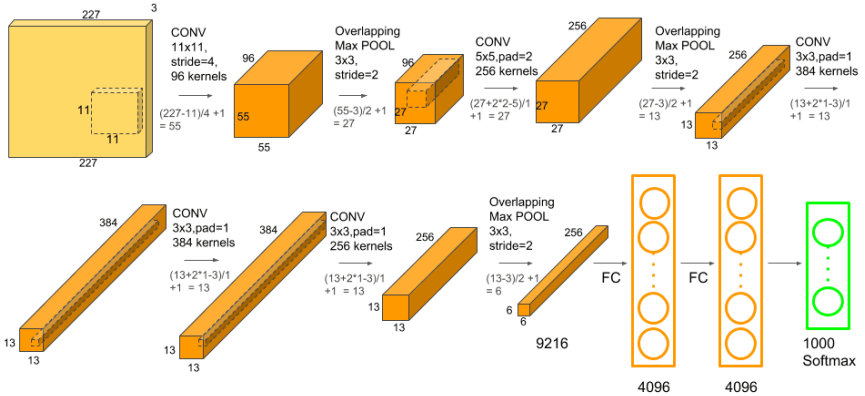

Learning Deep Features for Discriminative Localization
Keywords
Class Activation Mapping
Global Average Pooling
Object Localization
Introduction
Typical CNN
Conv Layer
Pooling Layer
Fully Connected (FC) Layer
e.g. AlexNet
Related Work
Obeject detectors w/ FC
pros
cons
GoogLeNet (Fully Conv)
Weakly Supervised Object Localization
Global Max Pooling
Visulizing CNN
Network in Networks w/ GAP
Concepts
Global Average Pooling (GAP)
Class Activation Mapping (CAM)
Results
Weakly Supervised Object Localization
Generic Localization
Fine-grained Recognition
Pattern Discovery
Disvocering Informative objects
Concept localization
Weakly supervised text detector
Interpreting visual question answering
VIsualizing Class-Specific Units

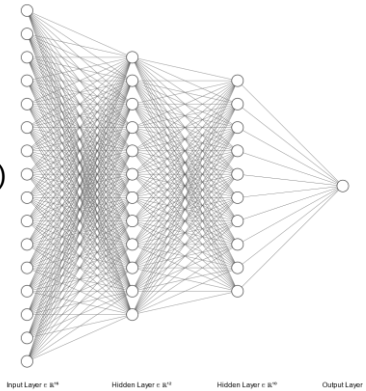
Introduction

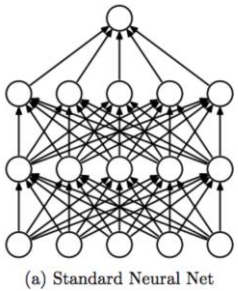
- CNNs are used in image classification.
  e.g. AlexNet
- Conv Layer: convolution, extract features

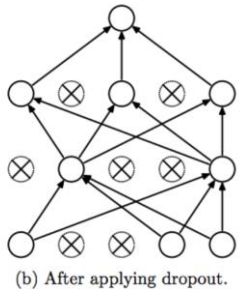
- Pooling
  down-sampling-> reduce parameter
  regularizer -> prevent overfitting
  Max vs Average
  Local vs Global

- Fully connected: summarize
  - Cons: overfitting! Localization!
  - Use Dropout (simple and effective, but only solved first)
  - Use Global Average Pooling









Related Work

- Convolutional units work as object detectors, despite no location supervision.[1]

- Fully Connected layer removes localization ability.
- Yet, we need a regularizer to 'sum' feature maps before classification, if not FC layer.

[1] Bolei Zhou. OBJECT DETECTORS EMERGE IN DEEP SCENE CNNS.2014. https://arxiv.org/pdf/1412.6856.pdf

Network In Network

- Use GAP to replace FC layer in CNNs as a regularizer.[2]

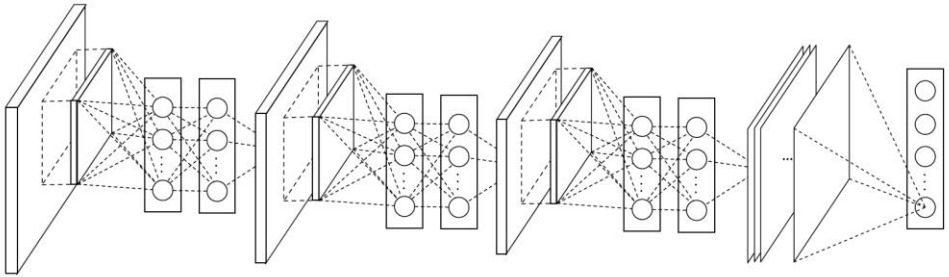


Figure 2: The overall structure of Network In Network. In this paper the NINs include the stacking of three mlpconv layers and one global average pooling layer.

- Generate one feature map for each corresponding category,
- Then take the average of each feature map, resulting in vector to classification.

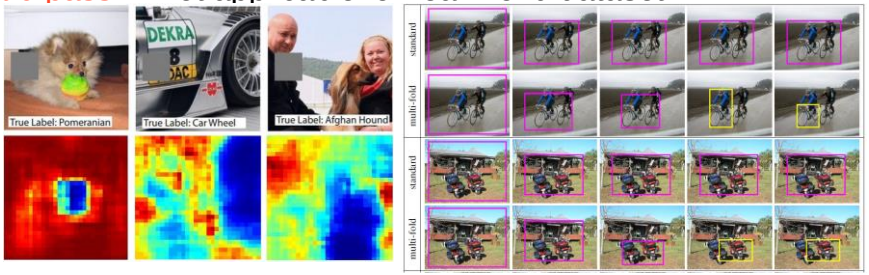
[2] Min Lin et al. Network In Network.2014. https://arxiv.org/pdf/1312.4400.pdf

Weakly-Supervised Object Localization

- One uses self-taught object localization by masking out image regions find the maximal activations .[3]
- Another combines multiple-instance-learning with CNN features to localize.[4]
- Oquab et al transfers mid-level image representations, localize object by evaluating the output of CNNs on several overlapping patchs.[5]
- Cons: Not trained end-to-end, requires multiple forward pass -> not applicable for real-world dataset.

- A similar study uses global max pooling GMP.
- Cons: limited to boundary points.

  Intuition: average considers all. VS max finds one.



- End-to-End Learning:
  - Omitting and hand-crafted intermediary algorithms, directly learning the solution for a given problem from the sampled dataset.

  E.g. OCR: instead of classifying and clustering them into words, directly use CNNs to do regression.

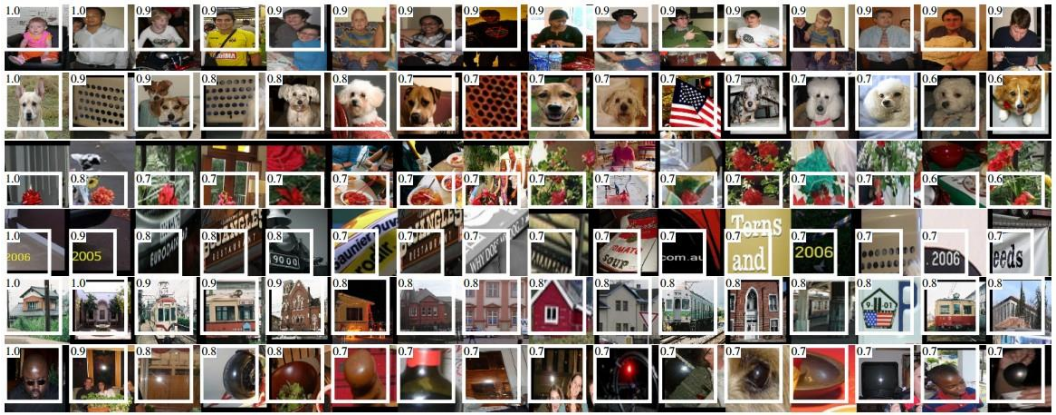
  Self Driving Car: directly train network to learn how to drive.

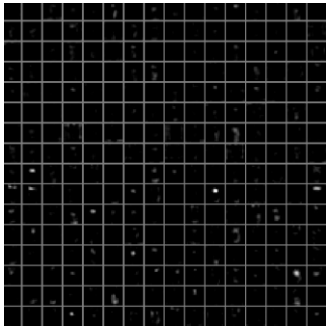
  Benefits: reduces human effort and performs better.

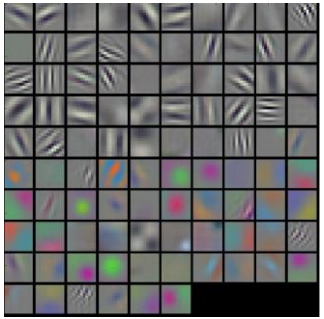
[3] Bergamo et al. Self-taught Object Localization with Deep Networks. 2014. https://arxiv.org/pdf/1409.3964
[4] Cinbis et al. Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning. 2015. https://arxiv.org/pdf/1503.00949
[5] Oquab et al. "Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks," CVPR2014

Visualizing CNN

- CNNs are magic. We need to see what's going on inside. Visualize!
- Many methods, all considered conv layers.
  Cons: FC layer is not considered -> incomplete
- Others [6] analyze the encoding of CNNs by inverting deep features at each layers.
  Applicable to FC but only focus on what information preserved, not the relative importance







[6] Mahendran et al. Understanding Deep Image Representations by Inverting Them. 2014. https://arxiv.org/pdf/1412.0035

Brief

- Propose: A simple modification of the GAP layer + Class Activation Mapping
- Localize the discriminative image regions using CAM in a single forward-pass
- Achieve 37.1% top-5 error for object localization on ILSVRC2014
  (compared to 34.2% top5 error achieved by fully supervised AlexNet)
- A generic localizable deep representation that can be applied to a variety tasks
  (transferred to other recognition datasets for generic classification, localization, concept discovery)

What is GAP

$$F_k = \sum_{x,y} f_k(x, y)$$

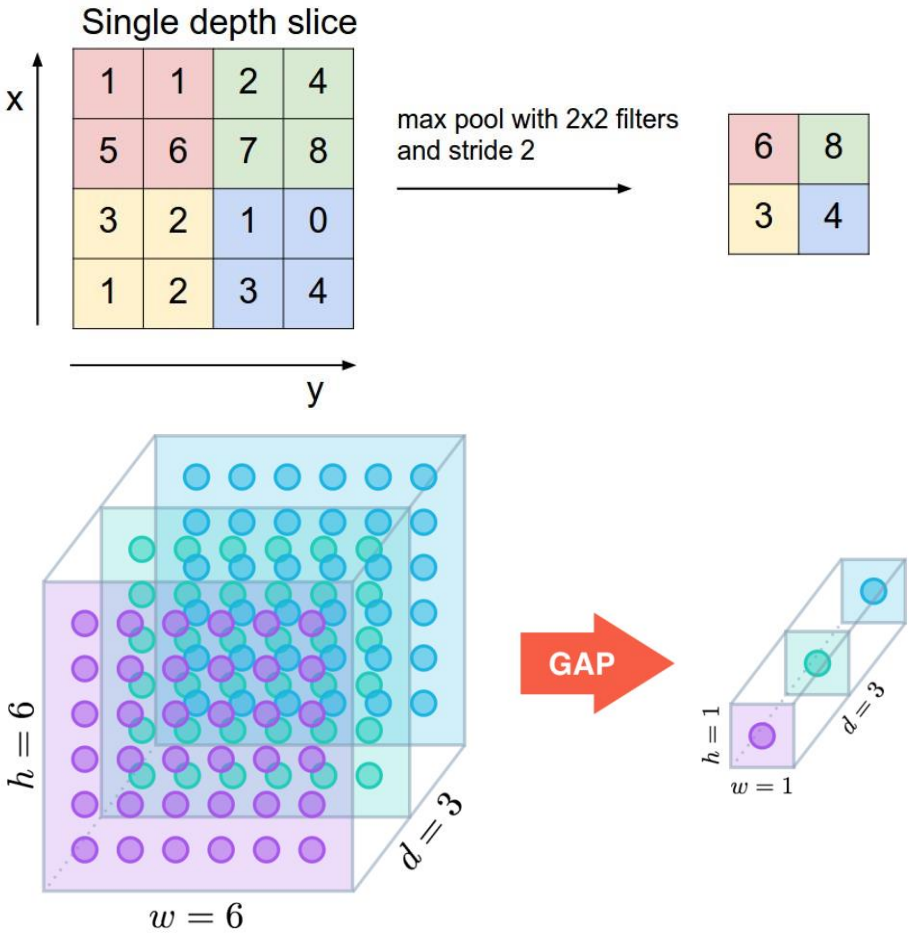
- Assume input: 3x6x6 tensor

- Local Pooling/ Pooling-> 3x3x3 tensor
- Benefits:
  - Translational tolerance
  - Computation power saving

- Global Pooling -> 3x1x1 tensor.
- Extreme but understandable as the end of CNNs

- More benefits:
  - No Overfitting since no parameter.
  - More robust to spatial translation
  - More native to convolution structure.

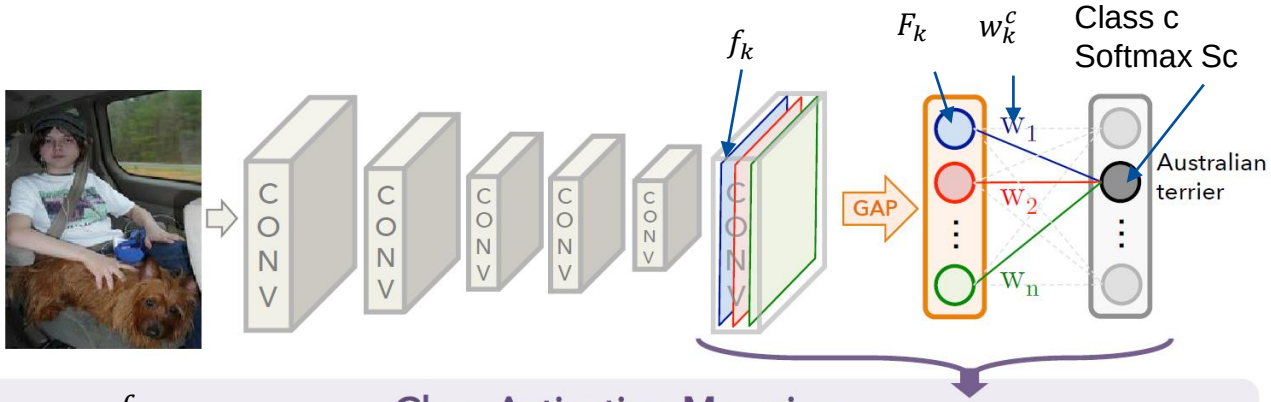
  (forcing connection between feature maps and catagories, can be easily interpreted as class confidence maps)



Class Activation Mapping CAM

- Indicates the discriminative image regions

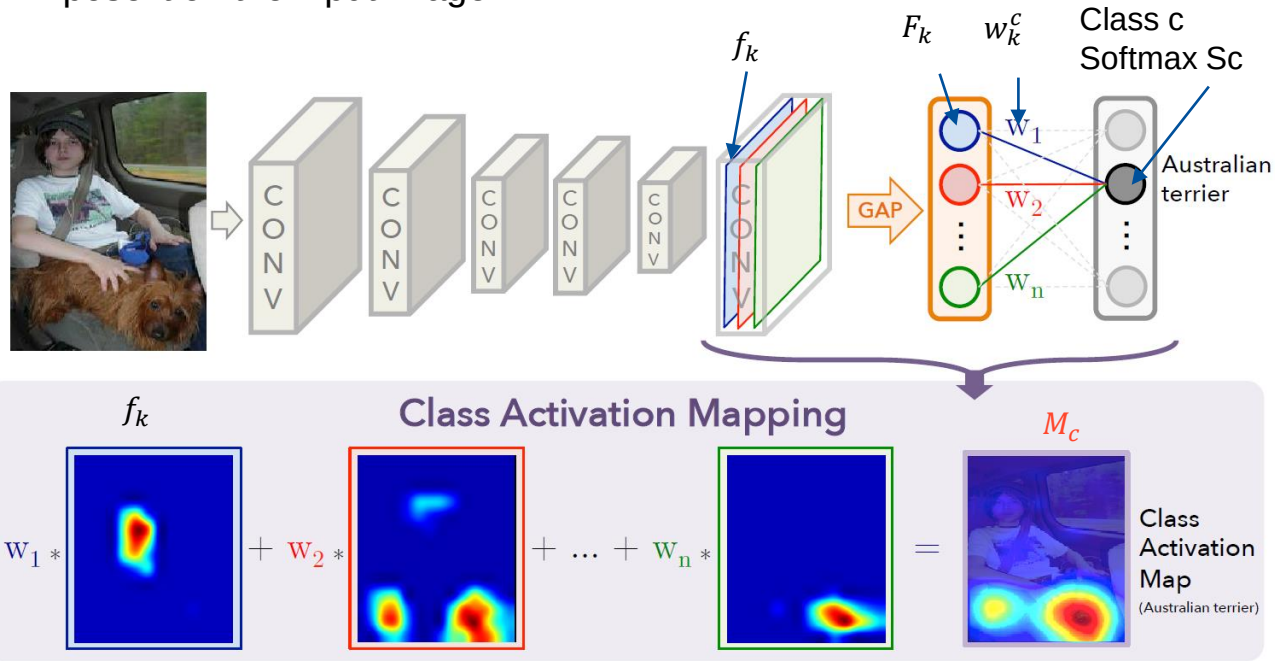
$$F_k = \sum_{x,y} f_k(x,y) \qquad S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \overbrace{\sum_k w_k^c f_k(x,y)}^{M_c}$$



M_c indicates the importance of the activation at spatial grid, Which contributes to the classification

Class Activation Mapping CAM

- 1. Get all weights between 'FC' and softmax output.
- 2. Then, do weighted sum (dot product) -> heatmap of class c
- 3. Impose it on the input image.

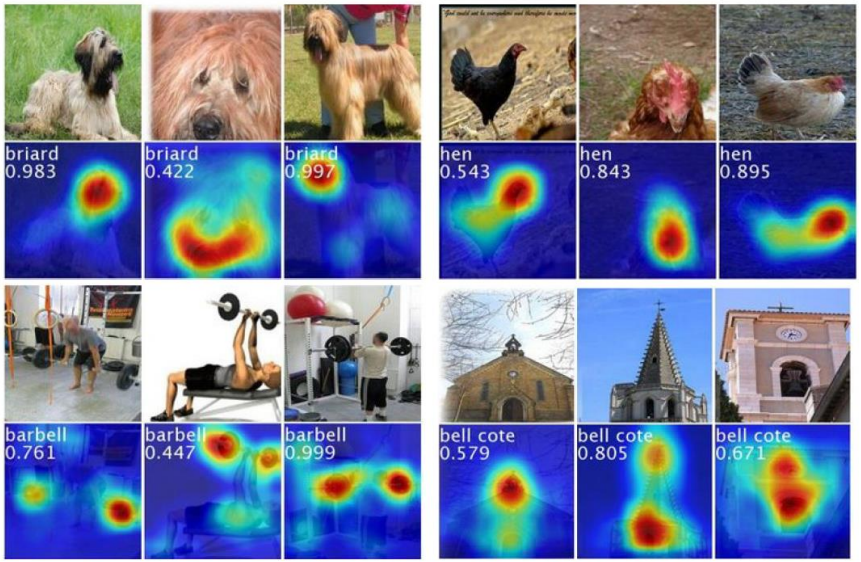


Intuition:

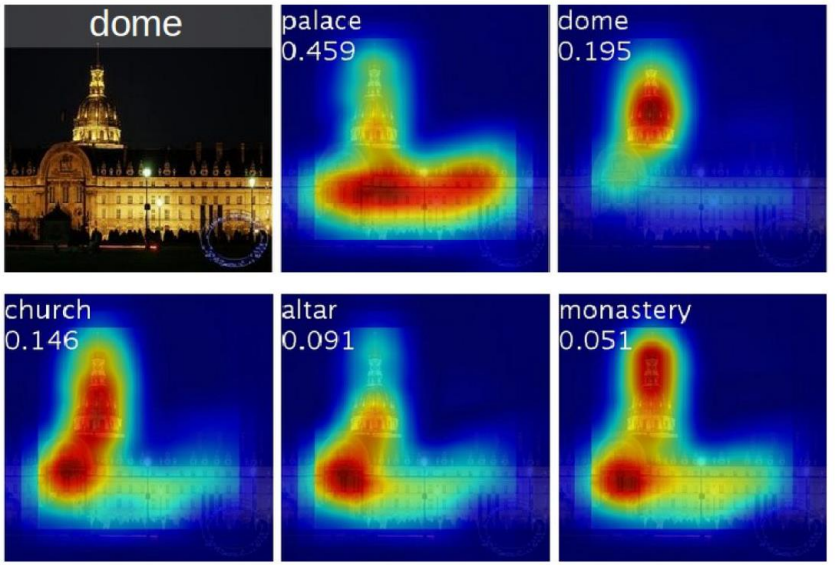
Each unit is activated by certain pattern, CAM is the weighted sum of the presence of these patterns.

Class Activation Mapping CAM

- Same Class Different Images



Same image Different Classes



Weakly-Supervised Object Localization

- **Setup**
  - AlexNet, VGGnet, GoogLeNet -> replace FC layers by GAP + softmax
- Observation: Higher Localization ability when last conv layer has higher spatial resolution
- Modify: remove layers,
  - AlexNet-GAP: after conv5. -> resolution 13x13
  - VGGnet-GAP: after conv5-3. -> resolution 14x14
  - GoogLeNet-GAP: after inception4e. -> resolution 14x14

  Add conv layer 3x3, stride=1, pad=1 with 1024 units, followed by GAP + softmax.
  Finetuned on ILSVRC 1000 classes.
- Classification:
  - Compare against original CNNs, including NIN.
- Localization:
  - Compare against GoogLeNet, NIN and using backpropagation.
  - Compare GMP vs GAP using GoogLeNet.
- Metric: top-1 and top-5 error.

Weakly-Supervised Object Localization

- **Classification Results**
  - A small performance drop of 1~2% when removing conv layer.
  - GMP and GAP have similar performance

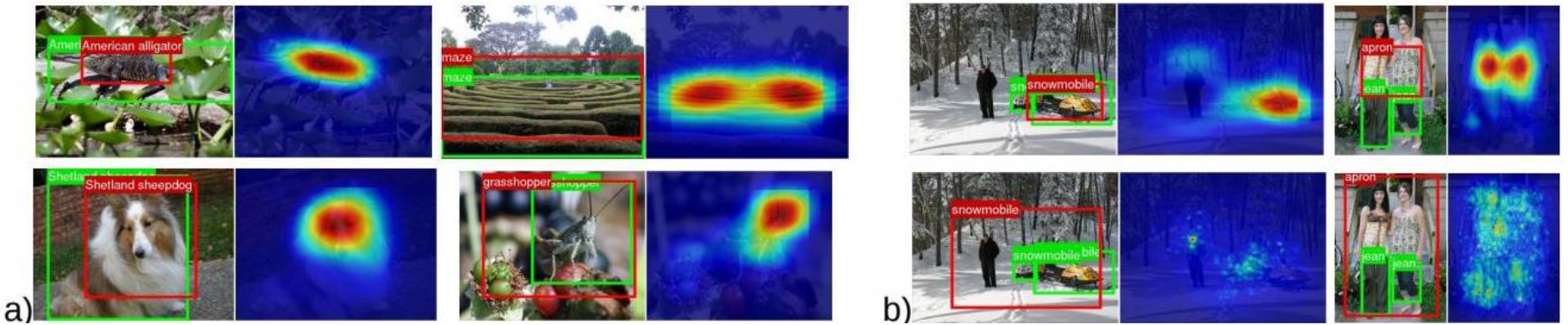
Table 1. Classification error on the ILSVRC validation set.

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

Weakly-Supervised Object Localization

- **Localization Results**
  - bounding box: thresholding 20% of max in CAM, covers largest connected component (a)

  (b) GoogLeNet-GAP vs backpropagation with AlexNet



Weakly-Supervised Object Localization

- **Localization Results**

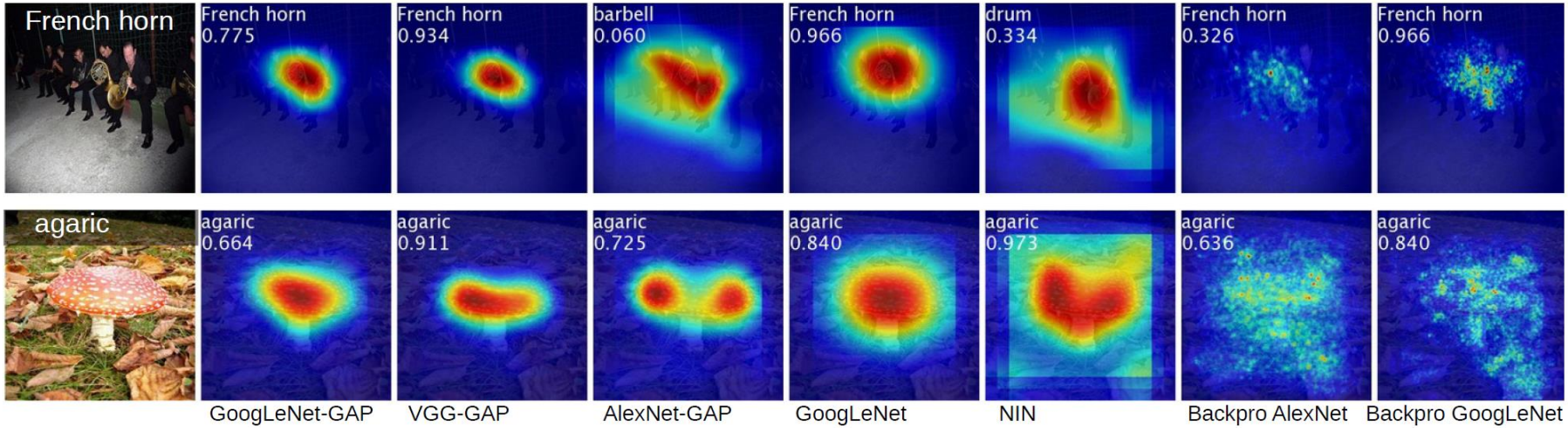


Figure 5. Class activation maps from CNN-GAPs and the class-specific saliency map from the backpropagation methods.

Weakly-Supervised Object Localization

- **Localization Results**
  - GAP outperforms all baseline networks. Remarkable!
  - GAP outperforms GMP

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

Weakly-Supervised Object Localization

- **Localization Results**
  - One tight (top-1st 2nd) and one lose(top-3rd) bounding boxes.

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

Tight+lose

Impressively close to Fully supervised result

Deep Features for Generic Localization

- **Higher-level layer + linear SVM**
  - GAP-CNNs performs well as generic features.
  - Plus, identify discriminative regions despite not being trained for
  - Use SVM after GAP.
- Compared against several classification benchmarks.

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 and GoogLeNet significantly outperforms AlexNet.
- Fewer conv layers. Competitive with the state-of-art.

Deep Features for Generic Localization

- **Higher-level layer + linear SVM**

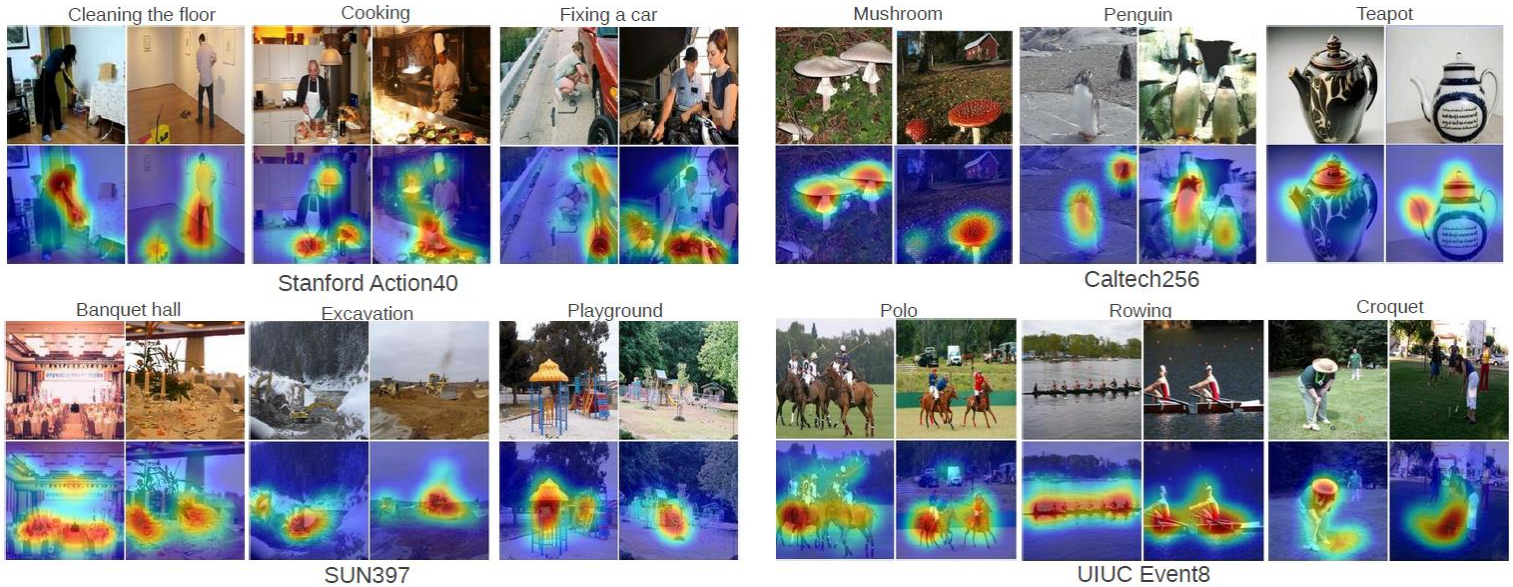


Figure 8. Generic discriminative localization using our GoogLeNet-GAP deep features (which have been trained to recognize objects). We show 2 images each from 3 classes for 4 datasets, and their class activation maps below them. We observe that the discriminative regions of the images are often highlighted 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.

Deep Features for Generic Localization

- **Fine-grained Recognition**
  - Dataset: 200 bird species, 11788 images(5994+5794), Bounding box anotation.

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%

GoogLeNet-GAP performs comparably to existing ones.
Intuition: A focused sub-image allows for better discrimination.

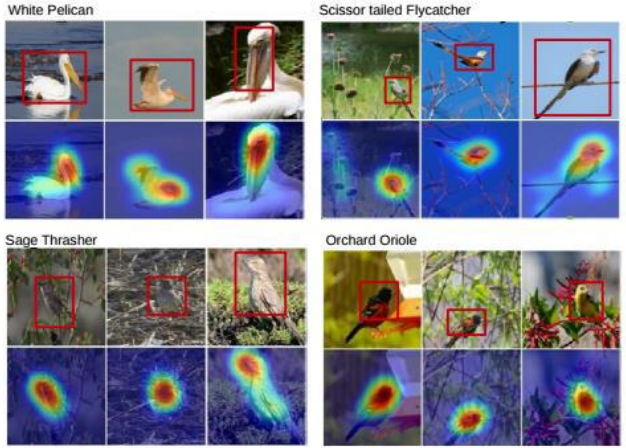
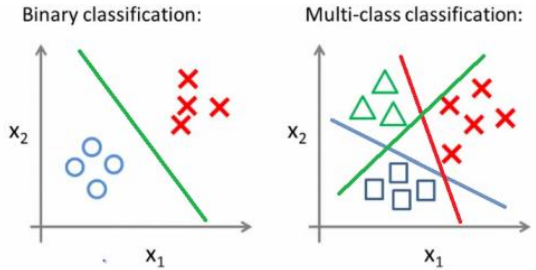


Figure 7. CAMs and the inferred bounding boxes (in red) for selected images from four bird categories in CUB200. In Sec. 4.1 we quantitatively evaluate the quality of the bounding boxes (41.0% accuracy for 0.5 IoU). We find that extracting GoogLeNet-GAP features in these CAM bounding boxes and re-training the SVM improves bird classification accuracy by about 5% (Tbl. 4).

Deep Features for Generic Localization

- **Pattern Discovery**
  - Identify common patterns in images beyond objects, such as text or high-level concepts.
- Given images containing a same concept -> find important regions/patterns
  - 1. train a linear SVM on the GAP layer.
  - 2. apply CAM technique to find important regions
- Discovering informative objects in the scenes
  - 10 scene categories from SUN dataset (4675 annotated images)
  - GAP layer + one-vs-all linear SVM for each scene category



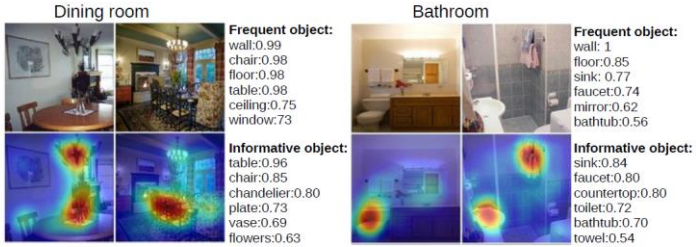


Figure 9. Informative objects for two scene categories. For the dining room and bathroom categories, we show examples of original images (top), and list of the 6 most frequent objects in that scene category with the corresponding frequency of appearance. At the bottom: the CAMs and a list of the 6 objects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

Deep Features for Generic Localization

- **Pattern Discovery**
- Concept Localization in weakly labeled images
  - Using hard-negative-mining algorithm -> concept detector + localize concept.

Train a concept detector for a short phrase:
1. Positive set: images contains it in caption
2. Negative set: random images without relevant words in caption

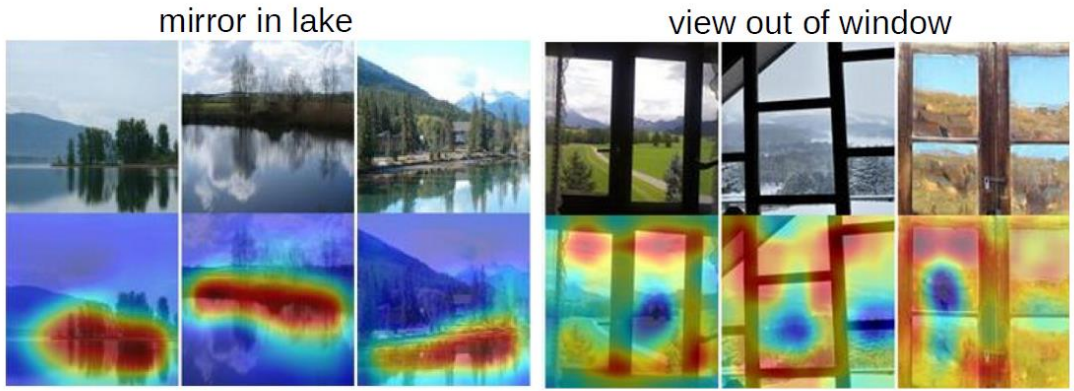


Figure 10. Informative regions for the concept learned from weakly labeled images. Despite being fairly abstract, the concepts are adequately localized by our GoogLeNet-GAP network.

Deep Features for Generic Localization

- **Pattern Discovery**
- Weakly supervised text detector
  - Using 350 Google StreetView Images as positive set
  - Using randomly sampled images in the SUN dataset as negative set

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.

Deep Features for Generic Localization

- **Pattern Discovery**
- Interpreting visual question answering
  - Overall 55.89% accuracy

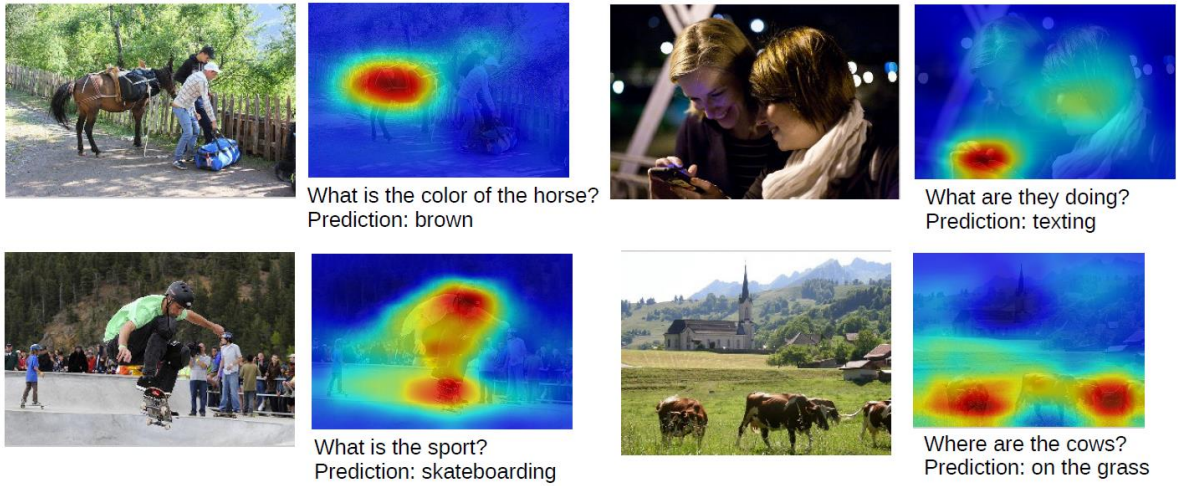


Figure 12. Examples of highlighted image regions for the predicted answer class in the visual question answering.

Visualizing Class-Specific Units

- Units that are most discriminative for a given class
- Using GAP and ranked softmax weight.
- Observation: CNN learns a bag of words, where each is a discriminative class-specific unit.

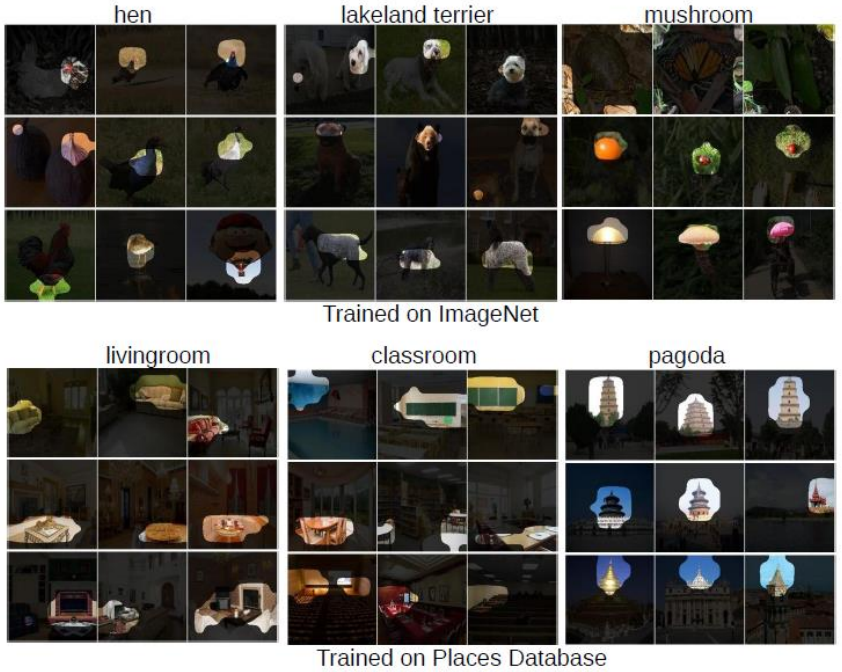


Figure 13. Visualization of the class-specific units for AlexNet*-GAP trained on ImageNet (top) and Places (bottom) respectively. The top 3 units for three selected classes are shown for each dataset. Each row shows the most confident images segmented by the receptive field of that unit. For example, units detecting blackboard, chairs, and tables are important to the classification of *classroom* for the network trained for scene recognition.

Conclusion

- Class Activation Mapping with Global Average Pooling are proposed as a way to visualize CNN.
- Trained with classification, but able to do object localization.
- Generalized to other visual recognition tasks.
- Limitations:
  - Need GAP layer to do CAM
  - Can only visualize the final layer heatmap
- Grad-CAM deals with these limitations.

GradCAM Source: https://arxiv.org/pdf/1610.02391.pdf
GradCAM++ Source: https://arxiv.org/pdf/1710.11063.pdf

Q & A

UF Herbert Wertheim College of Engineering
UNIVERSITY of FLORIDA

Learning Deep Features for Discriminative Localization

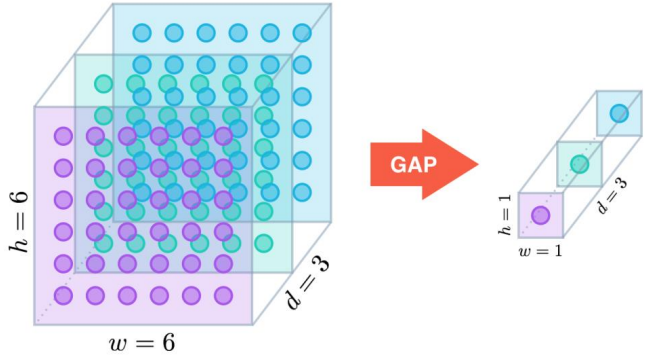
DEMO

Presentor: Zhiyun Ling

Recall

- Global Average Pooling + Class Activation Mapping
- Localize the discriminative image regions using CAM in a single forward-pass
- Achieve 37.1% top-5 error for object localization on ILSVRC2014
  (compared to 34.2% top5 error achieved by fully supervised AlexNet)
- A generic localizable deep representation that can be applied to a variety tasks
  (transferred to other recognition datasets for generic classification, localization, concept discovery)

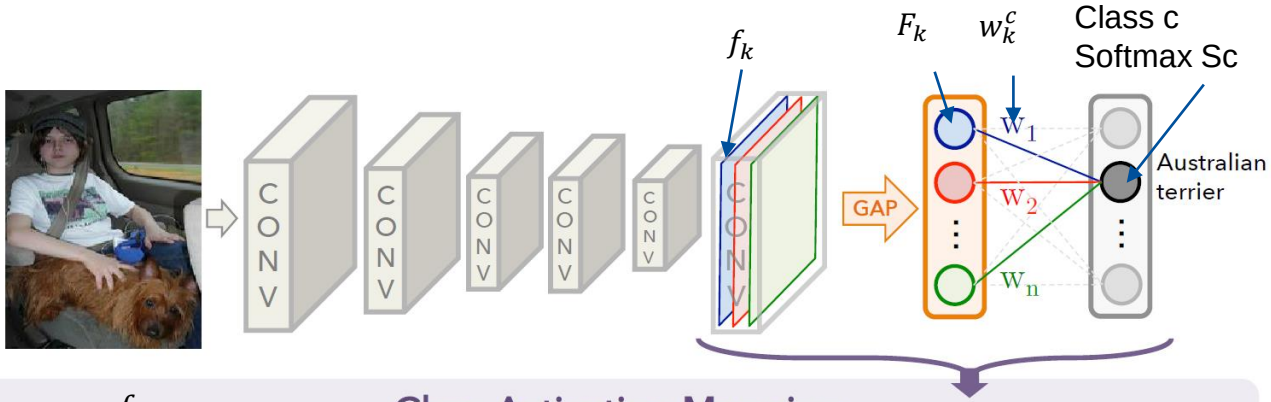
$$F_k = \sum_{x,y} f_k(x, y)$$



Class Activation Mapping CAM

- Indicates the discriminative image regions

$$F_k = \sum_{x,y} f_k(x,y) \qquad S_c = \sum_k w_k^c \sum_{x,y} f_k(x,y) = \sum_{x,y} \overbrace{\sum_k w_k^c f_k(x,y)}^{M_c}$$



M_c indicates the importance of the activation at spatial grid, Which contributes to the classification

Our Work CAM-dMRI

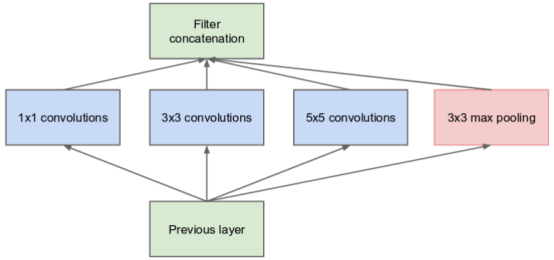
- Age classification (young vs old) based on diffusion MRI images
- Localize the discriminative image regions using CAM
- (Maintaining a good performance on classification)
- Weakly Object Localization, only image labels provided (age group)
- Models: GoogLeNet, GoogLeNet-GMP, VGG16, AlexNet, Inception V3
- Classification evaluation based on Precision-Recall Curves, Accuracy Curve, Loss Curve, Test set Accuracy
- Localization Evaluation, visual comparison and bounding box

Difference CAM vs CAM-dMRI

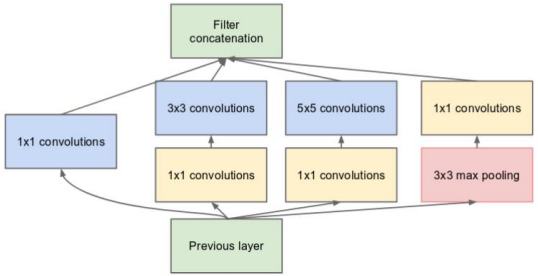
- Color Image: 3 x h x w vs Diffusion Tensor Image: 4D Volume = 3D voxel x direction
- 1. Use DiPY to load nifti volume images, convert to Fractional Anisotropy map (3D)
- 2. Use 2.5D representation, get multiple slices from 3D images in 3 directions.
- 3. modify all networks input to N x 1 x h x w.

- Dataset: 1.3M vs 9984 (8448+1536) = (26 subjects x (RL,LR) x (dir80,dir81) x (gdc, raw) x 2.5D)
- 1. Augmentation: Horizontal Flip, Gaussian noise (0, 0.3), Resizing, Random Cropping, Random Erasing, 2.5D.
- 2. Dropout, batch normalization to prevent overfitting.
- 2. Keep validation set(0.2) and test set to verify results.

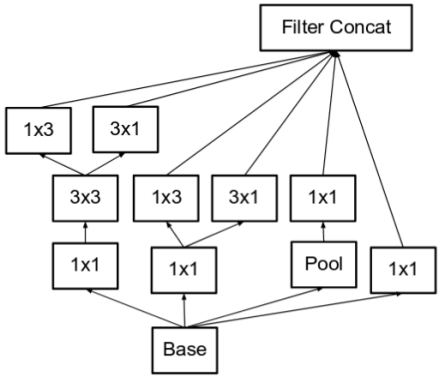
- Models: Added Inception V3 (Improved GoogLeNet).
- Implementation: Caffe + MATLAB vs PyTorch



(a) Inception module, naïve version



(b) Inception module with dimension reductions



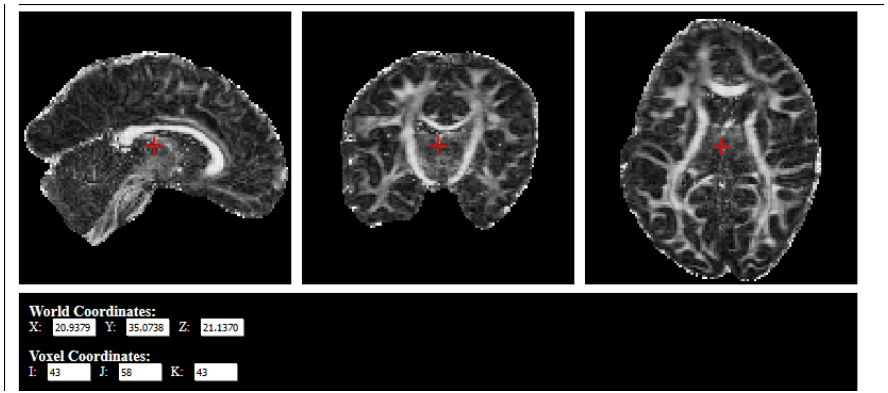
Preprocess Data

- Dataset: Human Connectome Project **WU-Minn HCP Lifespan Pilot Data**

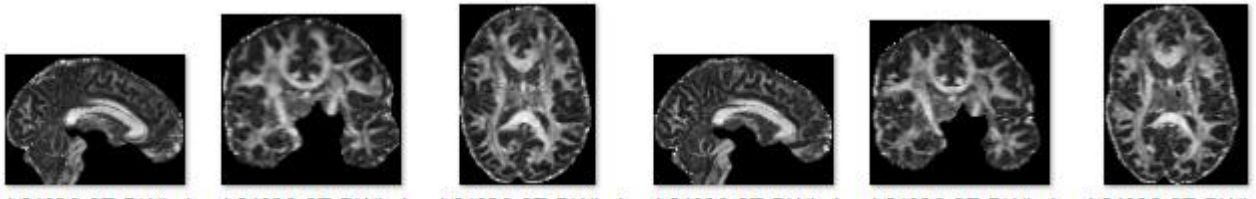
Imaging and behavioral data from 6 age groups (4-6, 8-9, 14-15, 25-35, 45-55, 65-75 years old) 27-1 = 26 subjects
Form 2 group: (12)young 8-15, (14)old 25-75. (4-6 yr subject is missing)

- Brain MRI viewer - 4D Volume https://brainbrowser.cbrain.mcgill.ca/volume-viewer

- FA image - 3D



- 2.5D Slices, center 16 slices at each FA , each direction, each configuration



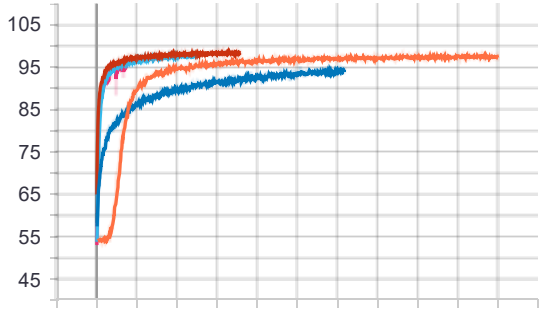
Weakly-Supervised Object Localization

- Observation: Higher Localization ability when last conv layer has higher spatial resolution
- **Setup**
- Modify: remove layers,
  - AlexNet: after conv5. -> resolution 13x13
  - VGGnet: after conv5-3. -> resolution 14x14
  - GoogLeNet: after inception4e. -> resolution 14x14
  - Inception V3: after inception4e -> resolution 14x14

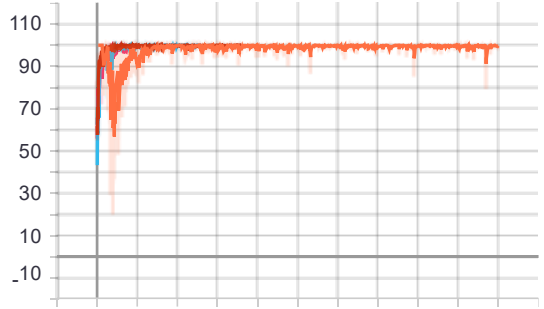
  Add conv layer 3x3, stride=1, pad=1 with 1024 units, followed by GAP + softmax.
- Classification:
  - Compare across models: Accuracy Curve, Loss Curve, Test set Accuracy
- Localization:
  - Compare across models.
  - Compare GMP vs GAP using GoogLeNet.
- Metric: top-1. (only two classes)

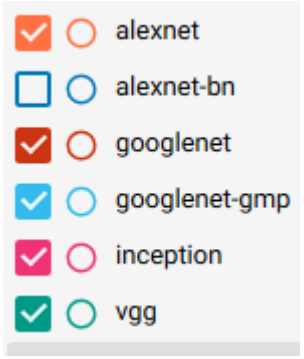
Results

- **Classification** http://192.168.1.3:6007/#scalars
- Training accuracy

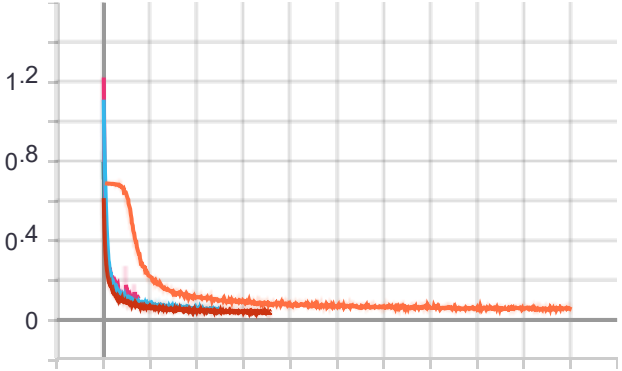


- Validation accuracy

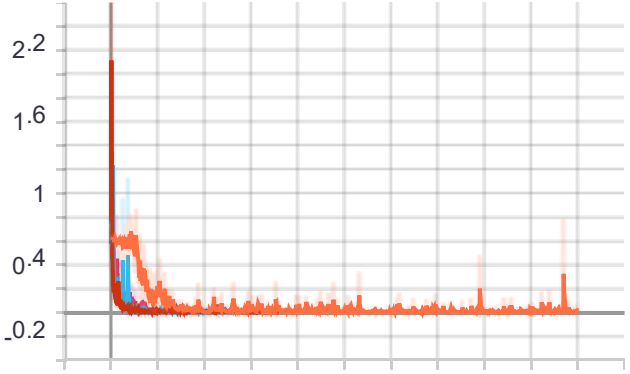




- Training loss



- Validation Loss



Results

- **Classification**
- Accuracy comparison

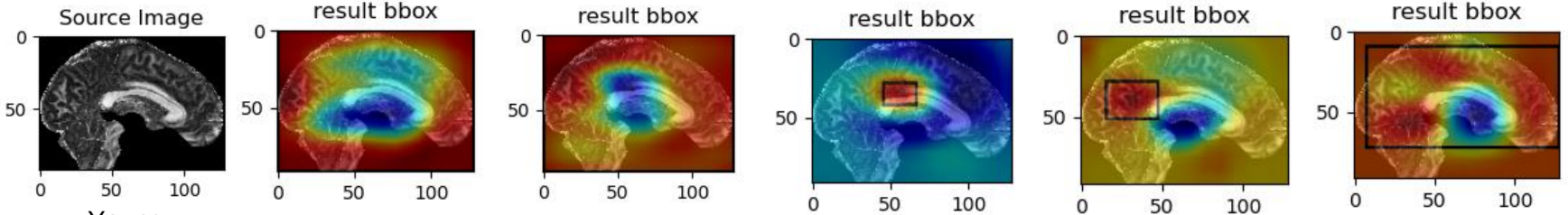
ACC %	AlexNet	VGG	GoogLeNet	GoogLeNet-GMP	Inception V3	AlexNet-BN
Validation	98.08	95.16	99.86	99.75	97.03	NaN
Test	96.03	89.9	94.53	94.92	88.41	77

- AlexNet with batch normalization (AlexNet-BN) performs bad during validation, so not considered.

Results

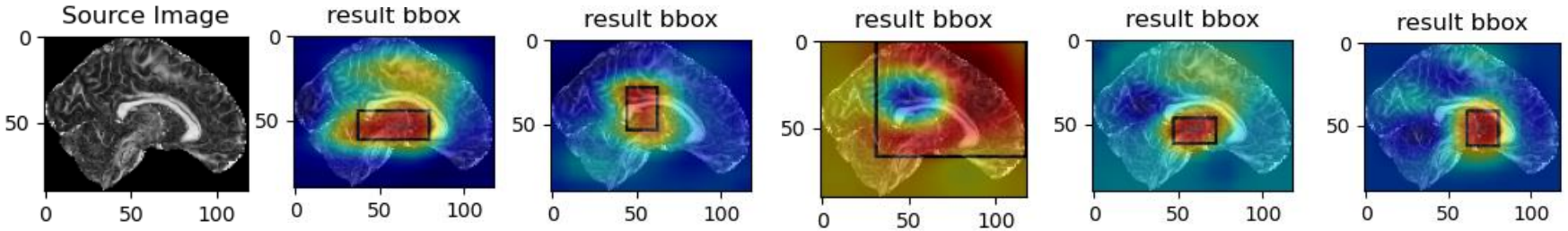
- **Localization**
- bounding box: thresholding 20% of max in CAM http://192.168.1.3:6007/#images
- Same Image Different Network

Old



GoogLeNet GoogLeNet-GMP InceptionV3 VGG AlexNet

- Young



- AlexNet gives best classification results, and performs good at localization
- VGG gives good localization result at a cost of classification performance drop.

Conclusion

- Class Activation Mapping in Diffusion MRI image.
- Trained with age classification, localize most discriminative regions across different age groups.
- Limitations:
  - Can only visualize the final layer heatmap
  - Dataset size is limited
  - Baseline comparison needed
  - Quantification on localization ability needed
- Further Improvements
  - Compare against other localization methods.
  - Use Grad-CAM to enable visualization at each layer
  - Use 3D-CNN to train on 4D volume data
  - More Data,1200subjects in HCP (~1TB)
  - More #Groups, describe where brain changes with age

Q & A


UF
Herbert Wertheim
College of Engineering
Department of Computer & Information
Science & Engineering
UNIVERSITY of FLORIDA
POWERING THE NEW ENGINEER TO TRANSFORM THE FUTURE