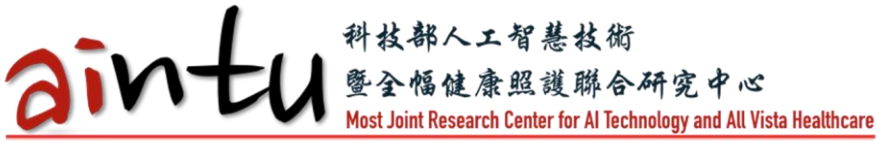
Incorporating External Textual Knowledge for Life Event Recognition and Retrieval

NTUnlg at NTCIR-14 Lifelog-3

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

[1]National Taiwan University, [2]National Chengchi University, [3]AI NTU

國立臺灣大學
National Taiwan University



NLP Lab

Introduction

- Lifelog semantic access task (LSAT)
  - Retrieve specific moments in a lifelogger's life (a known-item search task)
  - Example: *Find the moment when u1 was eating ice cream beside the sea. Find the moment when u1 was eating fast food alone in a restaurant.*

- Lifelog activity detection task (LADT)
  - Detect and recognize life event from 16 types of daily activities (a multi-label classification task)
  - Example: *traveling, face-to-face interaction, using a computer, cooking, eating, relaxing, house working, reading, socializing, shopping …*

Introduction (cont'd)

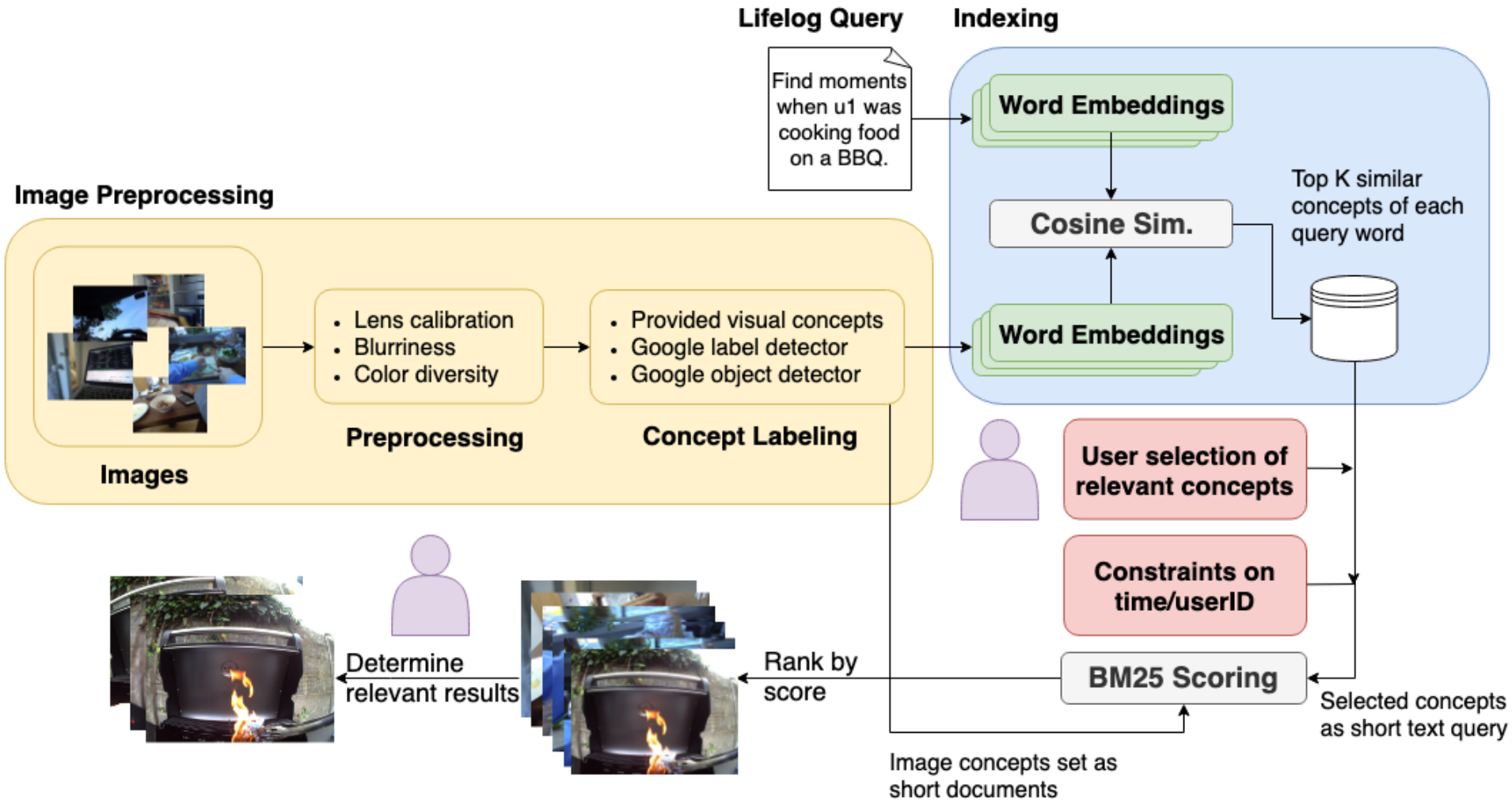
- A huge challenge for multimedia lifelog access: the semantic gap between visual and textual domains
  - Lifelogs are stored as multimedia archives (visual domain)
  - We want to retrieve life events using verbal expressions (textual domain)
- Intuitively we may exploit CV models to obtain visual concepts for lifelog images, but there is still gap between topics and concepts
- We incorporate word embeddings as external textual knowledge for both subtasks; specifically, we try to:
  - Suggest concept words related to life event topics for LSAT task
  - Enrich the training data of supervised learning for LADT task

Preprocessing

- Besides the official concepts, 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 advanced CV models
  - We further filter out images with low quality based on blurriness and color diversity detection

- We use the following visual concepts in this work:
  - Place attributes and categories from PlaceCNN (official)
  - Visual labels and objects from Google API

LSAT Framework



LSAT framework (cont'd)

- In our retrieval framework, lifelog images are represented as short documents consisting of associated concept words
- For each word in the event topic, the retrieval system suggests a list of semantically similar concept words to the user
- Users can **select concepts to formulate the query**, then our system will perform retrieval with BM25 ranking
- In the **refinement** stage, users can manually remove irrelevant images

LSAT result

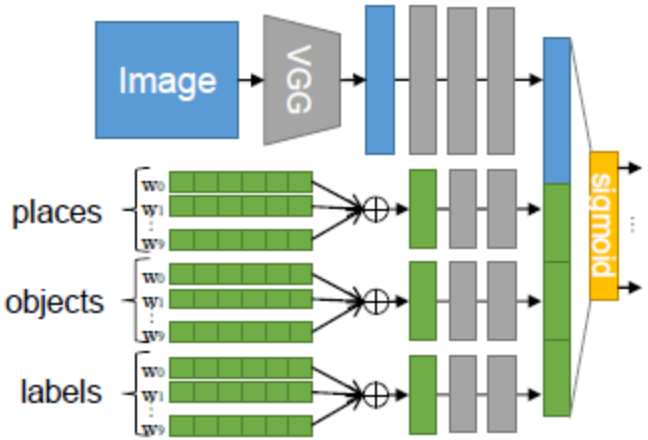
- Our interactive approach largely outperforms the automatic baseline that uses top-10 related concepts to all topic words as query
- We observed the total number of relevant documents retrieved has slightly decreased after the user refinement
  - This may result from that the user of our system is not the lifelogger himself, and possibly make wrong deletions of the relevant retrieval results

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 & Run03; the average interaction time of Run03 for each topic is *159.5 s*

LADT approach

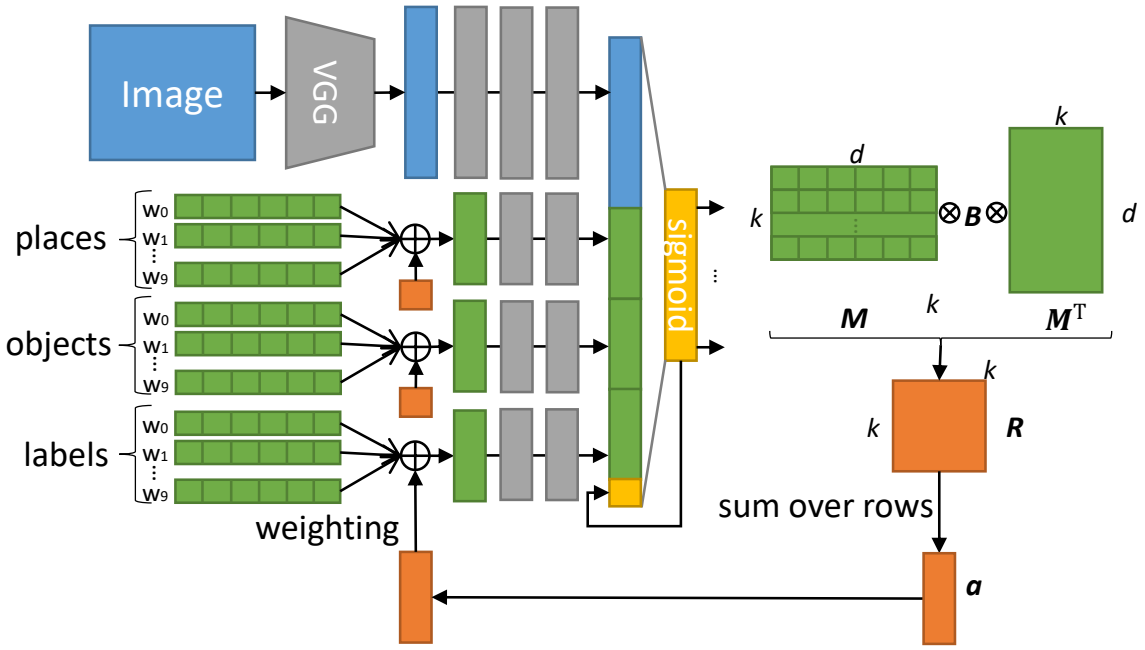
- We address LADT subtask as multi-label classification annotate partial dataset as training data
- Our proposed DNN model takes as input the visual features extracted by VGG-19 (512D) and the textual features encoded by GloVe (300D)



- One challenge to include unordered set of vectors as NN's input is that common network structures for ordered text are hardly applicable
- We adopt a similar structure to the Deep Averaging Network (DAN) to deal with the unordered input, but use weighted average instead

LADT approach (cont'd)

- We include semantic relatedness as the weighting factor
  - Concept that is more related to other concepts associated to the same image is considered more important
  - We may also measure the relatedness between concept words and activity description instead
- Self-feedback: the model can also accept its prediction in previous K time steps as additional input



LADT result

- The recall score of the model increases when we adopt proper aggregation strategies for concept words, while the precision score does not necessarily increase

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

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

Thank you!