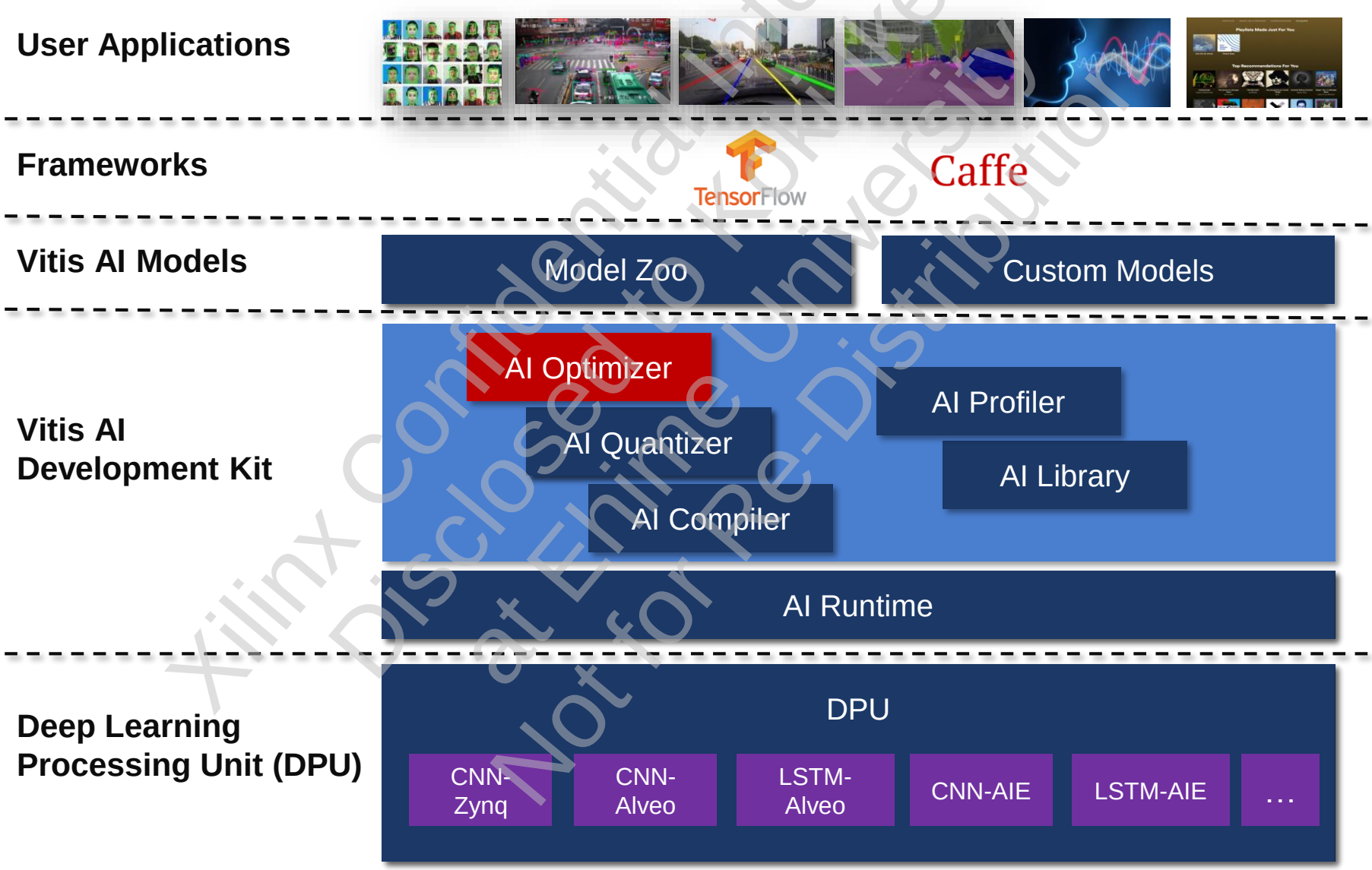
XILINX

Vitis AI Optimizer

XILINX

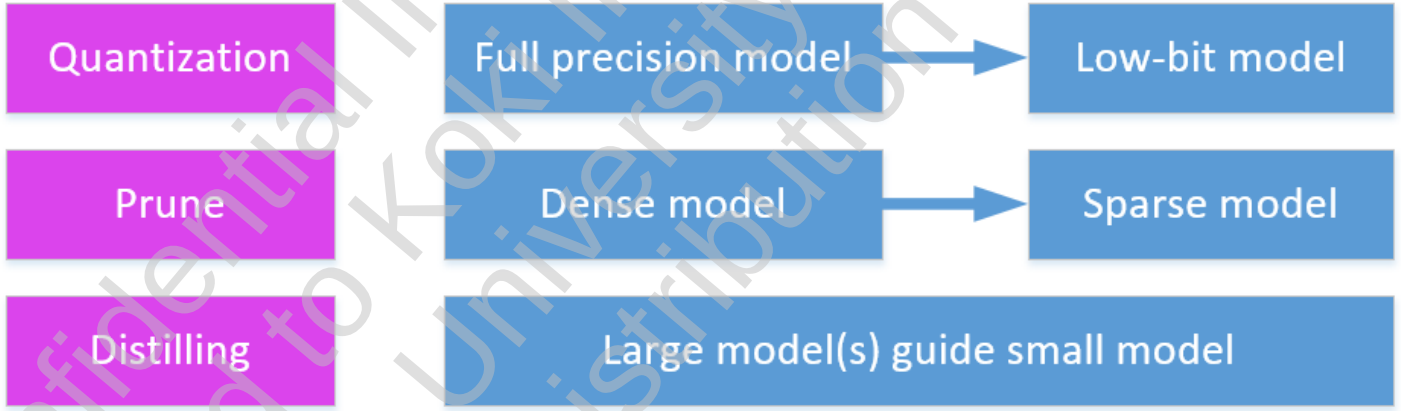
Overview

AI Optimizer in Vitis AI



Compression

- Quantization
- Pruning
- Distilling



- Designed Small Networks
  - MobileNets / SqueezeNet / ShuffleNet …

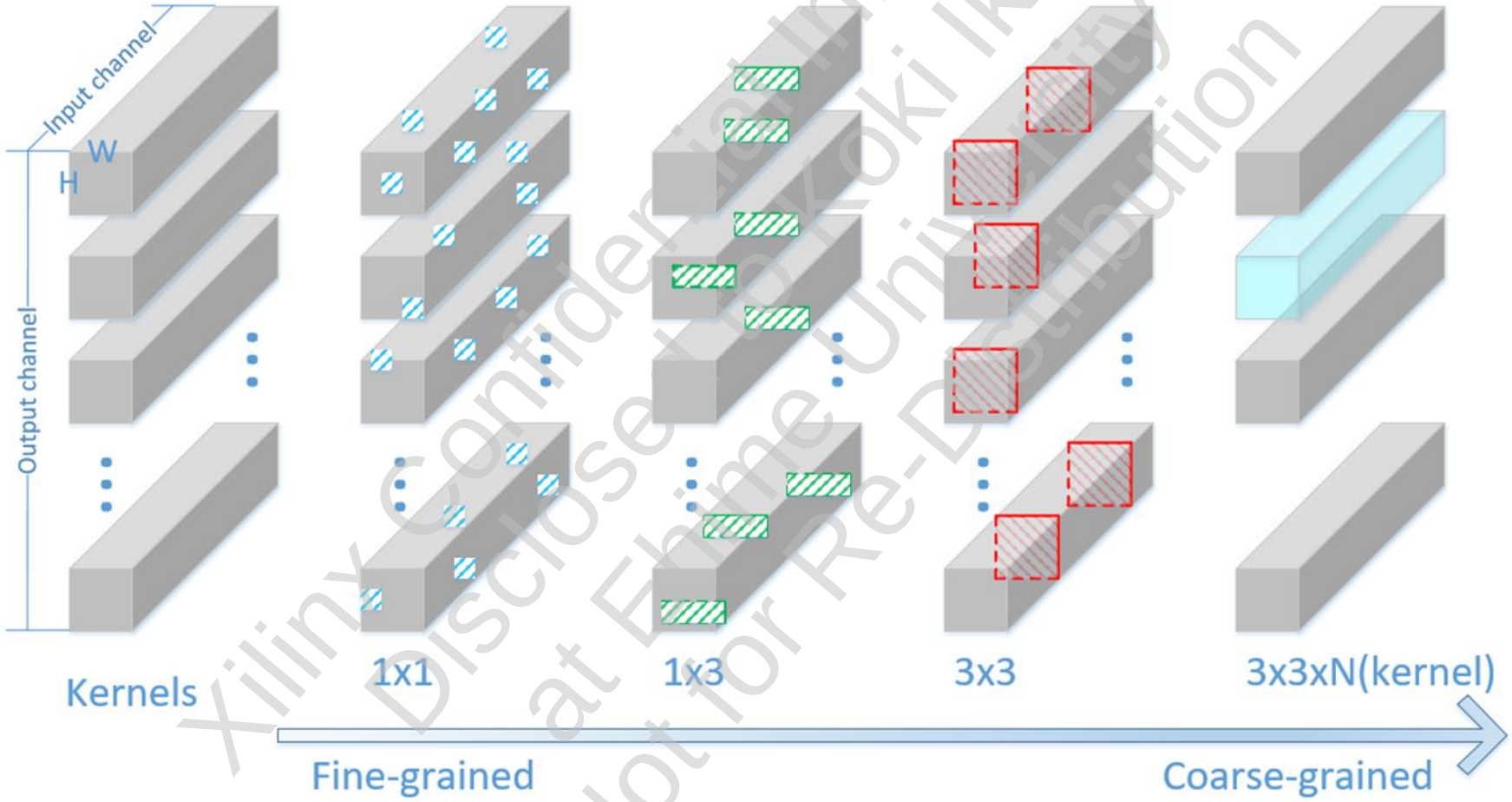
AI Optimizer - Pruning

- Benefits
  - Reduce model size 5 – 100x
  - Reduce running time 1.5 – 10x
- Supported framework
  - Caffe, Darknet, Tensorflow

XILINX

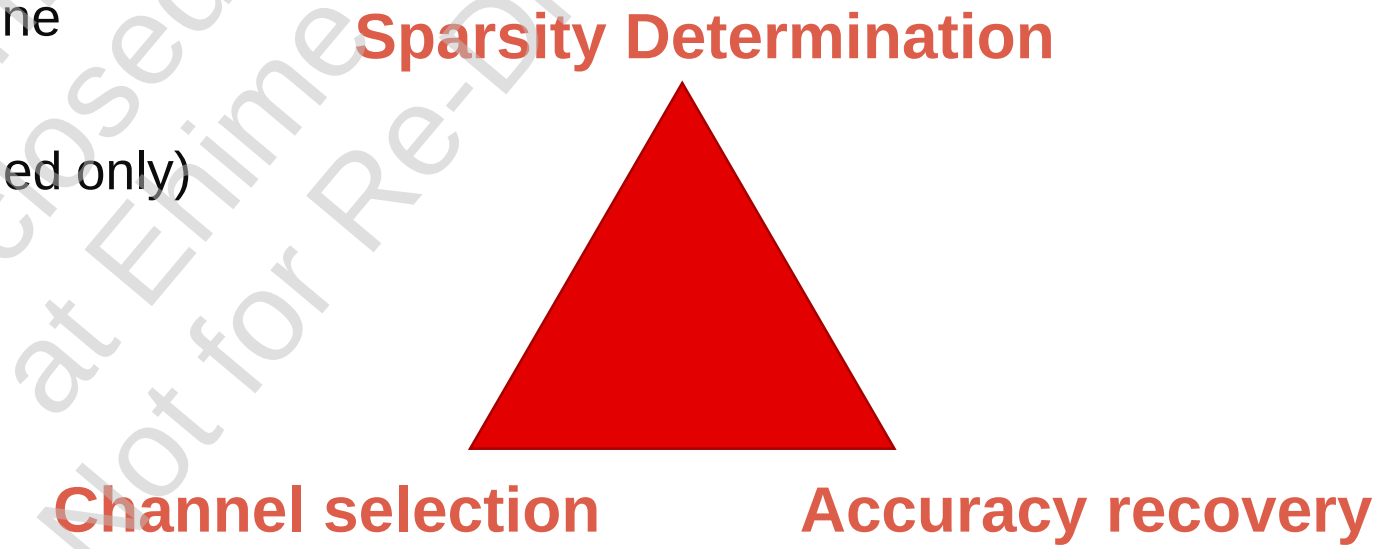
Pruning

Pruning from fine-grained to coarse-grained



Three aspects in coarse-grained pruning

- Sparsity Determination
  - How many flops/weights to prune in each layer
  - Sensitive analysis
  - Automated Deep Compression by reinforcement learning
- Channel selection
  - In a layer, which channel to prune
  - Different criteria
  - Lasso regression (coarse-grained only)
- Accuracy recovery
  - How to recover accuracy
  - Calibration
  - Finetune



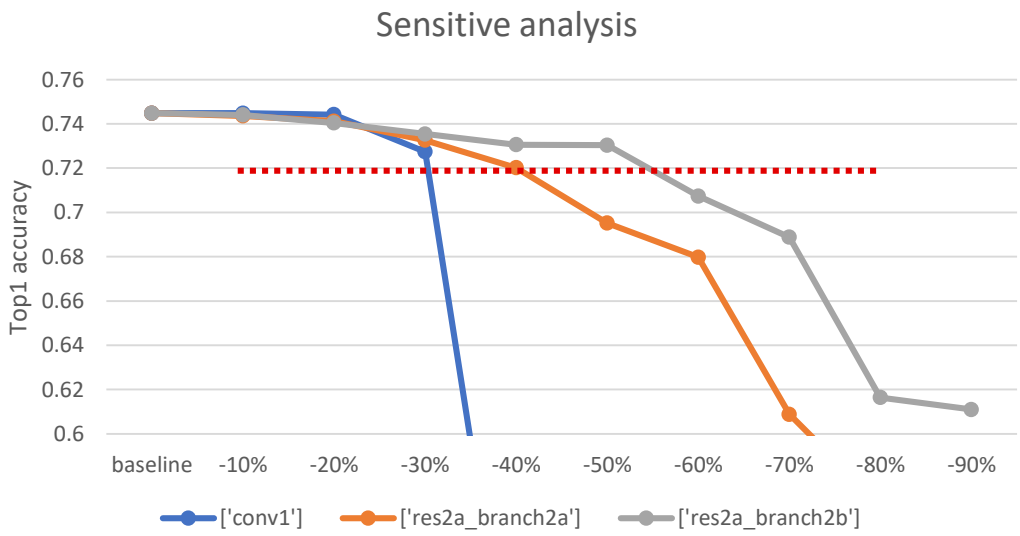
Three aspects in coarse-grained pruning

- Sparsity Determination
  - How many flops/weights to prune in each layer
  - Sensitive analysis / AutoML

Resnet-50	baseline	-10%	-20%	-30%	-40%	-50%	-60%	-70%	-80%	-90%
['conv1']	0.7448	0.7448	0.7442	0.7274	0.4662	0.0184	0.0032	0.0022	0.0014	0.0012
['res2a_branch2a']	0.7448	0.7436	0.7412	0.7326	0.7202	0.6952	0.6798	0.6088	0.5678	0.5358
['res2a_branch2b']	0.7448	0.744	0.7404	0.7354	0.7306	0.7304	0.7074	0.6888	0.6164	0.611
['res2b_branch2a']	0.7448	0.745	0.7416	0.7404	0.7368	0.7328	0.7286	0.724	0.7144	0.6546
['res2b_branch2b']	0.7448	0.7422	0.7396	0.738	0.7372	0.7322	0.7302	0.7158	0.6982	0.67
['res2c_branch2a']	0.7448	0.7438	0.741	0.7354	0.729	0.7228	0.7088	0.6742	0.635	0.588

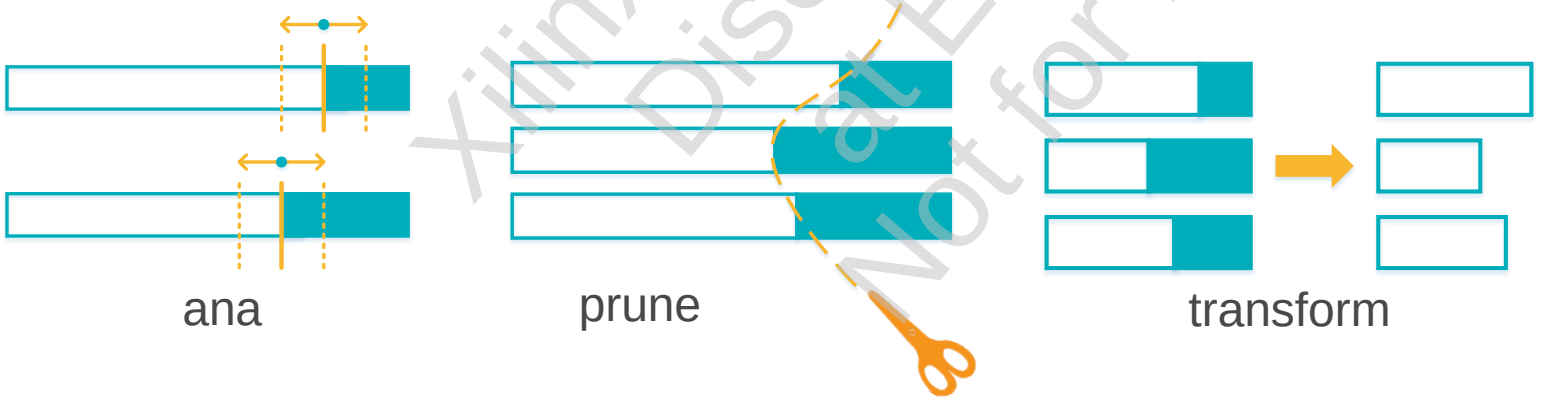
th=0.01

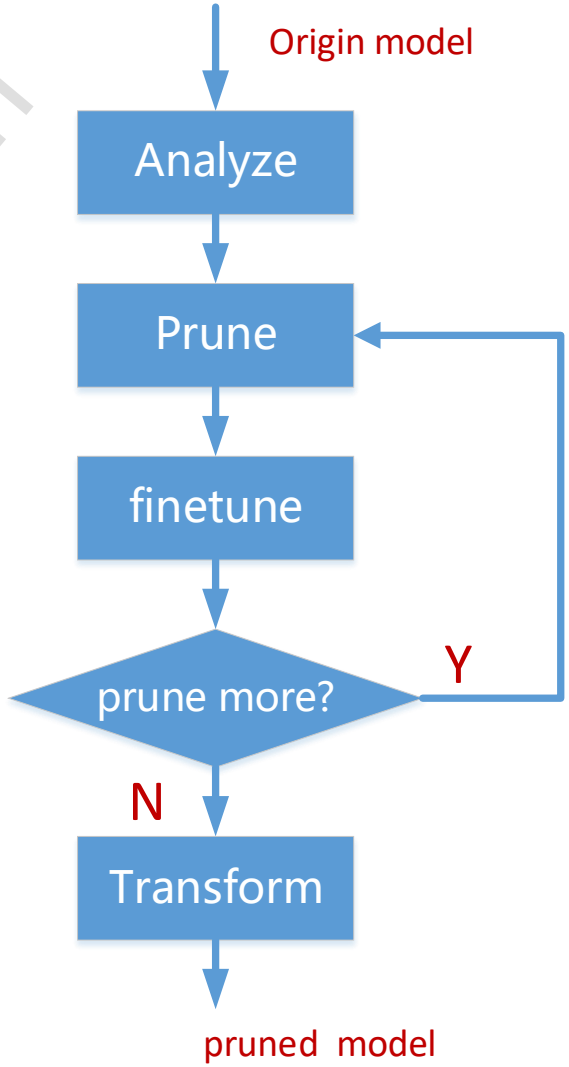
- Channel selection
  - In a layer, which channel to prune
  - Different criteria (L2norm, abs*gradient) / Lasso regression
- Accuracy recovery
  - How to recover accuracy
  - Finetune / Calibration



Three aspects in coarse-grained pruning (Cont)

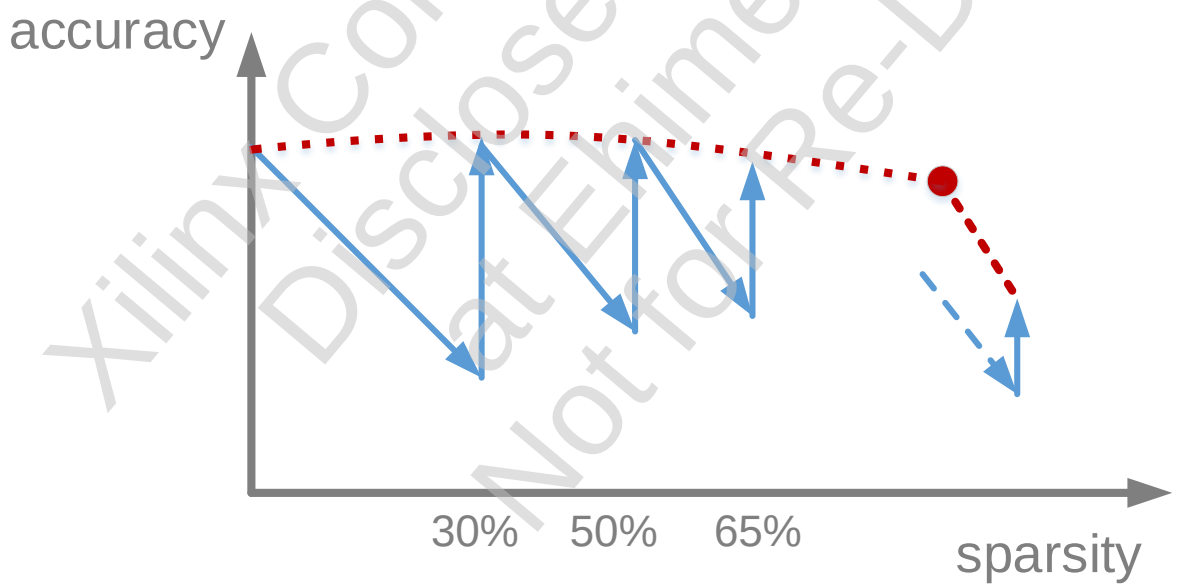
- vai_p
  - Automatically prune the network models to desired sparsity
  - Significantly reduce the Ops and parameters of CNN models without losing much accuracy
- 5 commands in vai_p
  - ana – run sensitivity analysis
  - prune – prune the network according to config
  - finetune – finetune the network to recovery accuracy
  - transform – transform the pruned model to regular model
  - stat – get flops and the number of parameters of a model





Iterative Pruning

- Iterative pruning
  - Iterative pruning generally works better than pruning directly
  - When facing accuracy drops, try a small step
  - For a full pruning work, set steps from large to small
  - Experience on model training is very helpful
- A typical pruning process



Pruning Results - Classification

- Classification networks on ImageNet

Classification Networks	Framework	Operations	Input size	Baseline	Pruning Result 1			Pruning Result 2		
				Top-1 / Top-5	Top-1 / Top-5	Δtop-1/5	ratio	Top-1 / Top-5	ΔTop1/5	ratio
VGG-16	TensorFlow	31G	224x224	70.89 / 89.84	70.80 / 89.97	-0.09 / +0.13	69%	69.29 / 89.04	-1.60 / -0.80	55%
Resnet-50	TensorFlow	7.0G	224x224	75.20 / 92.20	74.15 / 91.76	-1.05 / -0.44	47%	72.99 / 91.16	-2.21 / -1.04	33%
Inception_v3	TensorFlow	11.4G	299x299	78.00 / 93.93	77.34 / 93.46	-0.64 / -0.47	65%	76.68 / 93.17	-1.32 / -0.76	60%
VGG-16	Caffe	30G	224x224	70.96 / 89.84	70.20 / 89.70	-0.76 / -0.14	50%	69.12 / 89.13	-1.84 / -0.71	33%
Resnet-50	Caffe	7.7G	224x224	74.83 / 92.14	73.89 / 91.68	-0.94 / -0.42	49%	73.11 / 91.14	-1.72 / -1.00	34%
Inception_v1	Caffe	3.2G	224x224	70.33 / 89.67	69.53 / 89.16	-0.80 / -0.51	80%	68.64 / 88.58	-1.69 / -1.09	72%
Inception_v2	Caffe	4.0G	224x224	72.78 / 91.07	72.30 / 90.85	-0.48 / -0.22	70%	71.17 / 90.18	-1.61 / -0.89	60%
SqueezeNet	Caffe	778M	224x224	61.41 / 83.19	60.73 / 82.46	-0.68 / -0.73	89%	59.38 / 81.57	-2.03 / -1.62	75%

Pruning Results - Detection & Segmentation

Detection Networks	Framework	Operations	Dataset	Input size	Baseline mAP	Pruning Result 1			Pruning Result 2		
						mAP	ΔmAP	ratio	mAP	ΔmAP	ratio
SSD+VGG	TensorFlow	70G	VOC 21 classes	300x300	79.3%	79.0%	-0.3%	36%	77.0%	-2.3%	20%
SSD+VGG	Caffe	117G	Deephi 4 classes	480x360	61.5%	62.0%	+0.5%	16%	60.4%	-1.1%	10%
SSD+VGG	Caffe	173G	Cityscapes 4 classes	500x500	57.1%	58.7%	+1.6%	40%	56.6%	-0.5%	12%
SSD+InceptionV2	Caffe	10G	VOC 19 classes	320x320	74.6%	73.8%	-0.8%	55%	73.2%	-1.4%	45%
YoloV2	Darknet	34G	VOC 21 classes	448x448	78.5%	77.0%	-1.5%	34%	76.7%	-1.8%	29%
YoloV3 variant	Darknet	53.7G	Cityscapes 3 classes	512x256	53.7%	56.1%	+2.4%	26%	55.2%	+1.5%	11%

Segmentation Networks	Framework	Operations	Dataset	Input size	Baseline mIoU	Pruning Result 1			Pruning Result 2		
						mIoU	ΔmIoU	ratio	mIoU	ΔmIoU	ratio
FPN	Caffe	163G	Cityscapes 19 classes	1024x2048	65.69%	65.21%	-0.48%	80%	64.07%	-1.62%	60%

XILINX

Pruning Usage & Examples

vai_p Usage

- Sensitivity analysis, searching for best pruning strategy
- Prune the network iteratively until the network's sparsity meets your needs
- Using the pruning tool
  - Prune the network using a certain pruning ratio
  - Fine-tune the pruned network
  - Increase pruning ratio
  - Prune again
  - Loop the pruning-retraining iteration for a few times
  - Generate a final dense model

vai_p_caffe Usage

```
workspace: "examples/"
gpu: "0,1,2,3"
test_iter: 2500
acc_name: "top-5"

model: "examples/vai_p/float.prototxt"
weights: "examples/vai_p/float.caffemodel"
solver: "examples/vai_p/solver.prototxt"

rate: 0.1

pruner {
  method: REGULAR
}
```

vai_p_caffe Usage (Cont)

- Run sensitivity analysis (may take several hours to ten hours depending on the size of network)

```
vai_p_caffe ana -config config.prototxt
```

- Start pruning

```
vai_p_caffe prune -config config.prototxt
```

- Fine-tune the pruned model

```
vai_p_caffe finetune -config config.prototxt
```

- Transform the pruned (sparse) model to a normal (dense) model

```
vai_p_caffe transform -model baseline.prototxt -weights pruned.caffemodel
```

For more details, see UG1333

vai_p_tensorflow Usage

- Export an Inference Graph
  - evaluation codes are used to export an inference GraphDef file and evaluate network's performance during pruning.
- Run sensitivity analysis (may take several hours to ten hours depending on the size of network)

```
vai_p_tensorflow --action=ana …
```

- Start pruning

```
vai_p_tensorflow --action=prune …
```

- Fine-tune the pruned model
  - Training model with the generate checkpoint file generated in the previous fine-tune process and increased value of --sparsity flag
- Transform the pruned (sparse) model to a normal (dense) model

```
vai_p_tensorflow --action=transform …
```

- Freezing Graph
  - Get the final output file named frozen.pb which can be used for inference

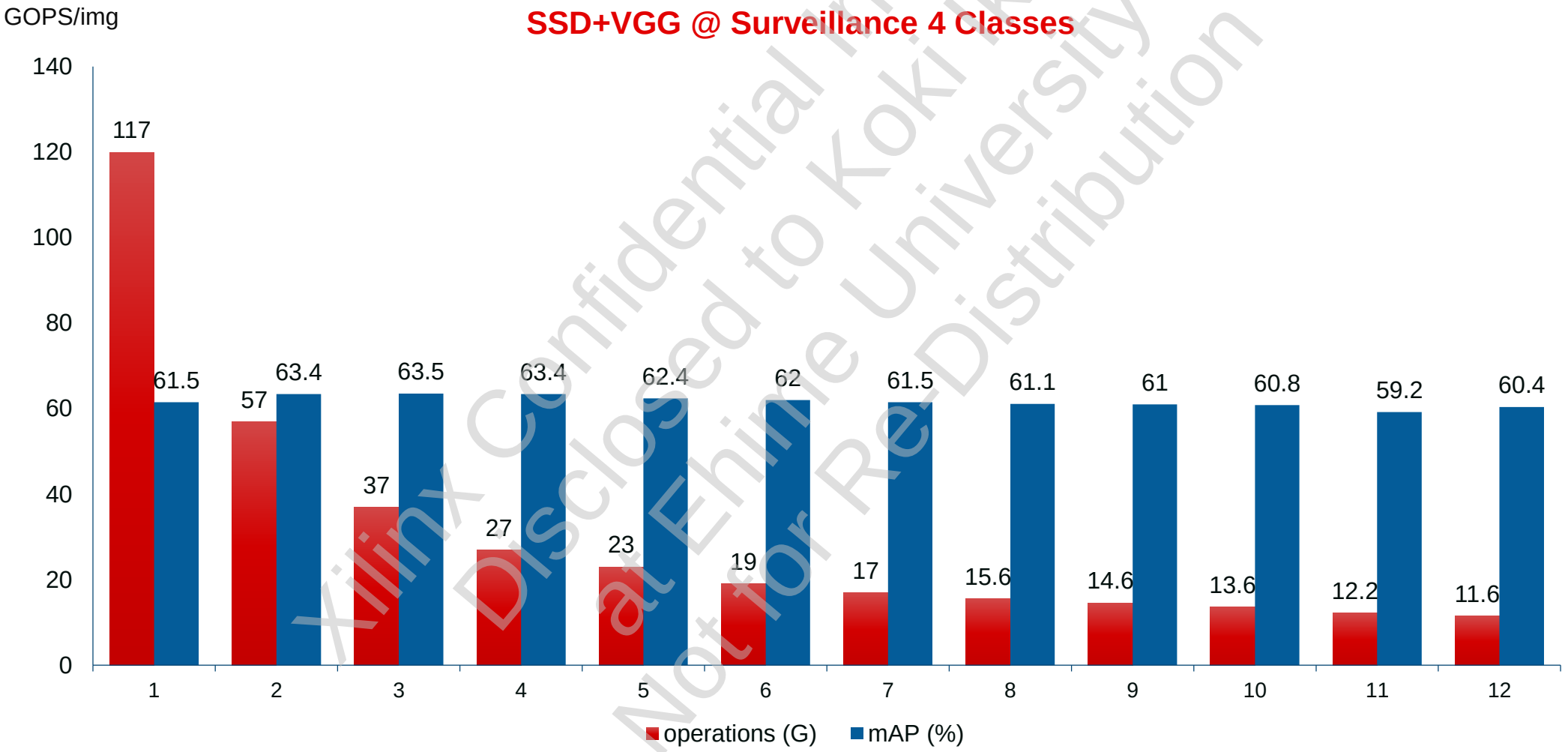
For more details, see UG1333

vai_p_darknet Usage

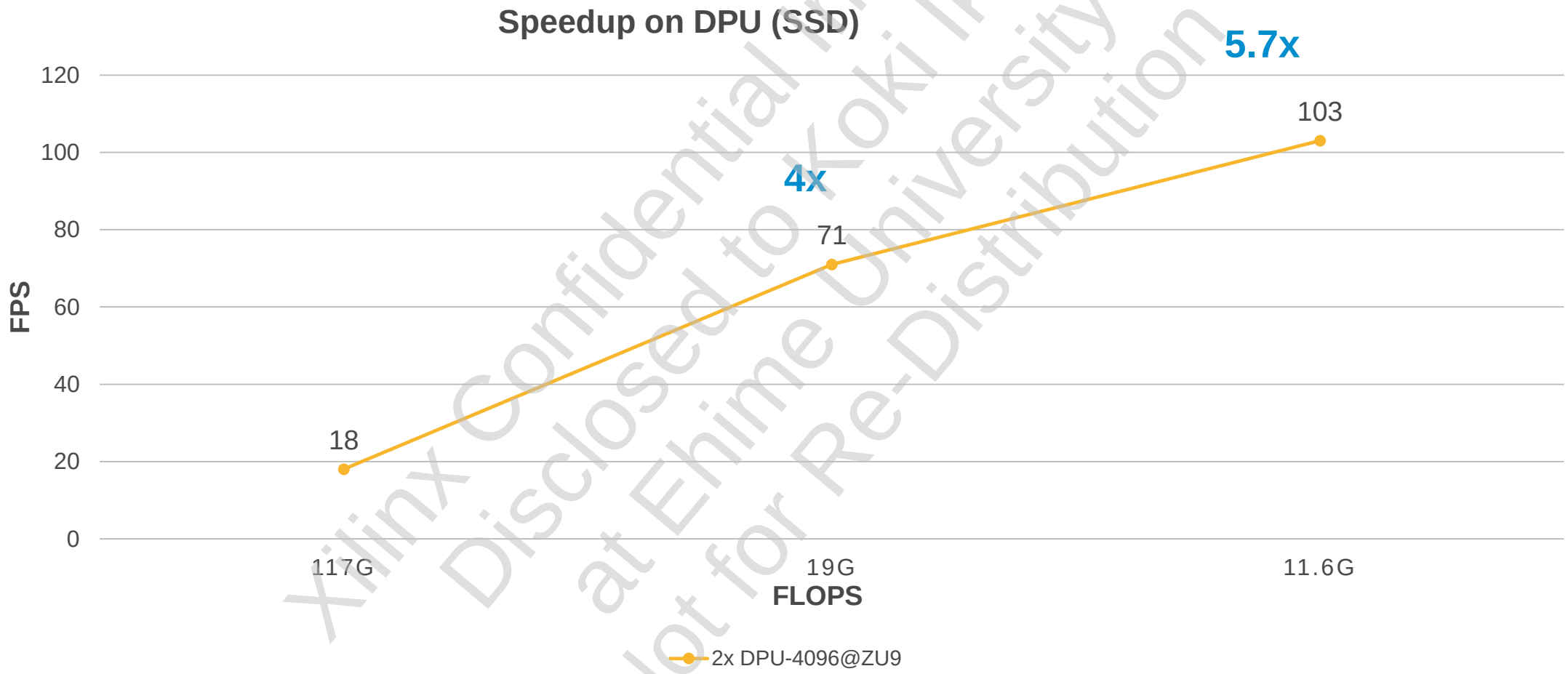
- Create a configuration file
  - A full list of main cfg options can be found in UG1333
- Run sensitivity analysis (may take several hours to ten hours depending on the size of network)
  - vai_p_darknet pruner ana pruning/cfg pruning/weights
- Start pruning
  - vai_p_darknet pruner prune pruning/cfg pruning/weights
- Fine-tune the pruned model
  - vai_p_darknet pruner finetune pruning/cfg
- Transform the pruned (sparse) model to a normal (dense) model
  - vai_p_darknet pruner transform pruning/cfg backup/*_final.weights
- Model Evaluation
  - vai_p_darknet pruner valid pruning/cfg weights.transform

For more details, see UG1333

Pruning Example for Embedded



Pruning Example for Embedded (Cont)



Pruning Example for Alveo

Detection(VOC dataset) - 416x416

Model	Ops	Weights	Model size	After finetune	After quantization	U200 fps - batch=1	U250 fps - batch = 1
Standard yolo v2	1	1	194M	76.95	75.99	250	334
Compress 22G	0.34	0.28	56M	75.32	74.6	703	937
Compress 24G	0.29	0.26	51M	75.1	74.1	901	1201
Compress 26G	0.23	0.19	37M	73.9	73.34	***1073***	***1431***

XILINX

Availability

Availability

- Available on Now on AI Optimizer Lounge
  https://www.xilinx.com/member/ai_optimizer.html
  - Both node-locked and floating license available

Documents & Files

Description	File	Version	Filesize	Checksum
Xilinx AI Optimizer Pruning Tool	vai_p.tar.gz	1.0	309 MB	2b8be4fdd1d5c18abbe22c63b313bb6c
Xilinx AI Optimizer User Guide	ug1333-ai-optimizer_WtMkX.pdf	1.0	--	--
Xilinx AI Optimizer Introduction	Xiinx_AI_Optimizer_intro_WtMkX.pdf	--	--	--

- Pricing
  - Please contact your Xilinx representative

XILINX

Support

Support

- Direct support from Xilinx
- Assign the support window from Xilinx – customer assign their contact window
  - Regular review meeting (if necessary)
  - Email/Skype/Webex on demand
- Provide guidance and support to set up the environment
- Provide pruning examples to get started
- On-site training (if necessary)
- Support pruning with custom dataset

Thank You