Pruned Lightweight Encoders for Computer Vision

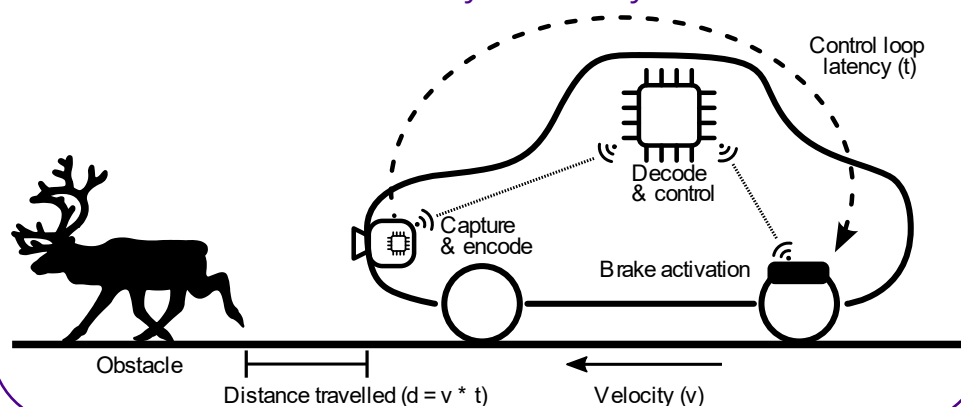Jakub Žádník, Markku Mäkitalo, Pekka Jääskeläinen, *Tampere University, Finland*

Model scenario: Low-latency control system

Assumptions:

- Requires fast compression
- Low computing power of encoding device
- Consumer is computer vision algorithm (NN), not human

Proposed Method:

- Prune encoding configurations of existing codecs to achieve low complexity
- Compensate low quality by retraining the NN with the recompressed dataset
- Two algorithms tested:
  - ASTC (texture compression)
  - JPEG XS (new mezzanine compression standard)

Conclusion:

- We achieved significant speedups by pruning the encoders
- Retraining allows recovering compression artifacts, but not completely.

Encoding time

- Encoding time of pruned ASTC encoder vs. optimized JPEG encoder (Samsung S10, single core)

	27:1	12:1
ASTC	5.8 ms	7.0 ms
JPEG	13.3 ms	16.7 ms

- Pruning reference JPEG XS encoder by disabling significance flag coding improved runtime by 22-23%.
  (Still too slow for real-time encoding: hundreds of ms.)

Computer Vision Performance (compr. ratio 27:1)

Classification accuracy vs. uncompressed

	Orig.	Retrain
ASTC	-15.1	-5.0
JPEG	-0.6	-0.7

Segmentation mIoU vs. uncompressed

	Orig.	Retrain
JPEG XS (main)	-6.5	-2.6
JPEG XS (main, pruned)	-7.4	-2.3
ASTC	-12.6	-4.0
JPEG	~-1.0	---