Fraunhofer IKS

Modular Concept Learning

Our Modular Concept Learning
helps companies with image classification tasks
to develop explainable and easily adaptable AI models for image classification
by breaking up a complex classification task into subtasks of recognizing relevant visual concepts, and combining these concepts into a transparent decision process.
Unlike other AI models in image classification, our Modular Concept Learning increases not only explainability but also adaptability to changes in the task or data through its modularity and reusability of concept models.

Common problem

Most state-of-the-art machine learning (ML) models treat the classification of an image as one complex task. Not only does this often require costly and time-consuming retraining of the entire ML model in the case of changes to the task or data, but also does it make the model's decision-making process difficult to explain. A lack in explainability reduces trust in the model's decision (output), which in turn makes a model less likely to be used in a safety-critical context, e.g., medical imaging.

Our solution

Instead of training an ML model directly on the entire task (e.g., traffic sign classification), we guide the model to follow a process like how humans would classify an image: They would usually look for the presence or absence of certain features or concepts in an image (e.g., shape, color, icon on the traffic sign) to make their decision. In a similar manner, our Modular Concept Learning breaks up an image classification task into subtasks (concept models of e.g., shape, texture, color) and combines the detected concepts into an overall classification decision. The mapping of concepts (e.g., color of the traffic sign) to the right class (red = stop/no entry sign) can either be manually modelled by a domain expert or trained on data. If the task or data changes after deployment, retraining can be limited to one or few concepts of the model instead of the entire model. In addition, it is easier to explain the cause of a wrong classification (e.g., the color was misclassified but the shape and icon were classified correctly), reducing both training and error investigation efforts.

Exemplary use cases

- Traffic sign classification: shape of the road sign, color, icon, etc.
- White blood cell analysis: cell size/shape, cytoplasm texture/color/vacuole, nucleus shape, etc.

Benefits

- ✓ **Interpretability:** The ML model's decision is explained using the detected concepts and their relationship to each other.
- ✓ **Flexibility:** The building blocks can be customized to the specific requirements of the use case (e.g., a combination of classical algorithms and AI models).
- ✓ **Adaptability:** Rather than retraining the entire model, our approach allows for targeted adaptation of parts of the model to respond to changes in the task and/or dataset.
- ✓ **Error Analysis:** The modularity enables a more detailed error analysis, e.g., which concept is difficult to detect, and the application of more targeted measures, e.g., collecting more data of a specific concept.

Interested? We are happy to share more insights.

Lena Heidemann
Dependable Perception & Imaging
lena.heidemann@iks.fraunhofer.de
www.iks.fraunhofer.de

Business Development
business.development@iks.fraunhofer.de