PAPER

Research on Lightweight Acoustic Scene Perception Method Based on Drunkard Methodology

Wenkai LIU[†], **Lin ZHANG**[†], ***Nonmembers***, **Menglong WU**[†a)], ***Member***, **Xichang CAI**[†], ***and*** **Hongxia DONG**[†], ***Nonmembers***

SUMMARY The goal of Acoustic Scene Classification (ASC) is to simulate human analysis of the surrounding environment and make accurate decisions promptly. Extracting useful information from audio signals in real-world scenarios is challenging and can lead to suboptimal performance in acoustic scene classification, especially in environments with relatively homogeneous backgrounds. To address this problem, we model the sobering-up process of "drunkards" in real-life and the guiding behavior of normal people, and construct a high-precision lightweight model implementation methodology called the "drunkard methodology". The core idea includes three parts: (1) designing a special feature transformation module based on the different mechanisms of information perception between drunkards and ordinary people, to simulate the process of gradually sobering up and the changes in feature perception ability; (2) studying a lightweight "drunken" model that matches the normal model's perception processing process. The model uses a multi-scale class residual block structure and can obtain finer feature representations by fusing information extracted at different scales; (3) introducing a guiding and fusion module of the conventional model to the "drunken" model to speed up the sobering-up process and achieve iterative optimization and accuracy improvement. Evaluation results on the official dataset of DCASE2022 Task1 demonstrate that our baseline system achieves 40.4% accuracy and 2.284 loss under the condition of 442.67K parameters and 19.40M MAC (multiply-accumulate operations). After adopting the "drunkard" mechanism, the accuracy is improved to 45.2%, and the loss is reduced by 0.634 under the condition of 551.89K parameters and 23.6M MAC.

key words: *acoustic scene classification, model compression, multi-scale module, knowledge distillation*

1. Introduction

Acoustic scene classification [1], as an important application of deep convolutional neural networks [2] in the audio field, simulates human perception of the external environment to make correct classifications of the surrounding environment and has been widely used in audio monitoring, intelligent driving assistance, voiceprint recognition, and other fields.

Most acoustic scene classification tasks adopt a top-down serial approach, directly inputting the extracted feature information into a neural network model for prediction. However, this approach has some limitations. Currently, the mainstream neural networks are still deep convolutional neural networks. In addition, some high-precision lightweight models [3], [4] have also been proposed. For example, Kim [5] used an efficient BC-ResNet architecture to extract two feature maps specific to the frequency and time dimensions by two-dimensional convolution on the frequency and one-dimensional convolution on the time, achieving excellent performance. Lee proposed BC-Res2Net [7] by fusing BC-ResNet and Res2Net [6] structures, which can effectively obtain features in the frequency and time dimensions through broadcast learning and can run at multiple scales with significant performance improvement. In recent years, the MobileNet series models [8] and ShuffleNet [9] have achieved lightweight and efficient networks by introducing deep convolution and shuffle operations.

However, the aforementioned networks are limited by their structure and the computational cost increases as the model deepens, which is not conducive to deployment on resource-constrained devices. Moreover, using a single neural network model may not fully extract key audio features [10], and there is currently no determined optimal model architecture and hyperparameter combination, which can lead to erroneous decisions on scene categories.

To address the aforementioned issues and inspired by the idea that a drunkard may not accurately perceive their environment like a sober person but can improve their judgment with guidance, we propose a lightweight and robust framework, the Drunkard Methodology. In the following, we provide a detailed explanation of three aspects:

1. This concept stems from our observations of everyday life. We have noticed that when it comes to perceiving the external environment, despite factors like impaired hearing and vision, the primary distinction lies in the working state of the brain between a normal individual and an intoxicated one. The brain of an intoxicated person, compared to that of a normal individual, exhibits reduced sensitivity and diminished information-processing capabilities. Inspired by this, we constructed two analogous models: the Normal Model and the Drunken Model, to simulate the perceptual abilities of a sober person and an intoxicated person's brain in the external environment. Additionally, we designed a Guide module to facilitate the transformation from the Normal Model to the Drunken Model. It's worth mentioning that the structural similarities between the two models are due to the relatively coarse granularity of tasks related to scene perception, where the differences in perceptual abilities between intoxicated and sober individuals are relatively minor.

2. Considering the numbing effect of alcohol on an intoxicated person, their brain's information reception ca-

Manuscript received May 29, 2023.
Manuscript revised September 4, 2023.
Manuscript publicized October 23, 2023.

[†]The authors are with North China University of Technology, Beijing, China.
a) E-mail: wumenglong@126.com (Corresponding author)

DOI: 10.1587/transinf.2023EDP7107

pabilities are impaired, leading to a reduced intake of information compared to a sober individual. To address this, we made design adjustments in the feature extraction component. Drawing an analogy to the broader spectrum of information a sober individual receives, we constructed the Normal Feature composed of three distinct features. Additionally, a Drunken Feature has been constructed, which has lower complexity and contains more concise feature information.

3. Furthermore, we hold the belief that the perceptual and information-processing capabilities of an intoxicated person can be enhanced through external guidance from a sober individual [11]. Therefore, in the feature extraction section, we designed a feature conversion module [12], serving as a guide from the Normal Feature to the Drunken Feature. The inspiration for this module comes from the Squeeze-and-Excitation (SE) attention mechanism [13], capable of removing redundant feature information from the Normal Feature, thereby enabling the Drunken Feature to focus more on crucial regions and become more lightweight. Additionally, we formulated a Fusion module to simulate the process of a sober individual guiding an intoxicated one. In summary, the Inebriated Paradigm we have devised comprises two parallel branches representing strong and weak environmental perception capabilities, incorporating several modules that interconnect these two branches, symbolizing the guidance provided by a sober individual to an intoxicated one. This parallel framework, compared to traditional serial structures, leverages the inherent connections between the two states, ensuring lightweight design and robustness while enhancing predictive accuracy and compensating for deficiencies in a singular model structure. A detailed explanation of the framework's structure will be provided in Sect. 2.1.

2. Proposed Method

This chapter presents the methodology of the current study in three parts. Section 2.1 provides an overview of the overall design architecture of the drunkard methodology, demonstrating its underlying design principles and explaining the interrelationships among its various components. Section 2.2 introduces the deep learning features employed in this study, as well as the feature conversion module. Section 2.3 provides a detailed description of the structures of the Drunken model, the Normal model, and the guidance and fusion modules.

2.1 Overall Framework

The overall architecture is illustrated in Fig. 1, which includes two branches and the modules connecting them.

The uppermost branch delineates the modeling process of normal behavior. Mapping to deep learning frameworks,the input data undergoes a designated feature extraction procedure to yield the Normal Feature, mirroring the scenario information reception that typifies the human experience. Following this, the Normal Feature is input into the Normal Model for further feature extraction, training, and subsequent scene prediction. In this context, the Normal Model simulates the perceptual and processing journey through which a person receives, comprehends, and processes information from their external environment.

Subsequently, the lowermost branch illustrates the modeling process of intoxicated behavior. To emulate the comparatively reduced external information intake by an intoxicated individual, the Normal Feature is subject to a feature conversion module, resulting in a more concise Drunken Feature. The Drunken Feature, in contrast to the Normal Feature, discards redundant feature information, concentrating on salient feature information and achieving a more lightweight representation. The ensuing step involves inputting the Drunken Feature into the Drunken Model for subsequent feature extraction, training, and prediction. Analogous to the prior case, this segment emulates the intoxicated individual's perception and processing of external scene information. The Drunken Model emerges from the transformation of the Normal Model through the Guide module, which primarily entails adding the Frequency Grouping Fusion Convolution layer to the Normal Model. This layer partitions the frequency dimension and reuses different-dimensional features, bolstering feature representation. Additionally, for lightweight design, the width of the Drunken Model is trimmed.

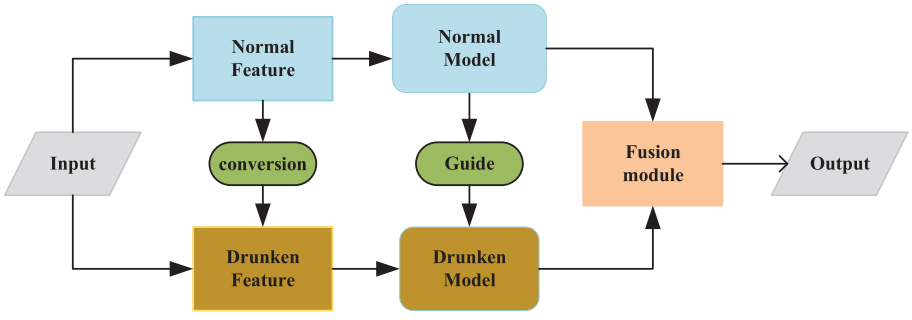


Fig. 1 The overall framework of the Drunkard Methodology

The Fusion module's role is to amalgamate the two branches. In this paper, the conventional knowledge distillation strategy is adopted. Knowledge distillation is a technique introduced by Hinton et al. [14] that transfers knowledge from a complex model to a simpler one, thereby enhancing the performance of the latter. The philosophy behind knowledge distillation aligns harmoniously with the overarching architecture proposed in this study, making it suitable as a transition module between normal and intoxicated behaviors. Furthermore, we reviewed literature about advancements in knowledge distillation techniques [15], [16]. It is undeniable that these improvements have yielded superior results, yet we acknowledge that such enhancements are often tailored to specific tasks. Consequently, the applicability of the enhanced knowledge distillation techniques might be limited in terms of generality. We sought a more universally compatible technique that aligns well with our method. Thus, we ultimately opted for the conventional knowledge distillation method put forth by Hinton et al., which can serve as a model compression approach, enhancing the performance of a simple model without incurring significant computational overhead.

2.2 Feature Extraction

2.2.1 Normal Feature

Although new advanced features are often designed for audio-related fields such as speech recognition, sound event detection, and information retrieval, we expect to choose mature and perceptually meaningful feature representations. Based on human auditory features, we are more sensitive to the different information in the low-frequency range, and the human ear cannot perceive frequency linearly. Therefore, we first consider using the logarithmic Mel spectrogram feature. The logarithmic Mel spectrogram is a commonly used audio feature extraction method, which contains time-domain and frequency-domain information as well as perceptually relevant amplitude information. Its core is the Mel scale, which better matches the auditory characteristics of the human ear.

In addition, since speech signals are temporally continuous, feature information extracted by frame-wise processing only reflects the characteristics of the current frame of the speech signal. To better capture the temporal continuity of the signal, the feature dimension can be increased by adding the dimensions of the preceding and succeeding frames, commonly achieved by first-order and second-order differences. Therefore, we applied first-order and second-order differences to the logarithmic Mel spectrogram and cascaded them along the channel dimension. The final Normal Feature is a splice of log-Mel spectral features, first-order difference features, and second-order difference features with three-dimensional channel dimensions.

2.2.2 Drunken Feature

Considering that the Normal Feature is obtained by concatenating three types of features along the channel dimension, distinct feature extraction methods might capture similar feature patterns, implying that the Normal Feature could potentially encompass redundant feature information. Moreover, intermediate feature maps generated during the model training process are often characterized by redundancy [17]. Originally devised to capture long-range dependencies and salient information in natural language processing tasks, attention mechanisms have found widespread application across various domains. In light of the aforementioned issues, we contemplate the introduction of attention mechanisms during the feature extraction phase, aiming to suppress redundant information within the Normal Feature while accentuating meaningful details. Drawing inspiration from the SE attention module and incorporating certain structural elements, we have formulated an attention-based feature conversion module. As a consequence of processing through this feature conversion module, the Normal Feature is refined to yield a de-redundant counterpart referred to as the Drunken Feature. In comparison to the Normal Feature, the Drunken Feature demonstrates a more streamlined profile, owing to the feature conversion module's capability to mitigate the presence of akin features.

As depicted in Fig. 2 below, the feature conversion module commences by subjecting each channel to a global average pooling operation, thereby reducing the spatial dimensions to scalar values, yielding the global average for each channel. The intent behind this operation is to compress each channel, capturing channel-wise global statistical information while concurrently diminishing computational overhead. Subsequently, channel-wise correlations are learned through fully connected layers. Eventually, by employing the sigmoid function, weights of different channel dimensions are derived to ascertain the significance of features. These weights are then multiplied with their corresponding original features, culminating in the derivation of the Drunken Feature.

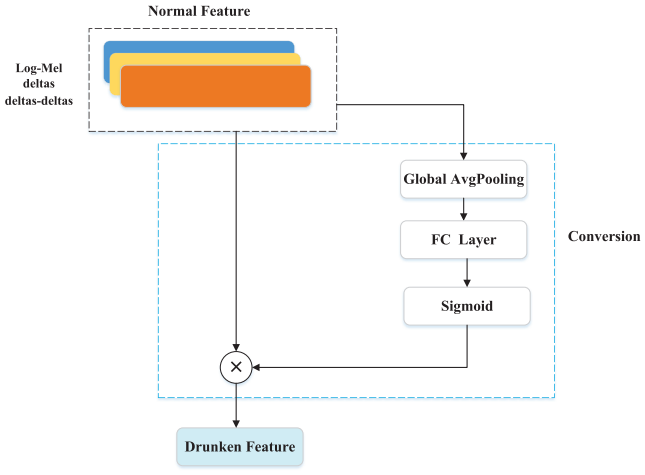


Fig. 2 Feature conversion module

2.3 Network Setup

2.3.1 Drunken Model

The datasets used in this study consist of 1-second audio files, which contain limited information. To address this issue, a multi-scale Drunken model is designed to capture more detailed and richer feature information at different scales. The model mainly consists of three modules, namely, Frequency Grouping Fusion Convolution (FreGroupConv2d), GroupConv2d, and Improved Bottleneck Block (FCresnet_block). We use the Drunken Model in Table 1 as the baseline model for this paper, which is roughly divided into six stages. First, the feature information is enriched through the FreGroupConv2d. Then, convolution and pooling layers are used for channel expansion and downsampling. The fourth stage contains multiple stacked FCresnet_blocks, with different channel numbers and step sizes, and Dropout is added to prevent overfitting. Finally, the classification results are output after global average pooling and fully connected layers. Figure 3 is a more visual presentation of the model structure in Table 1, with different colored blocks representing various modules in Table 1.

A. Frequency Grouping Fusion Convolution

The information contained within sound signals in different urban scenes varies across distinct frequency ranges. For instance, in the case of sound signals in a city center, the low-frequency component may encapsulate the baseline noise such as low-frequency vibrations from traffic flow and engine noises, while the high-frequency portion might comprise sharp sounds like braking and car horn honks. In contrast, sound signals within a park might encompass low-frequency natural sounds like bird calls and distant water flow, in the low-frequency range, and detailed sounds like bird chirps and small animal noises within the high-frequency range. Analyzing the auditory information across these distinct frequency segments can offer enhanced insights into activities within different scenes, rendering this aspect crucial for scene classification tasks.

Drawing from the aforementioned notions and inspired by the Res2Net model architecture, we partition the frequency dimension of convolutional layers. This partitioning permits the low-frequency component to capture global and coarse features, while the high-frequency component is adept at capturing local and intricate details. Such partitioning aids in augmenting the model's generalization capacity towards input data. Furthermore, the application of frequency-grouped convolutions enables the separate processing of features within different frequency ranges, thus enhancing the model's noise resistance to a certain extent.

Figure 4 illustrates the schematic principle of the Frequency Grouping Fusion Convolution. We uniformly partition the frequency dimension into four groups. Each group undergoes a 3×3 convolution. Except the first pathway, each subsequent pathway incorporates information from the preceding pathway. This progressive feature reuse mechanism, transitioning from low to high-frequency features, mirrors the stepwise feature extraction process observed in the human visual system from edges to textures, thereby enriching feature information to a certain extent. Ultimately, the outputs are concatenated along the channel dimension and forwarded to the subsequent stage. Given the subsequent downsampling along the frequency axis, the incorporation of this layer in the later stages of the model is disregarded.

Table 1 Overall architecture of the Drunken Model (our baseline).

Stage	Operator	Out_channels	Stride	Output
Stage 1	3×3 FreGroupConv2d	3	1	128×43×3
Stage 2	3×3 Conv2d	32	1	128×43×32
Stage 3	2×2 MaxPooling	32	2	64×22×32
Stage 4	FCresnet_block×2	64	1	64×22×64
	FCresnet_block×3	64	2	32×11×64
	FCresnet_block×5	128	1	32×11×128
	Dropout (0.3)	-	-	32×11×128
	FCresnet_block×2	128	2	16×6×128
Stage 5	GlobalAveragePooling	-	-	-
Stage 6	Dense	10	-	-

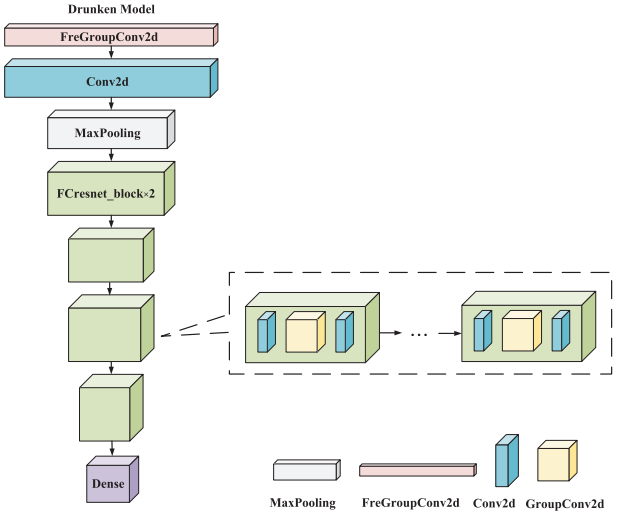


Fig. 3 Drunken Model (our baseline).

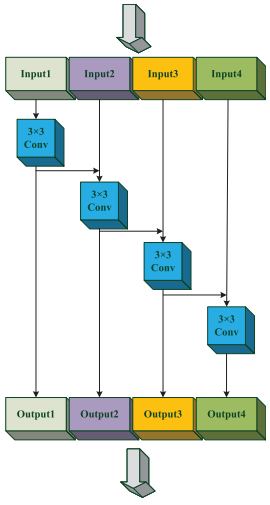


Fig. 4 Frequency grouping fusion convolution

Instead, the frequency-grouped fusion convolution is positioned solely at the first layer of the model.

B. FCresnet_block

This section constitutes the main body of the Drunken Model, which is modified based on the bottleneck in ResNet [18] and named FCresnet_block. The original bottleneck block consists of three convolutional layers and a shortcut connection that maps the input directly to the output. The three convolutional layers have kernel sizes of 1×1, 3×3, and 1×1, respectively. The 1×1 convolution mainly changes the number of channels and does not increase the spatial dimension of the feature information. Only a 3×3 convolutional layer is used to extract spatial features. To fully explore the deep information of audio features, we consider using grouped convolution to replace the 3×3 convolution in the bottleneck block. Grouped convolution evenly divides the channel dimension into multiple groups, and each path goes through a 3×3 convolutional layer. Finally, the outputs of all paths are concatenated along the channel dimension, ensuring that the output and input have the same number of channels. The parameter "Groups" is used to indicate the number of channel groups, and the baseline value for "Groups" in this paper is 8.

2.3.2 Normal Model

The overall architecture of the Normal Model is akin to the Drunken Model outlined in Table 1. The distinctions lie in several aspects: 1) The Normal Model lacks the Frequency Grouping Fusion Convolution, i.e., Stage 1 as delineated in Table 1; 2) The Normal Model exhibits higher complexity, primarily in terms of model width. Within the FCresnet_block module of the Normal Model, the value of "Groups" is set to 32, resulting in a channel count twice that of the Drunken Model; 3) The Normal Model comprises a depth of 37 layers, with a MAC (Multiply-Accumulate) index of 23.5M, slightly higher than that of the Drunken Model. It boasts approximately 1114.6k parameters, which is roughly three times that of the Drunken Model. Experimental outcomes reveal that the highest accuracy achieved by the Normal Model reaches 46.4%, with a loss of 1.619.

2.3.3 Guide and Fusion Module

The role of the Guide module is to transform the Normal Model into the Drunken Model through specific operations. One of these operations involves channel reduction. We conducted multiple experiments to fine-tune the channel count of the model and retained the optimal configuration, as detailed in Table 1. Additionally, we introduced Frequency Grouping Fusion Convolution to the Normal Model and decreased the number of groups for grouped convolutions, all building upon the foundation of the Normal Model.

The fusion module adopts the knowledge distillation strategy, the knowledge distillation strategy contains a teacher model and a student model, which improves the performance of the student model by teaching the knowledge in the higher-performing teacher model to the lower-performing student model. Combining the principle of knowledge distillation, we use the Normal Model in the Drunkard Methodology as the teacher model and the Drunken model as the student model, and we use the original form of knowledge distillation in this paper. First, the Normal model is trained under different configurations, and the parameter setting with the best performance is selected. Then, the soft label predictions of the Normal Model and the Drunken Model are computed at the same temperature T, and the distillation loss is calculated. The final loss is a weighted sum of the distillation loss and the hard label loss, as shown in Eq. (1):

$$L_{TOTAL} = \alpha L_{DIST} + (1 - \alpha) L_{LABEL} \tag{1}$$

The distillation loss is based on matching predictions using the soft targets given in Eq. (2) between the student and teacher, where z is the logarithm, q is the soft target, and T is the temperature that controls the softness of the probability distribution.

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_i/T)} \tag{2}$$

3. Experiments

3.1 Experimental Setup

3.1.1 Datasets and Evaluation Metrics

All experiments in this paper were conducted on the development set [19] of DCASE (Detection and Classification of Acoustic Scenes and Events) 2022 Task1, TAU Urban Acoustic Scenes 2022 Mobile, the audio data in the dataset is provided in monaural format at 44.1 kHz with 24-bit resolution. This dataset comprises a total of 230,359 audio clips, with a cumulative duration of 64 hours. We partitioned the dataset into training and testing subsets, allocating 70% for training and 30% for testing. Each audio clip has a duration of 1 second. The dataset comprises recordings from 10 cities and 9 devices: 3 real devices (A, B, C) and 6 simulated devices (S1 S6). It encompasses 10 distinct scenes, namely the airport, shopping mall, subway station, pedestrian street, public square, traffic street, tram, bus, subway, and park. For model evaluation, we employed the validation set provided by DCASE2022 to assess the performance of the trained model. The validation set comprises data from 12 cities, 10 acoustic scenes, and 11 devices, including five new devices (unavailable in the development set): real device D and simulated devices S7-S11. The evaluation data encompasses 22 hours of audio recordings, recorded at different locations compared to the development data.

We employed parameters and MAC count to measure the system's complexity. To validate the effectiveness of the proposed method, we utilized loss and accuracy as metrics. The cross-entropy loss function was employed to evaluate the effectiveness of the proposed method.

3.1.2 Training Settings

In terms of feature extraction, we use 128 Mel filters to process audio signals and perform fast Fourier transform (FFT) on them, with a Hamming window length and frame length of 0.04 seconds and 0.02 seconds, respectively, resulting in a 128×43 spectrogram. First-order and second-order differential features are then extracted, and the three types of feature spectrograms are concatenated along the channel dimension to form a 128×43×3 input feature spectrogram.

During the training phase, the experiment sets the batch size to 128 for both the Drunken model and the Normal model, with 256 iterations for each model and 200 iterations for the knowledge distillation experiment. In addition, Mixup and SpecAugment data augmentation techniques are introduced to optimize the training process, and an early stopping mechanism is added to prevent overfitting. Multiple ablation experiments are conducted to optimize the proposed Drunken model and save the best settings. Based on these settings, training for the Normal model and the knowledge distillation experiments are performed.

The baseline system in our paper uses logarithmic Mel spectra and first-order and second-order differential features as input features. The model adopts an FC_ResNet with a depth of 38 layers and Groups set to 8, training for 256 iterations on the entire dataset without data augmentation or coordinate attention mechanism. The experiments are conducted on an NVIDIA RTX2080 Ti using the Tensorflow and Keras frameworks, with a Windows 10 operating system.

3.1.3 Optimization Method

In this paper, several optimization methods are employed to effectively improve the model performance,which are described in detail as follows.

Mixup [21] randomly selects two different sample-label pairs from the same batch of data and generates a new sample-label pair by adding them together in proportion. By generating new samples, the generalization ability of the model can be improved and the problem of imbalanced data classes can be alleviated.

SpecAugment [22] involves random transformations in the time-frequency domain of speech signals, including masking in the time and frequency domain, time warping, and frequency masking. This technique improves the robustness and generalization ability of the model without requiring additional parameters or computational costs.

The positional attention mechanism [23] adds relative positional information to each position vector, enabling the model to better understand the relationship between different positions in a sequence. Compared to other position encoding methods, the computational complexity of the positional attention mechanism is relatively low, enabling efficient processing of long sequences.

In Sect. 3.1.1, we allocated 70% of the development dataset for training and 30% for testing, following the partitioning guidelines provided by the DCASE official documentation. However, we identified a potential drawback in this partitioning approach, as it could result in a relatively limited size of the training set, thereby affecting the effectiveness of model training. To address this concern, we introduced an optimization strategy by augmenting the training dataset. As part of this optimization, we performed a re-division of the development set, utilizing the entire development dataset for training purposes, while keeping the validation set unchanged.

3.2 Results and Analysis

3.2.1 Classification Results on the Drunken Model

Tables 2 and 3 investigate the impact of two hyperparameters, namely depth, and Groups, on the Drunken Model's performance in the ASC task. Initially, we varied the values of these two hyperparameters, iteratively training on the training set, and retained the Drunken Model with the best training results. Subsequently, we evaluated the corresponding best models saved under different hyperparameter settings on the validation set, yielding the experimental results presented in Tables 2 and 3.

To explore the optimal number of Groups, we maintained the Drunken Model's depth at 38 while tuning the Groups. With a stride of 2, we conducted multiple experiments within the range of Groups [2, 38], selecting representative data points for Table 2. Results reveal that an increase in Groups from 4 to 8 leads to a 1.9% accuracy improvement. However, within the range of Groups [8, 28], accuracy exhibited a gradual decline, with Groups set at 8 yielding the best performance. Although a slight uptick is observed in Group 32 compared to Group 28, this increase is within the margin of potential error, and the accuracy remains below the highest level by 1%. Furthermore, a substantial decrease in accuracy occurs when Groups is increased to 36. This observation substantiates that an excessive number of channels may lead to overfitting. Nevertheless, a judicious selection of channel grouping can introduce rich feature information and enhance predictive accuracy.

Turning to the exploration of the optimal model depth,

Table 2 Results with different numbers of groups.

Groups	Accuracy	Groups	Accuracy
4	39.6%	24	40.2%
8	**41.5%**	28	40.0%
12	40.6%	32	40.5%
16	40.1%	36	38.5%

Table 3 Results with different depths.

Depth	Accuracy	Loss	Parameters	MACs
20	38.0%	2.195	255.02K	14.48M
28	39.6%	1.925	300.12K	16.52M
38	**41.5%**	**1.86**	**442.67K**	**19.40M**
50	38.6%	1.829	732.10K	25.6M
56	36.5%	2.625	1236.25K	30.5M

we kept Groups fixed at 8 and conducted multiple experiments within the Depth range [16, 38], with a stride of 4. Representative data points are presented in Table 3. As model depth significantly influences complexity, we aimed to obtain a relatively lightweight Drunken Model with high accuracy. Consequently, we included model parameters and MAC values in Table 3. The results in Table 3 demonstrate that performance gradually improves with increasing model depth, but further deepening the model beyond a certain point yields diminishing returns, potentially due to overfitting. The model attains an accuracy of 41.5% at a depth of 38 layers, with a loss value of 1.86, parameter count of 442.67K, and MACs of 19.40M. Based on the aforementioned outcomes, we opt to proceed with experiments utilizing an FC_ResNet with Groups set at 8 and a depth of 38 layers.

We also used a series of optimization methods, and to verify their effectiveness, we conducted experiments under the optimal hyperparameter settings of Tables 2 and 3.Table 4 presents the performance of the Drunken Model with different optimization methods added, in the order from top to bottom of the table. The results indicate that each optimization method has improved the performance of the Drunken Model to some extent. Expanding the training set appropriately is the most direct way to improve the network performance, as it can provide diverse feature information for subsequent training. Mixup and Specaugment have been proven to be effective optimization methods, with an accuracy increase of 1% and a significant decrease in loss of 0.42 after data augmentation. The Coordinate Attention mechanism has performed well in improving accuracy, with an accuracy increase of 1.6% and a loss decrease of 0.123. This is due to the embedding of position information into the channel attention, enabling the network to obtain information from a larger range while avoiding significant overhead.

3.2.2 Feature Comparison Experiment

This section evaluates the effectiveness of the feature conversion module on both the Drunken Model and the Normal Model, with the results shown in Table 5. However, the expected improvement was not observed, as the accuracy decreased by approximately 1% after adding it. Considering that it is unreasonable to use global average pooling, our initial intention is to remove the redundant feature information, but GAP will greatly reduce the number of parameters, so some important feature parameters are also lost in the process of removing the redundant information. To address this issue, we will further investigate it in future research.

3.2.3 Knowledge Distillation Ablation Experiments

This section presents ablation experiments on knowledge distillation to verify the effectiveness of the proposed method and explore the optimal temperature and loss of weight. To eliminate the influence of other factors, we used logarithmic Mel spectrograms, first-order and second-order differences as input features, trained on the entire dataset with 200 iterations. The teacher model was the trained Normal Model with an accuracy of 46.4%, and the student model was FC_ResNet with 8 groups and 38 layers. We initially alter the values of two hyperparameters, iteratively iterating on the training set, and preserving the optimal configuration. Subsequently, an evaluation is conducted on the validation set, yielding the experimental outcomes presented in Tables 6 and 7.

We conducted several experiments both in the range of T as [1,8] and as [0.1,1], and selected representative data are reported in Tables 6 and 7.The best performance was achieved when the temperature was set to 2, indicating that appropriately softening the output of the label by the teacher model can reduce the degree of polarization in the results and provide more category information. The better the performance of knowledge distillation. The best result was obtained with a loss weight of 0.9, with an accuracy of 45.2% and a loss of 1.650, indicating that Drunken Model benefits significantly from the useful information obtained from the Normal Model, which is crucial for improving the accuracy of the Drunken Model.

Table 4 Optimization experiments on the Drunkard Model. The better results are darker colors.

Optimization Method		Accuracy	Loss
Training set expansion	**Use**	**40.4%(our baseline)**	**2.284**
	No use	39.3%	2.236
+Mixup&Specaugment	**Use**	**41.5%**	**1.864**
	No use	40.4%	2.284
+Coordinate attention(CA)	**Use**	**43.1%**	**1.741**
	No use	41.5%	1.864

Table 5 Feature comparison experiment on the Drunkard Model and Normal Model. The best results are darker colors.

Conversion	Druken Model	Normal Model
No use	**41.5%**	**46.4%**
use	40.2%	45.9%

Table 6 Accuracy and loss at different temperatures. Experimental setup α = 0.9.

Temperature	Accuracy	Loss
1	43.5%	1.768
2	**45.2%**	**1.650**
3	43.9%	1.721
4	43.1%	1.801
5	42.8%	1.854

Table 7 Accuracy and loss at different loss weights.

α	Accuracy	Loss
0.3	41.9%	1.870
0.4	41.8%	1.877
0.5	42.1%	1.832
0.6	42.0%	1.842
0.7	42.5%	1.774
0.8	43.4%	1.731
0.9	**45.2%**	**1.650**

Table 8 Accuracy on different devices.

Devices	our baseline	our baseline +Mixup +Specaugment	our baseline +Mixup +Specaugment +CA	KD T=2 α=0.9
A	58.2%	59.4%	59.1%	62.1%
B	46.2%	46.2%	46.3%	52.2%
C	49.3%	52.9%	54.2%	54.8%
S1	41.1%	42.9%	41.4%	45.7%
S2	37.5%	41.4%	39.1%	43.7%
S3	40.1%	40.4%	43.2%	44.3%
S4	30.4%	32.2%	35.6%	35.4%
S5	34.9%	33.3%	39.2%	40.1%
S6	26.4%	25.2%	30.0%	28.5%

Table 9 Accuracy on different scenes.

Scenes	our baseline	our baseline +Mixup +Specaugment	our baseline +Mixup +Specaugment +CA	KD T=2 α=0.9
Airport	35.6%	39.1%	42.0%	43.6%
Bus	34.8%	34.5%	45.3%	42.9%
Metro	40.9%	40.4%	40.4%	40.5%
Metro_station	38.8%	39.1%	29.5%	40.2%
Park	61.9%	67.1%	71.2%	68.1%
Public square	21.7%	18.7%	23.0%	18.6%
Shopping mall	38.8%	41.5%	39.1%	48.7%
Street pedestrain	29.7%	26.9%	24.7%	28.8%
Street traffic	66.4%	70.9%	65.1%	76.1%
tram	35.7%	37.3%	51.0%	44.5%

3.2.4 Accuracy Results on Different Devices and Scenes

This section presents the evaluation results of different model settings for nine devices and ten scenarios. Table 8 shows that data augmentation and coordinate attention mechanism are effective in improving model performance on most devices, especially the coordinate attention mechanism, which can further improve the accuracy of the model based on data augmentation. Both coordinate attention mechanism and Mixup can enhance the feature representation ability.

Furthermore, Table 8 shows that the effect of data augmentation and attention mechanism is not good on a small number of devices, such as the accuracy of S5 decreases by 1.6% after adding data augmentation, and the accuracy of S1 decreases by 2.3% after adding CA. These experimental results do not deny the effectiveness of CA and data augmentation. Besides the randomness of the experiment, we believe that this is closely related to the size of the dataset. Device A accounts for 70% of the total dataset, while the data of S1-S6 are transformed from the data of Device A and account for a relatively small proportion. Therefore, insufficient data volume may lead to a decrease in performance.

Compared with the previous two methods, knowledge distillation steadily improves the performance of each device, with an accuracy improvement of 2% 6.5% compared to the baseline, indicating that knowledge distillation can significantly improve the model's robustness and generalization ability, with a relatively low data requirement.

Table 9 reports the experimental accuracies under different scenarios for three configuration modes. There are accuracy drops in some scenarios for each mode. Besides the influence of experimental interference factors and insufficient data, we hypothesize that the short duration of audio may also contribute to the decrease in accuracy. All audios in the dataset used in this study are one second long, containing relatively limited scene information, and there may be similar sounds in different scenarios. We believe this is one of the key factors affecting the experimental results.

3.2.5 Comparison Experiments between the Drunken Model and Other Models

To validate the effectiveness of our designed Drunken Model, we compared it with state-of-the-art models, namely ShuffleNetV2 [9], ResNet [18], MobileNet [24], and GhostNet [17]. To ensure a fair comparison, we controlled the parameter count of the four models within the range of [440k, 465k]. We selected the Drunken Model named FC_ResNet with an accuracy of 41.5% from Table 4 and trained the remaining four models under the same training settings. The experimental results are presented in Table 10. The results indicate that our designed model achieves higher accuracy compared to ShuffleNetV2 and ResNet with similar parameter quantities. This can be attributed to the rich feature information extracted from various scales. In comparison to MobileNetV2, our approach yields a mere 1% increase in accuracy; however, it maintains a reduced parameter count. We believe this aspect underscores the balance between accuracy and parameter efficiency within our approach. Furthermore, in contrast to GhostNet, while our method presents a marginal accuracy improvement of 0.3%, it manages to reduce parameters by 5%. Thus, our approach demonstrates competitive performance comparable to GhostNet.

Table 10 Accuracy on different networks.

Method+Citation	Accuracy	Parameters
GhostNet [17]	41.2%	462.75K
MobileNetV2 [24]	40.5%	460.51K
RseNet [18]	39.1%	450.10K
ShuffleNetV2 [9]	40.1%	440.21K
FC_ResNet	41.5%	442.67K

4. Conclusion

In this paper, we first propose a lightweight Drunken Model, which achieves a 2.7% accuracy improvement and 0.543 loss reduction compared to the baseline system after tuning under the constraint of a parameter count of 442.67K. This meets the requirement of low complexity and demonstrates the feasibility of the proposed approach. Then, based on the optimal settings of the Drunken Model, a similar structured Normal Model is used and achieves an accuracy of 46.4%. Finally, we explore the idea of integrating knowledge distil-

lation as the fusion module in architecture. The proposed Drunkard Methodology achieves an accuracy of 45.2% on the DCASE2022 Task1 development dataset, which is 4.8% higher than the baseline system, demonstrating the effectiveness of the proposed approach. Of course, this module can also use other strategies such as multi-task learning and adversarial learning. We will conduct further research based on this methodology in the future.

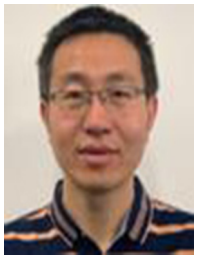
References

[1] I. Martín-Morató, F. Paissan, A. Ancilotto, et al., "Low-complexity acoustic scene classification in dcase 2022 challenge," arXiv preprint arXiv:2206.03835, 2022.
[2] W. Xie, Q. He, Z. Yu, and Y. Li, "Deep mutual attention network for acoustic scene classification," Digital Signal Processing, vol.123, p.103450, 2022.
[3] T. Choudhary, V. Mishra, A. Goswami, and J. Sarangapani, "A comprehensive survey on model compression and acceleration," Artificial Intelligence Review, vol.53, no.7, pp.5113–5155, 2020.
[4] M. Agarwal, S.K. Gupta, M. Biswas, and D. Garg, "Compression and acceleration of convolution neural network: a genetic algorithm based approach," Journal of Ambient Intelligence and Humanized Computing, vol.14, no.10, pp.13387–13397, 2022.
[5] B. Kim, S. Yang, J. Kim, et al., "QTI submission to DCASE 2021: Residual normalization for device-imbalanced acoustic scene classification with efficient design," arXiv preprint arXiv:2206.13909, 2022.
[6] J.H. Lee, J.H. Choi, P.M. Byun, et al., "HYU submission for the DCASE 2022: fine-tuning method using device-aware data-random-drop for device-imbalanced acoustic scene classification," 2022.
[7] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," IEEE transactions on pattern analysis and machine intelligence, vol.43, no.2, pp.652–662, 2021.
[8] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," Proceedings of the IEEE/CVF international conference on computer vision. 1314–1324, 2019.
[9] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient CNN architecture design," Proceedings of the European conference on computer vision (ECCV), pp.122–138, 2018.
[10] Y. Qu, X. Li, Z. Qin, and Q. Lu, "Acoustic scene classification based on three-dimensional multi-channel feature-correlated deep learning networks," Scientific Reports, vol.12, no.1, 2022.
[11] Y. Ding, Z. Zhang, X. Zhao, D. Hong, W. Cai, C. Yu, N. Yang, and W. Cai, "Multi-feature fusion: Graph neural network and CNN combining for hyperspectral image classification," Neurocomputing, vol.501, pp.246–257, 2022.
[12] W. Zou, D. Zhang, and D.-J. Lee, "A new multi-feature fusion based convolutional neural network for facial expression recognition," Applied Intelligence, vol.52, no.3, pp.2918–2929, 2022.
[13] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.7132–7141, 2018.
[14] G. Hinton, O. Vinyals and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[15] R. He, S. Sun, J. Yang, S. Bai, and X. Qi, "Knowledge distillation as efficient pre-training: Faster convergence, higher data-efficiency, and better transferability," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.9161–9171, 2022.
[16] G. Chen, J. Chen, F. Feng, S. Zhou, and X. He, "Unbiased Knowledge Distillation for Recommendation," Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, 2023.
[17] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.1580–1589, 2020.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770–778, 2016.
[19] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," arXiv preprint arXiv:1807.09840, 2018.
[20] Y. Li, W. Cao, W. Xie, Q. Huang, W. Pang, and Q. He, "Low-Complexity Acoustic Scene Classification Using Data Augmentation and Lightweight ResNet," 2022 16th IEEE International Conference on Signal Processing (ICSP). IEEE, pp.41–45, 2022.
[21] H. Zhang, M. Cisse, Y.N. Dauphin, et al., "Mixup: Beyond empirical risk minimization," arXiv preprint arXiv:1710.09412, 2017.
[22] D.S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E.D. Cubuk, and Q.V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019.
[23] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.13713–13722, 2021.
[24] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

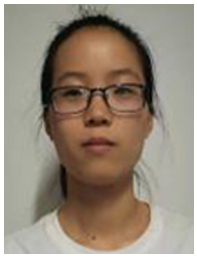
Wenkai Liu received the Ph.D. degree from the Institute of Semiconductors, Chinese Academy of Sciences in 2002. He is currently a Professor in North China University of Technology. His research interests include optical signal processing, machine learning, and neural networks.

Lin Zhang received B.S. degree in Communication Engineering from China University of Geosciences (Wuhan) in 2019. She is currently pursuing a Master's degree in Electronic Communication Engineering at North China University of Technology.

Menglong Wu received the Ph.D. degree in communications and information systems from Beijing University of Posts & Telecommunications of China in 2015. He is currently an Associate Professor with the School of Information Science and Technology, North China University of Technology, Beijing. His research interests include wireless communication, signal processing, machine learning, and neural networks.



Xichang Cai received the Ph.D. degree in mechanical and electronic engineering from Chinese Academy of Sciences. He is currently an Associate Professor with the School of Information Science and Technology, North China University of Technology, Beijing. His research interests include weak signal processing, machine learning, and embedded AI system.



Hongxia Dong received B.S. degree in electronic and information engineering from the Taiyuan Institute of Technology in 2020. She is currently pursuing a Master's degree in Electronic Communication Engineering at North China University of Technology.