Computer Vision – TP14
Explainable AI

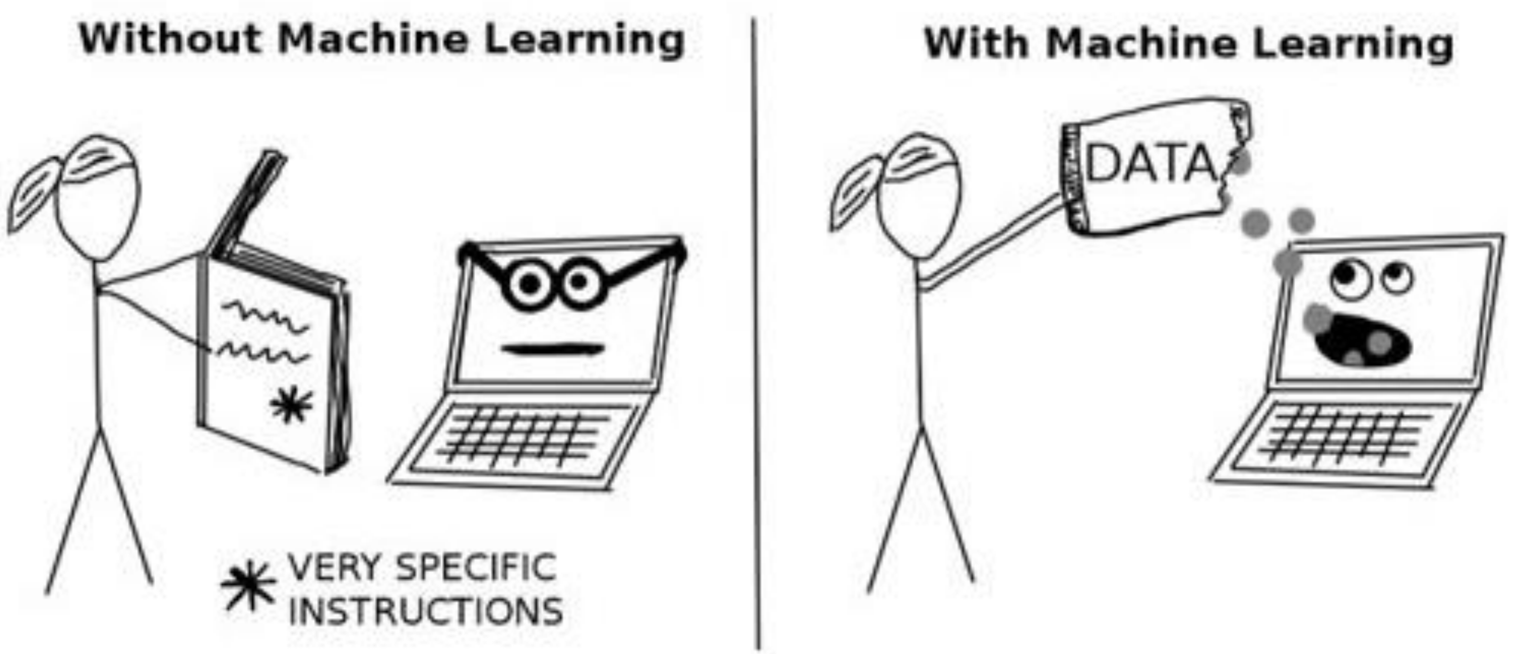
Miguel Coimbra, Hélder Oliveira

U.PORTO FC

Outline

- Explainable AI (XAI)
- Saliency Maps
- Class Activation Mapping
- Other examples

“Good” old times of AI



Christoph Molnar “Interpretable Machine Learning A Guide for Making Black Box Models Explainable”

“Good” old times of AI

- In the beginning, **artificial intelligence systems were based in algorithms**:
  - An algorithm is a **set of instructions** that the system will follow to **achieve a certain goal** (direct programming)
  - These **explicit** rules were often based on **domain knowledge**
  - Hence, they were “easy” to **explain** and to **understand**

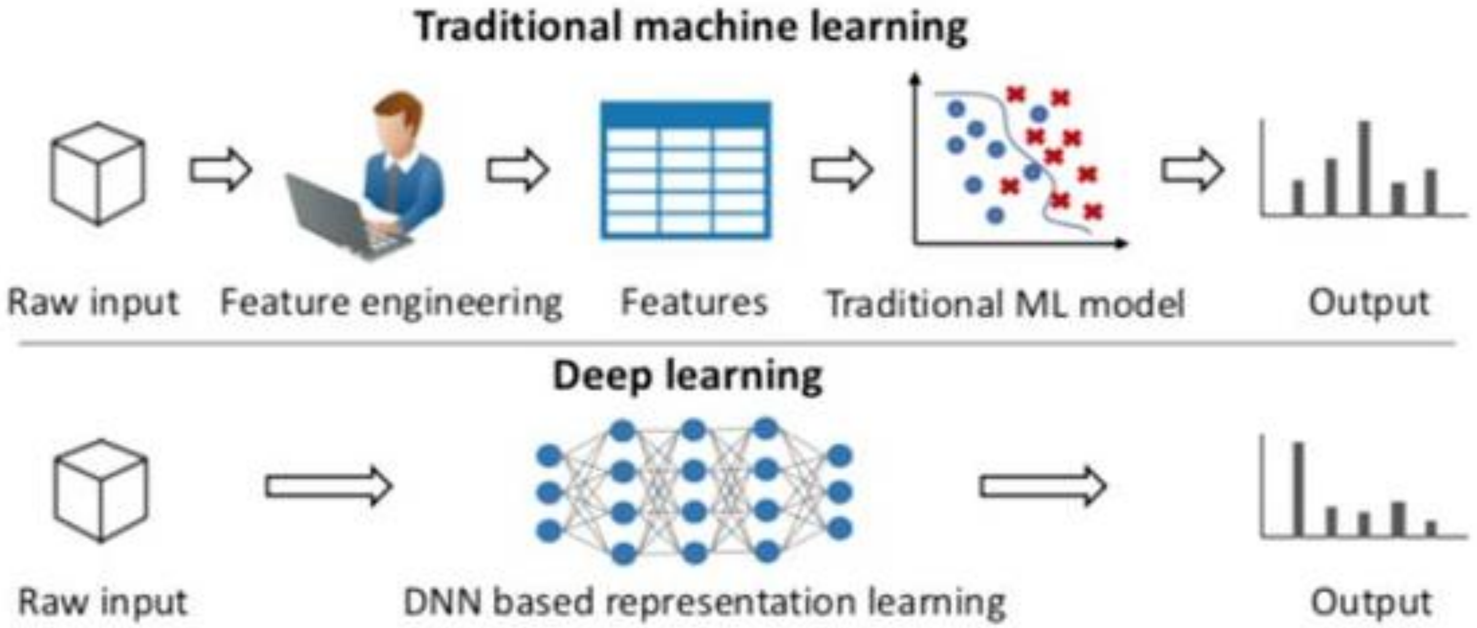
Christoph Molnar “Interpretable Machine Learning A Guide for Making Black Box Models Explainable”

“Good” old times of AI

- Nowadays, we use the available data to automatically learn **programs/functions**:
  - In machine learning, we **learn from data and make predictions** (indirect programming)
  - These algorithms work by **optimising an objective function**
  - Hence, the “rules” often are **implicit** and **difficult to understand**

Christoph Molnar “Interpretable Machine Learning A Guide for Making Black Box Models Explainable”

Deep learning versus traditional machine learning



[Lecun et al. "Deep learning"](), https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Deep learning versus traditional machine learning

- **Traditional machine learning**
  - required **experts to extract meaningful features** (*i.e.*, domain-specific features) from raw data and feed them into machine learning algorithms to obtain classification/regression models

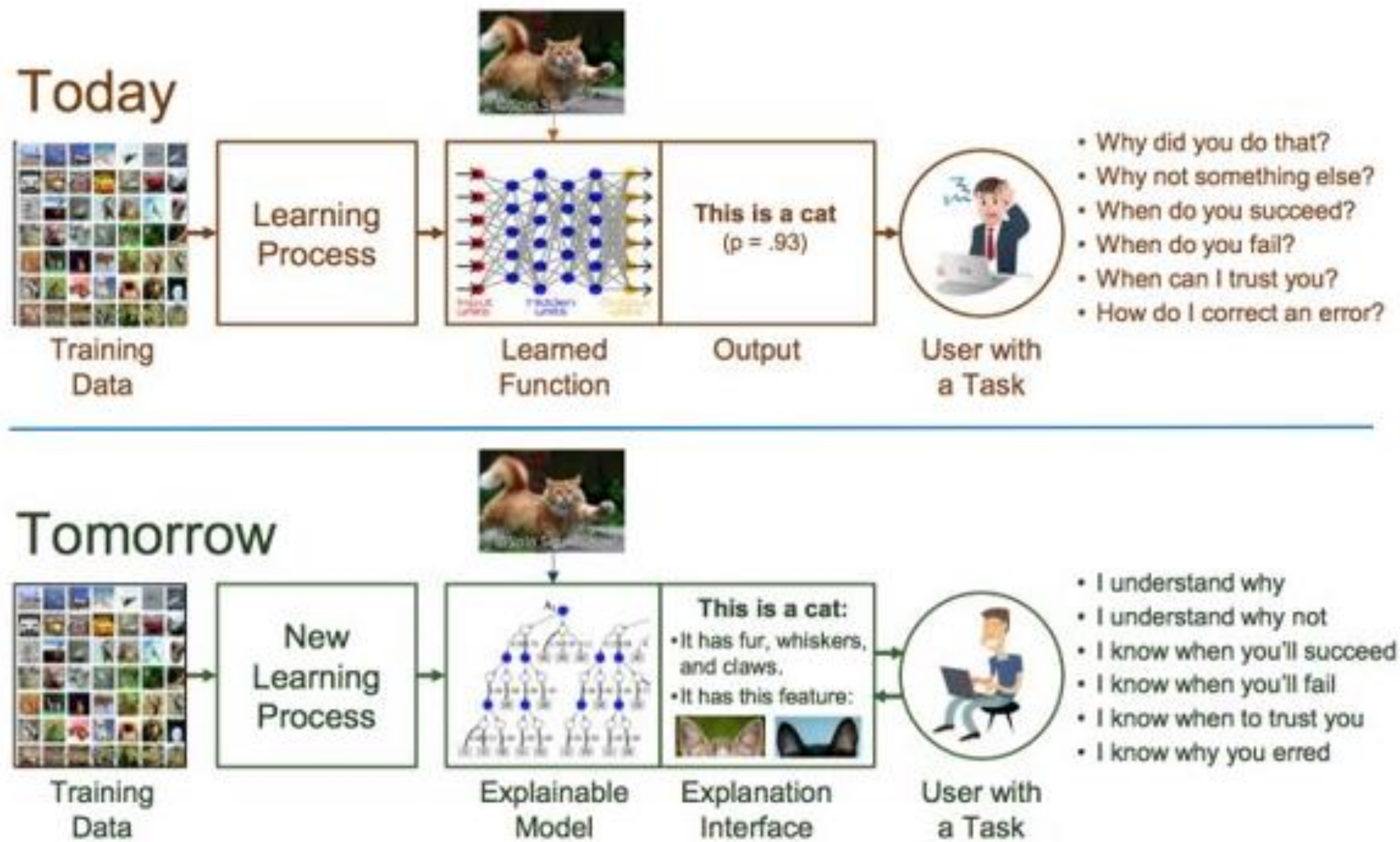
Lecun et al. "Deep learning", https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Deep learning versus traditional machine learning

- **Deep learning**
  - "only" requires **raw data and labels** to achieve high-performing models, since it **automatically extracts the patterns**
  - Deep learning algorithms are suitable for **representation learning**, i.e., finding the **best representation of the data** that optimises a given optimisation objective

Lecun et al. "Deep learning", https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Do we understand the features learned by these models?



https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Do we understand the features learned by these models?

- Even if the models achieve high performances, **it is not trivial to assure that they are learning features that are relevant for that domain** (i.e., **black box** behaviour)
  - Machine learning models **are good at extracting correlations**

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

Do we understand the features learned by these models?

- While this **may not be an issue in several domains** (e.g., recommendation systems), in others, it is of utmost importance that the **system is capable of transparently showing** the reasons behind its decisions (e.g., healthcare)

https://www.cc.gatech.edu/~alanwags/DLAI2016/(Gunning)%20IJCAI-16%20DLAI%20WS.pdf

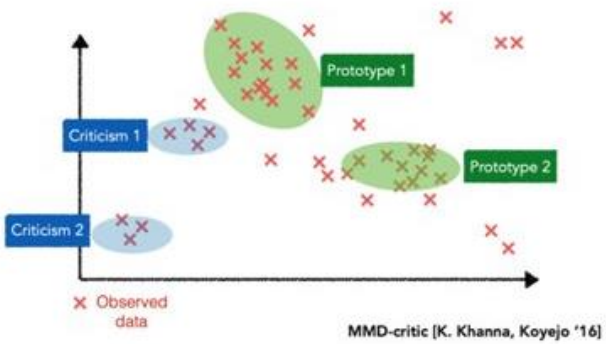
Types of Explainable AI (XAI)

Pre-Model	In-Model	Post-Model
(aim to understand the data before building the model)	(Seek to integrate Interpretability inside The model)	(Perform posterior Analysis of the Model predictions)

[Lipton "The Mythos of Model Interpretability"](), [Doshi-Velez and Kim "Towards A Rigorous Science of Interpretable Machine Learning"]()

Pre-model

- **Rely on data *exploratory analysis***
  - One may think of “K-Means Clustering”, “K-Nearest Neighbours”



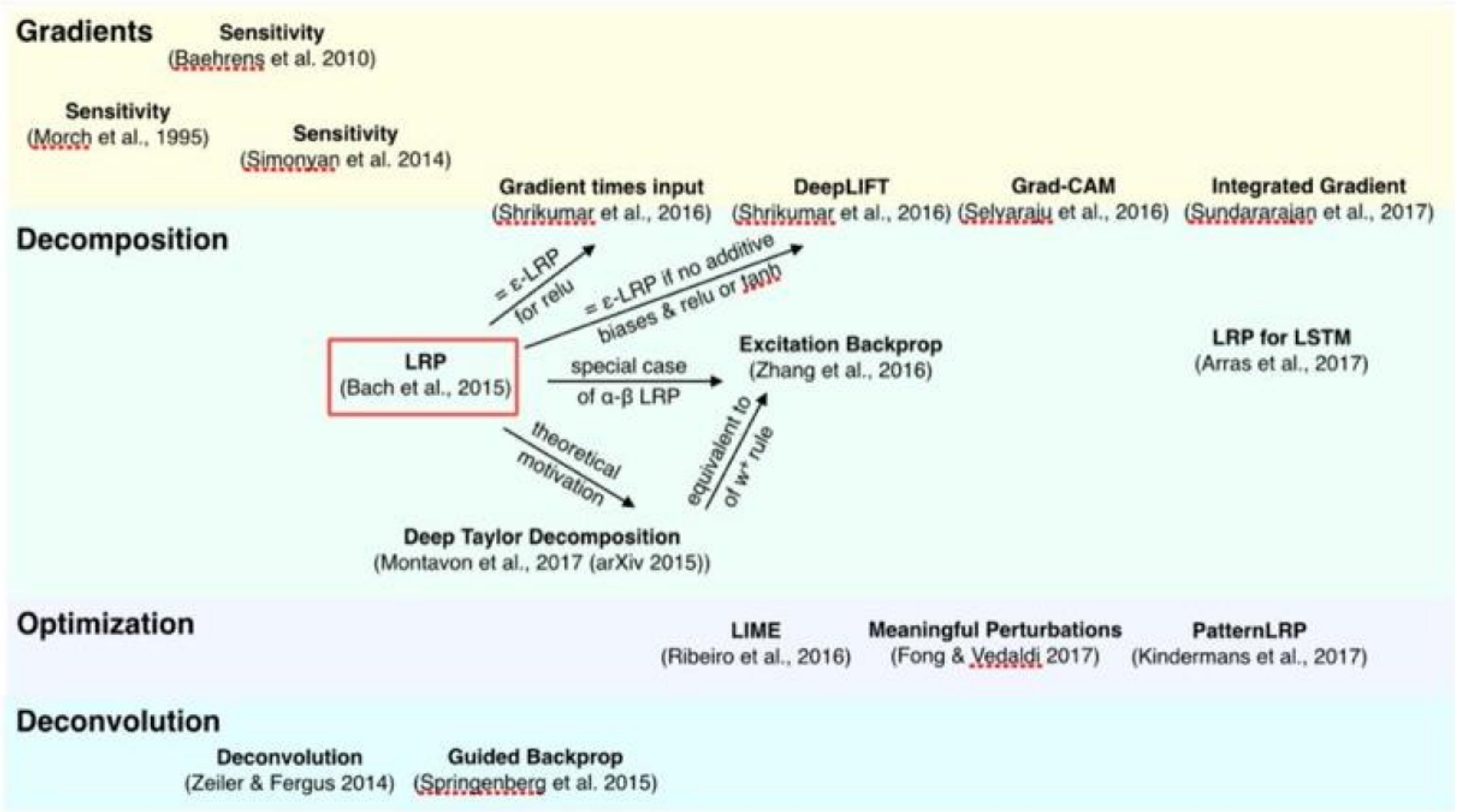
Tukey “Exploratory data analysis”, Kim et al.“Examples are not Enough, Learn to Criticize! Criticism for Interpretability”

Post-model

- In computer vision, one may think of methods based on **“Gradients”, “Decomposition”, “Optimisation” and “Deconvolution”**

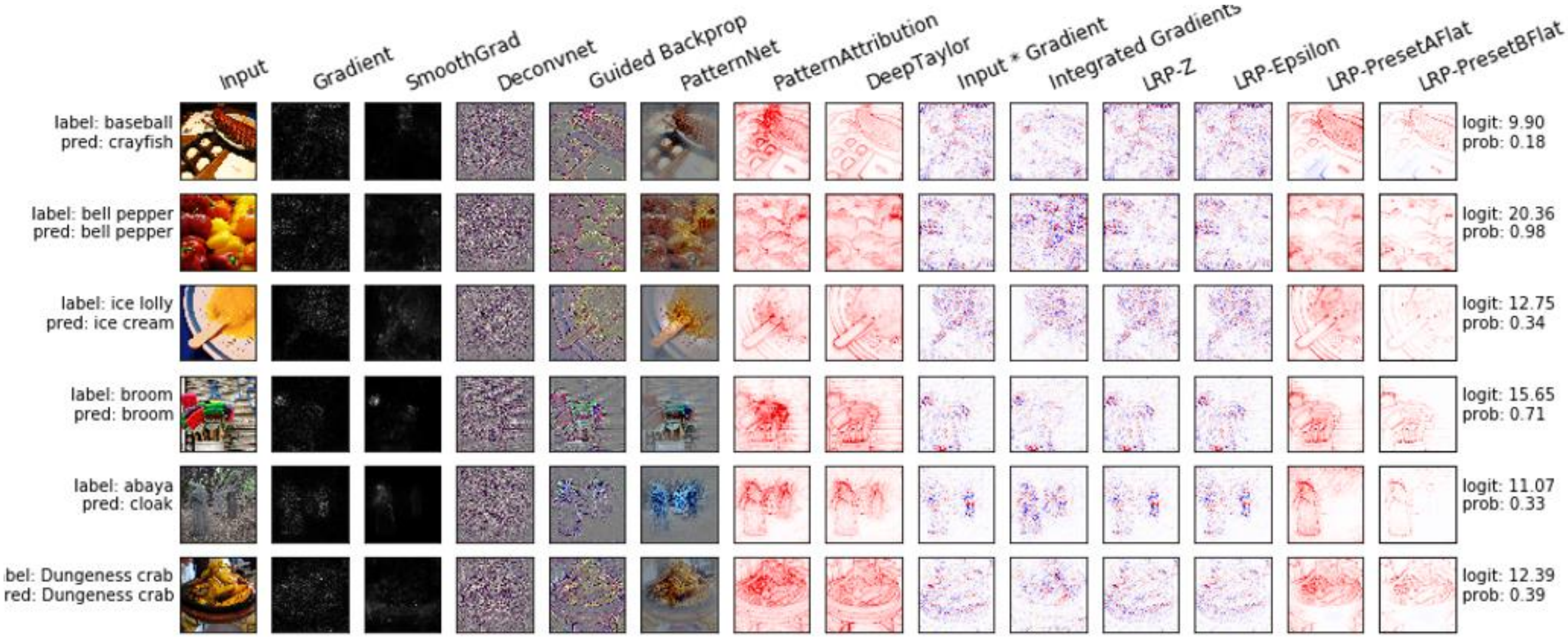
Samek “Interpreting Deep Neural Networks and their Predictions”, Alber et al. “iNNvestigate Neural Networks!”

Post-model



Samek "Interpreting Deep Neural Networks and their Predictions", Alber et al. "iNNvestigate Neural Networks!"

Post-model



[Alber et al. "iNNvestigate Neural Networks!"]

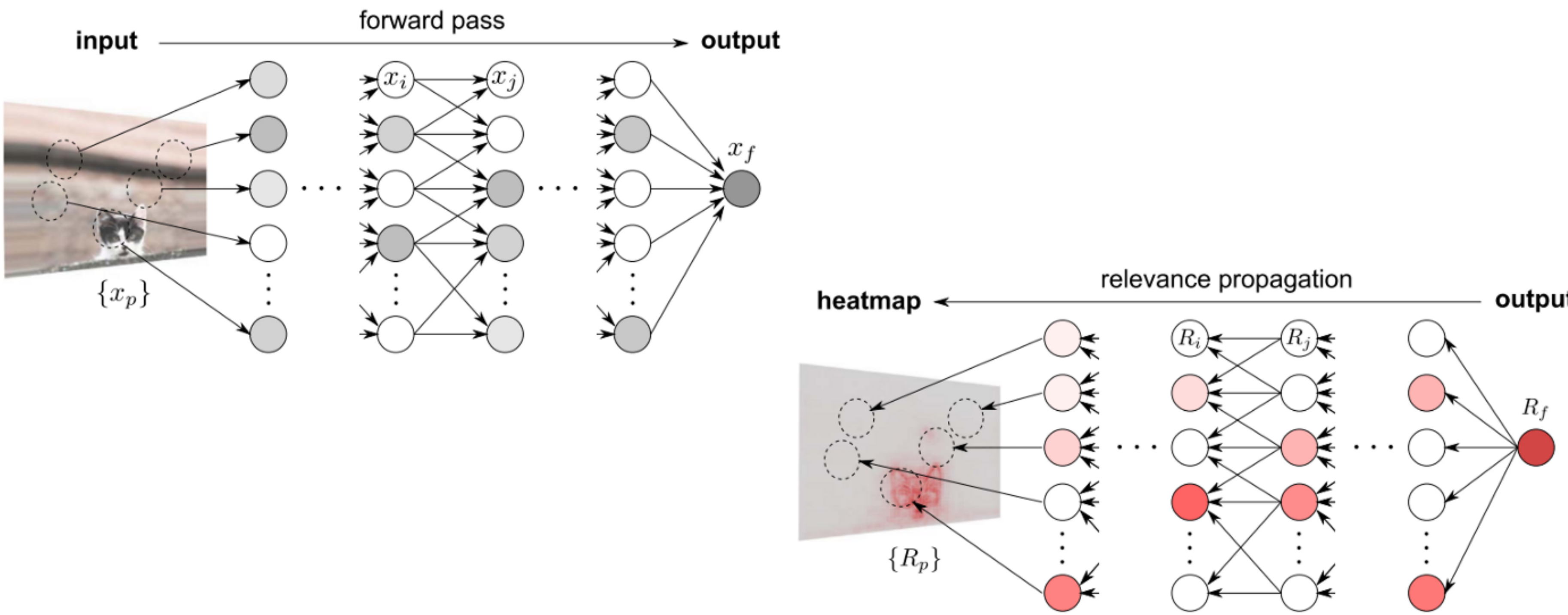
Post-model

- **Post-model explanations often do not make sense in a *human-understandable* manner**
  - One way or another, most of them produced some kind of ***saliency-maps***

Cynthia Rudin "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead"

Post-model

- **Reverse propagation**



DTD algorithm, from Bach et al.

Post-model

- **Saliency maps**
  - are a visualization technique to gain better insights into the decision-making of a neural network.
  - They also help in knowing ***what each layer of a convolutional layer focuses on***.
  - This helps us understand the decision making process a bit more clearly.

https://debuggercafe.com/saliency-maps-in-convolutional-neural-networks/

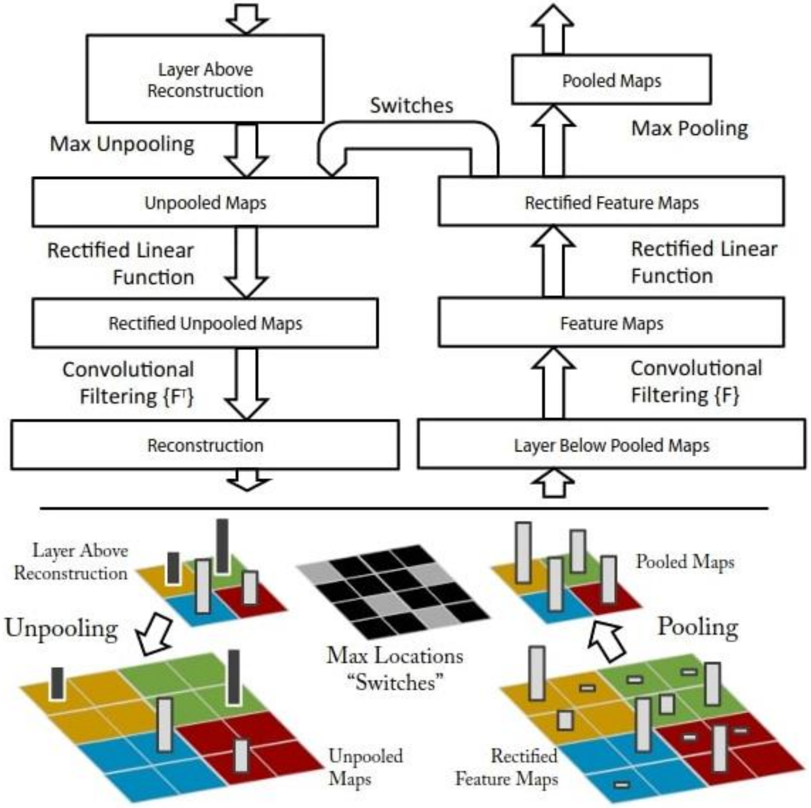
Post-model

- **Saliency maps**
  - help us visualize where the convolutional neural network is focusing in particular while making a prediction.
  - Generally, we visualize saliency maps as heatmaps overlayed on the original image. We can also visualize them as colored pixels concentrated around the area of interest of an object.

https://debuggercafe.com/saliency-maps-in-convolutional-neural-networks/

Saliency maps

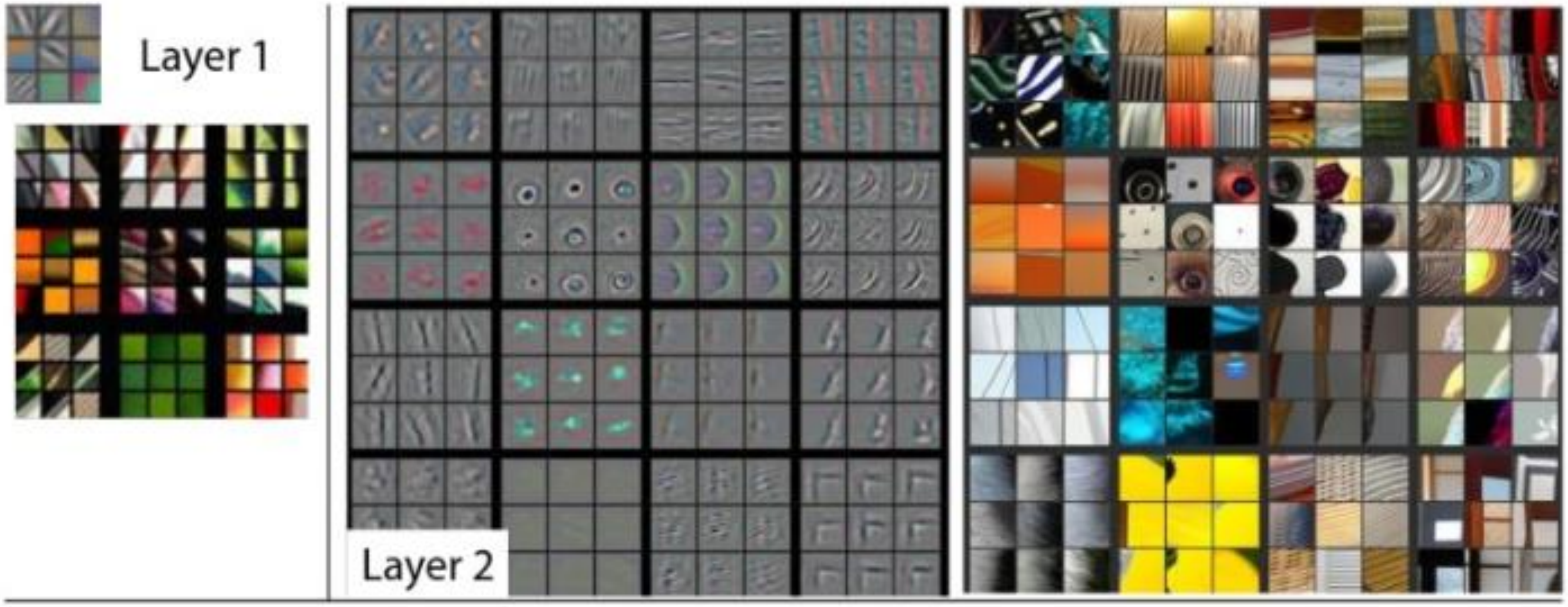
- Deconvolutional Network Approach



https://arxiv.org/pdf/1311.2901.pdf

Saliency maps

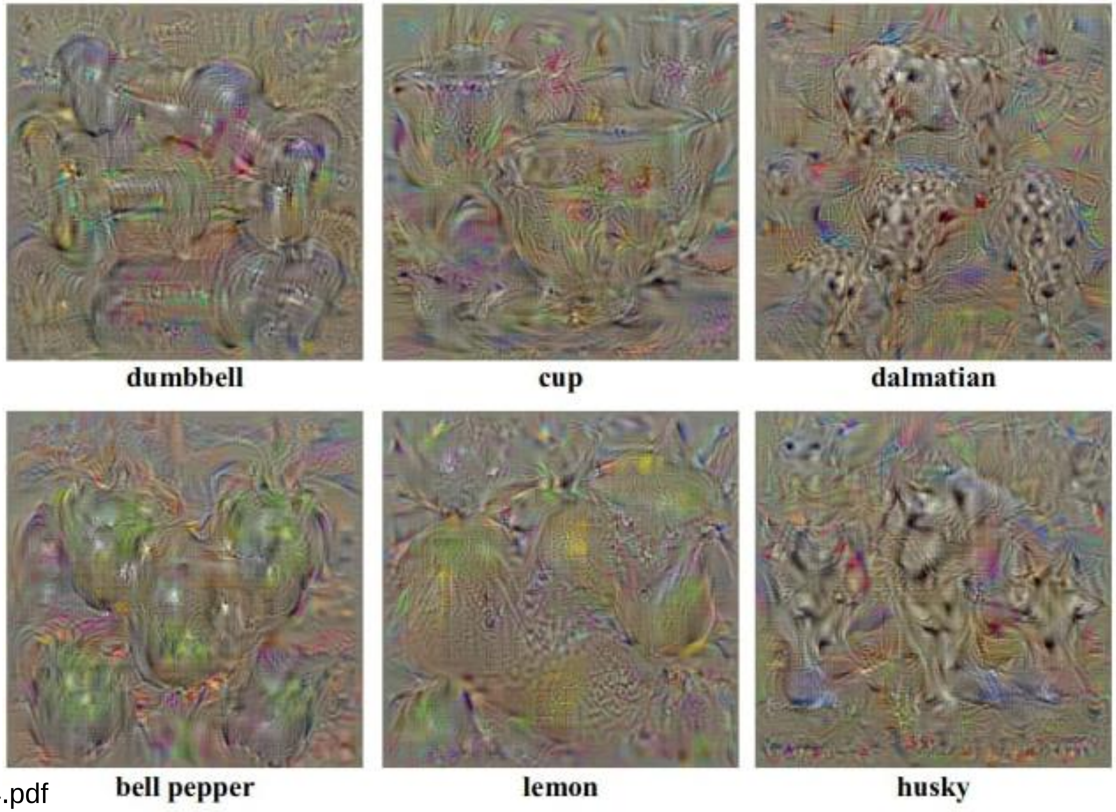
- Deconvolutional Network Approach



https://arxiv.org/pdf/1311.2901.pdf

Saliency maps

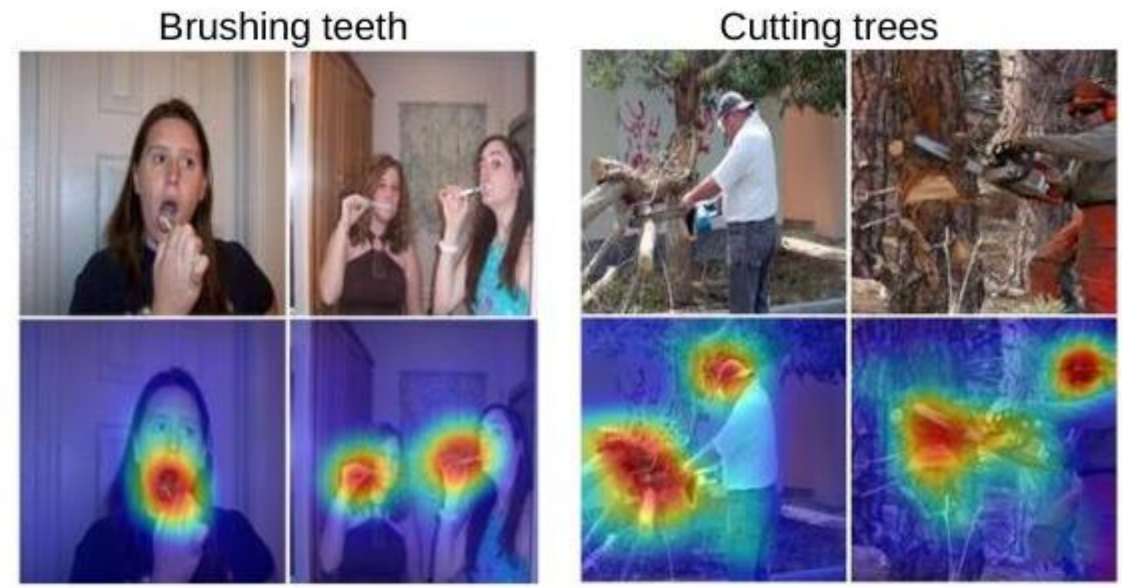
- Gradient Based Approach for Saliency Maps



https://arxiv.org/pdf/1312.6034.pdf

Saliency maps

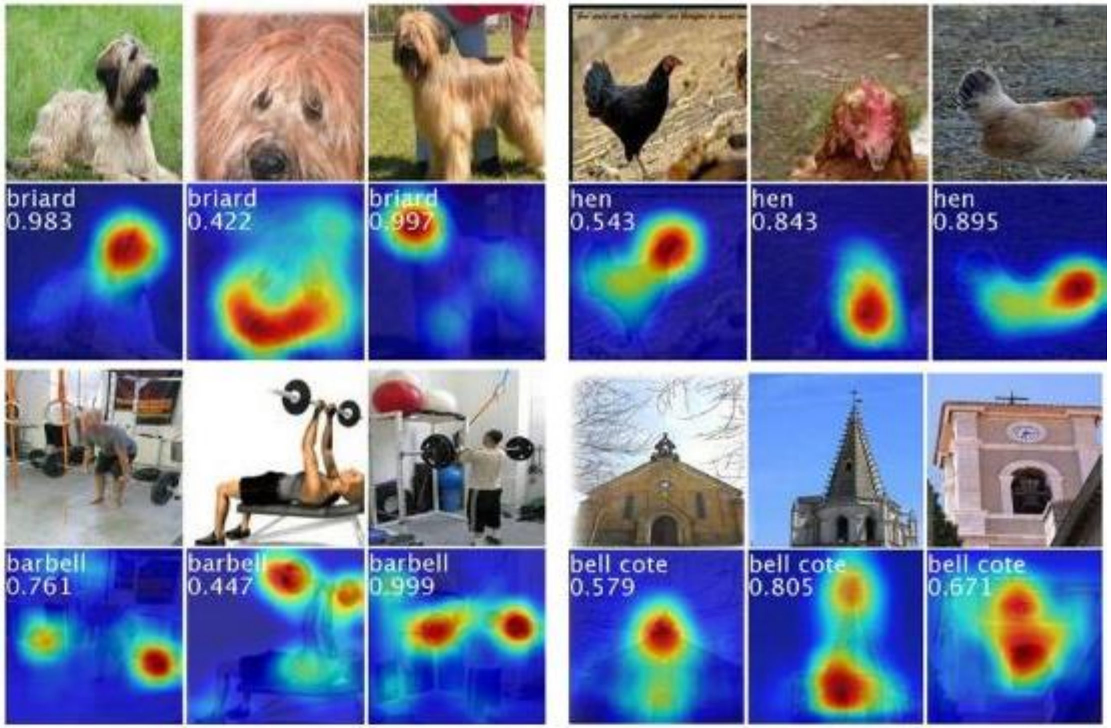
- Class Activation Mapping (CAM)



https://arxiv.org/pdf/1512.04150.pdf

Saliency maps

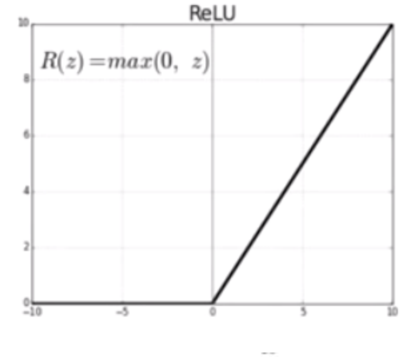
- Class Activation Mapping (CAM)

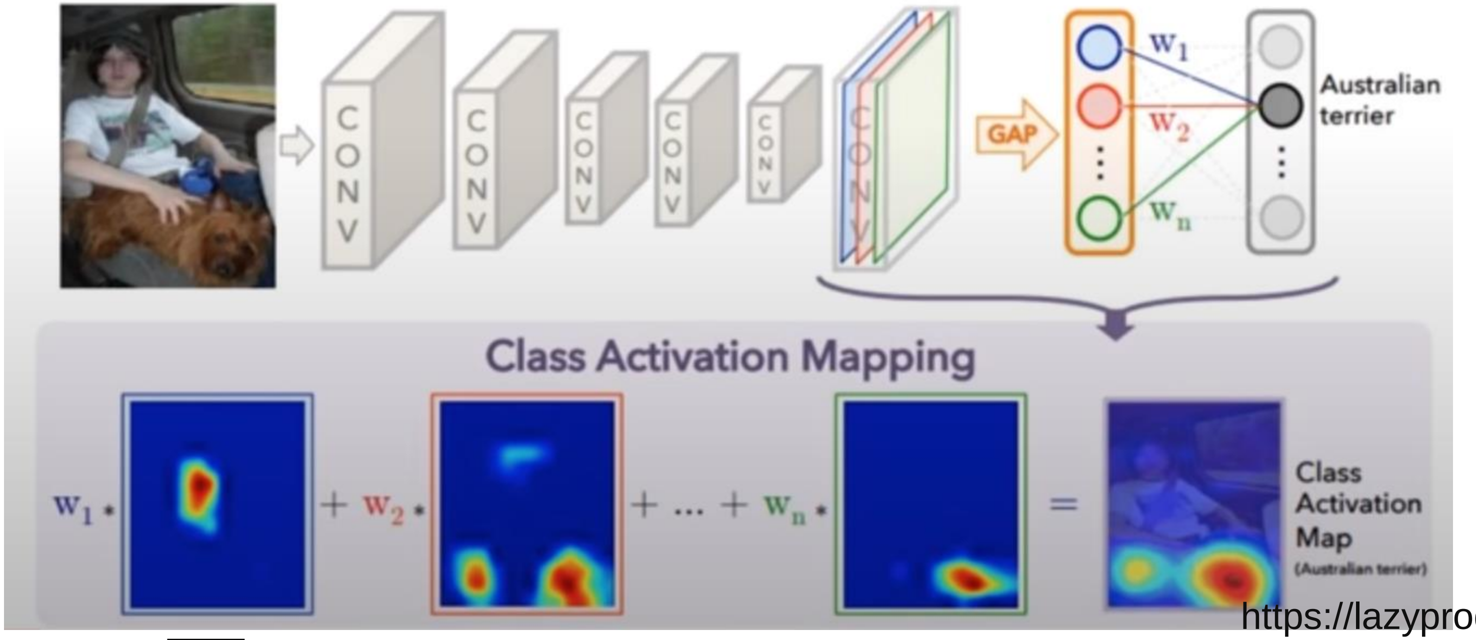


https://arxiv.org/pdf/1512.04150.pdf

Class Activation Mapping (CAM)

- Only need to do classification!
- Take any pre-trained CNN, e.g. ResNet
- Image shrinks, but # features increase
- ReLU: all features are +ve or zero

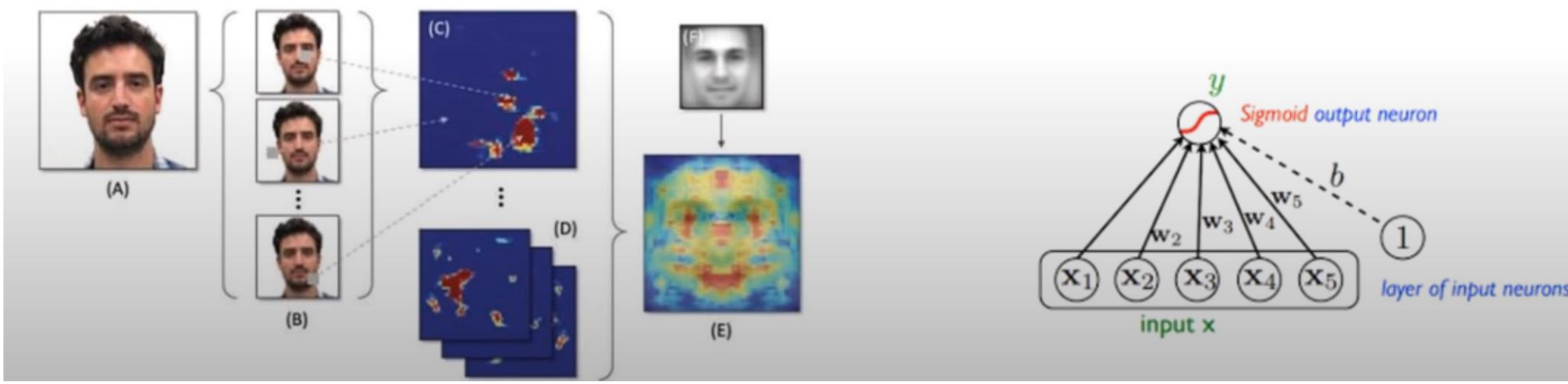




https://lazyprogrammer.me

Class Activation Mapping (CAM)

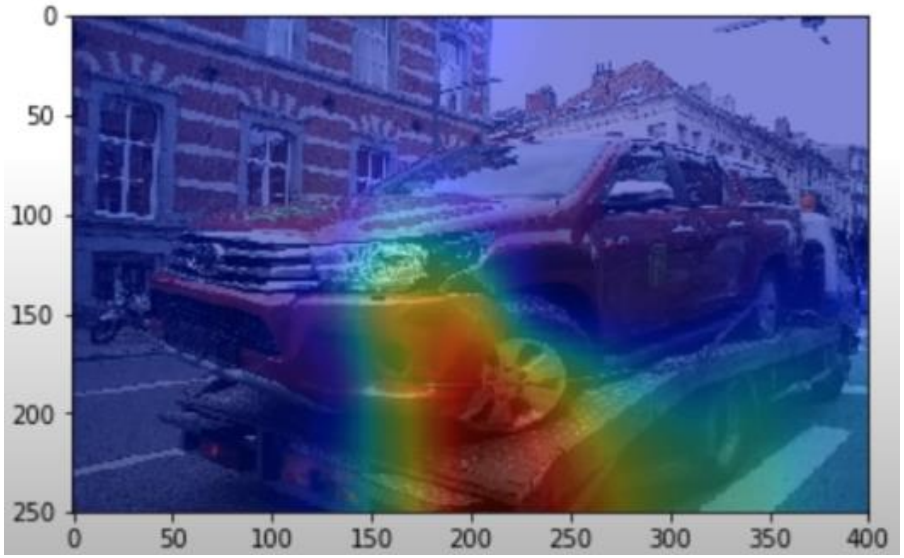
- Intuitively, you can think of a feature going into the Logistic Regression as a number denoting whether or not some “thing” appears in the image
  - E.g. one feature for nose, one for eyes, one for lips, hair, ears, etc.
  - +ve number if “thing” was found, 0 otherwise
    - E.g. the feature for “wheel” would be 0



https://lazyprogrammer.me

Class Activation Mapping (CAM)

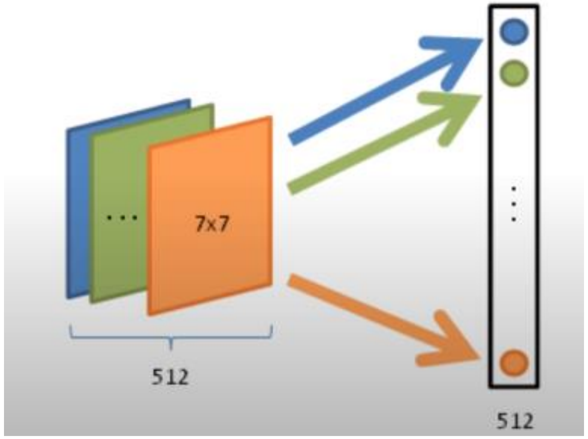
- In the picture if of a car, then the feature for “wheel” would be > 0, if a wheel was found
- Now the “nose” feature would be 0



https://lazyprogrammer.me

Class Activation Mapping (CAM)

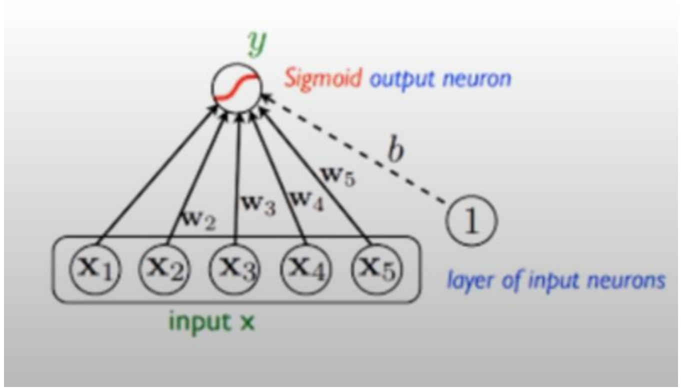
- If a feature is positive, that means the pooling operation must have found some +ve numbers in the final image (after going through several layers of convolutions)
- i.e. That feature must have been found “somewhere”
- If we simply looked at the image before pooling, the we would know where!



https://lazyprogrammer.me

Class Activation Mapping (CAM)

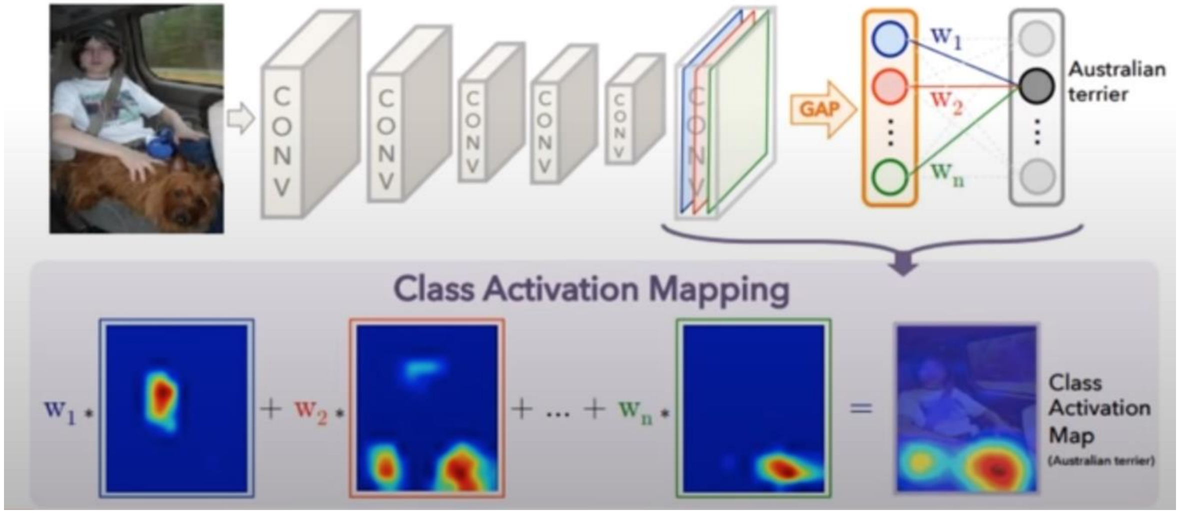
- If a weight is >0, then the corresponding feature is positively correlated with this class
- If it is 0, it has no effect
- If it is <0, the feature makes the image less likely to belong to this class



https://lazyprogrammer.me

Class Activation Mapping (CAM)

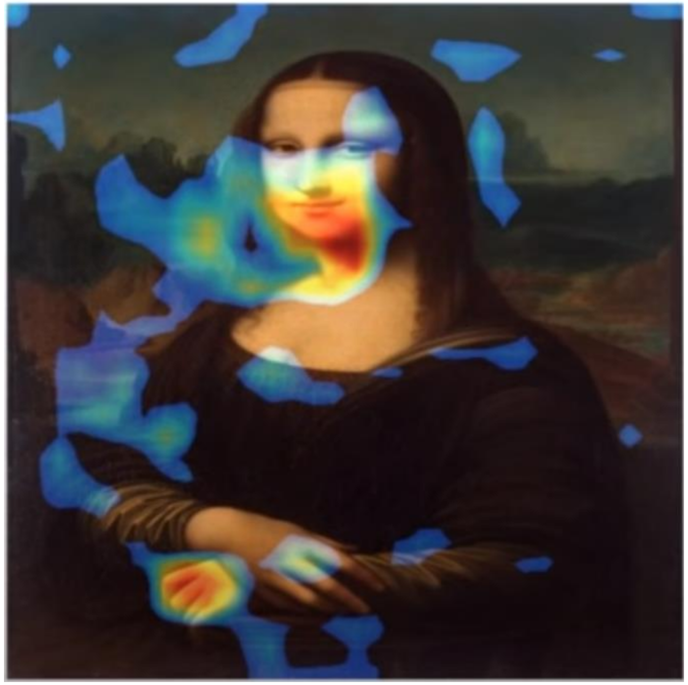
- We only consider 1 class at a time (usually the predicted class)
- E.g. w = W[:, human_face_índex] # size 2048
- F = 2048 7x7 images
- Class Activation Map = F[0]*w[0] + F[1]*w[1]+…+F[2047]*w[2047]
- Result is a 7x7 heat map



https://lazyprogrammer.me

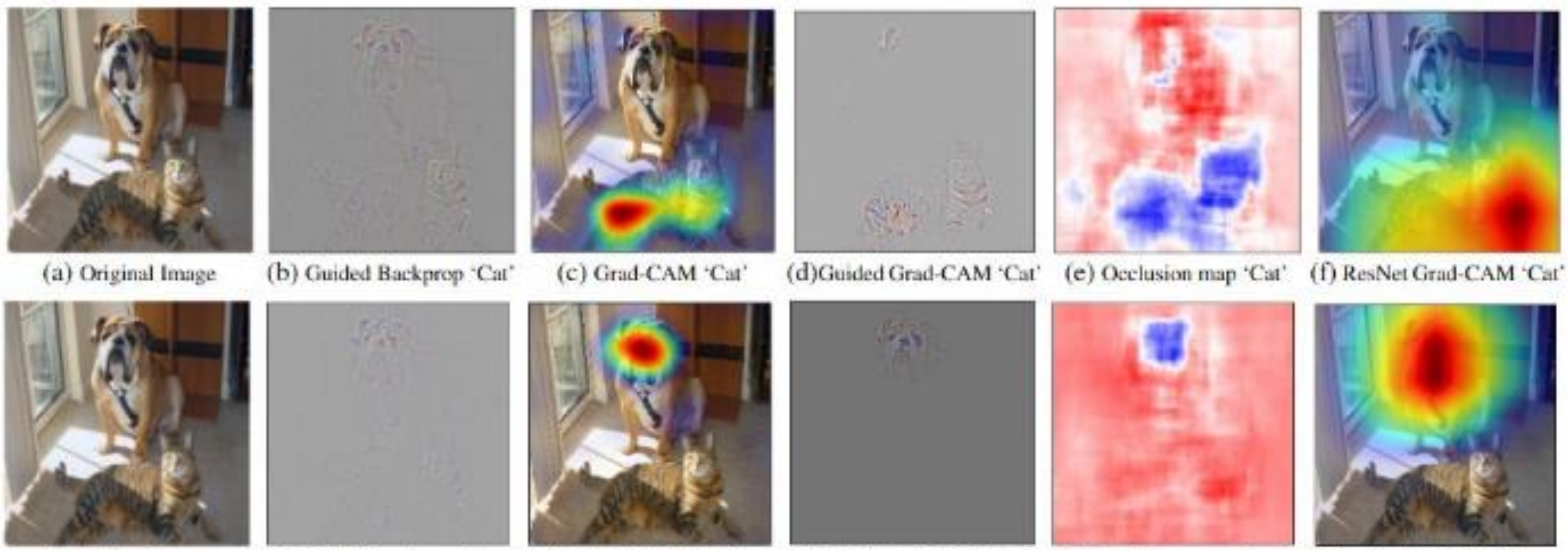
Class Activation Mapping (CAM)

- Final Step
  - Rescale the 7x7 image to the original image's size (224 x 224 for ResNet), and plot the 2 images over each other



https://lazyprogrammer.me

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

Other Examples

- Local Interpretable Model-agnostic Explanations (LIME)
  - https://homes.cs.washington.edu/~marcotcr/blog/lime/
- SHapley Additive exPlanations (SHAP)
  - https://shap.readthedocs.io/en/latest/
- Partial Dependence Plot (PDP)
  - https://scikit-learn.org/stable/modules/partial_dependence.html
- Accumulated Local Effects (ALE)
  - https://arxiv.org/pdf/1612.08468.pdf
- Individual Conditional Expectation (ICE)
  - https://arxiv.org/pdf/1309.6392.pdf