InfantCryNet: A Data-driven Framework for Intelligent Analysis of Infant Cries

Mengze Hong MENGZE.HONG@CONNECT.POLYU.HK
Hong Kong Polytechnic University

Chen Jason Zhang JASON-C.ZHANG@POLYU.EDU.HK
Hong Kong Polytechnic University

Lingxiao Yang LINGX.YANG@POLYU.EDU.HK
Hong Kong Polytechnic University

Yuanfeng Song YFSONG@WEBANK.COM
WeBank Co., Ltd.

Di Jiang DIJIANG@WEBANK.COM
WeBank Co., Ltd.

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Understanding the meaning of infant cries is a significant challenge for young parents in caring for their newborns. The presence of background noise and the lack of labeled data present practical challenges in developing systems that can detect crying and analyze its underlying reasons. In this paper, we present a novel data-driven framework, "InfantCryNet," for accomplishing these tasks. To address the issue of data scarcity, we employ pre-trained audio models to incorporate prior knowledge into our model. We propose the use of statistical pooling and multi-head attention pooling techniques to extract features more effectively. Additionally, knowledge distillation and model quantization are applied to enhance model efficiency and reduce the model size, better supporting industrial deployment in mobile devices. Experiments on real-life datasets demonstrate the superior performance of the proposed framework, outperforming state-of-the-art baselines by 4.4% in classification accuracy. The model compression effectively reduces the model size by 7% without compromising performance and by up to 28% with only an 8% decrease in accuracy, offering practical insights for model selection and system design.

Keywords: Convolutional Neural Networks, Model Compression, Infant Cry Classification

1. Introduction

It is reported that more than 380,000 babies are born each day globally. For these newborns, the cry is their fundamental mode of communication with the outside world, serving as a critical indicator of their biological and psychological needs (Lockhart-Bouron et al., 2023). Experienced medical practitioners have developed the ability to interpret various cry sounds, enabling them to discern an infant's physical condition and, in some instances, identify potential diseases solely based on the characteristics of the cry. However, acquiring such a nuanced understanding quickly is impractical for new parents due to the lack of specific knowledge and experience (Jiang et al., 2023).

There is a pressing need for the development of accessible, user-friendly tools that can assist parents in comprehending their newborns' cries and supporting the overall well-being

of the infant (Laguna et al., 2023). However, this is challenging due to several factors. First, due to the presence of background noise in the living environment, effectively detecting the infant's cry is difficult (Yao et al., 2022; Chen et al., 2024b). Second, the diversity of newborn cry patterns complicates pinpointing the precise reasons for crying. Third, audio records of infant cries are typically rare, and those with annotated information are even harder to obtain, leading to data scarcity issues. Lastly, the detection and analysis of infant cries need to be done with low computing resources to enable deployment on mobile devices such as smartphones or tablets.

In this paper, we propose a novel framework named **InfantCryNet** that detects and analyzes infant cries seamlessly. Specifically, detection aims to identify whether an infant is crying in an audio clip, and analysis aims to identify the underlying needs (or reasons) of the cry. The proposed framework employs a state-of-the-art pre-trained audio model. It incorporates methods such as feature extraction, statistic pooling layers, and model compression to offer an innovative, industry-standard solution. The contributions of this work are summarized as follows:

- To overcome the data scarcity issue, a pre-trained model is employed to provide prior knowledge for downstream task solving, resulting in significant improvements over baseline models.
- To effectively extract features from audio clips of varying lengths, we propose the use of statistic pooling and multi-head attention pooling methods, achieving state-of-the-art performance among global pooling techniques.
- In an effort to enhance model efficiency and facilitate industrial applications, we experimented with the integration of model compression using knowledge distillation and model quantization, resulting in lightweight models with satisfactory performance.

The rest of this paper is organized as follows. Section 2 reviews the relevant literature. Section 3 describes the model architecture and proposed techniques. Section 4 reports the experimental results and discussions, and Section 5 concludes with limitations and recommendations for further research.

2. Related Work

This study has close ties to three areas of research: audio classification, model pre-training, and model compression.

2.1. Audio Classification with CNNs

The task of infant cry classification has seen significant advancements in recent years, evolving from traditional machine learning techniques to deep learning approaches. Among these methods, Convolutional Neural Networks (CNNs) have gained increasing popularity due to their effectiveness in extracting local patterns. Numerous applications have been developed based on CNNs, including acoustic event detection (Bae et al., 2016), music start detection (Schlüter and Böck, 2014), automatic speech recognition (Abdel-Hamid et al., 2014; Song

et al., 2022, 2021), and causal inference (Du et al., 2024a,b). Additionally, Graph Convolutional Networks (GCNs) have garnered interest for their ability to improve accuracy with limited labeled training data (Chen et al., 2024a; Abbaskhah et al., 2023).

2.2. Pretraining Model for Audio Tasks

In computer vision, models commonly undergo pre-training on labeled datasets such as ImageNet (Deng et al., 2009), which has a massive sample size. In natural language processing, pretraining methods based on Transformers (Vaswani et al., 2017) have been proposed, leading towards the trending Large Language Models (LLMs) (Lin et al., 2024). Motivated by these advances, audio pretraining has become a popular research topic. For instance, Wav2vec (Schneider et al., 2019) can be used to transfer audio signals into vector representations, which can improve the training of acoustic models. Similar to Imagenet, AudioSet (Gemmeke et al., 2017) is a massive audio dataset comprising more than 2 million audio samples, each tagged with 527 sound event labels. By pretraining on AudioSet, researchers (Kong et al., 2018, 2020) have proposed a variety of deep neural networks for audio classification and achieved promising performance.

In infant cry classification, labeled datasets are scarce due to the sensitivity of data collection and the high cost of annotation by pediatricians. Extracting features from pretrained models can effectively incorporate prior knowledge. The transfer learning technique leverages the rich feature representations learned by deep neural networks to train classifiers (Pan and Yang, 2010), thereby enhancing the performance for more specific tasks.

2.3. Model Compression

In the evolving landscape of machine learning techniques, there has been a notable shift in research focus from the development of "best" performing models to those that offer better efficiency and practicality for industrial applications (Choudhary et al., 2020). The advent of novel compression techniques such as knowledge distillation, network pruning, and quantization have enabled the deployment of sophisticated models in many practical use cases (Hinton et al., 2015; Zhou et al., 2017). Moreover, with the emergence of large pre-trained models, Li et al. (2020) suggests the strategy of training very large models and then compressing them to obtain a relatively smaller model, which has higher accuracy than directing training a smaller model. Similar work has also shown that model compression techniques can effectively reduce the size of neural networks without significantly compromising accuracy (Polino et al., 2018), which is particularly beneficial for scenarios where computational resources are at a premium, such as in edge computing environments.

Our work diverges from existing infant cry systems by addressing data scarcity with pre-trained models, enhancing CNN pooling methods for better feature extraction, and providing a lightweight, mobile-deployable solution through model compression.

3. Framework Architecture

In this section, we first introduce the feature extraction methods for infant cry audio in Section 3.1. The two targeted tasks, namely crying detection and crying analysis, are

introduced in Section 3.2 and 3.3, together with the proposed pooling techniques. The model compression and compression techniques are introduced in Section 3.4.

3.1. Feature Extraction

In audio processing, the feature extraction can be divided into time and frequency domains. Time-domain features, such as amplitude and zero-crossing, provide straightforward insights but limited information for complex tasks due to their simplicity. In contrast, frequency domain features, including MFCCs, LPCCs, and LFCCs, have been proven to achieve superior results with more discriminative features to be learned by the model (Hertel et al., 2016). These features can be represented by two approaches: the **waveform** that reflects the pattern of sound pressure amplitude in the time domain, and the **spectrogram** which visually depicts a signal's frequency spectrum.

The comparison between infant cry sounds and adult voices is illustrated in Figure 1, using both waveform and spectrogram representations. Unlike adult speech, which tends to be more irregular and has lower amplitude, infant cries typically exhibit a more prosodic waveform and larger amplitude. This periodic nature makes infant cries particularly well-suited for combined analysis of prosodic and time-frequency domain features. The spectrogram, as a comprehensive visual representation, effectively displays both acoustic and prosodic characteristics, providing clear insights into how the signal's frequency content varies over time and highlighting the unique properties of infant cries.

With this feature representation, the problem of classifying audio has been transformed into an image classification problem, motivating the following discussions on constructing convolutional neural networks. Determining the right network architecture is crucial for developing the infant cry system. Thus, we present two distinct model architecture options in the subsequent sections, specifically focusing on the tasks of infant cry detection and classification. These proposed architectures aim to effectively capture relevant information and address the aforementioned challenges in the infant cry system.

3.2. Infant Cry Detection

For infant cry detection, as shown in Figure 2(a), we utilize the 10-layer CNNs (CNN-10) with four convolutional blocks, each consisting of two 3×3 kernel-sized convolutional layers, separated by batch normalization to improve training efficiency and stability. The ReLU activation function is utilized for this purpose. After each block, the spatial dimensions of the feature maps are reduced via a 2×2 average pooling operation. The final feature maps are summarized into a fixed-length vector through another pooling operation and fed into a softmax activation function to generate class probabilities for the binary classification of whether an audio clip contains an infant cry.

3.3. Infant Cry Analysis

Based on the available dataset, the infant cries can be classified into six distinct reasons: *awake*, *hug*, *sleepy*, *uncomfortable*, *diaper*, and *hungry*. Here, we utilize the 14-layer CNNs (CNN-14), as shown in Figure 2(b), which has two more convolutional blocks in comparison with CNN-10 for enhanced feature extraction and improved classification accuracy in identifying the distinct reasons behind infant cries.

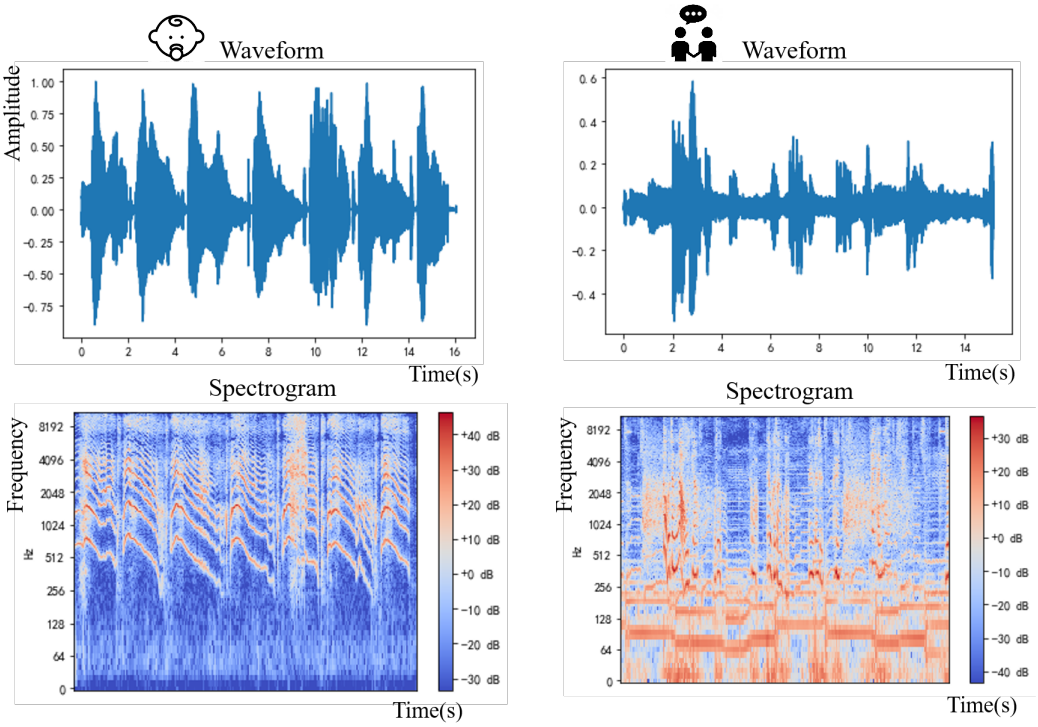

Figure 1: Infant cry (left) vs. adult voice (right) in waveform and spectrogram

Suppose the segment level embedding is $\boldsymbol{H} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_N\}$. The common pooling methods include max pooling and average pooling. Max pooling selects the maximum value in each patch, with the limitation that only one instance represents the whole audio while other instances are ignored. This can be described as follows:

$$h_{max} = max(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_N) \tag{1}$$

Average pooling calculates the average value of each patch, which considers all instances instead of the maximum. However, since each instance contributes equally, it fails to reflect the variance within the patch:

$$h_{avg} = \frac{\sum_{i=1}^{N} \boldsymbol{h}_i}{N} \tag{2}$$

In order to combine the advantages of max pooling and average pooling, the pre-trained audio neural networks (PANNs) calculated a fixed-length vector by adding the averaged and maximized vectors (Kong et al., 2020). However, the difference between instances is still ignored, and the method is formulated as follows:

$$h_{add} = max(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_N) + \frac{\sum_{i=1}^{N} \boldsymbol{h}_i}{N} \tag{3}$$

Recognizing the limitations of existing methods, we propose two kinds of global pooling methods: statistic pooling and multi-head attention pooling.

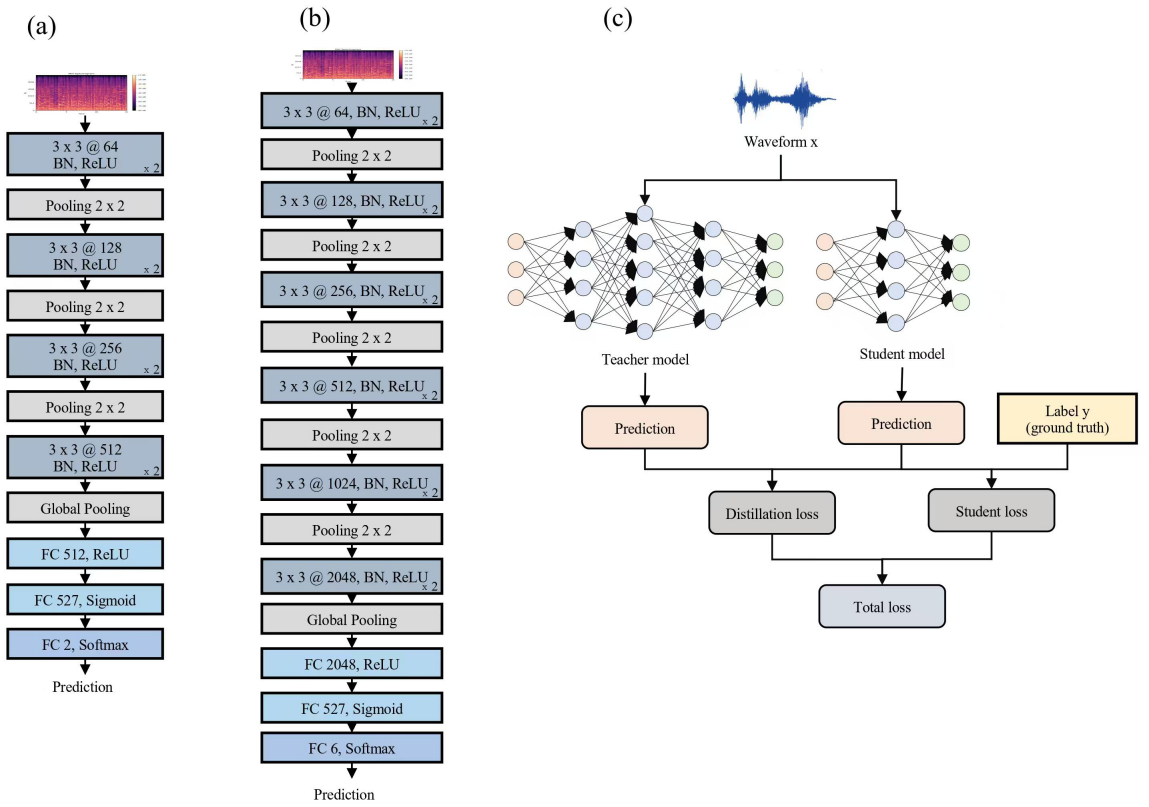

Figure 2: Model architect: (a) CNN10 for infant cry detection, (b) CNN14 for infant cry classification, (c) Knowledge Distillation for model compression.

- **Statistic pooling** calculates the mean and variance, then uses a linear layer to reduce two dimensions to one. Both the average and variability of values are leveraged to represent the feature. This can be summarized as:

$$s^2 = \frac{\sum_{i=1}^{N}(\boldsymbol{h}_i - h_{avg})^2}{N} \tag{4}$$

$$h_{stat} = fc(h_{avg}, s^2) \tag{5}$$

  where $fc(*)$ represents linear layer and h_{avg} denotes the average pooling.

- **Multi-head attention pooling** determines the attention distribution of each instance, allowing instances to be aggregated based on their significance rather than equally contributing to the feature representation, as shown below:

$$h_{attn} = \frac{\sum_{i=1}^{N}(\alpha_i \boldsymbol{h}_i)}{\sum_{i=1}^{N}(\alpha_i)} \tag{6}$$

  where $\alpha = Wh$, h denotes the output from the previous layer, and $W \in \mathbb{R}^{N \times N}$ is the learned attention parameter.

Task	Training	Testing
infant cry detection	6,000	600
infant cry classification	750	85

Table 1: Overview of dataset

3.4. Model Compression

For challenging tasks, models with complex network architectures typically perform better. However, the computation for such networks can be slow, posing the challenge of achieving good performance with a lightweight network. This paper addresses this problem by employing knowledge distillation (Hinton et al., 2015), which transfers knowledge from a complex "teacher" model to a simpler "student" model without sacrificing accuracy. It also considers model quantization technique (Jacob et al., 2018), which performs some or all operations on tensors using integers instead of floating-point values.

For knowledge distillation, we utilize ResNet22 (He et al., 2016) as the student model and CNN14 as the teacher model, with the distillation process illustrated in Figure 2(c). The loss function for the student model is carefully crafted to address two main discrepancies. First, it measures the divergence between the predictions made by the student model and the actual ground truth data, ensuring that the student learns to approximate the correct outputs effectively. Second, it accounts for the variance between the softened predictions of the student model and the softened labels provided by the teacher model. These softened labels are generated by applying a temperature scaling factor to the teacher's logits, which allows the student model to capture more nuanced information about the output distribution.

On the other hand, we implement dynamic quantization, which involves converting the weights and activation of the model from floating-point to integer values with scale factors determined dynamically at runtime, leading to a smaller model size and faster inference.

4. Experiments

In this section, we outline the experiments conducted to evaluate the performance of the proposed methods. We detail the datasets used, the baselines for comparison, and the implementation specifics. Additionally, we conduct several ablation studies to analyze the effectiveness of different components, suggesting best practices for implementing the system.

4.1. Experimental Setup

4.1.1. Dataset

In the task of infant cry detection, a dataset consisting of 6,600 audio clips is used, with 6,000 clips for training and 600 clips for validation (see Table 1). Each audio clip has a 15-second duration and a sampling rate of 16,000 Hz. Similarly, the dataset for the infant cry classification task contains 835 audio clips, each labeled by one of six reasons for crying (i.e., awake, hug, sleepy, uncomfortable, diaper, and hungry).

Model	Accuracy
CNN10	99.8%
CNN14	99.8%

Table 2: Performance of infant cry detection

4.1.2. Baselines

The performance of the proposed model is compared with various baselines and reported in terms of classification accuracy. The baseline models include:

- **CNN10** The 10-layer CNN, which features only four convolutional layers, is a simpler architecture compared to CNN14, which has six convolutional layers.
- **Resnet22** Each block in the ResNet comprises two convolutional layers with 3×3 kernel sizes and a shortcut connection between convolutional layers. Here, a 22-layer deep ResNet (Resnet22) with eight basic blocks is considered.
- **Wavegram-Logmel-CNN** In comparison to one-dimensional CNN that cannot capture frequency information, Kong et al. (2020) proposed Wavegram-Logmel-CNN, which combined wavegram (time-domain) and Log-Mel spectrogram (frequency-domain) as input to a CNN.

4.1.3. Implementation Details

The infant cry detection task involves identifying the presence of crying activity in an audio clip. To accomplish this, we first extracted the Log-Mel spectrogram, which serves as the input features for our model, and then fine-tuned a 10-layer CNN from PANNs for this purpose. The infant cry classification involves categorizing audio clips according to the reasons for crying. Similarly, we begin by extracting Log-Mel spectrograms, followed by fine-tuning a 14-layer CNN from PANNs to build an effective classification system. The Adam optimizer is used for both models, and the training was conducted on eight Nvidia Tesla V100 32GB GPUs with a batch size of 256.

4.2. Accuracy Analysis

For infant cry detection, the results of our proposed method and a baseline model are shown in Table 2, where both models exhibit similar accuracy. This can be attributed to the fact that crying sounds are relatively distinct and easy to detect compared to other sounds, enabling even a simple network to perform effectively. The high amplitude observed in Figure 1 contributes to the mitigation of the effect of background noise in the audio. Hence, it is advisable to choose computationally efficient models that can still achieve good performance in detecting infant cries.

For the classification task, the comparison between our method (i.e., CNN14 with pre-training) and various baselines is presented in Table 3. The proposed model outperforms all baselines, achieving an encouraging 4.4% improvement in accuracy over the best-performing

Model	Accuracy
Wavegram-Logmel-CNN with pretraining	70.33%
Resnet22 with pretraining	53.85%
CNN10 with pretraining	52.75%
CNN14	64.84%
CNN14 with pretraining	**74.73%**

Table 3: Performance of infant cry classification with different network architectures

Pooling Method	Accuracy
Max pooling	62.64%
Average pooling	70.33%
Max pooling + average pooling	71.42%
Statistic pooling (ours)	**74.73%**
Multi-head attention pooling (ours)	72.53%

Table 4: Performance of infant cry classification with different pooling methods

method. This demonstrates the effectiveness of the proposed network architecture for solving the classification problem, and these results support the hypothesis that a pre-trained model can significantly improve performance compared to those without pretraining.

4.3. Ablation Study

By focusing on the accuracy comparison of the same model (i.e., CNN14 with pretraining) under different pooling methods (see Table 4), our findings revealed that average pooling consistently outperformed max pooling among the baseline methods, underscoring the advantage of incorporating all instances in the calculation. While the combination of max pooling and average pooling provided only a small improvement, our proposed methods demonstrated superior performance. Notably, statistic pooling achieved the highest accuracy, which can be attributed to the periodic nature of infant cries, where varying attention to instances is not always necessary. Attention pooling ranked as the second-best option. Overall, our proposed methods outperformed the baselines, achieving significant improvements in accuracy over the best-performing baseline method.

4.4. Model Compression

The results in Table 5 depict the effectiveness of the model compression techniques. By applying model quantization, the model size is reduced by 7% without any loss of accuracy. Knowledge distillation reduces the model size by 21% but also lowers the accuracy by 7%. By combining knowledge distillation with model quantization, the model size is reduced by 28%, and the accuracy is only reduced by 8%.

Based on the observations, we conclude that model compression techniques can effectively reduce model size with a manageable compromise on accuracy, depending on the method chosen. Model quantization is ideal if the model size requirement is not strict, as

Model	Accuracy	Model Size
CNN14 (teacher)	73.63%	81M
Resnet22 (student)	53.85%	64M
Knowledge distillation	68.23%	64M
Model quantization	73.63%	75M
Model quantization + distillation	67.42%	58M

Table 5: Effectiveness of model compression in infant cry classification

it reduces size without losing accuracy. However, for significant size reduction, combining model quantization with knowledge distillation is effective, though it may slightly decrease accuracy. This paper suggests that the choice of compression technique should align closely with the application's specific needs. For the infant cry system, given the increasing capability of mobile devices, model quantization is recommended to retain optimal accuracy.

5. Conclusion

This paper presents "InfantCryNet," an innovative framework designed to detect and comprehend the meanings behind infant cries, addressing the challenges of background noise and limited labeled data. By utilizing pre-trained audio models, along with statistical and multi-head attention pooling techniques, we enhanced feature extraction capabilities and achieved significant improvement in classification accuracy. To facilitate the deployment on mobile devices, our approach incorporates knowledge distillation and model quantization to compress model size, offering practical solutions to real-world applications.

While the experiments present promising results, this study extensively focused on comparing neural network solutions with varying complexity levels. Although traditional machine learning methods generally lag in accuracy, their lightweight nature and training efficiency suggest the potential for exploring hybrid models, such as CNN-SVM and GMM-CNN, for developing more efficient infant cry classification systems. Furthermore, the challenge of data scarcity can be mitigated by adopting a federated learning approach, which enables the collection of more training data while protecting end-user privacy (Jiang et al., 2021; Tan et al., 2020; Jiang et al., 2019).

References

Ahmad Abbaskhah, Hamed Sedighi, and Hossein Marvi. Infant cry classification by mfcc feature extraction with mlp and cnn structures. *Biomedical Signal Processing and Control*, 86:105261, 2023. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2023.105261. URL https://www.sciencedirect.com/science/article/pii/S1746809423006948.

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10):1533–1545, 2014.

Soo Hyun Bae, Inkyu Choi, and Nam Soo Kim. Acoustic scene classification using parallel combination of lstm and cnn. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, pages 11–15, 2016.

Yifan Chen, Haiqi Zhu, and Zhiyuan Chen. Multi-dgi: Multi-head pooling deep graph infomax for human activity recognition. *Mobile Networks and Applications*, pages 1–12, 2024a.

Yixin Chen, Di Jiang, Conghui Tan, Yuanfeng Song, Chen Zhang, and Lei Chen. Neural moderation of asmr erotica content in social networks. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):275–280, 2024b. doi: 10.1109/TKDE.2023.3283501.

Tejalal Choudhary, Vipul Mishra, Anurag Goswami, and Jagannathan Sarangapani. A comprehensive survey on model compression and acceleration. *Artificial Intelligence Review*, 53:5113–5155, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Xiaojing Du, Jiuyong Li, Debo Cheng, Lin Liu, Wentao Gao, and Xiongren Chen. Estimating peer direct and indirect effects in observational network data, 2024a. URL https://arxiv.org/abs/2408.11492.

Xiaojing Du, Feiyu Yang, Wentao Gao, and Xiongren Chen. Causal gnns: A gnn-driven instrumental variable approach for causal inference in networks, 2024b. URL https://arxiv.org/abs/2409.08544.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Lars Hertel, Huy Phan, and Alfred Mertins. Comparing time and frequency domain for audio event recognition using deep learning. *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3407–3411, 2016. URL https://api.semanticscholar.org/CorpusID:7361616.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2704–2713, 2018.

Di Jiang, Yuanfeng Song, Yongxin Tong, Xueyang Wu, Weiwei Zhao, Qian Xu, and Qiang Yang. Federated topic modeling. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1071–1080, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450369763. doi: 10.1145/3357384.3357909. URL https://doi.org/10.1145/3357384.3357909.

Di Jiang, Conghui Tan, Jinhua Peng, Chaotao Chen, Xueyang Wu, Weiwei Zhao, Yuanfeng Song, Yongxin Tong, Chang Liu, Qian Xu, Qiang Yang, and Li Deng. A gdpr-compliant ecosystem for speech recognition with transfer, federated, and evolutionary learning. *ACM Transactions on Intelligent Systems and Technology*, 12(3), may 2021. ISSN 2157-6904. doi: 10.1145/3447687. URL https://doi.org/10.1145/3447687.

Di Jiang, Chen Zhang, and Yuanfeng Song. *Probabilistic topic models: Foundation and application*. Springer, 2023.

Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D Plumbley. Audio set classification with attention model: A probabilistic perspective. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 316–320. IEEE, 2018.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Ana Laguna, Sandra Pusil, Àngel Bazán, Jonathan Adrián Zegarra-Valdivia, Anna Lucia Paltrinieri, Paolo Piras, Clàudia Palomares i Perera, Alexandra Pardos Véglia, Oscar Garcia-Algar, and Silvia Orlandi. Multi-modal analysis of infant cry types characterization: Acoustics, body language and brain signals. *Computers in Biology and Medicine*, 167:107626, 2023. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2023.107626. URL https://www.sciencedirect.com/science/article/pii/S0010482523010910.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5958–5968. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/li20m.html.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *ArXiv*, abs/2403.20271, 2024. URL https://api.semanticscholar.org/CorpusID:268793990.

Marguerite Lockhart-Bouron, Andrey Anikin, Katarzyna Pisanski, Siloé Corvin, Clément Cornec, Léo Papet, Florence Levréro, Camille Fauchon, Hugues Patural, David Reby, and Nicolas Mathevon. Infant cries convey both stable and dynamic information about age and identity. *Communications Psychology*, 2023. URL https://api.semanticscholar.org/CorpusID:263665971.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. Model compression via distillation and quantization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S1XolQbRW.

Jan Schlüter and Sebastian Böck. Improved musical onset detection with convolutional neural networks. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 6979–6983. IEEE, 2014.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Yuanfeng Song, Xiaoling Huang, Xuefang Zhao, Di Jiang, and Raymond Chi-Wing Wong. Multimodal n-best list rescoring with weakly supervised pre-training in hybrid speech recognition. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 1336–1341, 2021. doi: 10.1109/ICDM51629.2021.00167.

Yuanfeng Song, Rongzhong Lian, Yixin Chen, Di Jiang, Xuefang Zhao, Conghui Tan, Qian Xu, and Raymond Chi-Wing Wong. A platform for deploying the tfe ecosystem of automatic speech recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6952–6954, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547731. URL https://doi.org/10.1145/3503161.3547731.

Conghui Tan, Di Jiang, Huaxiao Mo, Jinhua Peng, Yongxin Tong, Weiwei Zhao, Chaotao Chen, Rongzhong Lian, Yuanfeng Song, and Qian Xu. Federated acoustic model optimization for automatic speech recognition. In Yunmook Nah, Bin Cui, Sang-Won Lee, Jeffrey Xu Yu, Yang-Sae Moon, and Steven Euijong Whang, editors, *Database Systems for Advanced Applications*, pages 771–774, Cham, 2020. Springer International Publishing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Xuewen Yao, Megan Micheletti, Mckensey Johnson, Edison Thomaz, and Kaya de Barbaro. Infant crying detection in real-world environments. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2022. doi: 10.1109/ICASSP43922.2022.9746096.

Aojun Zhou, Anbang Yao, Yiwen Guo, Lin Xu, and Yurong Chen. Incremental network quantization: Towards lossless cnns with low-precision weights. *arXiv preprint arXiv:1702.03044*, 2017.