2023 RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing (RISP NCSP 2023), Feb. 28 - Mar. 3, 2023

Knowledge Distillation with Mask-based Relationship for Speech Enhancement

Jiachen Wang[1], Li Li[2], and Shoji Makino[1]

[1] Waseda University
E-mail: {jw723@ruri., s.makino@}waseda.jp

[2] Nippon Telegraph and Telephone Corporation
E-mail: lili-0805@ieee.org,

Abstract

Knowledge distillation methods aim at transferring encoded information from a large and complex teacher model to a small student model, which have been successfully applied in various fields like computer vision (CV) and nature language processing (NLP). However, current knowledge distillation methods mainly focus on image classification tasks and there exists few research on the knowledge distillation method for regression tasks like speech enhancement in audio field. In this paper, a novel mask-based relationship knowledge is proposed for speech enhancement models having U-net architectures. The proposed method works by utilizing the ideal ratio mask (IRM) calculated through the features from the corresponding encoder and decoder as the transferred knowledge. The result of the experiments demonstrated that the proposed method outperformed the response knowledge, which is the conventional knowledge distillation method in speech enhancement task. Besides, we also designed other experiments to evaluate whether the knowledge got distilled through the proposed method.

1. Introduction

With millions (even billions) of parameters, current deep neural networks (DNNs) provide splendid performance in different areas. However, it is still a challenge to implement these huge deep models on devices with limited storage and computing resources like a smartphone or hearing-aids equipment. Up to now, various knowledge distillation (KD) methods have been proposed to solve this problem, which utilize a huge teacher model to guide the training of a light-weighted small student model. Starting from dark knowledge [1], the response-based knowledge stands for the output of last layer of the teacher model. It provides student model with soft targets which contain more information compared to the original hard targets for the classification tasks, where the response-based KD has been originally proposed. Besides, later KD methods [2, 3, 6, 8] mainly focus on mimicking the intermediate layers of the teacher model. To be more specific, feature-based knowledge refers to the feature map of the intermediate layer and relationship-based knowledge transferring the knowledge through the relationship between intermediate layers.

Current KD approaches are mainly used for classification tasks and there exists few research concerned with regression tasks like speech enhancement in audio field. The difference between classification and regression tasks may hinder us from directly applying the methods proposed for classification tasks to regression tasks. For instance, the output of a classification model is the probability distribution of each sample. These soften targets could teach the student model the similarities between each class, which is not included in the original hard binary label. On the other hand, the output of a regression model is continuous values that cannot provide more information than the original ground truth. For example, the output of most speech enhancement task is the waveform or a mask. On top of that, there also exists huge difference between the model architectures. The classification models commonly only consist of encoders to extract high dimensional features while the speech enhancement models have extra decoders to recover the signals. Thus, we need to design the KD methods that are tailored to speech enhancement tasks.

In this paper, a novel mask-based relationship knowledge is proposed for the speech enhancement models with U-Net architectures. Considering about the symmetry of U-Net architecture and the features of corresponding encoders and decoders are the same size, which means they are at the same resolution, the proposed method is designed to utilize the relationship between the features from the corresponding encoder and decoder as the transferred knowledge. Besides, in speech enhancement, a mask is typically used to reflect the enhancement or the suppression at different time-frequency (T-F) bins of a spectrogram. This motivates us to design an ideal ratio mask (IRM) to represent the relationship between the features. In summary, the proposed method has the following advantages:

- Focus on the relationship between the features at the same resolution and dimension.
- Take the physical meaning of mask into consideration, which is more reasonable for speech enhancement tasks.

In addition, we also design experiments to verify whether teacher transfers helpful knowledge to students, such as whether the teacher model could help student model to reduce unseen noise.

2. Knowledge distillation methods

2.1 Formulation of knowledge distillation

Firstly, we start with formulating the KD for speech enhancement task. Let s and x denote the clean speech and noisy speech, respectively. The noisy speech x is fed into a teacher model F_T with fixed parameters and a student model F_S with parameters to be trained. We want the student model F_S to mimic the teacher model F_T with the help of KD methods. In the following part, we will show the general representation of three representative kinds of knowledge: response-based [1], feature-based [2] and relationship-based knowledge [3].

2.1.1 Response-based knowledge

The response-based knowledge directly utilizes the output of the teacher model as an additional label. The distillation loss is defined by

$$L_{distillation}(x) = L(e_T, e_S), \tag{1}$$

$$e_T = F_T(x), e_S = F_S(x), \tag{2}$$

where the e_T and e_S are the enhanced speech obtained by the teacher model and student model. L could be any loss function to measure the difference between e_T and e_S like L_1 or L_2 loss.

2.1.2 Feature-based knowledge

Feature-based distillation methods like Fitnet [2] train the student model to force the feature maps of student model more similar to the feature maps of teacher model at specific intermediate layers. It is realized by minimizing the distance between the feature maps:

$$L_{feature}(x) = L(\phi_T(F_T^l(x)), \phi_S(F_S^l(x))), \tag{3}$$

where $F_T^l(x)$ and $F_S^l(x)$ represent the output of teacher and student model at l_{th} layer. ϕ_T and ϕ_S are the transfer functions used to match the feature dimensions or change the feature maps into other more meaningful representations.

2.1.3 Relationship-based knowledge

Similar to feature-based knowledge, relationship-based knowledge also focuses on the intermediate layers but it aims at enforcing student to mimic the relationship between the intermediate layer pairs. Relationship-based knowledge distillation methods firstly generate the relationship matrix between intermediate layer pairs of teacher and student model respectively and then minimize the difference:

$$R_T = \sigma_T(F_T^{m_T}(x), F_T^{n_T}(x)), \tag{4}$$

$$R_S = \sigma_S(F_S^{m_S}(x), F_S^{n_S}(x)), \tag{5}$$

$$L_{relation}(x) = L(R_t, R_s), \tag{6}$$

where R_T and R_S are the relationship matrix, σ_T and σ_S are the function to calculate the relation, m_T, n_T, m_S and n_S denote the features are extracted from which layer.

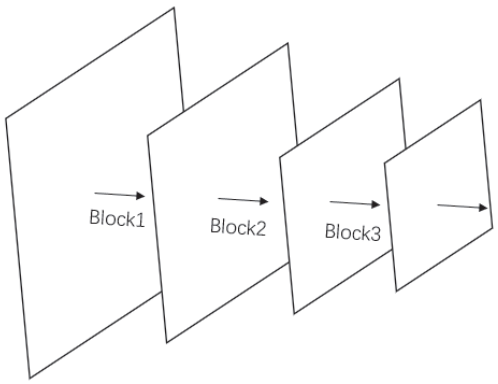

Figure 1: Diagram of pyramid architecture

2.2 Dimension mismatch problem

Currently, most image classification models [4, 5] are in a pyramid (hierarchical) structure, which is shown in Fig. 1. With this kind of structure, the dimension of the feature map increases while the resolution decreases as the network goes deeper. However, here arises a problem when we calculate the relationship matrix: the feature maps at initial layer and deeper layer are not in the same resolution and dimension. Though in [2, 6, 7], different methods like utilizing additional linear layers or other transform components (convolutional neural network, attention) have been proposed to match the dimension, these methods may bring extra parameters, which make the training of student model more complicated. To solve this problem, Yue et al. proposed Matching Guided Distillation (MGD) [8], a parameter-free KD method using linear assignment to match the channel. However, the three kinds of channel reduction methods presented in MGD (sparse matching, random drop and absolute max pooling) might be too simple to represent the connection between high dimensional and low dimensional features.

3. Proposed method: IRM-based relationship knowledge

In this paper, our proposed method solves the dimension mismatch problem in a completely different way. The proposed method is inspired by the symmetry of U-Net structure shown in Fig. 2, which is a common model structure for speech enhancement models. Different from pyramid structure, models with U-net structure firstly extracts high-dimensional features through downsampling path and then recovers the features to the original size through upsampling path. Due to this property, the features from i_{th} encoder E^i and decoder D^i are in the same resolution and dimension automatically. Thus, we decide to utilize the feature maps from the corresponding encoder and decoder pairs to calculate the relationship matrix.

Previous relation-based KD methods [3, 6] describe the correlation between feature maps through inner product and singular value decomposition (SVD). In speech enhancement

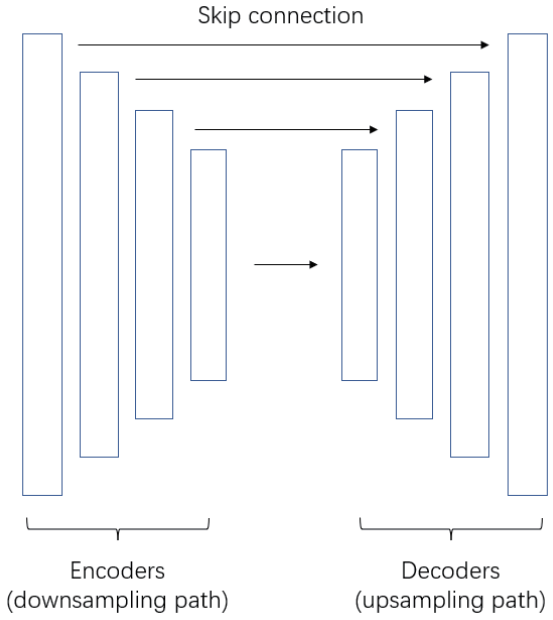

Figure 2: Diagram of U-net architecture

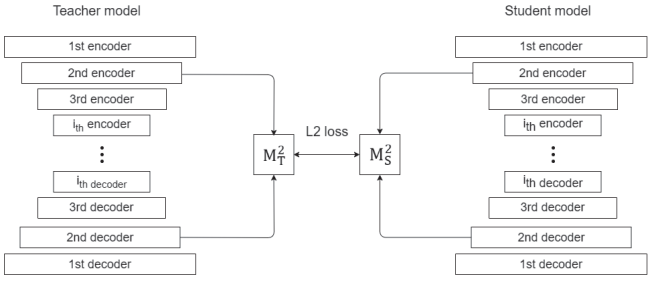

Figure 3: The architecture of proposed IRM-based relationship knowledge

task, a mask is typically used to reflect the enhancement or the suppression at different time-frequency (T-F) bins of a spectrogram. Similarly, we want to use IRM to reveal the change of the features when passing through the network. For the size of the feature maps from the corresponding encoder and decoder pair are the same, we could consider the feature map from encoders as noisy spectrogram and the feature map from decoders as enhanced spectrogram. Therefore, the IRM relationship matrix M could teach the student model the value should be enhanced or suppressed at each point for a feature map at specific resolution.

Fig. 3 shows the diagram of proposed KD module, the IRM relationship matrix is calculated as

$$M^i(k,n) = \frac{D^i(k,n)^2}{E^i(k,n)^2 + D^i(k,n)^2}, \tag{7}$$

where E^i and D^i are i_{th} encoder and decoder, k and n are indices of the feature dimensions of the axes corresponding to the frequency and time dimensions, respectively. Then the distillation loss could be expressed as

$$L_{distillation} = \sum_{k,n} \left\| M_T^i(k,n) - M_S^i(k,n) \right\|_2^2. \tag{8}$$

4. Experiment

4.1 Dataset

We evaluated the proposed method and compared it with response-based knowledge [1] and FitNet [2] with Voice-Bank+DEMAND [9], which is widely used for speech enhancement. In training set, there exists 11,572 samples uttered by 14 males and 14 females under four signal-to-noise ratios (0, 5, 10, 15 dB). The 824 testing samples are generated by utterance of 2 different speakers and 5 sorts of noise which are unseen during training process with other SNRs (2.5, 7.5, 12.5, 17.5 dB). Like other methods, we downsampled the audio data from 48 kHz to 16 kHz.

4.2 Teacher and student models

In this paper, we chose the pre-trained MANNER model (4 layers) as the teacher model, which had the same setting as that in [10]. The student model was based on MANNER-s with smaller depth (2 layers) and higher dimensions (twice as much as that of original MANNER-s). The detailed settings are summarized in Table 1. Besides, we only selected the first encoder and decoder as distillation layers.

Table 1: Detailed settings of the networks

Network	#layers	#params
T-MANNER	4	24M
S-MANNER	2	1.6M

4.3 Training procedure

For training, we opted for AdamW optimizer to train the student model for 350 epochs with the total loss in the form

$$L_{total} = L_{train} + \alpha L_{distillation}, \tag{9}$$

where α is the weight of distillation loss and it was set to decrease from 5 to 0.05 linearly during the training. Besides, we applied an OneCycleLR scheduler [11] to arrange the learning rate. The train loss L_{train} was the same as that in [10], which was the combination of L_1 loss and multi-resolution STFT loss.

4.4 Performance evaluation

We validated the effectiveness of the proposed method through four widely-used metrics: (1) Perceptual Evaluation of Speech Quality (PESQ) [12] for speech quality, (2) CSIG for signal distortion, (3) CBAK for background distortion, (4) COVL for overall quality. The proposed IRM-based knowledge is compared with the original response-based KD [1] and FitNet [2]. For response-based KD, we just utilized the

enhanced speech of teacher model as an extra label. On the other hand, we let the first encoder and decoder of student model to mimic the second encoder and decoder of teacher model when designing the experiment for FitNet. Besides, we used two distinct linear layers to map the features of student and teacher model to the same dimension. From the result in Table 2, we could find that the proposed method outperformed the student without KD and the FitNet for all the metrics. Besides, IRM-based KD outperformed response-based KD on PESQ, CSIG and COVL.

Table 2: Result on MANNER

model	PESQ	CSIG	CBAK	COVL
Student	2.98	4.39	3.53	3.70
Response-based KD	3.03	4.41	**3.56**	3.74
FitNet	2.95	4.34	3.51	3.67
IRM (proposed)	**3.04**	**4.42**	3.53	**3.76**

4.5 Study on whether teacher transferring useful information to student

In last section, the experimental results have demonstrated that huge performance improvement could be obtained by utilizing KD methods. However, it could not prove that the student model really mimics the teacher model. Ding et.al [13] pointed out that the response-based knowledge works as regularizers during training, which means regularization might be the reason for the performance improvement rather than the knowledge transferred from teacher model. To figure it out, we validated whether useful knowledge was distilled and transferred from teacher model to student model by testing whether student model could reduce unseen noise with the help of teacher model. If so, the student model can learn the implicit representation of unseen noise and have better performance.

Firstly, we removed speech-shaped noise (SSN) from the training set to make this kind of noise unseen. Then we trained several student model on this training set through different KD methods. Lastly, we evaluated the performance change on a test set which only contains SSN. The result is shown in Table 3, where Student (with SSN) and Student (without SSN) are the student models which were trained with and without SSN separately. According to the experimental result, we could find that only the proposed IRM-based knowledge got better performance compared with the student trained without SSN. However, the performance improvement is too little and there is still a huge gap between the performance of the student model trained with SSN. It seems that all the KD methods could not help student model learn how to suppress SSN from teacher model, which indicates that useful information actually was not gotten transferred.

Table 3: Performance under SSN

model	PESQ
Student(with SSN)	2.41
Student(without SSN)	2.20
Response-based KD	2.19
FitNet	2.17
IRM (proposed)	**2.22**

5. Conclusions

Based on the difference between classification tasks and speech enhancement tasks, we proposed an IRM-based KD method tailored to speech enhancement models with U-net architectures. Through the symmetry of U-net architecture, the proposed method could avoid dimension mismatch problem. Furthermore, the proposed method takes the physical meaning of mask into consideration. The experimental results demonstrated the effectiveness of the proposed method compared with other conventional KD methods. We also conducted experiments to demonstrate the issue of current KD methods, namely, KD methods could not help the student model to reduce noise that is unseen to the teacher model

References

[1] Hinton G et al., "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[2] Romero A et al., "Fitnets: Hints for thin deep nets," arXiv preprint arXiv:1412.6550, 2014.

[3] Yim J et al., "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," *in Proc. CVPR*, 2017, pp. 4133-4141.

[4] He K et al., "Deep residual learning for image recognition," *in Proc. CVPR*, 2016, pp. 770-778.

[5] Zhang P et al., "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proc. ICCV*, 2021, pp. 2998-3008.

[6] Lee S H et al., "Self-supervised knowledge distillation using singular value decomposition," in *Proc. ECCV*, 2018, pp. 335-350.

[7] Zagoruyko S et al., "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," arXiv preprint arXiv:1612.03928, 2016.

[8] Yue K et al., "Matching guided distillation," in *Proc. ECCV*, 2020, pp. 312-328.

[9] C. Valentini-Botinhao, "Noisy speech database for training speech enhancement algorithms and TTS models," Edinburgh DataShare, 2017.

[10] Park H J et al., "MANNER: Multi-View Attention Network For Noise Erasure," in *Proc. ICASSP*, 2022, pp. 7842-7846.

[11] Smith L N et al., "Super-convergence: Very fast training of neural networks using large learning rates," in Artificial intelligence and machine learning for multi-domain operations applications, vol. 11006, pp. 369-386, 2019.

[12] I.-T.Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Rec. ITU-T P. 862, 2001.

[13] Ding Q et al., "Adaptive regularization of labels," arXiv preprint arXiv:1908.05474, 2019.