[Supplementary] Zero-Shot Action Recognition with Transformer-based Video Semantic Embedding

Anonymous CVPR submission

Paper ID *****

1. Proposed Fair ZSL Test Setup

We pool the valid test classes from all benchmark datasets to form a novel test set. Altogether, there are 30 unique classes from the UCF-101, HMDB-51, and ActivityNet datasets, as shown in Table ??. We handpick each class carefully such that it does not violate the ZSL premise.

We next explain the rationale behind excluding the overlapping classes and completely irrelevant classes in the proposed test set.

2. Overlap between Datasets

In Fig. 1, we visualize the semantic embeddings of the classes in Kinetics, ActivityNet and UCF-101 datasets. We see that there are several classes in all the test datasets that directly overlap with the training dataset (Kinetics), which is a violation of the ZSL paradigm.

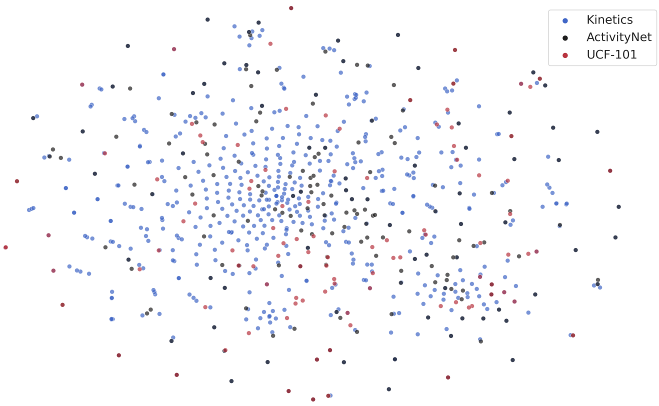

Figure 1. t-SNE visualization of the Kinetics, ActivityNet and UCF-101 classes. We see that several test classes directly overlap with the training classes in Kinetics, which violates the ZSL paradigm.

Dataset	Class
UCF	Pizza Tossing
UCF	Ice Dancing
UCF	Handstand Walking
UCF	Handstand Pushup
UCF	Mixing
UCF	Wall Pushups
UCF	Horse Race
UCF	Playing Dhol
HMDB	Draw Sword
HMDB	Sword Exercise
HMDB	Chew
ActivityNet	Applying sunscreen
ActivityNet	Beach soccer
ActivityNet	Cleaning shoes
ActivityNet	Cleaning sink
ActivityNet	Cutting the grass
ActivityNet	Doing karate
ActivityNet	Doing kickboxing
ActivityNet	Drinking beer
ActivityNet	Drinking coffee
ActivityNet	Fun sliding down
ActivityNet	Hand car wash
ActivityNet	Making an omelette
ActivityNet	Painting fence
ActivityNet	Playing water polo
ActivityNet	River tubing
ActivityNet	Snow tubing
ActivityNet	Starting a campfire
ActivityNet	Washing face
ActivityNet	Washing hands

Table 1. Classes in the proposed test set.