Which Features Contribute to Galaxy Morphology Classification?: A Novel Explainable Artificial Intelligence Approach

Xie, Mingxuan (School: Guangzhou Zhixin High School)

Galaxy morphology classification is an essential step in extragalactic astronomy to understand galaxy differences and uncover new structural features. With an enormous number of galaxies (up to 10^11) to classify, effective and reliable automated approaches become imperative. In recent years, thanks to the automated feature extraction ability of Deep Learning (DL) algorithms (e.g. Convolutional Neural Network, CNN), DL models have demonstrated great performance in astronomy tasks like galaxy morphology classification, potentially offering new fresh perspectives in astronomy. However, people found DL models have low interpretability and thus difficult to distinguish which features contribute to their predictions. Therefore, astronomers found it difficult to believe in their outcomes. In this research, I try to explain how a CNN-based galaxy morphology classifier makes prediction. I design a novel workflow combining Class Activation Mapping (CAM) with Unsupervised Semantic Segmentation (USS) to visualize its Regions Of Interest (ROI). By deriving the features that contribute to the model's prediction for each morphology class, a possible global explanation is provided. The results show that the CNN model's decision-making process has a high degree of alignment with human's (~88%), indicating its relative reliability. Meanwhile, an intriguing feature that contradicts human knowledge is discovered although its physical meaning remains unknown. In general, this research proposes a workflow that can effectively explain CNN-based galaxy classifiers and estimates their reliability. It also opens up promising new directions for discovering unexpected galaxy features from the perspective of CNN models.