Incorporating External Textual Knowledge for Life Event Recognition and Retrieval

Min-Huan Fu[1], Chia-Chun Chang[1], Hen-Hsen Huang[2,3] and Hsin-Hsi Chen[1,3]

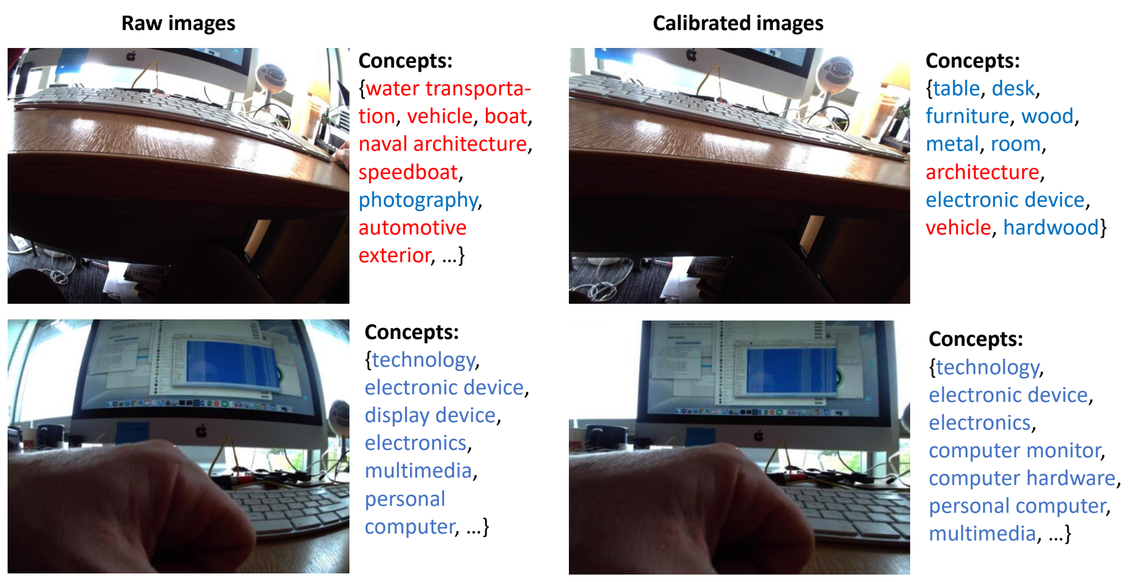
[1] Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
[2] Department of Computer Science, National Chengchi University, Taipei, Taiwan
[3] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan

Introduction

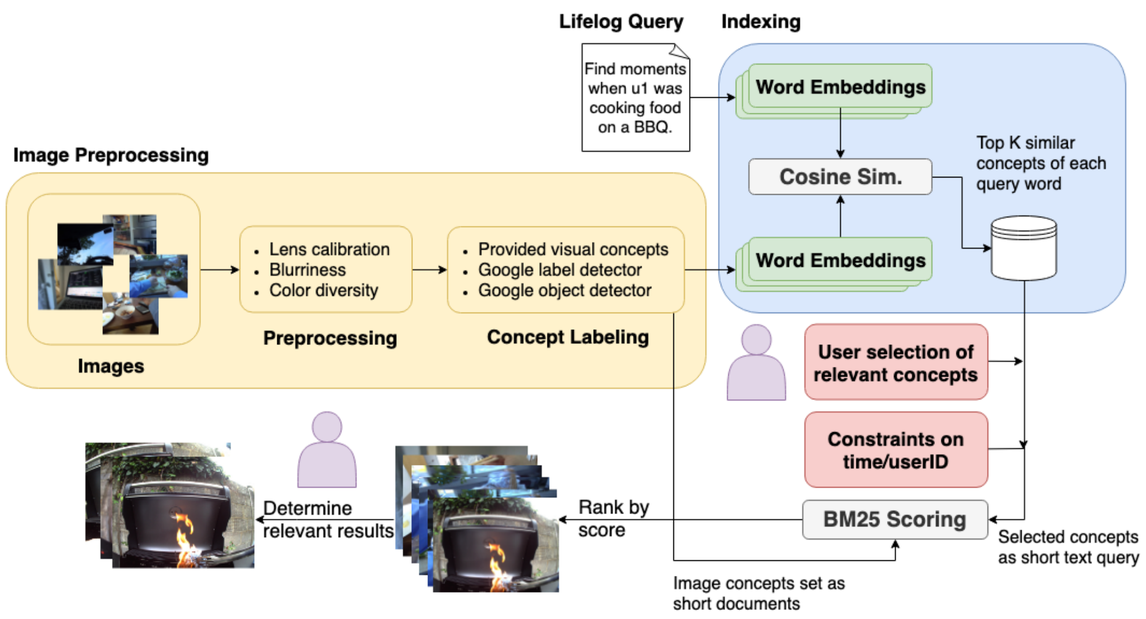
- Due to the increasing availability of dedicated lifelog devices, efficient method for organizing and accessing collected lifelog data is demanded
- Semantic gap between visual information from lifelogs and textual information from event-based queries is a challenge for multimedia lifelog access
- We 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

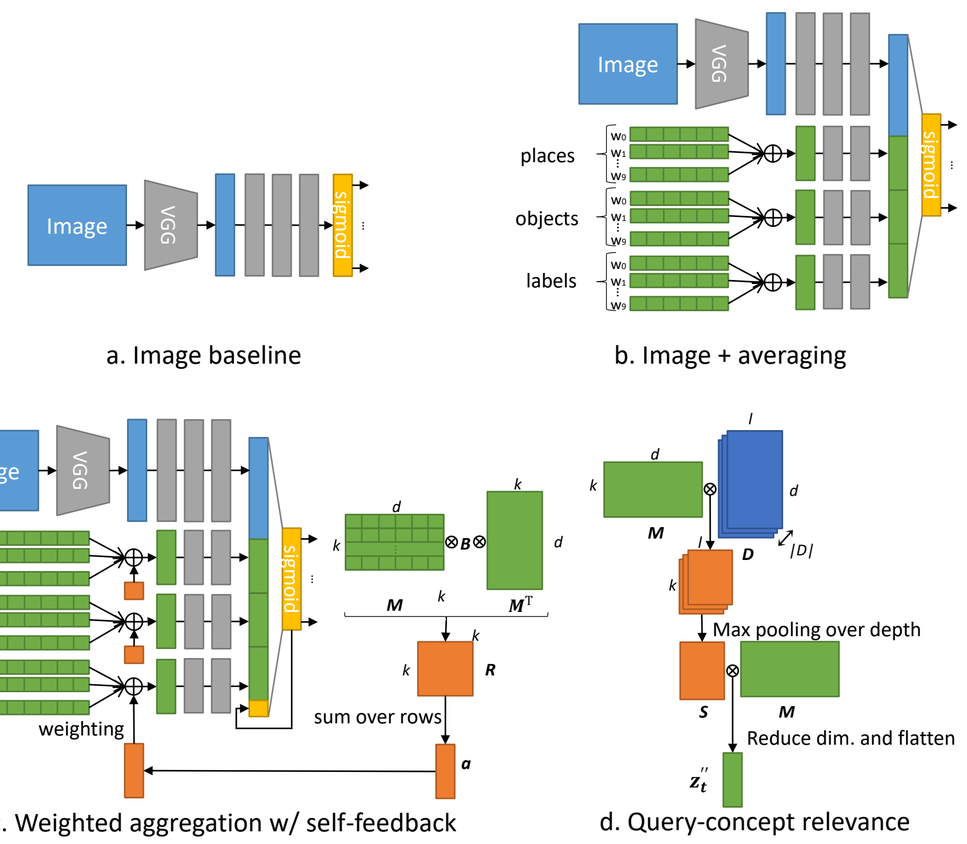


Run ID	mAP	P@10	RelRet
Run01: Automatic query expansion	0.0632	0.2375	293
Run02: Interactively selected query*	0.1108	0.3750	**464**
Run03: Selected query + refinement*	**0.1657**	**0.6833**	407

* 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



a. Image baseline
b. Image + averaging
c. Weighted aggregation w/ self-feedback
d. Query-concept relevance

LADT Experiment

Model	Precision	Recall	Micro-F1
Image (baseline)	0.7084	0.3606	0.4780
+ averaged words	0.7522	0.3840	0.5084
+ concept self-corr.	-	-	-
+ feedback	**0.7535**	0.4168	0.5367
+ concept-query corr.	0.7261	0.4023	0.5177
+ feedback	0.7307	**0.4332**	**0.5439**

- The recall score of the model increases when adopting adequate concept sets and aggregation strategies, while the precision score does not necessarily increase

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