VCI2R at the NTCIR-13 Lifelog-2 Lifelog Semantic Access Task

Jie Lin
Institute for Infocomm Research, A*STAR, Singapore
lin-j@i2r.a-star.edu.sg

Ana Garcia del Molino
Institute for Infocomm Research, A*STAR, Singapore
Nanyang Technological University, Singapore
stugdma@i2r.a-star.edu.sg

Qianli Xu*
Institute for Infocomm Research, A*STAR, Singapore
qxu@i2r.a-star.edu.sg

Fen Fang
Institute for Infocomm Research, A*STAR, Singapore
fang_fen@i2r.a-star.edu.sg

Vigneshwaran Subbaraju
Singapore Bioimaging Consortium, A*STAR, Singapore
Vigneshwaran_Subbaraju@sbic.a-star.edu.sg

Joo Hwee Lim
Institute for Infocomm Research, A*STAR, Singapore
joohwee@i2r.a-star.edu.sg

ABSTRACT

In this paper we describe our automatic approach for the NTCIR-13 Lifelog Semantic Access Task. The task is to query relevant lifelog images from user's daily life given an event topic. A major challenge is how to bridge the semantic gap between lifelog images and event-level topics. We propose a general framework to address this problem, with key components of various CNNs to translate lifelog images to object and scene features, relevant object/scene concepts searching for events, feature weighting adapted to events, and temporal smoothing to incorporate semantic coherence into the similarity between each image and query event. We achieved an official result 57.6% in terms of mean precision over 20 topics. We also analyze the effect of key components to the retrieval system.

Keywords

NTCIR, Lifelog, CNN, Faster R-CNN, Multi-modality, CRF

Team Name

VCI2R

Subtasks

Lifelog Semantic Access Task

1. INTRODUCTION

This paper addresses the problem of user-specific event retrieval from user's daily life, to answer the challenge of NTCIR-13 Lifelog-2 Lifelog Semantic Access Task (LSAT) [5]. Given a query event for a user, the task is to retrieve relevant moments from a set of lifelog images associated with meta-features (location, timestamp, etc.) collected by the user. Here, a moment is defined as sequential lifelog images that are relevant to the query event.

*Jie Lin, Ana Garcia del Molino and Qianli Xu contributed equally to this work.

The ultimate goal of LSAT is to bridge the so called semantic gap between lifelog images and event-level query topics specified by the organizers. To this end, the keys are "what" (634 visual concepts), "where" (user-given location tags for partial images) and "when" (image recorded time) information associated with lifelog images. Beyond that, external knowledge such as ImageNet1K and Places365 concepts may be complementary to the "what" and "where" information provided. In this work, we propose a deep learning based framework to integrate "what", "where" and "when" together towards better understanding user's lifelog data, and submitted automatic runs to LSAT. In Section 2, we give an overview of the proposed framework. Section 3 introduces the key components of the framework. We present experimental results and discussions in Section 4.

2. FRAMEWORK OVERVIEW

Figure 1 shows an overview of the framework, which consists of 4 key components. At offline stage:

- Lifelog images are represented as semantic feature vectors. Convolutional Neural Networks (CNN) based classifier and detector are respectively applied to translate each lifelog image to a set of feature vectors, with each element representing the probability that an object/scene is depicted in the image. These CNN models include object-centric classifier pre-trained on ImageNet1K, scene-centric classifier pre-trained on Places365, hybrid classifier fine-tuned on 634 visual concepts provided by the LSAT organizers and object detector pre-trained on MS COCO. Besides, location tags and time stamps (in hours) associated with lifelog images are also converted to 0/1 vectors, respectively.
- Second, for each type of features (objects, scenes, tags, etc.), concepts with high responses to an event are considered as relevant concepts to that event.
- Further, considering that events may rely on objects more than locations, or vice versa, feature importances

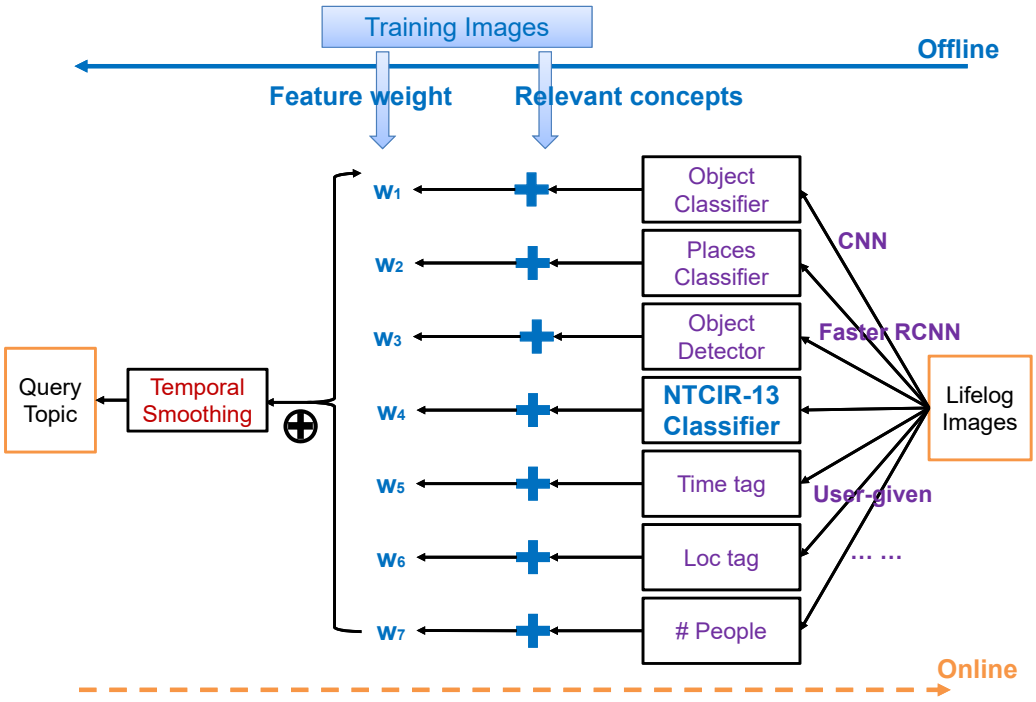

Figure 1: System framework. More details can be found in Section 2 and Section 3.

are learned with a Conditional Random Field (CRF) [11] model to weight the contributions of features for specific events.

- Since the similarity of an image to an event is derived without considering adjacent image frames, temporal smoothing is incorporated into the framework to refine the similarity, with the assumption that adjacent images are with semantic coherence.

At online stage, given a query event for a user, we need to compute the similarity between each of the user's lifelog images and that event, then rank all images in descending order of their similarities. In particular, the similarity is computed as follows. For each type of feature, a relevance score between an image and an event can be derived by average pooling over the probabilistic activations of this event's relevant concepts. Subsequently, the similarity is computed by aggregating weighted-sum of all relevance scores, followed by temporal smoothing to refine the similarity.

3. APPROACH

3.1 CNN Features

To associate lifelog images to events, one can make use of the objects present in an image and the location information of an image, (i.e. where the photo was taken). Similar to our previous work [3], we try to combine "what" and "where" of the lifelog images, which were extracted using deep CNN [7].

NTCIR-13 Classifier.

NTCIR-13 LSAT organizers provide 634 visual concepts annotated for 53,347 lifelog images, with 6 labels per image on average. One may note that these annotations are generated using the Computer Vision API from MS, which is not necessarily relevant to the lifelog images. We choose VGG-16 pre-trained on ImageNet1K as our starting point, replace the last layer (1000 neurons) with 634 neurons, followed by sigmoid as the activation function instead of softmax due to the multi-label settings. Finally, cross-entropy loss is used to optimize network weights. We fine-tune the VGG-16 with batch size 32 and learning rate 0.0001. We train the network for 20,000 iterations with SGD optimizer on a NVIDIA Titan X GPU.

CNN Classifiers.

For all the lifelog images, objects and places are predicted based on CNN that is pre-trained on ImageNet1K [2] and Places365 [14]. ImageNet1K is a dataset with 1.2 million images, each annotated according to 1000 object classes. Similarly, Place365 has 1.8 million images, each tagged against 365 place categories. ResNet with 152 layers [6] is pre-trained on each dataset respectively, which we refer to as ResNet152-ImageNet1K and ResNet152-Places365. Given a lifelog image, we pass it through ResNet152-ImageNet1K resulting in a 1000-dimensional probability vector. Similarliy, we use ResNet152-Place365 to extract a 365-dimensional probability vector to predict place information. Both vectors are extracted from the last layer of the network [3]. As in [4], data augmentation is performed to generate scaled and rotated versions for each lifelog image. Instead of average operation, the maximum activation value is chosen for

each class.

CNN Detector.

Next, we enhance object recognition by extracting location information based on object detection. To do so, we use a Faster R-CNN [10] with Inception-ResNet [12] as the base CNN architecture. The network is pre-trained on Microsoft Common Object in Context (MSCOCO) dataset [8]. The MSCOCO training set has over 200,000 images annotated against 80 object classes, with location information (bounding box over identified objects). Some of the categories are relevant to the lifelog images (e.g. cup and bowl, laptop, etc.). Using Faster R-CNN, a lifelog image is annotated with top 20 detections based on the maximum probability for each category [3].

People Counting.

For event queries like "Presenting/Lecturing", the number of people presented in an image is vital for determining the relevance of this image to the event. People counting has been studied in several earlier works and more recently it is available as a service such as the API provided by *Sighthound, Inc* (https://www.sighthound.com/docs/cloud/detection/). We have earlier used this API in our previous work [3], and we found that it provided many convenient features, e.g. detecting (along with a bounding box) and counting people and its performance was found to be good [1, 9].

3.2 Relevant Concepts Searching

Lifelog images are represented as semantic features including CNN predictions (NTCIR-13 concepts, objects, places, etc.), time and location vectors. To link semantic concepts to target event topics, we propose to automatically search relevant concepts from each type of semantic features to a given event. To this end, we manually annotate a subset of lifelog images with event labels, which is used to validate the effectiveness of relevant concepts searching. With the observation that concepts with high activations to an event are more likely to be relevant to that event, we average the activations of lifelog images annotated with a given event and binarize to 0/1 vector with a pre-defined threshold, with 1 denoting that the corresponding concept is relevant to the event. This is performed for each type of semantic features independently. The sub-optimal threshold is determined with greedy search, by testing the retrieval performance over the annotated training set. It is worth noting that we studied 2 thresholding strategies. One is that the threshold is adapted to user only; the other more advanced strategy adopts threshold tailored for each user and event.

3.3 Feature Weighting

There are 7 types of features (NTCIR-13, ImageNet1K, Places365, MSCOCO, Location, Time and # of People). For each of these features, there exist relevant concepts, *Rel.*, (*e.g. food* in *Eating Lunch*), and concepts to avoid, *Avoid*, (*e.g. kitchen* in *Hiking*). For each type of features, a relevance score is computed by averaging the activations from the last network layer over their respective relevant concepts.

As the query event probably favors one type of features more than the others, we use a probabilistic approach based on active inference in CRF [11] to learn adaptive feature importances tailored for event topics. In this formulation, there is one node per feature. The unaries are defined as $\phi_u(s_i) = mean(score_{rel}[gt = s_i])$ for $s_i = \{0, 1\}$, where gt is our annotation whether each image corresponds to the task. The pairwise potentials are defined to enforce that the nodes activation values be positive.

Table 1: Event Topics for User 1 and User 2.

Event Topic	User 1	User 2
Eating Lunch	Y	Y
Gardening	Y	N
Castle at Night	Y	N
Coffee	Y	N
Sunset	Y	N
Graveyard	Y	N
Presenting/Lecturing	Y	N
Grocery Shopping	Y	Y
Working Late	Y	Y
On the Computer	Y	Y
Cooking	Y	Y
Flying	Y	Y
Fruit or Vegetable Juice	Y	N
Photo of the Sea	Y	N
Having Beers in a Bar	Y	N
Greek Amphittheatre	N	Y
Television Recording	Y	N
Working in a Coffee Shop	Y	N
Painting Walls	Y	N
Eating Pasta	Y	Y
Exercises	Y	N
Benbulbin Mountain	Y	N
Hiking	N	Y
Turtles	N	Y

3.4 Temporal Smoothing

Adjacent lifelog images probably share similar event topics. Thus, temporal smoothing is proposed to ensure the semantic coherence along temporal dimension. In particular, the similarity between an image and an event is smoothed using a triangular window of size w, which is adaptive to event topics. Again, we perform greedy search to find the sub-optimal value of w, by testing their retrieval performances on the manually annotated lifelog images.

4. EXPERIMENTS

4.1 Training Dataset

There are 91,044 and 20,471 lifelog images for user 1 and user 2, respectively. The organizers identified 24 event topics, where 21 topics are applicable to user 1, and 10 topics are applicable to user 2. Table 1 shows the topic distributions for user 1 and user 2.

To identify relevant concepts, feature weights and temporal smoothing parameters for event topics, a subset of lifelog images are manually annotated with these topics. Specifically, we sampled 22,304 (18,209 for user 1, and 4,095 for user 2) lifelog images, which is 1/5 of the entire dataset. Among them, 4,857 images are relevant to the 24 topics, where 4,175 for user 1 and the rest for user 2. We randomly sample half of the relevant images for training, the remaining for test.

Table 2: Effect of thresholds for relevant concepts searching.

	User 1	User 2
Fixed	0.502	0.748
Ada (User)	0.528	0.761
Ada (User+Event)	0.654	0.826

Table 3: Effect of temporal smoothing.

Temporal Smoothing?	User 1	User 2
No	0.528	0.761
Yes	0.543	0.789

4.2 Official Results

The official evaluation evaluates the number of events detected in a given day (compared to the ground truth) as well as the accuracy of the event-detection process (given a sliding five minute window). The used metrics are precision and recall, and the official score is the mean of precisions over the topics. Due to the complexity and difficulty of the queries, topics 16, 20, 23, and 24 are discarded, and only evaluated the rest 20 topics. The official score reported for our team is 57.6%, which ranked at the first place.

4.3 Analysis

Besides the official results, we also use mean Average Precision (mAP) as our own evaluation metric to study the effect of key components in the proposed framework. The results are reported on the sampled test set with ground-truth annotated by our team.

Effect of Thresholds.

We explores the effect of thresholds for relevant concepts searching. Two configurations for the thresholds are tested: (1) the thresholds are adaptive to each user, and (2) the threhold are adaptive to both user and event, which is more advanced than the first configuration. As shown in the Table 2, both configurations outperforms fixed thresholds. Moreover, the advanced configuration improves the first one by a large margin.

Temporal Smoothing.

Table 3 studies the effect of temporal smoothing to the system, with thresholds for relevant concepts searching fixed. One may note that there are consistent improvements over both users.

Feature Importances.

Figure 2 compares how different features are important for the retrieval task. "All" denotes all features are used, while "- NTCIR-13" means the NTCIR-13 classifier feature is removed from "All" in the retrieval system, same for the other configurations. A lower score of "- NTCIR-13" means it causes more performance drop, indicating the repsective feature is important to the retrieval. We observe that "NTCIR-13" is the most important feature to the system, followed by time, MSCOCO, and location among all the CNN based features.

Event-level Results.

Figure 3 shows retrieval mAP for all event topics, using our best model. One can see that for user 1, our system performs worse on topics like "Gardening", "Grocery Shopping" and "Painting Walls".

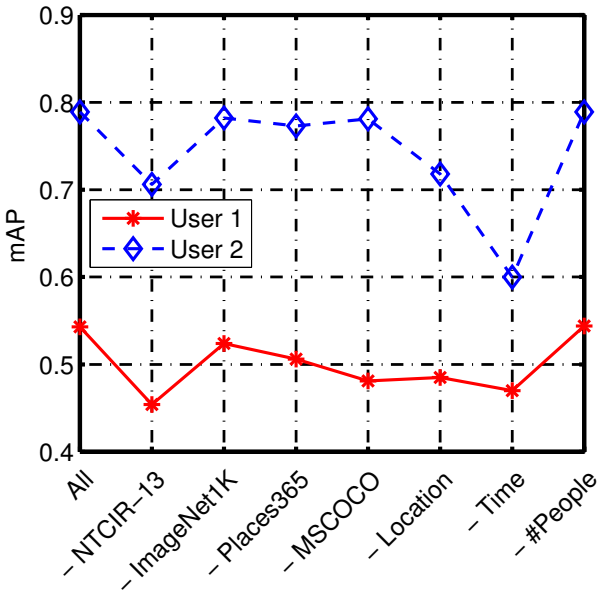

Figure 2: Comparison of feature importances to the retrieval system.

4.4 Application in LIT task

The method proposed in this paper is used in [13] for annotating activities with respect to the NTCIR-13 Lifelog-2 Lifelog Insight Task (LIT). Ten activities are defined for LIT, namely, eating, walking, running, hiking, gym/yoga, socializing, taking bus, driving a car/taking a taxi, taking train, and in a flight. Similar to the LSAT topics, the LIT activities have varying level of abstraction and the number of incidences ranges from a few to thousands. Our algorithm achieves similar level of precision and recall in the LIT activity retrieval. The result has been effectively used for insights generation.

5. CONCLUSIONS

This paper focuses on the problem of event driven lifelog image retrieval. We presented a general deep learning based framework to address a major challenge of the task - bridging the gap between visual images and high-level event concepts. We submitted the generated retrieval results to the NTCIR-13 Lifelog-2 Lifelog Semantic Access Task. Promising results has been officially reported, demonstrating the effectiveness of the proposed retrieval system.

6. ACKNOWLEDGEMENTS

The work is funded by the Singapore A*STAR JCO VIP REVIVE Project (1335h0009).

7. ADDITIONAL AUTHORS

Liyuan Li (Institute for Infocomm Research, A*STAR, Singapore. email: lyli@i2r.a-star.edu.sg) and Vijay Chandrasekhar (Institute for Infocomm Research, A*STAR, Sin-

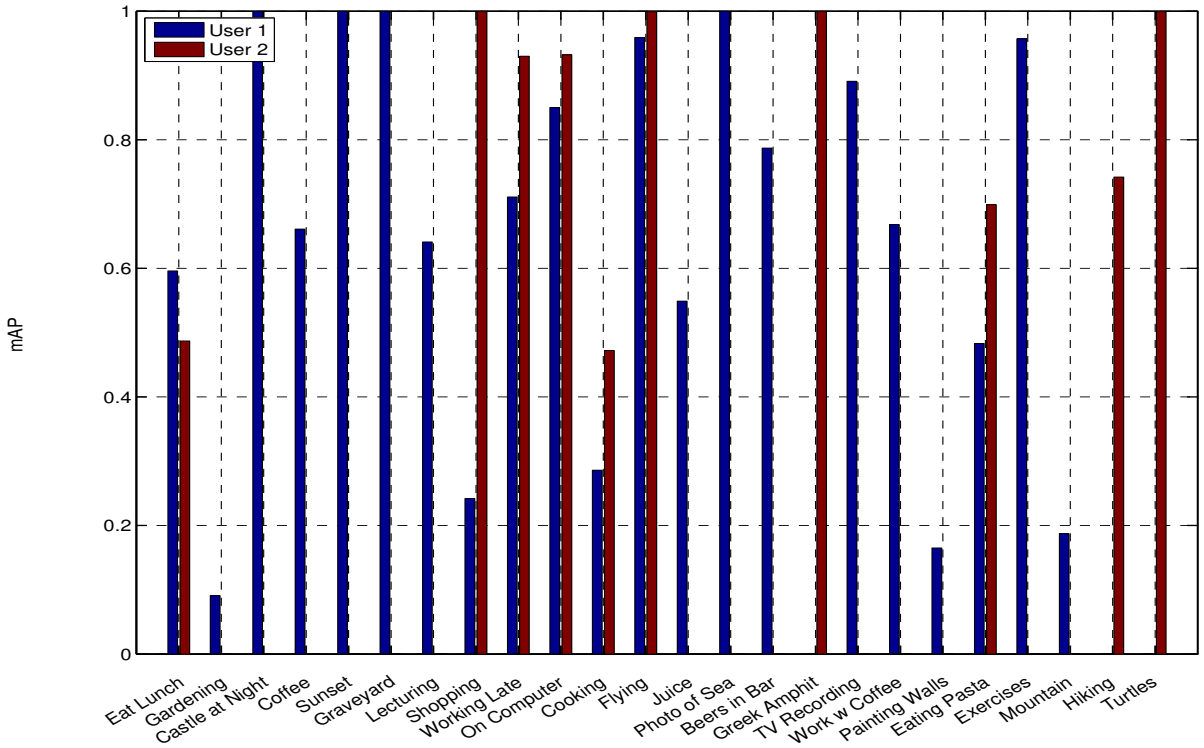

Figure 3: Event-level retrieval results.

gapore, Nanyang Technological University, Singapore. email: `vijay@i2r.a-star.edu.sg`).

8. REFERENCES

[1] A. Dehghan, E. G. Ortiz, G. Shu, and S. Z. Masood. Dager: Deep age, gender and emotion recognition using convolutional neural network. *arXiv preprint arXiv:1702.04280*, 2017.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[3] A. Garcia del Molino, M. Bappaditya, J. Lin, J.-H. Lim, S. Vigneshwaran, and C. Vijay. VC-I2R at ImageCLEF2017: Ensemble of deep learned features for lifelog video summarization. In *CLEF working notes, CEUR*, 2017.

[4] A. Garcia del Molino, Q. Xu, and J.-H. Lim. Describing lifelogs with convolutional neural networks: A comparative study. In *Proceedings of the first Workshop on Lifelogging Tools and Applications*, pages 39–44. ACM, 2016.

[5] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, D.-T. Dang-Nguyen, R. Gupta, and R. Albatal. Overview of NTCIR-13 lifelog-2 task. In *Proceedings of NTCIR-13, Tokyo, Japan*, 2017.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[8] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In *ECCV*, 2014.

[9] S. Z. Masood, G. Shu, A. Dehghan, and E. G. Ortiz. License plate detection and recognition using deeply learned convolutional neural networks. *arXiv preprint arXiv:1703.07330*, 2017.

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[11] G. Roig, X. Boix, R. De Nijs, S. Ramos, K. Kuhnlenz, and L. Van Gool. Active map inference in crfs for efficient semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2312–2319, 2013.

[12] C. Szegedy, S. Ioffe, and V. Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016.

[13] Q. Xu, S. Vigneshwaran, A. G. del Molino, J. Lin, F. Fang, J.-H. Lim, L. Li, and V. Chandrasekhar. Visualizing personal lifelog data for deeper insights at the NTCIR-13 Lifelog-2 task. In *Proceedings of NTCIR-13, Tokyo, Japan*, 2017.

[14] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. In *arXiv:1610.02055*, 2016.