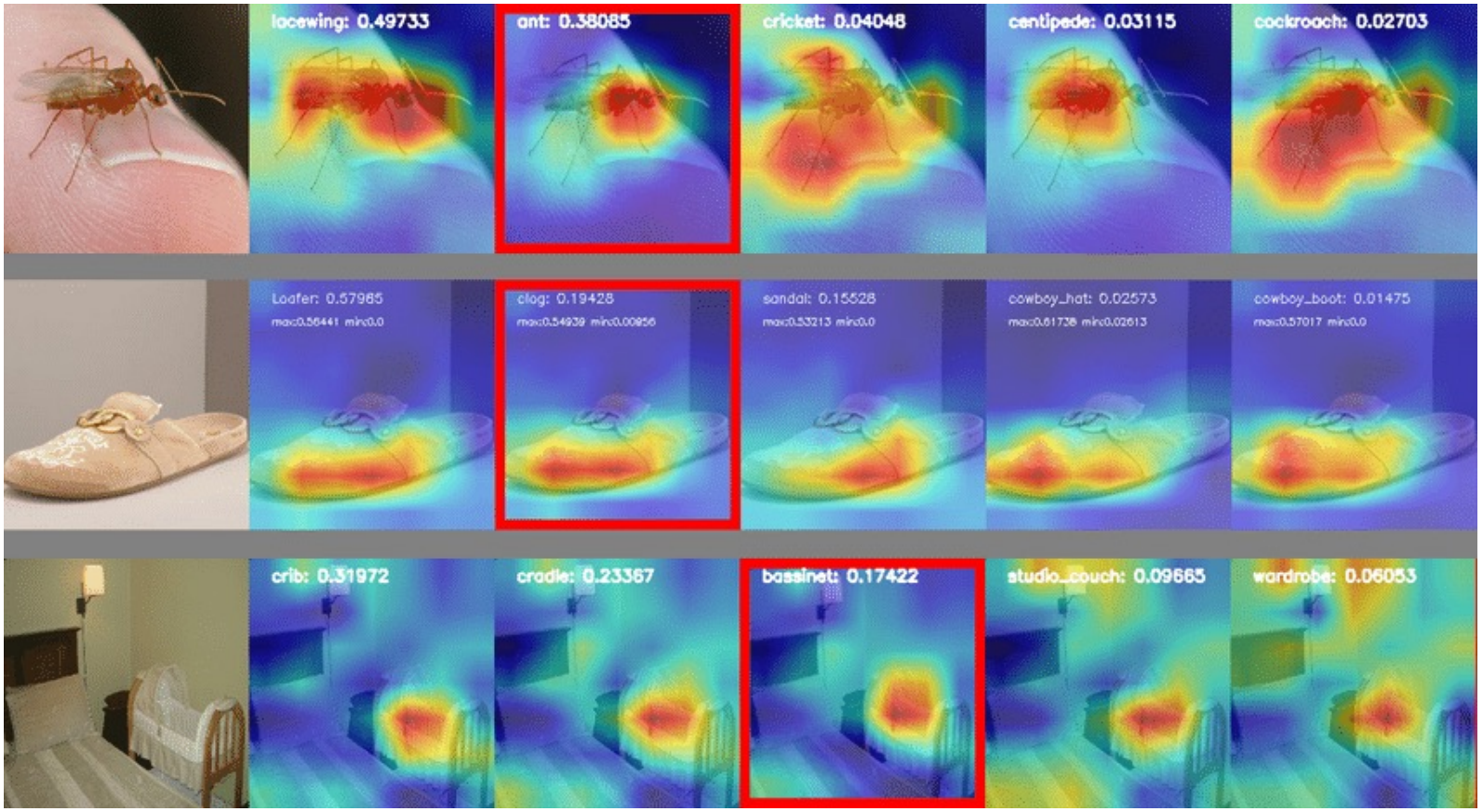
Interpretable AI



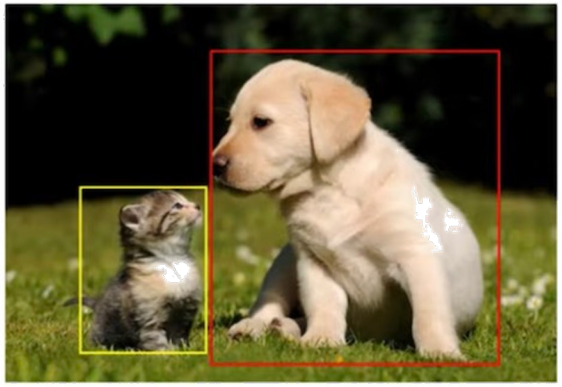
Computer Vision

Fall 2022, Lecture 20

Is this a dog?

Image Classification

What is there in image and where?



Object Detection

Which pixels belong to which object?

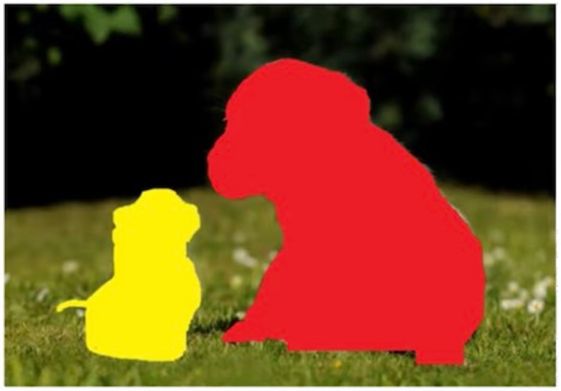
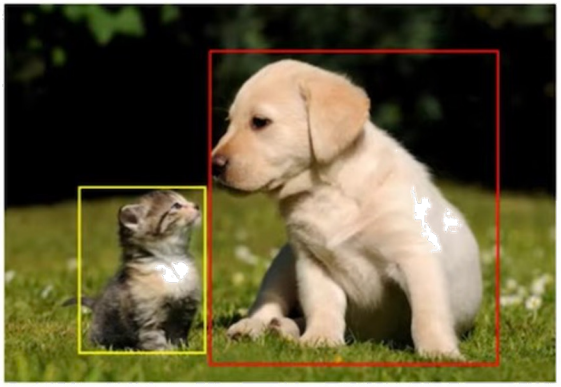


Image Segmentation

Is this a dog?

Image Classification

What is there in image and where?



Object Detection

Which pixels belong to which object?

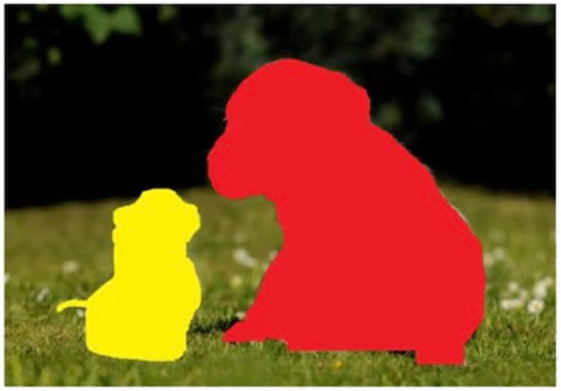
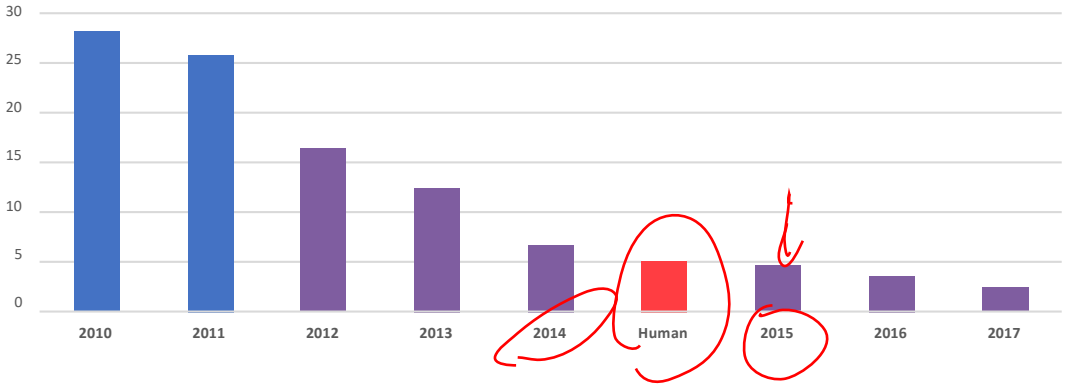


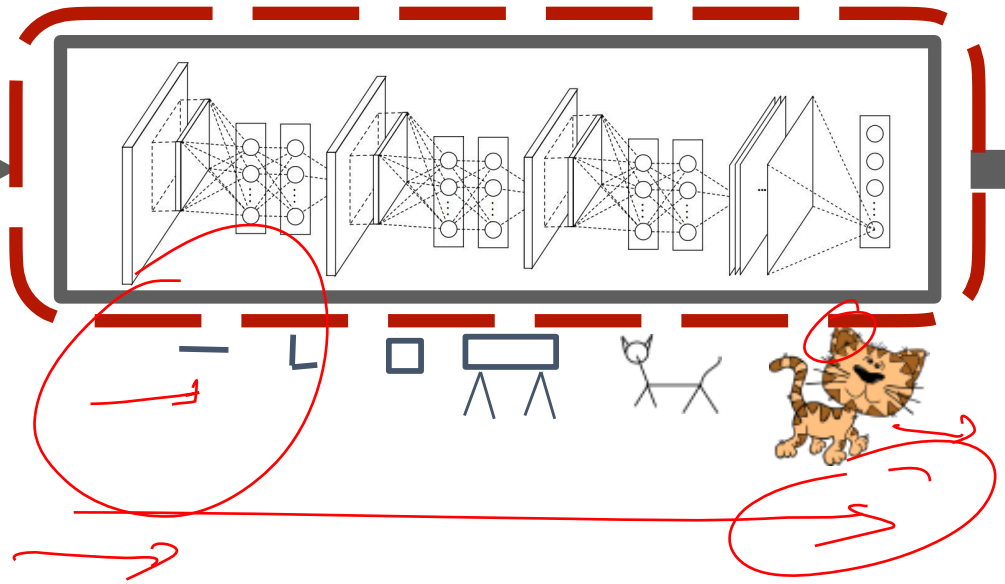
Image Segmentation

ILSVRC top-5 Error on ImageNet

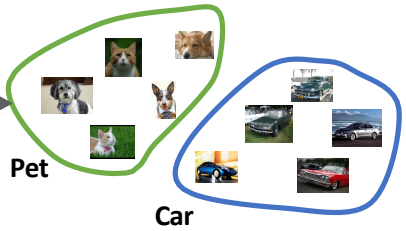


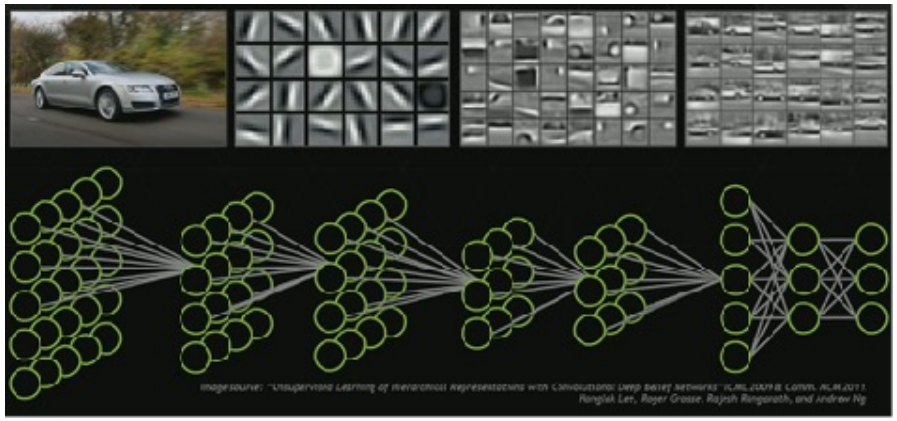
Why Do We Like Deep Learning?

High-dim input



Learned embedding



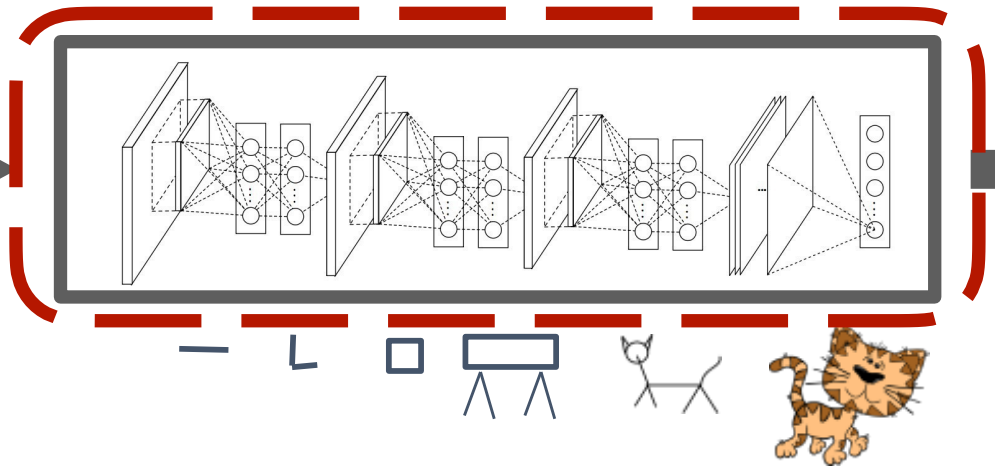


→ **Meaningful** data representations

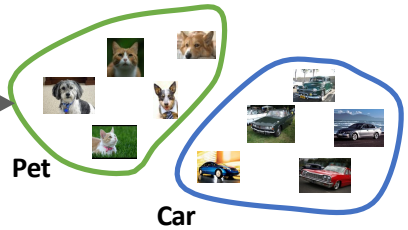
[NVIDIA GTC, 2019]

Why Do We Like Deep Learning?

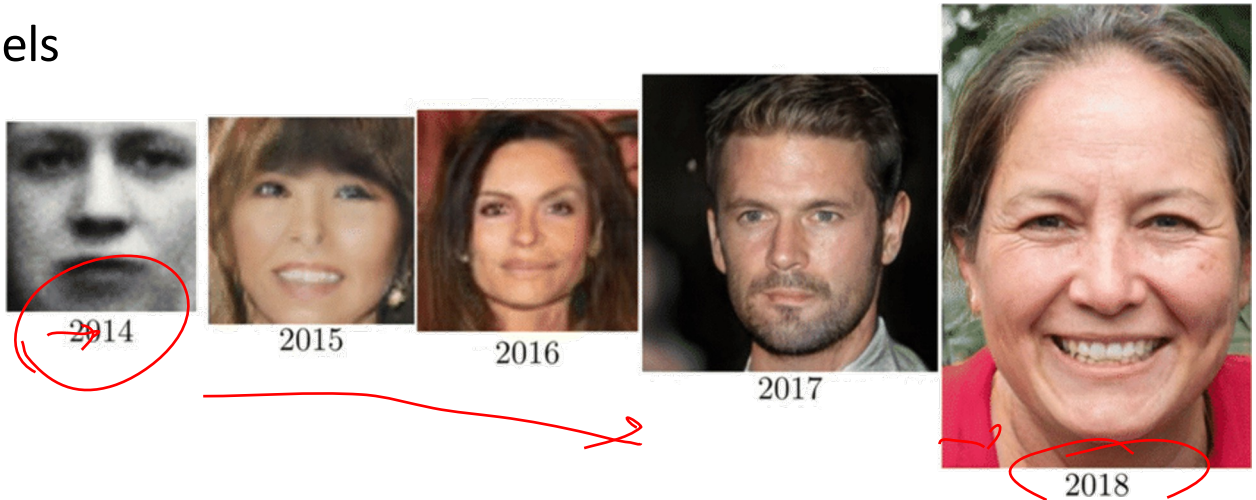
High-dim input



Learned embedding



→ **Better** generative models




REAL
FAKE

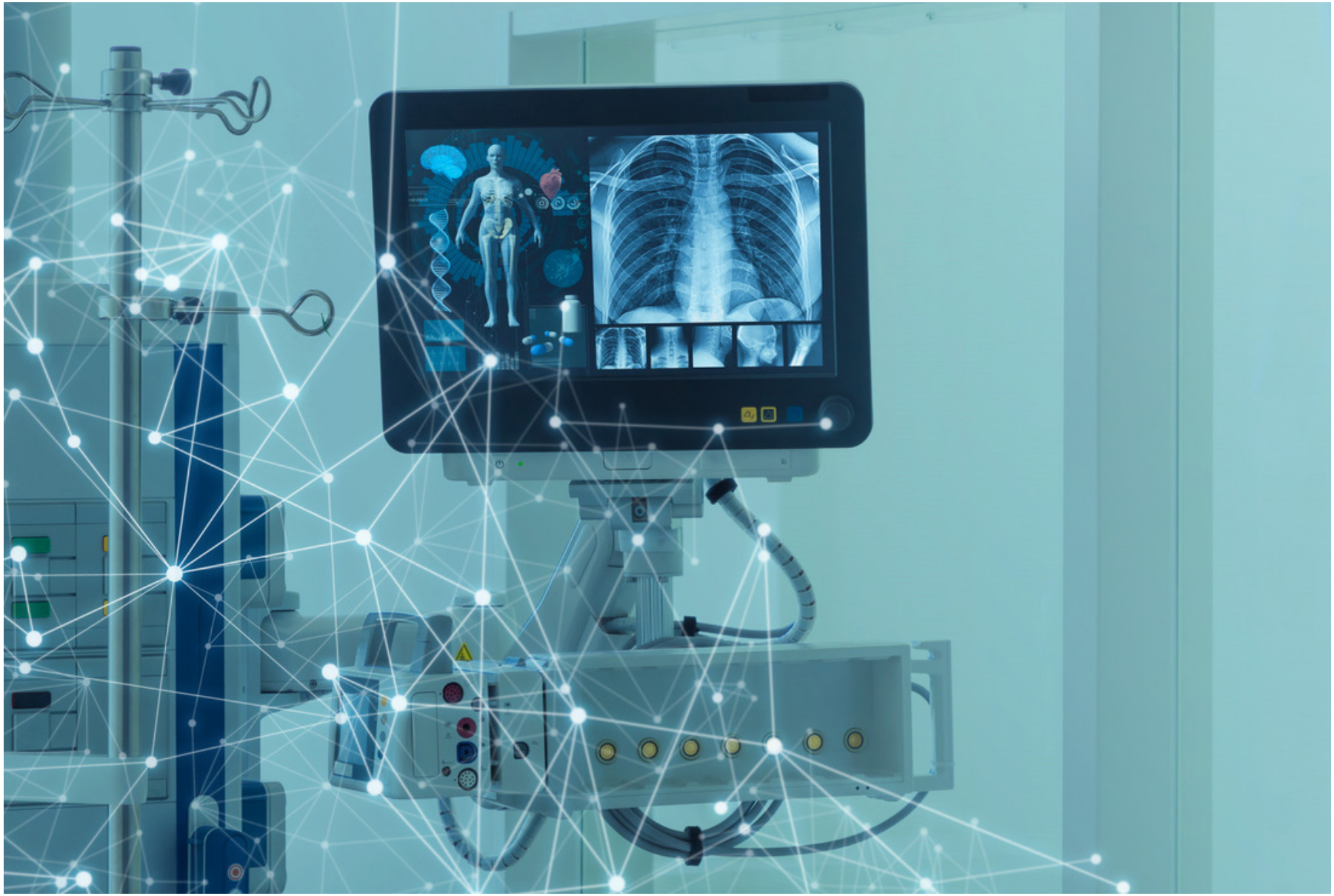
A small cactus wearing a straw hat and neon sunglasses in the Sahara desert.

A small cactus wearing a straw hat and neon sunglasses in the Sahara desert.

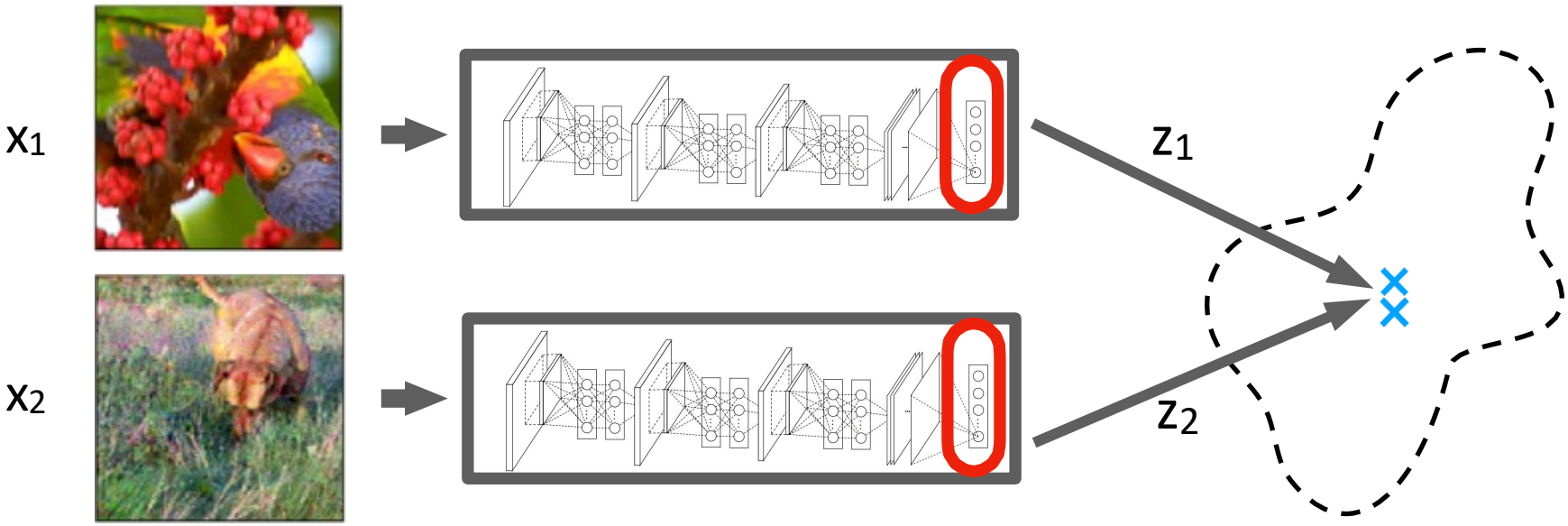
https://imagen.research.google/

Unfortunately...

Unfortunately...

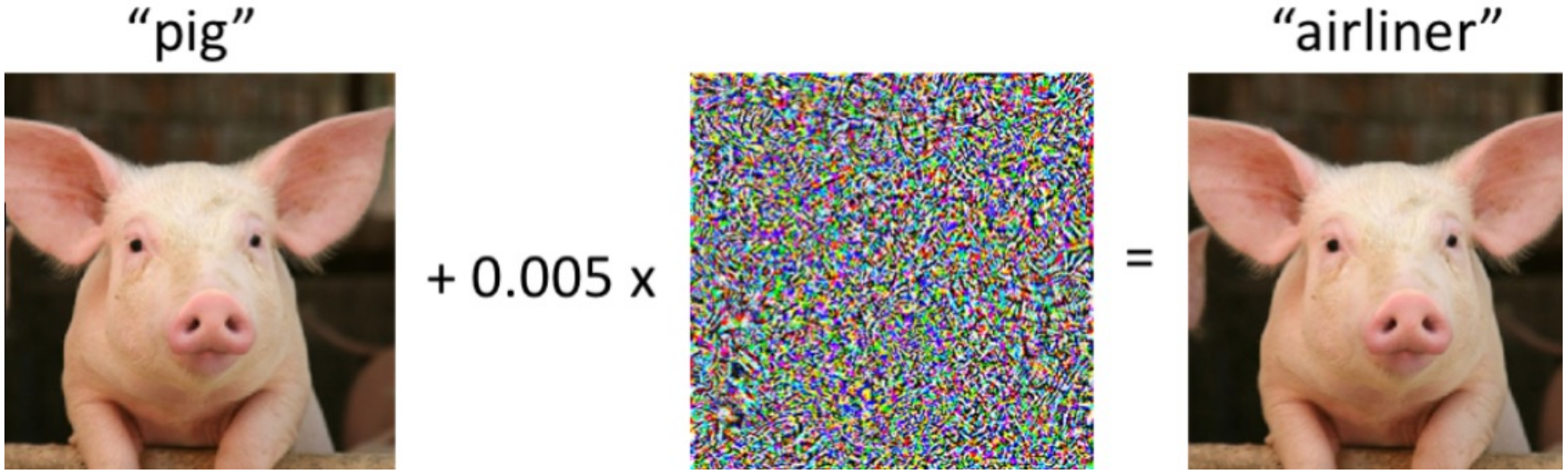


Story Behind



$x_1 \neq x_2$ but $z_1 \approx z_2$

Adversarial Examples

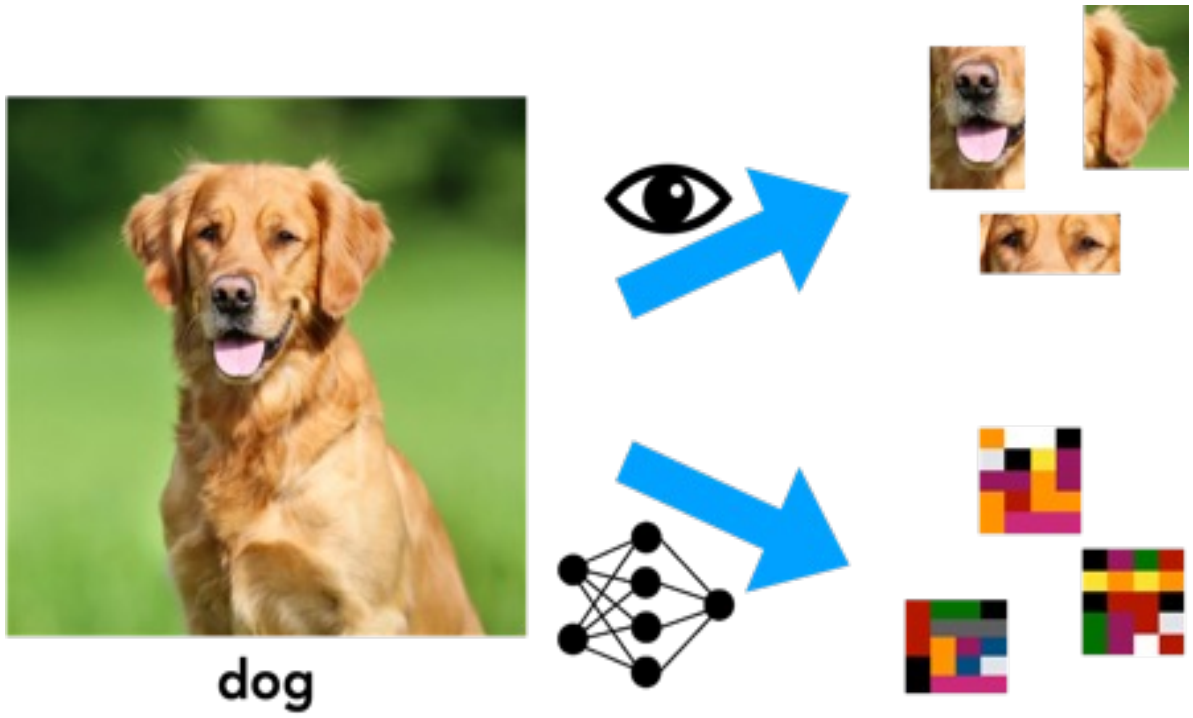


The image above shows an example of an adversarial attack, wherein adding a little bit of noise completely changes the class of the image. **A pig**, that was correctly classified as **a pig** by the classifier, is now classified as **an airliner** after the attack. The noise added looks random but causes the network to classify the image as an ‘airliner’.

AI and Human Comparison

dog

AI and Human Comparison



These are **equally valid** classification methods

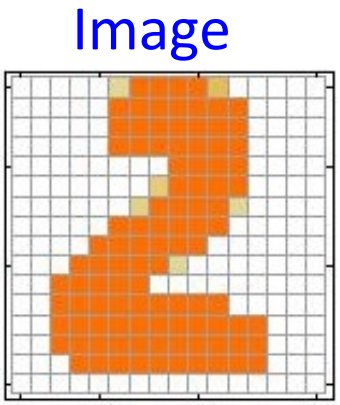
Why Care About Interpretability?

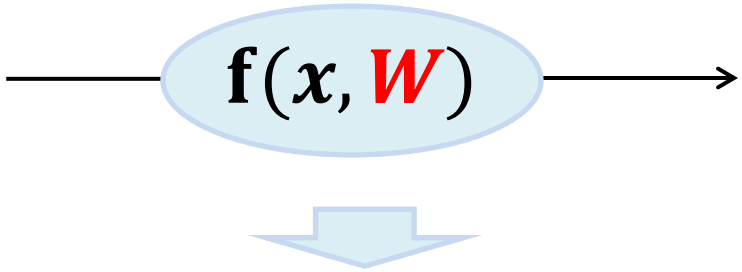
1. Help building trust:
    - Humans are reluctant to use AI for critical tasks
    - Fear of unknown when people confront new technologies

Why Care About Interpretability?

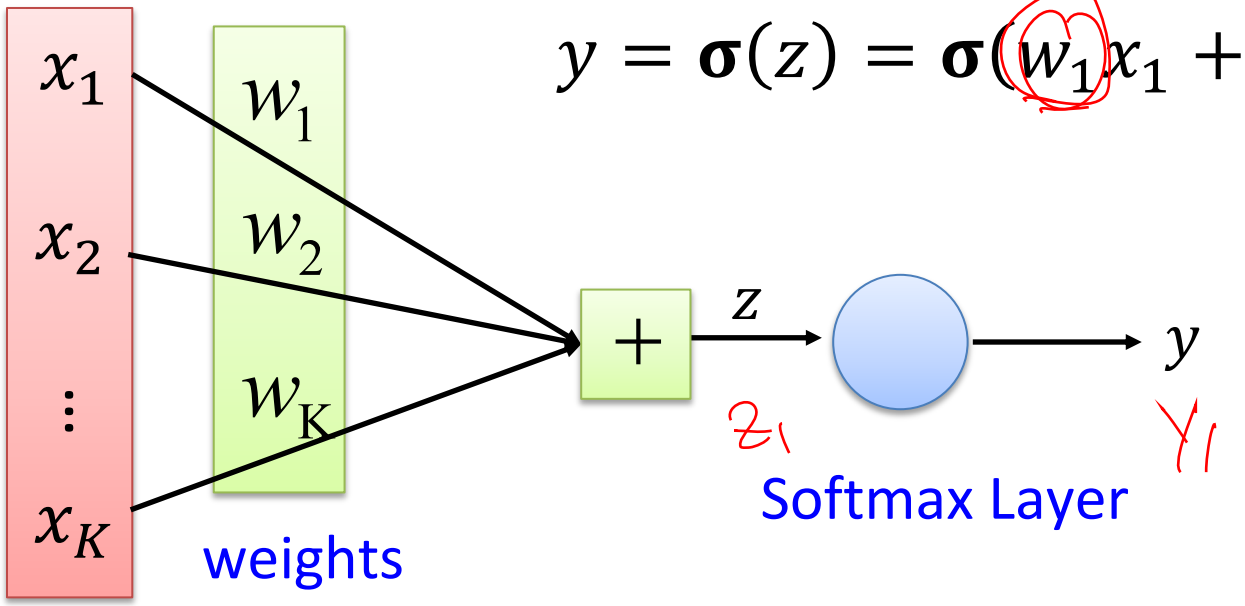
1. Help building trust:
    - Humans are reluctant to use AI for critical tasks
    - Fear of unknown when people confront new technologies
2. Promote safety:
    - Explain model's representation (i.e., important feature) providing opportunities to remedy the situation
3. Allow for contestability:
    - Black-box models don't decompose the decision into sub-models or illustrate a chain of reasoning

Linear Classifier – Score Function

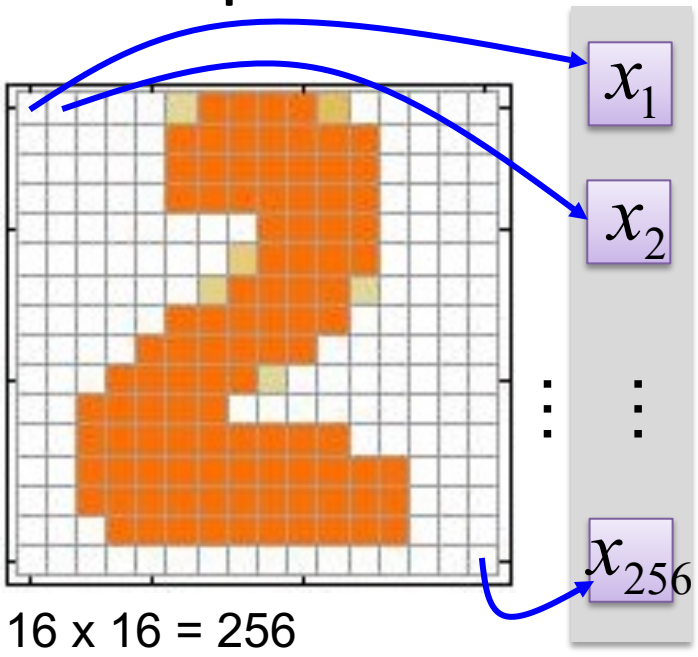


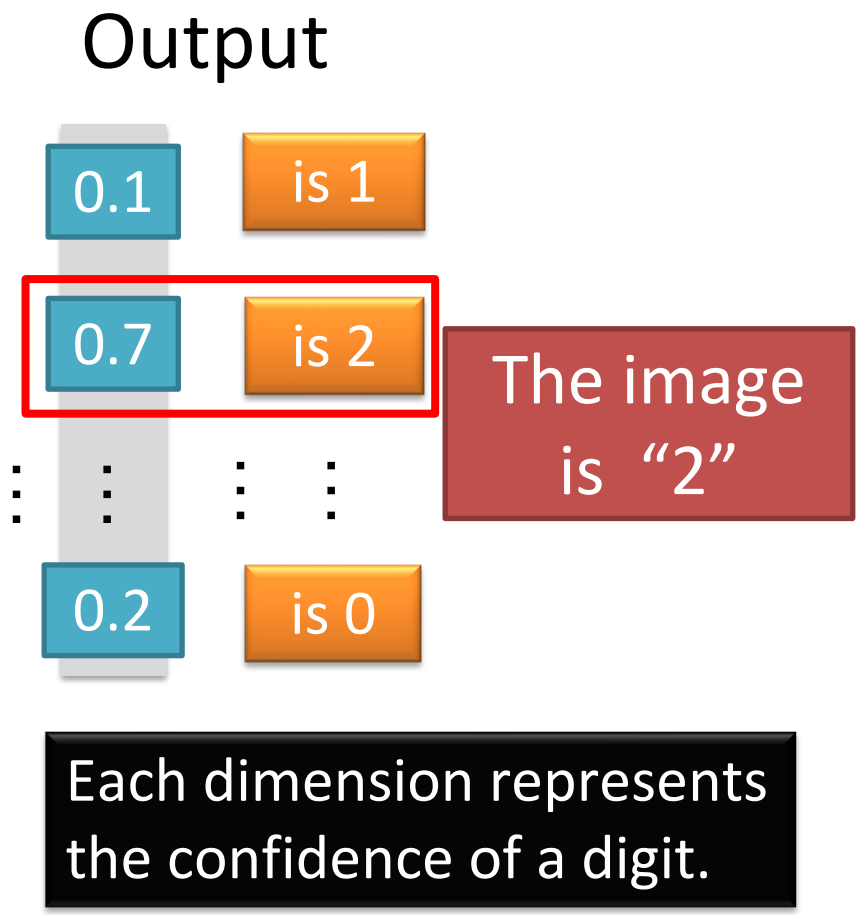


$$y = \sigma(z) = \sigma(w_1 x_1 + \cdots + w_K x_K)$$



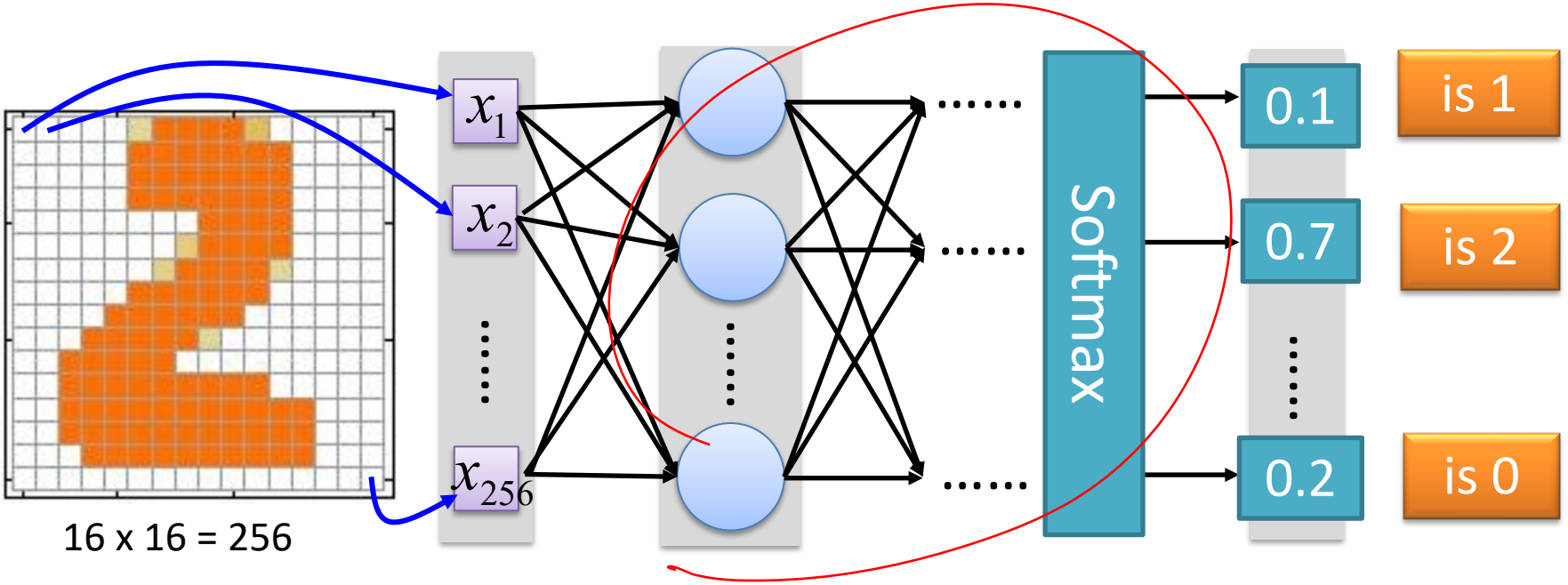
Linear Classifier – Score Function





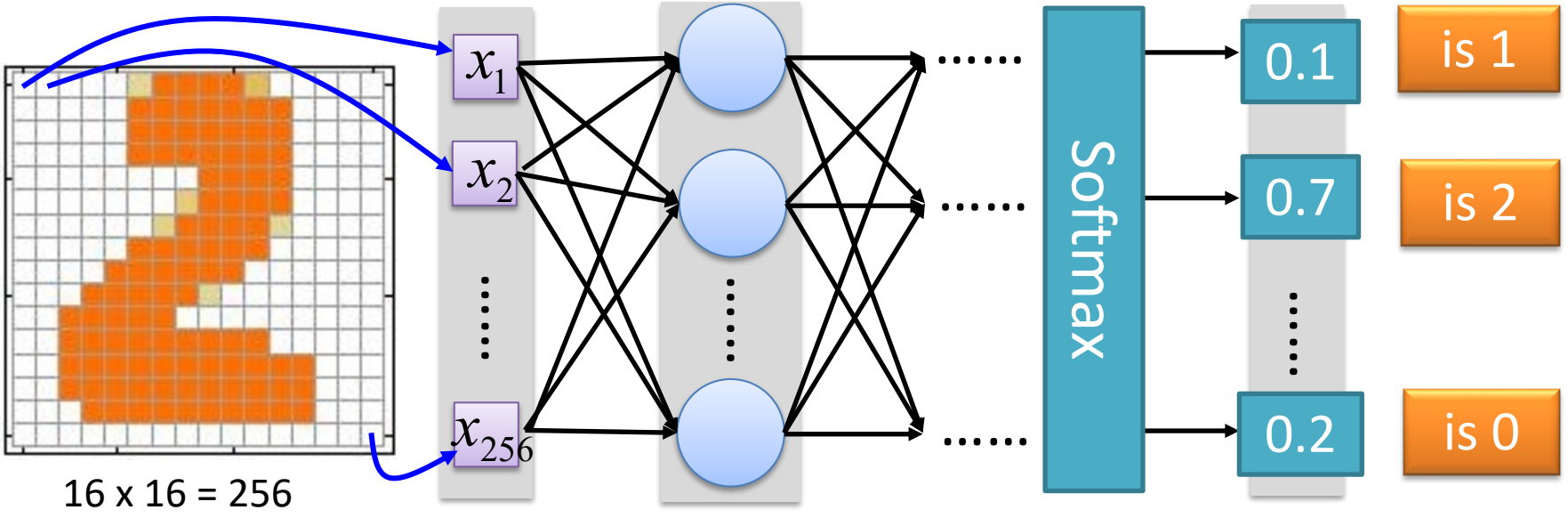
How about multi-layer perceptron?

$$\theta = \{W^1, b^1, W^2, b^2, \cdots W^L, b^L\}$$



How about multi-layer perceptron?

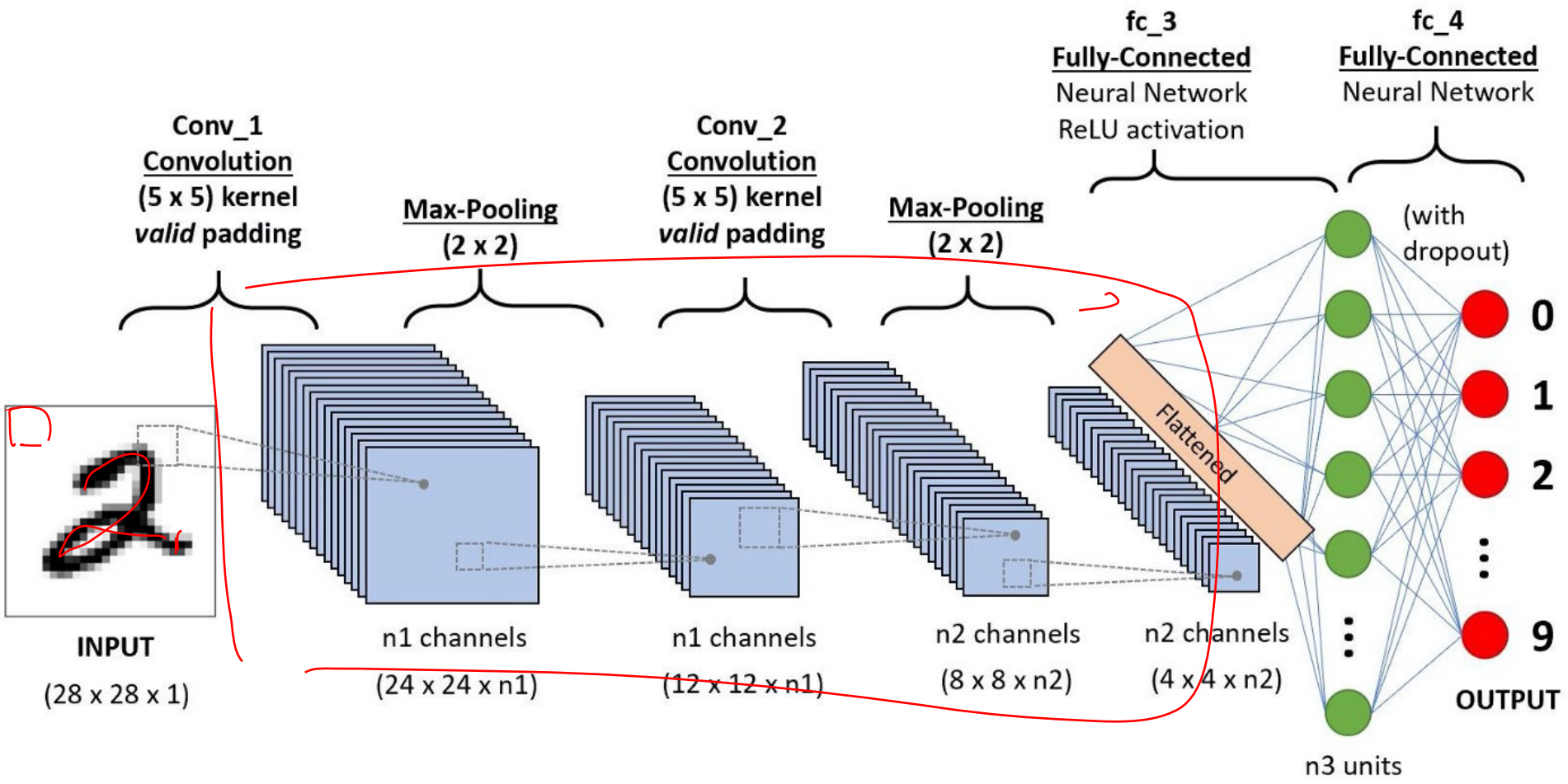
$$\theta = \{W^1, b^1, W^2, b^2, \cdots W^L, b^L\}$$



$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n$$

$$= f(0) + f'(0)\,x + \frac{f''(0)}{2!}x^2 + \frac{f'''(0)}{3!}x^3 + \cdots$$

How about ConvNets?



A Closer Look from Features

[Ilyas Santurkar Tsipras Engstrom Tran M 2019]

…depend unintuitively on linear directions [Jetley et al 2018]

…can be invariant to task-relevant features [Jacobsen et al 2019]

…depend too much on texture [Geirhos et al 2019]

…rely heavily on backgrounds [Xiao et al 2020]

…can learn from high-frequency components [Yin et al 2019]

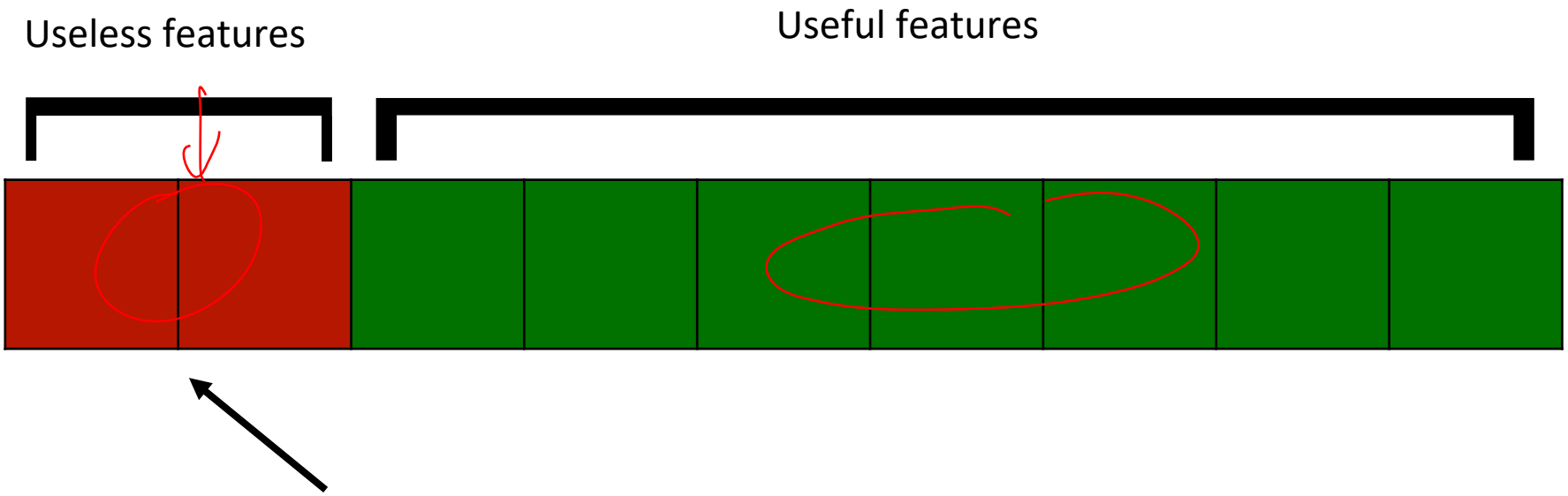
…can rely too much on image statistics [Jo & Bengio 2017]

A Closer Look from Features

http://ai.bu.edu/M3SDA/

A Closer Look from Features

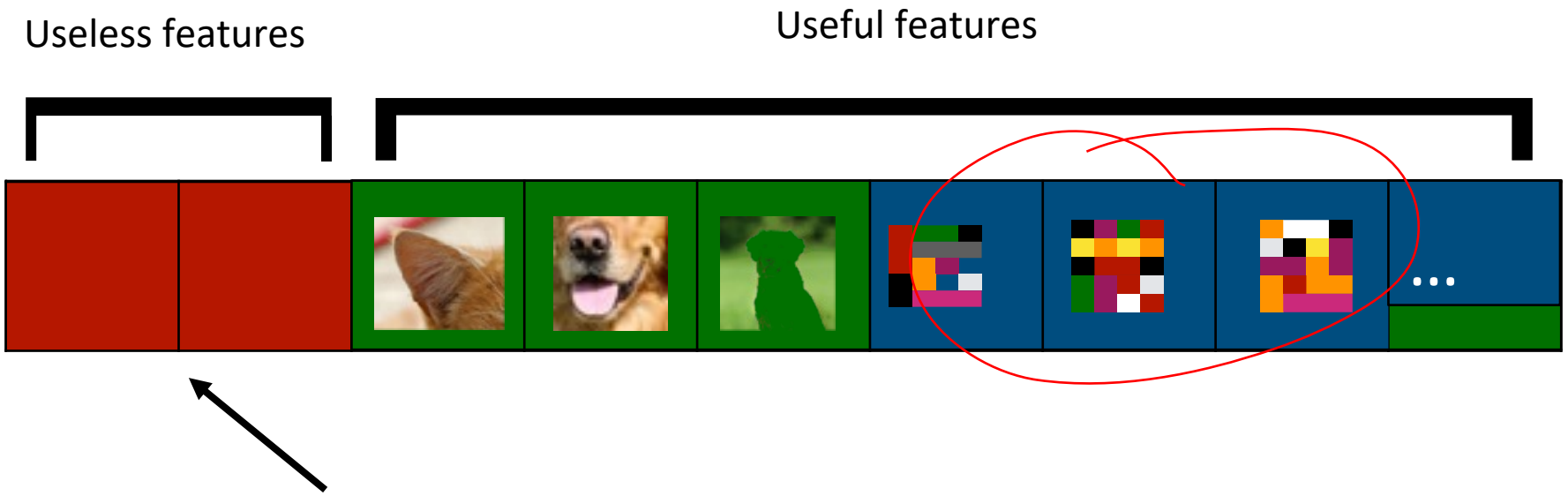
[Ilyas Santurkar Tsipras Engstrom Tran **M** 2019]



Natural hypothesis: Adv. examples correspond to manipulating these

A Closer Look from Features

[Ilyas Santurkar Tsipras Engstrom Tran **M** 2019]



Natural hypothesis: Adv. examples correspond to manipulating these

A Closer Look from Features

[Ilyas Santurkar Tsipras Engstrom Tran M 2019]

Useless features

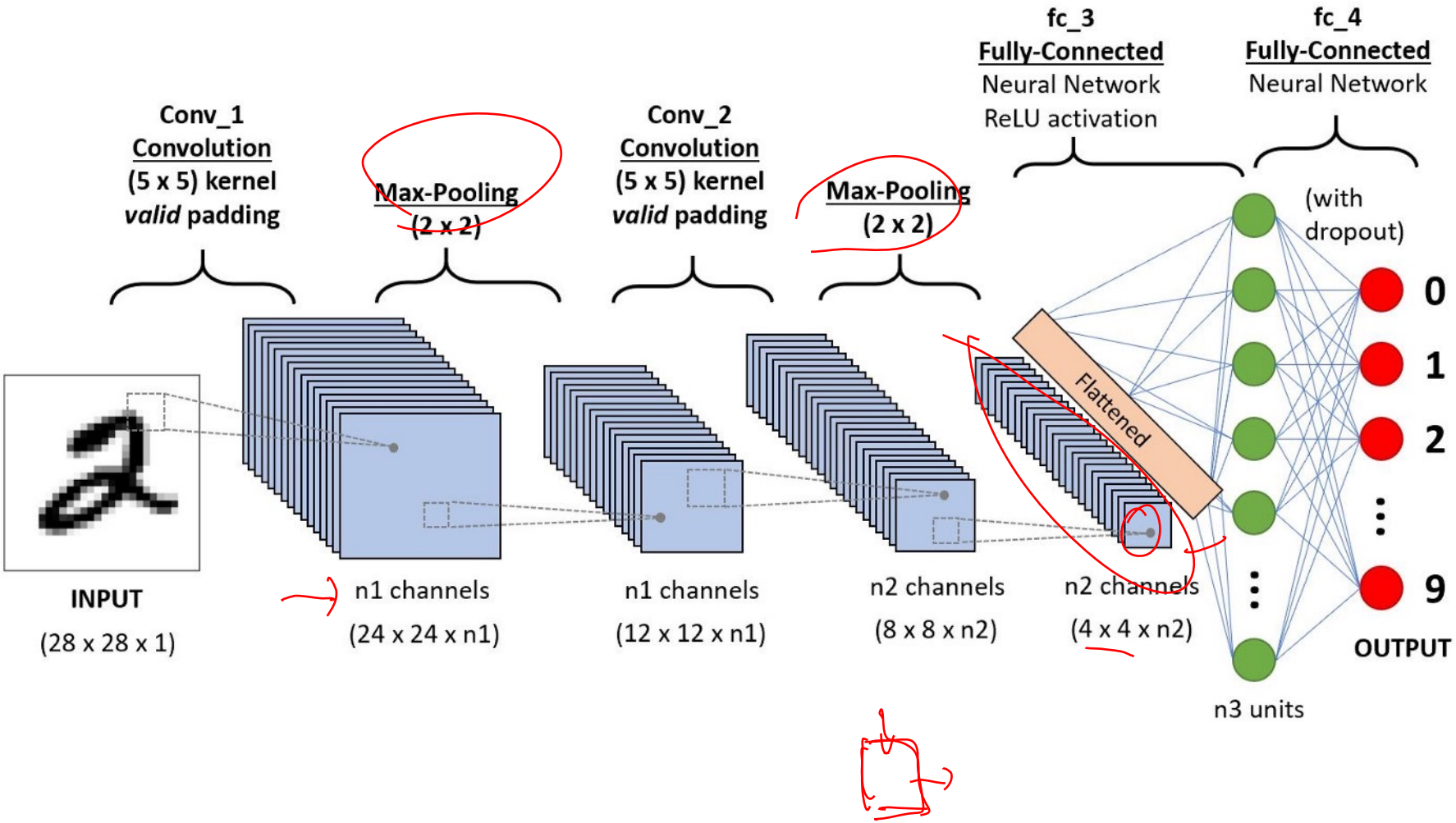
Robust features
Correlated with label even when perturbed

Non-robust features
Correlated with label, but can be flipped via perturbation

Natural hypothesis: Adv. examples correspond to manipulating these

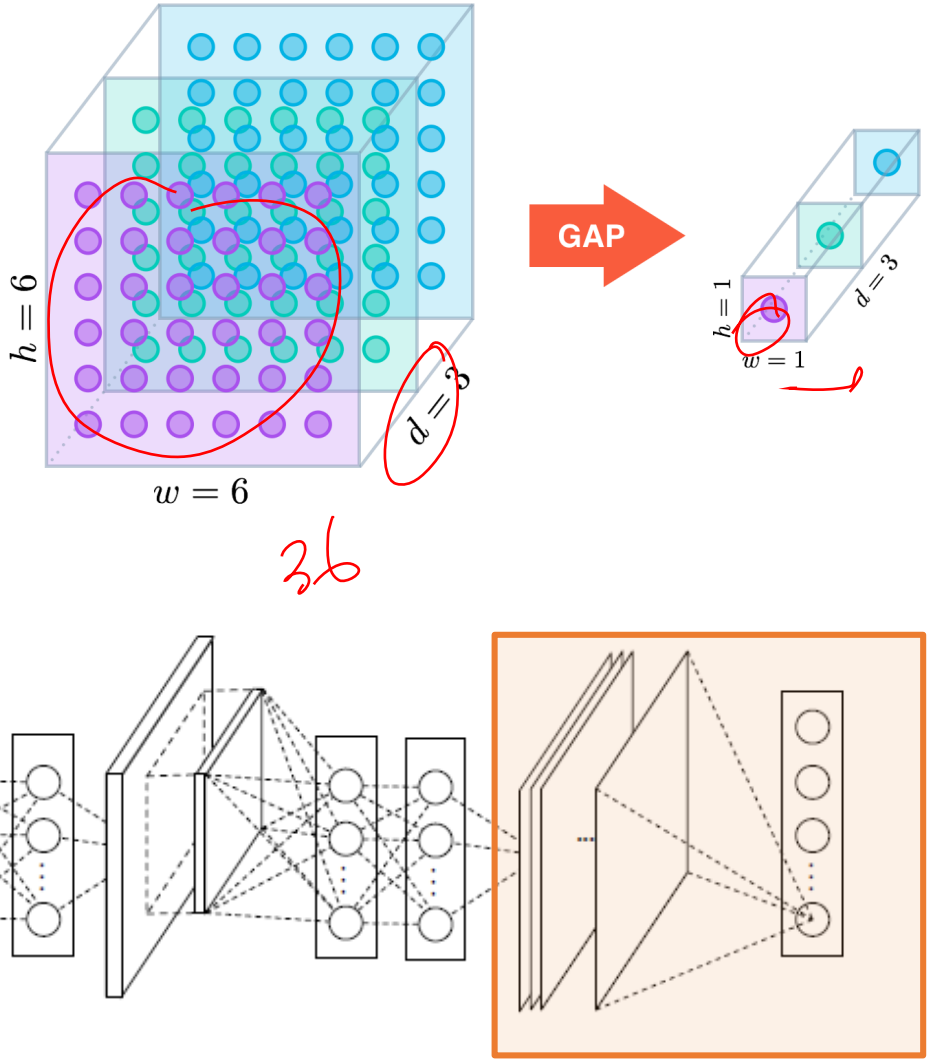
Actually: Adv. examples largely correspond to manipulating these

Revisit Convolutional Networks



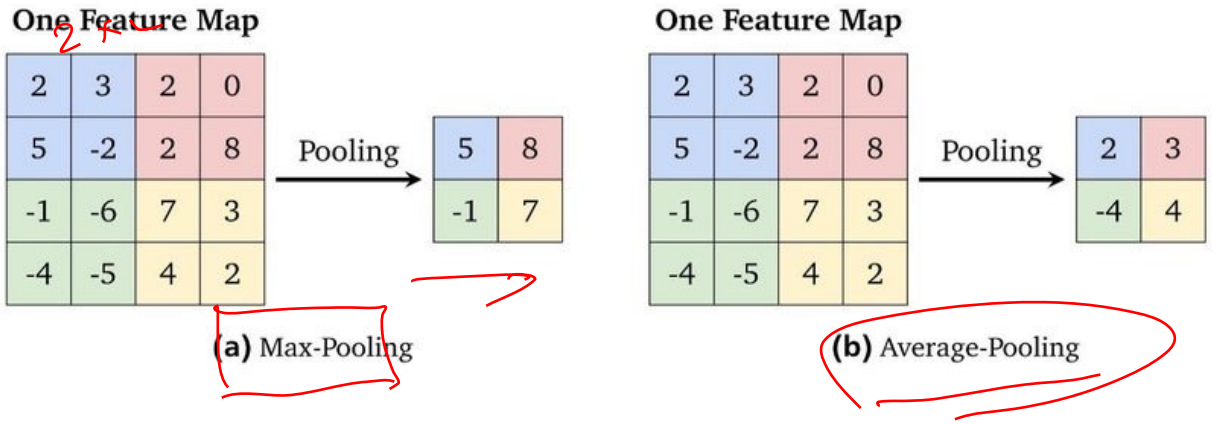
Global Average Pooling

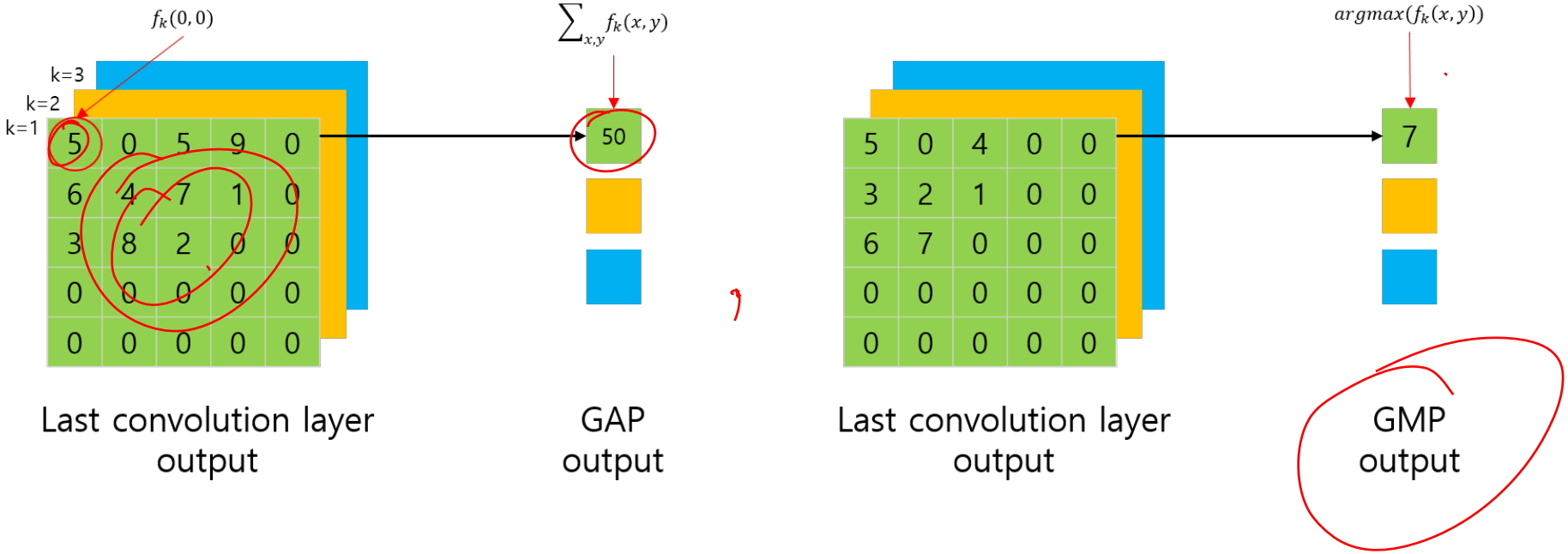
- The Network in Network (NIN) employed the GAP first for replacement of the FC layer.
- Without the FC layer, It shows the fine performance for the classification task.



M. Lin, Q. Chen, and S. Yan. Network in network. International Conference on Learning Representations, 2014

Comparison





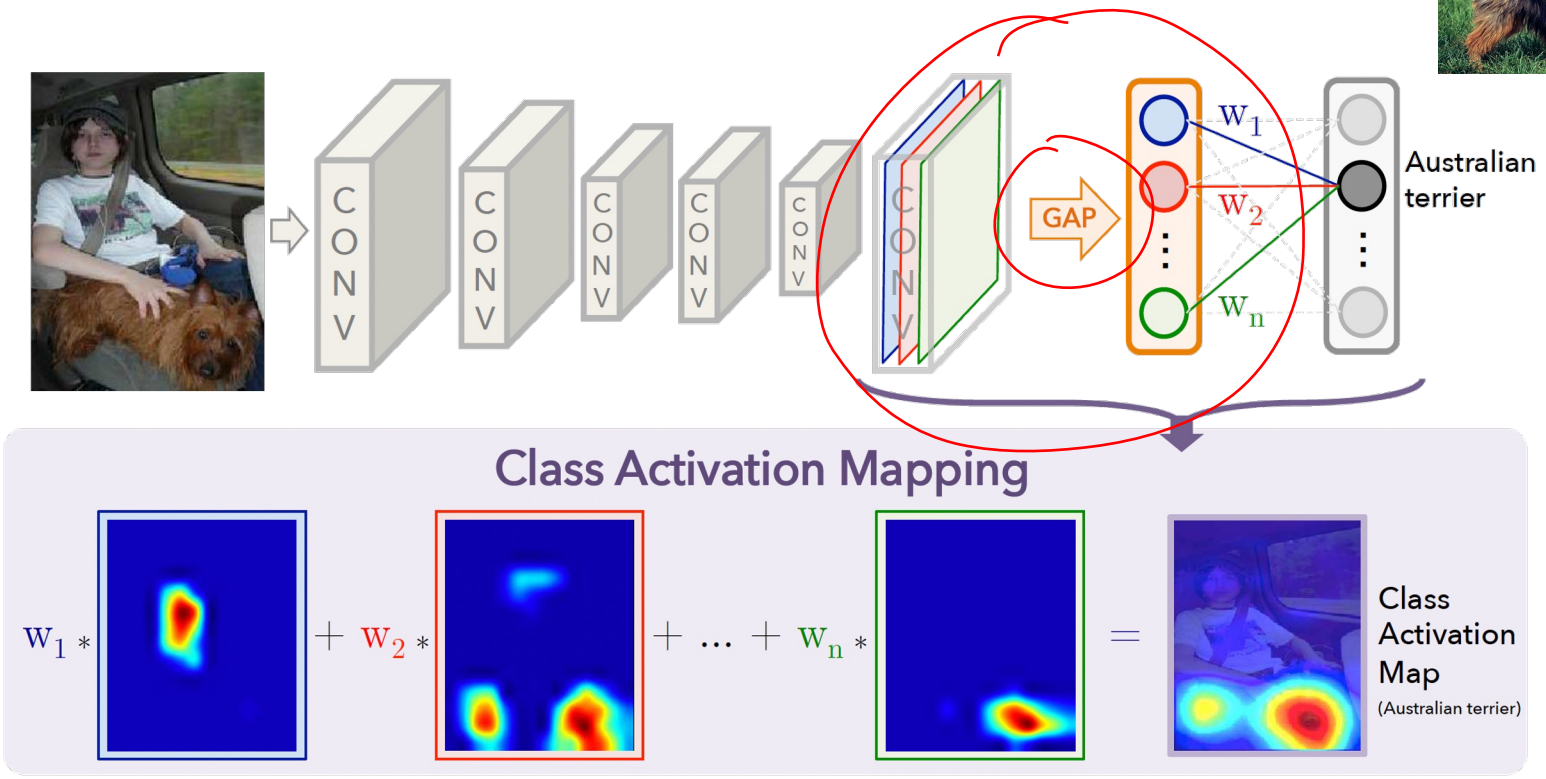
M. Lin, Q. Chen, and S. Yan. Network in network. International Conference on Learning Representations, 2014

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

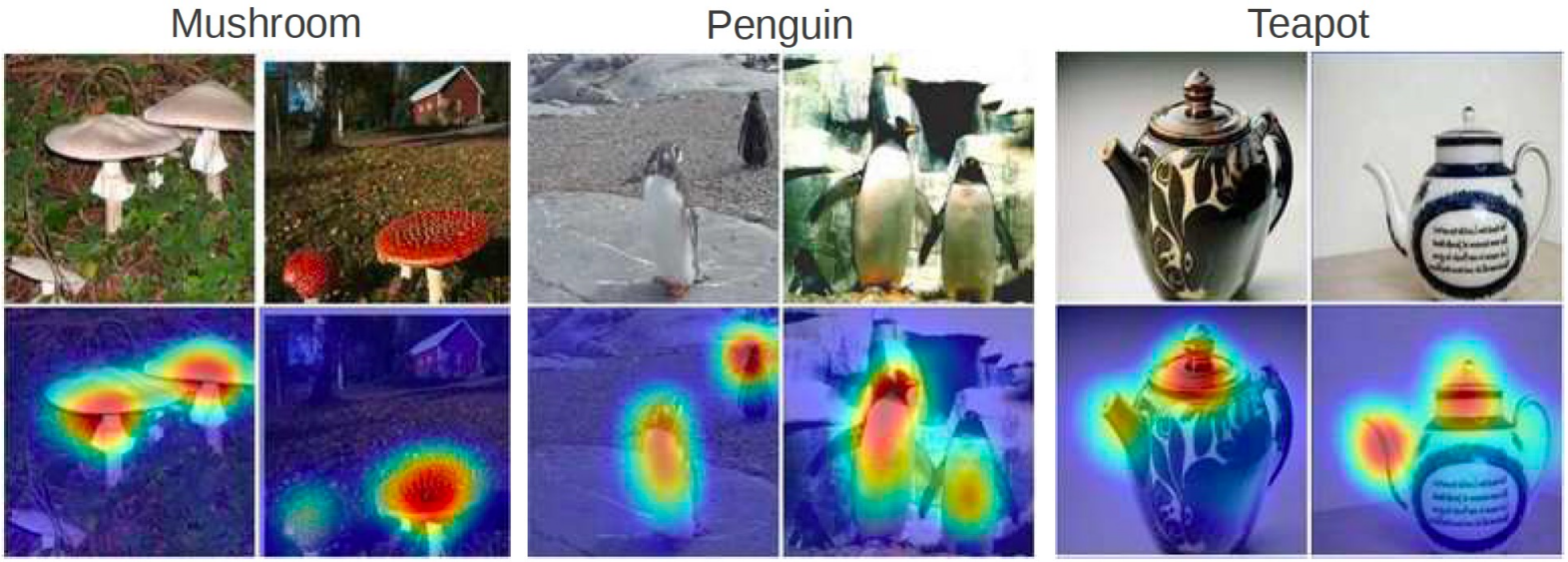
Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

Class Activation Maximization (CAM)

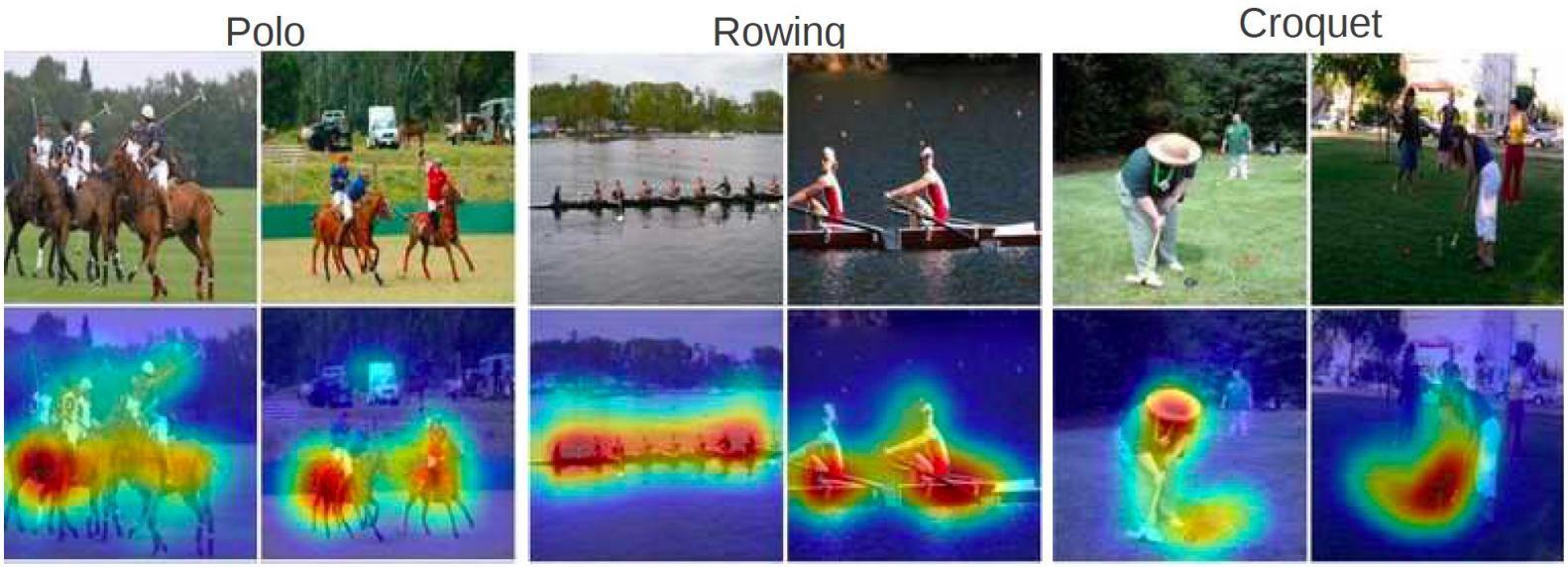


- Rid off the FC layer and attach the GAP layer right after the conv-layer & retrain.
- Calculate the weight of each image feature by Global Average Pooling (GAP).
- Conduct the weighted sum for every feature and normalize it.

Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.



Caltech256



UIUC Event8

Gradient Class Activation Maximization (Grad-CAM)

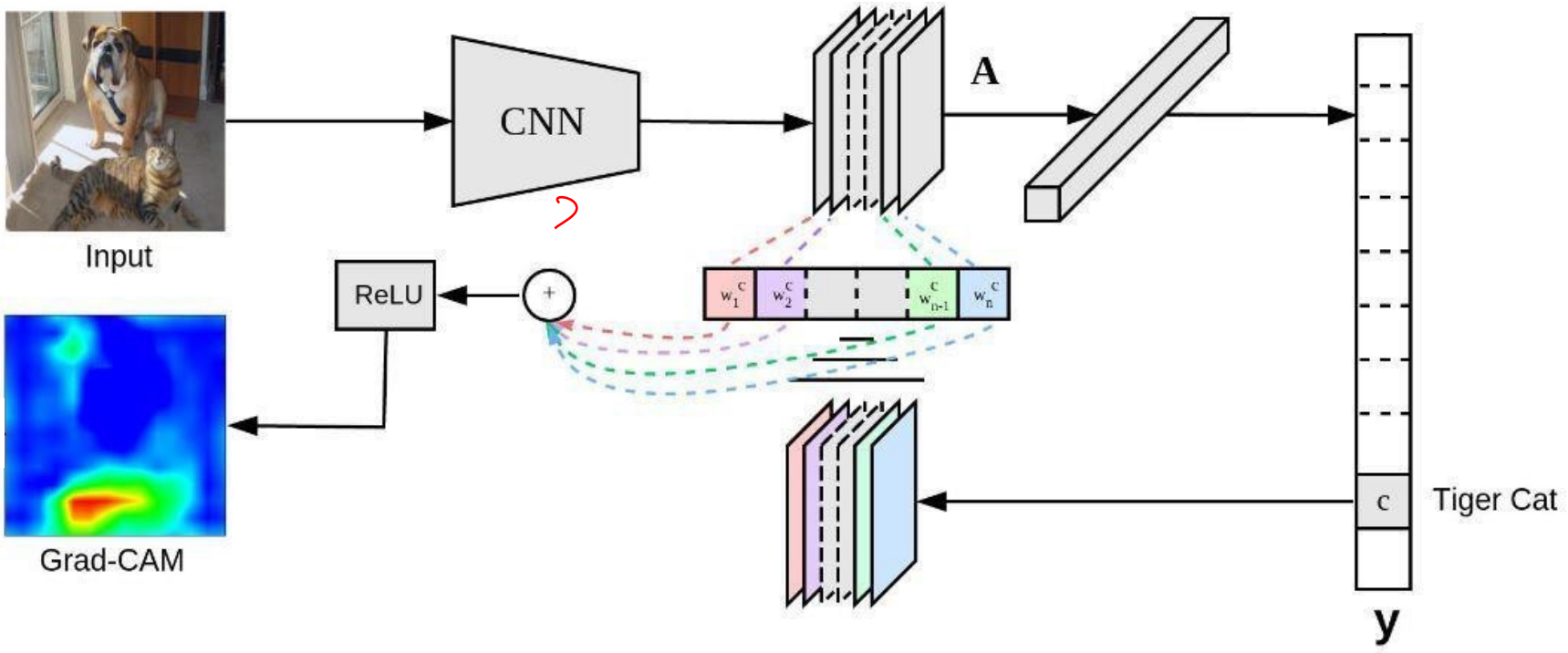
Cons for neat CAM (Solved by the Grad CAM)

- This method is only for the classification task. Therefore, it couldn't be applied on the other tasks such as segmentation.
- FC removal is required which brings the performance drop & GAP network has different structure from the original network.
- Grad-CAM can be applied for the Classification, Captioning, Question answering…

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

Grad-CAM

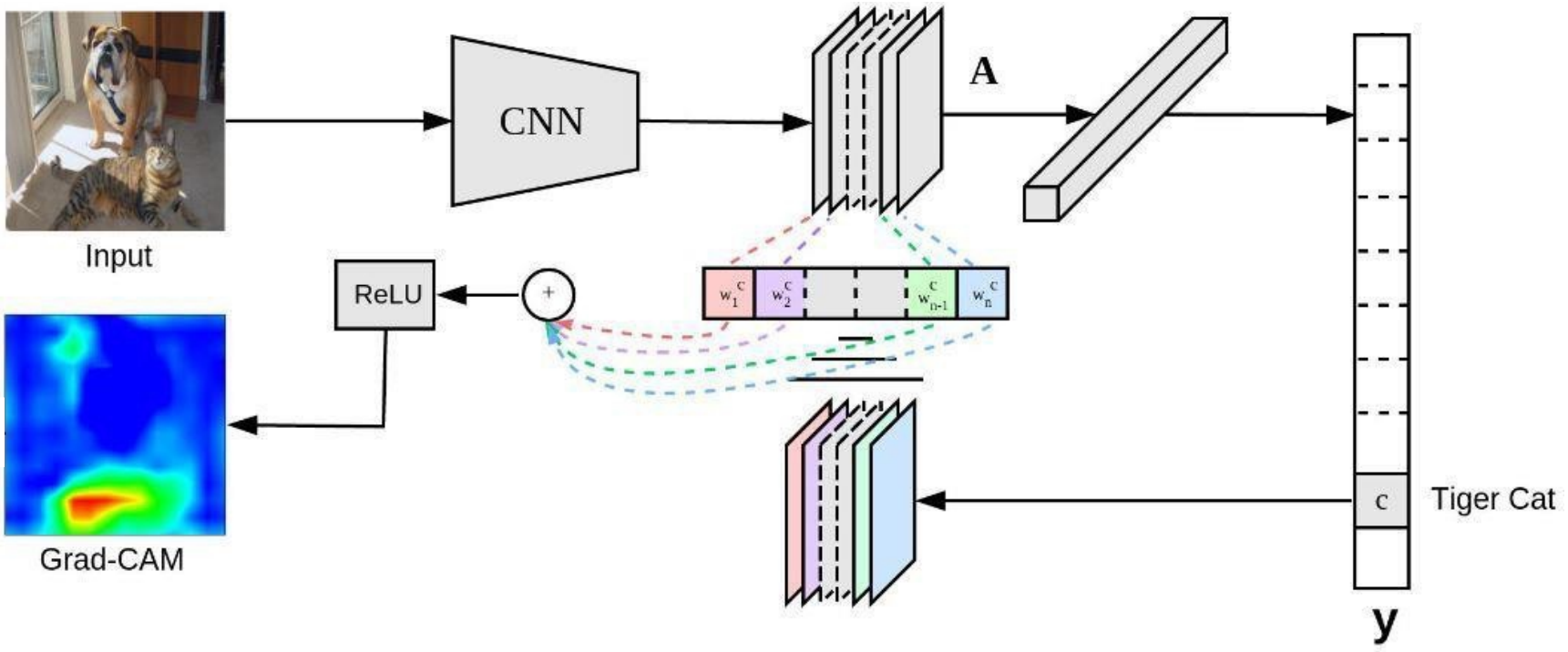
- Instead of the GAP, Grad CAM gets the gradient value for the weight.
- Grad-CAM does not need to rid off the FC layer.



Grad-CAM

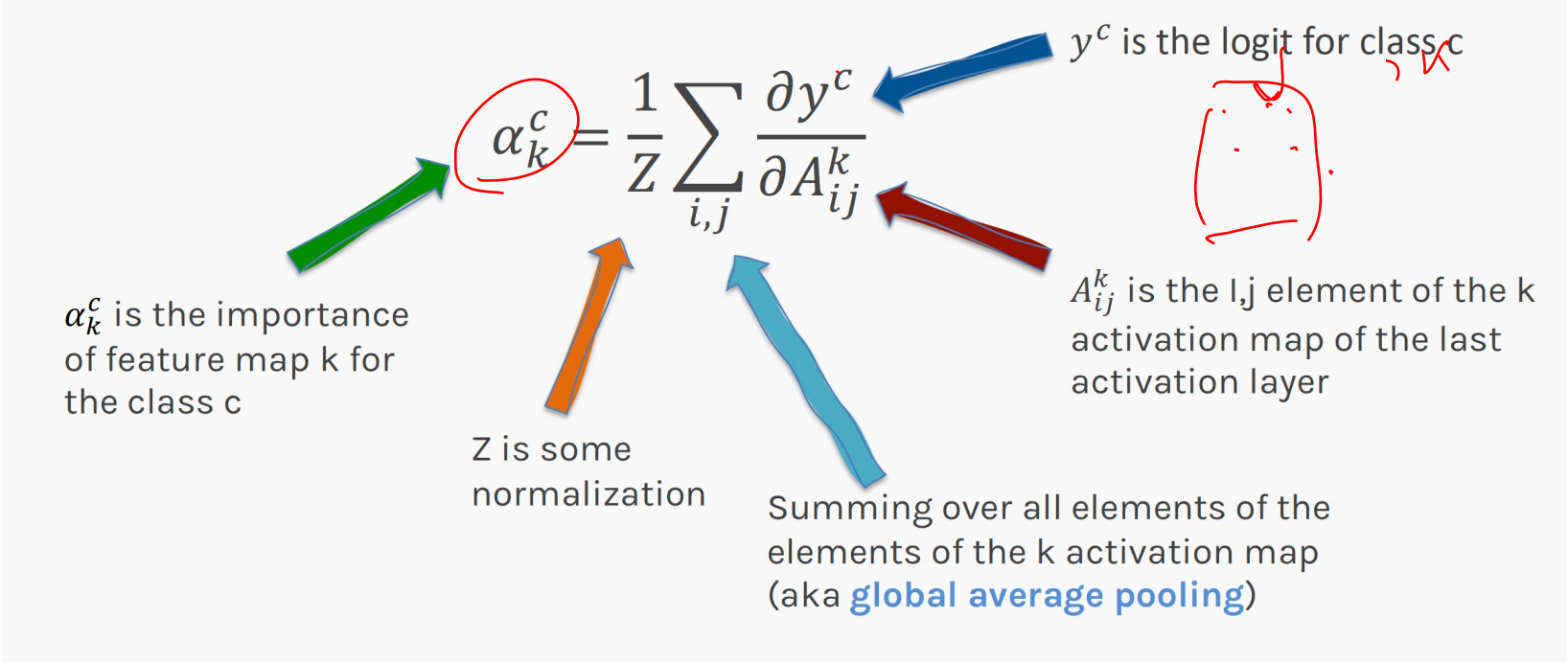
Why?

- Instead of the GAP, Grad CAM gets the gradient value for the weight.
- Grad-CAM does not need to rid off the FC layer.



Grad-CAM

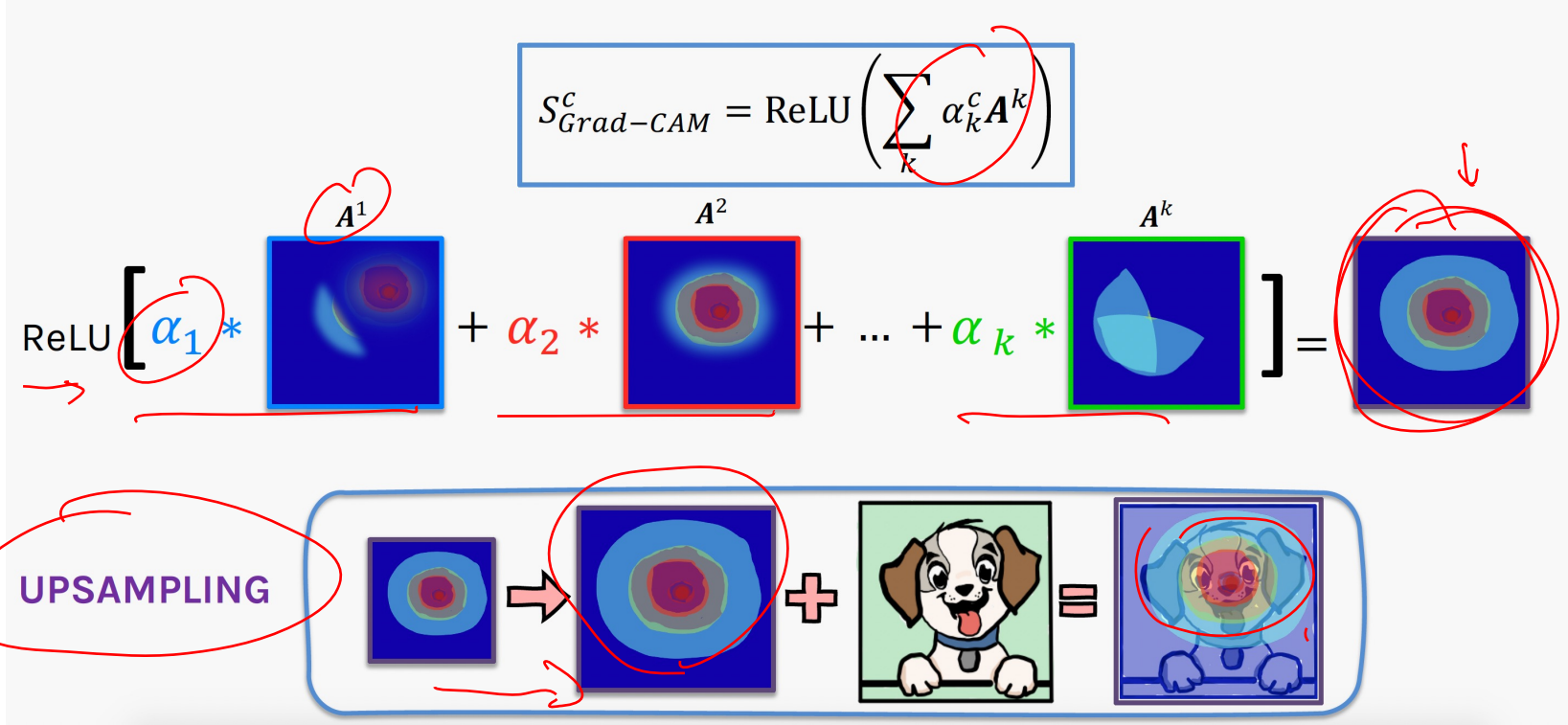
First, the gradient of the logits, y^c, of the class c w.r.t the activations maps of the final convolutional layer is computed and then the gradients are averaged across each feature map to give us an importance score.



Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

Grad-CAM

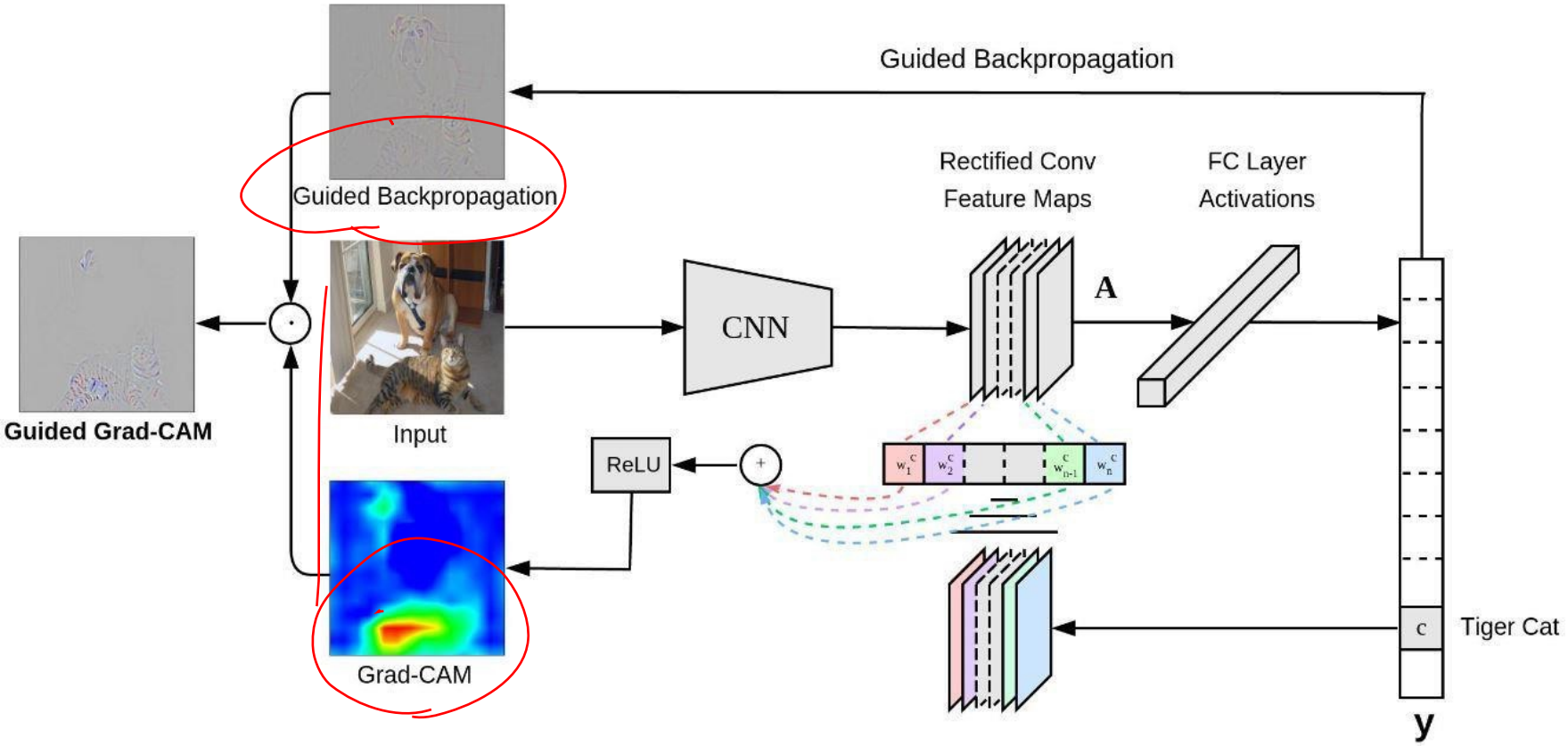
Combine the feature maps as we did in the CAM before except that here, we use the α's instead of the W and we also activate the



Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

Grad-CAM

As Grad-CAM can only produce coarse-grained visualizations, the authors have also combined guided-backpropagation with their method and propose Guided Grad-CAM.

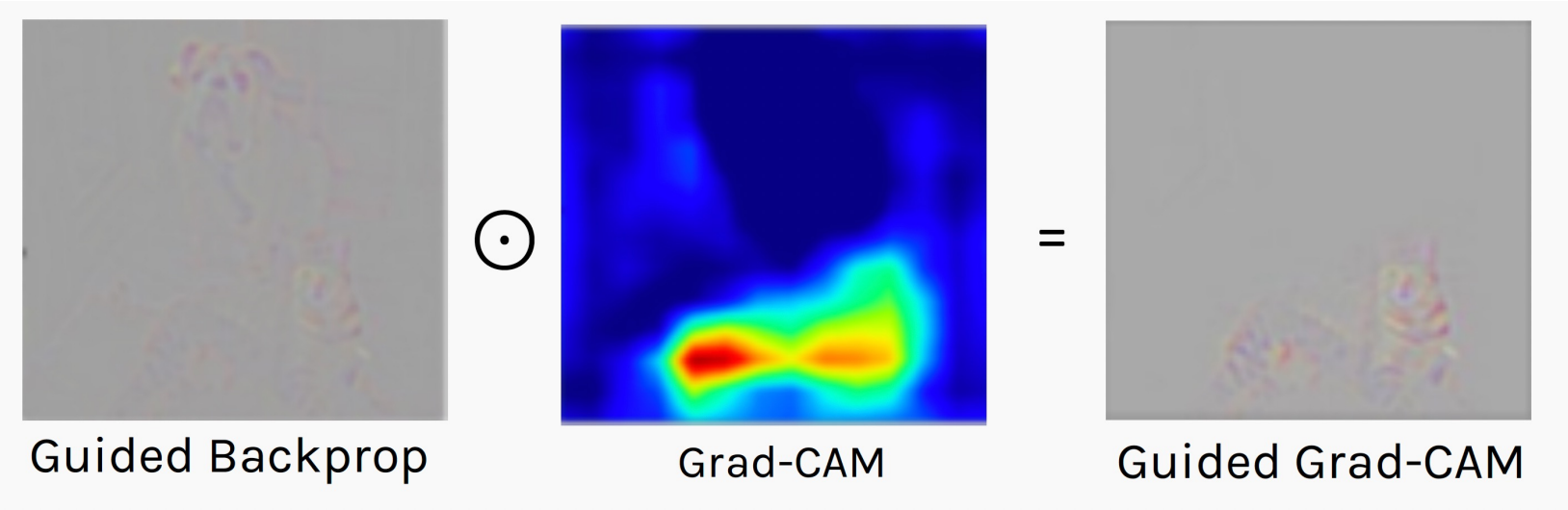


Guided Grad-CAM combines Guided-Backpropagation with Grad-CAM by simply perform an element-wise multiplication of Guided-Backpropagation with Grad-CAM

Grad-CAM

Grad-CAM can only produce coarse-grained visualizations.

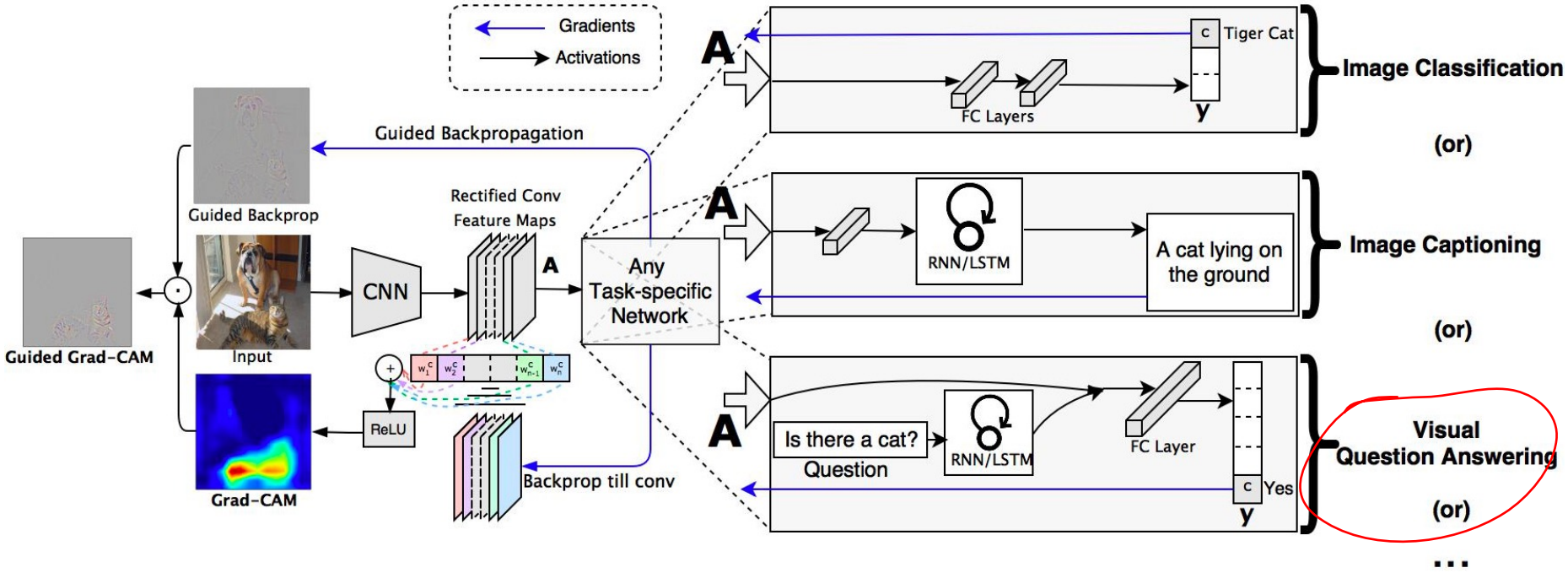
Guided Grad-CAM combines Guided-Backpropagation with Grad-CAM by simply perform an element-wise multiplication of Guided-Backpropagation with Grad-CAM.



Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

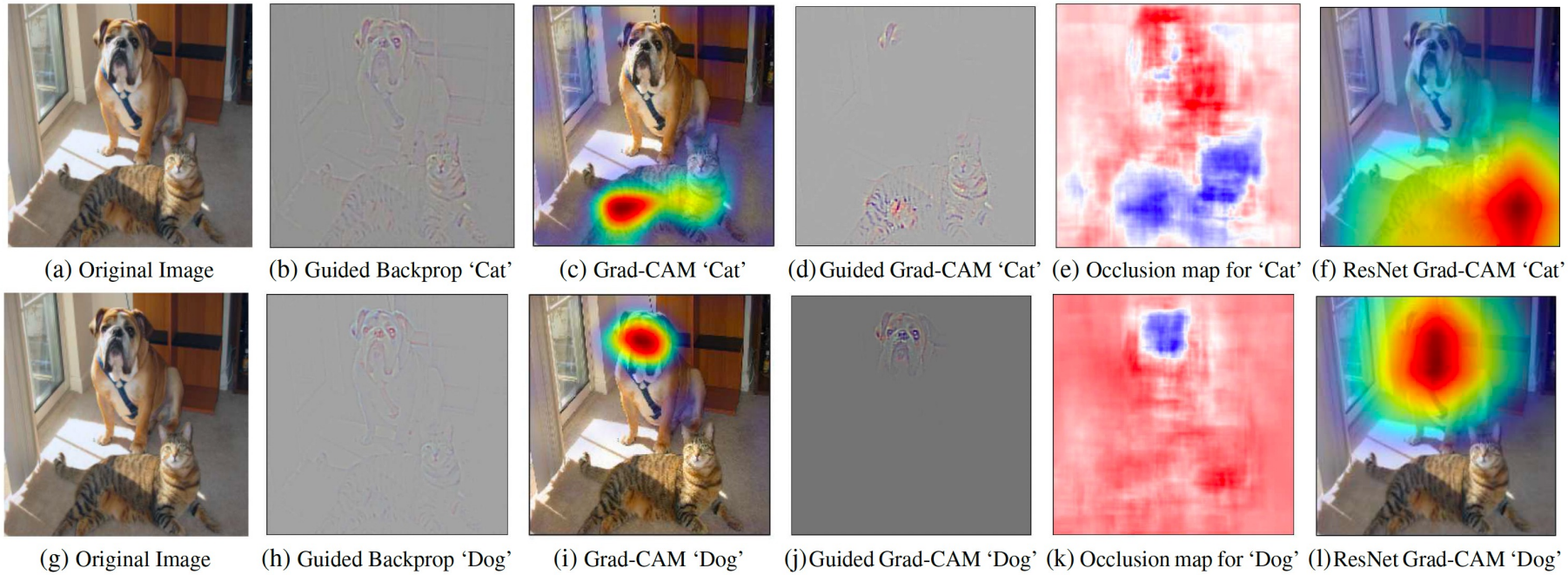
Grad-CAM

- Grad-CAM can be applied for every task where CAM could be only applied on the Classification.



Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization, ICCV 2017

Grad-CAM



(a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b) Guided Backpropagation [42]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (c, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class.

Analyzing Failure Modes with Grad-CAM



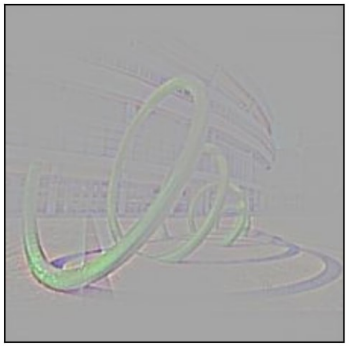



Ground truth: volcano


Ground truth: pineapple


Ground truth: polaroid camera

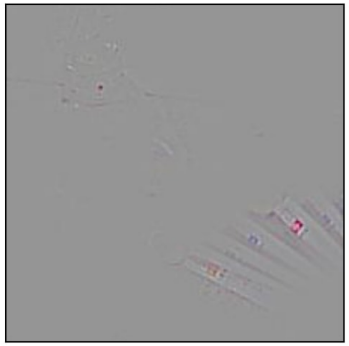

Ground truth: beaker

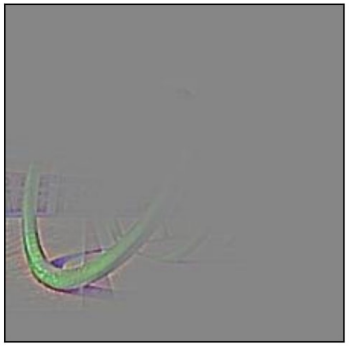

Ground truth: coil


Predicted: sandbar

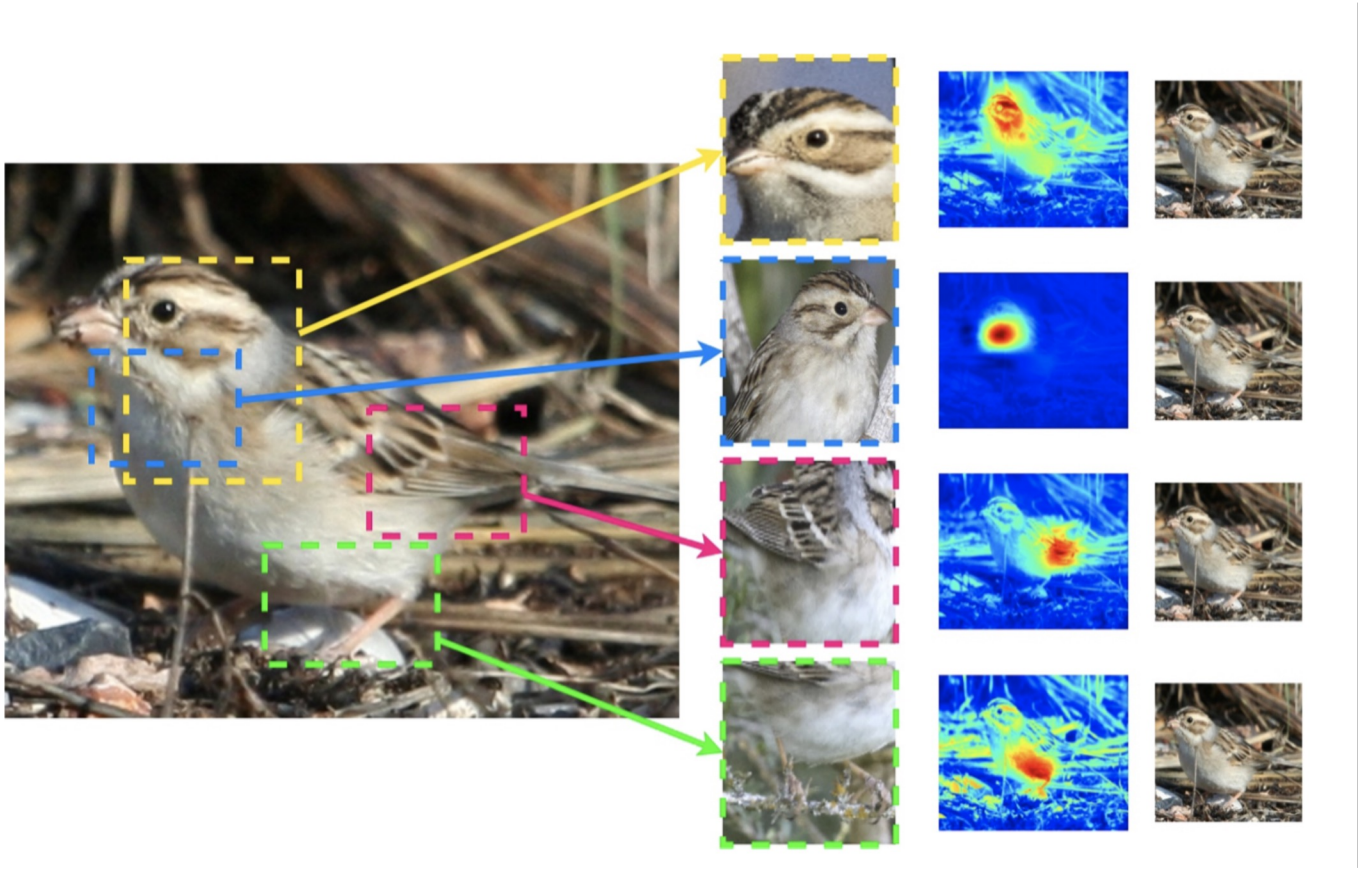

Predicted: patio


Predicted: pencil sharpener


Predicted: syringe

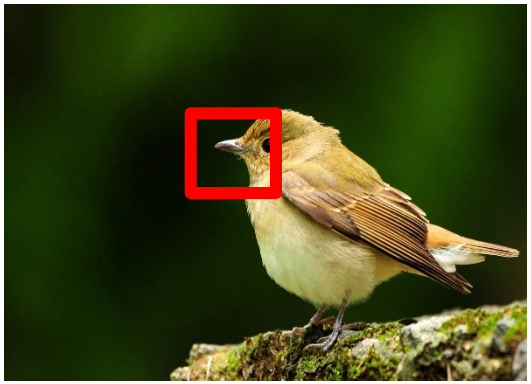

Predicted: vine snake

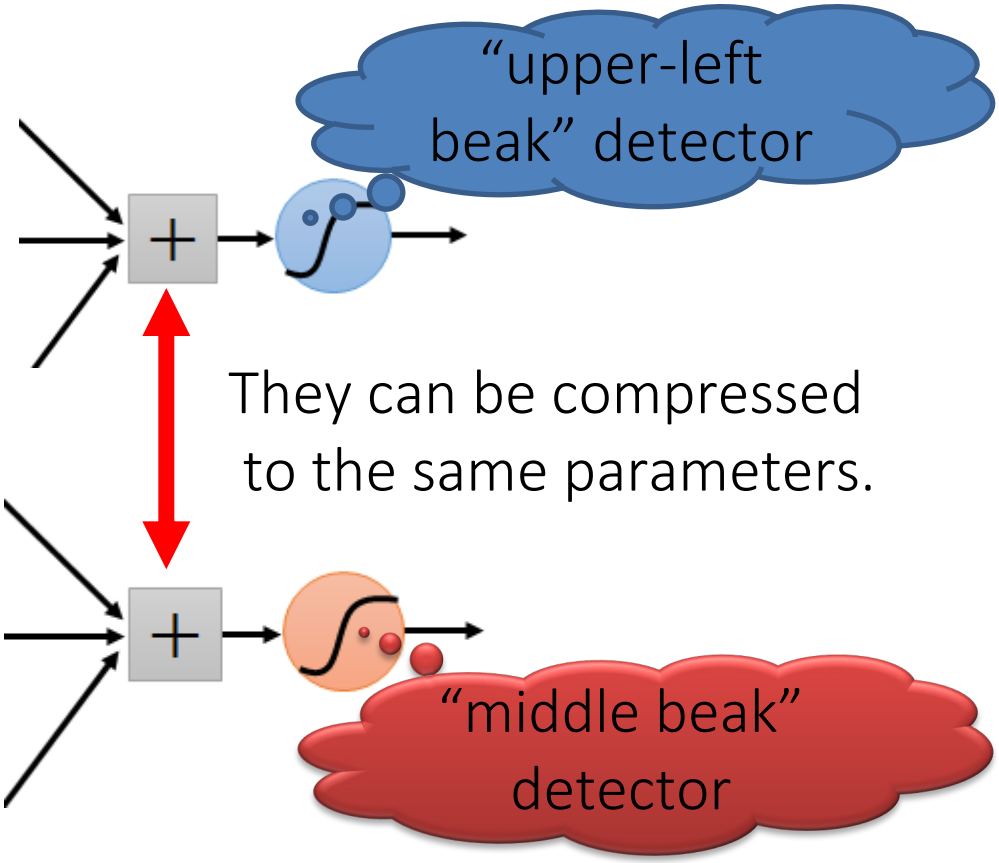
Case-based Reasoning



Chen, Chaofan, et al. “This looks like that: deep learning for interpretable image recognition.” Advances in neural information processing systems 32 (2019).

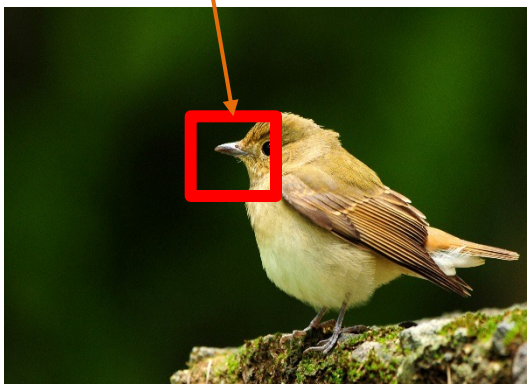
Intuition

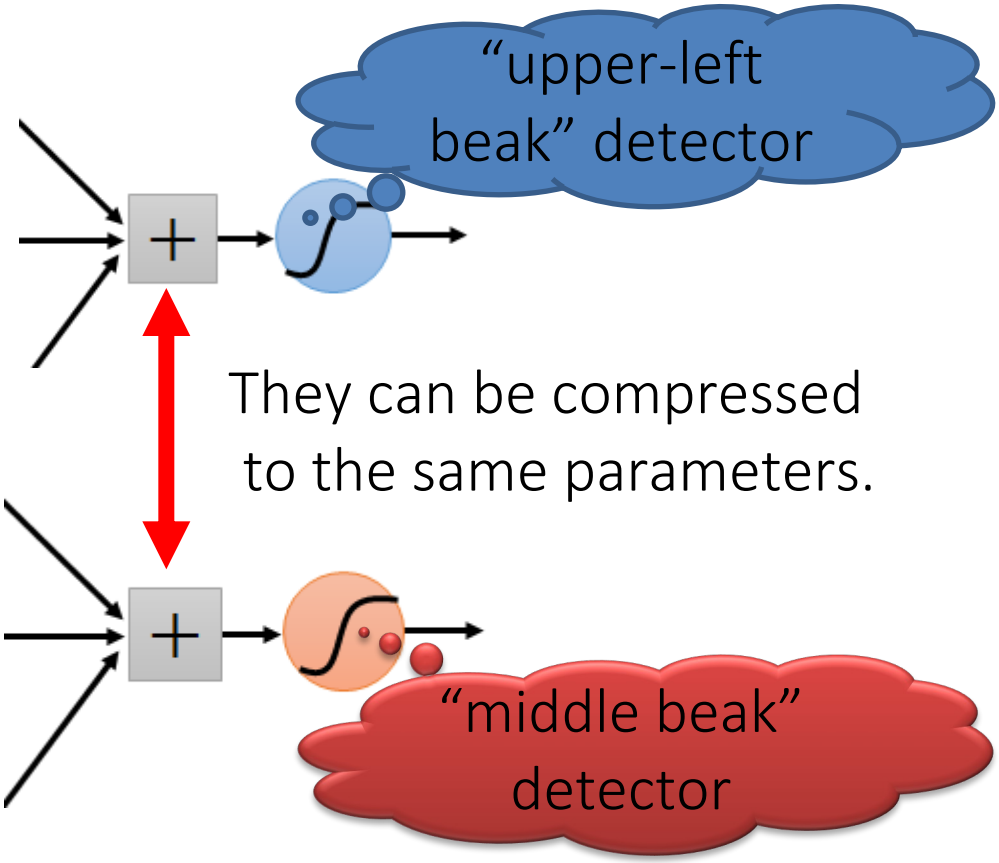



“upper-left beak” detector
They can be compressed to the same parameters.
“middle beak” detector

Intuition

This looks like that





Intuition

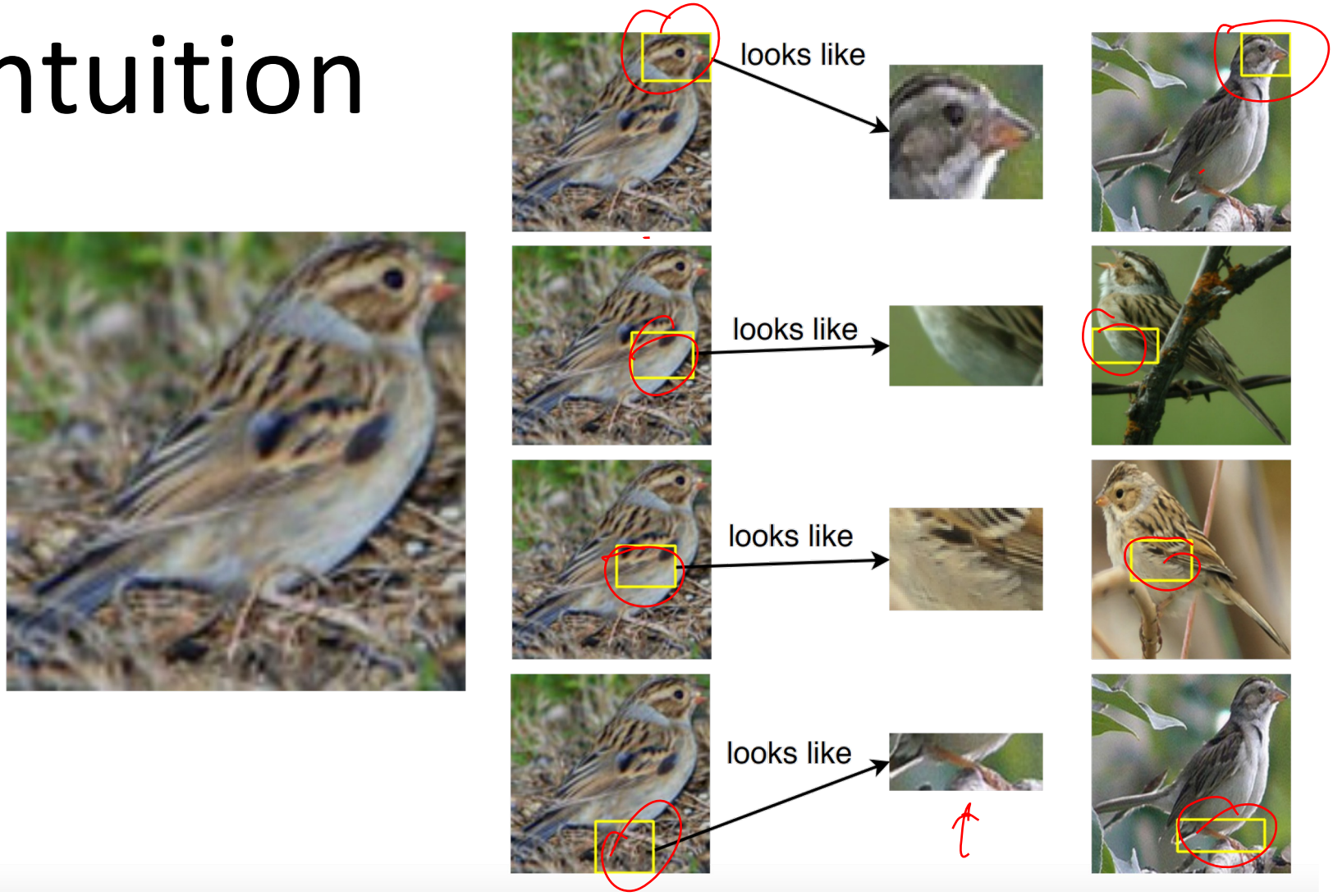
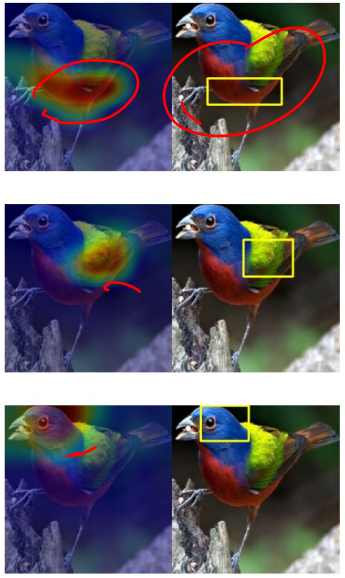


Image of a clay colored sparrow and how parts of it look like some learned prototypical parts of a clay colored sparrow used to classify the bird's species.

with richer explanations



(a) Object attention
(class activation map)

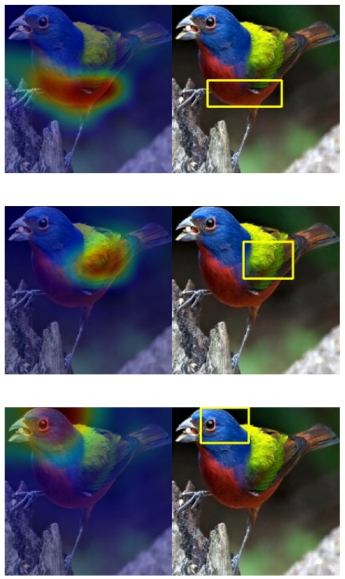
(b) Part attention
(attention-based models)

Previous methods

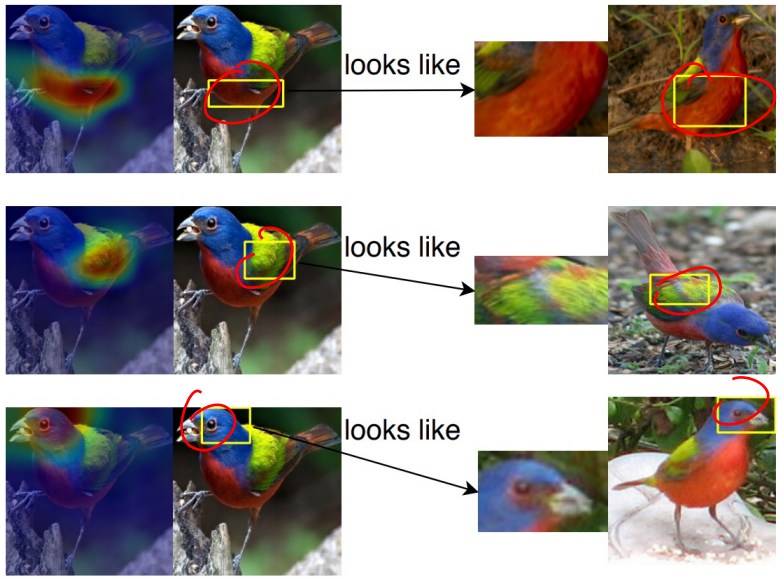
Chen, Chaofan, et al. “This looks like that: deep learning for interpretable image recognition.” Advances in neural information processing systems 32 (2019).

with richer explanations

(a) Object attention
(class activation map)



(b) Part attention
(attention-based models)

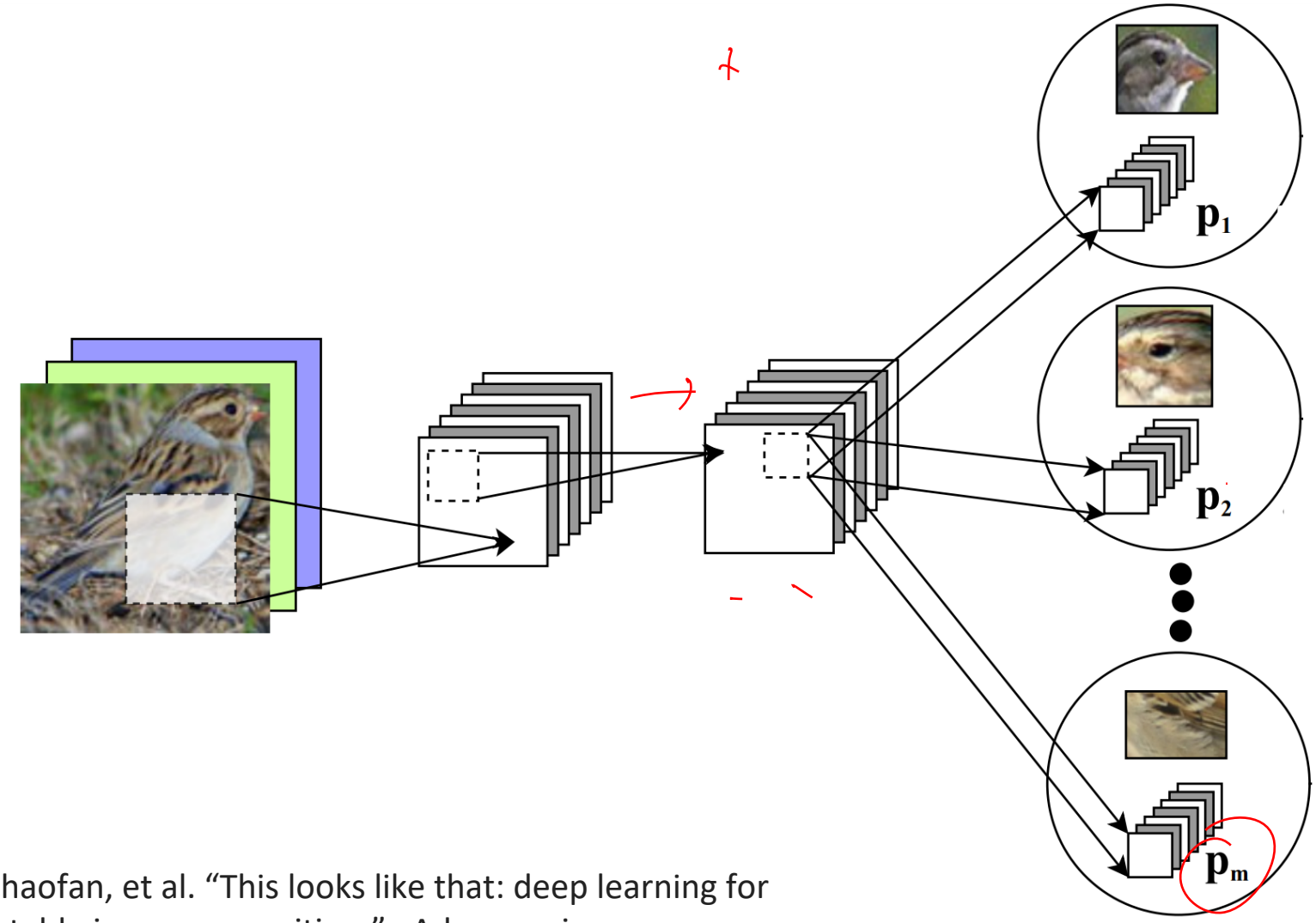


(c) Part attention + comparison with learned
prototypical parts (our model)

Previous methods

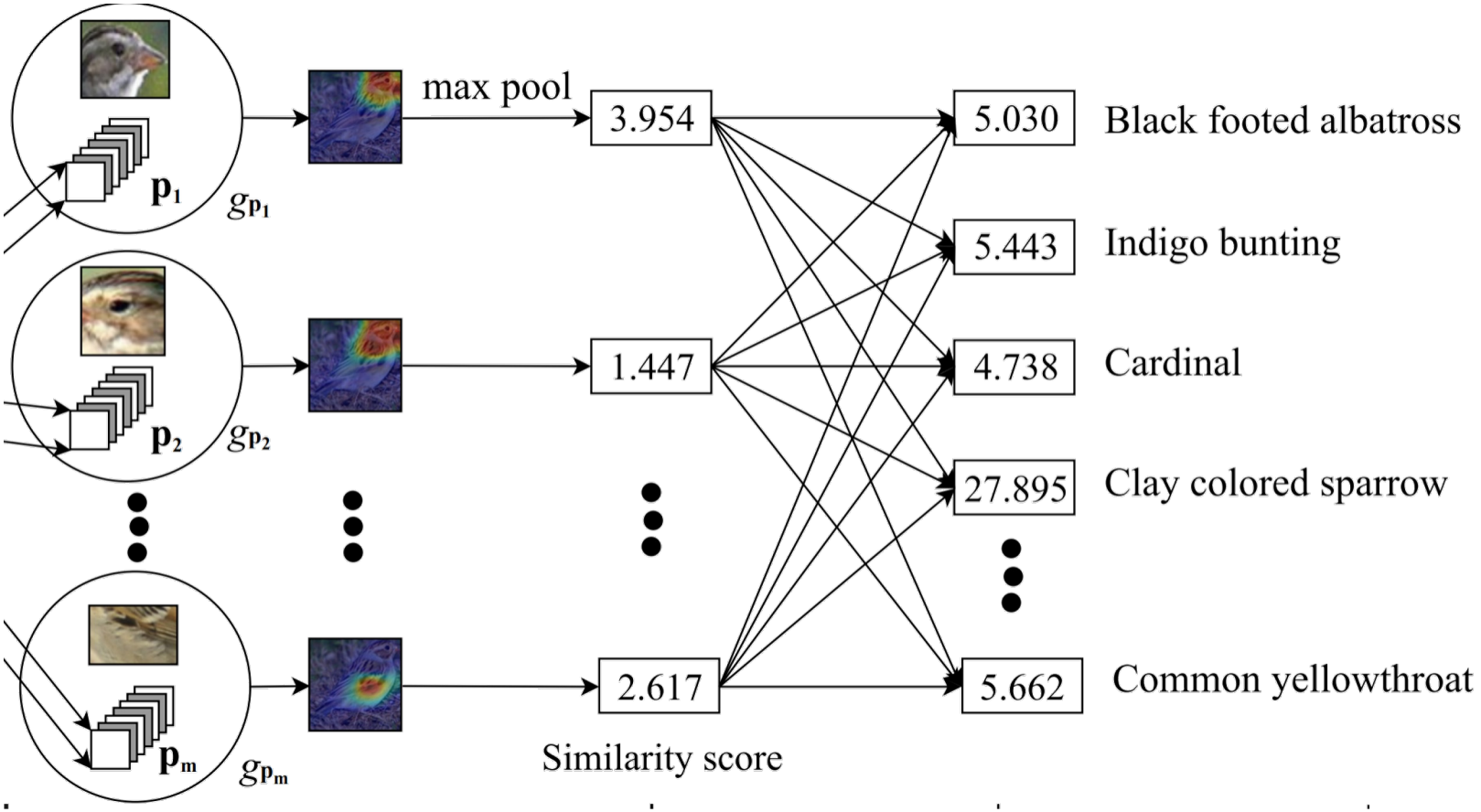
Chen, Chaofan, et al. “This looks like that: deep learning for interpretable image recognition.” Advances in neural information processing systems 32 (2019).

ProtoPNet Architecture

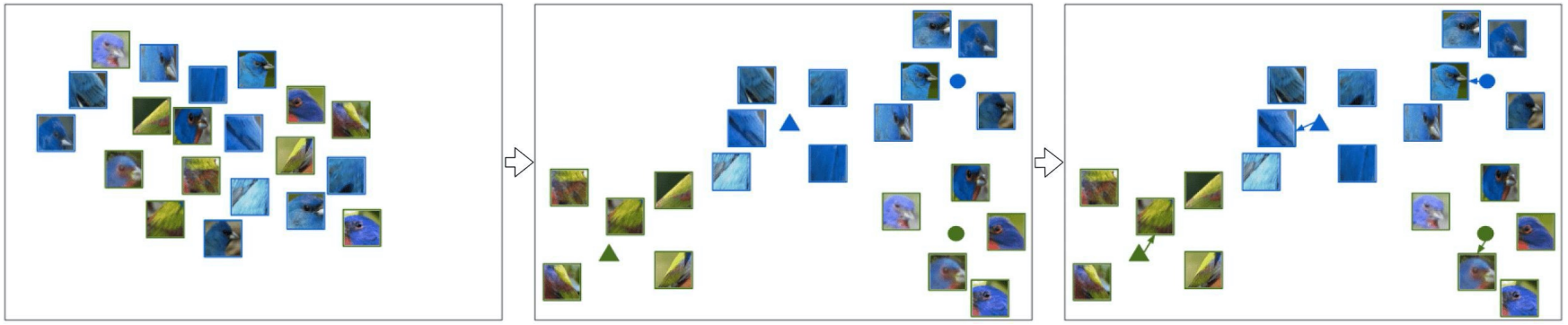


Chen, Chaofan, et al. “This looks like that: deep learning for interpretable image recognition.” Advances in neuralinformation processing systems 32 (2019).

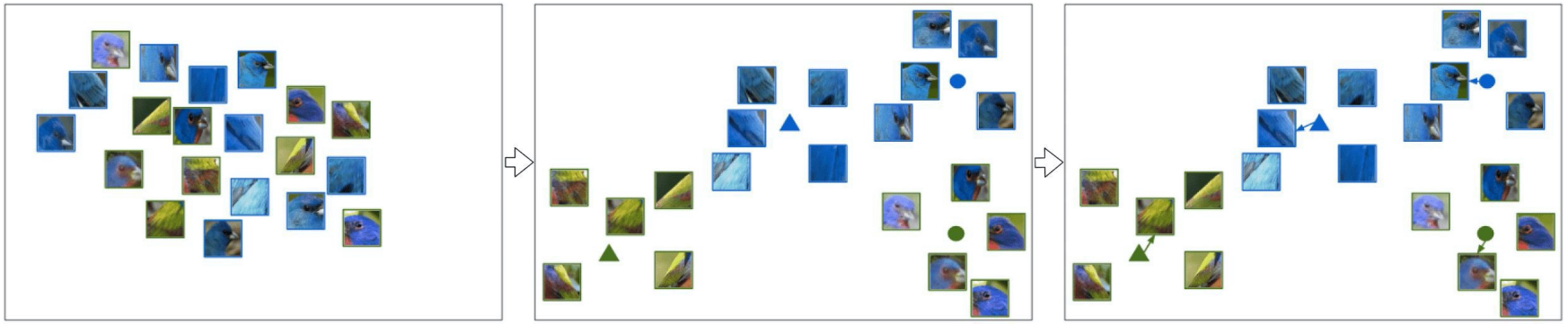
ProtoPNet Architecture



Chen, Chaofan, et al. “This looks like that: deep learning for interpretable image recognition.” Advances in neural information processing systems 32 (2019).



$$\min_{\mathbf{P}, w_{\text{conv}}} \frac{1}{n} \sum_{i=1}^{n} \text{CrsEnt}(h \circ g_{\mathbf{p}} \circ f(\mathbf{x_i}), \mathbf{y_i}) + \lambda_1 \text{Clst} + \lambda_2 \text{Sep}$$



$$\min_{\mathbf{P}, w_{\text{conv}}} \frac{1}{n} \sum_{i=1}^{n} \text{CrsEnt}(h \circ g_{\mathbf{p}} \circ f(\mathbf{x_i}), \mathbf{y_i}) + \lambda_1 \text{Clst} + \lambda_2 \text{Sep}$$

penalizes large distance between the closest patch and class k prototype, so each training image has at least one patch in its class prototypes

$$\text{Clst} = \frac{1}{n} \sum_{i=1}^{n} \min_{j:\mathbf{p}_j \in \mathbf{P}_{y_i}} \min_{\mathbf{z} \in \text{patches}(f(\mathbf{x}_i))} \|\mathbf{z} - \mathbf{p}_j\|_2^2$$

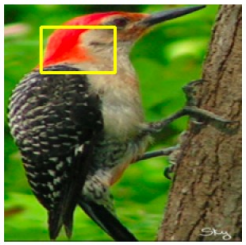
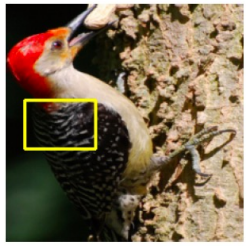
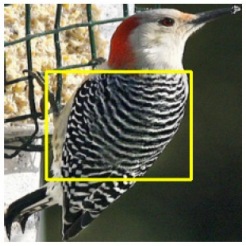
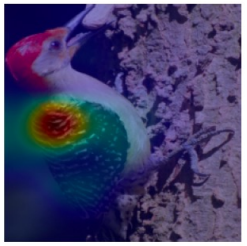
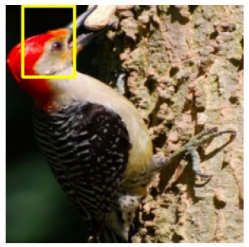
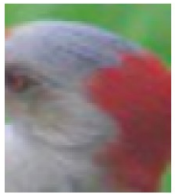
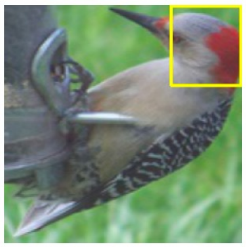
penalizes small distance between the closest patch and non-class k prototypes, so training images push away prototypes of other classes

$$\text{Sep} = -\frac{1}{n} \sum_{i=1}^{n} \min_{j:\mathbf{p}_j \notin \mathbf{P}_{y_i}} \min_{\mathbf{z} \in \text{patches}(f(\mathbf{x}_i))} \|\mathbf{z} - \mathbf{p}_j\|_2^2$$

This Looks Like That: Is this Interpretable?

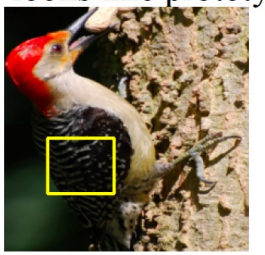
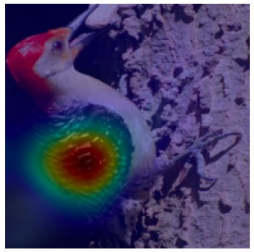
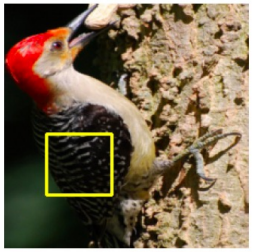
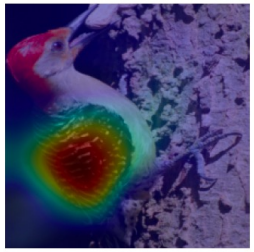
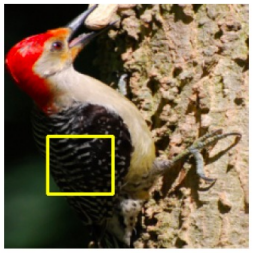
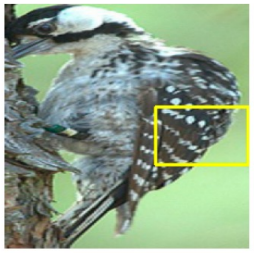
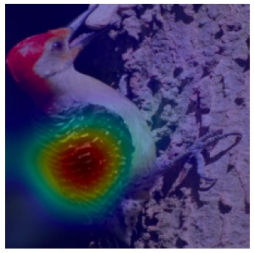
Why is this bird classfied as a red-bellied woodpecker?

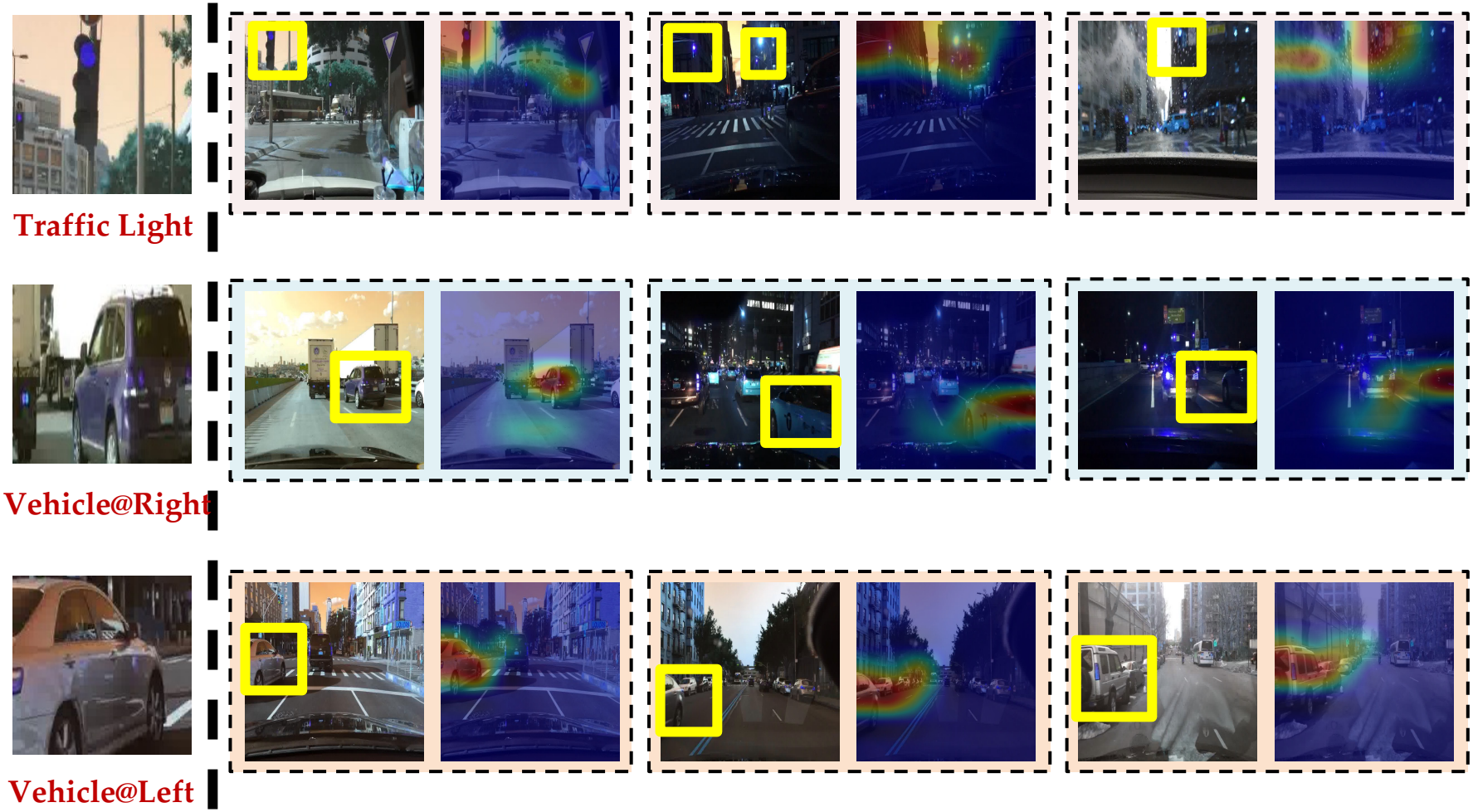
Evidence for this bird being a red-bellied woodpecker:

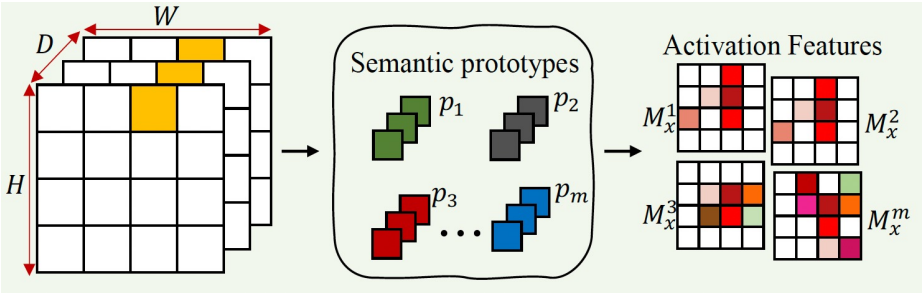
Original image (box showing part that looks like prototype)	Prototype	Training image where prototype comes from	Activation map	Similarity score	Class connection	Points contributed
				6.499	× 1.180	= 7.669
				4.392	× 1.127	= 4.950
				3.890	× 1.108	= 4.310

This Looks Like That: Is this Interpretable?

Evidence for this bird being a red-cockaded woodpecker:

Original image (box showing part that looks like prototype)	Prototype	Training image where prototype comes from	Activation map	Similarity score	Class connection	Points contributed
				2.452	×	1.046 = 2.565
				2.125	×	1.091 = 2.318
				1.945	×	1.069 = 2.079



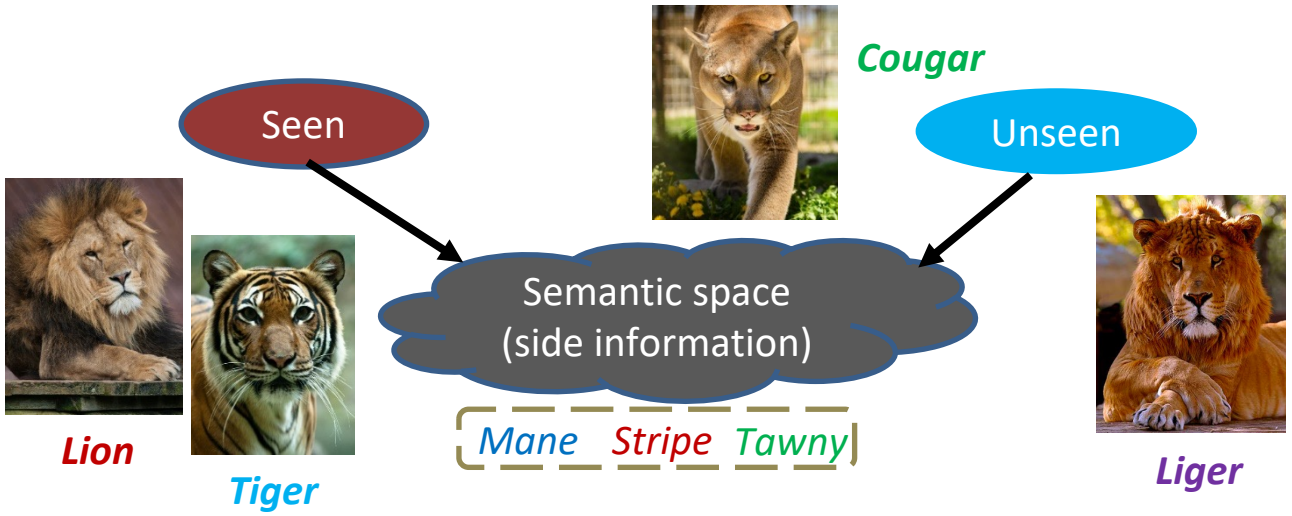


Taotao Jing, Haifeng Xia, Renran Tian, Haoran Ding, Xiao Luo, Joshua Domeyer, Rini Sherony, and Zhengming Ding. InAction: Interpretable Action Decision Making for Autonomous Driving. European Conference on Computer Vision (ECCV), 2022.

Can interpretability help us make networks generalize better?

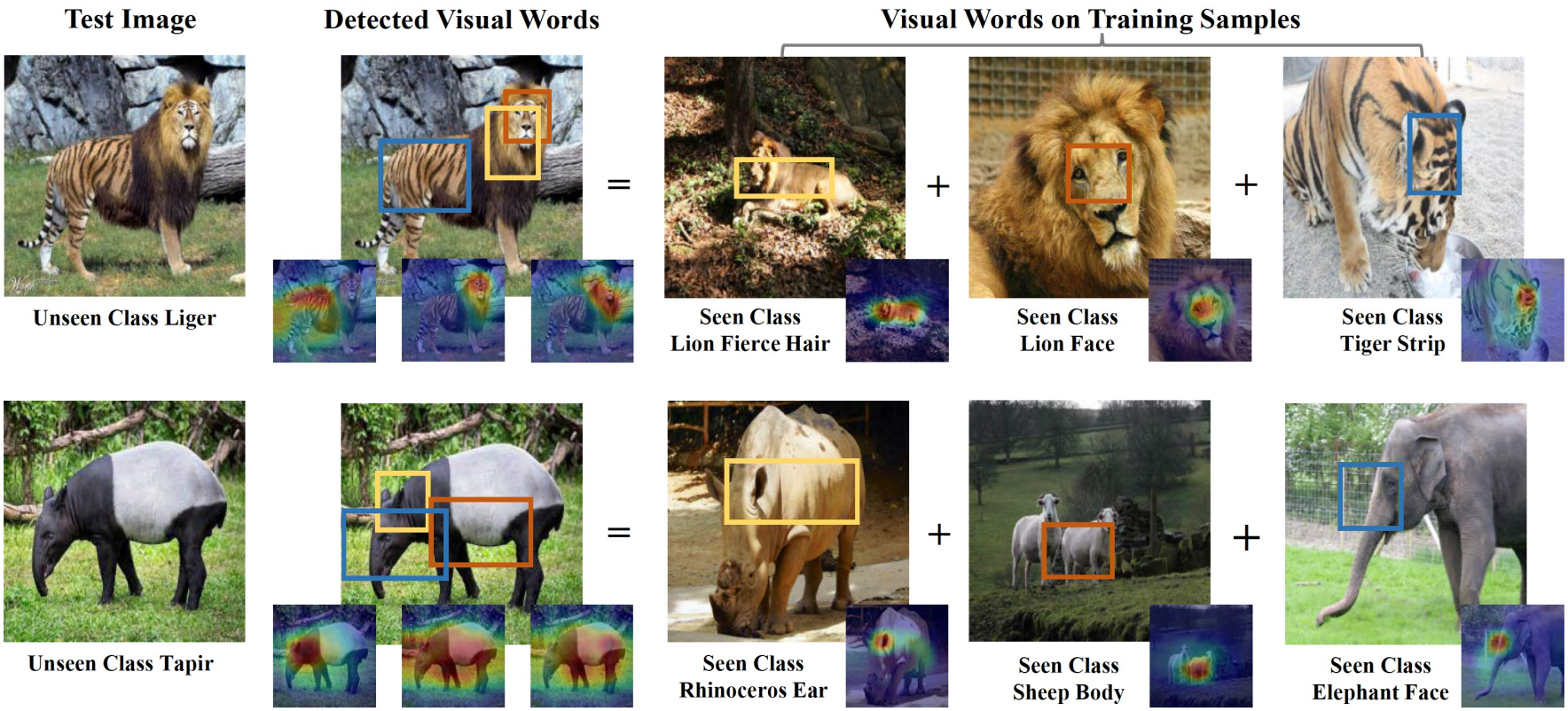
Use what you've already learnt → Generalize to unseen concepts

Zero-Shot Learning



Ruis, Frank, Gertjan Burghouts, and Doina Bucur. "Independent prototype propagation for zero-shot compositionality." *Advances in Neural Information Processing Systems* 34 (2021): 10641-10653.

Interpreting Unseen Species

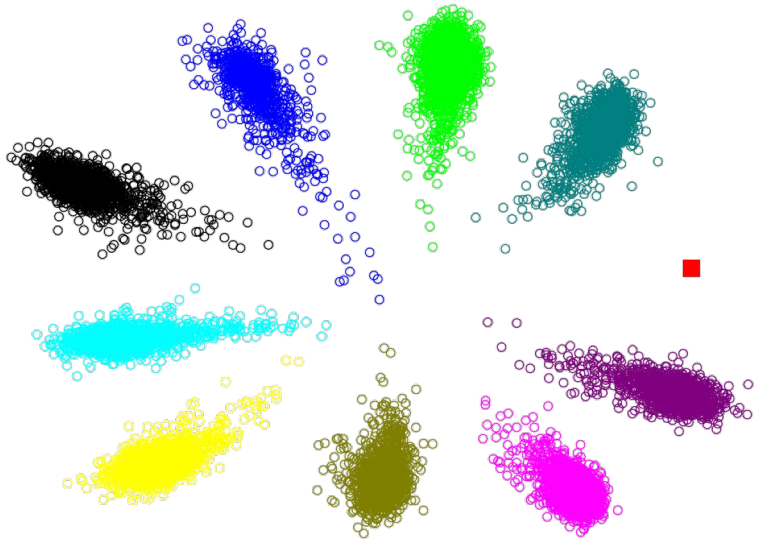


Xiao, Wenxiao, Zhengming Ding, and Hongfu Liu. "Learnable Visual Words for Interpretable Image Recognition." *arXiv preprint arXiv:2205.10724* (2022).

How about One-Shot Class?

Logistic regression loss is additive

$$L = \Sigma_{i=1}^{N} \mathbf{cross\_entropy}(\sigma(x_i), y_i)$$

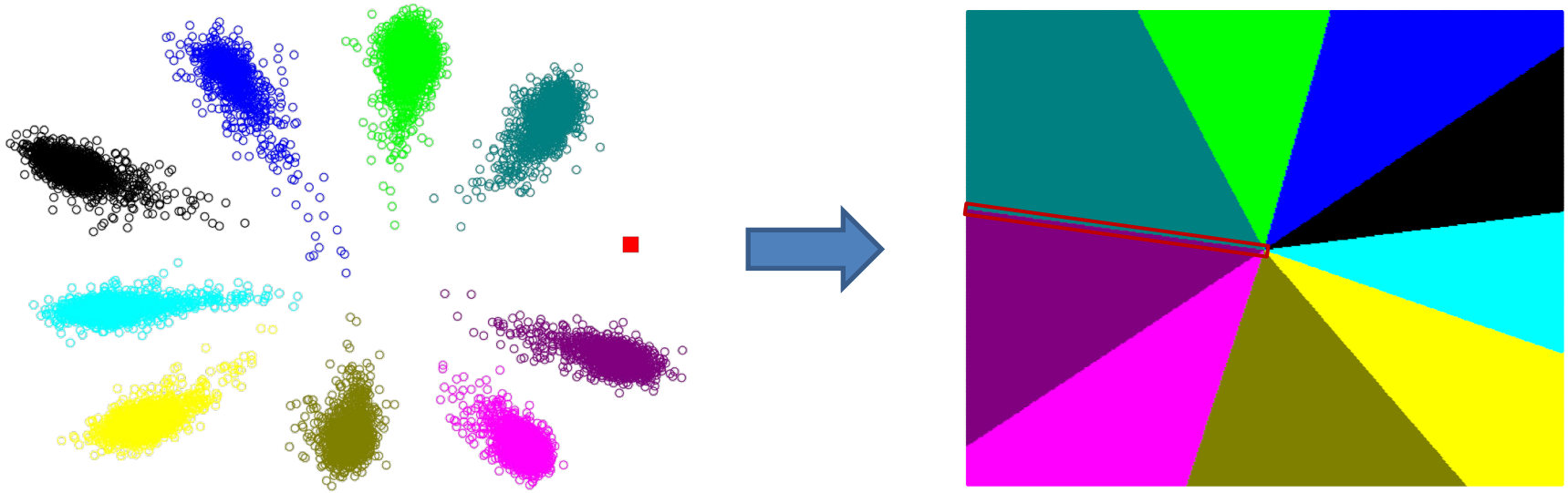


Lack of samples introduces smaller classification space
Accordingly, smaller classification space means smaller weighting vector norm for one-shot classes

How about One-Shot Class?

Logistic regression loss is additive

$$L = \Sigma_{i=1}^{N} \mathbf{cross\_entropy}(\sigma(x_i), y_i)$$



Lack of samples introduces smaller classification space
Accordingly, smaller classification space means smaller weighting vector norm for one-shot classes