Mateusz Gniewkowski

Rigorous Evaluation: Towards Trustworthy Representations in Machine Learning

Abstract

The starting point of this dissertation is the gap between "single-number" evaluation and the actual reliability of models, that is, their behavior in real-world conditions. High accuracy on data drawn from the same distribution as the training set (*Independent and Identically Distributed*, IID) does not tell us whether the model bases its decision on appropriate evidence, nor whether its representations are useful beyond a narrow task. We also do not know how the model will behave under a distribution shift or under small, semantically insignificant changes. In practice, real-world systems encounter environments full of spurious correlations, novelties, and stimuli that can easily "fool" a classifier. The gap between test and deployment arises in particular from: nonstationarity and domain shift; narrow, nonrepresentative validation procedures (no out-of-distribution assessment and no calibration); data risks (leakage, class imbalance, collection artifacts); lack of uncertainty handling (OOD detection); and the specifics of model training itself, which encourages spurious correlations and simplified decision rules. In response, we propose the discipline "audit → measure → improve", which combines explainability methods (*Explainable Artificial Intelligence*, XAI), representation-quality metrics, and out-of-distribution tests into a coherent, repeatable cycle: first we identify the "evidence" on which the model bases its decisions; next we quantify and verify it; finally, we introduce small, targeted modifications (based on previously detected problems) that minimize the cost in clean-data accuracy while improving the tool's practical utility.

Beyond the loop "audit → measure → improve", we propose a *portfolio* of methods for auditing and measuring the quality and robustness of representations: clustering as an early indicator of spurious correlations and the quality of semantic organization; OOD generalization and OOD detection; local and global XAI audits that reveal features potentially irrelevant causally; and multidimensional scaling together with a *numerical* assessment of projection quality. We complement this with adversarial evaluation based on explainability and with two improvement pathways: (i) for text – contrastive finetuning and distillation supported by a *human-in-the-loop* protocol; (ii) for images – a lightweight loss function aimed at increasing the *Class Robustness Score* (understood as the fraction of

attribution mass assigned to the annotated object) without degrading accuracy, while increasing robustness to black-box attacks. The whole is completed with practical guidance for open-world deployment of the produced models, enabling systematic assessment across the model lifecycle.

Having outlined the problem and introduced the methodology, we proceed to the first case study that illustrates the loop "audit → measure → improve" in practice. In Chapter 3, we propose a practical tool for vectorizing and classifying HTTP/URL traffic based on a language model. We then conduct a systematic XAI audit that exposes weaknesses in the solution. In parallel, we measure representation quality in clustering and OOD separation tasks, revealing issues invisible to standard classification metrics. Guided by the audit, we reshape the embedding geometry to improve representation quality, then rerun the measurements and compare them with the initial approach.

In the subsequent part of the dissertation, we propose weaponizing explainability, that is, deliberately repurposing explanation methods as tools for constructing adversarial attacks on text classifiers. The central idea is to use feature attributions as a mechanism for constraining the adversarial search space: instead of exploring the full combinatorial space of possible text perturbations, the attack operates only on those features that have the strongest influence on the model's decision. The proposed methods precisely modify – while preserving semantic meaning and grammatical correctness – those input components that are most important from the classifier's perspective, in order to induce misclassification. We consider both untargeted attacks (causing a transition to any other class) and targeted attacks (forcing prediction into a specific, chosen class). We demonstrate how this procedure applies both to classical text classifiers and to remotely accessible generative models, such as OpenChat, ChatGPT, and LLaMA, configured for *zero-shot* classification. In addition, we assess the perceptual quality of the generated adversarial samples in a human evaluation study.

We then propose a targeted method for improving the quality of visual representations in image classification models. Starting from a high-accuracy pretrained classifier, we audit its decisions in terms of the semantic validity of the evidence it relies on, assessed qualitatively through the correspondence between attribution maps and the ground-truth object. We adopt a localization-consistency measure – defined as the fraction of attribution mass overlapping the annotated object – not only as a diagnostic tool, but as a training signal that directly guides model refinement. Building on this insight, we propose a lightweight, localization-aware training objective that reduces reliance on spurious background correlations and shifts representations toward object-centric features. The resulting procedure is architecture-agnostic and introduces only minimal additional supervision. It yields consistent improvements in localization consistency and, as a secondary effect, increases robustness to black-box adversarial attacks, while preserving classification performance on clean data.

This dissertation shows that the trustworthiness of models cannot be assessed by a single number – such as an F1 point or test accuracy. That still-dominant paradigm in the literature is too narrow to guarantee reliability in practice. Instead, models should be embedded in the cycle "audit → measure → improve", which makes system behavior visible in a broader perspective and enables targeted improvements. High values of classical metrics remain a good starting point, but they say little about whether – and which – model will perform best in a real application. Models designed with unexpected conditions in mind (including out-of-distribution settings and related tasks) fail less often and can be granted greater trust where the stakes are highest.