NEURAL NETWORK SEARCH METHOD AND RELATED DEVICE

ABSTRACT

This application relates to the artificial intelligence field, and discloses a neural network search method and a related apparatus. The neural network search method includes: constructing attention heads in transformer layers by sampling a plurality of candidate operators during model search, to construct a plurality of candidate neural networks, and comparing performance of the plurality of candidate neural networks to select a target neural network with higher performance. In this application, a transformer model is constructed with reference to model search, so that a new attention structure with better performance than an original self-attention mechanism can be generated, and effect in a wide range of downstream tasks is significantly improved.

FIG. 12