Cross-Modal Knowledge Distillation with Dropout-Based Confidence

Won Ik Cho[1], Jeunghun Kim[2], and Nam Soo Kim[2]
Samsung Advanced Institute of Technology, Samsung Electronics[1]
Department of Electrical and Computer Engineering and INMC, Seoul National University[2]
E-mail: wonik.cho@samsung.com, jhkim@hi.snu.ac.kr, nkim@snu.ac.kr

***Abstract*—In cross-modal distillation, e.g., from text-based inference modules to spoken language understanding module, it is difficult to determine the teacher's influence due to the different nature of both modalities that bring the heterogeneity in the aspect of uncertainty. Though error rate or entropy-based schemes have been suggested to cope with the heuristics of time-based scheduling, the confidence of the teacher inference has not been necessarily taken into deciding the teacher's influence. In this paper, we propose a dropout-based confidence that decides the teacher's confidence and to-student influence of the loss. On the widely used spoken language understanding dataset, Fluent Speech Command, we show that our weight decision scheme enhances performance in combination with the conventional scheduling strategies, displaying a maximum 20% relative error reduction concerning the model with no distillation.**

***Index Terms*—dropout, uncertainty, cross-modal distillation, spoken language understanding**

I. Introduction

Knowledge distillation (KD) is mainly adopted to transfer information from a large-scale high-performance module to a small-scale low-performance module [1]. For the case where the teacher and student have the same input modality, it has been widely discussed on the proper information that should be conveyed in the transfer process. This has been handled not only in image processing [2], but also in speech processing such as automatic speech recognition (ASR) [3] and in textual comprehension tasks such as question answering [4].

Some prior arts have tackled this problem in a cross-modal manner. Take speech and text as an example; though one regards audio signals and the other concerns (digitized) letters, they constitute a case in which the different modalities convey the relevant semantics. This includes when the student is fed speech while the teacher is trained with text or vice versa. For instance, in spoken language understanding (SLU) [5], the conventional natural language understanding (NLU) module shares the same objective with the original SLU task. This makes the problem more a 'cross-modal' especially given that some SLU systems are implemented in an end-to-end manner, free from the ASR transcription.

However, in these cross-modal tasks, how the knowledge should be distilled is not easily decidable. If the teacher and student share the output, the distillation can first be considered as conveying logit-level information. The point here is that the uncertainty aspect of the speech-based model's inference may differ from that is observed in the text-based one (Fig. 1), due to the different nature of the two modalities. Therefore, for the knowledge transfer between the two modules to be guidance rather than a one-way projection, it is necessary to control the teacher influence.

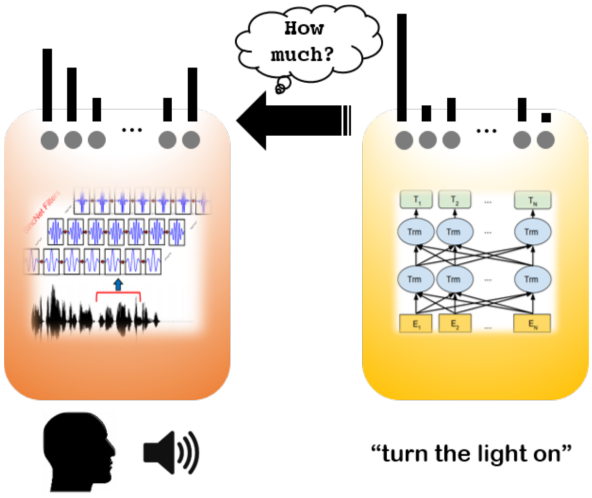

Fig. 1. Concept of deciding the teacher's influence in cross-modal knowledge distillation. The figure in each module is from SincNet [6] and BERT [7] respectively.

In the quantitative approach to this problem [8], various strategies such as exponential decaying, triangular scheduling, and error rate [3] were described. Such strategies are based on the heuristic hyper-parameter setup or the student model's performance. However, they hardly provide information on whether the probability distribution is reliable, or how confident the teacher prediction itself can be. We surmised that more attention on the teacher output is required, which had frequently been treated as a source of uncertainty.

In this paper, we investigate the methodologies that automatically determine the influence of the teacher model in a cross-modal scenario, concerning the teacher inference. We first introduce the concept of cross-modal KD in SLU, and then check the temporal scheduling and weight decision schemes available. Based on the viewpoints not handled in the above approaches, we scrutinize the methodology that quantifies the confidence the teacher's inference incorporates.

The contribution of this research is as follows:

- We propose dropout-based confidence modeling that can automatically decide the teacher's influence in cross-modal KD.
- We show that the proposed scheme works solid and boosts the scheduling strategies better than error rate and entropy-based methodologies.

II. Related Work

In general, KD loss is added to the original loss [2], [9], [10].

$$\begin{aligned} (a) \quad L &= (1-\lambda) * L_{orig} + \lambda * L_{kd} \\ (b) \quad L &= L_{orig} + \lambda * L_{kd} \end{aligned} \tag{1}$$

where either (1a) by a weighted sum or (1b) merely using a simple addition. The settings with fixed λ generally worked well in previous text-based or cross-modal studies [10], [11]. However, considering that KD is sensitive to various contexts such as data, task, and loss function, one may prefer to adopt a time-varying λ. It then becomes essential to understand how such scheduling affects distillation.

Accordingly, in recent distillation schemes, there have been attempts that employ diverse teachers [12], [13] or manage the influence of teachers using scheduling [13]. Beyond the popular vision domain [12], the trend is widely observed in applications that use large models, such as ASR [3] and natural language tasks [13] which benefits from recent high-performance pre-trained models such as BERT [7].

In the above cases, the input data modalities are usually identical, so it is advantageous for the student model to follow the observable uncertainty in the teacher inference. However, in an environment where input modalities are different, the student model may not benefit directly from the teacher inference [8]. For example, the distribution of SLU module inference when speech input contains much noise, and the tendency of its teacher induced by ambiguous or confusing text input, will not necessarily be similar. Here, setting λ by just scheduling may not be ideal in that it does not take into account the teacher or student's performance. One solution of adopting the student performance such as word error rate (WER) was suggested in Kim et al. [3] and was used for SLU in Cho et al. [8], but the approach does not concern teacher uncertainty and can inadvertently ignore such different natures of modalities.

A fundamental way to reflect the teacher uncertainty in distillation was tackled in Kwon et al. [14] by finding the entropy of the output probability distribution of the teacher model. Adapting the methodology to fit with our task, the entropy is defined as follows:

$$H(T) = -\sum_i t_i log(t_i) \tag{2}$$

for t_i in T, where T is a softmax teacher output. Though it is deemed sufficient to reflect the uncertainty of teacher inference, entropy is not a normalized concept, and in the case of classification it depends on the number of output classes. Thus, we rewrite as:

$$H_C(T) = \frac{H(T)}{log(C)} \tag{3}$$

where C is the number of output classes. Nevertheless, it also seems to be difficult to compensate for the difference in tendency from the variance of the number of output classes.

III. Proposed Method

To address the issues, we measure the reliability of prediction by assigning noise to the teacher output and checking the robustness accordingly. In detail, we investigate how the teacher output can become distant from the original distribution when exposed to perturbation.

The perturbation defined here adopts dropout layer [15]. In general, dropout is utilized to prevent overfitting in the training process of the neural network model. However, the characteristic that the weight of certain nodes are redistributed to the rest allows the dropout to be exploited as a kind of perturbation layer. We surmised that if a dropout is augmented to the teacher model's output layer, the resulting final layer will be a sort of skewed output, where the perturbation is applied to the original teacher inference. This kind of formulation had been done with dropout distillation in Bulo et al. [16] and distilled dropout network (DDN) in Gurau et al. [17], to prevent the overconfidence of the model training, and we want to claim that it can also be useful in cross-modal KD, in modeling the teacher 'confidence'.

However, it is difficult to grasp the uncertainty of teacher prediction with only one perturbation. Thus, we build multiple and parallel dropout layers, to make up various skewed outputs and compare them with the centroid (the original teacher prediction). In detail, for teacher output T, given N dropout layers q_n, we can write each skewed output as $q_n(T)$. At this time, if we interpret the Kullback-Leibler divergence (KLD) [18] between $q_n(T)$ and T to the extent that the teacher inference is vulnerable to perturbation, we obtain the following; for a classification problem of C output classes and a dropout layer set Q with a dropout rate p:

$$D_{C,Q,p}(T) = \frac{\sum_n KLD(T, q_n(T))}{N} \tag{4}$$

The factors that can affect D are $C, Q,$, and p. In order to see how these factors affect D, we simulated several constraints with a toy distribution set, where N comes from Q.

The results indicate that D is relatively robust to C and N (Fig. 2). This shows that the dropout-based scheme does not depend on the number of the output classes, as when using entropy. Besides, as N increases, the curve is smoothed, while the overall tendency is not affected. The only variable affecting D is p, which can be tuned empirically. We set $N = 100$ and $p = 0.1$ throughout the experiment.

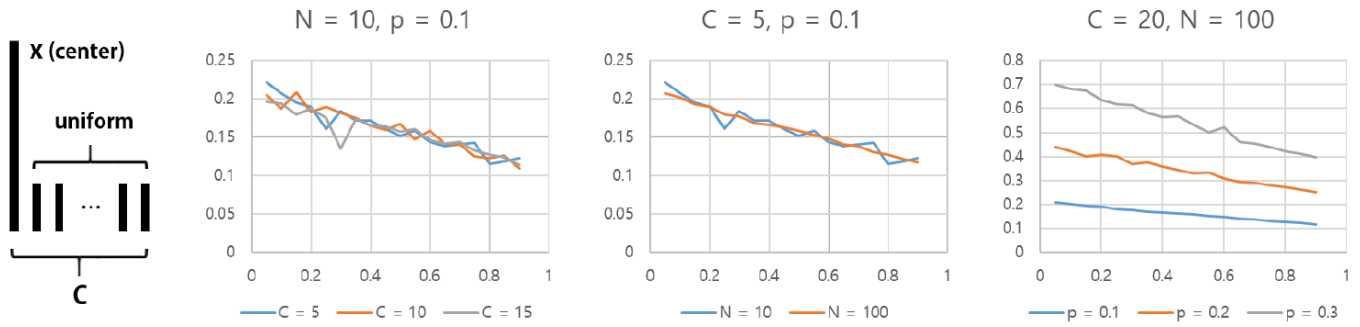

Fig. 2. Simulation for the proposed scheme, with y-axis as $D_{C,Q,p}$ for T the distribution on the left-hand side. The high value of the x-axis approximately denotes how low the entropy of the distribution is. Change in C and N does not necessarily show the variance, except for some smoothing displayed in $N = 100$ compared to $N = 10$.

IV. Implementation

A. Dataset

We exploit Fluent Speech Command (FSC) [5], an SLU dataset that incorporates 248 phrasings and 31 unique intentions, with 97 speakers and about 30K utterances. Three slots are subsequently predicted to comprise a command. Though the baseline performance is already decent [5], [19], [8], we want to note that training and evaluating the dataset may fit our goal. Also, we considered the comparison with relative error rate reduction (RERR) to be meaningful concerning the adequate size of the test set (about 3K).

B. Settings

The experiment was done by referring to two public implementations: (i) a vanilla SLU system presented in Lugosch et al. [5] and (ii) a pre-trained language model fine-tuned with the ground truth (GT) text script. We set (i) as a backbone and set the total loss by adding L_{kd}, a KD loss coming from (ii), to L_{orig}, the original cross-entropy (CE) loss. L_{kd} adopts L_1 (mean average error, MAE) loss function referring to [8]:

$$L_{kd} = L_1(f_{student}, f_{teacher}) \tag{5}$$

where $f(\cdot)$ is a logit representation.

C. Schemes

Backbone model. The backbone consists of an ASR pre-trained module and the intent identification module further trained upon it. The SincNet [6]-based ASR module incorporates phoneme and word-level seq2seq [20] submodules, where only the word posterior is utilized[1]. The RNN-based SLU module, trained with the word posterior outcome, yields slot-wisely calculated CE as a loss.

Teacher training. In training the teacher using text input, we utilized a fine-tuned model provided by Cho et al. [8], which was trained with publicly available BERT-Base PyTorch [21] wrapper released in Wolf et al. [22]. The dense layer of width 256 was augmented to the `[CLS]` representation of 768 dimension, to yield a CE-based multi-class inference at the end. Despite similar training accuracy as the backbone, the test accuracy was much higher (Table 1). We assume that this is because the variance of phrasings is fixed, and also the noise in the GT is considerably smaller than in speech data.

Scheduling. We refer to the following concepts of scheduling studied in previous work (6a-b) [8]. In exponential decaying (6a), where t is an epoch, the teacher model's influence can be transferred to the student by warming up. Contrarily, in triangular scheduling (6b) inspired by Yang et al. [13] and implemented in Cho et al. [8], the teacher's influence is increased in the mid-phase of the training. For (6b), $\mu = T/2$ and T is the maximum epoch, set 100 here.

$$\begin{aligned} (a) \quad & \lambda_t = exp(1-t) \\ (b) \quad & \lambda_t = 0.1 * max(0, -|t-\mu|/(0.5*\mu) + 1) \end{aligned} \tag{6}$$

Weight decision. For the above scheduling strategies, the control is temporary so that the student model can finally adapt to speech again. In contrast, deciding the teacher's influence regardless of the training step might bring us another result. Kim et al. [3] decide teacher influenced based on the student performance, represented by the WER-based approach in ASR, adopted here as batch-wise intent error rate (7a)[2]. The other two attain the influence of the teacher information. For entropy-based uncertainty, we follow the definition we stated in Section 2, where *sample* denotes each of the teacher's inference. For the proposed, dropout-based confidence, to guarantee the robustness, we fixed N to 100 and p to 0.1. Since smaller H or D implies more faithful teacher, we formulated the final weight as (7b) and (7c), utilizing $recip(x) = \frac{1}{1+x}$, a simple normalizing function.

$$\begin{aligned} (a) \quad & \lambda_{t,batch} = err_{batch} (= 1 - acc_{batch}) \\ (b) \quad & \lambda_{t,sample} = recip(H_C(T_{sample})) \\ (c) \quad & \lambda_{t,sample} = recip(D_p(T_{sample})) \end{aligned} \tag{7}$$

[1]In the implementation, we have adopted full SLU frameworks to see mainly how the distillation scheme works, that the pre-trained ASR module does not explicitly yield the WER. Also, the pre-trained module equals for all the implementations.

[2]The batch-wise processing is implemented here since the calculation of error rate grounds on the prediction regarding a chunk of utterances.

TABLE I
EVALUATION RESULTS. FOR MODELS WITH EXPERIMENT DONE (EXCEPT *Phoneme posterior*), WE CHOSE THE BEST AMONG WHOLE STEPS.

Teacher (text)		Text-based test error rate: 0.00% (Train: 3.78%)		
Models (speech)		**Test error rate (%)**		
No Distillation	*Vanilla*	1.16		
	Phoneme posterior	1.05 (BERT) **0.98 (ERNIE)**		
Baselines (with distil.)	*Decaying (6a)*	1.05		
	Triangular (6b)	**1.00**		
Automatic Decision		*Error rate (7a)*	*Entropy (7b)*	*Dropout (7c, Proposed)*
Scheduling	-	**1.00**	1.05	1.05
	Decaying (6a)	1.05	1.08	**0.97**
	Triangular (6b)	1.02	1.00	**0.92**

All the total loss of (6a-b) and (7a-c) follow the format of (1a).

V. EXPERIMENT

A. Results

At a glance, the baseline schemes perform well, with triangular scheduling (6b, 1.00%) and error rate (7a, 1.00%) being significant. The proposed method, in contrast, does not seem effective alone, reaching only the marginal place (1.05%) along with decaying (6a, 1.05%) and entropy (7b, 1.05%).

However, our experiment on scheduling and automatic weight decision differs from Cho et al. [8] in that we distinguish them. Scheduling is a rather mechanical control that relies on temporal factors, while the others decide the teacher's influence based on the student performance or teacher inference that depends on the input. Thus, we also experimented with a combination of those methods. The results are at the bottom of Table I, where the proposed model (7c), adopted with (6b), shows the highest gain. Though the results tell that the performance of the proposed model is confined to the accompaniment of the scheduling, we regard that it reveals another characteristics of the cross-modal distillation. Two observations are remarked.

B. Analysis

Confidence modeling works. In (6a-b), the student finally adapts to speech data, while in (7a-c), the amount of influence is decided regardless of the steps. Specifically, (7a-c) can be interpreted as the representative of student performance, teacher inference distribution, and teacher confidence, respectively. However, unlike (7b-c), (7a) more adapts the student to the gold label, whereas (7b-c) decide the weight free from the varying student performance. The former approach well captures the way for the high performance, but in a real-world scenario, the gold label might not be the 'answer' since overfitting is probable. We remark on the potential of the confidence-based decision given the decent distillation performance (1.16% to 1.05%), albeit the challenge that no moment was provided to fully adhere to the ground truth.

Confidence ameliorates scheduled KD. We took note of how our confidence modeling benefits the scheduling strategies than other automatic decisions. Though not significant if used alone probably due to the innate difference of the distillation objective, combination of (7c) with (6a-b) are all successful, at the same time reflecting the tendency that (6b) surpasses (6a). It was exhibited that the confidence modeling reduces the error rate to about 20% compared to *Vanilla*, and also reaches Wang et al. [19] that adopts the phoneme posterior-level pre-trained model [7], [23]. This does not hold for (7a-b), implying that the proposed model is more apt to preserve the positive influence of the scheduling strategies while exhibiting the advantage of the scheme itself. The result also suggests the potential of independently applying scheduling and weight decision, which may bring more flexibility in deciding the teacher's influence in distillation. Given that our model adopts an easily obtainable fine-tuned language model, the proposed scheme is expected to be a useful ingredient in determining the amount of distillation in speech-text cross-modal KD.

C. Limitation

Our study incorporates a limitation that the proposed scheme was validated only with a single benchmark instead of a set of known datasets. Also, we only handle speech-text cross-modality, not concentrating on other domains such as visuo-linguistic cues. However, given that speech and language is a scarce modality pair that conveys the relevant message in an unambiguous way, our experiment is to be meaningful for future SLU study that deals with more complicated text input.

VI. CONCLUSION

In this paper, we sought the proper scheme to manage the teacher's influence in cross-modal distillation. The proposed method, which utilizes the teacher output's dropout-based confidence, automatically induces the weight that decides the influence of KD loss. Its utility is verified on the public SLU dataset. We plan to check this scheme also works in the model training phase, beyond distillation, which is expected to separate training variance from the bias.

ACKNOWLEDGMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2021-0-00456, Development of Ultra-high Speech Quality Technology for Remote Multi-speaker Conference System).

REFERENCES

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *Stat*, vol. 1050, pp. 9, 2015.

[2] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4133–4141.

[3] Ho-Gyeong Kim, Hwidong Na, Hoshik Lee, Jihyun Lee, Tae Gyoon Kang, Min-Joong Lee, and Young Sang Choi, "Knowledge distillation using output errors for self-attention end-to-end models," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6181–6185.

[4] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu, "TinyBERT: Distilling BERT for natural language understanding," *arXiv preprint arXiv:1909.10351*, 2019.

[5] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," 2019, pp. 814–818.

[6] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with SincNet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[8] Won Ik Cho, Donghyun Kwak, Jiwon Yoon, and Nam Soo Kim, "Speech to text adaptation: Towards an efficient cross-modal distillation," *arXiv preprint arXiv:2005.08213*, 2020.

[9] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin, "Learning what and where to transfer," in *International Conference on Machine Learning*, 2019, pp. 3030–3039.

[10] Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong, "End-to-end speech translation with knowledge distillation," *arXiv preprint arXiv:1904.08075*, 2019.

[11] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin, "Distilling task-specific knowledge from BERT into simple neural networks," *arXiv preprint arXiv:1903.12136*, 2019.

[12] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, and Chun Chen, "Online knowledge distillation with diverse peers," *arXiv preprint arXiv:1912.00350*, 2019.

[13] Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Yong Yu, Weinan Zhang, and Lei Li, "Towards making the most of BERT in neural machine translation," *arXiv preprint arXiv:1908.05672*, 2019.

[14] Kisoo Kwon, Hwidong Na, Hoshik Lee, and Nam Soo Kim, "Adaptive knowledge distillation based on entropy," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7409–7413.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] Samuel Rota Bulò, Lorenzo Porzi, and Peter Kontschieder, "Dropout distillation," in *International Conference on Machine Learning*, 2016, pp. 99–107.

[17] Corina Gurau, Alex Bewley, and Ingmar Posner, "Dropout distillation for efficiently estimating model confidence," *arXiv preprint arXiv:1809.10562*, 2018.

[18] Solomon Kullback, *Information theory and statistics*, Courier Corporation, 1997.

[19] Pengwei Wang, Liangchen Wei, Yong Cao, Jinghui Xie, and Zaiqing Nie, "Large-scale unsupervised pre-training for end-to-end spoken language understanding," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7999–8003.

[20] Ilya Sutskever, Oriol Vinyals, and Quoc V Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, 2014, pp. 3104–3112.

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8026–8037.

[22] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Transformers: State-of-the-art natural language processing," *arXiv preprint arXiv:1910.03771*, 2019.

[23] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu, "ERNIE: Enhanced language representation with informative entities," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 1441–1451.