Learn a semantic video representation from videos + stories

- Contact: a.habibian@uva.nl
- A dataset of videos and stories
- A convex objective for learning semantic representation from stories
- Stories as a new source of supervision to learn video representation
- Filters to remove low quality videos

**Goal and Novelties**

- Learn a semantic video representation from videos + stories
- YouTube 46K dataset
- We collect 46K videos and their titles from YouTube
- Seeded from video event descriptions
- Filters to remove low quality videos

**Novelties:**
- Stories as a new source of supervision to learn video representation
- A convex objective for learning semantic representation from stories
- A dataset of videos and stories

**Learning from Videos and Stories**

We learn an embedding from videos into their stories

- **Visual embedding:** extracts term groups from stories
- **Textual embedding:** predicts term groups from video features

**Key contribution**

Instead of predicting each term individually we predict *term groups*

- All grouped
  - **High predictability**
  - **Low descriptiveness**
- **The optimal grouping** (*VideoStory*)
- No grouping
  - **Low predictability**
  - **High descriptiveness**

**VideoStory** is learnt by jointly minimizing two loss functions:

- **Descriptiveness:** error in reconstructing stories from VideoStories
- **Predictability:** error in predicting VideoStories from videos

**Convex loss function minimized by stochastic gradient descent**

**Experiments**

- **YouTube46K for learning representation**
- **TRECVID MED 2013 for event detection experiments**
- **SVM with RBF kernel for event classification**

**Event Detection Experiments**

- VideoStory improves the learnt representation
- VideoStory improves low-level (CNN feat.)
- Term grouping improves low-level (FV + MBH feat.)

- Attributes (Jiang et al. CVPR’14) .135
- Low-Level MBH (Jiang et al. KDD’13) .174
- Term embedding - MBH .172
- VideoStory embedding - MBH .196
- CNN features (Jiang et al. CVPR’14) .198
- VideoStory embedding - CNN .243

**Event Translation Experiments**

- Ground-truth: Manual video descriptions
- More accurate video translations by VideoStory

**Conclusions**

- Improved event detection and translation by learning the semantic representation from videos + stories