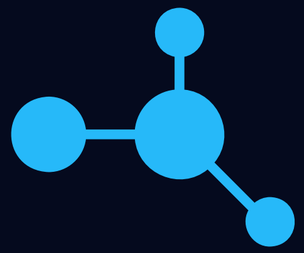
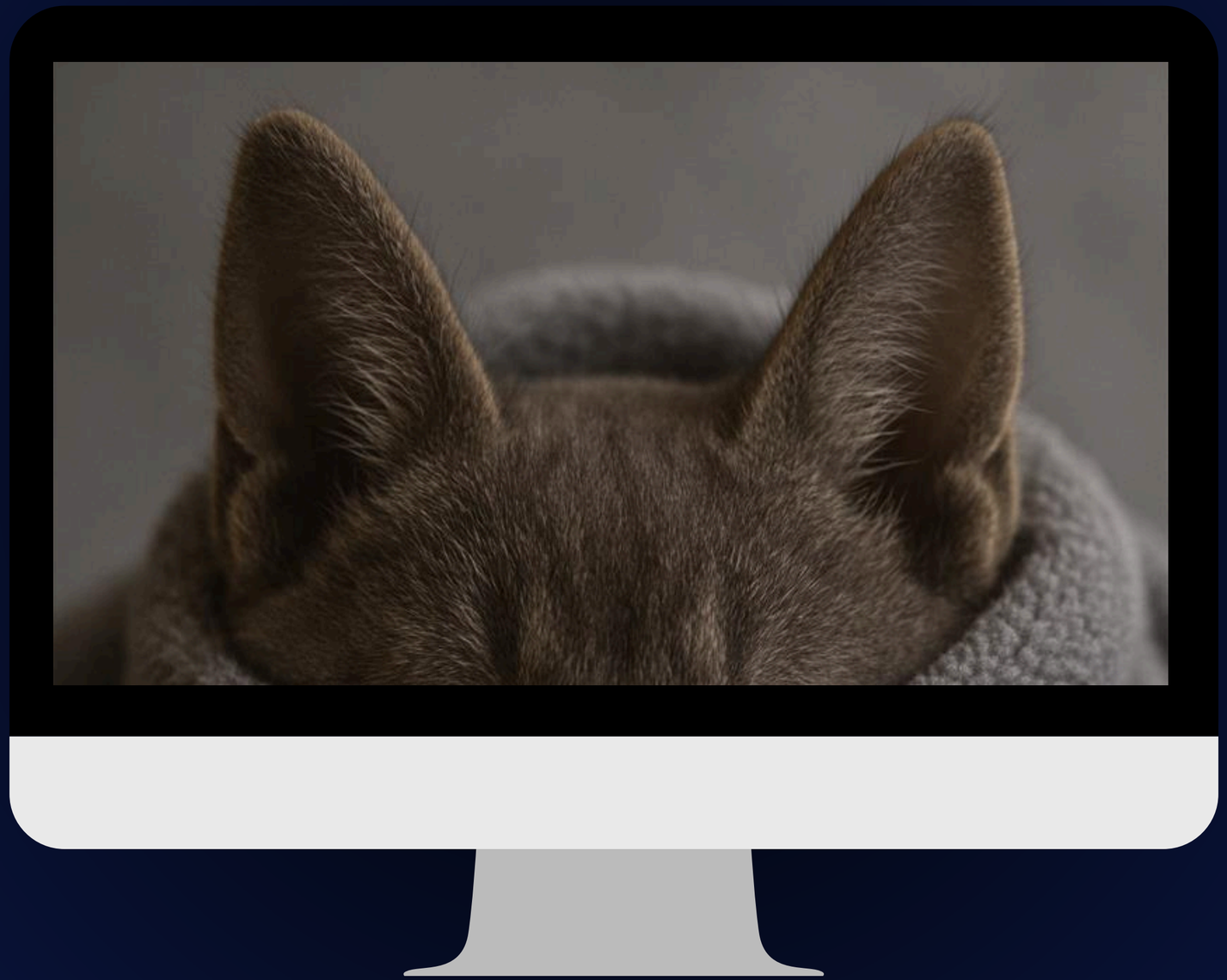
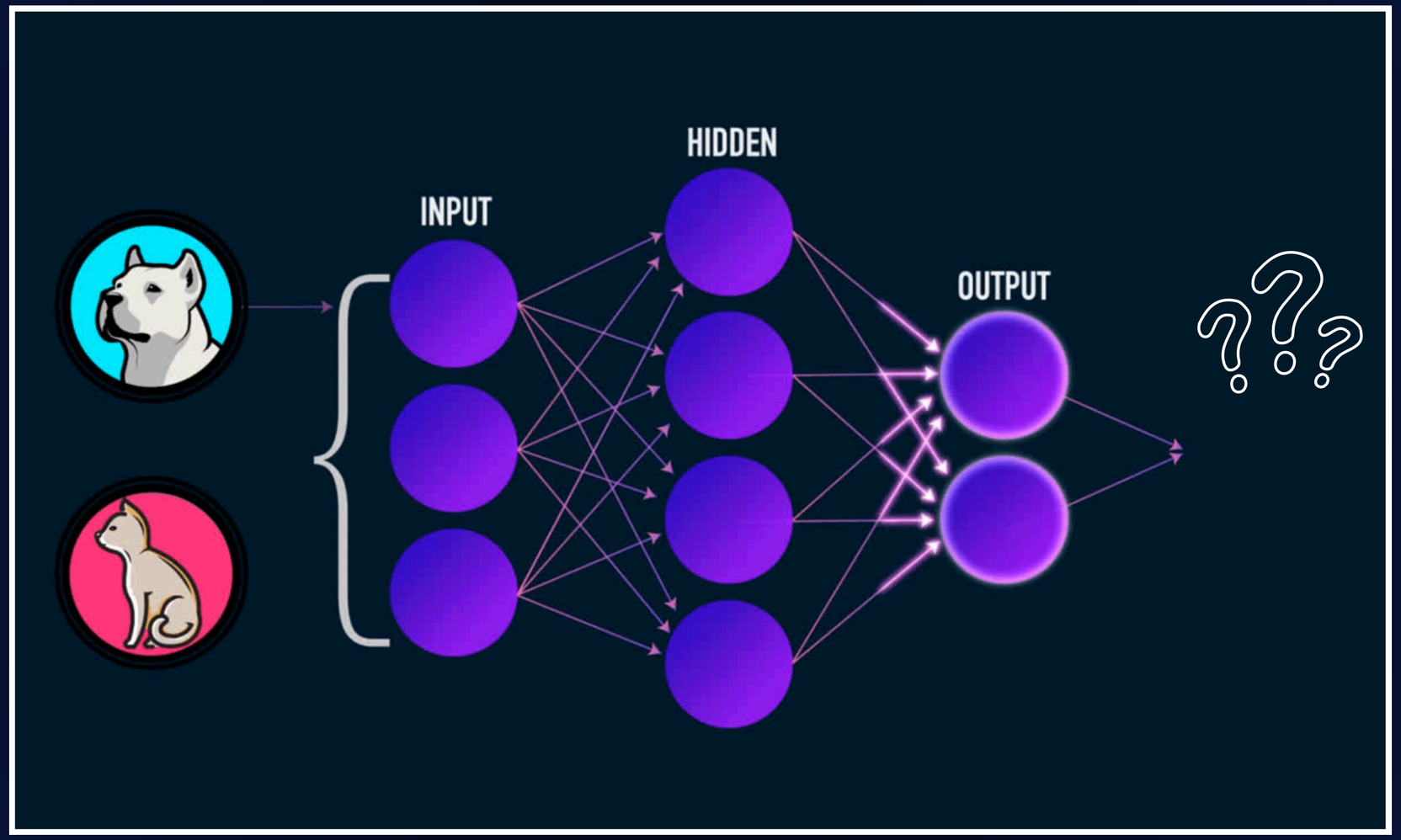


INTERPRETABILIDADE E EXPLICABILIDADE DE REDES NEURAIS

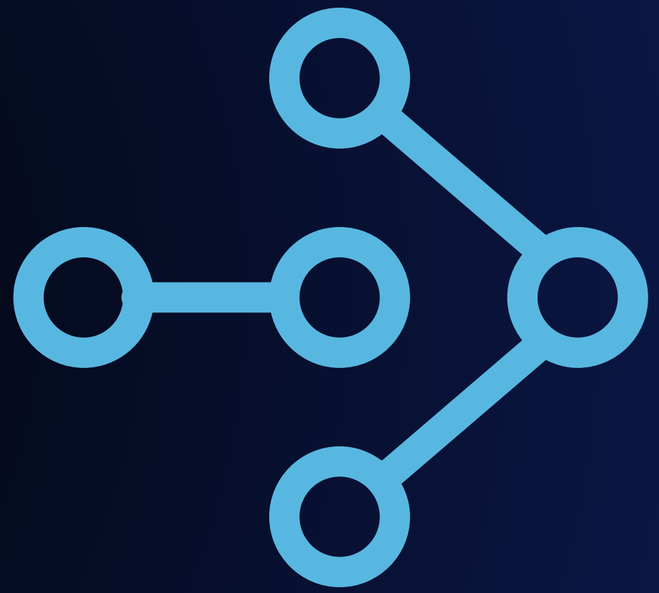
Ana Luz, Caio Matheus, Rafael Anis

Lumi Talks | Ilum - Escola de Ciência

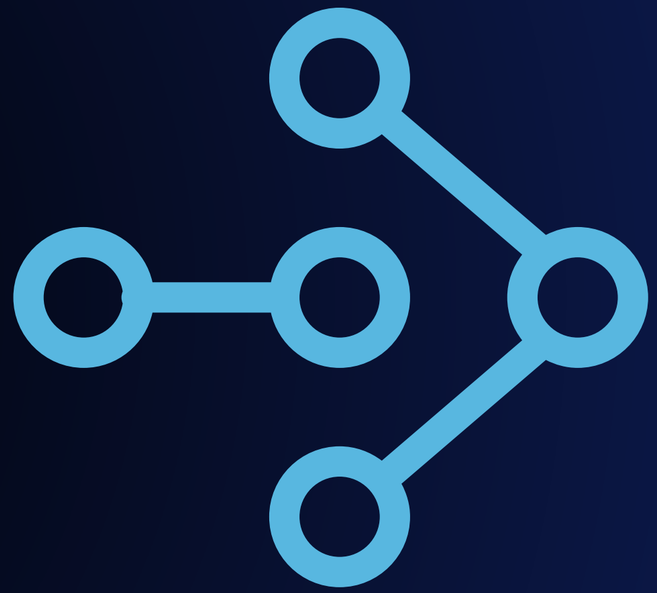



INPUT
HIDDEN
OUTPUT
???

XAI

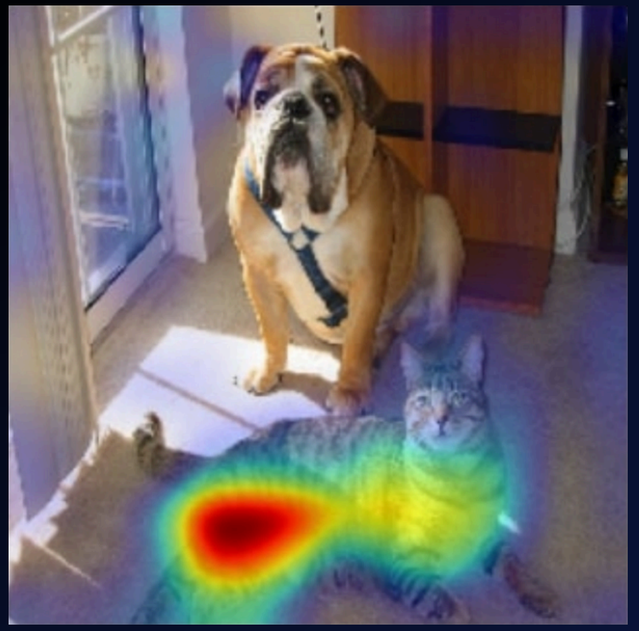


GRAD-CAM

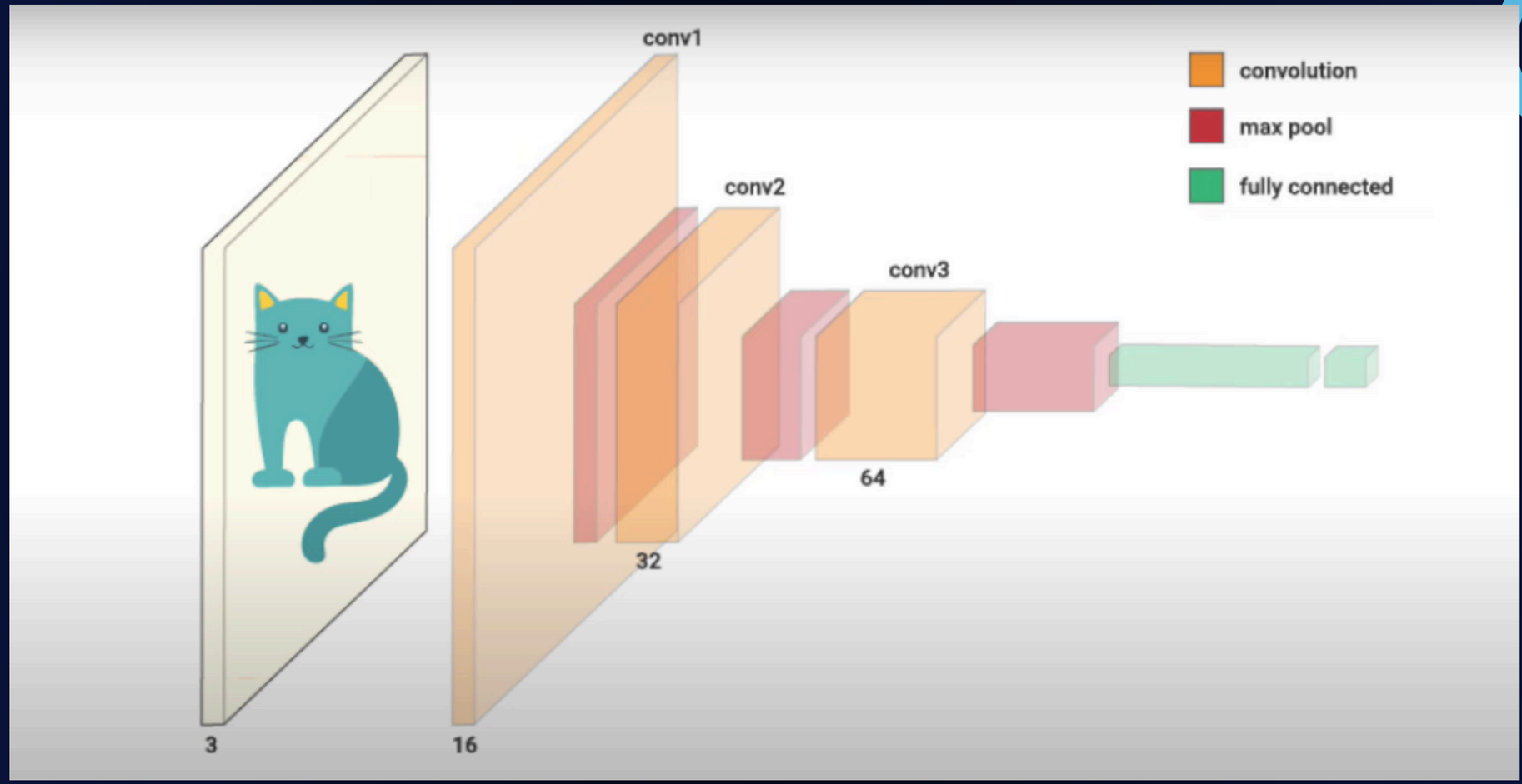


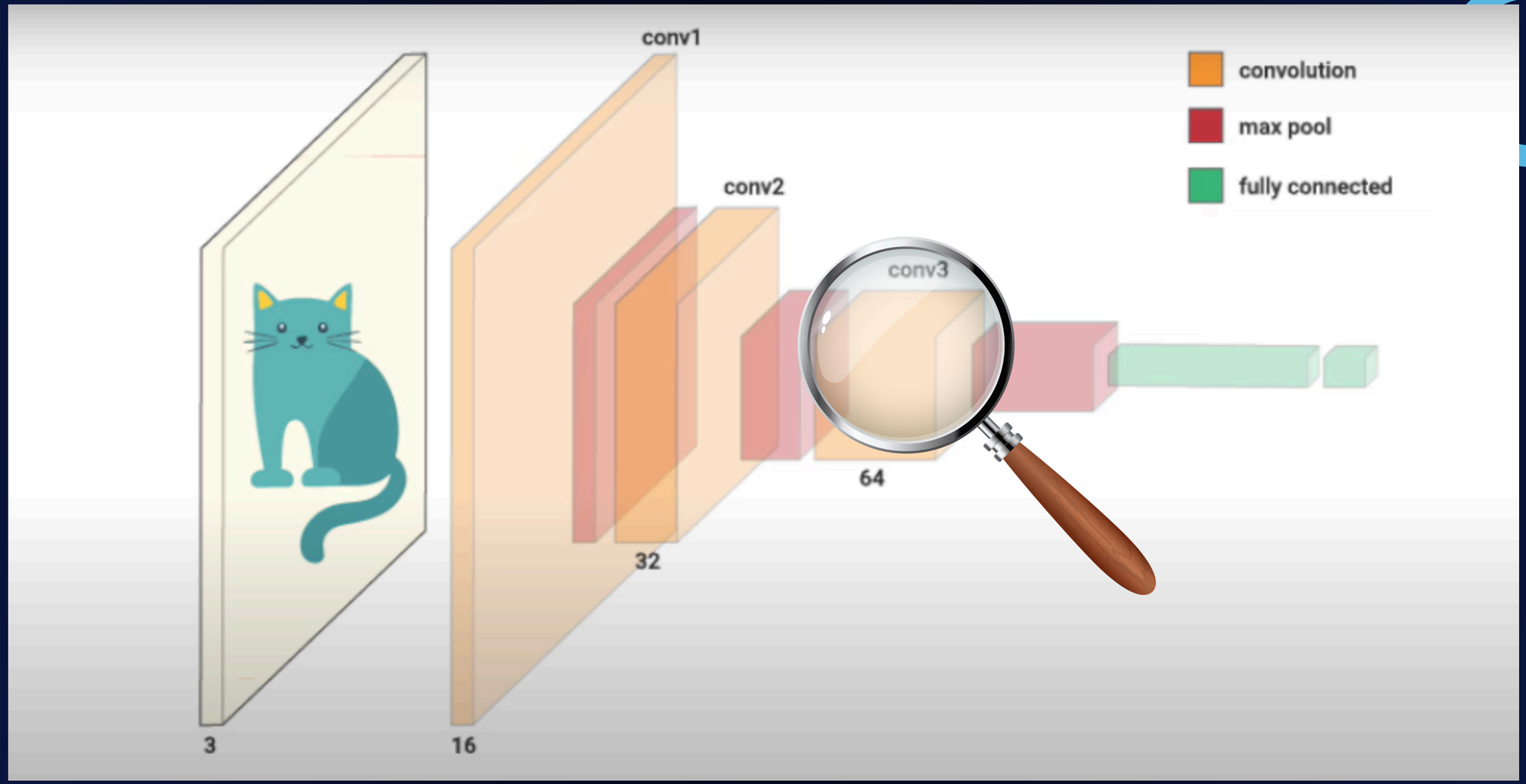
CACHORRO

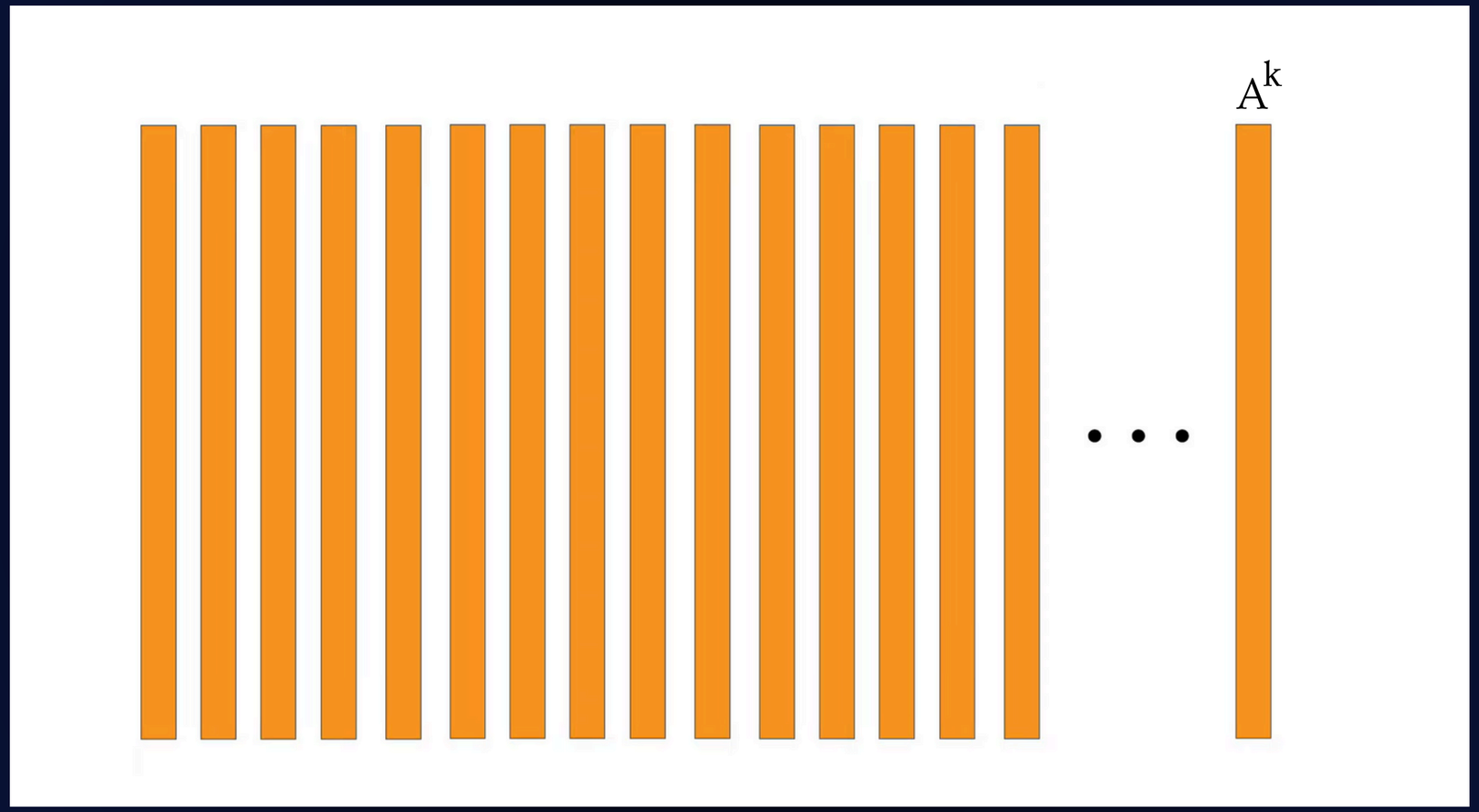
GATO

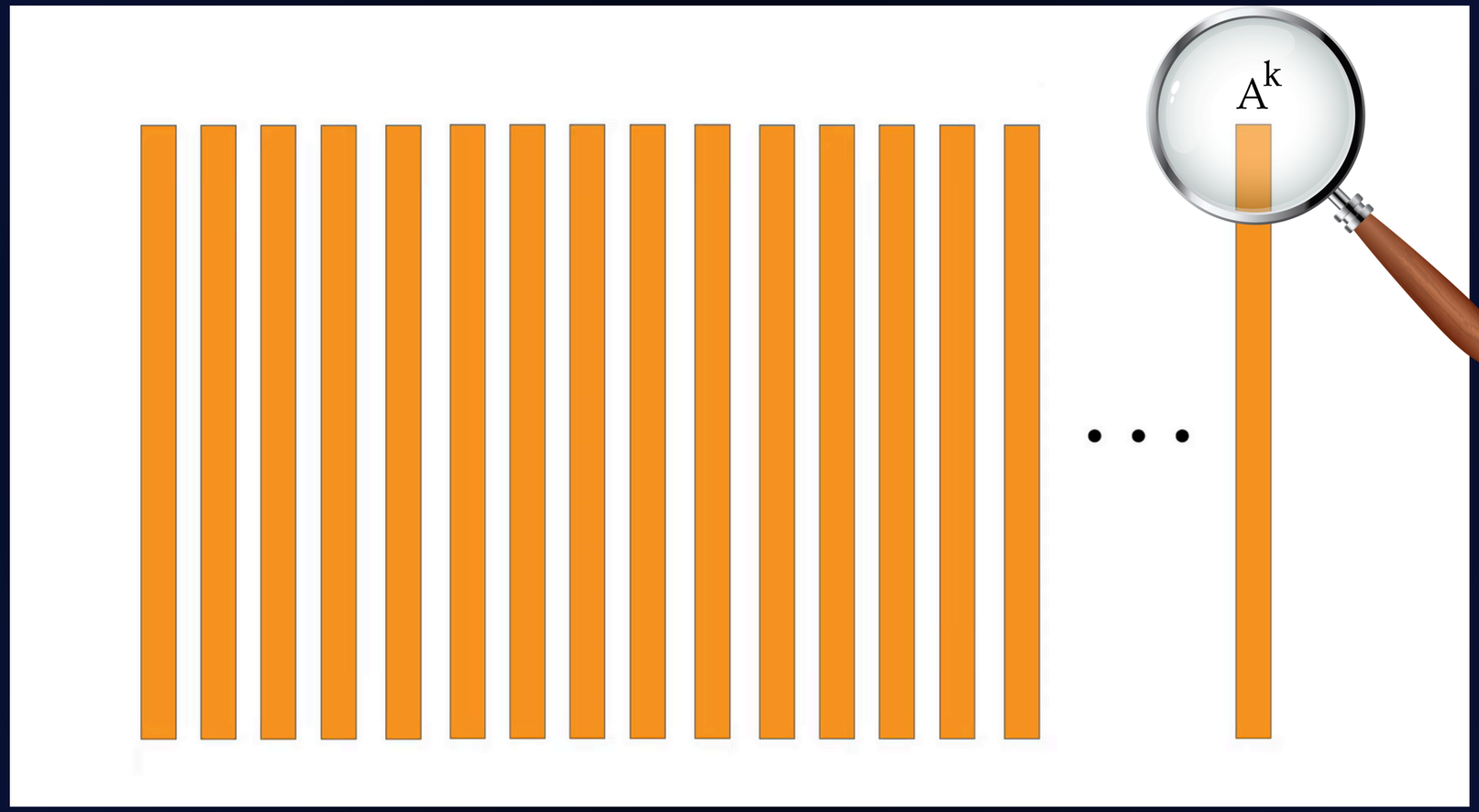


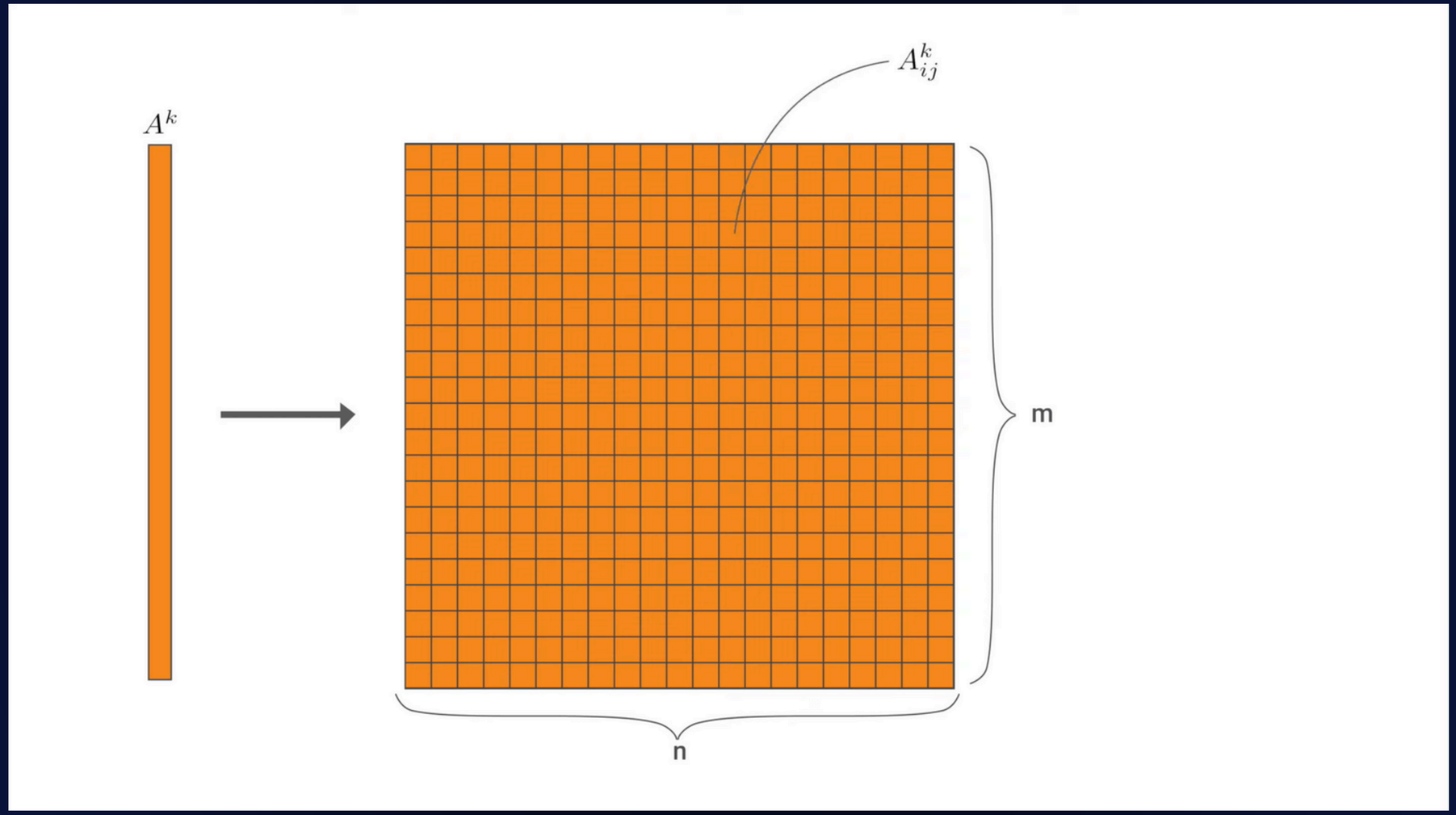
CNN CLASSIFICADORA



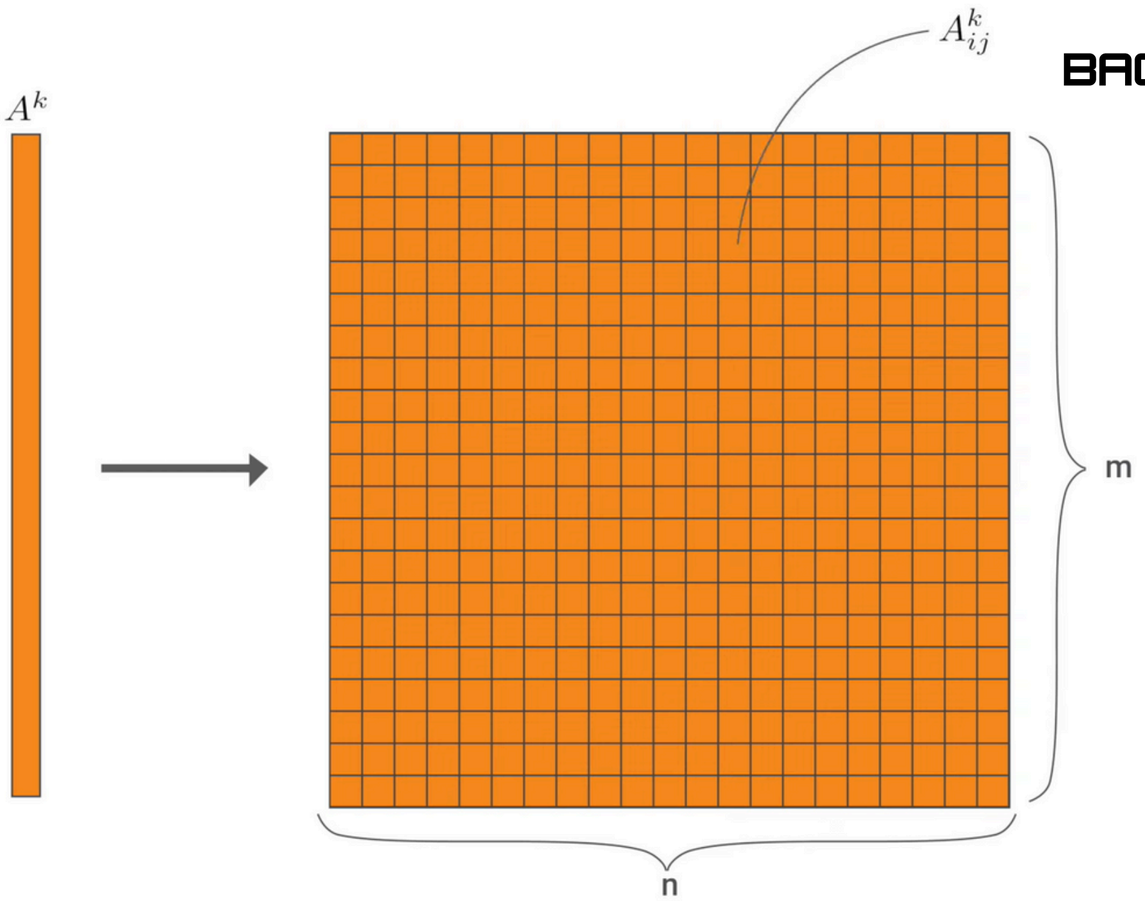

conv1
convolution
max pool
fully connected
conv2
conv3
64
32
3
16


A^k

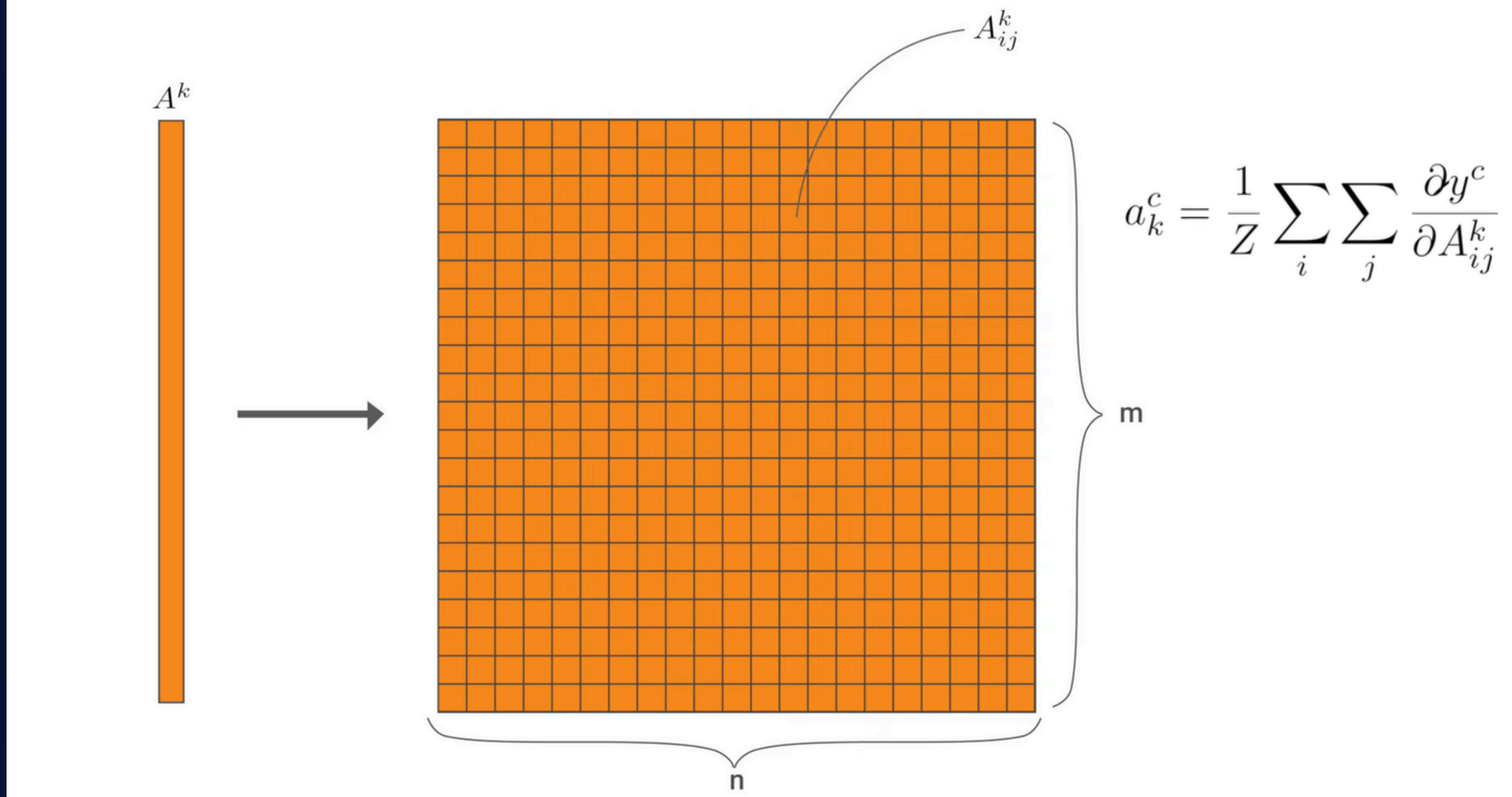

A^k


A^k
A^k_{ij}
m
n

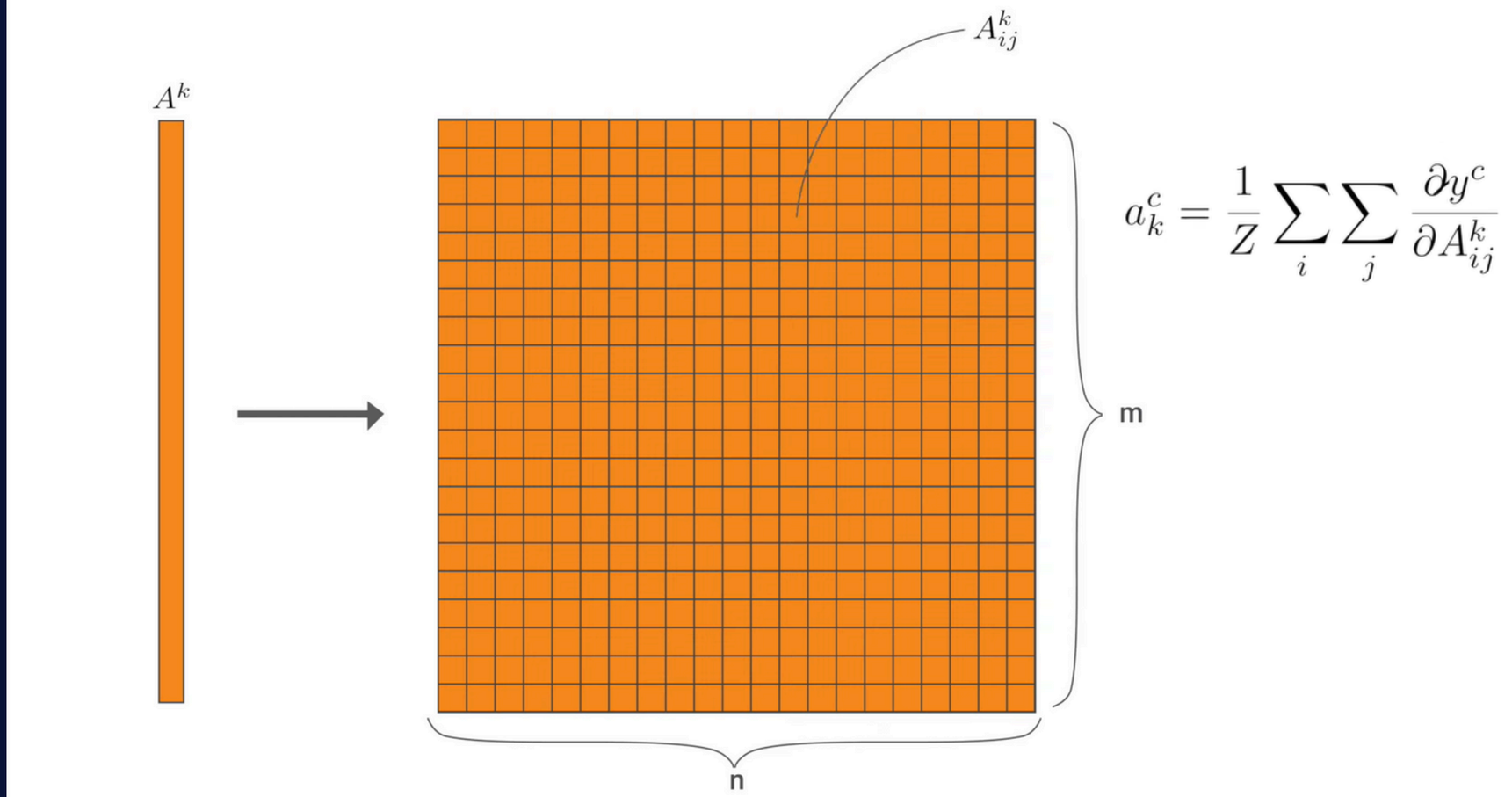
BACKPROPAGATION



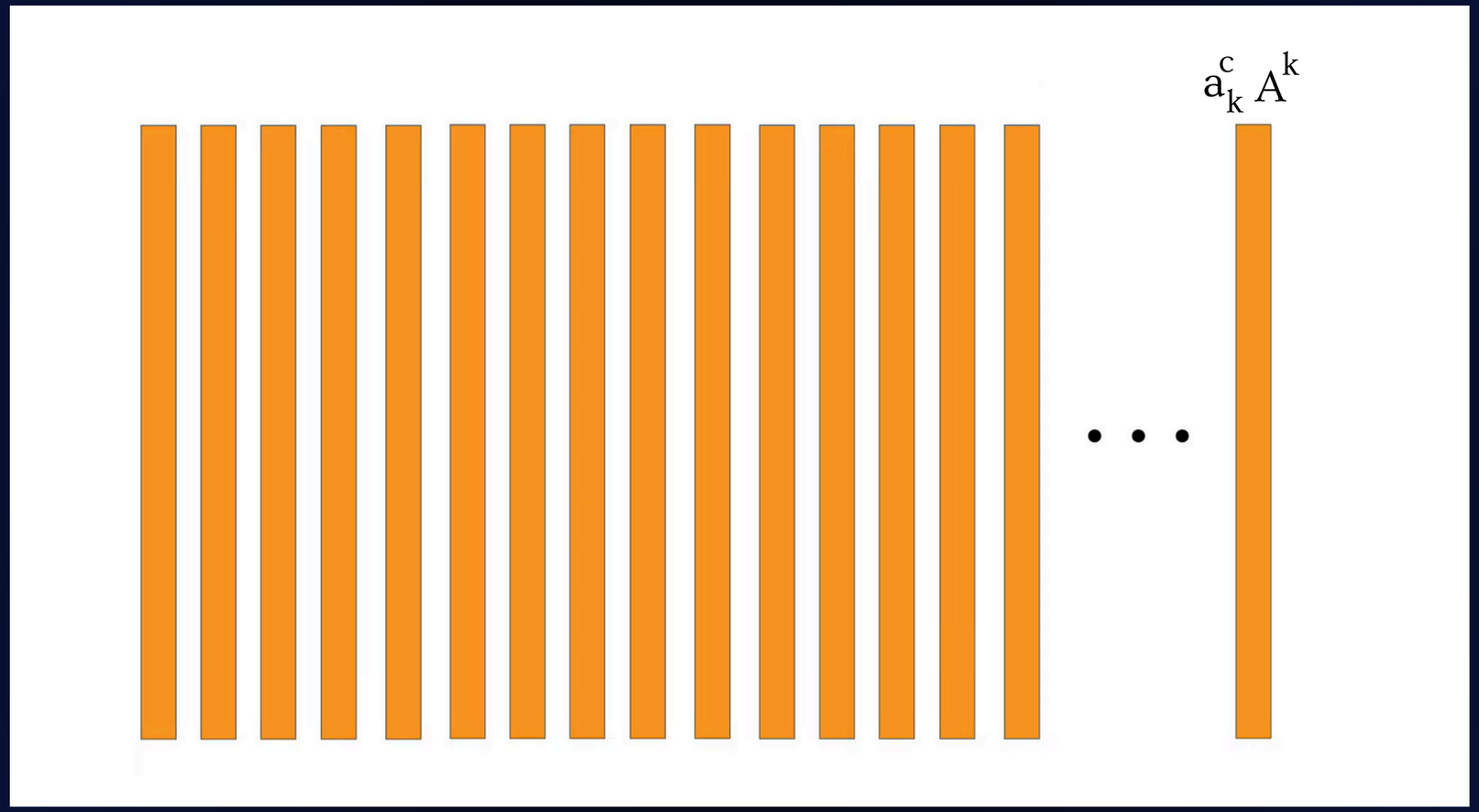
$$\frac{\partial y^c}{\partial A^k_{ij}}$$



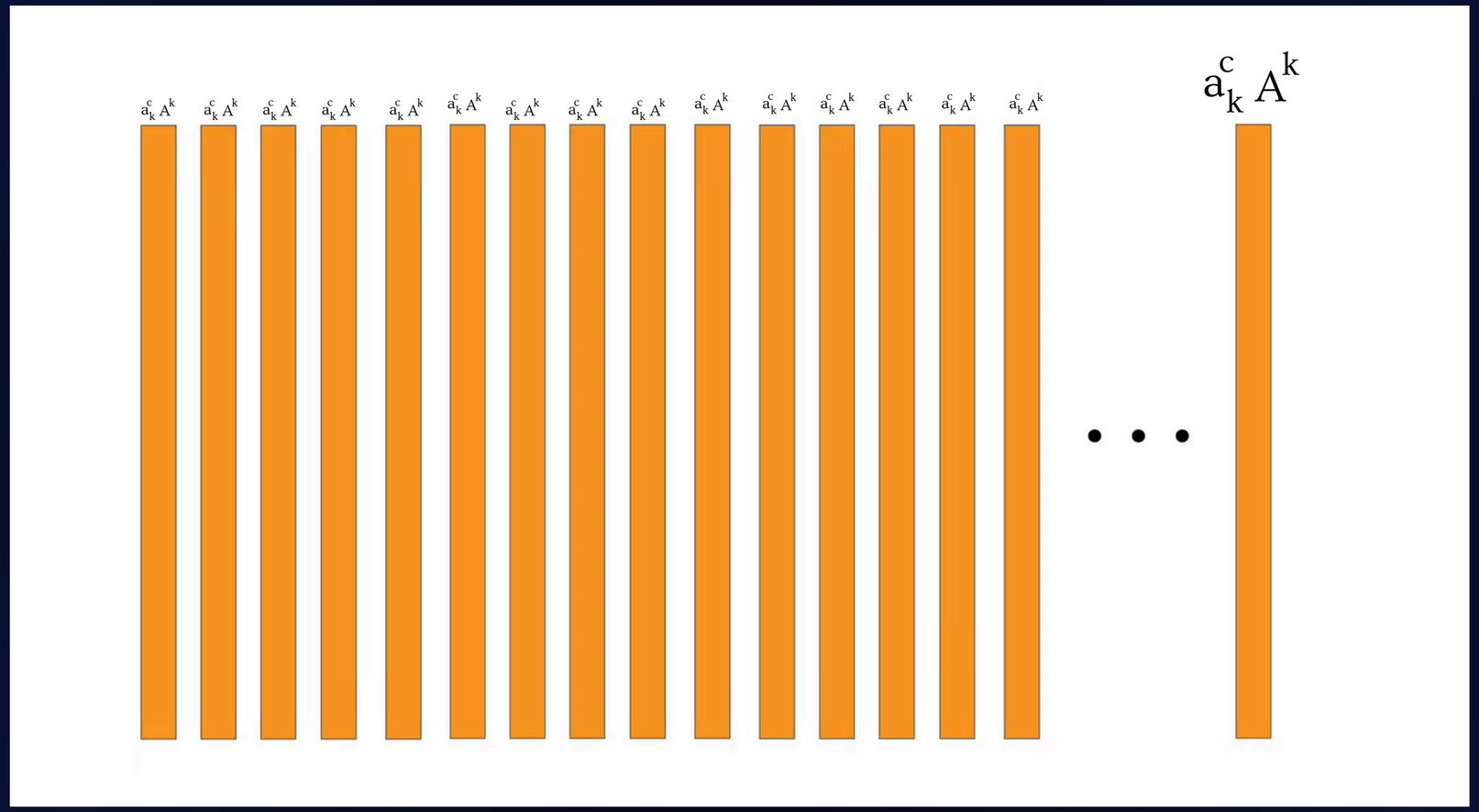
$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

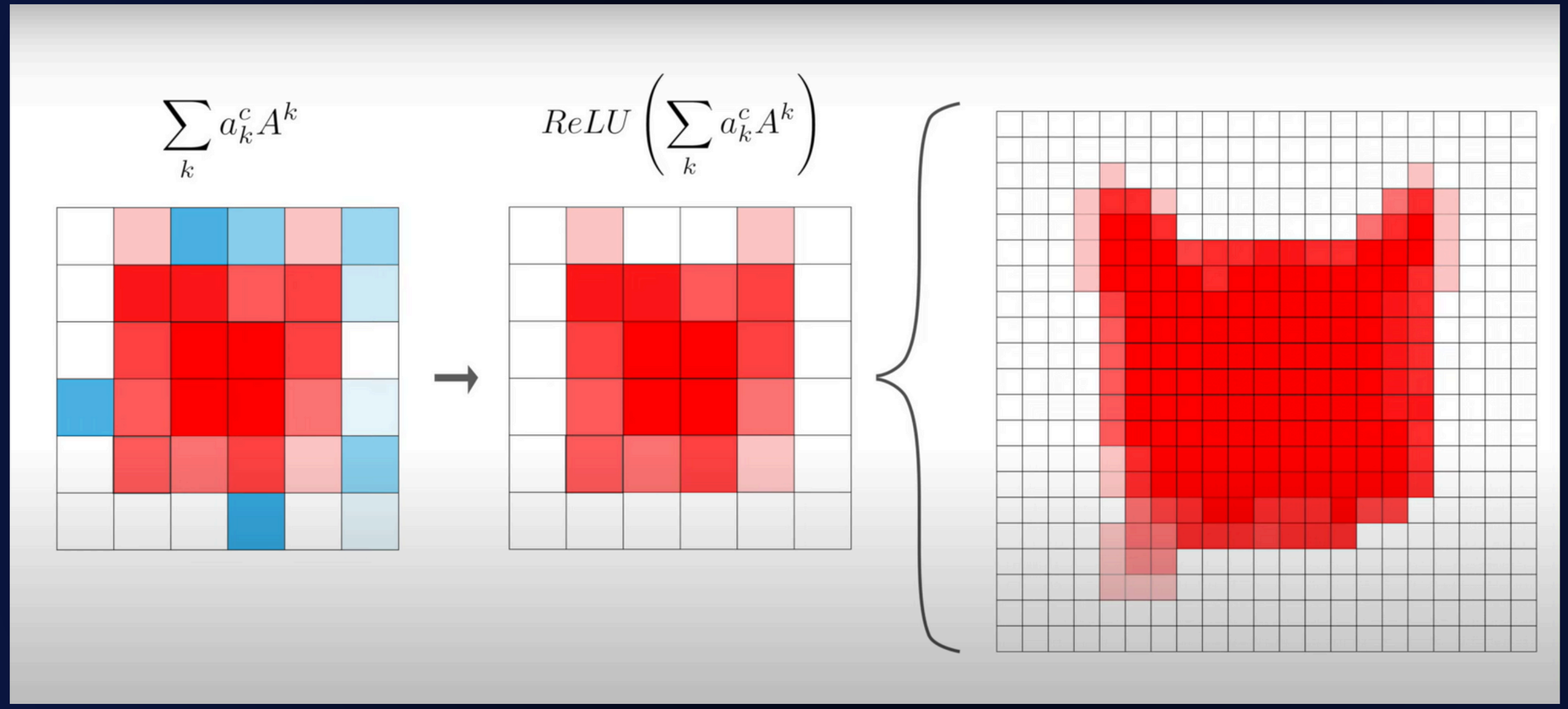


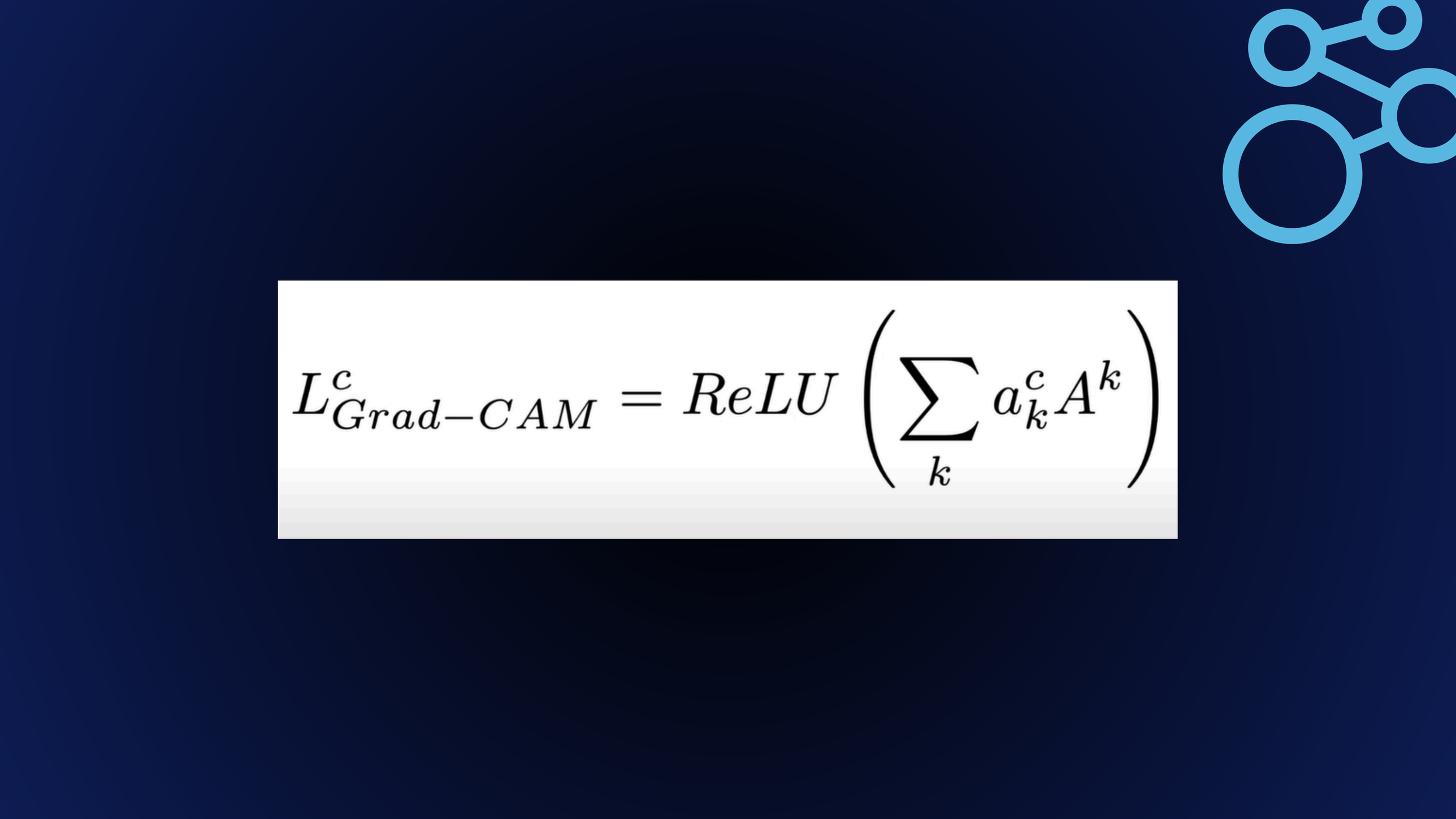
$$a_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$


a_k^c A^k

$$\left(\sum_k a_k^c A^k \right)$$




$\sum_k a_k^c A^k$
$ReLU\left(\sum_k a_k^c A^k\right)$

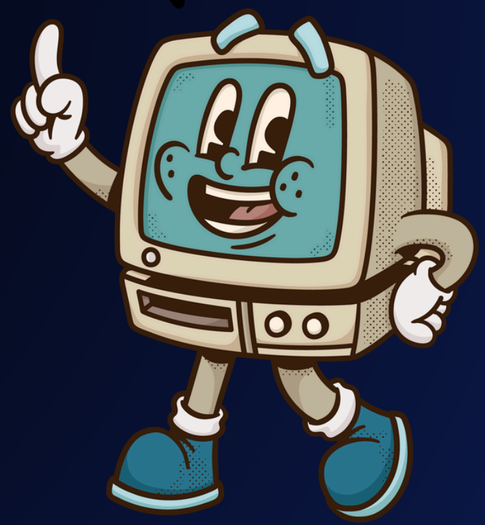

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} a^{c}_{k} A^{k}\right)$$

$$\frac{1}{Z}\sum_{i}\sum_{j}\frac{\partial y^{c}}{\partial A_{ij}^{k}}$$

$$L_{Grad-CAM}^{c} = ReLU\left(\sum_{k} a_{k}^{c} A^{k}\right)$$

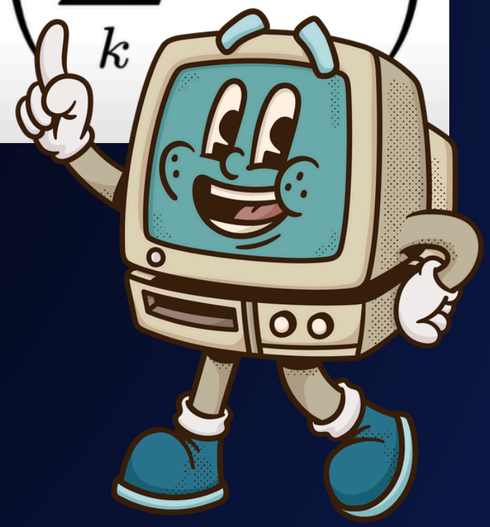
QUEM SOU EU?

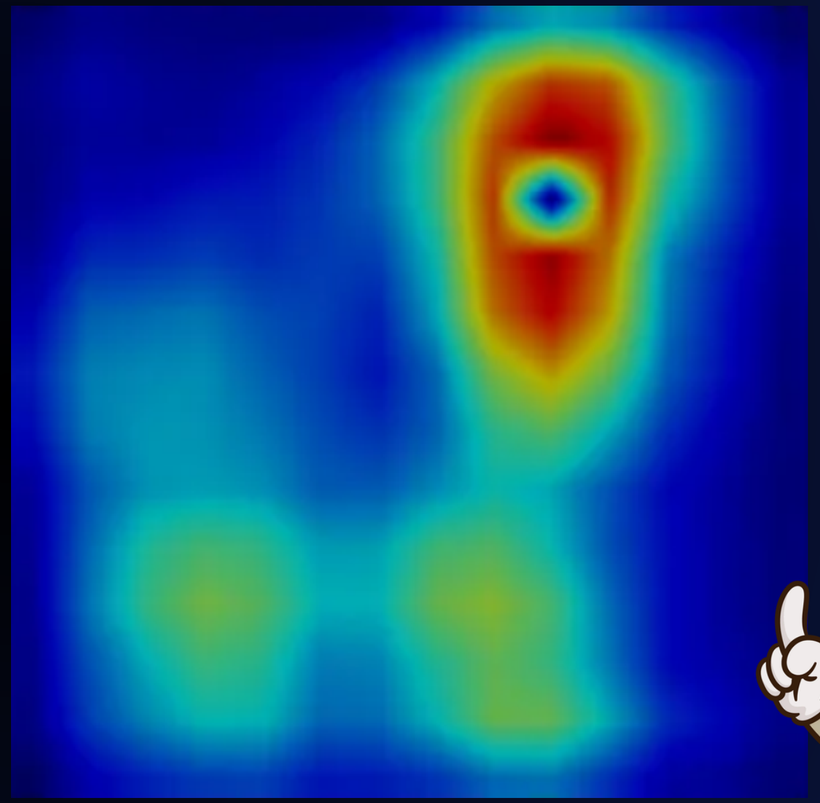
LEÃO



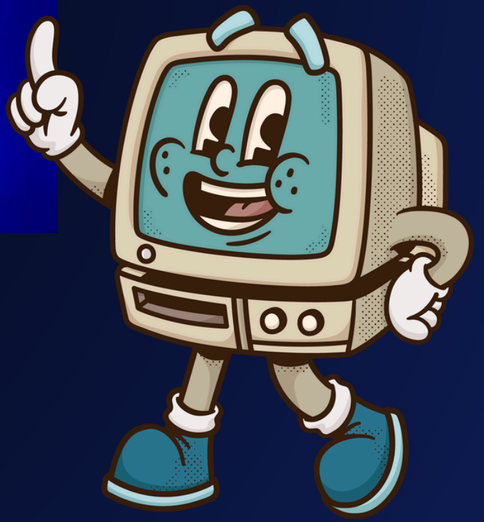
POR QUÊ?

$$L^{c}_{Grad-CAM} = ReLU\left(\sum_{k} a^{c}_{k} A^{k}\right)$$

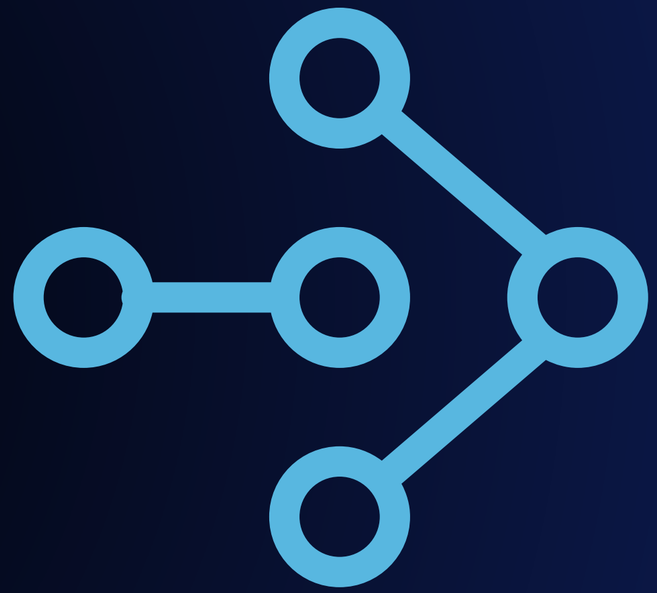




AQUI OH



LIME



Como obtemos essa explicação

LOCAL

A explicação é simples o suficiente para um humano entender

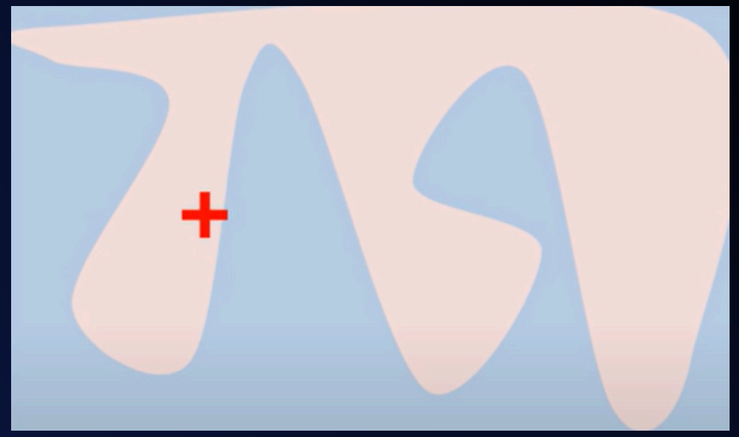
INTERPRETABLE

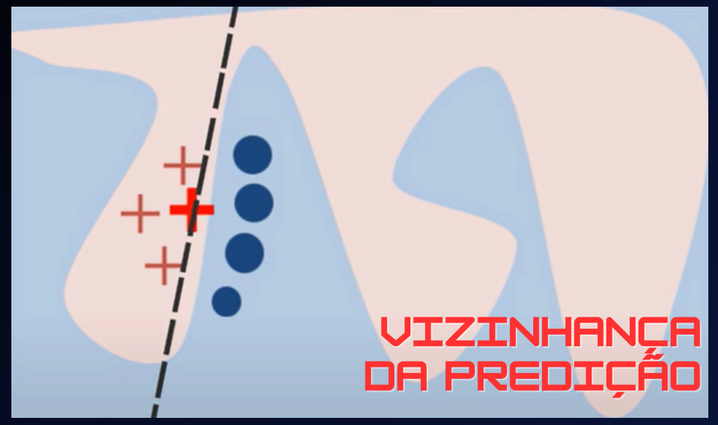
Agnóstico significa que ele pode ser aplicado para qualquer modelo de ML

MODEL-AGNOSTIC

Explicável

EXPLANATIONS





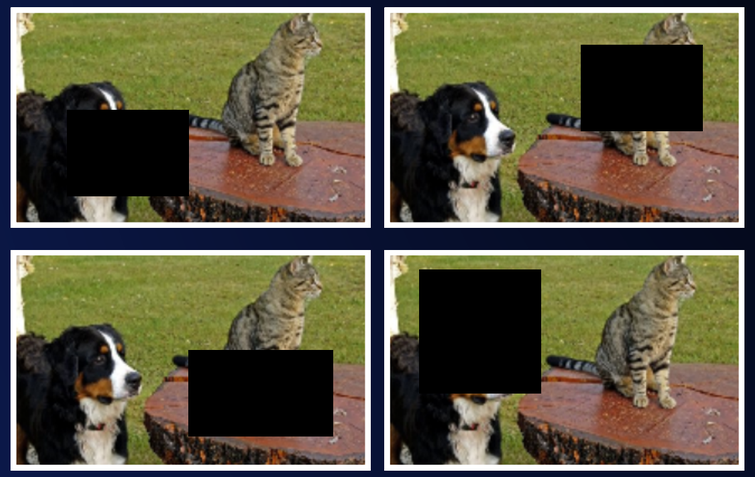
LIME

PREDIÇÃO ESPECÍFICA

BLACK BOX

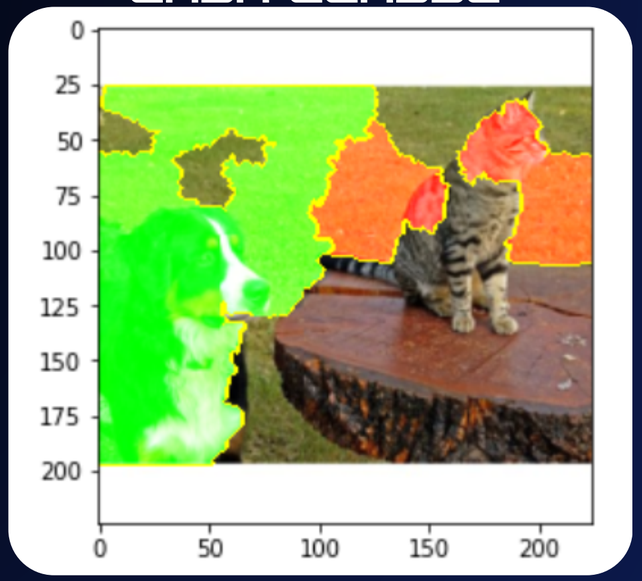
CAT

DADOS PERTURBADOS

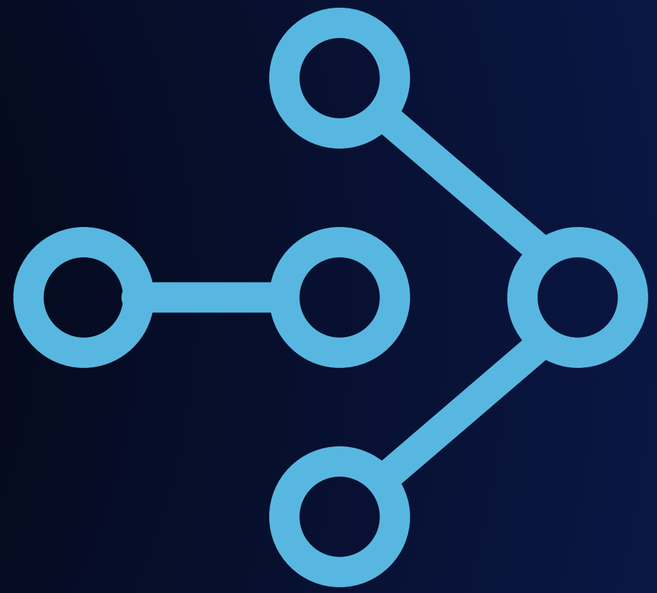


BLACK BOX

PROBABILIDADE PARA CADA CLASSE



SHAP



SHAPLEY ADDITIVE EXPLANATIONS

VALORES DE SHAPLEY

Mede a contribuição média marginal de cada feature do modelo, analisando todas as ordens possíveis de entradas.

INTERPRETABILIDADE

A interpretabilidade exata do Shap acontece de forma que a soma dos valores SHAPs preditos é igual a saída real do modelo.

FÓRMULA DO VALOR DE SHAPLEY PARA UMA FEATURE

$$\phi_i(p) = \sum_{S \subseteq N/i} \frac{|S|!(n - |S| - 1)!}{n!} (p(S \cup i) - p(S))$$

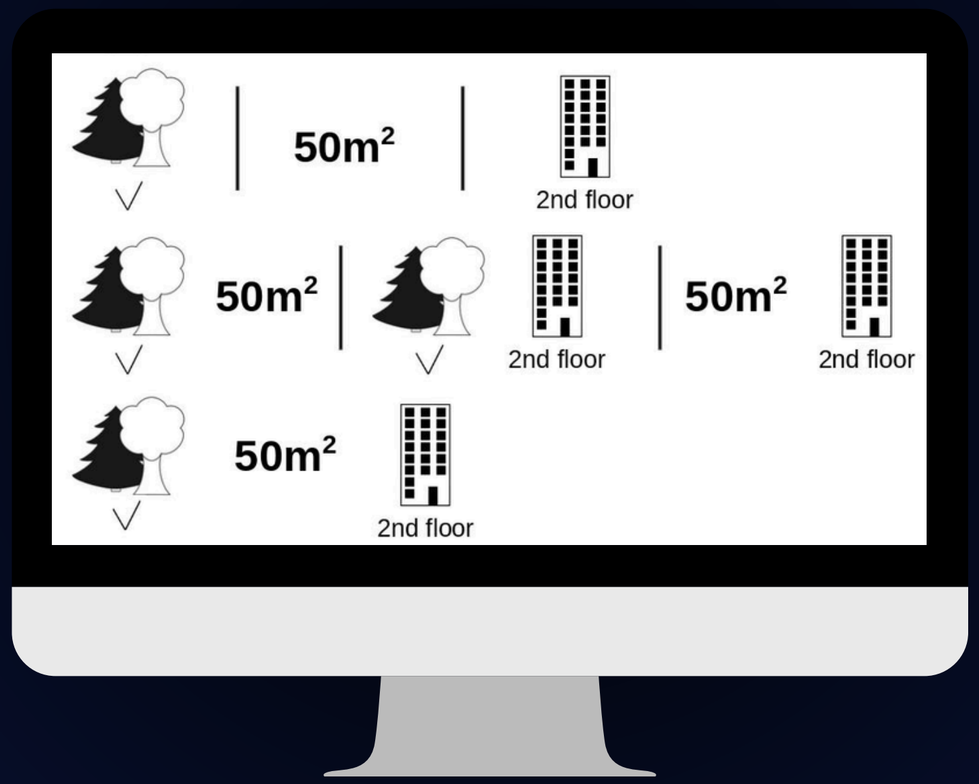
R$ 310.000,00

~~R$ 300.000,00~~

R$ 10.000,00

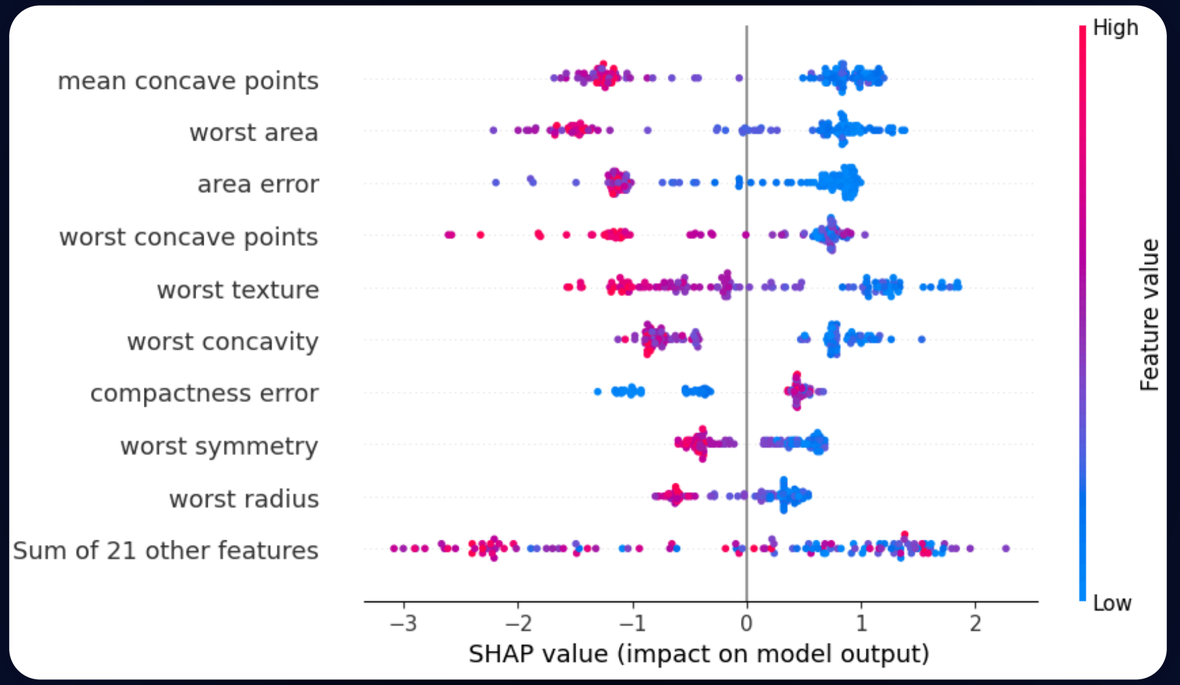
R$ 310.000,00

R$ 320.000,00


50m²
2nd floor
50m²
2nd floor
50m²
2nd floor
50m²
2nd floor

BEESWARM PLOT

Mostra a relação do valor da feature com a sua influência no modelo e se ela tende a empurrar a previsão para valores mais altos ou baixos.

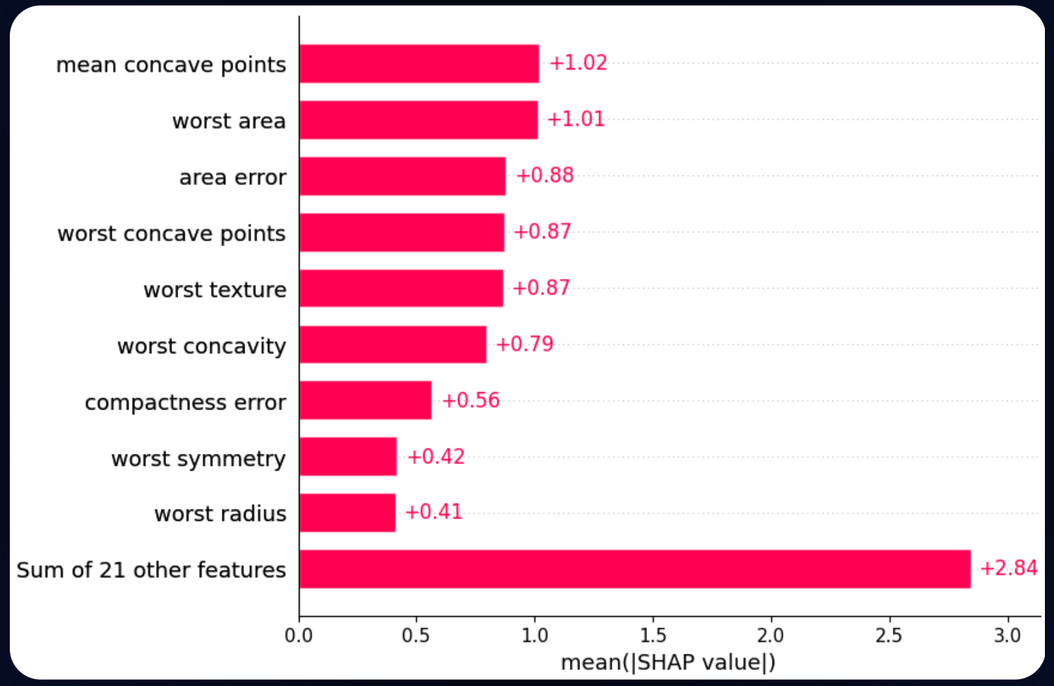


BEESWARM PLOT

Mostra a relação do valor da feature com a sua influência no modelo e se ela tende a empurrar a previsão para valores mais altos ou baixos.

BAR PLOT

Mostra quais features são mais importantes para o modelo na média em seu valor absoluto

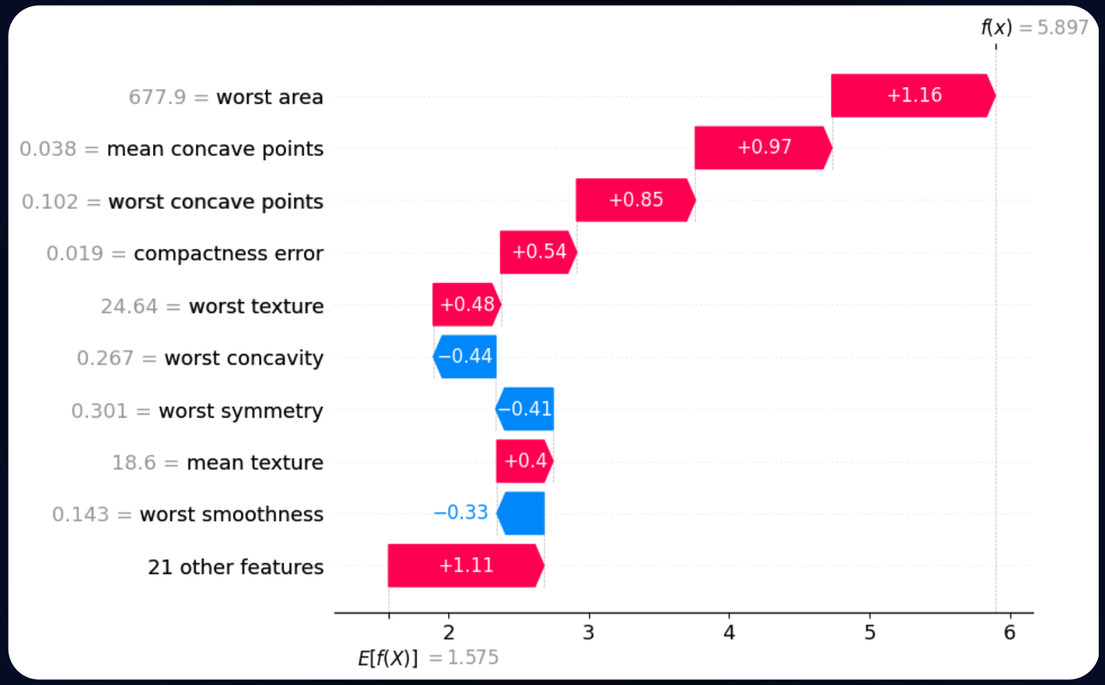


BEESWARM PLOT

Mostra a relação do valor da feature com a sua influência no modelo e se ela tende a empurrar a previsão para valores mais altos ou baixos.

BAR PLOT

Mostra quais features são mais importantes para o modelo na média em seu valor absoluto



WATERFALL PLOT

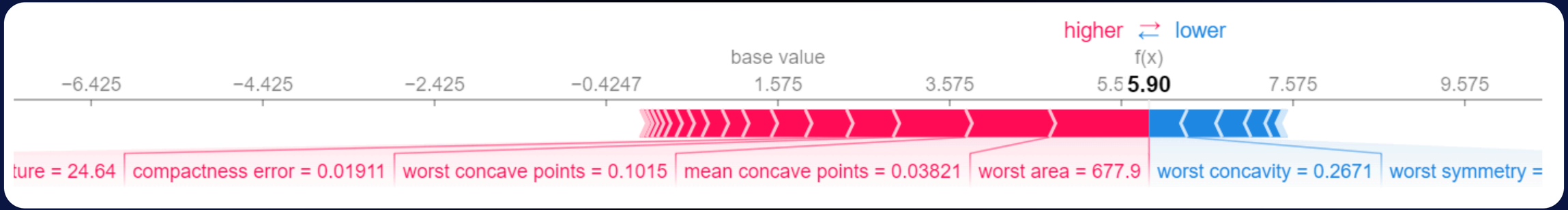
Mostra quais features ajudaram ou prejudicaram a saída e como que cada uma impactou, chegando assim em um resultado final

BEESWARM PLOT

Mostra a relação do valor da feature com a sua influência no modelo e se ela tende a empurrar a previsão para valores mais altos ou baixos.

BAR PLOT

Mostra quais features são mais importantes para o modelo na média em seu valor absoluto

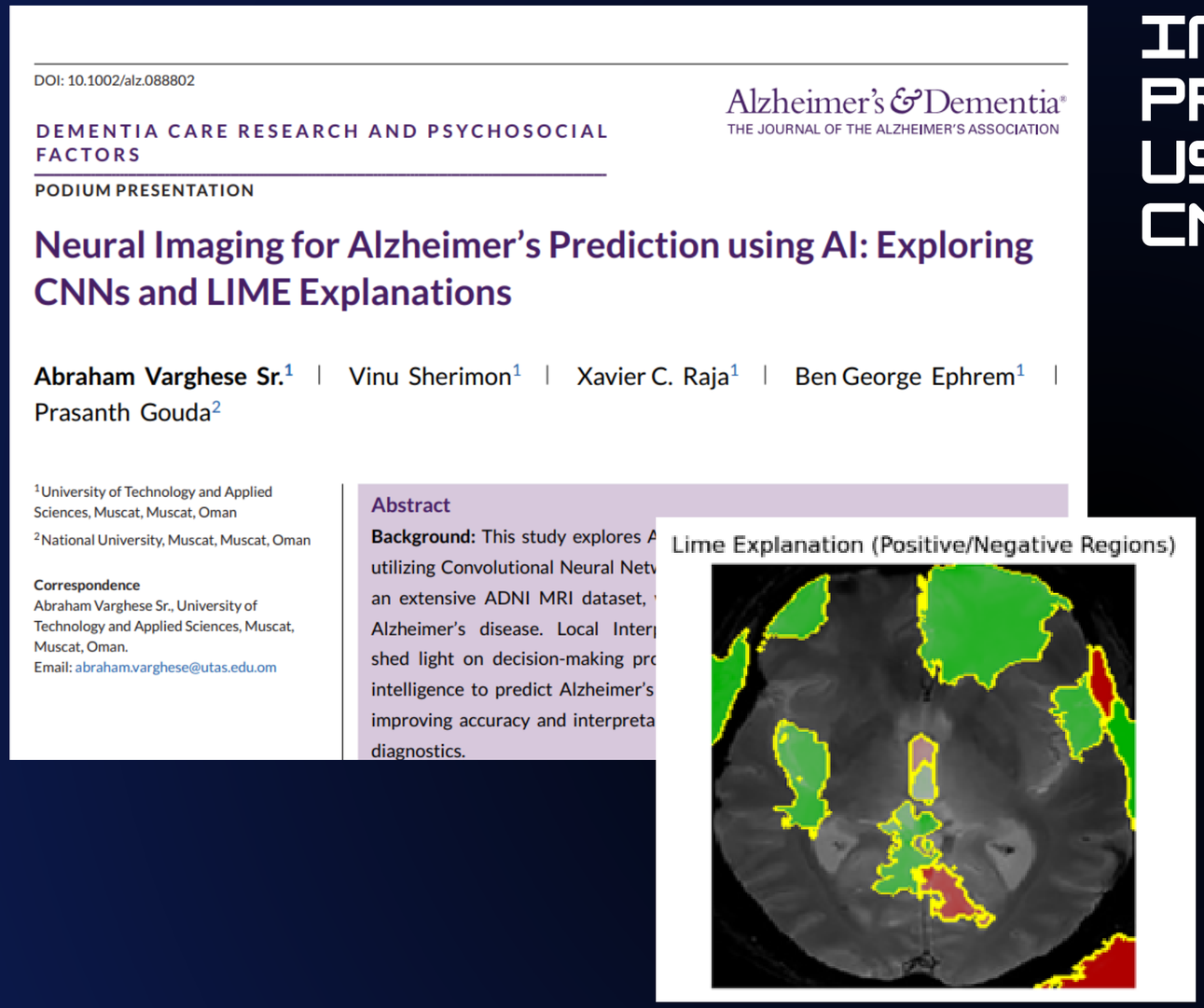


FORCE PLOT

Atua mostrando como uma “balança de forças”, indicando quais dos features aumentam a predição e quais diminuem, de forma que a soma resulta no valor previsto.

WATERFALL PLOT

Mostra quais features ajudaram ou prejudicaram a saída e como que cada uma impactou, chegando assim em um resultado final



DOI: 10.1002/alz.088802

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

DEMENTIA CARE RESEARCH AND PSYCHOSOCIAL FACTORS

PODIUM PRESENTATION

Neural Imaging for Alzheimer's Prediction using AI: Exploring CNNs and LIME Explanations

Abraham Varghese Sr.[1] | Vinu Sherimon[1] | Xavier C. Raja[1] | Ben George Ephrem[1] | Prasanth Gouda[2]

[1]University of Technology and Applied Sciences, Muscat, Muscat, Oman

[2]National University, Muscat, Muscat, Oman

Correspondence
Abraham Varghese Sr., University of Technology and Applied Sciences, Muscat, Muscat, Oman.
Email: abraham.varghese@utas.edu.om

Abstract

Background: This study explores A
utilizing Convolutional Neural Netv
an extensive ADNI MRI dataset,
Alzheimer's disease. Local Interp
shed light on decision-making pro
intelligence to predict Alzheimer's
improving accuracy and interpreta
diagnostics.

IMAGEM NEURAL PARA PREDIÇÃO DO ALZHEIMER USANDÓ IA: EXPLORANDO CNNS E EXPLICAÇÕES LIME

OBJETIVO

O estudo visa prever a doença de Alzheimer por meio de imagens de ressonância magnética cerebral, utilizando redes neurais convolucionais (CNNs) para classificar diferentes estágios da doença.

MODELOS TESTADOS

LIME

CONCLUSÃO

o aplicar o LIME, foi possível identificar regiões cerebrais específicas que influenciaram as previsões do modelo, proporcionando uma compreensão mais clara das decisões da IA

CLASSIFICAÇÃO INTERPRETÁVEL DO CÂNCER DE MAMA USANDO CNNS EM IMAGENS MAMOGRÁFICAS

OBJETIVO

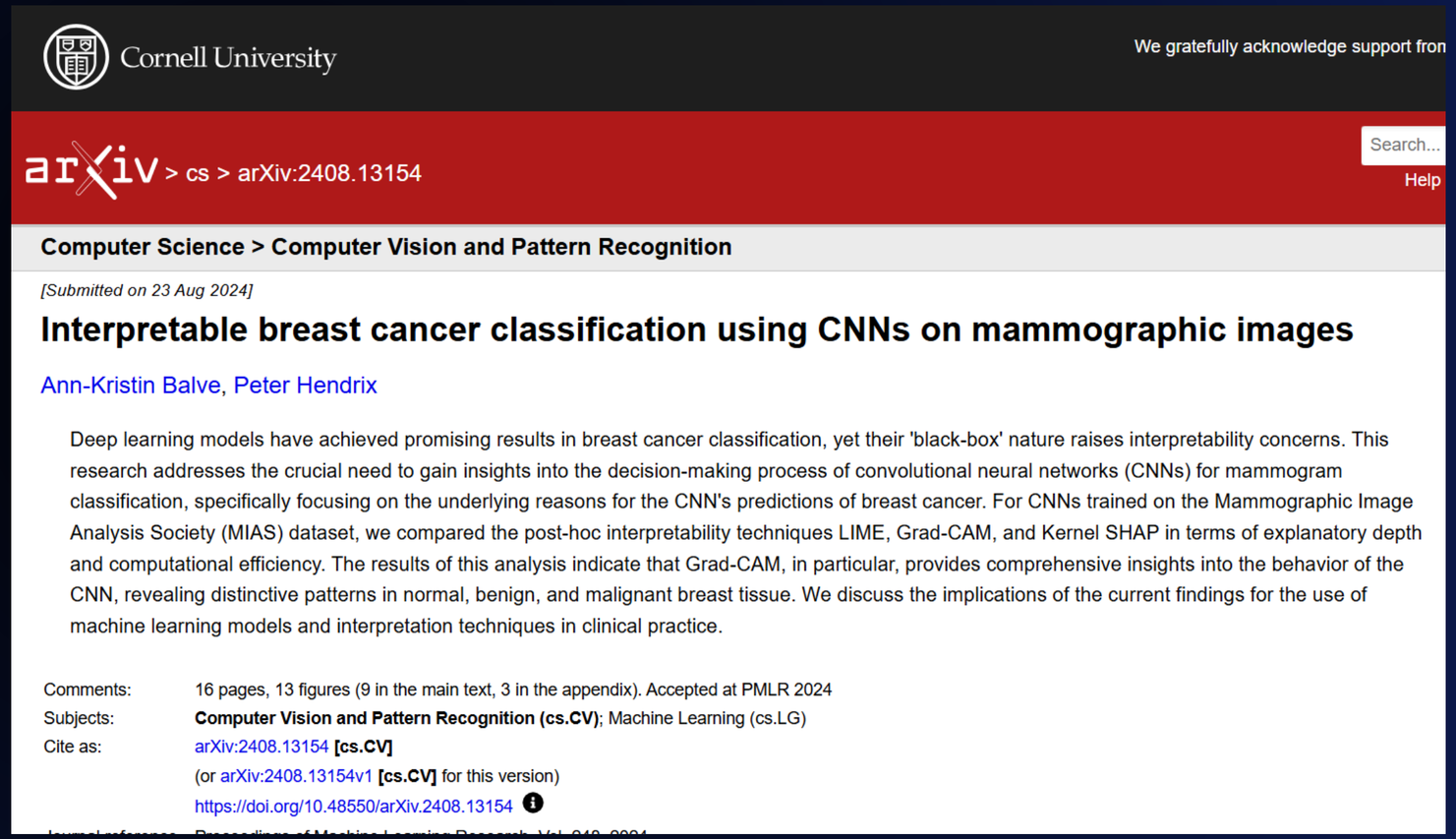
Avaliar e comparar técnicas de interpretabilidade aplicadas a redes neurais convolucionais (CNNs) treinadas para classificar imagens de mamografias em normal, benigno ou maligno.

MODELOS TESTADOS

Grad-CAM, LIME, Kernel SHAP

CONCLUSÃO

- Grad-CAM: forneceu explicações visuais mais coerentes com as regiões relevantes da imagem.
- LIME e Kernel SHAP: também ofereceram insights, mas com limitações em granularidade espacial ou eficiência



Cornell University

We gratefully acknowledge support fron

arXiv > cs > arXiv:2408.13154

Search...

Help

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 23 Aug 2024]

Interpretable breast cancer classification using CNNs on mammographic images

Ann-Kristin Balve, Peter Hendrix

Deep learning models have achieved promising results in breast cancer classification, yet their 'black-box' nature raises interpretability concerns. This research addresses the crucial need to gain insights into the decision-making process of convolutional neural networks (CNNs) for mammogram classification, specifically focusing on the underlying reasons for the CNN's predictions of breast cancer. For CNNs trained on the Mammographic Image Analysis Society (MIAS) dataset, we compared the post-hoc interpretability techniques LIME, Grad-CAM, and Kernel SHAP in terms of explanatory depth and computational efficiency. The results of this analysis indicate that Grad-CAM, in particular, provides comprehensive insights into the behavior of the CNN, revealing distinctive patterns in normal, benign, and malignant breast tissue. We discuss the implications of the current findings for the use of machine learning models and interpretation techniques in clinical practice.

Comments: 16 pages, 13 figures (9 in the main text, 3 in the appendix). Accepted at PMLR 2024

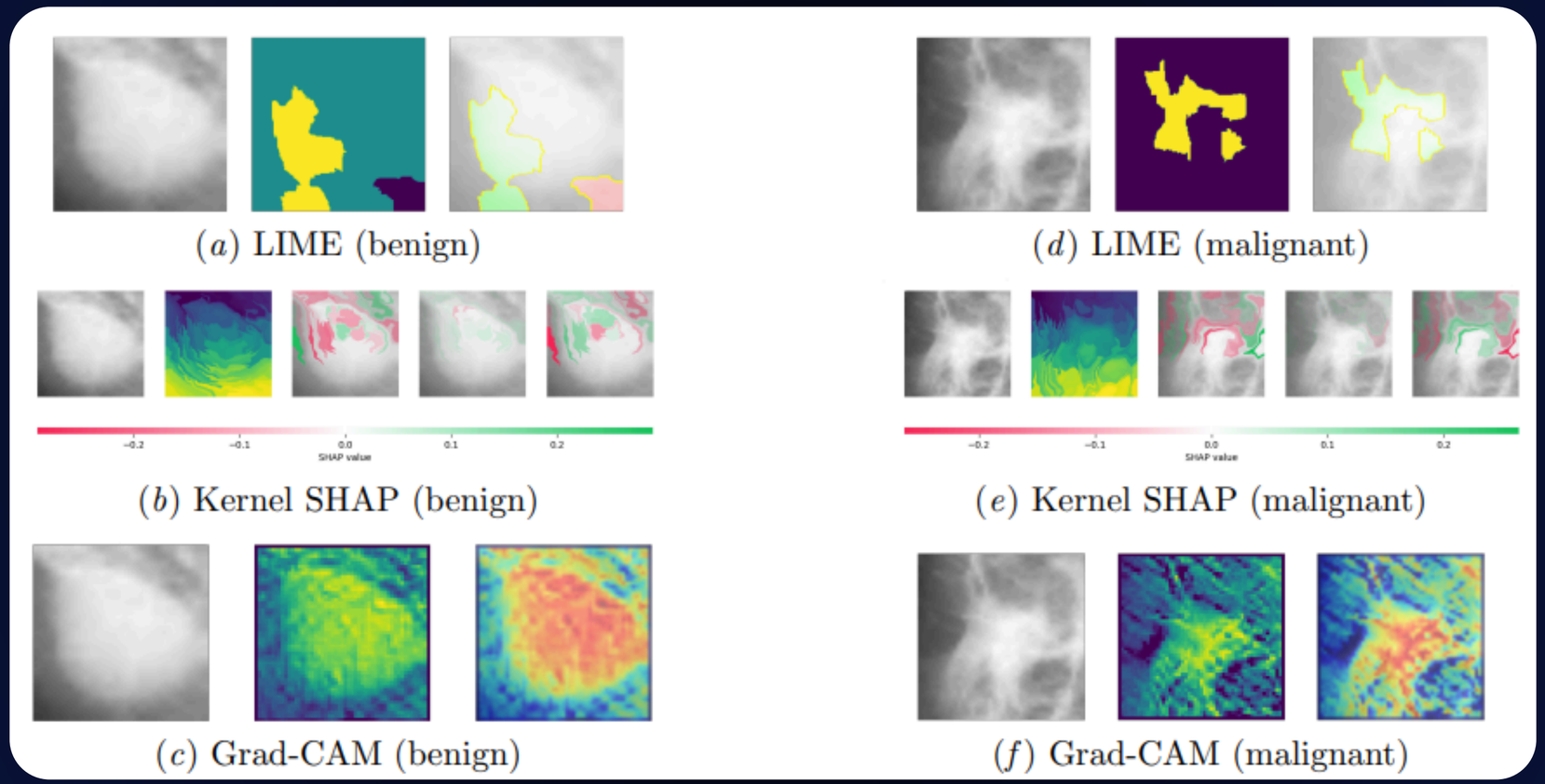
Subjects: **Computer Vision and Pattern Recognition (cs.CV)**; Machine Learning (cs.LG)

Cite as: arXiv:2408.13154 **[cs.CV]**

(or arXiv:2408.13154v1 **[cs.CV]** for this version)

https://doi.org/10.48550/arXiv.2408.13154

CLASSIFICAÇÃO INTERPRETÁVEL DO CÂNCER DE MAMA USANDO CNNS EM IMAGENS MAMOGRÁFICAS



GRAD-CAM SE DESTACOU POR FORNECER INTERPRETAÇÕES VISUAIS CLARAS E ÚTEIS NA DIFERENCIAÇÃO ENTRE PADRÕES DE TECIDOS NORMAIS, BENIGNOS E MALIGNOS.

VISÃO GERAL

	LIME	GRAD-CAM	SHAP
Tipo de modelo	Modelo-agnóstico	Específico para redes neurais convolucionais (CNNs)	Modelo-agnóstico
Tipo de dado	Tabular, texto, imagem	Imagem	Tabular, texto, imagem
Forma da explicação	Modelo linear com pesos interpretáveis	Heatmap sobreposto na imagem original	Valores de contribuição (baseados na teoria dos jogos)
Complexidade computacional	Média (gera várias amostras locais)	Baixa (uso eficiente de gradientes)	Alta (especialmente com muitas features)

VISÃO GERAL

LIME

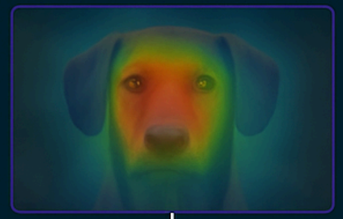
Input Image

Perturbed Samples

Prediction: Dog

Grad-CAM

Convolutional Neural netwok



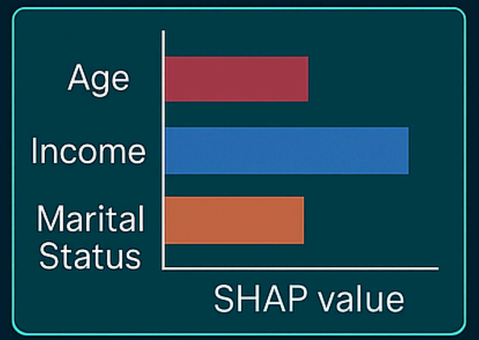
Gradients

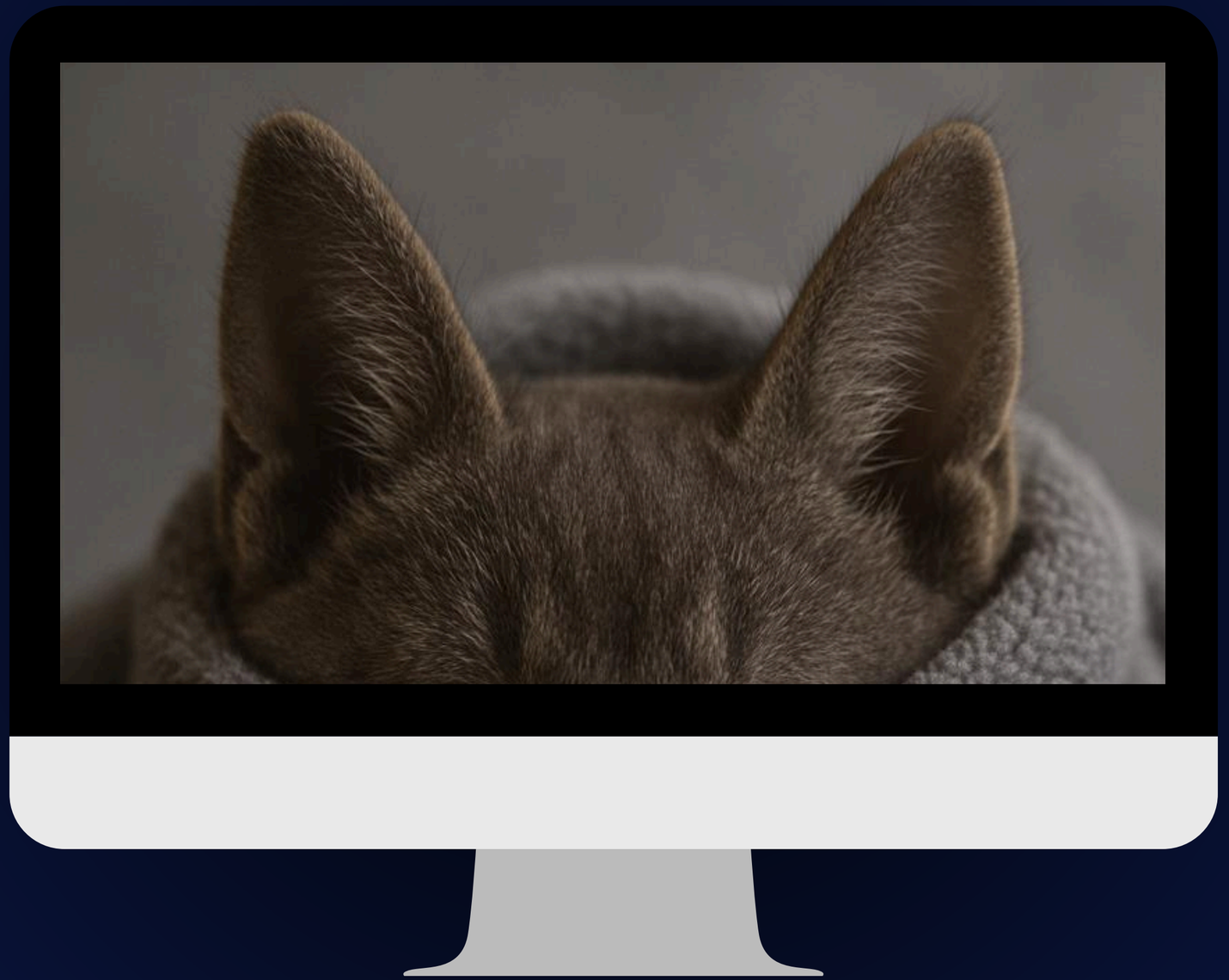
Heatmap

SHAP

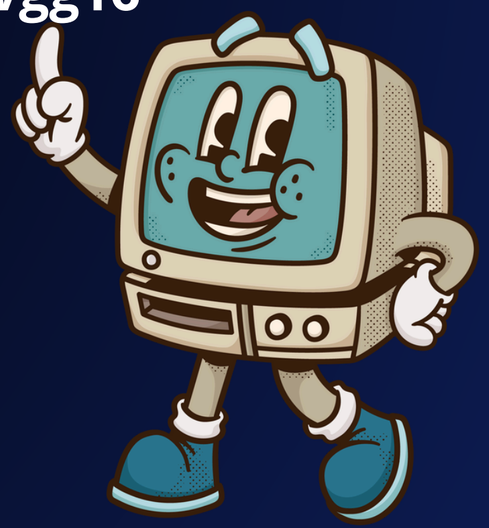
Data	Income
45	$80.000
25	$20.000

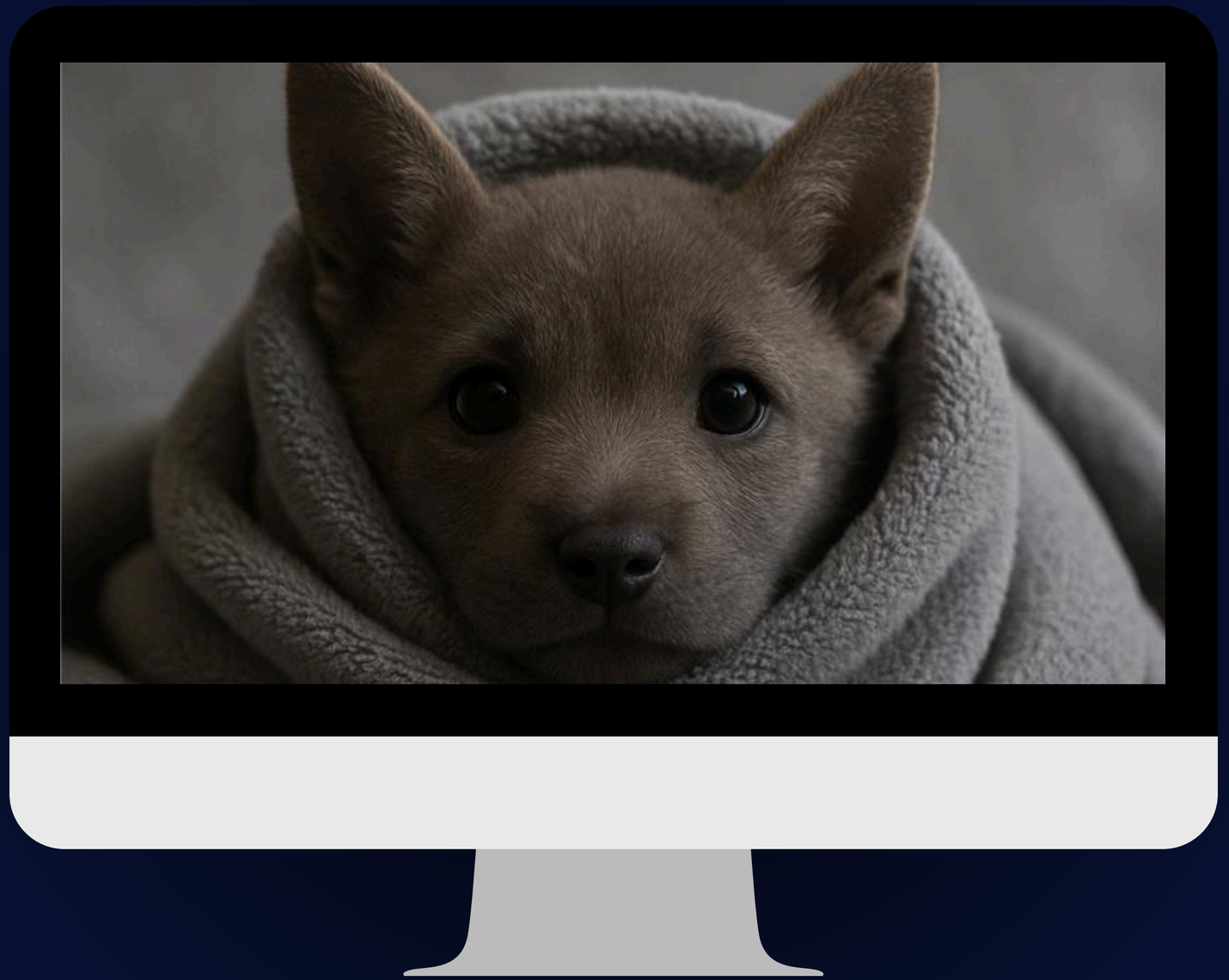
Explanation





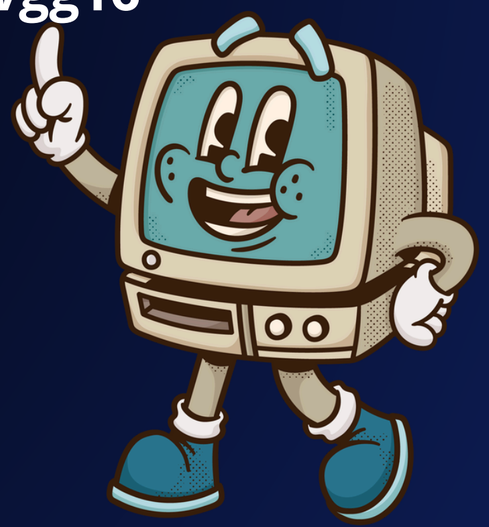
vgg16



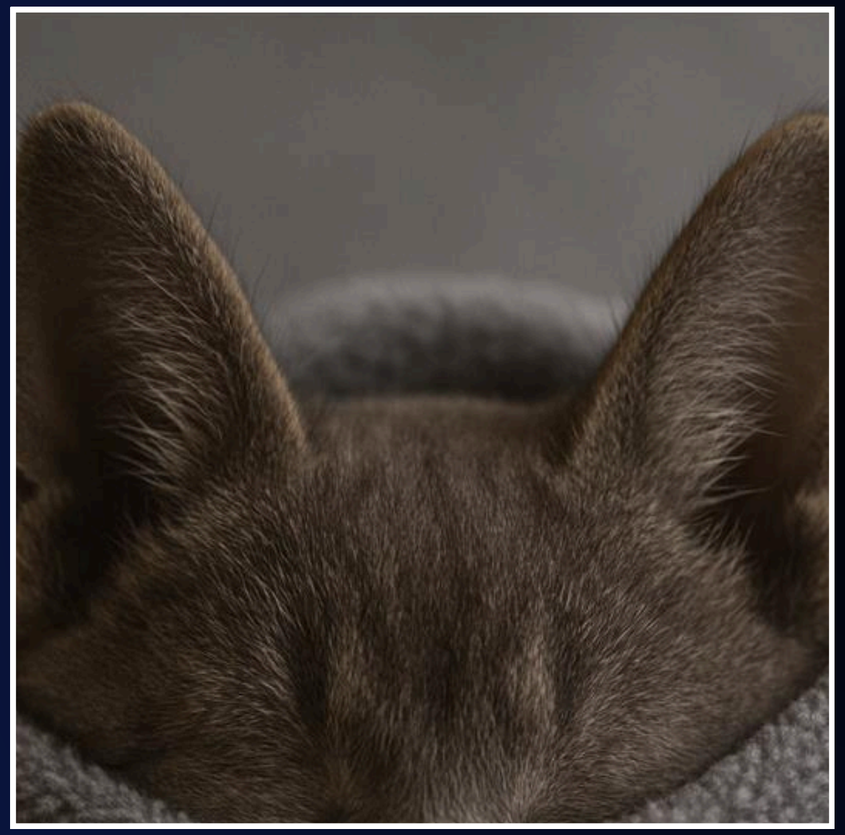


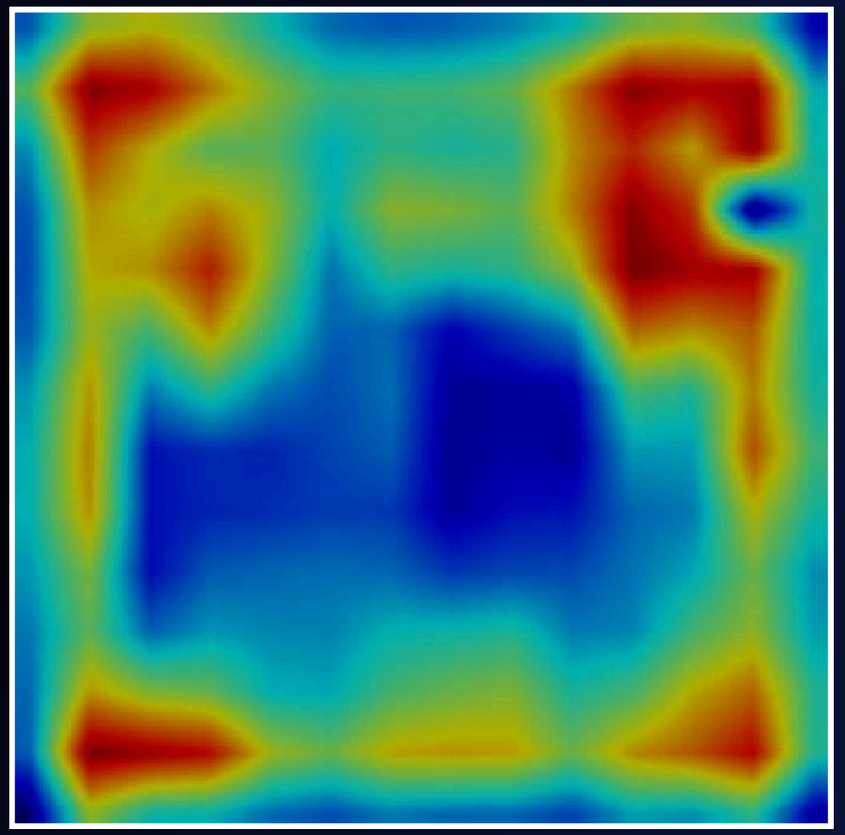
CACHORRO!

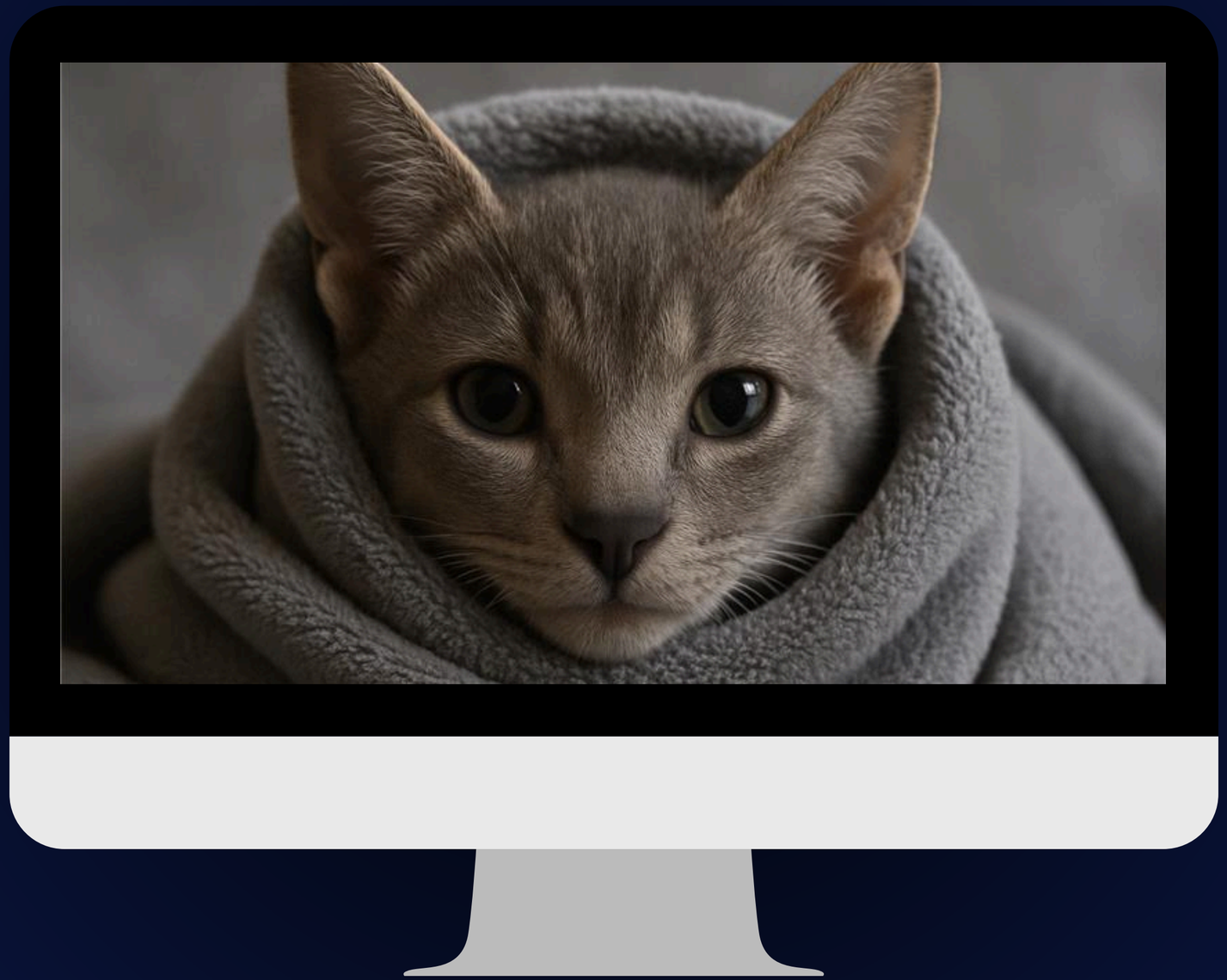
vgg16



GRAD-CAM







GITHUBS

LIME

GRAD-CAM

SHAP

REFERÊNCIAS

RIBEIRO, Marco Túlio; SINGH, Sameer; GUESTRIN, Carlos. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. arXiv preprint, 2016. Disponível em: Deep learning, a specialized form of machine learning, utilizes artificial neural networks with multiple layers to extract complex patterns from data.

YANG, Shiyue et al. A comprehensive survey of explainable artificial intelligence: A multimodal perspective. arXiv preprint, 2021. Disponível em: Deep learning, a specialized form of machine learning, utilizes artificial neural networks with multiple layers to extract complex patterns from data.

YIN, Pingping et al. A comprehensive survey on explainable artificial intelligence: From shallow to deep learning. National Center for Biotechnology Information, 2023. Disponível em: Deep learning, a specialized form of machine learning, utilizes artificial neural networks with multiple layers to extract complex patterns from data.

GOOGLE CLOUD TECH. Interpreting Machine Learning Models with LIME and SHAP. YouTube, 2019. Disponível em: https://www.youtube.com/watch?v=hUnRCxnydCc.

CAMPUSX. LIME | Machine Learning Interpretability. YouTube, 2021. Disponível em: https://www.youtube.com/watch?v=jFHPEQi55Ko.

REFERÊNCIAS

BLONDIEBYTES. SHAP Values | Explain ML Models. YouTube, 2023. Disponível em: https://www.youtube.com/watch?v=dQ_jvRkzN1Q.

GILL, Shagun. The Explainable Neural Network. Medium, 2020. Disponível em: https://medium.com/@shagunm1210/the-explainable-neural-network-8f95256dcddb. Acesso em: 22 abr. 2025.

LIME: Explain Machine Learning Predictions. Medium, 2018. Disponível em: https://medium.com/data-science/lime-explain-machine-learning-predictions-af8f18189bfe. Acesso em: 22 abr. 2025.

YOUR DATA TEACHER. How to explain neural networks using SHAP. YourDataTeacher, 17 maio 2021. Disponível em: https://www.yourdatateacher.com/2021/05/17/how-to-explain-neural-networks-using-shap/. Acesso em: 21 abr. 2025.

AKBARI, Kevin. Explaining Neural Network Models with SHAP Values – A Mathematical Perspective. Medium, 15 abr. 2021. Disponível em: https://akbarikevin.medium.com/explaining-neural-network-models-with-shap-values-a-mathematical-perspective-a57732d1ff0e. Acesso em: 21 abr. 2025.

LUNDENBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. arXiv preprint arXiv:1610.02391, 2017. Disponível em: https://arxiv.org/abs/1610.02391. Acesso em: 21 abr. 2025.

REFERÊNCIAS

MUSKAN, B. Grad-CAM: A Beginner's Guide. Medium, 2021. Disponível em: https://medium.com/@bmuskan007/grad-cam-a-beginners-guide-adf68e80f4bb. Acesso em: 21 abr. 2025.

DEEPFINDR. Grad-CAM Explained Visually - Computer Vision for Beginners. YouTube, 27 set. 2023. Disponível em: https://www.youtube.com/watch?v=_QiebC9WxOc. Acesso em: 21 abr. 2025.

SHAGUN, M. The Explainable Neural Network. Medium, 2021. Disponível em: https://medium.com/@shagunm1210/the-explainable-neural-network-8f95256dcddb. Acesso em: 21 abr. 2025.

IBM. Explainable AI (XAI). IBM Think, 2024. Disponível em: https://www.ibm.com/think/topics/explainable-ai. Acesso em: 21 abr. 2025.

XAI FOUNDATION. XAI for Neural Networks. XAI Foundation, 2024. Disponível em: https://www.xaifoundation.org/xai-for-neural-networks. Acesso em: 21 abr. 2025.

XAI EXPLAINABLE AI. Explainable AI (XAI). YouTube, 2023. Disponível em: https://youtu.be/YuDijSIR9iM?si=5EUT0rqxz-uN3BLg. Acesso em: 21 abr. 2025.

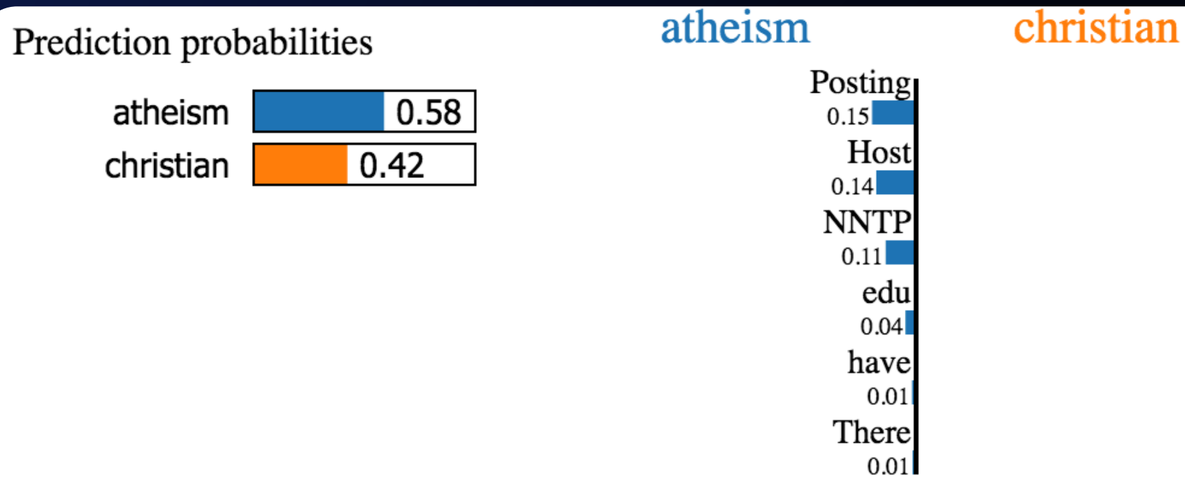
MUITO OBRIGADO

MATERIAL EXTRA

VAI QUE PRECISA NE...

LIME

CLASSIFICADOR DE TEXTO : MODELO QUE PROPÕE CLASSIFICAR O EMAIL EM ATEU OU CRISTÃO



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

ARTIGO SOBRE O SHAP

APROXIMAÇÕES EFICIENTES

O método de SHAP pode se adequar a vários modelos dadas as aproximações para os valores de Shapley em cada contexto

"PAI" DE MODELOS

Alguns vários métodos de explicação são frutos de casos especiais e aplicáveis do SHAP.

PROPRIEDADES DESEJADAS

Definição de 3 propriedades fundamentais para um bom método de explicação:

- Consistência;
- Exatidão local;
- "Irrelevância"

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2)

REFERÊNCIAS DO SHAP

GITHUB

ARTIGO

MEDIUM