SMARTSURV :

A 3D CNN that recognizes actions in video surveillance

Godfred S. Gyimah

PROJECT MOTIVATION

- CCTV operators typically monitor 16 to 64 cameras concurrently on the same screen.

- Humans lose ~95% of their attention after focusing on a screen for 20+ mins (Green, 1999).

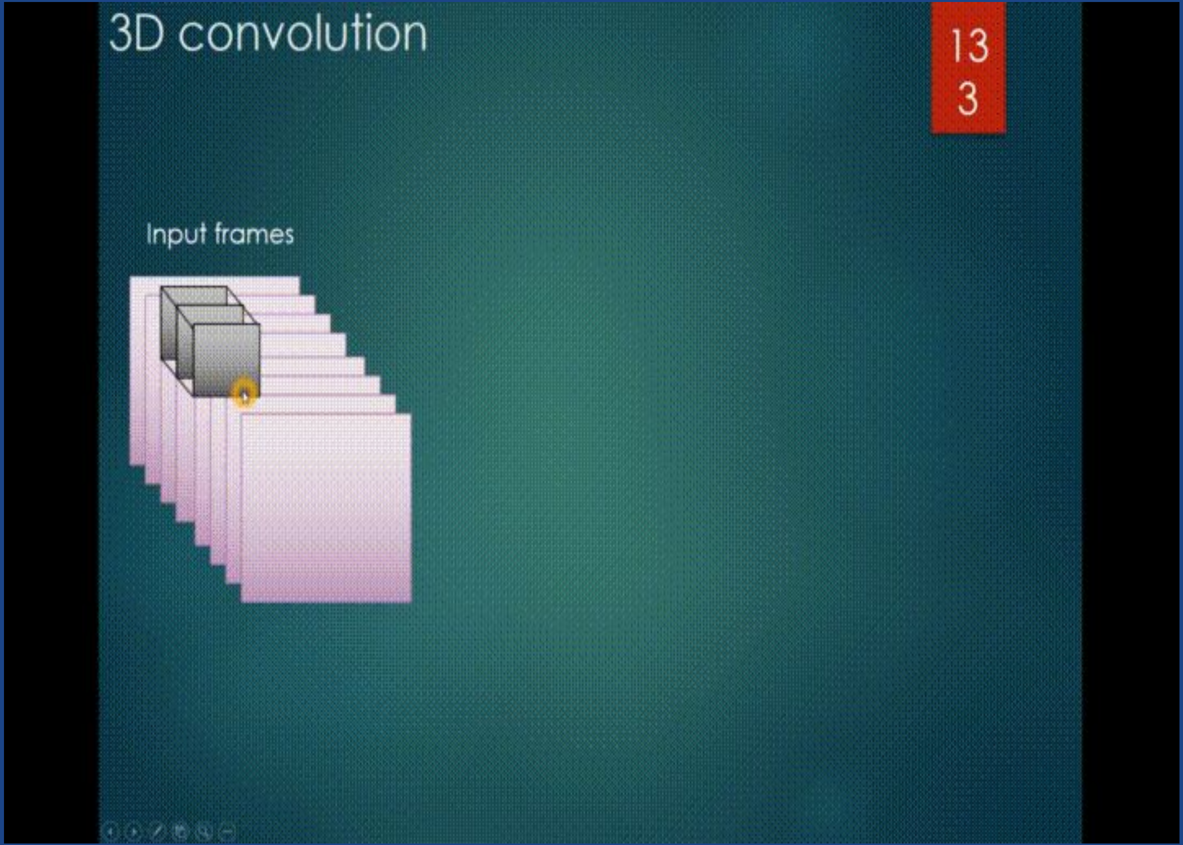
- Some CCTVs are not monitored.

- **Solution ??:** A system which recognizes certain actions and flags them.

BENCHMARKS FOR ACTION RECOGNITION

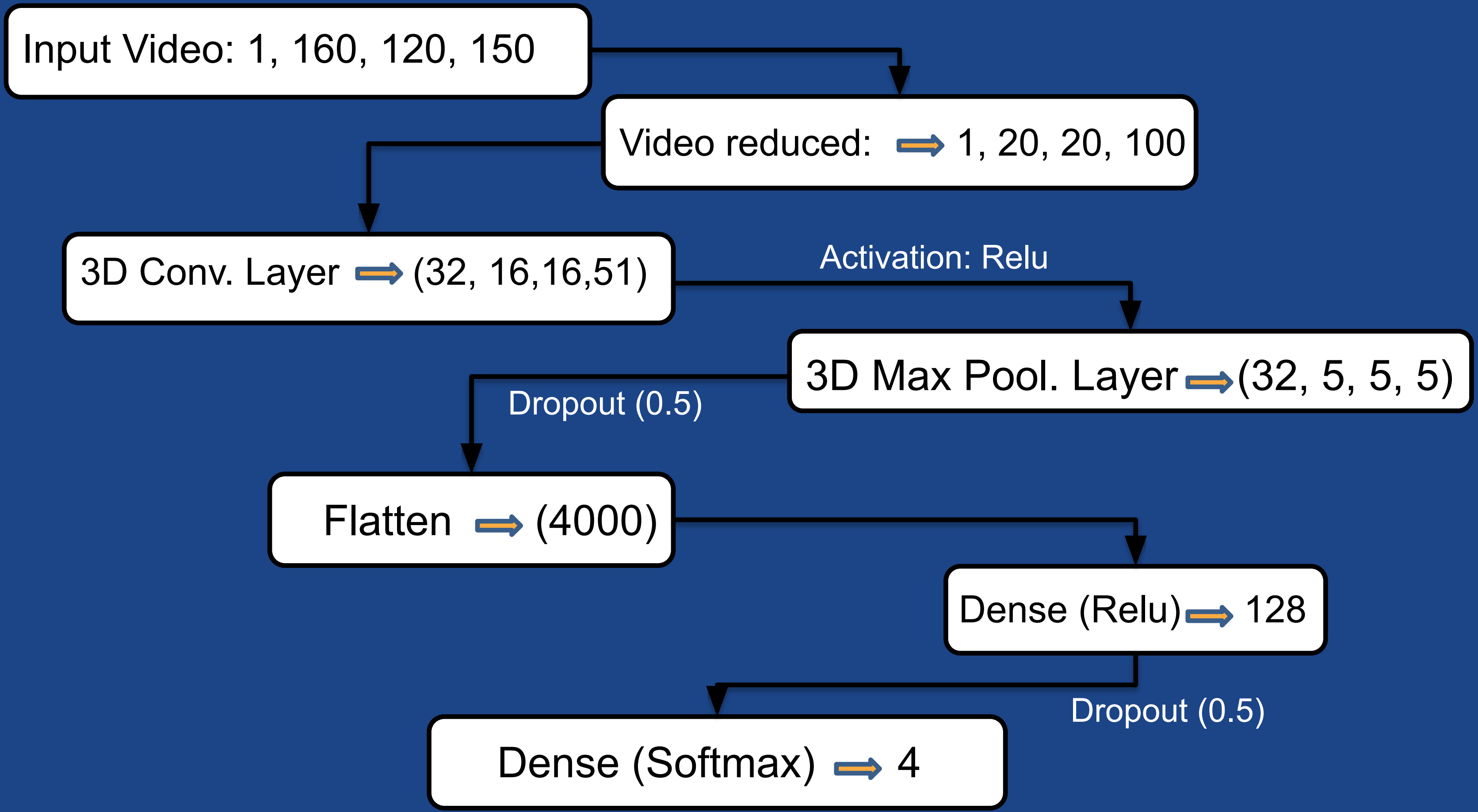
MODEL	METHOD	ACCURACY	RECALL	PRECISION
Schindler et al. 2008	3D CNN (10 frames: 0.5s)	92	88	89
Jhuang et al. 2007	Multi-class SVM (frame-by-frame)	91	-	-
Niebles et al. 2008	LDA & pLSA	83	83	84
Dollar et al. 2005	kNN & SVM	81	81	83
Schludt et al. 2005	SVM + Conv. Kernel	71	72	77
Smartsurv	3D CNN (50 frames: 5s)	??	??	??

So, how do we teach a machine to understand actions in videos?



3D convolution

MODEL PIPELINE



Input Video: 1, 160, 120, 150

Video reduced: → 1, 20, 20, 100

3D Conv. Layer → (32, 16,16,51)

Activation: Relu

3D Max Pool. Layer → (32, 5, 5, 5)

Dropout (0.5)

Flatten → (4000)

Dense (Relu) → 128

Dropout (0.5)

Dense (Softmax) → 4

MODEL INPUTS: KTH DATASET

- 4 action classes
- 400 videos
  - **Boxing / Fighting**
  - **Running**
  - **Walking**
  - **Waving**
- **Train: 80%**
- **Validation: 10%**
- **Test: 10%**

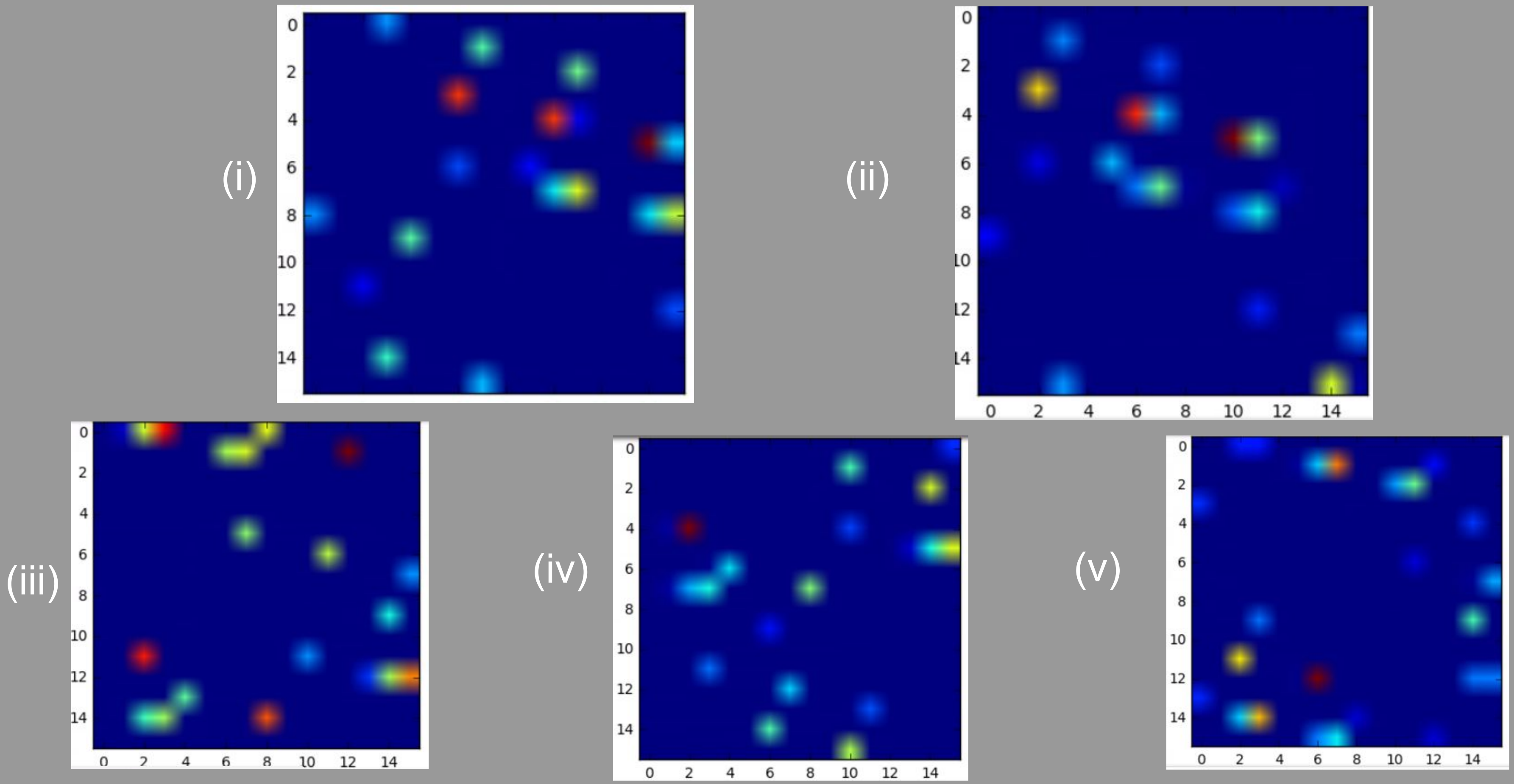
MODEL INPUTS: Train Parameters

- Optimizer: RMSprop (lr: 0.001)
- Loss function: Cross Entropy
- Training Time: 1000 epochs
- 1 hour on a 16GB P5000 GPU

MODEL RESULTS

RESULTS: 3D CONVOLUTIONS

5/51 Feature map samples from 1/32 filters for video 5 (boxing)



(i)

(ii)

(iii)

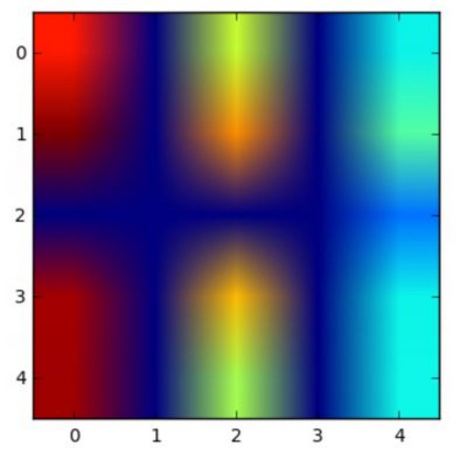
(iv)

(v)

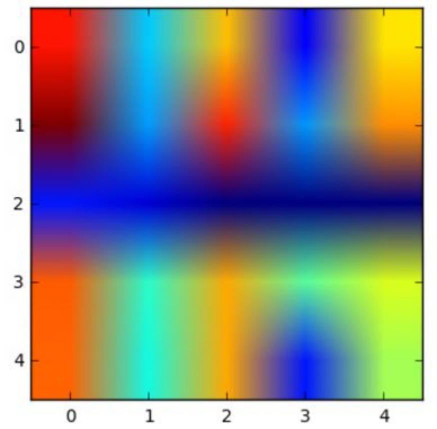
RESULTS: 3D MAXPOOLING STAGE

Feature maps from 5 seconds of a ‘walking’ class action

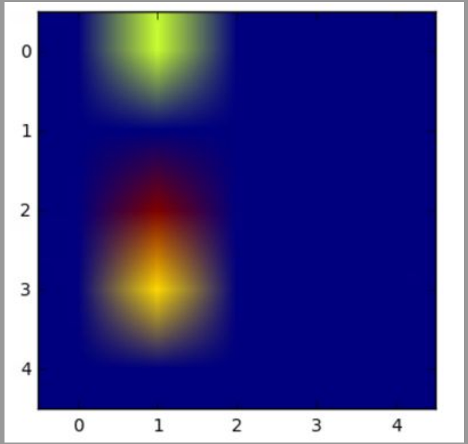
(i)



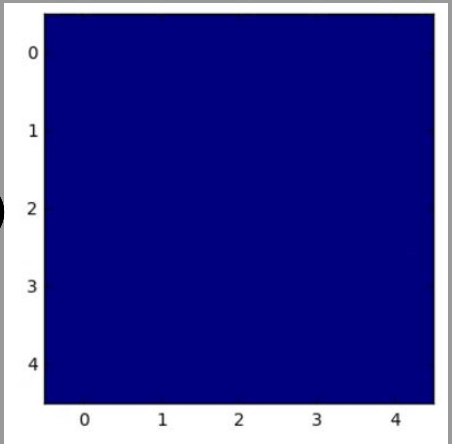
(ii)



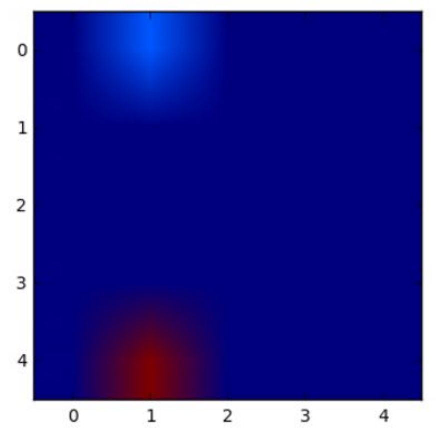
(iii)



(iv)



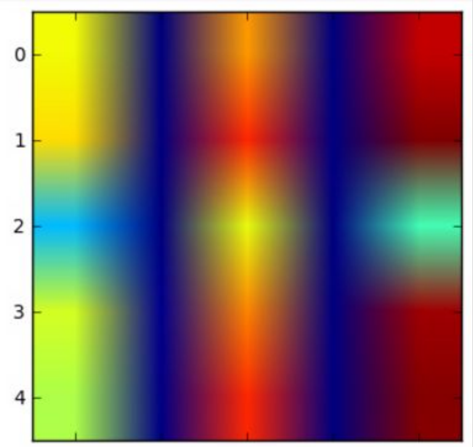
(v)



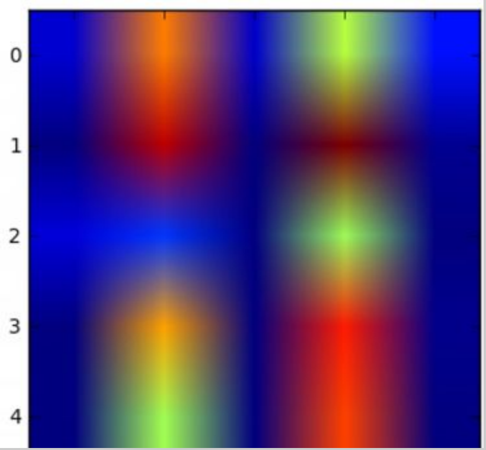
RESULTS: 3D MAXPOOLING STAGE

Feature maps for sample 3-D Maxpool filters

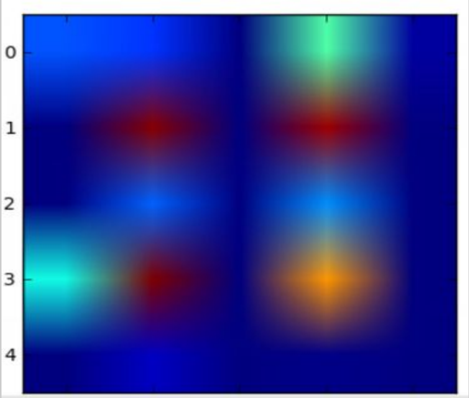
Boxing / Fighting



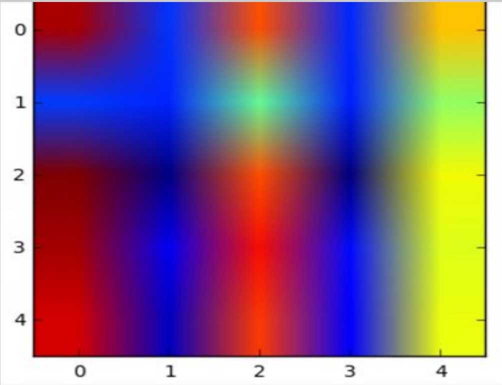
Walking



Running

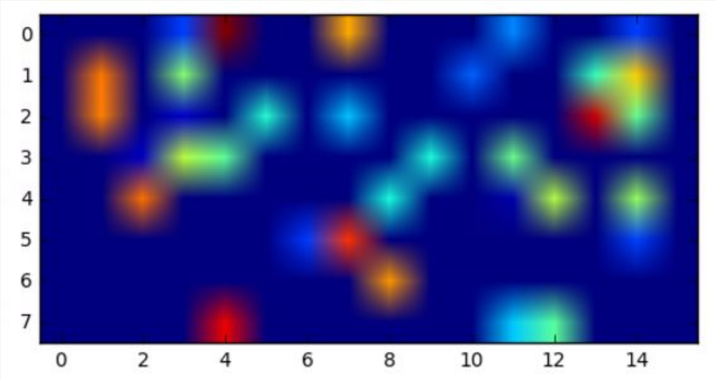


Waving

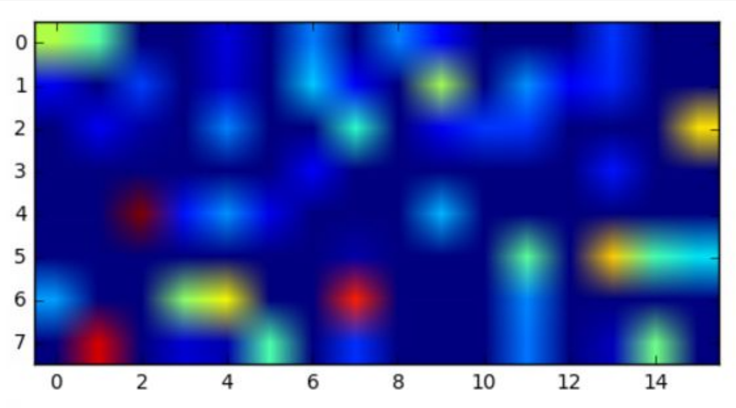


RESULTS: 128-node DENSE LAYER

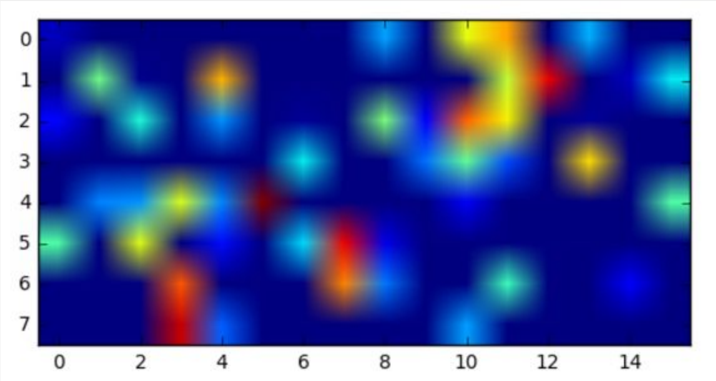
Boxing / Fighting



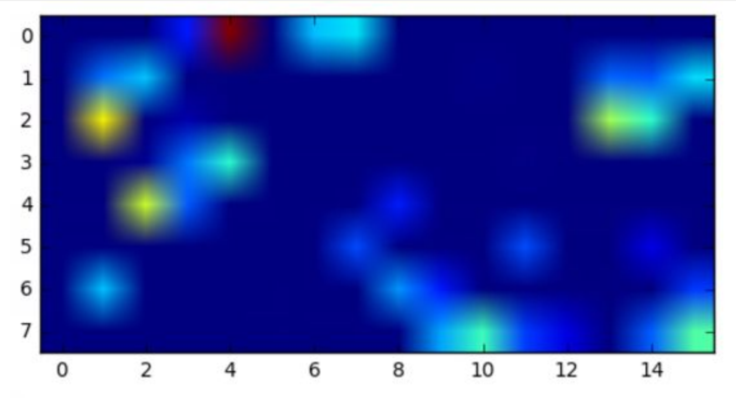
Walking



Running

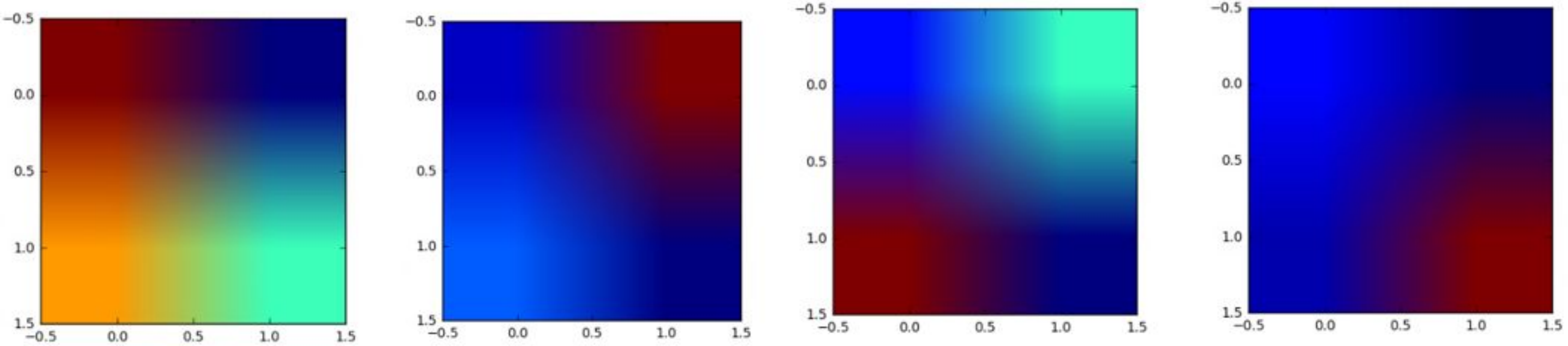


Waving

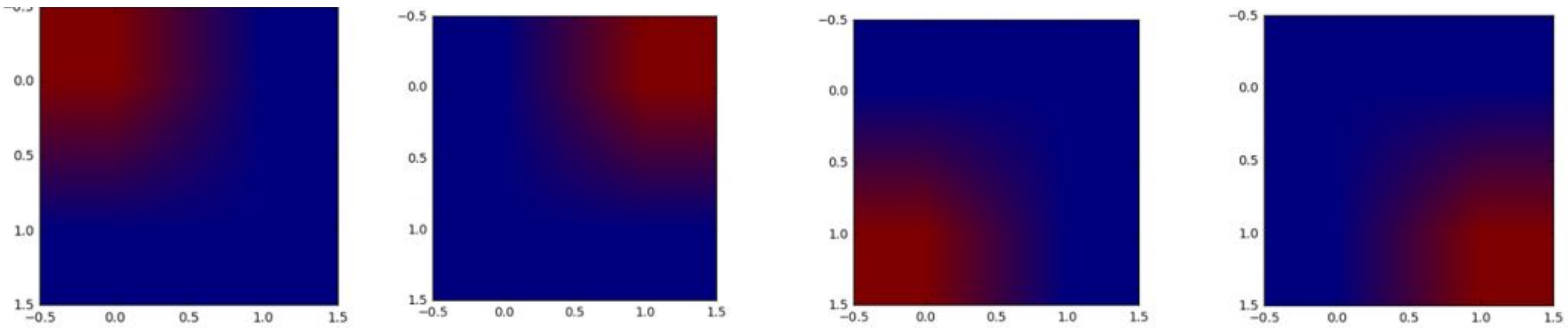


RESULTS: 4-node DENSE LAYER

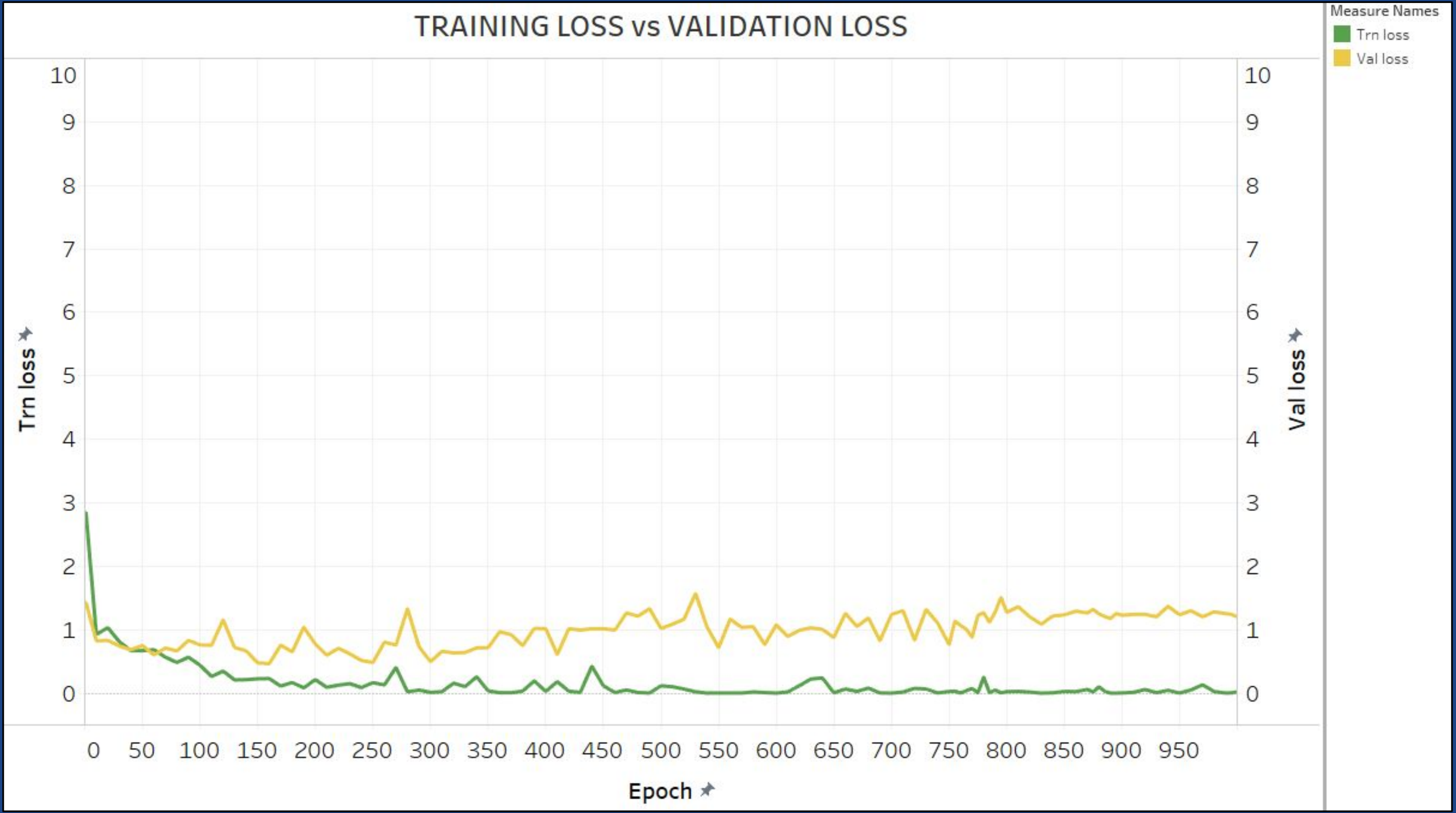
Boxing / Fighting | Running | Walking | Waving



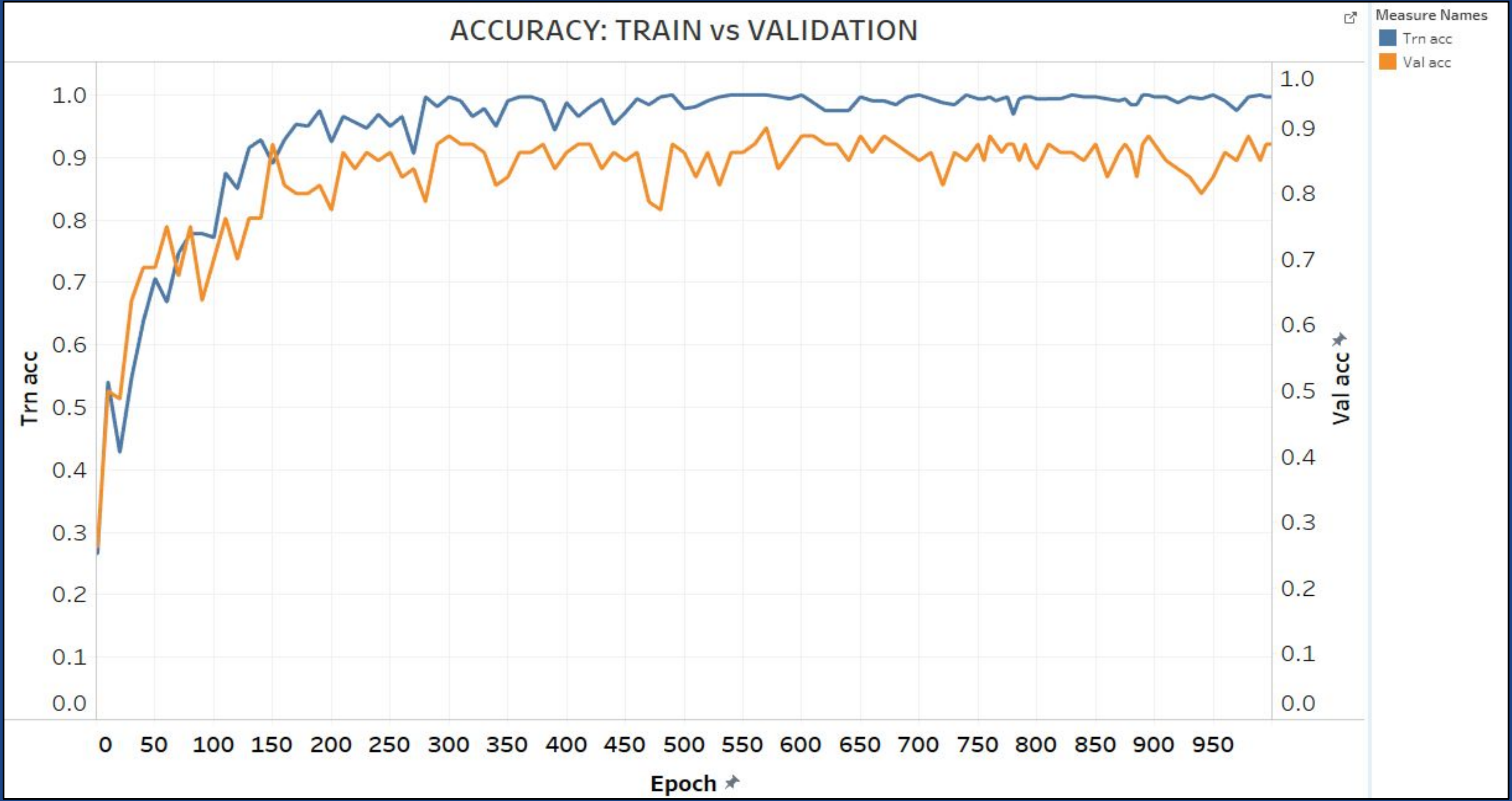
FINAL LAYER: Softmax



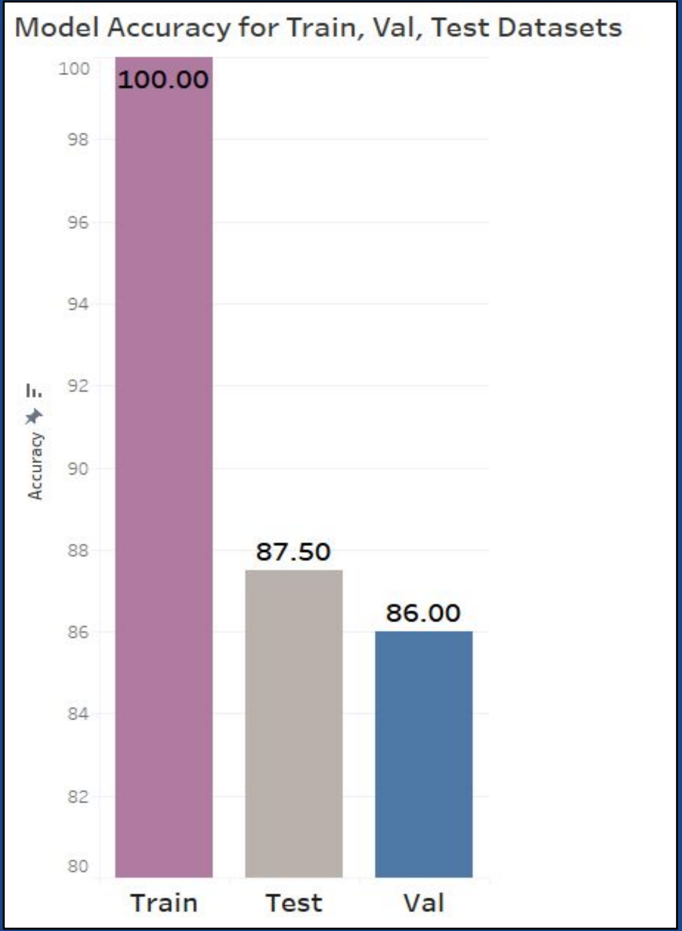
MODEL RESULTS: LOSSES

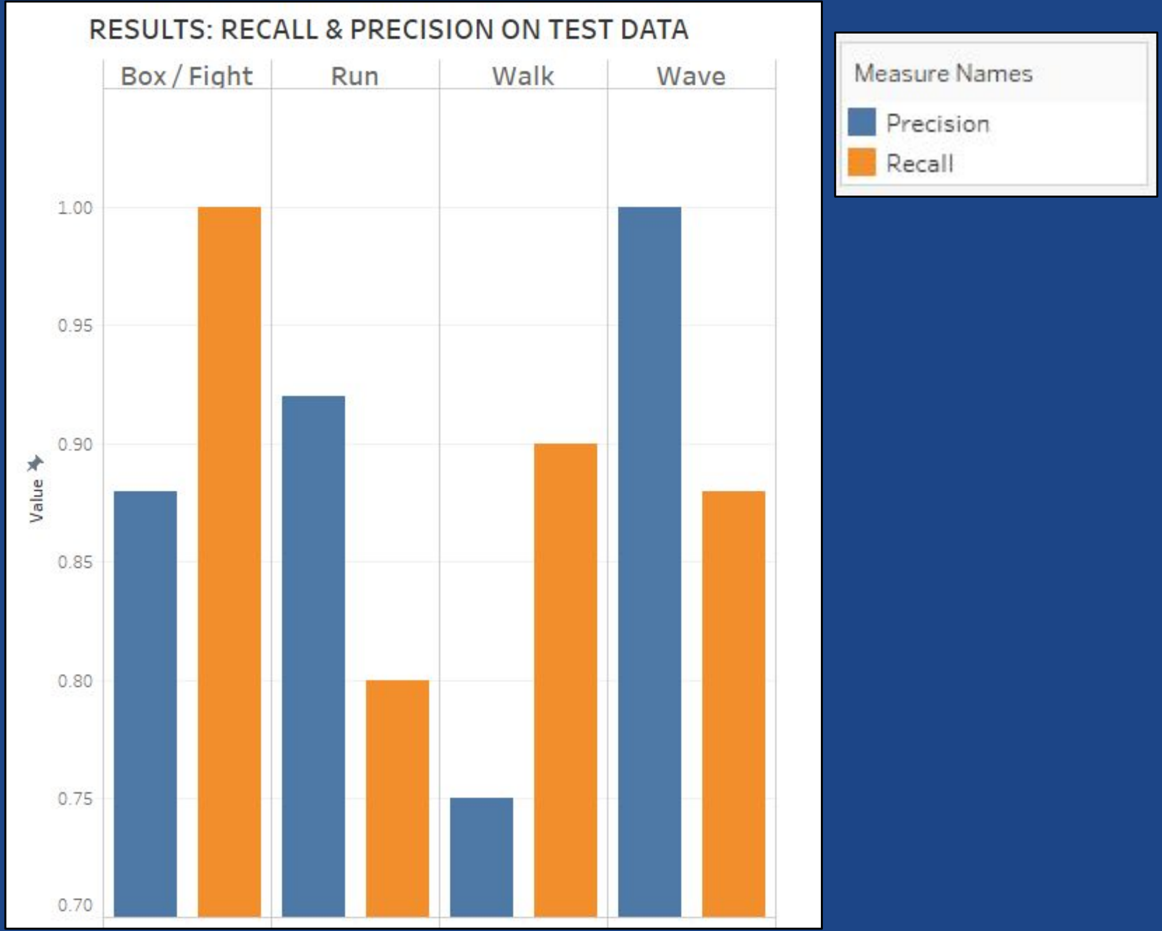


MODEL RESULTS:



MODEL RESULTS ON UNSEEN DATA


Model Accuracy for Train, Val, Test Datasets


RESULTS: RECALL & PRECISION ON TEST DATA

COMPARING MODELS ON SAME DATASET

MODEL	METHOD	ACCURACY	RECALL	PRECISION
Schindler et al. 2008	3D CNN (10 frames: 0.5s)	92	88	89
Jhuang et al. 2007	Multi-class SVM (frame-by-frame)	91	-	-
Smartsurv	3D CNN (50 frames: 5s)	88	89	90
Niebles et al. 2008	LDA & pLSA	83	83	84
Dollar et al. 2005	kNN & SVM	81	81	83
Schludt et al. 2005	SVM + Conv. Kernel	71	72	77

THE MODELING METHOD: OTHER BUSINESS CASES

❖ Self-driving Vehicles

❖ Autonomous Excavation

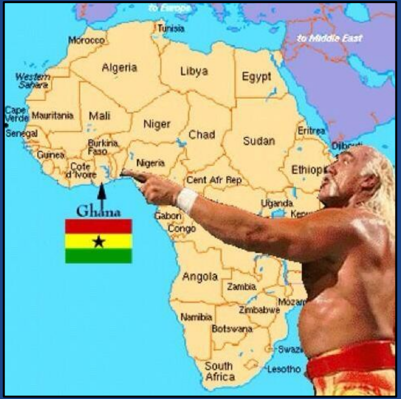
❖ Content-based video search engines

- Youtube : 20 hours of new videos per minute.

❖ Generally, for many problems involving:

- video data
- images (in a sequence)

ABOUT GODFRED...

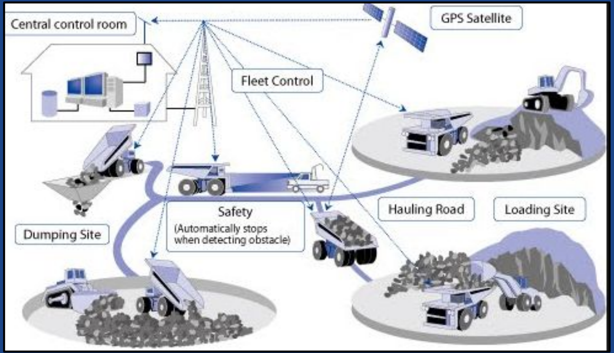


HUAWEI

- ❖ Civil Engineer
- ❖ Engineering Analytics

  (3 years)

- ❖ Geological Scientist



- ❖ PhD Research (Autonomous Excavation using Computer Vision & Machine Learning)