# Visual Concept Selection with Textual Knowledge for Understanding Activities of Daily Living and Life Moment Retrieval

Tsun-Hsien Tang[12]*, Min-Huan Fu[1]*, Hen-Hsen Huang[1],
Kuan-Ta Chen[2] and Hsin-Hsi Chen[13]

[1] Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan
[2] Institute of Information Science, Academia Sinica, Taipei, Taiwan
[3] MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
{thtang, mhfu, hhhuang}@nlg.csie.ntu.edu.tw;
swc@iis.sinica.edu.tw; hhchen@ntu.edu.tw

**Abstract.** This paper presents our approach to the task of ImageCLEFlifelog 2018. Two subtasks, activities of daily living understanding (ADLT) and life moment retrieval (LMRT) are addressed. We attempt to reduce the user involvement during the retrieval stage by using natural language processing technologies. The two subtasks are conducted with dedicated pipelines, while similar methodology is shared. We first obtain visual concepts from the images with a wide range of computer vision tools and propose a concept selection method to prune the noisy concepts with word embeddings in which textual knowledge is inherent. For ADLT, the retrieved images of a given topic are sorted by time, and the frequency and duration are further calculated. For LMRT, the retrieval is based on the ranking of similarity between image concepts and user queries. In terms of the performance, our systems achieve 47.87% of percentage dissimilarity in ADLT and 39.5% of F1@10 in LMRT.

**Keywords:** Visual Concept Selection, Distributed Word Representation, Lifelog.

## 1 Introduction

Wearable devices for personalized multimodal recording, together with dedicated lifelogging applications for smartphones, become more popular nowadays. For example, gadgets like GoPro and Google Lens have already attracted consumers' attention, and new kinds of media like Video Weblog (VLog) emerged on Youtube heavily rely on these devices. On the other hand, numerous personalized data that are acquired, recorded, and stored still remain challenging to access by their owners. As a result, a system that supports human to make summarization and recap precious life moments is highly demanded.

* Equal Contribution

In ImageCLEFlifelog 2018 [1,2], two subtasks are conducted to address the issue of image-based lifelog retrieval. The first subtask, activities of daily living understanding (ADLT), is aimed at providing a summarization of certain life events for a lifelogger. The second subtask, lifelog moment retrieval (LMRT), is aimed at retrieving specific moments in a lifelog such as shopping in a wine store. A key challenge in both subtasks is the semantic gap between the textual user queries and the visual lifelog data. Users tend to express their information needs in higher-level, abstract descriptions such as shopping, dating, and having a coffee, while the visual concepts that computer vision (CV) tools extract from images are usually a set of concrete objects such as cup, table, and television. The approaches proposed by previous work [3,4] focus on dealing with the visual information. In this work, we attempt to reduce the semantic gap by using both visual information and textual knowledge. We propose a framework that integrates visual and textual information extracted from advanced CV and natural language processing (NLP) technologies.

## 2 Related Work

In this section, we briefly discuss recent works on lifelog retrieval. For a retrieval model, relevance and diversity are two major criteria to achieve. The retrieval of relevant lifelog data is usually based on modeling the similarity between the textual concepts, from the user query, and the visual features, from the lifelog data. Diversity, on the other hand, can be improved by image clustering. For example, Liting et al. [3] propose a method based on textual concept matching and hierarchical agglomerative clustering. Ana et al. [4] propose a method based on both visual and metadata information along with clustering on results. So far, various techniques for image processing and image retrieval have been applied to lifelog retrieval, but relatively few NLP techniques are explored in this area.

As deep neural networks have achieved remarkable success in computer vision, it is also tempting to use deeply learned features in lifelog retrieval. For example, features generated from CNN are adopted to the lifelog retrieval task [3]. Models for textual concept extraction include image classification model [5] or object detection model [6,7]. For textual knowledge modeling, deep neural networks also benefit distributed word representations, also known as word embeddings, where every word is represented in a vector in a dense space. There are different implementations for learning word embeddings [8,9,10,11,12,13], which can be further used in multi-modal applications.

## 3 Retrieval Framework

Fig. 1 illustrates our proposed system. The two subtasks, ADLT and LMRT, share a similar framework. The details of our framework are introduced in the following subsections.

### 3.1 Visual Concept Extraction

Both subtasks rely on the information from the provided image set. We extract the visual concepts from each image by using a wide range of image recognition tools. Before that, preprocessing is performed to improve the image recognition. The images in the lifelog data are automatically taken with a wearable camera so that many of them suffer from the poor quality such as overexposed, underexposed, out of focus, or ill-composed. We apply blurriness detection and pixel-wise color histogram [4] to filter out those uninformative images. For the qualified images, several image recognition tools are integrated to extract visual concepts from a variety of aspects.

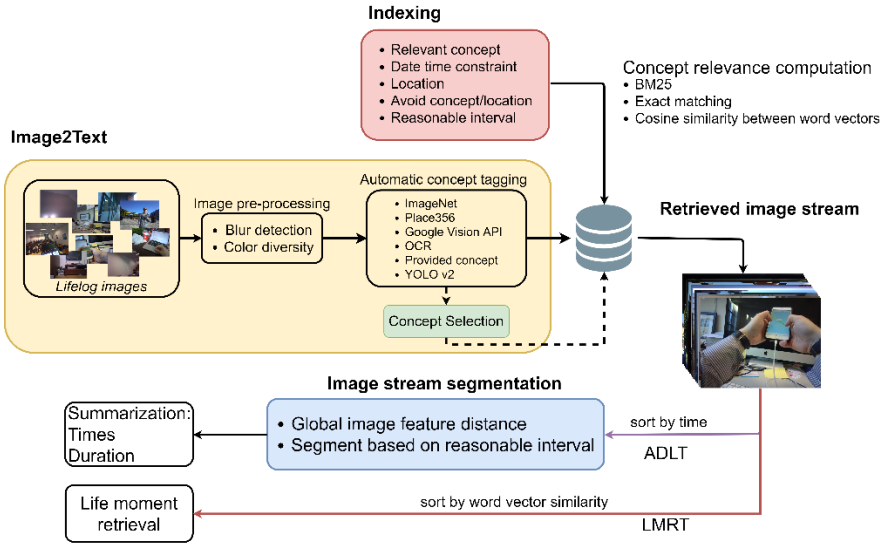



**Fig. 1.** System framework.

**Image Filtering.** We prune low quality images with blurriness and color diversity detection. The blurriness metric is defined based on the variation of the Laplacian. We perform convolution on each image with the Laplacian filter (3x3 kernel), and calculate the blurriness score as the variance of the convolved result. The images with a variance below a threshold are considered blurry and undesirable. Moreover, images with a high color homogeneity are also considered uninformative, and can be detected with quantized color histograms.

**Concept Labeling.** In order to retrieve lifelog data according to the query style defined in the two subtasks, effective textual representation for images is crucial. For a given photo, we would like to know where it was being taken, what objects are in it, and even what action the lifelogger took at that moment. To extract visual concepts from different aspects, deep learning models have shown breakthrough results in recent years.

Basically, general concepts and scene of images can be captured by two DenseNet [5] classifiers pre-trained on ImageNet1K [14] and Place365 [15], respectively. We

consider classes with output probability beyond the threshold as labels. The threshold is moderate to ensure the recall rate. For details present in images, object detection techniques Yolo-v2 [5] and Faster RCNN [7] are used. Both tools are pre-trained on MS COCO [16] and Open Images [17] datasets.

Open Images dataset, which consists of 15,440,132 boxes on 600 categories in a domain closer to human daily life, covers most of the topics in the subtasks. MS COCO Dataset consists of 91 objects types with 2.5 million labels in 328k images. We also utilize the image analysis function provided by Google Cloud Vision API[1]. The online service provides not only fruitful labels but also supports optical character recognition (OCR), which helps detect and extract text from images. Note that the image concepts provided by organizer are also added in. After going through the above tools, an image would be tagged with concepts present in various aspects, as shown in Fig. 2.

Provided concept: ['indoor', 'man', 'person']
ImageNet concept: ['patio', 'torch', 'bucket', 'Christmas_stocking', 'lab_coat']
Place365 concept: ['indoor', 'airplane_cabin', 'fastfood_restaurant', 'socializing', 'working']
Object detection: ['person', 'cup', 'person']
Google vision api:
- Labeling: ['car', 'man', 'fun', 'male', 'vehicle', 'recreation', 'world', 'vacation', 'tree']
- text detection: ['cafe', 'the']

(a) Image ID: 20160815_150731_000


Provided concept: None
ImageNet concept: [ 'shoe_shop', 'confectionery', 'shopping_cart', 'grocery_store', 'shopping_basket']
Place365 concept: ['indoor', 'supermarket', 'shopping', 'waiting in line', 'paper']
Object detection: ['person', 'person']
Google cloud vision api:
- Labeling: ['supermarket', 'grocery store', 'retail', 'product', 'convenience store', product', 'convenience food', 'aisle', 'outlet store', 'grocer']
- Text detection: ['k', 'ware', '16', 'home', 'baking']

(b) Image ID: 20160820_101541_000

Provided concept: None
ImageNet concept: ['notebook', 'monitor', 'dining_table', 'desk', 'goblet', 'laptop', 'desktop_computer', 'screen', 'mouse', ]
Place365 concept: ['indoor', 'home_office', 'computer_room', 'office', 'office_cubicles', 'conference_room', 'wood', 'reading', 'working', 'studying']
Object detection: ['laptop', 'wine glass', 'dining_table']
Google cloud vision api:
- Labeling:['laptop', 'technology', 'electronic device', 'table', 'personal_computer', 'furniture', 'desk', 'display device', 'product design', 'office']
- Text detection: None

(c) Image ID: 20160827_093525_000

**Fig. 2.** Lifelog images with their corresponding visual concepts labeled by various image recognition models.

Image recognizers are hardly to perfectly label the concepts. For instance, most of the deep learning tools cannot capture the place of Fig. 1(a) except OCR, which identifies the keyword *Cafe*. In our framework, the ensemble of the outputs of all image recognizers is considered as a set of candidate visual concepts.

[1] https://cloud.google.com/vision/

**Concept Filtering.** We leverage a number of state-of-the-art tools to depict an image in a set of candidate visual concepts. However, false positive concepts generated by those tools result in redundancy and noise. For example, an image relevant to interior like "bedroom" or "living room" would be supported by the visual concepts such as "couch", "bed", and "table". By contrast, the concept "church" would have lower similarities to those terms. Based on this idea, we prune the set of candidate visual concepts by removing the concepts that are less supported by other concepts, and produce a set of visual concepts for each image. We compute the semantic similarity between candidate visual concepts by using pre-trained word embeddings and construct a similarity matrix, which represents the similarity of each concept pair. We discard those concepts that would accumulate low correlations with other concepts. The procedure is illustrated in Fig. 3 with a real example.

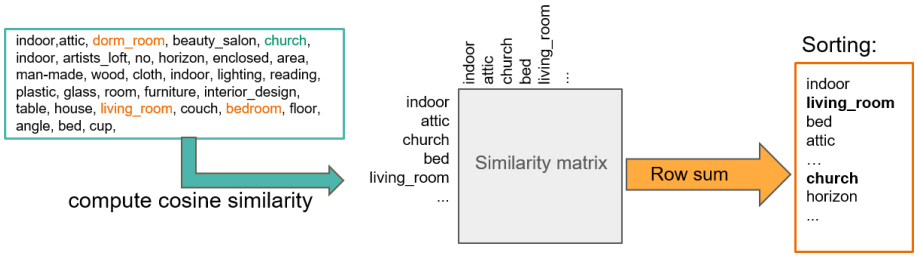



**Fig. 3.** Illustration of concept filtering.

### 3.2 Indexing

Our framework fetches images based on a given query. In ADLT, we require users to specify concepts that highly related to the given topics and keep off confusing terms as a query term set for each topic. Moreover, time span according to the topics would be considered. To ensure the quality and usability of the retrieval system in practice, preprocessing on textual data and the retrieval algorithms are other crucial issues.

**Metadata Preprocessing.** For generalization of the retrieval task, tags available in metadata like "Home" and "Work" in the location field and "transport" and "walking" in the activity field are extracted as attributes of images, instead of using all the location information. To proceed further with locational information, we calculate the average moving speed according to GPS coordinates to infer the type of transportation (e.g. car, airplane). For any pair of points $\{(lat1, lon1), (lat2, lon2)\}$ on the geographical coordinate system, the distance $d$ is given by the great-circle distance formula as follows.

$$d = \cos^{-1}(\sin(lat1)\sin(lat2) + \cos(lat1)\cos(lat2)\cos(|lon2 - lon1|)) \times 6371km \quad (1)$$

and the average speed is calculated according to the distance $d$ and the difference between timestamps.

**Retrieval Model.** The similarity between the user query and an individual image, represented as a set of visual concepts, is measured with three schemes as follows.

*Exact Matching.* Given a list of concepts in the user query that are combined with logical operators AND, OR, and NOT, the visual concepts of the image should meet the condition. This approach returns accurate results if the topic is explicit, e.g., watching TV or use cell phone in the car.
*BM25 Model.* Exact matching suffers from low recall rate. Here, we perform the partial matching by using a classic retrieval model, BM25 [18]. The BM25 scheme measures the relatedness between two sets of concepts based on term frequency (TF) and inverse document frequency (IDF). In this way, the specific concepts are more likely to be extracted. For example, the concept "grocery", which provides specific information than the general concept "indoor" does, has a higher BM25 score due to its higher inverse document frequency.
*Word Embedding.* Word embeddings (distributed word representations) have been widely used in text similarity measurement. For fuzzy matching, we adopt the word embeddings to measure the semantic relatedness between the concepts in the query and the concepts extracted from an image. The information of semantic relatedness would be helpful when similar but not identical concepts are present in both sides. We first obtain distributed representations of concepts with word embeddings that are pre-trained on a large-scale corpus, and aggregate concept-level semantics by taking the element-wise mean for each query/image. In this way, the relatedness between the query and an image can be computed by using the cosine similarity. Note that by using the pre-trained word embeddings, external knowledge is inherent in the retrieval model.

### 3.3 Image Stream Segmentation

**Deeply Learned Features.** Pre-trained convolutional neural networks have been shown to be beneficial for various computer vision tasks as generic image feature extractors. In this sense, we apply pre-trained CNN tools [5] to extract dense feature vectors, and estimate the level of change between consecutive images by measuring the dissimilarity of their representations using Euclidean distance and cosine similarity. Another neural network-based approach for this purpose is to compress the image by autoencoders. These methods are tempting that feature extraction can be done automatically, and the obtained feature can be integrated with other features simply.

In this work, we obtain the dense vector for each image with a pre-trained DenseNet and with a deep autoencoder trained on the provided images. For each pair of two consecutive images, a threshold of dissimilarity is heuristically tuned to determine whether the two images belong to the same event cluster or not. Besides, smoothing methods such as moving average should also be adopted to prevent the consecutive boundary occur in a short time period.
**Event Interval.** In addition to the features learned by models, human knowledge is also involved. Due to the property of lifelog data and given topics, human beings can easily figure out how long it takes for a daily activity. Therefore, our other approach is grouping the images by reasonable interval of shooting times. That is, two consecutively retrieved images would treat as a single event if the difference of their timestamps is smaller than a threshold according to the event topic. The thresholds for each topic are

intuitive defined by human. For example, the reasonable event interval would be about 60 minutes for having a lunch.

### 3.4 Daily Activities Summarization

For ADLT, the system output should be two real values indicating the frequency and total duration of the given topic, respectively. The frequency could be calculated by summing up the number of segmented retrieved events. The duration is obtained by summing up the time differences between the first frame and last frame of each event.

### 3.5 Life Moment Retrieval

For LMRT, the model is designed to retrieve a number of images indicating specific life moments with respect to the target query. There are total 10 topics, each consists of a title, a short description, and a longer narrative in detail. For query processing, we extract key terms in the title and in the description only, since the intention in the narrative is often too complicated to extract without human assistance. This can be done by simply removing function words in each query. Note that further processing such as stemming is unnecessary due to the use of word embeddings. The resulting concepts can be further improved by human, according to the narrative or human's general knowledge. Finally, the query is transformed into vector representation and compared with pre-computed vector representation of each image.

## 4 Experiments

For the two subtasks, automatic runs are submitted. For each of the subtasks, we first describe the parameter settings, and show the experimental results. We only report the execution time of our system for testing since we exploit pre-trained tools.

### 4.1 Activities of Daily Living Understanding Task

In the first trial of ADLT, we use the query consisting of the concepts automatically parsed from the given topics. We seize nouns and gerunds as concepts and list them for query. For the time condition, information is directly provided with the  tag. However, some of the required concepts like *socializing* or *having party* seldom appear in the visual concept sets since most of CV tools are trained with shallow semantic image descriptions. To deal with this issue, the topics consisting of abstract ideas are refined by human.

As mentioned in previous sections, visual concepts of each image labeled by CV tools are far from perfect. For this reason, visual concept filtering is applied to discard the concepts with a low relevance to other concepts in the same image. After we obtain a list of visual concepts sorted by the row sum of similarity matrix for each image, the top half of the concepts are retained. To our observation, this is a flexible threshold. Besides, two kinds of pre-trained word embeddings are brought in here, including

GloVe [9] trained on Common Crawl with 840B tokens and ConceptNet Numberbatch [8]. The comparison in percentage dissimilarity [1] is shown in Table 1, where (G) and (N) denote GloVe and ConceptNet Numberbatch word vectors, respectively.

**Table 1.** Performances in ADLT.

| Pipeline | Percentage Dissimilarity |
|---|---|
| vanilla retrieval | 0.2434 |
| fine-tuned query | 0.2850 |
| fine-tuned query + concept selection (N) | 0.3850 |
| fine-tuned query + concept selection (G) + coffee capturing | 0.4592 |
| fine-tuned query + concept selection (N) + coffee capturing | **0.4787** (Rank 2) |

For summarization, two types of image stream segmentation are tried on the development set. We find out that the segmentation based on deeply learned features results unstable boundaries and comes out a limited performance. Basically, it is hard to determine the threshold of dissimilarity between two images. On the other hand, using human defined reasonable interval for each event type, deemed an intuitive parameter for daily activities, provides more sensible boundaries. As a result, we apply the latter method on the test set.

In general, our framework achieves a percentage dissimilarity of 0.3850 with human refined query and visual concept filtering reported by the online evaluation platform. The inference time for each topic is 5.59s on average. Error analysis shows that our system fetches a lot of unnecessary images for topic 1 because the CV tools cannot identify “coffee cup” in the lifelog data. For this reason, we refine the query by using only “cup”, instead of “coffee cup”, to ensure the recall rate. It turns out that images with bottles or mug are retrieved (under the *Office* scenario).

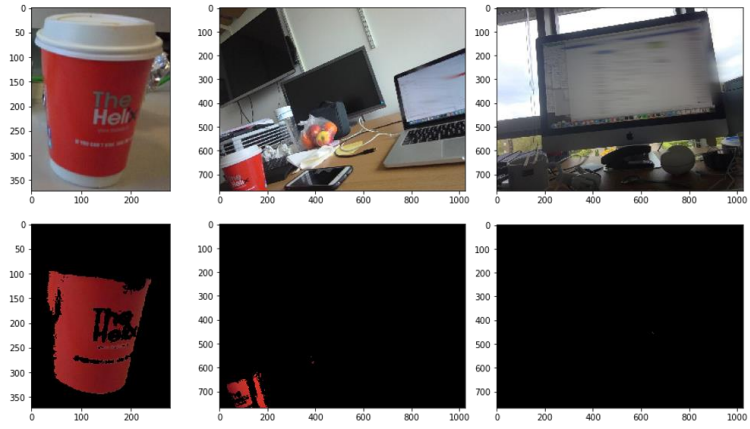
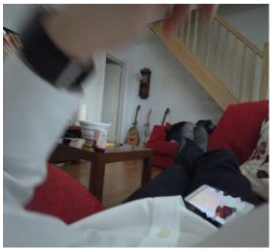


**Fig. 4.** Examples of coffee capturing based on RGB constraints

To enhance the precision of topic 1, an ad-hoc method is further introduced to filter out surplus retrieved items. First, we observe that the coffees might be bought by the

lifelogger from the same shop within the same red cup. So we specify the upper and the lower bounds of RGB value to capture red objects in a given shoot. For the consideration of spatial verification, we preserve images containing red area larger than a threshold. The procedures of the above operations are shown in Fig. 4. By distinguishing the red coffee cup from other types of cup, the results of topic 1 are greatly enhanced, and the overall performance is also increased as shown in Table 1.

The best result among our all submissions, i.e., percentage dissimilarity of 0.4787, is achieved by the combination of fine-tuned query, concept selection mechanism with ConceptNet Numberbatch word embedding and coffee capturing trick. The ad-hoc image filtering for topic 1 produces a significant improvement. For the construction of the similarity matrix used in visual concept filtering, the ConceptNet Numberbatch embeddings, built using an ensemble that combines data from ConceptNet [8], word2vec [10], GloVe [9], and OpenSubtitles2016 [19], provide better results comparing to GloVe. The effectiveness of concept filtering is shown in Fig. 5.



Original concepts:
[attic, dorm, beauty, salon, church, indoor, artists, loft, horizon, enclosed, area, man-made, wood, cloth, lighting, reading, plastic, glass, furniture, interior, design, table, house, living, room, couch, bedroom, floor, angle, bed, cup]

Refined concepts (sorted):
[lighting, enclosed, couch, room, floor, furniture, bedroom, house, bed, table, loft, wood, interior, living, area]

**Fig. 5.** Example of concept filtering on image 20160830_181833_000.

### 4.2 Life Moment Retrieval Task

For LMRT, we submit a total of four successful runs, which are either fully automatic or semi-automatic with human fine-tuned queries. In this subtask, we use a subset of CV tools based on a preliminary evaluation on the development set. The relatedness between images and queries are measured with the word embedding method as described in Section 3.2. We employ the 300-dimenstional word embeddings pre-trained on 16B tokens with fastText [11]. Out-of-vocabulary words are ignored. The first run is using automatically extracted query terms to match image concepts. We consider this as the baseline method.

In the rest of runs, we add time and location constraints to each query and re-rank the retrieved images using provided metadata. This information can be either inferred automatically by NLP APIs[2] in Run 3, or given by human in Run 4 and Run 5. We use a heuristic method for re-ranking that increases the score by a weight $w_l$ if the location constraints are satisfied, and decreases the score by a weight $w_t$ if the time constraints are not satisfied. The weight $w_l$ is set as 0.8, and $w_t$ is 0.1 in the experiments.

[2] https://cloud.google.com/natural-language/

In Run 4, we fine-tune the queries to get better results. Each query is rewritten into the form consisting of the following fields: positively related concepts, negatively related concepts, location constraint, and time constraint. For negatively related concepts, we assign a weight of -1 to their word representations before vector aggregation. Table 2 shows an example that the automatic one is extracted from the title and the description. The fine-tuned one is obtained by removing relatively useless query terms and manually adding additional concepts, according to narrative and retrieved images. Once the query is modified, the system retrieves new results, and the query can be further improved based on new results. In the last run, we perform clustering on the retrieved images based on the event intervals described in Section 3.3.

**Table 2.** Query Formulation for LMRT.

| | |
|---|---|
| ● Title: Interviewed by a TV presenter<br>● Description: Find all the moments when I was interviewed by TV presenter.<br>● Narrative: The moment must show the cameras or cameramen in front of the lifelogger. The interviews can occur at the lifelogger's home or in the office environment. | |
| Automatic | positive = {*interviewed, TV, presenter*}; negative = {};<br>location = {}; time = {} |
| Fine-tuned | positive = {*camera, person*}; negative = {};<br>location = {*home, work*}; time = {} |

Table 3 and Fig. 6 show the performance of our method for LMRT. The inference time (without clustering) for each topic is 2.46s on average. The best result on the test set is achieved by the combination of the fine-tuned query with time/location constraints, showing that it is crucial to have human to give more precise query expressions. We also notice that there is no improvement with clustering. A possible reason is that some queries in the test set ask for rare events such as assembling furniture. Under the scoring metric, cluster recall at 10 (CR@10) [1], used in this subtask, there is no benefit to cluster events that occur less than ten times. Instead, inaccurate results may be introduced and decrease the score.

For LMRT, we employ pre-trained fastText for the official runs. As word embedding playing a crucial role in this task, we try a number of off-the-shelf word embeddings for similarity computation. Details of these embeddings can be found in [8,9,10,11, 12,13]. These models are designed in quite different ways, but all produce semantic representations for individual words in a latent space. Fig. 7 compares different word embeddings. We report F1@10 scores per query in the development set with a fully automated approach. The results suggest that it would be beneficial to use word embeddings that associate with additional contextual information such as syntactic dependency or lexical ontology.

Another advantage of adopting word embeddings is that sometimes query words are absent while relative or similar concepts are present in the desired images. With word embeddings, we have an opportunity to capture this kind of relations. For example, it is possible to match an image with the concepts {*bowl, food, knife, meal, person, salad,*

*table*} to the query {*salad*, *kitchen*} with a high similarity, though the keyword "*kitchen*" is absent from the set of visual concepts.

**Table 3.** Performances in LMRT.

| Run ID | Description | F1@10 |
|---|---|---|
| Run 1 | automatic query (baseline) | 0.177 |
| Run 3 | automatic query + constraints | 0.223 |
| Run 4 | fine-tuned query + fine-tuned constraints | **0.395** (Rank 4) |
| Run 5 | fine-tuned query + fine-tuned constraints + clustering | 0.354 |

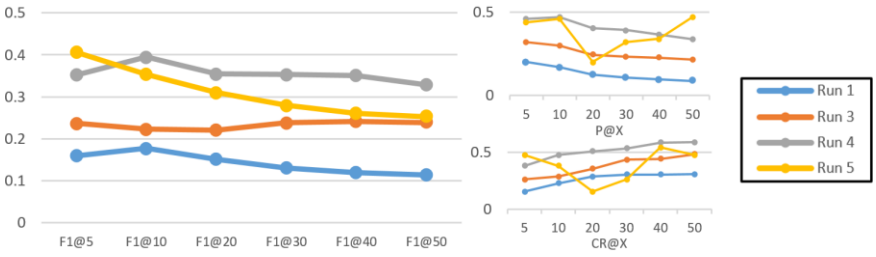



**Fig. 6.** Official results on the test set.

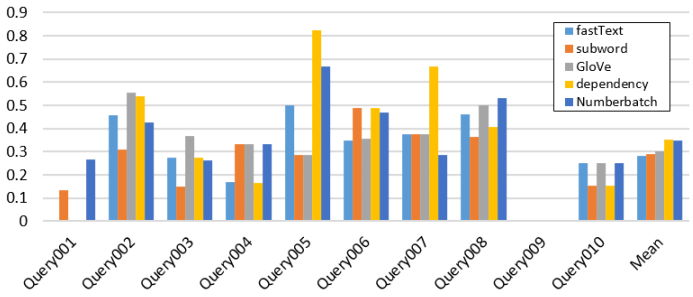



**Fig. 7.** F1@10 Scores on the development set with different embeddings used.

## 5 Conclusion

This paper presents our approaches to daily activities summarization and moment retrieval for lifelogging. In both subtasks, we introduce the external textual knowledge to reduce the semantic gap between the user query and the visual concepts extracted by the latest CV tools. Experimental results show the ensemble distributed word model, ConceptNet Numberbatch, provides effective word embeddings in both two subtasks.

Experimental results also suggest that better performances can be achieved by using fine-tuned queries. That means there still exists a room for improvement on bridging the gap between the abstract human intentions and the concrete visual concepts.

## References


1. Dang-Nguyen, D.T., Piras, L., Riegler, M, Zhou, L., Lux, M., and Gurrin, C.: Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval. In: CLEF2018 Working Notes (CEUR Workshop Proceedings).
2. Ionescu, B., Müller, H., Villegas, M., Herrera, A.G.S., Eickhoff, C., Andrearczyk, V., Cid, Y. D., Liauchuk, V., Kovalev, V., Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Dang-Nguyen, D.-T., Piras, L., Riegler, M., Zhou, L., Lux, M., and Gurrin, C.: Overview of ImageCLEF 2018: Challenges, datasets and evaluation. In proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018), Avignon, France. LNCS, Springer. (September 10-14 2018).
3. Zhou, L., Piras, L., Riegler, M., Boato, G., Dang-Nguyen, D.T., Gurrin, C.: Organizer Team at ImageCLEFlifelog 2017: Baseline Approaches for Lifelog Retrieval and Summarization. In proceedings of CLEF (2017).
4. Garcia del Molino, A., Mandal, B., Lin, J., Hwee Lim, J., Subbaraju, V., Chandrasekhar, V.: VC-I2R@ImageCLEF2017: Ensemble of Deep Learned Features for Lifelog Video Summarization. In proceedings of CLEF (2017).
5. Huang, G., Liu, Z., Maaten, L., Weinberger, K.: Densely Connected Convolutional Networks. In proceedings of CVPR (2017).
6. Redmon, J., Farhadi, A.: Yolo9000: Better, faster, stronger. In: arXiv:1612.08242 (2016)
7. Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., Murphy, K.: Speed/accuracy trade-offs for modern convolutional object detectors. In proceedings of CVPR (2017).
8. Speer, R., Chin, J, Havasi, C.: ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In proceedings of AAAI (2017).
9. Pennington, J., Socher, R., and Manning, C.: GloVe: Global Vectors for Word Representation. In proceedings of EMNLP (2014).
10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In proceedings of NIPS (2013).
11. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification, In proceedings arXiv:1607.04606 (2016).
12. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., Enriching Word Vectors with Subword Information, In arXiv:1607.04606 (2016).
13. Levy, O., Goldberg, Y.: Dependency-Based Word Embeddings. In proceedings of ACL (2014).
14. J., Deng, W., Dong, R., Socher, L.J., Li, K., Li, L., Fei-Fei: ImageNet: A LargeScale Hierarchical Image Database. In proceedings of CVPR (2009).
15. Bolei, Z., Agata, L., Aditya, K., Aude, O., Antonio, T.: Places: A 10 Million Image Database for Scene Recognition. In proceedings of TPAMI (2017).
16. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Common objects in context. In proceedings of ECCV (2014).
17. Krasin I., Duerig T., Alldrin N., Ferrari V., Abu-El-Haija S., Kuznetsova A., Rom H., Uijlings J., Popov S., Kamali S., Malloci M., Pont-Tuset J., Veit A., Belongie S., Gomes V., Gupta A., Sun C., Chechik G., Cai D., Feng Z., Narayanan D., Murphy K. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. (2017).
18. Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. In: Foundations and Trends in Information Retrieval archive, Vol 3 Issue 4. ACM (2009).
19. Lison, P., Tiedemann, J.: OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In proceedings of LREC (2016).