SEMI-SUPERVISED ENSEMBLE DNN ACOUSTIC MODEL TRAINING

Sheng Li[1]*, Xugang Lu*[2]*, Shinsuke Sakai*[1]*, Masato Mimura*[1] *and Tatsuya Kawahara*[1]

[1] School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[2] National Institute of Information and Communications Technology, Kyoto, Japan
lisheng@sap.ist.i.kyoto-u.ac.jp

ABSTRACT

It is very important to exploit abundant unlabeled speech for improving the acoustic model training in automatic speech recognition (ASR). Semi-supervised training methods incorporate unlabeled data in addition to labeled data to enhance the model training, but it encounters the error-prone label problem. The ensemble training scheme trains a set of models and combines them to make the model more general and robust, but it has not been applied to the unlabeled data. In this work, we propose an effective semi-supervised training of deep neural network (DNN) acoustic models by incorporating the diversity among the ensemble of models. The resultant model improved the performance in the lecture transcription task. Moreover, the proposed method has also shown a potential for DNN adaptation.

Index Terms— Speech recogntion, Acoustic model, DNN, Semi-supervised training

1. INTRODUCTION

For many automatic speech recogntion (ASR) tasks, there are usually limited amount of labeled speech for training acoustic models but often abundant unlabeled speech which requires much human efforts and expertise to label it. It is important to utilize the unlabeled speech for improving the acoustic model. In this work, we focus on effective training of deep neural network (DNN) acoustic models with limited labeled speech and abundant unlabeled speech.

Semi-supervised learning methods [1] exploit unlabeled data in addition to labeled data, where no human intervention is assumed, thus it has become an important topic. Typically, a seed model is trained with the labeled data to automatically transcribe the unlabeled data. In this setting, multiple seed models can be used, and their hypotheses are combined to improve the transcription accuracy [2].

The ensemble training scheme trains a set of models with the same data set and combines them or their prediction results during testing [3]. By incorporating diversity (=prediction differences) among the models during training, the resultant model will be generalized and robust [4]. However, conventional ensemble training has not been used to utilize the unlabeled data.

In this paper, we propose a generalized ensemble training method using both labeled and unlabeled data. Ensemble training can enhance the semi-supervised training by regarding the inconsistency of labels among different seed models as diversity to be leveraged. An ensemble of models is trained in parallel using diverse labels for unlabeled data. Together with the standard cross-entropy, the KL divergence between these models in ensemble is incorporated into the training objective function. The proposed method is evaluated on the lecture transcription task and the DNN adaptation task.

The rest of this paper is organized as follows. Section 2 formulates the proposed semi-supervised ensemble training. Section 3 describes its implementation. Section 4 and Section 5 presents evaluations of the proposed method. Conclusions are given in Section 6.

2. SEMI-SUPERVISED ENSEMBLE MODEL TRAINING

2.1. Semi-supervised Training and Ensemble Training

In the most commonly-used semi-supervised training of DNN acoustic model [2, 5, 6, 7, 8, 9], a seed model is trained with the labeled data to automatically transcribe the unlabeled data. In this setting, the ASR result is essentially error-prone and its quality significantly affects the overall performance. One of the solutions is to set up multiple seed models, so that a multi-system combination can improve the transcription accuracy [2].

In addition, data selection is also adopted, because the DNN training is sensitive to the errors in the label. The data selection can be conducted in different levels. The most widely-used method is the utterance-level selection, which sorts the utterances by an utterance-level confidence measure score (CMS) and selects a certain percentage of top utterances for model training [5].

When we have a frame-level CMS, it is possible to perform frame-level data selection (frame dropping [7]) in the fine-tuning step of DNN over frame-level mini-batches. Alternatively, by viewing the high-confidence data and the low-

confidence data as different resources, a multi-task training architecture inspired from multi-lingual modeling [10, 11, 12, 13, 14, 15] can be applied for semi-supervised training. However, these filtering methods cannot entirely solve the problem of errors in the label, and they will significantly reduce the amount of usable training data.

The ensemble training method can effectively improve the performance by training a set of individual models on the same data set, and either combining the models or their prediction results during testing [3]. Diversity is an important factor in ensemble learning. It represents the prediction differences between any pair of component models in the ensemble of models. It can be measured by either Euclidean distances, KL divergence, or some other metrics. By incorporating diversity among models during training, the resultant model becomes more general and robust [4].

Deng et al. [16] developed the linear and log-linear stacking methods for ensemble learning with class posterior probabilities computed by the convolutional, recurrent, and fully-connected deep neural networks. Experimental results demonstrated a significant increase in phone recognition accuracy. Recently, Zhang et al. [17] trained an ensemble of differently randomly initialized networks by introducing a penalty term of KL divergence between each individual DNN output and their average output. The method shows good results for a low-resource speech recogntion task and the TED spoken lecture transcription task. According to [18, 19], the prediction error for labeled data in the objective function can be regularized by the unlabeled data. We introduce diversity among models through the use of different ASR-based labels for unlabeled training data. On the other hand, individual models in the ensemble are encouraged by the objective function to get closer to the ensemble average in the training time as in [17], in order to stabilize the training. There is past work showing that it is possible to train a small DNN model to learn the predictions of a large DNN model [20] or RNN model [21] by minimizing the KL divergence between the output distributions on unlabeled data. Our work can be regarded as an extension by using a set of models to train a single DNN model.

2.2. Proposed Method

We propose a generalized ensemble training method using both labeled and unlabeled data as shown in Fig. 1.

We have two data sets L and U as follows: L=$\{(x_1, y_1), (x_2, y_2), \cdots, (x_{|L|}, y_{|L|})\}$ is the labeled data set, where $|L|$ is the number of labeled samples. For any $(x_i, y_i) \in L$, $(i = 1, 2, \cdots, |L|)$, y_i is the unique label for the feature vector x_i. Here y_i is represented with an N-dimentional vector $y_i = (y_{i1}, \cdots, y_{iN})^T$, where N is the the number of classes (senones) and only one element is 1 and others are 0. U=$\{x_1^{'}, x_2^{'}, \cdots, x_{|U|}^{'}\}$ is the unlabeled training data set, where $|U|$ is the number of unlabeled samples. For any feature vector $x_j^{'} \in U$, $(j = 1, 2, \cdots, |U|)$, we have M labels $(\hat{y}_{1j}, \hat{y}_{2j}, \cdots, \hat{y}_{Mj})$ generated by M different seed ASR systems trained using L.

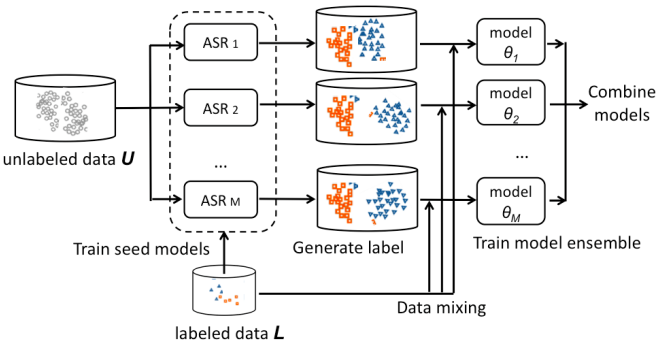

Fig. 1: Flow-chart of the proposed method.

Then we train a model ensemble $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_M)$ with the complete data set $L \cup U$. Each of θ_m is the parameters of a DNN model $(m = 1, \cdots, M)$, and it can be regarded as a non-linear function mapping the input feature vector x_i to the senone class posterior probabilities $\boldsymbol{s}(x_i; \theta_m)$. $\boldsymbol{s}(x_i; \theta_m) \triangleq (P(s_1|x_i; \theta_m), \cdots, P(s_N|x_i; \theta_m))^T$ is the posterior probability distribution over all senone classes given feature x_i and model θ_m $(m = 1, \cdots, M)$.

Unlike using differently initialized DNN models in [17], the model diversities are derived from the diverse labels of U. In addition, these model parameters (both weight matrices and bias vectors) are periodically averaged,

$$\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \theta_m. \tag{1}$$

This averaged model is used for a new start of parallel training of each model.

We try to find a set of parameters to minimize the objective function,

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, L, U), \tag{2}$$

which is defined as:

$$V(\boldsymbol{\theta}, L, U) = V_e(\boldsymbol{\theta}, L) + (1 - \lambda) V_e(\boldsymbol{\theta}, U) + \lambda V_d(\boldsymbol{\theta}, U), \tag{3}$$

where V_e is defined as the empirical loss measured on the training set, V_d is defined as the diversity loss measured between the ensemble classifiers and λ is a tunable tradeoff parameter. They are further defined as follows:

$$\begin{aligned} V_e(\boldsymbol{\theta}, L) &= \sum_{m=1}^{M} \{ \sum_{(x_i, y_i) \in L} CE(y_i, \boldsymbol{s}(x_i; \theta_m)) \} \\ &= \sum_{m=1}^{M} \{ - \sum_{(x_i, y_i) \in L} \sum_{n=1}^{N} y_{in} \log P(s_n|x_i; \theta_m) \} \end{aligned}, \tag{4}$$

where Eq. (4) is the standard cross-entropy objective for the training set L.

Similarly, we have Eq. (5) for unlabeled data set U.

$$V_e(\boldsymbol{\theta}, U) = \sum_{m=1}^{M} \{ \sum_{x_j^{'} \in U} CE(\hat{y}_{mj}, \boldsymbol{s}(x_j^{'}; \theta_m)) \}, \tag{5}$$

where $\hat{y}_{mj}$ is the label of feature $x_j^{'}$ automatically generated by the m-th seed ASR system.

We incorporate the regularization term of divergence between ensemble models, which would in effect penalize the inconsistency between the labels of M different systems. Instead of directly using the KL divergences or the Euclidean distances between the pair of component models, we use a one-sided KL divergence between each model and the averaged model,

$$V_d(\boldsymbol{\theta}, U) = \sum_{m=1}^{M} \{ \sum_{x_j^{'} \in U} KL(\boldsymbol{s}(x_j^{'}; \bar{\theta}) || \boldsymbol{s}(x_j^{'}; \theta_m)) \}, \quad (6)$$

where the $\bar{\theta}$ is the newly averaged model.

Suppose $h(x)$ and $g(x)$ represent N-class output distributions, we have their KL divergence as follows:

$$KL(h(x)||g(x)) = CE(h(x), g(x)) - H(h(x)), \quad (7)$$

where the KL divergence can be expressed with the cross-entropy and the self-entropy. Thus, we have

$$\arg\min_{g(x)} KL(h(x)||g(x)) = \arg\min_{g(x)} CE(h(x), g(x)). \quad (8)$$

So we rewrite Eq. (6) just using the cross-entropy term in the objective function.

$$V_d(\boldsymbol{\theta}, U) = \sum_{m=1}^{M} \{ \sum_{x_j^{'} \in U} CE(\boldsymbol{s}(x_j^{'}; \bar{\theta}), \boldsymbol{s}(x_j^{'}; \theta_m)) \}. \quad (9)$$

3. EXPERIMENTAL IMPLEMENTATION

The proposed semi-supervised ensemble DNN model training is implemented as shown in Fig. 2. It consists of two stages: diverse label generation stage and ensemble model training stage.

3.1. Diverse Label Generation Stage

In the proposed method, model diversity is realized by providing different label sets for unlabeled training data. We first train M different seed ASR systems using the labeled data set L. Then, the ASR systems generate different labels for the unlabeled data U with decoding and forced alignment. Finally, M sets of unlabeled data U with different unfaithful label sets are shuffled into mini-batches with the labeled data set L.

3.2. Ensemble Model Training Stage

The ensemble of models are trained with different labels in different GPUs. And the model parameters are updated every mini-batch in parallel and periodically averaged every ten mini-batches, then redistributed to different GPUs as a new start of parallel training. After 15 to 20 epochs of training, the averaged model is used as the final model. The ensemble training framework is modified from Kaldi toolkit (nnet2) [22].

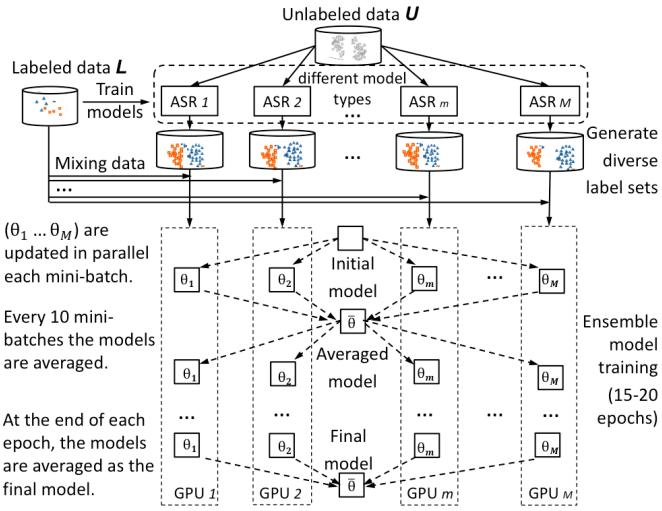

Fig. 2: Implementation of the proposed method. (the solid arrow is data flow, the dashed arrow is model copy)

4. EVALUATION ON LECTURE TRANSCRIPTION

The proposed method is evaluated on the Corpus of Chinese Lecture Room (CCLR) [23]. All the data sets are listed in Table 1.

The dictionary consists of 53K lexical entries. A word trigram language model (LM) was built for decoding by using transcriptions of the training data of CCLR with the LDC corpora and the Phoenix lecture archive. We first build a GMM-HMM system and then a DNN-HMM system. The GMM system uses PLP features. It is trained with the MPE criterion. In addition, we conduct unsupervised speaker adaptation using MLLR for each lecture, which is effective for long lecture speech. The baseline DNN model uses filterbank features. We use Kaldi toolkit (nnet1) [24] and the training is based on the CE criterion. For decoding, we use Julius ver.4.3.1 (DNN version) [25] using the state transition probabilities of the GMM-HMM.

In this paper, we train an ensemble of two models using the different labels generated from GMM-HMM and DNN baseline systems. It is also possible to use more seed ASR systems using different types of DNN and/or different acoustic features, but it turned out these two (GMM and baseline DNN) have the largest diversity (edit distance of 24.3%). This set is also an economical choice considering the computational resources. In our previous work [14], we used unlabeled data of 114.7 hours to enhance the baseline system by semi-supervised training. We first combine the hypotheses generated from the two baseline systems (**hypothesis combination**), and then conduct several data selection methods (**utterance selection**, **frame dropping**, **multi-task training**) as explained in Section 2.1. More detail is found in [14].

We tested a variation of weight λ, which controls the two terms in Eq. (3). ASR performances of the models enhanced

Table 1: Data Sets of CCLR

	#Lectures	Hours
Training set (labeled)	184	97.2
Training set (unlabeled)	184	114.7
Development set	12	7.2
Test set	19	11.9

by these methods are evaluated on both of the development set and the test set. Table 2 shows that our proposed method (**Ensemble**) outperforms other semi-supervised methods significantly in any values of λ=0.1, $\cdots$, 0.7. The performance becomes best when λ=0.5. We also notice that the ensemble training without regularization (λ=0.0) is comparable to the hypothesis combination method (HC) without data selection.

Table 2: ASR performances (CER%) of different semi-supervised training methods of DNN

	Training sets (Hours)		CER%	
	labeled	unlabeled	Dev	Test
Baseline GMM	97.2	0	24.2	27.5
Baseline DNN	97.2	0	22.7	25.7
Hypo. Combine (HC)	97.2	114.7	21.5	24.4
HC+utterance selection	97.2	71.5	21.3	24.2
HC+frame dropping	97.2	90.4	21.4	24.3
HC+multitask training	97.2	114.7	21.3	24.3
Ensemble (λ=0.0)	97.2	114.7	21.5	24.3
Ensemble (λ=0.1)	97.2	114.7	20.9	23.7
Ensemble (λ=0.3)	97.2	114.7	20.7	23.6
Ensemble (λ=0.5)	97.2	114.7	**20.6**	**23.6**
Ensemble (λ=0.7)	97.2	114.7	20.7	23.8

We also train an ensemble of four models using the different labels generated from a CNN model and a RNN model with GMM-HMM and DNN baseline systems. The CNN (filterbank, 2CNN+4DNN, Dev 21.5%, Test 24.4%) and the RNN (filterbank, 2LSTM, Dev 22.0%, Test 25.6%) model are trained from the labeled data of 97.2 hours. We obtained a slight improvement. CER% is Dev 20.3% and Test 23.2% when λ=0.5.

5. EVALUATION ON SPEAKER ADAPTATION

Next, we investigate the proposed method in the speaker adaptation setting. Since the simplest and most effective speaker adaptation (**Conservative retraining**) for DNN models is retraining the DNN over the adaptation data of the speaker [26], we apply our proposed method to unsupervised DNN adaptation (**Ensemble**).

We use the CHiME-3 challenge [27] data set. The training set is tr05-multi-noisy, which consists of 1600 real noisy utterances from 4 speakers in 4 noisy environments, and 7138 simulated noisy utterances from the 83 speakers forming the WSJ0 SI-84 training set in the 4 noisy environments. For testing, we use et05-real-noisy, which consists of 1320 utterances from 4 different speakers on 4 environments.

A DNN-HMM hybrid system is trained using the 40-dimension fMLLR transformed feature (MFCC feature transformed using LDA+MLLT before SAT training). The DNN was trained based on the sMBR criterion. We also trained a CNN model based on the CE criterion with the 40-dimension filterbank feature. The edit distance of their recognition results is 22.6%. CHiME challenge provided only MLE-trained GMM-HMM model as a baseline, but its accuracy is too low (WER of 32.9%) for unsupervised adaptation. We applied our proposed ensemble training method to retrain the baseline model (sMBR DNN) for each speaker using the test data and the baseline ASR results. An ensemble of two models were trained with diverse label sets generated by the baseline sMBR DNN and CNN models.

We compare our proposed method (**Ensemble**) to the baseline and the **Conservative retraining** method with and without KL divergence regularization [28]. For each testing speaker, we retrain the baseline sMBR DNN model with a fixed learning rate (0.02) [29]. We applied KL divergence regularization [28] with an weight of 0.1.

In Table 3, the results are significantly improved by the proposed method (**Ensemble**) compared with **Conservative retraining**. This demonstrates the proposed method is also effective for unsupervised speaker adaptation.

Table 3: ASR performance (WER%) of DNN adaptation

	et05-real-noisy
Baseline DNN (7DNN, sMBR)	22.7
Baseline CNN (1CNN+4DNN)	22.6
Conservative retraining	22.0
Conservative retraining (w/ KLD reg.)	21.7
Ensemble (λ=0.0)	21.8
Ensemble (λ=0.1)	21.5
Ensemble (λ=0.3)	**21.1**
Ensemble (λ=0.5)	21.5

6. CONCLUSION

We investigate an application of ensemble training to semi-superviesd training of DNN acoustic models. By incorporating a diversity term in the objective function, the resultant model mitigates the inconsistency of the labels and improved the performance. Moreover, the proposed method also has a potential for DNN adaptation. For the future planning, we will investigate effective use of a larger number of models for ensemble training.

7. REFERENCES

[1] O. Chapelle, B. Schlkopf, and editors A. Zien, *Semi-Supervised Learning*, MIT Press, Cambridge, MA, 2006.

[2] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration," in *Proc. INTERSPEECH*, 2013, pp. 2360–2364.

[3] T. G. Dietterich, *Ensemble methods in machine learning*, Springer, 2000.

[4] Z. Zhou, *Ensemble Methods: Foundations and Algorithms*, Chapman & Hall/CRC Press, 2012.

[5] K. Yu, M. Gales, L. Wang, and P. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7, pp. 652–663, 2010.

[6] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *Proc. IEEE-ASRU*, 2013, pp. 368–373.

[7] K. Vesely, M. Hannemann, and L. Burget, "Semi-supervised training of deep neural networks," in *Proc. IEEE-ASRU*, 2013, pp. 267–272.

[8] D. Imseng, B. Potard, P. Motlicek, A. Nanchen, and H. Bourlard, "Exploiting un-transcribed foreign data for speech recognition in well-resourced languages," in *Proc. IEEE-ICASSP*, 2014.

[9] P. Zhang, Y. Liu, and T. Hain, "Semi-supervised DNN training in meeting recognition," in *Proc. IEEE-SLT*, 2014.

[10] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Proc. IEEE-ICASSP, pp. 8619-8623*, 2013.

[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," in *Proc. IEEE-ICASSP, pp. 7304-7308*, 2013.

[12] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "Semi-supervised maximum mutual information training of deep neural network acoustic models," in *Proc. INTERSPEECH*, 2008.

[13] K. Vesely, M. Karafit, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proc. IEEE-SLT*, 2012.

[14] S. Li, Y. Akita, and T. Kawahara, "Semi-supervised acoustic model training by discriminative data selection from multiple ASR systems hypotheses," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 24, no. 9, pp. 1520–1530, 2016.

[15] H. Su and H. Xu, "Multi-softmax deep neural network for semi-supervised training," in *Proc. INTERSPEECH*, 2015.

[16] L. Deng and J. Platt, "Ensemble Deep Learning for Speech Recognition," in *Proc. INTERSPEECH*, 2014.

[17] X. Zhang, D. Povey, and S. Khudanpur, "A diversity penalizing ensemble training method for deep learning," in *Proc. INTERSPEECH*, 2015.

[18] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[19] P. Niyogi, "Manifold regularization and semi-supervised learning: Some theoretical analyses," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 1229–1250, 2013.

[20] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, "Learning small-size DNN with output-distribution-based criteria," in *Proc. INTERSPEECH*, 2014.

[21] W. Chan, N. Rosemary Ke, and I. Lane, "Transferring knowledge from a RNN to a DNN," in *Proc. INTERSPEECH*, 2015.

[22] D. Povey, X. Zhang, and S. Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," in *Proc. ICLR Workshop*, 2015.

[23] S. Li, Y. Akita, and T. Kawahara, "Corpus and transcription system of Chinese lecture room," in *Proc. ISCSLP*, 2014.

[24] D. Povey et al., "The Kaldi speech recognition toolkit," in *Proc. IEEE-ASRU*, 2011.

[25] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proc. APSIPA ASC*, 2009, pp. 131–137.

[26] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer, 2012.

[27] T. Hori, Z. Chen, H. Erdogan, J. R. Hershey, J. Le Roux, V. Mitra, and S. Watanabe, "The MERL/SRI system for the 3rd CHiME challenge using beamforming, robust feature extraction, and advanced speech recognition," in *Proc. IEEE-ASRU*, 2015.

[28] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. IEEE-ICASSP*, 2013, pp. 7893–7897.

[29] T. Yoshioka et al., "The NTT CHiME-3 system: advances in speech enhancement and recognition for mobile multi-microphone devices," in *Proc. IEEE-ASRU*, 2015.