CONTRASIM – A SIMILARITY MEASURE BASED ON CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work has compared neural network representations via similarity-based analyses, shedding light on how different aspects (architecture, training data, etc.) affect models' internal representations. The quality of a similarity measure is typically evaluated by its success in assigning a high score to representations that are expected to be matched. However, existing similarity measures perform mediocrely on standard benchmarks. In this work, we develop a new similarity measure, dubbed ContraSim, based on contrastive learning. In contrast to common closed-form similarity measures, ContraSim learns a parameterized measure by using both similar and dissimilar examples. We perform an extensive experimental evaluation of our method, with both language and vision models, on the standard layer prediction benchmark and two new benchmarks that we develop: the multilingual benchmark and the image–caption benchmark. In all cases, ContraSim achieves much higher accuracy than previous similarity measures, even when presented with challenging examples.