NTCIR 17 Conference: Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies, December 12-15, 2023, Tokyo, Japan
DOI: https://doi.org/10.20736/0002001337

DCU at the NTCIR17 Lifelog-5 Task

Naushad Alam
SFI Insight Centre for Data Analytics
Dublin City University
Ireland
naushad.alam2@mail.dcu.ie

Yvette Graham
School of Computer Science and Statistics
Trinity College
Ireland
ygraham@tcd.ie

Cathal Gurrin
School Of Computing
Dublin City University
Ireland
cathal.gurrin@dcu.ie

ABSTRACT

In this work, we present our system DCUMemento as part of our team 'DCU' to participate in the NTCIR-17 Lifelog-5 task. Our system leverages a suite of CLIP models developed by OpenAI as well as larger models from the OpenCLIP model suite which are trained on ~5x more data as compared to the OpenAI models. We also discuss the query data structure for the task as well as the models/ensemble approaches used in our system. Finally, we present the results from our submitted runs providing a comparative analysis of the approaches as well as discuss future work in this direction.

KEYWORDS

lifelog, information retrieval, quantified self, personal data

TEAM NAME

DCU

SUBTASKS

Lifelog Semantic Access Task (LSAT) - Automatic

1 INTRODUCTION

Lifelogs are longitudinal multimodal archives of continuous personal data recorded passively using wearable cameras and sensor devices such as FitBit, sleep trackers, mood trackers etc. The broader goal of recording oneself is to capture one's personal life experiences in totality on a daily basis which can later be utilised for address use cases like event retrieval or memory reminiscence.

Lifelogging has been made more accessible and comprehensive in the digital age due to the availability of various apps, wearable devices, and cheaper cloud storage solutions. However, passive form of lifelogging is still a very data intensive task and the size of data collected could be in the ballpark of 1 TB or more per individual annually [2]. With such a scale of data generation, the need to have technologies which can index, organise and search the archives efficiently is paramount.

Retrieving relevant information from lifelogs is a significantly challenging task given the nature of the dataset which not only consists of egocentric images but also textual metadata and other data modalities such as sensor data. Efficient information retrieval from lifelogs broadly entails two challenges, the first is comprehending the information from the egocentric images which is difficult as passive in-the-wild captured images do not always convey the full-context due to continously shifting point of view (POV) and also results in blurry or occluded images. Off-the-shelf large pre-trained vision models do not generalise very well on such a noisy dataset as they have been mostly pre-trained on third-person POV images which are abundant on the internet unlike first-person images. The second challenge here is to comprehend all the data modalities together to effectively retrieve relevant information from the dataset. For example, a search query *"Sunday morning in Ireland"* requires the search engine to retrieve the *'morning'* images from the dataset while retrieving the other required pieces of information such as day and location from the metadata and sensor data respectively.

Several challenges such as NTCIR-Lifelog task [11, 12], ImageCLEFlifelog [7] and Lifelog Search Challenge [4] [3] have been organized in recent few years aiming to advance the state of the art in multimodal information retrieval. The NTCIR-Lifelog task is a core task of the NTCIR-17 Conference[1] which includes a single subtask i.e. the Lifelog Semantic Access Task (LSAT) running both Automatic and Interactive modes as discussed in Section 2 and in more detail in [11].

In this paper, we present our system called DCUMemento which participated in the NTCIR-17 Lifelog-5 Automatic LSAT subtask. Our system used image-text representations derived from the CLIP models [9] developed by OpenAI as well as from the larger OpenCLIP models [6] to perform search and ranking of the images. The rest of the paper is structured as follows: Section 2 discusses the Lifelog Semantic Access Task in more detail while Section 3 briefly discusses the dataset associated with the task. Then, in Section 4, we discuss the DCUMemento system in detail, covering the core aspects of the system such as query reformulation, search engine and evaluation results.Finally in Section 5, we conclude our paper and discuss future work in the development of our system.

2 LIFELOG SEMANTIC ACCESS TASK

The Lifelog Semantic Access Task (LSAT) is an item search task that can be undertaken in an interactive or automatic manner, where the participants are required to retrieve a number of specific moments from the lifelogger's life [11].

- **Automatic LSAT**: The automatic run is intended to operate independently of any user involvement during the search process beyond specification of the initial query which can happen once for each topic at the start of the search.The process is not time-bound and once finished should return a ranked list of 100 images for each of the topic.
- **Interactive LSAT**: The interactive run allows user involvement during the search process with single or multiple phases of query reformulation or relevance feedback until the user is satisfied with the results. While interactive running also expects a ranked list of 100 images for each topic, automatic

[1] http://research.nii.ac.jp/ntcir/ntcir-16/conference.html

running is time-bound allowing a maximum of 300 seconds for each topic.

The subtask contains 41 topics out of which 17 are recall focused requiring as many relevant items as possible to be present in the ranked list of output images. For example, *"Find examples of when I was eating avocado for breakfast"* asks to rank all moments when the person was eating avocado during breakfast hours. On the other hand, the other 24 topics are precision focused with only 1 or a small number of relevant items in the collection e.g *"Find the moments when the lifelogger made their second visit to the house with the stone shed or hovel"* looking for a specific moment from that person's life.

3 DATASET

The NTCIR-17 Lifelog5 [11] task uses the same dataset from the Lifelog Search Challenge 2023 [5] which consists of:

- **Egocentric Images**: Images from a wearable camera which is fully redacted and anonymised captured using a Narrative Clip device collected during 2019-2020. All faces and readable text have been removed, as well as certain scenes and activities manually filtered out to respect local privacy requirements.
- **Metadata**: The metadata file consists of general user information like location, activity, elevation, etc. as well as biometric information like calories burnt, heart rate, step count, etc. captured using a wearable device.
- **Visual Concepts**: The visual concepts file contains scene descriptions, object tags with their confidence scores, object bounding boxes, etc., for each image in the dataset.

4 SYSTEM OVERVIEW

In this section, we present an overview of the system outlining the search and ranking approaches as well as the query structuring and processing aspects of the system.

4.1 Query Structure and Processing

Similar to our previous work [1], we reformulate the queries from the NTCIR-17 Lifelog5 task as well in a dictionary format allowing us to process it in an automatic manner as is the requirement of the LSAT task. Basically, we separate out the visual hints mentioned in the query from the other hints concerning with metadata such day, time, location etc. This allows for a stage-wise search process where initially, the search and ranking of the images is done solely based on visual hint. In the subsequent stage, other specific pieces of information are used to apply relevant filters over the ranked results from the previous step and hence obtain the final result. For example, we reformulated the query, *"Find the times I was getting a taxi to the airport before 6am."* to separate the visual description from other specific details like day and time,

- **Search query**: *getting a taxi*
- **Filters**:
  - **Hour**: *<6*

4.2 Search and Ranking Methods

Our search and ranking methodology uses image-text embeddings from the CLIP models [9] developed by OpenAI and as well models from OpenCLIP [6] which is an open source implementation of the OpenAI models and are trained on ~5x more data. These models are trained in a contrastive fashion on image-text pairs with goal of maximising the similarity scores of the correct image-caption pairs while minimising it for the incorrect pairs. The OpenAI models have been trained on 400 million image-text pairs while the OpenCLIP models [6] are trained on a subset (2 billion images) of LAION-5B [10] which is a corpus of 5 billion image-text pairs. The large-scale pre-training of these models makes them capable of zero-shot transfer to out-of-domain datasets like Lifelogs which is a first-person POV dataset with fairly reasonable accuracy.

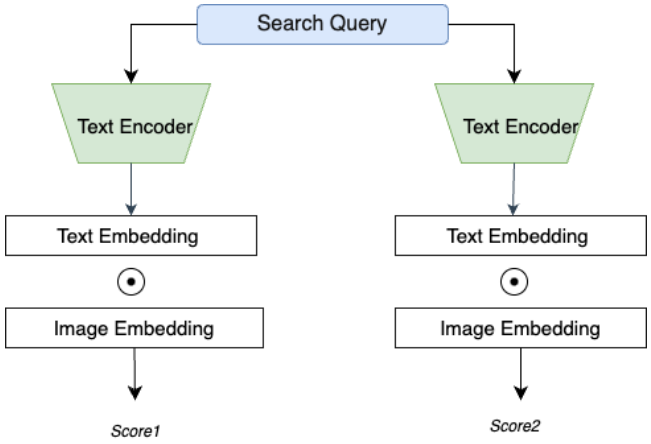

Figure 1: High-level architecture of the weighted ensemble approach where initially the search query sent as input to the respective text encoders of Model 1 (left) and Model 2 (right). The cosine similarity scores are derived by taking the dot product of text embedding with the image embedding (Score1 and Score2). A final ranking score is then derived by combining the two scores in 3:1 ratio *(0.75 Score1 + 0.25 Score 2)*.

The search and ranking in done in stages with the following execution steps:

- **Search using Visual Description**: Initially, the system tries to rank the images based only on the visual information available to it given the CLIP models are agnostic of any information besides visual descriptions.
- **Apply Relevant Filters**: The system then applies all filters (if mentioned) to generate a final ranked list of images maximum upto 100 images.

We submitted 9 runs in total for the Lifelog Semantic Access Task (LSAT) out of which 5 runs leveraged embeddings derived from the various CLIP models (OpenAI and OpenCLIP) independently while 4 runs were based on the weighted ensemble of cosine similarity scores from the models. Figure 1 shows a high-level architecture of the weighted ensemble approach. We discuss the details of the runs and the models used in section 4.3.

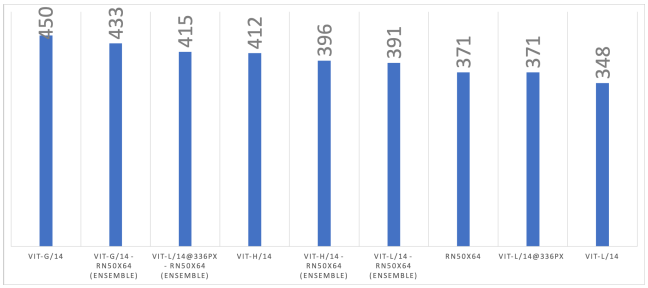

Figure 2: Total number of retrieved relevant images. Models are ordered left to right based on decreasing number of relevant images retrieved.

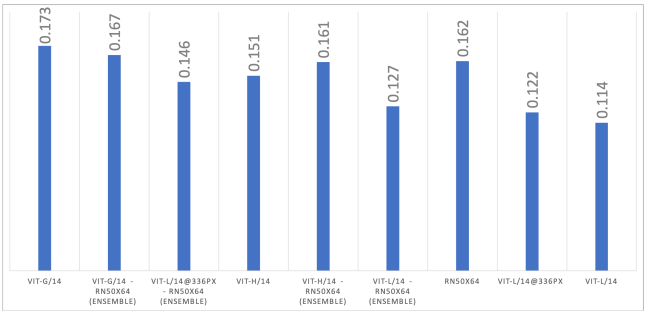

Figure 3: Mean Average Precision of the models. Models are ordered left to right based on decreasing number of relevant images retrieved.

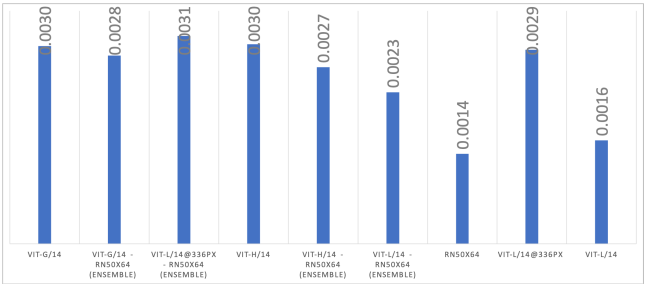

Figure 4: Geometric Mean Average Precision of the models. Models are ordered left to right based on decreasing number of relevant images retrieved.

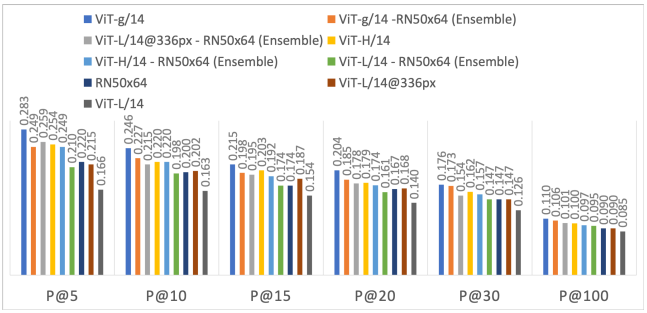

Figure 5: Precision@K. For each K, models are ordered left to right based on decreasing number of relevant images retrieved.

4.3 Evaluation Results

TRECEVAL [8] was used to evaluate runs for the subtask (LSAT) based on the comma separated files submitted by the participating teams. We submitted 9 runs in total for the LSAT task. The details of a the approaches/models are discussed below:

(1) **ViT-L/14 (OpenAI):** A large Vision Transformer model released by OpenAI. It generates 768-dimensional image-text embeddings.
(2) **ResNet50x64 (OpenAI):** OpenAI ResNet-50 model using 64x the compute of a ResNet-50. It generates 1024-dimensional image-text embeddings.
(3) **ViT-L/14@336px (OpenAI):** ViT-L/14 model pre-trained at a higher 336 pixel resolution for one additional epoch to boost performance.
(4) **ViT-g/14 (OpenCLIP):** Trained on ~5x more data and has ~3x parameters as compared to the ViT-L/14 model. It generates 1024-dimensional image-text embeddings.
(5) **ViT-H/14 (OpenCLIP):** Trained on ~5x more data and has ~2x parameters as compared tothe ViT-L/14 model. It generates 1024-dimensional image-text embeddings.
(6) **ViT-L/14 - RN50x64 (Ensemble):** Weighted sum of cosine similarity scores from ViT-L/14 and ResNet50x64 in a 3:1 ratio.
(7) **ViT-g/14 - RN50x64 (Ensemble):** Weighted sum of cosine similarity scores from ViT-g/14 and ResNet50x64 in a 3:1 ratio.
(8) **ViT-H/14 - RN50x64 (Ensemble):** Weighted sum of cosine similarity scores from ViT-H/14 and ResNet50x64 in a 3:1 ratio.

Figure 2 shows the total number of relevant items retrieved by each model. The ViT-g/14 model retrieves the highest number of relevant images overall with ViT-g/14-RN50x64 (Ensemble) model being at the close second position. The total of relevant items in the corpus were **7494** out of which **450 (6%)** could be retrieved by the best performing system.

The Mean Average Precision (MAP) metric (Figure 3) too mimicks the trend with ViT-g/14 leading the chart followed closely by ViT-g/14-RN50x64 (Ensemble). The Geometric Mean Average Precision (GMAP) metric (Figure 4) however does not follow this trend in its entirety where models with relatively low MAP such as ViT-H/14 and ViT-L/14@336px fare higher in terms of GMAP. Figure 5 shows average precision at multiple K values.

In terms of Mean Reciprocal Rank (Figure 6), models ViT-g/14 and ViT-g/14-ResNet50x64 (Ensemble) occupy the 1st and 2nd position respectively with with MRR values of **0.43** and **0.41**. This signifies that on an average the models were able to rank the first relevant image either at the 2nd or 3rd position. Figure 7 shows the Normalized discounted cumulative gain at multiple K values for

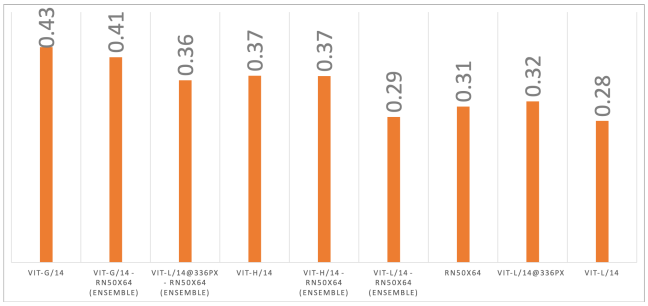

Figure 6: Mean Reciprocal Rank of the models. Models are ordered left to right based on decreasing number of relevant images retrieved.

the models which is a metric focused on the relevance and ranking quality of the search results. The ViT-g/14 model significantly outperforms other models on this metric as well.

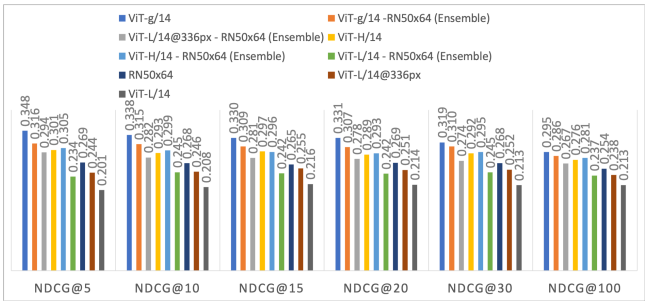

Figure 7: Normalised Discounted Cumulative Gain@K. For each k, models are ordered left to right based on decreasing number of relevant images retrieved.

Overall, the ViT-g/14 model shows the best overall performance beating other models on multiple evaluation metrics.

5 CONCLUSION AND FUTURE WORK

In this paper, we described our system DCUMemento which has earlier participated in the previous editions of the Lifelog Search Challenge [3–5] as well the NTCIR-16 Lifelog4 Task [12]. In this work, we experimented with multiple CLIP models where the ViT-g/14 was the best performing model overall in terms of precision and relevance ranking of images. We adopted a score based cutoff to decide the final number of images in the output for a given query as opposed to imposing a hard cutoff. In the future, it would be interesting to investigate if the cutoff score is model agnostic and whether the cutoff score can be assigned to each of the models in an optimal way separately.

ACKNOWLEDGMENTS

This publication has emanated from research supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289/P2, co-funded by the European Regional Development Fund.

REFERENCES

[1] Naushad Alam, Ahmed Alateeq, Yvette Graham, Mark Roantree, and Cathal Gurrin. 2022. DCU at the NTCIR16 Lifelog-4 Task. (2022), 5.

[2] Cathal Gurrin. 2021. Personal Data Matters: New Opportunities from Lifelogs. In *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. 1–3. https://doi.org/10.1109/iSAI-NLP54397.2021.9678155

[3] Cathal Gurrin, Klaus Schoeffmann, Bjorn Thor Jonsson, Duc Tien Dang Nguyen, Jakub Lokoc, Luca Rossetto, Minh-Triet Tran, Wolfgang Hurst, and Graham Healy. 2021. An Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *ICMR '21, The 2021 International Conference on Multimedia Retrieval*. ACM, Taipei, Taiwan, 690–691.

[4] Cathal Gurrin, Liting Zhou, Graham Healy, Bjorn Thor Jonsson, Duc Tien Dang Nguyen, Jakub Lokoc, Minh-Triet Tran, Wolfgang Hurst, Luca Rossetto, and Klaus Schoeffmann. 2022. An Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *ICMR '22, The 2022 International Conference on Multimedia Retrieval*. ACM, Newark, NJ, USA.

[5] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proc. International Conference on Multimedia Retrieval (ICMR'23)* (Thessaloniki, Greece) *(ICMR '23)*. Association for Computing Machinery, New York, NY, USA.

[6] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773 If you use this software, please cite it as below..

[7] Van-Tu Ninh, Tu-Khiem Le, Liting Zhou, Luca Piras, Michael Riegler, Pål Halvorsen, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, and Duc-Tien Dang-Nguyen. 2020. Overview of ImageCLEF Lifelog 2020:Lifelog Moment Retrieval and Sport Performance Lifelog. In *CLEF2020 Working Notes (CEUR Workshop Proceedings)*. CEUR-WS.org <http://ceur-ws.org>, Thessaloniki, Greece.

[8] NIST. 2021. *The Text REtrieval Conference (TREC)*. Retrieved April 25, 2022 from https://trec.nist.gov/

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]* (Feb. 2021). http://arxiv.org/abs/2103.00020 arXiv: 2103.00020.

[10] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs.CV]

[11] Liting Zhou, Cathal Gurrin, Duc-Tien Dang-Nguyen, Graham Healy, Chenyang Lyu, Tianbo Ji, Longyue Wang, Joho Hideo, Ly-Duyen Tran, and Naushad Alam. 2023. Overview of the NTCIR-17 Lifelog-5 Task. In *Proceedings of the 17th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-17)*. Tokyo, Japan.

[12] Liting Zhou, Cathal Gurrin, Graham Healy, Hideo Joho, Thanh-Binh Nguyen, Rami Albatal, and Frank Hopfgartner. 2022. Overview of the NTCIR-16 Lifelog-4 Task. In *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-16)*. Tokyo, Japan.