Improving Word Embeddings for Zero-Shot Event Localisation by Combining Relational Knowledge with Distributional Semantics

L.J. Pascha

Abstract

Temporal event localisation of natural language text queries is a novel task in computer vision. Thus far, no consensus has been reached on how to predict the temporal boundaries of action segments precisely. While most attention in literature has been dedicated towards the representation of vision, here we attempt to improve the representation of language for event localisation by applying Graph Convolutions (GraphSAGE) on ConceptNet with distributional node embedding features. We argue that due to the large vocabulary size of language and currently small temporally sentence annotated datasets in scale and size, a high dependency is placed upon zero-shot performance. We hypothesise that our approach leads to more visually centred and structured language embeddings beneï¬•cial for this task. To test this, we design a wide-scale zero-shot dataset based on ImageNet to optimise our embeddings on and compare to other language embedding methods.
State-of-the-art results are obtained on 5/17 popular intrinsic evaluation benchmarks, but with slightly lower performance on the TACoS dataset. Due to the almost complete overlap in train- and testset vocabulary, we deem additional testing necessary on a dataset that places more emphasis on word-relatedness; hypernyms, hyponyms and synonyms, which arguably makes language representation learning difi¬•cult.