



# Deep Learning Aplicado al procesamiento de Imágenes

Día 3 – Inteligencia Artificial Explicable

Dr. Rodrigo Salas Fuentes  
[rodrigo.salas@uv.cl](mailto:rodrigo.salas@uv.cl)



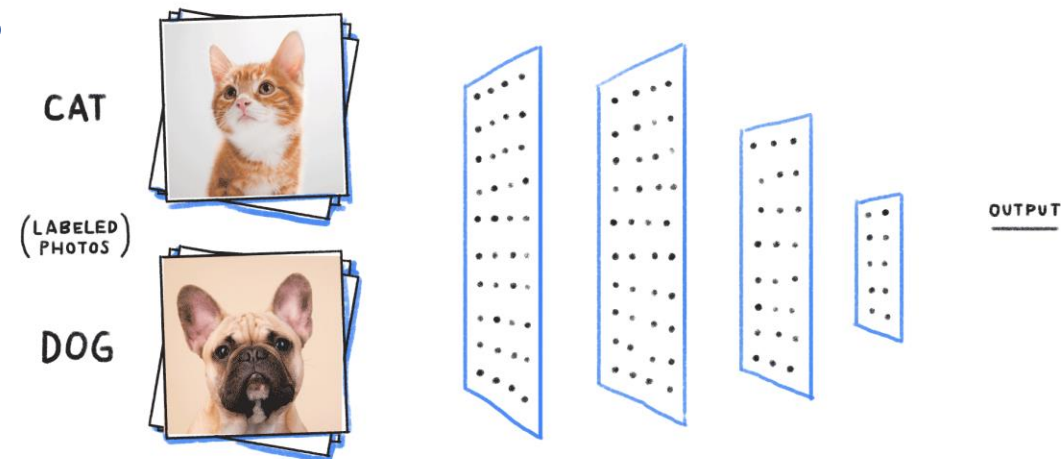
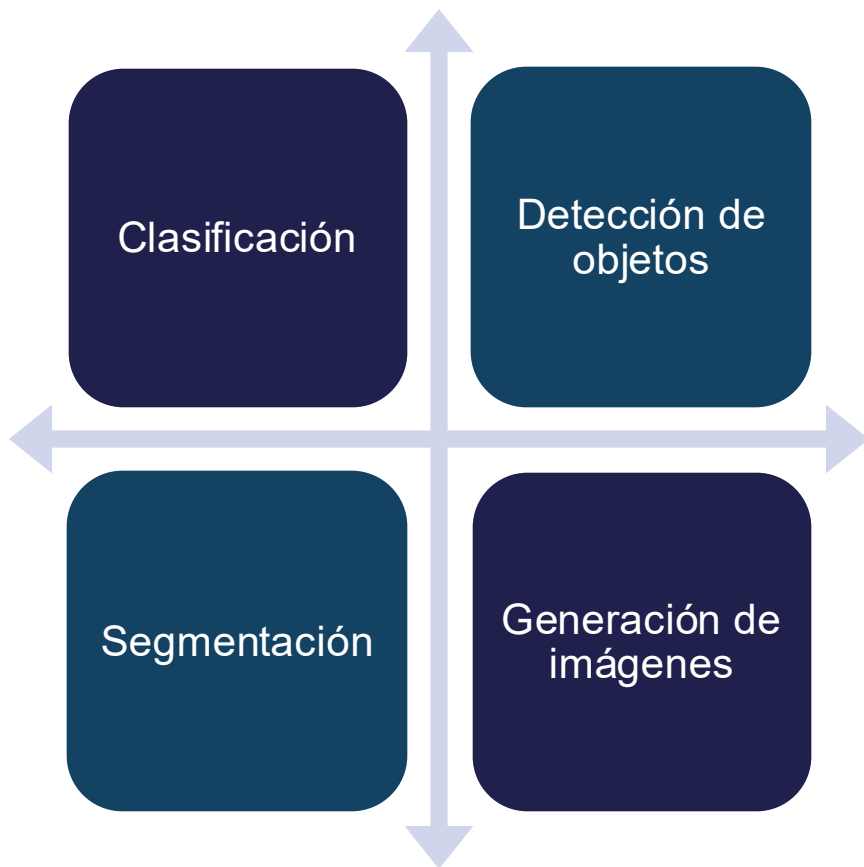
# Explicabilidad en Inteligencia Artificial



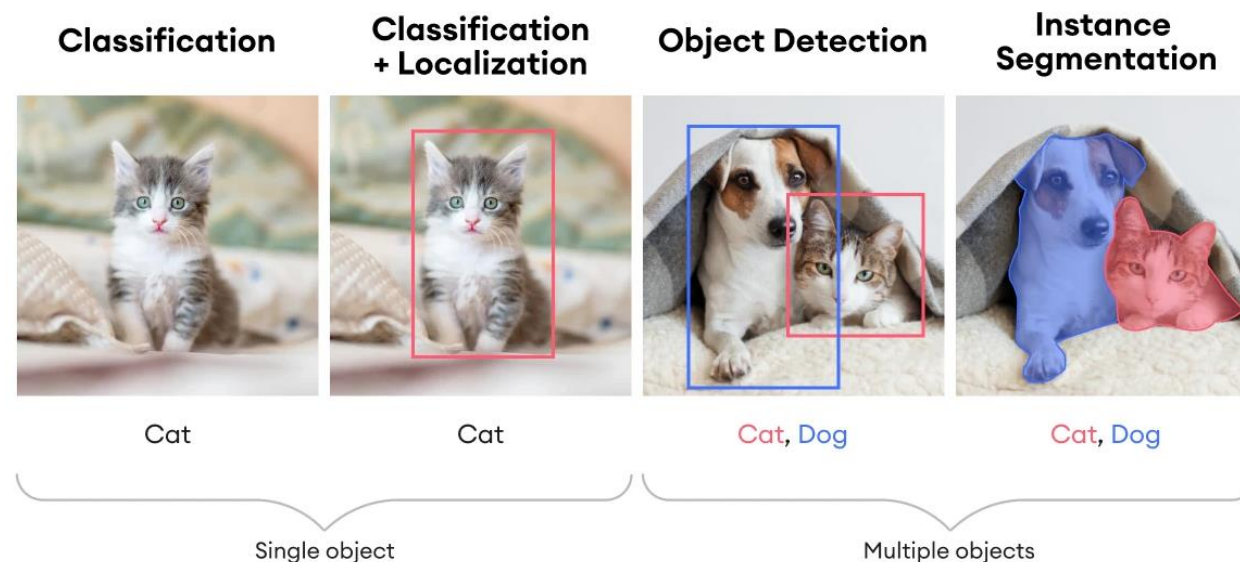
MSc. Student  
Esthefanía Astargo

Material desarrollado con el apoyo de la estudiante  
del Magister en Ciencias e Ingeniería para la Salud,  
Universidad de Valparaíso  
Sr. Esthefanía Astargo

# Usos de Deep learning en Imágenes



<https://medium.com/@DodoSirirat/image-classification-108ebbb19514>



<https://www.superannotate.com/blog/image-segmentation-for-machine-learning>

## Limitaciones



### Falta de transparencia

- No hay una explicación clara de cómo se generan las predicciones.



### Confianza limitada

- Es difícil confiar en salidas que no se pueden comprender.



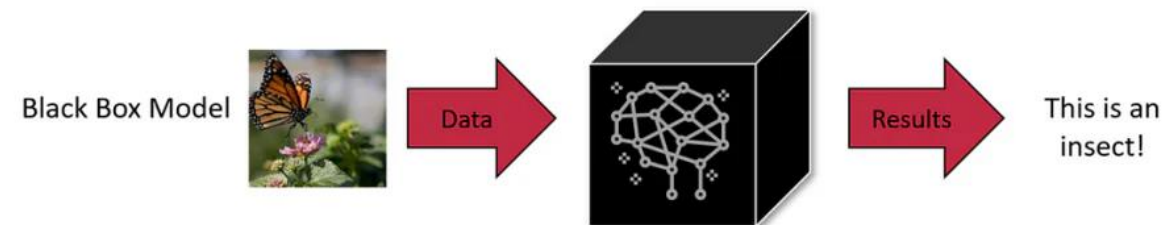
### Sesgos y errores ocultos

- Puede ser peligroso si las decisiones incorrectas pasan desapercibidas.

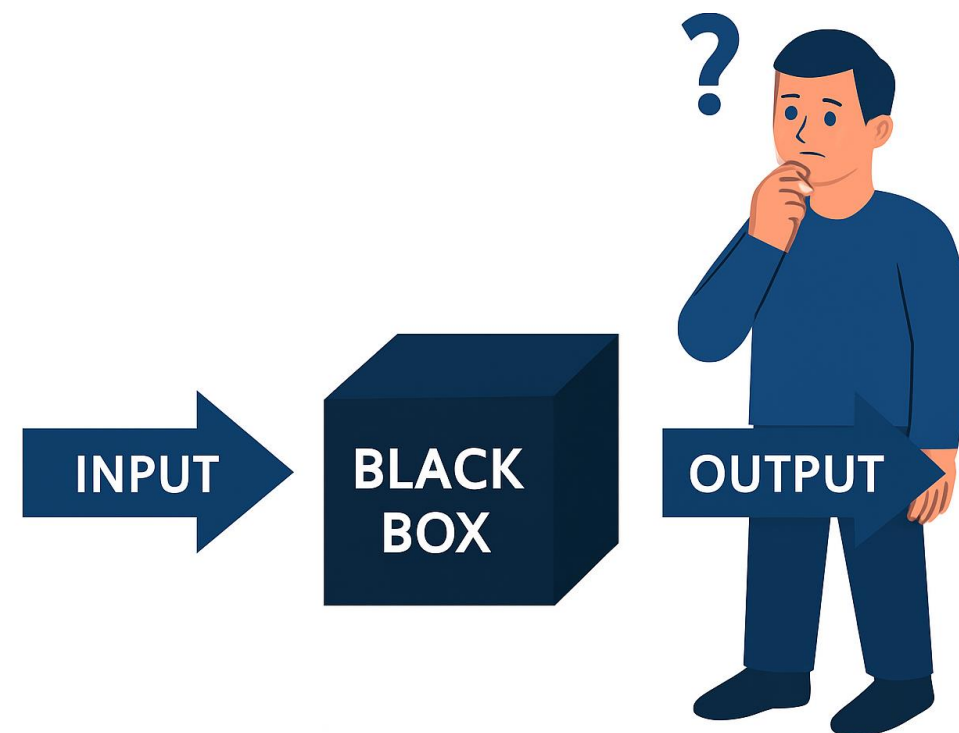


### Preocupaciones éticas

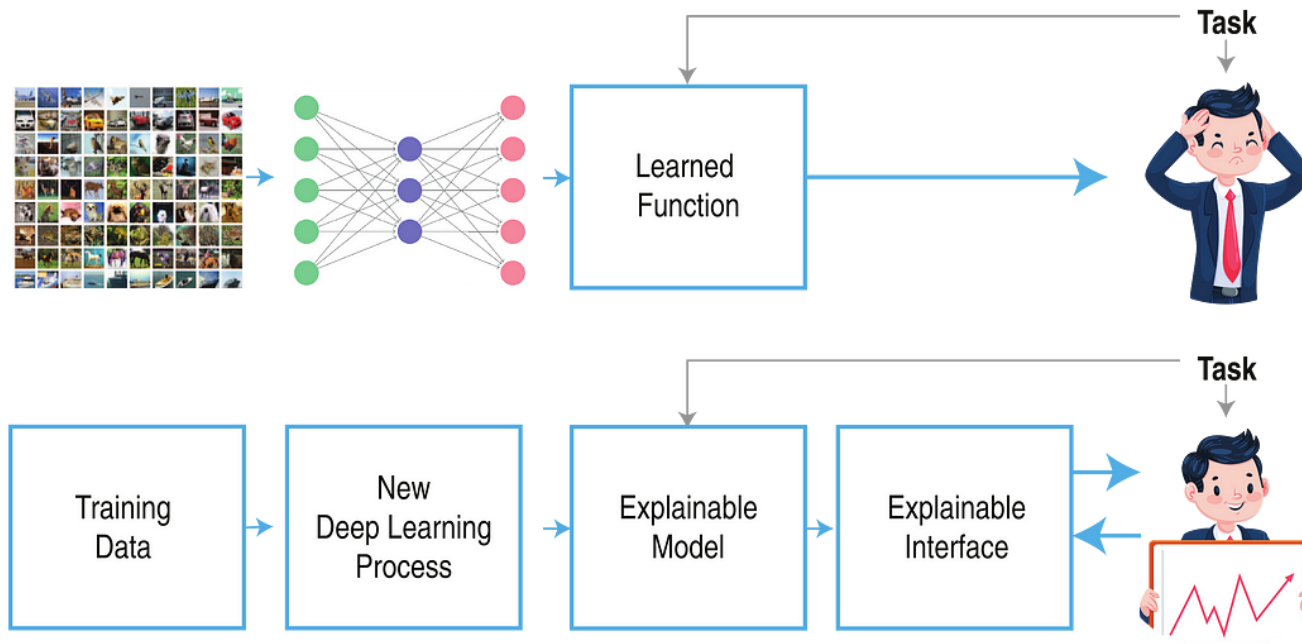
- Los usuarios necesitan razones comprensibles detrás de decisiones críticas.



<https://www.unite.ai/the-black-box-problem-in-llms-challenges-and-emerging-solutions/>



# Inteligencia Artificial Explicable (XAI)



## Beneficios

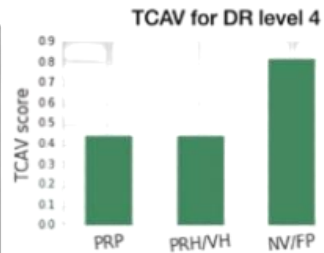
- ➡ Permite el desarrollo de modelos de IA confiables y comprensibles.
- ➡ Facilita la evaluación continua para obtener resultados más rápidos de la IA.
- ➡ Reduce los riesgos, los sesgos y los costos de auditoría en los sistemas de IA.

<https://medium.com/deepviz/what-is-xai-explainable-ai-and-visualization-part-10-da41c981c5fa>

La **Inteligencia Artificial Explicable (XAI)** es un conjunto de procesos y métodos que permite a los usuarios humanos comprender y confiar en los resultados generados por los algoritmos de inteligencia artificial.

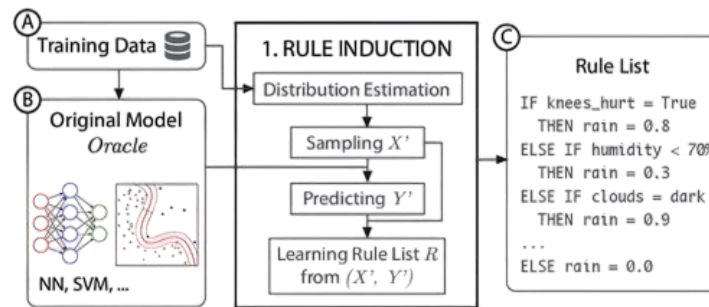


# Enfoques en Explicabilidad Post-Hoc



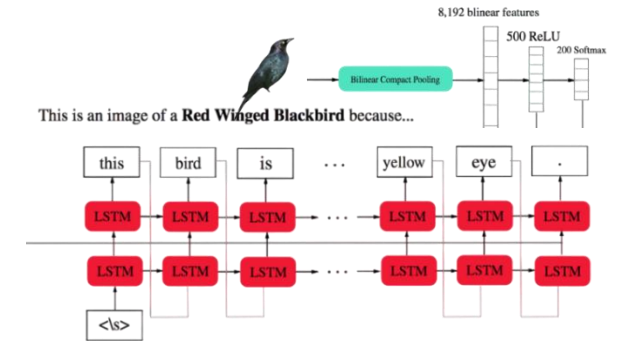
**Numeric Explanations**

<https://doi.org/10.48550/arXiv.1711.11279>



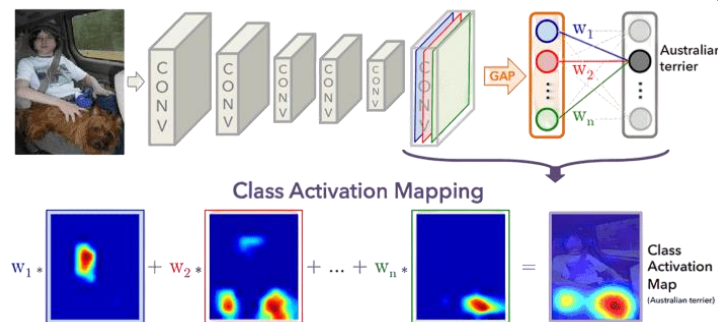
**Rule-Based Explanations**

<http://dx.doi.org/10.1109/TVCG.2018.2864812>



**Textual Explanations**

<https://doi.org/10.3390/make3030032>



**Visual Explanations**

<https://doi.org/10.48550/arXiv.1512.04150>

**Q: Is this a healthy meal?**



**A: No**

**Textual Justification**

...because it is a hot dog with a lot of toppings.



**A: Yes**

...because it contains a variety of vegetables on the table.



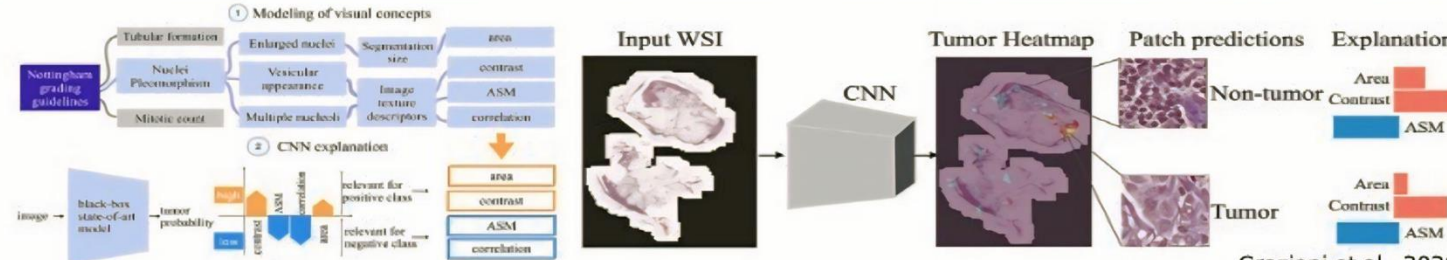
**Visual Pointing**



**Mixed Explanations**

<https://doi.org/10.3390/make3040048>

# XAI en Imágenes

Image	Saliency Map Visualization	 <p>Simonyan et al., 2013</p>
	Shapley Value Importance	 <p>Ghorbani et al., 2020</p>
	Concept Attribution	 <p>Graziani et al., 2020</p>

<https://spectra.mathpix.com/article/2021.09.00007/demystify-post-hoc-explainability>

# Explicabilidad Visual en Deep Learning



MSc. Student  
Esthefanía Astargo

Material desarrollado con el apoyo de la estudiante  
del Magister en Ciencias e Ingeniería para la Salud,  
Universidad de Valparaíso  
Sr. Esthefanía Astargo



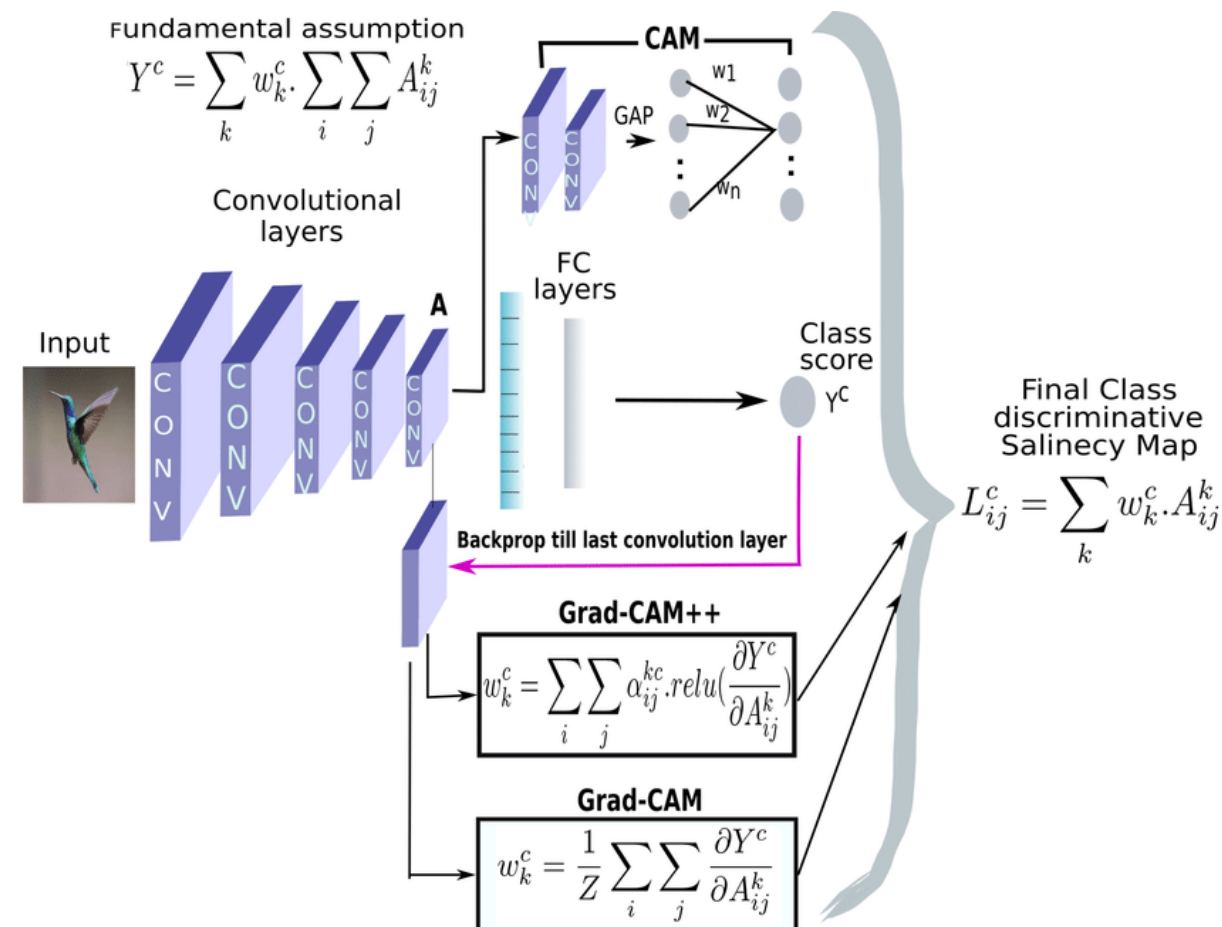
# Grad-CAM: Gradient-weighted Class Activation Mapping

Grad-CAM es un método visual post-hoc y específico de modelo, diseñado para redes convolucionales profundas (CNNs).

Se basa en los gradientes del modelo para identificar qué regiones del input activaron más una clase específica.

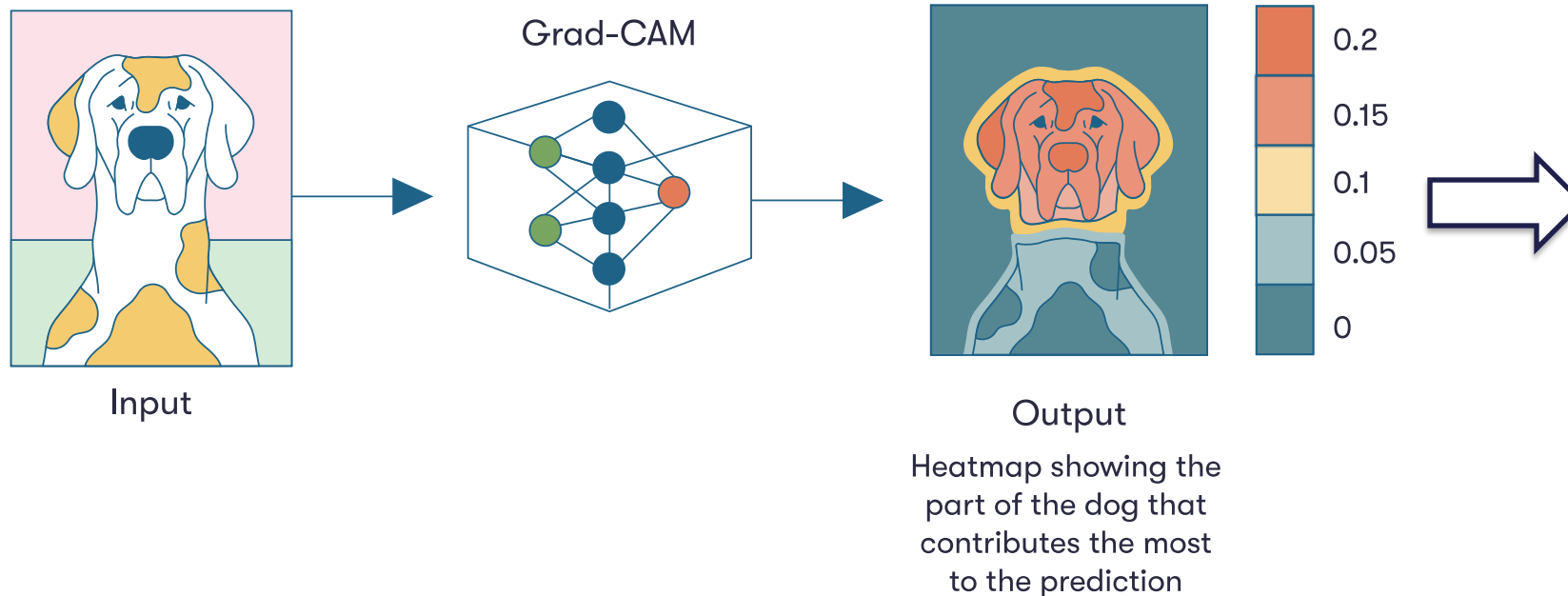
Utiliza los mapas de activación de la última capa convolucional, conservando la información espacial relevante.

Calcula un mapa ponderado de relevancia combinando gradientes y activaciones, destacando zonas clave.



<https://doi.org/10.48550/arXiv.1710.11063>

# Grad-CAM: Gradient-weighted Class Activation Mapping

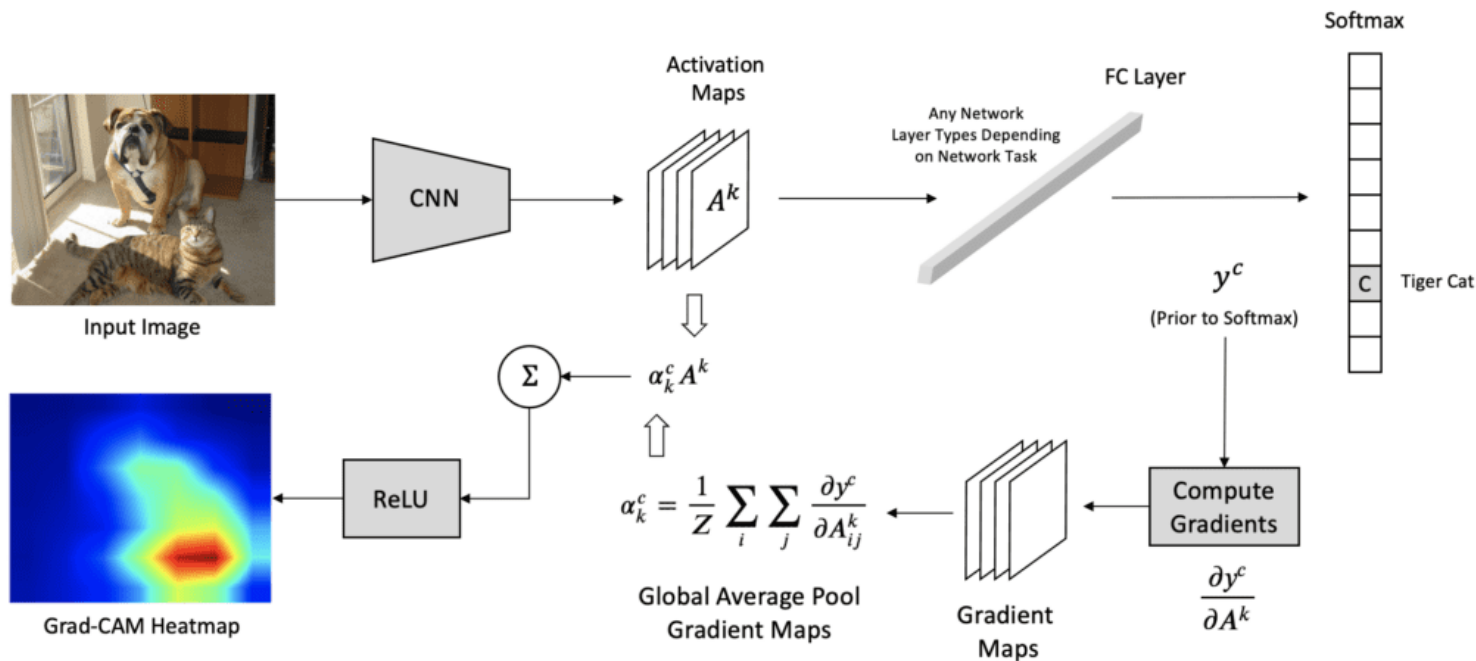


<https://courses.minnaleam.com/en/courses/trustworthy-ai/preview/explainability/types-of-explainable-ai/>

## Grad-CAM aplicado a clasificación de imágenes

- El modelo recibe como entrada la imagen de un perro y realiza una predicción de clase.
- Grad-CAM propaga hacia atrás los gradientes desde la salida hasta la última capa convolucional.
- A partir de los mapas de activación y gradientes, se genera un mapa de calor.
- El heatmap superpuesto indica qué regiones del perro fueron más relevantes para la predicción.

# Grad-CAM paso a paso



[https://xai-tutorials-readthedocs-io.translate.goog/en/latest/\\_model\\_specific\\_xai/Grad-CAM.html?\\_x\\_tr\\_sl=en&\\_x\\_tr\\_tl=es&\\_x\\_tr\\_hl=es&\\_x\\_tr\\_pto=tc](https://xai-tutorials-readthedocs-io.translate.goog/en/latest/_model_specific_xai/Grad-CAM.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc)

## Paso 1: Forward Pass

- Obtener los mapas de características de la última capa convolucional y las salidas sin procesar antes de la función softmax.
- Estos mapas de características se denominan  $A^k$ , donde  $k$  es un mapa de características específico dentro de una capa convolucional.

## Paso 2: Selección de la clase objetivo

- Se elige la clase  $c$  que se desea explicar (normalmente la clase predicha con la puntuación más alta) y se toma su valor de activación antes de la función softmax.
- $y^c$  representa la activación asociada a la clase  $c$  antes de aplicar la función softmax.

## Grad-CAM paso a paso

### Paso 3: Calcular los gradientes

- Se calculan los gradientes de  $y^c$  con respecto a los mapas de características  $A^k$ , es decir:

$$\frac{\partial y^c}{\partial A^k}$$

### Paso 4: Calcular el mapa Grad-CAM

- Se calcula un peso por cada mapa (media global de gradientes):

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

- Luego se combinan todos los mapas con sus pesos:

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

- La función ReLU conserva solo las regiones que contribuyen positivamente.

[https://xai--tutorials-readthedocs-io.translate.goog/en/latest/\\_model\\_specific\\_xai/Grad-CAM.html?\\_x\\_tr\\_sl=en&\\_x\\_tr\\_tl=es&\\_x\\_tr\\_hl=es&\\_x\\_tr\\_pto=tc](https://xai--tutorials-readthedocs-io.translate.goog/en/latest/_model_specific_xai/Grad-CAM.html?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc)

## Taller 5 – Uso de Grad-CAM para Explicabilidad en Clasificación de Imágenes



[https://colab.research.google.com/drive/1WISel\\_ilJlwhj\\_fx4kl2ufswH\\_ni-9\\_9?usp=sharing](https://colab.research.google.com/drive/1WISel_ilJlwhj_fx4kl2ufswH_ni-9_9?usp=sharing)



# Explicabilidad Numérica en Deep Learning



MSc. Student  
Esthefanía Astargo

Material desarrollado con el apoyo de la estudiante  
del Magister en Ciencias e Ingeniería para la Salud,  
Universidad de Valparaíso  
Sr. Esthefanía Astargo

# LIME: Local Interpretable Model-agnostic Explanations

Construye un modelo sustituto que aproxima al modelo complejo solo en torno a una predicción específica.

Genera variaciones del dato de entrada y observa cómo el modelo original responde a estos cambios.

Utiliza esas observaciones para entrenar un modelo simple que imita al modelo complejo localmente.

Este modelo sustituto revela qué características influyen más en esa predicción puntual.

Puede aplicarse a texto, imágenes, o datos tabulares, sin requerir acceso interno al modelo.

Matemáticamente, los modelos sustitutos se pueden expresar de la siguiente manera:

$$\xi(x) = \arg \min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$f$  es el modelo complejo original.

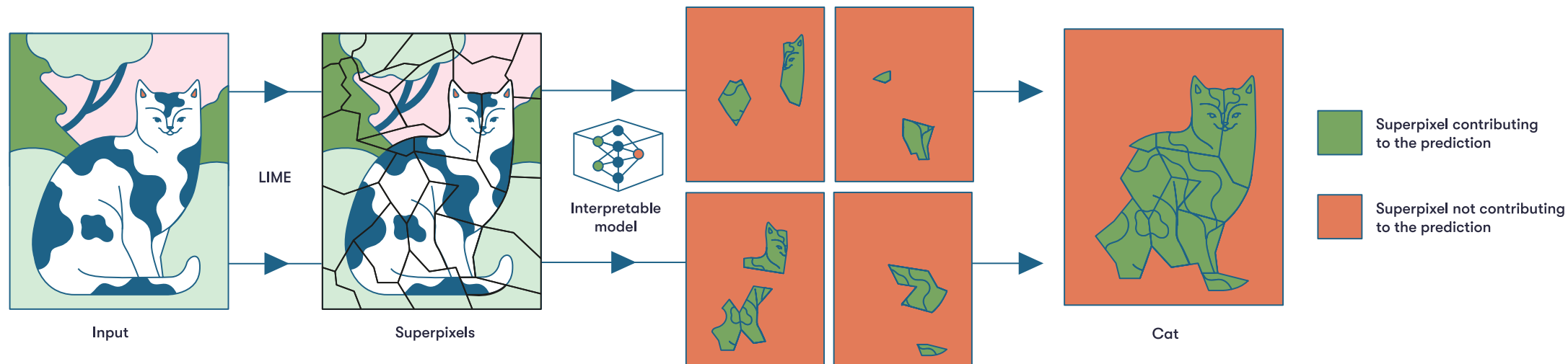
$g$  es el modelo interpretable local.

$\mathcal{L}(f, g, \pi_x)$  mide cuán bien  $g$  aproxima a  $f$  cerca de  $x$ .

$\pi_x$  define la vecindad local ponderando las instancias perturbadas.

$\Omega(g)$  penaliza la complejidad de  $g$ .

# LIME: Local Interpretable Model-agnostic Explanations



<https://courses.minnalearn.com/en/courses/trustworthy-ai/preview/explainability/types-of-explainable-ai/>

## LIME aplicado a clasificación de imágenes

Divide la imagen en superpíxeles, que agrupan regiones visuales coherentes y comprensibles.

Se generan múltiples versiones de la imagen desactivando distintas combinaciones de superpíxeles

El modelo original predice sobre cada versión modificada y LIME observa cómo cambian las salidas.

Luego, entrena un modelo interpretable para estimar qué superpíxeles son claves en la decisión.

Ejemplo: los superpíxeles verdes explican por qué el modelo clasificó la imagen como *gato*.

## Taller 6 - Uso de LIME para Explicabilidad en Clasificación de Imágenes



<https://colab.research.google.com/drive/1DNT8MrL63xLRvN3G02bdhNLYZvuY6F8y?usp=sharing>





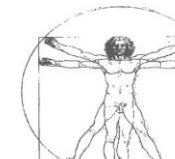


**Dr. Rodrigo Salas  
Fuentes**  
[rodrigo.salas@uv.cl](mailto:rodrigo.salas@uv.cl)



**Thanks for your Attention**

-  LinkedIn: [linkedin.com/in/rodrigo-salas-fuentes/](https://linkedin.com/in/rodrigo-salas-fuentes/)
-  Google Scholar: [scholar.google.com/citations?user=ZaqDIPcAAAAJ](https://scholar.google.com/citations?user=ZaqDIPcAAAAJ)
-  ORCID: [orcid.org/0000-0002-0350-6811](https://orcid.org/0000-0002-0350-6811)
-  Email: [rodrigo.salas@uv.cl](mailto:rodrigo.salas@uv.cl)



Ingeniería Civil Biomédica  
Facultad de Ingeniería  
Universidad de Valparaíso  
[www.biomedica.uv.cl](http://www.biomedica.uv.cl)