Explainable AI for Strategic Hybrid Operations

Felix Govaers
Fraunhofer FKIE
GERMANY

felix.govaers@fkie.fraunhofer.de

ABSTRACT

Explainable AI (XAI) provides means to overcome this issue based on additional supplemental information regarding the results of Deep Learning (DL) algorithms. While full transparency remains infeasible for complex DL algorithms, explanations help the user to judge on AI information products in critical situations. It should be noted that XAI is an umbrella term for aspects of transparency, causality, trustworthiness, confidence, fairness, confidence, and privacy. Therefore, the underlying methodologies are manifold. An approach, which has become popular, is the Local Interpretable Model-Agnostic Explanations (LIME) method, since it can well be applied for different models in various applications. In this paper, the LIME algorithm is investigated in the context of decision proposal for strategic operations. After a brief introduction to its concept, applications from the literature are presented. Then, a strategic gaming scenario is considered as a surrogate environment for military warfare. A DL-based chess AI is made "explainable" in order to evaluate the value of information for a human decider. Conclusions with respect to strategic hybrid operations are drawn, which reflect the limitations of the proposed approach.

INTRODUCTION

It is envisioned that decisions in future strategic warfare will heavily be influenced by information products based on methods of Artificial Intelligence (AI). Hybrid operations, in particular, take place in a high dimensional and variational environment, in which the assessment of potential threats and opportunities are difficult to grasp for human operators and where strategic planning must incorporate heterogenous, versatile and high volume data sources. Therefore, algorithmically produced classifications, predictions and suggestions based on AI methods are becoming increasingly important in such complex scenarios. In the last few years, methods of AI have gained significant momentum with a large number of innovations and respectable results for obtaining higher level information from large data sets. A major drawback of Deep Learning (DL) approaches, however, is their inherent black-box property, that is, the opaqueness of its results due to the complexity of the computing model. The latter, for instance, can have hundreds of layers and millions of parameters, which are found and optimized algorithmically during the training phase. As a consequence, even if the results are exact, there is no chance for the user to either comprehend it nor to grasp the causal parts of the input data. This, in turn, can affect the trust of the user to assisting devices heavily in both directions. This issue plays a minor role in certain civil applications such as voice recognition for instance, which is often applied for interaction with devices, since there is no potential risk other than decent disappointment. For other, very specific tasks, such as hand written character recognition, the performance of DL algorithms is beyond the human average, which implies that failures are highly unlikely such that the question regarding causality might become subsidiary. However, in many military applications, human trust is a key issue when it comes to the interaction with AI, since wrong decisions might have severe consequences and the user always remains accountable. This actually is two-fold. On the one hand, the operator often needs to understand the background of AI products, in particular, if those are against his or her own instincts. On the other hand, incomprehensible technology can create a bias towards algorithmic information products since it is hard to determine under which conditions it fails. Thus, the appropriate level of trust can be hard to figure.

Explainable AI (XAI) is the collections of approaches to provide “transparency”, “interpretability”, or “explainability” to the user of a black-box AI model. A joint definition for those terms is hardly available, but many publications refer to

- *transparency* as the degree of possible comprehension for a human to track and understand the process of model creation. That is the information extraction from the data into the manifestation of parameters for inference. A DL feed-forward network lacks this property due to its iterative learning process based on large data sets and the recursive propagation of errors to individual layers.
- *interpretability* as the degree of comprehension of the model itself such that the information flow from input data to the prediction result can be understood. This is infeasible for standard networks due to the number of parameters involved and the hierarchical structure of the layers.
- *explainability* as the degree of possibility to elucidate on a specific prediction result. That is, the coherence to the input data is made visible to the user and to some extent one can see whether a causal relation exists.

XAI cannot “explain” a DL model in its full extent, however, it provides means for the engineer or the operator to better understand the causality behind a given AI product. And quite often this can help to see, whether the model is sensible (or not) in the sense that a reasonable chain of causality implied the algorithmic decision or prediction. Therefore XAI can be an important tool for the engineering of AI models, for their validation with respect to safety or even for in certification processes as well as for providing additional information to an operator in order to support well informed decisions.

While most publications in the literature on XAI are focused on methods for image recognition, such results are difficult to transform onto the domain of tactical and strategic decision making based on a given challenging competitive situation. In this paper, we investigate *explainability* for AI models for a chess board evaluation. Some implications on more complex military strategic simulations are discussed.

This paper is structured as follows. In the next section, a brief overview of selected XAI methods is provided. Then, one of those methods (LIME) is applied to the problem of chess board evaluation to demonstrate the quality of the explanation in terms of supporting information. In the last section, conclusions are drawn and a generalization to more complex war gaming and simulation is discussed.

METHODS OF XAI

In this section, a selection of relevant methods in the wide research field of XAI is presented. We limit ourselves to approaches for uncertainty estimation by means of *Bayesian Deep Learning* and two model agnostic explanation methods (LIME and RISE), since those are considered[1] to be the most important for AI-based assistance systems in hybrid military operations. In particular, the fact that those are model agnostic makes them flexible in general and particular important for military applications.

BAYESIAN DEEP LEARNING *(BDL)*

Bayesian DL networks refer to a family of architectures, which are able to estimate the uncertainty of the AI result. There are different types of uncertainty (epistemic and aleatoric) and even more approaches to estimate them from the data:

- By an application of the Bayes theorem, one can estimate the posterior distribution for the prediction. To this end, it is required to either have a good model on the epistemic uncertainty in the data (in terms of a likelihood function) or to have a joint optimization process within the training

[1] By the humble author’s opinion.

step. The latter can easily be achieved by including the data variance in the loss function as a negative log-normal likelihood density. Instead of the standard loss function for regression given by $||\mathbf{y}-\hat{\mathbf{y}}||^2$, the negative log-likelihood of the normal density yields $\frac{||\mathbf{y}-\hat{\mathbf{y}}||^2}{2\sigma^2}+\frac{1}{2}\log\sigma^2$. By optimizing the prediction error and the variance σ^2 at the same time, a consistent model is obtained and can be used for the Bayes update step.

- Gaussian Processes can be used in order to replace fixed weights of a trained model by a normal probability density functions (pdf), which are propagated through the network by means of the linear and non-linear transformations. To this end, the non-linear transformations are approximated in terms of sufficient statistics (mean and variance) such that the resulting normal distribution can be inferred. As a result, the variance of the prediction can be obtained from the posterior distribution at the output layer [3].
- Another approach to capture the uncertainty of a model is to use stochastic dropout for Bayesian learning [4]. Usually, dropout is a method for regularization, that is, to avoid overfitting of the model and enhance generalization. In the stochastic context, it is applied in order to obtain a set of different models, which in total are a Monte-Carlo representation of the uncertainty. The squared distances of the predictions of each of the model to the joint result can be used in order to calculate the variance.

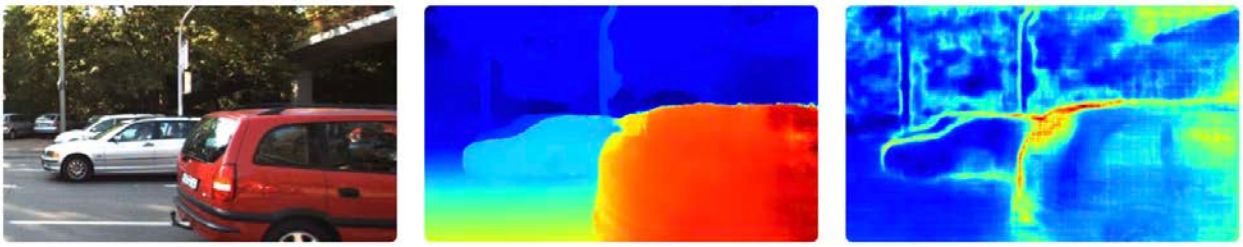


Figure 1: Example of a Bayesian DL for the depth estimation (middle) of scene (left). The back window of the red car is associatted to the wrong depth. However, the varaince estimate (right) indicates that the network is aware of the uncertainty in this area (red field). Picture and example are from [2].

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATION *(LIME)*

The LIME approach [5] has become quite popular due to the fact that it's model agnostic, quite efficient and easy to implement. It's an representative of the class of *Local Surrogate Models*, where the (non-linear) black-box model is locally[2] approximated by a direct interpretable model. The latter can be for instance a linear regression, logistic regression, a decision tree, or a *Support Vector Machine* (SVM) since those algorithms directly yield correlations or decision thresholds which indicate the relevance between input parameters and the computed output result. The approximation itself is achieved by a weighted sampling of data points within a neighboring region. Corresponding labels are created by the black-box model such that a local training data set is obtained to feed the surrogate model. Even the choice of the surrogate model can be automated by an application of all valid candidates and choosing the one with the lowest loss function score. This concept of local approximation of a non-linear black-box model is shown in Figure 2, which is taken from [5].

[2] „Locally“ means, that the surrogate model only approximates the original model in a neighboring region around a given sampling point.

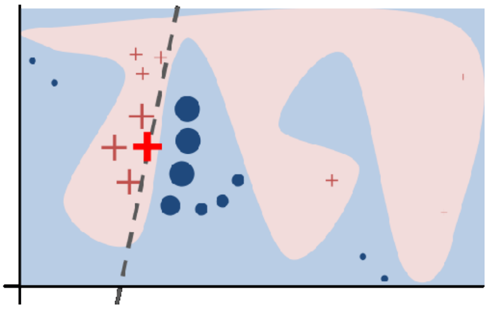
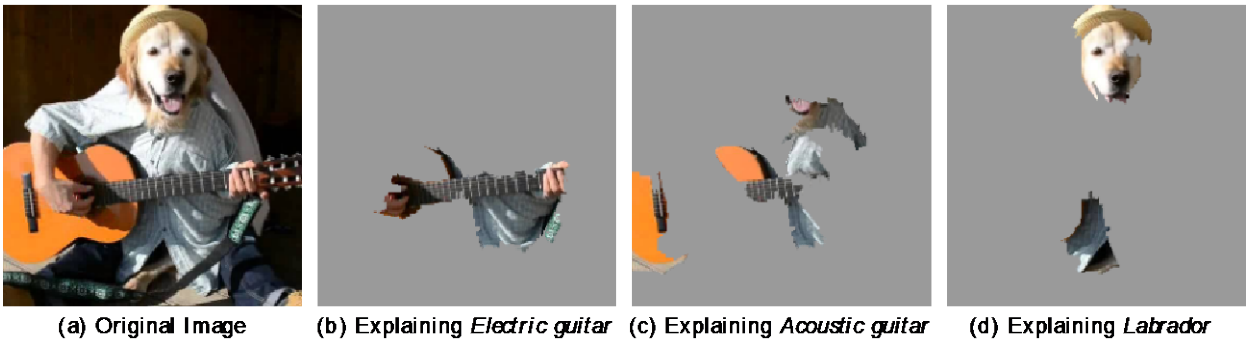


Figure 2: Local approximation of a non-linear binary classifier (red/blue regions) by a linear surrogate model (dashed line) around a given sampling point (red cross) [5].

The saliency of the relevant features in pictures is aggregated in so-called super-pixels, such that the relevant parts of an image can be shown to the engineer or user. An example is given for Google's Inception network in Figure 3, also from the LIME paper [5]:

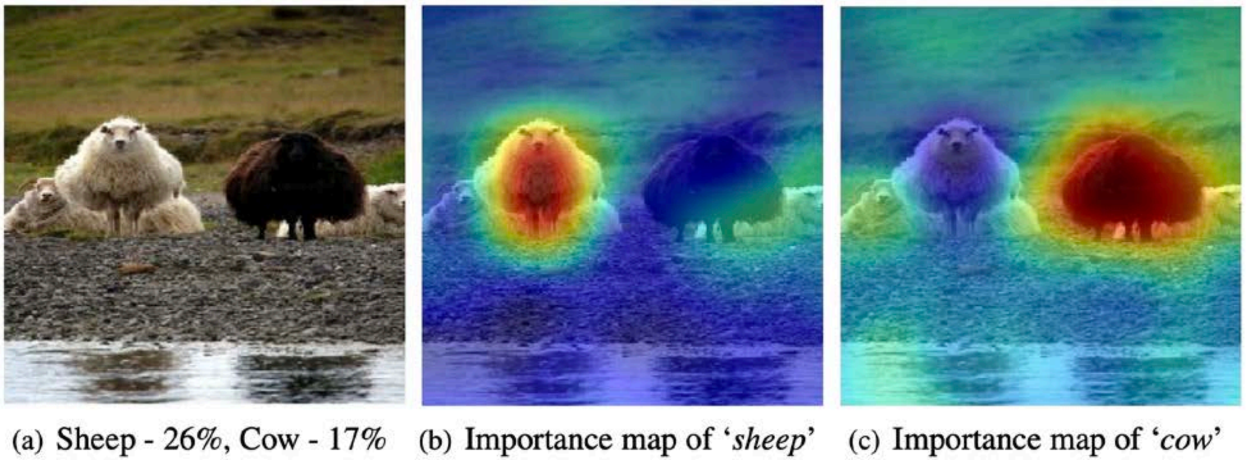


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 3: Explanation of the most relevant classes in terms of super-pixels for the input image (left) with increasing relevance from left to right [5].

RANDOMIZED INPUT SAMPLING FOR EXPLANATION *(RISE)*

The RISE method [6] is similar to LIME, since also stochastic sampling in a local region is used to provide an explanation of an AI result. However, there are differences in the details such that instead of local points RISE uses randomly created masks to blank parts of the input image. The partly obscured images are then fed to the black-box model to obtain the weights for each class. As a consequence, a saliency map (heat map) can be computed as a weighted sum of the stochastic masks such that the relevant parts with respect to a given class are highlighted. An example from [6] is shown in Figure 4.



(a) Sheep - 26%, Cow - 17% (b) Importance map of '*sheep*' (c) Importance map of '*cow*'

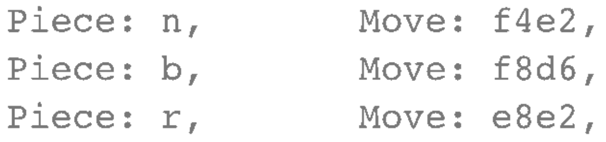
Figure 4: Example saliency map computed from the RISE method to explain the class "sheep" (middle) and "cow" (right) of the input image (left) [6].

NUMERICAL APPLICATION EXAMPLE: EVALUATION OF A CHESS BOARD

In this section, the application of XAI is demonstrated in a proxy chess gaming scenario. Due to its tactical and strategical components, chess is close to hybrid warfare, though of course the dimensionality of the action space is highly reduced. Thus, even though a hybrid warfare scenario is much more complex, some of the information quality provided by XAI could be transferred on an abstract level.

It is well-known that Monte-Carlo chess engines as well as DL methods based on *Markov Decision Processes* (MDP) can be superior to human players even on expert level. However, in highly critical applications where human lives might be at stake, the sole recommendation for the next tactical move can be insufficient in order to guarantee *meaningful human control*, that is the proper exposition of information to ensure accountability, moral responsibility, and controllability for the operator. Therefore, supplemental information is required. XAI provides the necessary means in this case. Though full transparency cannot be achieved, insights for the operator regarding the estimated error variance and the most relevant data features are crucial in critical scenarios. As described in the previous section, various approaches such as BDL and algorithmic feature explanations exist and can be applied.

As a reduced simulative example, a chess game situation is considered. To this end, a fully connected multi layer perceptron[3] was trained on a large chess dataset such that it was able to compete with a standard chess engine [7]. Afterwards, a chess game was stopped after 60 moves with quite even chances on both sides (see the board in Figure 6). The recommendation of the AI engine was now to move the white bishop from e2 to b5, which is a bit offensive. The reason for the recommendation can be revealed by the LIME approach, which indicates the top-relevant features in the data space. In particular, the top attacking moves from black could be extracted as shown in Figure 5.



```
Piece: n,      Move: f4e2,
Piece: b,      Move: f8d6,
Piece: r,      Move: e8e2,
```

Figure 5: Top 3 attacking moves from the opponent revealed by XAI (n = knight, b = bishop, r = rook).

[3] In our simulations, we used a 12-layer network.

The experienced chess player will of course directly see this. In a strategic hybrid military operation, the options might be less obvious, but the methodology can be applied, too.

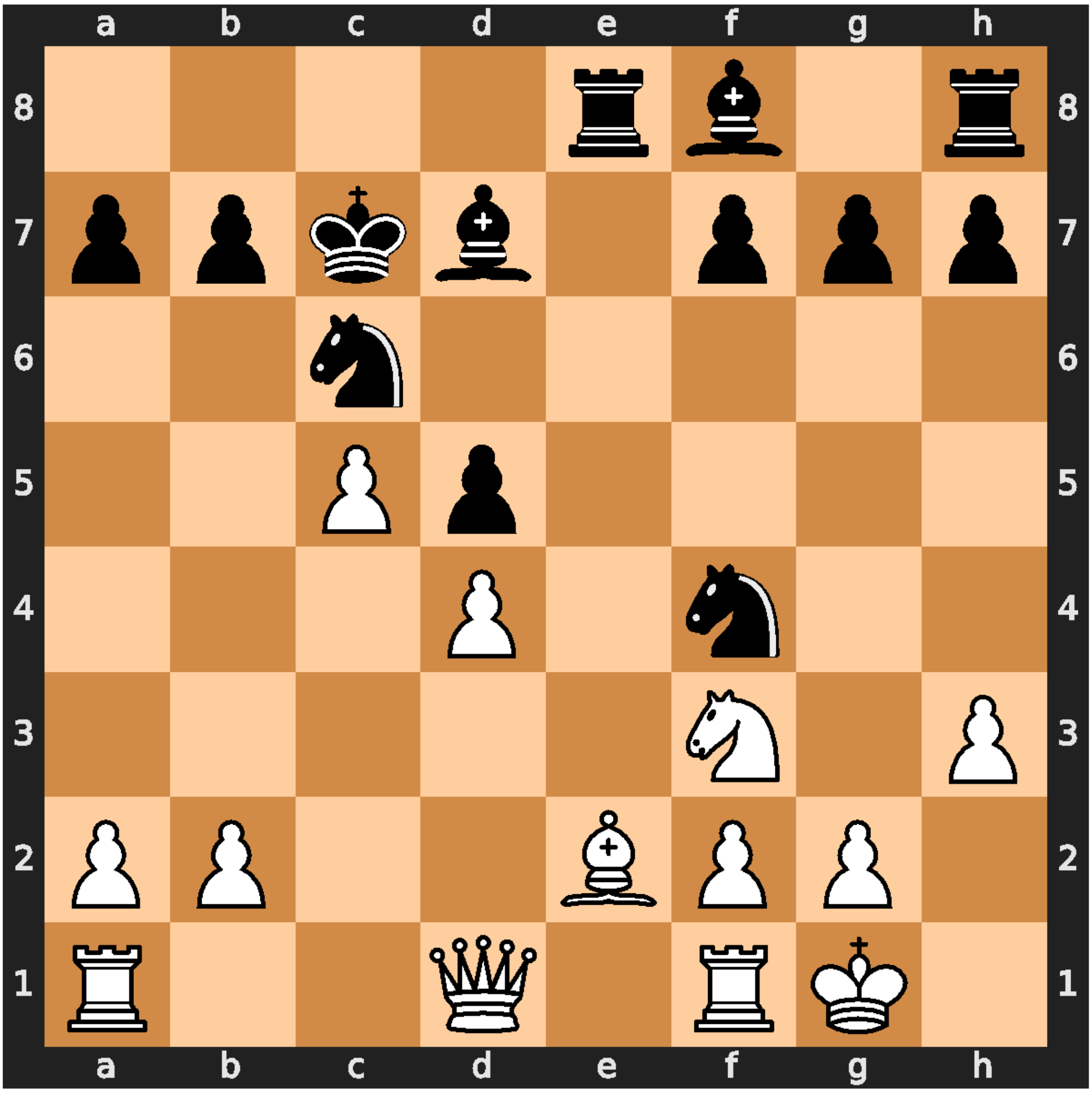


Figure 6: Chess board after 60 moves where the XAI information was evaluated.

CONCLUSION

In this paper, we have revisited the some of the relevant methods of Explainable AI to provide supplemental information in hybrid military operations. Those include approaches to compute the error variance of a computed AI product as well as algorithms for obtaining some degree of transparency by a calculation of the most relevant features in the data space with respect to the outcome of the AI. Though real war (gaming) scenarios are much more complex than chess, the game was used to demonstrate an illustrious example of XAI to a tactical scene on a fixed board. As one can see in the example, the methods are well able to automatically indicate the most relevant possible moves by the opponent. Though this scenario is absolutely not challenging for chess experts, it demonstrates that the approach can be applied with abstract tactical data.

REFERENCES

[1] Alex Kendall and Yarin Gal, “What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?”, Computer Vision and Pattern Recognition, 2017.

[2] Alex Kendall, Computer Vision & Robotics Researcher. Online: https://alexgkendall.com/computer_vision/bayesian_deep_learning_for_safe_ai/

[3] Mahed Javed, Lyudmila Mihaylova and Nidhal Bouaynaya: „Variance Guided Continual Learning in a Convolutional Neural Network Gaussian Process Single Classifier Approach in Noisy Images“. IEEE International Conference on Information Fusion (FUSION), 2021.

[4] Yarin Gal and Zoubin Ghahramani: “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”, Machine Learning, 2015.

[5] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin: “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”, Machine Learning, arXiv:1602.04938, 2016.

[6] Vitali Petsiuk, Abir Das, Kate Saenko: „RISE: Randomized Input Sampling for Explanation of Black-box Models“, Computer Vision and Pattern Recognition, arXive: 1806.07421, 2018.

[7] Stockfish 14, open source chess engine. Online: https://stockfishchess.org