Semantic Embeddings are a popular way to represent knowledge in the field of zero-shot learning. We observe their interpretability and discuss their potential utility in a safety-critical context. Concretely, we propose to use them to add introspection and error detection capabilities to neural network classifiers. First, we show how to create embeddings from symbolic domain knowledge. We discuss how to use them for interpreting mispredictions and propose a simple error detection scheme. We then introduce the concept of semantic distance: a real-valued score that measures confidence in the semantic space. We evaluate this score on a traffic sign classifier and find that it achieves near state-of-the-art performance, while being significantly faster to compute than other confidence scores. Our approach requires no changes to the original network and is thus applicable to any task for which domain knowledge is available.

Kongress- / Buchtitel: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops

Jahr: 2019

Monat: Jun