TMU19 at the NTCIR-15 Micro-activity Retrieval Task

Duy-Duc Le Nguyen[†]
Graduate Institute of Data Science
Taipei Medical University
Taipei, Taiwan
m946108007@tmu.edu.tw

Yu-Chi Lang
Graduate Institute of Data Science
Taipei Medical University
Taipei, Taiwan
m946108001@tmu.edu.tw

Yung-Chun Chang
Graduate Institute of Data Science
Taipei Medical University
Taipei, Taiwan
changyc@tmu.edu.tw

ABSTRACT

This paper proposes a new method for predicting user activities at the NTCIR-15 Micro-activity Retrieval Task. Additional concepts from ResNet generated features following with a Bidirectional Long-Short Term Memory block helps our neural network paying more attention in the corresponding class. Our model received an upright result on the scoreboard.

TEAM NAME

TMU19

SUBTASKS

Retrieval Task

KEYWORDS

Rich Multi-modal Data, Feature Extraction, Activity Recognition

1 INTRODUCTION

The **N**ii **T**estbed and **C**ommunity for **I**nformation access **R**esearch (NTCIR) project has promoted research efforts for enhancing Information Access (IA) technologies such as Information Retrieval (IR), Text Summarization, Information Extraction (IE), and Question Answering (QA) techniques since 1997. At the end of 2019, the NTCIR-15 was announced to focus on two topics on IA technology.

MART (Micro-activity Retrieval Task) [3] is a pilot task in the NTCIR-15 workshop data challenge task. The NTCIR15-MART aims to motivate the development of the first generation of techniques for high-precision micro-activity detection and retrieval of daily living micro-activities to help identify and retrieve activities that occur over short time-scales, such as minutes, rather than the long-duration event segmentation tasks of the past work. The provided data was gathered by equipping the individual to gather a detailed lifelong of activities as an individual progresses through a protocol of real-world activities (e.g., using a computer, solving a problem, eating, daydreaming, etc.). The chosen sensors captured a lifelong camera data stream, bio-signal activity (EOG, HR), and computer accesses to record interaction with digital artifacts. This task presents a new challenging set of micro-activity-based topics.

We live in a multi-modal data world — we hear sounds, smell scents, touch surfaces, see things, and taste flavors. A query will be defined as a multi-modal problem if it covers multiple modalities. As this research area is relatively fresh, we have decided to step into the NTCIR-15 Micro-activity Retrieval Task to find a novel method. This project can bring lots of practical application to the real world. In a clinic environment, AI applications will diagnose the patient's state and activity based on their biological data, breathing rate, voice & images. Hence, doctors and nurses can promptly meddle in case of emergencies.

2 RELATED WORK

With artificial intelligence development, various relevant information of the human body can be obtained through wearable devices, and human body activity recognition can be done using this information. In recent years, research related to distinguishing human activity has become more popular. A proposed method [1] to process the original data through feature processing and then use machine learning models to identify human activities. In this research, the author has integrated different types of data suitable for feature extraction methods, such as time-related data suitable for the use of mean, standard deviation, and variance. For frequency-related data, Fourier Transform (FT) and Discrete Cosine Transform (DCT) are suitable. Other data types can also be extracted using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). In addition to feature selection, this research also proposes many machine learning models such as Decision Trees, Naive Bayes, and Support Vector Machines (SVM). However, in most human activity recognition tasks, machine learning algorithms rely heavily on manual feature extraction techniques, requiring human body-related knowledge. A deep learning model can learn more meaningful features through its neural network and significantly reduce the workload required for feature extraction. Jindong Wang et al. [9] introduced why deep learning is used for human activity recognition and integrated the research using these deep learning models.

Andrey Ignatov [6] proposed to use Convolutional Neural Networks (CNN) to solve the problem of human activity recognition. The advantage of using CNN is that it almost does not require additional data pre-processing and feature extraction steps, and it can save a lot of computing time compared to the previous model CNN for human activity recognition. According to the author's experimental results, the CNN-based model is significantly better than other models such as random forest, KNN.

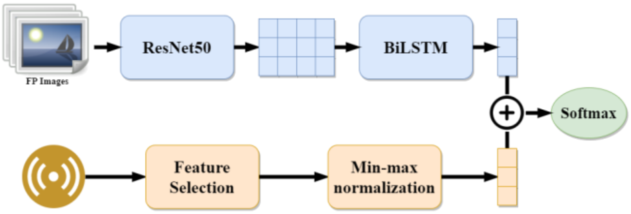

Figure 1: Our model architecture. FP images are first-person perspective images

3 METHODOLOGY

Given an input that insists on a set of first-person perspective images and rich time-aligned multi-modal sensor data, our model in Figure 1 will classify it into the most related activity among 20 predefined ones.

Image Feature Extraction: We extract m features from each image. There are various methods that generate crucial insights from images such as DenseNet [5], AlexNet [8], and ResNet [2]. We chose ResNet50 to be equivalent to provided ResNet probability outputs from the organizer. These features are stacked into a $m \times n$ matrix. Since the model has to deal with limited labeled data, all of the parameters in the Image Feature Extraction block are locked and cannot be learnable.

Bidirectional Long Short-Term Memory (BiLSTM): Two independent LSTM [4] layers putting together to form a BiLSTM. This structure allows the model to have both backward and forward information about time-series data and tackle the vanishing gradient problem from traditional recurrent neural network architectures. Using bidirectional will run an input in two ways, one from past to future and one from future to past. The differs this approach from unidirectional is that BiLSTM can learn how and when to forget and when not to use gates. As a result, the model can understand the context better. We set the hidden size as m to match with the previous block. After processing through the BiLSTM block, its output with 2m scalar-value will be concatenated with 3,000-length rich multi-modal data.

Rich Multi-modal Data: During 90 seconds of an activity, a summary - which includes median, min, max, average, standard error length from sensors (108 scalar-value) - was generated. Following up with sensor data is 3000 scalar-value from a pre-trained ResNet50 with maximum, minimum, and average probabilities (1000 categories). Since there are only 280 training data, we primarily select crucial features that maximum likelihood reflects its activity. Besides, each feature needs to be normalized. Normalization ensures that the network can learn regards all input features to a similar extent. Considering that the units of the data are inconsistent, the numbers in each column are very different. Before putting the data into the model, we normalize the data. The original data with different units and large differences in number size can be converted into data that falls in a specific interval and has no unit through normalization and standardization. There are two standard methods as below, namely Z-Score Standardization and Min-Max Normalization. Z-Score Stan-dardization uses the formula Equation 1 to convert data with an average μ and a standard deviation σ into a standard normal distribution (the average is 0 and the standard deviation is 1).

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Min-Max Normalization is to use the formula in Equation 2 in order to scale the data to a specific interval where the maximum value is one and the minimum value is 0.|

$$z_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} \tag{2}$$

In this MART task, we choose Min-Max Normalization to convert the original column data. Scaling data with different original units and large differences in number size to the interval with the maximum value of 1 and the minimum value of 0 through Min-Max Normalization has a significant impact on the model's performance.

As the Retrieval Task is a multi-class classification. The cross-entropy loss was obtained, although the submission result needs to be ranked in probability. Given a 20-scalar value output y with the correct class c, the loss function is defined as below:

$$loss_{y,c} = \log \frac{\exp(y_c)}{\sum_i \exp(y_i)}$$

4 EXPERIMENT & RESULT

4.1 Dataset

Rich multi-modal data will be made available from 7 experimental participants that have performed 20 different activities (with three repetitions). This data has been pre-processed to facilitate participating organizations to develop interactive/automatic retrieval systems. Raw data for corresponding periods is available upon request. Since each experimental participant will complete 20 different activities per session (60 per participant), the dataset comprises data for 420 activities in total. 280 of these activities will be released with labels as a training set (upon which participants can build an automatic/interactive retrieval system). The data for the remaining 140 activities are released as a testing set (without labels).

The participating organizations' performance will be ranked on the mAP (mean average precision) score. It is expected that many promising approaches that may perform sub-optimally to others in terms of mAP alone may offer other advantages in terms of speed, insight generation, interpretability, and cross-subject applicability.

4.2 Experiment Setting

As ResNet's output m is 2048 features, we set the hidden size of BiLSTM block matching with its input dimension $L = m = 2048$. Each entity in the dataset does not have the same number of images. Therefore, we initialize $n = 16$, the largest number of images in the dataset. If an entity has less than 16 images, it will be padded with a set of zero-value images. Before feeding to ResNet block, all images are resized to 256×256, then center cropped and normalized with $\mu = (0.485, 0.456, 0.406)$, $\sigma =$

$(0.229, 0.224, 0.225)$. This step was done by *transform* function from *torchvision*[1] package. We obtain Adam optimizer [7] with the weight decay factor. After trying several learning rates, the optimal learning rate with high slope and lowest training loss is $2e-5$. We try to use three methods commonly used in linear regression models to select independent variables for the raw data. The three methods are forward selection, backward selection, and stepwise selection. Forward selection is to select significant independent variables into the model until all independent variables that significantly impact the model are selected. Backward selection is the opposite of forwarding selection. It deletes the independent variables that have no significant impact on the model one by one until all the insignificant independent variables are deleted. Stepwise selection combines forward selection and backward selection. It considers each variable's influence on the model and selects independent variables that influence the model. After comparing these three methods, only fifteen independent variables are remaining. We observed an unpredictable scenario that the duration of sensor data (89 or 90 seconds) while the sensor's min, max, or average value does not correlate with its activity. According to the model's performance, we found that deleting the variable will make the performance decrease.

We assume that the model's decreasing performance might lack critical information for the model to learn and the limited labeled data during training thoroughly. Therefore, all of the features have remained. Seven traditional machine learning methods are acquired to be baseline models: Naive Bayer, k-Nearest Neighbor, Random Forest, Decision Tree, Support Vector Machine, Logistic Regression, and XGBoost. The top three performances are shown in Section 4.3.

Table 1: Model's performance in mAP score

Methods	Leave-One-Out	10-folds	Scoreboard
Random Forest	0.181	0.568	-
SVM	0.297	0.377	-
XGBoost	0.395	0.625	0.399
Our model	**0.540**	**0.638**	**0.465**

4.3 Result

In both leave-one-out and 10-fold cross-validation, our model performance in Table 1 grants the first place, significantly outperforming all comparisons within leave-one-out. Our proposed method in extracting additional features from time-series first-person perspective images with recurrent neural networks helps boost prediction's precision. Within the baseline models, XGBoost easily beats other traditional machine learning methods as it is the most evolutionary in decision-tree-based ensemble algorithms. Not only does XGBoost achieve a high mAP score, but also it is faster 2 to 5 times than the rest comparisons. Therefore, our proposed model and XGBoost are used to submit to the Retrieval Task scoreboard system. Our model's submission returns a 0.465 score in the organizer evaluation system, while the XGBoost result is almost 0.4.

4.4 Discussion

In summary, our approach in separately splitting difference modality data into two neural networks can achieve a satisfactory mean average precision score in the NTCIR-15 Micro-activity Retrieval Task. There is much room to optimize within our approach. Hence, we will keep configuring this model in the future.

ACKNOWLEDGMENTS

This research was supported by the Ministry of Science and Technology of Taiwan under grant MOST 107-2410-H-038-017-MY3, MOST 109-2410-H-038 -012 -MY2, and MOST 107-2634-F-001-005

REFERENCES

[1] Boris Ginsburg, Patrice Castonguay, Oleksii Hrinchuk, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, Huyen Nguyen, Yang Zhang, and Jonathan M. Cohen. 2019. Stochastic gradient methods with layer-wise adaptive moments for training of deep networks. *arXiv preprint arXiv:1905.11286* (2019).

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *arXiv:1512.03385 [cs]* (December 2015). Retrieved November 1, 2020 from http://arxiv.org/abs/1512.03385

[3] Graham Healy, Tu-Khiem Le, Hideo Joho, Frank Hopfgartner, and Cathal Gurrin. Overview of NTCIR-15 MART. In: Proceedings of the NTCIR-15 Conference, Tokyo, Japan (2020).

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (November 1997), 1735–1780. DOI:https://doi.org/10.1162/neco.1997.9.8.1735

[5] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2018. Densely Connected Convolutional Networks. *arXiv:1608.06993 [cs]* (January 2018). Retrieved November 1, 2020 from http://arxiv.org/abs/1608.06993

[6] Andrey Ignatov. 2018. Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Applied Soft Computing* 62, (2018), 915–922.

[7] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (January 2017). Retrieved November 1, 2020 from http://arxiv.org/abs/1412.6980

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90. DOI:https://doi.org/10.1145/3065386

[9] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119, (2019), 3–11.

[1] https://pytorch.org/docs/stable/torchvision/