ZJUTCVR Team at ImageCLEFlifelog2019 Lifelog Moment Retrieval Task*

Pengfei Zhou, Cong Bai*, and Jie Xia

College of Computer Science and Technology,
Zhejiang University of Technology, Hangzhou, China
zpf4wp@outlook.com, congbai@zjut.edu.cn, jiexiaXX@outlook.com
*Corresponding author: Cong Bai, congbai@zjut.edu.cn

Abstract. This paper describes our approach to the task of ImageCLEFlifelog 2019. Totally five runs are presented here, which contribute to Lifelog Moment Retrieval (LMRT) task. We use several different methods to provide a flexible retrieval system based on the huge amounts of multi-modal dataset. The work proposes a supervised learning approach with high precision and two exploratory approaches. The first run based on pre-trained Alexnet is the only run we summit through the ImageCLEFlifelog 2019 evaluation system, which reaches the second rank with F1-measure@10=0.44. We then improve our pipeline and get better results of retrieval.

Keywords: Lifelog Moment Retrieval, cross-modal retrieval, Convolutional neural network

1 Introduction

Lifelog is described as a phenomenon whereby people can digitally record their own daily lives in varying amounts of detail, for a variety of purposes [1]. With the rapid development of Internet of things (IOT) and the increasing popularity of sensors and wearable devices that can sense and record biological characteristics [2], data is ready to be captured and combined with more and more personal information in the form of a digital diary. Individuals can now use digital technology to track the details of their daily activities, such as eating, commuting, exercising, working and sleeping. Lifelog data also includes all kinds of data created in daily interactions between individuals and mobile phones and PCs, such as shopping records and music listening records [3].

As part of the ImageCLEF 2019 evaluation campaign [4], The ImageCLEFlifelog2019 task [5] aims to automatically analyze the data in order to categorize, summarize and also to retrieve the information as the users' need.

The task is divided into two subtasks: Solve my life puzzle (Puzzle), Lifelog moment retrieval (LMRT). The main demand of LMRT subtask is to retrieve a number of

CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

specific predefined moments in a lifelogger's normal life. For example, the retrieval result of the query "find my breakfast time" should be images that show the moments of user having breakfast at home in the morning.

And the results of retrieval should not only be relevant, the diversification should also be taken into account. The definition of diversification is that the results of retrieval should cover different proper moments as much as possible. And multiply methods can be used to diversify the retrieval results. To be specific, implementing cluster on textual or visual properties can improve the diversification of the selected moments with respect to the target scenario.

In this paper, we present three approaches to LMRT challenge. Two rely on visual concepts using respectively fine-tuned Googlenet [6] and Alexnet [7]. One relies on both the segmentation and visual concepts with fine-tuned Resnet18 [8]. Related works are discussed in section 2. The proposed methods are described in section 3. In section 4 we analyze the results of our experiments. And we conclude this paper in section 5.

2 Related works

In this section, we briefly discuss recent works on lifelog retrieval. The researchers in this field have proposed different strategies and models to explore inherent law. What's more, the personal devices provided more personal data that can be applied to lifelog retrieval could help users to find the need of multimedia data more efficiently.

Liting et al. [9] proposed retrieval system LIFER, an interactive life record retrieval system developed by ImageCLEFlifelog2018 organizing team [10], in the spirit of the MyLifeBits [11] seminal lifelog database. It provides effective interface with users according to different requirements. The method is to segment dataset based on time and concepts of metadata, and the pipeline is summarized into query, retrieval, filtering and diversification through hierarchical clustering. Minh-Triet et al. [12] proposed a novel method using conception coded feature augmentation to generate text descriptions to exploit further semantics of images; Bernd et al. [13] developed LifeXplore search system, a search and discovery tool that serves Lifelog domain researchers.

Ergina et al. [14] presented a method only based on visual concept using fine-tuned CNN with human-in-the-loop, and get a pretty good performance. The approaches basically consider the tasks as the problem of classification. In their methods, the training procedure are able be simplified with proper preprocessing methods. So, we propose three approaches of preprocessing. What's more, we attempt to present a new method of clustering for classification strategy to improve the performance of diversity.

3 Proposed method

3.1 Overview

The basic pipeline is described in Figure 1. The three approaches follow the same pipeline in earlier stage, but different in the methods of diversification. With the pipeline

and the proper threshold, the output of the system is formed by generating a list of numbered images , which are both relevant and diversity to the query.

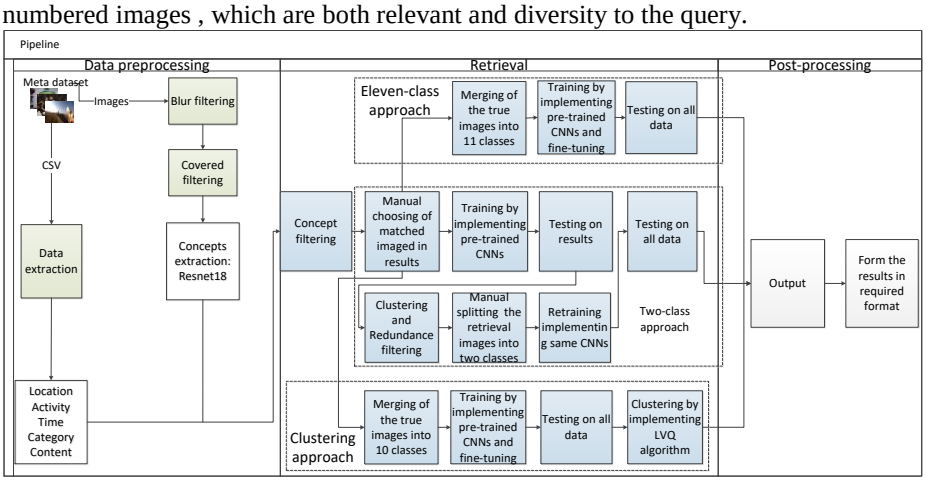

Fig. 1. Proposed pipeline

3.2 Preprocessing

We apply three filtering metrics over the dataset. Two of them filter the blur images, another one filter images that are covered by any objects.

Blur filtering. The first filter is the Laplacian filter (3x3 kernel) with OpenCV implementation to calculate the blur as the variance of convolution result and set the threshold as 30 to avoid misjudging and removing the true positive images [15].

After first filtering, we demand a more refined metric to cut down the amount of images. Then a Fast Fourier Transform is applied to images. Once this step is completed, the average value in the transformed image is obtained and then scaled according to the size of the image to compensate for the tearing effect. The average value is then used for thresholding the image with the larger value representing the focused image and the lower value representing the blurred image.

Covered filtering. To detect if an image is covered by something or facing the ceiling or wall, we use detectors with maximum connected area calculator to calculate the proportion of subjects in the image. Then we remove the images that have a subject's size over 90% of the whole area. The method used is described as follows:

Step 1 Convert the images into grayscale images.
Step 2 Convert grayscale images into binary images.
Step 3 Convert the binary images into matrixes.
Step 4 Find the largest pattern in matrixes and calculate the proportion of it.
Step 5 Remove the images according to the result of matrix calculate.

3.3 The two-class approach

In this approach, we consider each query topic independently to retrieve lifelog image according to the defined topic. For each topic, we get a two-class classifier that can classify the True images and False images after training. The topics defined by organizers are listed in Table 1 [16].

Table 1. The topics of LMRT

Topic ID	Topic Title	Topic Description
001	In a Toyshop	Find the moment when u1 was looking at items in a toyshop
002	Driving home	Find any moment when u1 was driving home from the office
003	Seeking Food in a Fridge	Find the moments when u1 was looking inside a refrigerator at home
004	Watching Football	Find the moments when either u1 or u2 was watching football on the TV
005	Coffee time	Find the moment when u1 was having coffee in a cafe
006	Having breakfast at home	Find the moment when u1 was having breakfast at home.
007	Having coffee with two person	Find the moment when u1 was having coffee with two person.
008	Using smartphone outside	Find the moment when u1 was using smartphone when he was walking or standing outside.
009	Wearing a red plaid shirt	Find the moment when U1 was wearing a red plaid shirt
010	Having a meeting in China	Find all moments when u1 was attending a meeting in China.

The performance of classifiers depends on the preprocessing methods and the training of CNNs. We propose to use supervised learning methods based on Alexnet or Googlenet, for what a pre-trained CNN has a better performance rather than training a CNN from the beginning. The details of the pipeline are described as follows:

1.Preprocessing the dataset as mentioned before. After all the blur filtering and covered filtering, it leaves us with 51.8K true images.
2.Directories choosing and concept filtering. For each query topic, we divide them into directories based on different specific categories. Then we make the corresponding directories for each query by exerting the constraint on the specific topic, and manually select the proper directories by imposing restriction on the concepts. Besides the visual concept, we adapt the time stamp into concept that represent the period of a day and also the binary concept that shows if the music is played in specific moment [17].

The directories we discussed are shown in table 2, the nulls in table mean that the constraint is not considered for avoid misjudging. The topic 009 is an exception corresponding to all images with no constraint.

Table 2. Selected Images per topic.

Topic ID	Location	Activity	Time	Concept
001	Not DCU Not Restaurant Not Home			Not outdoor
002		transport	Not morning	
003	Home			
004	Not Park	Not walking Not transport		TV, unknown, No music
005	Cafe, Unknown	Not transport		
006	Home	Not transport		Morning
007	Not DCU	Not walking Not transport		
008	Not Home	Not transport		Not enclosed area
009				
010	Not DCU Not Home Not Work			enclosed area

Besides, several state-of-art tools is used to filter the irrelevant images and the redundant images. For example, the relevant score is calculated by the concepts based on the referenced directories [18]. We apply VisiPics as a trick to remove the duplicate images for improving the diversification and the generalization performance [19], and also lightening the load. All these tricks are implemented for a better performance.

3.Manual choosing several images as true. We manually select 8 to 20 True images for each queries topic from the corresponding directory. The choosing procedure via the categories that predicted by using the Place CNN [20], and the number of True images depends on the official preprocessing clustering result of meta dataset.

4.Trainning by implementing pre-trained CNN. Pre-trained convolutional neural network Alexnet or Googlenet (trained on ImageNet) is adopted.

5.Testing on the chosen directory by fine-tuned CNN. After training, we use the fine-tuned CNN to test on the corresponding directory. And the results are used as the training dataset for retrained fine-tuning CNN.

6.Splitting the results into two classes by relevance to the query topic. We move the results into two directories by images batching processing (we notice that the true images also have principle of locality). One of the directories is True, then the other is False.

7.Training by implementing the same pre-trained CNN. We use the same trained CNN that we used in step2 as well.

8.Testing on all data. The retrained CNN is applied to all images of entire dataset.

9.Saving the results in required format. The procedure is designed to automatically transfer the results as the collecting into CSV format in MATLAB.

3.4 The eleven-class approach

With the basic pipeline proposed before, we apply the entire pipeline to all ten topics at once. We classify the True images of each topic into 10 classes. And the images which do not belong to these ten classes are classified as False. The 11 classes are 10 True classes of each topic and the only one False class. The pipeline before merging is the same as the former two-class approach, and the procedures after merging are presented below:

1.Trainning by implementing pre-trained CNN. The Alexnet or Googlenet is trained on the eleven classes.
2.Testing on all data. The retrained CNN is applied to all images of entire dataset.
3.Saving the results in the required format. The results are converted into required format automatically.

3.5 The clustering approach

The approach is quite same to the former approaches, the difference is that we propose a procedure of clustering right after the first-round retrieval. In clustering approach, we just merge the True images of each topic into 10 classes in spite of the False class for training. The pipeline before merging is the same as the Two-class approach, and the procedures after merging are presented below:

1.Trainning by implementing pre-trained CNN. The Alexnet or Googlenet is trained on the ten classes.
2.Testing on all data. The retrained CNN is applied to all images of entire dataset.
3.Clustering by implementing LVQ algorithm. In this process of work, learning vector quantization (LVQ) algorithm is used for clustering. The main steps of algorithm include: initializing the prototype vector; iterative optimization, update the prototype vector. The details of algorithm are introduced below (A set of prototype vectors is initialized by randomly selecting a sample labeled t_q from the q cluster firstly):

Algorithm 1 LVQ algorithm

Repeat
  Pick a random sample
  Calculate the Euclidean distance from the sample to each prototype vector
  Find the shortest distance
  If $y_j = t_i$(same class)
    $p' = p_j + a\,(x_j - p_j)$ (reduce the distance so that p become closer to the sample point)
  Else
    $p' = p_j - a\,(x_j - p_j)$ (increase the distance so that p become farther from the sample point)
  $p_j = p'$(update)
Until the condition to end is met.

4.Saving the results in the required format. The results are converted into required format automatically.

4 Experiment

4.1 Results discussion

Due to time limit and other force majeure factors, we just submit one run follow basic two-class approach during the competition, however, we send our new results to the organizers after the deadline of submission and get an evaluation of our whole three retrieval approaches. The only run that we submitted during the competition follows the pipeline described in Section 3.

To evaluate the performance, the metrics used are F1@X (X=10), which measure the harmonic mean between Precision at X (P@X) and Cluster recall at X (CR@X), with X representing the top X results are taken into consideration in evaluation. So the diversity (via CR@10) and relevance (via P@10) are both taken into account. The results of our runs are displayed in Table3. The runs with * are submitted after the competition.

Table 3. Result on ImageCLEFlifelog 2019 - LMRT challenge.

Run ID	Description	P@10	CR@10	F1@10
Run 1	Two-class approach with Alexnet	0.71	0.380	**0.440**
Run 2*	Two-class approach with Googlenet	0.74	0.342	0.428
Run 3*	Eleven-class approach with Alexnet	0.41	0.305	0.330
Run 4*	Eleven-class approach with Googlenet	0.48	0.348	0.363
Run 5*	Clustering approach	0.59	0.498	0.481

The run 1 (with the details in table 4) is the only run we submit through the evaluation system, while the run 5 has the best performance in diversification (CR@X), which is one of the most notable performance in lifelog moment retrieval. The details of last 4 runs are shown in Figure 2. Compared with run 3 and run 4, run 2 and run 5 have better performance. Which apparently shows that the eleven-class approach needs to be improved in precision.

From the charts we summarize that the query 8(Using smartphone outside) is difficult to retrieve, however, when we manually verify the retrieval results in query 8, we find that most of retrieval results are related to phone using and outdoor. So, we consider whether the problem is about the definition of outside? Using smartphone in the yards probably is not affirmed as true. Besides, the vital signs devices that included in dataset are noise for smartphone retrieval.

In the proposed pipeline, the manual operation is taken into consideration, so the results are fluctuating in a certain range. And to summarize from the results, the method of clustering is a significant influence factor, a better method of clustering improves the result rapidly in the performance in diversity, and also can improve the performance of relevance within limits.

Table 4. Detail results for run1

RUN1	P@5	CR@ 5	F1@ 5	P@10	CR@ 10	F1@ 10	P@50	CR@ 50	F1@ 50
Query 1	0.8	0.5	0.615	0.9	1	0.947	0.24	1	0.387
Query 2	1	0.095	0.174	1	0.095	0.174	1	0.143	0.25
Query 3	1	0.167	0.286	0.7	0.278	0.398	0.56	0.778	0.651
Query 4	0.8	0.25	0.381	0.7	0.25	0.368	0.72	0.5	0.590
Query 5	0.8	0.333	0.471	0.8	0.333	0.471	0.74	0.333	0.46
Query 6	0.8	0.222	0.348	0.8	0.222	0.348	0.68	0.444	0.538
Query 7	1	1	1	1	1	1	0.54	1	0.701
Query 8	0	0	0	0	0	0	0.08	0.333	0.129
Query 9	1	0.286	0.444	1	0.286	0.444	1	0.286	0.444
Query 10	0	0	0	0.2	0.333	0.25	0.52	1	0.684

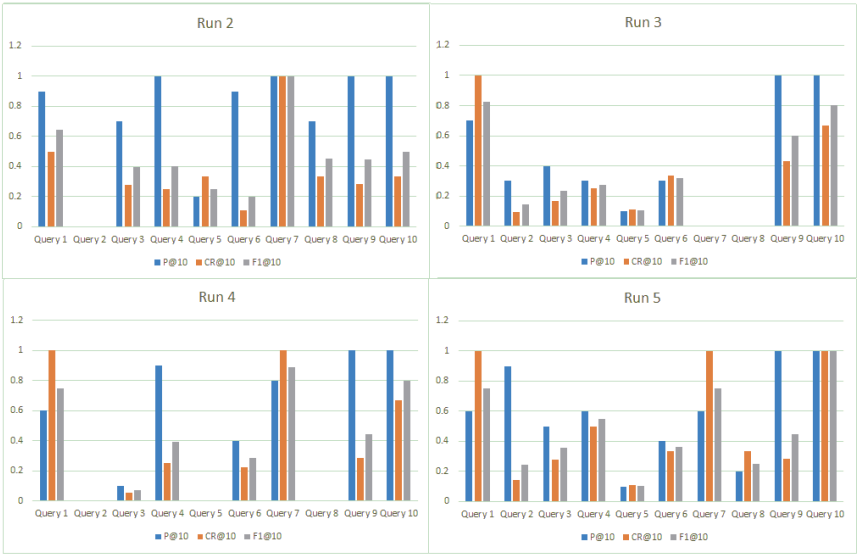

Fig. 3. Detail results for last 4 runs

4.2 Resources

Our approach is implemented using Intel(R) Xeon(R) CPU @3.50Ghz with 16G RAM. We work on Ubuntu16.04 using MATLAB 2019a. We use Neural Network Toolbox with GPU coder which generates CUDA from MATLAB code for deep learning. The OpenCV and PyTorch are also used.

5 Conclusion

This paper presents our proposal for lifelog retrieval system on Lifelog moment retrieval (LMRT)subtask, which is corresponding to different types of queries, such as location, time, activity, and additional biometric data. As for procedure of preprocessing, we present three different methods of preprocessing to adapt the meta dataset to dataset which is appropriate in quantity and quality. During the retrieval process, we implement three different methods to compare the relevant and diversified performance of retrieval results.

Experimental results show that our proposal is of high precision and reactiveness in official ranking metric F1-measure, while the relevant score is far better (P@10) than the diversified score (CR@10). Therefore, the ability of diversifying the results could be improved further, which helps to achieve a comprehensive and complete view of the query. The clustering method of our system is able to be updated to improve the performance of entire system.

As for future work, we will improve the pipeline of our proposal using the natural language processing approach such as RNN and LSTM to automatically match the query topics with visual concept. What's more, we will develop a user-friendly graphics interface for our proposal.

Acknowledgement:

This work is supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LY18F020032 and Natural Science Foundation of China under Grant No. 61502424.

References

1. C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. Foundations and Trends in Information Retrieval, 8(1):1–125. (2014)
2. Duc Tien Dang Nguyen, LitingZhou, Rashmi Gupta, Michael Riegler, andCathalGurrin. Building a Disclosed Lifelog Dataset: Challenges, Principles and Processes. 1-6. 10.1145/3095713.3095736. (2017)
3. Aiden R. Doherty, Alan F. Smeaton, Keansub Lee, and Daniel P.W. Ellis, Multimodal Segmentation of Lifelog Data Centre for Digital Video Proessing & Adaptive Information Cluster, Dublin City University, Ireland LabROSA, Columbia University, New York, USA In: Procs. Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (RIAO '07), pp. 21–38 Paris, France (2007)
4. Bogdan Ionescu, Henning Müller, Renaud Péteri, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Dzmitri Klimuk, Aleh Tarasau, Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Mathias Lux, Cathal Gurrin, Obioma Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, Narciso Garcia, Ergina Kavallieratou, Carlos Roberto del Blanco, Carlos Cuevas Rodríguez, Nikos Vasillopoulos, Konstantinos Karampidis, Jon Chamberlain, Adrian Clark, Antonio Campello, ImageCLEF 2019: Multimedia Retrieval in

Medicine, Lifelogging, Security and Nature In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), Lugano, Switzerland, LNCS Lecture Notes in Computer Science, Springer (September 9-12 2019)
5. Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Minh-Triet Tran, Liting Zhou, Mathias Lux, Tu-Khiem Le, Van-Tu Ninh and Cathal Gurrin. 2019. Overview of ImageCLEFlifelog 2019: Solve my life puzzle and Lifelog Moment Retrieval. In: CLEF 2019 Working Notes. CEUR Workshop Proceedings (CEUR- WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2380/. Lugano, Switzerland. (2019)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. pp. 1097–1105. NIPS'12, Curran Associates Inc., USA (2012)
7. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Computer Vision and Pattern Recognition (CVPR) (2015)
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778 (2016)
9. Liting Zhou, Luca Piras, Michael Riegler, Mathias Lux, Duc-Tien Dang-Nguyen, and CathalGurrin. An Interactive Lifelog Retrieval System for Activities of Daily Living Understanding. In: CLEF (2018)
10. Liting Zhou, Zaher Hinbarji, Duc-Tien Dang-Nguyen, CathalGurrin. LIFER: An Interactive Lifelog Retrieval System. 9-14. 10.1145/3210539.3210542. In: ICMR (2018)
11. J. Gemmell, A. Aris, and R. Lueder. Telling Stories with Mylifebits. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, pages 1536–1539. (2005)
12. Minh-Triet Tran, Tung Dinh-Duy, Thanh-Dat Truong, Viet-Khoa Vo-Ho, Quốc Lương, and Vinh-Tiep Nguyen. Lifelog Moment Retrieval with Visual Concept Fusion and Text-based Query Expansion. In: CLEF (2018).
13. Bernd Münzer, Andreas Leibetseder, Sabrina Kletz, Manfred Primus and Klaus Schoeffmann. lifeXplore at the Lifelog Search Challenge 2018. 3-8. 10.1145/3210539.3210541. (2018)
14. Ergina Kavallieratou, Carlos R. del-Blanco, Carlos Cuevas and Narciso García. Retrieving Events in Life Logging. In: CLEF (2018)
15. Renting Liu, Zhaorong Li and Jiaya Jia. Image partial blur detection and classification. IEEE International Conference on Computer Vision Pattern Recognition. 1-8. 10.1109/CVPR.2008.4587465. (2008)
16. Liting Zhou, Luca Piras, Michael Riegler, Giulia Boato, Duc-Tien Dang-Nguyen, and CathalGurrin. Organizer Team at ImageCLEFlifelog 2017: Baseline Approaches for Lifelog Retrieval and Summarization. In: CLEF (2017)
17. Tsun-Hsien Tang, Min-Huan Fu, Hen-Hsen Huang, Kuan-Ta Chen, and Hsin-Hsi Chen. Visual Concept Selection with Textual Knowledge for Understanding Activities of Daily Living and Life Moment Retrieval. In: CLEF (2018)
18. Ruth Fong, Andrea Vedaldi. Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks. (2018)
19. Arora R., Trelogan J., Ba T.N. Using High Performance Computing for Detecting Duplicate, Similar and Related Images in a Large Data Collection. In: Arora R. (eds) Conquering Big Data with High Performance Computing. Springer, Cham. (2016)

20. B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. A 10 million Image Database for Scene Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2017)