VCI^{2}R at the NTCIR-13 Lifelog-2 LSAT Task

Presented by: Qianli Xu

Co-authors: Jie Lin, Ana del Molino, Qianli Xu, Fen Fang, V. Subbaraju, Joo-Hwee Lim, Liyuan Li, V. Chandrasekhar

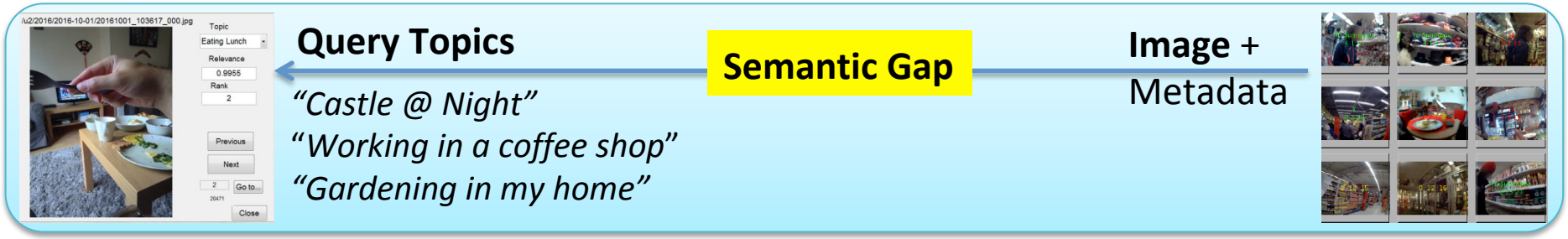
Organization: Institute for Infocomm Research, A*STAR, Singapore

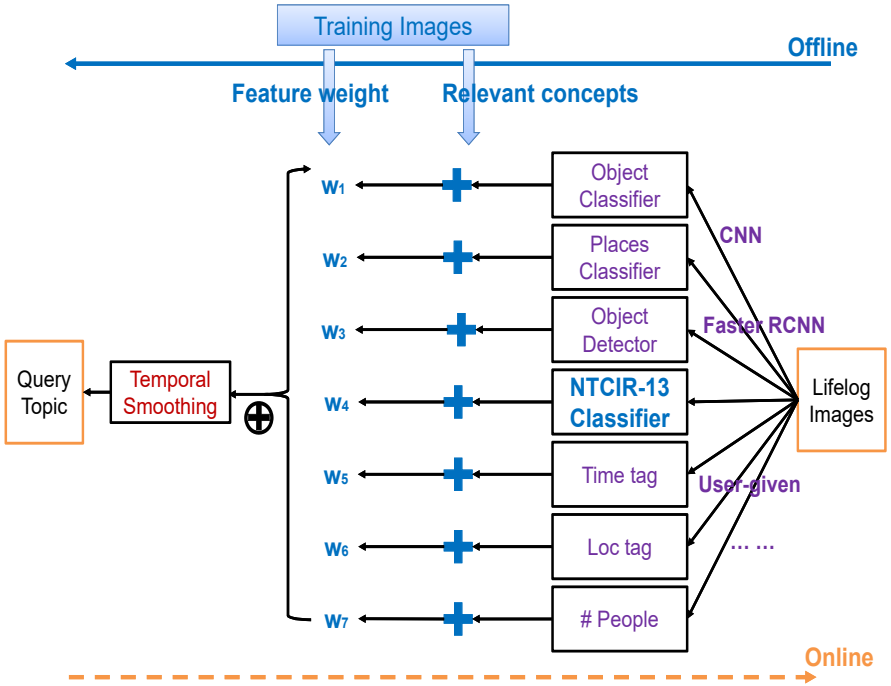
About VCI^{2}R

- Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
  - Visual Computing
  - Human Language Tech
  - Data Analytics
  - Neural Biomedical Tech
  - etc.
- Visual Computing Department
  - Video/image analytics & search
  - Augmented visual intelligence
  - Visual inspection

Website: www.a-star.edu.sg/i2r/

LSAT Framework



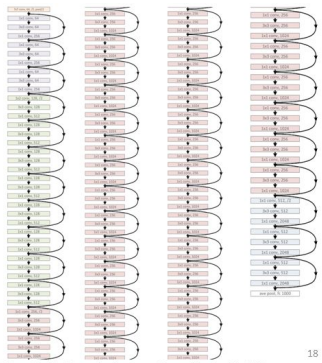


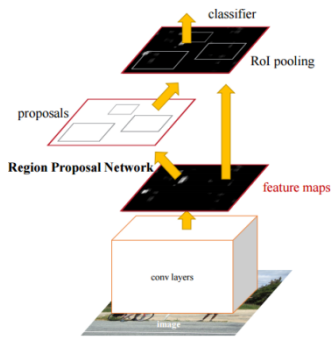
- **Relevant concepts**: What are the CNN predications relevant to query topics?
- **Feature weighting**: Which features contribute the most?
- **Temporal smoothing**: Temporal coherence, remove outliers
- **Post filtering**: refine search using location (GPS) and Time

del Molino, et al., 2017, VC-I2R at ImageCLEF2017: Ensemble of deep learned features for lifelog video summarization. *CLEF Working Notes*, *CEUR*.

1. Getting the Basic Semantics

- CNN classifiers
  - Object: ResNet152 – ImageNet1K
  - Place: ResNet152 – Place365
- CNN detector
  - Faster R-CNN – MSCOCO (80)
- NTCIR-13 classifier
  - VGG-16 – ImageNet1K
  - Replace the last layer (1K neurons) with 634 neurons
  - Sigmoid as the activation function
- Human detection and counting
  - Sighthound (https://www.sighthound.com)





COCO
Common Objects in Context

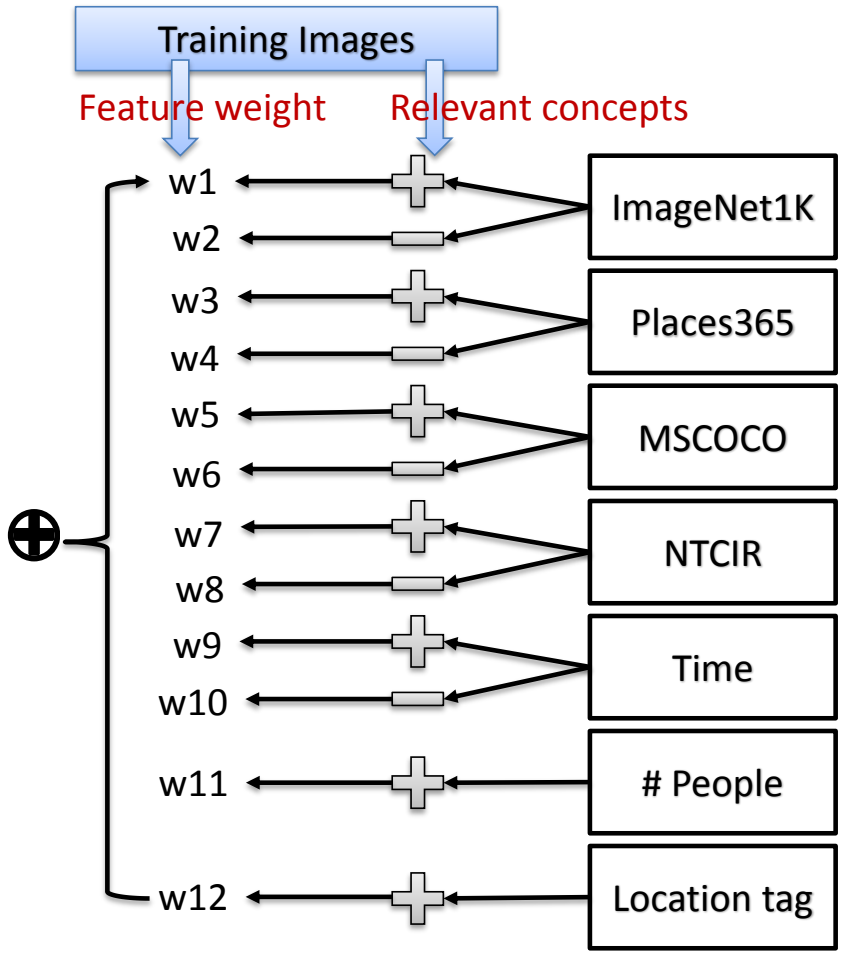
2. Aggregating & Weighing Features

Relevance mapping for each topic

Task	Objects Relevant	Objects Avoid	Places Relevant	Places Avoid	MSCOCO Relevant
1	computer group meeting	-	computer group meeting *etc.*	-	laptop keyboard
2	television food glass	computer group meeting	living room television room *etc.*	conference room lecture room *etc.*	tv remote *etc.*
3	computer group meeting	office	coffee shop living room *etc.*	conference room office *etc.*	laptop keyboard
4	computer pencil notebook	office	living room hotel room *etc.*	conference room office *etc.*	laptop book *etc.*
5	food glass	drum white goods menu'	food court restaurant *etc.*	-	fork sandwich *etc.*

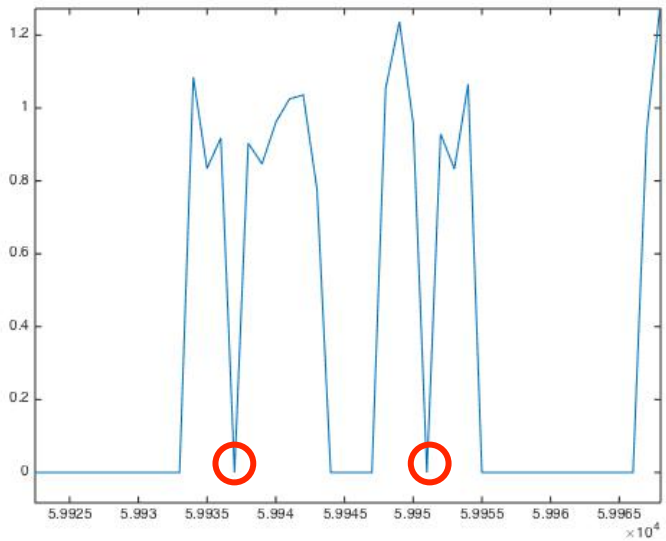
CRF for Feature weighing that accommodates individual differences

$$E_{\boldsymbol{\theta}}(\mathbf{s}) = \lambda \sum_{i} \underbrace{\phi_u(s_i)}_{\text{unary}} + \sum_{ij} \underbrace{\phi_p(s_i, s_j)}_{\text{pairwise}},$$



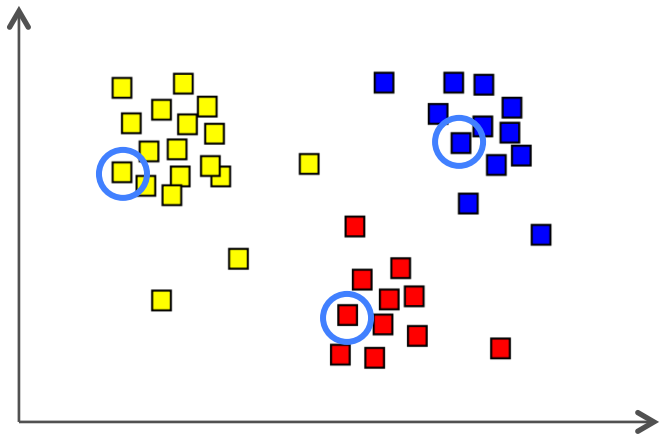
3. Temporal Smoothing

- Adjacent lifelog images may share similar event.
- Temporal smoothing is used to ensure the semantic coherence.
- A triangular window of size ***w*** is used. ***w*** is adaptive to event topics.



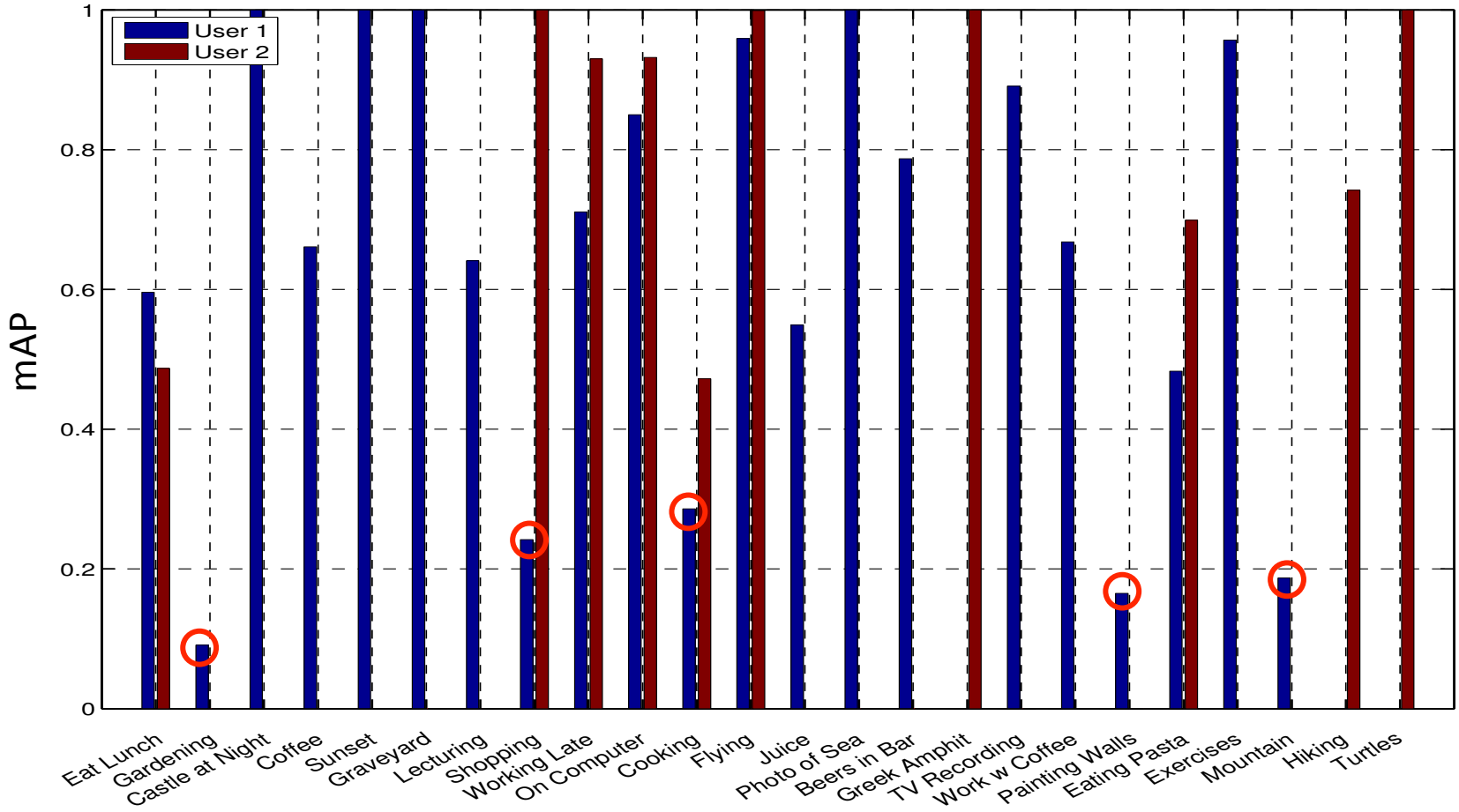
4. Post-filtering

- Increase diversity of retrieved images (avoid retrieving images of the same event)
- Use time and location (GPS) to filter images
- Exclude images that are closer in time and location.

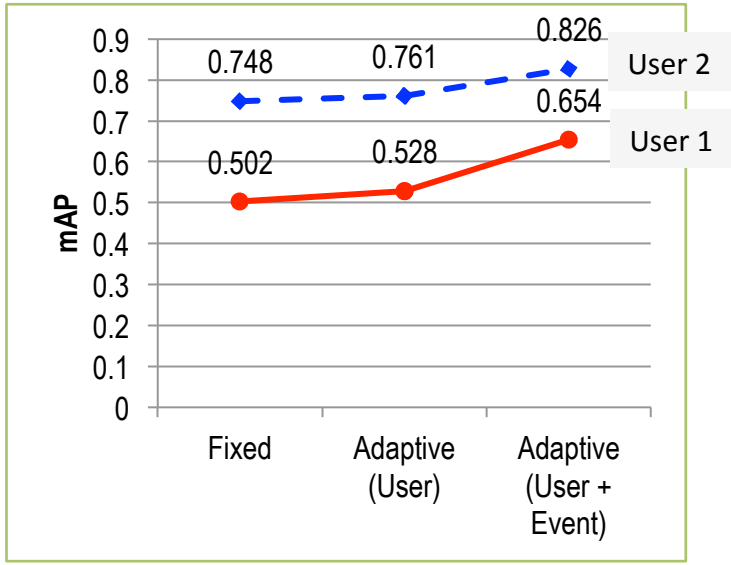


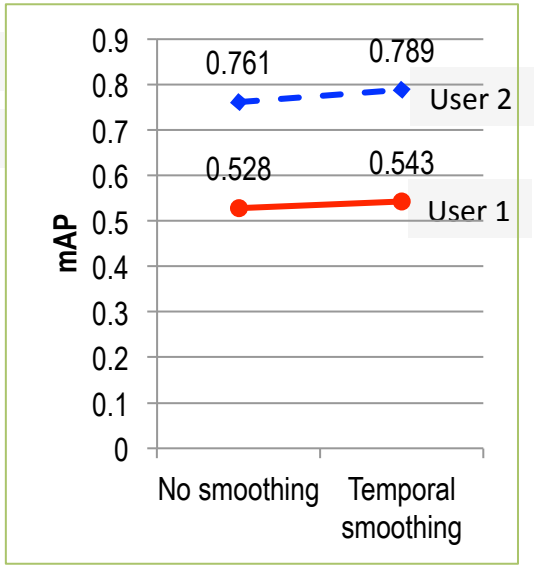
Result

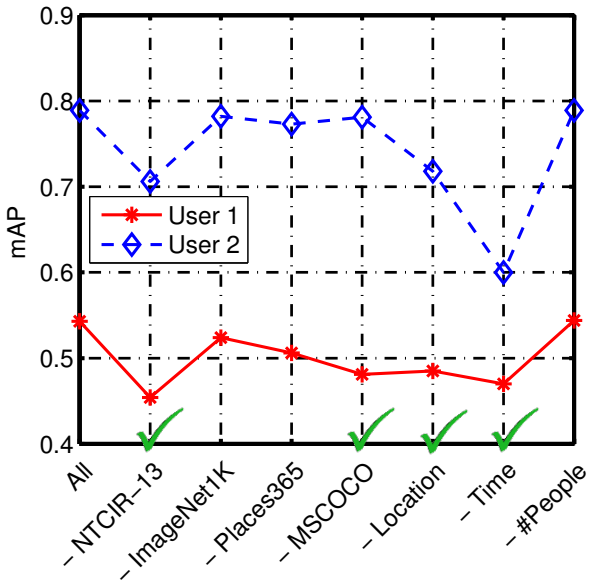
- Official score (precision): 57.6%



Analysis (Fine-tuning)







Effect of threshold for relevant concept searching

Semantic concepts which activation level is above the threshold is considered relevant to the query topic

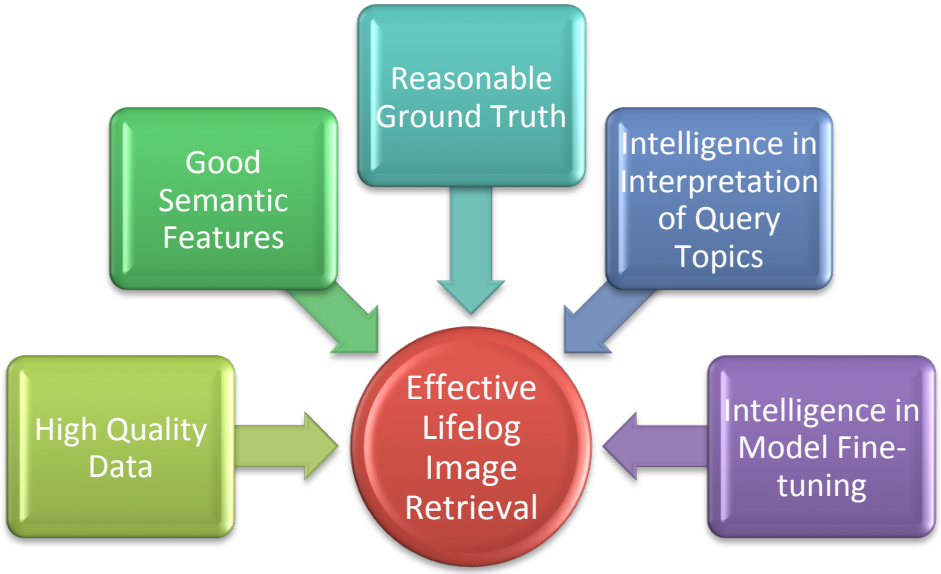
Effect of temporal smoothing

Whether temporal smoothing is performed or not

Feature importance

Decrease in performance when we remove one type of feature. The bigger the decrease, the more important the feature.

Summary



- A lot of fine-tuning and manual intervention are involved in the retrieval → Over-fitting?
- "Relevant" concepts may not be contributing, and *vice verse*.
- Interactive retrieval is probably a good intermediate solution.

LIT

Email: qxu@i2r.a-star.edu.sg