AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake

23年 TIFS（尚未开源）

1. 目标：多模态，伪造检测

2. 数据集：DefakeAVMiT（自建，尚未开源），FakeAVCeleb， DFDC，均是v+a

3. 方法：AVoiD-DF：时空编码器（TSE）+ 多模态联合解码器（MMD）+ 跨模态分类器

①TSE 以双流结构提取视听特征，将时间序列和空间位置嵌入到同步的音频和视频中，做为多模态融合的特征。

时空编码器TSE=时间编码器TE+空间编码器SE，空间标记从单帧的空间小块中收集信息，时间标记从跨帧的小块中收集信息；

②MMD 用于共同学习视听不一致性，基于双向交叉注意力模块融合视听信号，并共同学习模态间的不和谐性以进行深度伪造检测；

③跨模态分类器，$cls^{a,v} = \mathrm{FC}(concat(cls^a, cls^v))$,

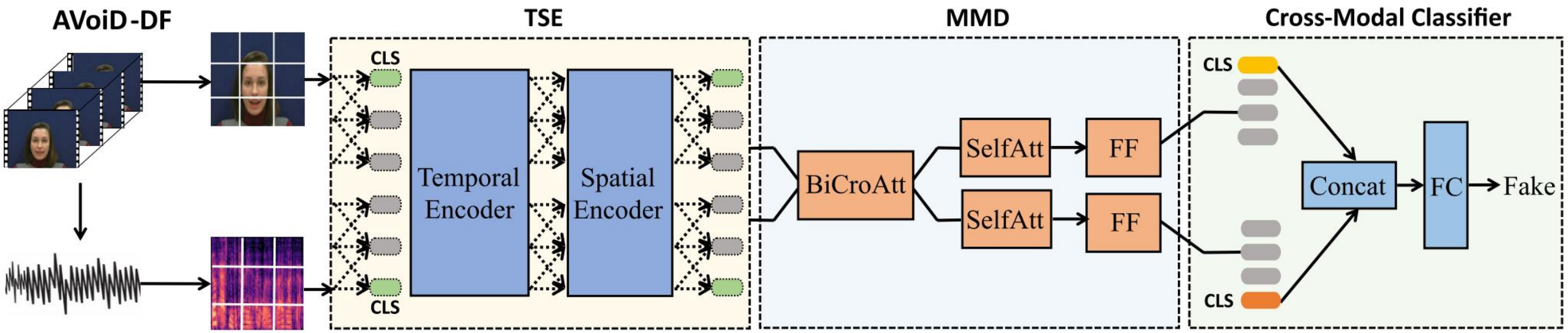
$L_{all} = \alpha L_{cro} + \beta L_{con} + \zeta L_{aam}$,

对$cls^{a,v}$使用对比损失👇，对cls^a和cls^v分别使用交叉熵损失👇，对真实标签使用加性角度 margin 损失👇，总损失👆

$$L_{con} = \frac{1}{N^2}\sum_{i}^{N}[\sum_{y_i=y_j}^{N}(1-\mathrm{Sim}(cls_i, cls_j)) + \sum_{y_i\neq y_j}^{N} max((\mathrm{Sim}(cls_i, cls_j)-\alpha), 0)],$$

$$L_{cro} = \frac{1}{S^v}\sum_{k=1}^{S^v} l'(cls_k^v, y_k^v) + \frac{1}{S^a}\sum_{k=1}^{S^a} l'(cls_k^a, y_k^a),$$

$$L_{aam} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\cdot cos(\theta_{y_i}+m)}}{e^{s\cdot cos(\theta_{y_i}+m)} + \sum_{j=1, y_j\neq y_i}^{N} e^{s\cdot(cos(\theta_j))}},$$



AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake

23年 TIFS（尚未开源）

4. TSE=时间编码器TE+空间编码器SE：

①整体结构：L个Transformer层的双序列组成。

每层=层归一化LN+多头自注意力MSA+前馈FF

$$\mathbf{I}^v = \frac{1}{t}[\mathbf{V}_{1|T} + \mathbf{V}_{2|T} + \cdots + \mathbf{V}_{t|T}],$$

②TE：- 对视觉模态应用时间池化，帧t在时间维度T上进行平均

- 在 MSA 中向视听输入特征添加时间索引

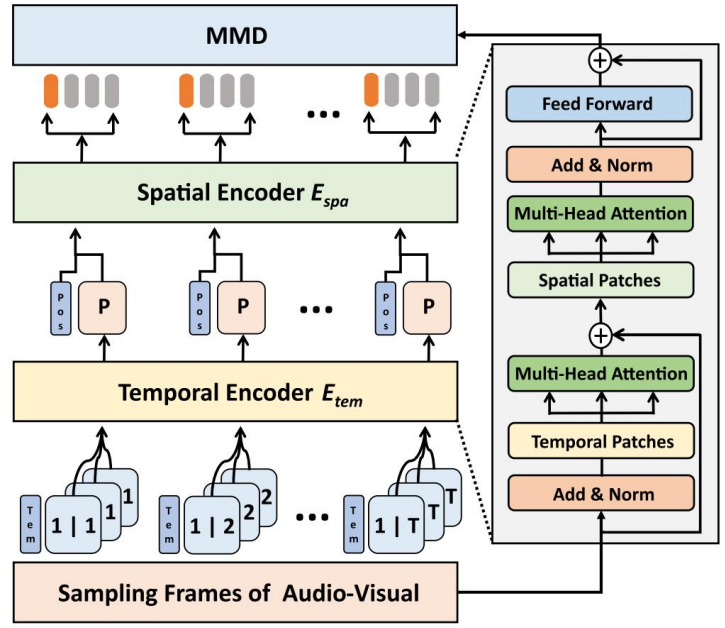
$$\mathbf{T}_l^v = \mathrm{MSA}(\mathrm{LN}(\mathbf{T}_{l-1}^v), Tem_i) + \mathbf{T}_{l-1}^v,$$
$$\mathbf{T}_l^a = \mathrm{MSA}(\mathrm{LN}(\mathbf{T}_{l-1}^a), Tem_i) + \mathbf{T}_{l-1}^a,$$

- l 表示第 l 层编码器，Tem_i是同一剪辑内视听的时间标记，

$$\mathbf{P}_l^v = [\mathbf{P}_1^v, \mathbf{P}_2^v, \ldots, \mathbf{P}_n^v] + Pos_j,$$
$$\mathbf{P}_l^a = [\mathbf{P}_1^a, \mathbf{P}_2^a, \ldots, \mathbf{P}_n^a] + Pos_k,$$

③SE：- 将视觉的关键帧和音频的梅尔频谱图分割成小块并添加位置标记

- P_i表示在相同位置的第 i 个通道的特征，共享相同的位置标记嵌入，SE的输出为 $\mathbf{P}^v = \mathrm{FF}(\mathrm{MSA}(\mathrm{LN}(\mathbf{P}_{l-1}^v))) + \mathbf{P}_{l-1}^v,$



5. MMD=（2x自注意力层 + 1x双向交叉注意力模块 + 2x前馈层）x L

①BiCroAtt模块的过程可以定义为：

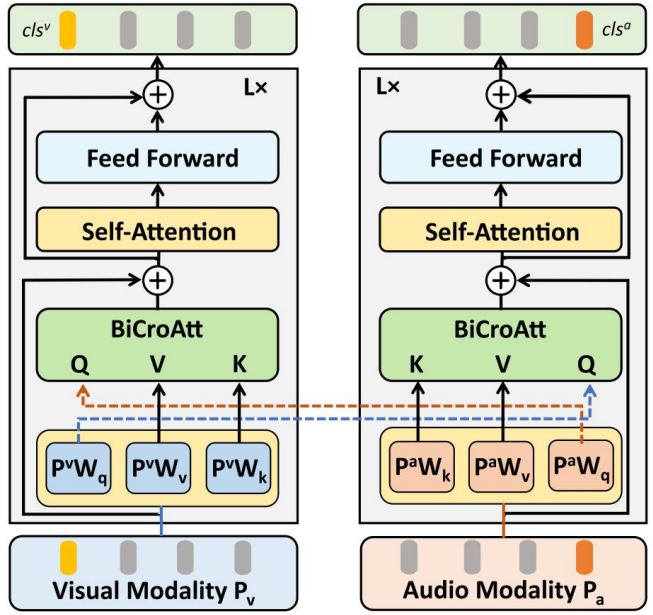
$$\mathbf{Z}^a = \mathrm{BiCroAtt}(\mathbf{Q}^v, \mathbf{K}^a, \mathbf{V}^a) = \mathrm{softmax}(\frac{\mathbf{P}^v\mathbf{W}_q \cdot (\mathbf{P}^a\mathbf{W}_k)^{\mathrm{T}}}{\sqrt{d_k}})\mathbf{P}^a\mathbf{W}_v,$$

$$\mathbf{Z}^v = \mathrm{BiCroAtt}(\mathbf{Q}^a, \mathbf{K}^v, \mathbf{V}^v) = \mathrm{softmax}(\frac{\mathbf{P}^a\mathbf{W}_q \cdot (\mathbf{P}^v\mathbf{W}_k)^{\mathrm{T}}}{\sqrt{d_k}})\mathbf{P}^v\mathbf{W}_v,$$

②BiCroAtt的输出为 $\mathbf{Z}_l^v = \mathrm{LN}(\mathbf{Z}_{l-1}^v + \mathrm{BiCroAtt}(\mathbf{Z}_{l-1}^v, \mathbf{Z}_{l-1}^a)),$

③SelfAtt 的输出为 $\mathbf{Z}_l^{\hat{v}} = \mathrm{LN}(\mathbf{Z}_l^v + \mathrm{SelfAtt}(\mathbf{Z}_l^v)),$

④联合解码器第 L 层的输出为 $\mathbf{v}^l = \mathrm{LN}(\mathbf{v}^{l-1} + \mathrm{FF}(\mathbf{Z}_l^{\hat{v}})),$



AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake

23年 TIFS（尚未开源）

Methods	Modality	DefakeAVMiT		FakeAVCeleb		DFDC	
		ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
MesoNet [61] (2018)	visual	80.1	82.5	57.3	60.9	71.7	75.3
Capsule [62] (2019)	visual	68.5	71.4	68.8	70.9	50.2	53.3
Head Pose [3] (2019)	visual	50.6	52.2	45.6	49.2	51.4	55.9
VA-MLP [63] (2019)	visual	56.9	59.8	65.0	67.1	59.3	61.9
Xception [50] (2019)	visual	49.4	51.0	67.9	70.5	46.5	49.9
LipForensics [8] (2021)	visual	73.1	77.2	80.1	82.4	71.3	73.5
DeFakeHop [64] (2021)	visual	83.6	86.1	68.3	71.6	78.9	81.1
CViT [65] (2021)	visual	72.9	74.5	69.7	71.8	62.1	63.7
Multiple-Attention [18] (2021)	visual	85.1	87.3	77.6	79.3	82.5	84.8
SLADD [66] (2022)	visual	73.3	77.5	70.5	72.1	73.6	75.2
LFCC-LCNN [27] (2019)	audio	48.5	50.1	47.4	50.3	44.3	47.8
RawNet [28] (2019)	audio	56.7	59.3	46.8	51.2	54.1	56.2
ECAPA-TDNN [4] (2020)	audio	70.2	72.6	59.8	62.7	67.3	69.8
AASIST [25] (2021)	audio	67.5	69.8	53.4	55.1	64.8	68.4
AVN-J [44] (2021)	audio-visual	84.4	86.2	73.2	77.6	81.1	83.3
MDS [10] (2020)	audio-visual	92.0	94.3	82.8	86.5	89.8	91.6
Emotions Don't Lie [9] (2020)	audio-visual	88.7	91.9	78.1	79.8	80.6	84.4
AVFakeNet [32] (2022)	audio-visual	91.8	93.7	78.4	83.4	82.8	86.2
VFD [33] (2022)	audio-visual	93.4	95.6	81.5	86.1	80.9	85.1
BA-TFD [34] (2022)	audio-visual	92.1	94.9	80.8	84.9	79.1	84.6
AVoiD-DF (Ours)	**audio-visual**	**95.3**	**97.6**	**83.7**	**89.2**	**91.4**	**94.8**

6. AVoiD - DF 与现有技术的总体比较 （左表）

7. MMD 与其他多模态融合方法的总体比较（右表）

8. 消融研究（下表）

Fusion Methods	Fusion Level	DefakeAVMiT		FakeAVCeleb	
		ACC	AUC	ACC	AUC
Ensemble	score-level	70.6	72.4	59.7	61.1
Contrastive	score-level	71.9	74.6	61.6	65.3
PC	pixel-level	66.0	68.1	56.4	58.9
MFB [67]	feature-level	72.1	75.8	65.3	67.8
AVN-F [44]	feature-level	72.5	73.2	63.0	66.3
AVN-E [44]	embedding-level	79.8	83.6	72.1	74.5
Ours	**embedding-level**	**95.3**	**97.6**	**83.7**	**89.2**

Ablation Model	DefakeAVMiT		FakeAVCeleb		DFDC	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)	ACC(%)	AUC(%)
visual only	75.6	77.5	70.3	72.4	74.2	75.8
audio only	62.5	64.3	55.8	57.2	57.4	60.9
w/o TSE	80.2	82.4	70.3	73.7	78.1	82.2
w/o MMD	76.8	79.2	68.5	70.2	74.6	77.8
w/o Classifier	85.8	87.6	74.1	79.4	82.5	86.0
w/o L_{cro}	94.6	96.9	83.1	88.5	90.2	94.1
w/o L_{con}	92.7	94.2	81.4	86.3	89.8	92.1
w/o L_{aam}	93.1	95.2	82.0	87.6	90.5	93.9
AVoiD-DF (Ours)	**95.3**	**97.6**	**83.7**	**89.2**	**91.4**	**95.8**