Unified Embedding and Metric Learning for Zero-Exemplar Event Detection

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

**Problem**

Zero-exemplar 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.

**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.

**Approach**

We use video samples from EventNet and event articles from WikiHow.

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.

**Results**

Out textual embedding \(v_f\) maps the text description of MED events to EventNet events better than off-the-shelf LDA, LG or Doc2vec.

**Experiments**

Model overview of baseline methods: Top: visual embedding models (med2vec). Bottom: separate embedding models (med2vec). Loss functions used to train the baseline models: visual embedding models (med2vec), contrastive visual models (med2vec), separate embedding models (med2vec), and non-metric embedding models (med2vec).

**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?