Interpretabilidad del Deep Learning

Explicabilidad específica

Christian Oliva Moya

Contenido del curso

- Segunda semana:
  - Explicabilidad específica para redes neuronales
  - Métodos basados en gradientes: *Gradient x Input*
  - Métodos basados en relevancias: Layerwise Relevance Propagation
  - ¿Es LRP equivalente a *Gradient x Input*?
  - **Dificultades del Deep Learning**

Dificultades de la Explicabilidad en DL (I)

- No todo es color de rosas en la explicabilidad del Deep Learning
- Hemos trabajado todo el rato con un dataset con datos tabulares
- ¿Qué pasa con el procesamiento de imágenes? ¿Y series temporales?

Dificultades de la Explicabilidad en DL (II)

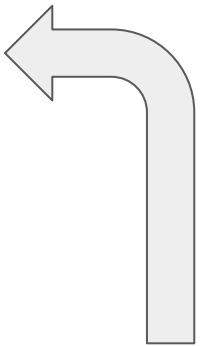
- El **coste computacional** de los algoritmos de XAI como SHAP o LIME es un problema. Si para 21 atributos es costoso, ¿qué sucede con imágenes?
- MNIST por ejemplo tiene 28x28 = 784 atributos de entrada
- Ni mucho menos pensemos en calcular el EPC, que hay que repetirlo para cada imagen de entrada (60.000 imágenes de train tiene MNIST)

Dificultades de la Explicabilidad en DL (III)

- Descartando los métodos genéricos, nos quedan:
    - Análisis de pesos como suma de valores absolutos (con MLPs)
    - Métodos basados en gradientes: Gradient x Input y LRP

Dificultades de la Explicabilidad en DL (IV)

- Descartando los métodos genéricos, nos quedan:
  - Análisis de pesos como suma de valores absolutos (con MLPs)
  - Métodos basados en gradientes: Gradient x Input y LRP



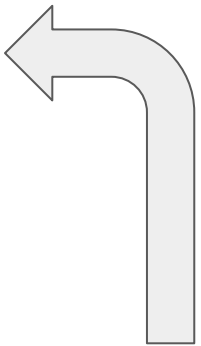
$$c^{(l)}_{m,(x,y)} = \sum_{n=1}^{N} \sum_{i=1}^{N_w} \sum_{j=1}^{N_w} o^{(l-1)}_{n,((x-1+i),(y-1+j))} w^{(l)}_{n,m,(i,j)} + b^{(l)}_{m}$$

$$o^{(l)}_{m,(x,y)} = f(c^{(l)}_{m,(x,y)})$$

¿Qué pasa con CNNs?
¿Y con transformers?
¿Y con RNNs?

Dificultades de la Explicabilidad en DL (V)

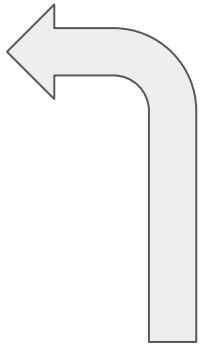
- Descartando los métodos genéricos, nos quedan:
  - ~~Análisis de pesos como suma de valores absolutos (con MLPs)~~
  - Métodos basados en gradientes: Gradient x Input y LRP



Mejor ni lo pensamos

Dificultades de la Explicabilidad en DL (VI)

- Descartando los métodos genéricos, nos quedan:
  - Métodos basados en gradientes: Gradient x Input y LRP



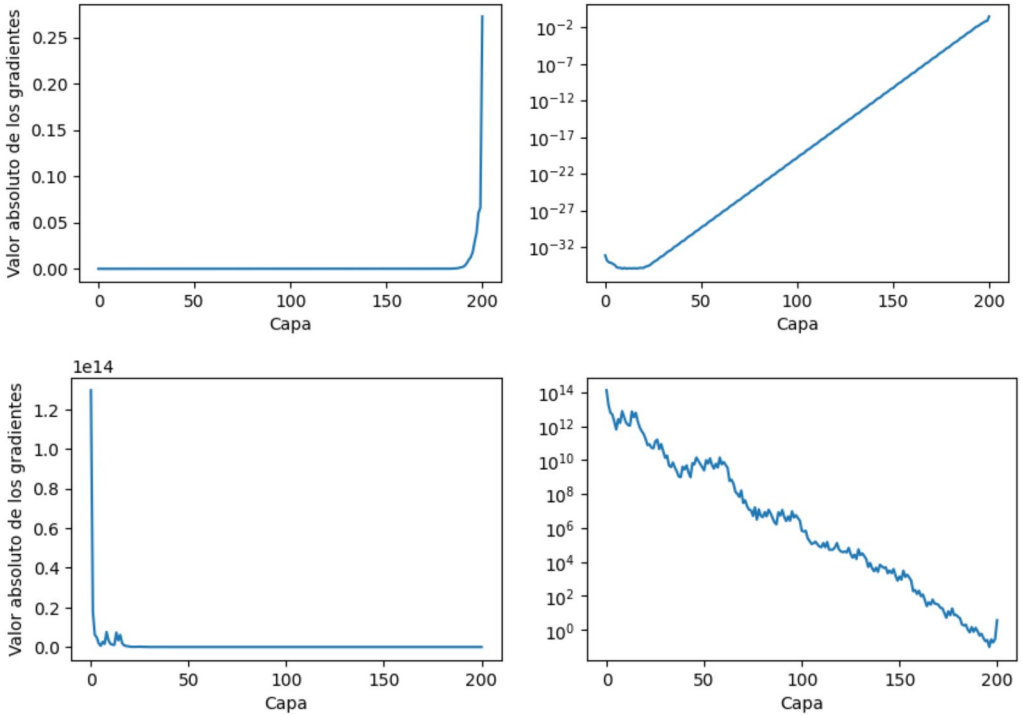
¿Qué pasa con CNNs?
¿Y con transformers?
¿Y con RNNs?

Dificultades de la Explicabilidad en DL (VII)

- Descartando los métodos genéricos, nos quedan:
    - Métodos basados en gradientes: Gradient x Input y LRP
- Podemos seguir calculando los gradientes
- Tenemos un nuevo problema: **Vanishing y Exploding Gradients** sobre todo en RNNs

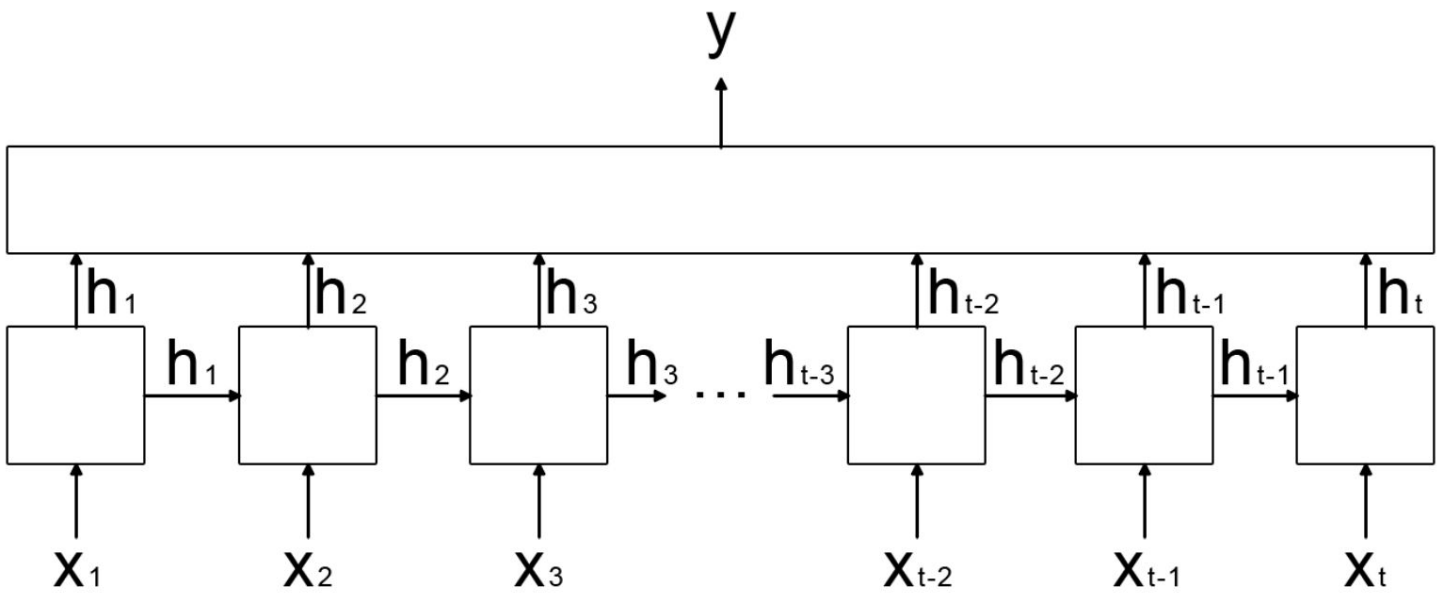
Dificultades de la Explicabilidad en DL (VIII)

- Si ocurre **Vanishing o Exploding Gradients**



Dificultades de la Explicabilidad en DL (IX)

- Intentemos algo para no tener pérdidas en la relevancia. ¿Y si…?



Taller 1: Explicabilidad de MNIST