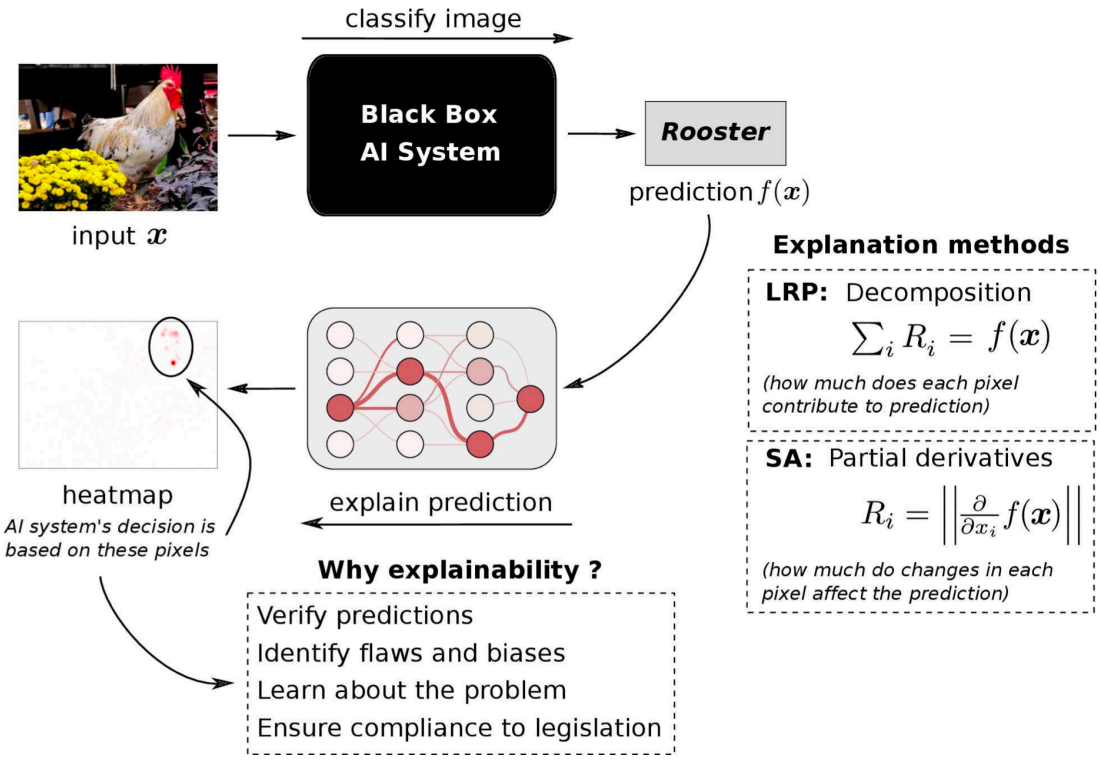
EXPLAINABLE ARTIFICIAL INTELLIGENCE: UNDERSTANDING, VISUALIZING AND INTERPRETING DEEP LEARNING MODELS

Wojciech Samek1, Thomas Wiegand1,2, Klaus-Robert Müller2,3,4

- dovody preco je dolezite aby bol AI model interpretovatelny a vysvetlitelny
- why black- box models are not acceptable for certain applications, e.g. in the medical domain where wrong decisions of the system can be very harmful.
-
- METHODS FOR VISUALIZING, INTERPRETING AND EXPLAINING DEEP LEARNING MODELS

LRP a Sensitivity analysis

- dobry obrazok do dp



LRP: ako velmi psipieva kazdy pixel do predikcie

- Na rozdiel od analýzy citlivosti táto metóda vysvetľuje predpovede týkajúce sa stavu maximálnej neistoty, t.j. identifikuje pixely, ktoré sú kľúčové pre predpoveď „kohút“.
  - Recent work [26] also shows close relations to Taylor decomposition, which is a general function analysis tool in mathematics. - pozriet zdroj 26!
- Vysvetluje rozhodnutia klasifikatora pomocou dekompozicie
- Matematicky redistribuuje predpoveď f (x) spatne pomocou lokalnych pravidiel redistribúcie, až kým každej vstupnej premennej (napr. Obrazový

pixel) nepriradí skóre relevantnosti Ri.
- Kľúčová vlastnosť tohto procesu redistribúcie sa označuje ako relevance conservation a možno ju zhrnúť ako

$$\sum_i R_i = \ldots = \sum_j R_j = \sum_k R_k = \ldots = f(\mathbf{x})$$

- Táto vlastnosť hovorí, že v každom kroku procesu redistribúcie (napr. V každej vrstve hlbokej neurónovej siete) je zachovaná celková relevantnosť, t.j predikcia je zachovan8
- počas redistribúcie sa umelo nepridáva ani neodstraňuje žiadna relevantnosť
- Skóre relevantnosti R každej vstupnej premennej určuje, do akej miery táto premenná prispela k predikcii.
- Na rozdiel od analýzy citlivosti teda LRP skutočne rozkladá funkčnú hodnotu f (x).
- www.explain-ai.org.
-

Sensitivity analysis: ako velmi zmena pixla ovplyvni vyslednu predikciu
- vysvetluje predikciu zalozenu na lokalne vyhodnocovanom gradiente modelu - pomocou parcialnych derivacii
- matematicky, SA kvantifikuje dolezitost kazdej vstupnej premennej - napr pixla na obrazku ako:

$$R_i = \left|\left| \frac{\partial}{\partial x_i} f(\mathbf{x}) \right|\right|$$

- tento vzorec hovori o tom, že najdôležitejšie vstupné vlastnosti sú tie, na ktoré je výstup najcitlivejší.
- Na rozdiel od prístupu uvedeného v nasledujúcom článku analýza citlivosti nevysvetľuje funkčnú hodnotu f (x) samotnú, ale skôr jej variáciu.
- Heatmapa vypočítaná s analýzou citlivosti naznačuje, ktoré pixely je potrebné zmeniť, aby sa obraz (z pohľadu systému AI) viac či menej podobal predpovedanej triede.
- priklad obrazok s kohutom

For instance, in the example shown in Fig. 1 these pixels would be the yellow flowers which occlude part of the rooster. Changing these pixels in a specific way would reconstruct the occluded parts of the rooster, which most probably would also increase the classification score, because more of the rooster would be visible in the image.

Note that such a heatmap would not indicate which pixels are actually pivotal for the prediction “rooster”. The presence of yellow flowers is certainly not indicative of the presence of a rooster in the image.

EVALUATING THE QUALITY OF EXPLANATIONS

- Na to aby me vedeli porovnat vysledne heatmapy z LRP a SA potrebujeme nejaku objektivnu metriku na zaklade ktorej by sme vedeli vyhodnotit kvalitu vysvetlenia
- autori v zdroji 31 predstavuju priklad takejto metriky kvality zalozenej na **analyze poruch - perturbation analysis**
  - Porucha vstupných premenných, ktoré sú veľmi dôležité pre predpoveď, vedie k prudšiemu poklesu predikčného skóre ako pri narušeni vstupných premennych ktoré sú menej dôležité.
  - Vysvetľovacie metódy ako SA a LRP poskytujú skóre pre každú vstupnú premennú. Preto môžu byť vstupné premenné usporiadané podľa tohto skóre relevantnosti.
  - Je možné iteratívne narušiť vstupné premenné (počnúc od najrelevantnejších) a sledovať predikčné skóre po každom kroku poruchy. Priemerný pokles skóre predikcie (alebo pokles presnosti predikcie) sa môže použiť ako **objektívne meradlo kvality** vysvetlenia, pretože veľký pokles naznačuje, že metóda vysvetlenia bola úspešná pri identifikácii skutočne relevantných vstupných premenných.

Zaujimave zdroje:

- G. Montavon, S. Bach, A. Binder, W. Samek, and K.R. Müller. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition,65:211–222, 2017.
- L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek. Explaining predictions of non- linear classifiers in nlp. In Proceedings of the 1st Workshop on Representation Learning for NLP, pages 1–7. ACL, 2016.
- S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE, 10(7):e0130140, 2015.
- W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller. Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems, 2017, in press.
- A. Mahendran and A. Vedaldi. Visualizing deep convolutional neural networks using natural pre- images. International Journal of Computer Vision, 120(3):233–255, 2016.
- L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595, 2017.
- M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European Conference Computer Vision - ECCV 2014, pages

818–833, 2014.

Starsi:
K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034, 2013.