ECCV 2016 Amsterdam

Top-down Neural Attention by Excitation Backprop

Jianming Zhang[1], Zhe Lin[1], Jonathan Brandt[1],
Xiaohui Shen[1], Stan Sclaroff[2]

[1]Adobe Research
[2]Boston University

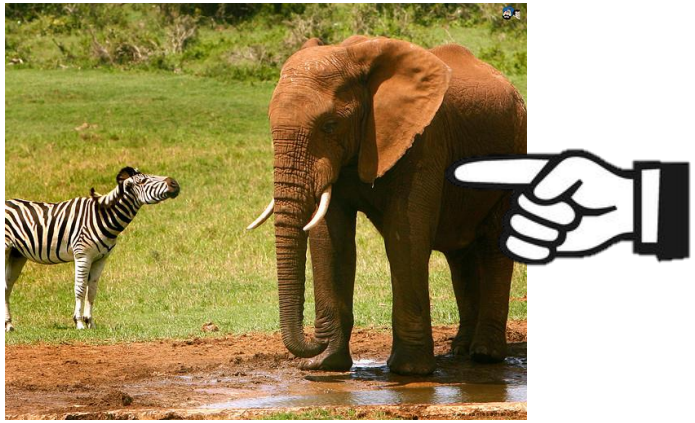
Motivation

Artificial Neural Networks

- Object Categories
- Captions
- Stories

Motivation

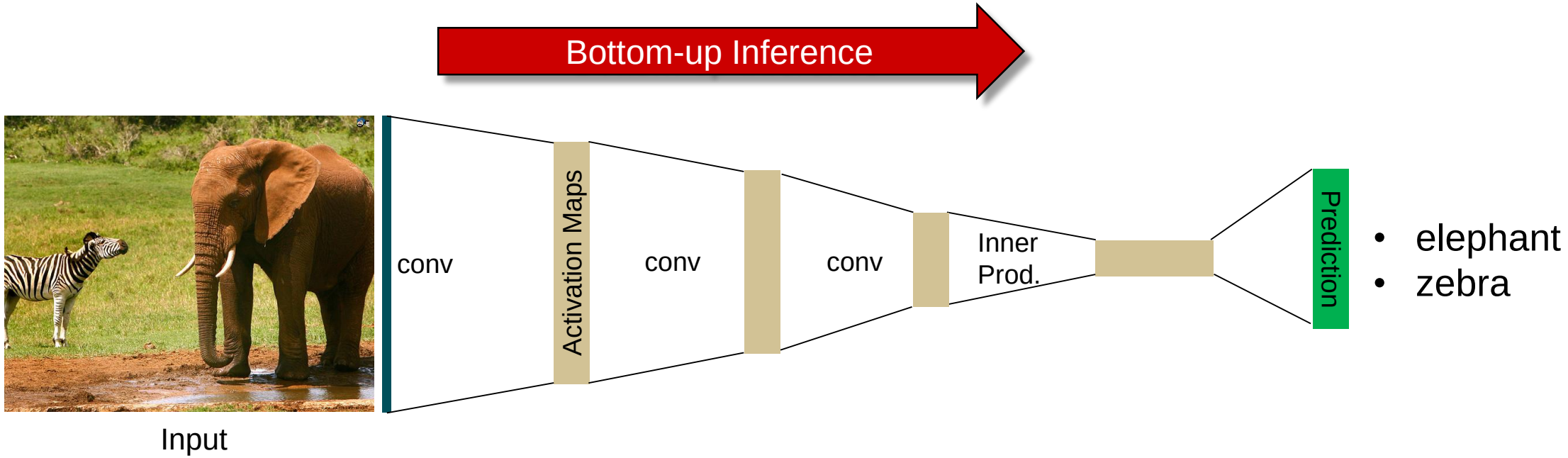
Artificial Neural Networks



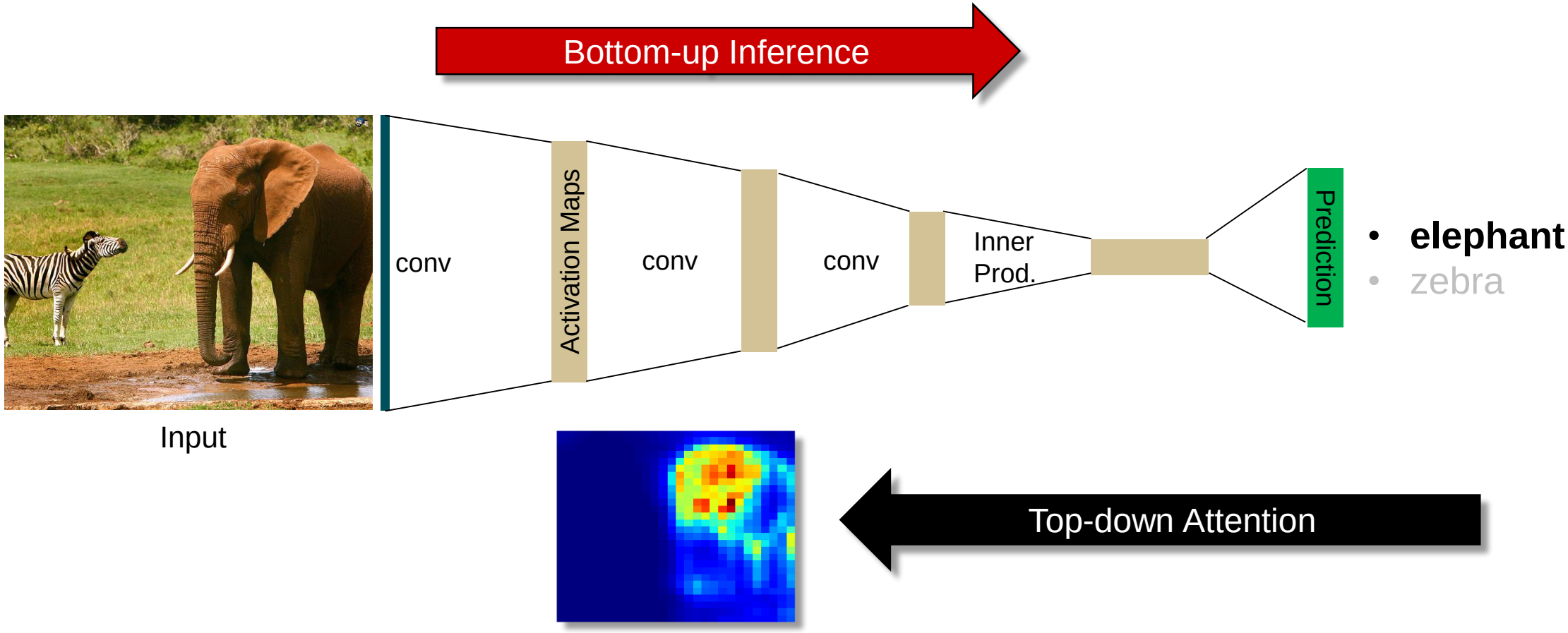
- Object Categories
- Captions
- Stories

Can these models ground their own predictions?

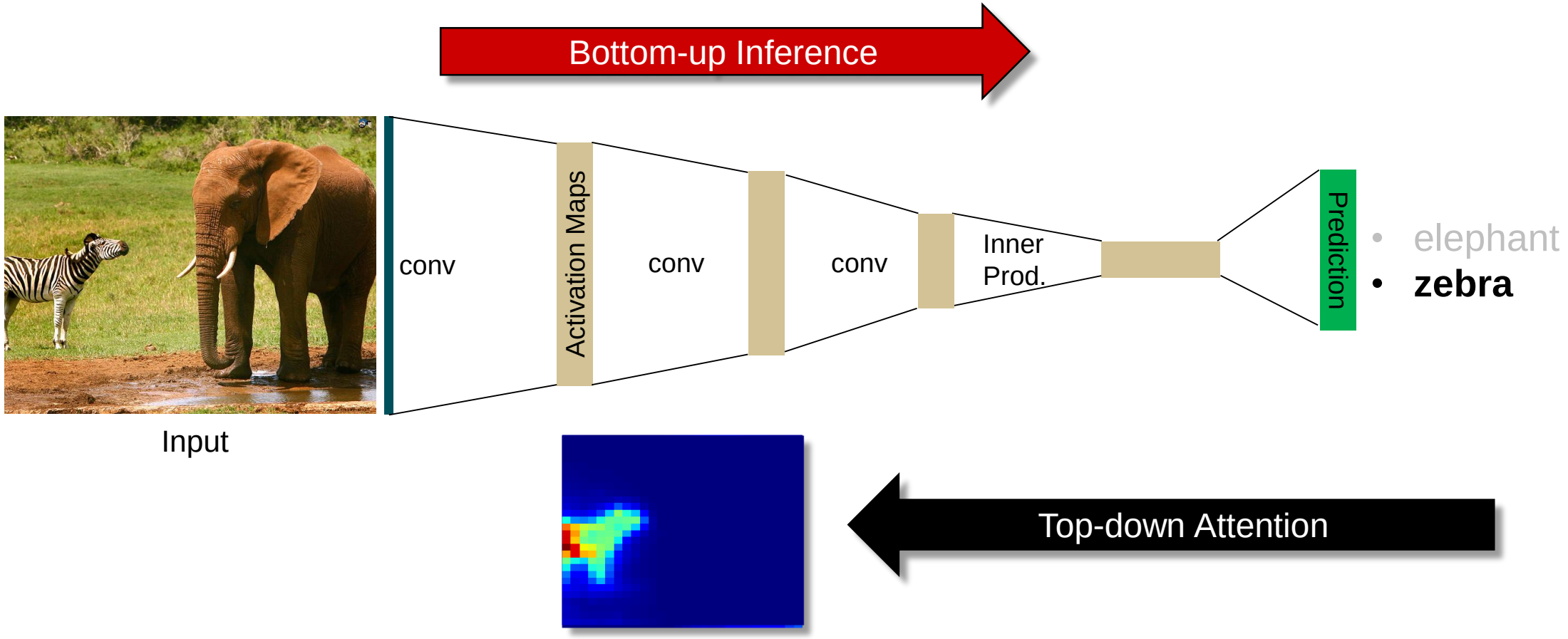
Goal: Generate Top-Down Attention Maps



Goal: Generate Top-Down Attention Maps



Goal: Generate Top-Down Attention Maps



Related Work

- Masking-based [1, 2]
- Optimization-based [3]
- Fully-conv-based [4, 5]
- Backprop-based [6, 7, 8]

[1] Zhou et al. “Object detectors emerge in deep scene CNNs.” *ICLR,* 2015.

[2] Bergamo et al. “Self-taught object localization with deep networks.” *arXiv preprint arXiv:1409.3964*, 2014.

[3] Cao et al. “*Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks*.” ICCV, 2015.

[4] Sermanet et al. “*Overfeat: Integrated recognition, localization and detection using convolutional networks*.” ICLR, 2014.

[5] Zhou et al. “Learning Deep Features for Discriminative Localization.” CVPR, 2016.

[6] Zeiler et al. “*Visualizing and understanding convolutional networks.*” ECCV, 2014.

[7] Simonyan et al. “*Deep inside convolutional networks: Visualizing image classification models and saliency maps*.” ICLRW, 2014.

[8] Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.” PloS One, 2015.

Related Work

Masking-based [1, 2]

Optimization-based [3]

Fully-conv-based [4, 5]

Backprop-based [6, 7, 8]

- **General:** is applicable to a wide variety of DNNs
- **Simple:** can generate an attention map in a single backward pass

[1] Zhou et al. “Object detectors emerge in deep scene CNNs.” *ICLR,* 2015.

[2] Bergamo et al. “Self-taught object localization with deep networks.” *arXiv preprint arXiv:1409.3964*, 2014.

[3] Cao et al. “*Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks*.” ICCV, 2015.

[4] Sermanet et al. “*Overfeat: Integrated recognition, localization and detection using convolutional networks*.” ICLR, 2014.

[5] Zhou et al. “Learning Deep Features for Discriminative Localization.” CVPR, 2016.

[6] Zeiler et al. “*Visualizing and understanding convolutional networks.*” ECCV, 2014.

[7] Simonyan et al. “*Deep inside convolutional networks: Visualizing image classification models and saliency maps*.” ICLRW, 2014.

[8] Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.” PloS One, 2015.

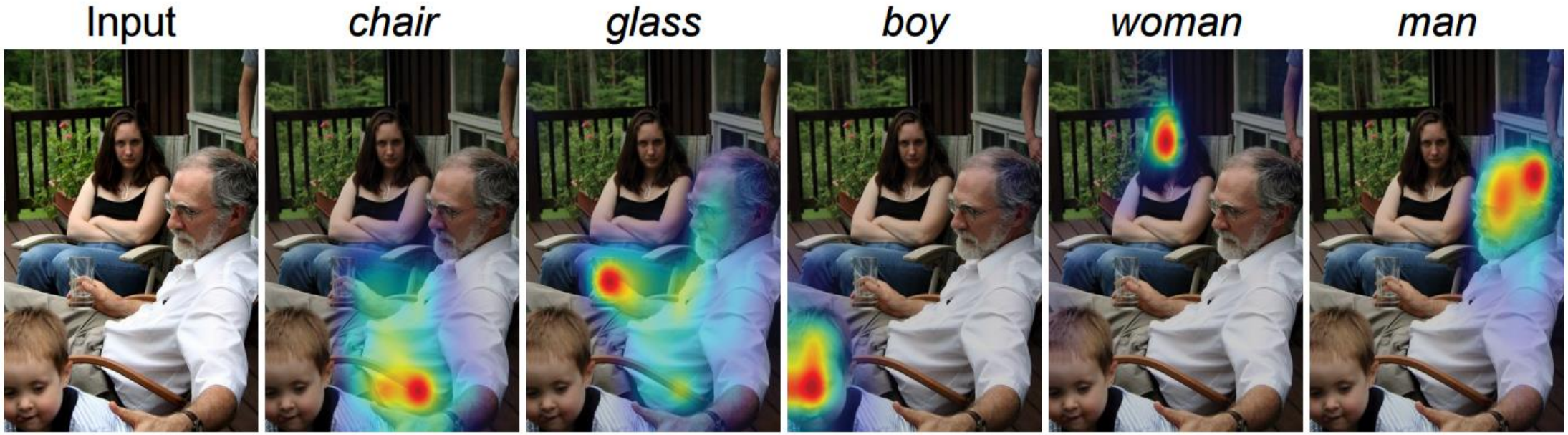
Contributions

Excitation Backprop

- Based on the biologically-inspired Selective Tuning model of visual attention
- Probabilistic Winner-Take-All scheme that is applicable to modern DNNs

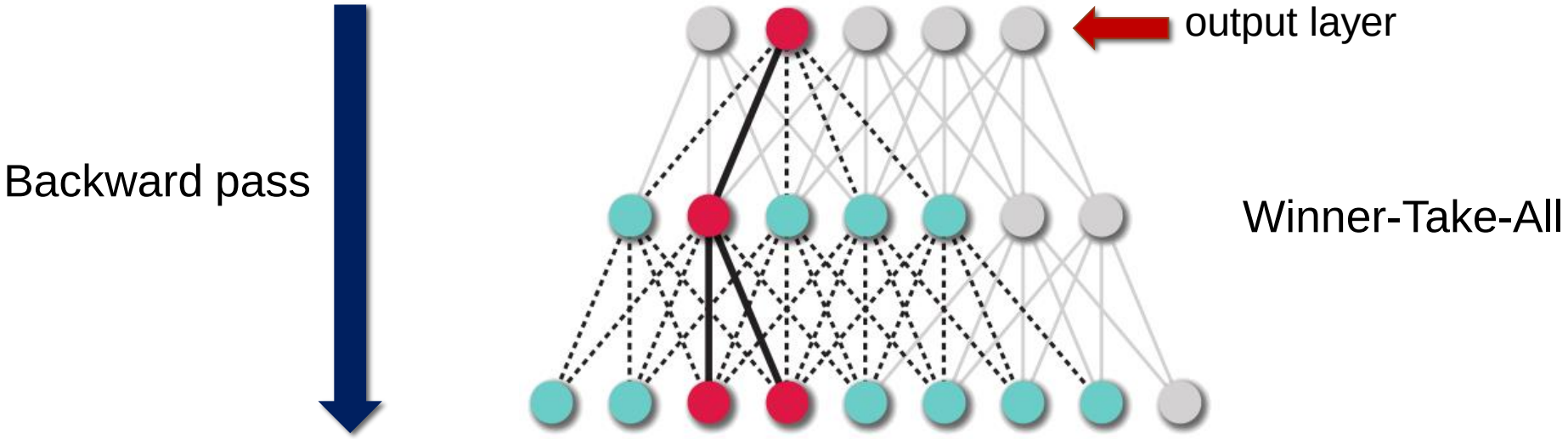
Contrastive Top-down Attention Formulation

- Significantly improves the discriminativeness of our attention maps



The Selective Tuning Model [Tsotsos et al. 1995]

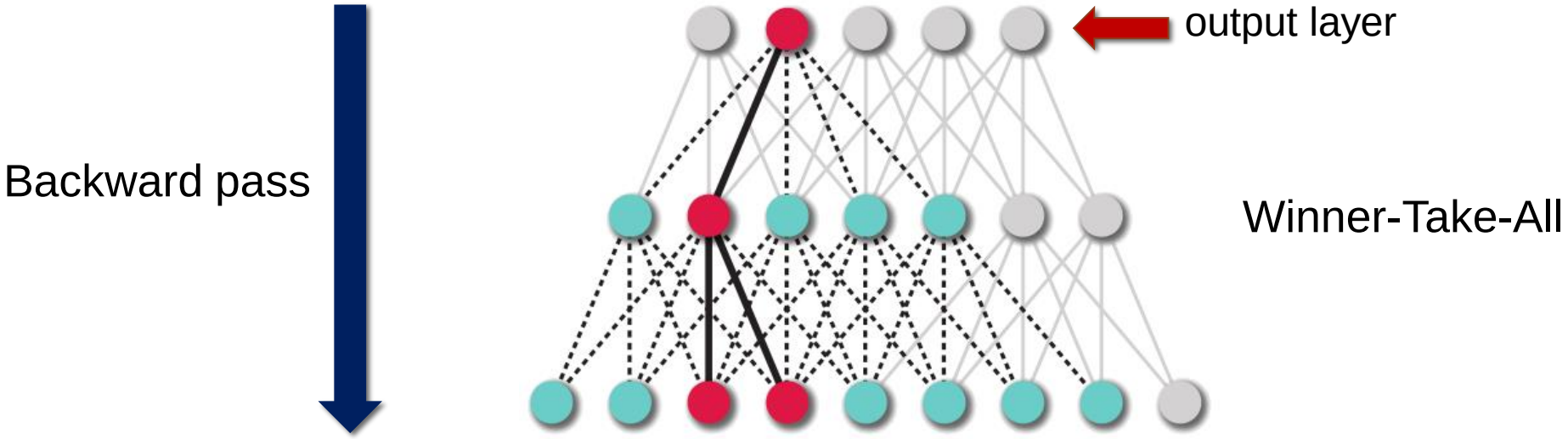
- Forward pass to compute the feature values at each layer, as well as predictions
- Backward pass to localize relevant regions



[1] Tsotsos et al. “Modeling Visual Attention via Selective Tuning.” Artificial Intelligence, 1995.

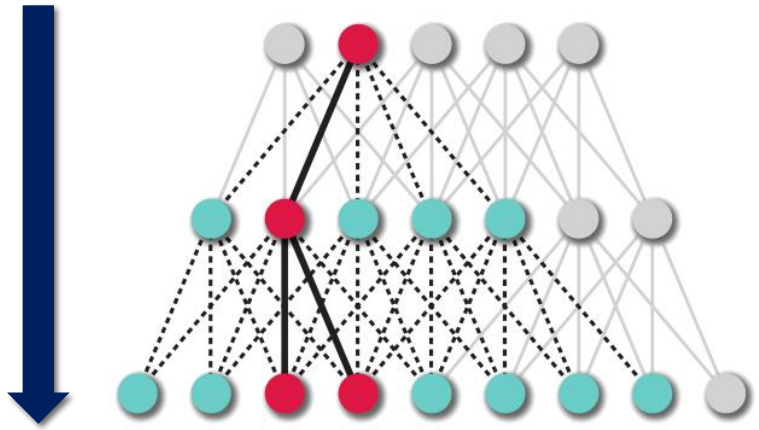
The Selective Tuning Model [Tsotsos et al. 1995]

- Forward pass to compute the feature values at each layer, as well as predictions
- Backward pass to localize relevant regions

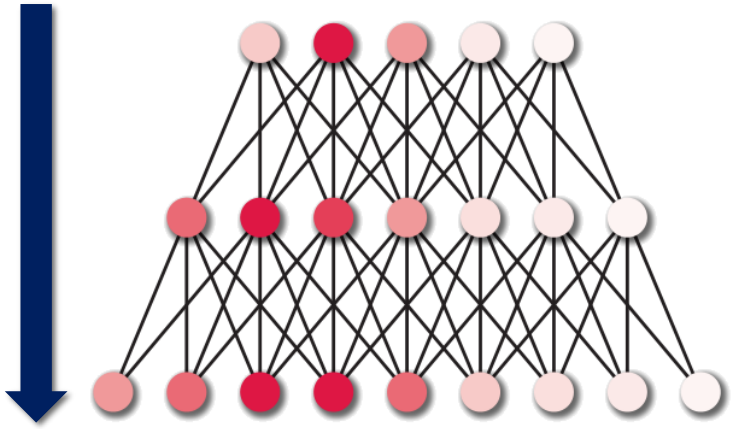


For deep neural networks, this greedy, winner-take-all method produces very sparse binary maps, and only uses information of a very small portion of the whole network.

Our Approach: Probabilistic Winner-Take-All



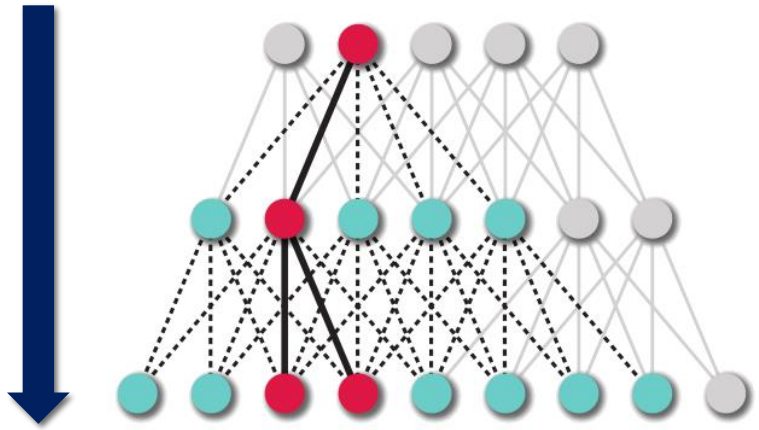
Winner-Take-All [1]



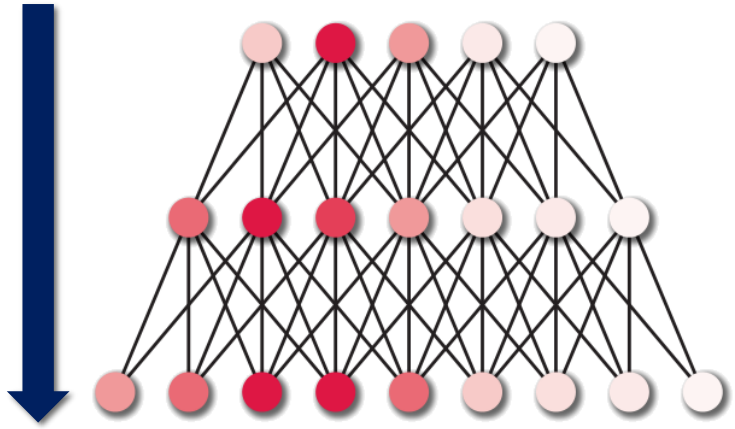
Winner Sampling

[1] Tsotsos et al. “Modeling Visual Attention via Selective Tuning.” Artificial Intelligence, 1995.

Our Approach: Probabilistic Winner-Take-All



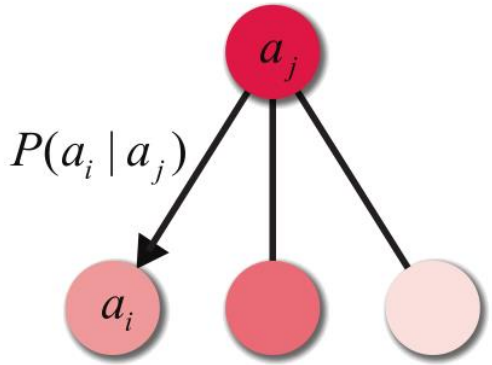
Winner-Take-All [1]



Winner Sampling

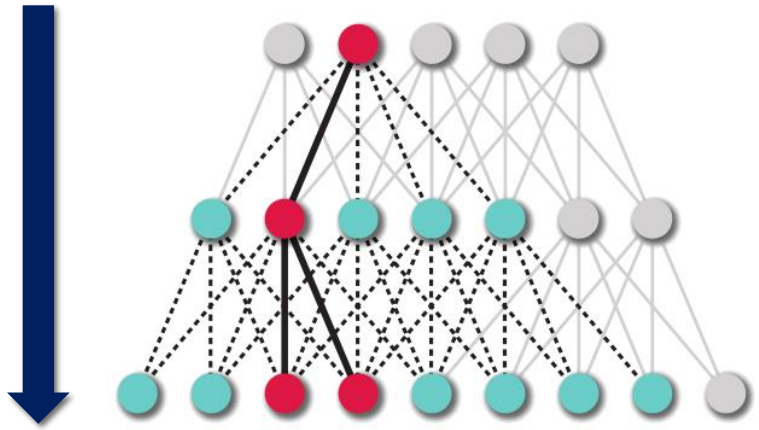
Marginal Winning Probability (MWP):

$$P(a_j) = \sum_{a_i \in \mathcal{P}_j} P(a_j|a_i)P(a_i)$$

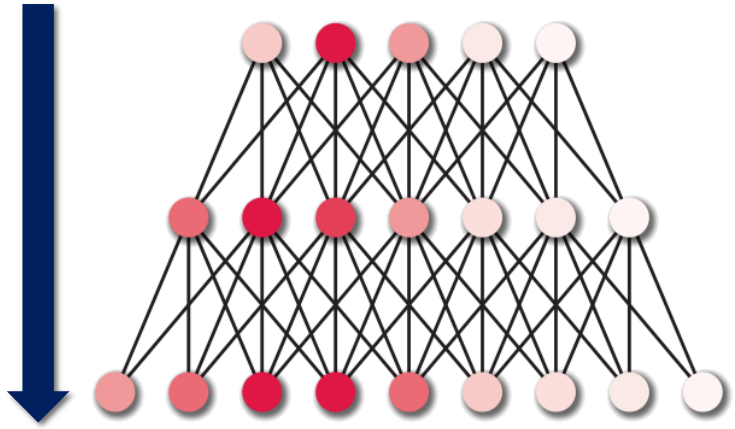


[1] Tsotsos et al. “Modeling Visual Attention via Selective Tuning.” Artificial Intelligence, 1995.

Our Approach: Probabilistic Winner-Take-All



Winner-Take-All [1]

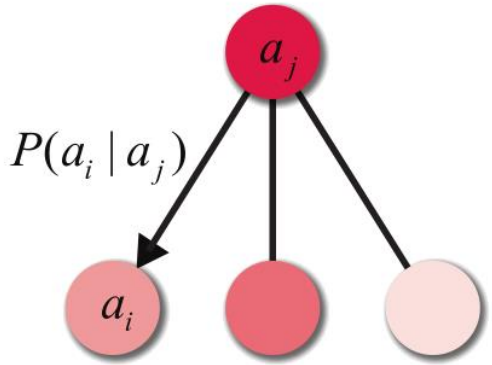


Winner Sampling

Equivalent to an Absorbing Markov Chain process.

Marginal Winning Probability (MWP):

$$P(a_j) = \sum_{a_i \in \mathcal{P}_j} P(a_j|a_i)P(a_i)$$



[1] Tsotsos et al. “Modeling Visual Attention via Selective Tuning.” Artificial Intelligence, 1995.

Excitation Backprop

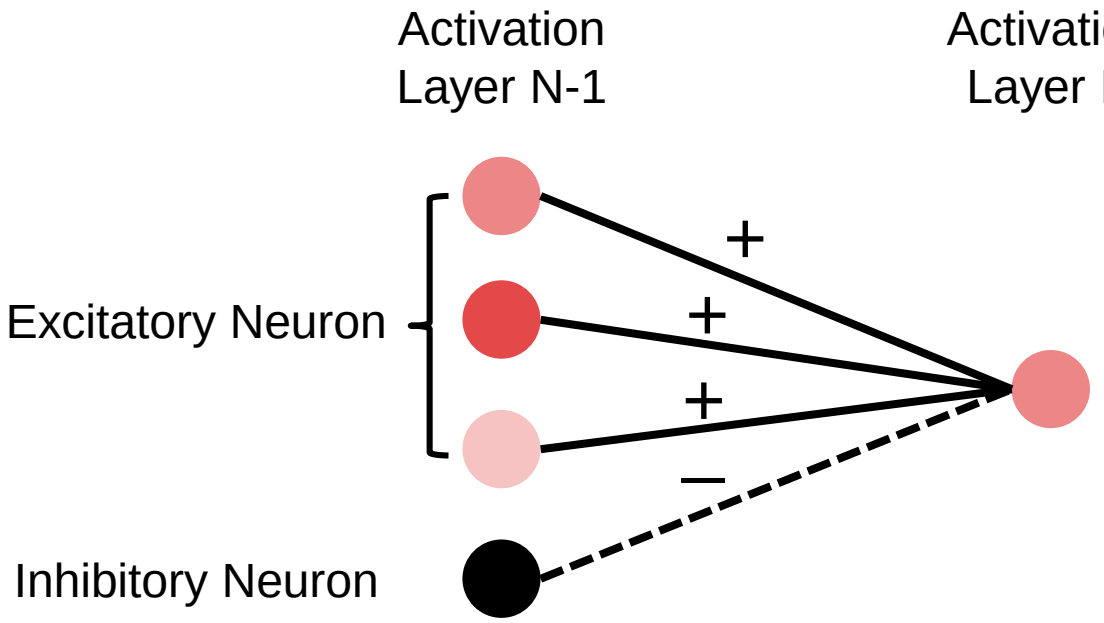
Assumptions:

- The responses of the activation neurons are non-negative.
- An activation neuron is tuned to detect certain visual features. Its response is positively correlated to its confidence of the detection.

Excitation Backprop

Assumptions:

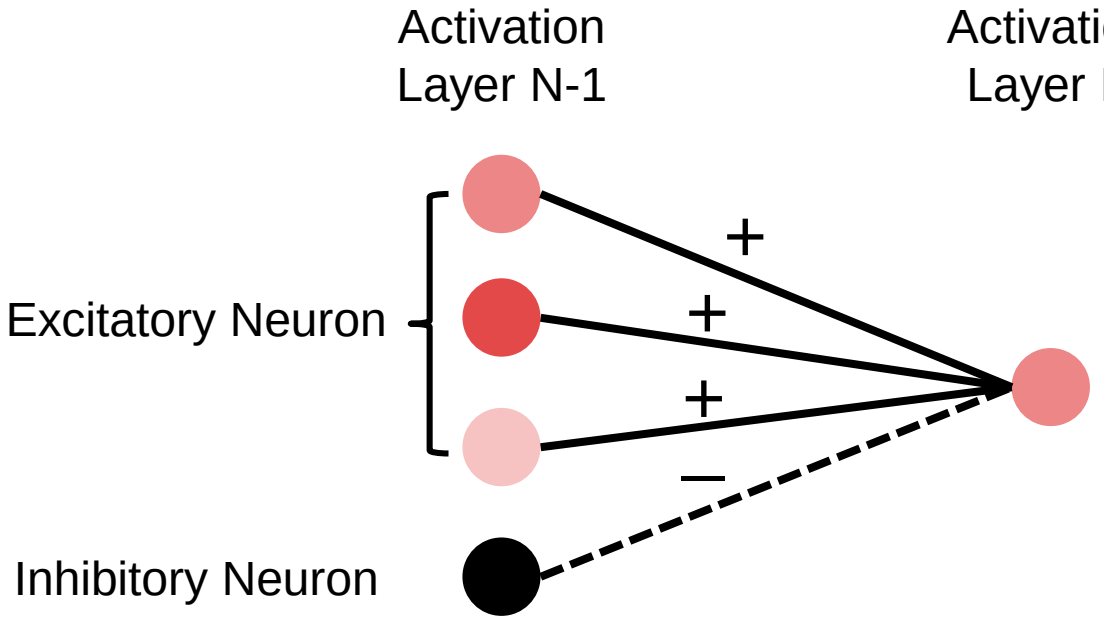
- The responses of the activation neurons are non-negative.
- An activation neuron is tuned to detect certain visual features. Its response is positively correlated to its confidence of the detection.



Excitation Backprop

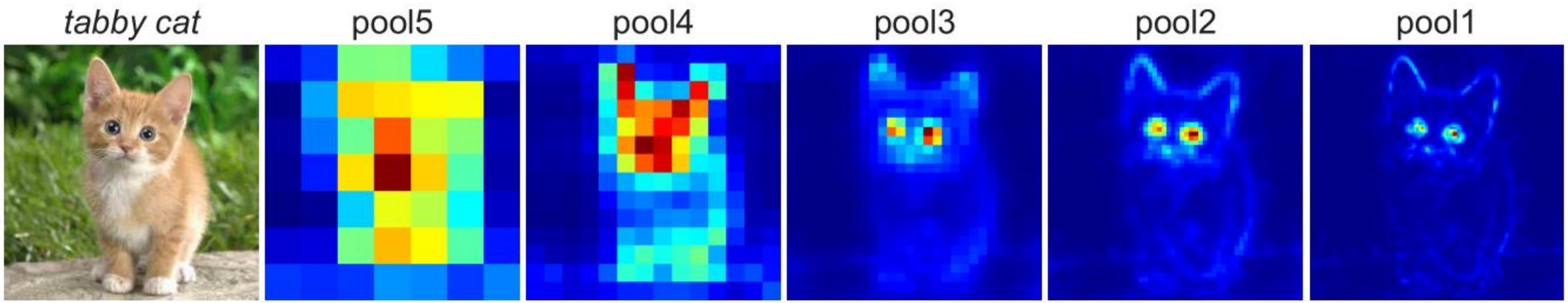
Assumptions:

- The responses of the activation neurons are non-negative.
- An activation neuron is tuned to detect certain visual features. Its response is positively correlated to its confidence of the detection.



$$P(a_j|a_i) = \begin{cases} Z_i \widehat{a}_j w_{ji} & \text{if } w_{ji} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

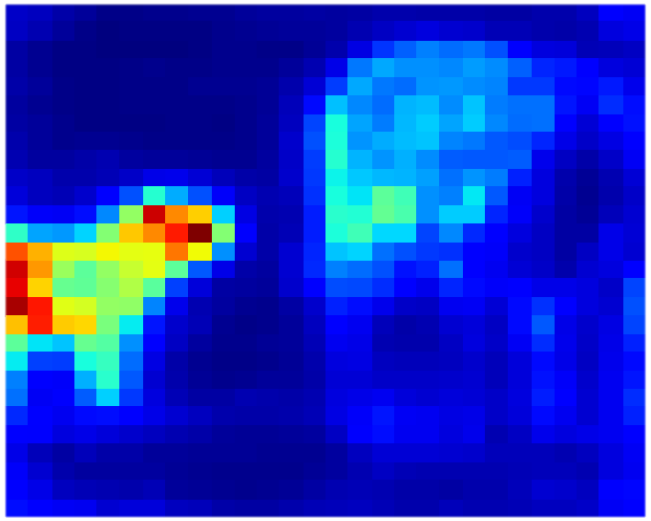
Excitation Backprop



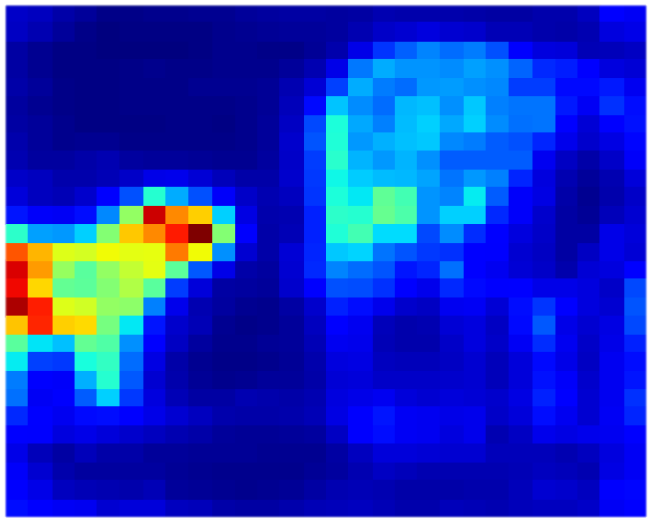
Running excitation backprop, we can extract attention maps from different layers.
Lower layers can generate maps that highlight features of smaller scale.

Challenge: Responsive to Top-down Signals?

zebra



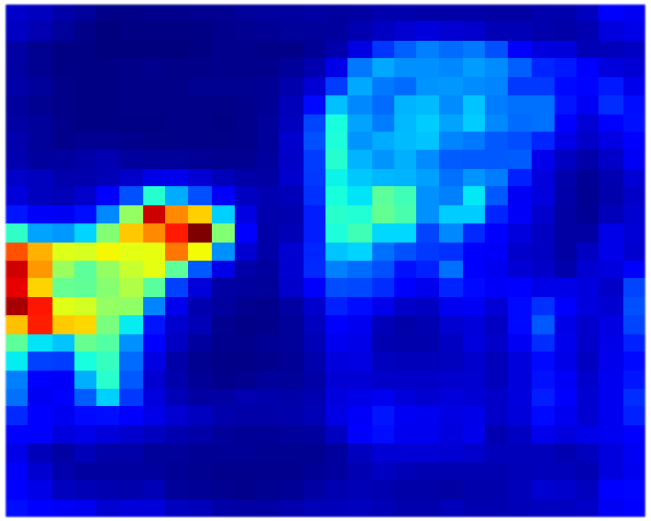
elephant



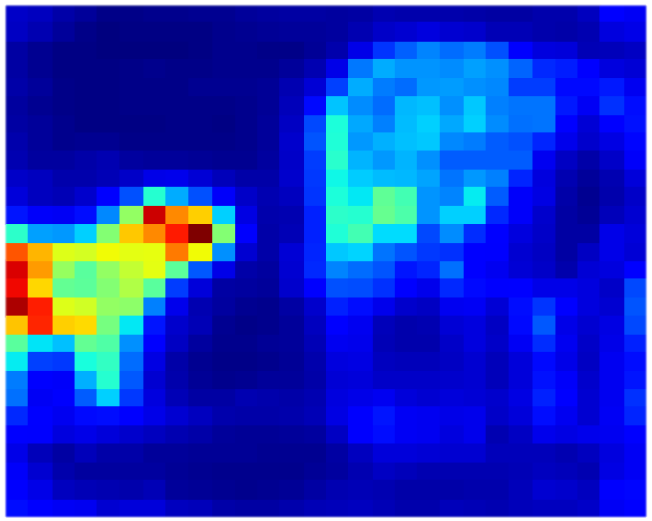
Maps obtained using VGG16 pool3

Challenge: Responsive to Top-down Signals?

zebra



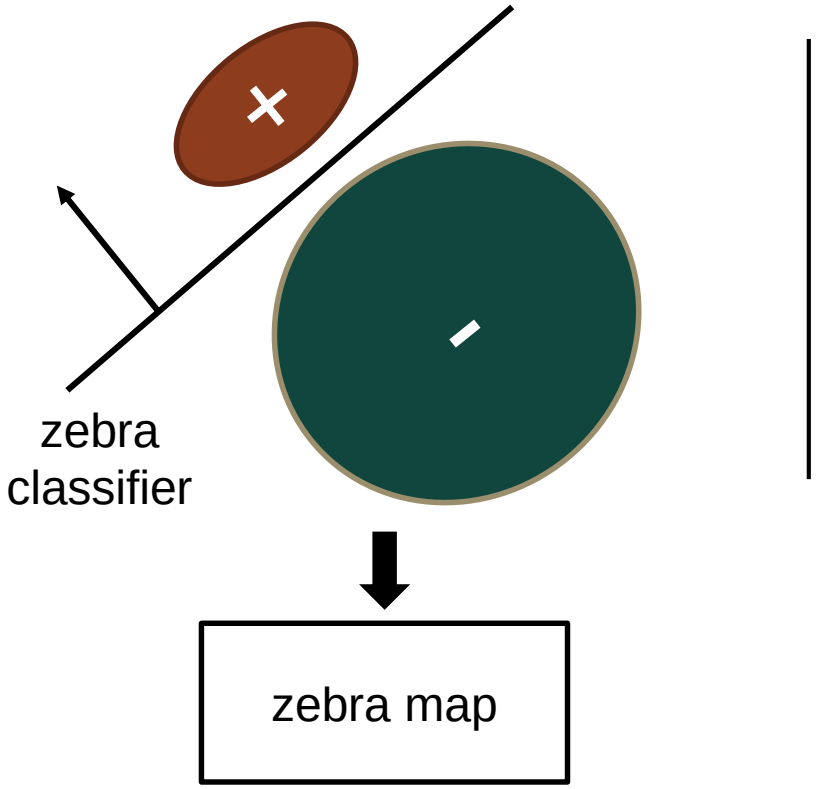
elephant



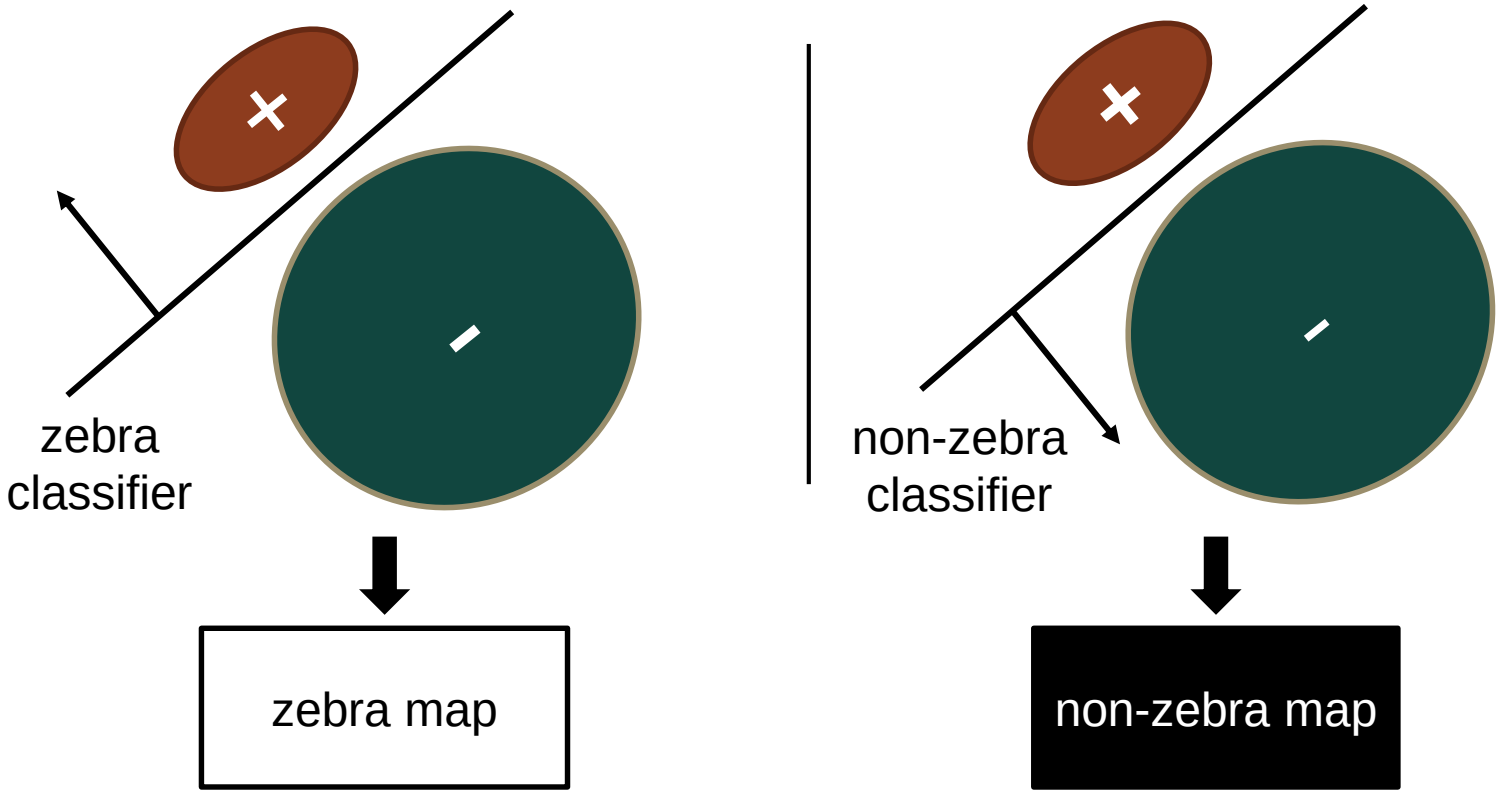
Maps obtained using VGG16 pool3

Dominant neurons always win!

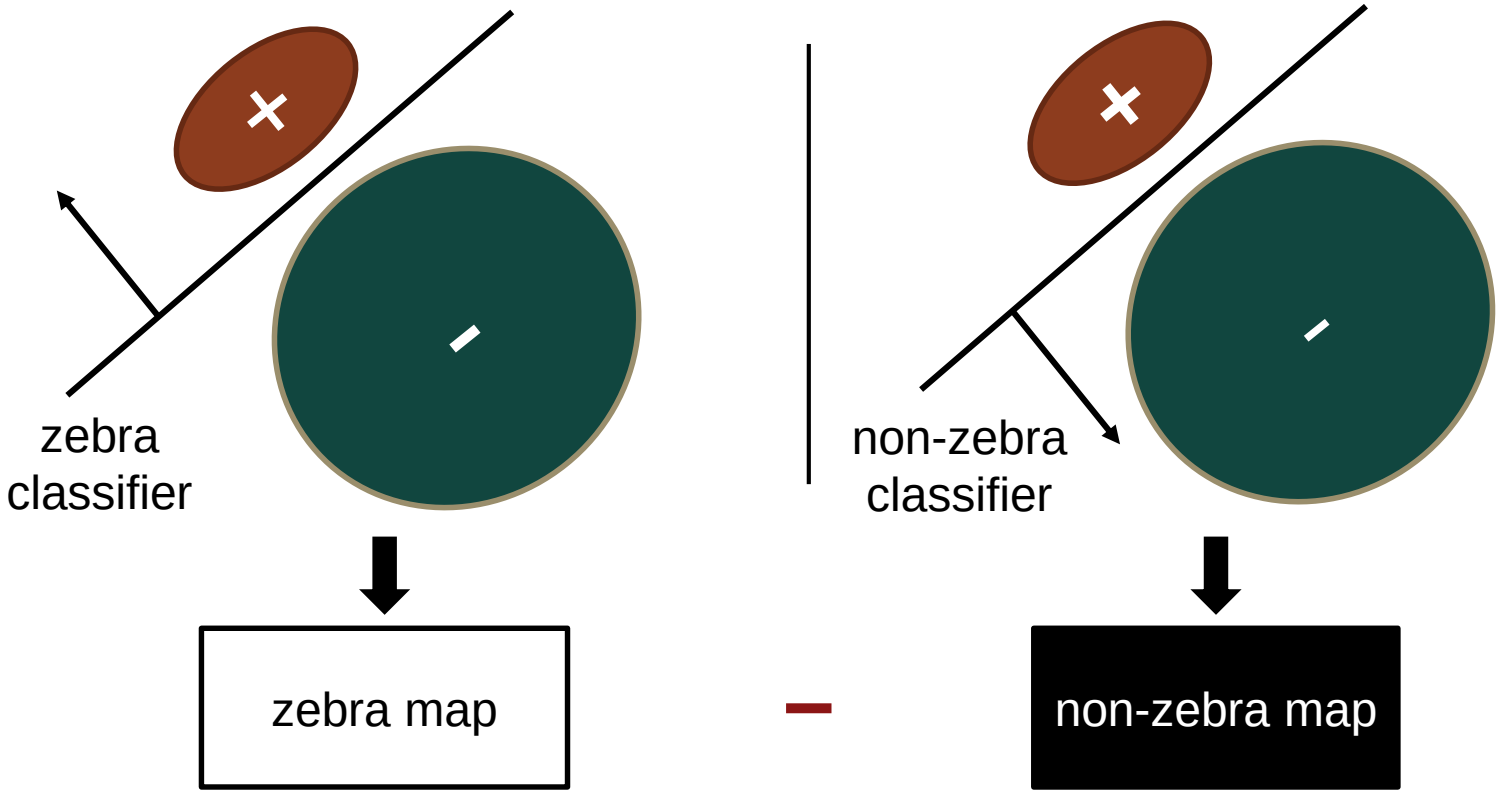
Negating the Output Layer for Contrastive Signals



Negating the Output Layer for Contrastive Signals

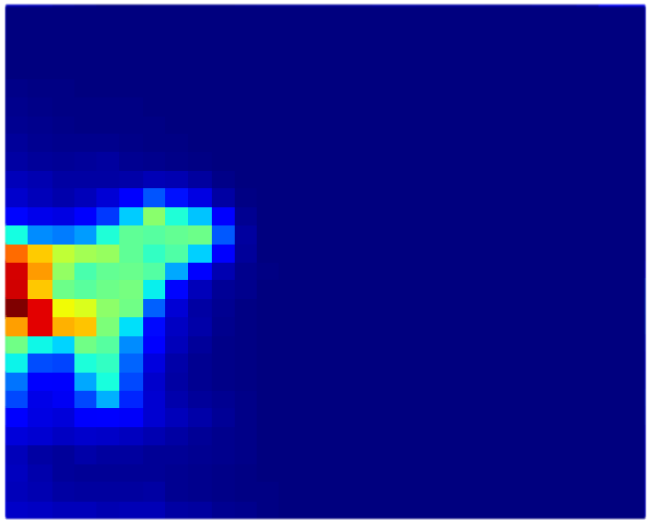


Negating the Output Layer for Contrastive Signals

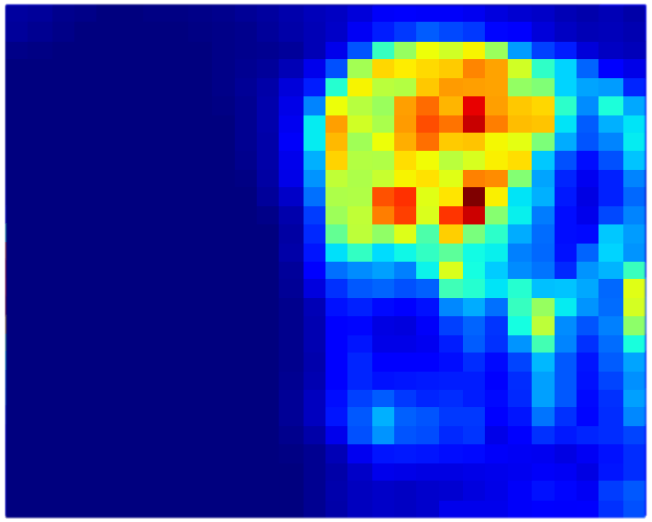


Contrastive Maps

zebra



elephant



- Negative values truncated to 0 and image values rescaled (for visualization)
- Contrastive attention map can be computed by a single pass

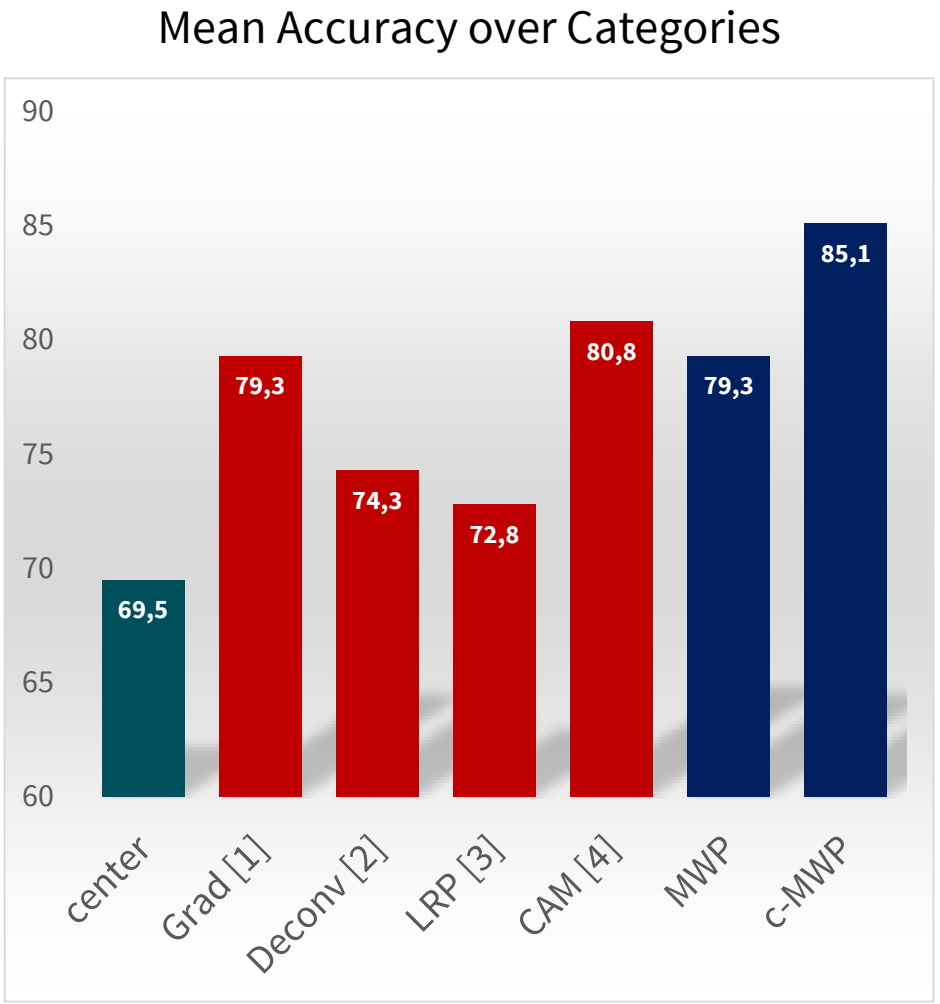
credit: elena milevska / stock.adobe.com

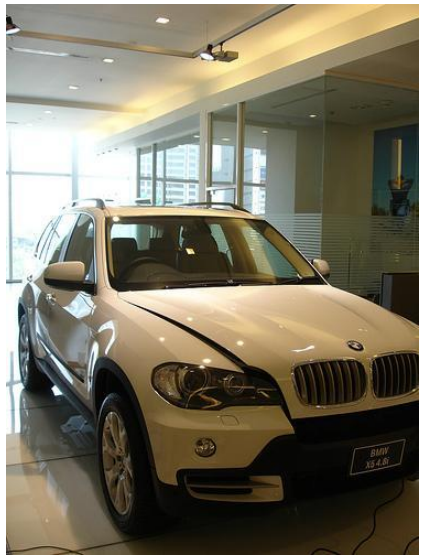
Evaluation: The Pointing Game

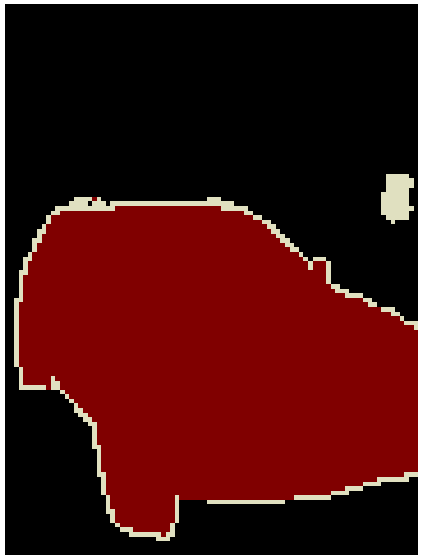
- **Task**:
  - Given an image and an object category, point to the targets.
- **Evaluation Metric**:
  - Mean pointing accuracy across categories
  - Pointing anywhere on the targets is fine
- **CNN Models Tested:**
  - CNN-S [Chatfield et al. BMVC'14]
  - VGG16 [Simonyan et al. ICLR'15]
  - GoogleNet [Szegedy et al. CVPR'15]
- **Model Training:**
  - Multi-label cross-entropy loss
  - Do not use any localization annotations

credit: howtomontessori.com

Results on VOC07 (GoogleNet)







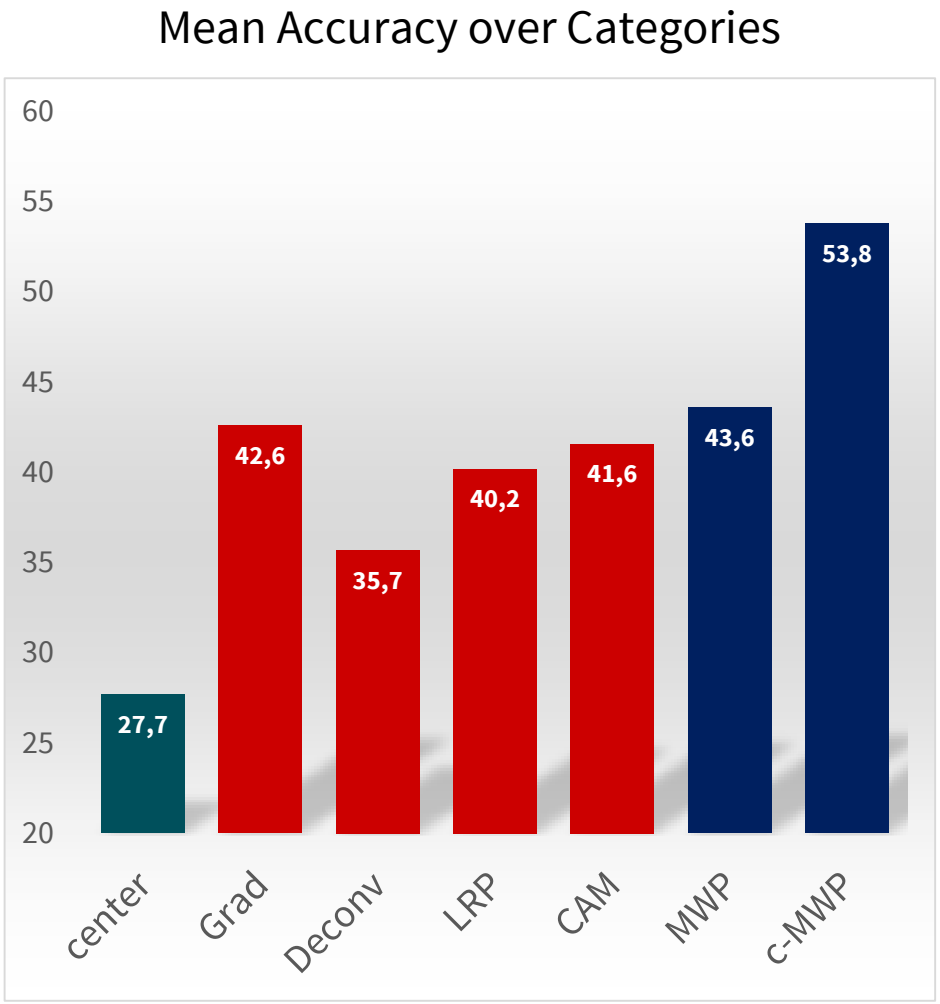
[1] Simonyan et al. "*Deep inside convolutional networks: Visualizing image classification models and saliency maps*." ICLRW, 2014.

[2] Zeiler et al. "*Visualizing and understanding convolutional networks.*" ECCV, 2014.

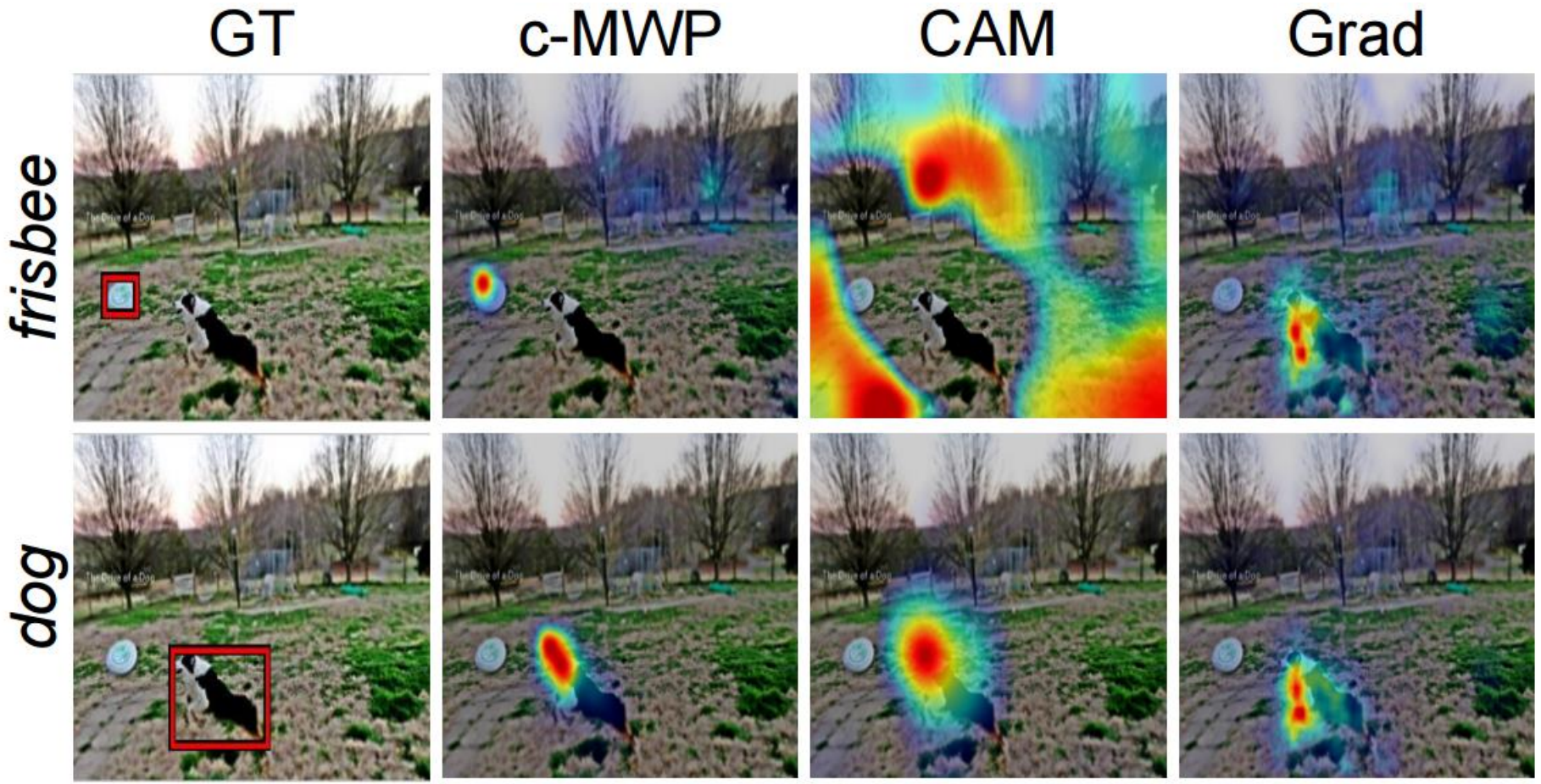
[3] Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS One, 2015.

[4] Zhou et al. "Learning Deep Features for Discriminative Localization." CVPR, 2016.

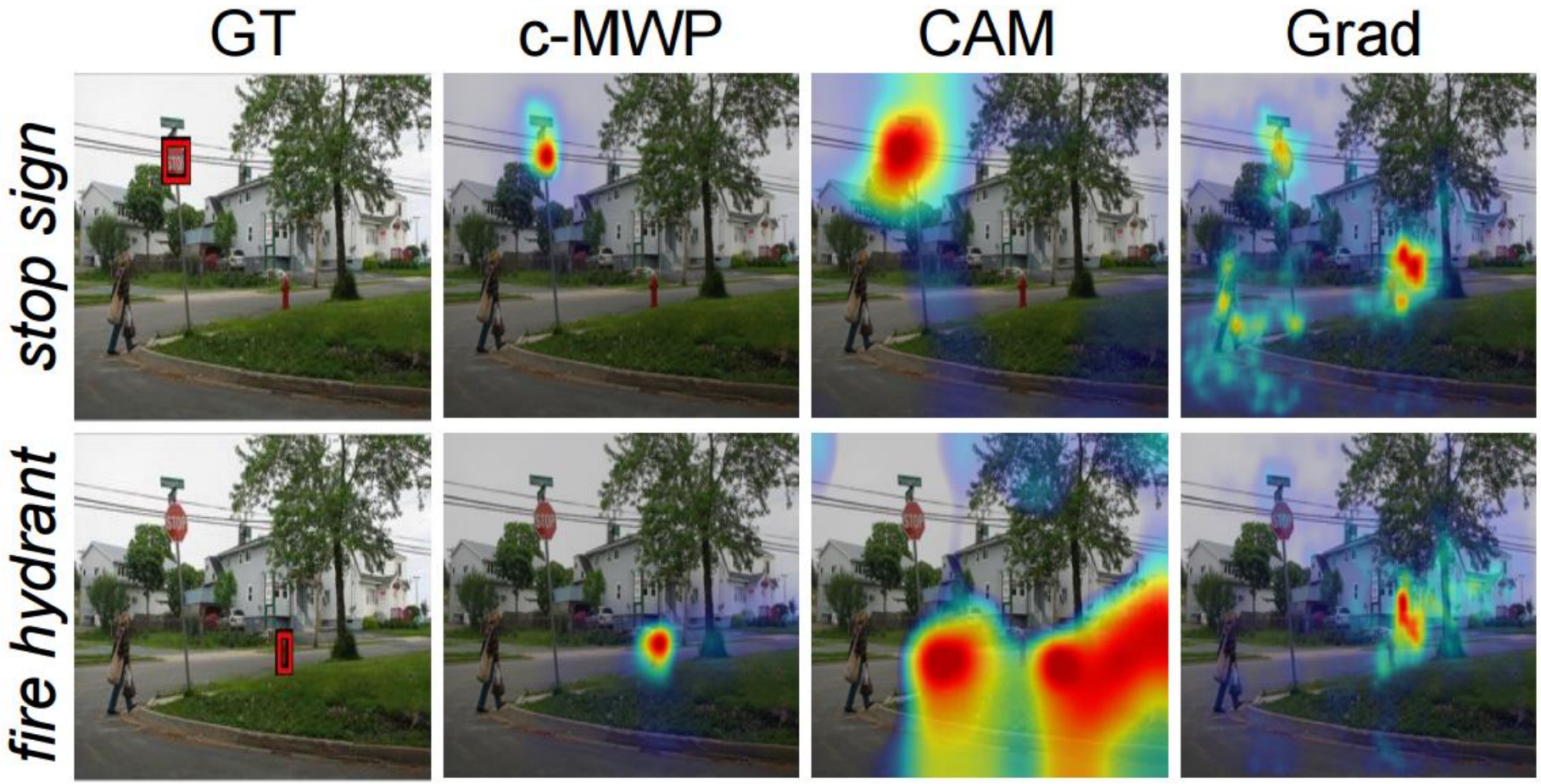
Results on MS COCO (GoogleNet)



Qualitative Comparison



Qualitative Comparison

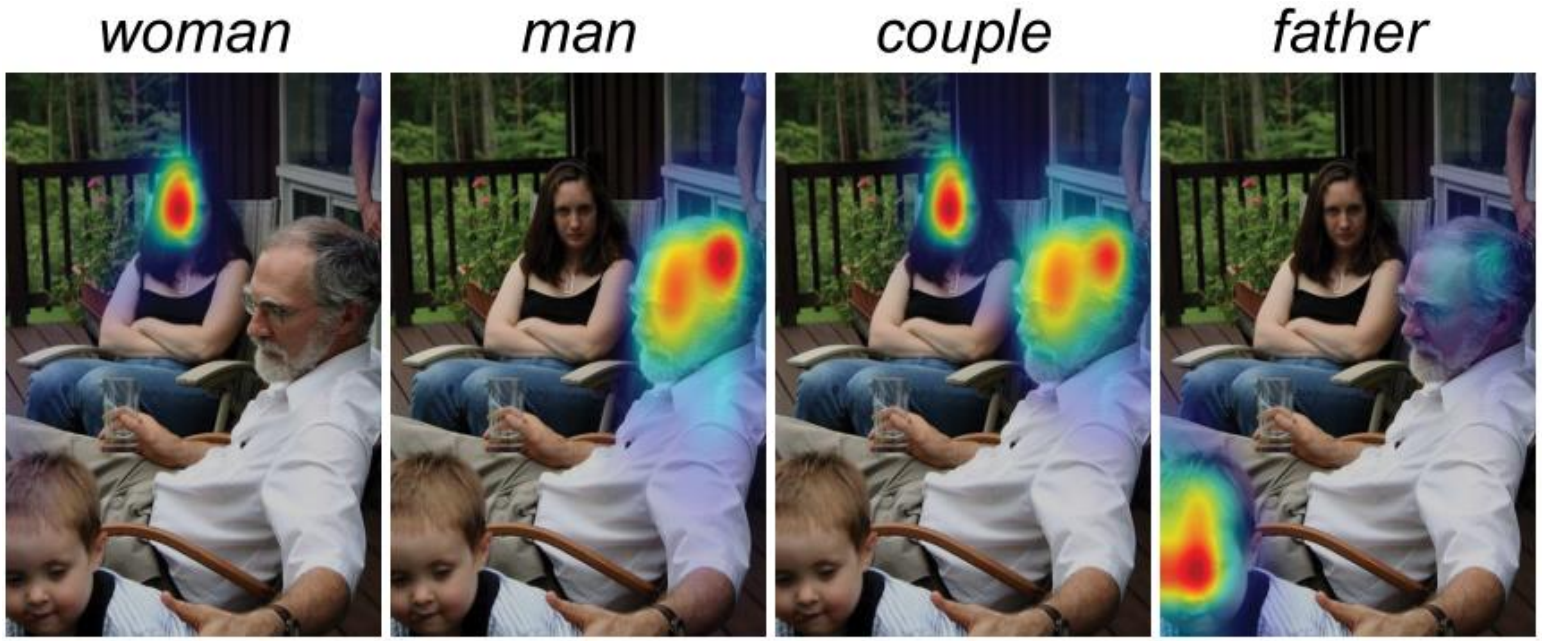


Top-down Attention from an 18K-Tag Classifier

Train an image tag classifier for ~18K tags

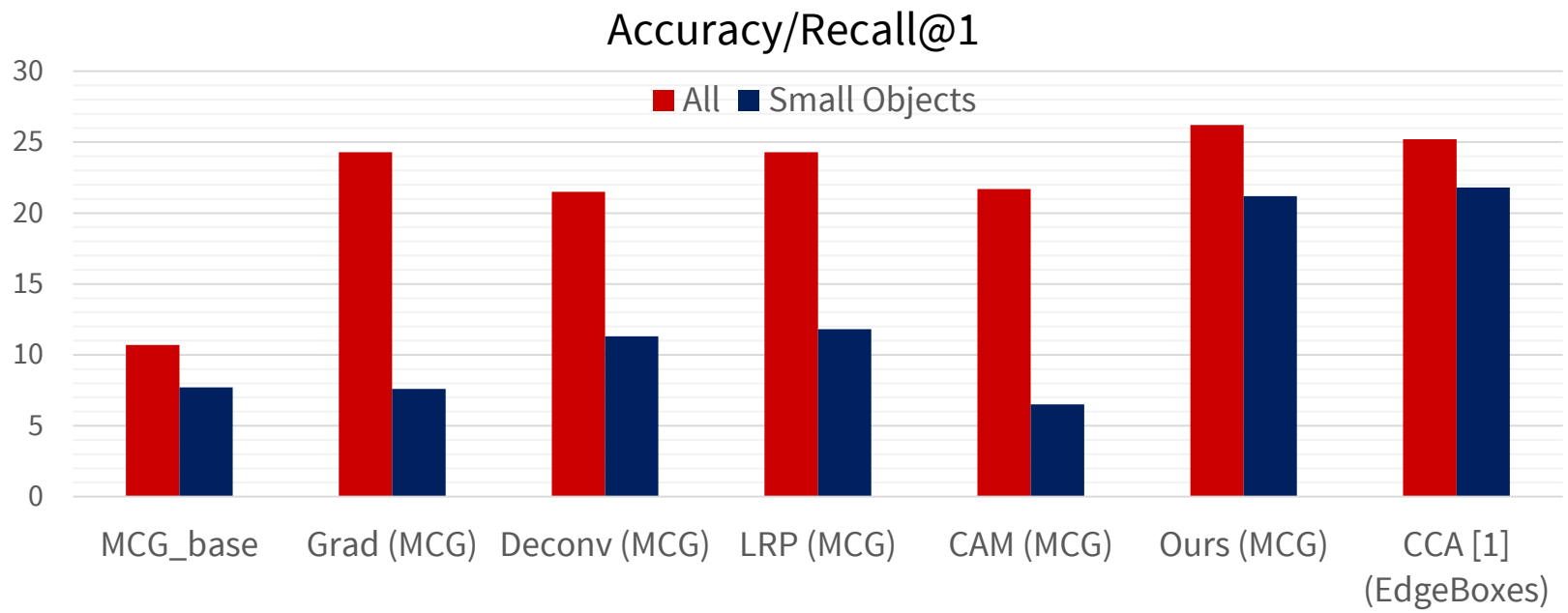
- 6M Stock images with user tags
- Pre-trained GoogleNet model from Caffe Model Zoo
- Cross entropy multi-label loss

An Interesting Case



Phrase Localization

- Follow the evaluation protocol of the Flickr30K entities dataset
- Localization based on top-down attention maps:
  - Take the average of word attention maps to get the phrase attention map
  - Compute object proposals
  - Re-rank proposals using the top-down phrase map



[1] Plummer, et al. "*Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models*." ICCV, 2015.

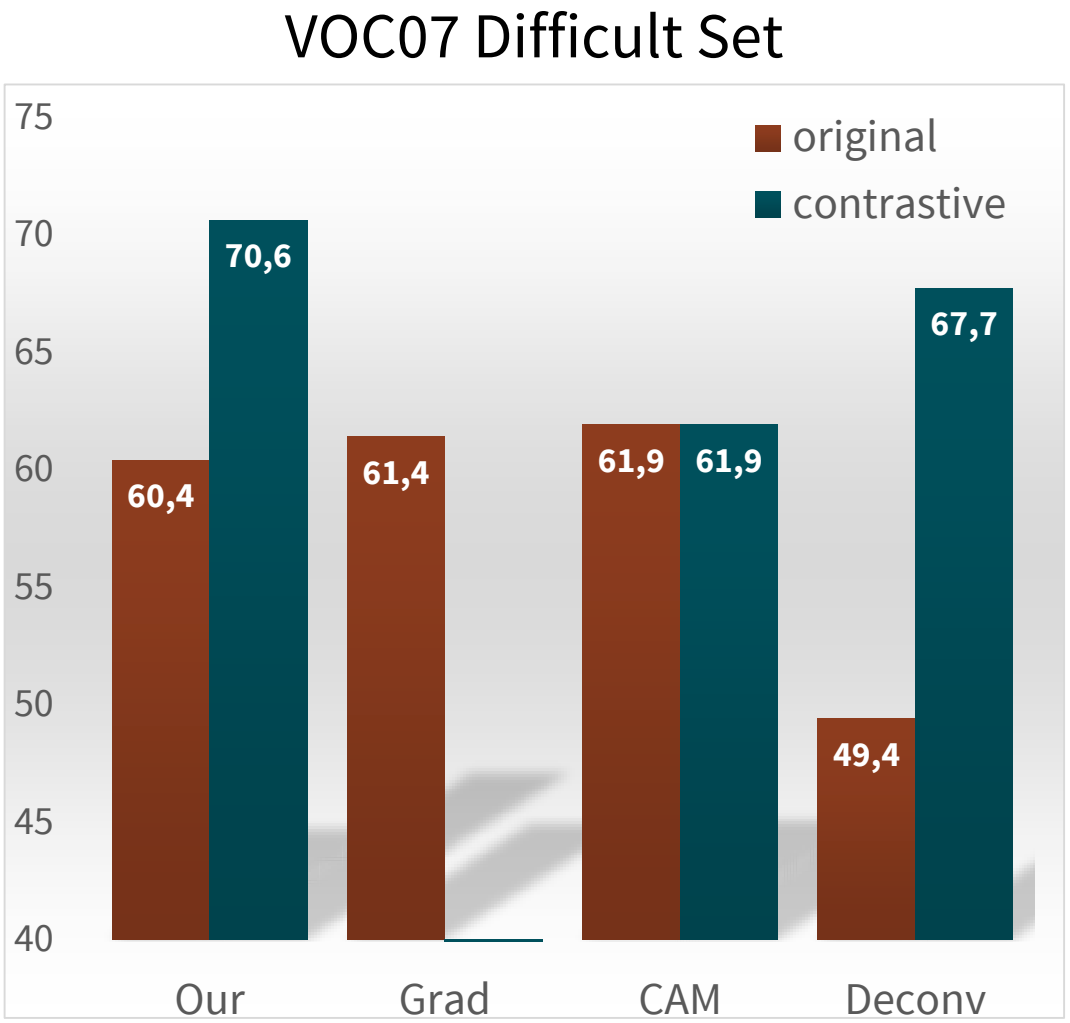
Conclusion

Excitation Backprop + Contrastive Attention = Discriminative Top-down Attention Map

GPU&CPU Implementation in Caffe
https://github.com/jimmie33/Caffe-ExcitationBP

Backup Slides

Does the Contrastive Attention Formulation Work for Other Methods?



Deconv:

+ Truncates negative signals
- Requires normalization
- Requires two backward passes
- Does not use the activation values in the backpropgation

Phrase Localization on the Flickr30K Entity Dataset [1]

	opt. γ	R@1	R@5	R@10	mAP (Group)	mAP (Phrase)
MCG_base	–	10.7/ 7.7	30.3/22.4	40.5/30.3	6.9/ 4.5	16.8/12.9
Grad (MCG)	0.50	24.3/ 7.6	49.6/32.9	59.7/45.8	10.2/ 3.8	28.8/15.6
Deconv (MCG)	0.50	21.5/11.3	48.4/34.5	58.5/46.0	10.0/ 4.0	26.5/16.7
LRP (MCG)	0.50	24.3/11.8	51.6/36.8	61.3/48.5	10.3/ 4.3	28.9/18.1
CAM (MCG)	0.75	21.7/ 6.5	47.1/27.9	56.1/39.1	7.5/ 2.0	26.0/11.9
MWP (MCG)	0.50	**28.5**/15.0	52.7/39.1	61.3/49.8	11.8/ 5.3	**31.1**/20.3
c-MWP (MCG)	0.50	26.2/**21.2**	**54.3**/**43.4**	**62.2**/**51.7**	**15.2**/**10.8**	30.8/**24.0**
CCA* [156] (EB)	–	25.2/**21.8**	**50.3**/**41.0**	58.1/**47.3**	12.8/**11.5**	28.8/**23.6**
CCA [156] (EB)	–	25.3/ –	–	**59.7**/ –	11.2/ –	–
c-MWP (EB)	0.25	**27.0**/18.4	49.9/35.2	57.7/43.9	**13.2**/ 8.1	**29.4**/20.0

[1] Plummer et al. “Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models.” ICCV, 2015.

Results

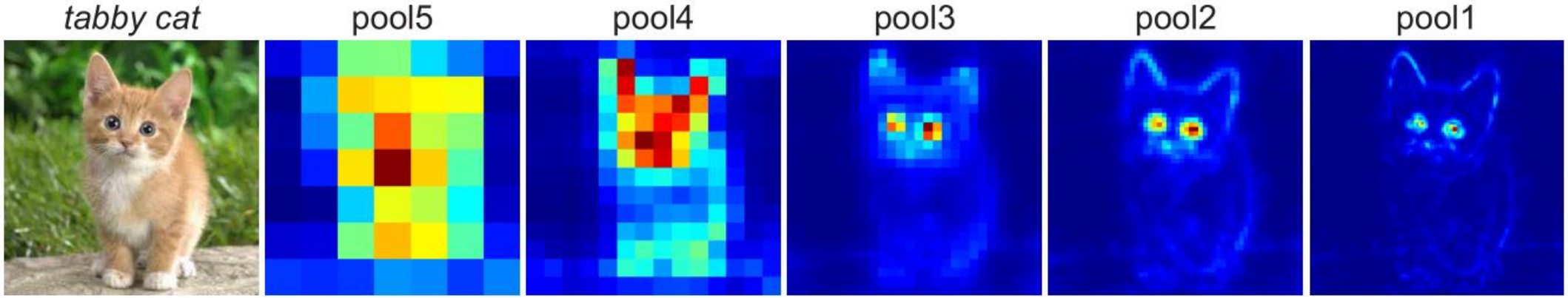
Mean Accuracy over Object Categories in the Pointing Game

	VOC07 Test (All/Diff.)			COCO Val. (All/Diff.)		
	CNN-S	VGG16	GoogleNet	CNN-S	VGG16	GoogleNet
Center	69.5/42.6	69.5/42.6	69.5/42.6	27.7/19.4	27.7/19.4	27.7/19.4
Grad [179]	78.6/59.8	76.0/56.8	79.3/61.4	38.7/30.1	37.1/30.7	42.6/36.3
Deconv [223]	73.1/45.9	75.5/52.8	74.3/49.4	36.4/28.4	38.6/30.8	35.7/27.9
LRP [11]	68.1/41.3	-	72.8/50.2	32.5/24.0	-	40.2/32.7
CAM [235]	-	-	80.8/61.9	-	-	41.6/35.0
MWP	73.7/52.9	76.9/55.1	79.3/60.4	35.0/27.7	39.5/32.5	43.6/37.1
c-MWP	**78.7/61.7**	**80.0/66.8**	**85.1/72.3**	**43.0/37.0**	**49.6/44.2**	**53.8/48.3**

Excitation Backprop

Assumptions:

- The responses of the activation neurons are non-negative.
- An activation neuron is tuned to detect certain visual features. Its response is positively correlated to its confidence of the detection.

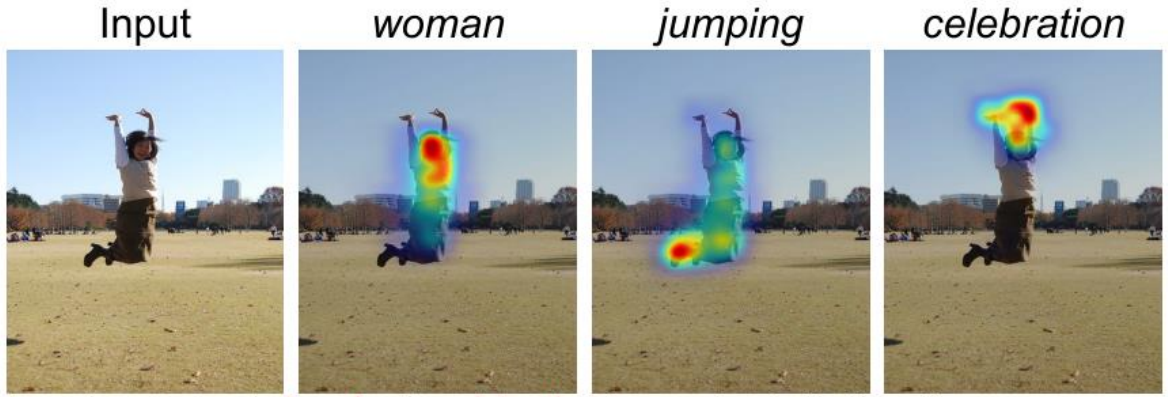


Running excitation backprop, we can extract attention maps from different layers.
Lower layers can generate maps that highlight features of smaller scale.

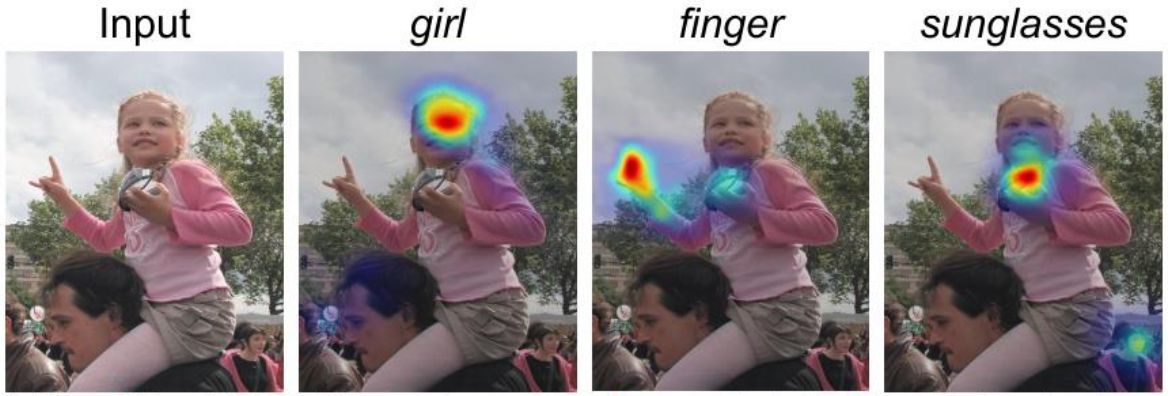
Example Results

A young lady wearing blue and black is running past an orange cone.

A woman sits with a boy in an orange hat with a cookie in his hand as he makes a funny face.



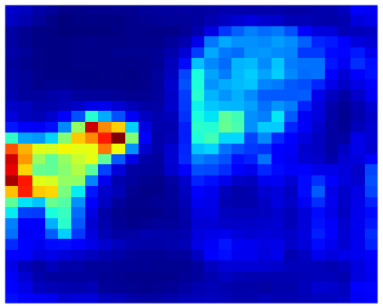
An asian woman is jumping in celebration in a park outside the city.



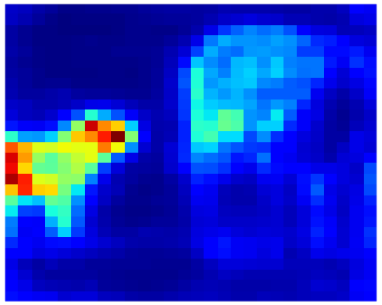
A little girl in pink is holding up her pointer finger and pinkie finger in the air while holding her sunglasses in the other hand.

Contrastive Attention

zebra

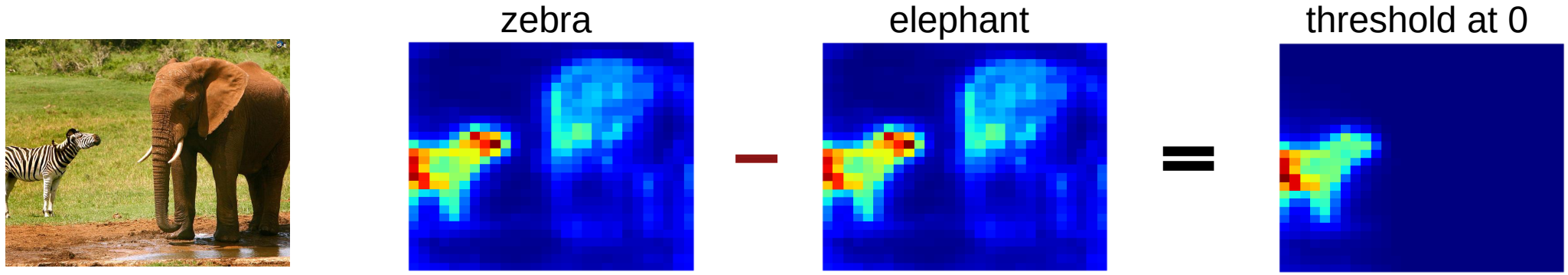


elephant



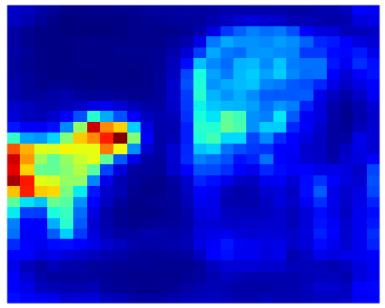
threshold at 0

Contrastive Attention

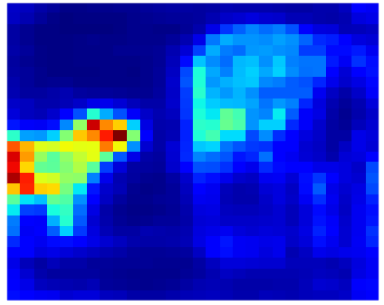


Contrastive Attention

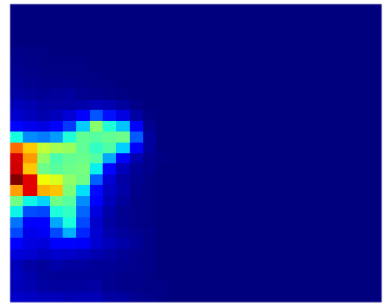
zebra



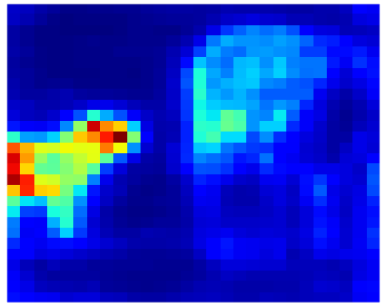
elephant



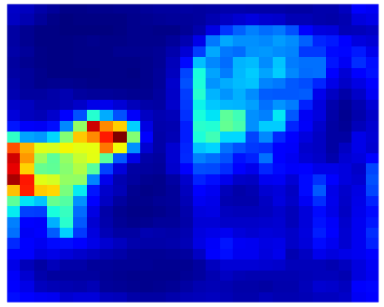
threshold at 0

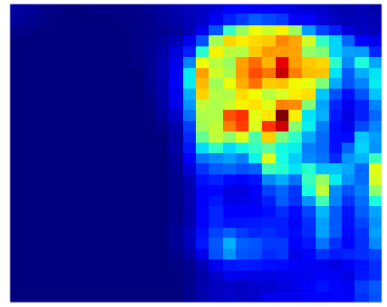


elephant

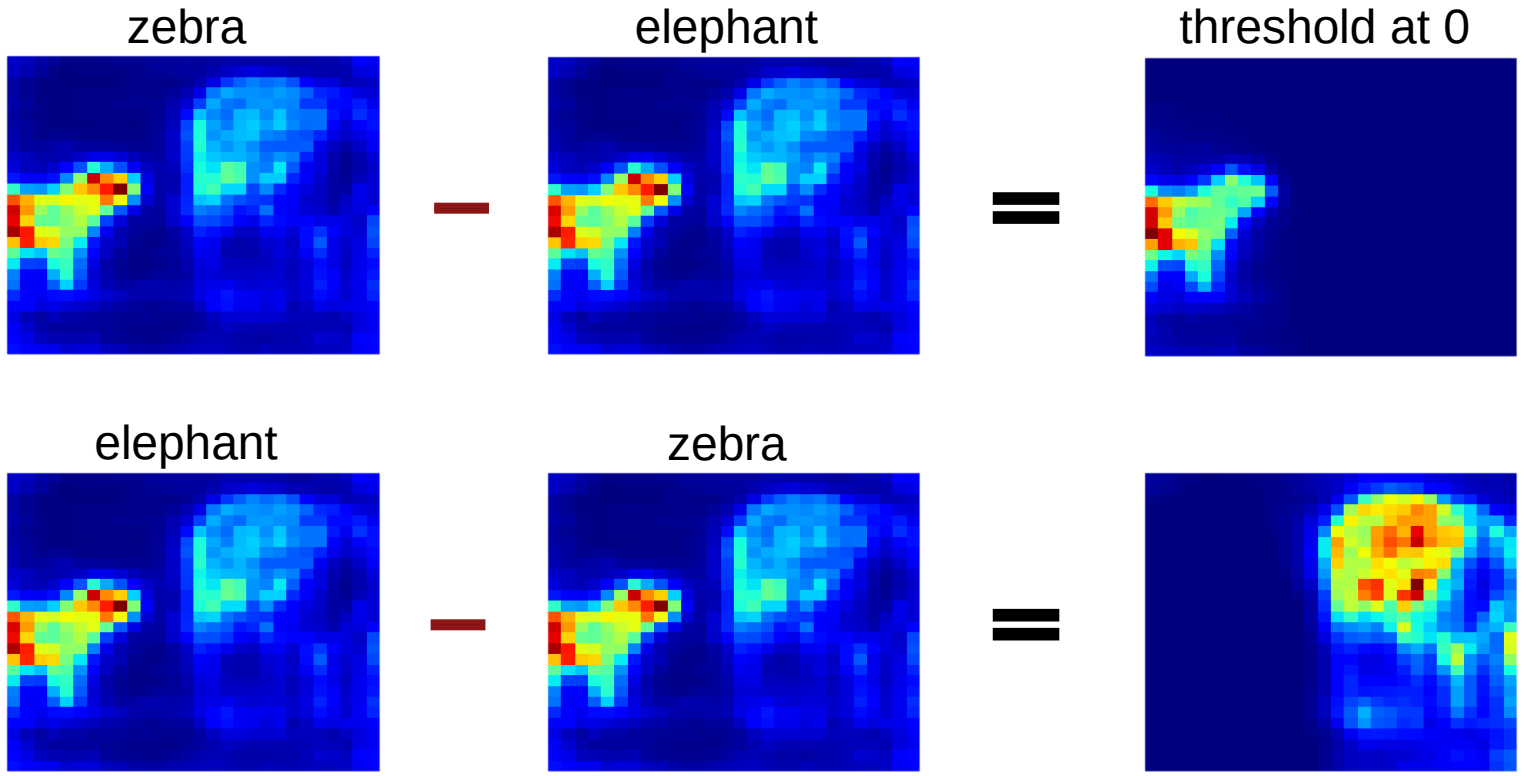


zebra





Contrastive Attention



- The pair of maps are well normalized using our probabilistic framework
- Contrastive attention map can be computed by a single pass