Update 02 Report

In previous update 01 we used YOLOv8 model.

In update 02 we used some CNN models along with SOTA (State-Of-The-Art) YOLO-NAS model.

CNN models we used:

- VGG16

- VGG19

- ResNet152V2

- InceptionV3

- MobileNetV2

- DenseNet201

Among those six models VGG16, VGG19 and MobileNetV2 passed to annotate after training process but ResNet152V2, InceptionV3, DenseNet201 these three models failed to annotate after training. We have test our annotation code to see if it is working correctly or not and several modifications and observation have been done while performing the training , validation, testing process.

Now let's discuss about VGG16, VGG19, MobileNetV2 performance:

Training process:

In the training process, we used SmoothL1Loss. As we are training our model for annotation bounding box compared to MSE, SmoothL1Loss performance on object detection annotation task is very well. Total 25 epoch is used. Where we observed the mean IOU (Intersection over Union). It is a commonly used metric to evaluate the performance of object detection and semantic segmentation models. IOU measures the overlap between the predicted bounding box or segmentation mask and the ground truth bounding box or mask. It is used to quantify how well the predicted object aligns with the actual object in the image.

Model Name	Highest training mean IoU	Highest valid mean IoU	Total training time(second)
VGG16	0.9304	0.8446	1640.69s
VGG19	0.9365	0.8452	1962.71s
MobileNetV2	0.9412	0.7811	1023.70s

Testing the model:

While testing the model we have observed that three of the model performance is average compared to previous update 01 YOLOv8 model. These models couldn't properly annotate violence and non-violence images or videos.

YOLO-NAS model:

Training process:

YOLO-NAS training process took a very long time compared to other mentioned model. Each epoch took 2:47 to 2:30 minutes and each epoch has validation time of 15-20 second. While training we used yolo_nas_s model. It is a small size model. A total number of 25 epoch along with 16 batch size and mAP@0.50 matric is observed and tensorboard is used to see the detail result.

Testing the model:

In testing we observed that it performed well compared to mentioned CNN models but the comparison between YOLOv8, it did not performed well. Performance of annotation and prediction has been observed by using supervision annotation and prediction side by side and same videos is used in order to compare.

YOLOv8 model:

In previous update 01 we trained YOLOv8 model. Which did very well compared to YOLO-NAS and Other mentioned CNN models.

Conclusion:

CNN models did not perform well for many reasons. Major reasons are annotation, dataset, and the task where we are using it. CNN model can perform well in classification task but when it comes to object detection it won't do well because of multiple object annotation. In our dataset multiple annotation is present which leads to poor performance of CNN models. On the other hand, YOLOv8, YOLO-NAS, these two models we used are specifically built to do this type of task easily. These can easily do the task of detection, classification, and segmentation. So, the best model till now for our violence and nonviolence task is YOLOv8.