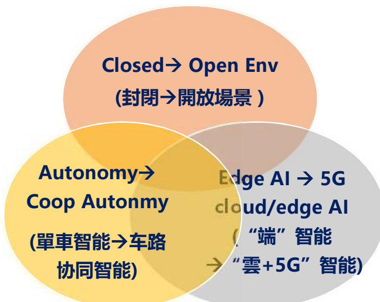
MoCAD Objectives目標

❑ Challenges in AD

- Open and Uncertain Env: hard to fully consider variety of driving scenarios
- Limits of vehicular autonomy: cooperative cloud/edge intelligence help augment vehicular autonomy

❑ 當前無人駕駛產業化和廣泛應用還存在諸多挑戰

- 封閉場景無法窮盡現實世界的各種複雜場景
- 單車智能駕駛技術無法解決大規模車群的協同控制問題

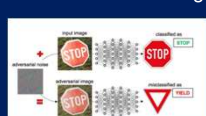


Obj I: Robustness and Adaptivity of self-driving in open and uncertain Env
開放不確定環境下智能駕駛的高魯棒性及自適應性

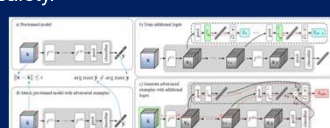
Obj II: V2I-enabled collaborative control in hybrid human/self-driving scenarios
大規模混合場景下無人網聯車的群體智能協同控制

Representative Work I:
Robustness in Autonomous Driving

❑ Problem: Perception, cognition and decision-making of autonomated vehicles in complex scenes are often realized through deep learning models. When such models are interfered and attacked, it is very easy to produce wrong results, which will affect the stability of AD. Therefore, improving the system's anti-interference ability is the cornerstone of driving safety.



Misjudgment of traffic signs caused by adversarial attacks



LAFEAT model structure with strong defense strategy evaluation

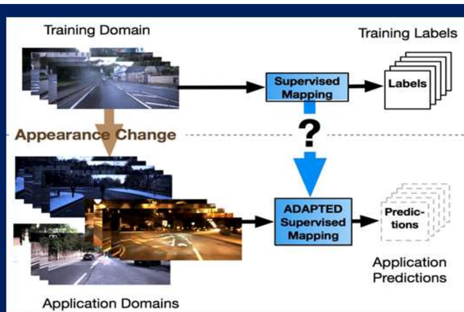
- UM PhD student Yunrui Yu et al proposed LAFEAT approaches and won the 2nd prize in "CVPR Security AI Challenge" (organized by UIUC/Tsinghua/Alibaba with 1681 team participation) and cash award of $10K.
- Results were published in AI conf CVPR 2021 in Oral Presentation (acceptance rate is 4.59%)

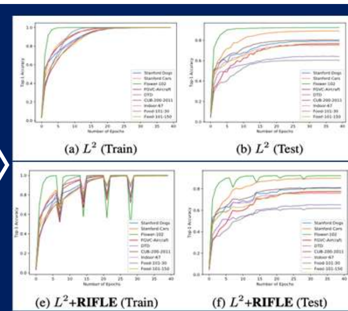
❑ Method and innovation: A new sub-network branch is added to the traditional deep model architecture, which realizes the efficient use of model feature information, thereby improving the accuracy of defense strategy evaluation.

LAFEAT: Piercing Through Adversarial Defenses with Latent Features

Representative Work II：

Transfer Learning for Scene Adaptation in AD





Problem: Model trained on typical scenes (e.g. city at daytime) may fail to adapt new scenes (e.g. countryside, night)

新任務樣本不足
和預訓練場景差異較大
懶惰訓練問題：新模型被預訓練模型『吸引』

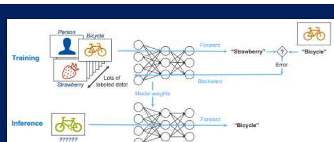
RIFLE brings deeper and reactivated gradient updates for adequate adaptation to the target task
有效提升新任務上的遷移效果
多種模型結構（**ResNet/MobileNet/EfficientNet**）通用，
準確率普遍提升**1-2%**
訓練高效，幾乎不需額外計算代價

RIFLE: Backpropagation in Depth for Deep Transfer Learning through Re-Initializing the Fully-connected LayEr (Li， et al, ICML 2020) Li is a PhD of UM

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

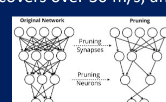
Representative Work III：

Efficeint Model Inference in AD: 1/3 (Dynamic Pruning)



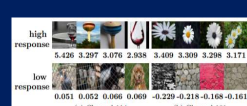
深度学习推理效率决定智能驾驶的速度

Running at 100 KPH covers over 30 m/s, and requires millisec for on-site real-time perception



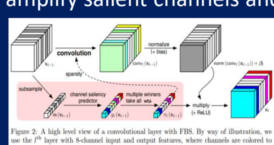
Model Compression & Acceleration

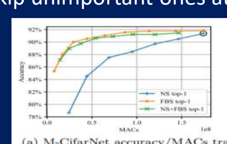
Key Observation：
Importance of features produced by deep models is highly input-dependent

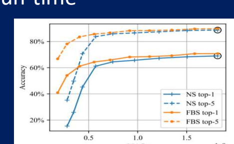


Images to excite neurons of ResNet-18 model and outputs high/low weights

Proposed Feature Boosting and Suppression (FBS):
predictively amplify salient channels and skip unimportant ones at run-time







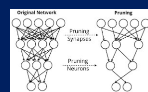
Dynamic Channel Pruning (Gao, et al. ICLR’2019)

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Representative Work III：

Efficeint Model Inference in AD: 2/3 (Shift Quantization)

深度学习推理效率决定智能驾驶的速度



Key Observation:

- Channel pruning introduces various degrees of sparsity to different layers
- Shift quantization 位移量化: quantize weights values in a model to powers-of-2 (…,2^{-1},0,2^{1},…) so that multipl be reduced to shift op
- But, shift quantization becomes a poor choice for certain layers in sparse models, as most near-zero quantization levels are under-utilized.

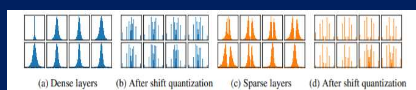
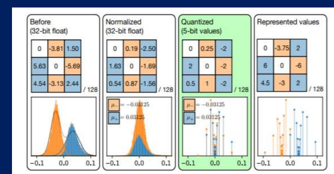


Figure 1: The weight distributions of the first 8 layers of ResNet-18 on ImageNet. (a) shows the weight

Dynamic Channel Pruning (Gao, et al. ICLR’2019)
Focused Quantization for Sparse CNNs (Gao, et al. NeurIPS’2019)

Proposed Focused Quantization --- exploit the statistical properties of weights in pruned models to quantize them efficiently and effectively



ResNet-18	Top-1	Top-5	Size (MB)	CR (×)
TTQ [27]	66.00	87.10	2.92*	16.00*
INQ (2 bits) [26]	66.60	87.20	2.92*	16.00*
INQ (3 bits) [26]	68.08	88.36	4.38*	10.67*
ADMM (2 bits) [14]	67.0	87.5	2.92*	16.00*
ADMM (3 bits) [14]	68.0	88.3	4.38*	10.67*
ABC-Net (5 bases, or 5 bits) [15]	67.30	87.90	7.30*	6.4 *
LQ-Net (preact, 2 bits) [23]	68.00	88.00	2.92*	16.00*
D&Q (large) [20]	**73.10**	**91.17**	21.98	2.13*
Coreset [3]	68.00	—	3.11*	15.00
Focused compression (5 bits, sparse)	68.36	88.45	**2.86**	**16.33**
ResNet-50	**Top-1**	**Top-5**	**Size (MB)**	**CR (×)**
INQ (5 bits) [26]	74.81	92.45	14.64*	6.40*
ADMM (3 bits) [14]	74.0	91.6	8.78*	10.67*
ThiNet [17]	72.04	90.67	16.94	5.53*
Clip-Q [22]	73.70	—	6.70	14.00*
Coreset [3]	74.00	—	5.93*	15.80
Focused compression (5 bits, sparse)	**74.86**	**92.59**	**5.19**	**18.08**

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

Representative Work III：

Efficient Model Inference in AD: 3/3 (HW/SW Co-Design)

深度学习推理效率决定智能驾驶的速度

Key Observations:

- Shift op facilitates HW impl. HW design tends to use flattened streaming arch (vs systolic arrays) for inference acceleration.
- Flatten streaming accelerators isolate layer-wise computations, offering the chance to use different arithmetic and precisions for each layer’s computation

Proposed Tomato HW/SW Co-Design:

- HW: Multi-Precision Multi-Arith accelerator on Multi-FPGAs
- SW: Hybrid quantization to automate the selection of arithmetic and precisions for different layers of the model, so as to map all the layers onto a single or multiple FPGAs.

	Implementation	Quantisation(s) Weights	Quantisation(s) Acts	Platform	Frequency (MHz)	Latency (ms)	Throughput (FPS)	Arithmetic perf. (GOP/s)
VGG16	Throughput-Opt [33]	FXP8	FXP16	Intel Stratix V	120	262.9	3.8*	117.8
VGG16	fpgaConvNet [34]	FXP16	FXP16	Xilinx Zynq XC7Z045	125	197*	5.07	156
VGG16	Angel-Eye [9]	BFP8	BFP8	Xilinx Zynq XC7Z045	150	163*	6.12*	188
VGG16	Going Deeper [25]	FXP16	FXP16	Xilinx Zynq XC7Z045	150	224*	4.45	137
VGG16	Shen *et al.* [31]	FXP16	FXP16	Xilinx Virtex US XCVU440	200	49.1	26.7	821
VGG16	HARPv2 [23]	BIN	BIN	Intel HARPv2	–	8.77*	114	3500
VGG16	GPU [23]	FP32	FP32	Nvidia Titan X	–	–	121	3590
MobileNet	Ours	Mixed	FXP8	Intel Stratix 10	156	0.32	3109	3536
MobileNet	Ours	Mixed	FXP8	Xilinx Virtex US+ XCVU9P	125	0.40	2491	2833
MobileNet	Zhao *et al.* [41]	FXP16	FXP16	Intel Stratix V	200	0.88	1131	1287
MobileNet	Zhao *et al.* [42]	FXP8	FXP8	Intel Stratix V	150	4.33	231	264
MobileNet	GPU	FP32	FP32	Nvidia GTX 1080Ti	–	279.4	515	586

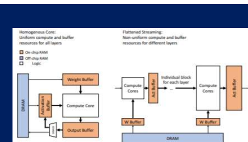
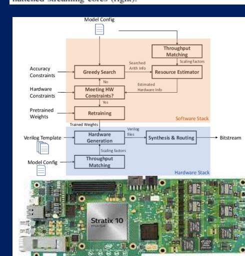


Fig. 1: An illustration of a homogenous core (left) and flattened streaming cores (right).



- Dynamic Channel Pruning (Gao, et al. ICLR’2019)
- Focused Quantization for Sparce CNNs (Gao, et al. NeurIPS’2019)
- FPGA Implementation for CNN acceleration (Gao, et al. FTS’2019)

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

澳門大學
UNIVERSIDADE DE MACAU
UNIVERSITY OF MACAU

科技學院
Faculdade de Ciências e Tecnologia
Faculty of Science and Technology

Mostly Cited Venues in Google

(google scholar metrics)

Categories ▾ English ▾

	Publication	h5-index	h5-median
1.	Nature	414	607
2.	The New England Journal of Medicine	410	704
3.	Science	391	564
4.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	356	583
5.	The Lancet	345	600
6.	Advanced Materials	294	406
7.	Cell	288	459
8.	Nature Communications	287	389
9.	Chemical Reviews	270	434
10.	International Conference on Learning Representations	253	470
11.	JAMA	253	446
12.	Neural Information Processing Systems	245	422
13.	Proceedings of the National Academy of Sciences	245	337
14.	Journal of the American Chemical Society	245	330
15.	Angewandte Chemie	235	314
16.	Chemical Society Reviews	234	339
17.	Nucleic Acids Research	233	512
18.	Renewable and Sustainable Energy Reviews	225	294
19.	Journal of Clinical Oncology	213	297
20.	Physical Review Letters	209	297
21.	Advanced Energy Materials	206	267
22.	Nature Medicine	205	356
23.	International Conference on Machine Learning	204	370
24.	Energy & Environmental Science	202	306