



E9: 309 ADL 16-12-2020

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

# Recap from previous lectures

- \* Analyzing trained neural networks

- ✓ Hierarchical representations

- Maximizing activations

- Visualizing representations

- Reconstruction of input patterns from hidden layers.



# Today's lecture

- \* Why models predict what they predict



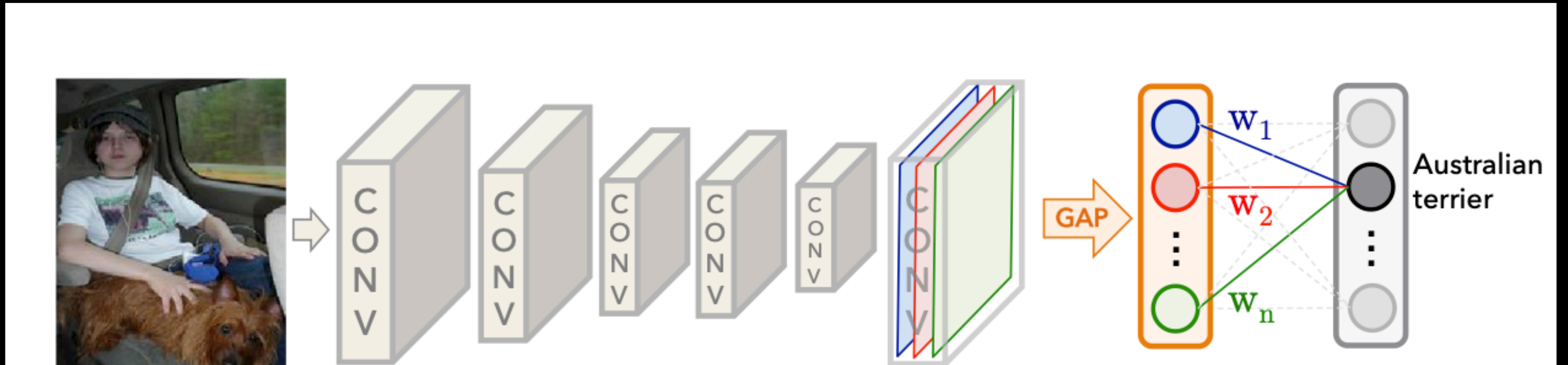
# Architecture updates for interpretability

## 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



# Learning the input pattern of a trained network



# Architecture updates for interpretability

- \* Global average pooling

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

- \* Last layer mapping to classes

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

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



# Architecture updates for interpretability

\* Rearranging the terms

$$\begin{aligned} a^{c,L} &= \sum_{k=1}^K w_c^{k,L} \sum_{i,j} \mathbf{f}^{k,L}(i,j) \\ &= \sum_{i,j} \sum_{k=1}^K w_c^{k,L} \mathbf{f}^{k,L}(i,j) \end{aligned}$$

\* Defining the map

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



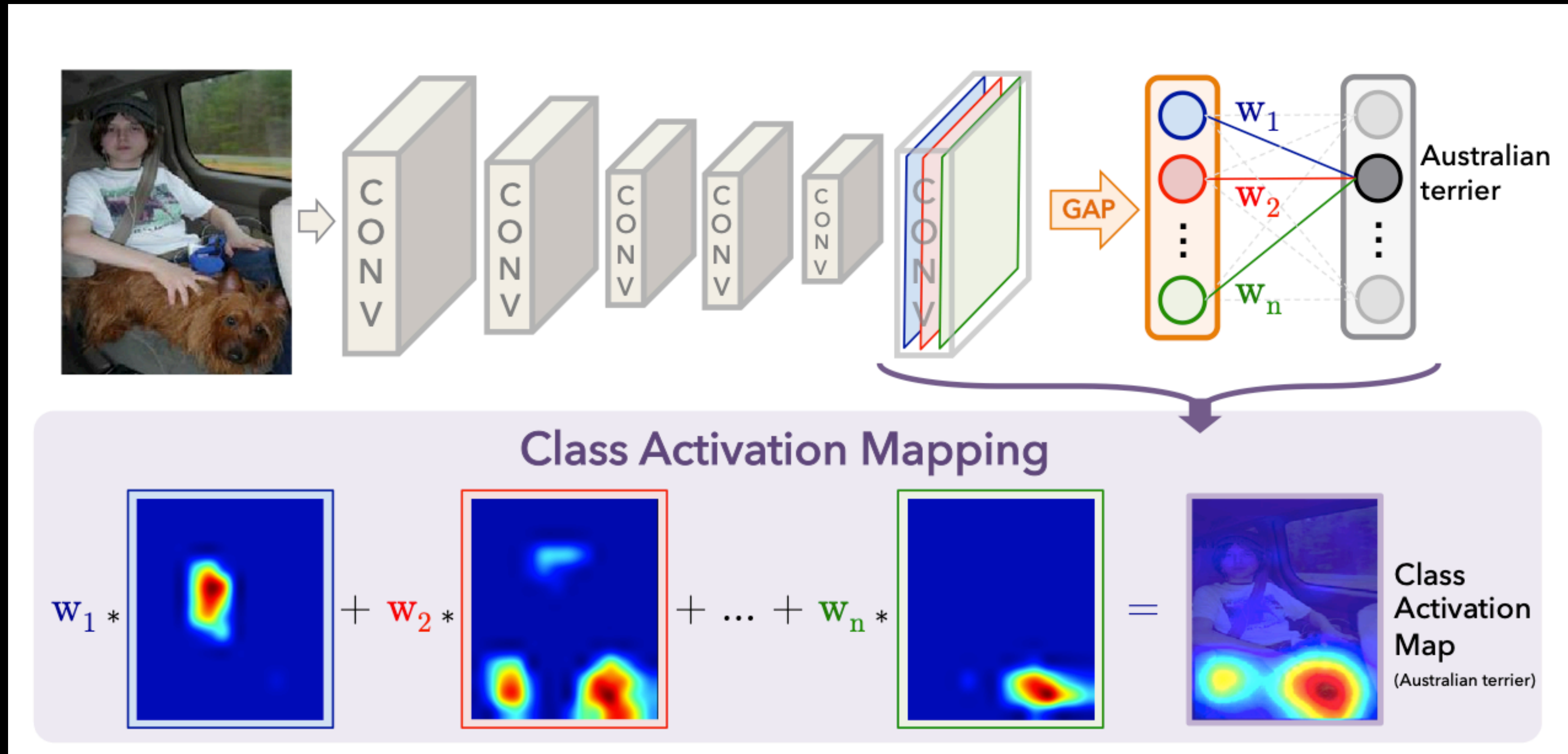
# Architecture updates for interpretability

- \* 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

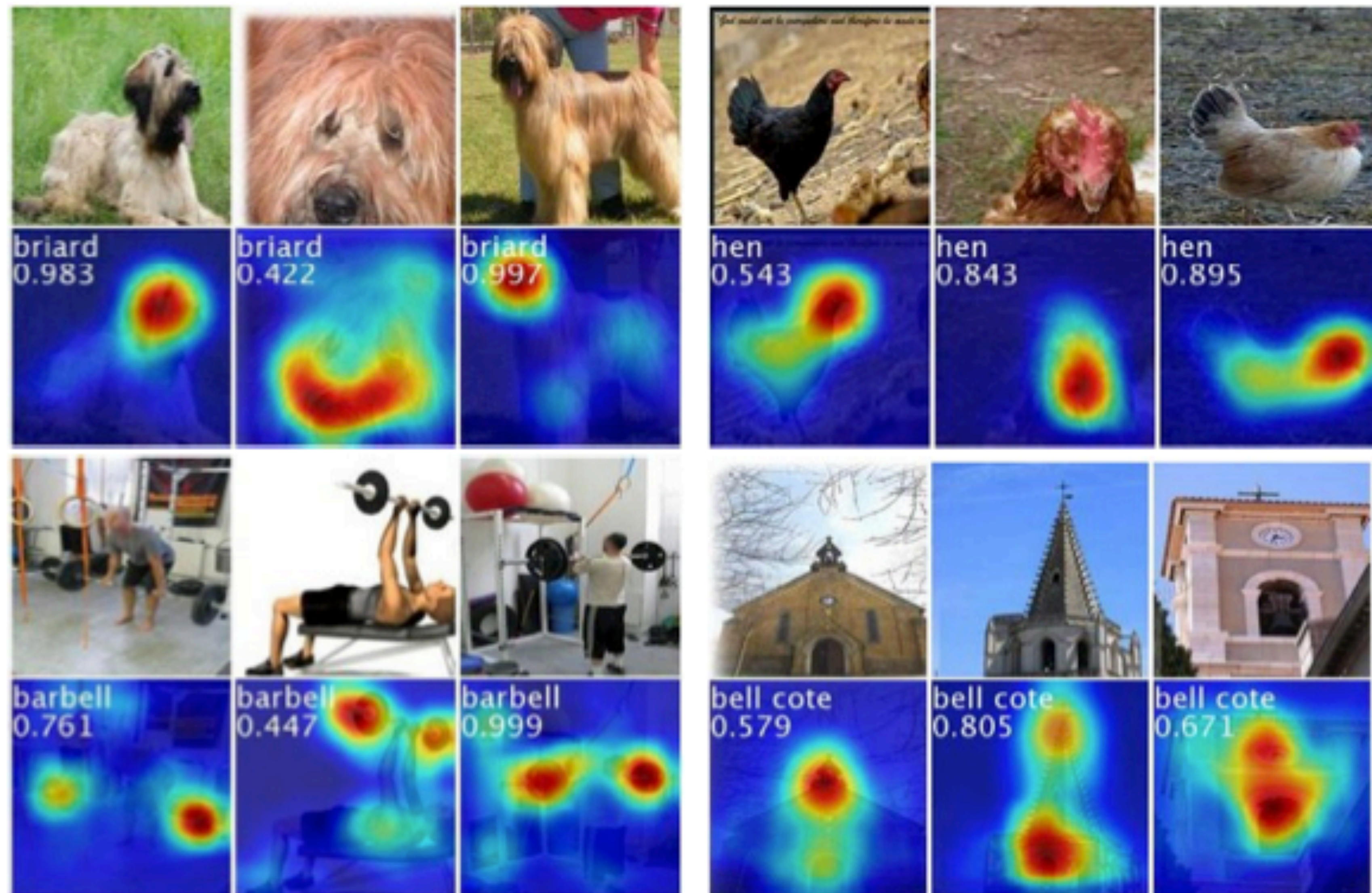




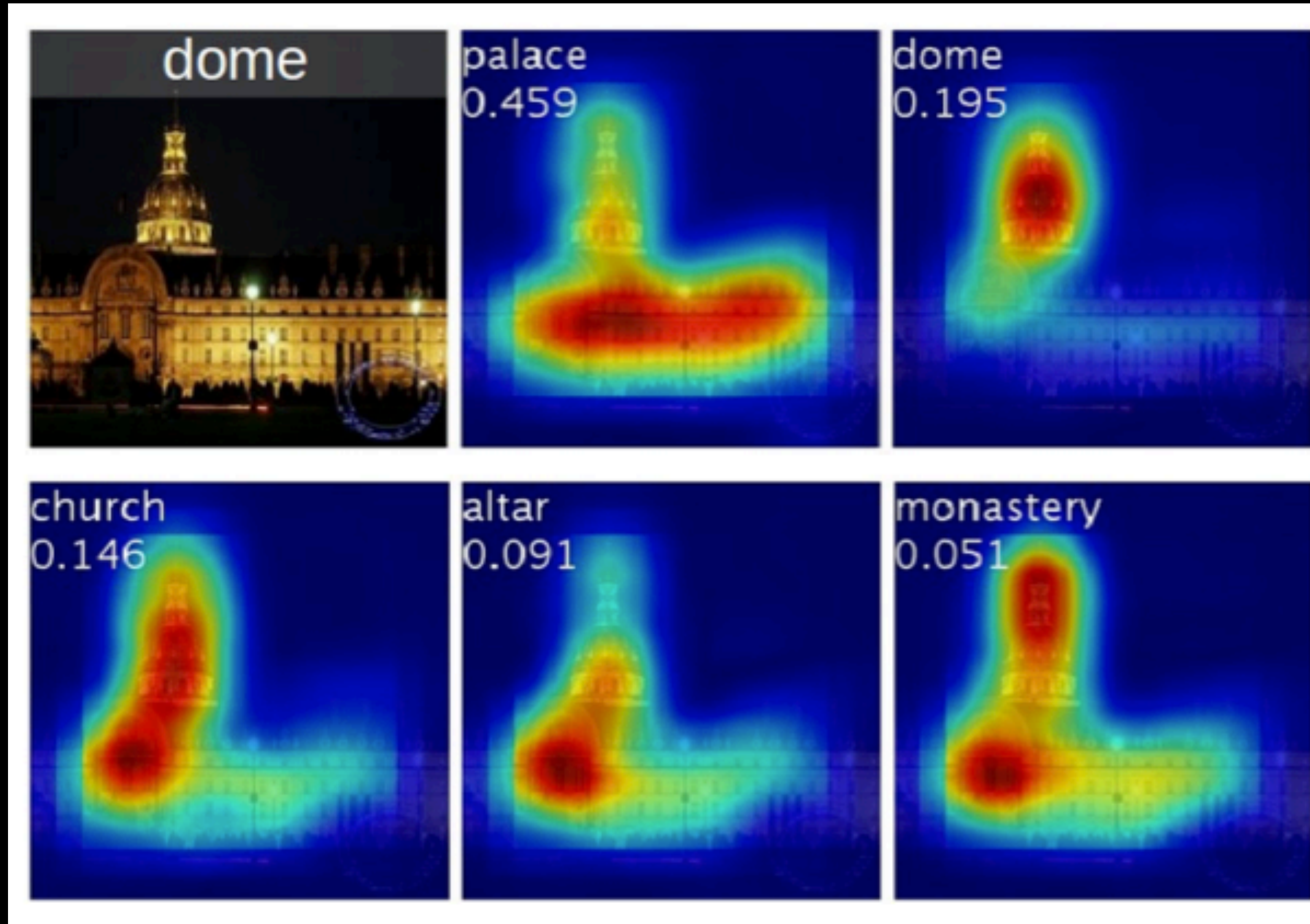
# Architecture updates for interpretability



# 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



# Improving CAM without compromising architecture

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

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

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

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



# Improving CAM without compromising architecture

- \* Gradient based activation maps (Grad-CAM)

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

- \* 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

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

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

$$\frac{\partial a^{c,L}}{\partial F^{k,L}} = w_c^{k,L}$$

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



# 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

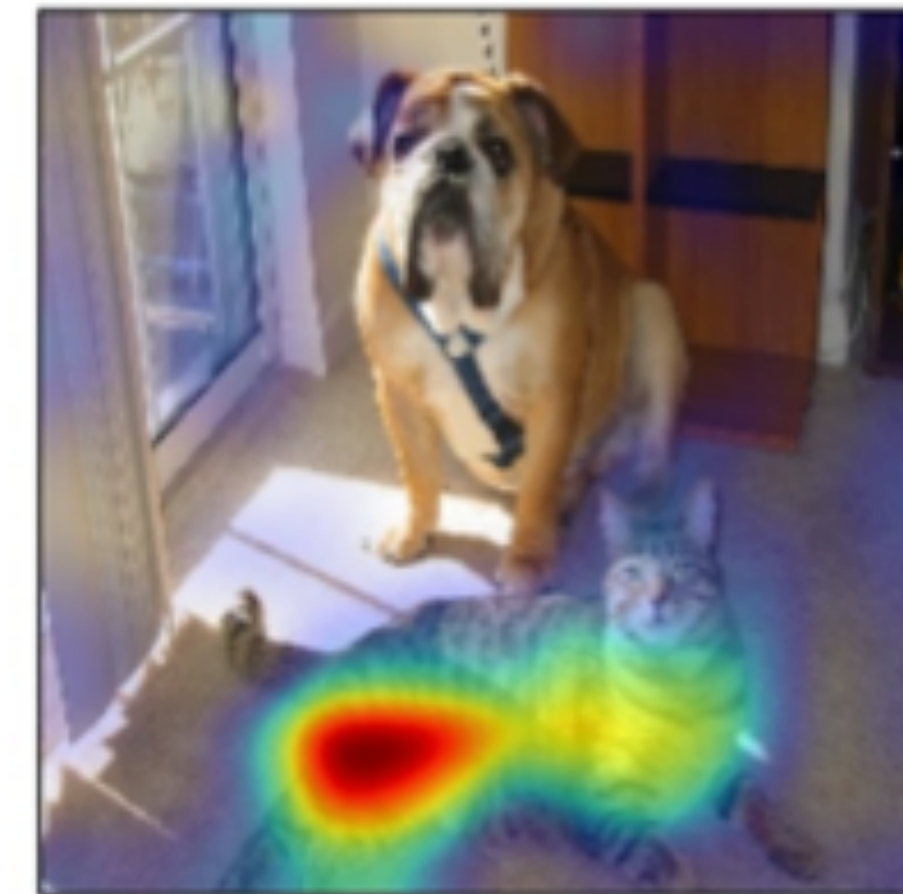
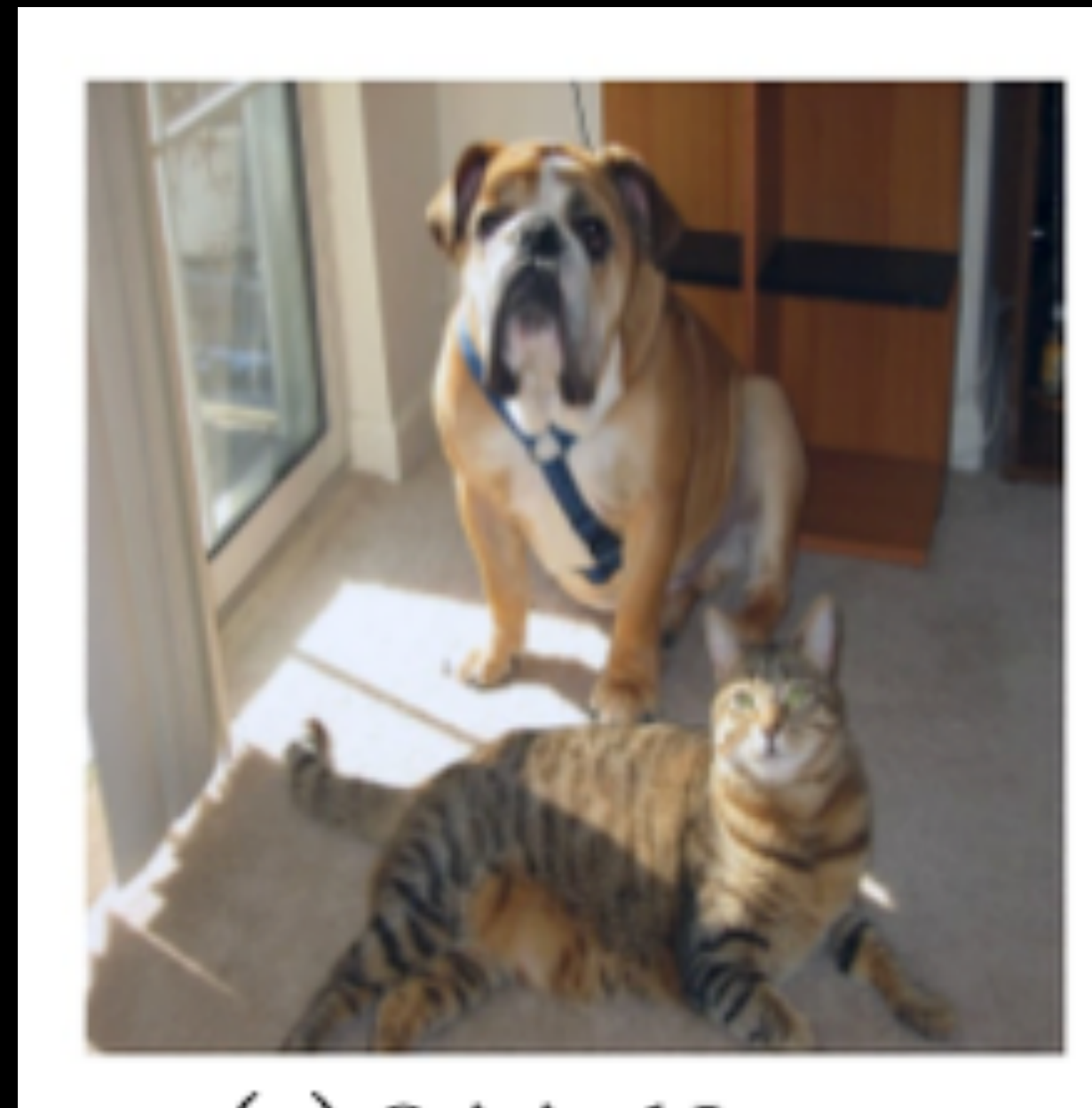
$$\alpha_k^c = \frac{1}{S} \sum_{i,j} \frac{\partial a^{c,L}}{\partial \mathbf{f}^{k,l}(i,j)} \longrightarrow \alpha_c^k = w_c^{k,L}$$

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





# Visualizing Grad-CAM outputs

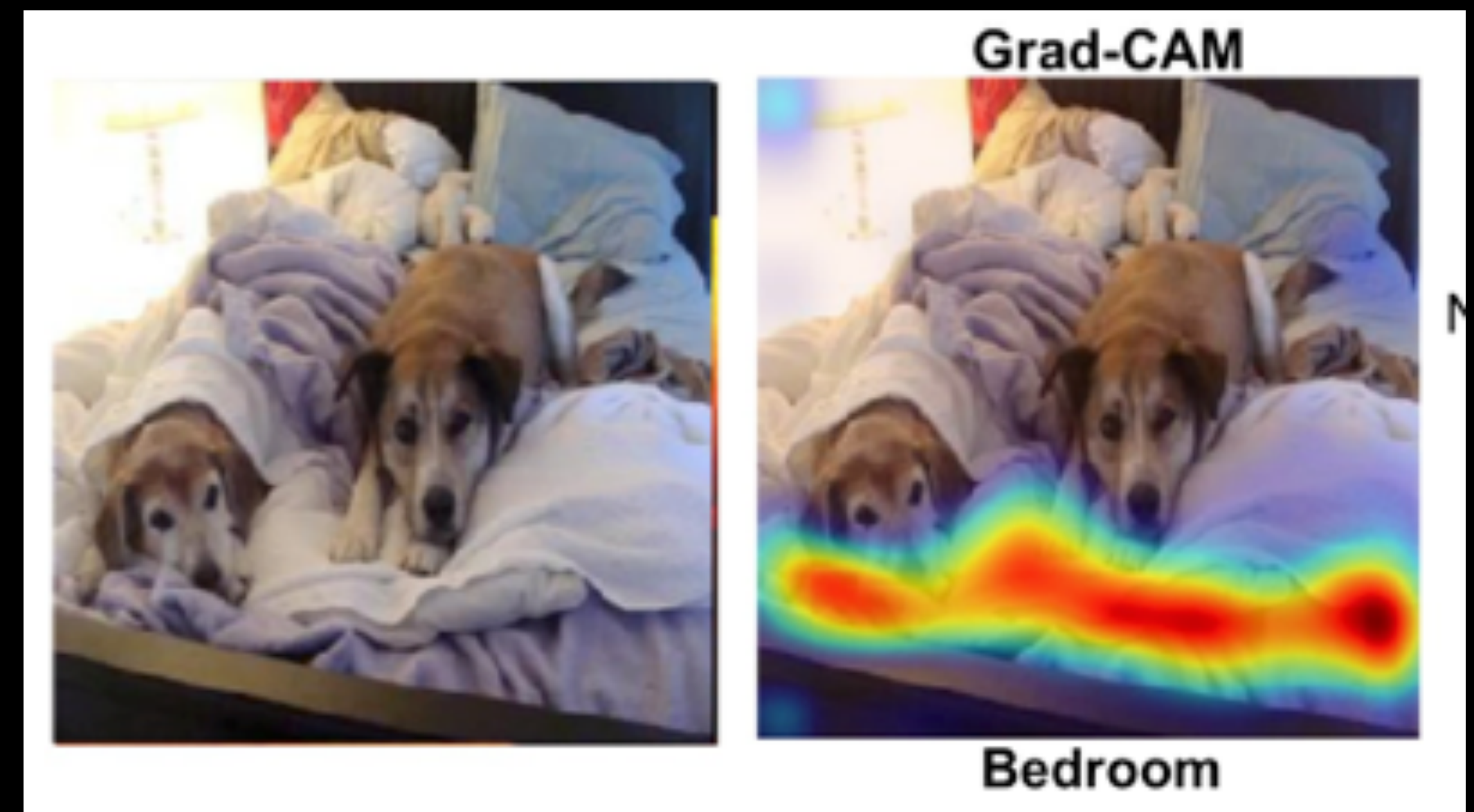
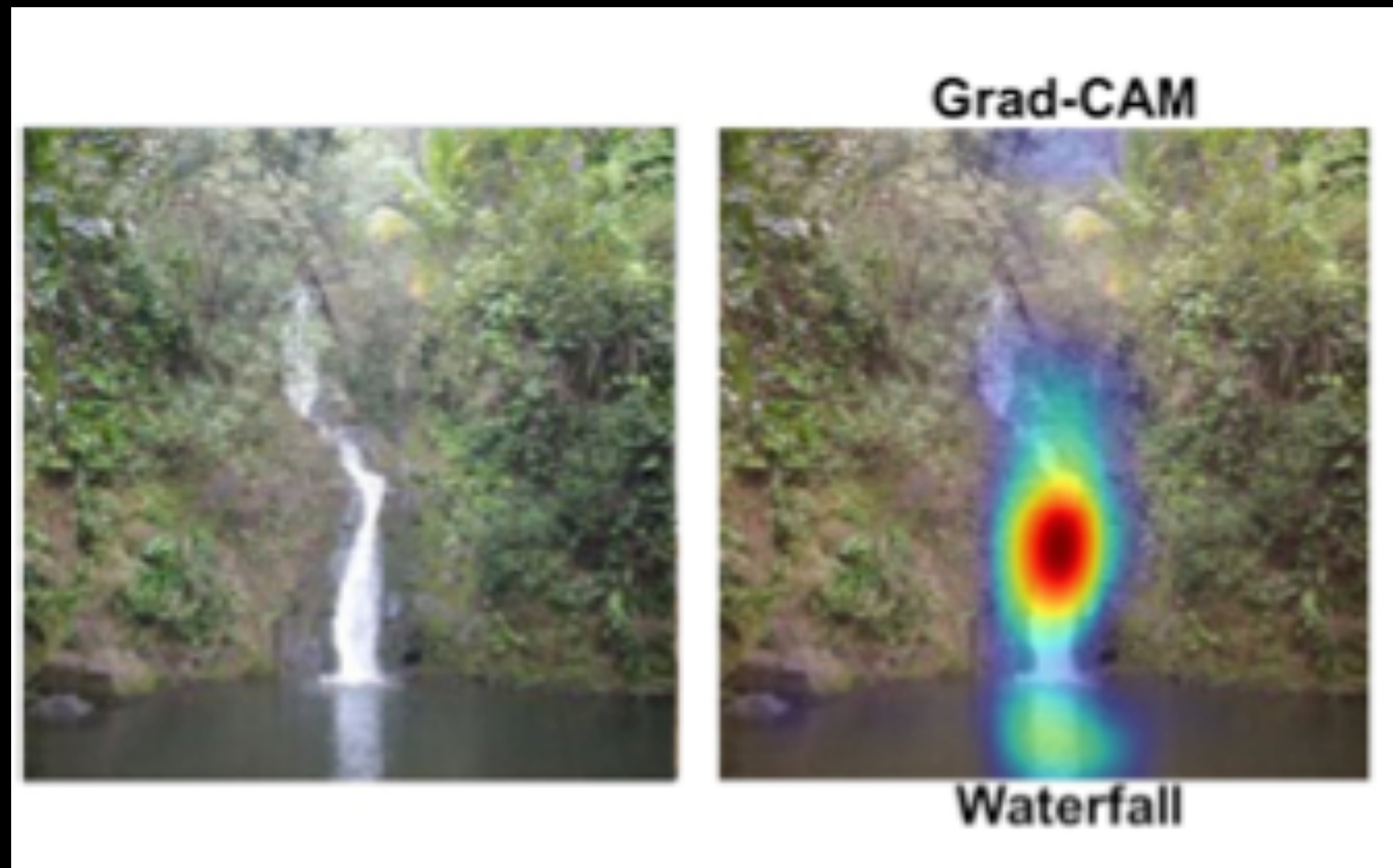


(c) Grad-CAM 'Cat'

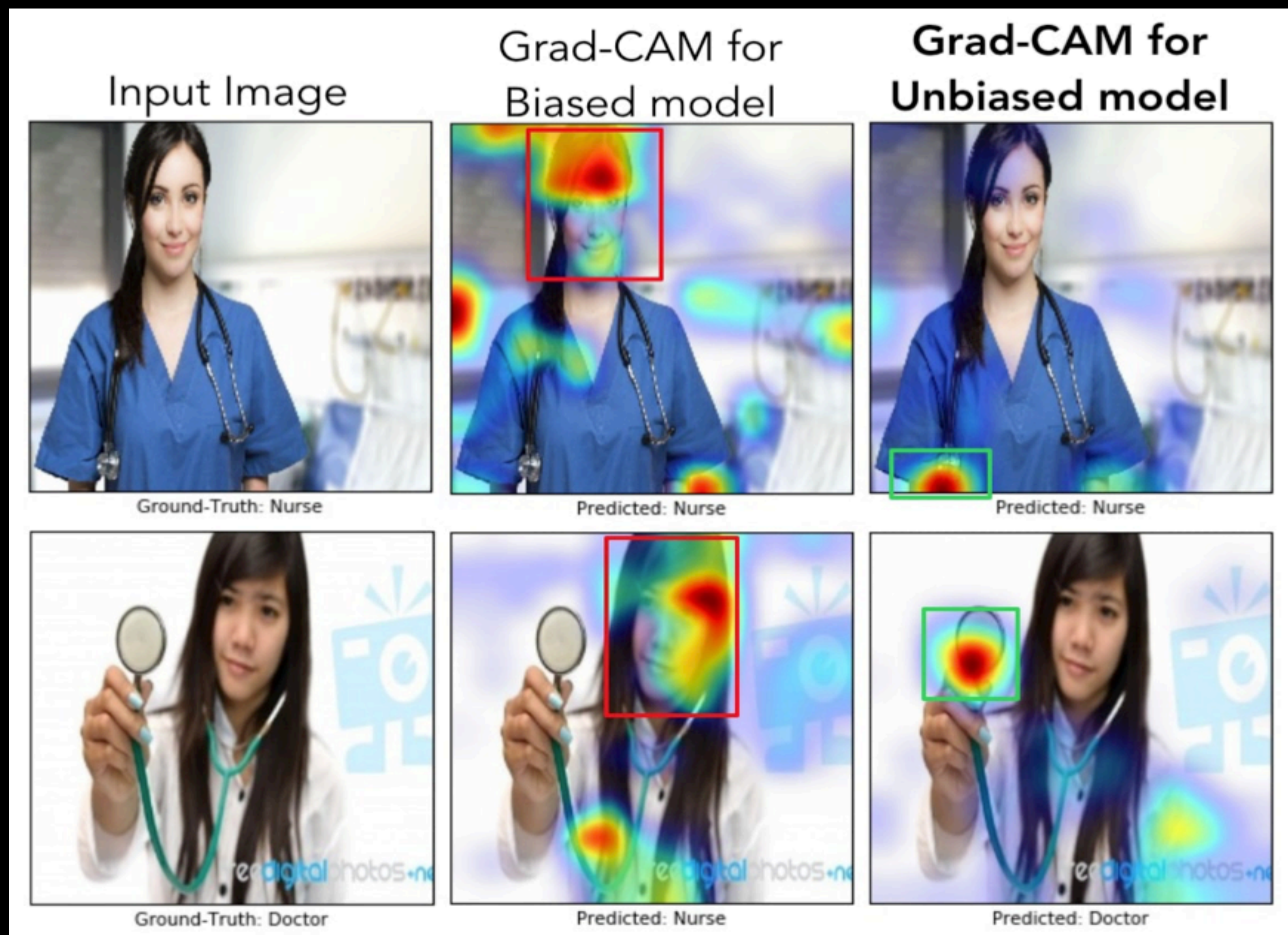


(i) Grad-CAM 'Dog'

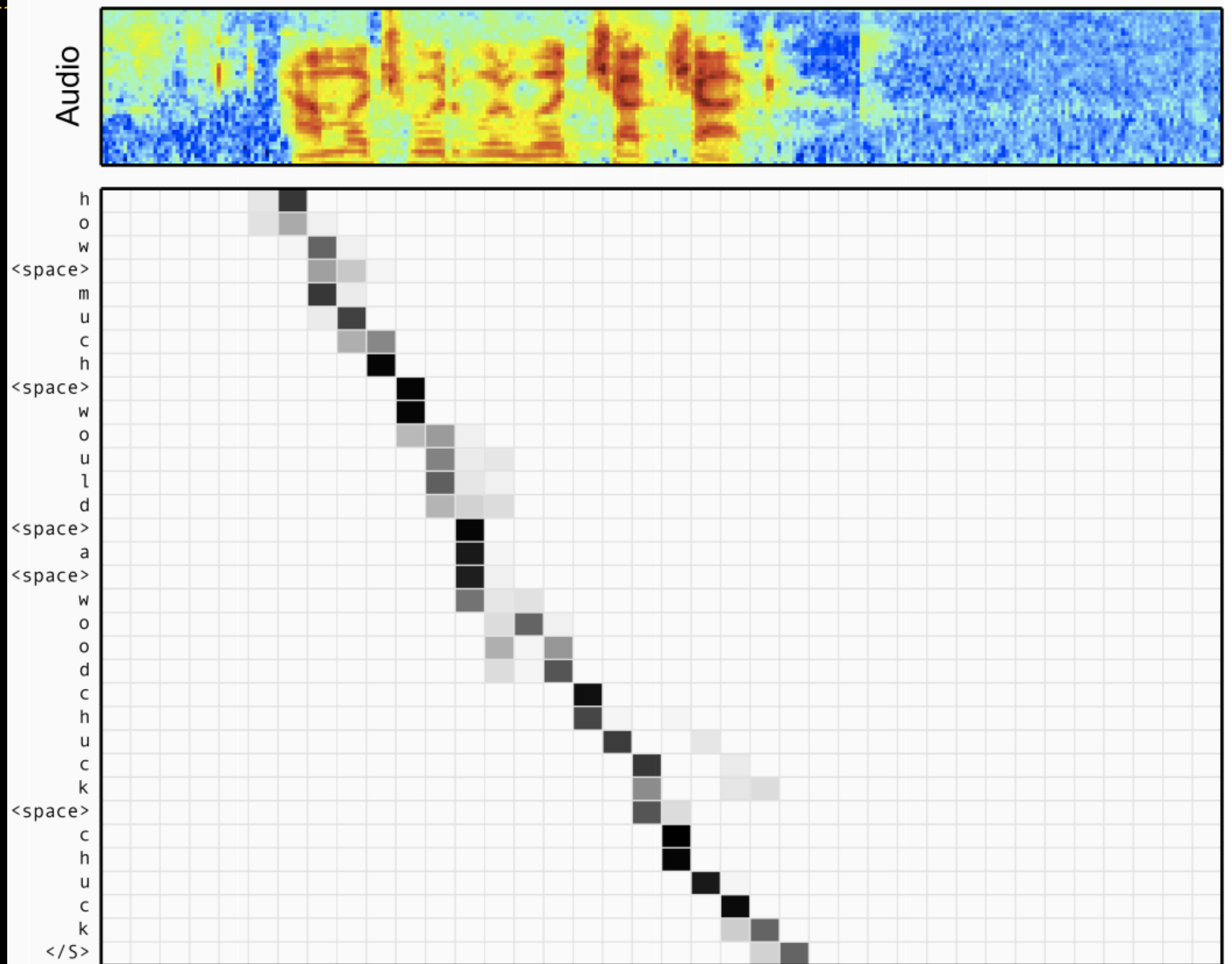
# Grad-CAM in image captioning



# Grad-CAM for identifying model biases



# Using attention for visualization



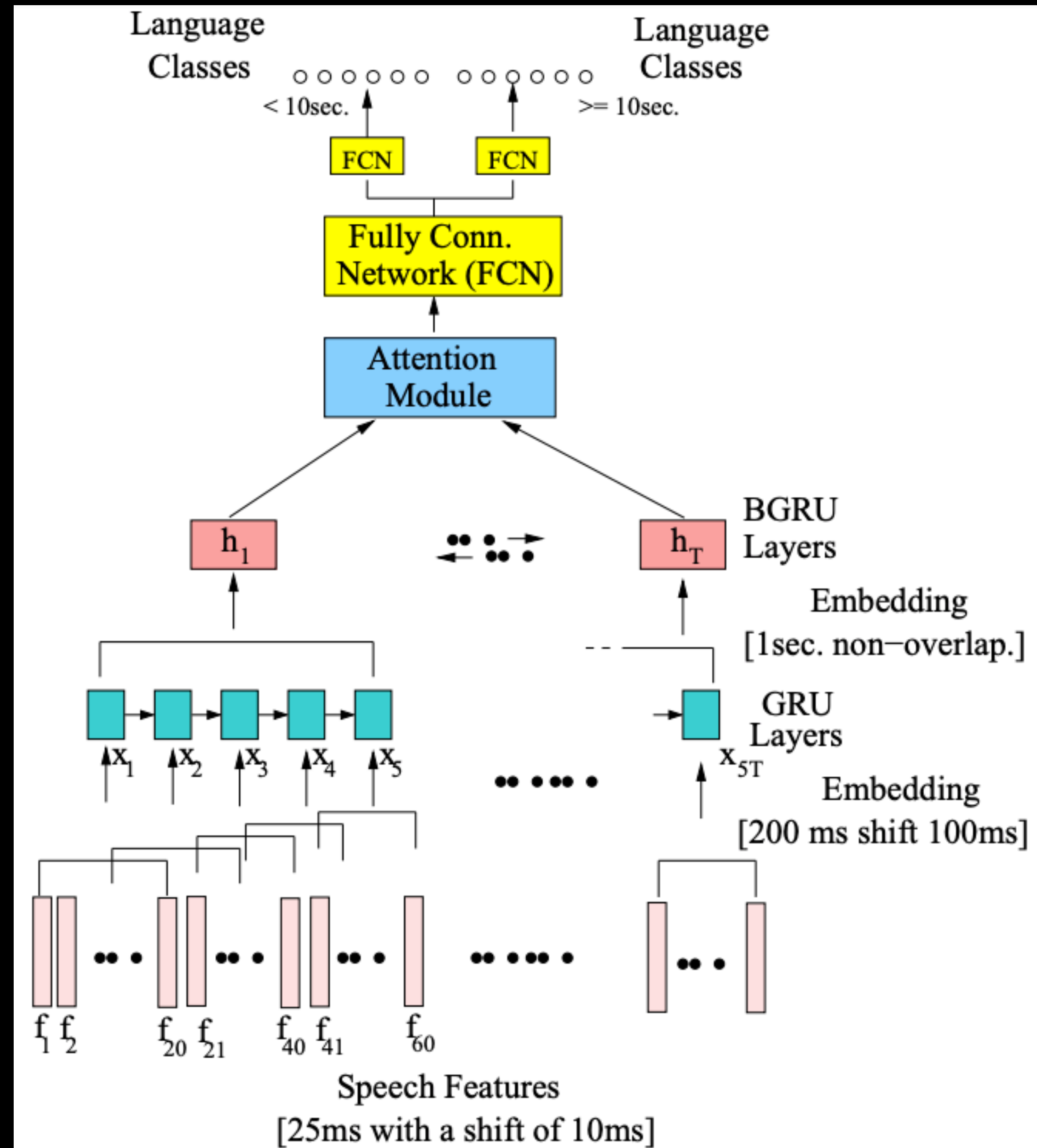
# Using attention mechanism for explainability

## Towards Relevance and Sequence Modeling in Language Recognition

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

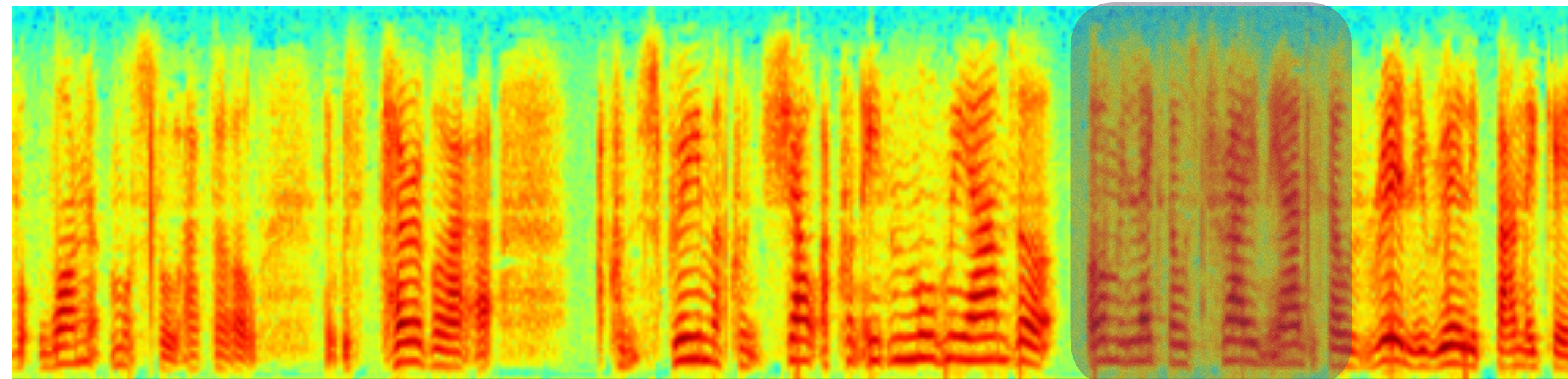


# Using attention mechanism for explainability



# Using attention mechanism for explainability

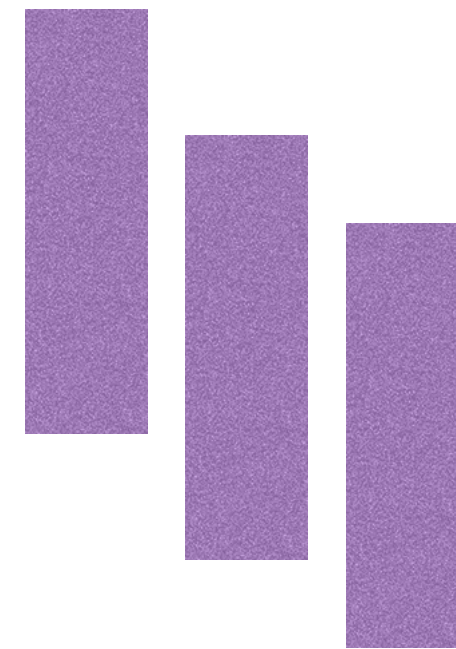
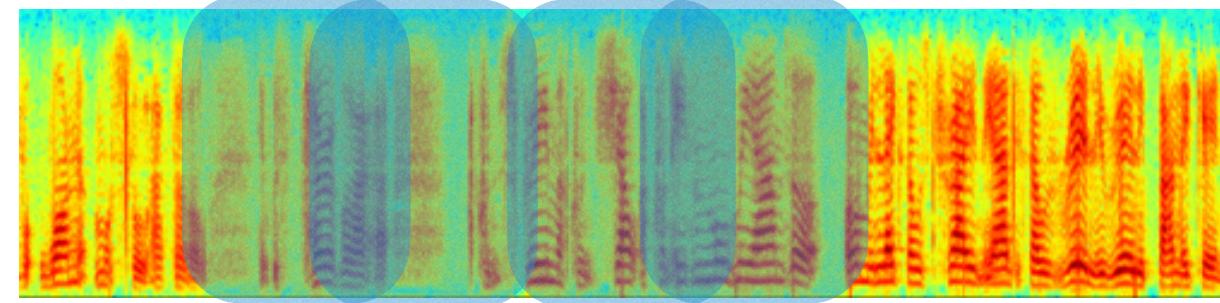
- Certain regions of the audio signal may have more information for the task than the rest.
  - ✓ May also have more signal quality than rest.
  - ✓ For example, language identification involving Eng-UK v/s Eng. US.



- 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

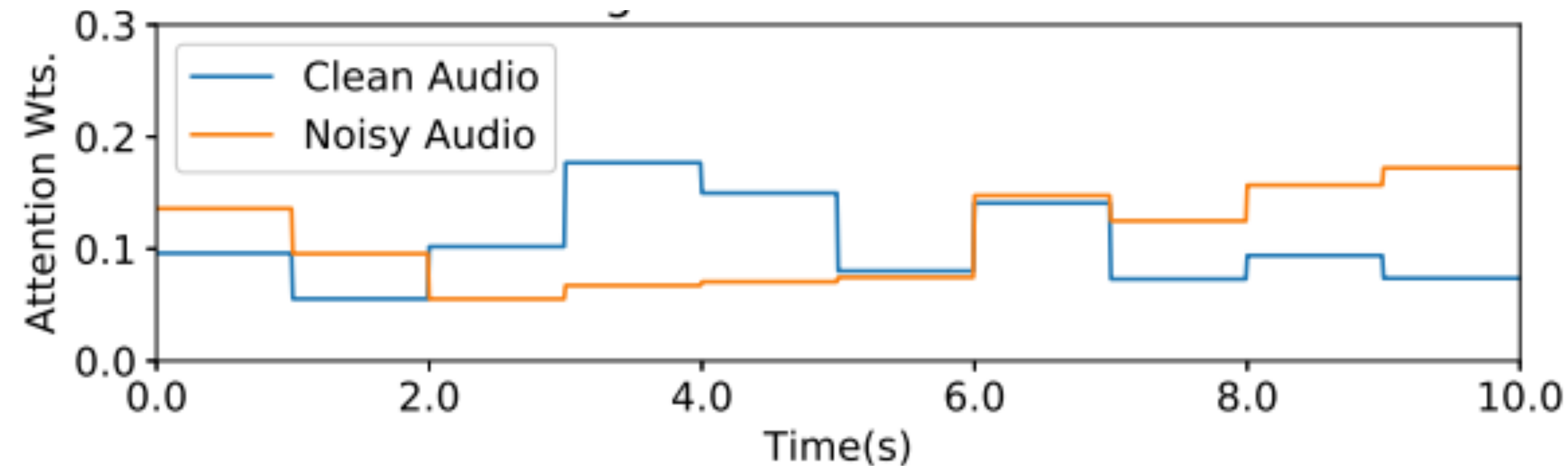
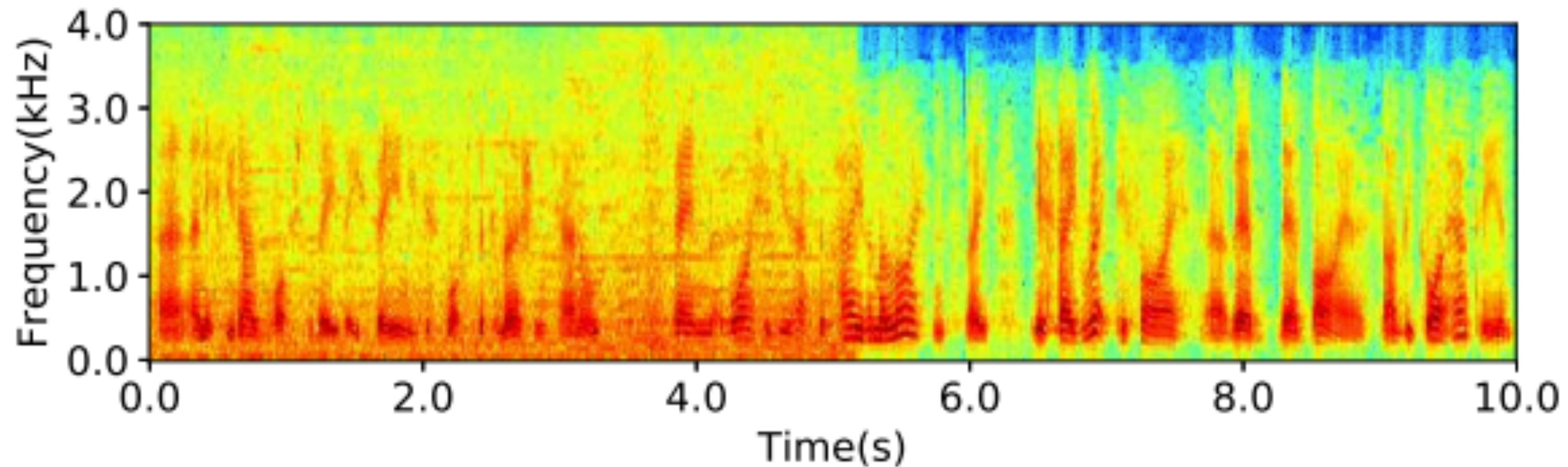
3s-4s

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

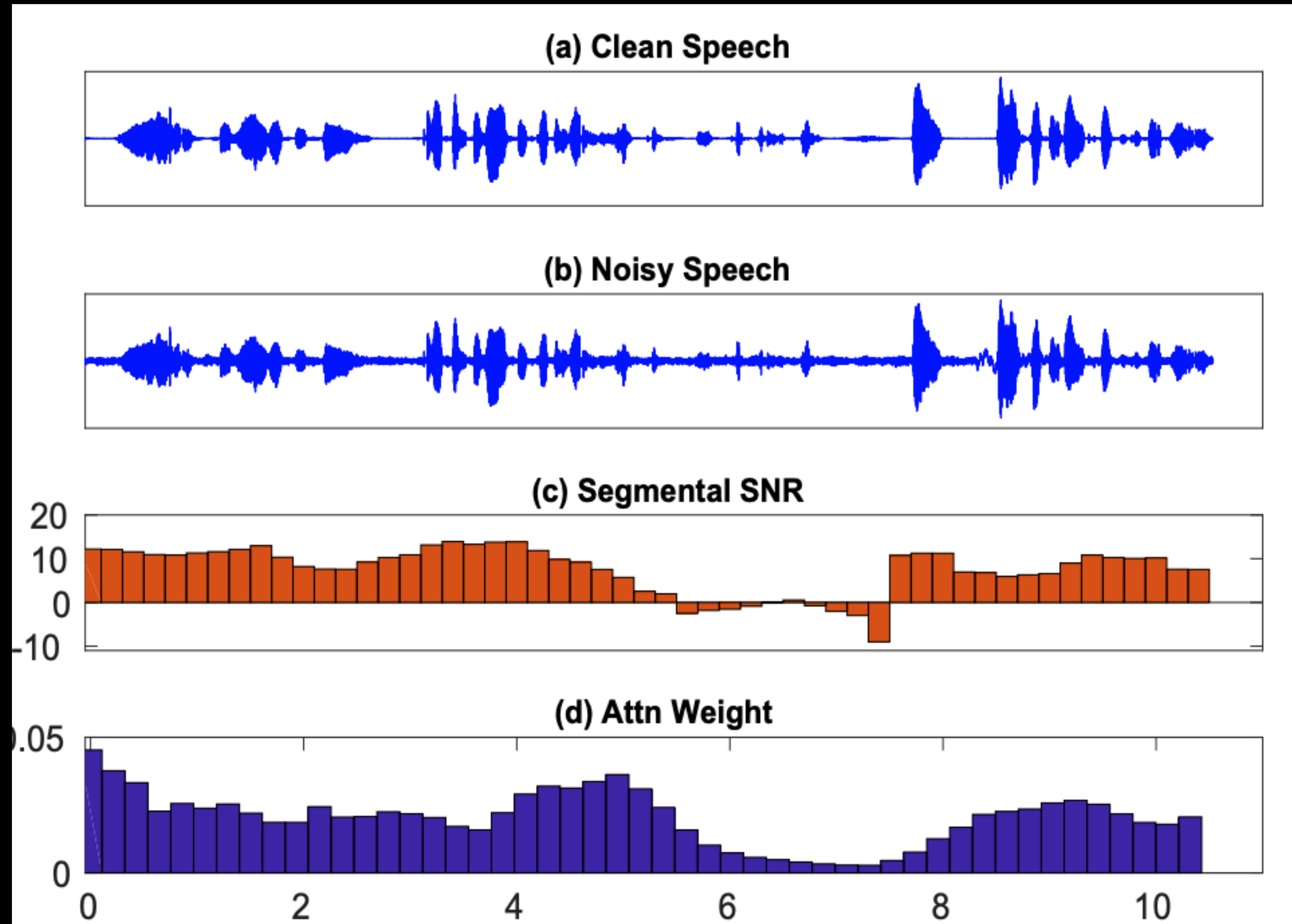
9s -11s



# Using attention mechanism for explainability



# Using attention mechanism for visualization



# Summary thus far

- \* Analyzing trained neural networks
  - ✓ Hierarchical representations
  - ✓ Activation maps to determine saliency
  - ✓ Incorporating attention mechanism for improved explainability

