Pruning Implementation

Created on : 2024-06-14
tags: []
References:

Ultralytics prune yolov5 : https://docs.ultralytics.com/yolov5/tutorials/model_pruning_and_sparsity/

Pruning tutorial PyTorch: https://pytorch.org/tutorials/intermediate/pruning_tutorial.html

https://arxiv.org/pdf/2204.13699

Modoptima : https://github.com/VikasOjha666/modoptima
Sparse ML

Performance of Tiny YOLO on Wider Face with Pruning

Model Version	mAP	Parameters	Description
Unpruned	0.123	4.70 M	Baseline model without pruning, used as a reference point.
Pruned TinyYOLO	0.10	3.00 M	Pruned model, showing a reduction in size with some accuracy loss.

Performance of YOLOv5 Small on Pascal VOC with Pruning

Magnitude Based Pruning(Structured)

Pruning Level	mAP50	Parameters	epochs	Description
0% Prune	0.74	7.07 M	3	Baseline model with no pruning applied.
25% Prune	0.664	5.30 M	2	Moderate pruning with a slight drop in accuracy.
50% Prune	0.3	2.47 M	3	Significant pruning leading to a major accuracy drop. (retrained 3 epochs)

Total Prune per: 70% of channels, iterative steps = 6

Pruning Level	mAP50	Parameters	epochs	Description
0% Prune	0.792	7.07 M	6	Baseline model with no pruning applied. Finetuned on pascal VOC.
21% Prune	0.79	5.53 M	3	Moderate pruning with a almost no loss accuracy.
36% Prune	0.71	4.46 M	4	Some drop in accuracy of the model
52% Prune	0.66	3.50 M	5	
60% Prune	0.47	2.68M	6	
70% prune	0.25	2.04M	7	Sharp drop in accuracy
	0.08	1.57M	8	Very Sharp drop in accuracy.

Magnitude based pruning (Group Regularization Pruner)

Total Prune per: 70% of channels, iterative steps = 6

Pruning Level	mAP50	Parameters	epochs	Description
0% Prune	0.792	7.07 M	6	Baseline model with no pruning applied. Finetuned on pascal VOC.
21% Prune	0.76	4.20 M	3	Moderate pruning with a almost no loss accuracy.
36% Prune	0.72	4.46 M	4	Some drop in accuracy of the model
52% Prune	0.65	3.0 M	5	
60% Prune	0.38	2.05M	6	
70% prune	0.12	1.27M	7	Sharp drop in accuracy
	0.08	0.70M	8	Very Sharp drop in accuracy.

Vitis AI optimizer (Pruning)

- Fine grained - Unstructured
- Coarse Based Pruning - Structured Pruning

Recommendations

- Use as much data as possible to perform model analysis. Ideally, you should use all the data in the validation dataset, but this can be time-consuming. You can also use partial validation set data to ensure that at least half of the dataset is used.
- During the fine-tuning stage, experiment with a few hyperparameters, including the initial learning rate and the learning rate decay policy. Use the best result as the input for the next iteration.
- The data used in fine-tuning should be a subset of the original dataset used to train the baseline model.
- If the accuracy does not improve sufficiently after several fine-tuning experiments, try reducing the pruning rate and re-run pruning and fine-tuning.