



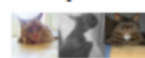


Εργασίες AIDA

Εαρινό εξάμηνο 2022

eXplainable **A**rtificial **I**ntelligence - XAI

How to XAI

Table 1: Examples of explanations divided for different data type and explanation

TABULAR	IMAGE	TEXT																																																									
Rule-Based (RB) A set of premises that the record must satisfy in order to meet the rule's consequence. $r = Education \leq College$ $\rightarrow \leq 50k$	Saliency Maps (SM) A map which highlight the contribution of each pixel at the prediction. 	Sentence Highlighting (SH) A map which highlight the contribution of each word at the prediction. the movie is not that bad																																																									
Feature Importance (FI) A vector containing a value for each feature. Each value indicates the importance of the feature for the classification. <table data-bbox="355 860 533 972"><tr><td>capitalgain</td><td>0.00</td></tr><tr><td>education-num</td><td>14.00</td></tr><tr><td>relationship</td><td>1.00</td></tr><tr><td>hoursperweek</td><td>3.00</td></tr></table>	capitalgain	0.00	education-num	14.00	relationship	1.00	hoursperweek	3.00	Concept Attribution (CA) Compute attribution to a target "concept" given by the user. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)? 	Attention Based (AB) This type of explanation gives a matrix of scores which reveal how the word in the sentence are related to each other. <table data-bbox="1019 860 1299 983"><tr><td></td><td>the</td><td>movie</td><td>is</td><td>not</td><td>that</td><td>bad</td></tr><tr><td>the</td><td>0.8</td><td>0.2</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td></tr><tr><td>movie</td><td>0.2</td><td>0.8</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td></tr><tr><td>is</td><td>0.1</td><td>0.1</td><td>0.8</td><td>0.1</td><td>0.1</td><td>0.1</td></tr><tr><td>not</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.8</td><td>0.1</td><td>0.1</td></tr><tr><td>that</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.8</td><td>0.1</td></tr><tr><td>bad</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.1</td><td>0.8</td></tr></table>		the	movie	is	not	that	bad	the	0.8	0.2	0.1	0.1	0.1	0.1	movie	0.2	0.8	0.1	0.1	0.1	0.1	is	0.1	0.1	0.8	0.1	0.1	0.1	not	0.1	0.1	0.1	0.8	0.1	0.1	that	0.1	0.1	0.1	0.1	0.8	0.1	bad	0.1	0.1	0.1	0.1	0.1	0.8
capitalgain	0.00																																																										
education-num	14.00																																																										
relationship	1.00																																																										
hoursperweek	3.00																																																										
	the	movie	is	not	that	bad																																																					
the	0.8	0.2	0.1	0.1	0.1	0.1																																																					
movie	0.2	0.8	0.1	0.1	0.1	0.1																																																					
is	0.1	0.1	0.8	0.1	0.1	0.1																																																					
not	0.1	0.1	0.1	0.8	0.1	0.1																																																					
that	0.1	0.1	0.1	0.1	0.8	0.1																																																					
bad	0.1	0.1	0.1	0.1	0.1	0.8																																																					
Prototypes (PR) The user is provided with a series of examples that characterize a class of the black box $p = Age \in [35, 60], Education \in [College, Master] \rightarrow \geq 50k$ $p = $  \rightarrow "cat" $p = "... not bad ..."$ \rightarrow "positive"																																																											
Counterfactuals (CF) The user is provided with a series of examples similar to the input query but with different class prediction $q = Education \leq College \rightarrow \leq 50k$ $c = Education \geq Master \rightarrow \geq 50k$ $q = $  $\rightarrow "3"$ $c = $  $\rightarrow "8"$ $q =$ The movie is not that bad \rightarrow "positive" $c =$ The movie is that bad \rightarrow "negative"																																																											

Εκφώνηση

Σκοπός της εργασίας είναι η δημιουργία ενός εμπλουτισμένου dataset πάνω σε εικόνες, με σκοπό να βοηθήσει στην ερμηνεία προεκπαιδευμένων classifiers για την κατηγοριοποίηση των εικόνων με χρήση XAI αλγορίθμων.

Φόρμα δήλωσης ομάδας/θέματος:

<https://docs.google.com/forms/d/e/1FAIpQLSdz97QihFmY7avNb1kGzHB4PwPFppBg8Era9B5vPmYbPg2n3g/viewform>

Dataset

Βήμα 1: Δημιουργία dataset: Επιλέγετε η κάθε ομάδα 2 κλάσεις του places

(<http://places2.csail.mit.edu/>), οι οποίες μπορούν να αντιστοιχιστούν σε υποσύνολα του COCO dataset. Διαλέξτε τα υποσύνολα του COCO μέσω του interface <https://cocodataset.org/#explore>
Οι κατηγορίες του places μπορούν να βρεθούν εδώ:

https://github.com/CSAILVision/places365/blob/master/categories_places365.txt

Βήμα 2: Δημιουργία γνώσης: Ενώστε το dataset με το WordNet

<https://www.nltk.org/howto/wordnet.html> ή με άλλη γνώση (ConceptNet πχ - δείτε More Info)

Βήμα 3: Αναπαράσταση WordNet και individuals σε .owl.

Enrichment

Βήμα 4: Εμπλουτισμός της γνώσης:

- Προσθήκη αξιωμάτων συμβατών με το domain
- Scene-graph generation για προσθήκη ρόλων. Εδώ μπορεί να γίνει χρήση προεκπαιδευμένων μοντέλων, κατά προτίμηση αυτών με τις καλύτερες μετρικές και που να έχουν διαθέσιμο κώδικα.

Βήμα 5: Ορίστε την κατηγορία με αξιώματα (πχ pizzeria = has.Pizza)

Βήμα 6: Τρέχετε τον έτοιμο [places classifier](#) στο dataset σας. Συγκρίνετε τα αποτελέσματα των αξιωμάτων με τον classifier.

Explanation

Βήμα 7: Εφαρμόστε μεθόδους XAI πχ LIME, Rule Matrix, Scope Rule (pixel level - saliency maps, semantic level - feature importance) και άλλα που να ταιριάζουν στα αποτελέσματά σας.

Repeat Βήματα 4,5,6, 7

Analysis

Βήμα 8: Αναλύστε τα συμπεράσματά σας μετά την επανάληψη των βημάτων

Βήμα 9: Προτείνετε δικές σας ιδέες σχετικά με το πώς μπορεί να αξιοποιηθεί το dataset που δημιουργήσατε για XAI ή για άλλη εφαρμογή. Πώς συνεισφέρει η γνώση στις εξηγήσεις?

Παραδοτέο: jupyter notebook ή ακόμα καλύτερα Github pages

<https://medium.com/analytics-vidhya/convert-your-jupyter-notebook-to-github-pages-with-github-action-fa2ce9b4182a> .

COCO dataset

Πώς επιλέγουμε κατηγορίες στο COCO?

Μέσω του COCO explorer <https://cocodataset.org/#explore>

COCO Explorer

COCO 2017 train/val browser (123,287 images, 886,284 instances). Crowd labels not shown.



More info

Γράφοι γνώσης που μπορούν να χρησιμοποιηθούν για εμπλουτισμό:

Widely used structured KGs:

- Wordnet (hierarchical)
- ConceptNet (commonsense)
- DBPedia (Hierarchical, Encyclopedic/Factual)
- Wikidata (Encyclopedic/Factual)
- WebChild (Commonsense)
- HasPartKB (Commonsense, part-whole)
- Visual Genome (Visual)