Illuminating Neural Networks: A Cycle of Explainable AI for Gravitational Waves

Miguel Cárdenas-Montes CIEMAT (Spain)

Neural Networks

- Neural Networks has become an integral component of data analyses pipelines in contemporary experiments within Particle Physics and Astrophysics, including Gravitational Waves analysis.
- Their application emerges as a promising solution to address the escalating volume, speed and complexity of experimental data.



- Within this context, **Explainable Artificial** Intelligence (XAI) assumes a crucial role in deciphering the intricate decisions of Deep Learning, uncovering biases, and facilitating iterative improvements.
- XAI algorithms serve as interpreters, offering researchers and practitioners a glimpse into the mechanisms guiding the models' decisions for later improvements.
- Not more deeper and deeper NN for improving the performance!

XAI, intuition



XAI, intuition and biased learning

Dog correctly classified as dog (class 1): [4.8e-11 (cat), 1.0e+00 (dog)]

Image segmented in homogeneous patches (250).

Randomly patches are switched on/off and then predicted.



Degradation of prediction illuminates the importance of the patches. And repeat ... Segmented Image with Top Superpixels



The 10 most important patches are mostly grass.

Perturbed Image (Top Features)



Linear regressor with patches on/off as input and the performance degradation as output.

Gravity Spy (Kaggle)

- time-frequency images (Q-transform),
- 22 labels (chirp + 21 glitches),
- 31.9k files: train, test, validation;
- 4 images per event, time window of 0.5,
 1, 2, and 4 seconds,
- strongly unbalanced labels,
- public data set.
- Ideal for testing ideas.
- Images ideal for Convolutional NN (CNN).
- Classification per image or per event?





Gravity Spy (Kaggle)

(2 or 3)X(Conv2D+MaxPooling) + output layer (from 6 to 2 [irreducible] chirp errors).

- Trainable parameters: 1M-400k
- ~190 per epoch
- EarlyStopping patience=3
- Most of the chirp errors go to blip label.
- All the chirp errors are for 4 s time window.

1080Lines	183	4	0	5	0	0	0	0	1	0	0	2	0	0	0	0	1	1	0	3	0	0
1400Ripples	0	35	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Air_Compressor	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Blip	0	0	0	1066	0	0	0	0	0	0	8	14	0	0	0	3	0	0	1	0	0	0
Chirp	0	0	0	6	34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Extremely_Loud	0	0	1	0	0	158	0	95	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Helix	0	0	0	9	0	0	139	0	0	0	0	18	0	0	0	0	0	2	0	0	0	0
Koi_Fish	0	0	0	54	0	1	0	347	1	0	0	0	0	0	0	0	0	0	5	0	0	0
Light_Modulation	0	0	3	25	0	0	1	23	226	11	0	0	7	0	0	13	0	1	2	0	0	0
Low_Frequency_Burst	0	0	1	9	0	0	2	0	5	278	42	18	1	0	1	0	2	1	0	0	0	0
Low_Frequency_Lines	0	0	0	0	0	0	0	0	0	2	247	15	0	0	0	0	0	0	0	0	0	0
No_Glitch	1	0	0	1	0	4	0	0	0	0	0	78	0	0	0	0	0	0	0	0	0	0
None_of_the_Above	0	0	3	7	0	0	0	2	4	0	0	0	19	0	3	0	0	2	4	0	0	0
Paired_Doves	0	0	0	2	0	8	0	0	0	0	0	0	1	3	0	1	1	0	0	0	0	0
Power_Line	0	0	3	0	0	0	0	0	0	1	9	1	0	0	257	0	0	1	0	0	0	0
Repeating_Blips	0	0	0	31	0	0	0	1	1	0	0	0	0	0	0	113	0	1	0	1	0	0
Scattered_Light	0	0	6	6	0	0	0	1	0	4	1	1	0	0	1	0	245	3	0	0	0	0
Scratchy	0	0	0	9	0	0	0	0	0	0	0	2	0	0	0	0	4	185	0	0	0	0
Tomte	0	0	0	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	0	0	0
Violin_Mode	5	2	0	5	0	0	4	0	0	4	0	14	0	0	0	5	0	1	0	212	0	4
Wandering_Line	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	5	0	0	7	13	1
Whistle	8	1	1	10	0	0	0	0	0	0	0	13	0	0	0	0	0	0	0	0	1	146
	1080Lines	1400Ripples	Air_Compressor	Blip	Chirp	Extremely_Loud	Helix	Koi_Fish	Light_Modulation	ow_Frequency_Burst	ow_Frequency_Lines	No_Glitch	None_of_the_Above	Paired_Doves	Power_Line	Repeating_Blips	Scattered_Light	Scratchy	Tomte	Violin_Mode	Wandering_Line	Whistle

Predicted label

800

600

400

200

XAI (GradCAM), correctly classified events

Blip

Chirp



Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization <u>https://doi.org/10.48550/arXiv.1610.02391</u>

XAI (GradCAM), correctly classified events



Chirp event (correctly classified) from 0.5 s to 4 s. As the window is larger, the image is more stretched vertically. Horizontal part of the chirp becomes less relevant, and more prone to misclassification.

XAI (GradCAM), incorrectly classified events



XAI (GradCAM), incorrectly classified events



XAI (GradCAM), incorrectly classified events

How to modify the input for remove confusing features (pixels)? XAI teaches us how.





Original Image

150

200

XAI (LIME), correctly classified events

Most relevant patches for correctly classify.

In chirps, ramp-up in frequency is relevant (asymmetric horizontal).

10 parches

5 parches





Chirp



XAI (LIME), correctly classified events

Most relevant patches for correctly classify.

In chirps, ramp-up in frequency is relevant (asymmetric horizontal).

10 parches

5 parches



Chirp



"Why Should I Trust You?": Explaining the Predictions of Any Classifier: <u>https://doi.org/10.48550/arXiv.1602.04938</u>

XAI (LIME), correctly classified events

In blips, vertical-lateral to stronger part of the signal is relevant.



14

Events as sequences of images (time windows)



Convert 4 images into a sequence (event) -> event classifier

input shape = (4, 479, 569, 3)

model.add(layers.Conv3D(8, kernel_size=(1, 3, 3), activation='relu'))

instead of:

model.add(layers.Conv2D(filters=128, kernel_size=(3,3), activation = 'relu'))

Input: multitime tensor data

- No errors on chirps or 1 single error.
- 2D
 - 3 errors
 - 190 s /epoch, ~24 epochs, patience 3
 - 1M-400k trainable parameters.
- 3D
 - 0-2 errors (depends on 2D or 3D SpaAtt)
 - 15 s /epoch, ~4 epochs, patience 3
 - ~ 150k par. Lower carbon footprint
- Spatial Attention layer for interpretability.

Spatial Attention



inputs = keras.Input(shape=sample_shape)

smartly remove this segment if not spatial attention SpaAtt = SpatialAttention()(inputs)

conv1 = layers.Conv3D(8, kernel_size=(1, 3, 3), activation='relu')(SpaAtt)
maxpool1 = layers.MaxPooling3D(pool_size=(1, 3, 3))(conv1)

self.kernel = self.add_weight(name='kernel', shape = (1,1,input_shape[-1],1), initializer=Juniform', trainable=True)cla

Spatial attention 2D

Spatial attention per time layer.

It learns to amplify the pixels of interest, minimizing others.

Spatial Attention

1080Lines	92	1	1	0	0	0	1	0	0		_									_		
1400Ripples	0	22	0	0	0	0	0	0	0		Sp	Da	tia	ll á	att	e	nti	or	13	3D		
Air_Compressor	0	0	16	0	0	0	0	0	0													
Blip	0	0	0	534	0	1	0	7	0	0	2	0	0	0	0	2	0	0	2	0	0	0
Chirp	0	0	0	2	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Helix	0	0	0	0	0	123	0	6	0	0	0	0	0	0	0	0	0	2	0	0	0	0
Extremely_Loud	0	0	0	0	0	0	75	0	0	1	6	1	0	1	0	0	0	0	0	0	0	0
Koi_Fish	0	0	0	1	0	1	0	204	2	0	0	0	0	0	0	0	0	0	0	0	0	0
Light_Modulation	0	0	0	4	0	0	0	4	134	2	0	0	1	0	0	5	0	1	0	0	0	0
Low_Frequency_Lines	0	0	0	0	0	0	4	0	4	142	18	12	2	0	0	1	0	1	0	0	0	0
No_Glitch	0	0	0	0	0	0	0	0	0	6	101	24	0	0	0	0	1	0	0	0	0	0
ខ្ទី Low_Frequency_Burst	0	0	0	0	0	4	1	0	0	11	11	15	0	0	0	0	0	0	0	1	0	0
None_of_the_Above	0	0	1	0	0	0	1	1	2	0	0	0	12	0	0	0	0	2	2	0	2	1
Paired_Doves	0	0	0	0	0	3	0	0	0	2	0	1	1	1	0	0	0	0	0	0	0	0
Power_Line	0	0	0	0	0	0	0	0	0	0	0	0	0	0	135	0	0	0	0	0	0	0
Repeating_Blips	0	0	0	11	0	0	3	3	3	0	0	2	0	0	0	54	0	2	0	0	0	0
Scattered_Light	0	0	0	1	0	0	0	0	0	3	1	0	0	0	0	0	128	2	0	0	0	0
Scratchy	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	98	0	0	0	0
Tomte	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	0	0
Violin_Mode	0	1	0	0	0	0	1	0	0	2	1	1	0	0	0	1	0	0	0	120	0	1
Wandering_Line	2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	7	3
Whistle	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2	1	85
	1080Lines	1400Ripples	Air_Compressor	Blip	Chirp	Helix	Extremely_Loud	Koi_Fish	Light_Modulation	Low_Frequency_Lines	No_Glitch	Low_Frequency_Burst	None_of_the_Above	Paired_Doves	Power_Line	Repeating_Blips	Scattered_Light	Scratchy	Tomte	Violin_Mode	Wandering_Line	Whistle

- 500

- 400

- 300

- 200

100

										_												~ ~	
1080Lines	92	1	0	0	0	0	2	0	0		_								_				
1400Ripples	0	22	0	0	0	0	0	0	0		Sp	ba	tia	ll a	att	er	nti	on	12	ĽD			
Air_Compressor	0	0	16	0	0	0	0	0	0														
Blip	0	0	0	531	0	1	0	2	2	1	2	0	1	0	0	7	0	0	1	0	0	0	
Chirp	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Helix	0	0	0	0	0	125	0	2	1	0	0	0	2	0	0	0	0	0	0	1	0	0	
Extremely_Loud	0	0	0	0	0	0	77	0	0	0	3	0	0	0	0	1	0	2	0	0	0	1	
Koi_Fish	0	0	0	1	0	2	0	202	2	0	0	0	0	0	0	1	0	0	0	0	0	0	
Light_Modulation	0	0	0	2	0	з	0	з	134	1	0	0	з	0	0	3	0	1	0	0	0	1	
Low_Frequency_Lines	0	0	0	1	0	0	5	0	4	158	11	2	1	0	0	0	2	0	0	0	0	0	
No_Glitch	0	0	0	0	0	0	14	0	0	13	94	10	1	0	0	0	0	0	0	0	0	0	
2 Low_Frequency_Burst	0	0	0	0	0	4	4	0	0	12	6	14	0	0	0	2	0	0	0	1	0	0	
None_of_the_Above	0	0	0	0	0	0	0	2	1	0	0	0	14	0	0	1	0	3	0	0	2	1	
Paired_Doves	0	0	0	0	0	3	0	0	1	1	0	0	0	1	0	0	0	0	1	1	0	0	
Power_Line	0	0	0	0	0	0	0	0	0	0	0	0	0	0	135	0	0	0	0	0	0	0	
Repeating_Blips	0	0	0	9	0	1	1	0	2	0	0	0	0	0	0	64	0	0	0	0	0	1	
Scattered_Light	0	0	0	0	0	1	0	2	0	2	1	0	1	0	0	0	125	3	0	0	0	0	
Scratchy	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	1	0	94	0	0	0	0	
Tomte	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	29	0	0	0	
Violin_Mode	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	121	3	1	
Wandering_Line	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	9	2	
Whistle	0	1	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	86	
	1080Lines	1400Ripples	Air_Compressor	Blip	Chirp	Helix	Extremely_Loud	Koi_Fish	Light_Modulation	ow_Frequency_Lines	No_Glitch	w_Frequency_Burst	None_of_the_Above	Paired_Doves	Power_Line	Repeating_Blips	Scattered_Light	Scratchy	Tomte	Violin_Mode	Wandering_Line	Whistle	

- 500

- 400

- 300

- 200

- 100

Spatial Attention 2D vs. 3D





Label: Chirp

Spatial Attention 2D vs. 3D





Label: Blip

Spatial Attention 2D vs. 3D





Label: Koi_Fish

Preliminary resu Spatial Attention in the two first layers 0.5s ls 2s 4s Attention Mask 150 Attention Intensity - 25 - 0 Spatial attention 2D L0 Label: Chirp 0.5s 2s 1s 4s Attention Mask - 14 Attention Intensity - 2 Spatial attention 2D L1

Label: Chirp



Label: Chirp



Label: Blip





Spatial Attention in the two first layers





Label: Koi_Fish

Preliminary resu. Spatial Attention in the two first layers 0.5s 2s 1s 4s Attention Mask - 80 Attention Intensity L 0 Spatial attention 2D L0 Label: Low_Frequency_Lines 0.5s 2s 1s 4s Attention Mask - 6 Attention Intensity Spatial attention 2D L1

Label: Low_Frequency_Lines

Recap

- Neural Networks are **not inherently black boxes**.
- Explainable AI (XAI) enables the visualization of key input features.
- This helps practitioners address misclassifications without solely relying on deeper networks to enhance performance.
- XAI is the **connection** between AI practitioners and physicists. XAI is a dictionary between AI and another discipline!

