Evaluation of Error and Correlation-Based Loss Functions For Multitask Learning Dimensional Speech Emotion Recognition

Bagus Tris Atmaja
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Japan
bagus@jaist.ac.jp

Masato Akagi
School of Information Science
Japan Advanced Institute of Science and Technology
Nomi, Japan
akagi@jaist.ac.jp

***Abstract*—The choice of a loss function is a critical part in machine learning. This paper evaluated two different loss functions commonly used in regression-task dimensional speech emotion recognition, an error-based and a correlation-based loss functions. We found that using correlation-based loss function with a concordance correlation coefficient (CCC) loss resulted better performance than error-based loss function with a mean squared error (MSE) loss, in terms of the averaged CCC score. The results are consistent with two input feature sets and two datasets. The scatter plots of test prediction by those two loss functions also confirmed the results measured by CCC scores.**

***Index Terms*—loss function, multitask learning, mean squared error, concordance correlation coefficient, dimensional speech emotion recognition**

I. Introduction

Dimensional emotion recognition is scientifically more challenging than categorical emotion recognition. In dimensional emotion recognition, the goal is to predict the continuous degree of emotional attributes, while in categorical emotion recognition, the task is to predict emotion category of speakers, whether they are angry, happy, sad, fear, disgust, or surprise. Although the practical application of this dimensional emotion recognition is not clear yet, Russel [1] argued that categorical emotion can be derived from two-space dimensional emotion, i.e., valence (positive or negative) and arousal (high or low).

In the dimensional emotion model, several models have been introduced by psychological researchers, including 2D, 3D, and 4D models. In the 3D model, either dominance (power control) or liking is used as the third attribute. In 4D model, either expectancy or unpredictability is used, such as in [2] and [3]. This research used 3D emotions with valence, arousal, and dominance (VAD) model, as suggested in [4].

In the 3D emotion model, the emotion recognizer system (classifier) needs to predict three emotional attributes. This task is often performed by simultaneous or jointly learning prediction of VAD. This simultaneous learning technique is known as multitask learning (MTL). Compared to single-task learning (STL), MTL tries to optimize three parameters at the same time, while STL only tries to optimize one parameter (either valence, arousal, or dominance). To train the model over those three attributes, the choice of loss function for the MTL is vital for the performance of the system. The traditional regression task used mean squared error (MSE) as both loss function and evaluation metric. Dimensional emotion recognition, as a regression task, conventionally follow that rule by applying MSE for both loss and evaluation metric.

Recently, affective computing researchers argued that using correlation-based metric to evaluate the performance of dimensional emotion recognition is more appropriate than calculating its errors [5]–[7]. The concordance correlation coefficient (CCC) [8] is often used to measure the performance of dimensional emotion recognition since it takes the bias into Pearson's correlation coefficient (CC). Hence, we hypothesized that using CCC loss ($1 - CCC$) is more relevant than using MSE as loss function for MTL dimensional speech emotion recognition. This paper aims to evaluate this hypothesis.

To the best of our knowledge, there is no research reporting direct comparison of the impact of using MSE vs. CCC for dimensional speech emotion recognition. Some authors used MSE loss, such as in [9], [10] while the others used CCC loss, such as in [7], [11]. Both groups reported the performance of the evaluated method using CCC. We choose speech emotion recognition as our task since the target application is speech-based apps like voice assistant and call center service.

II. Problem Statement

We focused our work to evaluate which loss function performs better on multitask learning dimensional emotion recognition, MSE or CCC loss. To achieve this goal, we used two different datasets and two different acoustic feature sets. We expected consistent results across four scenarios or parts (2 datasets $\times$ 2 feature sets). The same experiment condition (i.e., the same architecture with the same parameters) is used to evaluate both MSE and CCC loss functions in four scenarios. Although the main metric is CCC, MSE scores are also given as additional metrics. The averaged CCC score among three emotion dimensions is used to evaluate the performance of evaluated loss function on each scenario.

III. Evaluation Methods

In this section, we present the core idea of the research: how to evaluate two loss functions for multitask learning dimensional speech emotion recognition. First, we describe data and feature sets to evaluate the loss functions. Second, we describe MSE and CCC-based loss functions. Finally, we show the architecture of dimensional speech emotion recognition to evaluate those loss functions.

A. Data and Feature Sets

Two datasets and two acoustic sets are used to evaluate two different loss functions.

Datasets: IEMOCAP and MSP-IMPROV datasets are utilized to evaluate error and correlation-based loss functions. Among many modalilites provided by both datasets, only speech data is used to extract acoustic feature sets. The first dataset consists of 10039 turns while the second consists of 8438 utterances. For both datasets, only dimensional labels are used, i.e., valence, arousal, and dominance, in the range [1, 5]. We scaled those labels into the range [-1, 1], following the work in [9] when fed it into deep learning-based dimensional speech emotion recognition system. The detail of IEMOCAP dataset is given in [12], while for MSP-IMPROV dataset is available in [13]. All scenarios in the two datasets are performed in speaker-independent configuration for test data, i.e., the last one session is left out for test partition (LOSO, leave one session out). For IEMOCAP data, the number of training partition is 7869 utterances, and the rest 2170 utterances (session fifth) are used for the test partition. On the MSP-IMPROV dataset, 6816 utterances are used for the training partition, and the rest 1622 utterances (session sixth) are used for the test partition. On both test partitions, 20% of data is used for validation (development).

Acoustic Features: High-level statistical function (HSF) of two feature sets are used. The first is HSF from Geneva minimalistic acoustic and parameter set (GeMAPS), as described in [14]. The second is HSF from pyAudioAnalysis (pAA) [15]. Note that the definition of HSF referred here is only mean and standard deviation (Mean+Std) from low-level descriptor (LLD) listed in both feature sets. GeMAPS feature set consists of 23 LLDs, while pAA contains 34 LLDs. A list of LLDs in those two feature sets is presented in table I. The use of Mean+Std in this research follows the finding in [16]. Additionally, we implement the Mean+Std of LLDs from pAA to observe its difference from GeMAPS.

TABLE I
Acoustic features used to evaluate the loss functions (only Mean+Std of those LLDs are used as input features).

Feature set	LLDs
GeMAPS	loudness, alpha ratio, Hammarberg index, spectral slope 0-500 Hz, spectral slope 500-1500 Hz, spectral flux, 4 MFCCs, F0, jitter, shimmer, Harmonics-to-Noise Ratio (HNR), harmonic difference H1-H2, harmonic difference H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.
pAA	zero crossing rate, energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectra flux, spectral roll-off, 13 MFCCs, 12 chroma vectors, chroma deviation.

B. MSE-based Loss Function

A mean squared error to measure the deviation between predicted emotion degree x and gold-standard label y is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2. \quad (1)$$

where n is number of measurement (or batch size). For three emotion dimensions, the total MSE is the sum of MSE from valence, arousal, and dominance.

$$MSE_T = MSE_V + MSE_A + MSE_D \quad (2)$$

Following the work of [9], we added weighting factors for valence and arousal. Hence, the MSE total became,

$$MSE_T = \alpha MSE_V + \beta MSE_A + (1-\alpha-\beta)MSE_D \quad (3)$$

where α and β are weighting factors for valence and arousal. The weighting factor for dominance is obtained by subtracting 1 with those two variables.

C. CCC-based Loss Function

CCC is a common metric in dimensional emotion recognition to measure the agreement between true emotion dimension with predicted emotion degree. If the predictions shifted in value, the score is penalized in proportion to deviation [5]. It becomes *de facto* metric to measure the performance of dimensional speech emotion recognition. CCC is formulated as

$$CCC = \frac{2\rho_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \quad (4)$$

where ρ_{xy} is the Pearson coefficient correlation between x and y, σ is the standard deviation, and μ is a mean value. This CCC is based on Lin's calculation [8]. The range of CCC is from -1 (perfect disagreement) to 1 (perfect agreement). Therefore, the CCC loss function (CCCL) to maximize the agreement between true value and prediction emotion can be defined as

$$CCCL = 1 - CCC \quad (5)$$

Similar to what in MSE, we accommodate the loss functions from arousal ($CCCL_V$), valence ($CCCL_A$), and dominance ($CCCL_D$). The $CCCL_T$ is a combination of those three CCC loss functions.

$$CCCL_T = \alpha\ CCCL_V + \beta\ CCCL_A + (1-\alpha-\beta)\ CCCL_D \quad (6)$$

where α are β are the weighting factors for each emotion dimension loss function. The same weighing factors are used for both MSE and CCC losses, i.e., 0.1 and 0.5 for IEMOCAP dataset, 0.3 and 0.6 for MSP-IMPROV dataset. Those weighting factors are obtained via linear search.

D. Architecture of Dimensional Speech Emotion Recognition

We used deep learning-based architecture to evaluate two loss functions, i.e., three layers of stacked LSTM networks [17]. For the input, either 46 HSFs from GeMAPS or 68 HSFs from pAA are fed into the network. A batch normalization layer is performed to speed up the computation process [18]. Three LSTM layers are stacked; the first two layers return all sequences while the last LSTM layer returns final outputs only. A dense network with 64 nodes is coupled after the last LSTM layer. Three dense layers with one unit each ended the network to predict the degree of valence, arousal, and dominance. Either MSE or CCC loss is used as the loss function with RMSprop optimizer [19]. The architecture of this dimensional speech emotion recognition is shown in Fig. 1.

As an additional analysis tool, we used scatter plots of predicted valence and arousal degrees compared to the gold-standard labels. These plots will show how similar or different between labels and predicted degrees. This similarity between labels and predictions can be used to confirm obtained CCC scores by different loss functions.

The implementation of the evaluation methods is available in the following repository, https://bagustris.github.com/ccc_mse_ser. This LSTM-based dimensional speech emotion recognition is implemented using Keras toolkit [20] with TensorFlow backend [21].

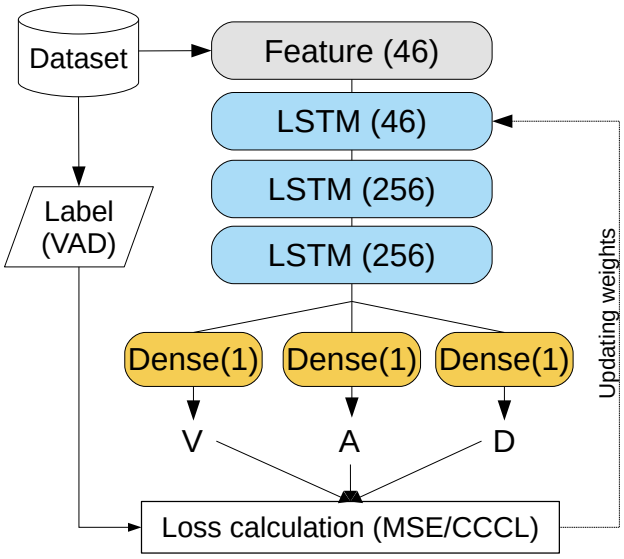

Fig. 1. Architecture of dimensional speeh emotion recognition system to evaluate loss functions.

IV. Results and Discussion

A. Performance of Evaluated Loss Functions

The main question of this research aimed to find which loss function work better for multitask learning dimensional emotion recognition. The following results report evaluation of two lost functions in two datasets and two features sets.

Table II shows the result of using different loss function to the same dataset using the same network architecture. We can divide the results into four parts: two datasets with two different feature sets for each dataset. The first part is the IEMOCAP dataset with HSF of GeMAPS as the input feature. Using CCC loss, the obtained CCC score for each dimension is higher than the obtained using MSE loss. The resulted averaged CCC score is 0.400 for CCCL and 0.310 for MSE loss. As a comparison, the obtained MSE scores are 0.163 for CCCL and 0.121 for MSE loss. Clearly, it is shown that CCCL obtained better performance than MSE in this part and continued to other three parts.

We evaluated HSFs from pAA on the second part of the table. Although the feature set is not designed specifically for an affective application, however, it showed a similar performance to the result obtained by affective-designed GeMAPS feature set. In this IEMOCAP dataset, HSF of pAA even performed marginally better than HSFs of GeMAPS for both CCC and MSE losses. The comparison of performance between CCCL and MSE is similar to what is obtained by GeMAPS with the averaged CCC score of 0.401 for CCCL and 0.383 for MSE loss. Scores of MSE for both losses are 0.163 and 0.125 for CCCL and MSE loss.

Moving to the MSP-IMPROV dataset, a similar trend was observed. On the third part with MSP-IMPROV and GeMAPS feature set, CCCL obtained the averaged CCC score of 0.363 compared to MSE with an averaged CCC score of 0.327. Finally, on the fourth part with MSP-IMPROV and pAA feature set, the CCCL obtained 0.34 of the averaged CCC while MSE obtained 0.305. The obtained MSE scores for both losses are similar, i.e., 0.164 and 0.122 for MSP-IMPROV with GeMAPS feature and 0.161 and 0.124 for MSP-IMPROV with pAA features.

The overall results above suggest that, in terms of CCC, CCC loss is better than MSE loss for multitask learning dimensional emotion recognition. Four scenarios with CCC loss function obtained higher score, on both individual emotion dimensions scores (CCC of valence, arousal, and dominance) and the averaged score, than other four scenarios with MSE loss. We extend the discussion to the results obtained by MSE scores and different feature sets.

We found that the averaged MSE scores across data and feature sets are almost identical (last column in Table II). If so, the MSE metrics might be more stable to generalize the model generated by dimensional speech emotion recognition system across different datasets. However, this consistent error for the generalization of dimensional speech emotion recognition system needs to be investigated with other datasets in different scale of labels. Another possible cause for the consistent error is the small range of the output, i.e., 0-1 scale after squared by the MSE.

On the use of different feature sets, we observed no remarkable difference between results obtained by HSF of affective-designed GeMAPS and general-purpose pAA feature sets. The results obtained by those two feature sets are quite similar for the same loss function. Not only on CCC scores, but the similarity of performance is also observed on MSE scores. This results can be viewed as the generalization from the

TABLE II
EVALUATION RESULTS OF CCC AND MSE LOSSES ON IEMOCAP AND MSP-IMPROV DATASETS.

Feature	Loss	CCC				MSE			
		V	A	D	Mean	V	A	D	Mean
IEMOCAP dataset									
GeMAPS	CCCL	0.192	0.553	0.456	**0.400**	0.211	0.074	0.137	0.140
	MSE	0.121	0.451	0.358	0.310	0.182	0.069	0.113	0.121
pAA	CCCL	0.183	0.577	0.444	**0.401**	0.248	0.080	0.161	0.163
	MSE	0.093	0.522	0.383	0.333	0.196	0.065	0.115	0.125
MSP-IMPROV dataset									
GeMAPS	CCCL	0.204	0.525	0.361	**0.363**	0.303	0.098	0.091	0.164
	MSE	0.138	0.492	0.353	0.327	0.221	0.084	0.060	0.122
pAA	CCCL	0.150	0.496	0.374	**0.340**	0.268	0.129	0.087	0.161
	MSE	0.122	0.475	0.319	0.305	0.218	0.094	0.058	0.124

previous research [7] that not only Mean+Std of GeMAPS useful for dimensional emotion recognition but also Mean+Std of other feature sets, in this case, pAA feature set.

B. Scatter Plot of Predicted Emotion Degrees

We showed the scatter plots of prediction by CCC loss and MSE loss in Figs. 2 and 3. In those cases, the plots showed the prediction from GeMAPS test partition. From both plots, we can infer that the prediction from CCC loss is more similar to gold-standard labels than the prediction from MSE loss. This result confirms the obtained CCC scores from valence, arousal, and dominance and its average. Although we only showed the result from IEMOCAP with GeMAPS feature, the plots are consistent with other parts. Note in those scatter plots that a single dot may represent more than one label (overlapped), since there is a possibility to have the same labels (score of valence and arousal) for several utterances.

Comparing the shape of both predictions (orange color), it is clear both have a different scale. CCC loss works better because it takes account of the shifted values of prediction into concordance correlation calculation while MSE only counts its errors. We can conclude from both metric and visualization that CCC loss gained better performance than MSE loss.

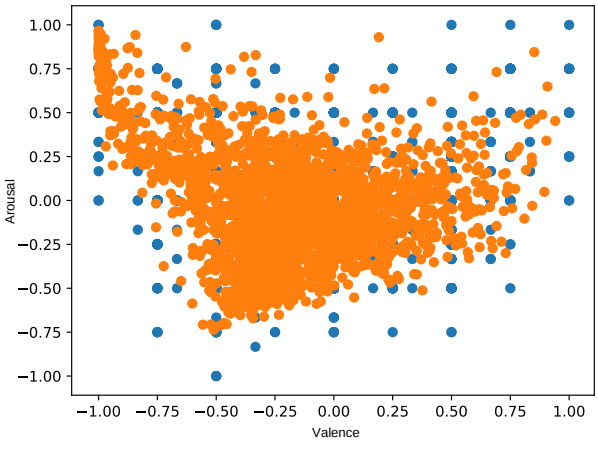

Fig. 2. Scatter plot of valence and arousal dimensions on test partition from IEMOCAP dataset with GeMAPS feature and CCC loss (blue: gold-standard label, orange: predicted degree).

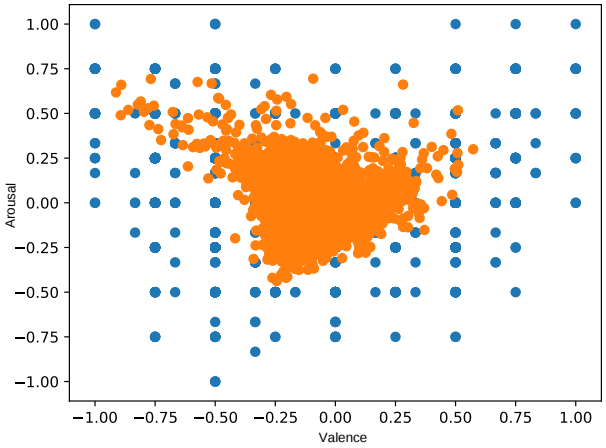

Fig. 3. Scatter plot of valence and arousal dimensions on test partition from IEMOCAP dataset with GeMAPS feature and MSE loss (blue: gold-standard label, orange: predicted degree).

V. CONCLUSIONS

This paper reported an evaluation of different loss functions for multitask learning dimensional emotion recognition. The result shows that the CCC loss obtained better performance than the MSE loss in terms of CCC scores across four scenarios. We are confident that this result is universal since we use two different datasets and two different feature sets that resulting consistent results and the process is straightforward (a CCC loss as the loss function with CCC scores as evaluation metrics). These results are also supported by scatter plots of the valence-arousal prediction from both losses, compared to gold-standard labels.

On the other side, we also found that the use of MSE as a metric resulting a more consistent errors across datasets and scenarios. Further study to investigate the correlation between error and correlation from both theoretical and practical approaches may improve our understanding on it. Although it is suggested to use CCC as the main metric for dimensional emotion recognition, additional metrics such as MSE and RMSE may be useful to accompany CCC measure for tracking the pattern of the performance across different datasets, feature sets, and methods, particularly in dimensional emotion recognition.

REFERENCES

[1] J. A. Russell, "Affective space is bipolar," *J. Pers. Soc. Psychol.*, 1979.

[2] J. D. Moore, L. Tian, and C. Lai, "Word-level emotion recognition using high-level features," in *Int. Conf. Intell. Text Process. Comput. Linguist.* Springer, 2014, pp. 17–31.

[3] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, C. Phoebe, J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The World of Emotions Is Not Two-Dimensional," vol. 18, no. 12, pp. 1050–1057, 2017.

[4] I. Bakker, T. van der Voordt, P. Vink, and J. de Boon, "Pleasure, Arousal, Dominance: Mehrabian and Russell revisited," *Curr. Psychol.*, vol. 33, no. 3, pp. 405–421, 2014.

[5] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "AV+EC 2015 - The first affect recognition challenge bridging across audio, video, and physiological data," in *AVEC 2015 - Proc. 5th Int. Work. Audio/Visual Emot. Challenge, co-Located with MM 2015*. Association for Computing Machinery, Inc, oct 2015, pp. 3–8.

[6] J. Han, Z. Zhang, M. Schmitt, B. Schuller, M. Pantic, and B. Schuller, "From Hard to Soft: Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty," in *ACMMM*, vol. 17, 2017, pp. 890–897. [Online]. Available: https://www.sewaproject.eu/files/c67f17f5-27d5-44d5-3f18-203d0198021f.pdf

[7] M. Schmitt, N. Cummins, and B. W. Schuller, "Continuous Emotion Recognition in Speech Do We Need Recurrence?" in *Interspeech 2019*. ISCA: ISCA, sep 2019, pp. 2808–2812.

[8] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–68, 1989.

[9] S. Parthasarathy and C. Busso, "Jointly Predicting Arousal, Valence and Dominance with Multi-Task Learning," in *Interspeech*, 2017, pp. 1103–1107.

[10] M. Abdelwahab and C. Busso, "Domain Adversarial for Acoustic Emotion Recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2423–2435, dec 2018.

[11] B. T. Atmaja and M. Akagi, "Multitask Learning and Multistage Fusion for Dimensional Audiovisual Emotion Recognition," in *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. (to Appear.*, 2020.

[12] C. Busso, M. Bulut, C.-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," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[13] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Trans. Affect. Comput.*, vol. 8, no. 1, pp. 67–80, jan 2017.

[14] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 190–202, apr 2016.

[15] T. Giannakopoulos, "PyAudioAnalysis: An open-source python library for audio signal analysis," *PLoS One*, vol. 10, no. 12, 2015. [Online]. Available: https://github.com/tyiannak/pyAudioAnalysis/

[16] M. Schmitt and B. Schuller, "Deep Recurrent Neural Networks for Emotion Recognition in Speech," in *DAGA*, 2018, pp. 1537–1540.

[17] S. Hochreiter and J. J. Urgen Schmidhuber, "LONG SHORT-TERM MEMORY," *Mem. Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997. [Online]. Available: http://www7.informatik.tu-muenchen.de/{~}hochreithttp://www.idsia.ch/{~}juergen

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *32nd Int. Conf. Mach. Learn. ICML 2015*, vol. 1, 2015, pp. 448–456.

[19] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural networks Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[20] F. Chollet and Others, "Keras," https://keras.io, 2015.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and Others, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI '16)*, 2016, pp. 265–283.