INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

# Hierarchical LSTMs with Joint Learning for Estimating Customer Satisfaction from Contact Center Calls

*Atsushi Ando, Ryo Masumura, Hosana Kamiyama, Satoshi Kobashikawa, Yushi Aono*

NTT Media Intelligence Laboratories, NTT Corporation, Japan

{ando.atsushi, masumura.ryo, kamiyama.hosana, kobashikawa.satoshi, aono.yushi}@lab.ntt.co.jp

## Abstract

This paper presents a joint modeling of both turn-level and call-level customer satisfaction in contact center dialogue. Our key idea is to directly apply turn-level estimation results to call-level estimation and optimize them jointly; previous work treated both estimations as being independent. Proposed joint modeling is achieved by stacking two types of long short-term memory recurrent neural networks (LSTM-RNNs). The lower layer employs LSTM-RNN for sequential labeling of turn-level customer satisfaction in which each label is estimated from context information extracted from not only the target turn but also the surrounding turns. The upper layer uses another LSTM-RNN to estimate call-level customer satisfaction labels from all information of estimated turn-level customer satisfaction. These two networks can be efficiently optimized by joint learning of both types of labels. Experiments show that the proposed method outperforms a conventional support vector machine based method in terms of both turn-level and call-level customer satisfaction with relative error reductions of over 20%.

**Index Terms**: customer satisfaction, long short-term memory (LSTM), joint learning, hierarchical LSTMs

## 1. Introduction

In contact centers, customer satisfaction is one of the most important quality metrics. It is strongly related to success in sales, loyalty and retention [1–3]. Most contact centers survey customer satisfaction manually by sampling and listening to calls, but costs are high and the survey size is limited. Therefore, automatic estimation of customer satisfaction is an urgent requirement.

Several studies have tackled the automatic estimation of customer satisfaction. There are two large tasks: assessing the entire call and assessment of each customer turn during the call. In this paper, we refer to the former as call-level customer satisfaction estimation and the latter as turn-level customer satisfaction estimation.

Call-level customer satisfaction estimation is the task of classifying individual calls into classes (e.g., *satisfied, neutral, dissatisfied*). Correct classes are decided by several annotators. The conventional methods utilize call-level heuristic features for classification, which include prosodic, lexical and contextual features such as call dominance, existence of product names and sentiment words, and gratitude at the end of the call [4]. Turn-taking features such as pauses and overlaps are also effective in improving classification performance [5]. In contrast to these heuristic features, automatic feature extraction based on convolutional neural networks was recently examined [6].

The other task, turn-level customer satisfaction estimation, categorizes individual turns in a call into several classes. The class categories and class definitions are almost the same as call-level ones. A turn represents a segment determined by speaker's change information. In order to estimate turn-level customer satisfaction, previous studies employed prosodic and lexical features individually extracted from each customer's turn [7–9].

Though conventional methods improved the performance by means of features, two problems remain. First, in both tasks, long-range sequential information is ignored. In turn-level customer satisfaction estimation, conventional methods treat individual turns as being independent even though customer satisfaction in the target turn is related to surrounding turns. In call-level estimation, conventional methods use the statistics of prosodic and lexical features as determined for the entire call, but not their sequences. Second, the relationship between call-level and turn-level customer satisfaction is not considered. In fact, call-level customer satisfaction is highly related to turn-level customer satisfaction. For example, call-level customer satisfaction tends to be positive if turn-level customer satisfaction raises towards the end of the call. However, all conventional methods regarded call-level and turn-level estimation as individual tasks and model them independently.

In this paper, we propose a new customer satisfaction estimation method that utilizes long-range sequential information in a call and the relationship between call-level and turn-level customer satisfaction. Different from conventional methods, we assume that both call-level and turn-level customer satisfaction labels are available. The proposed model is achieved by hierarchically stacking two types of long short-term memory recurrent neural networks (LSTM-RNNs). In the lower network, LSTM-RNN is employed for sequential labelling of turn-level customer satisfaction. This network can use multiple information extracted from not only the target-turn but also that of the surrounding turns for turn-level estimation. In the upper network, another LSTM-RNN uses all of the information of estimated turn-level customer satisfaction in estimating the call-level customer satisfaction label. In order to optimize the proposed model, this paper also presents a joint learning method in which turn-level and call-level estimation are mutually enhanced.

The structure of this paper is as follows. Section 2 presents the dataset used in this paper. The proposed method is described in Section 3. The experimental conditions and results are given in Section 4, and Section 5 provides a summary.

## 2. Dataset

### 2.1. Recording, Annotation and Segmentation

In this paper, we use newly recorded and annotated calls. All of the speakers in the dataset are operators working in a contact center. The task is frozen food selling and includes several sub-

http://dx.doi.org/10.21437/Interspeech.2017-725

Table 1: *Data distribution of training, development and test set.*

| | # of calls | | | # of turns | | |
|---|---|---|---|---|---|---|
| | *pos.* | *neu.* | *neg.* | *pos.* | *neu.* | *neg.* |
| Train. | 85 | 67 | 49 | 700 | 2994 | 1419 |
| Dev. | 8 | 7 | 4 | 83 | 262 | 135 |
| Test | 11 | 13 | 7 | 98 | 499 | 185 |
| Total | 104 | 87 | 60 | 881 | 3755 | 1739 |

tasks such as new orders, inquiries or canceling. First, we set the situation, desired results, and emotion information in each sub-task to make the scenarios. Emotional words like 'gladly' and 'with annoyance' were used as emotion information. These scenarios were checked by operators and the calls that lacked naturalness in terms of emotions were eliminated. We created 89 scenarios. Next, two speakers read the same scenario before talking and decided their roles as operator or customer. They talked via phone sets while following the scenarios but the speech content was created on the fly. The result was 251 recorded calls that included a variety of satisfaction and dissatisfaction types. Total length was 29.7 hours and each call had a talk time from 5 to 12 minutes. All were recorded in stereo, 8 kHz with 16 bit format.

Satisfaction labels for calls and intervals were annotated by three people. All were contact center supervisors, and so had to evaluate customer satisfaction on a daily basis. Every call was annotated by two persons. Each annotator listened to each call twice; first time to assign call label and the second time to give interval labels. Each label is a 5-point Likert degree: very positive, positive, neutral, negative, very negative. Positive includes satisfied, happy, pleased while negative includes dissatisfied, cold/hot anger and frustrated. To harmonize the criteria among the annotators, all listened to sample calls of each satisfaction degree before commencing the annotation.

Turn segmentation and correct label decisions of call-level and turn-level customer satisfaction were made after annotation. The turn segmentation step applied model-based Voice Activity Detection (VAD) to each channel to obtain operator and customer Inter-Pausal Units (IPUs) [10], i.e., automatic segmentation as in [8]. IPUs are defined as consecutive tokens with no gap greater than 200 ms. IPUs less than 1 second or that included contradictory IPU intervals were regarded as backchannels. A continuous string of IPUs without backchannels was taken as a turn. This yielded turns and backchannels of the customer and operator. The label decision step collapsed customer satisfaction categories to assign three labels: *positive* includes very positive and positive, *negative* includes very negative and negative and *neutral* includes neutral (all on 5-point Likert scale). With regard to turn-level customer satisfaction labels, in addition to that criteria, if both annotators assigned *positive* or *negative* to over 50 % of the length of a customer turn, we regarded the turn as *positive* or *negative*, respectively. Call-level customer satisfaction labels were defined similarly: If both annotators assigned *positive* or *negative* to a call, the call was regarded as *positive* or *negative* call, while all the rest were *neutral*. The Cohen's kappa of call-level and turn-level customer satisfaction degrees for the two annotators were 0.68 for call-level and 0.51 for turn-level, which suggests a moderate degree of matching.

These calls and two levels of customer satisfaction labels were divided into training, development, and test sets. Each set was speaker-open. The number of calls and turns in each set is shown in Table 1.

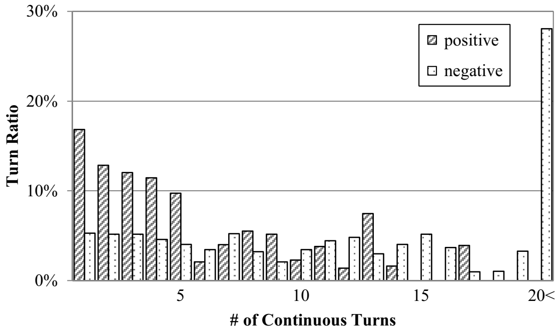

Figure 1: *The ratio of positive and negative turns in the numbers of continuous positive and negative turns (e.g. the ratio of 5 continuous positive turns is 9.7% , which means 9.7% of all positive turns appear within 5 continuous positive turns).*

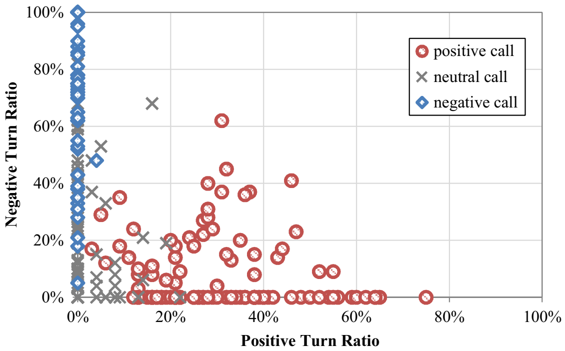

Figure 2: *The ratio of positive and negative turns in each call. Both axes mean the ratio of positive and negative turns to all turns in the call.*

### 2.2. Data analyses

Two analyses of call-level and turn-level customer satisfaction were conducted.

The first analyzed the continuity of turn-level customer satisfaction. The ratio of positive and negative turns in the numbers of continuous positive and negative turns are shown in Figure 1. This figure indicates that less than 20 % of positive turns and 10 % of negative turns were found in single-turn customer satisfaction appearances, while most appeared within continuous positive and negative turns. This indicates that not only the target customer turn but also the surrounding turns contain valid information for turn-level customer satisfaction estimation.

The second analysis examined the relationship between call-level and turn-level customer satisfaction. The ratio of positive and negative turns in each call is shown in Figure 2. This figure shows the clear relationship between call-level and turn-level customer satisfaction; some positive calls included negative turns, whereas very few negative calls contained any positive turns. Furthermore, there were several calls which the ratios of positive/negative turns were the same but call-level customer satisfaction was different. This is due to differences in where the positive/negative turns occurred in the calls. From this analysis, two hypotheses were obtained: First, the sequential information of turn-level customer satisfaction in a call can be directly utilized for call-level estimation. Second, modeling the relationship between turn-level and call-level customer satisfaction will mutually improve the two-level estimation of customer satisfaction.

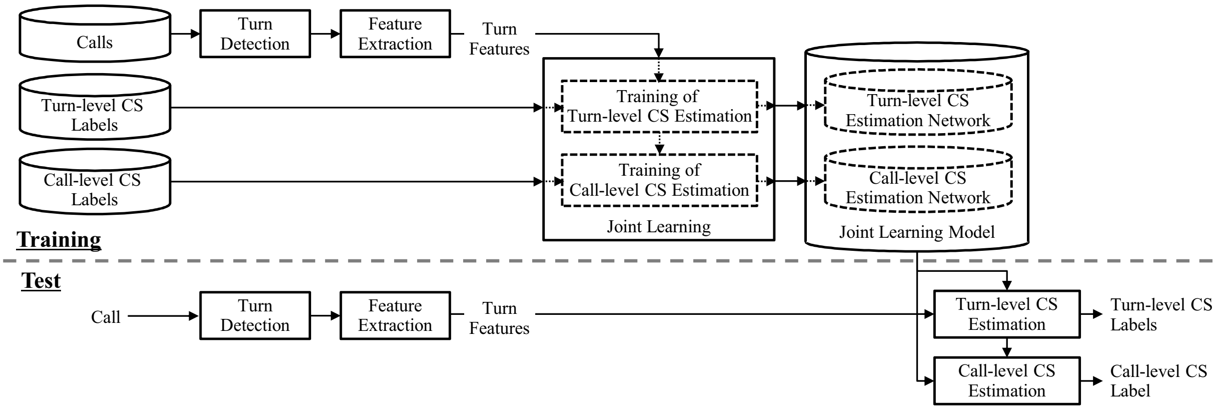

Figure 3: *Overview of the proposed method. CS means customer satisfaction.*

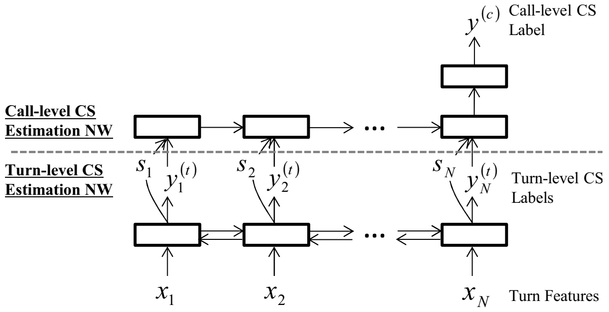

Figure 4: *Model structure of the hierarchical LSTMs.*

# 3. Proposed Method

We propose a new customer satisfaction estimation framework that utilizes contextual information and the relationship of call-level and turn-level customer satisfaction. In this section, we describe the proposed model, called hierarchical LSTMs, and its input features. An overview of the proposed method is shown in Figure 3.

## 3.1. Hierarchical LSTMs

The model of the proposed method is constructed by hierarchically stacking two types of long short-term memory recurrent neural networks (LSTM-RNNs), see Figure 4.

The lower network uses long-range contextual information to sequentially label turn-level customer satisfaction. LSTM-RNN [11] is employed for this network because it enables longer contexts to be modelled than conventional sequential models such as Hidden Markov Model (HMM). Thus it is widely used for sequence labeling problems such as language modeling [12]. The proposed model uses two types of LSTM-RNNs: unidirectional LSTM (uniLSTM) and bidirectional LSTM (biLSTM). The former has a left-to-right architecture and utilizes only previous information for estimation, while the latter has a bidirectional structure and so can deal with both prior and next information.

The upper network is for estimating call-level customer satisfaction from the sequential information of estimated turn-level customer satisfaction. The upper network uses LSTM-RNN with multiple inputs and one output. As shown in Figure 4, the upper network uses not only the estimated turn-level customer satisfaction degrees $y_1^{(t)}, \cdots, y_N^{(t)}$ but also states of the lower network $s_1, \cdots, s_N$, which contain more information than turn-level degrees.

These networks are fused into a single model and optimized by joint learning of both levels of labels to learn the relationship of turn-level and call-level customer satisfaction. Joint learning of multiple tasks is inspired by the Joint Many-task Model [13], which models hierarchical information like part-of-speech tags, words, and meanings. However, it is difficult to achieve stable learning of the turn-level and call-level customer satisfaction estimation networks. One reason is that backpropagation is generally stronger in the upper part of the network. Another is that call-level customer satisfaction is considered to be easier to estimate than turn-level satisfaction. For these reasons, the call-level customer satisfaction estimation network tends to converge before the turn-level network. To offset these convergence differences in each network parameter, we introduce two techniques. First, in the first few epochs, the proposed method trains only the lower network, after which the entire network is updated. This is similar to the pre-training and fine-tuning used by the training acoustic model for speech recognition. Second, the proposed method introduces loss weight $\lambda$ when calculating entire network loss:

$$\lambda L\left(\mathbf{x}, \mathbf{y}^{(t)}; \theta^{(t)}\right) + (1-\lambda)\, L\left(\mathbf{x}, y^{(c)}; \theta^{(t)}, \theta^{(c)}\right), \qquad (1)$$

where $L(\cdot)$ is the loss function, $\mathbf{x} = \{x_1, \cdots, x_N\}$ are turn feature vectors in a call, $\mathbf{y}^{(t)} = \{y_1^{(t)}, \cdots, y_N^{(t)}\}$ and $y^{(c)}$ are correct label(s) of turn-level and call-level customer satisfaction, and $\theta^{(t)}$, $\theta^{(c)}$ are weight parameters of the turn-level and call-level customer satisfaction estimation networks. In the proposed method, $\lambda$ is usually more than 0.5 because the lower network should be impacted more by turn-level labels rather than call-level labels.

## 3.2. Features

The features used in the proposed method include prosodic, lexical and interactive features extracted by both customer and operator. Different from several conventional methods [7, 8], we use manually-selected low dimensional features because it is difficult to train the sequence model when using conventional large dimensional features given the small training data sets available.

Prosodic features include the information of fundamental frequency (F0), loudness, and speech rate. It has 21 dimensions: mean, std., max, min, range and ratio of start/end 500 ms mean to entire mean of customer turn log F0, mean, std., and max of loudness, mean, std., max, min of first derivative of log

F0 and loudness, speech rate of customer turn and previous operator turn (mora/sec), and duration of the end of phoneme of the target customer turn. The speech rates and duration are obtained by speech recognition.

As the lexical features, we use Bag-of-Words (BoW) of specific words in customer or operator turns. To reduce task dependency, target words are selected by hand. The lexical features have 12 dimensions: total number of words in the target customer turn or previous operator turn, number of filler words, backchannel words, appreciation and its emphasis words (e.g. *'kindly'*), personal pronoun in the target customer turn, number of filler words, backchannel words, appreciation, humility, and apology words in the previous operator turn. Filler words include *'uh'* or *'hmm'* and backchannel words are *'hai (yes in English)'*, *'wakarimashita (I see)'*, etc. The words are identified in the speech recognizer's output.

Interactive features include turn taking, pause and backchannel information. In addition to the conventional methods [4, 5], the proposed method utilizes the characteristics of backchannels around target customer turn. The proposed method uses 11 dimensional interactive features: length of the target customer turn and the previous operator turn, length of pause between the target customer turn and the previous/next operator turn, length of interval between the target and previous customer turn, the ratio of length of the target customer turn to the sum of previous operator and target customer turn, frequency and average length of customer backchannels, average number of repeated words in backchannels, and the ratio of average F0 of customer backchannels to customer turn. Backchannels are automatically determined by the segmentation method shown in Section 2. The customer backchannels of the target turn are defined as those of between previous and target customer turns.

# 4. Experiments

## 4.1. Experimental Setup

The dataset shown in Section 2 was used to evaluate the performance of the proposed method. For feature extraction, frame length of F0 and loudness were set at 64 ms and 5 ms shift, respectively. F0 extraction method was based on dominant harmonic components [14]. A DNN acoustic model with a large-vocabulary WFST language model was used to obtain words for lexical features. Turn features were normalized to zero mean and unit variance by the training set.

A Support Vector Machine (SVM) with radial basis function kernel was used as the classifier of the baseline method. The inputs of the baseline were individual turn features for turn-level customer satisfaction estimation and statistics of turn features in a call for call-level estimation.

For the proposed method, 4 network variants were evaluated: uni/bi-LSTM for turn satisfaction network, and with/without joint learning. Without joint learning, we employed two-step training: First, the turn-level customer satisfaction estimation network was trained by turn-level labels. Second, the call-level estimation network was trained by call-level labels with posteriors and LSTM layer outputs of the turn-level network. In the second step, losses of call-level network were used only for training the upper network; they were not propagated to the lower network. The structure of the turn-level and call-level customer satisfaction estimation network is a 1 layer LSTM with 128 units and a 1 layer LSTM with 64 units, respectively. Minibatch size was 3 calls. Loss function is softmax cross entropy and Adam [15] is used for optimization. In joint learning, starting epoch of entire network updating was varied from 1 to 10. Loss weight ranged from 0.5 to 0.9. Final results were obtained by taking the average of 5 differential initial weights in all variants of the proposed method. Early-stopping was triggered by the losses of the development set. LIBSVM [16] and Chainer [17] were used to implement conventional SVM and proposed model, respectively. Performance comparisons used macro F1, which is the average F-measures of all classes.

Table 2: *Macro F1s of turn-level and call-level customer satisfaction estimation. The column of turn NW means the structure of the network of turn-level customer satisfaction estimation.*

| Methods | turn NW | joint learning | Turn | Call |
|---|---|---|---|---|
| *Random* | | | 0.299 | 0.328 |
| SVM | | | 0.490 | 0.534 |
| Proposed | uniLSTM | | 0.524 | 0.681 |
| | uniLSTM | ✓ | 0.552 | 0.696 |
| | biLSTM | | 0.590 | 0.671 |
| | biLSTM | ✓ | **0.611** | **0.710** |

## 4.2. Results

Results are shown in Table 2. For both variants with joint learning, starting epoch of entire network updating was 5 and loss weight was 0.8. *Random* is the method that selects turn-level and call-level labels randomly.

With regard to turn-level customer satisfaction estimation, all variants of the proposed method achieved higher performance than conventional SVM. This indicates that long-range sequential information is effective for turn-level customer satisfaction estimation, especially when using a bidirectional LSTM to utilize both previous and next information. In the case of call-level estimation, the proposed method showed better macro F1s than SVM, even without joint learning. It is considered that the sequences of estimated turn-level customer satisfaction are useful for call-level estimation. Finally, the proposed method with joint learning outperformed the variant without joint learning in terms of both turn-level customer satisfaction and call-level customer satisfaction. This indicates that sharing knowledge among hierarchical LSTMs with joint learning improves both turn-level and call-level customer satisfaction estimation. The highest improvement was achieved by the proposed method using bidirectional LSTM with joint learning, 0.176 in call-level customer satisfaction and 0.121 in turn-level customer satisfaction, which is a great improvement because the relative error reductions exceeded 20%.

# 5. Conclusion

In this paper, we presented a joint modeling of both turn-level and call-level customer satisfaction in contact center dialogue. The proposal was achieved by stacking two types of LSTM-RNNs. In the lower network, LSTM-RNN was employed for sequential labeling of turn-level customer satisfaction. In the upper network, another LSTM-RNN was used to estimate call-level customer satisfaction labels from the information of estimated turn-level customer satisfaction. These two networks were efficiently optimized by joint learning of both labels. Experiments showed that the proposed method outperformed the conventional SVM-based method in terms of both turn-level and call-level customer satisfaction estimation.

## 6. References

[1] R. Hallowell, "The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study," *International journal of service industry management*, vol. 7, no. 4, pp. 27–42, 1996.

[2] C. Ranaweera and J. Prabhu, "The influence of satisfaction, trust and switching barriers on customer retention in a continuous purchasing setting," *International journal of service industry management*, vol. 14, no. 4, pp. 374–395, 2003.

[3] M. Amin, Z. Isa, and R. Fontaine, "The role of customer satisfaction in enhancing customer loyalty in malaysian islamic banks," *The Service Industries Journal*, vol. 31, no. 9, pp. 1519–1532, 2011.

[4] Y. Park and S. C. Gates, "Towards real-time measurement of customer satisfaction using automatically generated call transcripts," in *Proc. of ACM*, 2009, pp. 1387–1396.

[5] S. A. Chowdhury, E. A. Stepanov, and G. Riccardi, "Predicting user satisfaction from turn-taking in spoken conversations," in *Proc. of INTERSPEECH*, 2016.

[6] C. Segura, D. Balcells, M. Umbert, J. Arias, and J. Luque, "Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls," in *Proc. of IberSPEECH*, 2016, pp. 255–265.

[7] L. Devillers, C. Vaudable, and C. Chastagnol, "Real-life emotion-related states detection in call centers: a cross-corpora study." in *Proc. of INTERSPEECH*, 2010, pp. 2350–2353.

[8] C. Vaudable and L. Devillers, "Negative emotions detection as an indicator of dialogs quality in call centers," in *Proc. of ICASSP*, 2012, pp. 5109–5112.

[9] N. Kamaruddin, A. W. A. Rahman, and A. N. R. Shah, "Measuring customer satisfaction through speech using valence-arousal approach," in *Proc. of ICT4M*, 2016, pp. 298–303.

[10] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," *Language and speech*, vol. 41, no. 3–4, pp. 295–321, 1998.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] M. Sundermeyer, R. Schlüter, and H. Ney, "LSTM neural networks for language modeling." in *Proc. of INTERSPEECH*, 2012, pp. 194–197.

[13] K. Hashimoto, C. Xiong, Y. Tsuruoka, and R. Socher, "A joint many-task model: Growing a neural network for multiple NLP tasks," *arXiv preprint arXiv:1611.01587*, 2016.

[14] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *The Journal of the Acoustical Society of America*, vol. 116, no. 6, pp. 3690–3700, 2004.

[15] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[17] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proc. of NIPS LearningSys*, 2015.