E9: 309 ADL 16-12-2020

http://leap.ee.iisc.ac.in/sriram/teaching/ADL2020/

Recap from previous lectures

- Analyzing trained neural networks
  - ✓ Hierachical representations
    - ◉ Maximizing activations
    - ◉ Visualizing representations
    - ◉ Reconstruction of input patterns from hidden layers.

* Transferability of representations

Today's lecture

- ✴️ Why models predict what they predict

Architecture updates for interpretability

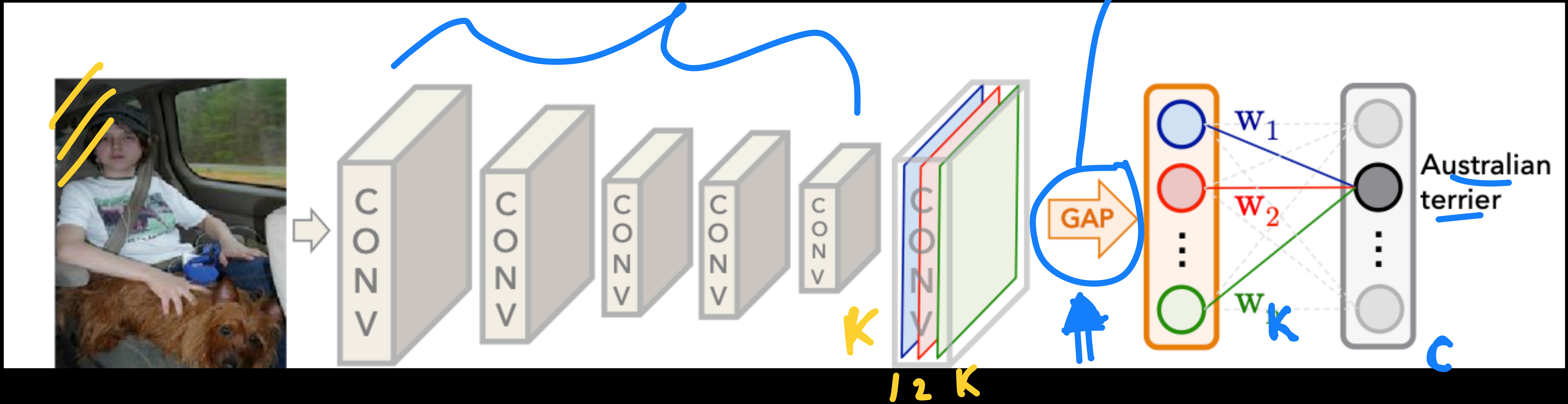
Learning Deep Features for Discriminative Localization

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba
Computer Science and Artificial Intelligence Laboratory, MIT
{bzhou,khosla,agata,oliva,torralba}@csail.mit.edu

2015

Learning the input pattern of a trained network

Global average pooling



Architecture updates for interpretability

- Global average pooling

$$\boxed{F^{k,L} = \sum_{i,j} \mathbf{f}^{k,L}(i,j)}$$

k - index of feature map

$k = 1 .. K$

L - last layer

- Last layer mapping to classes

$$a^{c,L} = \sum_{k=1}^{K} w_c^{k,L} F^{k,L}$$

$$\hat{\mathbf{y}} = softmax(\mathbf{a}^L)$$

$$a^L = \begin{bmatrix} a^{1,L} \\ a^{c,L} \\ \vdots \\ a^{C,L} \end{bmatrix}$$

Architecture updates for interpretability

- Rearranging the terms

$$a^{c,L} = \sum_{k=1}^{K} w_c^{k,L} \overbrace{\sum_{i,j} \mathbf{f}^{k,L}(i,j)}^{\text{GAP}}$$

$$= \sum_{i,j} \underbrace{\sum_{k=1}^{K} w_c^{k,L} \mathbf{f}^{k,L}(i,j)}_{\text{Image}}$$

$a^{c,L}$ → output activation

- Defining the map

$$\mathbf{m}^{c}(i,j) = \underbrace{\sum_{k=1}^{K} w_c^{k,L} \mathbf{f}^{k,L}(i,j)}$$

Class activation maps

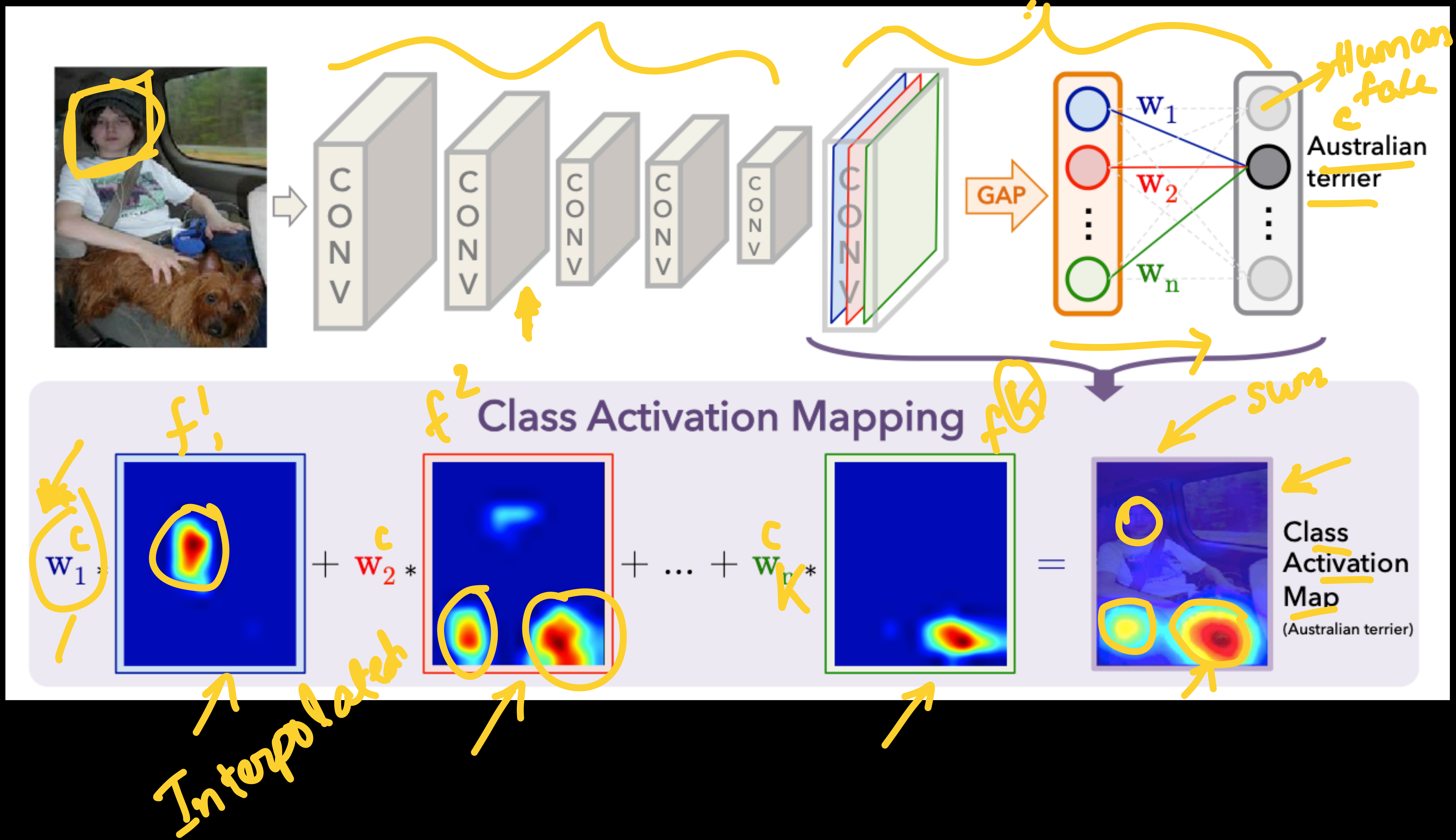
Architecture updates for interpretability

CAM

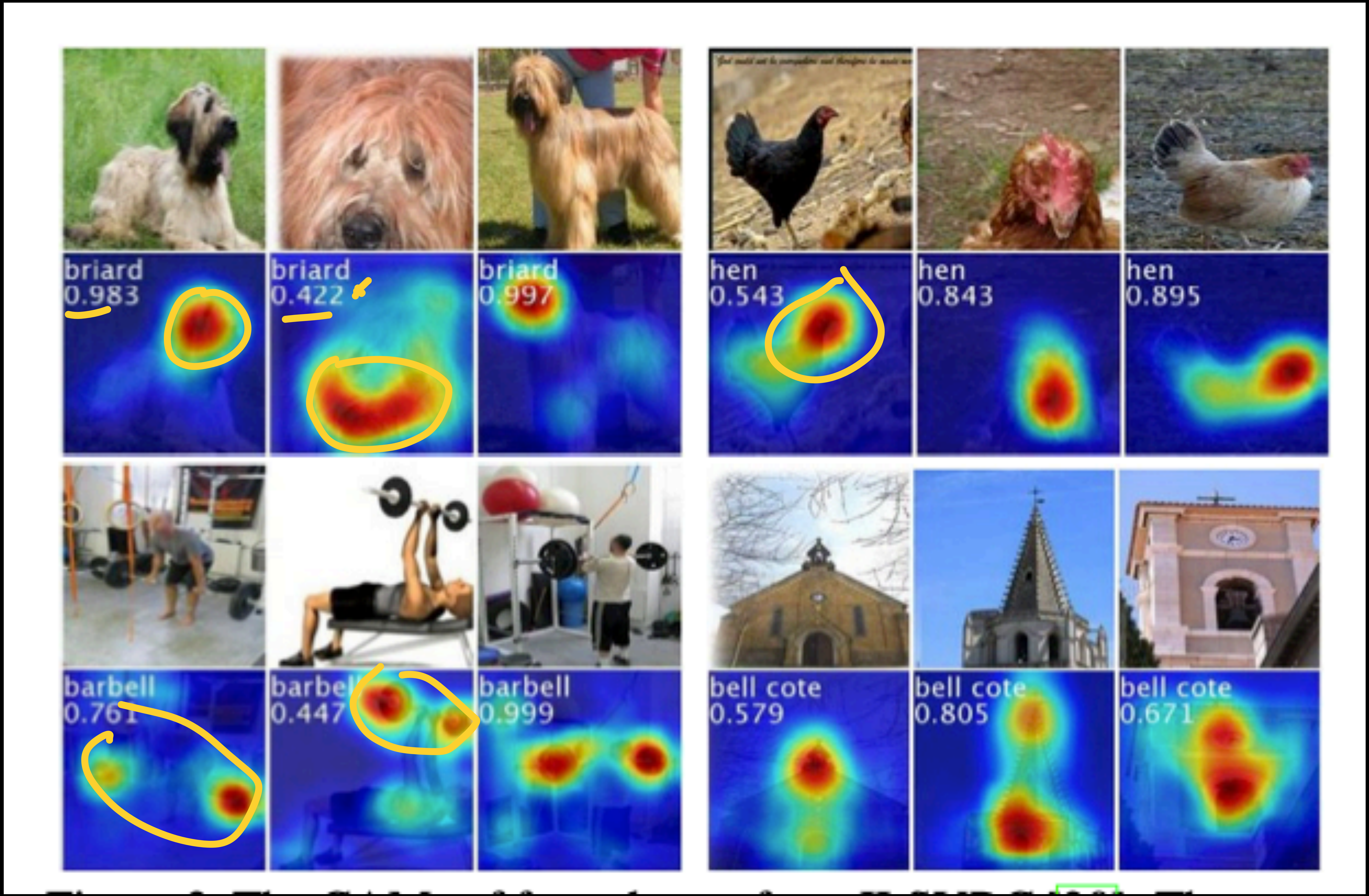
- The maps denote spatial patterns that define how important a particular pixel is for that particular class.

- The map can be interpolated to the original input dimensions to visualize the pattern

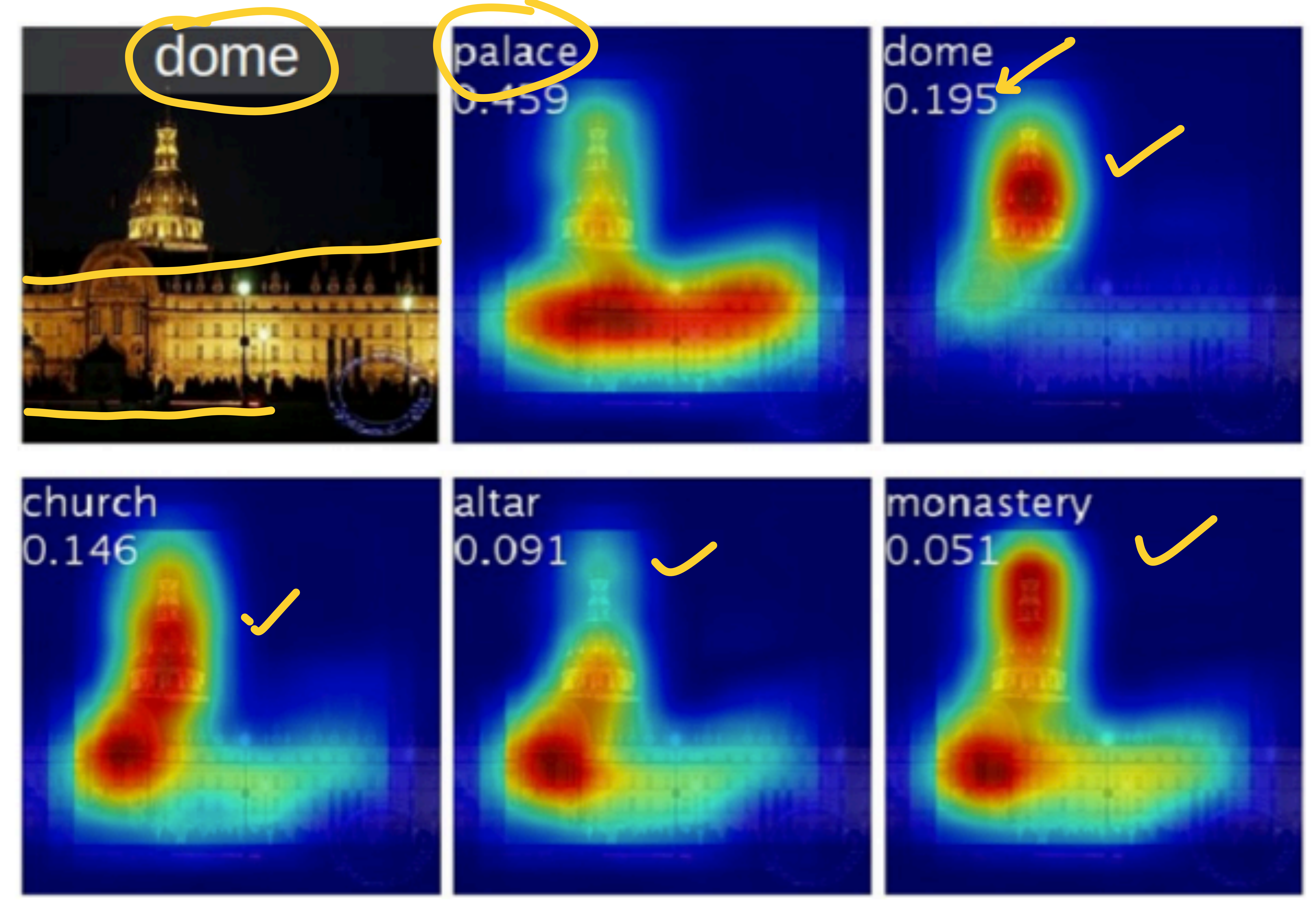
Architecture updates for interpretability



Architecture updates for interpretability



Architecture updates for interpretability



Architecture updates for interpretability

- Flaws in the approach
  - Requires Global average pooling based
  - The model may be inferior to the other models with fully connected layers after the CNN layers

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

Grad-CAM : Visual Explanations ....

Selvaraju et. al (2016 CVPR)

Improving CAM without compromising architecture

- Find the gradient of the output activation with respect to the feature maps of the last convolutional layer

$$\alpha_k^c = \frac{1}{S} \sum_{i,j} \frac{\partial a^{c,L}}{\partial \mathbf{f}^{k,l}(i,j)}$$

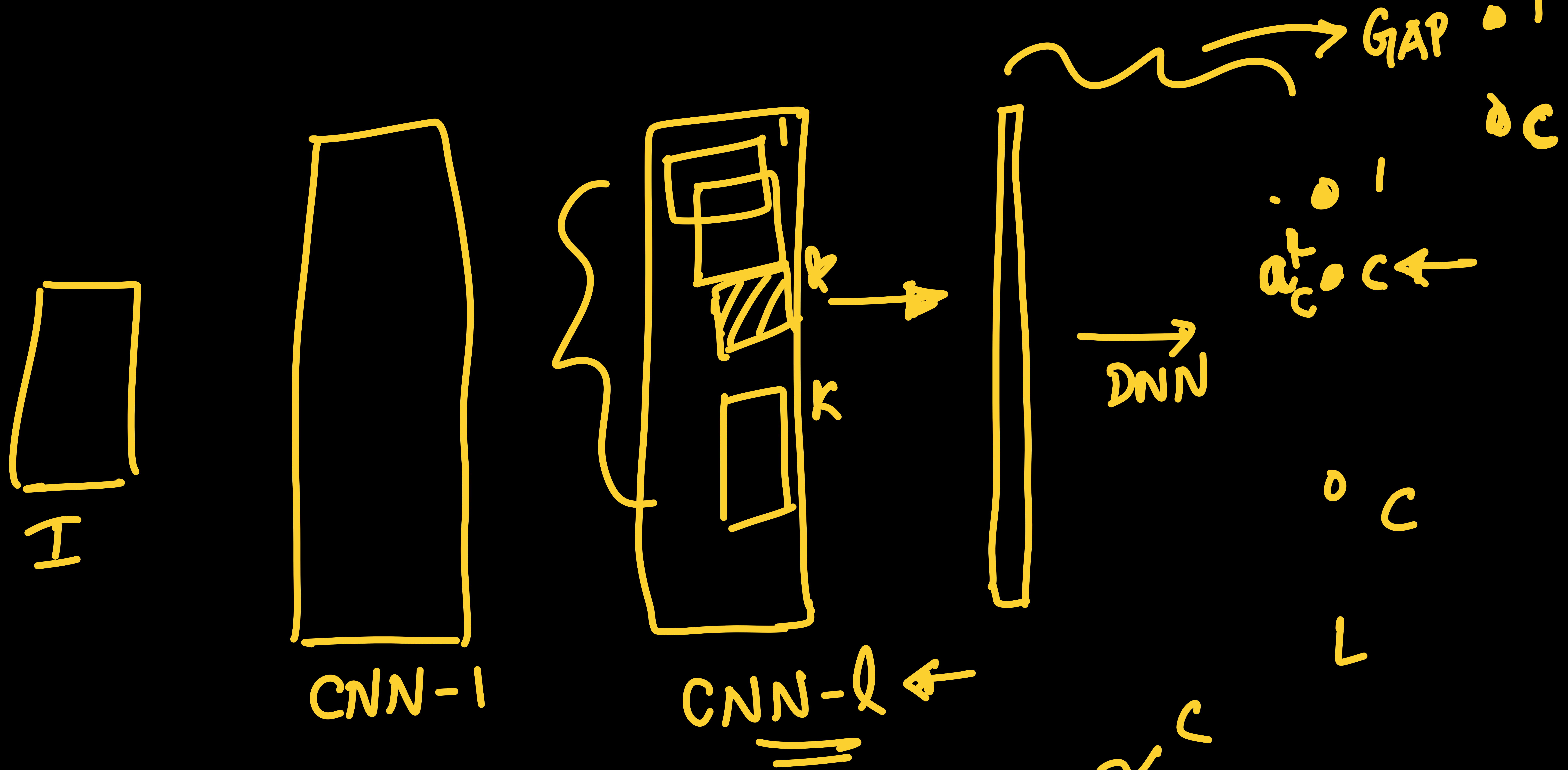
$f^{k,l}(i,j)$

- Assumption - The feature maps from the last convolutional layer capture the spatial information as well as the semantic information required for classification.

- $$S = \sum_{i,j} 1$$

l - last CNN layer

k - feature map index



$$\frac{\partial a_c^L}{\partial f_k^l(i,j)}$$

$$\alpha_k^c$$

Improving CAM without compromising architecture

- Gradient based activation maps (Grad-CAM)

CAM output

$$\mathbf{m}^c = ReLU\Big(\sum_k \alpha_k^c \mathbf{f}^{k,l}(i,j)\Big)$$

Interpolate

- Multiply the gradient based contribution to the individual pixels of the feature maps
- The ReLU operation preserves only the pixels that have a positive influence on the output activations.
  - ➡️ The negative pixels may belong some other class category

Relation between CAM and Grad-CAM

- If the last convolutional layer is followed by a global average pooling (GAP) and softmax layer then, Grad-CAM gives

$$F^{k,L} = \sum_{i,j} \mathbf{f}^{k,L}(i,j)$$

GAP

- The derivative output activation w.r.t GAP output

$$\frac{\partial a^{c,L}}{\partial F^{k,L}} = w_c^{k,L} \qquad \boxed{\frac{\partial a^{c,L}}{\partial \mathbf{f}^{k,L}(i,j)} = w_c^{k,L}}$$

Relation between CAM and Grad-CAM

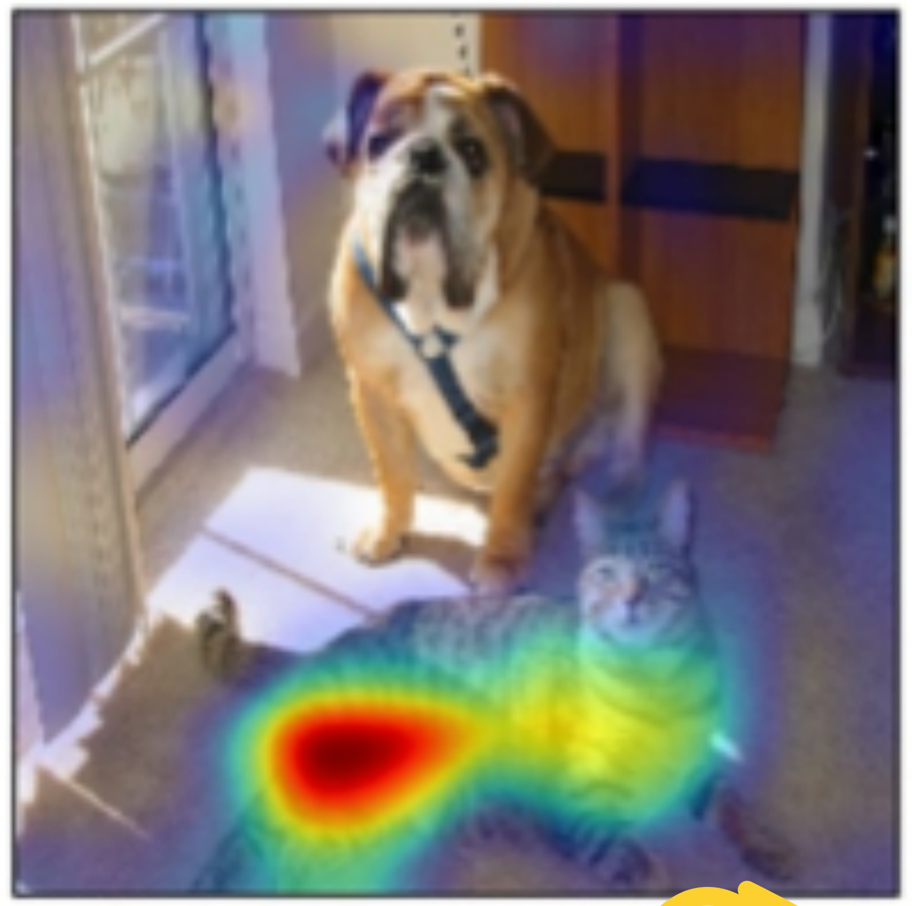
- If the last convolutional layer is followed by a global average pooling (GAP) and softmax layer then, Grad-CAM gives

$$\alpha_k^c = \frac{1}{S}\sum_{i,j}\frac{\partial a^{c,L}}{\partial \mathbf{f}^{k,l}(i,j)} \longrightarrow \alpha_c^k = w_c^{k,L}$$

- Grad-CAM generalizes the CAM framework to neural networks that can have convolutional networks followed by other architectures.

Visualizing Grad-CAM outputs

Tiger-cat



(c) Grad-CAM 'Cat'

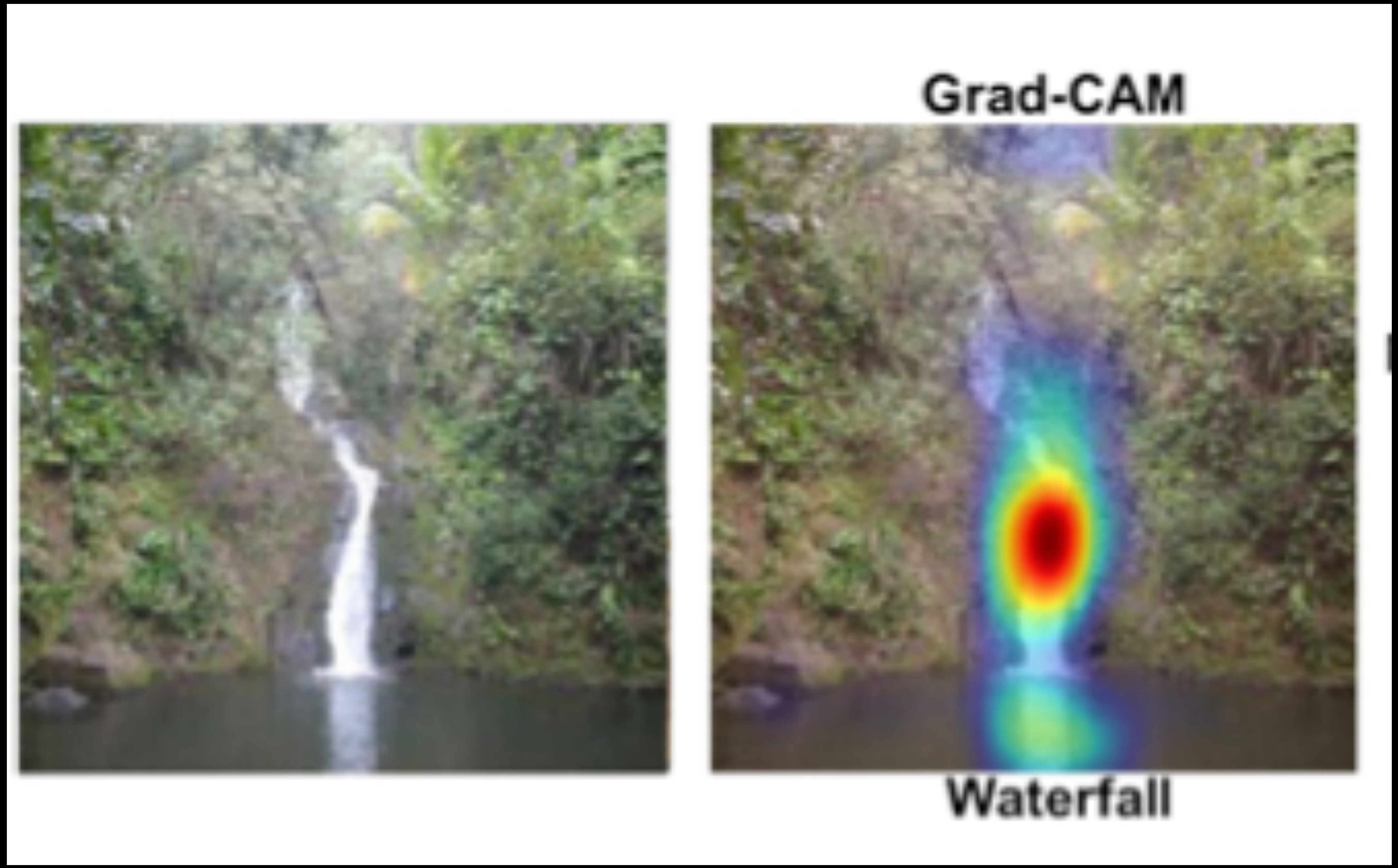
(i) Grad-CAM 'Dog'

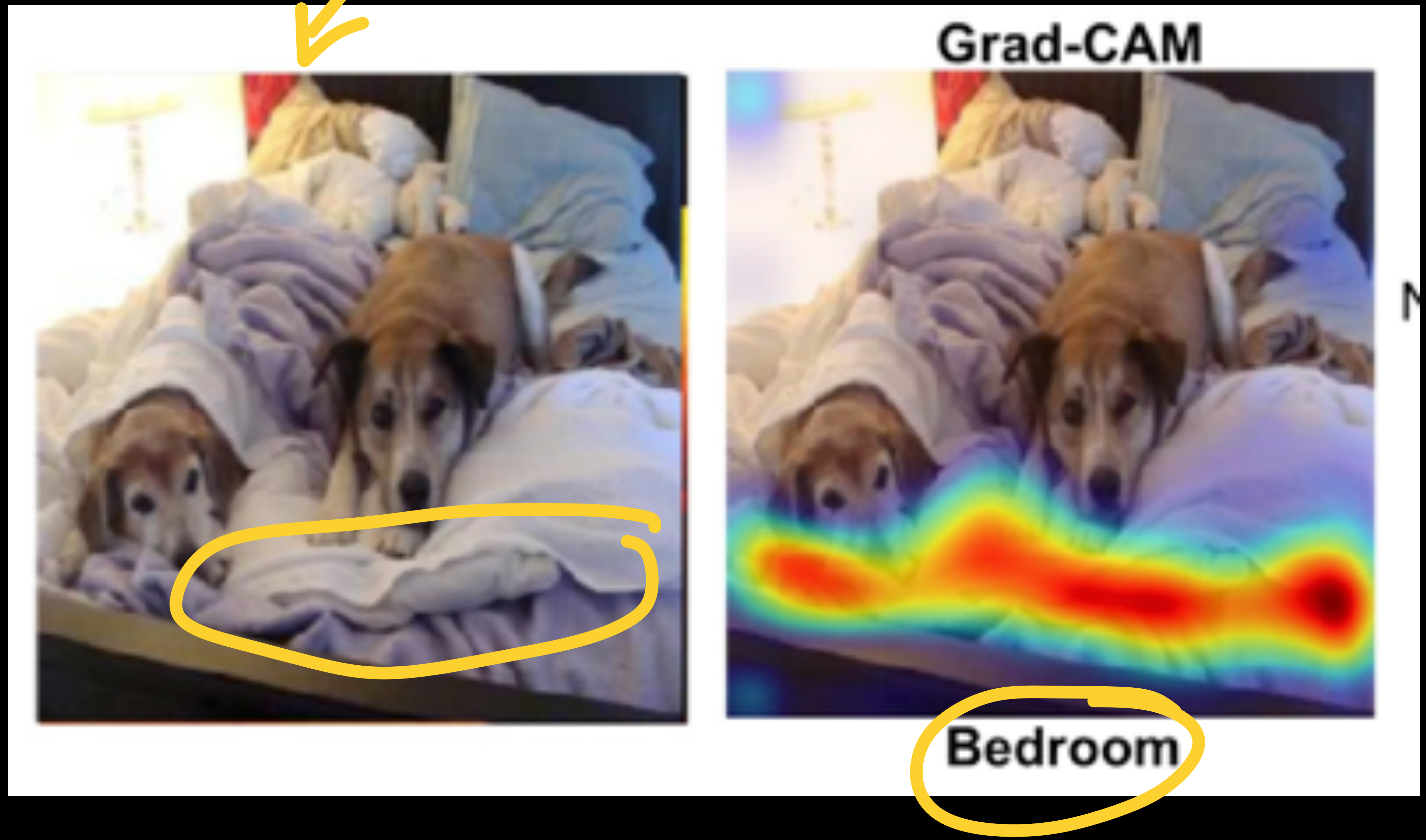
100

101

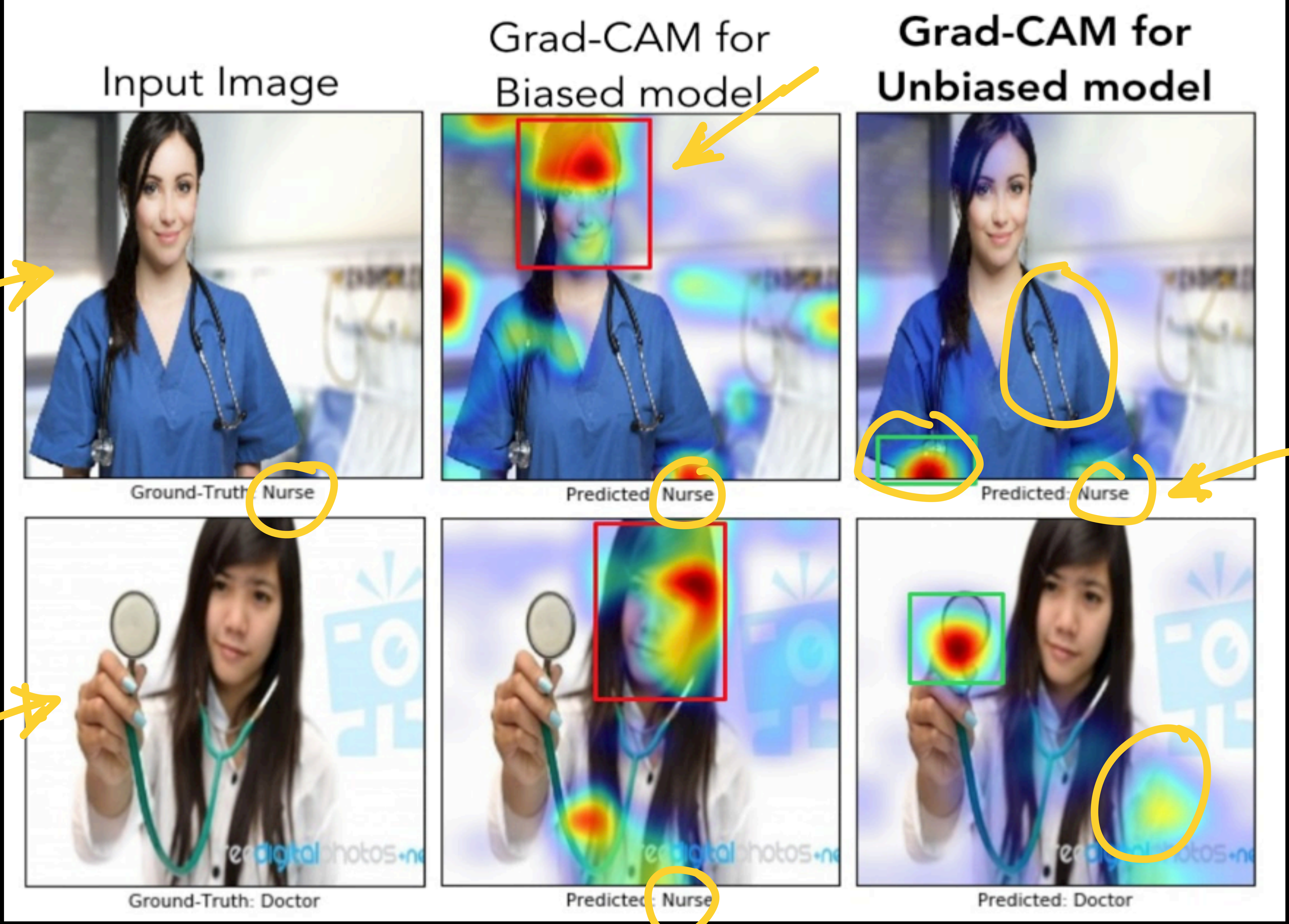
90

Grad-CAM in image captioning



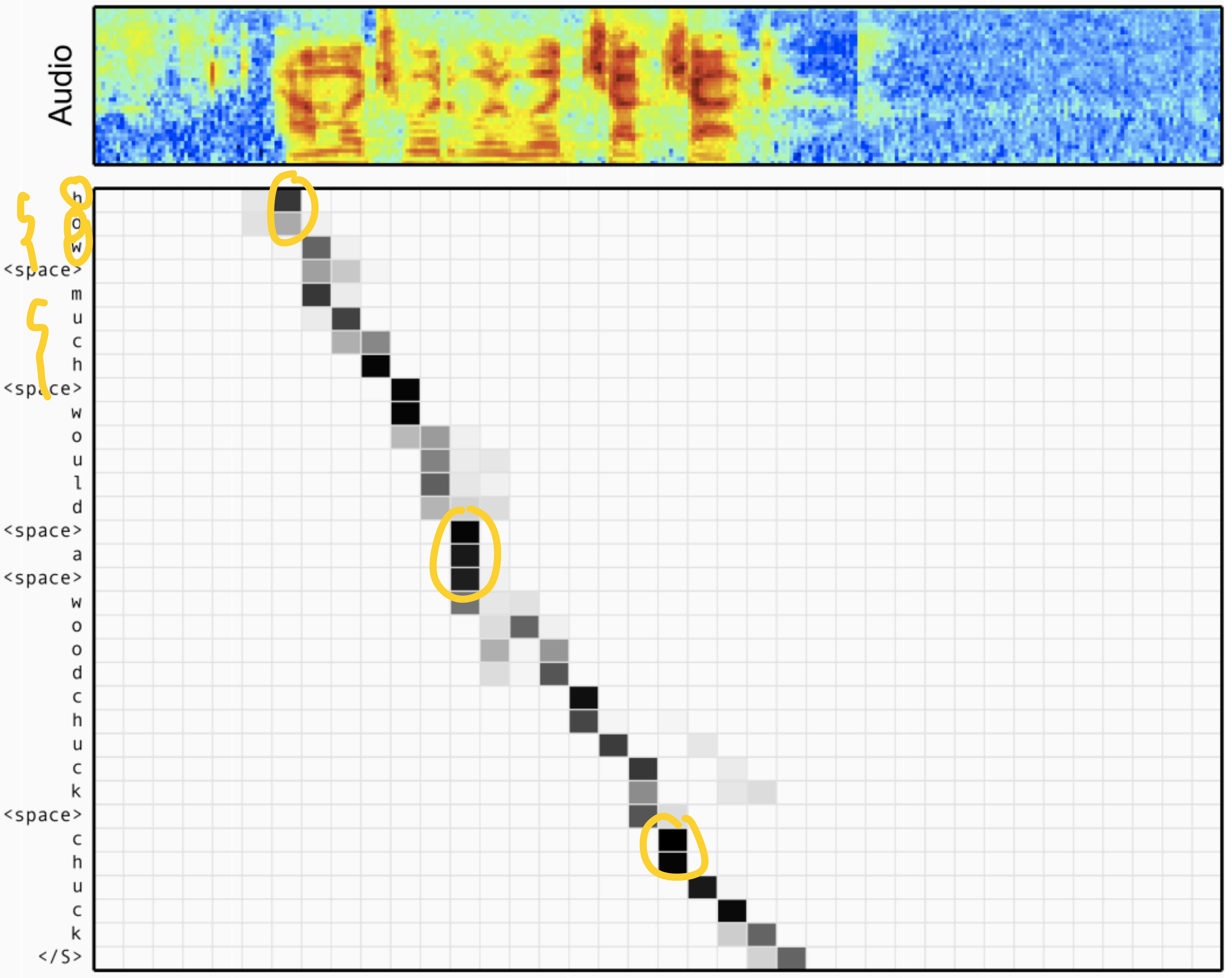


Grad-CAM for identifying model biases



Using attention for visualization

Audio

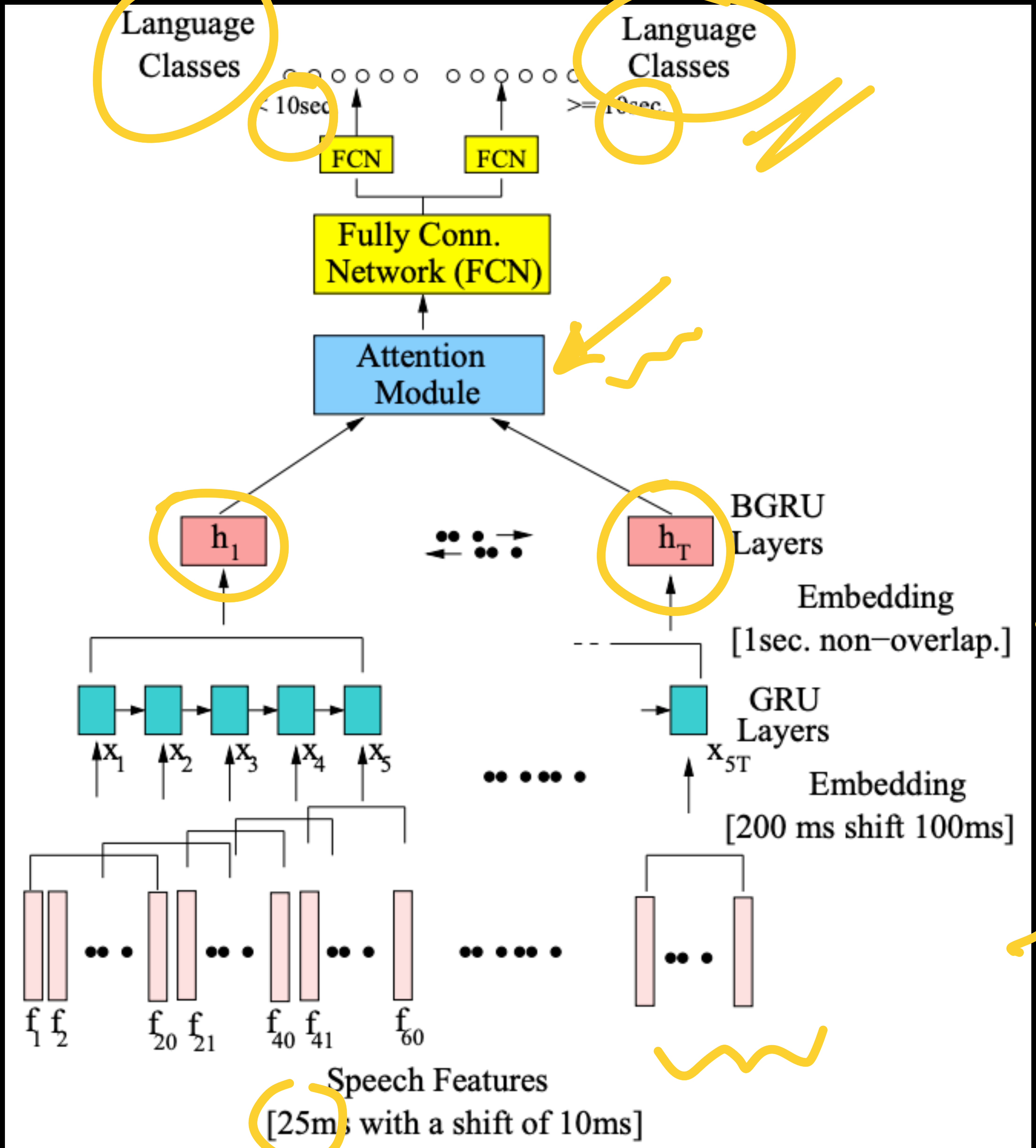


Using attention mechanism for explainability

Towards Relevance and Sequence Modeling in Language Recognition

Bharat Padi, Anand Mohan and Sriram Ganapathy, *Senior Member, IEEE*

Using attention mechanism for explainability



Speech

Eng

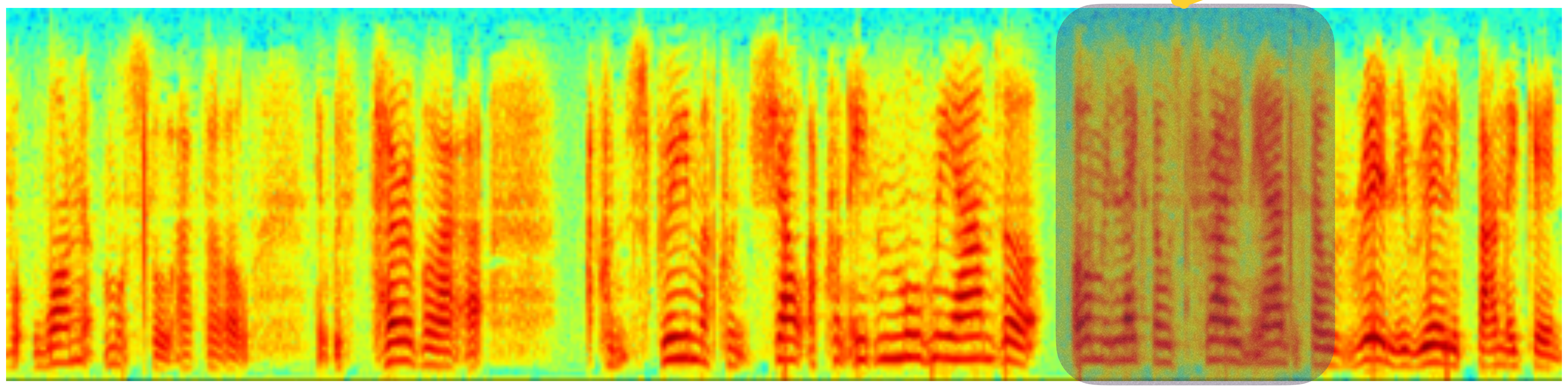
1s

1/5

1/40

Using attention mechanism for explainability

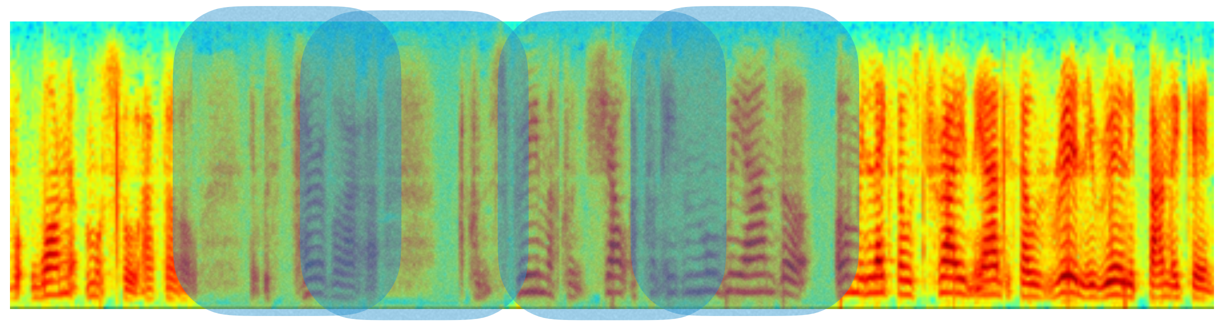
- Certain regions of the audio signal may have more information for the task than the rest.
  - ✓ May also have more signal quality than rest.
  - ✓ For example, language identification involving Eng-UK v/s Eng. US.

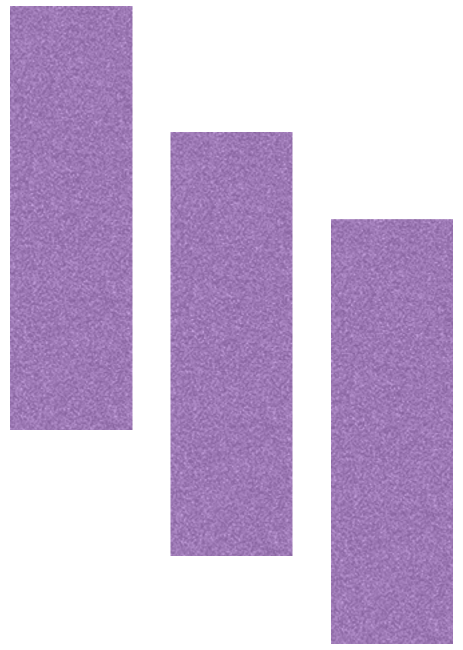


- Current models - use the information from audio with uniformity.

Using attention mechanism for explainability

- Derive short segment i-vectors





Sequence to label model using **attention**

- Attention weighs the importance of each short-term segment feature for the task.

Attention Weight

0-3s

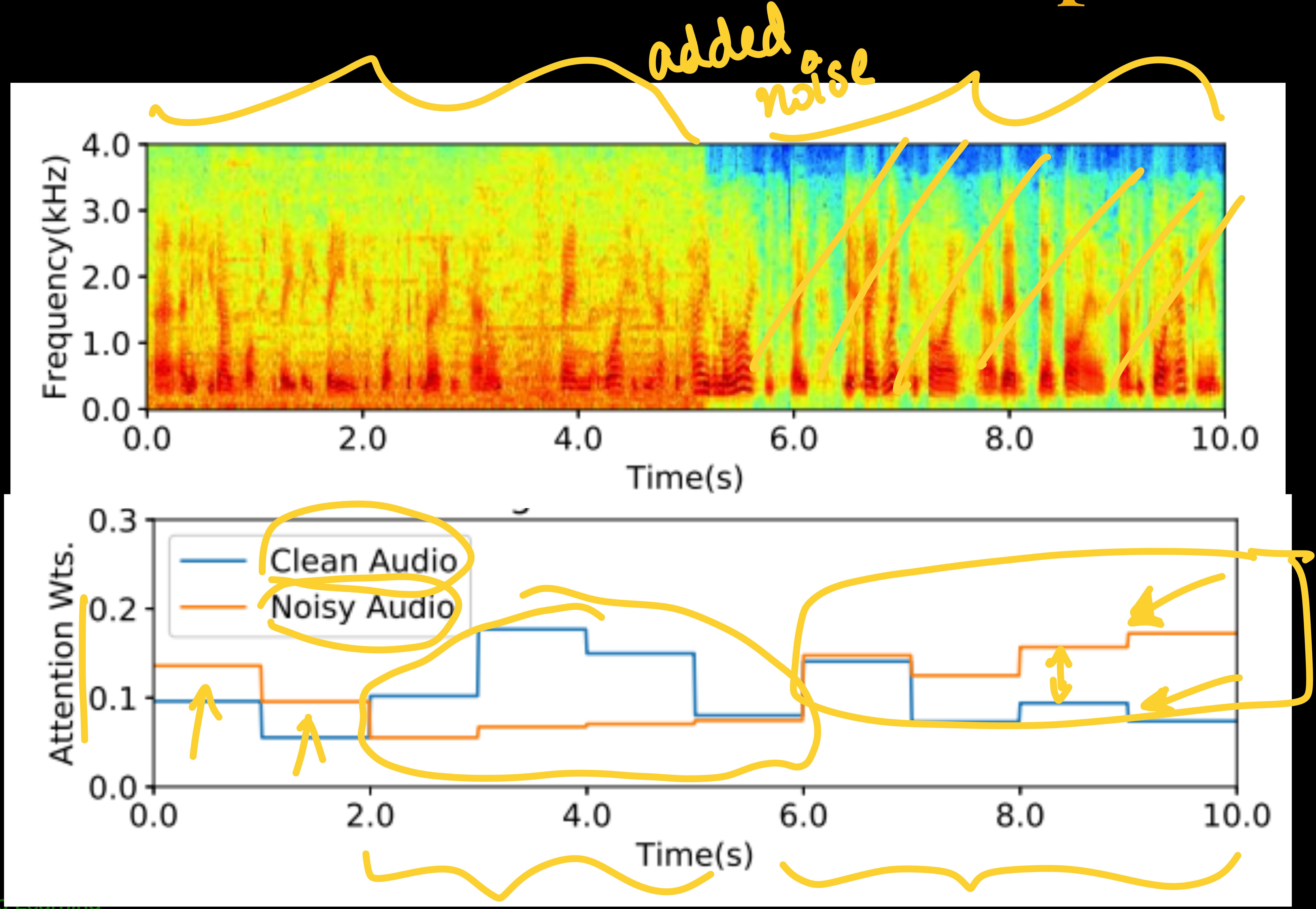
3s-4s

4s - 9s   I couldn't scream, I couldn't shout, I couldn't even move my arms up, or my legs

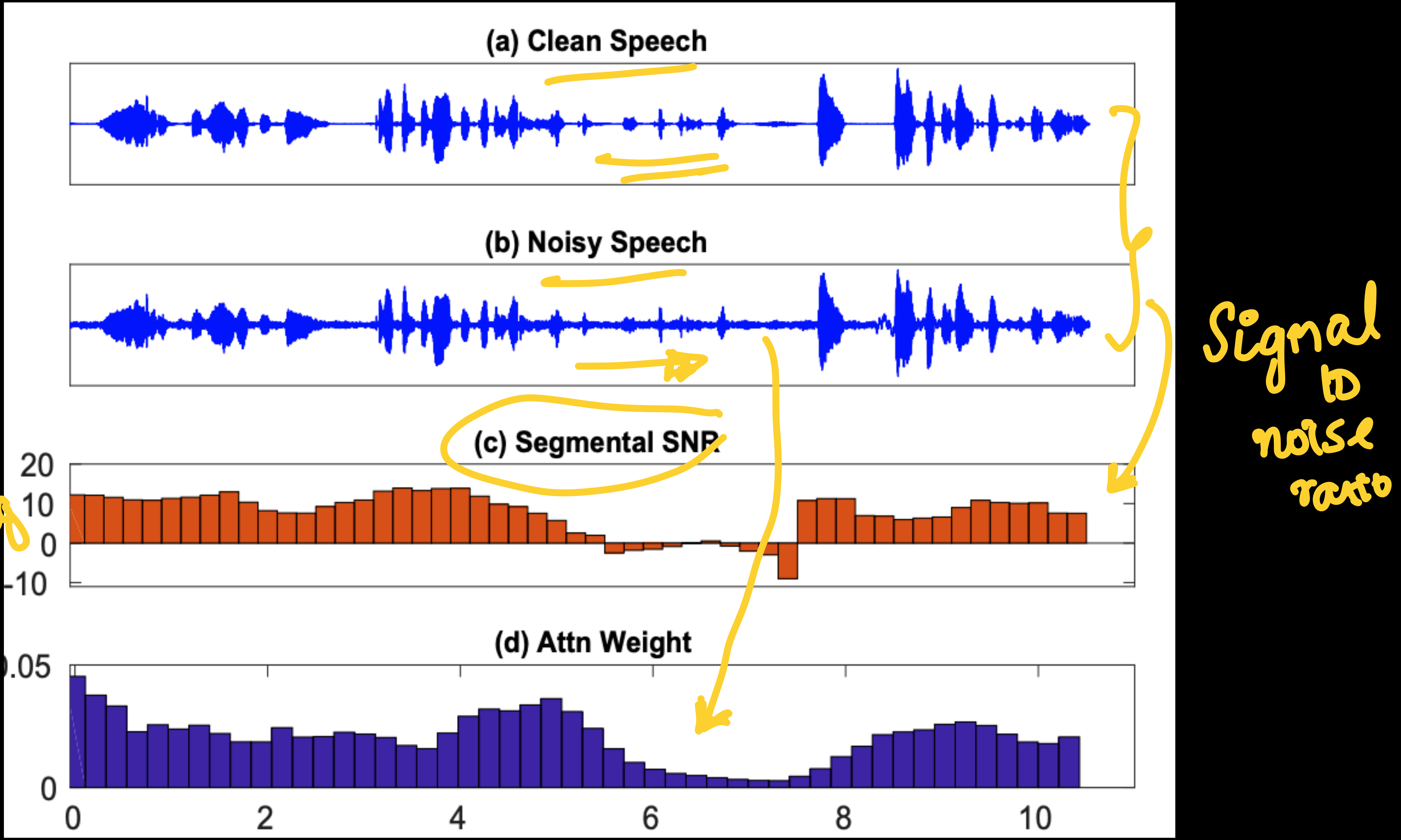
9s -11s

Bharat Padi, et al. 2018 in prep.

Using attention mechanism for explainability



Using attention mechanism for visualization



Summary thus far

- ✴ Analyzing trained neural networks
    - ✓ Hierachical representations
    - ✓ Activation maps to determine saliency
    - ✓ Incorporating attention mechanism for improved explainability