Recap: Positive values in feature map indicate patterns in the input image

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

6 x 6 image

-1	1	-1
-1	1	-1
-1	1	-1

Filter size: 3x3

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

1	0	0	0	0	1
0	1	0	0	1	0
0	0	1	1	0	0
1	0	0	0	1	0
0	1	0	0	1	0
0	0	1	0	1	0

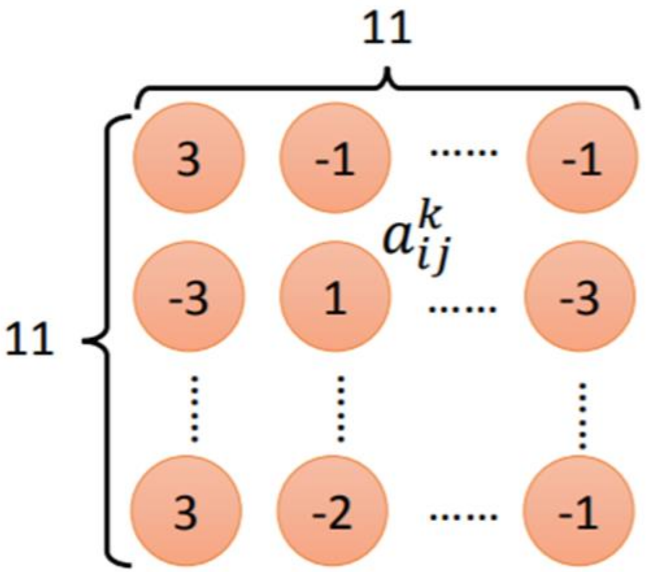
-1	-1	-1	-1
-1	-1	-2	1
-1	-1	-2	1
-1	0	-4	3

Degree of activation of the k^{th} filter

The output of the k-th filter is a 11 x 11 matrix.

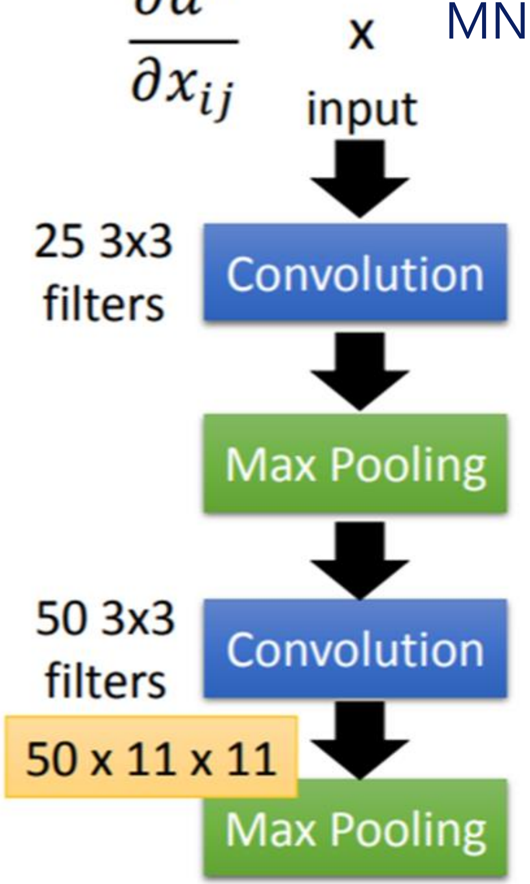
Degree of the activation of the k-th filter: $a^k = \sum_{i=1}^{11}\sum_{j=1}^{11} a_{ij}^k$

$x^* = arg\max_x a^k$ (gradient ascent)



$\frac{\partial a^k}{\partial x_{ij}}$

MNIST images



Reference: 李弘毅 ML Lecture 10 https://youtu.be/FrKWiRv254g

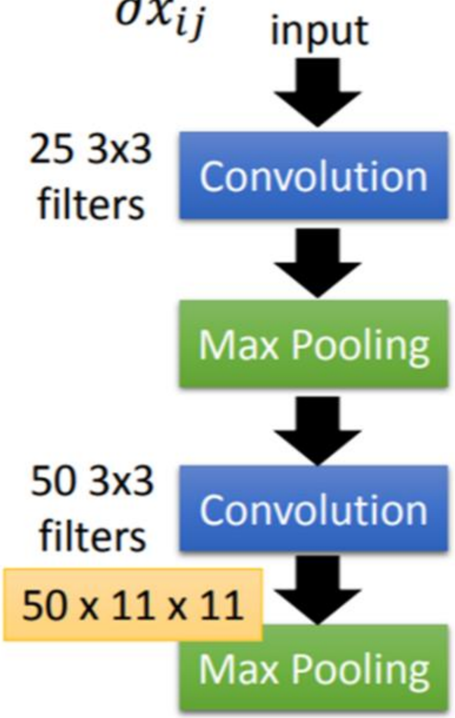
What input images result in higher activation degree?

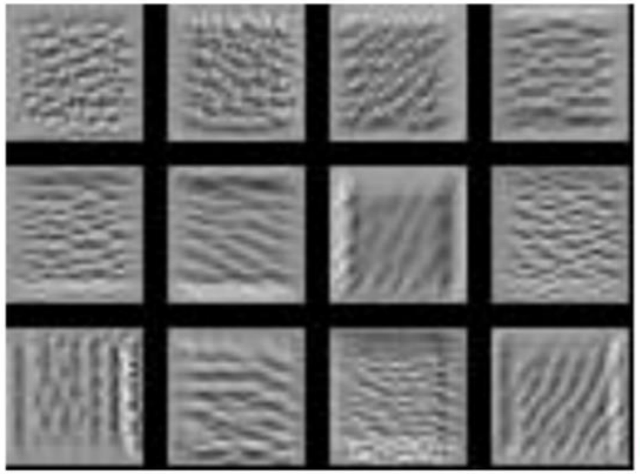
The output of the k-th filter is a 11 x 11 matrix.

Degree of the activation of the k-th filter: $a^k = \sum_{i=1}^{11} \sum_{j=1}^{11} a_{ij}^k$

$x^* = arg \max_x a^k$ (gradient ascent)

$\frac{\partial a^k}{\partial x_{ij}}$



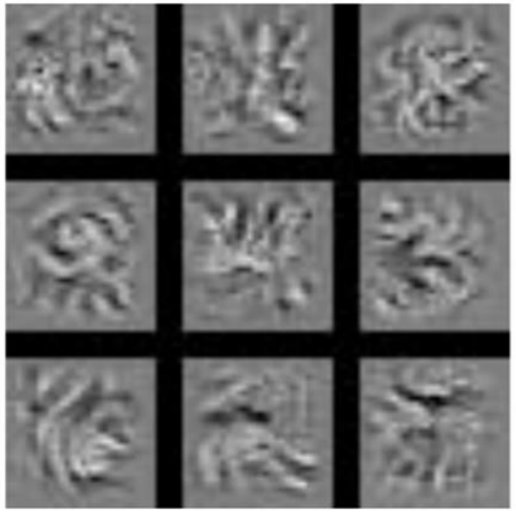


For each filter

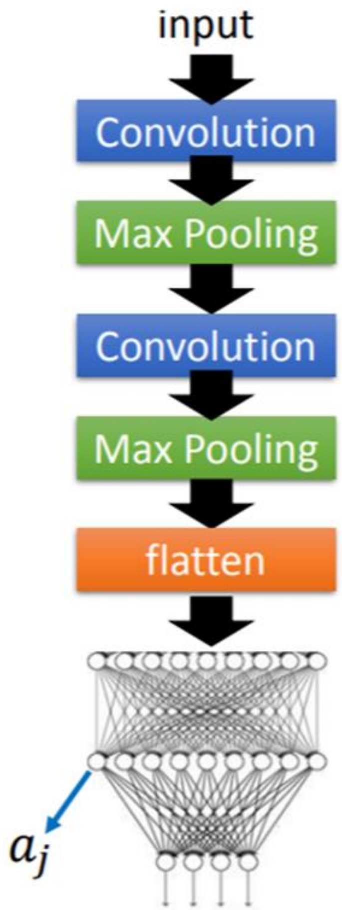
Input images that make the first 14 filters activate most

Find an image maximizing the output of neuron:

$x^* = arg \max_x a^j$



Each figure corresponds to a neuron



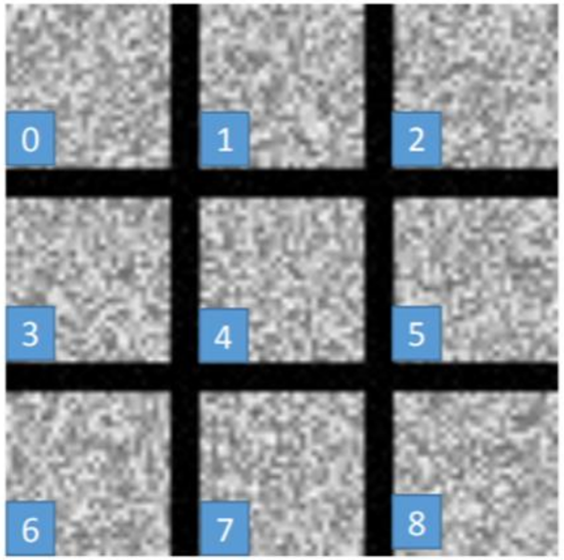
Input images that make the first 9 nodes in the fully connected layer activate the most

Reference: 李弘毅 ML Lecture 10 https://youtu.be/FrKWiRv254g

What input images result in higher activation degree?

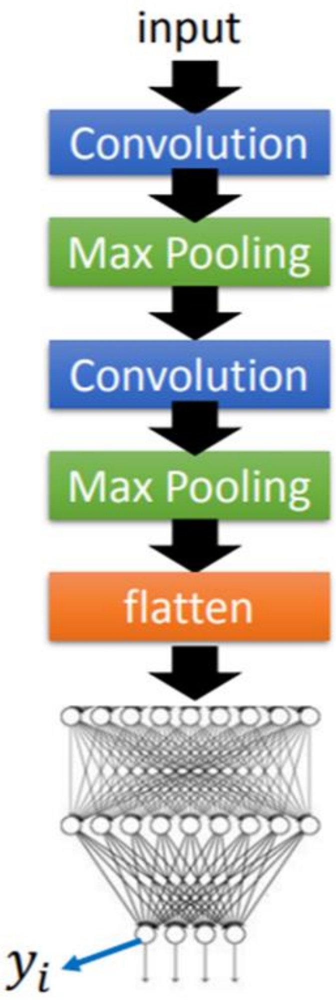
$$x^* = arg\max_x y^i$$

Can we see digits?



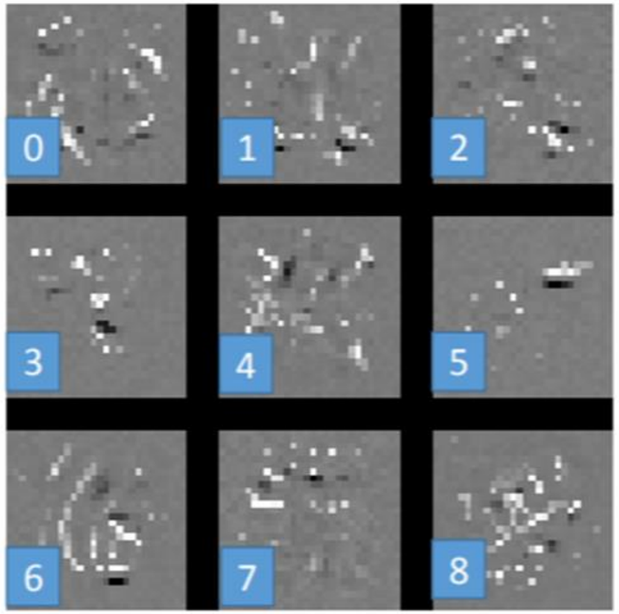
Deep Neural Networks are Easily Fooled
https://www.youtube.com/watch?v=M2IebCN9Ht4

Input images that make the 9 output classes activate the most



Over all pixel values

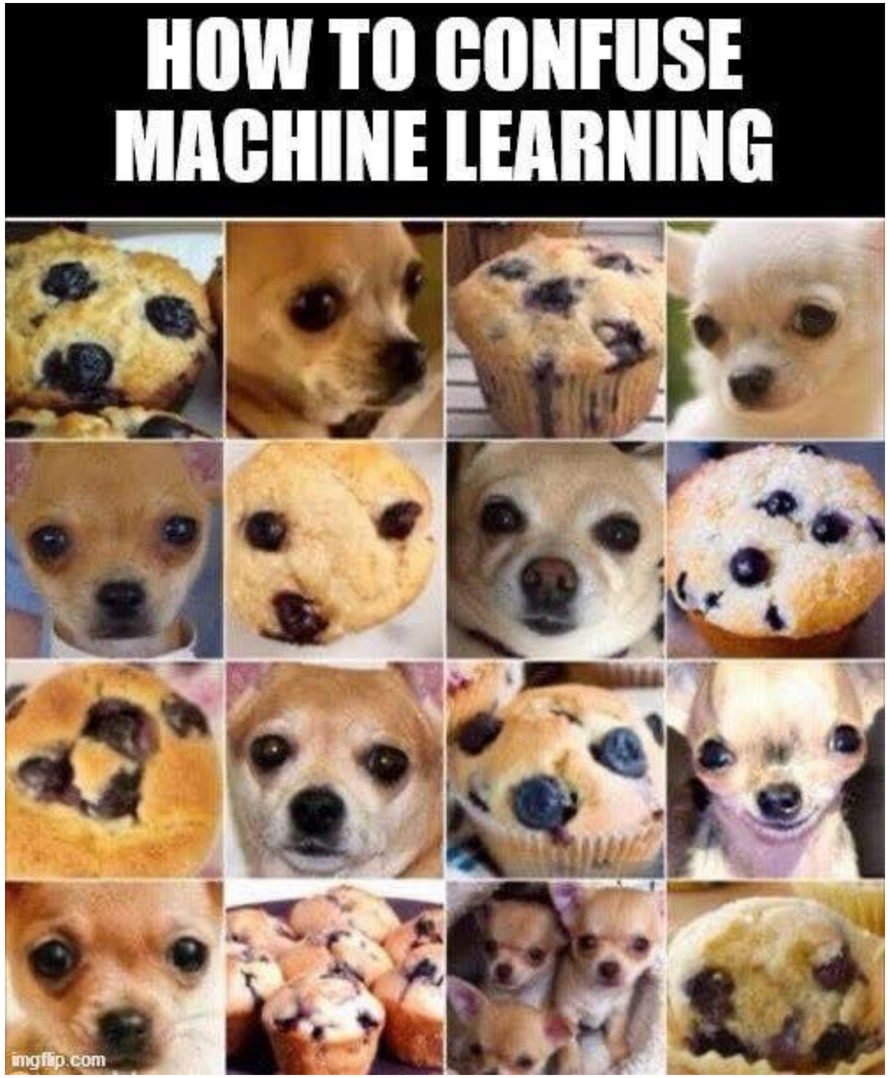
$$x^* = arg\max_x \left(y^i - \boxed{\sum_{i,j} |x_{ij}|} \right)$$



Force x_{ij}=0, i.e., force most pixels to NO INK (as only small part of the image has ink)

Reference: 李弘毅 ML Lecture 10 https://youtu.be/FrKWiRv254g

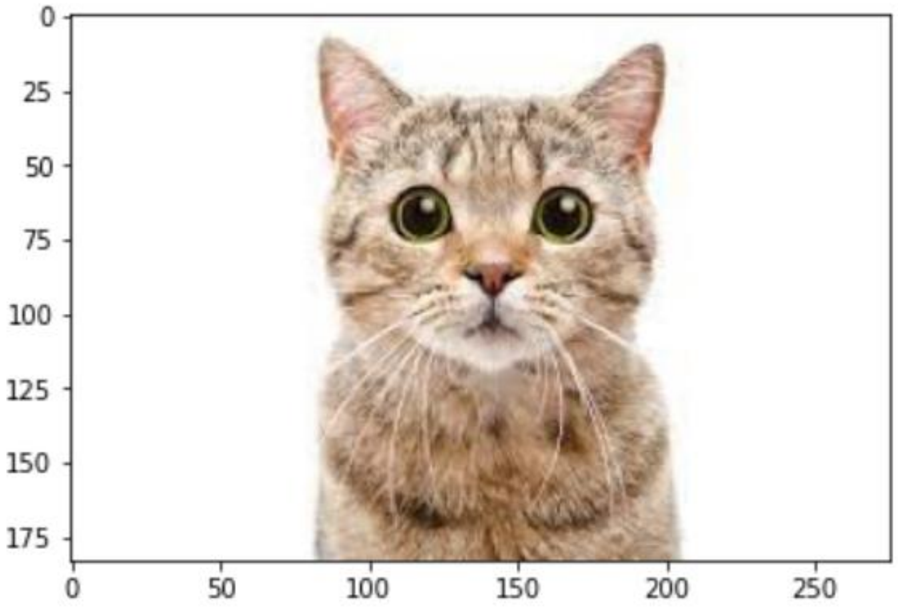
Neural networks are easily fooled

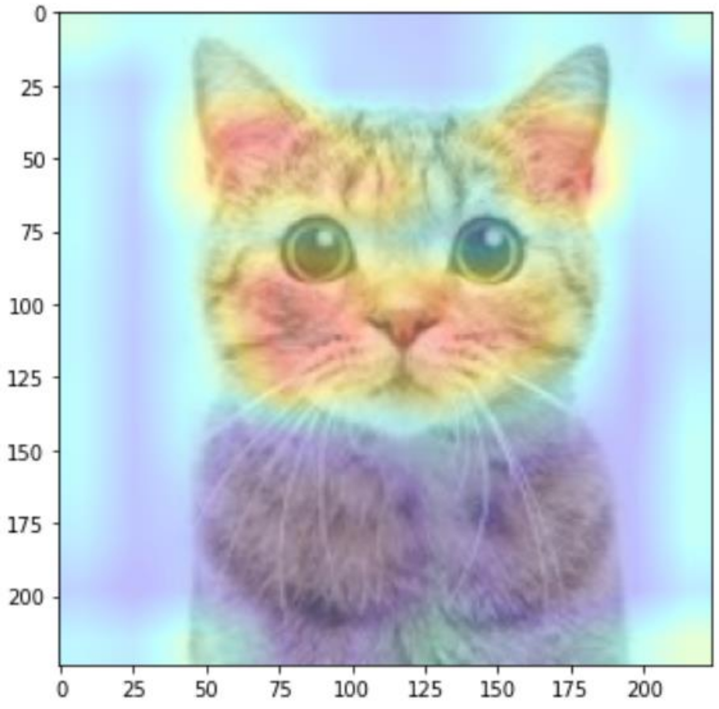


https://www.facebook.com/105366834575253/photos/a.106433371135266/217334733378462/

Gradient-weighted class activation map (Grad-CAM)

6.5 GradCAM.ipynb





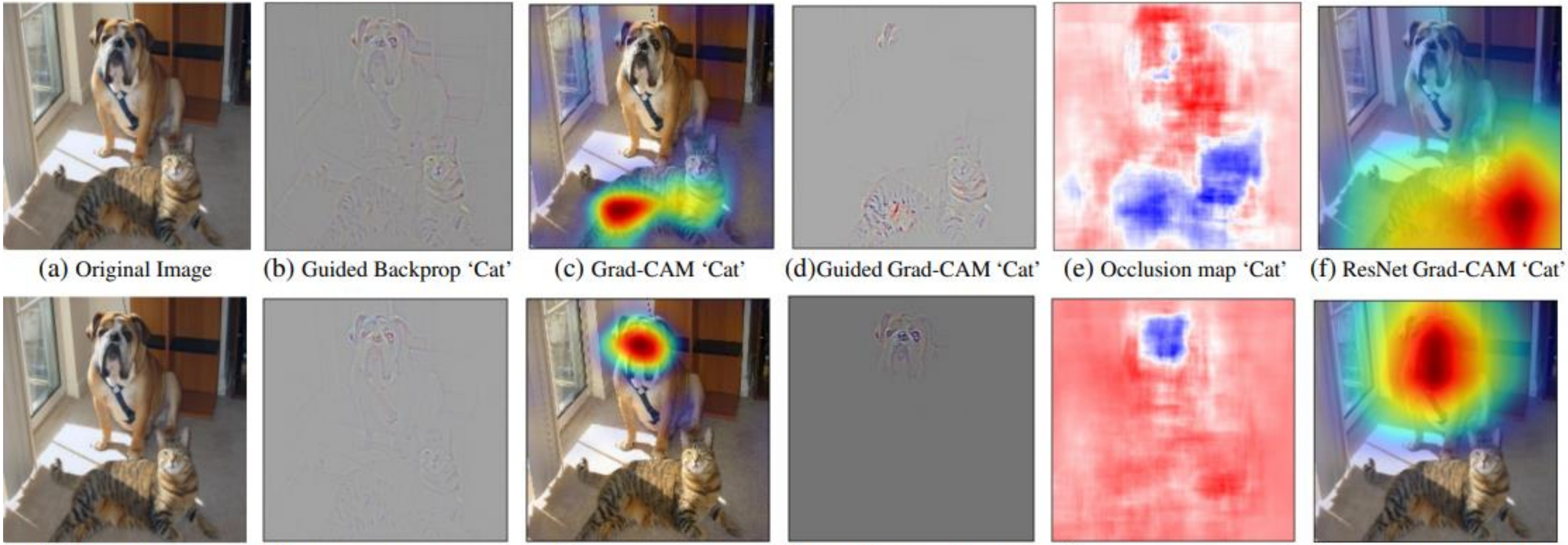
Gradient-weighted class activation map (Grad-CAM)

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra

https://arxiv.org/pdf/1610.02391.pdf

Grad-CAM

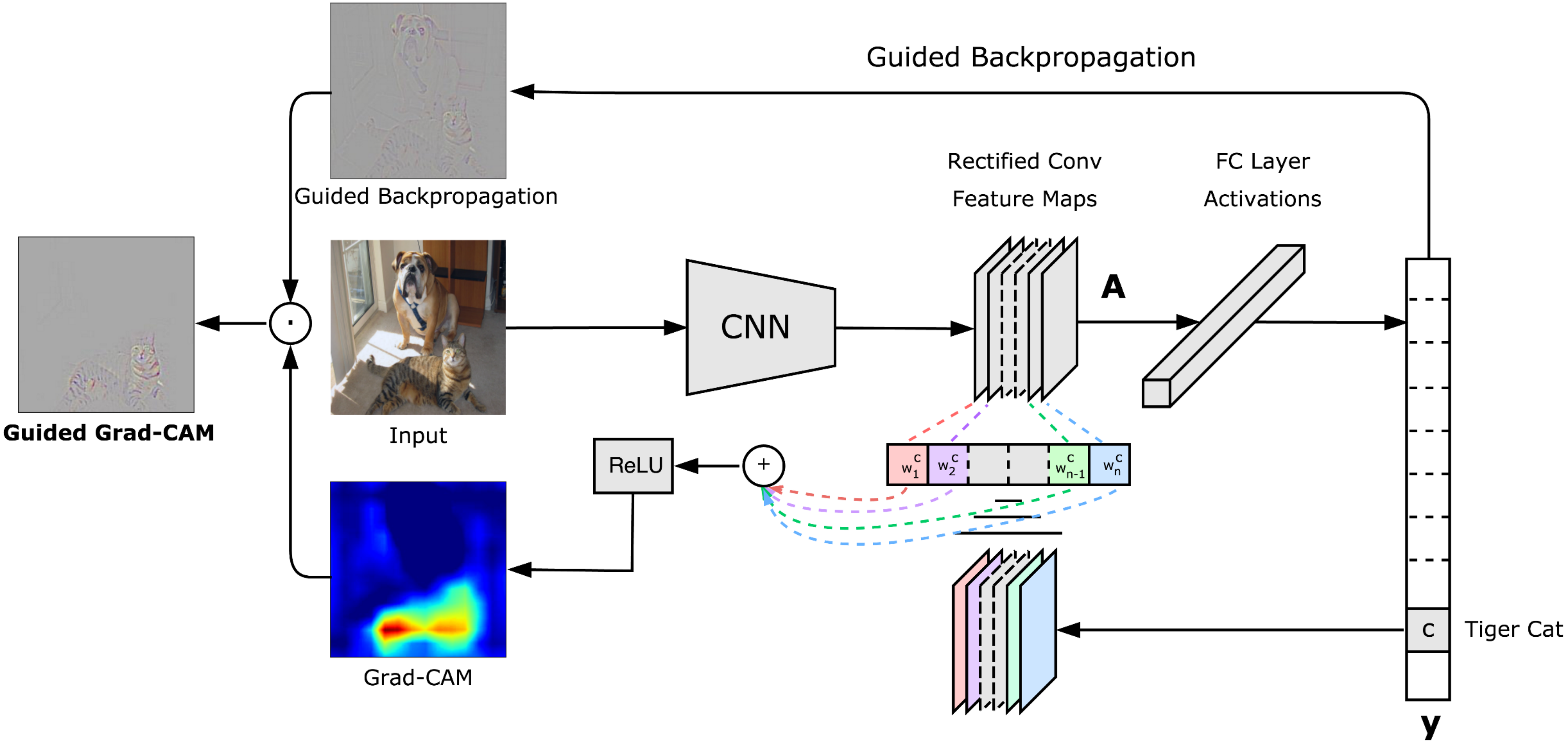


(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d)Guided Grad-CAM 'Cat' (e) Occlusion map 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j)Guided Grad-CAM 'Dog' (k) Occlusion map 'Dog' (l)ResNet Grad-CAM 'Dog'

https://arxiv.org/pdf/1610.02391.pdf

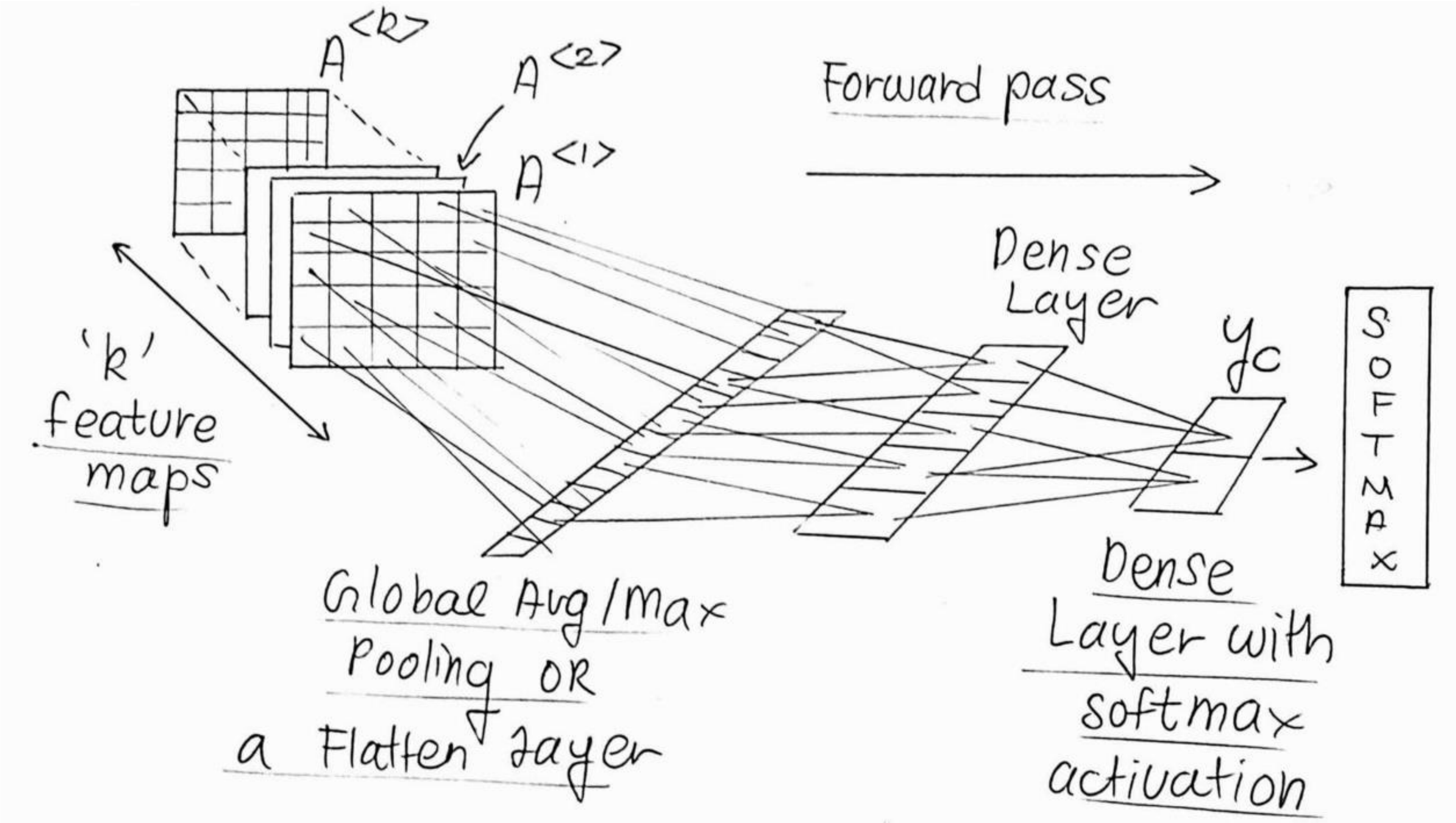
Overall architecture



http://gradcam.cloudcv.org/

Gradient of output versus bottom layers of a CNN

$A \in \mathbb{R}^{K \times W \times H}$
$A^k \in \mathbb{R}^{W \times H}, 1 \leq k \leq K$



$$\frac{\partial y^{cat}}{\partial A^k}$$

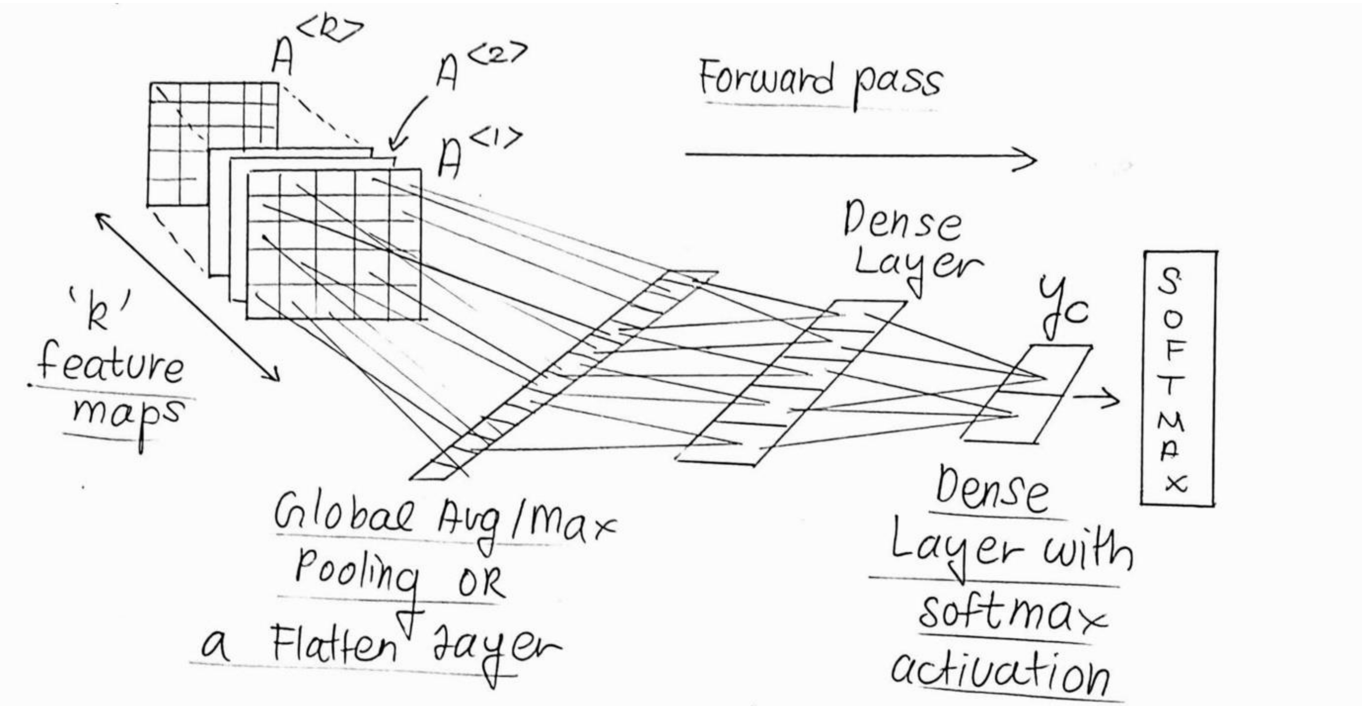
$$\frac{\partial y^{cat}}{\partial A^k_{i,j}}$$

$1 \leq i \leq W$
$1 \leq j \leq H$

https://towardsdatascience.com/grad-cam-camera-for-your-models-decision-1ef69aae8fe7

Generate a score for each feature map

$$A \in \mathbb{R}^{K \times W \times H}$$
$$A^k \in \mathbb{R}^{W \times H}, 1 \leq k \leq K$$



$$\frac{\partial y^{cat}}{\partial A_{i,j}^{k}}, 1 \leq i \leq W, 1 \leq j \leq H$$

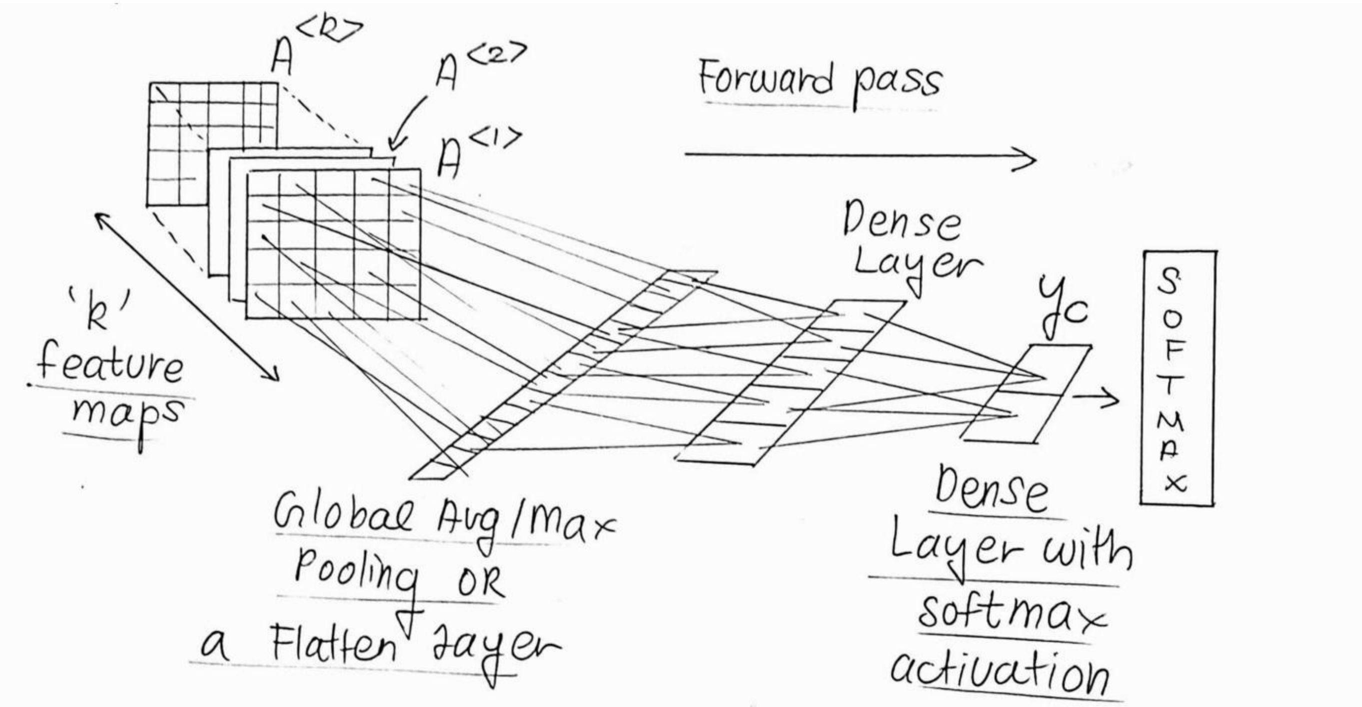
$$\alpha_k^C = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^{cat}}{\partial A_{i,j}^{k}}$$

$$Z = W \times H$$

https://towardsdatascience.com/grad-cam-camera-for-your-models-decision-1ef69aae8fe7

Generating the Grad-CAM heat map

$$A \in \mathbb{R}^{K\times W\times H}$$
$$A^k \in \mathbb{R}^{W\times H}, 1 \le k \le K$$



$$\frac{\partial y^{cat}}{\partial A^k_{i,j}}, 1 \le i \le W, 1 \le j \le H$$

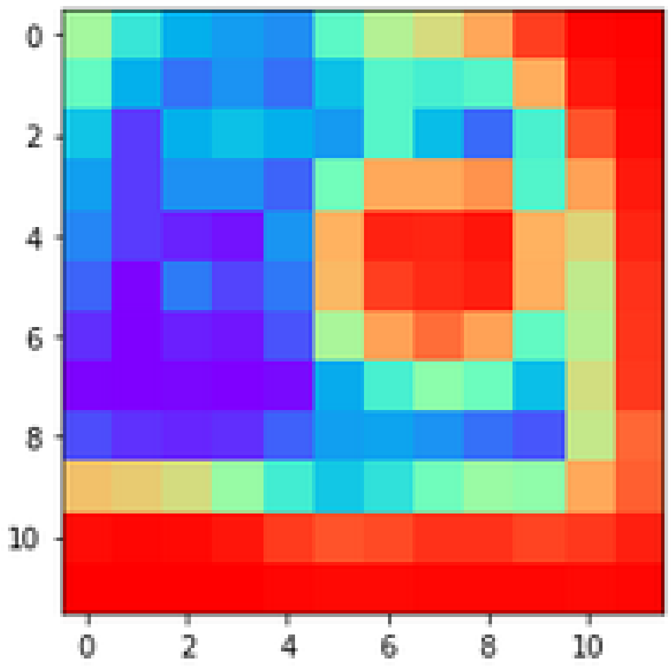
$$\alpha_k^C = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^{cat}}{\partial A^k_{i,j}}, Z = W \times H$$

$$s = \sum_{k=1}^{K} \alpha_k^C A^k$$

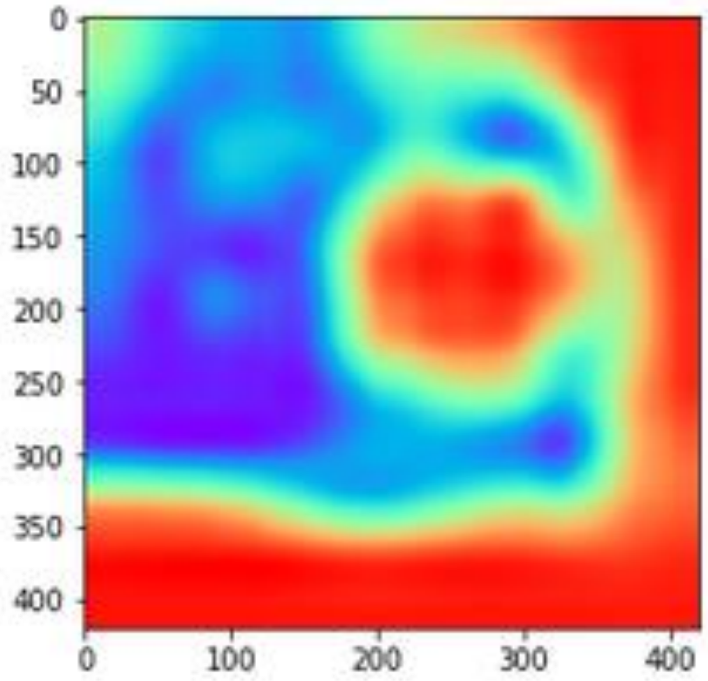
$$L^C_{GradCAM} = RELU(s)$$

Won't the Grad-CAM Heat map Be Too Small?

12 × 12 heat map

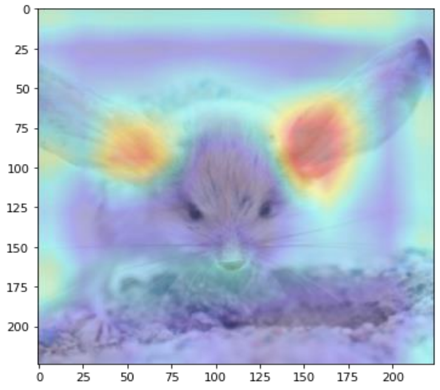
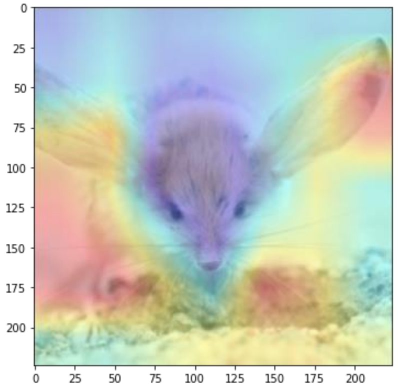
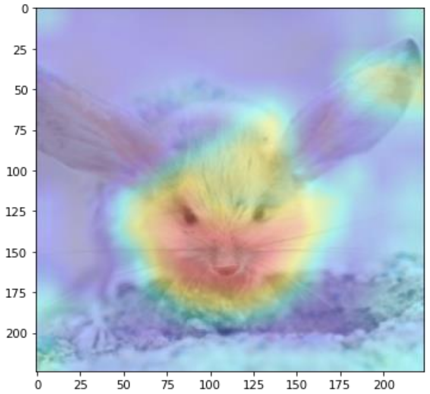
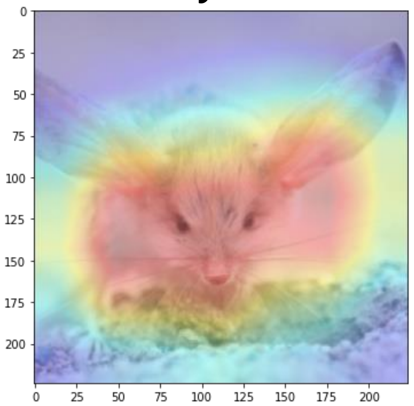
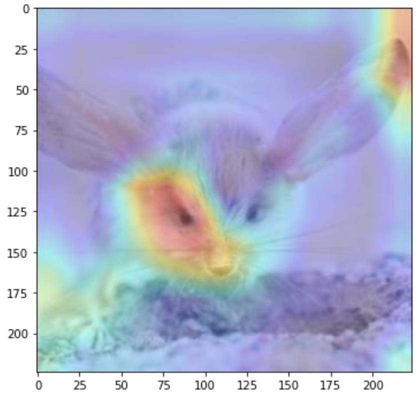
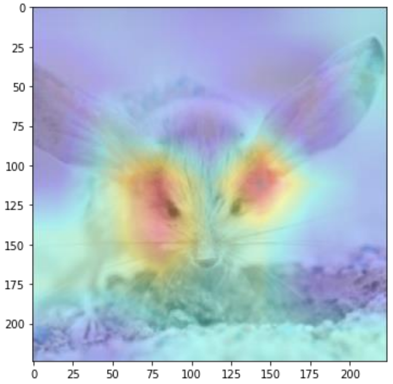
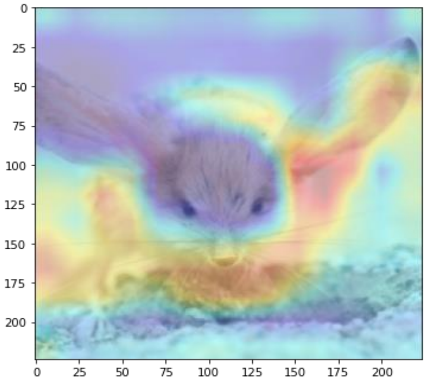
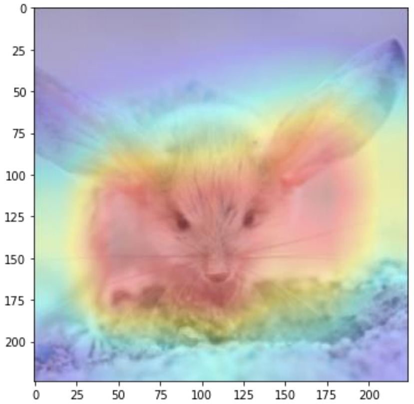


Same heat map upsampled to 420 × 420 using the Python package cv2



https://glassboxmedicine.com/2020/05/29/grad-cam-visual-explanations-from-deep-networks/

Use GradCAM to visualize focused area

	AlexNet (features_11)	VGG16 (features_30)	VGG19 (features_35)	ResNet18 (layer4)
Class predicted by NN	 (墨西哥鈍口螈 29)	 (野兔 331)	 (333 倉鼠)	 (330棉尾兔)
Class manually assigned	 (野兔 331)	 (333 倉鼠)	 (野兔 331)	 (野兔 331)

Ref: 1061307林家禾 (2021)