Unified Embedding and Metric Learning for Zero-Exemplar Event Detection

Noureldien Hussein, Efstratios Gavves, Arnold W. M. Smeulders | Quva Lab, University of Amsterdam

Problem

Zero-example Event Detection (ZED) is posed as a video retrieval task. Given test videos and a novel query, the model is required to rank the videos accordingly.

Approach

We pose ZED as learning from a set of predefined events. Given video exemplars of events “removing drywall” or “fit wall tiles”, one may detect a novel event “renovate home” as a probability distribution over the predefined events.

Novelties

1. **Unified embedding** for cross-modalities with metric loss for maximum discrimination between events.
2. **Textual embedding** poses a novel query a probability of predefined events.
3. **External data source**, of event articles and related videos, with end-to-end learning from cross-modal pairs.

Results

Out textual embedding $f_t$ maps the text description of MED events to EventNet events better than off-the-shelf LDA, LSI or Doc2vec.

• The unified embedding (c) is doing a better job in discriminating the test articles of the events.
• In (b), projecting the videos on their related events is much better than the baselines (a) and (c).

Experiments

Model overview of baseline methods. Top: visual embedding (model$^1$). Bottom: separate embedding (model$^2$).

Loss functions used to train the baseline models: visual embedding model$^1$, contrastive visual model$^3$, separate embedding model$^2$, and non-metric embedding model$^4$.

Take Home

The Good: external knowledge (EventNet, WikiHow) is leveraged for better zero-exemplar event detection.

The Bad: no fine-grained event detection, e.g. “fixing musical instrument” vs. “tuning musical instrument”.

The Ugly: is average pooling enough for video representation or temporal modeling is required?

GitHub: https://git.io/vS2Cf