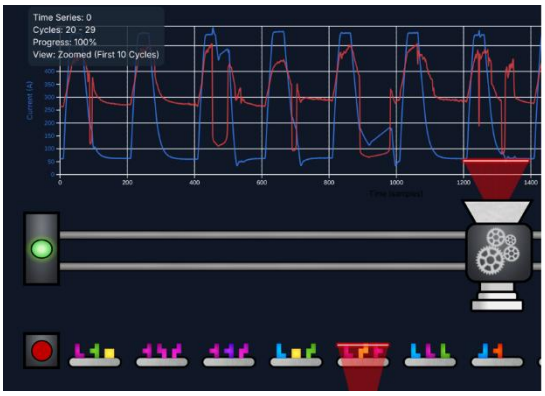
----- English version below -----

Ausschreibung Masterarbeit

Explainable Deep Learning: Analyzing Neural Network Decisions with DeEXCODER

Ausgangslage

Tiefe neuronale Netze erzielen in zahlreichen Anwendungsbereichen hervorragende Ergebnisse, ihre Entscheidungen sind jedoch häufig nur schwer nachvollziehbar. Insbesondere in sicherheitskritischen oder industriellen Anwendungen gewinnt die Erklärbarkeit (Explainable Artificial Intelligence, XAI) zunehmend an Bedeutung. Mit **DeEXCODER** wird ein neuer Ansatz entwickelt, der Entscheidungsprozesse neuronaler Netze transparent macht und deren interne Repräsentationen analysiert.



Problemstellung

Der bisherige ExCODER-Ansatz nutzt VQ-VAEs zur Interpretation gelernter Merkmalsrepräsentationen, beschränkt sich jedoch auf die Input-Feature-Ebene. Veränderungen der Repräsentationen über tiefere Netzwerkschichten bleiben dadurch verborgen.

Ziel dieser Arbeit ist die Entwicklung und Evaluation des DeEXCODER-Ansatzes, der für jede Schicht ein eigenes Codebook erzeugt. So soll analysiert werden, wie sich Merkmalsrepräsentationen entlang der Netzwerkhierarchie entwickeln und den Entscheidungsprozess des Modells beeinflussen. Die Evaluation erfolgt mit gängigen Architekturen auf etablierten Benchmark-Datensätzen (z. B. MNIST).

Vorgehensweise und Erwartete Ergebnisse

Zu Beginn erfolgt eine Einarbeitung in die Grundlagen von Explainable AI, Deep Learning und den EXCODER-Ansatz. Anschließend werden geeignete Benchmark-Datensätze und neuronale Netzwerkarchitekturen ausgewählt und für die Experimente vorbereitet. DeEXCODER wird auf die trainierten Modelle angewendet und die erzeugten Erklärungen systematisch analysiert. Abschließend werden die Ergebnisse hinsichtlich Interpretierbarkeit, Aussagekraft und möglicher Grenzen des Verfahrens bewertet und dokumentiert.

Ansprechpartner

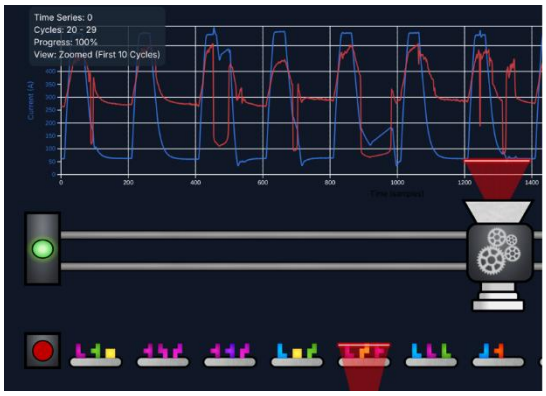
Yannik Hahn | **E-Mail:** yhahn@uni-wuppertal.de

Master's thesis

Explainable Deep Learning: Analyzing Neural Network Decisions with DeEXCODER

Initial Situation

Deep neural networks achieve excellent results in numerous application areas, but their decisions are often difficult to understand. Especially in safety-critical or industrial applications, explainability (Explainable Artificial Intelligence, XAI) is becoming increasingly important. **DeEXCODER** is a newly developed approach that makes the decision-making processes of neural networks transparent and analyzes their internal representations.



Problem Definition

The previous ExCODER approach uses VQ-VAEs to interpret learned feature representations but is limited to the input feature level. Consequently, changes in representations across deeper network layers remain hidden.

The aim of this work is to adapt and evaluate the DeEXCODER approach, which generates a separate codebook for each layer. This allows for an analysis of how feature representations evolve along the network hierarchy and influence the model's decision-making process. The evaluation is conducted using standard architectures on established benchmark datasets (e.g., MNIST).

Methods and Expected Results

The work begins with familiarization with the fundamentals of Explainable AI, Deep Learning, and the EXCODER approach. Next, suitable benchmark datasets and neural network architectures are selected and prepared for the experiments. DeEXCODER is applied to the trained models, and the generated explanations are systematically analyzed. Finally, the results are evaluated and documented with respect to interpretability, explanatory power, and potential limitations of the method.

Contact Person

Yannik Hahn | **E-Mail:** yhahn@uni-wuppertal.de