Increasing trust in complex machine learning systems: Studies in the music domain

Jaehun Kim
Delft University of Technology
j.h.kim@tudelft.nl

Abstract

Machine learning (ML) has become a core technology for many real-world applications. Modern ML models are applied to unprecedentedly complex and difficult challenges, including very large and subjective problems. For instance, applications towards multimedia understanding have been advanced substantially. Here, it is already prevalent that cultural/artistic objects such as music and videos are analyzed and served to users according to their preference, enabled through ML techniques.

One of the most recent breakthroughs in ML is Deep Learning (DL), which has been immensely adopted to tackle such complex problems. DL allows for higher learning capacity, making end-to-end learning possible, which reduces the need for substantial engineering effort, while achieving high effectiveness. At the same time, this also makes DL models more complex than conventional ML models. Reports in several domains indicate that such more complex ML models may have potentially critical hidden problems: various biases embedded in the training data can emerge in the prediction, extremely sensitive models can make unaccountable mistakes. Furthermore, the black-box nature of the DL models hinders the interpretation of the mechanisms behind them. Such unexpected drawbacks result in a significant impact on the trustworthiness of the systems in which the ML models are equipped as the core apparatus.

In this thesis, a series of studies investigates aspects of trustworthiness for complex ML applications, namely the reliability and explainability. Specifically, we focus on music as the primary domain of interest, considering its complexity and subjectivity. Due to this nature of music, ML models for music are necessarily complex for achieving meaningful effectiveness. As such, the reliability and explainability of music ML models are crucial in the field.

The first main chapter of the thesis investigates the transferability of the neural network in the Music Information Retrieval (MIR) context. Transfer learning, where the pre-trained ML models are used as off-the-shelf modules for the task at hand, has become one of the major ML practices. It is helpful since a substantial amount of the information is already encoded in the pre-trained models, which allows the model to achieve high effectiveness even when the amount of the dataset for the current task is scarce. However, this may not always be true if the "source" task which pre-trained the model shares little commonality with the "target" task at hand. An experiment including multiple "source" tasks and "target" tasks was conducted to examine the conditions which have a positive effect on the transferability. The result of the experiment suggests that

the number of source tasks is a major factor of transferability. Simultaneously, it is less evident that there is a single source task that is universally effective on multiple target tasks. Overall, we conclude that considering multiple pre-trained models or pre-training a model employing heterogeneous source tasks can increase the chance for successful transfer learning.

The second major work investigates the robustness of the DL models in the transfer learning context. The hypothesis is that the DL models can be susceptible to imperceptible noise on the input. This may drastically shift the analysis of similarity among inputs, which is undesirable for tasks such as information retrieval. Several DL models pre-trained in MIR tasks are examined for a set of plausible perturbations in a real-world setup. Based on a proposed sensitivity measure, the experimental results indicate that all the DL models were substantially vulnerable to perturbations, compared to a traditional feature encoder. They also suggest that the experimental framework can be used to test the pre-trained DL models for measuring robustness.

In the final main chapter, the explainability of black-box ML models is discussed. In particular, the chapter focuses on the evaluation of the explanation derived from model-agnostic explanation methods. With black-box ML models having become common practice, model-agnostic explanation methods have been developed to explain a prediction. However, the evaluation of such explanations is still an open problem. The work introduces an evaluation framework that measures the quality of the explanations employing fidelity and complexity. Fidelity refers to the explained mechanism's coherence to the black-box model, while complexity is the length of the explanation.

Throughout the thesis, we gave special attention to the experimental design, such that robust conclusions can be reached. Furthermore, we focused on delivering machine learning framework and evaluation frameworks. This is crucial, as we intend that the experimental design and results will be reusable in general ML practice. As it implies, we also aim our findings to be applicable beyond the music applications such as computer vision or natural language processing.

Trustworthiness in ML is not a domain-specific problem. Thus, it is vital for both researchers and practitioners from diverse problem spaces to increase awareness of complex ML systems' trustworthiness. We believe the research reported in this thesis provides meaningful stepping stones towards the trustworthiness of ML.

Awarded by: Delft University of Technology, Delft, The Netherlands **on** 19 May 2021.
Supervised by: Cynthia C.S. Liem.
Available at: https://doi.org/10.4233/uuid:01ba927d-28e7-4abd-8193-e4ebef3b8218.

Selected Publications

Jaehun Kim, Minz Won, Xavier Serra, and Cynthia C. S. Liem. Transfer learning of artist group factors to musical genre classification. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1929–1934. ACM, 2018.

Jaehun Kim, Julián Urbano, Cynthia C. S. Liem, and Alan Hanjalic. Are nearby neighbors

relatives? testing deep music embeddings. *Frontiers in Applied Mathematics and Statistics*, 5: 53, 2019.

Jaehun Kim, Julián Urbano, Cynthia C. S. Liem, and Alan Hanjalic. One deep music representation to rule them all? A comparative analysis of different representation learning strategies. *Neural Comput. Appl.*, 32(4):1067–1093, 2020.