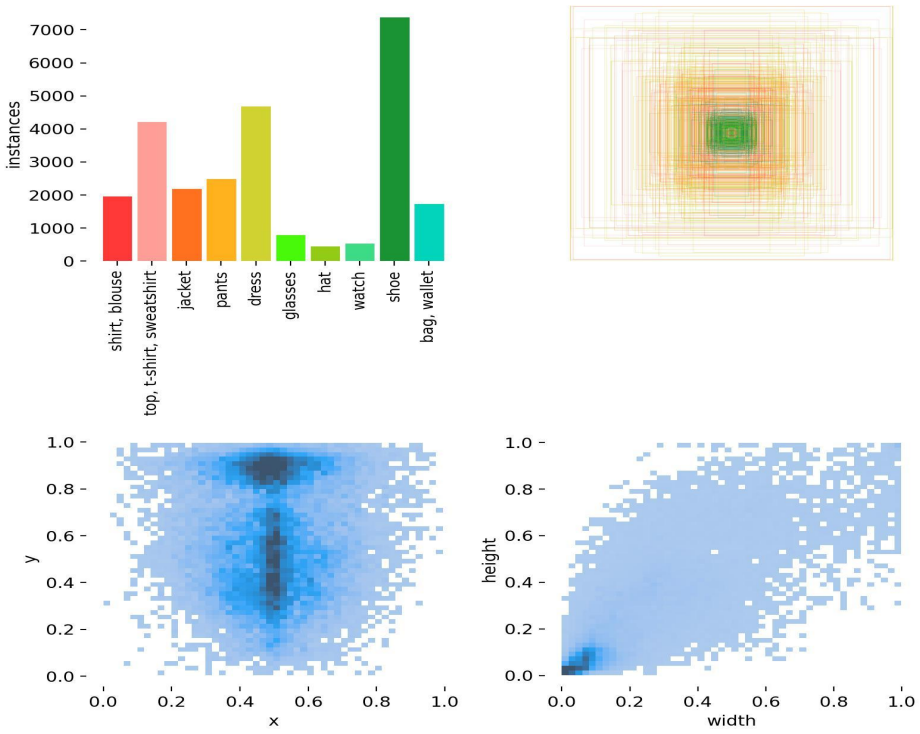
Model Comparison between RTMDet and YOLO v8

Dataset and annotations

Data used in these experiments is a subset of the Fashiopedia dataset. 10 classes are selected to be used in the training process.

Category	Counts
shirt, blouse	924
top, t-shirt, sweatshirt	2483
jacket	1183
pants	1899
dress	2816
glasses	727
hat	404
watch	516
shoe	6945
bag, wallet	1088

The data contains 6646 images for training and 972 images for validation. Data is a stratified sample of whole train data containing around 45000 images and 46 classes.



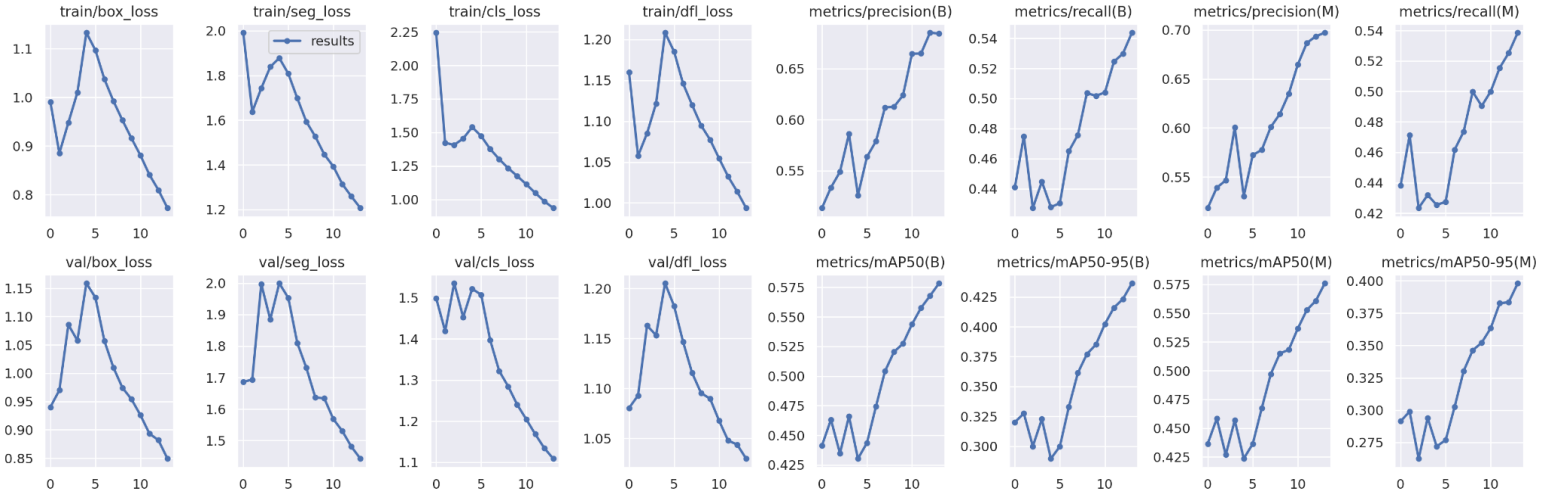
YOLO v8

YOLO-v8-small is used in these experiments , pretrained on coco and trained for 14 epochs.

- Batch-size : 12
- Image size : 640
- Optimizer : SGD (lr =0.01 , momentum = 0.973, weight decay = 0.0005)
- Some default augmentations (random flip, random resize,etc)
- Albumentation augmentations after epoch 8 : CLAHE , Blur, MedianBlur

Results after 14 epochs (best checkpoint) :

Box(P	R	mAP50	mAP50-95)	Mask(P	R	mAP50	mAP50-95):
0.684	0.544	0.578	0.436	0.697	0.539	0.576	0.398



Class	Images	Instances	Box(P	R	mAP50	mAP50-95)	Mask(P	R	mAP50	mAP50-95):
all	1455	5816	0.684	0.546	0.578	0.436	0.697	0.538	0.576	0.398
shirt, blouse	1455	446	0.529	0.271	0.335	0.256	0.552	0.267	0.337	0.225
top, t-shirt, sweatshirt	1455	949	0.644	0.414	0.464	0.368	0.647	0.404	0.464	0.356
jacket	1455	512	0.608	0.438	0.467	0.374	0.61	0.426	0.463	0.361
pants	1455	506	0.736	0.694	0.71	0.63	0.744	0.694	0.711	0.626
dress	1455	1077	0.669	0.497	0.533	0.434	0.677	0.496	0.533	0.434
glasses	1455	157	0.871	0.758	0.765	0.543	0.898	0.758	0.757	0.467
hat	1455	96	0.764	0.844	0.851	0.658	0.786	0.833	0.829	0.57
watch	1455	109	0.65	0.413	0.464	0.273	0.631	0.393	0.451	0.231
shoe	1455	1609	0.732	0.699	0.737	0.522	0.75	0.694	0.735	0.41
bag, wallet	1455	355	0.635	0.428	0.458	0.305	0.675	0.42	0.479	0.302

Inference time :

Inference time was inconsistent in different runs but as average it is around 25 ms on gpu as pytorch model (not with tensorRT) for single inference.

And Dataset inference time is 0.4ms preprocess, 4.4ms inference, 0.0ms loss, 2.5ms post-process per image.

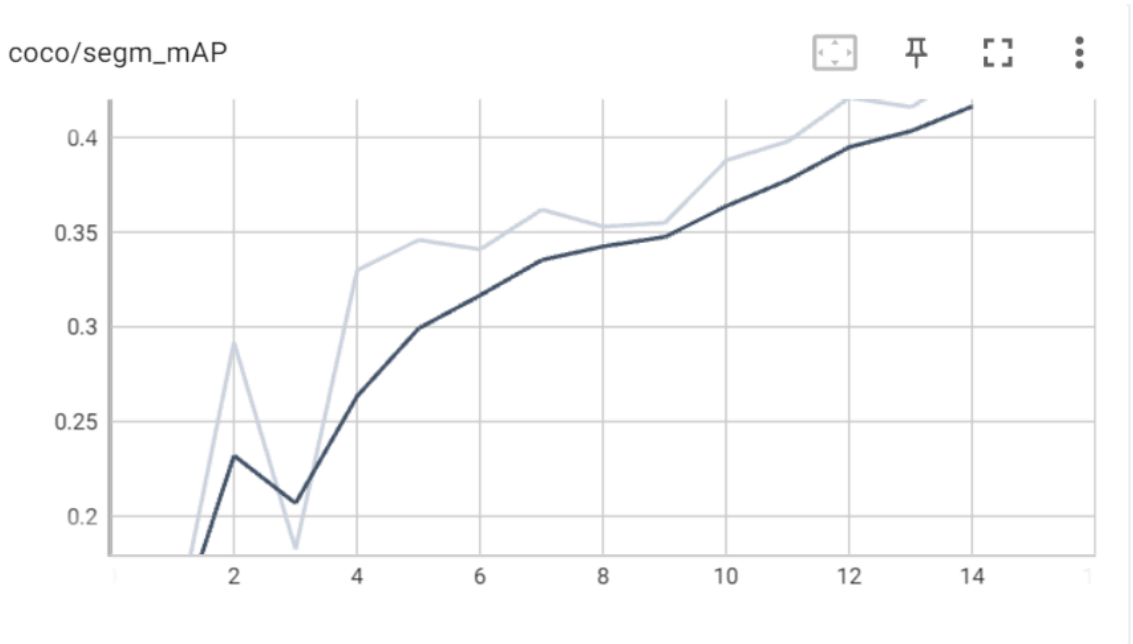
RTMDet

RTMDet-small is used in these experiments. Pretrained on coco dataset and trained for 14 epochs.

- Batch-size : 12
- Image size : 640
- Optimizer : AdamW(lr = 0.0005 , weight decay=0.05)
- Some default augmentations (random flip, random resize,etc)
- Post-processing changed to have a fair comparison with yolo

Results for segmentation after 14 epochs (best checkpoint) :

```
Average Precision  (AP) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.436
Average Precision  (AP) @[ IoU=0.50      | area=   all | maxDets=100 ] = 0.658
Average Precision  (AP) @[ IoU=0.75      | area=   all | maxDets=100 ] = 0.454
Average Precision  (AP) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.073
Average Precision  (AP) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.288
Average Precision  (AP) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.577
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=  1 ] = 0.538
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets= 10 ] = 0.607
Average Recall     (AR) @[ IoU=0.50:0.95 | area=   all | maxDets=100 ] = 0.612
Average Recall     (AR) @[ IoU=0.50:0.95 | area= small | maxDets=100 ] = 0.155
Average Recall     (AR) @[ IoU=0.50:0.95 | area=medium | maxDets=100 ] = 0.439
Average Recall     (AR) @[ IoU=0.50:0.95 | area= large | maxDets=100 ] = 0.750
```



Inference time :

Inference comparison :

RTMDet classes in the config file should be ordered by their id in annotations file but I didn’t know that and class names are false.

RTMDet :

YOLO v8 :