多注意力Deepfake检测

Hanqing Zhao1 Wenbo Zhou1, † 陈冬冬2
Tianyi Wei1 University of Science Weiming Zhang1, † Nenghai Yu1
and Technology of China1 {zhq2015@mail, 微软云 AI2
welbeckz@, bestwty@mail, zhangwm@, ynh@}.ustc.edu.cn
cddlyf@gmail.com

抽象的

Deepfake 的面部伪造在互联网上广泛传播,并引起了严重的社会关注。最近,
如何检测此类伪造内容已成为热门搜索话题,许多deepfake检测方法已经
被提议。他们中的大多数都对深度伪造检测进行建模
作为一个普通的二元分类问题,即首先使用一个
主干网络提取全局特征,然后馈送
它进入一个二元分类器（真/假）。但是由于这个任务中真假图像之间的差异通常是
微妙和局部,我们认为这种普通的解决方案不是最佳的。在本文中,我们将 deepfake 检测表述为细粒度分类问题,并提出了一个

新的多注意力深度伪造检测网络。具体来说,它由三个关键组件组成:1）多个空间注意力头,使网络关注不同的本地部分； 2)纹理特征增强块

放大浅层特征中的细微伪影； 3)聚合低级纹理特征和高级语义

由注意力图引导的特征。此外,为了解决
这个网络的学习难度,我们进一步介绍
新的区域独立性丧失和注意力引导
数据增强策略。通过广泛的实验
在不同的数据集上,我们证明了我们的优越性
普通二元分类器对应物的方法,以及
达到最先进的性能。模型将是
最近在https://github.com/yoctta/发布
多注意。

一、简介

受益于生成模型的巨大进步,deepfake 技术取得了重大成功

最近和各种面部伪造方法 [19, 41, 21, 31,
32, 44, 28, 38] 已被提出。像这样的技术

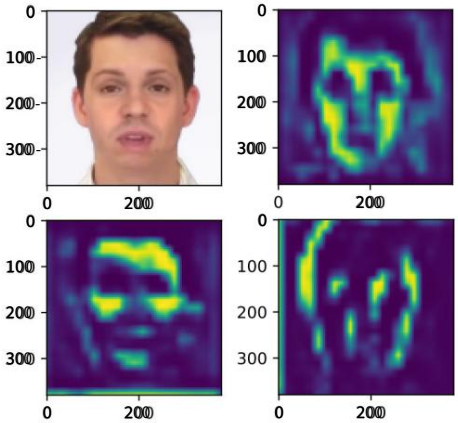

图 1:我们的方法获得的多个注意区域的示例。注意区域是分开的

并对不同的判别特征做出反应。

可以生成人眼无法分辨的高质量假视频,很容易被恶意用户滥用,造成严重的社会问题或政治问题

威胁。为了减轻这种风险,已经提出了许多深度伪造检测方法[27、34、22、33、26、45]。最多
其中将 deepfake 检测建模为一个普通的二进制分类问题（真/假）。基本上,他们经常首先使用
用于提取嫌疑人全局特征的骨干网络
图像,然后将它们输入二元分类器以区分真假。

然而,随着赝品变得越来越真实,真假之间的差异将变得更加微妙和局部,从而使得这种基于全局特征的香草解决方案效果不佳。但实际上,

这种微妙的地方财产与
细粒度分类问题。例如,在
细粒度的鸟类分类任务,有些物种看起来很
相似,并且只有一些小的区别
和局部差异,例如形状和颜色
喙。基于这一观察,我们建议将深度伪造检测建模为一种特殊的细粒度分类概率

†通讯作者。

lem 有两个类别。

受到基于零件模型的成功启发
细粒度分类领域,本文提出了一种新颖的
用于深度伪造检测的多注意力网络。首先,在
为了使网络关注不同的潜在工件区域,我们设计了多注意力头来预测

使用深度语义的多个空间注意力图
特征。其次,为了防止细微的差异出现在深层,我们增强了纹理特征

从浅层获得,然后聚合低级纹理特征和高级语义特征作为

每个局部部分的表示。最后,每个局部部分的特征表示将被独立池化

通过一个双线性注意力池层并融合为整个图像的表示。图1给出了一个例子

通过我们的方法获得的判别特征。

然而,训练这样一个多注意力网络并不是
一个微不足道的问题。这主要是因为,不像单一的注意力网络 [6],
它可以使用视频级标签
作为明确的指导并以有监督的方式接受培训,
多注意力结构只能以无监督或弱监督的方式进行训练。通过使用一个常见的

学习策略,我们发现多注意力头会降级为单注意力对应物,即只
有一个注意力区域会产生强烈的反应,而所有剩余的注意力区
域都会产生强烈的反应。
注意区域被抑制,无法捕获有用的信息。为了解决这个问题,我们进一步提出了
一种新的注意力引导数据增强机制。

具体来说,在训练的时候,我们会故意模糊一些
高响应注意力区域（软注意力下降）
并强制网络从其他注意力区域学习。
同时,我们引入了新的区域独立
损失以鼓励不同的注意力头关注不同的局部部分。

展示我们的多注意力机制的有效性
网络,我们对不同的
现有的数据集,包括 FaceForensics++[34]、Celeb DF[25] 和
DFDC[9]。这表明我们的方法是优越的
到香草二元分类器基线并实现最先进的性能。总之,

本文分为以下三部分:

- 我们将deepfake 检测重新定义为细粒度分类任务,为该领域带来了新的视角。

- 我们提出了一种新的多注意力网络架构,以从多个人脸注意力区域捕获局部判别特征。为了训练这个网络,我们

  还引入了区域独立性损失和设计
  一种注意力引导的数据增强机制
  协助网络训练进行对抗学习
  方法。

- 大量实验证明我们的方法
  优于普通二元分类基线
  并实现最先进的检测性能。

2. 相关作品

人脸伪造检测是计算机中的经典问题
视觉和图形。近期,深部进展迅速
生成模型使面部伪造技术“深入”
并且可以产生现实的结果,这提出了一个新的
deepfake 检测问题并带来重大挑战。大多数 deepfake 检测方法都能解决这个问题
然而,作为一种普通的二元分类,伪造人脸的细微和局部修改使其更类似于

细粒度的视觉分类问题。

2.1。深度伪造检测

由于面部伪造对社会安全造成巨大威胁,因此开发有效的面部伪造技术至关重要。
针对它的对策。许多作品 [46, 23, 4, 53, 34,
22, 33, 26, 45, 43] 已经被提出。早期作品 [46, 23]
通过视觉生物制品检测伪造品,例如,
不自然的眨眼或不一致的头部姿势。

随着基于学习的方法成为主流,
一些作品 [53, 34] 提出了从空间域中提取特征的框架,并在特定数据集上取得了出色的性能。最近,更

新兴方法已经考虑了数据域。
[45] 通过空间、隐写分析检测篡改的人脸
和时间特征。它添加了带有约束卷积层和 LSTM 的简化 Xception 流。

[26] 使用双分支表示提取器来组合
来自颜色域和频率的信息主要使用多尺度拉普拉斯高斯 (LoG) 算子。 [33]使用频率感知分解和局部

频率统计以在频率上暴露 deepfake 伪影
域并实现最先进的性能。

大多数现有方法将深度伪造检测视为
通用二元分类问题。他们专注于如何
构造复杂的特征提取器,然后进行二分法来区分真假面孔。然而,

逼真的仿冒品带来巨大挑战
到这个二元分类框架。在本文中,我们
将 deepfake 检测问题重新定义为细粒度
根据相似度的分类问题。

2.2.细粒度分类

细粒度分类 [50, 49, 13, 37, 12, 52, 47, 17,
10]是计算机视觉中一项具有挑战性的研究任务,
捕获局部判别特征以区分不同的细粒度类别。该领域的研究主要

专注于定位判别区域和学习
以弱监督的方式收集各种互补部分。以前的作品 [50, 49] 构建部分

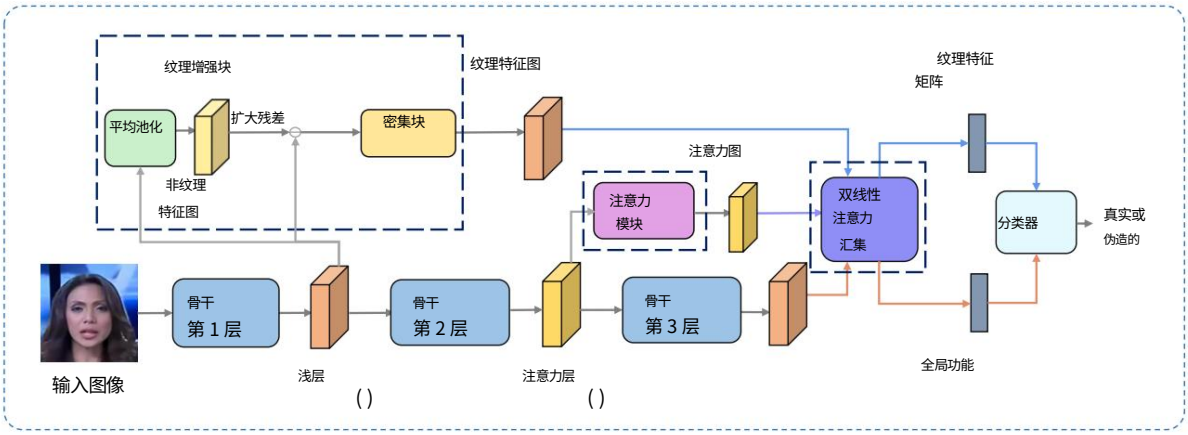

图 2:我们方法的框架。三个组件在我们的框架中发挥着重要作用： 用于生成的注意力模块
多个注意图,用于提取和增强纹理信息的纹理增强块和双向使用的
用于聚合纹理和语义特征的双线性注意力池。

模型来定位对象并平等地对待对象和语义部分。最近,几部作品 [52, 47, 10]
在多注意框架下提出，
这些方法的核心理想是同时学习多个尺度或图像部分的判别区域,并且
鼓励融合来自不同地区的这些特征。此外,[17] 设计了注意力裁剪和注意力下降以获得更平衡的注意力图。

在本文中,我们首次将深度伪造检测建模为一个特殊的细粒度分类问题。它共享
学习微妙和辨别特征的相同精神，
但只涉及真假两大类。

3. 方法

3.1。概述

在本节中,我们首先说明了
设计并简要概述我们的框架。
如前所述,真假之间的差异
面孔通常是微妙的并且出现在局部区域,这
不容易被单注意力结构捕捉到
网络。因此,我们认为分解注意力
进入多个区域可以更有效地收集
deepfake检测任务的局部特征。同时，
当前 deepfake 检测方法普遍采用的全局平均池化被替换为局部

我们框架中的注意力集中。这主要是因为不同的纹理图案差异很大

区域,从不同区域提取的特征可能
被全局池化操作平均,导致
可分辨性的丧失。另一方面,我们观察
由伪造方法造成的轻微伪影往往是
保留在浅层特征的纹理信息中。这里，
纹理信息代表浅层特征的高频分量,就像残差信息一样

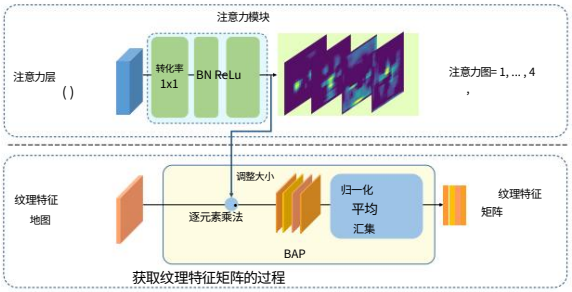

图 3:注意力模块的结构和过程
获得纹理特征矩阵P. 所提出的归一化
采用平均池化而不是全局平均池化。

RGB 图像的合成。因此,更浅的特征
应重点关注和加强
由当前最先进的检测方法考虑。
受这些观察的启发,我们提出了一个多注意力框架来解决深度伪造检测问题。
细粒度分类问题。在我们的框架中,三个
关键组件集成到骨干网中：
1)我们使用一个注意力模块来生成多个注意力图。 2)我们使用密集连接的卷积
层[18]作为纹理增强块,可以从浅层特征图中提取和增强纹理信息。
3) 我们将全局平均池化替换为

双线性注意力池（BAP）。我们使用BAP从浅层收集纹理特征矩阵

保留来自深层的语义特征。我们方法的框架如图 2 所示。

与基于单注意力结构的网络不同，
可以将视频级别的标签作为训练的明确指导,基于多注意力的网络只能被训练
以无监督或弱监督的方式,由于
缺乏区域级别的标签。它可能导致降级

多个注意力图集中在同一区域,而忽略其他可能提供判别信息的区域。为了解决这个问题,我们专门设计了一个区域独立损失,旨在确保每个注意力图都集中在一个特定区域上而不会重叠,并且所关注的区域在不同样本之间是一致的。此外,我们采用注意力引导数据增强（AGDA)机制来降低最具辨别力的特征的显着性,并迫使其他注意力图挖掘更多有用的信息。

3.2.多注意力框架

将网络的输入人脸图像记为 I,我们框架的主干网络记为 f,从第 t 层的中间阶段提取的特征图记为ft(I),大小为Ct × Ht × Wt。这里， Ct是通道数， Ht，Wt分别是特征图的高度和宽度。

多注意力图生成。如上所述,给定一张真/假人脸图像 I 作为输入,我们的框架首先使用一个注意块为 I 生成多个注意图。如图 3 所示,注意块是一个轻量级模型,由一个1×1 卷积层、批量归一化层和非线性激活层 ReLU。从特定层 SLa 中提取的特征图将被馈送到该注意力块中,得到 M 个大小为Ht × Wt 的注意力图 A,其中Ak ∈ R{Ht×Wt}表示第 k 个注意力图,对应一个特定的判别区域,例如眼睛、嘴巴甚至是[22]中定义的混合边界。 SLa 的确定将在第 4 节中讨论。

纹理特征增强。大多数深度伪造检测的二元分类框架都没有关注一个重要现象,即伪造方法造成的伪影通常在浅层特征图的纹理信息中很突出。这里的纹理信息代表了浅层特征的高频分量。因此,为了保留更多的纹理信息来捕获这些伪影,我们设计了一个纹理特征增强块,如图 3 所示。我们首先在补丁中应用局部平均池化来对来自特定层 SLt 的特征图进行下采样并获得池化特征图 D. 如何选择 SLt将在下面的实验部分讨论。然后类似于空间图像的纹理表示,我们在特征级别定义残差来表示纹理信息,如下所示：

$$TSLt = fSLt (I) - D \tag{1}$$

这里 T 包含fSLt (I) 的大部分纹理信息。
然后我们使用具有 3 层的密集连接卷积块来增强 T,输出记为 F ∈ RCF ×Hs×Ws ,定义为“文本特征图”。

,

双线性注意力池。在得到注意力图 A 和纹理特征图 F 后,我们使用双线性注意力池 (BAP) 来获得特征图。我们将 BAP 双向用于浅层特征图和深层特征图。如图 3 所示,为了提取浅层纹理特征,如果不匹配,我们首先使用双线性插值将注意力图调整为与特征图相同的比例。然后,我们分别将每个注意力图Ak逐元素乘以纹理特征图F ,得到部分纹理特征图Fk。

在这一步结束时,部分纹理特征图Fk应在全局池化后输入分类器。然而,考虑到不同区域范围之间的差异,如果使用传统的全局平均池化,池化的特征向量会受到注意力图强度的影响,这违背了关注纹理信息的目的。为了解决这个问题,我们设计了一个标准化的平均池：

$$vk = \frac{\sum_{m=0}^{Hs-1} \sum_{n=0}^{Ws-1} Fk,m,n}{\left\| \sum_{m=0}^{Hs-1} \sum_{n=0}^{Ws-1} Fk,m,n \right\|_2} \tag{2}$$

然后将归一化的注意力特征vk ∈ R1×N堆叠在一起以获得纹理特征矩阵 P ∈
RM×CF , 它将被输入分类器。

对于深度特征,我们首先拼接每个注意力图以获得单通道注意力图Asum。然后我们使用 BAP for Asum和网络最后一层的特征图得到全局深度特征 G，它也将被输入到分类器中。

3.3.注意力图正则化的区域独立性损失

如前所述,由于缺乏细粒度的级别标签,训练多注意力网络很容易陷入网络退化的情况。具体来说,不同的注意力图倾向于集中在同一个区域,如图 4 所示，这不利于网络捕获给定输入的丰富信息。此外,对于不同的输入图像,我们希望每个注意力图都位于固定的语义区域,例如,注意力图A1关注不同图像中的眼睛， A2关注嘴巴。因此,每个注意力图捕获信息的随机性会降低。

为了实现这些目标,我们提出了一个区域独立损失,它有助于减少注意力图之间的重叠并保持不同输入的一致性。我们将 BAP 应用于第 3.2 节中的池化特征图 D 以获得“语义特征向量”:V ∈ RM×N ,并且通过修改 [15] 中的中心损失,区域独立损失定义如下：

,

$$LRIL = \sum_{i=1}^{乙} \sum_{j=1}^{米} 最大值 \left(V_j^{它-c_j} \quad {}_2^2 - \min(y_i), 0\right) + \sum_{i,j\in(M,M),i6=j} mout - c - c_i \quad tt \max_j \left(\frac{2}{2}, 0\right) \tag{3}$$

其中B是batch size,M是attention的数量， min表示特征与对应特征中心之间的边距,当yi为0和1时设置为不同的值。mout是每个特征中心之间的边距。 c ∈ RM×N是 V 的特征中心,定义如下,并在每次迭代中更新：

,

$$_{tc\_} = c^{t-1} - \alpha c^{t-1} - \frac{1}{乙} \sum_{我=1}^{乙} 在_{我!} \tag{4}$$

这里 α 是特征中心的更新率,我们在每个训练 epoch 后衰减 α。LRIL的第一部分是将 V 拉近特征中心 c 的类内损失,第二部分是排斥分散的特征中心的类间损失。我们通过计算每个批次中 V 的梯度来优化 c。考虑到假人脸的纹理模式应该比真人脸更多样化,因为假人脸是通过多种方法生成的,因此我们将假人脸的部分特征限制在真人脸的特征中心但有更大的余量。通过这种方式,我们在类内给予更大的余量来搜索假脸中的有用信息。

对于我们框架的目标函数,我们将此区域独立性损失与传统的交叉熵损失相结合：

$$L = \lambda 1 * LCE + \lambda 2 * LRIL \tag{5}$$

LCE是交叉熵损失， λ1和λ2是这两项的平衡权重。默认情况下,我们在实验中设置λ1 = λ2 = 1。

3.4.注意力引导的数据增强

在区域独立损失的约束下,我们减少了不同注意力区域的重叠。然而,尽管可以很好地分离不同的注意力区域,但注意力图仍可能对相同的判别特征做出反应。例如,在图 5 中,注意力区域没有重叠,但它们都对输入人脸的地标有强烈反应。为了迫使不同的注意力图关注不同的信息,我们提出了注意力引导数据增强（AGDA)机制。

对于每个训练样本,随机选择一个注意力图Ak来指导数据增强过程,∈ RH×W 。并将其归一化为Augmentation Map A然后我们使用高斯模糊来生成退化的图像。

SLt候选人	SLa ACC 候选人(%)	
L2	L4	96.38
L2	L5	97.26
L3	L4	96.14
L3	L5	96.81

表 1:我们的方法基于 SLt和SLa 的不同组合的性能。

最后,我们使用A*作为原始图像和降级图像的权重：

$$_{我}^{0} = I \times (1 - A^{*}) + 身份证 \times A^{*} \tag{6}$$

注意力引导的数据增强有助于在两个方面训练模型。首先,它可以为某些区域添加模糊,从而确保模型从其他区域学习更强大的特征。或者,AGDA 可以偶然擦除最显着的区分区域,这迫使不同的注意力图将它们的反应集中在不同的目标上。此外,AGDA 机制可以防止单个注意力区域过度扩展,并鼓励注意力块探索各种注意力区域划分形式。

4. 实验

在本节中,我们首先探索我们提出的多注意力框架的最佳设置,然后展示广泛的实验结果来证明我们方法的有效性。

4.1。实施细节

对于所有真/假视频帧,我们使用最先进的人脸提取器 RetinaFace[8] 来检测人脸并将对齐的人脸图像保存为大小为 380 × 380 的输入。

我们在等式 4 中设置超参数 α = 0.05,并在每个 epoch 后衰减 0.9。等式 3 中的类间边距mout设置为 0.2。对于真图像和假图像,类内边距最小值分别设置为 0.05 和 0.1。

我们通过实验选择注意力图 M、SLa和 SLt的数量。在 AGDA 中,我们设置调整大小因子 0.3 和高斯模糊 σ = 7。我们的模型使用 Adam 优化器 [20] 进行训练,学习率为 0.001,权重衰减为 1e 6。我们在 4 个批量大小为 48 的 RTX 2080Ti GPU 上训练我们的模型。

4.2. SLa和SLt的测定

在本文中,我们采用 EfficientNet-b4[39] 作为我们的多注意力框架的骨干网络。

EfficientNet-b4 能够达到与 XceptionNet [3] 相当的性能,而只有一半的 FLOPs。主要有7个

方法	量产		总部	
	ACC AUC	ACC AUC		
Steg.Features[11]	55.98	-	70.97	
LD-CNN[5]	58.69	-	78.45	-
中观网[1]	70.47	-	83.10	-
面部X光[22]	-	61.60	-	87.40
异常[3]	86.86 89.30	95.73 96.30		
Xception-ELA[14]	79.63 82.90	93.86 94.80		
Xception-PAFilters[2]	87.16 90.20		-	-
F^3-Net[33]	90.43 93.30	97.52 98.10		
两支[26]	-	86.59 98.70	-	
EfficientNet-B4[39]	86.67 88.20	96.63 99.18		
我们的（异常）	86.95 87.26	96.37 98.97		
我们的（高效-B4）	88.69 90.40	97.60 99.29		

表 2:FaceForensics++ 的定量比较
分别具有高质量和低质量设置的数据集。最好的表现被标记为粗体。

EfficientNet 的总层数,由 L1- 表示
L7,分别。
如上所述,我们观察到细微的伪影
倾向于被浅层的纹理特征保留
网络,因此我们选择 L2 和 L3 作为候选
SLt 。相反,我们希望注意力图关注
不同区域的输入,这需要指导
某种程度的高级语义信息。所以,
我们使用更深的阶段 L4 和 L5 作为SLa 的候选。
默认设置 M = 1,我们在 FF++(HQ) 上训练具有四种组合的模型。从表 1 的结果中,我们发现
模型在使用 L2 时达到最佳性能
SLt和 L5 用于SLa。

4.3.与以往方法的比较

在本节中,我们将我们的框架与当前的框架进行比较
最先进的深度伪造检测方法。我们评估
分别在 FF++ [34] 和 DFDC [9] 上的性能。
我们进一步评估了跨数据集的性能
第 4 节中的 Celeb-DF [25]。我们采用 ACC（准确度)和
AUC（接收器操作特性曲线下的面积）
作为广泛实验的评估指标。

4.3.1 FaceForensics++的评估

FaceForensics++[34] 是目前使用最广泛的数据集
许多 deepfake 检测方法,它包含 1000
来自互联网的原始真实视频和每个真实视频
对应4个伪造的,被操纵的
由 Deepfakes、NeuralTextures[40]、FaceSwap[48] 和
Face2Face[41],分别。在训练过程中,我们
将原始帧增加 4 次以获得真实/虚假标签平衡。我们采用 EfficientNet-B4 作为我们的主干

方法	对数损失
塞利姆·塞费尔别科夫[35]	0.1983
西马[51]	0.1787
NTechLab[7]	0.1703
十八岁[36]	0.1882
医生[16]	0.2157
我们的	0.1679

表 3:与 DFDC 获胜团队方法的比较
在 DFDC 测试数据集上。我们作为WM队参加了比赛。

框架,并在 HQ (c23) 版本上测试性能
和 LQ (c40) 版本,分别。特别是,当在 LQ 上训练我们的模型时,参数由那些

在总部进行预训练以加速收敛。比较结果列于表 2。

表 2 中的结果表明我们的方法
在 HQ 版本上实现了最先进的性能
FF++。并且不同主干的性能验证了我们的框架不受主干的限制

网络。然而,在 LQ 版本上,与F3 -Net [33] 相比,性能下降了 1.5%,因为F3
-Net 是
一种专门设计的高压缩深度假视频检测方法。这主要是因为视频

在 FF++(LQ) 中被高度压缩并导致显着
纹理信息的丢失,这对我们的纹理增强设计来说是一场灾难。结果还揭示了我们框架的局限性,即我们的框架是敏感的

高压缩率模糊了空间域中大部分有用的信息。我们将制作我们的框架

将来对压缩更健壮。

4.3.2 评估DFDC数据集

DeepFake Detection Challenge (DFDC) 是最近发布的最大规模的
deepfake 检测数据集,
数据集在 Deepfake 检测挑战赛上公开,或由 Facebook 于 2020 年组织。由于该数据集中假视频的出色伪造质量,它是目前深度伪造检测任务中最具挑战性的数据集。很少

以前的方法已经在这个数据集上进行过,因此
我们在这个数据集的训练集上训练我们的模型,并且
仅将 logloss 分数与获胜团队的比较
DFDC竞赛的方法。这里提供的logloss
分数是在 DFDC 测试集上计算的（参考表
2 of [9]）,它是 DFDC 私有集的一部分。更小
logloss 代表更好的性能。表 3 中的结果表明,我们的框架在 DFDC 数据集上实现了最先进的性能。

方法	FF++	名人-DF
双流[53]	70.10	53.80
中纪4[1]	84.70	54.80
感悟4[1]	83.00	53.60
时间[24]	80.10	56.90
Xception-原始[25]	99.70	48.20
Xception-c23[25]	99.70	65.30
Xception-c40[25]	95.50	65.50
多任务[29]	76.30	54.30
胶囊[30]	96.60	57.50
DSP-FWA[24]	93.00	64.60
两支[26]	93.18	73.41
F^3-网络[33]	98.10	65.17
EfficientNet-B4[39]	99.70	64.29
我们的	99.80	67.44

表 4:Celeb-DF 的跨数据集评估（AUC(%)）
通过对 FF++ 的培训。其他一些方法的结果是
直接引用自[26]。我们的方法优于大多数
deepfake检测方法。

米	FF++(总部)	名人-DF
1	97.26	67.30
2	97.51	65.74
3	97.35	66.86
4	97.60	67.44
5	97.39	66.82

表 5:不同数量的注意力图在 FF++(HQ) (Acc %) 和 Celeb DF (AUC %) 上的消融结果。

4.3.3 Celeb-DF的跨数据集评价

在这一部分中,我们评估了我们的框架的可迁移性,该框架在 FF++(HQ) 上进行了多次伪造训练
方法,但在 Celeb-DF [25] 上进行了测试。我们采样 30 帧
为每个视频计算帧级 AUC 分数。
结果如表 4 所示。我们的方法显示出比大多数现有方法更好的可迁移性。两支

[26] 在传输能力方面实现了最先进的性能,但是,它的数据集内 AUC 远远落后于我们。

4.4.消融研究

4.4.1 多重注意力的有效性

为了确认使用多重注意力的有效性,
我们评估注意力图的数量如何影响
我们模型的准确性和可转移性。我们在我们的框架中用不同的注意力量 M 训练模型
在 FF++(HQ) 上,其他超参数保持与
表 2 中的设置。对于单注意力模型,我们做
不使用区域独立性损失和 AGDA。
FF++(HQ) 上的 Acc 结果和 AUC 上的结果

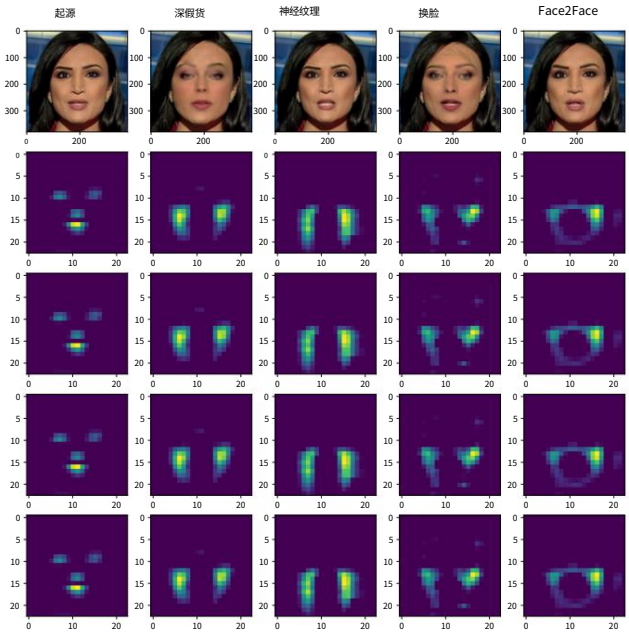

图 4:在没有区域独立损失 (RIL) 和 AGDA 的情况下训练的注意力图。如果没有 RIL 和 AGDA,则
网络很容易降级和多注意力图
定位相同的输入区域。

Celeb-DF 在表 5 中报告。在某些情况下,
基于多注意的模型比单注意模型表现更好,我们发现 M = 4 提供

最好的表现。

4.4.2 区域独立性损失消融研究和AGDA

如上所述,区域独立性丧失和
AGDA 在正则化多注意力图训练中发挥着重要作用。在这一部分中,我们进行了定量实验并给出了一些可视化来证明

这两个组件是必要的。
首先,为了证明我们的区域依赖损失的有效性,我们比较了使用不同辅助损失训练的模型的性能。我们保留所有

除了损失函数外,设置与之前相同。
出于设计辅助损失的相同动机,我们
用 Additive An gular Margin softmax(AMS)[42] 代替区域独立性损失,它也可以强制特征
靠近中心的向量。
然后我们验证我们的设计的有效性
阿格达。如前所述,我们模糊原始图像以降低所选输入区域的质量。因此策略

AGDA 可以看作是一种“软注意力下降”。在
这部分,我们或者采用“hard attention drop”,
它通过二进制直接擦除所选区域的像素

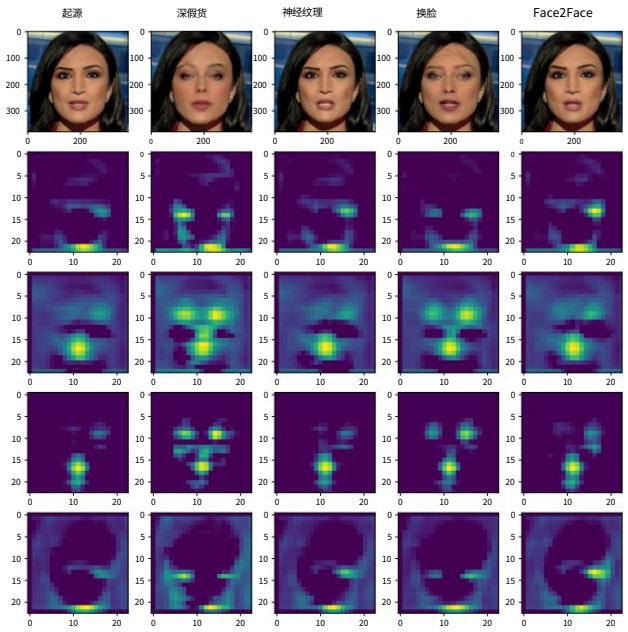

图 5:在没有 AGDA 的情况下训练的注意力图。虽然
地区独立性丧失迫使不同的关注
地图分开,他们倾向于对同一个突出点做出反应
没有 AGDA 帮助的功能。

损失型 AGDA 型	FF++(HQ)	Celeb-DF	
没有任何	没有任何	96.74	64.86
AMS	没有任何	96.49	64.23
RIL	没有任何	97.38	65.85
AMS	难的	96.53	63.73
RIL	难的	97.24	64.40
AMS	柔软的	96.78	66.42
RIL	柔软的	97.60	67.44

表 6:不同损失函数的消融结果和
AGDA 战略。该模型达到最佳性能
使用区域独立性损失和软 AGDA 时
机制。 FF++(HQ) 数据集的度量是 ACC,并且
在 Celeb-DF 上是 AUC。

张力面罩 BM:

$$BMk(i, j) = \begin{cases} 0, & \text{如果} A^{1}(i, j) > \theta d \\ 1, & \text{否则。} \end{cases} \tag{7}$$

在这个实验中,我们设置了注意力下降阈值θd = 0.5。该消融研究的比较结果是
如表 6 所示。结果验证了区域依赖损失（RIL)和注意力引导数据增强（软注意力下降)都有显着的贡献

以提高我们框架的性能。
进一步帮助了解区域功能

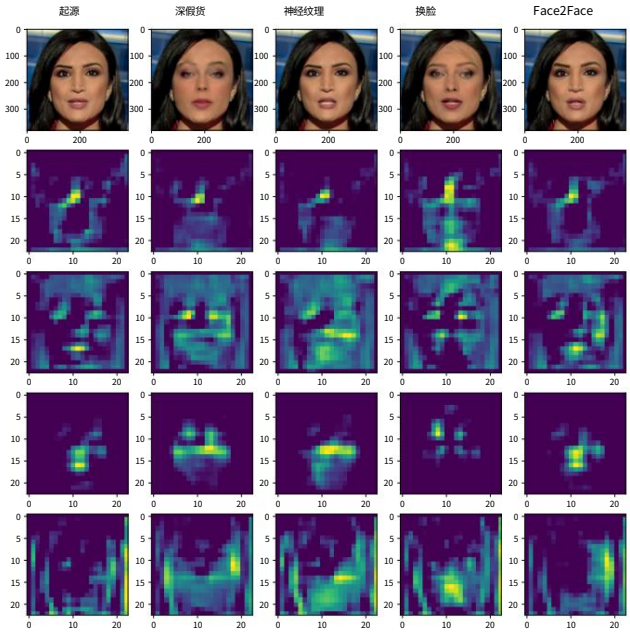

图 6:使用区域独立损失和 AGDA 训练的注意力图。注意图的位置和响应是正确分布的。

独立性损失和 AGDA 策略,我们可视化
有/没有这两个训练的模型的注意力图
成分。图 4 说明了没有注意图
RIL,它显示了一个明显的趋势,即所有注意力图都集中在同一区域上。图 5 表明,尽管
注意区域在重新训练下被分离
RIL,不同地区仍然表现出相似的反应
最显着的特征,例如地标。这不利于多个注意力图捕获来自不同区域的不同信息。虽然图 6 验证了

当同时采用 RIL 和软 AGDA 时,注意
地图显示了不同地区的反应
语义表示。

5. 结论

在本文中,我们研究了来自
一种新颖的观点,将深度伪造检测任务制定为细粒度分类问题。我们提出了一个多注意力的深度伪造检测框架。这

提议的框架探索了有区别的地方区域
通过多个注意力图,并增强纹理特征
从浅层捕捉更微妙的伪影。然后
低级纹理特征和高级语义特征由注意力图引导聚合。区域独立性损失函数和注意力引导数据

引入了增强机制来帮助训练解开的多重注意力。我们的方法效果很好

广泛指标的改进。

参考

[10] [redacted]

[20] [redacted]9/

https://github.com通讯 /

[51] Hanqing Zhao, Hao Cui, and Wenbo Zhou. https://
github.com/cuihaoleo/kaggle-dfdc。