Anchors vs Attention: Comparing XAI on a Real-Life Use Case

Gaëlle Jouis, Harold Mouchère, Fabien Picarougne, Alexandre Hardouin

Anchors vs Attention: comparing XAI on a real-life use case

Gaëlle Jouis[1,2], Harold Mouchère[1], Fabien Picarougne[1], and Alexandre Hardouin[2]

[1] LS2N, Université de Nantes, CNRS, F-44000 Nantes
[2] Pôle Emploi, Direction des Systèmes d'Information, Nantes
gaelle.jouis@pole-emploi.fr

Abstract. Recent advances in eXplainable Artificial Intelligence (XAI) led to many different methods in order to improve explainability of deep learning algorithms. With many options at hand, and maybe the need to adapt existing ones to new problems, one may find in a struggle to choose the right method to generate explanations. This paper presents an objective approach to compare two different existing XAI methods. These methods are applied to a use case from literature and to a real use case of a French administration.

Keywords: Machine Learning · Deep Learning · Explainability · Attention Networks.

1 Introduction

Artificial Intelligence, and specifically Machine Learning, has been thriving for the past few years. Deep Learning has proven its ability to perform in many tasks, such as image processing, object recognition, and natural language processing [14]. Unlike other approaches such as linear models, deep learning models are considered as *black boxes*, because their processing is quite opaque. Hence, explaining the result of a deep learning algorithm is a difficult task.

This paper focuses on a specific real-life use case of text classification for a French Institution, *Pôle Emploi*. The classification consists in the detection of uncompliant job offers with the use of machine learning, and will later be called "LEGO". The institution has a legal duty of transparency about its algorithms and seeks to provide explanations alongside the tool's results. The "Why" and the "How" of eXplainable Artificial Intelligence (XAI) are now being tackled by the scientific community. To the best of our knowledge, proposed solutions are not systematically evaluated. Thus, choosing which method would best suit a particular AI project is not a straightforward task.

This contribution compares two different explanation methods when applied to a use case. Methods are the Anchors, a black box method, and an Attention-based white box method, both being popular and based on distinct mechanisms. The specific LEGO use case and a second use case from literature are used in this qualification process, suggesting best practices for XAI method comparison.

2 Related Works

2.1 eXplainable Artificial Intelligence

One can regroup the numerous existing XAI approaches according to the global logic of survey papers [4, 5]. Hence are defined three categories:

1. Explain a *black box* model based on its inputs and outputs.
2. Observe internal mechanisms of a system (*grey box*) after it was trained.
3. Design a transparent solution (*white box*) explaining itself.

Explaining black box models induces the use of a proxy model. One well-known method is LIME [10] and its improvement: Anchors [11]. *LIME* (Local Interpretable Model-agnostic Explanations) is an approximation of a black box model with a linear regression. The regression weights the inputs by importance. Similarly, Anchor explanation explains a result with a rule. The rule presents a set of words leading to the decision of the model. These methods are designed to explain one instance at a time and are only accurate for close examples.

Convolutional Neural Networks (CNN) internal processing has been analyzed in [13]. The authors used a neural network they named *Deconvolutional Network* to visualize patterns that activate neurons layer-wise. Also based on CNN, authors of [12] combine their works to those of [13], to detect regions and patterns of an image helping on class detection. Similar work has been done on semantic analysis with LSTM (Long Short-Term Memory) networks [7].

On the other hand, transparent solutions are inherent to the developed model. In [8], the authors create an attention-based word embedding called *Structured self-attentive embedding*. Words associated with high Attention weights are used by the model to classify the text. Another attention-based visualization is shown in [9]. Compared to black box strategies, which are approximating the trained model, Attention is a core component of this model. After training, there is no need for more computing as inference will generate Attention weights used for meaningful visualizations.

2.2 Evaluate Explanations

Evaluating explanations can be done with two main approaches: 1) Criteria and metrics, 2) User evaluation.

Criteria and metrics Explanations or models generating them are often evaluated with criteria and metrics in the literature. To evaluate proxy models, one mostly used criteria is fidelity to the black box model, measured with accuracy, or f1 score [5, 11]. Interpretability is also measured, often as a size of the proxy model, such as number of weights [5]. The coverage can be measured as the number of instances that are in agreement with an explanation [11]. Metrics can also evaluate the explanation itself. In the case of natural language explanations, it is possible to use readability score such as the *Flesch-Reading-Ease* score, used

in [3]. When required explanations are available, expected and obtained explanations can be compared as sets of features. Computing the Intersection over Union (IoU) gives a score from 0 to 1. An IoU of 1 means explanations are identical. This metric is used in [1] to evaluate interpretability in image based problems. If only a few explanations are possible, the case can be considered as a classification problem, and usual accuracy metrics can be used [2].

Evaluation based on metrics allows to work on huge test datasets. Without the need of finding users, it is also faster and cheaper to develop quantitative evaluations on any XAI method. However, explanations are designed to be an interface between algorithms and humans, hence they need to be evaluated by or with humans.

User evaluation When conducting a user study on model explanations, evaluation can be objective or subjective. Subjective evaluation can be a poll asking users if they are satisfied with a given explanation, or which explanation do they prefer among a couple ones [11]. They might also be asked to choose between two classifiers, one being significantly better than the other, given only their explanations [10]. These evaluations are appropriate when the purpose is improving acceptance of a model. One the other hand, objective metrics can be extracted from user studies. In [6], users are given an explanation and must predict the next output of the system. Considering user's answers as results of binary classifiers, the authors compute a Roc Curve and its Area Under Curve to measure the success of their explanations. When the user must predict the output of the model, the response time of the user can be used as a measure of the user's confidence [11].

3 Experiments

We want here to compare two explanation methods: generation of Anchors upon any model from [11] and the use of attention with a transparent model from [8]. For each following use case, a transparent attention-based model will be trained, and Anchors will be generated on the predictions of this same model. Quoting [11] for the example, lets take the sentence "This movie is not bad", which is classified "positive" by an attention-based model. The Anchor explanation would be $A = \{not, bad\} \rightarrow Positive$. Every word in the sentence would have an attention weight, and "not" and "bad" would have the highest weights.

The generation of Anchors is made with the python library developed by the authors of [11]. Following the research of [8], a neural network with a bi-LSTM and the same Attention mechanism was trained. The architecture is described in the table below (cf. Table 1). Adapting the network for each use-case resulted in some differences of dimensions, which are detailed in the 3rd and 4th columns of Table 1. The Attention mechanism results in an Attention matrix A, which is the output of the layer 5 (cf. Table 1). Words of interest are filtered using a threshold t on attention values. For the LEGO use case, when the model predicts no reject, the explanation is forced to be empty.

Table 1. Network architecture specifications of YELP and LEGO classifiers.

ID	Layer type	YELP	LEGO	Comment
1	Input Layer	300	80	Size is the number of words in texts
2	Embedding	100	300	Embedding, respectively word2vec and glove
3	BLSTM	u =150	u =50	Output is Hidden states H
4	Dense 0	d_a =350	d_a =300	$tanh$ activation
5	Dense 1	r =1	r =1	Output is attention matrix A
6	Attention and average pooling	out: $[2u, r]$	out: $[2u, r]$	Combines attention and hidden states, $M = A^T * H$
7	Dense 2	1000		$ReLu$ activation, only for YELP
8	Dense 3	5	28	Output Layer

3.1 LEGO

Pôle Emploi is the french job center. One of its tools aims to automatically reject uncompliant job offers. Indeed, *Pôle Emploi* is legally bound to reject offers not complying with the Labor Code or being discriminative. Training dataset contains 480000 sentences extracted from real offers. Retrieving the reason for rejection is a multiclass classification task, with 28 topics being targeted in this study. Offers used for the training of the classifier are already labeled in *Pôle Emploi*'s database, with labels predicted by the existing rule-based system.

The classifier for this use case is similar to the one for the YELP dataset. The embedding matrix used is a 300-dimensional GloVe embedding.[3] The optimizer is Adam, with a learning rate of 0.0005. This network achieves an accuracy of 83.67% on its test set.

The existing system produces some errors. Thus, to accurately analyze explanations, a corrected test set was necessary. As the correction of labels is time-consuming, a subset of 208 sentences has been manually labeled and attributed the desired explanation. This explanation consists of highlighting keywords that led to rejection. As explanations are meaningful for uncompliant offers only, explanations for compliant offers are considered to be empty. The real world's class distribution has not been respected, and compliant explanations are underrepresented in this test set. Hence, the model's accuracy for this test set is lower and irrelevant (70.67%).

3.2 YELP

The YELP dataset contains user reviews about restaurants, associated with 1 to 5 star-ratings. The training set contains 453 600 reviews. As shown in Table 1, the embedding is a 100-dimensional English-based word2vec.[4] The optimizer is Adam, with a learning rate of 0.0005. This network achieves an accuracy of 74.63% on its test set. In comparison, authors of [8] present in their paper an

[3] https://dl.fbaipublicfiles.com/fasttext/vectors-crawl/cc.fr.300.vec.gz
[4] https://wikipedia2vec.github.io/wikipedia2vec/pretrained/

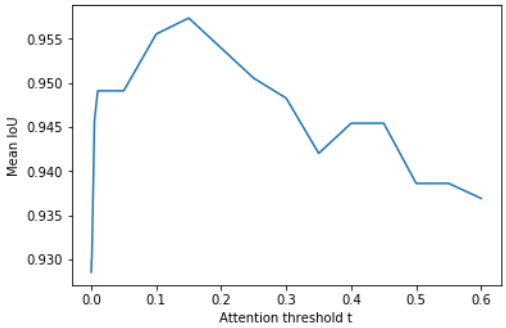

Fig. 1. Mean IoU between Attention and ground truth explanations in test set, for varying Attention threshold t. Words with Attention greater or equal to t are kept as explanation. First point is at 10^-4.

accuracy of 64.21% on their own test set. As anchors explanation on large texts often leads to memory issues, anchors have been applied to a subset of 1060 reviews over the 2653 reviews of the full test set.

4 Evaluating explanations

4.1 Quantitative analysis

In the context of the LEGO dataset, each method is compared with ground truth. To get fair measures, stop words are not taken into account. As the evaluation of the model is not the point in this experimentation, the test set is a subset of 147 sentences over 208, that have been correctly predicted. A threshold t is used in order to filter words. t is determined by optimizing IoU on the test set, as shown in the graph in Figure 1. The graph indicates that words with attention greater or equal to 0.15 are a good explanation in the LEGO use case.

As used in [1], IoU will allow comparing ground truth, to generated explanation, with Anchors or with Attention. Accuracy and F1-score are also displayed in Table 2, plus recall and precision used in the F1-score. Recall is an interesting metric as it is not impacted by true negatives. Anchors and Attention are also compared to one another, assuring they give similar results. With no ground truth, metrics such as accuracy and f1-score are irrelevant, as noted in Table 2. High IoU between anchors and attention explanations indicates that explanations are similar with both methods. Overall, when comparing to ground truth, Attention explanations are slightly better than Anchors, cf. Table 2.

For the YELP use case, there is no expected explanation. Comparison is only possible between Anchors and Attention explanations. IoU indicates whether the given explanations are similar. Mean IoU on the successful test set is 0.2292, which shows strong differences between the two explanation methods. This can be explained by long texts and vast expected vocabulary in explanations. Hence, to assess evaluation methods in YELP use case, qualitative analysis is needed.

Table 2. Evaluation of Anchors and Attention with ground truth, LEGO for correct prediction only, best result in bold.

Metric	Anchors	Attention	Anchors vs Attention
IoU	0.9377	**0.9573**	0.9471
Acc	0.9803	**0.9871**	*irrelevant*
Recall	**0.9696**	0.9641	*irrelevant*
Precision	0.9614	**0.9932**	*irrelevant*
F1	0.9540	**0.9688**	*irrelevant*

4.2 Qualitative analysis

As high IoU shows similarity between two explanations, qualitative analysis can be more efficient with filtering texts with low IoU in the test set. Doing so will point out if one explanation method is more accurate when being different. Hence this filtering will be used in the following qualitative analysis for both use cases.

For the LEGO use case, Anchors explanations are shorter than Attention-based explanations. Mean lengths are respectively 0.15 and 0.33 words in the test set. The mean value is low due to empty explanations. The following table (cf. Table 3) gives a few examples where IoU is lesser than 0.5. This qualitative analysis is in agreement with the quantitative analysis, and point out Attention weights as a better explanation method for this use case. On the YELP dataset, there is no possibility to compare explanations with any wished one. Still, it is interesting to have a look at explanations for extreme reviews (1 and 5 stars, well recognized) when IoU is 0, meaning explanations are very different.

Mean explanation lengths are similar, 2.34 and 2.13 words for Anchors and Attention respectively. The first two lines of Table 4 indicates a lack of meaningful words in Anchors, hence Attention-based explanations are the best choice. The last line of Table 4 show explanations based on different but meaningful words. As average lengths are similar and Attention seems more accurate when explanations are very different, this qualitative analysis indicates that Attention-based explanations are a safer choice in this particular use case.

Table 3. LEGO texts with different explanations. Text is above other information.

Reject	Ground truth	Anchor	Attention
"Contrat a duree indeterminee - Dfd Notre agence de Saint-Medard-en-Jalles recherche une Assistante Administrative pour completer son equipe."			
Gender	['assistante administrative']	['recherche', 'Assistante', 'Jalles']	['assistante', 'administrative']
"Nous recherchons actuellement un Teleconseiller FRANCAIS / NEERLANDAIS (H/F) pour le compte de notre client, a Marcq-en-Baroeul."			
Nationality	['francais / neerlandais']	['un', 'neerlandais', 'recherchons', 'francais']	['neerlandais']

Table 4. YELP texts with different explanations. Text is above other information.

Stars	**Anchor**	**Attention**
Wow! Superb Maids did an amazing job cleaning my house. They stayed as long as it took to make sure everything was immaculate. I will be using them on a regular basis.		
5	[]	['superb', 'amazing', 'everything']
For the record, this place is not gay friendly. Very homophobic and sad for 2019. Avoid at all costs		
1	['not']	['record', 'not', 'homophobic', 'sad', 'avoid']
Had the best experience buying my dress at brilliant bridal in jan 2018. Can't wait to wear my beautiful gown in oct 2018		
5	['brilliant']	['best', 'buying', 'can']

One interesting point is that both explanations are pointing out the same parts when reviewers mention their own rating, as shown in Table 5. For the first example, this even leads to a wrong prediction.

5 Conclusion

The upcoming multiplicity of XAI methods leads to the necessity of choosing one method that suits each specific use case. In this paper, two use cases have been developed. One is available for sharing with the community, and the second one is extracted from a real need in the french job center. The use of metrics as a way to evaluate explanations has proven to be quite useful when explanations test set is available. However, creating this data set can be expensive and needs the contribution of human experts. In this case, and if examined methods are not too many, user studies can be more relevant but are quite expensive themselves.

Another criterion that was not identified at first sight is the computing cost and time of explanations. As generating Anchors explanations was so costly, it led for one use-case to the filtering of the test set. As XAI often meets ethics, responsible AI, and even green IT, one can wonder if explanation methods such as attention mechanism can be preferred based upon efficiency criteria.

Finally, comparing various explanation methods and observing when explanations are similar or in disagreement can help to learn more about an AI model. This process might be used to evaluate the model itself.

Table 5. Mentions of ratings in explanations.

Text	Stars	Prediction	Anchor	Attention
[...] An this is the reason I gave them a mere 2 stars[...]	3	2	['2', 'stars']	['2']
3 Stars is about right. [...]	3	3	['3']	['3', 'decent']
[...] Perfect amount of sweet. 5/5 bobas.	5	5	['5/5', 'sweet', 'Perfect', 'great', 'all']	['5']

Bibliography

[1] Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proc of the IEEE Conf. on Computer Vision and Pattern Recognition. pp. 6541–6549 (2017)

[2] Codella, N.C., Hind, M., Ramamurthy, K.N., Campbell, M., Dhurandhar, A., Varshney, K.R., Wei, D., Mojsilovic, A.: Ted: Teaching ai to explain its decisions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019)

[3] Costa, F., Ouyang, S., Dolog, P., Lawlor, A.: Automatic generation of natural language explanations. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion. pp. 1–2 (2018)

[4] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89 (2018)

[5] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)

[6] Iyer, R., Li, Y., Li, H., Lewis, M., Sundar, R., Sycara, K.: Transparency and explanation in deep reinforcement learning neural networks. In: Proc. of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (2018)

[7] Karpathy, A., Johnson, J., Li, F.: Visualizing and understanding recurrent networks. CoRR **abs/1506.02078** (2015)

[8] Lin, Z., Feng, M., dos Santos, C.N., Yu, M., Xiang, B., Zhou, B., Bengio, Y.: A structured self-attentive sentence embedding. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings (2017)

[9] Olah, C., Carter, S.: Attention and augmented recurrent neural networks. Distill (2016). https://doi.org/10.23915/distill.00001

[10] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proc. of the 22Nd ACM Int. Conf. on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)

[11] Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18). pp. 1527–1535 (2018)

[12] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: The IEEE Int. Conf. on Computer Vision (2017)

[13] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European Conference on Computer Vision. pp. 818–833 (2014)

[14] Zou, Z., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055 (2019)