Human
Violence Detection

By: Ankur Goswami & Najmus Saqib

Approach

The approach, we used, is **Human Body Pose Tracking**. Our approach is divided into three phases.

Phase I: Human detection and Tracking

Phase II: Body Keypoints(joints) detection

Phase III: Classification (Violence or Non-Violence)

Phase I

Human Detection & Tracking

It is further divided into two subproblems, i.e., **detection & tracking**.

- You only look once (YOLO) is a state-of-the-art, real-time object detection system. Human Detection is done using **YOLO v3**. To be more specific, YOLOv3-320 can be used because its inference time is 22 ms per frame which is nearly equal to 45 frames per second. YOLOv3 is pre-trained on COCO dataset with 80 classes. We can retrain the network on same dataset with only 1 class, person or not. This will make our model light hence boosting its speed.
  Reference: You Only Look Once: Unified, Real-Time Object Detection

- Tracking is done using **DeepSORT.**
  This algorithm is so intuitive. In all of the available methods, one fundamental thing that is missing that we humans use all the time in tracking is a visual understanding of that bounding box. We track based on not just distance, velocity but also what that person looks like. Deep sort allows us to add this feature by computing deep features for every bounding box and using the similarity between deep features to also factor into the tracking logic.

  Reference: Simple Online and Realtime Tracking with a Deep Association Metric

Results

FPS: 6 on 720x1280 frame
GPU: Nvidia GeForce GTX 1070Ti

Phase II

Human Body Keypoint Detection

- **Approach used**: Body joints and limb detection

- **Reference**: OpenPose-Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields (v2018).

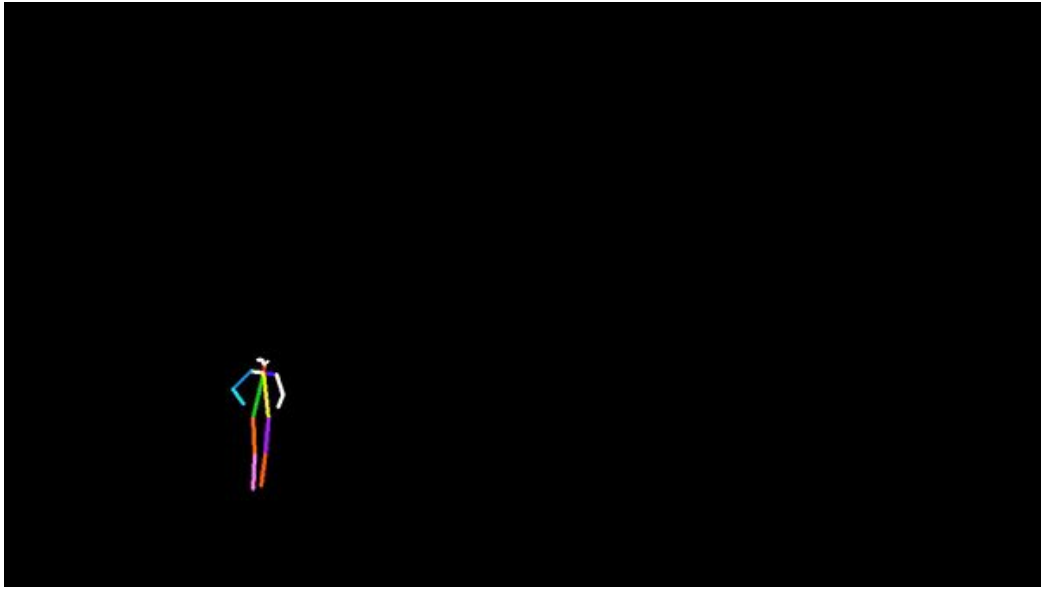
- **Description**: We use OpenPose Pose Estimation architecture. This architecture uses Part Affinity Fields(PAF) and Confusion maps to detect body keypoints. Part Affinity Fields(PAF) are vectors that represent limbs, connecting two joints whereas Confusion maps are heatmaps that represent probability of a keypoint in the region.

Results

Middle person in front row with keypoints.

Skeletal View of a Human

FPS: 0.18 (5 secs for each frame) on 720x1280 frame
GPU: Nvidia GeForce GTX 1070Ti



Association of Bounding boxes with keypoints

(Tracking of body keypoints)

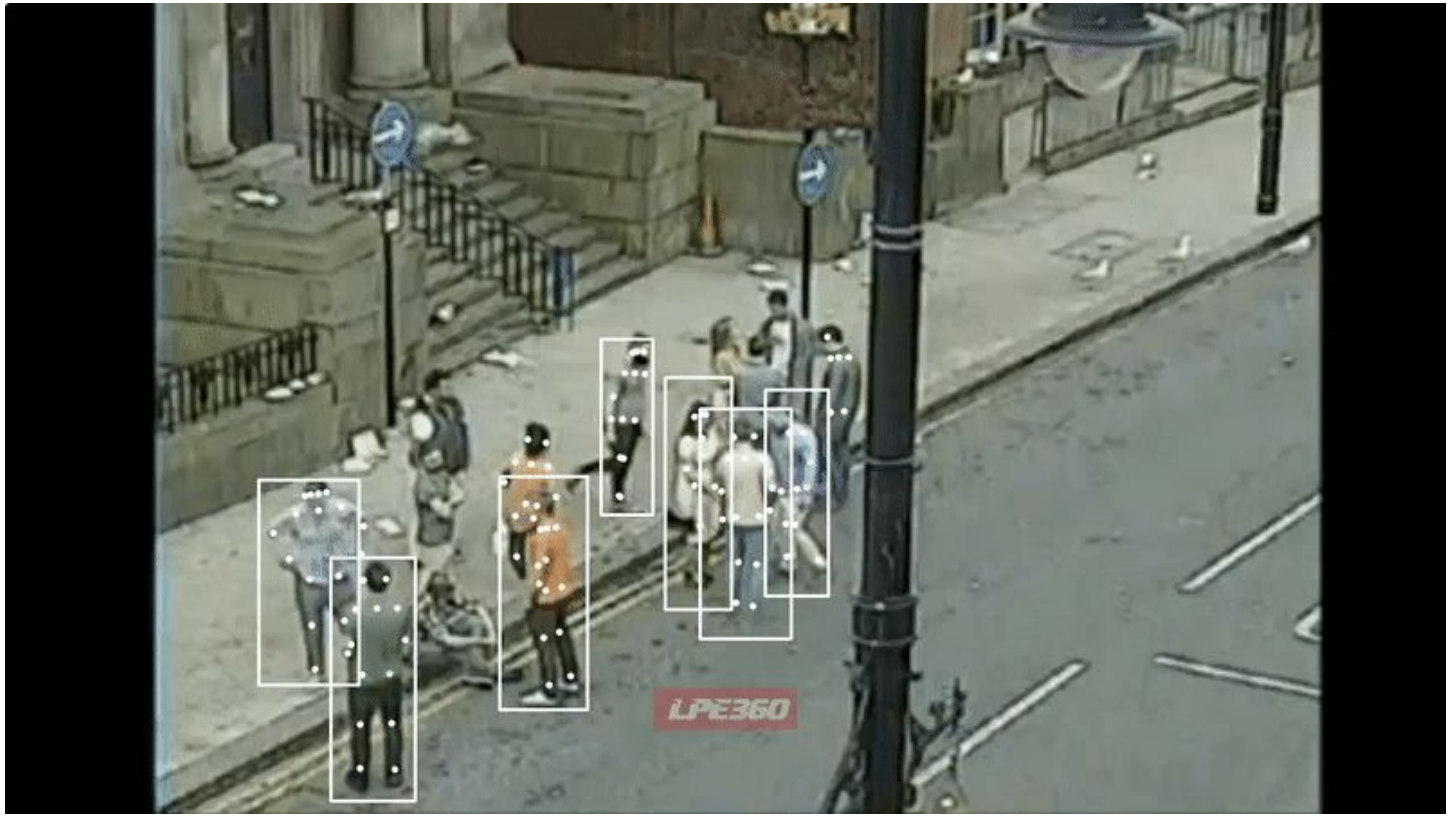
Association is done using a bruteforce method on following criterion:

Criteria 1: More than 70%* of keypoints must lie inside a given bounding box.

Criteria 2: The ratio of height of a body(distance b/w top and bottom keypoint coordinate) and height of a given bounding box must be more than 0.7*.

* These numbers are experimental thresholds.

Results



Phase III

Classification

Classification is done using **Recurrent Neural Networks(RNNs)**.
Recurrent neural networks (RNNs) are networks with loops, allowing information to persist [Rumelhart et al., 1986].

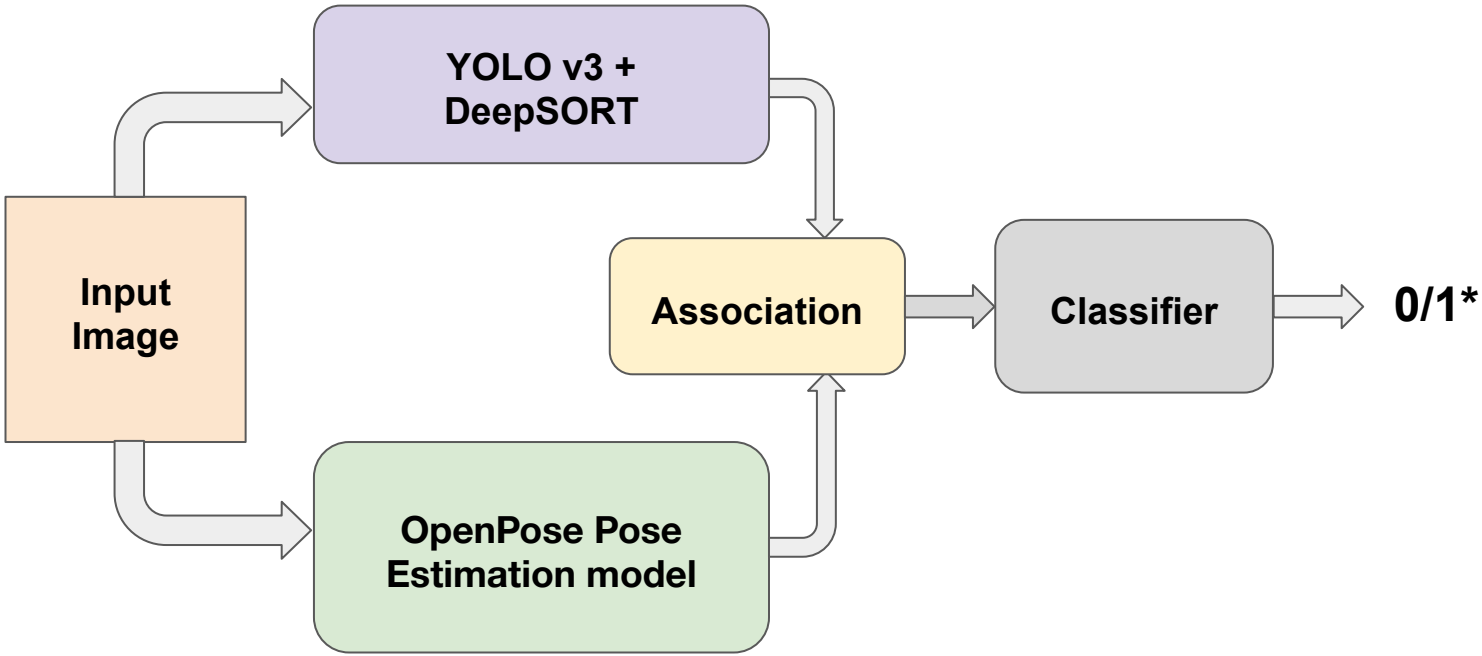
Features of RNN:

- RNN have memory that keeps track of information observed so far.
- RNN maps from the entire history of previous inputs to each output.
- RNN can handle sequential data.

The reason to prefer RNNs over CNNs or DNNs is that one cannot conclude that a subject is violent or non-violent on the basis of a single image(or a single body pose).

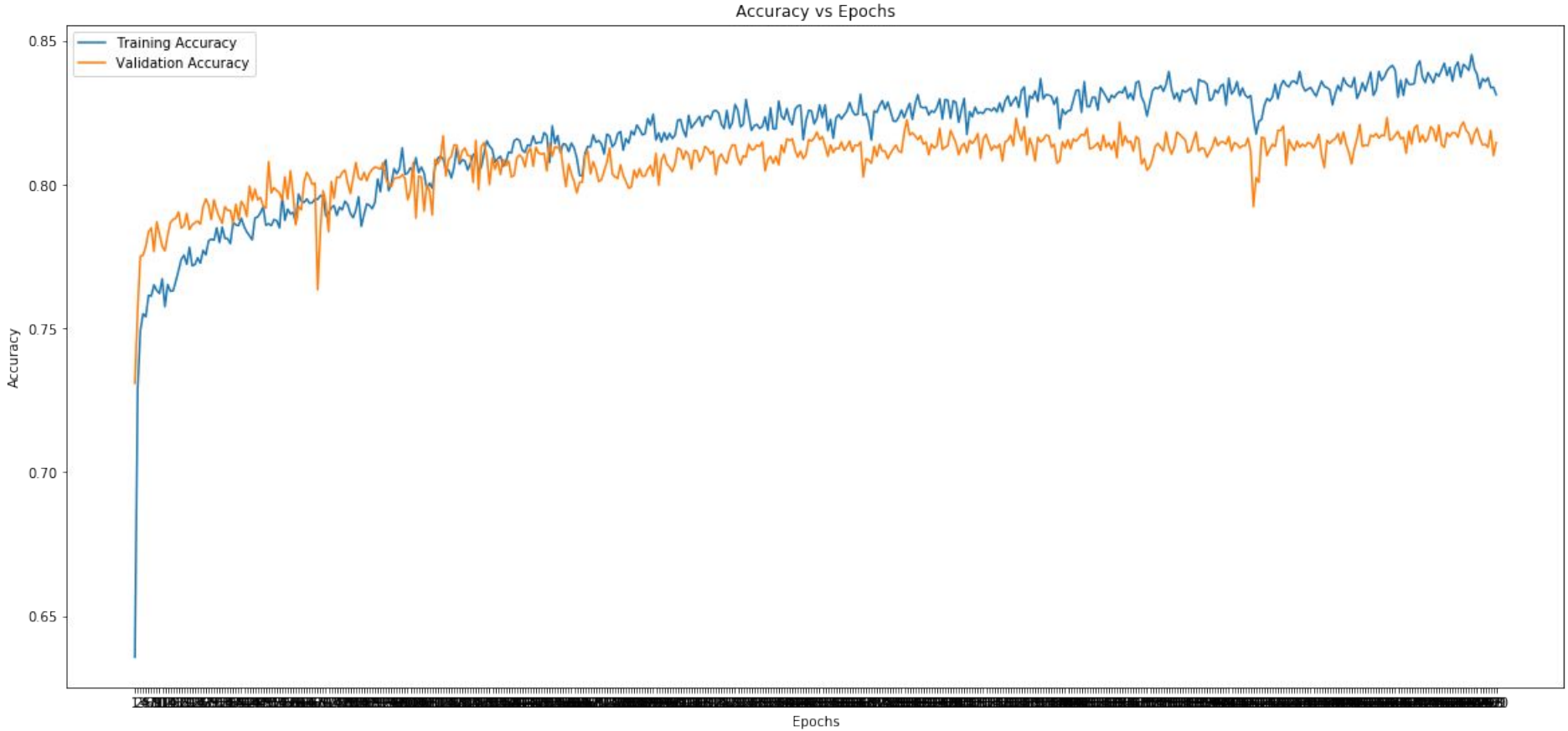
To classify with more accuracy there should be a method to analyze past as well as present body movements. Hence, Recurrent Neural Network is better and sensible option for the model.

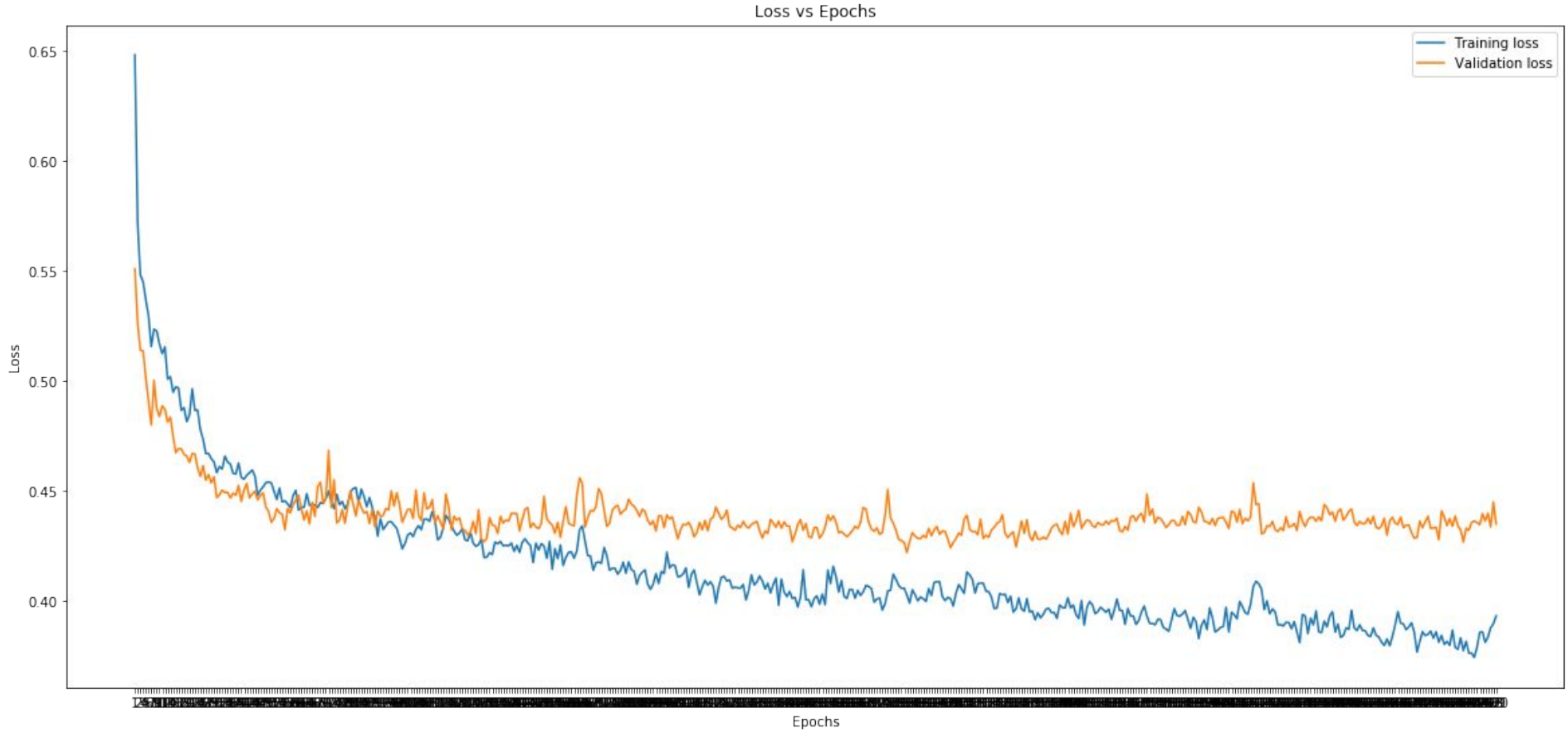
Model Architecture



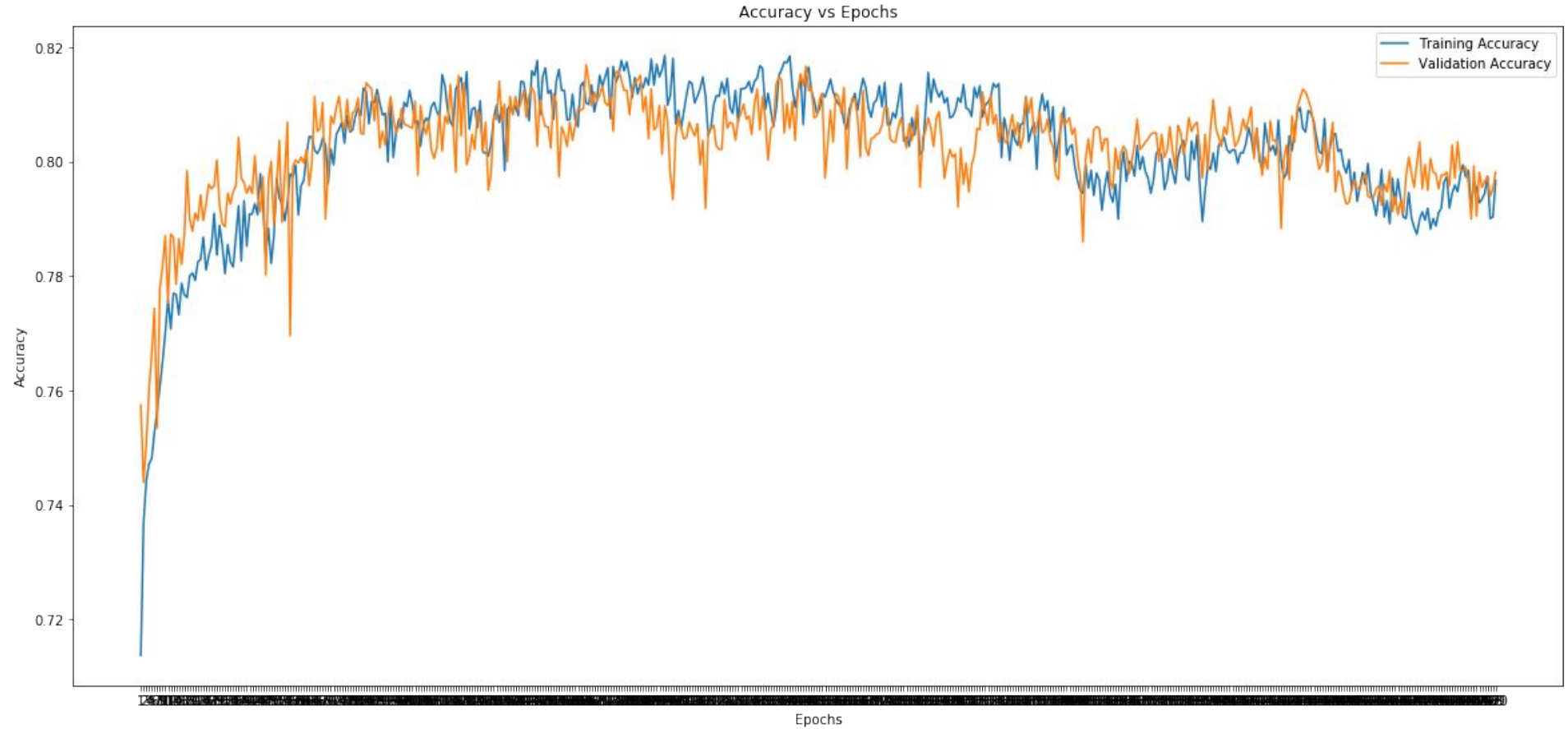
* 0 for Violence & 1 for Non-Violence

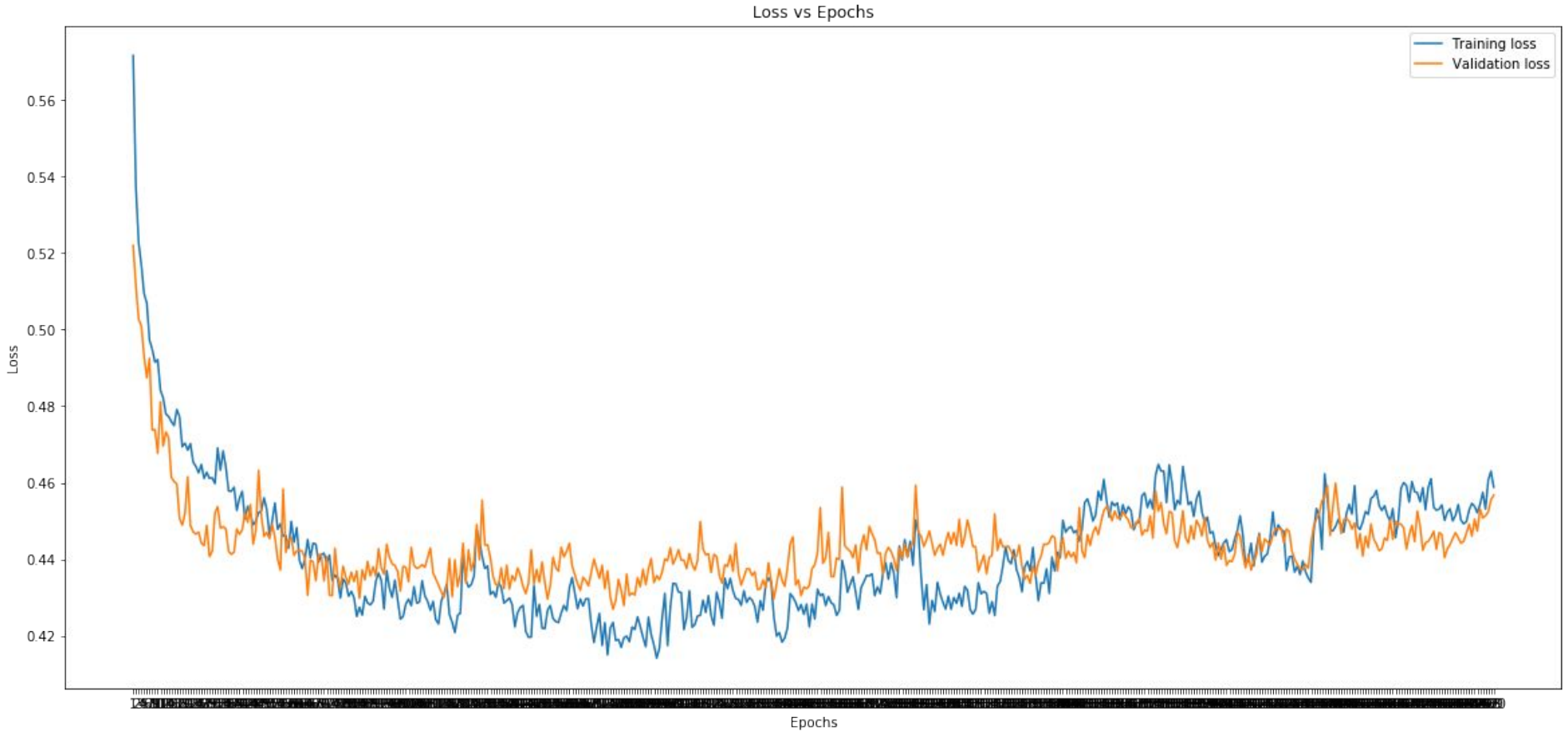
Results on 2-Layer deep GRU




Loss vs Epochs
Training loss
Validation loss
0.65
0.60
0.55
0.50
0.45
0.40
Loss
Epochs

Results on 3-Layer deep GRU




Loss vs Epochs
Training loss
Validation loss
0.56
0.54
0.52
0.50
0.48
0.46
0.44
0.42
Loss
Epochs

Results on Videos


10 N. ENT
Cyberlink
by PowerDirector


J2KV SharpView Camera 1
LM05 YNA
J2KV
19/10/07 14:51:56.2 6 IPS


by CyberLink PowerDirector