

Generic Interpretation Approach for Transformer Models Incorporating Heterogenous Attention Structures

Authors: Yongjin Cui, Xiaohui Fan, Huajun Chen, Xiaohui Fan

Date:

Abstract

Transformer has significantly propelled the development of artificial intelligence, and certainly the development of agents as well. We categorize attention structures of Transformer into two types based on the source of the input information: homogenous and heterogenous attention structures. Heterogenous attention structures, with co-attention as a typical example, process information from different sources. Heterogenous attention structure is the foundation for Transformer models to achieve more complex functions and integrate more modal information. Whether for research purposes or policy requirements, the interpretation of Transformer models with heterogenous attention structures is an important task. The fusion of information from different sources brings new challenges. Our work mainly includes two parts: method and experimentation. In terms of method, we propose an interpretation method for Transformer models with heterogenous attention structures. In terms of experimentation, based on our experimental analysis paradigm, we interpret the operating mechanisms of representative models, conduct semantic interpretation and logical interpretation.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv – Machine translation. Verify with original.