Multi-Task Learning for Improved Recognition of Multiple Types of Acoustic Information

Jae-Won KIM†, Nonmember and Hochong PARK†a, Member

SUMMARY We propose a new method for improving the recognition performance of phonemes, speech emotions, and music genres using multi-task learning. When tasks are closely related, multi-task learning can improve the performance of each task by learning common feature representation for all the tasks. However, the recognition tasks considered in this study demand different input signals of speech and music at different time scales, resulting in input features with different characteristics. In addition, a training dataset with multiple labels for all information sources is not available. Considering these issues, we conduct multi-task learning in a sequential training process using input features with a single label for one information source. A comparative evaluation confirms that the proposed method for multi-task learning provides higher performance for all recognition tasks than individual learning for each task as in conventional methods.

key words: phoneme recognition, emotion recognition, music genre recognition, multi-task learning, neural network

1. Introduction

In machine learning, multi-task learning (MTL) achieves common feature representation applicable to multiple tasks so that the learned model is less dependent on a specific task[1]–[3]. Therefore, it has the potential to improve the performance of all the tasks by learning more generalized model for them. In this study, we propose a method for improving the recognition performance of phonemes, speech emotions, and music genres using MTL based on a structure shown in Fig. 1. This structure resembles human audition that recognizes multiple sources of acoustic information from one input based on common feature modeling [4]. The proposed method improves all tasks simultaneously by using the same function obtained by joint learning for all the tasks, which is a different approach from conventional methods that are designed independently for each task [5]–[8].

The effect of MTL is maximized when tasks are closely related using the same input [3]. Accordingly, MTL has been successfully used to recognize multiple objects in images, where all tasks share the same input and low-level features [9], [10]. For speech, MTL is used for multilingual speech processing, where common acoustic model is learned from different languages [11]. It is also used in various fields of recognition for improved performance, such as speech recognition [12] and emotion recognition [13].

In most speech applications of MTL, all tasks share the same speech input and are closely related within one recognition category such as speech recognition or emotion recognition. In contrast, recognition tasks for phonemes, emotions, and music genres are not closely related except for the fact that they all recognize acoustic information, because they demand two different signal types of speech and music at different time scales. Furthermore, unlike in conventional MTL, input features with multiple labels for all the tasks are not available. Therefore, we consider a more complex setting of MTL, similar to disjoint MTL using single-task datasets in computer vision and language processing [14]–[16], and investigate the ability of MTL for a new scenario with single-labeled input features extracted from different signal types at different time scales, which has not been studied much for recognition of acoustic information.

The objective of this study is to improve the recognition performance using MTL for given input features and modeling structure, not to design new features and modeling structure for improved recognition. As a preliminary study to verify the feasibility of MTL, we adopt general input features that are widely used for analyzing acoustic information and a simple feedforward neural network (NN) with hard parameter sharing [2]. An advanced network with special devices for MTL, such as cross-stitch [9] and adaptive sharing [2], and the learned input features from autoencoder structures [17] will be investigated in our next study.

2. Proposed Method for Multi-Task Learning

2.1 Input Features

We obtain the 42-band mel-spectrogram and the 42 mel-frequency cepstral coefficients (MFCC) of input signal. Then, we compute their first- and second-order frame deltas, resulting in a feature set of $42 \times 3 = 126$ parameters based on the mel-spectrogram and MFCC, respectively. In addition, we obtain the 32-band spikegram of input using a Gammatone filter-bank [4], [8], and compute the magnitude sum of each band in a frame and the magnitude sum of all bands in each of ten subframes. Including the first- and second-order frame deltas, we obtain a feature set of 126 parameters based on the spikegram. Then, we conduct MTL and evaluate its performance using each feature set.

Phoneme recognition requires a short-time analysis of acoustic characteristics, whereas emotion and music genre
Fig. 1 Structure for phoneme, emotion, and music genre recognition based on multi-task learning.

Table 1 Structures of two frames for feature extraction.

| Frame | Frame length | Hop length | Information       |
|-------|--------------|------------|-------------------|
| T1    | 25 ms        | 10 ms      | Phoneme           |
| T2    | 100 ms       | 40 ms      | Emotion, music genre |

Fig. 2 Probability distribution of mel-spectrogram-based input features for each recognition task.

Fig. 3 Network architectures evaluated in this study.

2.2 Network Architectures and Training

Figure 3 shows the four NN architectures evaluated in this study with the number of neurons per layer; the number of output neurons is equal to the number of classes in each task. $S(n)$, where $n = 3, 4$ indicates the number of hidden layers, is a single-purpose NN that is individually trained for one task in a conventional way and provides baseline performance. $C(n)$ is a common NN for the three tasks with hard parameter sharing [2].

In $C(3)$, all the hidden layers are shared among the tasks and conduct common feature modeling in Fig. 1, yielding common feature representation applicable to all the tasks. The network parameters between the last hidden layer and each output sublayer are independently determined per task and conduct individual classification in Fig. 1. In $C(4)$, the last hidden layer is split into three sublayers for each task, providing one more task-specific layer than $C(3)$. Each NN layer, except for the output layer of $C(4)$, is a fully-connected layer; the output layer of $C(4)$ is fully connected to the last hidden layer on a sublayer basis. Comparing $S(n)$ and $C(n)$ for the same $n$, both have the same network structure and the same number of network parameters with respect to each task.

In normal MTL, a single input with multiple labels yields the outputs of all tasks and all costs are jointly optimized [3]. In this study, however, an input with a single label yields the output of one task and a sequential optimization across the tasks on a batch basis is conducted [11], [18]. For each batch to have the same time length, we set the batch size for phoneme, emotion, and music genre to 4096, 1024, and 1024, respectively.

Figure 4 illustrates the proposed procedure of NN training for all the tasks in MTL, applied to $C(4)$. We alternately train for each task on a batch basis. Specifically, one batch of 4096 $X_P$ features is input to $C(4)$, the phoneme recognition cost is computed at the output sublayer for phoneme, and the network parameters are updated

tional MTL. Instead, we aggregate datasets for phonemes, emotions, and music genres. In the training dataset for phonemes, the features $X_P$ are computed based on frame T1. In the training datasets for emotions and music genres, the features $X_E$ and $X_G$, respectively, are computed based on frame T2. Then, MTL is performed using a combination of $X_P$, $X_E$, and $X_G$ with labels in different spaces.
using backpropagation from this output sublayer to input layer. Then, one batch of 1024 \(X_E\) features is input to the \(C(4)\) composed of the parameters updated in the previous training stage for phoneme recognition, and the network parameters are updated based on the emotion recognition cost using backpropagation from the output sublayer for emotion to input layer. Finally, the same process is conducted using one batch of 1024 \(X_G\) features for music genre recognition. The same alternate training for the three tasks is repeated using new batches of \(X_P\), \(X_E\) and \(X_G\). The training for \(C(3)\) is conducted analogously.

Although \(X_P\), \(X_E\) and \(X_G\) have different distributions and the transition in feature distribution continues to occur when the task changes during alternate training, the proposed training procedure does not conduct any special operation to handle this situation. We will confirm that, despite this unique environment, MTL can learn the generalized model for all the tasks.

3. Performance Evaluation

We used the TIMIT corpus with 39 phoneme classes [19], the IEMOCAP database with four emotion classes (i.e., happy, sad, angry, and neutral) [20], and the GTZAN database with ten genre classes [5] for phoneme, emotion, and music genre recognition, respectively. We resampled the GTZAN database from 22.05 to 16 kHz for the sampling frequencies in all tasks to be the same. We evaluated the recognition performance based on 10-fold cross-validation.

In the evaluated NNs, the hidden layers used the rectified linear unit activation and the output layers used the softmax activation. For a fair performance comparison, all the NNs were trained under the same conditions. We used He initialization, Adam optimizer, drop-out with a probability of 0.5 for all the hidden layers, the cross-entropy cost function with no regularization terms, and the same early stopping condition. The training datasets across tasks were set to the same length of approximately 3.7 hours by truncating the longer datasets, and consisted of 329 batches with approximately \(1.35 \times 10^6\) frames for phoneme and \(337 \times 10^3\) frames for emotion and music genre. When training \(C(n)\), therefore, all tasks terminated one training epoch simultaneously. When all tasks satisfied the early stopping condition, they terminated the training simultaneously with the same number of epochs. As a result, we obtained the learned \(C(n)\) that conducted the same number of updates per task.

Figure 5 shows the learning curve of each task when training \(S(3)\) and \(C(3)\). \(S(3)\) was trained individually for each task in a conventional way, whereas \(C(3)\) was trained as shown in Fig. 4 including repetitive transitions between label spaces of 39, 4, and 10 dimensions. Nevertheless, there was no substantial difference between the learning curves of \(S(3)\) and \(C(3)\) per task. When further training after the early stopping condition was conducted, similar learning curves for \(S(3)\) and \(C(3)\) were still observed. Similar results were also observed in the learning curves for other feature sets and NN structures. Therefore, it was confirmed that \(C(n)\) successfully learned all tasks based on common feature representation.

We evaluated the recognition performance per task for all the combinations of feature sets and NNs. The accuracy of phoneme recognition was measured in each T1 frame, and the accuracy of emotion and music genre recognition was measured in each input file based on the average of the class probabilities at the output layer across the T2 frames. Table 2 lists the recognition performance in terms of mean accuracy (MA), measured as a ratio between the number of correctly recognized items and the number of all tested items. Each number in parenthesis for \(C(n)\) represents the amount of MA increase in a unit of % point, compared with that in \(S(n)\) for the same \(n\). We conducted the \(t\)-test for these results, and the significant MA increase by \(C(n)\) with the \(p\)-value less than 0.05 is indicated in bold.

For emotion and music genre recognition, \(C(n)\) provides the higher performance than \(S(n)\) in all test cases. The MA increase is due to the reduced overfitting by the generalized modeling for emotion and music genre recognition. In contrast, the MA for phoneme recognition in \(C(n)\) is lower than that in \(S(n)\) because this task has much more classes and more training epochs are required, compared with emotion and music genre recognition. Thus, when training three tasks under the same conditions, \(C(n)\) does not fully learn phoneme recognition and learning is biased toward emotion and music genre recognition.

To enhance generalization for phoneme recognition in MTL, we modified the training procedure to adopt dropout only for training phoneme recognition but omit it for training emotion and music genre recognition in Fig. 4. Furthermore, we extended the length of the training dataset for phoneme recognition by eight times, inducing more parameter updates for phoneme recognition. Then, at each training stage for phoneme recognition in Fig. 4, eight parameter updates for phoneme recognition. Then, at each training stage for phoneme recognition in Fig. 4, eight parameter updates for phoneme recognition.
updates using eight batches of 4096 \( X_P \) features were conducted in a row. In this way, all tasks terminated one training epoch simultaneously.

The NN trained with emphasis on phoneme recognition, denoted as \( C_P(n) \) in Table 2, exhibits the significant MA increase for phoneme recognition with \( p \)-value < 0.05 in the \( t \)-test, except for the test case of \( C_P(3) \)–mel-spectrogram. For emotion and music genre recognition, however, \( C_P(n) \) has lower MA than \( C(n) \) in some cases. This experiment confirms the effectiveness of emphasized training on a specific task in MTL.

Table 2 also shows the average MA per task across the three feature sets. Except for the case of \( C_P(3) \)–music genre recognition, \( C_P(n) \) has the significant increase in MA for every task. Finally, \( C_P(n) \) provides the significant increase in average MA across all the features sets and tasks.

4. Conclusion

We propose a new method to improve the recognition performance of phonemes, emotions, and music genres using MTL, which is differentiated from conventional methods developed for each recognition task. To conduct MTL, we adopt a sequential training process across the tasks, which enables generalized feature modeling applicable to every task. The comparative evaluation confirmed that MTL outperforms individual learning for each task, when using the same feature sets, NN structures, and training conditions.

Acknowledgments

This work was supported by the Research Grant of Kwangwoon University in 2021 and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2016R1D1A1B03930923).

References

[1] R. Caruana, “Multitask learning,” Machine Learning, vol.28, no.1, pp.41–75, 1997.
[2] S. Ruder, “An overview of multi-task learning in deep neural networks,” arXiv preprint arXiv:1706.05098, 2017.
[3] Y. Zhang and Q. Yang, “A survey on multi-task learning,” arXiv preprint arXiv:1707.08114, 2018.
[4] E.C. Smith and M.S. Lewicki, “Efficient auditory coding,” Nature, vol.439, pp.978–982, 2006.
[5] G. Tzanetakis and P. Cook, “Musical genre classification of audio signals,” IEEE Trans. Speech and Audio Processing, vol.10, no.5, pp.293–302, 2002.
[6] M.E. Ayadi, M.S. Kamel, and F. Karray, “Survey on speech emotion recognition: features, classification schemes, and databases,” Pattern Recognition, vol.44, no.3, pp.572–587, 2011.
[7] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” IEEE/ACM Trans. Audio, Speech and Language Processing, vol.22, no.10, pp.1533–1545, 2014.
[8] S.-H. Shin, H.-W. Yun, W.-J. Jang, and H. Park, “Extraction of acoustic features based on auditory spike code and its application to music genre classification,” IET Signal Processing, vol.13, no.2, pp.230–234, 2019.
[9] I. Misra, A. Shrivastava, A. Gupta, and M. Herbert, “Cross-stitch networks for multi-task learning,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.3994–4003, 2016.
[10] R. Ranjan, V.M. Patel, and R. Chellappa, “Hyperface: a deep multitask learning framework for face detection, landmark localization, pose estimation, and gender recognition,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol.41, no.1, pp.121–135, 2019.
[11] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, “Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc., pp.7304–7308, 2013.
[12] S. Kim, T. Hori, and S. Watanabe, “CTC-attention based end-to-end speech recognition using multi-task learning,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process., pp.4835–4839, 2017.
[13] Z. Zhang, B. Wu, and B. Schuller, “Attention-enhanced end-to-end multi-task learning for emotion prediction from speech,” Proc. IEEE Int. Conf. Acoustics, Speech and Signal Proc., pp.6705–6709, 2019.
[14] H. Chen, Y. Zhang, and Q. Liu, “Neural network for heterogeneous annotations,” Proc. Conf. on Empirical Methods in Natural Language Processing, pp.731–741, 2016.
[15] D.-J. Kim, J. Choi, T.-H. Oh, Y. Yoon, and S.S. Kweon, “Disjoint multi-task learning between heterogeneous human-centric tasks,” Proc. IEEE Winter Conf. on Applications of Computer Vision, pp.1699–1708, 2018.
[16] Y. Hong, L. Niu, J. Zhang, and L. Zhang, “Beyond without forgetting: multi-task learning for classification with disjoint datasets,” Proc. IEEE Int. Conf. on Multimedia and Expo, 2020.
[17] J. Chorowski, R.J. Weiss, S. Bengio, and A. van den Oord, “Unsupervised speech representation learning using WaveNet autoencoders,” IEEE/ACM Trans. Audio, Speech and Language Processing, vol.27, no.12, pp.2041–2053, 2019.
[18] J.-W. Kim and H. Park, “Performance enhancement of phoneme and emotion recognition by multi-task training of common neural network,” J. Broadcast Eng., vol.27, no.5, pp.742–749, 2020.
[19] V. Zue, S. Seneff, and J. Glass, “Speech database development at MIT: TIMIT and beyond,” Speech Communication, vol.9, no.4, pp.351–356, 1990.
[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, “IEMOCAP: interactive emotional dyadic motion capture database,” Language Resources and Evaluation, vol.42, no.4, pp.335–359, 2008.