KI ABSICHERUNG
Safe AI for Automated Driving

Using concept embeddings for safety assurance

Gesina Schwalbe, Martin Schels

DNN specific Safety Concern: Incomprehensible Behavior (SC-1.3)

Problem

Many safety related requirements are symbolic, i.e. they use (visual) semantic concepts from natural language. E.g. „arms are part of persons“. For perception tasks and standard DNNs, such concepts cannot be accessed for verification purposes: The inputs are non-symbolic, outputs are sparse (e.g. no arm), and DNNs are opaque, i.e. how such concepts are internally encoded is unknown.

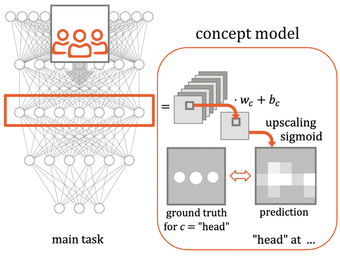


Figure 1: Concept analysis approach from [1]

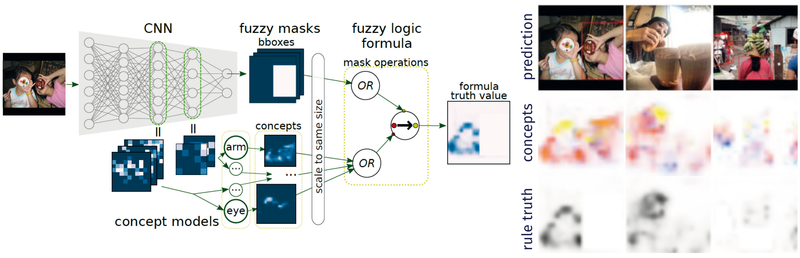
Concept Embedding Analysis

Literature suggests that visual semantic concepts, like body parts or colors, can be associated with vectors in DNN layer outputs, the concept (activation) vectors (CAV) [1]. Following prior work, our approach in [1] finds these by training for each layer a small linear model, the concept model (CM), to predict the presence of the concept from an activation map pixel in the layer output (Fig. 1). The weights of this model are the global CAV.

Towards Object Detection

Comparative studies discarded expensive activation map preprocessing, and found suitable hyperparameter settings for object detectors and object part concepts.

Figure 3: Framework from [2] (left) to use CMs for verification of DNNcompliance w/ fuzzy rules, for the rule „bodypart → person“; sample outputs on the right



Safety Hypothesis:
To verify compliance w/ symbolic background knowledge, one requires access to the representation of semantic concepts within the DNN. This method provides such access both locally for single samples and globally.

Applications

Several verification applications were investigated for object detectors:

- **Necessary concepts** are „known“ to DNN: Performance of CMs (Fig. 2, top)
- **CAV similarity** vs. true semantic similarity: Cosine similarity of CAVs (Fig. 2, right)
- **Bias** of DNN representations, e.g. wrt. size: CM performance on training/test subsets.
- **Inspection** of internal logics: Use CM outputs to create interpretable proxies.
- Compliance w/ (fuzzy) **logic rules**: Use CM outputs as inputs to logic formulas (Fig. 3).

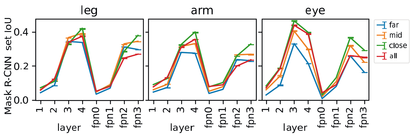
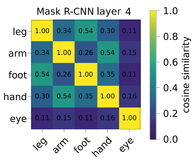


Figure 2: CM performances for object detector (top) and cosine similarities of CAVs (right)



References:

[1] Schwalbe, Gesina. 2021. “Verification of Size Invariance in DNN Activations Using Concept Embeddings.”
[2] Schwalbe, Gesina, Christian Wirth, and Ute Schmid. 2022. “Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings.”

Continental

For more information contact:
gesina.schwalbe@continental-corporation.com
martin.schels@continental-corporation.com

KI Absicherung is a project of the KI Familie. It was initiated and developed by the VDA Leitinitiative autonomous and connected driving and is funded by the Federal Ministry for Economic Affairs and Climate Action.

KI FAMILIE

Supported by:
Federal Ministry for Economic Affairs and Climate Action
on the basis of a decision by the German Bundestag