M3Net: Efficient Time-Frequency Integration Network with Mirror Attention for Audio Classification on Edge

Xuanming Jiang[1,3], Baoyi An[2,3], Guoshuai Zhao[1,4*], Xueming Qian[1,4]

[1]School of Software Engineering, Xi'an Jiaotong University, Xi'an, China [2]School of Physical Science and Technology, Lanzhou University, Lanzhou, China

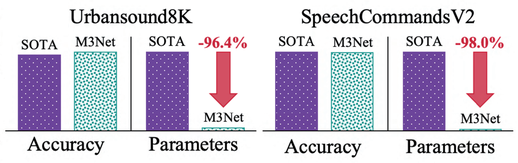
[3]Xi'an Jiyun Technology Co., Ltd., Xi'an, China [4]Shaanxi Yulan Jiuzhou Intelligent Optoelectronic Technology Co., Ltd., Xi'an, China

jiangxm24@stu.xjtu.edu.cn, anby20@lzu.edu.cn, {guoshuai.zhao, qianxm}@mail.xjtu.edu.cn

Overview

• We introduce the **Time-Frequency Separation Method** of extracting separated time-domain and frequency-domain features in parallel to enhance the effectiveness of audio content usage.

• We propose the **Mirror Attention Mechanism** for the integration of separated time-frequency features based on the robust features that extend the original audio pattern set like what humans do.



Performance Comparison Between M3Net and the Accuracy-Comparable SOTAs

• Experiments demonstrate the proposed **Mini Mirror Multi-View Network (M3Net)** with merely around 0.03M parameters can achieve SOTA-comparable performance on the UrbanSound8K and SpeechCommandsV2 datasets, with the parameter count being **less than 10%** of those in other accuracy-comparable SOTAs, and even without any data augmentation or pre-training.

• M3Net exhibits comparable classification performance to the other SOTAs with dozens of times more parameters in both environmental noise and speech commands datasets without data augmentation and pre-training. Achievements in SOTA comparisions and ablation studies indicate M3Net has the potentiality to facilitate the deployment of high-performance but lightweight models on edge devices.

Experimental Results

Dataset: UrbanSound8K		Accuracy	# Param	Wilcoxon	Data	Pre-training
Methods	Features	(%)	(×10^6)	p-value	augmentation	
AemNet-DW	Log-Mel	82.25	0.9	< 5.0e-2		✓
ULSED	Log-Mel	83.5	0.34	< 5.0e-2	✓	
2D CNN	GFCC	89	1.8	< 5.0e-2	✓	
PhiNets M40	Mel-Spectrogram	76.3	**0.027**	< 5.0e-2	✓	
SE-TCAM 1D CNN	Raw audio	94.04	0.81	< 5.0e-2		
M3Net (Ours)	Mel-Spectrogram	**97.44**	0.029	baseline		

Dataset: SpeechCommandsV2		Accuracy	# Param	Wilcoxon	Data	Pre-training
Methods	Features	(%)	(×10^6)	p-value	augmentation	
DeLoRes M	Log-Mel	89.7	5.3	< 5.0e-2	✓	✓
AdaptFormer	Log-Mel	92.3	1.43	< 5.0e-2		✓
DCLS-Delays	Mel-Spectrogram	95.35	2.5	< 5.0e-2	✓	
SeqBoat	Raw audio	97.35	0.293	2.2e-1		
EAT-S-GMME	Raw audio	**97.88**	1.54	1.0e-1	✓	
M3Net (Ours)	Mel-Spectrogram	97.03	**0.031**	baseline		

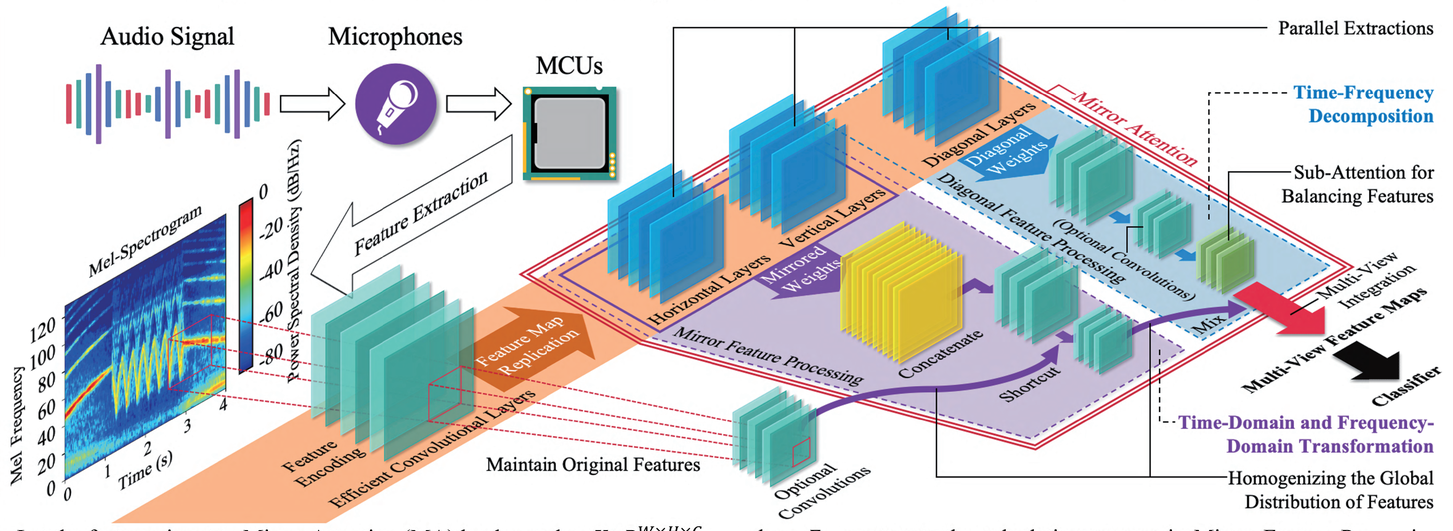
****The ablation study can be found in our paper.***

Acknowledgments

This work is jointly supported by National Natural Science Foundation of China (No. 62372364), Technical Innovation Guidance Plan of Shaanxi Province, China (No. 2024QCY-KXJ-199), and honored to receive initial support from Hui-Chun Chin and Tsung-Dao Lee Chinese Undergraduate Research Endowment (LZU-JZH2619, LZU-JZH2620). We gratefully honor the memory of ***Mrs. Chin*** and ***Prof. Lee***, whose immortal legacy will inspire generations of scholars.

Methodology

Edge-Based Audio Classification Through M3Net and Mel-Spectrogram Feature Extraction



Let the features input to Mirror Attention (MA) be denoted as $\boldsymbol{X}\in R^{W\times H\times C}$, where C represents the number of channels, W and H denote the width and height of the feature maps, respectively. For any element X in the input time-frequency features $\boldsymbol{X}$, their relationship can be defined as:

$$\boldsymbol{X}=\{X(w,h,c)|w\in[0,W-1],h\in[0,H-1],c\in[0,C-1]\}$$

where the width, height and channel index of X are denoted as w, h, and c.

The supervised time-domain and frequency-domain transformation results $X_T\in\boldsymbol{X}_T$ and $X_F\in\boldsymbol{X}_F$ that mimic human cognition can be expressed as:

$$X_{T,c\in[0,C-1]}(w,h)=M_{HMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\qquad M_{HMF}=\begin{bmatrix}-1&0&W\\0&1&0\\0&0&1\end{bmatrix}$$

$$X_{F,c\in[0,C-1]}(w,h)=M_{VMF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\qquad M_{VMF}=\begin{bmatrix}1&0&0\\0&-1&H\\0&0&1\end{bmatrix}$$

where θ represents the extent of transformation in the time and frequency domains of the audio (****The specific meaning can be found in our presentation***). The features transformed by M_{HMF} and M_{VMF} are concatenated along the channel dimension, and then perform a depth-wise separable convolution (DSConv) to align the feature dimensions with those from the shortcut connection originating from the original features. Finally, the outcome of the element-wise addition is convolved with another DSConv to obtain the time-frequency double-transformed features $\boldsymbol{X}_{MFP}$:

$$\boldsymbol{X}_{MFP}=F_M(M_{HMF}\boldsymbol{X},M_{VMF}\boldsymbol{X}),\qquad \boldsymbol{X}\in R^{W\times H\times C}$$

where F_M represents the calculation process in Mirror Feature Processing distinct from M_{HMF} and M_{VMF}.

For element $X(w,h,c)$ in the input features $\boldsymbol{X}$, the time-frequency separated result $X_D\in\boldsymbol{X}_D$ can be expressed as:

$$X_{D,c\in[0,C-1]}(w,h)=M_{CSF}\begin{bmatrix}w\\h\\\theta\end{bmatrix},\qquad M_{CSF}=\begin{bmatrix}-1&0&W\\0&-1&H\\0&0&1\end{bmatrix}$$

As demonstrated in above figure, $M_{CSF}\boldsymbol{X}$ is first individually process additional DSConv and one-dimensional Efficient Channel Attention to obtain the separated time-frequency feature weights that are equivalent to those in the other branch of MA. The resulting features $\boldsymbol{X}_{DFP}$ is capable of extracting the symmetry of global and local features that are distinct from $\boldsymbol{X}_{MFP}$ and $\boldsymbol{X}$, and can be expressed as:

$$\boldsymbol{X}_{DFP}=F_D(M_{CSF}\boldsymbol{X})=F_D(M_{HMF}M_{VMF}\boldsymbol{X}),\qquad \boldsymbol{X}\in R^{W\times H\times C}$$

where F_D is the calculation process in Diagonal Feature Processing beyond M_{CSF}. The final operation in MA is to multiply the separated features output from the two branches above. The output $\boldsymbol{X}_{MA}$ can be expressed as:

$$\boldsymbol{X}_{MA}(w,h,c)=\boldsymbol{X}_{MFP}(w,h,c)\odot\boldsymbol{X}_{DFP}(w,h,c)$$

where $\odot$ denotes the Hadamard product of matrices.

M3Net transcends the conventional emulation of human auditory perception by CNNs and ventures into the field of replicating human counterintuitive supervised cognitive processes through Mirror Attention.

The 39th Annual AAAI Conference on Artificial Intelligence

FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, PENNSYLVANIA, USA

西安交通大學 XI'AN JIAOTONG UNIVERSITY

蘭州大學 LANZHOU UNIVERSITY

极韵 | DREAM

SMILES LAB 域 览 九 州