Video-based Deception Detection and Corresponding Feature Discovery

Anastasis Stathopoulos, Ligong Han, Dimitris Metaxas

Center for Computational Biomedicine, Imaging and Modeling (CBIM)

Rutgers University

Outline

- Problem Definition
- Literature Review
  - Video Modeling
  - Datasets for video-based Deception Detection
  - Video-based Deception Detection
- Method
- Results

Problem Definition

- Video-based Deception Detection

- Given an input video, classify it as positive when the person exhibited deceptive behavior at some point

- Temporally localize (for positive samples) when deception took place
  - Debug the system
  - Enable scientist study the act of deception

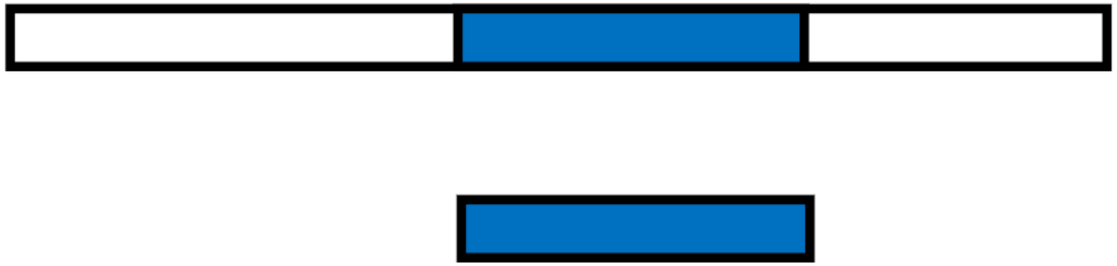
Literature Review: Video Modeling

Video Classification

- Input: 1 video
- Output: 1 categorical label

→ Model → Skateboarding

Trimmed vs Untrimmed



Untrimmed Video Classification

Trimmed Video Classification

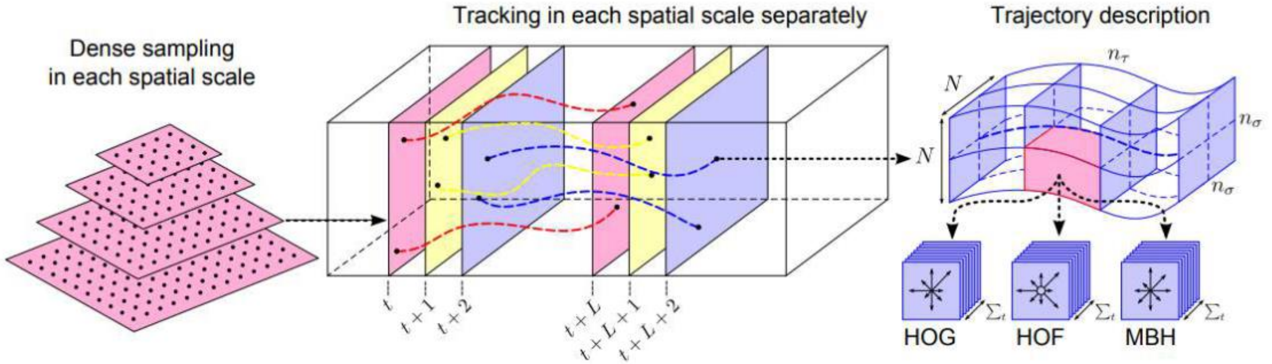
- Rich spatiotemporal information in videos
- How to extract the useful information to make a prediction?
- Prediction in untrimmed videos is a harder task
- Real-world application of trimmed videos is limited

Methods for Video Representations

- Hand-crafted Spatiotemporal Features
  - Space-time bag of features
  - Dense Trajectories
  - Improved Dense Trajectories (iDT)
- Deep Features
  - Deep Neural Networks to extract video representations

Dense Trajectories

- Dense Trajectories [1]



- Improved Dense Trajectories (iDT) [2]
  - Camera Motion
  - Human Mask (center feature extraction around the person)

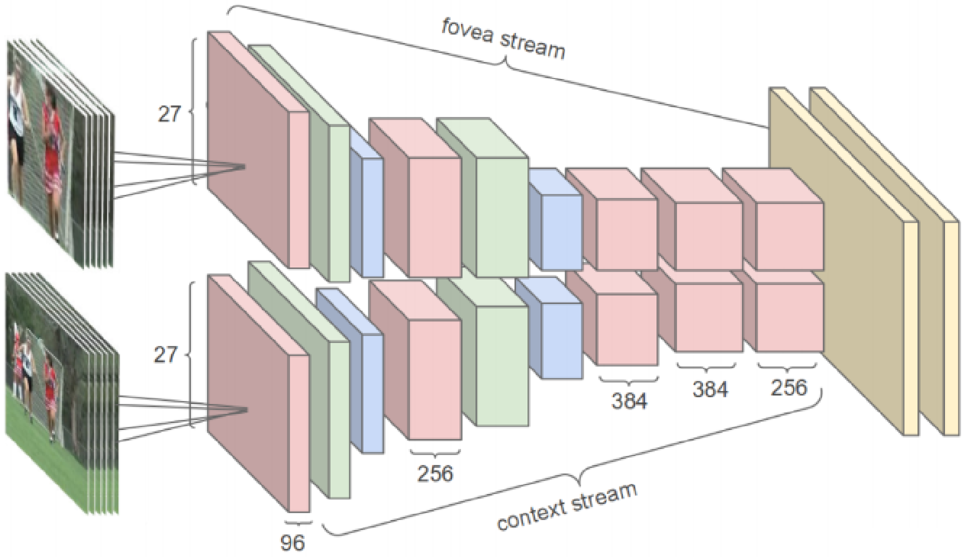
[1] Heng Wang et al., **Dense trajectories and motion boundary descriptors for action recognition**, IJCV 2013

[2] Heng Wang et al., **Action recognition with improved trajectories**, ICCV 2013

Hand-crafted Features

- Heavy computational cost
- Hard to scale and deploy

Single Stream Network



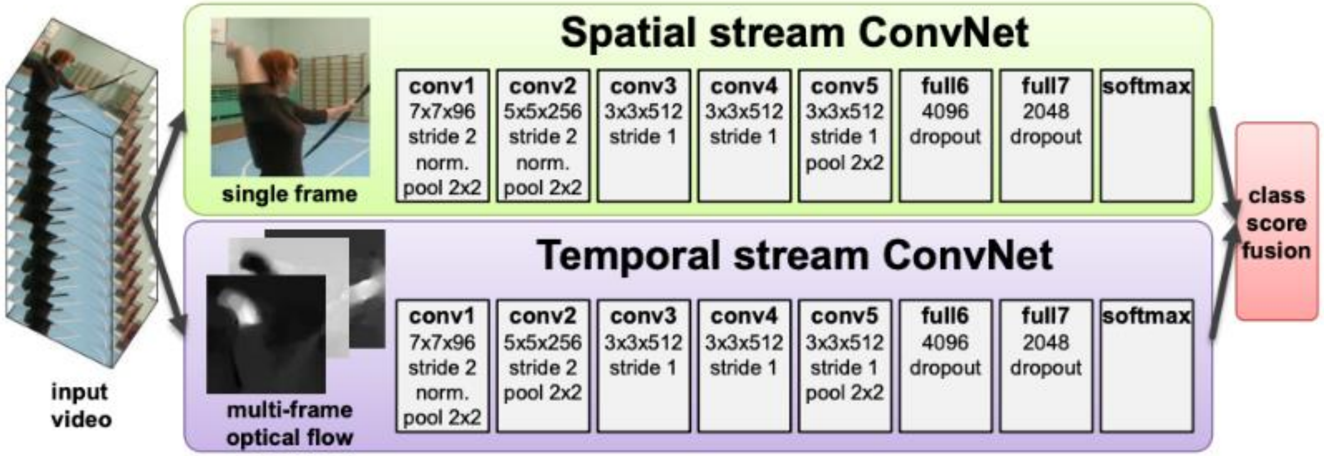
	UCF-101
IDT	87.9%
DeepVideo	65.4%

Average Classification Accuracy

DeepVideo lacks motion modeling

Andrej Karpathy et al., **Large-scale Video Classification with Convolutional Neural Networks**, CVPR 2014

Two-Stream Network



	UCF-101
iDT	87.9%
DeepVideo	65.4%
Two-Stream	88.0%

- First time that a DL approach achieves similar performance to hand-crafted features

Simonyan et al., **Two-Stream Convolutional Networks for Action Recognition in Videos**, NeurIPS 2014

Two-Stream Network Follow-up

➢ A lot of follow-up papers based on two-stream networks

[1] Limin Wang et al., **Action Recognition with Trajectory-Pooled Deep-Convolutional Descriptors**, CVPR 2015

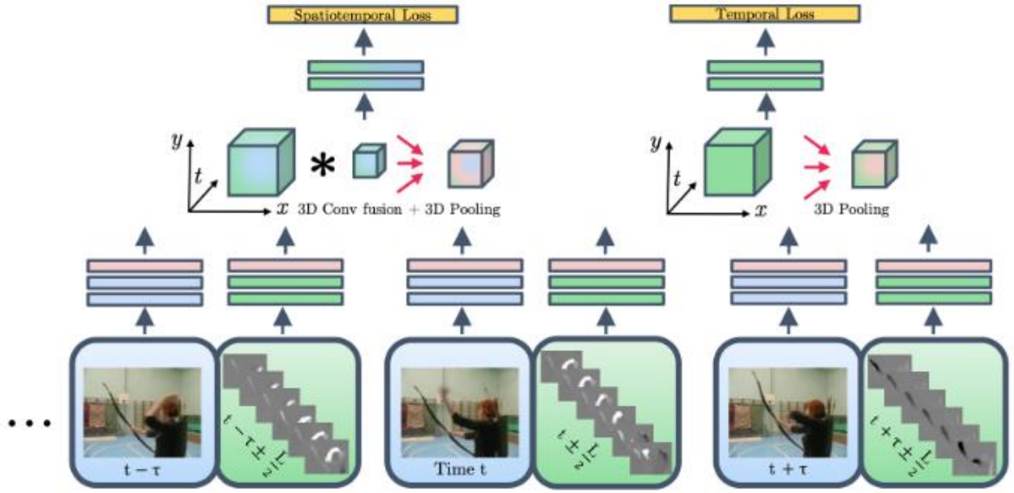
[2] Joe Yue-Hei Ng, **Beyond Short Snippets: Deep Networks for Video Classification**, CVPR 2015

[3] Christoph Feichtenhofer, **Convolutional Two-Stream Network Fusion for Video Action Recognition**, CVPR 2016

[4] Limin Wang et al., **Temporal Segment Networks,** ECCV 2016

[5] Abi Diba et al., **Temporal Linear Encoding Networks**, CVPR 2017

Two-Stream Fusion

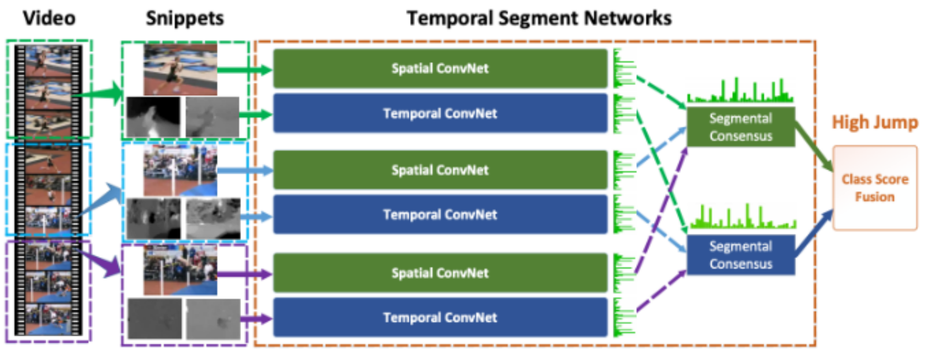


	UCF-101
iDT	87.9%
DeepVideo	65.4%
Two-Stream	88.0%
Two-Stream Fusion	92.5 %

Christoph Feichtenhofer, **Convolutional Two-Stream Network Fusion for Video Action Recognition**, CVPR 2016

Temporal Segment Networks (TSN)

- ➢Divide video into segments
- ➢Consensus to aggregate information about clips
- ➢Model long-range temporal structure over the entire video



	UCF-101
iDT	87.9%
DeepVideo	65.4%
Two-Stream	88.9%
Two-Stream Fusion	92.5%
TSN	94.0%

Limin Wang et al., **Temporal Segment Networks,** ECCV 2016

Two-Stream Networks Follow-up

- Performance on UCF-101 is saturated

- Drawback: Usage of optical flow
  - Precomputing optical flow is **computationally intensive** and **storage demanding**
  - Not ideal for large-scale training or real-time deployment

3D CNNs

- **C3D** [1]
  - Replace the 2D kernels of VGG-16 [2] with 3D kernels
  - Lower performance than two-stream networks

	UCF-101
iDT	87.9%
DeepVideo	65.4%
Two-Stream	88.0%
C3D	82.3%

[1] Tran et al., **Learning Spatiotemporal Features with 3D Convolutional Network**, ICCV 2015

[2] Simonyan, Zisserman, **Very Deep Convolutional Networks for Large-Scale Image Recognition**, ICLR 2015

I3D

- C3D trained from scratch: hard to optimize
- I3D initialize 3D model weights by utilizing 2D weights trained on ImageNet

	UCF-101
iDT	87.9%
DeepVideo	65.4%
Two-Stream	88.0%
C3D	82.3%
I3D	**95.6%**

Kinetics-400

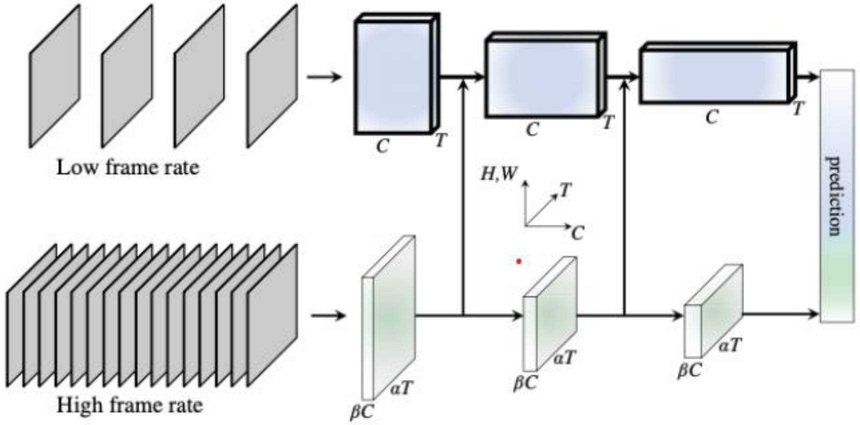
- Performance on UCF-101 is saturated

- Kinetics-400 [1] is used to benchmark models

[1] Zisserman et al., **The Kinetics Human Action Video Dataset**, arXiv 2017

SlowFast Network

- **Slow Pathway**: capture detailed semantic information
- **Fast Pathway**: rapidly changing motion



	Kinetics-400
C3D	59.5%
I3D	71.1%
SlowFast	**78.0%**

Feichtenhofer et al., **SlowFast Networks for Video Recognition**, ICCV 2019

Datasets for video-based Deception Detection

Real-life Trial (RLT)

- **RLT** [1]
  - Publicly available database with 121 videos from real-life court room trials
  - Only **104 videos** are used in practice
  - Label for someone telling a truthful fact or not
  - **Duration**: few second clips
  - "**Trimmed**" Videos

[1] Perez-Rozas et al., **Deception detection using real-life trial data**, ICMI 2015

Bag-of-Lies (BoL)

- **BoL** [1]
  - Publicly available dataset
  - 35 subjects, each of whom are shown some images and asked to describe them
  - Subjects describe some images honestly, while other deceptively
  - Answers recorded in a video
  - **Duration:** [4 – 42] seconds
  - “**Trimmed**” videos

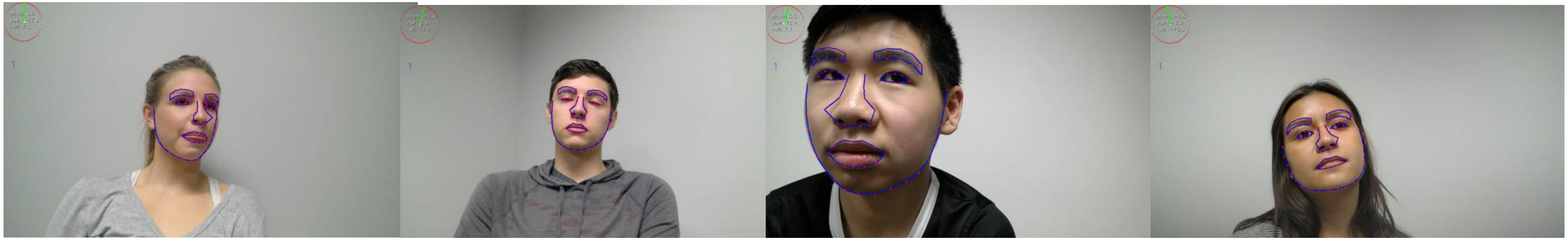
[1] Gupta et al., **Bag-of-lies: A multimodal dataset for deception detection**. CVPR workshops, 2019

Problems with public datasets

- Both datasets are **trimmed**
- They contain a **single act of deception**
- Need extra steps to be done if we wish to build a real-world application
- Need to introduce a new dataset to study the deception detection

Resistance Game

- Dataset with videos from a social role-playing game
- Players are given one of two roles
  - **Deceivers** or **Truthtellers**
- **Untrimmed** videos: average duration **46 minutes**



Truth-teller   Deceiver   Truth-teller   Truth-teller

Literature Review:
Video-based Deception Detection

Deception Detection in Videos (DDiV) [1]

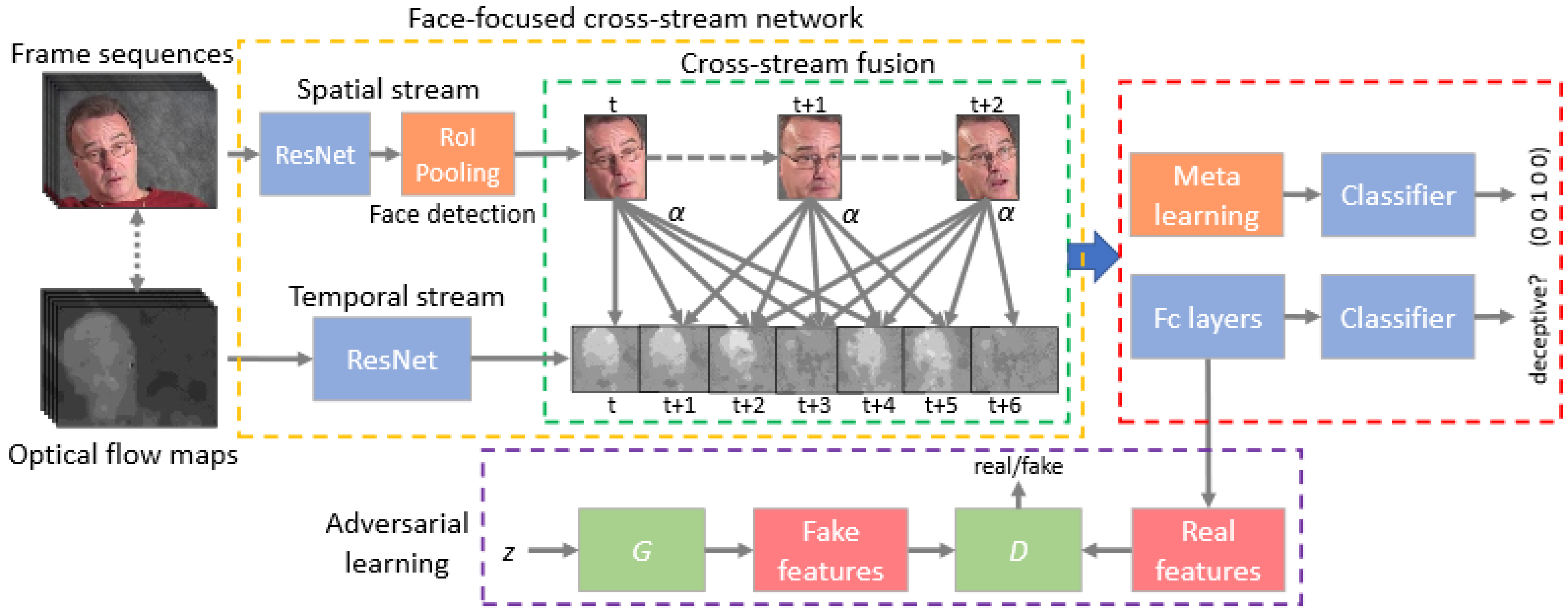
- Get iDT features
- Fisher Vector [2] encoding to aggregate features to a fixed length vector (low-level features)
- Use features to predict micro-expression detectors (high-level features)
- Combine low-level and high-level features for **binary classification**
- Hand-crafted features

[1] Wu et al., **Deception Detection in Videos**, AAAI 2018

[2] Jaakola et al., **Exploiting generative models in discriminative classifiers**, NeurIPS 1999

Face-Focused Cross-Stream Network (FFCSN)

- Two-Stream Network



Mingyu Ding et al., **Face-Focused Cross-Stream Network for Deception Detection in Videos**, CVPR 2019

Problems with current methods

- Are tested only on **trimmed** videos
  - Real-world application limited

- Overfit to background (training samples are limited)
  - Experienced overfitting issues when using off-the self video modeling deep architectures mentioned before

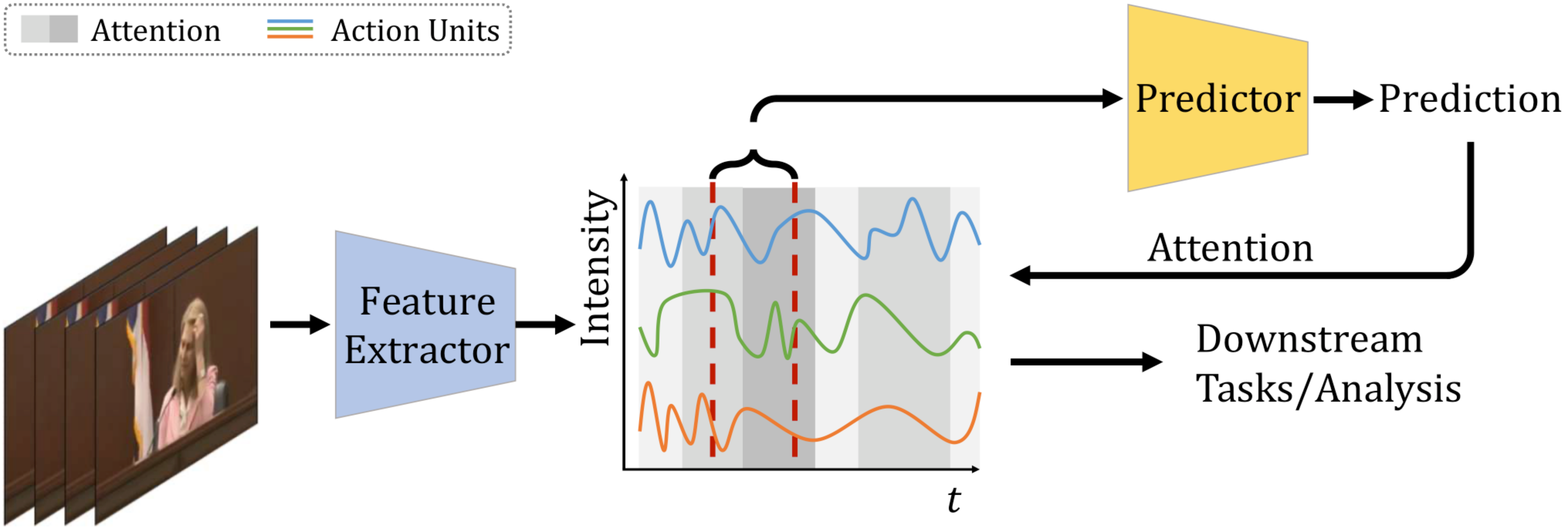
- Their predictions are not easy to interpret

Proposed Approach

Method

- Propose a two-stage approach
- Extract identity invariant and robust facial features (17 Facial Action Units, or FAUs, normalized with the parameters of the morphable model fitted to subjects' face; gaze angles, etc.)
- Those measurements define a set of 1-D signals (over time); Concatenate those 1-D signals **channel-wise**
- Feed input waveform to a Temporal Convolution Network (TCN)
- Use labels to train the model for **binary classification**

Pipeline



Attention
Action Units
Feature Extractor
Intensity
t
Predictor
Prediction
Attention
Downstream Tasks/Analysis

Contributions

- Achieves state-of-the-art performance on video-based deception detection on several benchmarks.

- The proposed framework is modular, lightweight and robust to the identity of a person by nature.

- Allows a framework for retrospective analysis of deceptive behavior.

Baseline

- Temporal Segments Networks (TSN) [1]

- Two-Stream architecture
  - **Appearance Stream**: RGB frames
  - **Motion Stream**: Optical Flow maps

[1] Limin Wang et al., **Temporal Segment Networks,** ECCV 2016

Results: RLT

Methods	ACC (%)	AUC (%)
TSN	77.5	81.78
DDiV	-	83.47
FFCSN	89.16	91.89
Ours	**92.36**	**97.27**

Results: BoL

Method	ACC (%)	AUC (%)
LBP	55.12	55.32
TSN	56.94	57.62
Ours	**64.47**	**67.08**

Results: Resistance Game

Method	ACC (%)	AUC (%)
LBP	49.56	49.56
TSN	51.15	51.15
Ours	**71.08**	**71.08**

Analysis of Deceptive Behavior

- Adapt Grad-CAM [1] to find the find the attention of the model in the time domain
- For positive samples we can compute the key time-steps for the decision of the detection model
- Utilize the gradient of the model w.r.t. a feature layer
- Framework for retrospective analysis of deceptive behavior by domain experts

[1] Selvaraju et al., **Grad-Cam: Visual explanations from deep networks via gradient-based localization,** ICCV 2017

Analysis of Deceptive Behavior

Closing Remarks

- Off-the self video classification architectures overfit due to small number of samples available

- Providing high-level information to the model helps
  - Do not model pixel-level nuances

- Framework for retrospective analysis of deceptive by utilizing the gradients of the model

Thank you