An exploration into the benefits of the CLIP model for lifelog retrieval

Ly-Duyen Tran*, Naushad Alam*, Linh Khanh Vo†, Nghiem Tuong Diep†,
Binh Nguyen†, Yvette Graham¶, Liting Zhou*, and Cathal Gurrin*
*School of Computing, Dublin City University, Dublin, Ireland
Email: ly.tran2@mail.dcu.ie
†AISIA Research Lab, Ho Chi Minh, Vietnam
‡University of Science, Ho Chi Minh, Vietnam
§Vietnam National University, Ho Chi Minh, Vietnam
¶School of Computer Science and Statistics, Trinity College, Dublin, Ireland

Abstract—**In this paper, we attempt to fine-tune the CLIP (Contrastive Language-Image Pre-Training) model on the Lifelog Question Answering dataset (LLQA) to investigate retrieval performance of the fine-tuned model over the zero-shot baseline model. We train the model adopting a weight space ensembling approach using a modified loss function to take into account the differences in our dataset (LLQA) when compared with the dataset the CLIP model was originally pretrained on. We further evaluate our fine-tuned model using visual as well as multimodal queries on multiple retrieval tasks demonstrating improved performance over the zero-shot baseline model.**

Index Terms—**lifelogging, image retrieval, pretrained models**

I. Introduction

"Where are my car keys?". Human memory can be fallible and unreliable, though it is undoubtedly a vital cognitive ability [1]. We humans tend to constantly forget trivial things, such as failing to remember the location of our things, details about a recent event, or simply struggling to remember the name of the person we just met. In this work, we are interested in augmenting human memory by building a digital twin of an individual that can answer all such daily informaiton needs.

Lifelogging, as defined by [2], is the process of passively capturing a personal digital collection of daily life experiences using a variety of devices such as wearable cameras, tracking devices such as Fitbit and other wearable sensor devices. As a concept, lifelogging was introduced in Vannevar Bush's 1945 article 'As We May Think' [3] where he discusses about a "future mechanised device" which acts as an "enlarged intimate supplement of an individual's memory" storing all his books, records, communications and can be consulted with "exceeding speed and flexibility".

The last two decades have witnessed growing attention to lifelogging after MyLifeBits [4] was proposed by Gemmell and Bell in the early 2000s. Further, with the advancement of sensor technology, the availability of cheap storage facilities and cost-efficient wearable devices recording one's life passively has become feasible, and hence lifelogging has witnessed a surge in interest from the research community over time. Due to the sheer variety of data collected, lifelog data have been used to address various use cases in research domains such as signal processing [5], [6], natural language processing [7], [8], computer vision [8], [9], and human-computer interactions [10].

Information retrieval from lifelogs to realise the goal of memory augmentation is, however, a very challenging problem as human memory is pervasive and immediate while retrieval from lifelogs using implicit queries is an iterative and cumbersome process. The multimodal characteristics of lifelog data, which includes data from multiple sources such as egocentric images, textual data specifying details like location, time, date and biometrics as well as it being a noisy and repetitive archive due to passive data collection over longer periods of time, further adds to the challenges of developing an effective retrieval system.

Recent models like Contrastive Language-Image Pre-Training (CLIP) [11], A Large-scale ImaGe and Noisy-text embedding (ALIGN) [12], etc. which leverages the supervision inherent in natural language texts to learn generalised vision-language representations that can further be used to solve multiple downstream tasks such as information retrieval, object recognition, scene recognition have seen tremendous success on multiple benchmarks. The zero-shot CLIP model [11] beats several supervised baselines on multiple datasets, showing robust transfer capability when applied to out-of-domain datasets. However, as discussed in Section II-B the model fares poorly when applied to certain specialised datasets, which motivated us to experiment with model fine-tuning on lifelogs and compare the performance with the zero-shot model.

This work aims to investigate whether fine-tuning the CLIP model on domain-specific data improves the model's performance when compared with the zero-shot baseline model to solve the task of lifelog information retrieval. Our contributions in this paper are as follows.

- We attempt to fine-tune the CLIP model on an in-the-wild egocentric multimodal dataset (Lifelogs), which to the best of our knowledge is the first work done in this direction.
- We devise a modified loss function to accommodate the structure of the dataset we use to fine-tune the CLIP model.

- We evaluate our fine-tuned model on multiple retrieval tasks using both visual and multimodal queries demonstrating superior performance over the zero-shot baseline model.

The results of this study will provide insight into the design decisions of lifelog retrieval systems in the future. The rest of the paper is structured as follows: Section II discusses the efforts carried out so far to fine-tune the CLIP models on various niche datasets, as well as briefly covering the major milestones achieved so far at a high level in the area of transfer learning. Subsequently, Section III discusses the question answering LLQA dataset which has been used to fine-tune the CLIP model followed by a detailed discussion on our adopted methodology. Finally, in Section IV, we evaluate the performance of our fine-tuned model over the zero-shot baseline using both visual queries as well as multimodal queries as input.

II. Related Work

A. Transfer Learning

Curating a high-quality, large-scale annotated dataset is a challenge in many specialised research domains, making it hard to train large deep learning models from scratch. Transfer learning aims to solve this issue of insufficient training data by 'transferring' knowledge from a data-abundant source domain to a data-scarce target domain [13]. For a long time now, researchers have used pretrained features from ImageNet [14] to solve several downstream computer vision tasks such as image classification, object detection, action recognition, image segmentation, etc. In recent years, transfer learning in the form of pretrained language models [15] [16] [17] based on the tranformer architecture [18] has become quite ubiquitous in the field of natural language processing as well, achieving state-of-the-art performance in areas like machine translation, natural language inference, etc.

Recently, models like CLIP [11] and ALIGN [12] generating zero-shot transferable representations have made a leap forward towards generalised models which can work without any data-specific fine-tuning. However, as discussed in Section II-B, zero-shot transfer to few specific domains is still very challenging. Consequently, several recent works have tried to leverage the pretrained CLIP model to further improve its performance by fine-tuning it on various specialised datasets and have demonstrated competitive performance over the zero-shot model.

B. Fine-tuning CLIP

As discussed in [11], the zero-shot CLIP model has shown significant gains over the performance of fully supervised ResNet-50 [19] baselines trained on several datasets. However, the zero-shot model fails to surpass the performance of supervised models on a few specialised datasets such as EuroSAT [20] and RESISC45 [21] (satellite image classification), PatchCamelyon [22] (lymph node tumour detection), CLEVR-Counts [23] (counting objects in synthetic scenes), GTSRB [24] (German traffic sign recognition), KITTI Distance [25] (recognising distance to the nearest car), which shows the room for improvement for such complex and niche datasets.

Several recent works have tried to fine-tune the network to improve its performance over specialised datasets in order to address abstract problems. Clip-Art [26] aims to solve the problem of retrieving and classifying fine-grained attributes of artwork images by fine-tuning the network on the iMet dataset. PointCLIP [27] fine-tuned the model with the objective of transfer learning across different modalities. It learns to efficiently transfer representations learnt from 2D images to do cross-modality zero-shot recognition on a 3D point cloud. Arutiunian et al. [28] fine-tuned the model on satellite images and captions from the RSICD dataset [29] to support satellite image retrieval using queries in natural language. In addition, ActionCLIP [30] applied the model to perform video action recognition.

Another line of work focusses on strategies and techniques to robustly fine-tune the CLIP model. CLIP Adapter [31] adopts a lightweight bottleneck architecture to prevent the potential overfitting problem of few-shot learning by reducing the number of parameters and only fine-tuning a small number of additional weights instead of optimising all CLIP parameters. Tip-Adapter [32] (Training free CLIP-Adapter) further improves over CLIP-Adapter [31] by doing away with stochastic gradient descent to train the adapter and instead constructing a query-key cache model from few-shot supervisions to obtain the weights of the adapter. WiSE-FT [33] proposed to do weight space ensembling between zero-shot and fine-tune models to preserve the model's accuracy under data distribution shift.

We attempt to fine-tune the CLIP model on the lifelog dataset, which is a in-the-wild egocentric multimodal dataset adopting the weight-space ensembling approch from [33] demonstrating encouraging results.

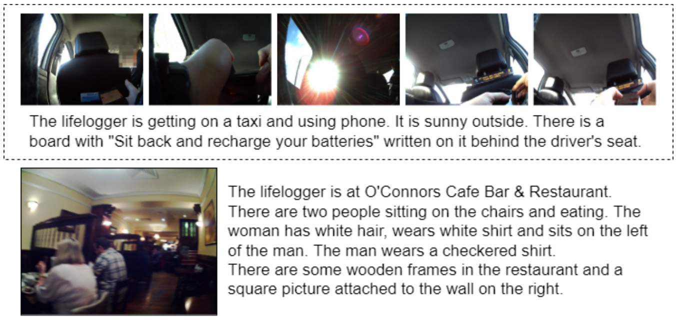
C. Lifelog Retrieval

Effective information retrieval from lifelogs has been a long-standing challenge given the multimodal nature and size of the dataset as well as the very specific nature of information users would want to retrieve from it. In the recent few years, several benchmarking challenges have been organized like the Lifelog Search Challenge [34], ImageCLEF Lifelog tasks [35]–[38], NTCIR-Lifelog tasks [39]–[42] to advance the state of the art in lifelog information retrieval.

LSC over the years has seen participation of many video retrieval systems having previously participated in the VBS Challenge and tweaked to support lifelog retrieval [43]–[45]. Systems like Exquisitor [46], THUIR [47] proposed to use relavance feedback from users to guide the search process. Virtual reality systems [48]–[50] providing a fully-immersive experience to the search process have also been popular due to their unique interactivity. Likewise several past systems leveraged visual concepts derived from object detection models to build their retrieval engines [51]–[54]. Recently, approches leveraging multimodal embeddings [55] and in particular the zero-shot CLIP model for lifelog retrival have surpassed prior state-of-the-art techniques in the field [56]–[60].

III. Experiment Setups

A. Lifelog captions



The lifelogger is getting on a taxi and using phone. It is sunny outside. There is a board with "Sit back and recharge your batteries" written on it behind the driver's seat.

The lifelogger is at O'Connors Cafe Bar & Restaurant. There are two people sitting on the chairs and eating. The woman has white hair, wears white shirt and sits on the left of the man. The man wears a checkered shirt. There are some wooden frames in the restaurant and a square picture attached to the wall on the right.

Fig. 1. Examples of annotated descriptions in the updated LLQA dataset. The first example feature general narrative descriptions for a longer episode of activity. The second one include details in a single image, usually when the lifelogger was moving and the surroundings change considerably. Each sentence in these descriptions count as one caption in the LLQA dataset used to fine-tune CLIP models.

TABLE I
Statistics of the number of images that one caption describes.

	#captions	mean	min	25th	50th	75th	max
Original	11398	15.89	1	5	8	15	297
Updated	1919	3.39	1	1	1	1	112

To generate a dataset to fine-tune the CLIP models, we utilise the lifelog question answering dataset LLQA [61], which includes questions and answers automatically generated from human-annotated captions. Although LLQA was not created for the purpose of this paper, the number of lifelog captions collected is the largest to the best of our knowledge. A total of 11,398 captions are available in the dataset, describing daily activities on 85 days of lifelog. However, since this work was an initial attempt at lifelog captioning and question answering, a large portion of the descriptions are too vague and not specific enough for some use cases (as previously seen in some queries in the Lifelog Search Challenges (LSC) [34], [62], [63]). For this reason, since the publication, we have added more captions to the original dataset, including more detailed descriptions that are more suitable. An addition of 1,919 captions is added, which describe details of each image instead of vague, general activities. The comparison between these two parts of the dataset is presented in Table I. Furthermore, Figure 1 shows two examples from the dataset.

One challenge in adapting this dataset is that the description can describe a period of activity, including multiple images, some of which do not match the caption individually. Thus, we filter the dataset to choose only instances where the caption covers at most 15 images to reduce the possibility of ill-matched pairs of caption and image. The chosen instances are then divided into a training set and a validating set, as detailed in Table II to fine-tune the CLIP models, making sure there are no overlapping images between the two sets.

TABLE II
LLQA dataset splits to fine-tune CLIP models.

Split	#image-caption pair	#unique images	#unique captions
Train	11982	6328	2916
Validate	1234	421	328

B. Fine-tuning CLIP models

CLIP models [11] were originally designed to match a single image with a single caption. However, since the descriptions we use often span across multiple images, we modify the loss function accordingly to take into account this characteristic. For every mini-batch, with T as text embedding matrix, I as image embedding matrix, we calculate text similarity S_T and image similarity S_I using the cosine similarity function in Equation 5. The pairwise similarities between text and image, which are $Logits$, are aimed to match with the mean self similarity (of text and image) $Target$ using cross-entropy loss,

$$\mathbf{S_T} = \frac{\mathbf{T} \cdot \mathbf{T}^{\tau}}{\|\mathbf{T}\| \|\mathbf{T}^{\tau}\|}; \; \mathbf{S_I} = \frac{\mathbf{I} \cdot \mathbf{I}^{\tau}}{\|\mathbf{I}\| \|\mathbf{I}^{\tau}\|} \tag{1}$$

$$Logits = \frac{\mathbf{T} \cdot \mathbf{I}^{\tau}}{\|\mathbf{T}\| \|\mathbf{I}^{\tau}\|}; \; Target = \sigma\Big(c \cdot \frac{\mathbf{S_T} + \mathbf{S_I}}{2}\Big) \tag{2}$$

$$Loss = crossEntropy(Logits, Target) \tag{3}$$

where σ is the softmax function and c is the logit scale.

Due to our limitation of GPU power, we could not fine-tune the best performing model, 'ViT-L/14', amongst the public releases. Instead, we choose to use the pretrained 'ViT-B/32' and 'ViT-B/16' for our experiments. In order to prevent overfitting, we selected the largest minibatch size possible on our machine, which is 48 and 24 for the two models, respectively. We use Adam Optimiser [64] with a weight decay regularization [65] of 0.01, except for gains or biases, and decay the learning rate using a cosine scheduler [66]. Large pretrained CLIP models can perform zero-shot inference with consistent accuracy across a variety of data, which is a valuable characteristic that we want to maintain. For this reason, Wortsman et al. [33] suggested the idea of interpolating the weights between the fine-tuned model and the original to improve robustness. In other words, the final weights of the model are as follows.

$$\theta_{\text{final}} = (1 - \alpha) \cdot \theta_{\text{original}} + \alpha \cdot \theta_{\text{fine-tuned}} \tag{4}$$

The authors suggest choosing $\alpha = 0.5$ as it produced near optimal performance in various experiments. More details on how α affects the performance of the models are provided in Section IV.

IV. Evaluation and Results

Despite a large proportion of lifelog data are images, lifelog data are intrinsically multimodal. As CLIP models are incapable of explicit information, such as time or date, in this section, we will adopt CLIP models in two ways:

TABLE III
SOME EXAMPLES OF THE TASKS IN LSC'21 AND THE CORRESPONDING INPUTS FOR THE CLIP MODELS.

Task ID	Original Hints	Transformed Hints
4	(1) I needed to buy a blood pressure monitor. (2) So I was looking in a pharmacy (3) that sold Omron and Braun devices. (4) Afterwards, I waited for a long time in my dentist office, (5) before getting a coffee/bagel and driving to my own office. (6) It was in 2016.	(1) Looking to buy a blood pressure monitor (2) in a pharmacy (3) that sold Omron and Braun devices.
18	(1) I was looking at small computer chips on rolls. (2) It was in a small university electronics laboratory in China. (3) There were at least 100 rolls of small computer chips. (4) It was part of a tour of computing and engineering facilities (5) and I was with a small delegation of people. (6) It was in May 2018.	(1) Looking at small computer chips on rolls (2) in a small university electronics laboratory (3) which had at least 100 rolls.

- **Image-only**: we simplify the queries and include only content-based description;
- **Multimodal**: we incorporate CLIP model with query parsing, automatically extract non-visual information from the query and apply corresponding other search operations.

Two metrics are used to evaluate the models:

- Hit rate at K ($H@K$): $H@K = 1$ means that one of the target images appears in the top K of the result set. Otherwise, $H@K = 0$;
- Average Precision ($AP@K$): the mean of the precision scores after each relevant document is retrieved, where K is the total of relevant documents.

A. Image-only

The most recent iteration of the Lifelog Search Challenge, LSC'21 [34], presented a total of 23 queries with various difficulty levels. Each query was gradually revealed over a 30-second time interval, showing more hints of visual descriptions, time, location, etc. For this experiment, we simplify the queries and consider only three time steps, in a way similar to the approach in [56]. Some examples of the original queries and the simplified hints are shown in Table III.

For each query, we encode the query using the textual encoder of the CLIP model and calculate the similarity score of each image with the text embedding. The images are then ranked on the basis of their similarity score. With $\mathbf{q}$ as the encoded search query, and $\mathbf{c}$ as the encoded image, the similarity is defined as:

$$\cos(\mathbf{q}, \mathbf{c}) = \frac{\mathbf{q}\mathbf{c}}{\|\mathbf{q}\|\|\mathbf{c}\|} = \frac{\sum_{i=1}^{n} \mathbf{q}_i \mathbf{c}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{q}_i)^2}\sqrt{\sum_{i=1}^{n} (\mathbf{c}_i)^2}} \qquad (5)$$

The task of LSC is to find *one* instance of the lifelog moment that matches the search query. Thus, we use the Hit rate at K to measure the performance on these queries. The performance of the public release version of ViT-B/16 can be seen in Table IV, which shows the average $H@K$ of each time step.

As mentioned in the previous section, we assemble the fine-tuned CLIP model with the original pretrained weights. To choose the best value for the interpolation parameter α, we recorded H@5 scores across all interpolated models to assess the influence of α. In general, the fine-tuned models increased the retrieval result after fine-tuning, as can be seen when $\alpha = 1.0$. For the Vit-B/32 model, the increase in H@5 is mostly positively correlated with alpha. However, the pattern for the interpolated weights for ViT-B/16 models is less definite. However, the H@5 scores tend to be higher around the middle point when using two and three hints. For this reason, with the original suggestion from [33], from this point on, we choose to evaluate the fine-tuned ViT-B/16 with $\alpha = 0.5$ on different tasks and address it as **LifelogCLIP** for the sake of simplicity.

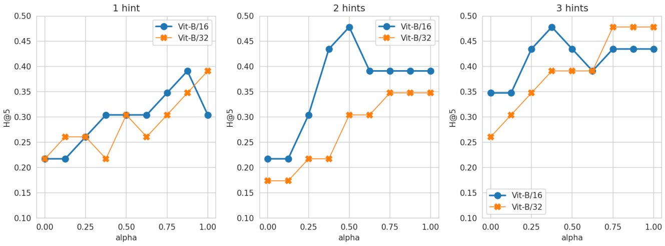


Fig. 2. Affect of α on H@5 when interpolation the fine-tuned models.

TABLE IV
ORIGINAL CLIP VIT-B/16 PERFORMANCE ON 23 QUERIES OF LSC'21

	H@1	H@3	H@5	H@10	H@20	H@50	H@100
h=1	0.17	0.22	0.22	0.26	0.35	0.48	0.52
h=2	0.17	0.22	0.22	0.35	0.43	0.52	0.52
h=3	0.26	0.30	0.35	0.35	0.48	0.57	0.61

The performance of LifelogCLIP is detailed in Table VI. The table shows an increase in almost all hit rates, compared to the original result in Table IV. Surprisingly, the H@1 score for $h = 2$ is lower than that of $h = 1$, considering the intuitive assumption that more hints should increase the score, as seen in other cases. This can be explained by the fact that CLIP classifiers can be sensitive to wording or phrasing [11]. Hence, adding more information, which changes the phrasing, does not always improve the performance. Other than that, the most significant improvements tend to be in the first row where $h = 1$ and in lower values K. This proves that the fine-tuned

TABLE V
INPUTS OF THE SAME TASKS IN TABLE III FOR LIFELOGCLIP IN E-MYSCÉAL.
* INDICATES OMITTING INFORMATION DUE TO THE LIMITATION OF THE SYSTEM.

Task ID	Before	Main	After
4		Buying a blood pressure monitor in a pharmacy that sold ”Omron” and ”Braun” device in 2016.	I waited for a long time in my dentist office.*
18	A tour of computing and engineering facilities and I was with a small delegation of people	I was looking at small computer chips on rolls. There were at least a hundred rolls of small computer chips. It was in a small university electronics laboratory in China in May 2018	

TABLE VI
LIFELOGCLIP (FINE-TUNED CLIP VIT-B/16 WITH $\alpha = 0.5$) PERFORMANCE ON 23 QUERIES OF LSC’21

	H@1	H@3	H@5	H@10	H@20	H@50	H@100
h=1	0.26	0.30	0.30	0.35	0.35	0.52	0.65
h=2	0.21	0.35	0.43	0.48	0.48	0.57	0.65
h=3	0.30	0.43	0.48	0.48	0.52	0.61	0.65

model more effective in ranking the results.

B. Multimodal

Taking into account temporal and spatial clues, we incorporate the LifelogCLIP model with the query parsing unit from E-MyScéal [51], a state-of-the-art interactive lifelog retrieval system. To facilitate free-text querying (as opposed to using multimodal faceted filters), the component detects location names, as well as date and time format using part-of-speech tagging, semantic role labelling, and regex matching.

1) LSC’21 queries: LSC’21 queries are complex and usually include multiple temporal-related events. To address these queries, the user interface of E-MyScéal accepts *up to* three temporal hints as ‘before’, ‘main’, and ‘after’ queries to address the temporal context in the original search query. Thus, we manually split the LSC’21 queries into temporal queries if needed. Examples of transformed hints are shown in Table V. Note that in some cases, there are more than one ‘before’ events or more than one ‘after’ events. Due to the limitation of E-MyScéal, we only use the first event in order of appearance. For example, the fifth clue of ‘*before getting a coffee/bagel and driving to my own office*’ in Task 4 (Table III) is omitted.

TABLE VII
MEAN $H@K$ FOR LSC’21 QUERIES, USING ALL HINTS.

H@1	H@3	H@5	H@10	H@20	H@50	H@100
0.52	0.65	0.74	0.74	0.74	0.74	0.74

Using all hints, E-MyScéal’s LifelogCLIP can find the answer to more than half of the queries in the first result as seen in Table VII. Interestingly, there is no difference in the hit rates when $k \geq 10$. Since the interface of E-MyScéal can accommodate 12 events at once, this minimises the user’s effort to scroll further down the result page.

2) Comparing with baselines on NTCIR-13 lifelog queries: Since LSC’21 queries are aimed at retrieving a specific moment in lifelog data, hit rate is a suitable metric for evaluation. We also want to assess LifelogCLIP for a different type of lifelog retrieval task with another conventional information retrieval metric: average precision (AP@K). About half of the queries in the NTCIR-13 lifelog challenge focus on retrieving many instances of an activity. We choose the first ten queries and compare LifelogCLIP with the reported performance of two state-of-the-art embedding models from [8]. Similarly, we also used the cut-off point at 10 to calculate AP.

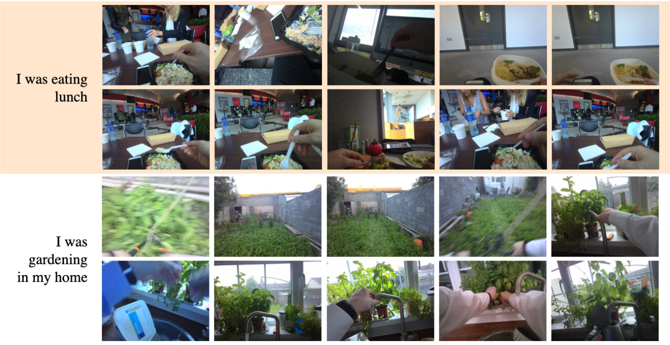


Fig. 3. Top 10 retrieved result from the first two tasks in NTCIR-13.

Table VIII details the task descriptions and the performance of the baseline models and LifelogCLIP. For each description, we remove the first part of *”Find the moment when”* or similar phraseing and use only the action (*”I was eating lunch”*) as the search query. As we can see from the table, LifelogCLIP achieved a higher score on most tasks, an equal score on one task, and a lower score on two tasks. Figure 3 illustrates the retrieval results using LifelogCLIP with query parsing. Since we are using LifelogCLIP on an image level, several results in the figure belong to the same event cluster.

The automatic retrieval results of multimodal LifelogCLIP on LSC’21 queries and NTCIR-13 lifelog queries demonstrate that the incorporation of CLIP models can increase the performance of lifelog moment retrieval on both metrics that we proposed at the beginning of the section.

TABLE VIII

THE AP@10 EVALUATED ON THE FIRST 10 TASKS OF NTCIR-13, COMPARED TO THE BASELINE APPROACHES IN [8]

Task	Description	Caption	Joint embedding	LifelogCLIP
1	Find the moments when I was eating lunch	0.65	0.88	**0.88**
2	Find moments when I was gardening in my home	0.12	0.23	**0.40**
3	Find the moment when I was visiting a castle at night	0.51	0.67	**0.78**
4	Find the moments when I was drinking coffee in a cafe	0.60	0.70	**0.88**
5	Find the moments when I was outside at sunset	0.56	**0.64**	0.51
6	Find the moments when I visited a graveyard	0.54	0.43	**1.00**
7	Find the moments when I was lecturing to a group of people in a classroom environment	0.35	0.55	**0.58**
8	Find all the moments when I was grocery shopping	0.62	0.68	**1.00**
9	Find the moments when I worked at home late at night	0.67	**0.71**	0.66
10	Find the moments when I was working on the computer at my office desk	0.57	0.85	**1.00**

V. CONCLUSION

This paper has described our efforts to fine-tune the CLIP models by collecting annotated lifelog descriptions, modifying a loss function for fine-tuning, and evaluating the fine-tuned model on different lifelog retrieval tasks. Its performance is also compared with the baseline multimodal embedding models for lifelog. In summary, we have obtained encouraging results, demonstrating that integrating the fine-tuned CLIP model with query parsing can comparatively enhance the retrieval performance. However, some limitations should be considered. First, the LLQA [61] dataset used for fine-tuning, despite being the best free-form collection of lifelog, is not in the format where CLIP models are usually trained (i.e. having exact matching image-caption pairs). Second, the size of the dataset is tremendously small for a deep learning task and might not have introduced enough difference for the model to better adapt to lifelog data. Lastly, CLIP models are incapable of taking into consideration the temporal aspect of lifelog. More studies are needed to explore these points. In particular, research on solving the last point is already in progress. Additionally, CLIP models, especially the more powerful pretrained versions, are being integrated in more systems in the next Lifelog Search Challenge [63] in various approaches. This provides us a great oppurtunity to ascertain the performance of CLIP models in the future.

ACKNOWLEDGMENT

This publication has emanated from research supported in party by research grants from Science Foundation Ireland under grant numbers SFI/12/RC/2289, SFI/12/RC/2289-P2, SFI/13/RC/2106, 18/CRT/6223 and 18/CRT/6224.

REFERENCES

[1] M. Harvey, M. Langheinrich, and G. Ward, "Remembering through lifelogging: A survey of human memory augmentation," *Pervasive and Mobile Computing*, vol. 27, pp. 14–26, 2016.

[2] C. Gurrin, A. F. Smeaton, and A. R. Doherty, "Lifelogging: Personal big data," *Foundations and trends in information retrieval*, vol. 8, no. 1, pp. 1–125, 2014.

[3] BushVannevar, "As we may think," *ACM Sigpc Notes*, 1979.

[4] J. Gemmell, C. Bell, and R. Lueder, "Mylifebits: A personal database for everything," *Communications of the ACM*, vol. 49, pp. 89–95, 01 2006.

[5] C. Dobbins and S. Fairclough, "Signal processing of multimodal mobile lifelogging data towards detecting stress in real-world driving," pp. 632–644, 2019.

[6] M. Shah, B. Mears, C. Chakrabarti, and A. Spanias, "Lifelogging: Archival and retrieval of continuously recorded audio using wearable devices," in *2012 IEEE International Conference on Emerging Signal Processing Applications*, 2012, pp. 99–102.

[7] P.-W. Kao, A.-Z. Yen, H.-H. Huang, and H.-H. Chen, "Convlogminer: A real-time conversational lifelog miner," in *Proceedings of the 30th International Joint Conference on Artificial Intelligence*, 2021.

[8] L. Zhou and C. Gurrin, "Multimodal embedding for lifelog retrieval," in *International Conference on Multimedia Modeling*. Springer, 2022, pp. 416–427.

[9] T. Ye, "Visual object detection from lifelogs using visual non-lifelog data," Ph.D. dissertation, Dublin City University, 2018.

[10] Y. Yang, H. Lee, and C. Gurrin, "Visualizing lifelog data for different interaction platforms," in *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, 2013, pp. 1785–1790.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv:2103.00020 [cs]*, Feb. 2021, arXiv: 2103.00020. [Online]. Available: http://arxiv.org/abs/2103.00020

[12] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," *arXiv:2102.05918 [cs]*, Jun. 2021, arXiv: 2102.05918. [Online]. Available: http://arxiv.org/abs/2102.05918

[13] T. Ye, "Visual object detection from lifelogs using visual non-lifelog data," Ph.D. dissertation, Dublin City University, 2018.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692 [cs]*, Jul. 2019, arXiv: 1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692

[17] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *arXiv:1910.01108 [cs]*, Feb. 2020, arXiv: 1910.01108. [Online]. Available: http://arxiv.org/abs/1910.01108

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. [Online]. Available: http://ieeexplore.ieee.org/document/7780459/

[20] P. Helber, B. Bischke, A. Dengel, and D. Borth, "EuroSAT: A Novel Dataset and Deep Learning Benchmark for Land Use and Land Cover Classification," *arXiv:1709.00029 [cs]*, Feb. 2019, arXiv: 1709.00029 version: 2. [Online]. Available: http://arxiv.org/abs/1709.00029

[21] G. Cheng, J. Han, and X. Lu, "Remote Sensing Image Scene Classification: Benchmark and State of the Art," *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017, conference Name: Proceedings of the IEEE.

[22] B. S. Veeling, J. Linmans, J. Winkens, T. Cohen, and M. Welling, "Rotation Equivariant CNNs for Digital Pathology," *arXiv:1806.03962 [cs, stat]*, Jun. 2018, arXiv: 1806.03962. [Online]. Available: http://arxiv.org/abs/1806.03962

[23] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," *arXiv:1612.06890 [cs]*, Dec. 2016, arXiv: 1612.06890. [Online]. Available: http://arxiv.org/abs/1612.06890

[24] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition," *Neural Networks*, no. 0, pp. –, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0893608012000457

[25] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[26] M. V. Conde and K. Turgutlu, "CLIP-Art: Contrastive Pre-training for Fine-Grained Art Classification," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 3951–3955. [Online]. Available: https://ieeexplore.ieee.org/document/9522786/

[27] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "PointCLIP: Point Cloud Understanding by CLIP," *arXiv:2112.02413 [cs]*, Dec. 2021, arXiv: 2112.02413. [Online]. Available: http://arxiv.org/abs/2112.02413

[28] A. Arutiunian, D. Vidhani, G. Venkatesh, M. Bhaskar, R. Ghosh, and S. Pal, "Fine tuning clip with remote sensing (satellite) images and captions," https://huggingface.co/blog/fine-tune-clip-rsicd, 2021.

[29] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring Models and Data for Remote Sensing Image Caption Generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018, arXiv: 1712.07835 version: 1. [Online]. Available: http://arxiv.org/abs/1712.07835

[30] M. Wang, J. Xing, and Y. Liu, "ActionCLIP: A New Paradigm for Video Action Recognition," *arXiv:2109.08472 [cs]*, Sep. 2021, arXiv: 2109.08472. [Online]. Available: http://arxiv.org/abs/2109.08472

[31] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," *arXiv:2110.04544 [cs]*, Oct. 2021, arXiv: 2110.04544. [Online]. Available: http://arxiv.org/abs/2110.04544

[32] R. Zhang, R. Fang, W. Zhang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling," *arXiv:2111.03930 [cs]*, Nov. 2021, arXiv: 2111.03930. [Online]. Available: http://arxiv.org/abs/2111.03930

[33] M. Wortsman, G. Ilharco, M. Li, J. W. Kim, H. Hajishirzi, A. Farhadi, H. Namkoong, and L. Schmidt, "Robust fine-tuning of zero-shot models," *CoRR*, vol. abs/2109.01903, 2021. [Online]. Available: https://arxiv.org/abs/2109.01903

[34] C. Gurrin, B. . Jónsson, K. Schöffmann, D.-T. Dang-Nguyen, J. Lokoč, M.-T. Tran, W. Hürst, L. Rossetto, and G. Healy, "Introduction to the fourth annual lifelog search challenge, lsc'21," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 690–691.

[35] D.-T. Dang-Nguyen, L. Piras, M. Riegler, G. Boato, L. Zhou, and C. Gurrin, "Overview of ImageCLEF lifelog 2017: lifelog retrieval and summarization," vol. 1866. Dublin: CEUR-WS, Sep. 2017. [Online]. Available: http://ceur-ws.org/Vol-1866/invited_paper_10.pdf

[36] D.-T. Dang-Nguyen, L. Piras, M. Riegler, L. Zhou, M. Lux, and C. Gurrin, "Overview of ImageCLEFlifelog 2018: Daily Living Understanding and Lifelog Moment Retrieval," p. 19.

[37] D.-T. Dang-Nguyen, L. Piras, M. Riegler, L. Zhou, M. Lux, M.-T. Tran, T.-K. Le, V.-T. Ninh, and C. Gurrin, "Overview of ImageCLEFlifelog 2019: Solve My Life Puzzle and Lifelog Moment Retrieval," p. 17.

[38] V.-T. Ninh, T.-K. Le, L. Zhou, L. Piras, and M. Riegler, "Overview of ImageCLEFlifelog 2020: Lifelog Moment Retrieval and Sport Performance Lifelog," p. 17.

[39] C. Gurrin, H. Joho, and F. Hopfgartner, "Overview of NTCIR-12 Lifelog Task," p. 7, 2016.

[40] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, R. Gupta, R. Albatal, and D.-T. Dang-Nguyen, "Overview of NTCIR-13 Lifelog-2 Task," p. 6, 2017.

[41] C. Gurrin, H. Joho, F. Hopfgartner, L. Zhou, V.-T. Ninh, T.-K. Le, R. Albatal, D.-T. Dang-Nguyen, and G. Healy, "Overview of the NTCIR-14 Lifelog-3 Task," p. 13, 2019.

[42] L. Zhou, C. Gurrin, G. Healy, H. Joho, T.-B. Nguyen, R. Albatal, and F. Hopfgartner, "Overview of the ntcir-16 lifelog-4 task," in *Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies*, ser. NTCIR-16, Tokyo, Japan, 2022.

[43] S. Heller, L. Rossetto, L. Sauter, and H. Schuldt, "Vitrivr at the lifelog search challenge 2022," in *Proceedings of the 5th Annual on Lifelog Search Challenge*, ser. LSC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 27–31. [Online]. Available: https://doi.org/10.1145/3512729.3533003

[44] J. Lokoč, F. Mejzlik, P. Veselý, and T. Souček, "Enhanced somhunter for known-item search in lifelog data," in *Proceedings of the 4th Annual on Lifelog Search Challenge*, ser. LSC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 71–73. [Online]. Available: https://doi.org/10.1145/3463948.3469074

[45] G. Kovalčík, V. Škrhak, T. Souček, and J. Lokoč, "VIRET Tool with Advanced Visual Browsing and Feedback," in *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. Dublin Ireland: ACM, Jun. 2020, pp. 63–66. [Online]. Available: https://dl.acm.org/doi/10.1145/3379172.3391725

[46] O. S. Khan, A. Duane, B. T. Jónsson, J. Zahálka, S. Rudinac, and M. Worring, "Exquisitor at the lifelog search challenge 2021: Relationships between semantic classifiers," in *Proceedings of the 4th Annual on Lifelog Search Challenge*, ser. LSC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 3–6. [Online]. Available: https://doi.org/10.1145/3463948.3469255

[47] J. Li, M. Zhang, W. Ma, Y. Liu, and S. Ma, "A Multi-level Interactive Lifelog Search Engine with User Feedback," in *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. Dublin Ireland: ACM, Jun. 2020, pp. 29–35. [Online]. Available: https://dl.acm.org/doi/10.1145/3379172.3391720

[48] A. Duane, B. Thor Jónsson, and C. Gurrin, "VRLE: Lifelog Interaction Prototype in Virtual Reality: Lifelog Search Challenge at ACM ICMR 2020," in *Proceedings of the Third Annual Workshop on Lifelog Search Challenge*. Dublin Ireland: ACM, Jun. 2020, pp. 7–12. [Online]. Available: https://dl.acm.org/doi/10.1145/3379172.3391716

[49] F. Spiess and H. Schuldt, "Multimodal interactive lifelog retrieval with vitrivr-vr," in *Proceedings of the 5th Annual on Lifelog Search Challenge*, ser. LSC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 38–42. [Online]. Available: https://doi.org/10.1145/3512729.3533008

[50] A. Duane and B. T. Jónsson, "Virma: Virtual reality multimedia analytics at lsc 2021," in *Proceedings of the 4th Annual on Lifelog Search Challenge*, ser. LSC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 29–34. [Online]. Available: https://doi.org/10.1145/3463948.3469067

[51] L.-D. Tran, M.-D. Nguyen, B. Nguyen, H. Lee, L. Zhou, and C. Gurrin, "E-myscéal: Embedding-based interactive lifelog retrieval system for lsc'22," in *Proceedings of the 5th Annual on Lifelog Search Challenge*, 2022, pp. 32–37.

[52] T.-N. Nguyen, T.-K. Le, V.-T. Ninh, M.-T. Tran, N. Thanh Binh, G. Healy, A. Caputo, and C. Gurrin, "Lifeseeker 3.0: An interactive lifelog search engine for lsc'21," in *Proceedings of the 4th Annual on Lifelog Search Challenge*, ser. LSC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 41–46. [Online]. Available: https://doi.org/10.1145/3463948.3469065

[53] W.-H. Ang, A.-Z. Yen, T.-T. Chu, H.-H. Huang, and H.-H. Chen, "Lifeconcept: An interactive approach for multimodal lifelog retrieval through concept recommendation," in *Proceedings of the 4th Annual on Lifelog Search Challenge*, ser. LSC '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 47–51. [Online]. Available: https://doi.org/10.1145/3463948.3469070

[54] R. Ribiero, A. Trifan, and A. J. R. Neves, "Memoria: A memory enhancement and moment retrieval application for lsc 2022," in *Proceedings of the 5th Annual on Lifelog Search Challenge*, ser. LSC '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 8–13. [Online]. Available: https://doi.org/10.1145/3512729.3533011

[55] L. Zhou and C. Gurrin, “Multimodal Embedding for Lifelog Retrieval,” in *MultiMedia Modeling*. Cham: Springer International Publishing, 2022, pp. 416–427.

[56] N. Alam, Y. Graham, and C. Gurrin, “Memento: A prototype lifelog search engine for lsc’21,” in *Proceedings of the 4th Annual on Lifelog Search Challenge*, 2021, pp. 53–58.

[57] L.-D. Tran, M.-D. Nguyen, B. Nguyen, H. Lee, L. Zhou, and C. Gurrin, “E-myscéal: Embedding-based interactive lifelog retrieval system for lsc’22,” in *Proceedings of the 5th Annual on Lifelog Search Challenge*, ser. LSC ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 32–37. [Online]. Available: https://doi.org/10.1145/3512729.3533012

[58] T.-N. Nguyen, T.-K. Le, V.-T. Ninh, M.-T. Tran, T. B. Nguyen, G. Healy, S. Smyth, A. Caputo, and C. Gurrin, “Lifeseeker 4.0: An interactive lifelog search engine for lsc’22,” in *Proceedings of the 5th Annual on Lifelog Search Challenge*, ser. LSC ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 14–19. [Online]. Available: https://doi.org/10.1145/3512729.3533014

[59] A. Alateeq, M. Roantree, and C. Gurrin, “Voxento 3.0: A prototype voice-controlled interactive search engine for lifelog,” in *Proceedings of the 5th Annual on Lifelog Search Challenge*, ser. LSC ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 43–47. [Online]. Available: https://doi.org/10.1145/3512729.3533009

[60] N. Alam, A. Alateeq, Y. Graham, M. Roantree, and C. Gurrin, “DCU at the NTCIR16 Lifelog-4 Task,” p. 5, 2022.

[61] L.-D. Tran, T. C. Ho, L. A. Pham, B. Nguyen, C. Gurrin, and L. Zhou, “LLQA - Lifelog Question Answering Dataset,” in *MultiMedia Modeling*. Cham: Springer International Publishing, 2022, pp. 217–228.

[62] C. Gurrin, T.-K. Le, V.-T. Ninh, D.-T. Dang-Nguyen, B. . Jónsson, J. Lokoč, W. Hürst, M.-T. Tran, and K. Schoeffmann, “Introduction to the third annual lifelog search challenge (lsc’20),” in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 2020, pp. 584–585.

[63] C. Gurrin, L. Zhou, G. Healy, B. T. Jonsson, D. T. Dang Nguyen, J. Lokoc, M.-T. Tran, W. Hurst, L. Rossetto, and K. Schoeffmann, “An Introduction to the Fifth Annual Lifelog Search Challenge, LSC’22,” in *ICMR ’22, The 2022 International Conference on Multimedia Retrieval*. Newark, NJ, USA: ACM, 2022.

[64] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR (Poster)*, 2015.

[65] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *International Conference on Learning Representations*, 2018.

[66] ——, “Sgdr: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983*, 2016.