Synthesize & Learn: Jointly Optimizing Generative and Classifier Networks for Improved Drowsiness Detection

Paper # 4321

:) Affectiva

Sandipan Banerjee[1], Ajjen Joshi[1], Ahmed Ghoneim[1], Survi Kyal[1], Taniya Mishra[2]

SureStart

[1]Affectiva Inc., Boston, USA

[2]SureStart Inc., New York, USA

OVERVIEW

- Equipping vehicles with robust, automatic, real-time systems that can estimate a **driver's drowsiness state** from their **facial signals** can have potentially life-saving impact.
- Training **deep learning systems** based on real-world drowsy driving behavior can enable the development of such systems.
- However, **real-world drowsy driving datasets are unbalanced**, due to the sparsity of drowsy driving events, posing a challenge to building accurate drowsiness prediction systems.

CONTRIBUTIONS

- We attempt to **alleviate the class imbalance** problem by using GANs to synthesize examples of sparse classes directly in the feature-space.
- Our GAN-based framework **simultaneously generates** realistic examples of sparse drowsiness classes, and **learns** a separate drowsiness classifier using the generated samples.
- We **validate this approach in a real-world drowsiness dataset**, where we demonstrate a classifier trained using this approach outperforms a stand-alone classifier trained without any augmentations.

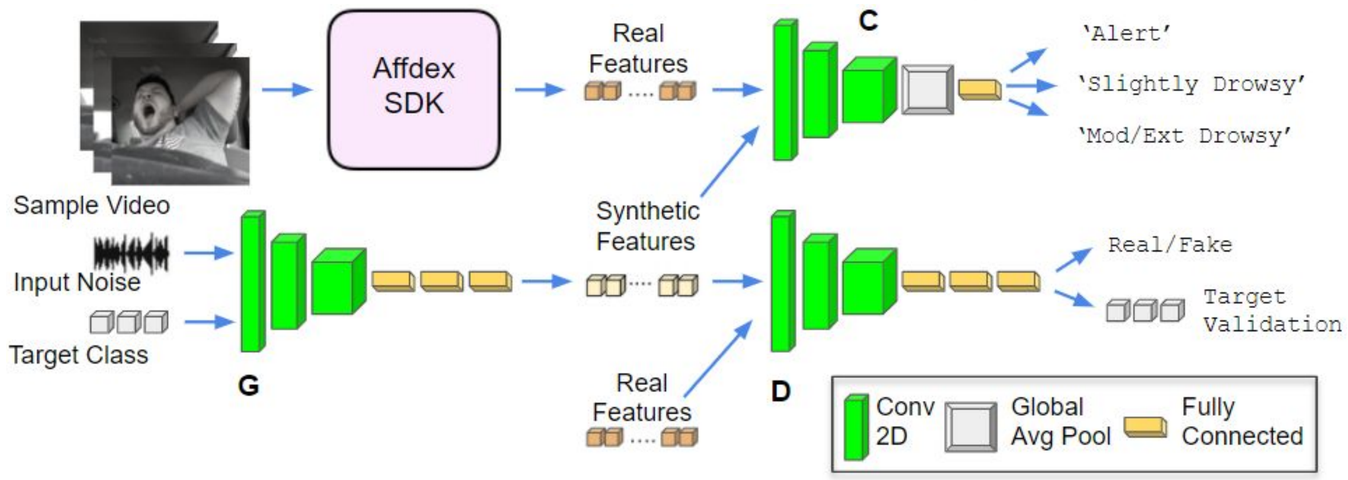
DATASET

- We used the **naturalistic drowsiness dataset** presented in [1], consisting of **1450 5-minute videos** of real-world driving behavior annotated into 4 drowsiness intensity classes: '**Alert**', '**Slightly Drowsy**', '**Moderately Drowsy**', '**Extremely Drowsy**'.
- We used the **Affdex SDK** [2] to generate a low-dimensional embedding for every frame, based on the driver's **head-pose**, **facial expressions** and **emotions** (anger, disgust, joy, surprise, valence).
- The dataset is **highly skewed** towards the 'Alert' class, with only ~3% of the data in the extremely drowsy classes.

PROPOSED GAN FRAMEWORK

- **Architecture**: Our GAN framework is composed of generator G and discriminator D networks jointly optimized for improving the drowsiness classifier C during training. During testing, only C is used.

  (1) **G**: takes a **100-D noise vector** & an embedded target class vector, and generates an **1800-D synthetic sample**.
  (2) **D**: takes the synthetic sample and generates predictions based on its **realness** and **target class association**.
  (3) **C**: is trained with real data, **augmented** with synthetic samples, and used to estimate drowsiness of validation set.



- **Loss Function**: The full loss is a weighted sum of following:

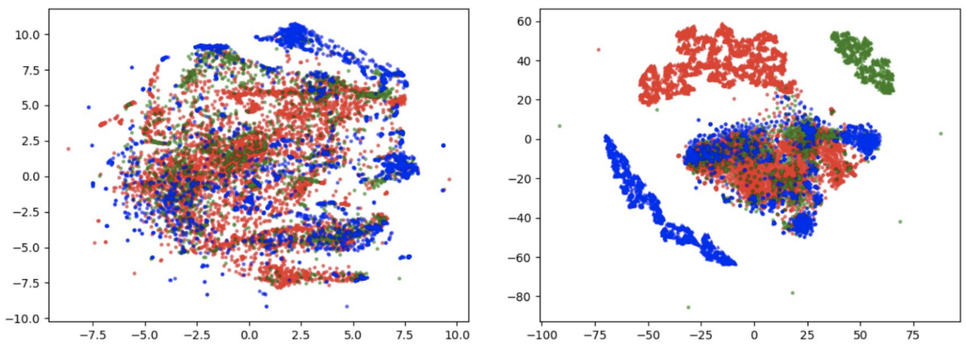
  (1) L_G: **D**'s weights are leveraged to tune **G**'s hallucinations to match distribution of real data and **produce realistic samples** as training progresses.
  (2) L_{cls}: Ensures the **target class association** of a synthetic vector is preserved, using cross entropy over **D**'s softmax prediction.
  (3) L_{corr}: To **generate variations** in synthetic data, while preserving ``realness", we calculate the Pearson correlation between real samples and synthetic samples for a particular drowsiness class. Ideally, the correlation should be as close to 1 as possible.
  (4) L_{opt}: Formulated as the cross-entropy over C's predictions of the level of drowsiness. On top of the explicit task of improving drowsiness prediction, this joint optimization loss **implicitly refines the quality of the generated synthetic samples**.

EXPERIMENTAL RESULTS

- **Training modes**: Along with initializing C **from scratch** for joint training, we also try fine-tuning it from a **pre-trained snapshot**. To estimate the optimal unfreezing point of C's weights, we fine-tuning it after 0, 25, 50 and 75 epochs.
- **GAN-based augmentation**: improves C's performance metrics (accuracy, ROC-AUC) for both the drowsy classes.
- **Well separated class boundaries**: Surprisingly, joint training also improves C's performance on the Alert class, when synthetic samples only for the drowsy classes are used in training.

Model	Alert	Slightly Drowsy	Mod/Extremely Drowsy	Macro Average
Classifier only (wo/ GAN) [4]	0.747,0.901	0.481,0.772	0.820,0.936	0.683,0.869
Joint training from scratch (proposed)	**0.841**,0.905	0.497,**0.785**	0.813,**0.956**	**0.717**,**0.882**
Joint training from snapshot, unfreeze after 0 epochs (proposed)	0.737,0.945	**0.540**,0.754	**0.849**,0.894	0.709,0.865
Joint training from snapshot, unfreeze after 25 epochs (proposed)	0.762,**0.950**	0.527,0.767	0.829,0.902	0.706,0.873
Joint training from snapshot, unfreeze after 50 epochs (proposed)	0.828,0.901	0.450,0.769	0.835,0.952	0.704,0.874
Joint training from snapshot, unfreeze after 75 epochs (proposed)	0.821,0.900	0.455,0.767	0.839,0.947	0.705,0.871

- **Feature visualization**: The synthetic samples stretch the feature manifold and push real samples from different classes away from each other, improving inter-class separation.
- **Initializing C from scratch is better**: During the initial epochs, C learns representations from noisy samples generated by G. As training progresses, D improves G, helping C to gradually learn the authentic representation of the drowsy classes.
- **Fine-tuning C from a snapshot**: When unfreezing early (0 or 25 epochs), the noisy synthetic samples from the early iterations distort C's representations. When unfreezing late (after 75 epochs), C learns better features from mature samples but does not experience the full variance produced by G through the whole 100 epochs.
- **Ablation Study**: We ablate each model component individually to gauge their contribution. L_G and L_{opt} are key in generating realistic synthetic samples while the L_{cls} and correlation L_{corr} losses act as regularizers.



Model	Macro Avg. Accuracy	Macro Avg. ROC-AUC
wo/ L_G	0.676	0.861
wo/ L_{cls}	0.680	0.860
wo/ L_{corr}	0.681	0.863
wo/ L_{opt}	0.673	0.860
Full model	**0.717**	**0.882**

[1] A. Joshi, et al. "In-the-wild drowsiness detection from facial expressions," in Human Sensing and Intelligent Mobility Workshop, IEEE Intelligent Vehicles Symposium, 2020
[2] D. McDuff, et al. "Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit," in CHI Extended Abstracts, 2016