Abstract

Understanding the internal representations of neural networks is a fundamental challenge in artificial intelligence. Despite achieving remarkable performance across various domains, these models often lack interpretability, transparency, and explainability, hindering their deployment in critical applications where insight into their decision-making processes is essential. This thesis investigates and extends methods to interpret neural networks' predictions by transforming their internal representations into more human-understandable forms, striving to minimize biases or distortions. The thesis is structured around three key contributions:

First, in the domain of computer vision, we propose the Bi-Source Class Visualization (BSCV) method to address challenges in feature visualization. BSCV leverages an adversarial neural network framework, using the input layer as a shared interface between a discriminator and a classifier. By allowing gradients to flow through this interface, BSCV generates realistic and interpretable class-specific visualizations without relying on handcrafted biases, thereby overcoming limitations of previous methods.

Second, we explore how pretraining imparts morphosyntactic knowledge to language models by probing the internal representations of multilingual Bidirectional Encoder Representations from Transformers (BERT) models. Diagnostic classifiers, or probes, are trained to predict linguistic properties from the models' internal states that are not fine-tuned on morphosyntactic tasks. Extensive ablation studies and probing controls ensure the validity of the findings, revealing that pretraining leads to highly abstract internal representations encoding rich linguistic information.

Third, in a case study involving the Migration Media Discourse (MIMEDIS) project, we analyze media discourse on migration using cross-lingual and monolingual models. By developing a specialized application of Shapley values for natural language processing tasks, the models' predictions were systematically decomposed to quantify the contribution of individual input tokens. This methodology enabled a detailed examination of the models' internal decision-making processes, uncovering significant differences in the internal representations of architecturally similar models trained on different data, and thereby demonstrating the impact of training data on model behavior.