Revisiting the D-vlog Dataset: an Input Attribution Approach

The final project of *AI in MH 2023*

Yifan Huang, Zhewen Yao

December 27, 2023

School of Computer Science, Peking University

Table of contents

Introduction

The D-vlog Dataset [Yoon et al., 2022]

D-vlog: YouTube videos (posted between 1st January 2020 and 31st January 2021) are collected, by searching the following keywords:

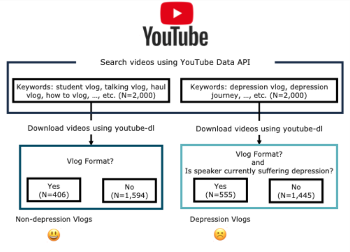
- depression: 'depression daily vlog', 'depression diary', . . .
- non-depression: 'daily vlog', . . .

4 college students are recruited to label the dataset.

D-vlog: Multimodal Vlog Dataset for Depression Detection

Jeewoo Yoon[1,3], **Chaewon Kang**[1], **Seungbae Kim**[2], **Jinyoung Han**[1,3,*]

[1] Department of Applied Artificial Intelligence, Sungkyunkwan University, Seoul, Korea
[2] Department of Computer Science, University of California, Los Angeles, USA
[3] RAONDATA, Seoul, Korea
{yoonjeewoo, codnjs3}@g.skku.edu, sbkim@cs.ucla.edu, jinyounghan@skku.edu

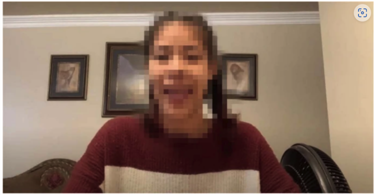


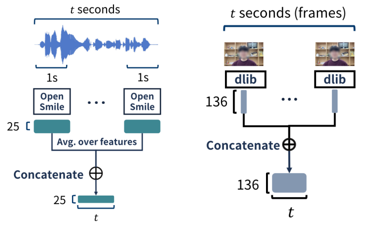
	Gender	# Samples	Avg. Duration
Depression	Male	182	583.74s
	Female	373	667.63s
Non-depression	Male	140	438.77s
	Female	266	587.76s

Acoustic and visual features are extracted from the videos and provided in the public dataset.

- audio: 25 low-level acoustic features
- visual: 68 facial landmarks (x_i, y_i), 136 features in total
- all these features are non-verbal; no semantic information is included

Sampling rate: 1 segment or frame per second.





(a) Acoustic features (b) Visual features

Motivations

Depression Detection Task: since the diagnosis of major depression disorder requires a high level of expertise, it's meaningful to develop an automatic method for efficient screening of depression.

We choose **D-vlog** dataset for the following reasons:

- real-life scenes
- good generalizability
- convenient to use

Train	Test	Precision	Recall	F1-Score
DW	DV	60.14	60.38	60.24
DV	DV	65.40	65.57	63.50
DW	DW	62.57	52.63	55.45
DV	DW	69.45	55.26	57.73

Table 7: Cross-corpus validation results between D-Vlog and DAIC-WOZ datasets. DV and DW denote D-Vlog and DAIC-WOZ, respectively.

Depression Detection on D-vlog

Task Formulation

We have:

- two modalities $\phi = \{\mathcal{A}, \mathcal{V}\}$
- a set of N labeled depression or non-depression samples
- each sample x_i is a sequence of T_i frames (not equally long)
- each frame $x_i[t]$ is a feature vector of dimension $M = M_{\mathcal{V}} + M_{\mathcal{A}} = 136 + 25 = 161$

Binary classification setting: train a binary classifier f, which can predict the label (depressed, not depressed) of a sample given its feature vector.

Model 1. TMeanNet

For a given input sample $x_i \in \mathbb{R}^{T_i \times M}$,

1. Temporal Average Pooling: average over the temporal dimension to obtain a feature vector $z_i \in \mathbb{R}^M$;
2. Binary Classification: use a 3-layer MLP to predict the label.

Model 2. Depression Detector [Yoon et al., 2022]

Key point:

- multi-modal fusion through cross-attention.

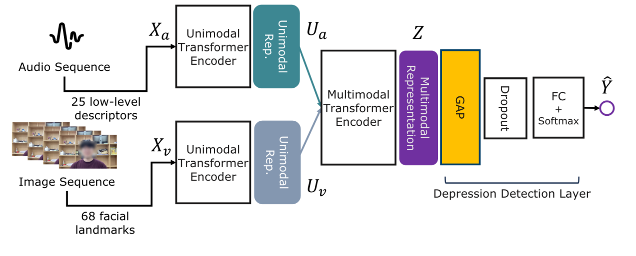
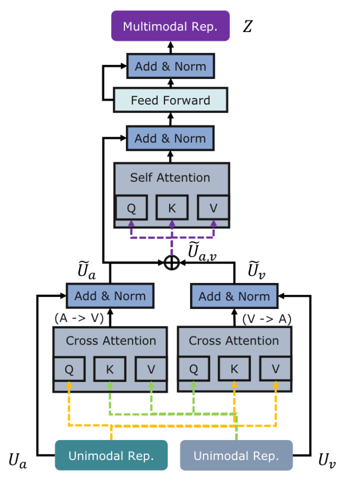


Figure 3: An illustration of the proposed model, *Depression Detector*.



Model 3. TAMFN [Zhou et al., 2023]

Time-aware Attention-based Multimodal Fusion Network

Key points:

- temporal convolutional network with global information: extract acoustic and visual features
- inter-modal feature extraction: integrates early acoustic and visual interaction features
- time-aware attention multimodel fusion: fuse multiple features through time-aware attention

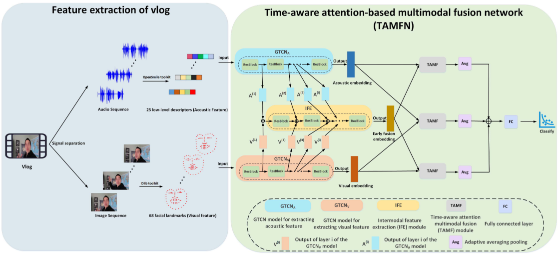


Fig. 2. The detailed structure of TAMFN.

Our Results

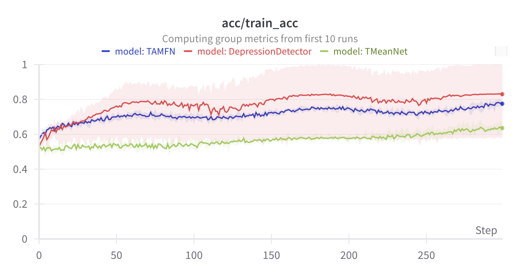
	Precision	Recall	F1	Accuracy
TMeanNet	0.6352	0.7902	0.7035	0.6151
DepressionDetector	**0.6900**	0.7512	0.7172	0.6575
TAMFN	0.6889	**0.7935**	**0.7360**	**0.6708**

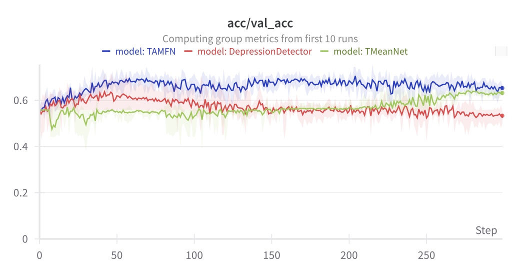
Table 1: Test performance metrics for TMeanNet, DepressionDetector, and TAMFN on D-vlog. Average of 5 runs.

Here, performance of DepressionDetector and TAMFN are much better than those reported in the original papers. (DepressionDetector: F1=0.635; TAMFN: F1=0.6582)

- We use cos annealing learning rate scheduler?
- We test the performance using the model with the best validation accuracy?

Training Curves





Input Attribution

Question

[Sun et al., 2024]: F1=0.9505, Accuracy=0.9391

- use only acoustic features!
- based on GNN

Table 1
Depression detection results of the method proposed in this paper and other comparative methods on the data set DAIC-WOZ, MODMA, and D-Vlog, where the best results are shown in bold.

Dataset	Method	Acc (%)	Pre (%)	Rec (%)	F1 (%)
DAIC-WOZ	SVM [17]	58.93 ± 2.69	59.86 ± 2.73	57.42 ± 4.18	57.42 ± 4.18
	Depaudionnet [38]	72.13 ± 2.35	70.97 ± 3.27	75.72 ± 2.64	73.26 ± 2.95
	Yoon et al. [36]	–	62.57	52.63	55.45
	MSCDR [39]	77.1	–	–	66.0
	DEPA [40]	–	91.0	89.0	90.0
	Ghadiri et al. [32]	61.0	61.1	66.7	63.4
	HCAG [23]	–	77.0	83.0	80.0
	MS2-GNN [24]	89.13	80.0	85.71	82.76
	Ours	**92.21** ± **1.86**	**92.36** ± **2.53**	**92.18** ± **1.55**	**92.23** ± **2.01**
MODMA	Chen et al. [41]	83.4	83.5	76.8	80.0
	MSCDR [39]	85.7	–	–	84.0
	MS2-GNN [24]	86.49	82.35	87.5	84.85
	Ours	**90.35** ± **2.46**	**88.25** ± **3.26**	**90.33** ± **3.67**	**89.15** ± **2.89**
D-Vlog	Yoon et al. [36]	–	65.4	65.57	63.5
	TAMFN [42]	–	66.02	66.5	65.82
	CAIINET [43]	–	66.57	66.98	66.56
	Ours	**93.91** ± **1.43**	**91.9** ± **1.92**	**98.48** ± **1.34**	**95.05** ± **1.19**

Question: what are the most important features for depression detection?

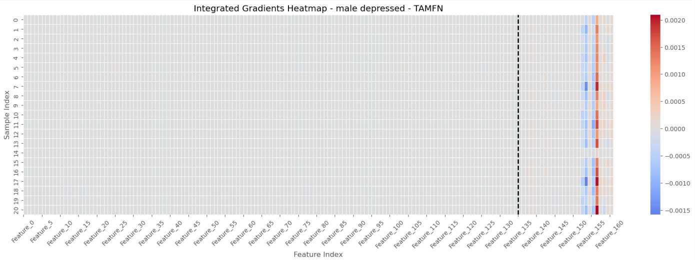
Integrated Gradients

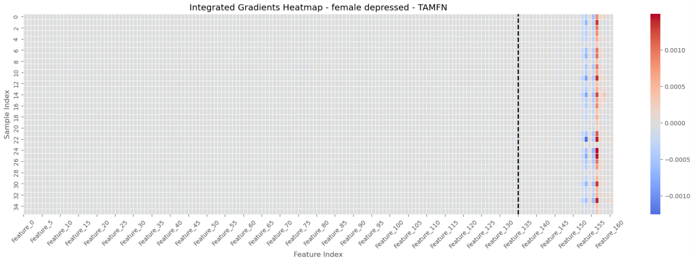
Input Attribution: given a trained model f, a sample x, we want to know how much x_i contributes to the prediction $f(x)$.

Integrated Gradients: a method for computing input attributions for any differentiable model [Sundararajan et al., 2017].

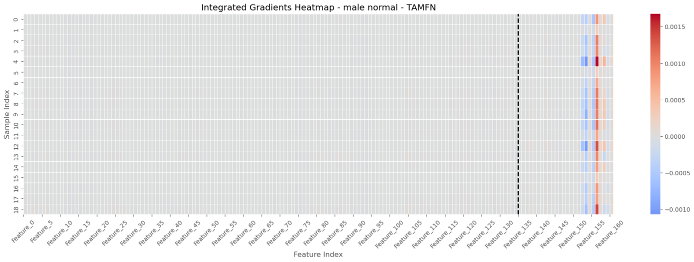
- Consider the straightline path (in Rn) from the baseline x_0 to the input x, and compute the gradients at all points along the path.
- Integrated gradients are obtained by cumulating these gradients.

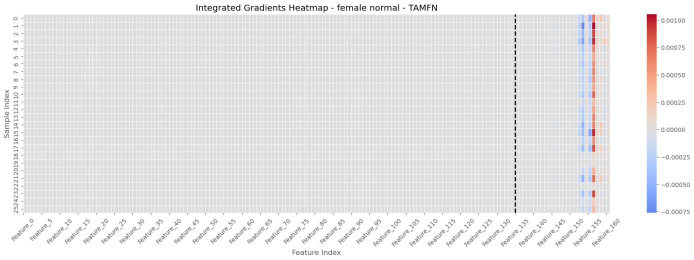
Depression Samples



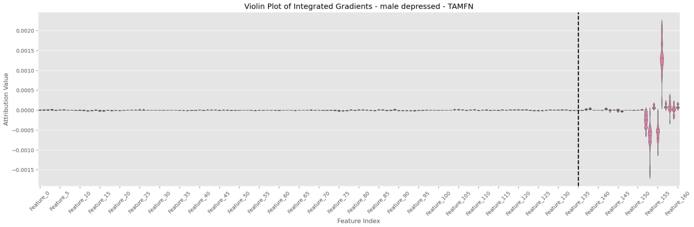


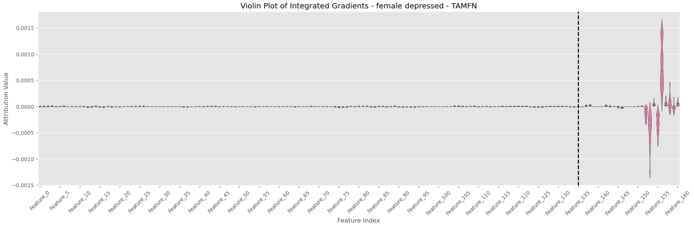
Normal Samples



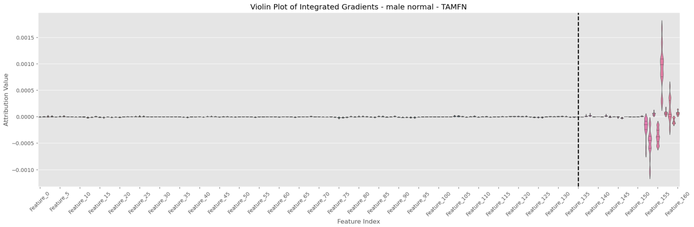


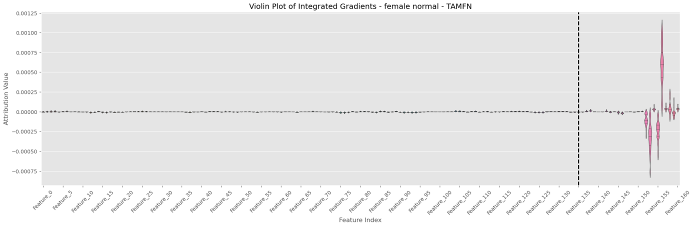
Depression Samples





Normal Samples





Discussion

Our Findings

The visual features in the D-vlog dataset seem to be less important than the acoustic features.

They can even be discarded without much performance loss!

We call for improvements on the dataset!

Thanks for listening!

Any questions?

Sun, C., Jiang, M., Gao, L., Xin, Y., and Dong, Y. (2024).
A novel study for depression detecting using audio signals based on graph neural network.
Biomedical Signal Processing and Control, 88:105675.

Sundararajan, M., Taly, A., and Yan, Q. (2017).
Axiomatic Attribution for Deep Networks.

Yoon, J., Kang, C., Kim, S., and Han, J. (2022).
D-vlog: Multimodal Vlog Dataset for Depression Detection.
Proceedings of the AAAI Conference on Artificial Intelligence, 36(11):12226–12234.

Zhou, L., Liu, Z., Shangguan, Z., Yuan, X., Li, Y., and Hu, B. (2023).
TAMFN: Time-Aware Attention Multimodal Fusion Network for Depression Detection.
IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31:669–679.