DCU at the NTCIR16 Lifelog-4 Task

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

Ahmed Alateeq*
School Of Computing
Dublin City University
Ireland
ahmed.alateeq2@mail.dcu.ie

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

Mark Roantree
SFI Insight Centre for Data Analytics
Dublin City University
Ireland
mark.roantree@dcu.ie

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

ABSTRACT

In this paper, we present two systems from DCU named DCUMemento and DCUVOX that earlier participated in the 2021 edition of the Lifelog Search Challenge and were redeveloped to participate in the NTCIR-16 Lifelog-4 task. Both systems use image-text embeddings from various CLIP models to build their search backend with DCUVOX using the ViT-B/32 model while DCUMemento uses a weighted ensemble of scores from ViT-L/14 and ResNet-50x64 models. The paper also discusses the query reformulation strategy used by the systems in addition to the system architecture. Finally, we present the results of our evaluation and discuss limitations of both systems with details of improvements planned for future iterations.

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. Lifelogging as a concept is not new and was introduced by Bush [5] in the 1940s and later popularized by Gemmell and Bell [7]. However, the recent fast-paced growth in storage and sensor technology in line with Moore's law has spurred significant interest from the research community owing to potential use cases of lifelogging.

Information retrieval of lifelog data is a challenging problem, given the data is multi-modal with information scattered between images, textual metadata, and other sensor data. The egocentric nature of the images further aggravates the problem as they are at times occluded, blurry and repetitive. Solving specific use cases such as augmenting human memory using lifelogs or using them for memory reminiscence therapy to aid people suffering from neurodegenerative diseases like Alzheimer's depends on how well we can retrieve information from lifelogs. Hence this is an important topic for the research community.

Several challenges such as NTCIR-Lifelog task [14], ImageCLEFlifelog [11] and Lifelog Search Challenge [9] [8] 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-16 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 [14].

In this paper, we present two systems from DCU, DCUMemento and DCUVOX, that participated in the NTCIR-16 Lifelog-4 Automatic LSAT subtask. Both systems leveraged image-text embeddings from various CLIP [13] models to develop their respective search and ranking functionality. DCUVOX's backend was supported by the ViT-B/32 model from CLIP while DCUMemento used a weighted ensemble of scores from ViT-L/14 and ResNet50x64 models to rank 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. Similarly, Section 5 discusses the functionality of DCUVOX, covering the methodology and evaluation results. Finally in Section 6, we conclude our paper and discuss future work in the development of these systems.

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 [14].

- **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

*The two authors contributed equally to this paper

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

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

The subtask contains 48 topics out of which 24 are recall focused requiring as many relevant items as possible to be present in the top-100 images. For example, *"Find examples of when I was in an antiques store"* asks to rank all moments when the person was in an antiques store. 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 examples of when I was having coffee in a cafe with a friend on Saturday mornings"* looking for a specific moment from that person's life.

3 DATASET

The NTCIR-16 Lifelog4 [14] task uses the same dataset from the Lifelog Search Challenge 2021 [8] which consists of:

- **Egocentric Images**: The dataset consists of ~191k images collected from a single lifelogger captured in 2015, 2016 and 2018, spanning 114 days using wearable cameras such as OMG Autographer and Narrative Clip.
- **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 - DCUMEMENTO

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

4.1 DCUMemento Query Reformulation

We reformulated the queries from the NTCIR-16 Lifelog4 task in an easy to process dictionary form to suit the needs of our search engine. The search engine of DCUMemento which leverages image-text embeddings generated from the CLIP [13] model accepts queries in natural language. However, it can only process visual concepts well and fails to comprehend details like time, date, day, month, etc. To mitigate these shortcomings, we manually restructured queries to extract the visual concepts from metadata such as date, time, etc. to allow us to perform a stage-wise search process.

At the initial stage, the search engine ranks the images based solely on visual concepts. It subsequently searches for other specific pieces of information to apply relevant filters over the ranked results and initiate a temporal search if required. For example, we reformulated the query, *"Find examples of when I was waiting at the baggage carousel at an airport after a flight"* to separate the visual concepts and temporal information in the following manner:

- **Search query**: *waiting at the baggage carousel at an airport*
- **Temporal Information**:
  - **Before**: *in an airplane*
  - **After**:

Similarly, we reformulated the query, *"Find examples of when I was having coffee in a cafe with a friend on Saturday mornings"* to separate the visual description from other specific details like day and time,

- **Search query**: *having coffee in a cafe with someone*
- **Filters**:
  - **Day Name**: *Saturday*
  - **Hour**: *<10*

To apply filters easily, we borrowed the enhanced metadata from [1] which has separate columns for day name, hour, and month.

4.2 DCUMemento Search Engine

The search engine of DCUMemento is powered by the CLIP [13] Model from OpenAI and accessed via an API endpoint deployed using the Flask Framework. The model has been trained in a contrastive manner on 400 million image-caption pairs gathered from the internet. This type of large scale pre-training allows it to learn generalized visual concepts which can later be used to solve multiple computer vision downstream tasks such as object recognition, scene recognition, optical character recognition, etc.

Our system, like [2], leverages image-text embeddings from two recently released CLIP models, one using Vision Transformer [6] backbone (ViT-L/14) while the other one using a ResNet-50 [10] backbone (ResNet-50x64) to devise a weighted ensemble approach to rank images. The methodology combines the scores from ViT-L/14 and ResNet-50x64 in a 3:1 ratio to obtain a final score for every image in the corpus which is then used to create the final rankings. The rationale to combine the model scores in a 3:1 ratio comes from the evaluation results presented in [2] which showed significantly superior performance of the ensemble model over the two individual models.

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 as described in Figure 1.
- **Temporal Search**: In the next stage, if the query has a prompt for temporal events (past, future or both), a temporal search is initiated based on the algorithm proposed by [2] which then reranks the images accordingly.
- **Apply Relevant Filters**: Finally, the system applies all filters which are explicitly specified in the query text to generate a final list of top-100 images.

4.3 Evaluation Results of DCUMemento

TRECEVAL [12] was used to evaluate runs for the subtask (LSAT) based on the comma separated files submitted by the participating teams. Table 1 presents an overview of the evaluation results for DCUMemento.

DCUMemento was successful in finding **1201** (40.1 %) relevant images out of the total **2993** relevant images in the entire dataset.

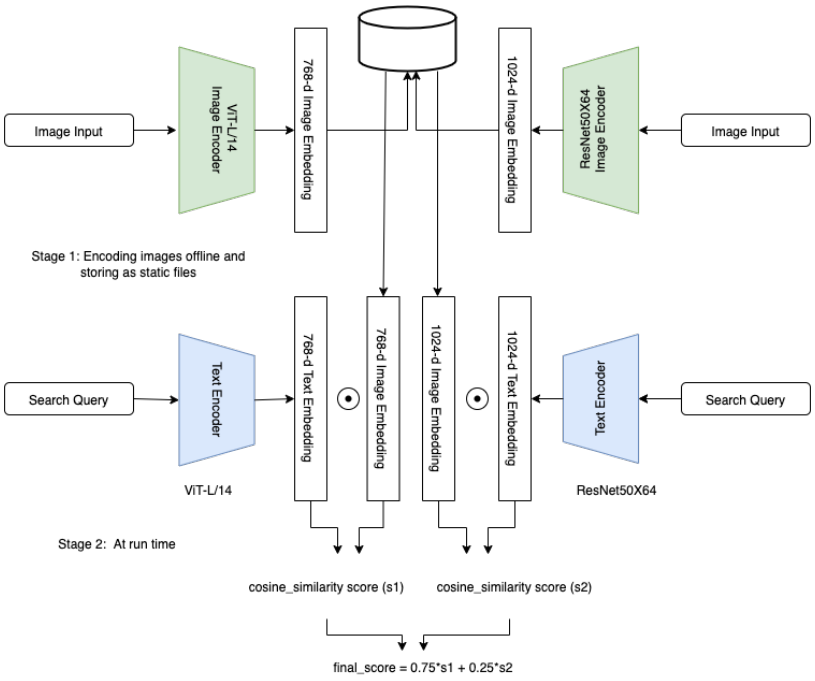

Figure 1: System Architecture of DCUMemento

Total number of topics	48
Total number of retrieved images	4800
Total number of relevant images in the corpus	2993
Total number of retrieved images which were relevant	1201
Mean Average Precision	0.3605
Geometric Mean Average Precision	0.1321
Binary Preference	0.6279
Mean Reciprocal Rank	0.7199

Table 1: DCUMemento - Overview of evaluation results

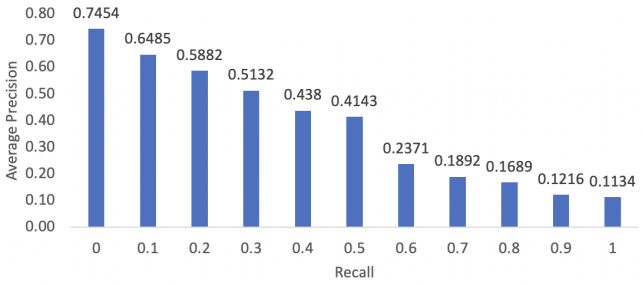

Figure 2: Average Precision at Recall - DCUMemento

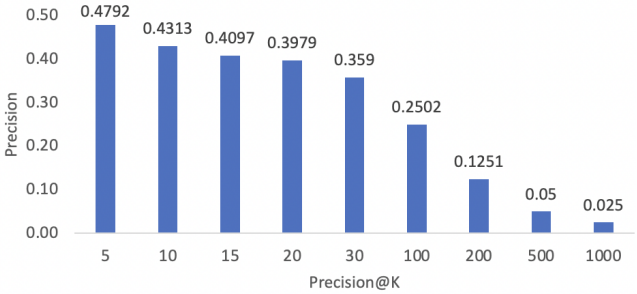

Figure 3: Precision@K - DCUMemento

The mean reciprocal rank (MRR) of the system was **0.7199** which means that on average, the system ranks the 1st relevant image for any given topic either at the 1st or 2nd position while the Mean Average Precision (MAP) was at **0.3605**. MAP unlike MRR is concerned with the ranks of *all* retrieved images instead of ranking only the 1st relevant image, penalizing systems which rank relevant images lower down the order.

We observe a sharp decline in our geometric mean average precision (GMAP) score when compared to MAP as GMAP tends to heavily penalize low-performing topics even if they are few. We suspect that our system might have fared poorly in a few recall-focussed topics as if was not optimized to tackle those ultimately pulling down our GMAP score. Figure 2 plots average precision at specific recall values while Figure 3 shows precision of the system at top-K retrieved results.

5 SYSTEM OVERVIEW - DCUVOX

This version of the DCUVOX system is based on the second iteration of Voxento [3] that participated in the 4th Lifelog Search

Challenge held at ACM ICMR Conference 2021 where it was ranked 4th while competing against 17 other systems. Voxento 2.0 used the backend retrieval API provided by [1] to perform search and ranking of images leveraging image-text embeddings derived from the CLIP [13] model besides providing functionalities such as enhanced metadata and event segmentation.

In this DCUVOX implementation, we attempt to further enhance and improve the metadata as well as the efficiency of our retrieval functionality. This system focuses solely on the automatic task to enable an investigation into how well the CLIP model performs in this setting.

Both DCUVOX and DCUMemento systems include a relatively similar backend in NTCIR-16 with differences in the CLIP model versions to generate image-text features and also in metadata size. A further difference is in the search algorithm implemented in DCUMemento to rank images. Hence, we briefly explain the efforts made in DCUVOX without delving too deeply in details of the backend and the search engine, as both are explained in Section 4 and [1]. Moreover, both system performed query reformulation with a different application of filters.

The first goal was to enhance the metadata as it plays a crucial role in retrieving results accurately and efficiently. The first step was to reduce the number of images as much as possible. The backend API from [1] has enhanced metadata such as the identification of blurred images. The number of blurred images was approximately 40k. We decided to exclude these from the dataset but then, after some testing, we found that some of the blurred images were relevant to certain queries and were thus, not excluded. Indeed, this investigation led to an enhancement in the metadata preparation task in the third version of Voxento [4].

5.1 DCUVOX Query Reformulation

Based on our experience with the Lifelog Search Challenge participation, as experts users, we reformulated the queries from understanding the NTCIR-16 Lifelog4 task's queries in a form that most suits the need of the search engine. The search engine employs the CLIP [13] model that is based on text-image features, and aims to interpret queries expressed in natural language. However, based on multiple testing, we found that summarised queries show better results than detailed queries in similar context. For example, we reformulated this query *"Find examples of when I had keys in my hand and was about to either open or close my front door"* to *"I had keys in my hand and was open or close my front door"*. Although at times shortened or summarised queries can have some grammatical mistakes, the results can still be relevant because the system appears to overcome the lack of grammar. The provided list of queries for the LSAT task provides a high-level description of the event as well a detailed narrative explaining the situation. However, these topics are not direct queries, which gives us the opportunity to reformulate them to work better with the search engine. It is mentioned in Section 4.1 that the CLIP model [13] is good at finding similarities between text and images using only visual concepts while unable to properly identify specific information such as date and place. We hence devised a strategy to exclude the filter words from our initial query and use them to apply relevant filters, such as city name, date, day name and time, during the search process.

5.2 DCUVOX Search Engine

In brief, we derived the image features using the newly released package of the CLIP model [13] as of January 2021. Comparing the recent release with the older one from last year which our system Voxento [3] also used when participating in the LSC 2021, we found the results to be marginally better with the newer release of CLIP. We believe that the difference in package version of libraries like PyTorch or difference in CUDA runtime can impact results. For example with some queries, we observed that relevant images ranked higher when using the newer version. However, it can not be legitimately stated that the newer version will impact all queries in the same manner. Thus, we retained the older CLIP model unlike DCUMemento which used ViT-L/14 and ResNet-50x64 models, whereas we used ViT-B/32 for our current system. It is worth noting that the model ViT-B/32 was also employed in Voxento 2.0 [3] in LSC 2021. An explanation of the workings of the search engine can be referred to in the section 4.2.

The search task was carried out as follows: we initially do query reformulation followed by search based on the visual description leveraging the text-image from the CLIP model. Next, appropriate filters such as place name, date, and time, are applied to the result set. Some frontend features such as image search using event segmentation as well as temporal search are not used for this version of the system. Instead, we will develop these features in interactive runs in the near future.

5.3 Evaluation Results of DCUVOX

The NTCIR organizers employ the TREC evaluation [12] to generate result scores for each run. Relevance judgements for the queries are generated using a pooled approach for up to a maximum of 100 images per topic, per run, and per participant. Table 2 shows an overview of the evaluation results for DCUVOX.

Total number of topics	48
Total number of retrieved images	4800
Total number of relevant images in the corpus	2993
Total number of retrieved images which were relevant	644
Mean Average Precision	0.1267
Geometric Mean Average Precision	0.0050
Binary Preference	0.2941
Mean Reciprocal Rank	0.3544

Table 2: DCUVOX - Overview of the evaluation results

As shown Table 2, the system, DCUVOX, was able to find **644** relevant images out of a total **2993** relevant images in the dataset which represents a precision of (21.5 %). Regarding the ranking of relevant images at first place for any topic, the system has a Mean Reciprocal Rank (MRR) of **0.3544**. The Mean Average Precision (MAP) was at **0.1267** which indicates the average ranking of all relevant images listed in order in the results set. For the geometric mean average precision (GMAP), it has a low score which means that a set of topics did not perform well. The reasons for this may be the due to the absence of some features or the system being unable to solve those tasks in the entirety. Figure 4 illustrates the

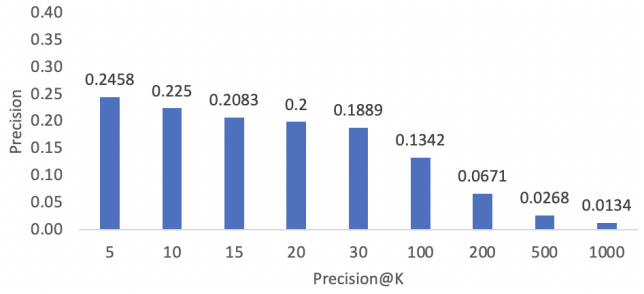

Figure 4: Interpolated Precision at Recall - DCUVOX

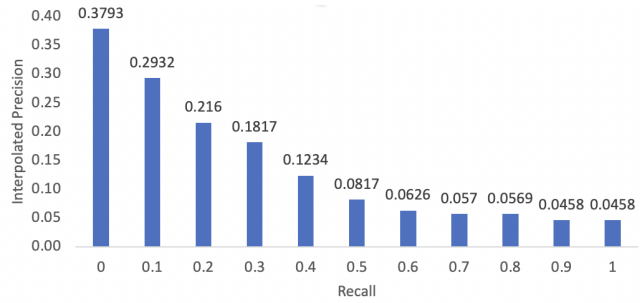

Figure 5: Precision@K - DCUVOX

average interpolated precision at specific recall values while Figure 5 shows precision of the system for the top-k retrieved results.

6 CONCLUSION AND FUTURE WORK

In this paper, we described the DCUMemento and DCUVOX systems which earlier participated in the 2021 edition of the Lifelog Search Challenge [8], both re-developed to participate in the NTCIR-16 Lifelog-4 task in an automated manner. With effective query reformulation and ranking functionality, both DCUMemento and DCUVOX showed competitive performance retrieving about 40% and 22% relevant images respectively despite no human intervention during the run. This shows the robustness of the zero-shot CLIP models when applied to challenging out of domain datasets like Lifelogs.

Neither systems was optimized to solve recall-focused topics which expects only 1 or a few images as output from the system. Currently, the systems output 100 images (maximum allowed) for every topic which leads to poor evaluation scores on metrics like GMAP. Our current research is focused on efficient handling of both precision-focused and recall-focused queries. Furthermore, for DCUVOX it is planned to use larger models from CLIP and compare their performance with the DCUMemento system.

ACKNOWLEDGMENTS

We acknowledge the support of Science Foundation Ireland and the Insight Centre for Data Analytics through the grant number SFI/12/RC/2289-P2 to support the PhD research of the authors Naushad Alam and Ahmed Alateeq. We also acknowledge the support of Ministry of Education in Saudi Arabia for sponsoring the research work of the author Ahmed Alateeq.

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* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3512729.3533006

[3] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2021. Voxento 2.0: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *LSC 2021 - Proceedings of the 4th Annual Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 65–70. https://doi.org/10.1145/3463948.3469071

[4] Ahmed Alateeq, Mark Roantree, and Cathal Gurrin. 2022. Voxento 3.0: A Prototype Voice-controlled Interactive Search Engine for Lifelogs. In *Proceedings of the 5th Annual Lifelog Search Challenge (LSC '22)* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3512729.3533009

[5] Vannevar Bush. 1945. As We May Think. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/ Section: Technology.

[6] 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.

[7] Jim Gemmell, Gordon Bell, and Roger Lueder. 2006. MyLifeBits: a personal database for everything. *Commun. ACM* 49, 1 (Jan. 2006), 88–95. https://doi.org/10.1145/1107458.1107460

[8] 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.

[9] 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.

[10] 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

[11] 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.

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

[13] 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.

[14] 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.