Auditing Deep Learning processes through Kernel-based Explanatory Models

Danilo Croce
croce@info.uniroma2.it
Dpt. of Enterprise Engineering
University of Roma, Tor Vergata
Roma, Italy

Daniele Rossini
rossini.danie@gmail.com
Dpt. of Enterprise Engineering
University of Roma, Tor Vergata
Roma, Italy

Roberto Basili
basili@info.uniroma2.it
Dpt. of Enterprise Engineering
University of Roma, Tor Vergata
Roma, Italy

KEYWORDS

Explainable Neural Networks, Kernel Methods, Nystrom Method

AI systems are currently used in a wide variety of applications, with several levels of societal impact, and are expected to be soon deployed in safety-critical fields, e.g., autonomous driving. Hence, a natural need for ethical accountability of such systems is gaining importance. A central issue lies in designing systems whose decisions are *transparent* [6], i.e., they must be easily interpretable by humans, as users must be able to suitably weight and trust their assistance. Deep neural networks are clearly problematic in this regard: their high non-linearity, despite allowing for state-of-the-art performances in several challenging problems also amplifies the epistemological opaqueness of the decision-flow and limits its interpretability. The concept of transparency of a machine learning model spans multiple definitions, focusing on different aspects, from the simplicity of the model, e.g., the number of nodes in a decision tree, to the intuitiveness of its parameters and computations [4]. In this context, an important capability of an AI system is the ability of providing *post-hoc explanations* in terms of evidences supporting the provided decisions: although they usually do not formally elucidate how a model works, post-hoc explanations often have the nice property of being quite intuitive, conveying useful information also to end-users without any AI or machine learning expertise [8]. In semantic inference tasks (e.g., text classification), an *explanation model* producing post-hoc explanations should hence be able to trace back connections between the output categories and the semantic and syntactic properties of the input texts. Such models should have three desired properties: *semantic transparency*, *informativeness* w.r.t. the system decision and *effectiveness* in enabling auditing processes against the system.

In this work we focus on a specific post-hoc mechanism which is to provide, along with the prediction, a comparison with one or more other examples, namely *landmarks*, that share task-relevant linguistic properties with the input. From an argument theory perspective, this corresponds to supporting decisions through an "argument by analogy" schema [9]: a user exposed to such a kind of argument will endow a different level of trust into the machine decision according to the linguistic plausibility of the analogy. In fact, he/she will implicitly gauge the evidence from the linguistic properties shared between the input sentence (or its parts) and the one used for comparison as well their importance with respect to the output decision. Let us consider, for example, the following prediction in question classification (QC) [7]: *"What is the capital of Zimbabwe?" refers to a* `Location`. We would like the system to motivate its decision with an argument such as: *...since it recalls me of "What is the capital of California?" which also refers to a* `Location`. Notice that a decision explaining task is quite different from relevance ranking, and semantic similarity plays here a minor role: clear and trustful analogies may exist between training examples that are semantically different but such that their properties imply similar causal relationships between the input and the decision. Recent work has been inspired by efforts in improving model's interpretability in image processing tasks, in particular by the *Layerwise Relevance Propagation* (LRP) [3]. In LRP, the classification decision of a deep neural network is decomposed backward across the network layers and evidence about the contribution to the final decision brought by individual input fragments (i.e., pixels of the input image) is gathered. We propose here to extend the LRP application to a linguistically motivated network architecture, known as Kernel-Based Deep Architecture (KDA) [5], which frames semantic information captured by linguistic Tree Kernel [2] methods within the neural-based learning paradigm. The result is a mechanism that, for each system's prediction such as in question classification, generates an argument-by-analogy explanation based on real training examples, not necessarily similar to the input.

We also propose here a novel approach to evaluate numerically the interpretability of any explanation-enriched model applied in semantic inference tasks. By defining a specific audit process, we derive a synthetic metric, i.e. *Auditing Accuracy*, that takes into account the properties of transparency, informativeness and effectiveness. The evaluation of the proposed methodology shows the meaningful impact of LRP-based explanation models: users faced with explanations are systematically oriented to accept (or reject) the system decisions, so that *post-hoc* judgments may even help in improving the overall application accuracy.

This work has been accepted for publication at the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP2019 [1]).

Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
IIR 2019, September 16–18, 2019, Padova, Italy

REFERENCES

[1] 2019. *Proceedings of the 2019 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Hong Kong, China. https://www.emnlp-ijcnlp2019.org/program/accepted/

[2] Paolo Annesi, Danilo Croce, and Roberto Basili. 2014. Semantic Compositionality in Tree Kernels. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014*. 1029–1038. https://doi.org/10.1145/2661829.2661955

CEUR-WS.org/Vol-2441/paper16.pdf

[3] Sebastian Bach, Alexander Binder, Gregoire Montavon, Frederick Klauschen, Klaus-Robert MÃijller, and Wojciech Samek. 2015. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. *PLOS ONE* 10, 7 (2015).

[4] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, and Moustafa Alzantot et al. 2017. Interpretability of deep learning models: A survey of results. *2017 SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI* (2017), 1–6.

[5] Danilo Croce, Simone Filice, Giuseppe Castellucci, and Roberto Basili. 2017. Deep Learning in Semantic Kernel Spaces. In *Proceedings of ACL 2017*. Vancouver, Canada, 345–354.

[6] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI Under the Law: The Role of Explanation. *CoRR* abs/1711.01134 (2017).

[7] Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering* 12, 3 (2006), 229–249.

[8] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Queue* 16, 3, Article 30 (June 2018), 27 pages.

[9] Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.