If an image is worth a thousand words, why ask machine-learning models to describe images with a single word?

Esla Timothy Anzaku, Arnout Van Messem, Wesley De Neve

March 2023

1 Abstract

Deep learning-based image classification is a critical task that drives advancements in various computer vision tasks. However, despite the use of large models and datasets, performance on key benchmarks has plateaued. In this study, we investigate compelling examples of how machine learning models label and describe images. This prompts us to question the suitability of the current approach of describing images with a single label, given the adage that a picture is worth a thousand words. We argue that revisiting fundamental assumptions and simplifications in the model creation process is necessary to create reliable and trustworthy artificial intelligence models. By doing so, we can improve the accuracy of deep learning-based image classification and enhance the performance of related tasks in computer vision.