Abstract

Transformer models, which utilize the self-attention mechanism, have achieved unparalleled performance in the field of computer vision. However, the underlying principles of this mechanism and the reasons for transformers' reliance on extensive datasets remain elusive. This study aims to elucidate these aspects by conducting a comparative analysis of various configurations of transformers, including certain modifications. Through this approach, we seek to deepen our understanding of the self-attention mechanism and explore the dependency of transformers on large datasets.