ViT-DD: Multi-Task Vision Transformer for Semi-Supervised Driver Distraction Detection

Yunsheng Ma and Ziran Wang (Purdue University College of Engineering)

NEURAL INFORMATION PROCESSING SYSTEMS

One-sentence Summary

We propose a multi-modal Vision Transformer (termed ViT-DD) for driver distraction detection, as well as a pseudo-labeled multi-task training algorithm to train ViT-DD.

Introduction

- **What is distracted driving?**

Distracted driving is defined by NHTSA as any activity that diverts attention away from safe driving.

- **What did existing works do?**

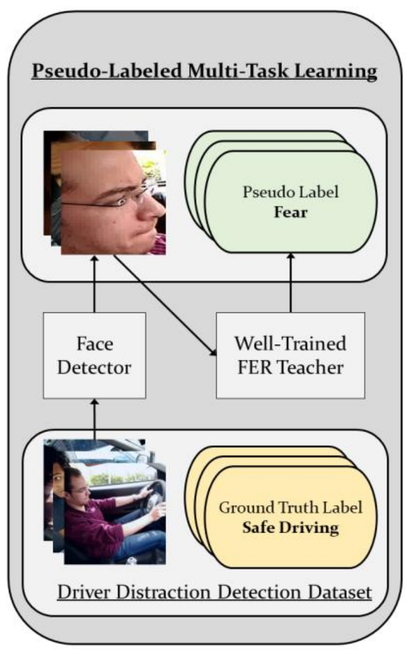
Driving automation systems, which can assist the driver in navigating the vehicle when he or she is distracted, have experienced rapid development over the past decade.

- **What do we do?**

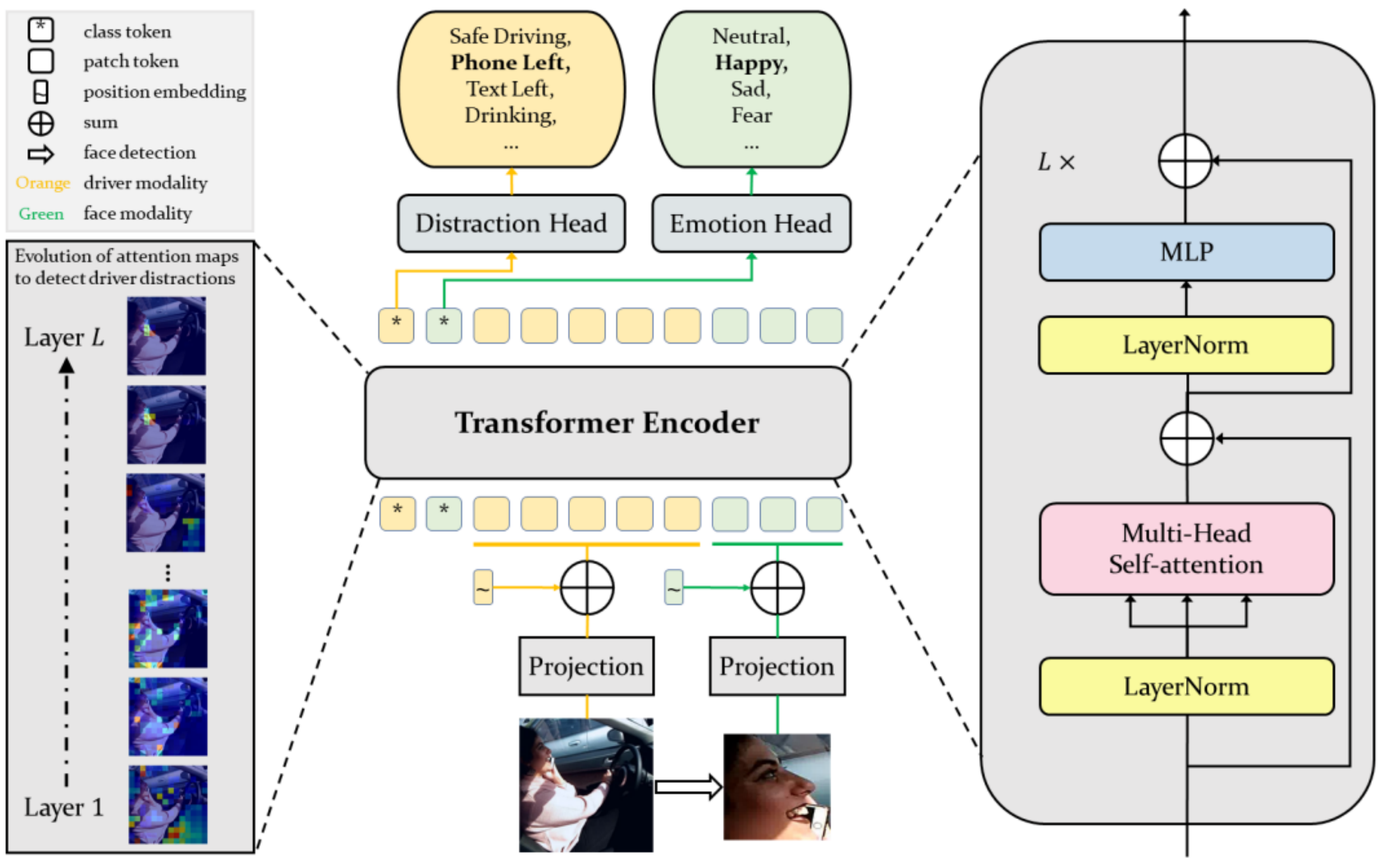
We propose a multi-modal Vision Transformer (termed ViT-DD) to exploit inductive information contained in the training signals of both emotion recognition and distraction detection.

Pseudo-Labeled Multi-Task Training

- A teacher ViT is trained on a large facial expression recognition dataset
- A face detector is used to crop face images from an in-cabin camera
- The FER teacher model is used to label the unlabeled face images
- The driver dataset now contains both ground-truth labels for distraction detection and pseudo labels for emotion recognition



Overall Architecture of ViT-DD



$$
\begin{aligned}
\bar{\mathbf{x}}^{(i)} &= [\mathbf{t}_1^{(i)} E^{(i)}; \cdots; \mathbf{t}_{N_i}^{(i)} E^{(i)}] + E_{\text{pos}}^{(i)}, && i = 0, 1 \\
z^0 &= [\mathbf{t}_{\text{class}}^{(0)}; \mathbf{t}_{\text{class}}^{(1)}; \bar{\mathbf{x}}^{(0)}, \bar{\mathbf{x}}^{(1)}] \\
\mathbf{z}'_\ell &= \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}, && \ell = 1 \ldots L \\
\mathbf{z}_\ell &= \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell, && \ell = 1 \ldots L \\
\mathbf{y}^{(i)} &= \text{LN}(\mathbf{z}_L^i), && i = 0, 1
\end{aligned}
$$

- A face detector is applied to the input signal from an in-cabin camera
- The driver and face images are divided into patches and independently embedded
- The embeddings are added with their respective position embeddings
- The resulting sequence is concatenated
- Tokens representing distractions and emotions are prepended
- The sequence of class and visual tokens is fed into the Transformer encoder
- The class tokens from the final sequence are used for classification

Experiments

Comparison with State-of-the-Art

Experiment	Method	Accuracy (↑)	NLL (↓)
AUCDD	GA-Weighted Ensemble*[5]	0.9006	0.6400
	ADNet*[27]	0.9022	–
	C-SLSTM*[28]	0.9270	0.2793
	ViT-DD (ours)	**0.9359**	**0.2399**
SFDDD Split-by-Driver	DD-RCNN*[29]	0.8600	**0.3900**
	ViT-DD (ours)	**0.9251**	0.3972
SFDDD Split-by-Image	ViTConv* [30]	0.9790	0.0800
	Inception+ResNet+HRNN* [31]	0.9930	–
	LWANet* [32]	0.9937	0.0260
	ViT-DD (ours)	**0.9963**	**0.0171**

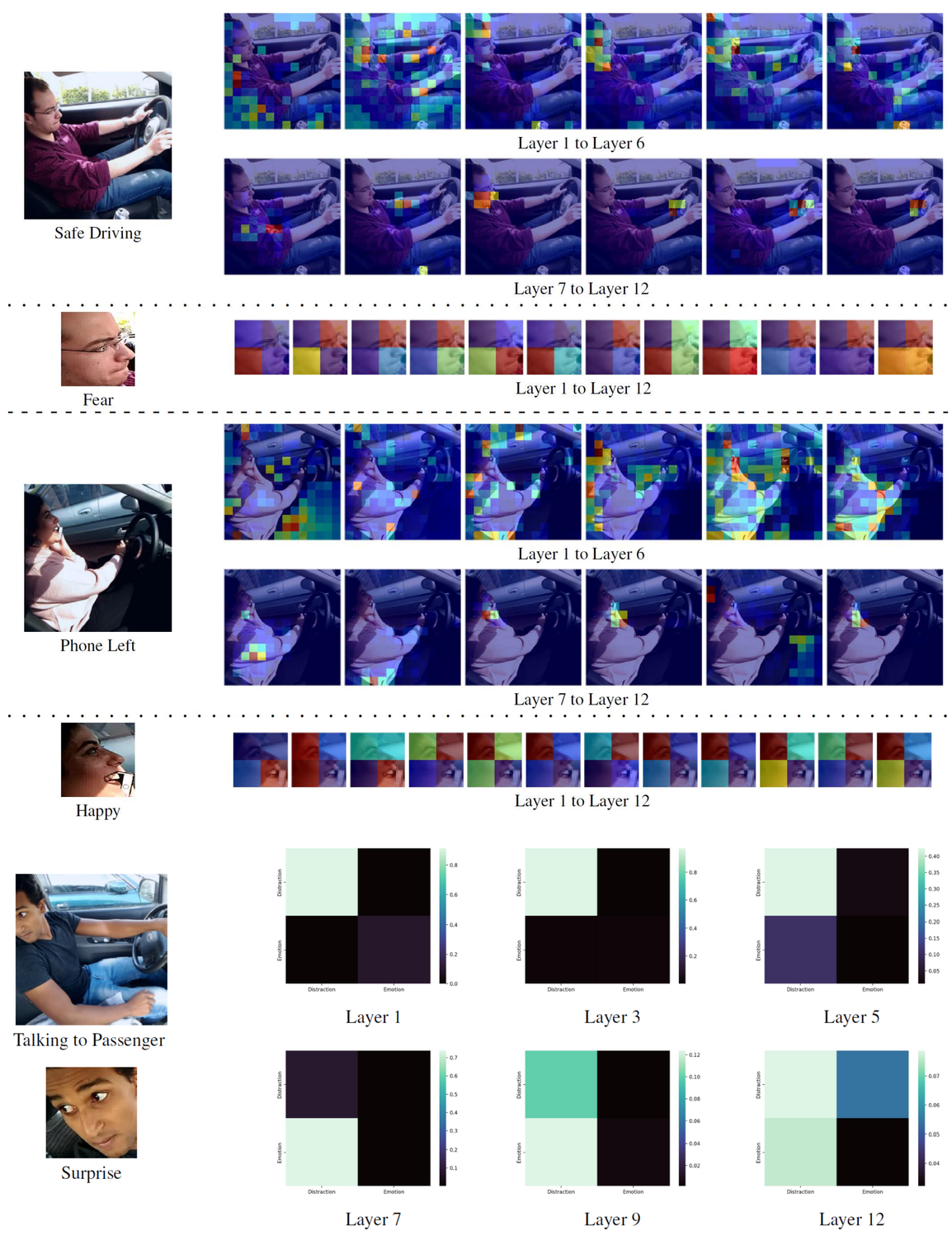
- Experiments are conducted on SFDDD and AUCDD benchmarks
- ViT-DD achieves 6.5% and 0.9% performance improvements as compared to the best state of the art methods

Ablation Study

Dataset	Method	Accuracy (↑)	NLL (↓)
SFDDD	Standard ViT[10]	0.9036	0.5355
	ViT-DD	**0.9251**	**0.3972**
AUCDD	Standard ViT[10]	0.9092	0.2895
	ViT-DD	**0.9359**	**0.2399**

- An ablation study is conducted in comparison to the standard ViT
- The accuracy improvements of ViT-DD are 2.2% and 2.7% on the SFDDD and AUCDD datasets, respectively

Visualization



Contact Information

Code: https://github.com/PurdueDigitalTwin/ViT-DD
Email: ma801@purdue.edu

SCAN ME