Incorporating External Textual Knowledge for Life Event Recognition and Retrieval

Min-Huan Fu¹, Chia-Chun Chang¹, Hen-Hsen Huang²,³ and Hsin-Hsi Chen¹,³

¹ Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
² Department of Computer Science, National Ch Cheng Kung University, Tainan, Taiwan
³ MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan

Introduction

- Due to the increasing availability of dedicated lifelog devices, efficient method for organizing and accessing collected lifelog data is demanded
- Semantic gap between visual information from lifelogs and textual information from event-based queries is a challenge for multimedia lifelog access
- We incorporate semantic word embeddings to reduce the gap between queries and visual concepts for LSAT task and to enrich training data of supervised learning for LADT task

Image Indexing

- Each image is associated with additional visual concepts extracted by Google Cloud Vision API
- Lens calibration is performed on all images to prevent erroneous outputs from the CV models
- We further filter out images with low quality based on blurriness and color diversity detection

LSAT Framework

- Raw images
- Concepts: vehicle, boat, naval architecture, photography, automotive exterior…
- Calibrated images
- Concepts: table, desk, furniture, wood, metal, room, architecture, electronic device, vehicle, hardware…
- LSAT Framework
- Life event indexing
- Word Embeddings
- Image Preprocessing
- Image enhancement
- Color diversity
- Less calibration
- Provided visual concepts
- Single object detector
- Top K similar concepts of each query word
- User selection of relevant concepts
- Constraints on time
- BM25 Scoring
- Image concepts set as short text query
- Run ID
- RelRet
- Run01: Automatic query expansion 0.0632 0.2375 293
- Run02: Interactively selected query* 0.1108 0.3750 464
- Run03: Selected query + refinement* 0.1657 0.6833 407
- * We use the same queries for Run02 and Run03; the average interaction time of Run03 for each topic is 159.5 s

LADT Approach

- The LADT subtask is aimed at detecting and recognizing life events from sixteen types of daily activities
- We address the problem as multi-label classification and manually annotated partial dataset as training data
- Our model takes as input the visual features extracted by VGG-19 and the textual features encoded by GloVe
- To maximally exploit the knowledge inherent in word embeddings, we include semantic word similarities as weighting factors when aggregating concept words
- Self-feedback mechanism: the model can also accept its prediction in previous K time steps as additional input

LADT Experiment

- Model
  - Precision: 0.7084, 0.7535, 0.7307
  - Recall: 0.3606, 0.4332, 0.4780
  - Micro-F1: 0.4780, 0.5084, 0.5367

Conclusion

- For life moment retrieval, we introduce external textual knowledge to reduce the semantic gap between textual queries and visual concepts extracted by CV models
- For activity detection and recognition, we incorporate textual features aggregated in an unordered fashion to enrich the training data for supervised DNN models