NLP301 at the NTCIR-15 Micro-activity Retrieval Task: Incorporating Region of Interest Features into Supervised Encoder

Tai-Te Chu
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
tdjhu@nlg.csie.ntu.edu.tw

Yi-Ting Liu
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
ytliou@nlg.csie.ntu.edu.tw

Chia-Chung Chang
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
ccchang@nlg.csie.ntu.edu.tw

An-Zi Yen
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
azyen@nlg.csie.ntu.edu.tw

Hen-Hsen Huang
Department of Computer Science, National Chengchi University, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
hhhuang@nccu.edu.tw

Hsin-Hsi Chen
Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan
MOST Joint Research Center for AI Technology and All Vista Healthcare, Taipei, Taiwan
hhchen@ntu.edu.tw

ABSTRACT

This paper presents our approach to the NTCIR-15 MART task. The task is divided into two subtasks, micro-activity retrieval task (MART) and insights task. We participated in micro-activity retrieval task, where the goal is to retrieve the detailed lifelog of activities from multi-modal data, such as first-person perspective images, screenshots, and bio signals. The major challenge in the task is the semantic gap between textual descriptions and the visual concepts in images. To reduce the semantic gap, we propose a supervised model that encodes activity description as vectors. Our model incorporates the visual features and bio signal features from seven users to classify the pre-defined twenty micro-activities. In order to recognize computer activities (e.g., reading text on screen and browsing news websites), we utilize RoI (Region of Interest) features. After encoding visual features by employing the pre-trained computer vision model, we use Gated Recurrent Unit (GRU) [4] to capture the slight variation of users' movement in time-series. Experimental results show that our system is effective in the micro-activity retrieval task. In terms of performance, our system achieved 0.85050 of MAP score and won the third place.

TEAM NAME

NLP301

SUBTASKS

Retrieval task (English)

1 INTRODUCTION

With the advance of technology, people are used to recording their daily life by taking photos or filming videos. These personalized data captured by lifelogging devices provide rich information for supporting human living assistance services, such as lifestyle understanding [5], diet monitoring [14], and visual lifelog retrieval [3, 7]. Since the devices can capture a large amount of multimedia data, we need an efficient system to categorize the activities and access the desired information. Activity retrieval, which is aimed at identifying real-world activities of an individual (e.g., using a computer, eating, having a conversation with others, etc.), has become a popular task for users who record their life through digital devices.

In general, activity retrieval focuses on retrieving coarse-grained activities from longer, untrimmed videos [1, 2, 6, 10], such as using a computer and doing housework. Comparing with general activity retrieval, micro-activity retrieval task (MART) in NTCIR-15 [12] targets at retrieving fine-grained activities from images in short time period, such as writing/replying to an e-mail and cleaning with a broom, hoover or cloth. MART focuses on finding the patterns in different activities and reducing the semantic gap between micro-activity description and the visual representation. NTCIR-15 MART dataset consists of multimodal information, including images, screenshots, body/eye movement, and object detection information. Specifically, the queries in NTCIR-15 MART dataset are natural language descriptions for finding consecutive images over a period of time corresponding to the micro-activity described in the query. Fig 1 shows an example query and its answer.

To identify the activity occurs over short time-scales, we propose a model that takes the content of the images in the time-series into account. We utilize the residual neural network (ResNet) [11] to extract visual features from the images taken in a period. Note that people describe their questions with textual expressions, while the personalized data captured by the camera are visual data. That is, the main challenge of MART is to reduce the semantic gap between the visual and textual domains. In this work, we tackle the challenging issue of reducing the semantic gap between textual information in screenshot images and activity description by cropping the useful part, URL (Uniform Resource Locator), for example, of screenshot images. We encode the cropped screenshot images by using ResNet. Then, the features are fed into the gate recurrent unit (GRU) layer to derive the sequential features. The detail of model construction is described in Section 3.3.

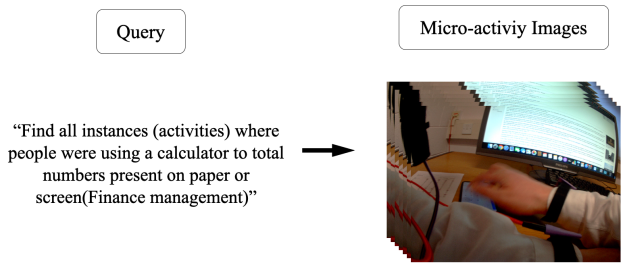

Figure 1: Example of the Query and Its Answer in the MART Dataset.

The contributions of this paper are threefold as follows.

(1) We create new RoI features from screenshot images to help our model better categorize distinct computer micro-activities.
(2) We construct a user lookup table and add additional information in the final stage of our model to personalize the seven users' actions.
(3) The GRU structure encodes the time-series information in consecutive images and improves performances in the micro-activity retrieval task.

2 RELATED WORK

Activity retrieval can be subdivided into two types. One is personal activity retrieval, and the other is general activity retrieval. Recently, there has been significant work in general activity retrieval. One major challenge in activity retrieval is to retrieve the corresponding activity images/videos by natural language queries. Hence, how to build an efficient system to assist human to retrieve the information they want is essential. Overall, an activity retrieval process could be classified into two parts: multi-modal learning and retrieval model construction.

Multi-modal Learning Yang et al. [16] propose a correlational recurrent neural network (RNN) to fuse multiple input modalities that are inherently temporal in nature. Inspired by Yang et al. [16], we construct a model to extract temporal features for recognizing activity in consecutive images. Zhang et al. [17] discuss the pros and cons of several multi-modal learning methods. According to the characteristic of the dataset, we create a more reasonable feature called RoI feature to help our model know the patterns of various activities. Besides, we learn the time series feature in images by using the RNN structure.

Activity Retrieval Models In recent years, activity retrieval has been discussed in many novel ways. Zhang et al. [18] propose a novel task setting called zero-shot temporal activity detection, whose goal is to detect the activities never seen in training data. Gao et al. [8] align language query to video clips using multi-processed interaction between visual representation and sentence representation.

Compared to general activity retrieval, personal activity retrieval is a special research topic in lifelogging. Gurrin et al. [9] propose a lifelog task in 2014, where the goal is to extract and retrieve specific moments in a lifelogger's life. NTCIR-15 MART, different from the tasks mentioned above, is a novel task that addresses the research topic of retrieving personal micro-activities.

Following multi-modal method mentioned above, we add additional information called bio signal features. That is, the bio signals are used to construct a lookup table of seven users to detect their behaviors.

3 RETRIEVAL TASK

3.1 Data Analysis

The statistics of the twenty activities numbered from 1 to 20 in the training data are shown in Table 1. Each activity is collected from seven different users.

Table 1: Activities Distribution of the Seven Users.

Activity	Frequency
1. Writing/replying to an e-mail	97
2. Reading text on screen	108
3. Editing a presentation	100
4. Zoning out while staring at the screen	115
5. Finance management (using a calculator)	88
6. Physical precision task (using a circuit board)	85
7. Document organisation task	78
8. Reading text on an printed A4 sheets	105
9. Counting/arranging physical currency (money)	76
10. Writing with pen on paper	94
11. Watching a youtube video	111
12. Browsing news websites	106
13. A face-to-face conversation	80
14. Making a telephone call	97
15. Drinking/eating	85
16. Closing eyes	110
17. Cleaning (with a broom/hoover)	78
18. Physical exercise	72
19. Hand-eye corrdination activity	76
20. Walking around	69

From the statistics of the training data, we find that the distribution of each activity in the dataset is balanced. We adopt the idea of supervised learning multi-modal RNN as mentioned in the work of Yang et al. [16], the detailed model structure will be elaborated in Section 3.3.

3.2 Data Preprocessing

The NTCIR-15 MART dataset includes photos taken with Autographer, images of computer screenshot, and metadata of each activity performed by seven users. Each user had performed two round of 20 activities as training data, and one round of 20 activities as test data. Each training instance we input to the model is a pack of images labeled with the same ID in MART, representing an activity performed by an individual.

For extracting visual features from images, we use a pre-trained computer vision model, ResNet. The visual features of screenshot

images are also extracted by ResNet. For computer-related micro-activities, we crop the URL of screenshot images into several RoI regions. We then encode them using CNN model as RoI features.

The bio-signal data are built with seven users' vectors $\boldsymbol{x}^{u,a} \in \mathbb{R}^{108}$, where u is user id, a is activity id and the number 108 represents 108 sorts of bio signals. In total, $N = 7$ represents the number of users and $M = 20$ stands for the amount of micro-activities. To remove the bias of different bio signal features, we normalize them using the equations as follows:

$$\boldsymbol{f}^{u,a} = \frac{\boldsymbol{x}^{u,a} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \tag{1}$$

$$\boldsymbol{\mu} = \frac{1}{NM} \sum_{u=1}^{N} \sum_{a=1}^{M} \boldsymbol{x}^{u,a} \tag{2}$$

$$\boldsymbol{\sigma} = \sqrt{\frac{1}{NM} \sum_{u=1}^{N} \sum_{a=1}^{M} (\boldsymbol{x}^{u,a} - \boldsymbol{\mu})^2} \tag{3}$$

3.3 Proposed Model

In this section, we introduce our model, which is capable of retrieving micro-activity from multi-modal data. Our model is composed of ResNet and GRU. ResNet is utilized to extract visual features from images. Each image is encoded as 1,000-dimension vectors by the ResNet. For better understanding the context in which the activity occurs, we connect the output from the ResNet with the GRU layer. We perform mean pooling at each timestamp output from the GRU layer to encode the features of the time series images. Fig 2 shows the model structure.

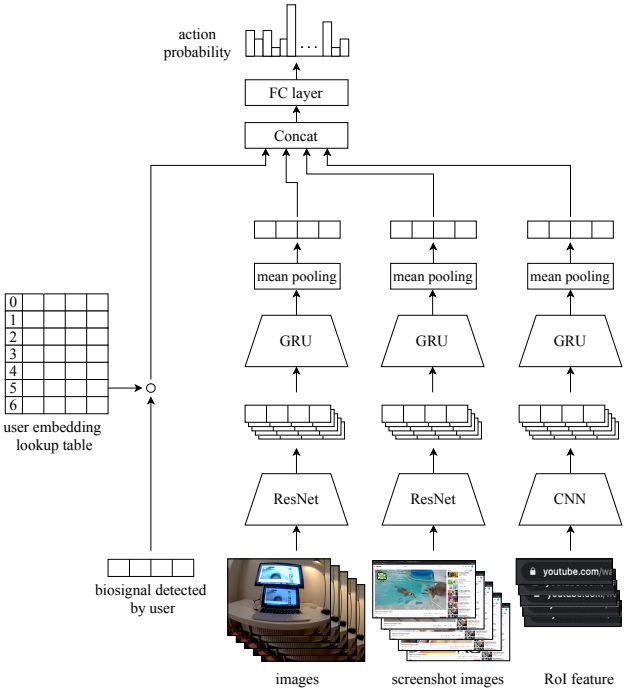

Figure 2: Structure of Our Proposed Model.

As a result of observation on screenshot images, we discover some textual information that could be used to recognize activities related to "using computer". For instance, the textual information of describing the activities of "reading text on screen" and "writing/replying an e-mail" are different. Therefore, we crop the screenshot images with positions that may contain the important textual information, such as URL. The experimental results show that the RoI features could improve model performance. In addition, we also incorporate the bio signal features for distinguishing dynamic activities from static ones. The bio-signals are multiplied with 7 trainable dimensional user embedding from user embedding lookup table to better recognize the micro-activities. Experimental results of using different features are shown and discussed in the following section.

Finally, we concatenate all the aforementioned features and fed them into a fully connected layer to compute the probabilities of the 20 activities. The model will select the activity with the highest probability as the answer.

To sum up, the whole process contains three steps as follows:

(1) Preprocess the input images, screenshot images, and bio signals and pack them.
(2) Feed the preprocessed images and bio signals into the model.
(3) Use softmax function to determine to which activity it belongs.

3.4 Experiment Settings

To optimize our model, we apply Adam optimizer [13] on the cross-entropy loss with a learning rate of 1×10^{-4}. The batch size is 4, and the number of training epochs is at most 75. We implement the model using PyTorch [15].

3.5 Official Results

We submitted 10 runs to the retrieval task, including different ways to incorporate the features with our baseline model. The best model is shown in Fig 2. The symbol 'Share" in Table 2 denotes the images, screenshot images and RoI features are encoded by using the same ResNet. Excluding the runs for evaluating the performances of different epochs , we report the 6 settings in Table 2.

Table 2: Model Settings for each Run of Retrieval Task.

Run ID	Model	Features				
		Image	Screenshot	Crop	User	Share
1	Baseline	✓	✓	✓	✓	✓
2, 3	Our Model	✓		✓		
4		✓		✓	✓	
5		✓	✓	✓		
6		✓	✓	✓	✓	
7		✓	✓		✓	✓
8, 9, 10		✓	✓	✓	✓	✓

4 EXPERIMENT

We further compare our best model (i.e. the one with the run id 10) with the baseline model composed of ResNet only. Both of them are trained with all features. Table 3 shows the performance of

the baseline model and our best model. We observe that the mean average precision (MAP) increases when the GRU layer is adopted. Comparing the baseline model that uses the ResNet only, our model obtains an improvement of MAP score by 0.20059. Experimental results show that the GRU layer is effective in the task of micro-activity retrieval.

Table 3: Performances of the Baseline and the Best Model.

Model	MAP score
Run ID 1	0.64991
Run ID 10	0.85050

Table 4: Performances of Our Proposed Model.

Run ID	Features					MAP Score
	Image	Screenshot	Crop	User	Share	
3	✓		✓			0.72088
4	✓		✓	✓		0.73806
5	✓	✓	✓			0.75513
6	✓	✓	✓	✓		0.78897
7	✓	✓		✓	✓	0.81305
10	✓	✓	✓	✓	✓	0.85050

Table 4 reports the results of our proposed model under different feature settings. Comparing the results of the run id 5 and the run id 6, adding the feature of user embedding significantly improves the performance. The models with the run id 6 and the run id 10 achieve the MAP scores of 0.77898 and 0.85050, respectively. It shows that sharing the parameter of ResNet benefits to learning more robust visual representations.

5 DISCUSSION

In the task of MART, we find that our model can successfully recognize some micro-activities that even humans cannot. For instance, as shown in Fig 3, the images of “zoning out while starring at the screen”, “closing eyes”, and “making a telephone call” are similar.

Figure 3: Examples of Three Micro-Activities with Similar Scenes.

In the confusion matrix of the baseline model shown in Fig 4, the numbers of labels "True Label" and "Predicted Label" are referred to the numbers in Table 1. The activity of “zoning out while starring at the screen” may be predicted as “closing eyes”. We observe that the values of “data_EOG_UD_by_activity_median” provided in the MART dataset are significantly different between the activity of

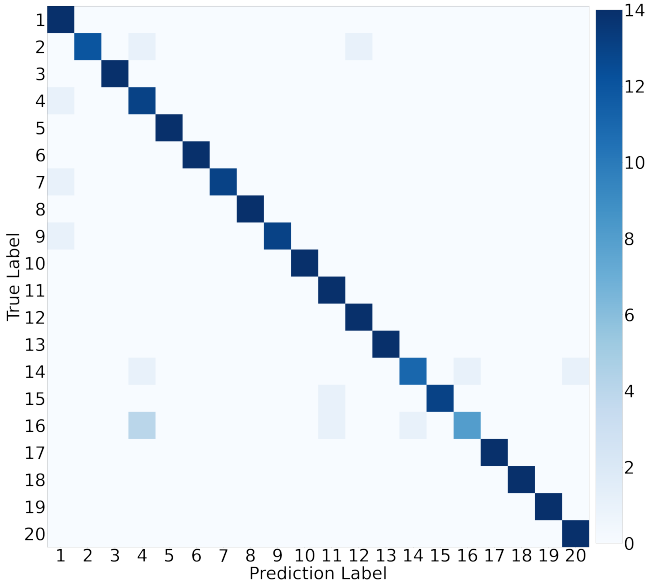

Figure 4: The Confusion Matrix of the Baseline Model.

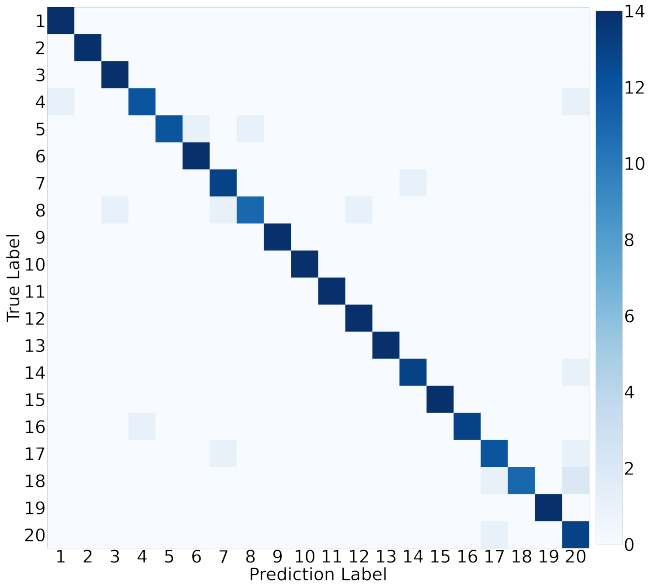

Figure 5: The Confusion Matrix of the Best Model.

“zoning out while starring at the screen” and “closing eyes”. Besides, the activity of “making a telephone call” is similar to the activities of "zoning out while starring at the screen" and "closing eyes", since the images of these three activities only contain the scene in front of the user. That is, the information of user’s bio signal in a period is important.

On the other hand, some activities are recognized correctly by our best model that considers the content of consecutive images. Comparing with Fig 4, Fig 5 shows our best model that introduces

the GRU layer recognizes the micro-activities more precisely, especially "closing eyes". Fig 6 shows the images of the activity of "walking around" at three timestamps, where *n* denotes the time interval. In this case, the structure of GRU captures the trajectory of the user's movement, and improves the performance of micro-activity recognition.

Figure 6: Example of Continuous Scenes of an User Walking Around.

6 CONCLUSION AND FUTURE WORK

This paper presents our work in NTCIR-15 MART task. To better recognize the micro-activity that occurs in the context of consecutive images, we propose a model composed of ResNet and GRU. We find that RoI features improve the performance of classifying the micro-activities about using computer, such as reading text on screen and browsing news websites. In addition, we incorporate the information of bio signals in our model for classifying the static and dynamic activities. Surprisingly, the information of bio signals also benefits to distinguish the static activities zoning out while staring at the screen and closing eyes. Experimental results show that encoding visual features extracted from the ResNet by using the GRU layer improves the performance. Our model is capable of recognizing the similar micro-activities by capturing the user's slightly different movement in the short time-scales, such as walking around.

From this task, we figure that different users have their own habits of doing some micro-activities. In the future, we may construct personal knowledge base with micro-activities of each user.

7 ACKNOWLEDGMENTS

This research was partially supported by Ministry of Science and Technology, Taiwan, under grants MOST-109-2634-F-002-040-, MOST-109-2634-F-002-034-, MOST 109-2218-E-009-014-, and by Academia Sinica, Taiwan, under grant AS-TP-107-M05.

REFERENCES

[1] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. 2018. Action Search: Spotting Actions in Videos and Its Application to Temporal Action Localization. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX (Lecture Notes in Computer Science)*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11213. Springer, 253–269. https://doi.org/10.1007/978-3-030-01240-3_16

[2] Yunlong Bian, Chuang Gan, Xiao Liu, Fu Li, Xiang Long, Yandong Li, Heng Qi, Jie Zhou, Shilei Wen, and Yuanqing Lin. 2017. Revisiting the Effectiveness of Off-the-shelf Temporal Modeling Approaches for Large-scale Video Classification. *CoRR* abs/1708.03805 (2017). arXiv:1708.03805 http://arxiv.org/abs/1708.03805

[3] Tzu-Hsuan Chu, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Image Recall on Image-Text Intertwined Lifelogs. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. IEEE, 398–402.

[4] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR* abs/1412.3555 (2014). arXiv:1412.3555 http://arxiv.org/abs/1412.3555

[5] Aiden R Doherty, Niamh Caprani, Vaiva Kalnikaite, Cathal Gurrin, Alan F Smeaton, Noel E O'Connor, et al. 2011. Passively recognising human activities through lifelogging. *Computers in Human Behavior* 27, 5 (2011), 1948–1958.

[6] Min-Huan Fu, Chia-Chun Chang, Hen-Hsen Huang, and Hsin-Hsi Chen. 2019. Incorporating external textual knowledge for life event recognition and retrieval. In *Proceedings of the 14th NTCIR Conference on Evaluation of Information Access Technologies*. 61–71.

[7] Min-Huan Fu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Incorporating Semantic Knowledge for Visual Lifelog Activity Recognition. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. 450–456.

[8] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: Temporal Activity Localization via Language Query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 5277–5285. https://doi.org/10.1109/ICCV.2017.563

[9] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Found. Trends Inf. Retr.* 8, 1 (2014), 1–125. https://doi.org/10.1561/1500000033

[10] Meera Hahn, Asim Kadav, James M. Rehg, and Hans Peter Graf. 2019. Tripping through time: Efficient Localization of Activities in Videos. *CoRR* abs/1904.09936 (2019). arXiv:1904.09936 http://arxiv.org/abs/1904.09936

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *CoRR* abs/1512.03385 (2015). arXiv:1512.03385 http://arxiv.org/abs/1512.03385

[12] Graham Healy, Tu-Khiem Le, Hideo Joho, Frank Hopfgartner, and Cathal Gurrin. 2020. Overview of NTCIR-15 MART. In *NTCIR-15: Evaluation of Information Access Technologies*.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Takuya Maekawa. 2013. A sensor device for automatic food lifelogging that is embedded in home ceiling light: A preliminary investigation. In *2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops*. IEEE, 405–407.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. (2017).

[16] Xitong Yang, Palghat Ramesh, Radha Chitta, Sriganesh Madhvanath, Edgar A. Bernal, and Jiebo Luo. 2017. Deep Multimodal Representation Learning from Temporal Data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 5066–5074. https://doi.org/10.1109/CVPR.2017.538

[17] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal Intelligence: Representation Learning, Information Fusion, and Applications. *IEEE J. Sel. Top. Signal Process.* 14, 3 (2020), 478–493. https://doi.org/10.1109/JSTSP.2020.2987728

[18] Lingling Zhang, Xiaojun Chang, Jun Liu, Minnan Luo, Sen Wang, Zongyuan Ge, and Alexander G. Hauptmann. 2020. ZSTAD: Zero-Shot Temporal Activity Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. IEEE, 876–885. https://doi.org/10.1109/CVPR42600.2020.00096