Explainability of Vision Transformers: A Comprehensive Review and New Perspectives

Transformers have had a significant impact on natural language processing and have recently demonstrated their potential in computer vision. They have shown promising results over convolution neural networks in fundamental computer vision tasks. However, the scientific community has not fully grasped the inner workings of vision transformers, nor the basis for their decision-making, which underscores the importance of explainability methods. Understanding how these models arrive at their decisions not only improves their performance but also builds trust in AI systems. This study explores different explainability methods proposed for visual transformers and presents a taxonomy for organizing them according to their motivations, structures, and application scenarios. In addition, it provides a comprehensive review of evaluation criteria that can be used for comparing explanation results, as well as explainability tools and frameworks. Finally, the paper highlights essential but unexplored aspects that can enhance the explainability of visual transformers, and promising research directions are suggested for future investment.