Toward Projection Learning between Sensor Data and Semantic Word Vector for Zero-shot Learning

Author
An affiliation
A country
a@email

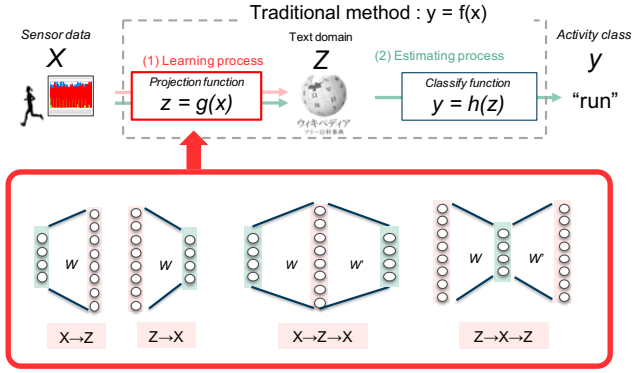

Fig. 1. The top of figure indicate overview of Zero-shot learning. The traditional method learn from a sensor data X to an activity class Y directly. The idea of Zero-shot learning project a sensor data X to text domain Z and then classfy for recognition activity class Y. In this paper, we focus on the projection function$z = g(x)$.

***Abstract*—In this paper, we compare 4 learning projection models between sensor domain and text domain for Zero-shot learning(ZSL). In traditional activity recognition with sensor data, the task of collecting training dataset is too tough and costly to apply for social. Our challenge is making the task efficient. The Zero-shot learning's purpose is to recognize the unknown activity which is activity class out of training dataset. In our previous research, we propose the Zero-shot learning method using the word vectors made from Wikipedia courpus for recognizing the human living activities like breakfast, watching TV etc. We found that this method success to recognize unknown activities and need to improve the projection function for performance. In this paper, we construct 4 learning models for projection and evaluate them with accelerometer sensor data annotated simple activities. As the result, we realize that (1) the learning method with twice projection is usuful for performance.(2) The pattern is difficult to identify that distance of unknown activity vector are closer than the distance between unknown activity and known activity.**

***Index Terms*—feature learning, activity recognition, accelerometer**

I. Introduction

Human activity recognition with sensor data is important factor in context awareness such as watching human living, helping working and care elders field [1]. Almost method use an supervised machine learning. It learns how to recognize activity class y from sensor data x and then we can represent $y = f(x)$ as estimation function. In this traditional method, the task of collecting training dataset, which we need to use the method is too tough and costly to apply for social. So our challenge is making the task efficient. For that we focus on the Zero-shot learning. The Zero-shot learning's purpose is to recognize the unknown activity which is activity class out of training dataset. The point of the way to recognize the unknown activity classes is using the word vector z describing the activity class y. They have vectors describing both of unknown activities and known activities. In the basic Zero-shot learning, firstly it learn how to project the sensor samples to the word vector z which is annotated same activity class [5]. By doing so, we can translate new sensor data to the word domain. And then the word vector projected searches for the closest word vector z which is annotated activity class. If the word vector estimate nearest to unknown activity word vector, it can recognize the unknown activity class.

[5] proposed using an attribute as word vector domain for basic Zero-shot learning method for human activity recognition. However the method take time and works. To solve the problems, we proposed using an semantic vectors instead of the attribute vecotors [3]. We found 2 points on the method. The first one is that it is possible to recognize unknown activities even using the semantic vectors. The second one is that the method need to improve projection function for good performance [3].

In this paper, we construct 4 learning models for projection and evaluate them with accelerometer sensor data annotated simple activities. The difference between them is the direction of projection and the number of projections. The direction of projection means whthere it projects the sample from feature domain X to word domain Z or from Z to X. The number of projection means that how many it projects for learning. We constract 4 projection models which $X \rightarrow Z$, $Z \rightarrow X$, $X \rightarrow Z \rightarrow X$ and $Z \rightarrow X \rightarrow Z$. And then we evaluate the 4 methods with dataset that has accerelometer labeled simple activities. As the result, we realize that (1)the learning method with twice projection is usuful for performance.(2)The pattern is difficult to identify that distance of unknown activity vector are closer than the distance between unknown activity and known activity.

II. Related Work

The 3 categories for learning projecting models are proposed for image recognition.

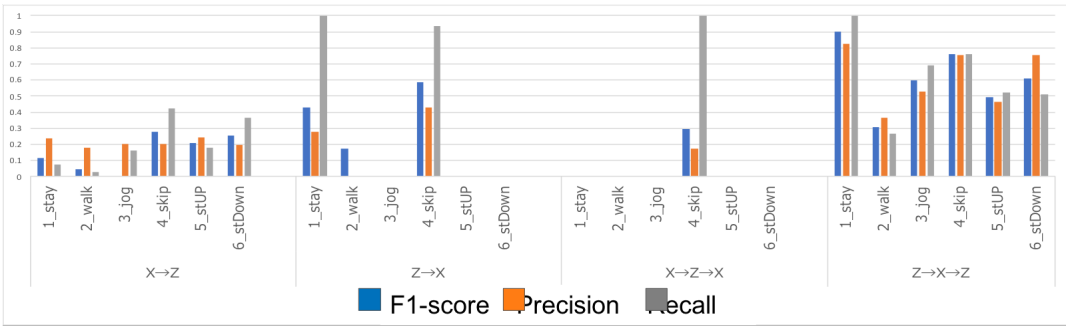

Fig. 2. The accuracy of each model in case 1 which is in the case normal situation.

(1) The first is to project feature data to text domain [6]. (2) The second is to project word vector sample to sensor domain [6]. (3) The third is to learn both derection [2]. And then we use only encoder function in estimate process. This method project sensor sample to word domain and then project to feature domain again. And we can construct almost same method but opposite projection. We called first category $X \rightarrow Z$ model, second category $Z \rightarrow X$ model and third category $X \rightarrow Z \rightarrow X$ model. And we call another third category $Z \rightarrow X \rightarrow Z$. We compare these 4 learning models and analyze them.

III. Evaluation

We evaluate the methods and analyze the accuracy of them. For evauation, we prepare two cases; (case 1) in the case of without unknown activities, and (case 2) in the case of with 2 unknown activities which are not overlap with known activities.

1) Sensor Dataset: We use the sensor dataset we collected. This dataset is accerelometer data labeled simple activities, "stay", "walk", "skip", "jog", "stair up" and "stair down". We make subjects wear a smartphone on thier left arm. A subject keep the activity for 20 seconds per a one series and it repeats 5 times. And then we extract the feature vector by using the slide time window method.

2) Word Vector: We use the word vectors from English Wikipedia by using the word2vec [4].

IV. Result

A figure2 indicate the rate of the number of collect samples out of the each activities for each models. We can see that $X \rightarrow Z$ has the low accuracy for each activities. The $Z \rightarrow X$ and $X \rightarrow Z \rightarrow X$ have the 0% accuracy for some activities. However, the $X \rightarrow Z \rightarrow X$ has accuracy for all activities. This indicate that the estimate projection works well for performance. So and then we evauate the $X \rightarrow Z \rightarrow X$ in the case there are 2 unknown activities.

A tableI indicate the rate of the number of collect samples out of the each activities for $Z \rightarrow X \rightarrow Z$ with 2 unknown activity classes. We can see from the table that there are 3 good accuracy in that case, when unknown activity classes are "stay" and "stair up" "stay" and "stair down" and "stay" and "jog".

TABLE I
THE TABLE INDICATE THE ACCURACY FOR $Z \rightarrow X \rightarrow Z$ PROJECTION MODEL IN CASE 2. THE ACTIVITY INDICATE UNKNOWN ACTIVITY CLASSES.

	stay	skip	jog	walk	stair up
skip	50				
jog	**100**	54.10			
walk	49.66	50.03	55.55		
stair up	**93.51**	59.09	55.59	50.03	
stair down	**98.21**	40.41	55.55	50	50.26

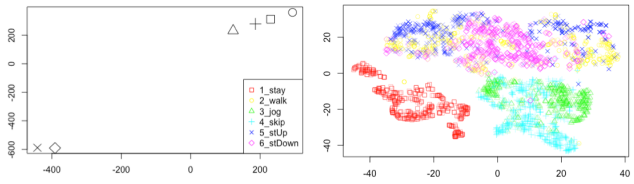

Fig. 3. The visualizing both domain by using the t-SNE. The left figure is the word domain and the right figure is the sensor domain. The sharp of the points indicates the each activities.

V. Discussion

According to figure2, $X \rightarrow Z$ model recognize activities randomly because almost same accuracy for all activities. $Z \rightarrow X$ and $X \rightarrow Z \rightarrow X$ model occur the hubness problem because they recognize almost samples to one class. And we analyze the trend unknown activities which it can recognize from figureI and tableI.

figure3.

When we focus on the "stay" activity class, this method could not recognize when combination of "walk" and "skip". As same time, when we focus on the word domain in the figure 3, we can see that combination of the unknown activities are closer than distance to known activity points. This means that the pattern is difficult to identify that distance of unknown activity vector are closer than the distance between unknown activity and known activity.

As the future work, we will try to search how to construct the word vectors for recognition unknown activity classes with the considering this second result.

VI. Conclusion

We compare 4 learning projection models between sensor domain and text domain for Zero-shot learning to improve the accuracy of recognizing unknown activities. As the result, we realize that (1)the learning method with twice projection is usuful for performance.(2)The pattern is difficult to identify that distance of unknown activity vector are closer than the distance between unknown activity and known activity. As the future work, we will try to search how to construct the word vectors for recognition unknown activity classes with the considering this second result.

References

[1] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):33, 2014.

[2] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4447–4456. IEEE, 2017.

[3] Moe Matsuki and Sozo Inoue. Recognizing unknown activities using semantic word vectors and twitter timestamps. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, pages 823–830. ACM, 2016.

[4] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[5] Minh-Tien Nguyen, Quang-Thuy Ha, Thi-Dung Nguyen, Tri-Thanh Nguyen, and Le-Minh Nguyen. Recognizing textual entailment in vietnamese text: An experimental study. In *Knowledge and Systems Engineering (KSE), 2015 Seventh International Conference on*, pages 108–113. IEEE, 2015.

[6] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.