Towards complementary explanations using Deep Neural Networks

Wilson Silva [1,2] Kelwin Fernandes[1] Maria J. Cardoso [1,3,4]
Jaime S. Cardoso [1,2]

[1]Faculdade de Engenharia, Universidade do Porto, Portugal

[2]INESC TEC, Porto, Portugal

[3]Faculdade de Ciências Médicas, Universidade NOVA de Lisboa, Portugal

[4]Breast Unit, Champalimaud Foundation, Portugal

September 11, 2018

Overview

Introduction - Machine Learning Interpretability

Strategies to generate reasonable and perceptible explanations can be grouped in three clusters of interpretable models.
[Kim and Doshi-Velez, 2017]

Pre-Model

Visualization; Exploratory Data Analysis

In-Model

Rules; Cases; Sparsity; Monotonicity

Post-Model

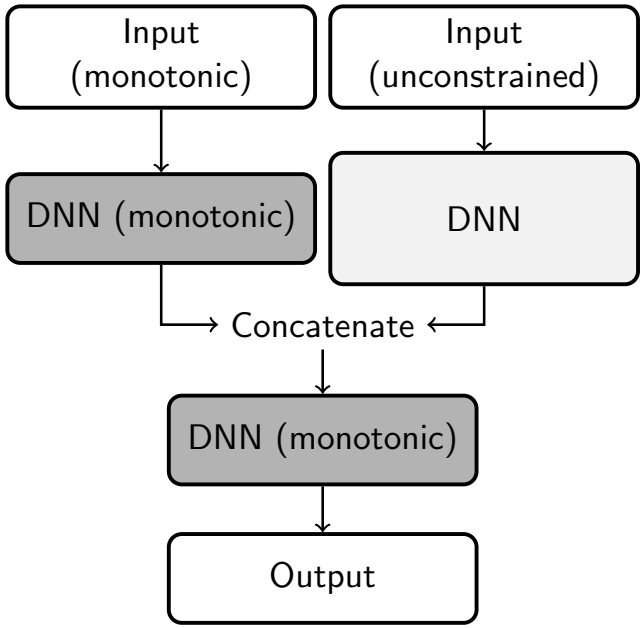
Sensitivity Analysis; Mimic Models; Investigation on hidden layers

Introduction - Satisfying the Curiosity of Decision Makers

- Human beings have different ways of thinking and learning [Pashler et al., 2013]
- Visual Explanation vs. Verbal Explanation
- As many explanations as needed to convince audience
- Some observation require more complex explanations

Complementary Explanations using Deep Neural Networks

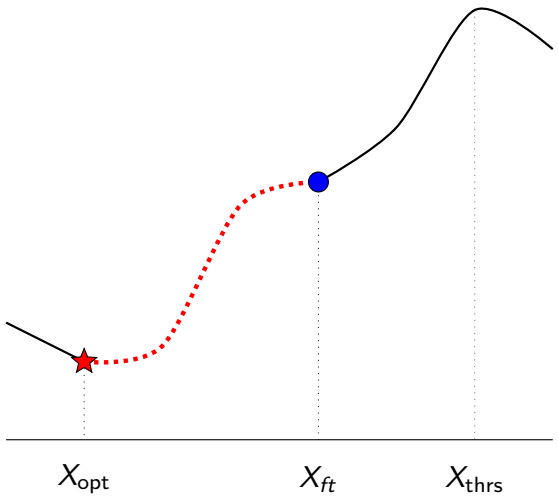
- DNN have the ability to jointly integrate different strategies of interpretability: case-based, monotonicity and sensitivity analysis.



Explanation by local contribution

- To measure the contribution, C_{ft}, of a feature, ft, on the prediction, y, we can find the assignment X_{opt} that approximates X to an adversarial example, i.e., that minimizes the following equation, in which $\bar{y}$ is the opposite class and $f(X)$ the estimated probability.

$$\mathscr{L} = (\bar{y} - f(X))^2 \tag{1}$$



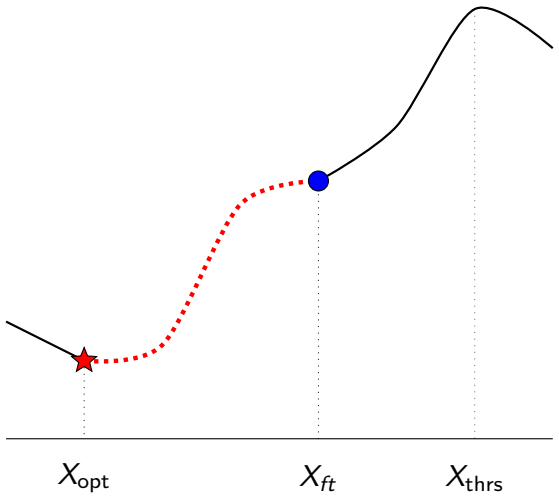
Explanation by local contribution

- Since some features may have a generalized higher contribution than others we balanced the contribution on the target variable with the range of the feature domain traversed from the initial value to the local/global minimum, X_{opt}.

$$C_{ft} = |f(X) - f(X')| \cdot \frac{X_{ft} - X_{opt}}{X_{\max} - X_{\min}} \tag{2}$$

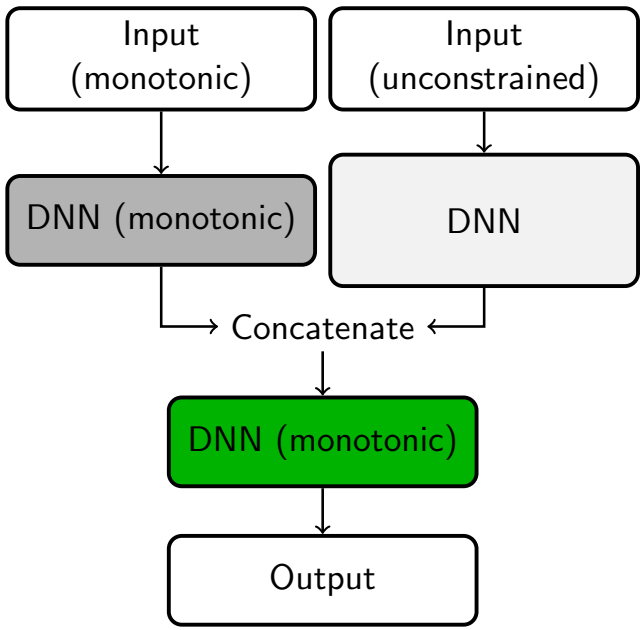
Explanation by local contribution

- The inductive rule constructed for ft covers the space between X_{ft} and the value X_{thrs} where the probability of the predicted class is maximum.



Explanation by similar examples

- DNN are able to learn intermediate representations adapted to the predictive task. Thus, we can use the nearest neighbors in the learned semantic space as an explanation for the decision.



Explanation by similar examples

While the **latent space is not fully interpretable**, we can evaluate which **features** (and at which degree) impact the distance between two observations using **sensitivity analysis**. In this sense, two types of explanations can be extracted:

- **Same Class Example**: the nearest neighbor from the same class in the latent space and what features make them similar.
- **Opponent Class Example**: the nearest neighbor from the opponent class in the latent space and what features make them different.

The three Cs of interpretability

A good explanation should maximize the following properties:

1. **Completeness**: It should be susceptible of being applied in other cases where the audience can verify the validity of that explanation.
2. **Correctness**: It should generate trust, i.e., it should be accurate.
3. **Compactness**: It should be succinct.

The three Cs - Illustration of explanation quality

Example for decision rules:

1. **Completeness**: Where the decision rule precondition holds / blue rows.
2. **Correctness**: Label agreement between the blue rows.
3. **Compactness**: Number of conditions in the decision rule.

$$\texttt{If } A \geq 1 \wedge B \leq 5 \wedge B \geq 2 \texttt{ then } y = 1$$

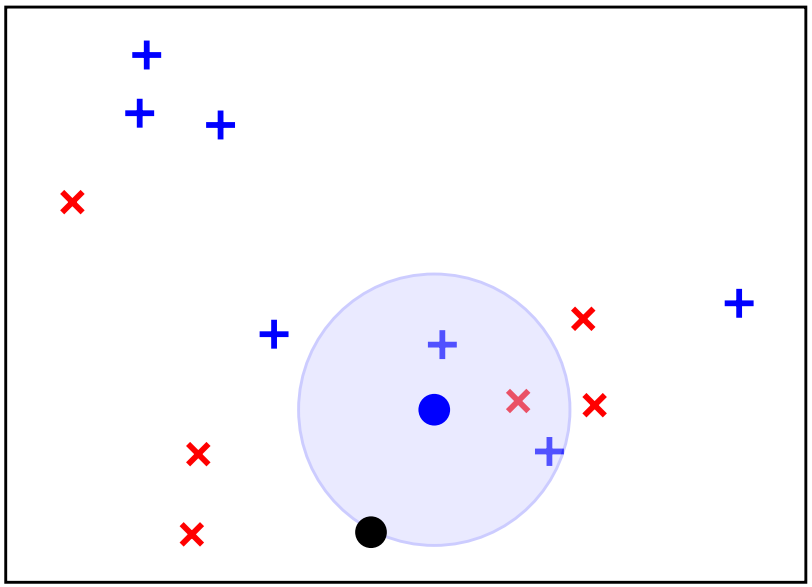
A	B	Y
0	4	1
2	3	1
4	2	0
3	6	1
3	3	1

Completeness: 3/5
Correctness: 2/3
Comptactness: 3

The three Cs - Illustration of explanation quality

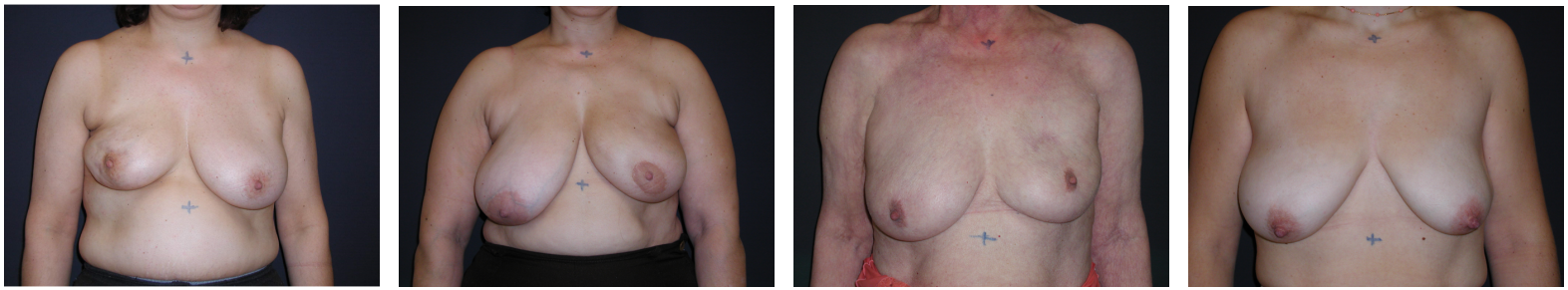
Example for kNN (black dot is the new observation and the blue dot is the nearest-neighbor):

1. **Completeness**: Observations within the same distance of the neighbor explanation.
2. **Correctness**: Label agreement between the points inside the *n*-sphere.
3. **Compactness**: Feature dimensionality of a neighbor-based explanation.



Completeness: 4/14
Correctness: 3/4
Comptactness: 2

Experimental Assessment - Post-surgical Aesthetic Evaluation [Cardoso and Cardoso, 2007]



- 143 images
- 23 high-level features describing breast asymmetry in terms of shape, and global and local color (i.e., scars)
- 4 classes (Poor, Fair, Good, and Excellent)

Experimental Assessment - Classification of Dermoscopy Images [Mendonça et al., 2013]

- 200 images
- 14 high-level features describing the presence of certain colors on the nervus and abnormal patterns such as asymmetry, dots, streaks, among others
- 3 classes (Melanoma, Atypical Nervus, and Common Nervus)

Quantitative Assessment - Breast Aesthetics

Binarization	Model	Predictions		Explanations			
		ROC	PR	Type	Corr	Compl	Compt
Excellent *vs.* Good, Fair, Poor	DT	71.96	92.19	Rule	75.52	3.82	31.97
	1-NN	67.37	90.74	Similar	89.27	3.25	95.94
				Opponent	72.96	80.84	96.00
	DNN	**80.61**	**96.55**	Similar	85.69	95.20	124.94
				Opponent	92.04	46.87	149.68
				Rule	99.91	3.69	62.59
Excellent, Good *vs.* Fair, Poor	DT	85.18	75.20	Rule	51.75	3.16	30.00
	1-NN	52.81	39.49	Similar	85.69	2.98	95.94
				Opponent	54.76	91.26	95.97
	DNN	**86.78**	**82.82**	Similar	72.52	17.34	80.36
				Opponent	81.16	31.28	138.00
				Rule	98.89	2.33	48.59
Excellent, Good Fair *vs.* Poor	DT	**94.20**	**74.92**	Rule	76.92	6.71	17.08
	1-NN	54.42	20.63	Similar	94.45	3.01	95.94
				Opponent	84.42	85.33	96.00
	DNN	91.03	73.00	Similar	87.25	1.46	79.79
				Opponent	92.82	67.86	157.81
				Rule	99.88	5.48	58.44

Quantitative Assessment - Dermoscopy Images

Binarization	Model	Predictions		Explanations			
		ROC	PR	Type	Corr	Compl	Compt
Common *vs.* Atypical, Melanoma	DT	97.60	97.90	Rule	43.00	5.03	13.10
	1-NN	94.37	94.29	Similar	94.97	5.56	15.29
				Opponent	59.42	81.38	15.94
	DNN	**99.74**	**99.83**	Similar	97.11	39.00	19.32
				Opponent	74.59	70.61	37.69
				Rule	98.86	38.83	16.27
Common, Atypical *vs.* Melanoma	DT	95.55	81.63	Rule	82.00	5.82	19.00
	1-NN	80.94	63.67	Similar	94.81	5.70	15.23
				Opponent	69.75	86.98	21.25
	DNN	**96.02**	**89.30**	Similar	91.49	8.15	33.27
				Opponent	84.02	62.12	46.24
				Rule	97.89	44.84	23.65

Qualitative Assessment - Breast Aesthetics

Input image

Prediction:
{Poor, Fair}

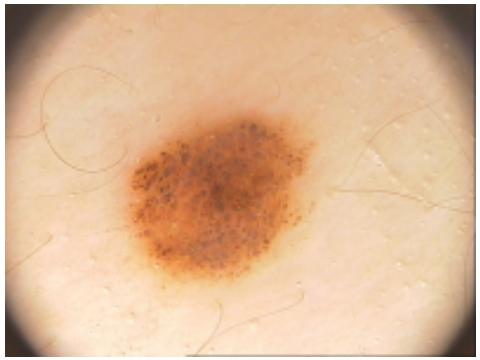
Rule: High visibility of the scar ($sX2a > 0.98$), low inter-breast overlap ($\overline{pBOD} \leq 0.9$), low inter-breast compliance ($\overline{pBCE} \leq 0.43$) and high upward nipple retraction ($pUNR > 0.71$).

Similar case. Why?: Similar scar ($sEMDL$), inter-breast overlap ($pBOD$), color ($cEMDb$), contour difference ($pBCD$) and upward nipple retraction ($pUNR$).

Opponent case. Why?: Strong difference on the scar visibility ($sX2a$), breast overlap ($pBOD$), upward nipple retraction ($pUNR$), compliance evaluation ($pBCE$) and lower contour ($pLBC$).

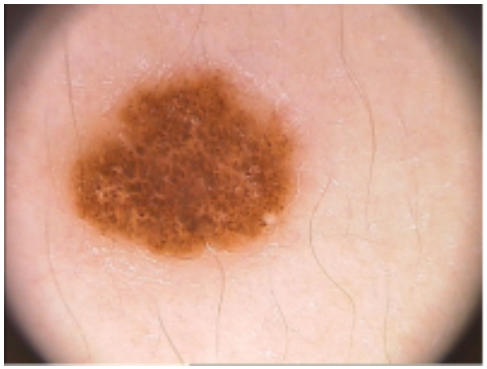
Qualitative Assessment - Dermoscopy Images

Input image

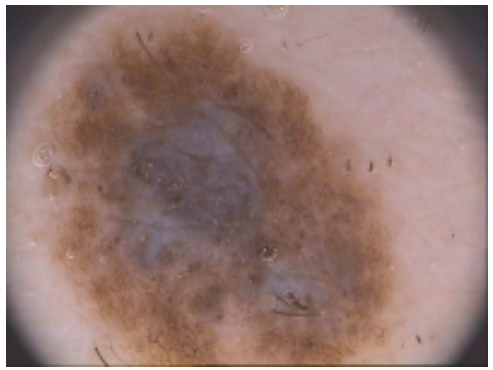


Prediction:
{Common, Atypical}

Rule: It is symmetric, doesn't have black color, blue whitish veil, atypical pigmented network or streaks.



Similar case. Why?: Both images have light and dark brown color and atypical presence of dots/globules.



Opponent case. Why?: It doesn't have light brown color or atypical dots/globules. It has blue whitish veil and pigmented network.

Conclusions

- DNN model able to generate complementary explanations both in terms of type and granularity
- Objective framework to evaluate explanations
- Two biomedical applications: Post-surgical Aesthetic Evaluation and Dermoscopy Images
- Quantitative and qualitative results show an improvement in the quality of the explanations generated compared to other interpretable models

References

Been Kim and Finale Doshi-Velez

Interpretable machine learning: The fuss, the concrete and the questions.

ICML Tutorial on interpretable machine learning (2017)

Harold Pashler, Mark McDaniel, Doug Rohrer, and Robert Bjork

Learning styles: Concepts and evidence.

Psychological Science in the Public Interest **9**(3), 105-119 (2008)

Jaime S. Cardoso and Maria J. Cardoso

Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment.

Artificial Intelligence in Medicine **40**, 115-126 (2007)

Teresa Mendonça, Pedro M. Ferreira, Jorge S. Marques, André R. S. Marcal, and Jorge Rozeira

PH2 - a dermoscopic image database for research and benchmarking.

International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) 5437-5440 (2013)

Thank you. Questions?