Current Research Progress

Advisor: Dr. Chih-Yu Wang
Presenter: Shao-Heng Chen
Date: August 17, 2022

1. Previous reproduction summary

(1) Simulation environment

* Hardware Specifications
  - CPU. Intel i7-8700
  - GPU. Geforce RTX 2060
  - RAM. DDR5-3600 32GB

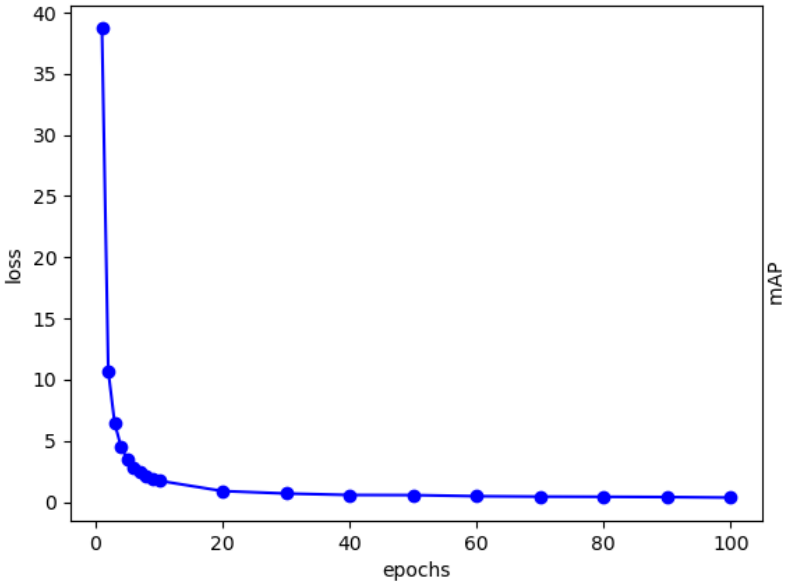
* Software Versions
  - python 3.7.13
  - torch 1.12.0
  - torchvision 0.13.0

(2) Simulation results

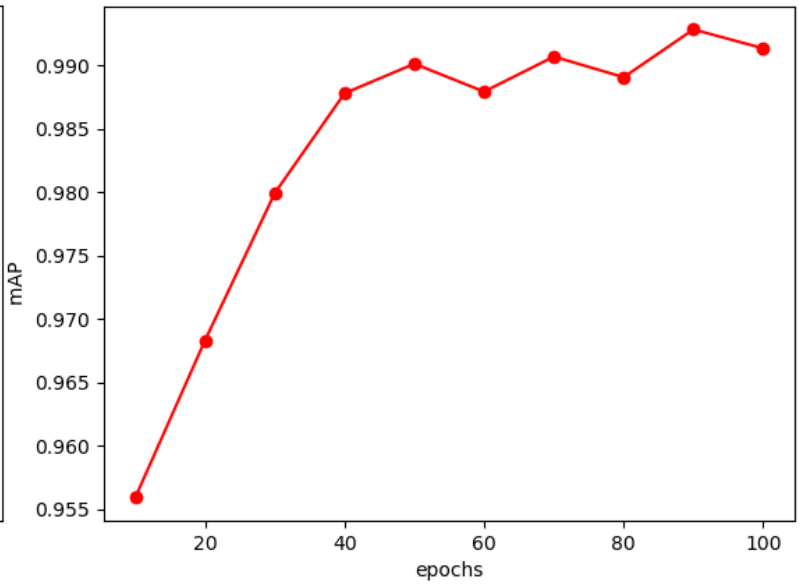
- Inference speed = 8*16 / 3 = 128 / 3 = 42.66 fps

```
On Test loader:
100%|██████████| 7/7 [00:03<00:00, 1.81it/s]
Class accuracy is: 100.000000%
No obj accuracy is: 73.054192%
Obj accuracy is: 100.000000%
```

- Training loss



- mAP (mean Average Precision)



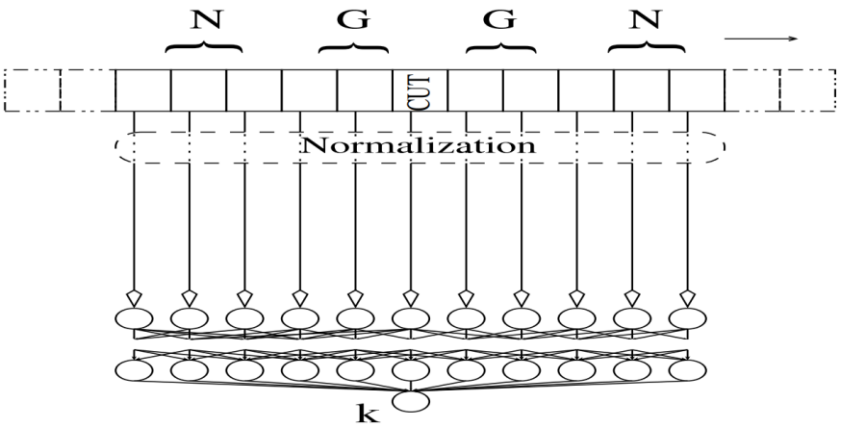
2. CFAR related works

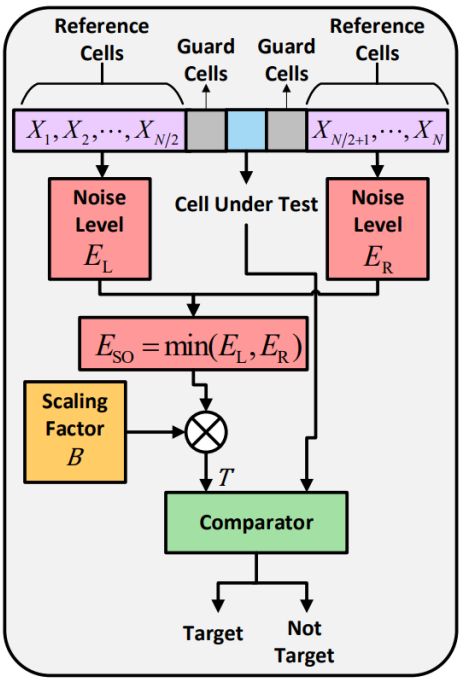
(1) J. Akhtar, "Training of Neural Network Target Detectors Mentored by SO-CFAR," *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 1522-1526

(2) J. Akhtar and K. E. Olsen, "GO-CFAR Trained Neural Network Target Detectors," *2019 IEEE Radar Conference (RadarConf)*, 2019, pp. 1-5

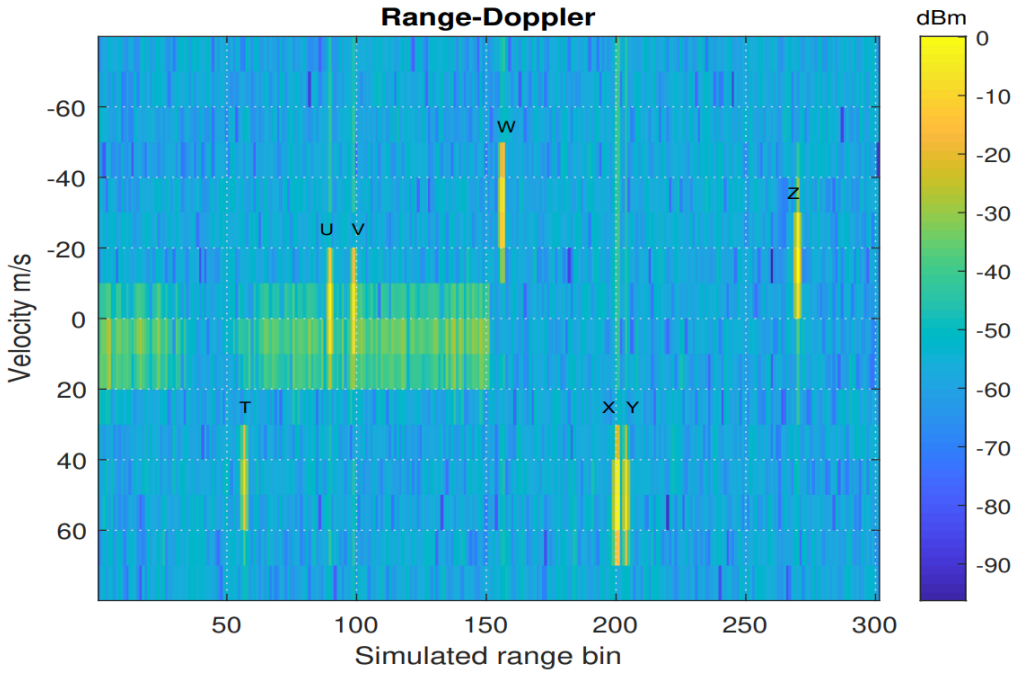
(3) J. Akhtar and K. E. Olsen, "A Neural Network Target Detector with Partial CA-CFAR Supervised Training," *2018 International Conference on Radar (RADAR)*, 2018, pp. 1-6

- Neural network detector

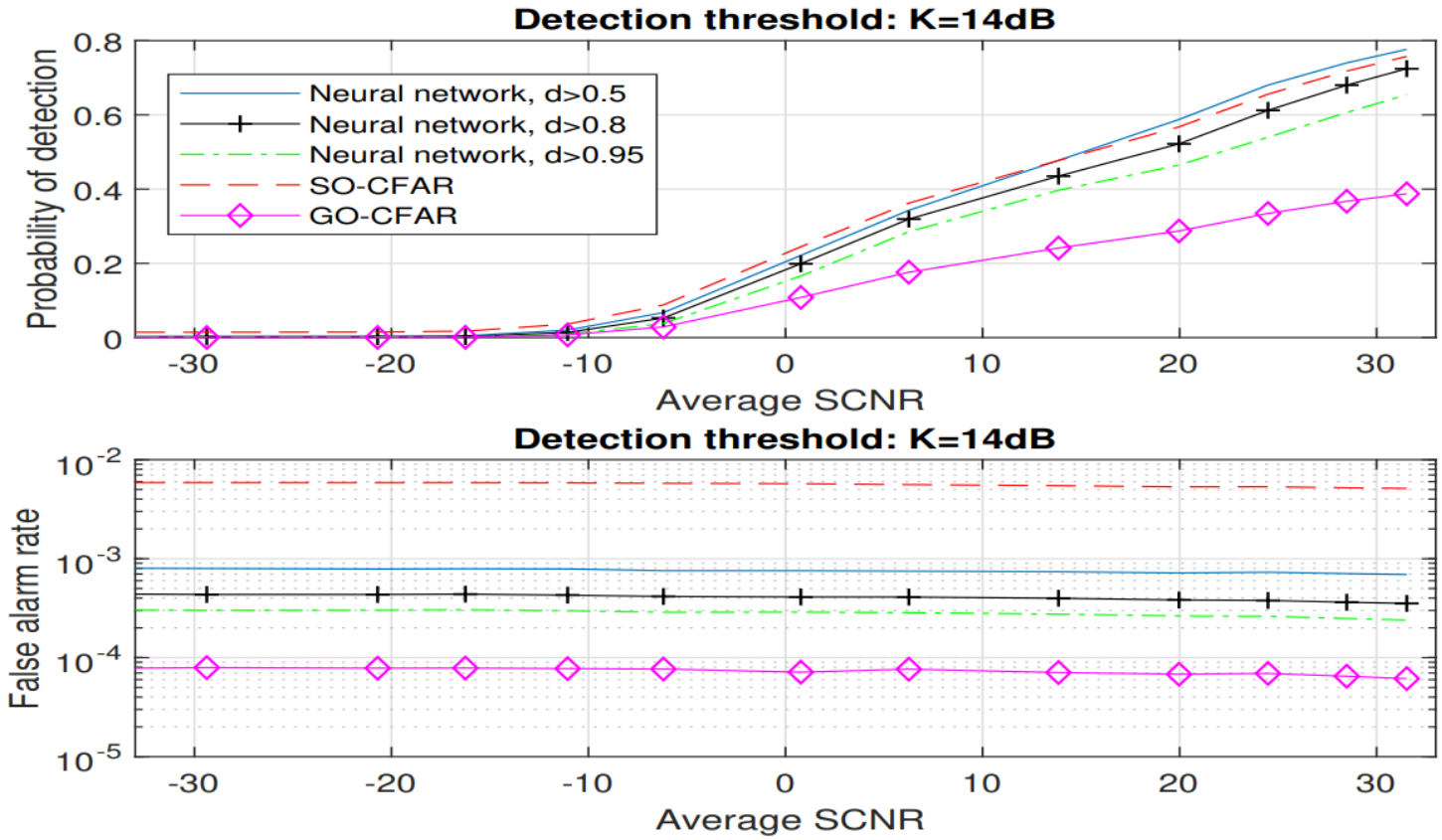




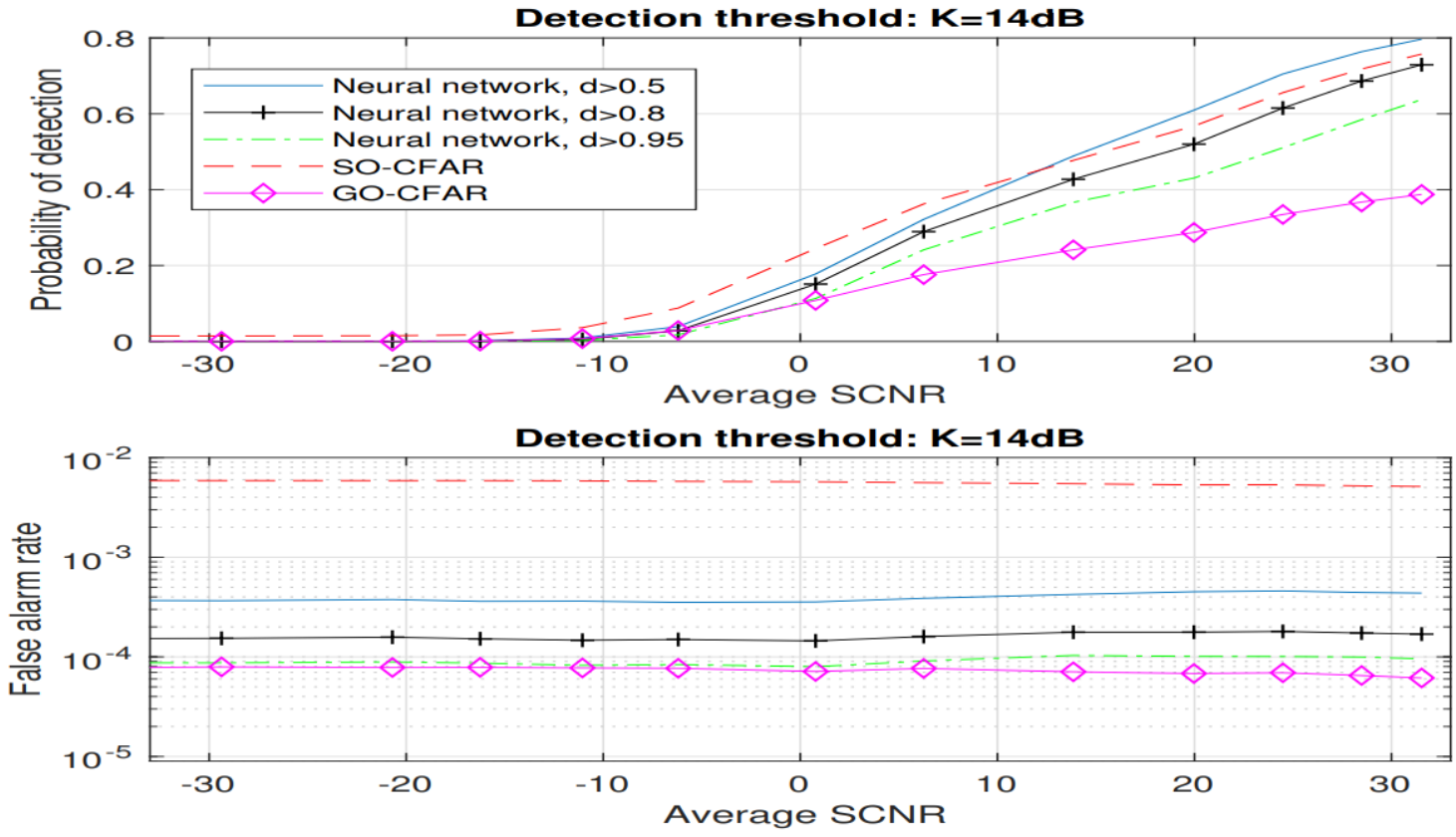
- Simulated range-Doppler map



- showing that adding an additional NN classifier at the end can reduce false alarm rate performance.



- showing that taken account more reference cells, say 2 cells above CUT and 2 cells below CUT, resulted in better false alarm rate performance.

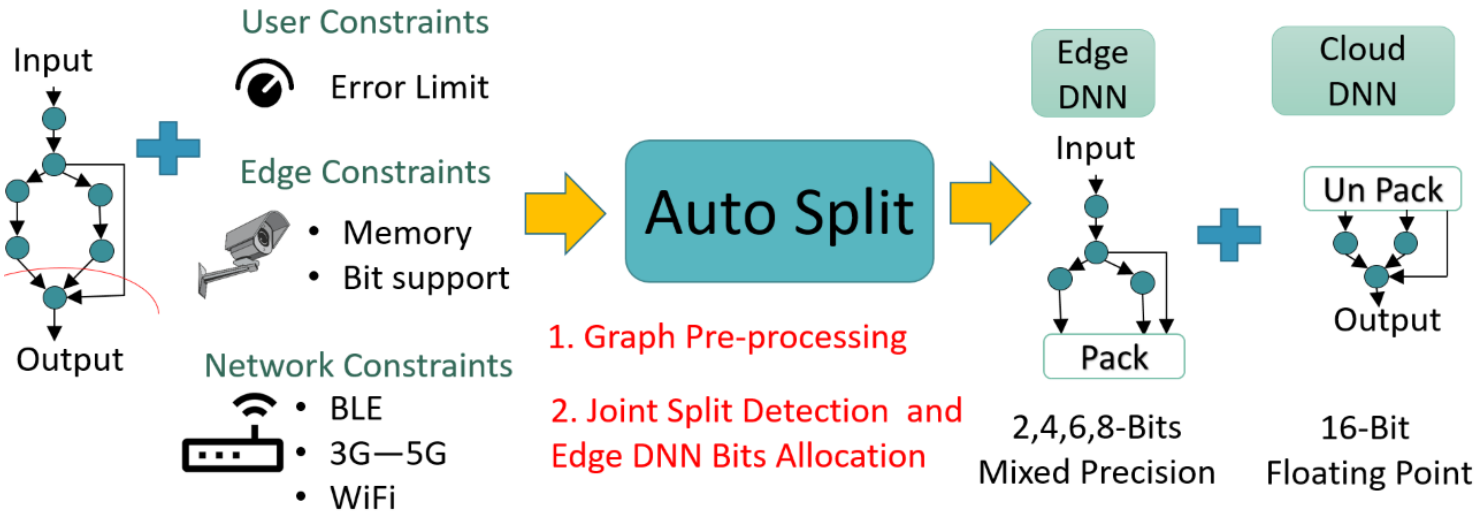


3. YOLO related works

(1) Amin Banitalebi-Dehkordi, Naveen Vedula, Jian Pei, Fei Xia, Lanjun Wang, and Yong Zhang. 2021. “Auto-Split: A General Framework of Collaborative Edge-Cloud AI[1].” *In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*.

- Overview

Auto split engine takes a DNN as input and finds the best split point to execute the initial DNN on the edge device and the later DNN on the cloud device to minimize end-to-end latency.



- Quantization Support

For edge device, it also finds quantization statistics, and bit-width per layer to apply various post training quantization techniques such as Loss Aware Post-training Quantization (LAPQ)[2], ACIQ: Analytical Clipping for Integer Quantization of neural networks (ACIQ)[3].

- Reduced Transmission Cost

At the split point, the auto-split engine parses the DNN graph and collects the features to be transmitted to the cloud device at lower precision reducing the total transmission cost.

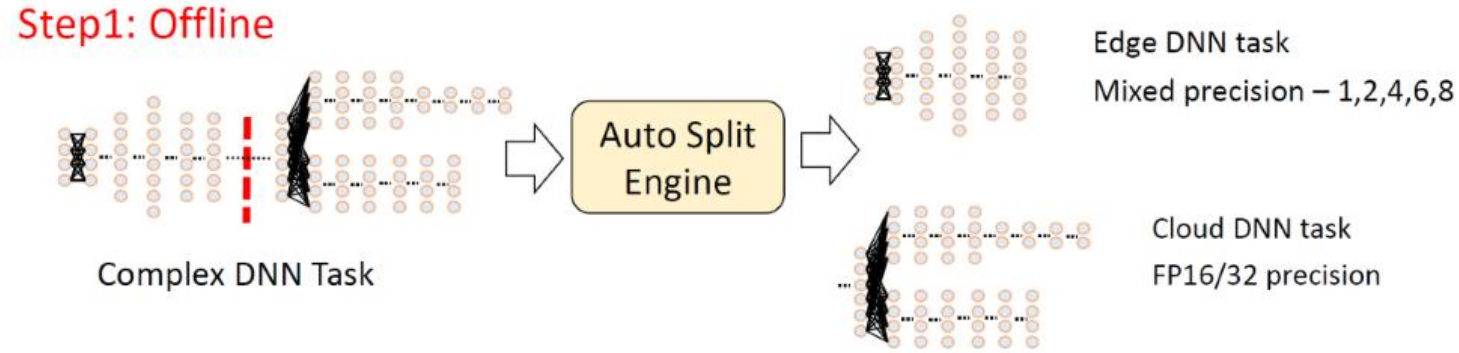
---

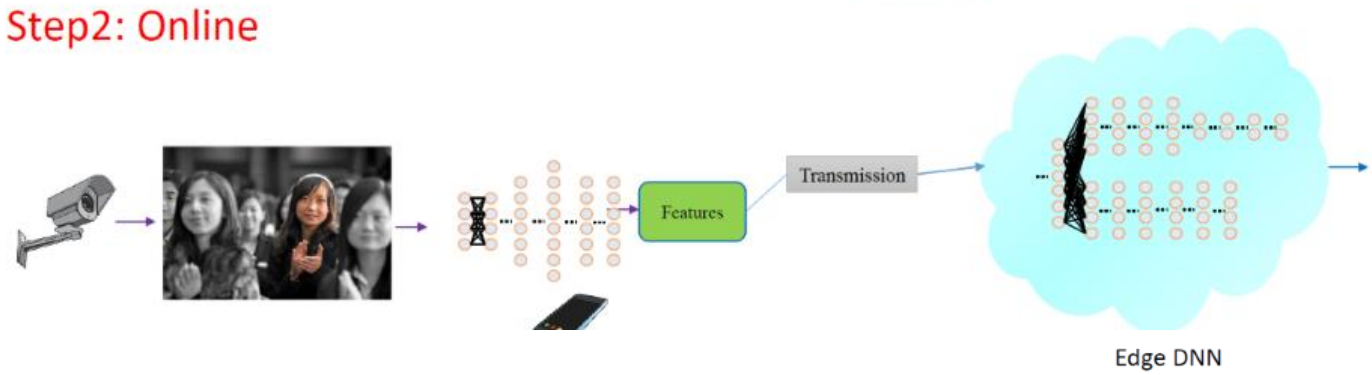
[1] Amin Banitalebi-Dehkordi, Naveen Vedula, Jian Pei, Fei Xia, Lanjun Wang, and Yong Zhang. 2021. Auto-Split: A General Framework of Collaborative Edge-Cloud AI. *In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21)*. https://dl.acm.org/doi/abs/10.1145/3447548.3467078

[2] Ron Banner, Yury Nahshan, Elad Hoffer, Daniel Soudry. ACIQ: Analytical Clipping for Integer Quantization of neural networks. *ICLR 2019 Conference*. https://github.com/submission2019/cnn-quantization

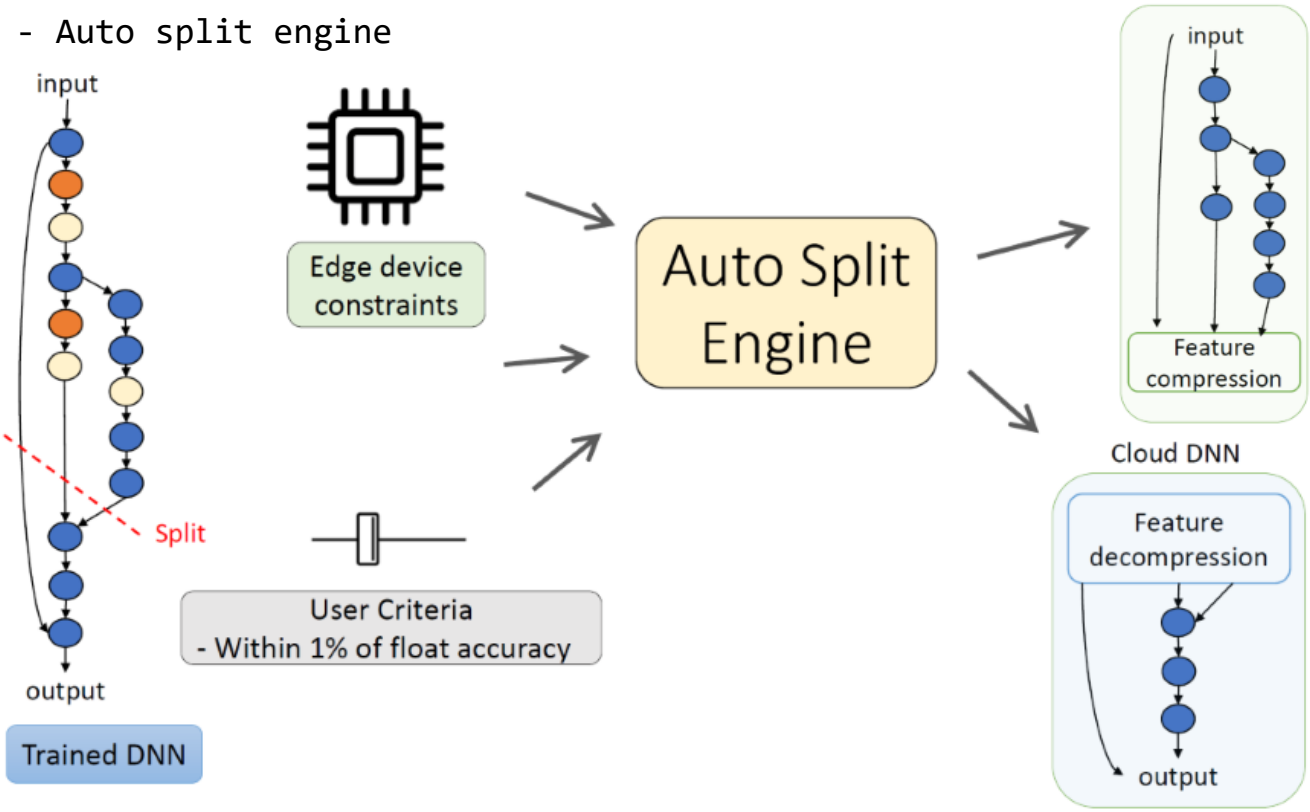
[3] Nahshan, Y., Chmiel, B., Baskin, C. et al. Loss aware post-training quantization. *Mach Learn 110*, 3245–3262 (2021). https://github.com/ynahshan/nn-quantization-pytorch/tree/master/lapq

- Cloud device executes in floating point precision[4]





- Auto split engine



- Auto-Split Solution

For Auto-Split solution, the Darknet based Yolov3 object detection is split into two parts, the edge DNN and cloud DNN.

[4] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, Hao Wu. Mixed Precision Training. *ICLR 2018 Conference*. https://arxiv.org/abs/1710.03740

The split point was suggested by the Auto-split algorithm. They manually partition the YOLOv3 into edge DNN and cloud DNN.

For the demo, the video executes each frame on the edge DNN and transmits activations to an IP address, which is then received by the cloud DNN on the GPU of that IP address.
In their example, they use the same device to transmit the data.

Appendix

- Artificial Intelligence[5]

	Publication	h5-index	h5-median
1.	International Conference on Learning Representations	286	533
2.	Neural Information Processing Systems	278	436
3.	International Conference on Machine Learning	237	421
4.	AAAI Conference on Artificial Intelligence	180	296
5.	IEEE Transactions On Systems, Man And Cybernetics Part B, Cybernetics	142	186
6.	Expert Systems with Applications	132	190
7.	IEEE Transactions on Neural Networks and Learning Systems	131	187
8.	Neurocomputing	123	187
9.	International Joint Conference on Artificial Intelligence (IJCAI)	120	186
10.	Applied Soft Computing	112	150
11.	Knowledge-Based Systems	107	143
12.	IEEE Transactions on Fuzzy Systems	101	151
13.	Neural Computing and Applications	99	137
14.	Journal of Machine Learning Research	98	162
15.	International Conference on Artificial Intelligence and Statistics	85	119
16.	Neural Networks	81	112
17.	Engineering Applications of Artificial Intelligence	76	117
18.	Artificial Intelligence Review	70	131
19.	Applied Intelligence	65	95
20.	Conference on Robot Learning	64	114

[5] https://scholar.google.es/citations?view_op=top_venues&hl=en&vq=eng_artificialintelligence