Memento 3.0: An Enhanced Lifelog Search Engine for LSC'23

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

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

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

ABSTRACT

In this work, we present our system Memento 3.0 for participation in the Lifelog Search Challenge 2023, which is a successor to the previous 2 iterations of our system called Memento 1.0 [1] and Memento 2.0 [2]. Memento 3.0 employs image-text embeddings derived from OpenAI CLIP models as well as larger OpenCLIP models trained on ~5x more data. Our system also significantly reduces the query processing time by almost 75% when compared to its predecessor systems by employing a cluster-based search technique. We additionally make important updates to the system's user interface to offer more flexibility to the user and at the same time be better suited to efficiently handle new query types introduced in the Lifelog Search Challenge.

CCS CONCEPTS

• **Information systems → Retrieval models and ranking**; **Search interfaces**.

KEYWORDS

lifelog, information retrieval, semantic image representation

ACM Reference Format:
Naushad Alam, Yvette Graham, and Cathal Gurrin. 2023. Memento 3.0: An Enhanced Lifelog Search Engine for LSC'23. In *6th Annual ACM Lifelog Search Challenge (LSC '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3592573.3593103

1 INTRODUCTION

Lifelogging is the process of digitally capturing and storing information about one's daily activities, experiences, and surroundings using various types of sensors and devices. The collected dataset called lifelogs is a rich in-the-wild multimodal dataset recorded using wearable devices such as cameras and fitness bands. The goal of lifelogging is to create a comprehensive and detailed record of one's life which can be used to support use cases, such as memory augmentation and retrieval, health analytics, activity detection, and developing customized applications to support elderly people in leading an independent life.

In recent times, lifelog data has been used to address several of these use cases in research domains such as memory augmentation and reminiscence [4, 7, 11], human activity recognition [6], health monitoring for elderly and people suffering from chronic diseases [21, 22]. The research on information retrieval from lifelogs has also significantly progressed in recent few years aided by benchmarking challenges such as the Lifelog search challenge [12–14].

The Lifelog search challenge since its inception in 2018 has garnered significant attention from across the globe and has become an important event for researchers and the lifelogging community. The challenge provides an opportunity to showcase the latest advances in lifelog retrieval and facilitates collaboration and knowledge sharing among researchers and practitioners. The Lifelog Search Challenge 2022 [12] where 9 global teams competed against each other, introduced 2 new query formats (Ad-hoc and QA task) in addition to the original known item search task which makes the competition even more exciting and challenging. The newer format of LSC poses a very unique information retrieval challenge where only a semantic understanding of an individual image would not be beneficial to address all the 3 query types. For example, with QA tasks a more fine-grained contextualized understanding of the data would be required to answer questions such as *what did I do after doing X on a specific day?* or *how many times did I do activity Y last month?*.

In this paper, we present our system Memento 3.0 to participate in the upcoming edition of Lifelog Search Challenge [14]. Our proposed system builds on top of its predecessor to further improve the search and ranking functionality in terms of a better semantic understanding of the data by using enhanced image-text embeddings. The system is also significantly faster in terms of search speed by using a cluster-based approximate search methodology over the image embedding space. Memento 3.0 further offers more flexibility to the user during the search process where the user can dynamically toggle between the backend models from the primary search interface as per requirements. The user interface of the system has further been modified to accommodate the newer query types in the challenge while borrowing the functionalities such as visual faceted filtering and image starring functionality as is from its predecessor systems.

2 RELATED WORK

The Lifelog search challenge which is in its 6th year, has driven the advancement in the state-of-the-art in lifelog information retrieval. Over the years since 2018, dozens of systems have participated in the competition proposing exciting and novel solutions to address the problem.

Overall 9 teams participated in the lifelog search challenge 2022. E-MySceal [31] won the last year's challenge and used image-text embeddings from the CLIP model [27] to perform the search as opposed to the concept-based search methodology they employed in the previous iterations of their system [32]. LifeSeeker 4.0 [25]

LSC '23, June 12–15, 2023, Thessaloniki, Greece
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0188-7/23/06.
https://doi.org/10.1145/3592573.3593103

proposed a novice-friendly system with an enhanced query parser that splits queries into concepts, time, and location. They also used CLIP embeddings to build their search backend besides the concept-based approach. Memento 2.0 [2] used an ensemble approach to rank images based on a query using image embeddings from two CLIP models.

FIRST 3.0 [17] attempted to enhance the CLIP embeddings using an attention-like approach where they encode the salient features of the image along with the information in the overall image to improve the semantic understanding of the image. Voxento 3.0 [3] presented an improved version of their earlier system incorporating functionalities such as text-based search, enhanced speech query, and new filters based on metadata. This system like its previous version also leverages image-text embeddings from the CLIP model. vitrivr [16] which has also participated in several previous iterations of LSC proposed a backend comprised of Cineast which is a feature extraction and query processing engine along with CottontailDB [10] as a database. They also extract textual embeddings from the lifelog images using a similar approach to W2VV++ [24], as well as the CLIP model. vitrivr-VR [29] is a virtual-reality based system built on top of vitrivr's [16] backend providing a fully-immersive experience to the search process. The system also supports multiple query types such as textual input, boolean queries, and geospatial queries. LifeXplore [23] which has been participating in LSC since 2018 presented an enhanced system that utilizes deep features such as YOLOv4 [5], The system also leverages embeddings from InceptionNetv3 [30] model to retrieve visually similar images.

Overall a majority of the systems participating in LSC'22 used the image-text embeddings from the CLIP model given its superior zero-shot performance and robust generalization capabilities to datasets such as lifelogs. Our proposed system Memento 3.0 like its predecessor system uses better image-text representations derived from larger CLIP [27] models and makes a significant improvement in query processing time.

3 SYSTEM OVERVIEW

In this section, we present an overview of the LSC'23 Dataset and discuss the improvements in our search and ranking functionality, as well as the modifications in the user interface to better accommodate the new query types introduce in the lifelog search challenge.

3.1 LSC'23 Data

The Lifelog Search Challenge 2023 reuses the dataset from last year's challenge. The dataset consists of ~724K first-person images collected using a narrative clip device from a single lifelogger for an 18-month period during 2019-2020. All the images in the dataset are fully redacted and anonymized as per GDPR norms.

- **Visual Concepts:** For each image in the dataset, the visual concepts consist of information such as detected objects within the image, image caption along with caption confidence score, and text detected from images using off-the-shelf OCR models.
- **Metadata:** The metadata for LSC'23 is similar to last year's data consisting of data points like biometrics (calories burnt, heart rate, step count, etc.), sleep information such as sleep stages as well as sleep efficiency and music data. The GPS location data, however, has been enhanced from last year which includes missing values imputation by leveraging the techniques discussed in [33].

3.2 CLIP Embeddings for Semantic Search

The backend of our system Memento 2.0 which participated in Lifelog Search Challenge 2022 consisted of an ensemble of scores from two CLIP models, one using a Vision Transformer [9] backbone (ViT-L/14) while the other one had a ResNet50 [15] architecture (RN50x64). The ensemble model showed better performance than the individual models when evaluated on queries from LSC 2019, however, the ViT-L/14 model seemed to work fairly well in practice as well for most of the LSC queries. Given the success of the CLIP model on out-of-domain datasets and on varied downstream tasks, a lot of effort recently has been carried out to train language-vision models on internet scale data [19, 26]. Recently [8] attempted to train a suite of open source CLIP [27] models (OpenCLIP [18]) using a subset (2 billion images) of the LIAON-5B [28] dataset which is a large-scale publicly available dataset consisting of 5 billion image-text pairs.

We compared the performance of our ensemble approach (Memento 2.0) and the CLIP models (ViT-L/14, ResNet50x64) from OpenAI with recent larger models from [8] which are trained on ~5x more data and are ~2-3x larger in terms of parameters to observe if scaling the model with respect to model size and training data translates into a tangible jump in zero-shot performance on a challenging dataset like lifelog. We evaluated all the models on queries from LSC 2019 on the Hit@K metric. The evaluation benchmark has been intentionally kept consistent with the last two iterations of our system to allow easy comparison between models.

We observed that despite larger training data and model size, the OpenCLIP models do not show a significant performance jump when compared with the OpenAI ViT-L/14 model which is the largest model in the OpenAI CLIP model suite.

The OpenCLIP ViT-L-14 model (same size as OpenAI ViT-L/14) and OpenCLIP ViT-H/14 model (~2x larger than OpenAI ViT-L/14) were not able to improve upon the OpenAI ViT-L/14 model in our evaluation results. However, we observed that the performance of OpenCLIP ViT-G-14 (~3x larger than OpenAI ViT-L/14) was better than the OpenAI ViT-L/14 model when evaluated at t=60 seconds while the ViT-L/14 model performs better at lower t values (0 seconds and 30 seconds). Our evaluation results (discussed in section 4) show a mixed picture where it is not possible to single out a concrete winner.

Our proposed system, Memento 3.0 hence adopts an approach where the backend model powering the search can be switched seamlessly by the user from the primary search UI as per the situation and requirements. As observed from LSC'22, the ViT-L/14 model works well in practice for a majority of the LSC tasks and is capable of ranking the relevant images higher up the order given adequate scene descriptions, however, for some scenarios where it fails to locate the target image or locates it lower down the order, the ViT-G/14 model or the ensemble models could be leveraged.

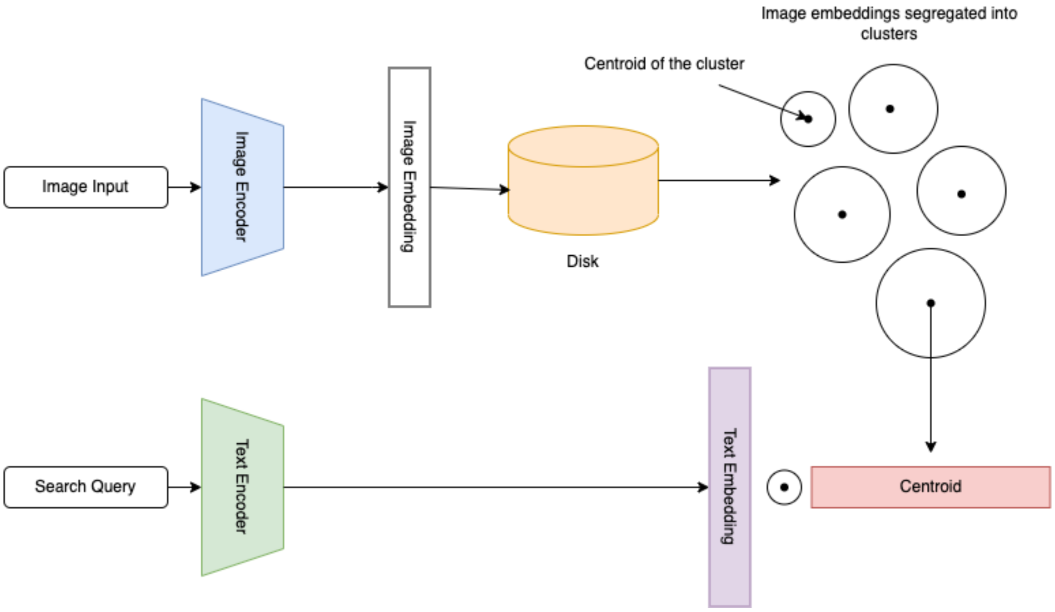


Figure 1: Search flow for Memento 3.0. Initially, the raw lifelog images are encoded using the image encoder and stored as a static file on the disk. Further, the indexing process creates clusters from the embeddings where each cluster has a centroid which acts as a cluster identifier. At run-time, the search query is passed through the text encoder, and the text embedding is first matched with the centroids of all the clusters. The cluster whose centroid matches best with the text embedding is chosen and the search then confines to that particular cluster instead of the entire dataset.

3.3 Improved Vector Similarity Search

The search mechanism of our system relies on computing vector similarity scores, where a vector representation of the search query is matched with vector representations of all the images in our dataset which are then sorted based on cosine similarity scores. A brute-force approach of matching the query vector with all image vectors is not scalable and significantly slows down the search process given that the lifelog dataset has grown in size by a factor of ~3.5 when compared with the dataset used for LSC 2020.

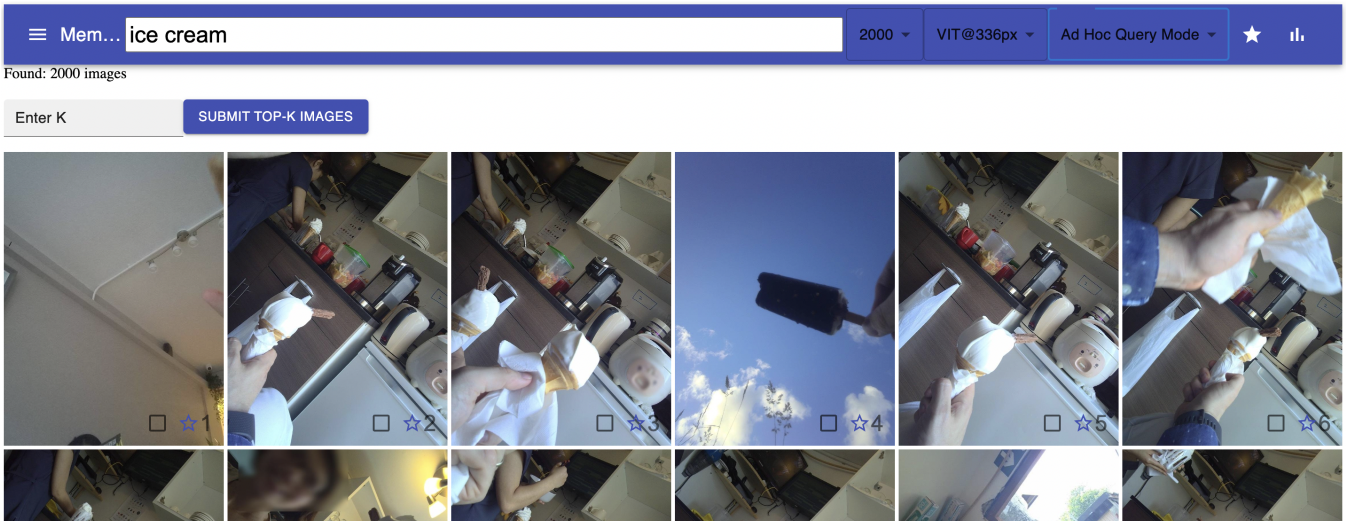
We leverage FAISS (Facebook AI Similarity Search) [20] which is an open-source library for efficient similarity search and clustering of large datasets. The library offers several indexing methods which basically partition the dataset into smaller subsets or clusters and search only the relevant subsets instead of the entire dataset.

For our use case, we use the inverted file index from the FAISS library which segregates the dataset into a set of clusters using a clustering algorithm such as k-means. The number of clusters is a hyperparameter and is empirically chosen to be 10. At run time the query vector is initially matched with the centroids of all 10 clusters and a cluster is chosen based on the highest centroid matching score. The search process is then confined to the chosen cluster instead of going over the entire dataset thus reducing the computations by a factor of 10 when compared with the previous 2 iterations of our system, Memento 1.0 [1] and Memento 2.0 [2]

3.4 Modifications to the User Interface

The user interface of Memento 3.0 is based on its previous versions but with added functionalities to efficiently handle the new query types introduced in the Lifelog Search Challenge.

- **User controls for flexible search:** In the current version of our system, the user has more flexibility in terms of choosing the backend model for search as well as choosing the number of images they want to display on the primary screen. Enabling the users to dynamically change these settings will result in increased efficiency given the new format of the competition. For example, for the Known item search task where one correct image is needed, the user can choose to display the top 100 images while for the Ad-hoc task that requires multiple correct submissions, it would be useful to look at the top 1000 or 2000 images as some relevant images may be found lower down the order. The functionality to dynamically switch the backend model was added to allow the user to opt for a backend model given the situation and the current requirements.
- **Multiple query modes** Further to facilitate the varied requirements of the 3 query types i.e Known item search, Ad-hoc queries and QA type queries our system supports different UI layouts/widgets which can be toggled using the drop-down menu from the search bar.
  - Known Item Search: For this query type the search and submission mechanism is similar to the previous system Memento 2.0 [2] which has functionalities to zoom and


Mem... ice cream

2000 VIT@336px Ad Hoc Query Mode

Found: 2000 images

Enter K

Figure 2: Memento 3.0: Primary Search Interface for Ad-Hoc LSC tasks.

submit a particular image from the main screen or save it and submit it later when more hints are revealed.

- Ad-hoc task: This query type requires the system to submit all correct images given a query within a limited time frame. To handle this efficiently we have added a functionality to submit the top-K images displayed on the primary search interface. Every image is assigned a number using which the user can decide the number of images he/she wants to submit. However, submitting a large bulk of images at once might not be a good idea in some scenarios and might include a lot of incorrect images as well. To handle this the system also supports individual manual submission from the main screen by clicking on the checkbox corresponding to that particular image. Figure 2 shows the user interface for Ad-hoc task types.
- QA task: For the QA query type, the system has a separate input box displayed on top of the screen below the primary search bar to submit the answer.

4 SYSTEM EVALUATION

Our evaluation approach for Memento 3.0 is consistent with the last two iterations of our system, Memento 1.0 [1] and Memento 2.0 [2]. We evaluate the 9 models on the HiT@K metric which can be defined as finding at least one target image among top-K images in the result set using multiple K values (1, 3, 5, 10, 20, 50, and 100) over information (hints) available to us at time= 0, 30 and 60 seconds. We reuse the same manually created evaluation queries used to evaluate our system's previous iterations.

We evaluate the following 9 models on 24 evaluation queries from LSC 2019:

(1) **ViT-B/32:** Baseline model which powered the backend of Memento 1.0 in LSC 2020
(2) **ViT-L/14:** A larger Vision Transformer model released by OpenAI as successor to (1). It generates 768-dimensional image-text embeddings.
(3) **ResNet50x64:** OpenAI ResNet-50 model using 64x the compute of a ResNet-50. It generates 1024-dimensional image-text embeddings.
(4) **ViT-L/14@336px:** OpenAI ViT-L/14 model pre-trained at a higher 336 pixel resolution for one additional epoch to boost performance.
(5) **ViT-L/14 (OpenCLIP):** Trained on 5x more data and is similar in size to the OpenAI ViT-L/14 model. It generates 768-dimensional image-text embeddings.
(6) **ViT-H/14 (OpenCLIP):** Trained on 5x more data and has 2x parameters as compared the OpenAI ViT-L/14 model. It generates 1024-dimensional image-text embeddings.
(7) **ViT-g/14 (OpenCLIP):** Trained on 5x more data and has 3x parameters as compared with the OpenAI ViT-L/14 model. It generates 1024-dimensional image-text embeddings.
(8) **Ensemble 3:1 (OpenAI ViT-L/14 and RN50x64):** Weighted sum of cosine scores from OpenAI ViT-L/14 and ResNet50x64 in a 3:1 ratio.
(9) **Ensemble 3:1 (OpenCLIP ViT-g/14 and RN50x64):** Weighted sum of cosine scores from OpenCLIP ViT-G/14 and ResNet50x64 in a 3:1 ratio.

Table 1 shows the hit percentages calculated from all 9 models for 24 LSC 2019 evaluation topics at different values of K and t. The OpenCLIP models ViT-L/14 and ViT-H-14 were unable to beat the benchmark ViT-L/14 model from OpenAI despite their larger training dataset and model size. The OpenCLIP ViT-G/14 model, however, performs better than the OpenAI ViT-L/14 model at higher values of t (60 seconds) while ViT-L/14 seems to beat all other models at lower t values (0 seconds and 30 seconds). Our proposed system Memento 3.0 hence adopts an approach where the backend model can be switched dynamically as per the situation and requirements of the user.

	t	@1	@3	@5	@10	@20	@50	@100
ViT-B/32 (Baseline)	0 sec	8.33	25.00	29.17	29.17	37.50	50.00	62.50
	30 sec	8.33	25.00	25.00	33.33	33.33	54.17	58.33
	60 sec	12.50	29.17	29.17	41.67	54.17	75.00	79.17
ViT-L/14	0 sec	20.83	33.33	41.67	50.00	54.17	62.50	83.33
	30 sec	33.33	41.67	41.67	45.83	58.33	66.67	75.00
	60 sec	37.50	45.83	45.83	54.17	62.50	79.17	83.33
ResNet50x64	0 sec	25.00	29.17	29.17	29.17	45.83	58.33	70.83
	30 sec	25.00	33.33	41.67	50.00	58.33	62.50	70.83
	60 sec	25.00	37.50	41.67	54.17	58.33	75.00	79.17
ViT-L/14@336px	0 sec	12.50	33.33	33.33	50.00	58.33	66.67	75.00
	30 sec	29.17	41.67	45.83	45.83	50.00	66.67	66.67
	60 sec	**45.83**	50.00	50.00	54.17	62.50	70.83	75.00
ViT-L-14 (Open CLIP)	0 sec	12.50	20.83	29.17	41.67	41.67	58.33	70.83
	30 sec	20.83	37.50	41.67	45.83	54.17	58.33	66.67
	60 sec	25.00	37.50	45.83	50.00	62.50	62.50	75.00
ViT-H-14 (Open CLIP)	0 sec	16.67	20.83	25.00	33.33	50.00	54.17	66.67
	30 sec	37.50	41.67	45.83	45.83	50.00	58.33	66.67
	60 sec	41.67	45.83	45.83	45.83	50.00	70.83	79.17
ViT-g-14 (Open CLIP)	0 sec	20.83	25.00	29.17	45.83	50.00	58.33	70.83
	30 sec	25.00	41.67	45.83	54.17	58.33	62.50	66.67
	60 sec	29.17	**54.17**	**54.17**	**66.67**	70.83	70.83	**87.50**
Ensemble 3:1 (OpenAI ViT-L-14 and RN50x64)	0 sec	29.17	29.17	33.33	54.17	58.33	70.83	79.17
	30 sec	37.50	41.67	45.83	54.17	58.33	70.83	75.00
	60 sec	41.67	45.83	45.83	62.50	70.83	**83.33**	83.33
Ensemble 3:1 (ViT-g-14 and RN50x64)	0 sec	16.67	33.33	33.33	37.50	50.00	62.50	75.00
	30 sec	25.00	45.83	50.00	54.17	58.33	62.50	70.83
	60 sec	29.17	45.83	50.00	62.50	**75.00**	**83.33**	**87.50**

Table 1: Hit@K calculated for all 9 models at different amounts of elapsed times, t, and K values across 24 evaluation topics for LSC'19. The highest value in each column is highlighted in bold

5 CONCLUSION AND FUTURE WORK

In this work, we present Memento 3.0, an enhanced version of our previous system Memento 2.0 [2]. We derive embeddings from both OpenAI CLIP models and larger OpenCLIP models which are trained on ~5x more data and evaluated their performance on the HiT@K metric using queries from LSC 2019. We further made improvements in the query processing time by adopting an approximate nearest neighbor search algorithm as opposed to the brute-force approach. Additionally, we modified the user interface of our system to accommodate the newer query types introduced in the lifelog search challenge.

In the future, it would be interesting to experiment with enhanced image embeddings which are not just good at scene understanding but also encapsulate other non-visual information within them.

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, Yvette Graham, and Cathal Gurrin. 2021. Memento: A Prototype Lifelog Search Engine for LSC'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 53–58. https://doi.org/10.1145/3463948.3469069

[2] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2022. Memento 2.0: An Improved Lifelog Search Engine for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. ACM, Newark NJ USA, 2–7. https://doi.org/10.1145/3512729.3533006

[3] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2022. Voxento 3.0: A Prototype Voice-Controlled Interactive Search Engine for Lifelog. In *Proceedings*

of the 5th Annual on Lifelog Search Challenge. ACM, Newark NJ USA, 43–47. https://doi.org/10.1145/3512729.3533009

[4] Seyed Ali Bahrainian and Fabio Crestani. 2018. Augmentation of Human Memory: Anticipating Topics that Continue in the Next Meeting. (2018), 10.

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. http://arxiv.org/abs/2004.10934 arXiv:2004.10934 [cs, eess].

[6] Alejandro Cartas, Juan Marín, Petia Radeva, and Mariella Dimiccoli. 2017. Recognizing Activities of Daily Living from Egocentric Images. *arXiv:1704.04097 [cs]* (April 2017). http://arxiv.org/abs/1704.04097 arXiv: 1704.04097.

[7] Mariona Carós, Maite Garolera, Petia Radeva, and Xavier Giro-i Nieto. 2020. Automatic Reminiscence Therapy for Dementia. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. ACM, Dublin Ireland, 383–387. https://doi.org/10.1145/3372278.3391927

[8] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. http://arxiv.org/abs/2212.07143 arXiv:2212.07143 [cs].

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929 [cs]* (June 2021). http://arxiv.org/abs/2010.11929 arXiv: 2010.11929.

[10] Ralph Gasser, Luca Rossetto, Silvan Heller, and Heiko Schuldt. 2020. Cottontail DB: An Open Source Database System for Multimedia Retrieval and Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 4465–4468. https://doi.org/10.1145/3394171.3414538

[11] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Foundations and Trends® in Information Retrieval* 8, 1 (2014), 1–125. https://doi.org/10.1561/1500000033

[12] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoć, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*. ACM, Newark NJ USA, 685–687. https://doi.org/10.1145/3512527.3531439

[13] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *Proc. International Conference on Multimedia Retrieval (ICMR'21)*. ACM, Taipei, Taiwan.

[14] 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.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[16] Silvan Heller, Luca Rossetto, Loris Sauter, and Heiko Schuldt. 2022. vitrivr at the Lifelog Search Challenge 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. ACM, Newark NJ USA, 27–31. https://doi.org/10.1145/3512729.3533003

[17] Nhat Hoang-Xuan, Hoang-Phuc Trang-Trung, E-Ro Nguyen, Thanh-Cong Le, Mai-Khiem Tran, Tu-Khiem Le, Van-Tu Ninh, Cathal Gurrin, and Minh-Triet Tran. 2022. Flexible Interactive Retrieval SysTem 3.0 for Visual Lifelog Exploration at LSC 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. ACM, Newark NJ USA, 20–26. https://doi.org/10.1145/3512729.3533013

[18] 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.

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. http://arxiv.org/abs/2102.05918 arXiv:2102.05918 [cs].

[20] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[21] Kenji Karako, Yu Chen, Peipei Song, and Wei Tang. [n.d.]. Super-aged society: Constructing an integrated information platform of self-recording lifelogs and medical records to support health care in Japan. *BioScience Trends.* ([n. d.]), 3.

[22] Seongjung Kim, Seongkyu Yeom, Oh-Jin Kwon, Dongil Shin, and Dongkyoo Shin. 2018. Ubiquitous Healthcare System for Analysis of Chronic Patients' Biological and Lifelog Data. *IEEE Access* 6 (2018), 8909–8915. https://doi.org/10.1109/ACCESS.2018.2805304 Conference Name: IEEE Access.

[23] Andreas Leibetseder, Daniela Stefanics, and Klaus Schoeffmann. 2022. lifeXplore at the Lifelog Search Challenge 2022. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. ACM, Newark NJ USA, 48–52. https://doi.org/10.1145/3512729.3533005

[24] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. 2019. W2VV++: Fully Deep Learning for Ad-hoc Video Search. In *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, Nice France, 1786–1794. https://doi.org/10.1145/3343031.3350906

[25] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh, Minh-Triet Tran, Thanh Binh Nguyen, Graham Healy, Sinéad Smyth, Annalina Caputo, and Cathal Gurrin. 2022. LifeSeeker 4.0: An Interactive Lifelog Search Engine for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. ACM, Newark NJ USA, 14–19. https://doi.org/10.1145/3512729.3533014

[26] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, Mingxing Tan, and Quoc V. Le. 2022. Combined Scaling for Open-Vocabulary Image Classification. http://arxiv.org/abs/2111.10050 arXiv:2111.10050 [cs].

[27] 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.

[28] 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. http://arxiv.org/abs/2210.08402 arXiv:2210.08402 [cs].

[29] Florian Spiess and Heiko Schuldt. 2022. Multimodal Interactive Lifelog Retrieval with vitrivr-VR. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. ACM, Newark NJ USA, 38–42. https://doi.org/10.1145/3512729.3533008

[30] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. http://arxiv.org/abs/1512.00567 arXiv:1512.00567 [cs].

[31] Ly-Duyen Tran, Manh-Duy Nguyen, Binh Nguyen, Hyowon Lee, Liting Zhou, and Cathal Gurrin. 2022. E-Myscéal: Embedding-based Interactive Lifelog Retrieval System for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge*. ACM, Newark NJ USA, 32–37. https://doi.org/10.1145/3512729.3533012

[32] Ly-Duyen Tran, Manh-Duy Nguyen, Nguyen Thanh Binh, Hyowon Lee, and Cathal Gurrin. 2021. Myscéal 2.0: A Revised Experimental Interactive Lifelog Retrieval System for LSC'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 11–16. https://doi.org/10.1145/3463948.3469064

[33] Ly-Duyen Tran, Dongyun Nie, Liting Zhou, Binh Nguyen, and Cathal Gurrin. 2023. VAISL: Visual-Aware Identification of Semantic Locations in Lifelog. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part II*. Springer. in press.