Human Activity Recognition Using Semi-supervised Multi-modal DEC for Instagram Data

Dongmin Kim(✉), Sumin Han, Heesuk Son, and Dongman Lee

School of Computing, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea
{dmkim25,suminhan,heesuk.son,dlee}@kaist.ac.kr

Abstract. *Human Activity Recognition (HAR)* using social media provides a solid basis for a variety of context-aware applications. Existing HAR approaches have adopted supervised machine learning algorithms using texts and their meta-data such as time, venue, and keywords. However, their recognition accuracy may decrease when applied to image-sharing social media where users mostly describe their daily activities and thoughts using both texts and images. In this paper, we propose a semi-supervised multi-modal deep embedding clustering method to recognize human activities on Instagram. Our proposed method learns multi-modal feature representations by alternating a supervised learning phase and an unsupervised learning phase. By utilizing a large number of unlabeled data, it learns a more generalized feature distribution for each HAR class and avoids overfitting to limited labeled data. Evaluation results show that leveraging multi-modality and unlabeled data is effective for HAR and our method outperforms existing approaches.

Keywords: Human activity recognition · Social media · Multi-modal · Deep learning · Deep embedded clustering · Semi-supervised learning

1 Introduction

Social media has been an ambient data platform on which people share their activities of daily living and memorable experience. Using the embedded behavioral patterns in the shared data, *Human Activity Recognition (HAR)* provides a solid basis for a variety of applications such as context-aware recommendation systems and health-care services [9–11]. To achieve better HAR performance using popular image-sharing social media such as *Instagram* and *Yelp*, various machine learning models have been presented [1–3]. Since these social media allow users to put their geolocation features when uploading the posts, human activities extracted from them have high potential to enhance awareness of real world dynamics.

Existing HAR approaches [1–3] leverage supervised machine learning models (e.g., SVM, LSTM) which take texts and meta-data of social media posts for

H. W. Lauw et al. (Eds.): PAKDD 2020, LNAI 12084, pp. 869–880, 2020.
https://doi.org/10.1007/978-3-030-47426-3_67

learning important patterns with respect to human activities. However, its use of uni-modal textual features cannot capture enough patterns of human activities shared on the social media because users mostly describe their daily activities and thoughts using both texts and images. Such a limitation can be relieved by incorporating the inherent multi-modality of social media into the learning process as in [4,5]. These multi-modal approaches adopt an early fusion technique which leverages concatenated features of text and image to their proposed classifiers. However, none of them has investigated applicability to HAR and it is not trivial to construct a labeled dataset which is large enough to train their multi-modal supervised methods; to the best of our knowledge, no such dataset has been published yet for multi-modal HAR using social media.

In this paper, we present a semi-supervised method for HAR using multi-modal Instagram data, that is, both image and text, which achieves a high recognition accuracy while using only a limited amount of labeled data. On social media, the number of unlabeled data exponentially increases every day while only a few labeled data for a specific task exists. In such a domain, *semi-supervised learning* methods which leverage a small amount of labeled data and a much larger set of unlabeled data together are effective alternatives to improve learning performance [13]. To devise a semi-supervised learning method for HAR, we adopt the state-of-the-art clustering method, *MultiDEC* [7], which can learn deep embedded feature representations of social media image and text, respectively and extend it into a semi-supervised model which incorporates a small portion of labeled data into its training procedure. The proposed method minimizes both cross-entropy loss and Kullback-Leibler divergence loss. This enables the proposed model to learn more generalized feature distribution by leveraging the feature distribution of a large unlabeled dataset while optimizing the learning results to the labeled features. Evaluation results show that our method achieves 71.58% of recognition accuracy which outperforms the best accuracy of the existing HAR methods (maximum 64.15%).

2 Related Work

2.1 Human Activity Recognition Using Social Media

In general, previous HAR methods using social media incorporate text features with metadata such as timestamps, venues, and keywords into their supervised machine learning models. Zack et al. [1,2] leverage linear SVM for their human activity classifier and train the model using the text features from *Twitter* and *Instagram* data which they have collected and labeled using crowd-sourcing. Besides, Gong et al. [3] leverage *Yelp* data for HAR: They split each caption of a Yelp post into word tokens and create a sequence of embedded Word2Vec features. In addition to the extracted features, they use keyword dictionary knowledge embedding and temporal information encoding (e.g., date and week) for training a Long Short-Term Memory (LSTM) model to classify Yelp posts to human activity classes based on Yelp taxonomy[1].

[1] https://www.yelp.com/developers/documentation/v3/all_category_list.

To improve the recognition performance of the existing methods, multi-modal features (i.e., image and caption) of social media posts can be leveraged; there have been several approaches published already. Roy et al. [4] concatenate image features from a CNN model and text features from a Doc2Vec model and use them together to train a fully connected neural networks to identify social media posts related to illicit drugs. Schifanella et al. [5] use visual semantics from a CNN model and text features from an NLP network model together to train traditional models, SVM and DNN, respectively. After that, they leverage the trained models to detect sarcastic social media posts.

2.2 Deep Embedding Clustering and Semi-supervised Learning

Deep embedded clustering (DEC) [6] is an unsupervised method which simultaneously learns optimal feature representations and cluster assignments of unlabeled data by iteratively optimizing clustering objective. During the training process, DEC minimizes the KL divergence between the cluster soft assignment probability and the proposed target distribution so that its embedding network can learn feature representation. MultiDEC [7] is an extended version of DEC to deal with multi-modal data such as image-caption pairs. While DEC has a single embedding network, MultiDEC is composed of two embedding networks which are jointly trained to simultaneously learn image and text representations having similar distributions.

While DEC and MultiDEC are powerful methods, they cannot be directly used for HAR because they are clustering algorithms. To leverage their proven effectiveness for dealing with classification tasks, a semi-supervised DEC, *SSLDEC* [8], has been recently presented. SSLDEC learns the target distribution of labeled data, iteratively estimates the class probability distribution of unlabeled data by measuring its feature distances from the clusters of labeled data and optimizes them during training. Thus, SSLDEC is a transductive learning method that retains recognition performance even with a relatively small amount of labeled dataset. However, the supervision mechanism of SSLDEC may fail to maximize the HAR performance when the labeled data cannot represent the target distribution correctly or its distribution cannot accommodate that of unlabeled data. Especially, when applied to social media data where such characteristics are evident, the applicability of SSLDEC may drastically decrease.

3 Proposed Method: Semi-supervised Multi-modal DEC

3.1 Overview

The proposed method aims to predict human activity classes of multi-modal Instagram posts while training the optimal embedding network. More specifically, models in the proposed method should be well trained by means of a large number of unlabeled data where a limited number of labeled data is available.

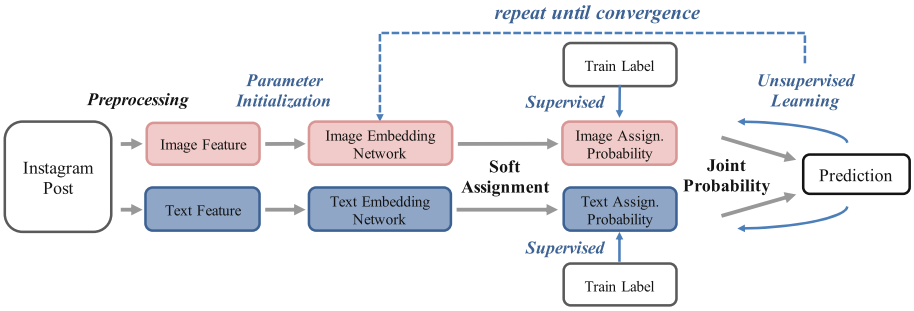
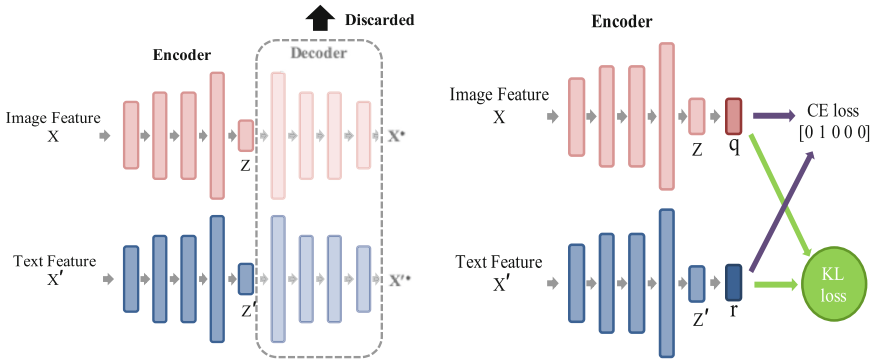


Fig. 1. An overview of our proposed semi-supervised multi-modal DEC

Figure 1 illustrates an overview of our proposed semi-supervised multi-modal DEC method to fulfill the key requirement. When multi-modal Instagram posts are given as a training dataset, our method preprocesses them to extract image and text features. Then, two embedding networks with parameters θ and θ' are initialized for learning deep representations of the image and text features, respectively. These networks are intended to embed image data, X, and text data, X', into the corresponding latent spaces, Z and Z'. Our method trains the embedding networks by alternating a supervised learning phase and an unsupervised learning phase. In the supervised learning phase, labeled multi-modal data is utilized for learning a class assignment. In the unsupervised learning phase, we leverage rich unlabeled data for computing cluster assignment probabilities for both features. After that, we compare them to the target joint probability distribution for adjusting cluster centroids, μ and μ', and eventually improving the cluster purity. This semi-supervised method helps us to learn the optimal representations of image and text features and apply Multi-modal DEC to HAR.

3.2 Multi-modal Data Pre-processing

Instagram data consists of image-text pairs where text is given as a mixture of a caption and multiple hashtags. For image preprocessing, we extract the 2048-dimensional feature representations of ResNet-50 [17] pretrained on ImageNet dataset [18]. We use this embedding method for image data because these features are known to be effective in image clustering as well as classification [19]. We compress the extracted 2048-dimensional features once again into 300-dimensions using Principal Component Analysis (PCA) [20]. For text preprocessing, we split a text into separate words using a Korean tokenizer [14], and embed them onto 300-dimensions of Doc2Vec [15] Skip-gram feature space. It is verified that Doc2vec trained in a large corpora can produce robust vector representations for long text paragraphs [16].



(a) Embedding Network Initialization (b) Semi-Supervised Learning

Fig. 2. Core components of semi-supervised multi-modal DEC

3.3 Embedding Network Initialization

Once the preprocessed image and text data points, X and X', are given, two embedding networks with initial parameters, θ and θ', are created as in the original MultiDEC [7]. For the embedding networks, we train two symmetric stacked autoencoders which contain encoding and decoding layers; since the networks are symmetric, we describe only one embedding network for image data hereafter, supposing that text data embedding is gone through the same procedure, simultaneously. The stacked autoencoder with parameter θ compresses the input data X into a latent space Z in the encoder of stacked DNN layers and regenerates X^* from Z in the decoder with the minimized mean square error loss between X and X^*. To leverage the embedded features for HAR with j human activity classes, we apply the K-means algorithm to Z where k equals to j and generate j initial clusters with centroids μ. Then, to associate the generated j clusters with the most relevant human activity classes, we generate a $j \times j$ confusion matrix. The (m, n) element of this matrix indicates how many data with the mth class label is contained in the nth cluster. We create a cost matrix by subtracting the maximum value of the confusion matrix from the value of each cell and find the class assignment that minimizes the cost by applying the Hungarian algorithm. Finally, we rearrange the centroids μ to follow the assignment we find (Fig. 2).

3.4 Supervised Learning

When the latent features are ready to be associated with the label-oriented supervision, we initiate the supervised learning phase. In this phase, we optimize the model by minimizing the cross-entropy between the soft assignment probabilities, q_{ij} and r_{ij}, of the image and text samples, x_i and x'_i, from the labeled training set $\mathcal{L}$ and the given class label indicators, y_{ij}, respectively. y_{ij} is a binary indicator and it is assigned by 1 if a data point x_i is assigned to the cluster of its correct class label j and closely located to its centroid, otherwise 0.

Soft Assignment. We calculate the image soft assignment probability, q_{ij}, which is the similarity between the image embedding, z_i, and the image cluster centroid, μ_j, by making use of the Student's t-distribution [22] on 1 degree of freedom (Eq. 1). Similarly, we calculate the text soft assignment probability, r_{ij}, using the text embedding, z'_i, and the text cluster centroid, μ'_j, (Eq. 2).

$$q_{ij} = \frac{(1 + ||z_i - \mu_j||^2)^{-1}}{\sum_{j'}(1 + ||z_i - \mu_{j'}||^2)^{-1}} \tag{1}$$

$$r_{ij} = \frac{(1 + ||z'_i - \mu'_j||^2)^{-1}}{\sum_{j'}(1 + ||z'_i - \mu'_{j'}||^2)^{-1}} \tag{2}$$

Cross Entropy Minimization. The supervised loss functions for image and text models are defined by a sum of cross-entropy values between the calculated soft assignment probability and the given class label indicator of each sample $x_i \in \mathcal{L}$ as follows:

$$SL_{img} = \sum_{x_i \in \mathcal{L}} H(\mathbf{y}_i, \mathbf{q}_i) = -\sum_{x_i \in \mathcal{L}} \sum_j y_{ij} log(q_{ij}) \tag{3}$$

$$SL_{txt} = \sum_{x_i \in \mathcal{L}} H(\mathbf{y}_i, \mathbf{r}_i) = -\sum_{x_i \in \mathcal{L}} \sum_j y_{ij} log(r_{ij}) \tag{4}$$

During the supervised learning phase, our model learns to locate the embedded feature z_i of the labeled data points $x_i \in \mathcal{L}$ as close as possible to the centroid μ_j of each labeled class j.

3.5 Unsupervised Learning

In the unsupervised learning phase, our model is trained by using deep embedded clustering both on the unlabeled and labeled datasets, $\mathcal{U} \cup \mathcal{L}$. This learning proceeds by minimizing the KL-divergence of $\mathbf{q}$ and $\mathbf{r}$ against the generated target probability distribution $\mathbf{p}$.

Joint Target Distribution. The target distribution p_{ij} is computed using q_{ij} and r_{ij} jointly. We apply the second power distribution to q_{ij} and r_{ij}, respectively, in order to improve cluster purity and give more emphasis on data points assigned with high confidence as proposed in the DEC [6]. We take the mean distribution of them to calculate both q_{ij} and r_{ij} evenly (i.e., late fusion), following the MultiDEC [7].

$$p_{ij} = \frac{1}{2}\left(\frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij'}^2/f_{j'}} + \frac{r_{ij}^2/g_j}{\sum_{j'} r_{ij'}^2/g_{j'}}\right) \tag{5}$$

KL Divergence Minimization. Once the joint target distribution is computed, we train our model by minimizing KL divergence with the unsupervised loss functions defined in Eq. 6 and 7. In addition to KL divergence minimization among $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{r}$, DEC models can be trained by introducing extra losses

between the mean $\mathbf{h}$ of the target probability distribution $\mathbf{p}$ and the prior knowledge $\mathbf{w}$ of the class distribution [7]. Here, our prior knowledge is obtained from the class distribution of $\mathcal{L}$.

$$UL_{img} = KL(\mathbf{p}||\mathbf{q}) + KL(\mathbf{h}||\mathbf{w}) \tag{6}$$

$$UL_{txt} = KL(\mathbf{p}||\mathbf{r}) + KL(\mathbf{h}||\mathbf{w}) \tag{7}$$

where $h_j = \sum_i p_{ij}/N$ and $w_j = \sum_i y_{ij}/N$.

Algorithm 1. Semi-Supervised Multi-Modal DEC

Require: models $M = (M_{img}, M_{txt})$, labeled set $\mathcal{L} = (\mathcal{L}_{img}, \mathcal{L}_{txt})$, unlabeled set $\mathcal{U} = (\mathcal{U}_{img}, \mathcal{U}_{txt})$, supervised learning rate η_s, unsupervised learning rate η_u.

Initialization:
$M_{img} \leftarrow$ encoder(autoencoder($\mathcal{L}_{img} \cup \mathcal{U}_{img}$))
$M_{txt} \leftarrow$ encoder(autoencoder($\mathcal{L}_{txt} \cup \mathcal{U}_{txt}$))
repeat
  Supervised Learning:
  $Q_1 \leftarrow$ predictions $(M_{img}, \mathcal{L}_{img})$ ▷ defined in (1)
  $R_1 \leftarrow$ predictions $(M_{txt}, \mathcal{L}_{txt})$ ▷ defined in (2)
  $M_{img} \leftarrow$ supervised_training $(\mathcal{L}_{img}, Q_1)$ ▷ based on loss defined in (3)
  $M_{txt} \leftarrow$ supervised_training $(\mathcal{L}_{txt}, R_1)$ ▷ based on loss defined in (4)
  Unsupervised Learning:
  $Q_2 \leftarrow$ predictions $(M_{img}, \mathcal{L}_{img} \cup \mathcal{U}_{img})$ ▷ defined in (1)
  $R_2 \leftarrow$ predictions $(M_{txt}, \mathcal{L}_{txt} \cup \mathcal{U}_{txt})$ ▷ defined in (2)
  $P \leftarrow$ target_distribution (Q_2, R_2) ▷ defined in (5)
  $M_{img} \leftarrow$ train_model (P, Q_2) ▷ based on loss defined in (6)
  $M_{txt} \leftarrow$ train_model (P, R_2) ▷ based on loss defined in (7)
until *end condition is met;*

In this phase, we set a learning rate, η_u, smaller than that in the supervised learning phase, η_s, so that $Z_{\mathcal{L}}$ is not affected too much by $\mathcal{U}$; in this work, we use $\eta_u = \kappa \times \eta_s$, where κ is an input parameter. Our learning method leverages the feature distributions of $\mathcal{U}$ and $\mathcal{L}$ together to make our model learn more generalized parameters, θ and μ, and prevent the trained model from being overfitted to $\mathcal{L}$. Algorithm 1 presents how the presented algorithms are used together through the entire procedure.

4 Evaluation

4.1 Dataset Construction

For dataset construction, we have collected geo-tagged Instagram posts containing various human activities in urban places nearby 25 stations on subway line 2 in Seoul from January 2015 to December 2018. Our collection is limited to Korean Instagram posts with non-empty captions. We have refined the captions

by removing URLs, numbers, email addresses, or emoticons. We filtered out spam posts created multiple times by the same author, with the same caption, or at the same location. If a single post has more than one image, we take only the first image. Eventually, we constructed a dataset of 967,598 image-text pairs.

Table 1. The number of labeled Instagram posts for each human activity classes

Class label	Count	Class label	Count
Eating & Drinking*	5,013	Educational Activities*	256
Arts & Entertainment*	2,721	No Activity	102
Socializing & Communicating*	2,609	Caring for HH Members	87
Traveling*	1,651	Household Activities	35
Personal Care*	1,114	Unknown	27
Relaxing & Leisure*	926	Telephone Calls	19
Sports, Exercise, Recreation*	465	Volunteer Activities	15
Consumer Purchases*	456	Religious & Spiritual Activities	8
Work-Related Activities*	409	Government Services	6
Advertisement	339	Caring for NonHH Members	6
Attending or Hosting Social Events*	337	Household Services	2
Professional & Personal Services*	291	**Total**	**16,894**

For data labeling, we use major human activity classes in *American Time Use Survey (ATUS) taxonomy* [12] which has been widely-used for HAR. To establish a qualified dataset with a consensus mechanism, we first gathered 27 participants and divided them into three groups. Then, each group is given the same set of Instagram posts and asked to annotate the most likely human activity that each post represents. We use only posts that more than two participants agreed on the same human activity class label.

Table 1 shows 23 human activity class labels and their corresponding counts. Out of 17 ATUS classes, *Socializing, Relaxing, and Leisure* class appears too frequently in social media. We divide its Instagram posts into their sub-classes, *Socializing & Communicating, Attending or Hosting Social Events, Relaxing & Leisure, and Arts & Entertainment (other than sports)*, defined in ATUS taxonomy. Additionally, we add *Advertisement* and *Unknown* classes for filtering out spam and ambiguous posts. Eventually, we establish an HAR dataset of 16,894 labeled posts. From the dataset, we use 16,248 posts of 12 most frequently appearing classes for this evaluation, where they are marked as asterisks.

4.2 Evaluation Setup

For evaluation metric, we adopt accuracy score, macro f1 score, and Normalized Mutual Information (NMI). Accuracy score indicates the straightforward recognition performance, macro f1 score is a normalized HAR performance metric,

and NMI indicates similarity between the probability distributions of the actual classes of the test set and the probability distributions of the predicted classes.

With these three standard metrics, we perform five-fold cross-testing and measure the average of the results. For training our model, we use unlabeled data with the training set of labeled data. In addition, we use the stochastic gradient descent (SGD) optimizer with a batch size of 256 and a learning rate of 0.01 (η_s). For training the stacked autoencoders in our model, we use the same configurations (i.e., layer structure and hyper-parameters) as in the original DEC model [6].

For a comparative evaluation, we implement baseline models with different data modalities. For text-only baseline models, we implement a linear SVM with a TF-IDF text input vector (TF-IDF+SVM) [1], an LSTM with a text Word2Vec input (Word2Vec+LSTM) [3], and an LSTM with dictionary embedding and a Word2Vec input (Word2Vec+LSTM+DE). For image-only baseline models, we implement a ResNet50 model which is pre-trained on ImageNet. This model is known to be robust to image classification [21] but no empirical results has been presented on HAR using social media data yet. To evaluate whether our proposed model utilizes the multi-modality of image-sharing social media data effectively and enhances the HAR performance, we also implement semi-supervised DEC models using uni-modality (i.e., text or image only) and compare their performance with that of the multi-modal one.

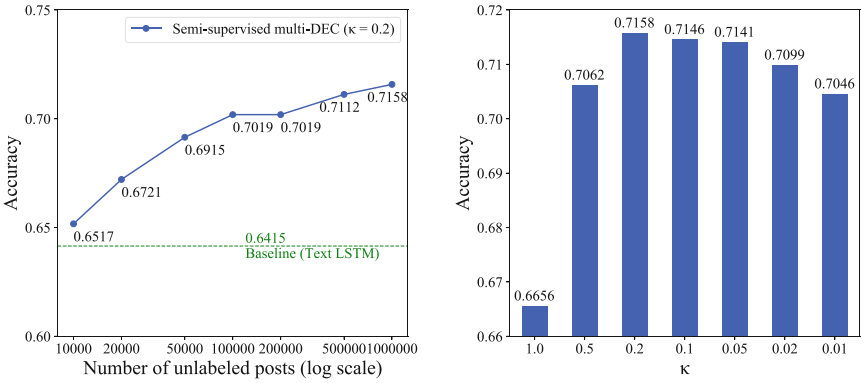
4.3 Evaluation Results

Evaluation results are summarized in Table 2. Above all, we find that our proposed model (Semi-supervised multi DEC) performs the best for HAR in all evaluation metrics. For example, in terms of accuracy score, its performance is 7.43% and 11.89% higher than the best performance of the text-only and image-only models, respectively. In terms of a normalized metric, macro f1 score, our proposed model performs 5.69% and 21.41% better than the text-only and image-only models. The higher NMI of our proposed model indicates that the semi-supervised multi-modal method is more effective for achieving an optimal clustering to the correct distribution than the baseline approaches. Considering that the semi-supervised uni-modal DEC performs worse than the baseline models in text-only case, we can deduce that our proposed model's achievement of the high accuracy is not attributed to the DEC component. In addition, by comparing the semi-supervised models with multi-modality and uni-modality, we can clearly find that incorporating multi-modal data helps to improve the HAR performance.

In order to verify the effect of semi-supervised learning using unlabeled samples, we measure the accuracy improvement as raising the number of unlabeled data while fixing the number of labeled data. Since our model generalizes itself by learning the feature distribution of the unlabeled data and avoids overfitting issue, its recognition accuracy is supposed to increase as the number of unlabeled data does. We fix the number of labeled data to 12,998 (i.e., 5-fold training data) and raise the number of unlabeled data $n(\mathcal{U})$ from 0 to 951,350.

Table 2. The result across different modals

Modality	Model	ACC	Macro F1	NMI
Text only	TF-IDF + SVM	0.6335	0.5283	0.3589
	Word2Vec + LSTM	0.6385	0.5372	0.3815
	Word2Vec + LSTM + DE	0.6415	0.5453	0.3804
	Semi-supervised text DEC	0.5939	0.5205	0.3394
Image only	ResNet50	0.5969	0.3881	0.3149
	Semi-supervised image DEC	0.5984	0.4484	0.3384
Text+Image	Semi-supervised multi DEC	**0.7158**	**0.6022**	**0.4790**



(a) different number of unlabeled posts (b) different value of κ ($n(\mathcal{U}) = 1$M)

Fig. 3. The result of semi-supervised multi-DEC accuracy

From the results in Fig. 3(a), we observe that the accuracy increases logarithmically as the number of unlabeled data does in general. Compared to the accuracy with no unlabeled data given, we achieve about 8.9% of the performance gain when we use all unlabeled data. The accuracy increases very rapidly up to 100K unlabeled data, but the improvement speed slows down when the number of unlabeled data is from 100K to 200K. This means that at least 100K to 200K unlabeled data is required to generalize the feature distribution of 12K labeled data based on our data set. Hence, we conclude that incorporation of unlabeled data into the model training is helpful to improve HAR performance and our proposed model is capable of taking the advantage effectively.

When we set the supervised learning rate and the unsupervised learning rate to the same value ($\kappa = 1$), the performance is 66.56%, which is 2.41% higher than that of the baseline model. We conduct a further experiment here by putting κ, an input parameter that adjusts the unsupervised learning rate, so that supervised learning is not overly influenced by unsupervised learning. As a result, our model shows the highest performance of 71.58% when κ is 0.2 on our

dataset when the number of unlabeled data is one million. This result assures that we can control the generalization effect of unsupervised learning with κ. We note that the value of κ varies according to the characteristics of the data such as the number of unlabeled data, the type of dataset, ambiguity among classes, etc. and we leave it for our future research.

5 Conclusion

In this paper, we present a semi-supervised multi-modal deep embedded clustering method for human activity recognition using social media. We adopt MultiDEC to leverage both image and text modalities and extend it into a semi-supervised learning method. By leveraging both labeled and unlabeled data, our model learns generalized feature representations and avoids being overfitted to the labeled features. Our proposed model achieves an improved HAR accuracy, compared to those of existing uni-modal approaches. In addition, we find that the incorporation of unlabeled data into the training procedure is helpful to improve HAR performance and our proposed model is capable of taking the advantage effectively.

Acknowledgement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01126, Self-learning based Autonomic IoT Edge Computing).

References

1. Zhu, Z., Blanke, U., Calatroni, A., Tröster, G.: Human activity recognition using social media data. In: Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia, p. 21. ACM (2013)
2. Zhu, Z., Blanke, U., Tröster, G.: Recognizing composite daily activities from crowd-labelled social media data. Pervasive Mob. Comput. **26**, 103–120 (2016)
3. Gong, J., Li, R., Yao, H., Kang, X., Li, S.: Recognizing human daily activity using social media sensors and deep learning. Int. J. Environ. Res. Public Health **16**(20), 3955 (2019)
4. Roy, A., Paul, A., Pirsiavash, H., Pan, S.: Automated detection of substance use-related social media posts based on image and text analysis. In: 2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI), pp. 772–779. IEEE (2017)
5. Schifanella, R., de Juan, P., Tetreault, J., Cao, L.: Detecting sarcasm in multimodal social platforms. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 1136–1145. ACM (2016)
6. Xie, J., Girshick, R., Farhadi, A.: Unsupervised deep embedding for clustering analysis. In: International Conference on Machine Learning, pp. 478–487 (2016)
7. Yang, S., Huang, K.H., Howe, B.: MultiDEC: multi-modal clustering of image-caption pairs. arXiv preprint arXiv:1901.01860 (2019)
8. Enguehard, J., O'Halloran, P., Gholipour, A.: Semi-supervised learning with deep embedded clustering for image classification and segmentation. IEEE Access **7**, 11093–11104 (2019)

9. Cain, J.: Social media in health care: the case for organizational policy and employee education. Am. J. Health-Syst. Pharm. **68**(11), 1036–1040 (2011)
10. Mittal, R., Sinha, V.: A personalized time-bound activity recommendation system. In: 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), pp. 1–7. IEEE (2017)
11. Wei, Y., Zhu, Y., Leung, C.W.K., Song, Y., Yang, Q.: Instilling social to physical: co-regularized heterogeneous transfer learning. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
12. Shelley, K.J.: Developing the American time use survey activity classification system. Monthly Lab. Rev. **128**, 3 (2005)
13. Zhu, X., Goldberg, A.B.: Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 3, no. 1, pp. 1–130 (2009)
14. Park, E.L., Cho, S.: KoNLPy: Korean natural language processing in Python. In: Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, vol. 6 (2014)
15. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: International Conference on Machine Learning, pp. 1188–1196 (2014)
16. Lau, J.H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368 (2016)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
19. Guérin, J., Gibaru, O., Thiery, S., Nyiri, E.: CNN features are also great at unsupervised classification. arXiv preprint arXiv:1707.01700 (2017)
20. Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **374**(2065), 20150202 (2016)
21. Su, D., Zhang, H., Chen, H., Yi, J., Chen, P.Y., Gao, Y.: Is robustness the cost of accuracy?-A comprehensive study on the robustness of 18 deep image classification models. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 631–648 (2018)
22. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)