

Lecture 13:

Explainable AI (XAI)

https://github.com/kaopanboonyuen/SC310005_ArtificialIntelligence_2025s1

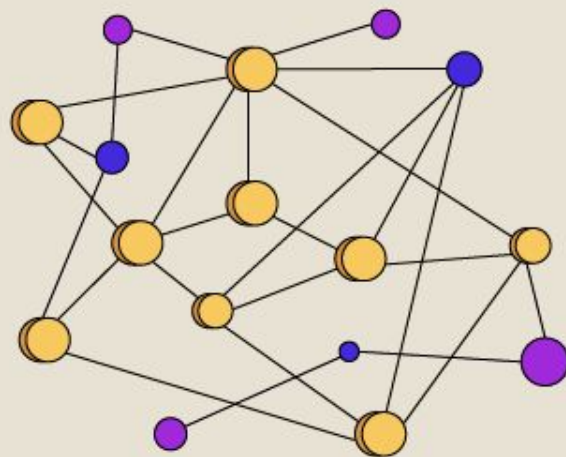
Teerapong Panboonyuen
<https://kaopanboonyuen.github.io>

Reference

- <https://www.geeksforgeeks.org/artificial-intelligence/explainable-artificial-intelligencexai/>
- <https://www.saltdatalabs.com/blog/ai-ethics/ai-ethics-101-what-is-explainable-ai>
- <https://www.techtarget.com/whatis/definition/explainable-AI-XAI>
- <https://ioe.engin.umich.edu/2025/06/13/new-ai-framework-increases-transparency-in-decision-making-systems/>

XAI

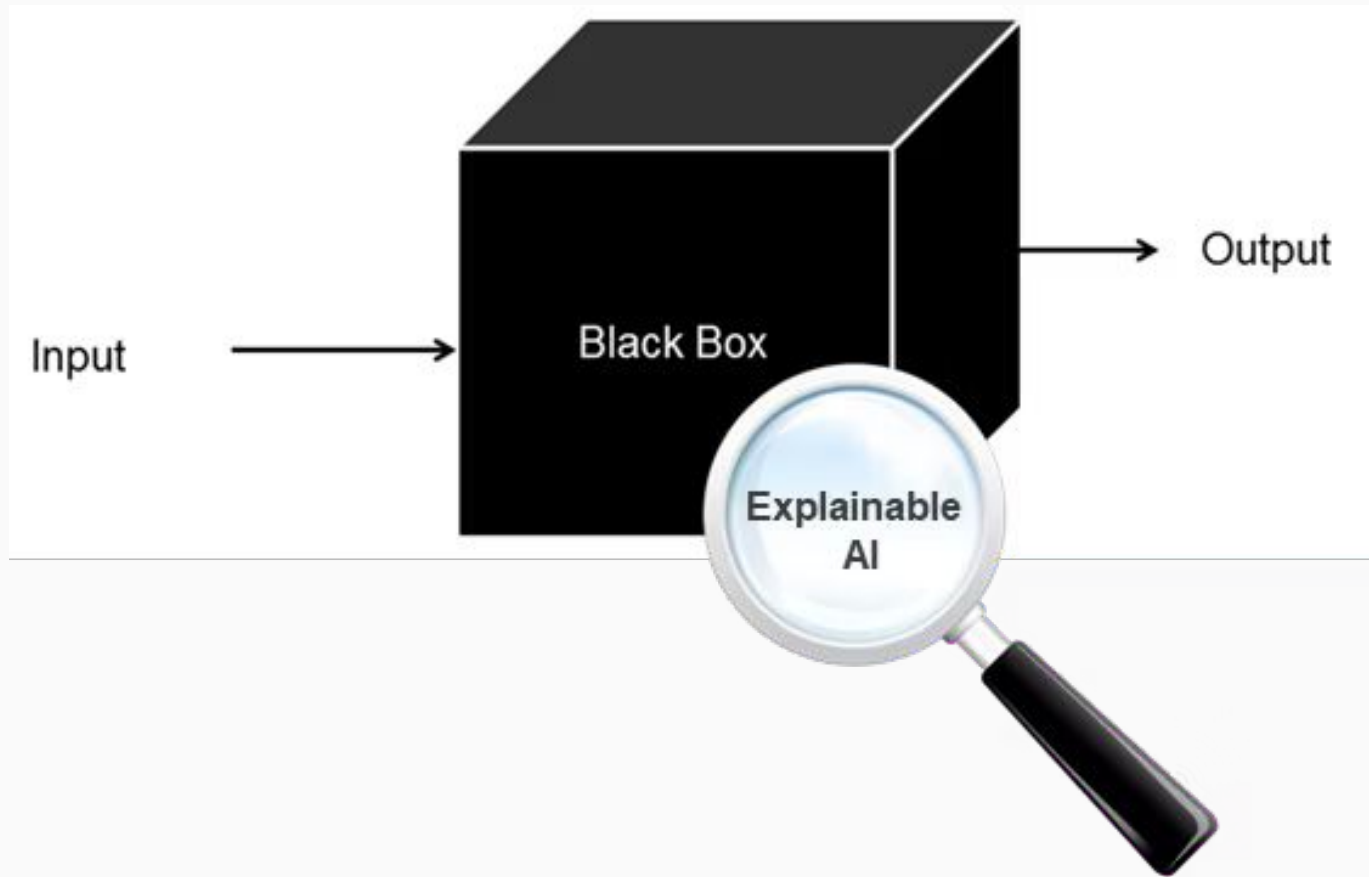
Explainable Artificial Intelligence



Explaining AI – Dive into Explainable AI (XAI) 🧠 ✨

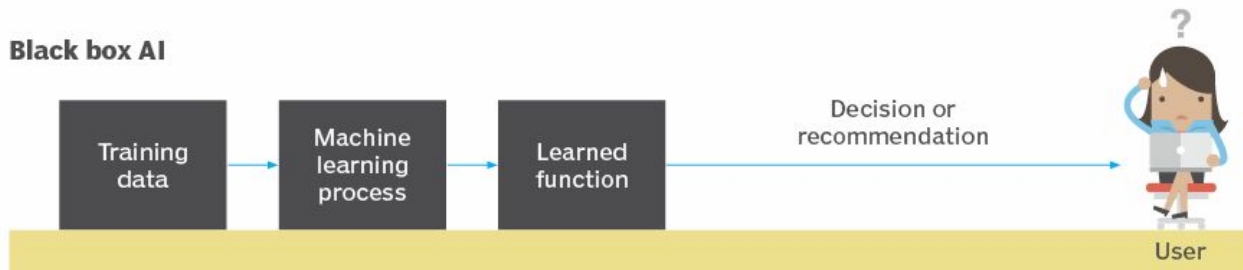
Introduction: Why Explainability Matters

- AI is everywhere: healthcare 🏥 , finance 💰 , autonomous driving 🚗 , and more.
- **Black-box models** (deep learning, ensemble methods) can be highly accurate but **hard to interpret**.
- XAI helps humans **understand, trust, and improve AI**.
- Example: Why did the AI classify a skin lesion as melanoma? 🔬

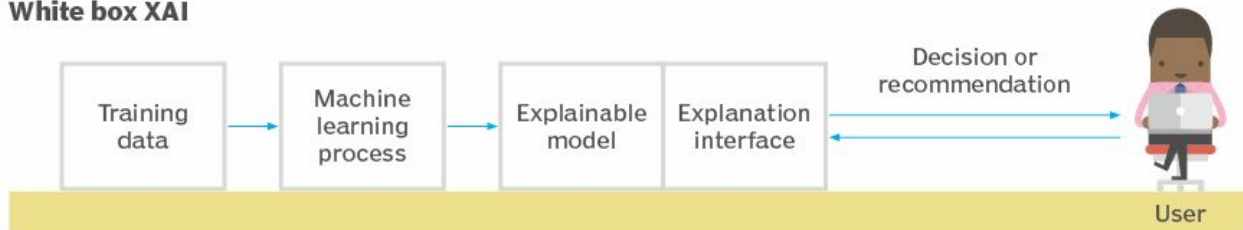


Black box AI vs. white box XAI

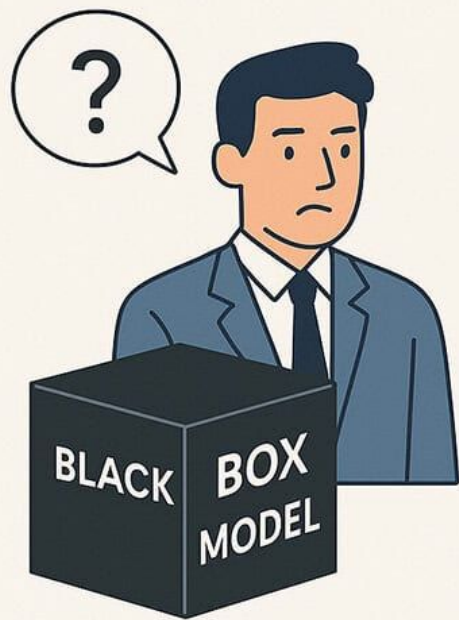
Black box AI



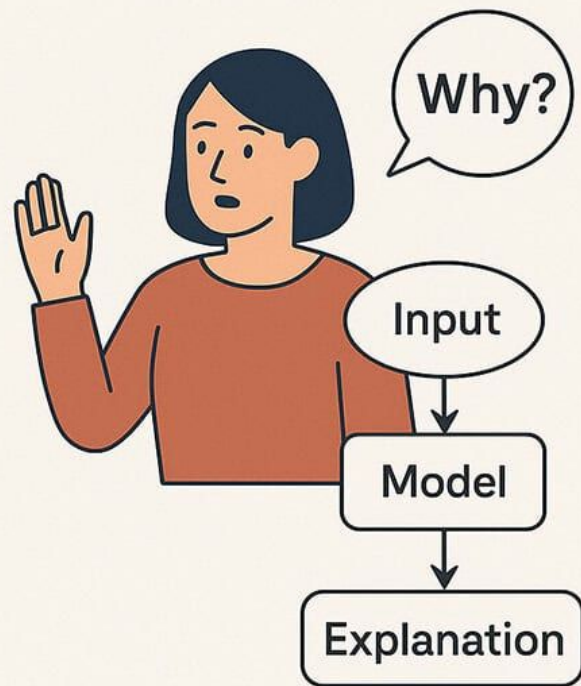
White box XAI

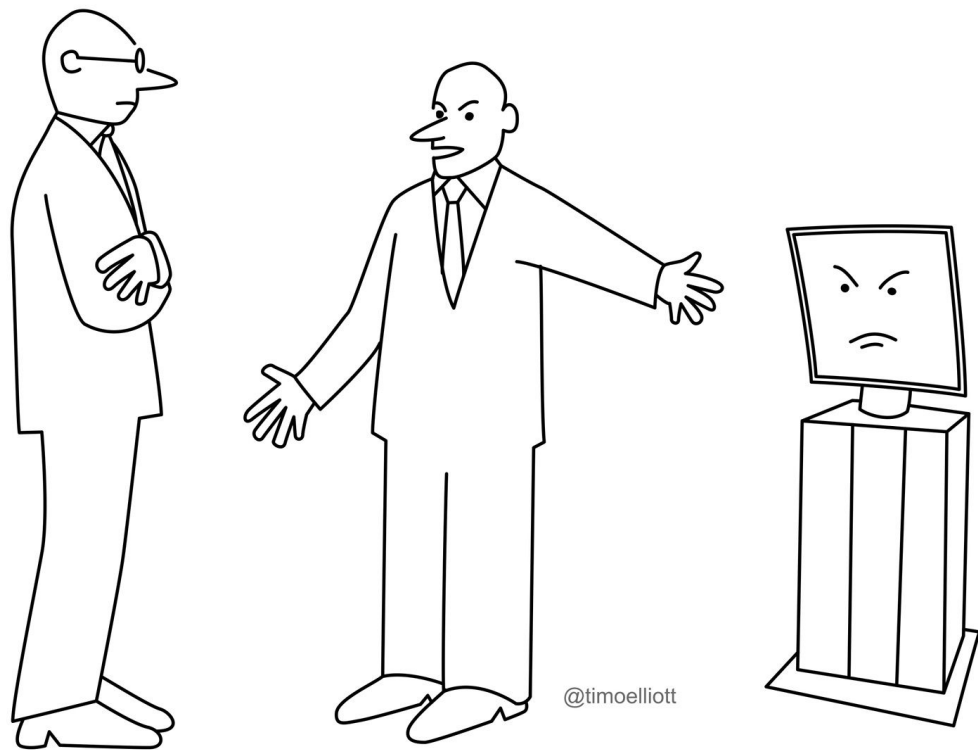


Need for Explainable-by-Design ML Methods



Prediction





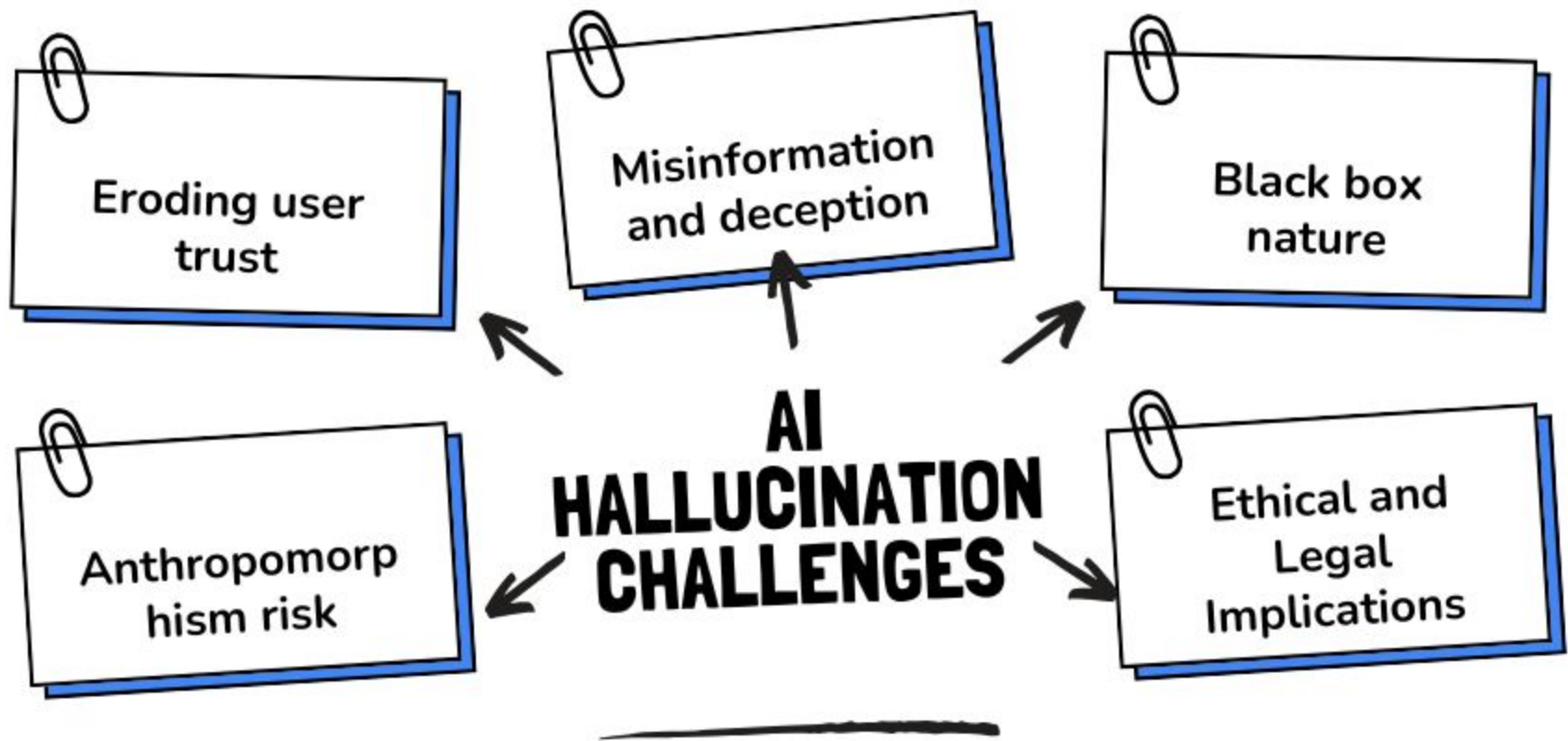
*His decisions aren't any better than yours
— but they're WAY faster...*

HOW DID YOU GET
BUDGET APPROVAL
FOR ALL THIS?

I JUST
TOLD THEM
THE NAME
OF THE
PROJECT.

AI

TOM
FISH
BURNE



Causes and Types of LLMs Hallucination

Causes of LLMs Hallucination

Source-Reference Divergence

Exploitation through Jailbreak Prompts

Reliance on Incomplete or Contradictory Datasets

Overfitting and Lack of Novelty

Guesswork from Vague or Insufficiently Detailed Prompts

Types of LLMs Hallucination



Sentence Contradiction



Prompt Contradiction



Factual Contradiction



Nonsensical Output



Irrelevant or Random Hallucinations

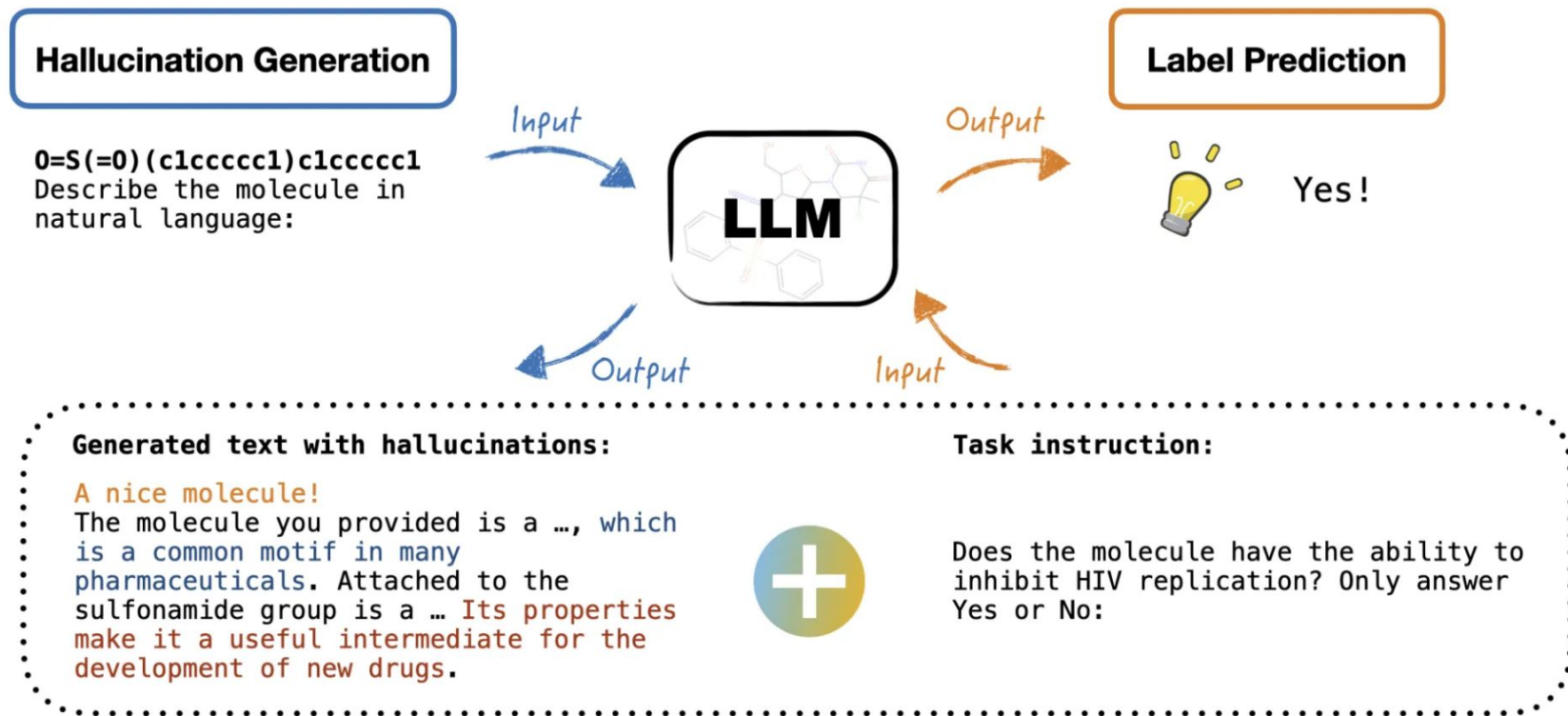


Figure 2: Illustration of the evaluation method with an example molecule from the HIV dataset: 1. We use LLMs to generate a textual description of the molecule based on its SMILES string. 2. The generated text, which contains hallucinations, is added to the prompt, and the LLM is tasked with predicting the specific property of the molecule. The answer is constrained to “Yes” or “No” to evaluate the LLM’s performance. We highlight obvious hallucinations that are unrelated to the input using colors.

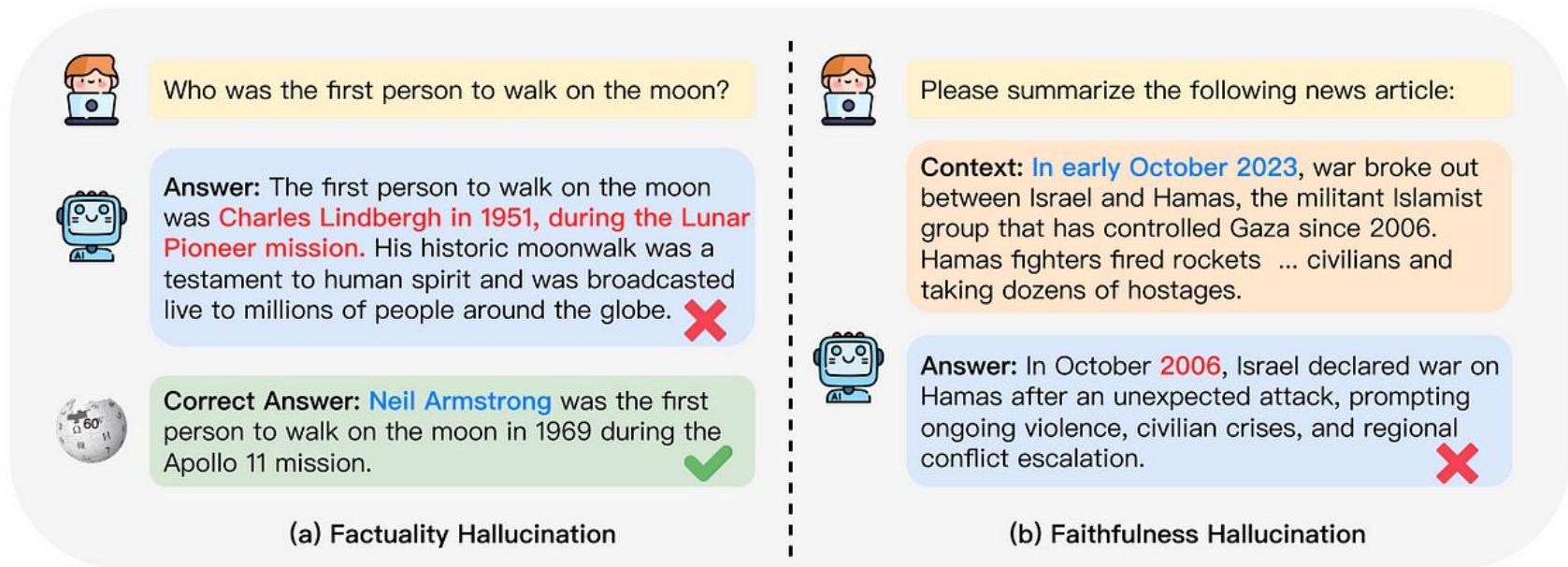
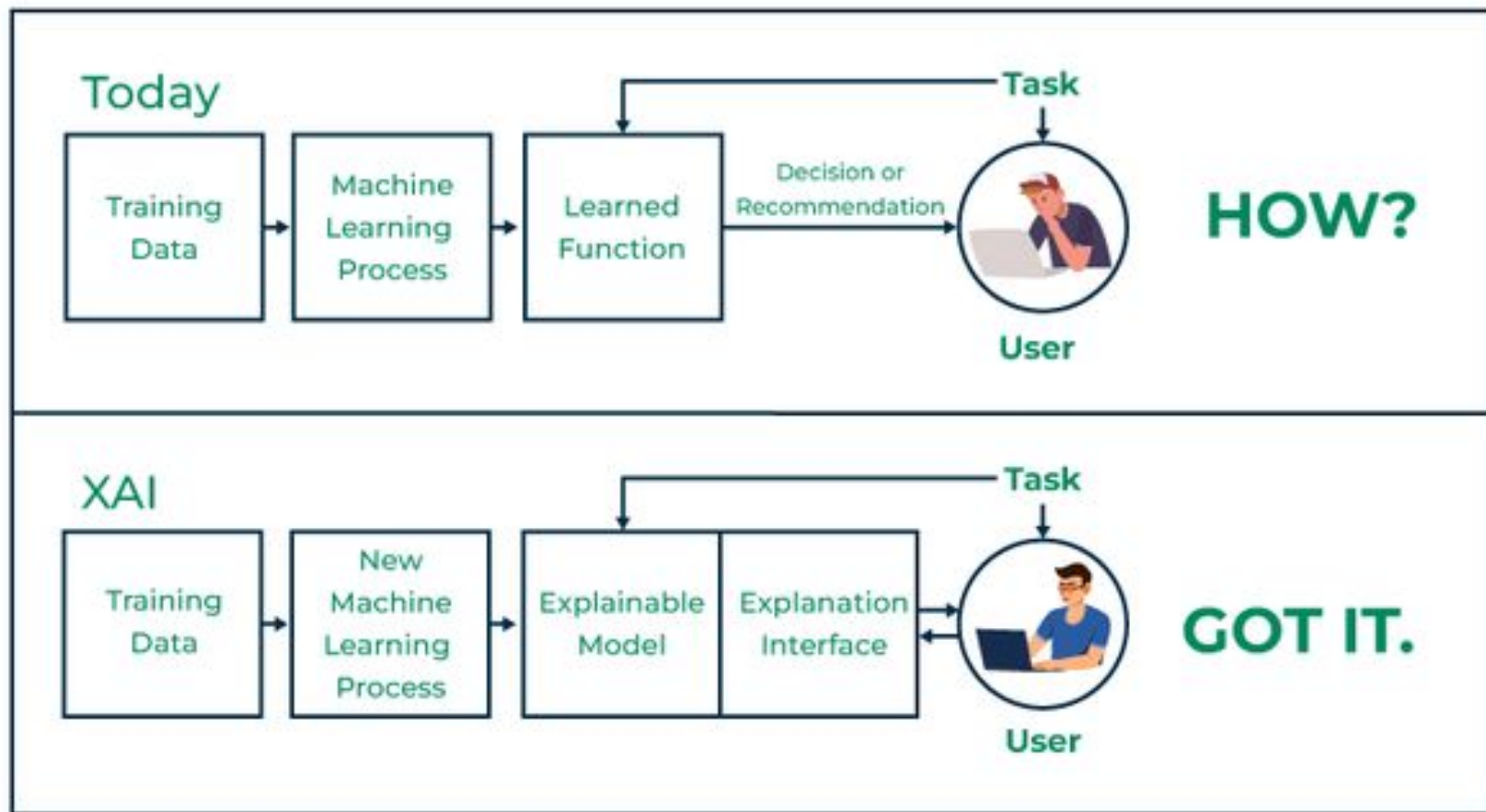
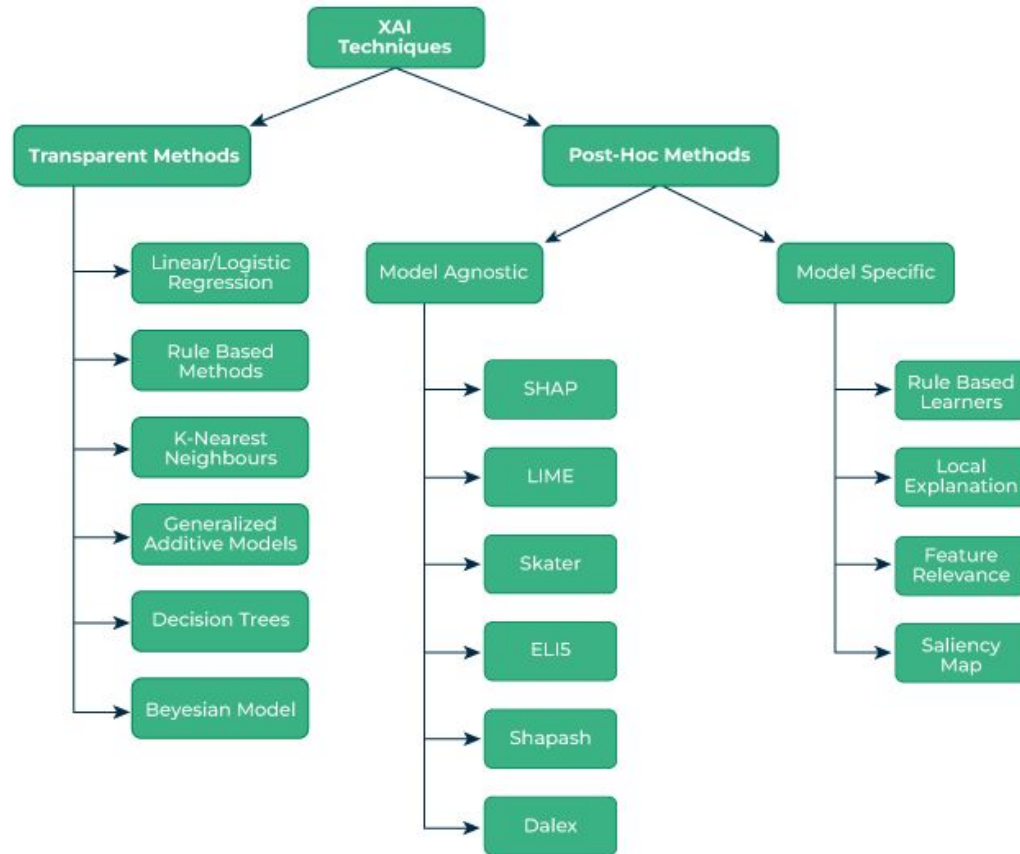
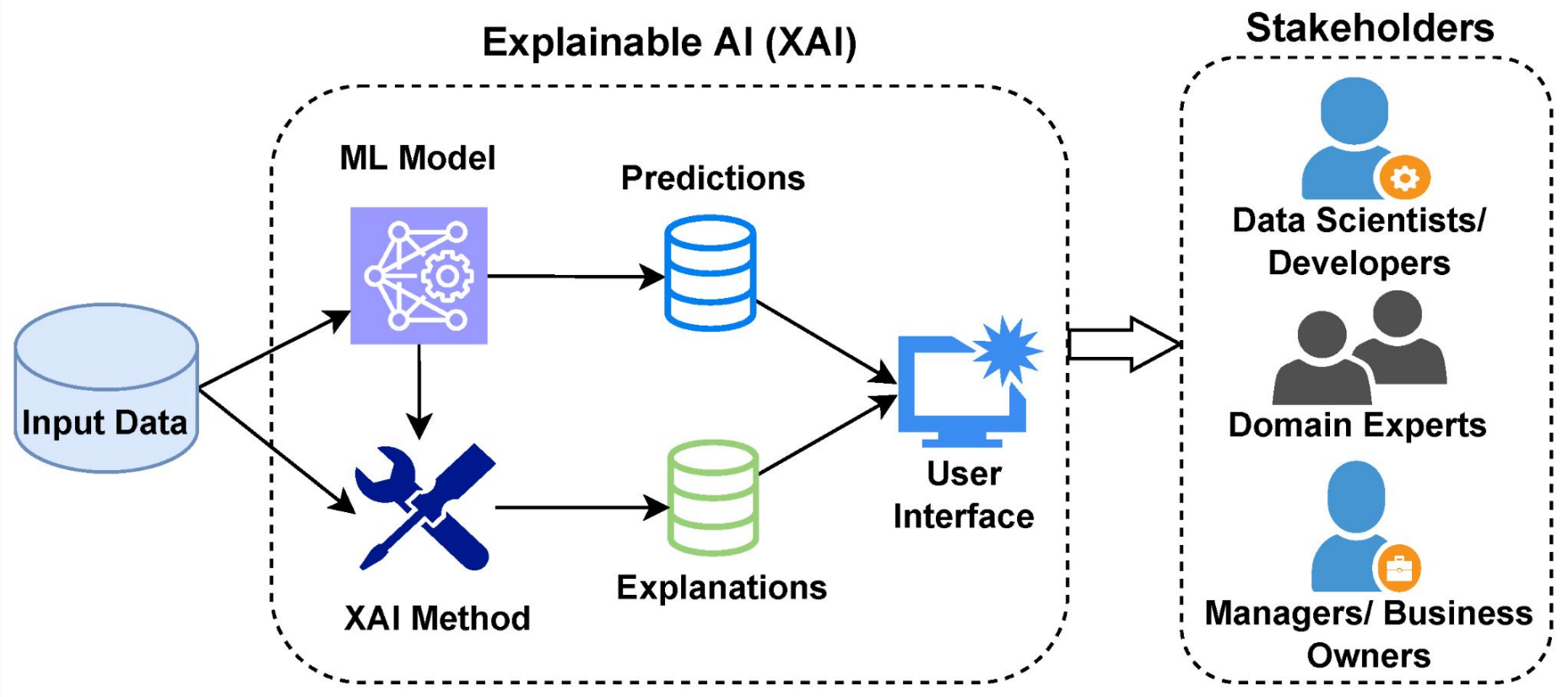


Figure 1: An intuitive example of LLM hallucination.



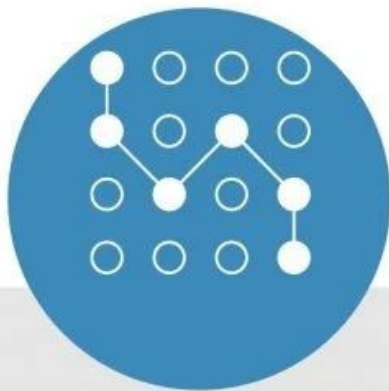






Explainable Data

What data was used to train the model and why?



Explainable Predictions

What features and weights were used for this particular prediction?



Explainable Algorithms

What are the individual layers and the thresholds used for a prediction?

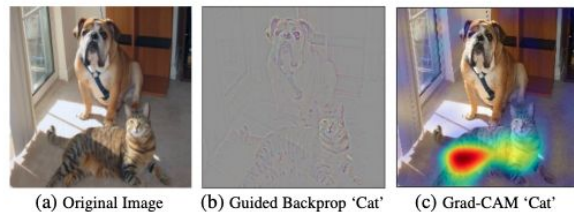
Questions around AI explainability help us understand how data, predictions and algorithms influence decisions.

Grad-CAM

(Gradient-weighted Class
Activation Mapping)

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

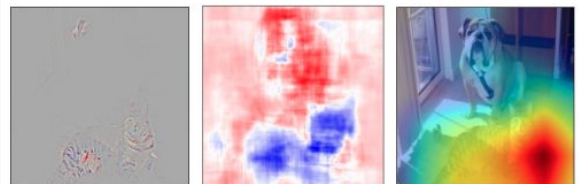
Ramprasaath R. Selvaraju · Michael Cogswell · Abhishek Das · Ramakrishna Vedantam · Devi Parikh · Dhruv Batra



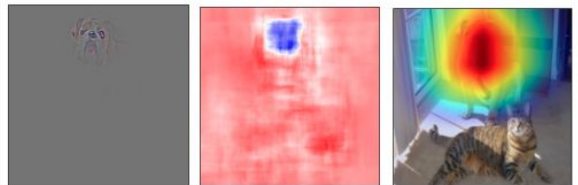
(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat'



(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog'



(d) Guided Grad-CAM 'Cat' (e) Occlusion map 'Cat' (f) ResNet Grad-CAM 'Cat'



(j) Guided Grad-CAM 'Dog' (k) Occlusion map 'Dog' (l) ResNet Grad-CAM 'Dog'

1610.02391v4 [cs.CV] 3 Dec 2019

Abstract We propose a technique for producing ‘visual explanations’ for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent and explainable.

Our approach – Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say ‘dog’ in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (*e.g.* VGG), (2) CNNs used for structured outputs (*e.g.* captioning), (3) CNNs used in tasks with multi-modal inputs (*e.g.* visual question answering) or reinforcement learning, all *without architectural changes or re-training*. We combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative vi-

ualization, Guided Grad-CAM, and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures.

In the context of image classification models, our visualizations (a) lend insights into failure modes of these models (showing that seemingly unreasonable predictions have reasonable explanations), (b) outperform previous methods on the ILSVRC-15 weakly-supervised localization task, (c) are robust to adversarial perturbations, (d) are more faithful to the underlying model, and (e) help achieve model generalization by identifying dataset bias.

For image captioning and VQA, our visualizations show that even non-attention based models learn to localize discriminative regions of input image.

We devise a way to identify important neurons through Grad-CAM and combine it with neuron names [4] to provide textual explanations for model decisions. Finally, we design and conduct human studies to measure if Grad-CAM explanations help users establish appropriate trust in predictions from deep networks and show that Grad-CAM helps untrained users successfully discern a ‘stronger’ deep net-

Ramprasaath R. Selvaraju
Georgia Institute of Technology, Atlanta, GA, USA
E-mail: ramprs@gatech.edu

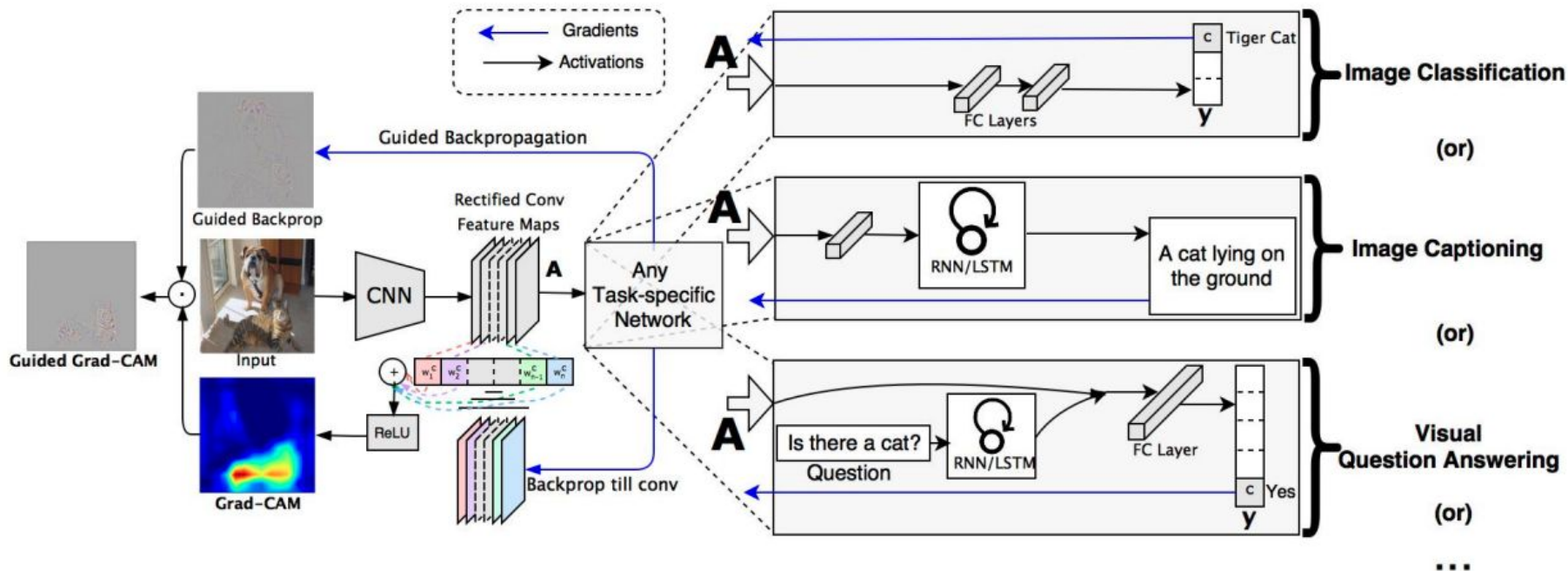


Fig. 2: Grad-CAM overview: Given an image and a class of interest (e.g., ‘tiger cat’ or any other type of differentiable output) as input, we forward propagate the image through the CNN part of the model and then through task-specific computations to obtain a raw score for the category. The gradients are set to zero for all classes except the desired class (tiger cat), which is set to 1. This signal is then backpropagated to the rectified convolutional feature maps of interest, which we combine to compute the coarse Grad-CAM localization (blue heatmap) which represents where the model has to look to make the particular decision. Finally, we pointwise multiply the heatmap with guided backpropagation to get Guided Grad-CAM visualizations which are both high-resolution and concept-specific.

Input Image



Ground-Truth: Nurse

Grad-CAM for
Biased model



Predicted: Nurse

Grad-CAM for
Unbiased model



Predicted: Nurse



Ground-Truth: Doctor



Predicted: Nurse



Predicted: Doctor

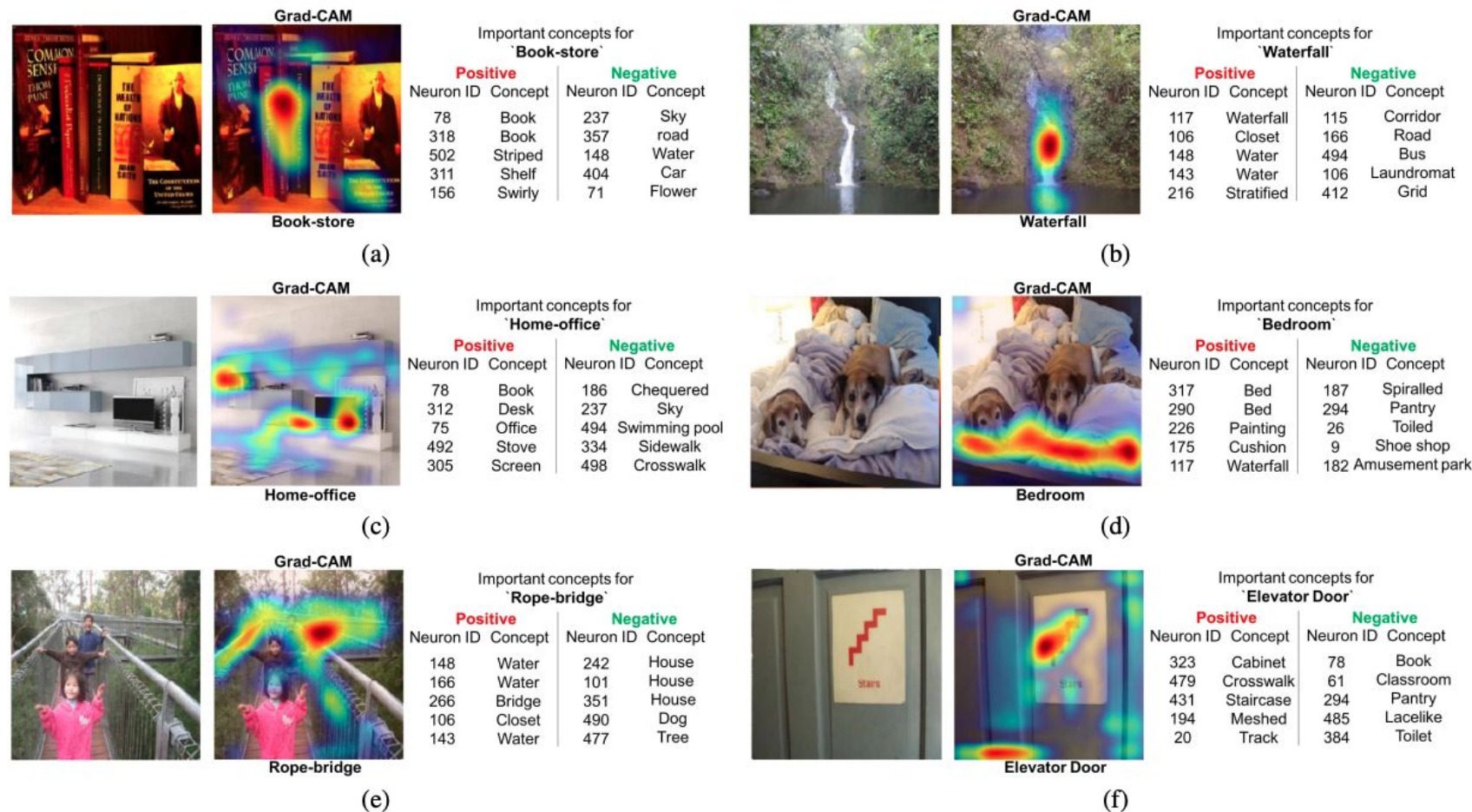


Fig. 9: Examples showing visual explanations and textual explanations for VGG-16 trained on Places365 dataset [61]. For textual explanations

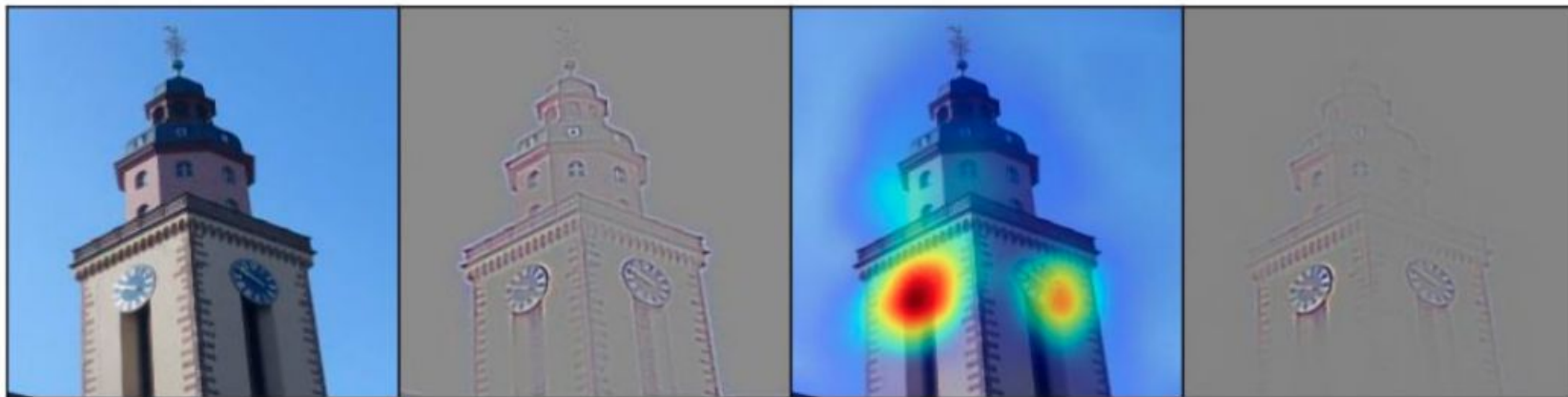
Guided Backprop

Grad-CAM

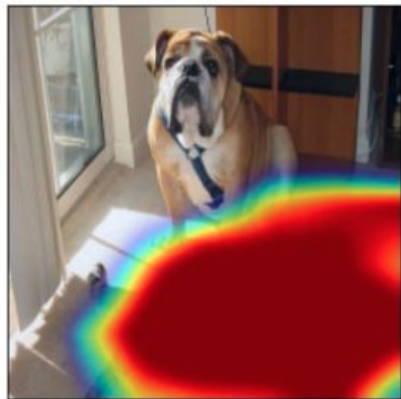
Guided Grad-CAM



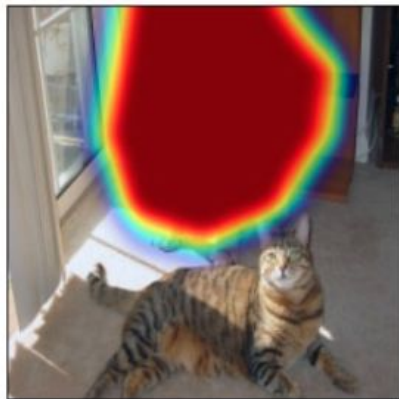
A man is holding a hot dog in his hand



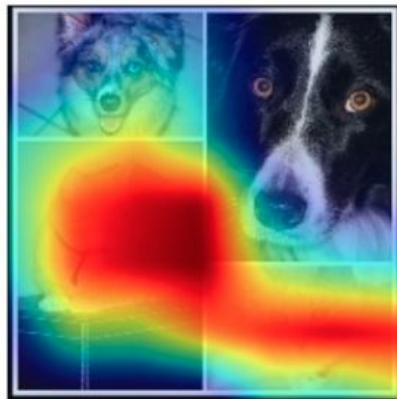
A large clock tower with a clock on the top of it



Grad-CAM for last Residual block



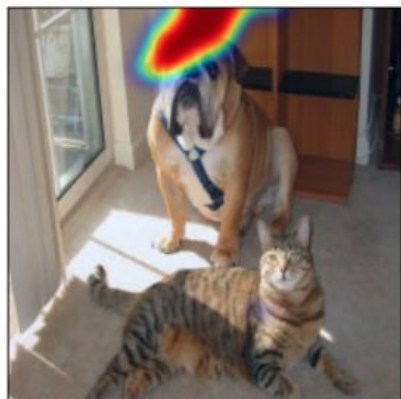
Grad-CAM for last Residual block



Grad-CAM for last Residual block



Grad-CAM for last Residual block



Grad-CAM for last downsampling layer



Grad-CAM for last downsampling layer



Grad-CAM for last downsampling layer



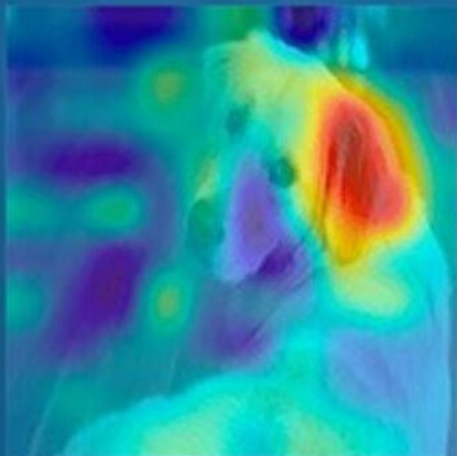
Grad-CAM for last downsampling layer

(a) Grad-CAM visualizations for the ResNet-200 layer architecture for 'tiger cat' (left) and 'boxer' (right) category.

(b) Grad-CAM visualizations for the ResNet-200 layer architecture for 'tabby cat' (left) and 'boxer' (right) category.

Different Technique – Different Explanation

LIME

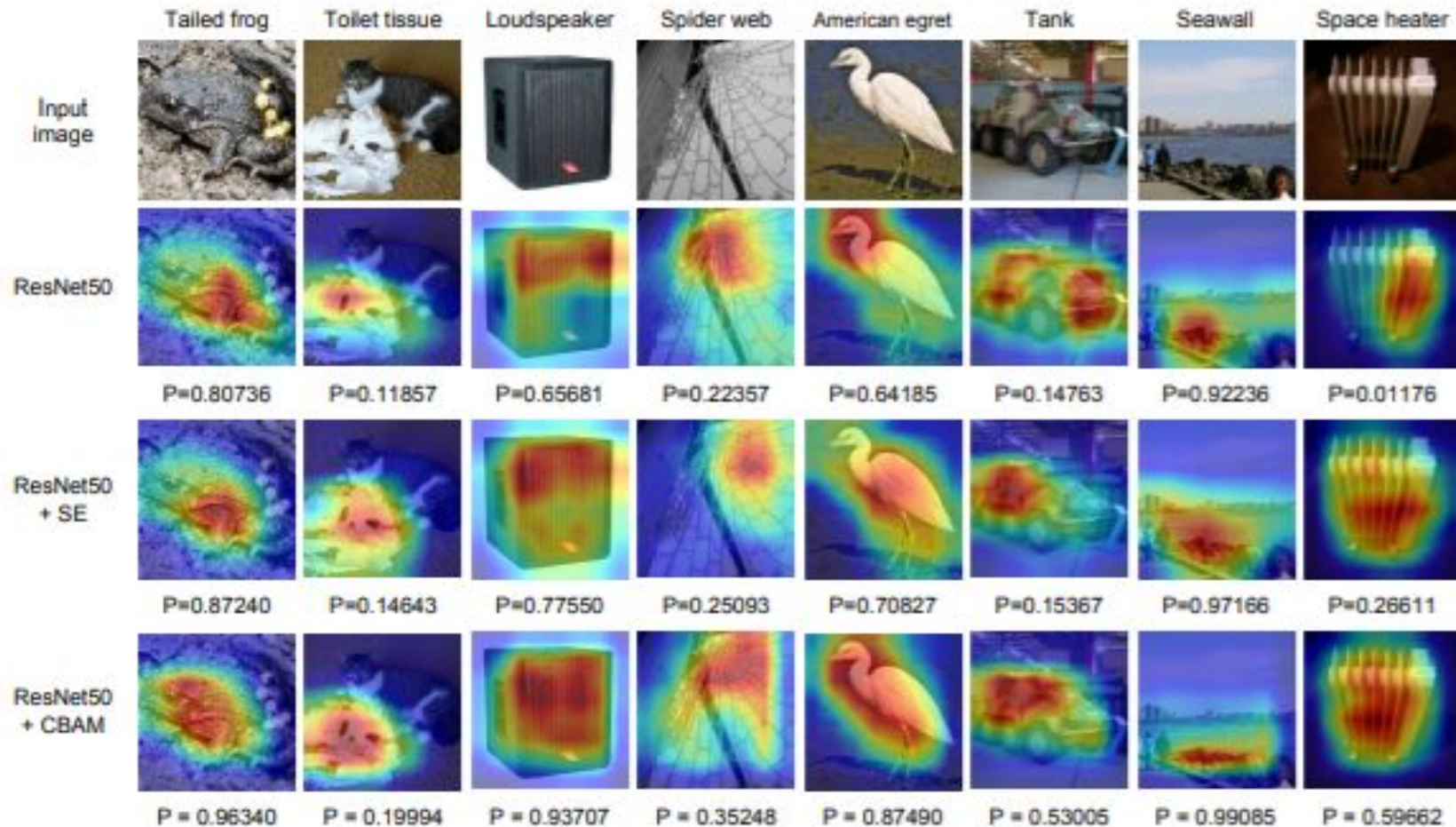


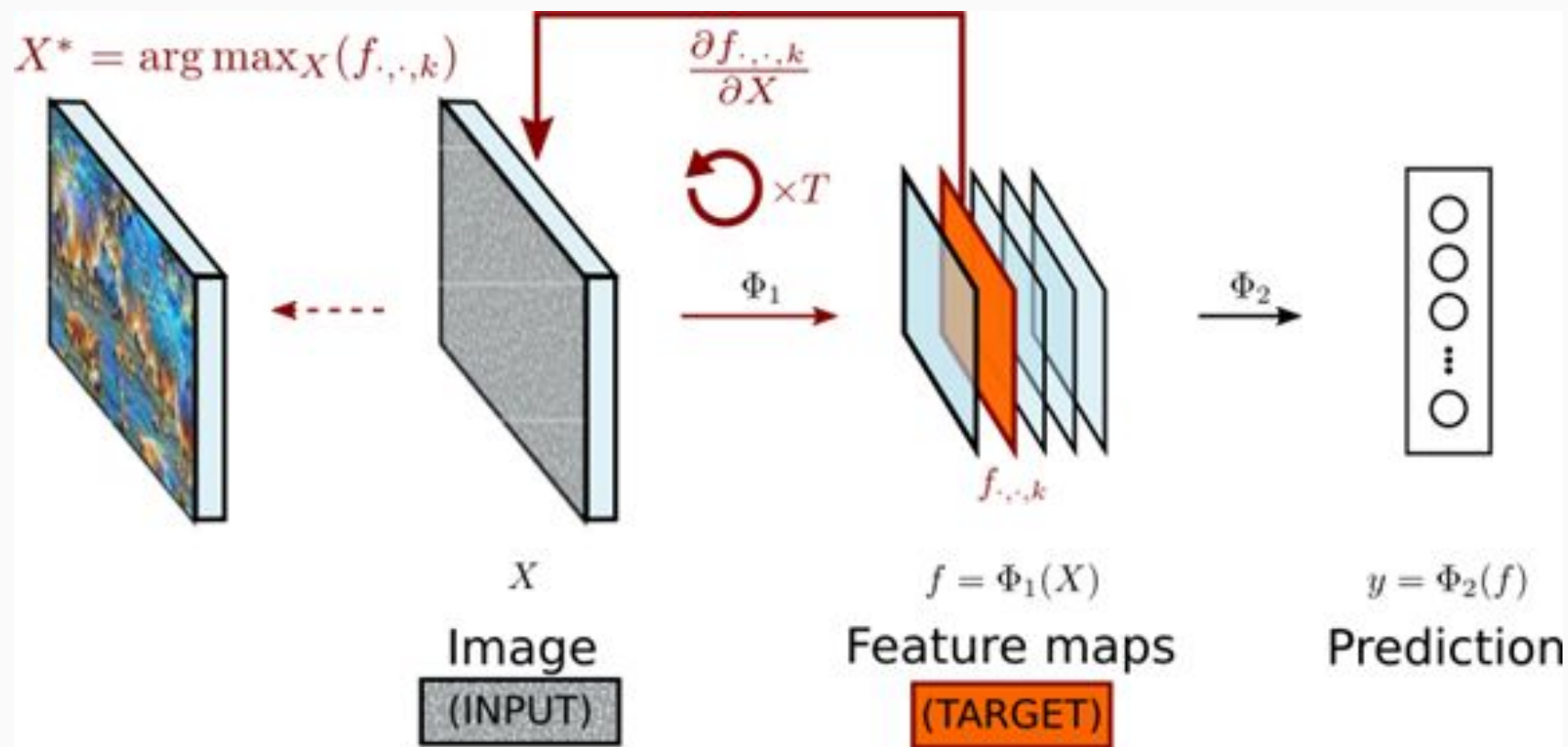
Grad-CAM

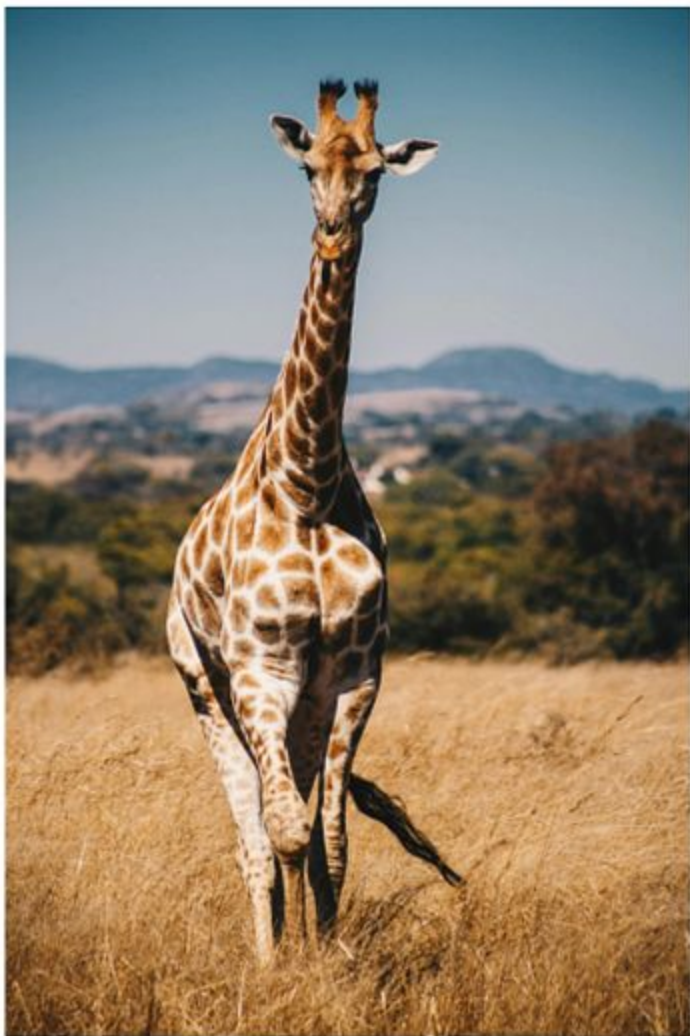


Occlusion Sensitivity



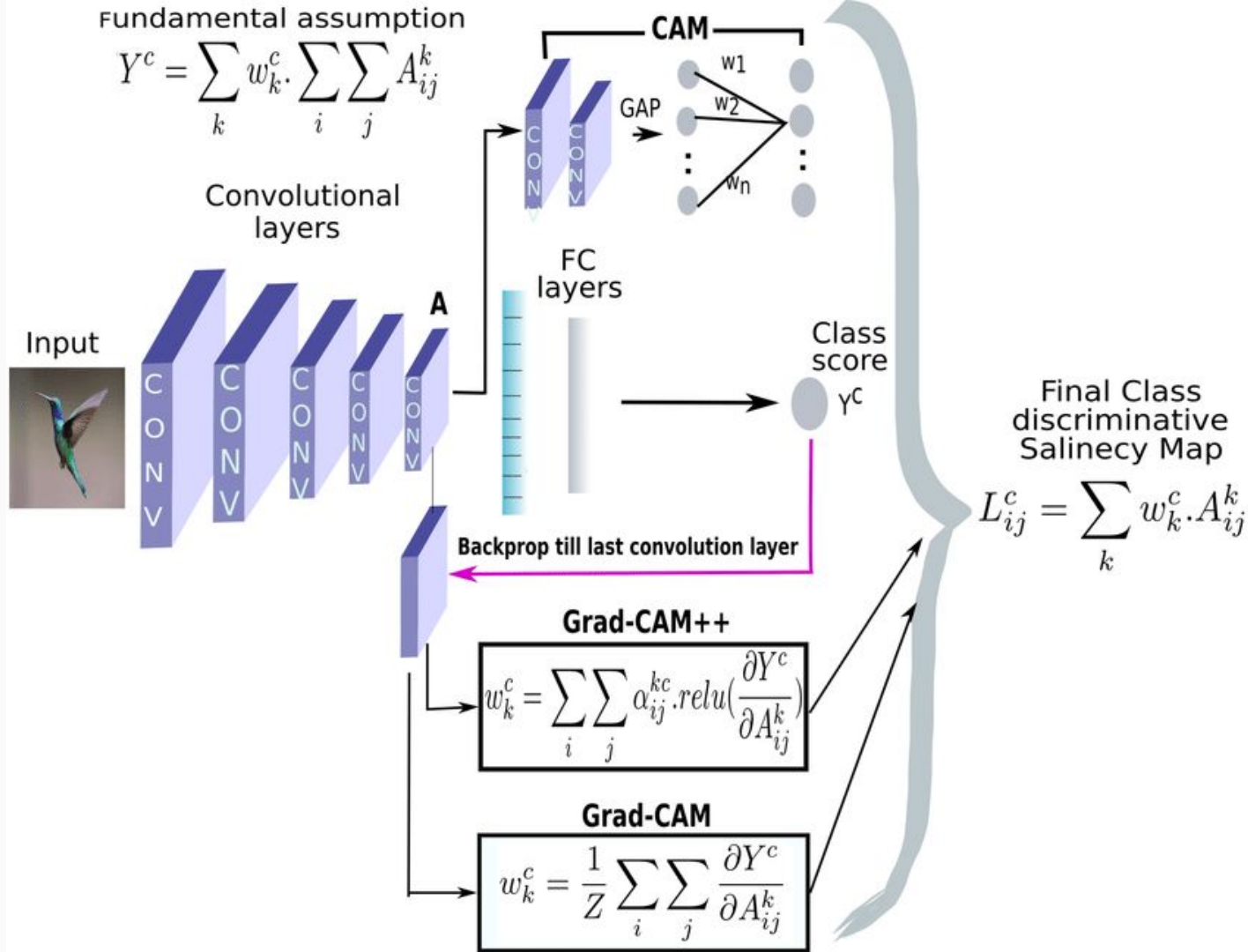






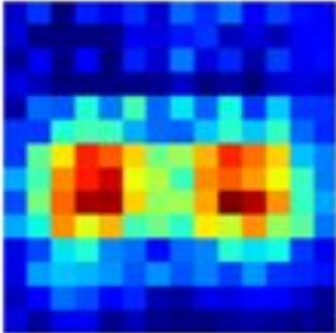
layer2 + layer3 + layer4



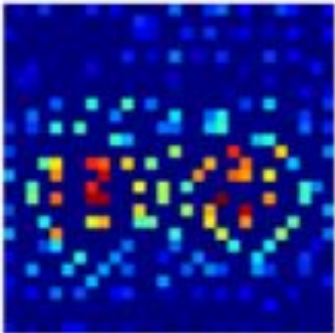




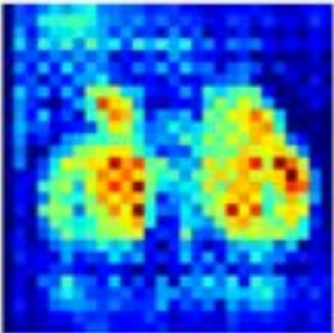
(a)



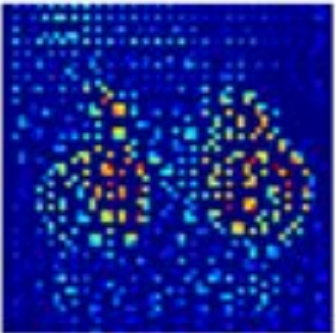
(b)



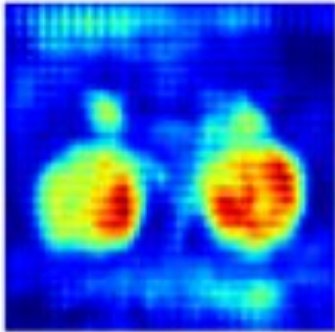
(c)



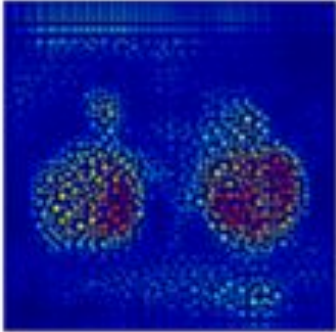
(d)



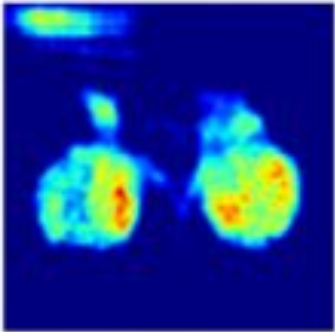
(e)



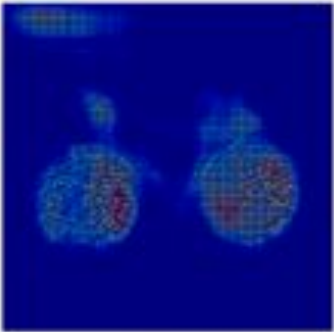
(f)



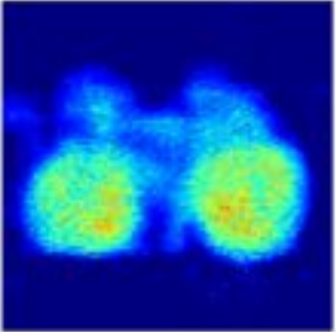
(g)



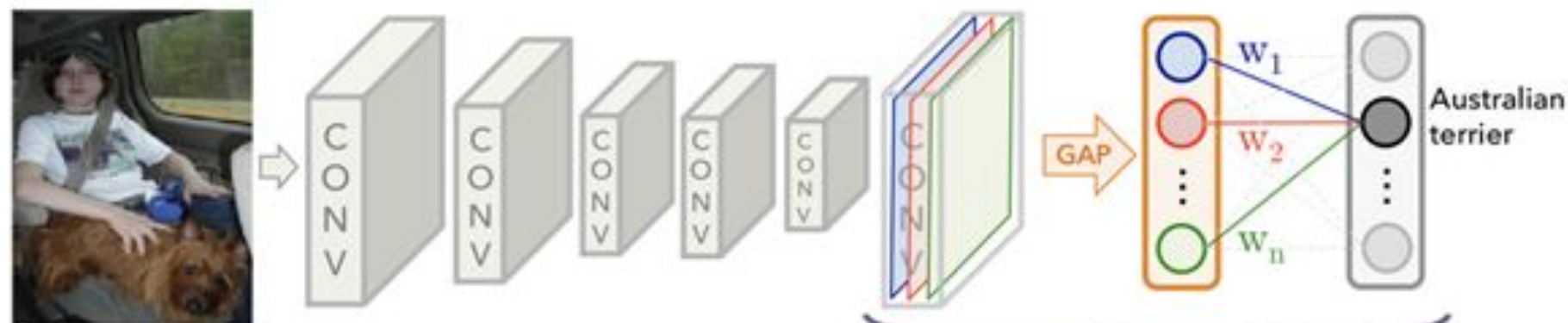
(h)



(i)



(j)

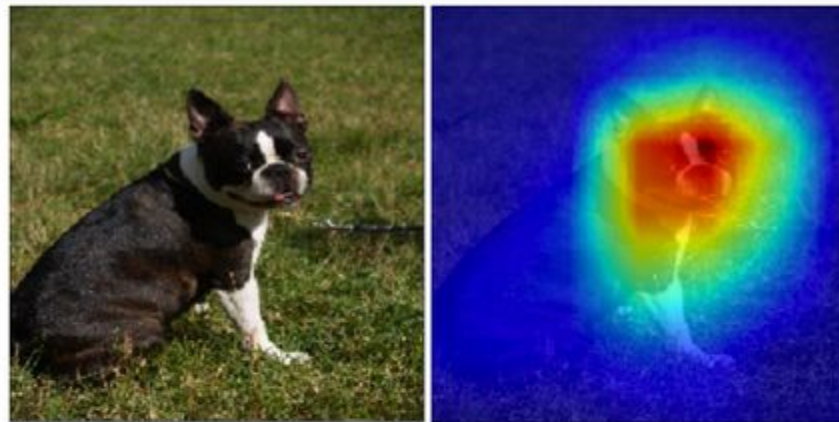


Class Activation Mapping





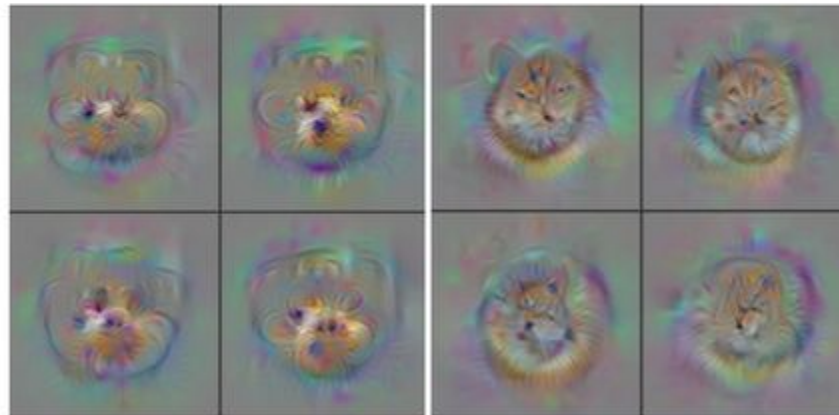
(1) Heatmap



(2) Grad-CAM



(3) Saliency map



(4) Activations visualization

AI

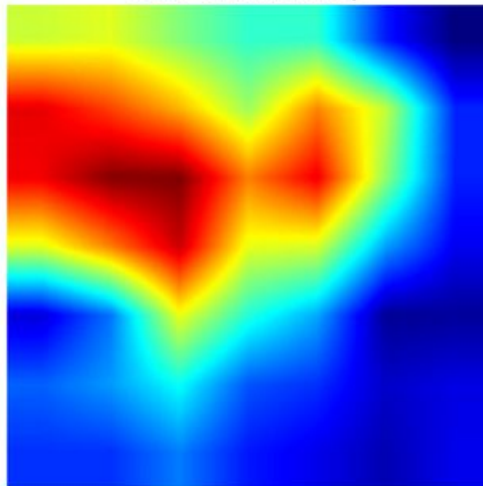
EXPLAINABLE AI

Explaining Neural Network Predictions
with PyTorch and Grad-CAM

Original Image



Grad-CAM Heatmap



Overlaid Image (Class: Labrador retriever, Prob: 0.9882)



What is Explainable AI (XAI)? 🤔

- XAI is a set of techniques that **make AI decisions transparent**.
- Bridges the gap between **accuracy and interpretability**.
- Types of explanations:
 - **Global explanations:** How the model behaves overall 🌐
 - **Local explanations:** Why the model made a specific prediction for one sample 🔍

Why Use XAI? 🌟

- **Build Trust:** Stakeholders can understand AI decisions. ✅
- **Debug Models:** Detect bias, errors, or overfitting 🐛
- **Regulatory Compliance:** Some industries require interpretable AI ⚖️
- **Better Communication:** Explain AI to non-technical users 💬
- **Improvement & Feedback:** Insights from explanations can guide model refinement 🔧

Popular XAI Techniques

1. LIME – Local Interpretable Model-Agnostic Explanations 🔍

- Explains predictions **locally** using **interpretable models**.
- Highlights which features contribute most to a prediction.
- Works for **tabular, text, and image data**.
- Visual example: Feature importance bar chart or text highlights.

2. SHAP – SHapley Additive exPlanations

- Based on **game theory**.
- Quantifies **how much each feature contributed** to a prediction.
- Consistent and theoretically grounded.
- Visual example: Waterfall plots, beeswarm plots.

3. Grad-CAM – Gradient-weighted Class Activation Map 🌈

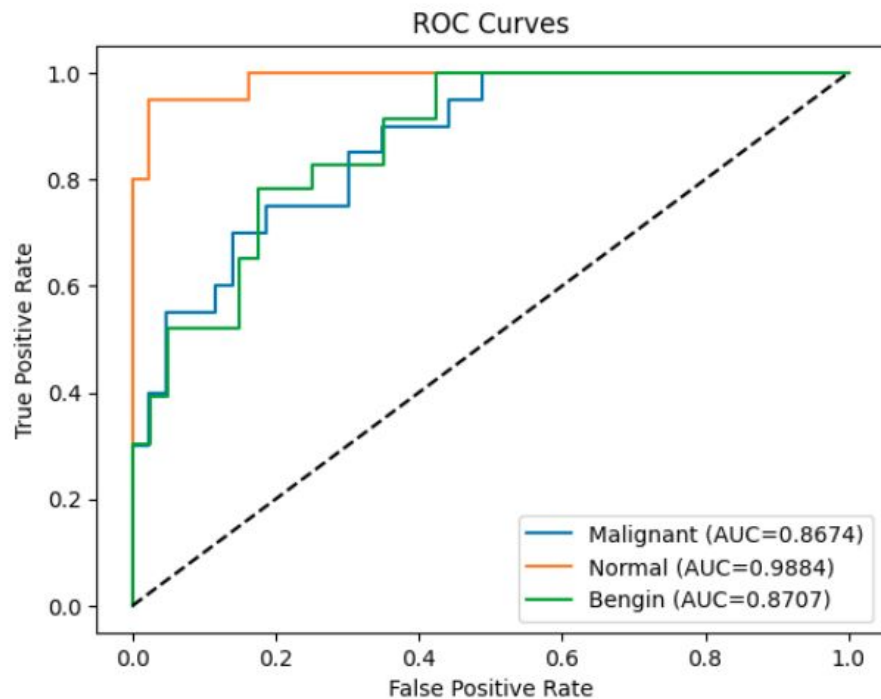
- Designed for **CNNs and image classification**.
- Visualizes **which parts of an image influenced the model prediction**.
- Overlay heatmap on image: red areas = high influence 🔥
- Great for medical AI, e.g., **highlighting lesions in skin cancer images**.

How to Use XAI in Practice

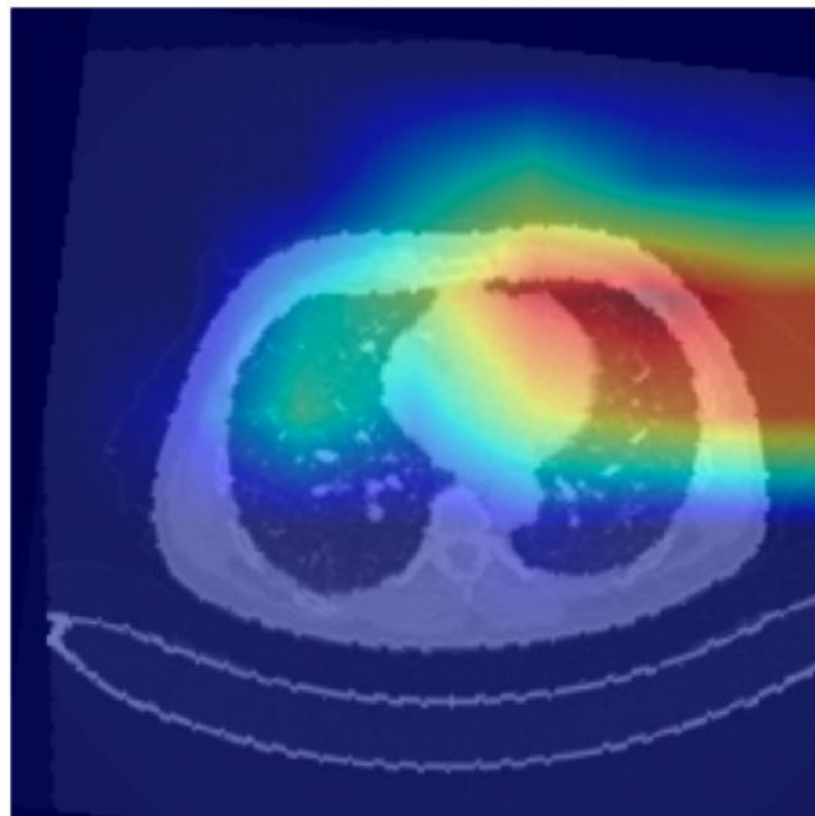
- **Step 1:** Train your model (CNN, ResNet, etc.)
- **Step 2:** Choose an XAI method suitable for your task:
 - LIME or SHAP for tabular/text
 - Grad-CAM for images
- **Step 3:** Generate explanations for predictions
- **Step 4:** Visualize results (bar charts, heatmaps, feature importance)
- **Step 5:** Interpret and communicate insights to stakeholders



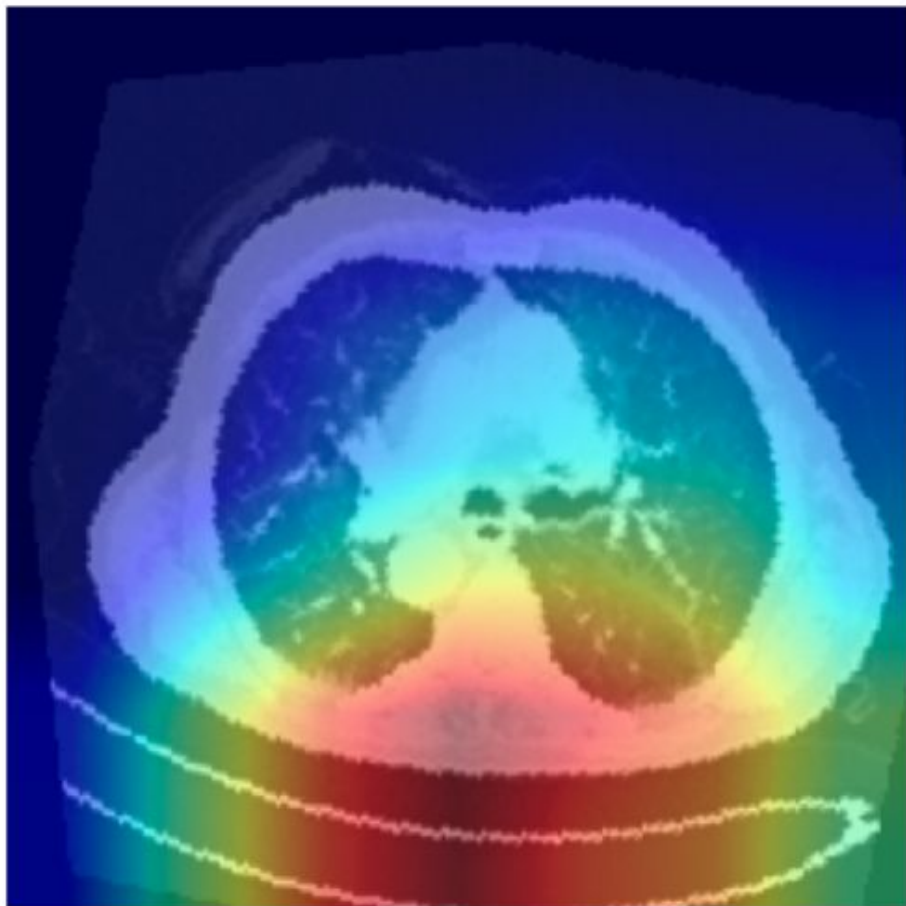
	precision	recall	f1-score	support
Malignant	0.7619	0.8000	0.7805	20
Normal	0.7407	1.0000	0.8511	20
Bengin	0.8667	0.5652	0.6842	23
accuracy			0.7778	63
macro avg	0.7898	0.7884	0.7719	63
weighted avg	0.7934	0.7778	0.7677	63



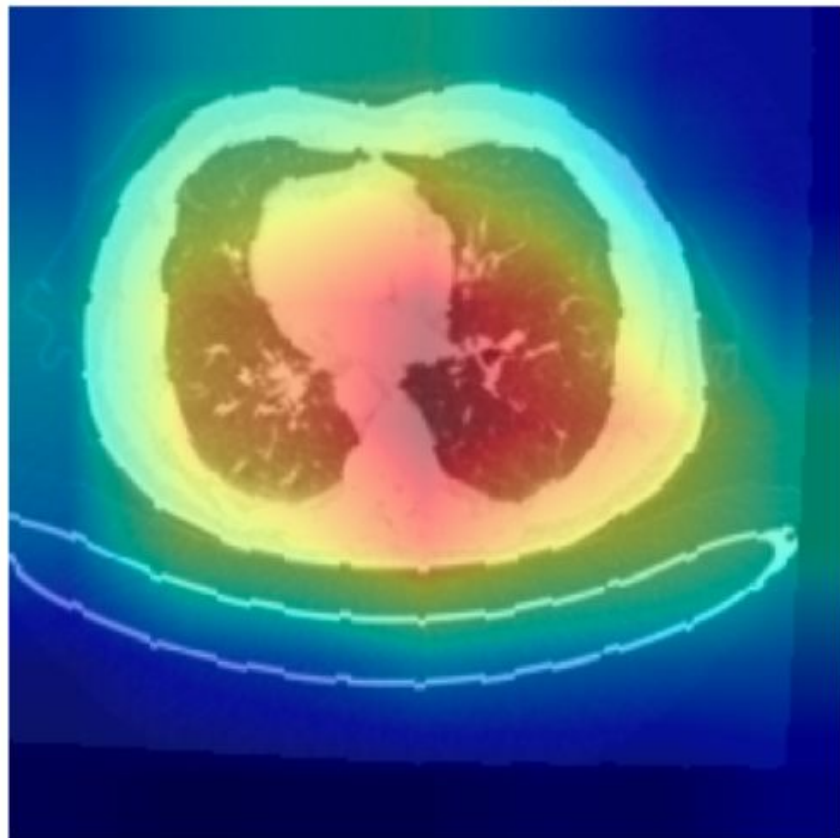
Grad-CAM for Bengin




Grad-CAM for Normal



Grad-CAM for Malignant



Best Practices & Tips

- Always **visualize explanations** – humans understand images/charts better than raw numbers.
- Combine **global and local explanations** for deeper insights.
- Use XAI not only to explain but also to **improve your models**.
- Avoid over-reliance: explanations are **tools, not absolute truths** 

XAI in Action – Example Scenarios 🌟

- **Healthcare:** Highlight tumor regions in X-ray or skin images 🩺
- **Finance:** Explain credit score predictions 🏠
- **Autonomous Driving:** Show which objects influenced steering decisions 🚦
- **Text Analysis:** Highlight words that drove sentiment classification 📄

Prediction probabilities

Benign	0.00
Malignant	1.00

Benign

Malignant

radius_mean <= 11.71	0.00
16.34 < texture_mean...	0.00
perimeter_mean <= ...	0.00
area_mean <= 420.40	0.00
0.10 < smoothness_m...	0.00
compactness_mean >...	0.00
concavity_mean > 0.13	0.00
0.03 < concave points...	0.00
symmetry_mean > 0.20	0.00
fractal_dimension_me...	0.00

Takeaway

- **XAI = Transparency + Trust + Improvement**
- Choosing the right method depends on your **data type and goal**
- Explaining AI is as important as building AI! 🏗️💡

Q&A in XAI?

