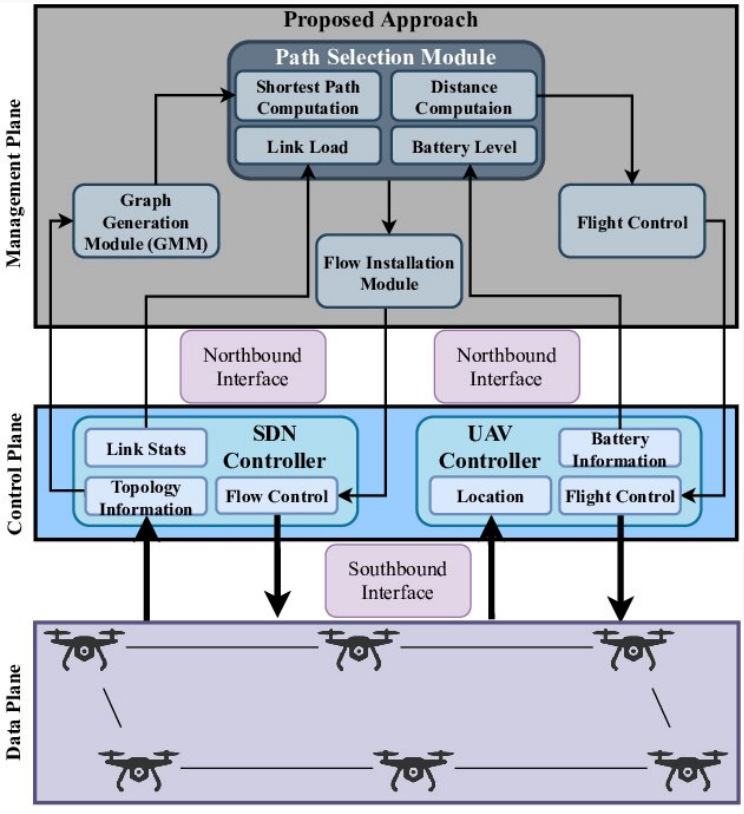
ISSUE: FEBUARY2023

CTSOC-NCT NEWS ON CONSUMER TECHNOLOGY



The proposed system architecture of SDN-Based framework for load balancing and flight control in UAV networks.

IEEE CTSoc CONSUMER TECHNOLOGY SOCIETY

TABLE OF CONTENTS

EDITOR'S NOTE

On behalf of the Editorial Board of IEEE CTSoc News on Consumer Technology (NCT) editor-in-chief Wen- Huang Cheng and editors Yafei Hou, Luca Romeo, Yuen Peng Loh, Chuan-Ju Wang, and I am delighted to introduce the February issue of the News on Consumer Technology (NCT).

For this issue, we begin with a cover story regarding a new framework for load balancing and flight control in Unmanned Aerial Vehicles (UAVs) published by IEEE Transactions on Consumer Electronics Magazine. This work proposes a software degined networking to conduct separate control logics for mutliples UAVs simutaneously. The system also puts battery limitation, collision avoiding, and traffic load into consideration for a better flight control planning.

Following the first part, there is a feature interview with Prof. Jiun-In Guo from National Yang Ming Chiao Tung University, who has been an IEEE valued member for 21 years. His research covers a broad range of areas, including images, multimedia, digital signal processing, SOC design, and self-driving vehicles, etc. His recent work on "Self-learning AI system" proposes a brand new solution which can help users to fine-tune existing developed AI models without labeling data in an automatic way.

Finally, the issue ends with a featured article from Dr. Jingkuan Song who is a full professor with the Univeristy of Electronic Science and Technology of China (UESTC). This article gives a detialed introduciton on Multimedia Compact Representation, including different approaches like hashing and quantization, applications like visual compression, fast retrieval, and challenges like optimization problems in learning comapct codes. This survey is comprehensive and easy to follow.

Have a nice read!

Jianlong Fu, Ph.D.

Senior Research Manager
Microsoft Research Asia

ARTICLE TITLE
An SDN-Based Framework for Load Balancing and Flight Control in UAV Networks

AUTHOR(S)
Zohaib Latif; Choonhwa Lee; Kashif Sharif; Fan Li; Saraju P. Mohanty

JOURNAL TITLE
IEEE Consumer Electronics Magazine

JOURNAL VOLUME AND ISSUE
Volume 12, Issue 1

DATE OF THE ARTICLE
August 2022

PAGE NUMBERS FOR THE ARTICLE
43 – 51

DOI
10.1109/MCE.2022.3200174

Unmanned Aerial Vehicles (UAVs) are gaining tremendous attention due to their flying nature. To complete the task efficiently, multi UAV systems are a good choice as compared to a single UAV system. However, multi-UAV systems introduce issues such as high dynamics, limited battery, and frequent changes in topology.

To solve these issues, Software defined networking (SDN) is an excellent candidate to separate control logic from forwarding elements and provide high-level programming abstractions. However, due to architectural constraints, applying SDN introduces some new challenges, including uneven load on multiple links between source and destination. This irregular load also affects UAVs' battery consumption, necessitating an adequate solution to meet these challenges fully.

The authors of this paper proposed an SDN-based framework for UAV elements that monitors frequent changes in the network topology. They also designed an algorithm to distribute traffic load evenly on different links of multi-UAV systems. The proposed UAV networks also considered the battery limitations, and traffic is shifted to a path where elements have more battery. Moreover, a flight control mechanism is proposed to avoid collisions due to the high dynamics of UAVs. Simulation results show that the traffic load is distributed evenly on multiple links connecting different systems with less battery consumption.

INTERVIEW WITH PROF. JIUN-IN GUO

National Yang Ming Chiao Tung University
IEEE valued member for 21 years

Prof. Jiun-In Guo

Prof. Jiun-In Guo received the B.S. and Ph.D. degrees in Electronics Engineering from National Chiao Tung University (NCTU), Hsinchu, Taiwan, in 1989 and 1993, respectively. He is currently a Distinguished Professor of the Institute of Electronics, and the Director of Wistron-NCTU Embedded Artificial Intelligence Research Center, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan. Prof. Guo has been the Associated Dean of Electrical and Computer Engineering during 2017-2023, the Director of Institute of Electronics during 2013-2015, and the Professor in Department of Electronics Engineering, NCTU since 2011. Before joining in NCTU, Prof. Guo was an Associate Professor of the Dept. of Computer Science and Information Engineering, National Chung-Cheng University (CCU), from 2001 to 2003 and get promoted as a Professor since 2003. Prof. Guo also served as the Director of the SOC Research Center, CCU, from 2005 to 2008, and the Director of the Dept. of Computer Science, CCU, Taiwan during 2009-2011 and the Research Distinguished Professor of CCU since 2008. During 1994 to 2001, Prof. Guo served as an Associate Professor of the Dept. of Electronics Engineering, National Lien-Ho Institute of Technology, Miaoli, Taiwan. His research interests include images, multimedia, and digital signal processing, VLSI algorithm/architecture design, digital SIP design, SOC design, and intelligent vision processing applications including ADAS/Self-driving vehicles.

Prof. Guo received the outstanding electrical engineering professor award from the Chinese Institute of Electrical Engineering in 2010, the outstanding engineering professor award from the Chinese Institute of Engineers in 2014, the outstanding research award of Minister of Science (MOST) in 2017, as well as the outstanding

technology transferring award of MOST in 2018 and 2020 with the topic of deep learning ADAS systems. Prof. Guo was also selected as the Elsevier 1960-2020 top 2% Scientist and 1960-2021 top 2% Scientist in Life-long Impact by Stanford University in 2021 and 2022, respectively.

Prof. Guo is the author of 258 technical papers on the research areas and has served as the PI of 102 research projects, 122 industrial projects, and 61 industrial technology transfer projects. Prof. Guo is also the inventor of 48 invention patents and the receiptent of 122 awards in the research areas. With all the cumulated research outcome, Prof. Guo starts up a company called eNeural Technologies, Inc. since March 2022, where eNeural Technologies Inc. is an embedded AI design service house in the area of automotive and AIOT applications to help customers to solve the pain points in developing quality light weight AI models on embedded computing platforms. For more information on the Prof. Guo and his established iVSLab, NYCU, as well as the eNeural Technologies Inc., please visit the official websites below: http://ivs.ee.nctu.edu.tw/ivs/ and https://www.eneuraltech.com/.

Since you have been with three universities in the past 28 years after you got the Ph.D degree, please summarize your research roadmap and highlight the representative research outcome.

It has been a long time (28 years) after I graduated from NCTU and got the Ph.D degree in Electronics Engineering. My Ph.D research is regrading to the transform module design and implementation in video coding, including discrete cosine transform (DCT), discrete sine transform (DST), discrete Fourier transform (DFT) and discrete Hadamard transform (DHT). All of these transform functions own similarity in algorithm such that it is possible to develop a universal architecture to realize them in an efficient way from hardware point of view, i.e. using systolic array architecture. With the research background, my research in the first 7 years when I stayed in National Lien-Ho Technology focuses on the design and implementation of the transform functions in video coding and compression.

After that when I stayed in National Chung Cheng University (CCU) during 2000 to 2010, my research area moved from component design to system design in the video compression and decompression, starting from MPEG-1/2/4 to H.264. During the 10 years in CCU, I established a research group to develop video encoder/decoder silicon IP (SIP) to support MPEG-1/2/4 and H.264 that can pass the compliance testing of the MPEG standards with the features of low-power consumption and high processing performance, which makes these SIPs not only to be published in quality IEEE journal papers [1-13] and ISSCC top conference papers [14-17], but also attracts more than 10 industrial companies to license our SIPs to develop multimedia SoC. The representative one is the H.264 video Codec SoC from Faraday/Grain Media launched in 2008, which adopted our H.264 video encoder/decoder IP and made this SoC as the first H.264 video Codec SoC in Taiwan.

Since 2011 I moved to National Chiao Tung University (NCTU), my mother university, and started research on intelligent vision processing system. The reason why I change my research areas from video Codec to vision system is due to the large and complex scale of designing and implementing video Codec SoC after H.264, which is not affordable in both man-power and budget of a research group in university. With the design experience in H.264 video Codec SoC, I decide to conduct the research related to the applications of H.264 video Codec. The topic we selected is the car camcorder with ADAS functions (i.e. LDWS/FCWS). During 2011-2015, there is a trend to install car camcorder in every vehicle to record the driving traffic for self-protection purpose. In addition to the video recording function, there is extra requirement to implement the ADAS functions like Lane Departure Warning System (LDWS), Forward Collision Warning System (FCWS), Pedestrian Detection System (PDS), Blind Spot Detection (BSD) and traffic light detection (TLD) in the embedded processor in car

camcorders. Therefore, we have developed lots of ADAS functions and realize them in embedded systems. Due to the design trend of ADAS and Autonomous driving system since 2011, our research in embedded ADAS functions did attract a lot of industrial collaboration projects and published in international journals and conference. Some representative publications are included in [18-26] and we have published a book chapter related to intelligent vision processing algorithms in ADAS applications in [27]. Since 2016, we developed our embedded ADAS functions based on deep learning methods to further improve the reliability, but suffered from the high computational complexity. In order to resolve this problem, we decided to develop some automatic design methods for data labeling (called ezLabel that can be used for free in https://www.aicreda.com/), model pruning (called ezModel) and model quantization (called ezQUANT) to help designing an efficient AI model that can be realized in an embedded AI SoC. Some representative research outcome can be found in reference [28-41].

From your research areas, you focus on the research of embedded AI technology design a lot. What are your observations on the key factors in embedded AI technology development?

When I look into the research to develop AI models for embedded system realization without using GPU, there are a lot of design challenges starting from labeled data preparation, model design and pruning, model quantization from gloating point model to fixed point model for the model inference by a dedicated hardware accelerator. Therefore, I would say that if you would like to design an efficient AI model for embedded AI SoC implementation, you must deal with the design challenges I mentioned above. To achieve this goal, we have proposed some solutions to deal with the design challenges.

First, in order to reduce the large manual efforts in data labeling, we have developed a fast/automatic image data label tool, called ezLabel, for users to speed-up the data labeling before they train the AI models. The ezLabel is a web-based labeling tool that provides users a friendly data labeling environment to speed-up the boring data labeling process. The ezLabel is free to use for academic research purpose and users can register an account in the website of https://www.aicreda.com/. More description on using ezlabel can be found in the reference paper [34] (https://www.mdpi.com/2072-4292/14/4/833). In addition to ezLabel, we also released a bunch of labeled objects, including vehicles, pedestrians, motorcycle riders and bicycle riders, in Taiwan road traffic for ADAS and Autonomous driving applications. The released dataset is called iVS-Dataset that can be downloaded in the following link: https://github.com/ivslabnctu/IVS-Dataset.

Second, in order to perform the model pruning efficiently and flexibly for any embedded systems, we proposed an automatic model pruning methodology and implemented it as an automatic model pruning tool, called ezModel. The ezModel has been evolved to the third generation that outperforms the previous versions in model pruning efficiency as shown in Figure 1. Take SSD as an example with PASCAL VOC benchmark, the ezModel 3.0 outperforms its previous version in showing higher model accuracy with the same percentages of model complexity reduction. It can even preserve the same model accuracy under 56% model complexity reduction without any quality degradation.

Model Architecture : SSD
Image size : 300 x 300
Dataset : PASCAL VOC (20 classes)

		DGM-AB (ezModel 3.0)								
		CPGM&T (ezModel 2.0)								
		S&T (ezModel 1.0)								
Ratio(%)	Original	P10	P20	P30	P40	P50	P60	P70	P80	P90
mAP(%)	78.78	79.32	79.16	79.37	79.68	79.49	79.56	78.53	75.9	68.79
	78.7	78.5	77.06	76.64	76.25	76.57	75.72	76.45	74.01	68.56
	78.7	77.1	77.87	78.0	77.49	78.02	77.51	77.23	75.15	61.4
FLOPs(G)	60.3	58.8	55.59	52.23	46.7	38.15	26.25	19.15	5.2	4.07
	61.05	60.52	57.32	51.78	44.35	35.85	26.13	17	9.14	3.51
	61.05	52.08	46.18	40.84	35.42	30.73	25.23	18.12	10.38	4.24
FLOPs Reduction Rate(%)	-	2.49	7.81	13.38	22.55	36.73	56.46	68.24	91.37	93.25
	-	0.86	6.10	15.18	27.35	41.27	57.19	72.15	85.02	94.25
	-	14.69	24.35	33.10	41.98	49.66	58.69	70.31	82.99	93.05

No quality loss Better accuracy

Figure 1: Performance of ezModel

Third, in order to make sure the AI model in fixed point format (8-bit) does not have much quality degradation compared with the floating

point one, we need to do model quantization analysis and retraining to prevent from the severe quality degradation. Figure 2 shows an example (tiny-yolo v2) of error accumulation of fixed-point model on the model output if there is no bit-accurate simulation. As you can see that the quantization error accumulation will cause inaccurate un-predictable AI detection results that cannot be resolved from model training.

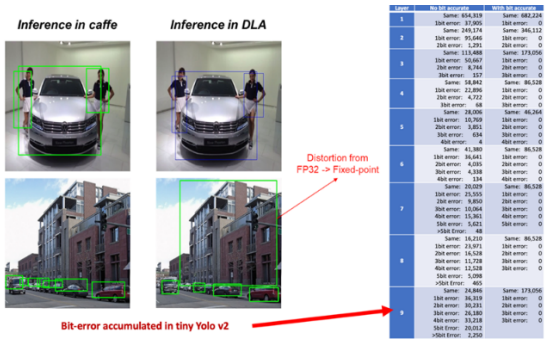


Figure 2: Error accumulation of tiny-yolov2 model

In order to solve this problem, we have developed model quantization and training tool for fixed point AI model, which is called ezQUANT [32]. The ezQUANT can support dynamic fixed point model quantization in both bit-accurate mode and layer-by-layer model. In bit-accurate mode, it can support bit-accurate model simulation on a fixed-point model to make sure the model simulation results and model inference results on AI chip are the same. For the layer-by-layer mode, it supports the fixed-point model simulation in layer input and output without taking care of the precision issue within a layer in CNN model, which is easy to be implemented, but might cause error propagation issue. In order to further reduce impact of the model quantization to model quality, we also proposed a multi-scale dynamic fixed point model quantization method on CNN architecture, which is referred to be ezQUANT 2.0 [41]. As shown in Figure 3, the multi-scale quantization outperforms the single-scale quantization in preserving model accuracy while doing model quantization on the examples of Yolov4, Yolov3, and Yolov3-tiny. Figure 4 shows the comparison of the proposed multi-scale quantization with the Qualcomm AIMET model quantization [42, 43]. As you can see that no matter in post-training quantization (PTQ) and Quantization aware training (QAT) modes, the proposed multi-scale quantization method [41] outperforms Qualcomm AIMET in preserving better model accuracy after performing quantization in both 8-bit feature maps and 4-b/8-b weights on different models under ImageNet-1000 benchmark. For more details on the proposed embedded AI design technologies, please visit my talk video in AutoCAS2022 invited speech with the topic of "Embedded AI Deep Learning Technology for ADAS/ADS Applications" in the following link: https://www.youtube.com/watch?v=hmORIRvzvQ8.

YOLOv4	FP	Uniform Quantization (W/A)	Multi-scale Quantization (W/A)
Precision (%)	49.1	35.8	27.5
Recall (%)	70.8	59.8	75.8
mAP@0.5 (%)	65.6	44.3 (-21.3)	62.2 (-3.4)
YOLOv3	**FP**	**Uniform Quantization (W/A)**	**Multi-scale Quantization (W/A)**
Precision (%)	45.8	39.7	38.3
Recall (%)	70.4	63.0	69.4
mAP@0.5 (%)	65.2	56.5 (-9.0)	61.6 (-3.6)
YOLOv3-Tiny	**FP**	**Uniform Quantization (W/A)**	**Multi-scale Quantization (W/A)**
Precision (%)	43.0	36.1	34.1
Recall (%)	38.7	31.2	36.2
mAP@0.5 (%)	35.7	27.3 (-8.4)	30.5 (-5.2)

Figure 3: Performance of the proposed multi-scale quantization method

▪ ImageNet1000 a8w8

a8w8	FP32	AIMET PTQ		AIMET QAT		ezQuant PTQ		ezQuant QAT	
	top1	top1	Drop	top1	Drop	top1	Drop	top1	Drop
Resnet18	69.76	69.52	-0.24	69.85	+0.09	69.64	-0.12	70.20	+0.44
Resnet50	76.13	75.81	-0.32	76.29	+0.16	75.88	-0.25	76.38	+0.25
Resnet101	77.37	77.03	-0.35	77.14	-0.23	75.41	-1.96	77.54	+0.17
Wide_resnet50	78.47	77.53	-0.94	78.29	-0.18	78.43	-0.04	78.49	+0.02
Wide_resnet101	78.85	77.93	-0.92	78.26	-0.59	72.92	-5.93	78.57	-0.28
MobilenetV2	71.88	70.88	-1.00	70.37	-1.51	70.92	-0.96	71.23	-0.65

▪ ImageNet1000 a8w4

a8w4	FP32	AIMET PTQ		AIMET QAT		ezQuant QAT	
	top1	top1	Drop	top1	Drop	top1	Drop
Resnet18	69.76	59.31	-10.45	67.38	-2.38	67.87	-1.89
Resnet50	76.13	64.31	-11.82	74.75	-1.38	75.42	-0.71
Resnet101	77.37	65.23	-12.14	74.57	-2.80	76.80	-0.57
Wide_resnet50	78.47	64.96	-13.50	67.89	-10.58	77.47	-1.00
Wide_resnet101	78.85	61.08	-17.77	53.51	-25.34	77.75	-1.10
MobilenetV2	71.88	48.32	-23.56	65.20	-6.68	65.52	-6.36

Figure 4: Comparing the proposed multi-scale quantization to Qualcomm AIMET quantization

Please talk about more on the open dataset you mentioned, iVS-Dataset. What is the impact of the open datasets to the researchers in the area of ADAS/Autonomous driving applications?

For overcoming the limitations in the standard datasets with the data such as wide-variety of scales and data captured in various conditions that are necessary to train the neural networks to yield efficient results in the ADAS applications, we delivered the self-built open iVS-Dataset and an open-to-free-use data annotation tool entitled 'ezLabel'. The iVS-Dataset comprises of various objects of different scales as seen in and around the real driving environments. The data in iVS-Dataset are collected employing a camcorder in vehicles driving under different conditions such as light, weather and traffic, and driving scenarios ranging from city traffic in peak and normal hours to freeway traffic in busy and normal conditions. Thus, the collected data are of wide-range and captured all the possible objects in all the various scales as appeared in the real-time. The data collected to build the dataset has to be annotated before used in training the CNNs and this paper presents an open-to-free-use data annotation tool, ezLabel, for the data annotation as well. Figure 5 shows some snapshots of the iVS-Dataset under different scenarios and weather conditions.

Figure 5: Some snapshots of the iVS-Dataset under different scenarios and weather conditions

With this open dataset, users can adopt them to fine tune the AI models together for ADAS applications with some open datasets in other countries for the field adaptation in Taiwan traffic. In order to further promote the iVS-Dataset, we have hosted embedded AI model development contest (called PAIR competition) in the past four years and this year 2023 as well, including IEEE MMSP2019 Grand Challenge (GC) PAIR competition (https://aidea-web.tw/topic/28f4dad1-420a-435c-8afd-fd161728f5db?focus=intro), IEEE ICME2020 GC PAIR competition (https://pairlabs.ai/icme2020-grand-challenge-pair-%e7%ab%b6%e8%b3%bd-%e5%9c%93%e6%bb%bf%e8%90%bd%e5%b9%95/), ACM ICMR2021 GC PAIR competition (https://pairlabs.ai/en/acm-icmr-2021-grand-challenge-pair-competition/) and IEEE ICME2022 GC PAIR competition (https://pairlabs.ai/icme-2022-grand-challenges/). The PAIR competition this year is entitled as: Low-power Deep Learning Object Detection and Semantic Segmentation Multitask Model Compression Competition for Traffic Scene in Asian Countries (https://aidea-web.tw/topic/c20c2d78-b199-48a4-ae03-6cf2ff50f569?focus=intro) associated with IEEE ICME2023 GC. If anyone who is interested in this topic, you are encouraged to sign up this competition. All the finished four competitions have attracted more than 720 teams to sign up the competition and download the iVS-Dataset for model training and performance evaluation, which is helpful in the embedded AI design community to cultivate more young talents in this field. More details on these four competitions can be found in the summary papers in the reference papers [44-47].

Do you have a plan to spin your scientific research out of the university and into a startup?

With over 28 years research in video coding, silicon IP/SoC design, intelligent vision, and embedded AI technologies, I have decided to start up an embedded AI IP design house, called eNeural Technologies Inc., in March 2022. Before making this decision, I have struggled for more than 3 years about starting up companies, since the culture to run a business is totally different from the academic research. The major reason why I decide to start up a company is that I think start up a company can fulfill the goal of commercialize the research outcome in universities to enable much more impact on our society. In the past 20 years, I have involved in more than 180 industrial collaboration projects. But, I think only less technologies are completely put into production for commercial usage by the collaborated

companies. Though there are a lot of factors influencing the successful probability of our research outcome, the major factor is the momentum of keep polishing and customizing the research outcome for commercial usage, which is not easily carried out by the collaborated companies. Therefore, with my own startup, it should be easier to overcome this challenge to commercialize the research outcome of universities.

Although the future of AI is bright and we are only at the dawn of exploring all the possible AI applications, deep learning technology is only heavily developed since 2016 and not quite well known to the high-tech industry. The AI modeling and model compression knowhow is only controlled by relatively few. The high-tech industry lacks the ability to deploy complicate AI models onto edge devices with limited computing power. This problem hinders the AIoT development and severely jeopardizes the competitiveness of technology companies.

eNueral Technologies Inc. aims as an embedded AI IP and service design house, where the IP here includes embedded AI model IP, as well as the silicon IP of the AI accelerator for SoC integration purpose. Our vision is "Bright and Light", which means to develop the brightest AI model to be realized in a light weight AI chip/NPU. To pursue this vision, our mission is to deliver a comprehensive AI solution – Cutting edge NPU IP empowered by embedded AI software stack, to help customers to bring AI into production. Figure 6 shows the provided technology service of eNeural Technologies Inc. to our potential customers.

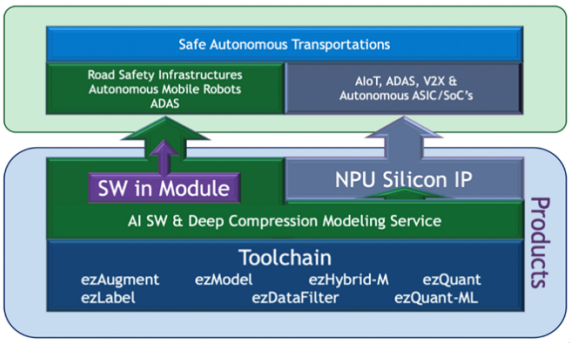


Figure 6: The provided technology service of eNeural Technologies Inc.

Next, I would like to introduce a brand new solution developed by eNeural Technologies Inc. that we announced in CES2023 TTA Eureka Park, i.e. "Self-learning AI system" that can help users to automatically fine tune your developed AI model without labeling data. Figure 7 shows the feature of the announced "Self-learning AI system" that can support different applications through the same methodology. Some use cases in eMirror and multi-functional ADAS system show that the AI model accuracy and recall can be improved by using the proposed self-learning AI system automatically in short period of time that can help speed up 6 times compared to the manual model fine tuning. For more details on the proposed "Self-learning AI system", please refer to the introducing video in the link: https://www.youtube.com/watch?v=HdTwAS5Cv0s&t=12s.

Figure 7: The features of the announced "Self-learning AI system" by eNeural Technologies Inc.

Reference
[1] J. I. Guo, Rei-Chin Ju, and Jia-Wei Chen, "An Efficient 2-D DCT/IDCT Core Design using Cyclic Convolution and Adder-based Realization," IEEE Transactions on Circuits and Systems for Video Technology, vol. 14, no. 4, pp. 416~428, April 2004.
[2] Kuan-Hung Chen, Jiun-In Guo, Jinn-Shyan Wang, Ching-Wei Yeh and Jia-Wei Chen, "A. Power-Aware IP Core Design for the Variable-Length DCT/IDCT Targeting at MPEG4 Shape-Adaptive Transforms," IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Integrated Multimedia Platforms, vol. 15, no. 5, pp. 704-715, May 2005.
[3] Hun-Chen Chen, Jiun-In Guo, Tian-Sheuan Chang, and Chein-Wei Jen, "A Memory Efficient Realization of Cyclic Convolution and its Application to Discrete Cosine Transform," IEEE Transactions on Circuits and Systems for Video Technology, vol. 15, no. 3, pp.445-453, March 2005.

[4] Kuan-Hung Chen, Jiun-In Guo, Jinn-Shyan Wang, "An Efficient Direct 2-D Transform Coding IP Design for MPEG-4 AVC/H.264", IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no. 4, pp.472-483, April 2006.
[5] Chih-Da Chien, Keng-Po Lu, Yu-Min Chen, Jiun-In Guo, Yuan-Sun Chu, and Ching-Lung Su,"An Efficient Variable Length Decoder IP Core Design for MPEG-1/2/4 Video Coding Applications," IEEE Transactions on Circuits and Systems for Video Technology, vol. 16, no.9, pp.1172-1178, September 2006.
[6] Chien-Chang Lin, Jia-Wei Chen, Hsiu-Cheng Chang, Yao-Chang Yang,Yi-Huan Ou Yang, Ming-Chih Tsai, Jiun-In Guo, and Jinn-Shyan Wang, "A 160K Gates/4.5KB SRAM H.264 Video Decoder for HDTV Applications," IEEE Journal of Solid-State Circuits, vol. 42, no.1, pp. 170~182, January 2007.
[7] Chih-Da Chien, Cheng-An Chien, Chien-Chang Lin, Ching-Hwa Cheng, and Jiun-In Guo," A 252Kgates/4.9Kbytes SRAM/71mW Multi-Standard Video Decoder for High Definition Video Applications," ACM Transactions on Design Automation of Electronic Systems, vol. 14, no. 1, January 2009.
[8] Yao-Chang Yang, and Jiun-In Guo, "A High Throughput H.264/AVC High Profile CABAC Decoder for HDTV Applications," IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 9, pp. 1395-1399, September 2009.
[9] Wei-Chun Ku, Shu-Hsuan Chou, Jui-Chin Chu, Chi-Lin Liu, Tien-Fu Chen, Jiun-In Guo and Jinn-Shyan Wang, "VisoMT: A Collaborative Multithreading Multicore Processor for Multimedia Applications with a Fast Data Switching Mechanism," IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 11, pp. 1633-1645, November 2009.
[10] Hsiu-Cheng Chang, Jia-Wei Chen, Bing-Tsung Wu, Ching-Lung Su, Jinn-Shyan Wang, and Jiun-In Guo, " A Dynamic Quality-Adjustable H.264 Video Encoder for Power-Aware Video Applications," IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, no. 12, pp. 1739-1754, Dec. 2009.
[11] Jinn-Shyan Wang, Pei-Yao Chang, Tai-Shin Tang, Jia-Wei Chen, and Jiun-In Guo, " Design of Subthreshold SRAMs for Energy-Efficient Quality-Scalable Video Applications," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 1, no. 2, pp. 183-192, June 2011.
[12] Jia-Wei Chen, Hsiu-Cheng Chang, Jinn-Shyan Wang, and Jiun-In Guo," A Dynamic Quality-Adjustable H.264 Intra Coder," IEEE Transactions on Consumer Electronics, vol. 57, no. 3, pp. 1203-1211, August, 2011.
[13] Cheng-An Chien, Guo-An Jian, Hsiu-Cheng Chang, Kuan-Hung Chen, and Jiun-In Guo,"High Efficiency Data Access System Architecture for Deblocking Filter Supporting Multiple Video Coding Standards," IEEE Transactions on Consumer Electronics, vol. 58, no. 2, pp. 670-678, May 2012.
[14] C. C. Lin, J. I. Guo, H. C. Chang, Y. C. Yang, J. W. Chen, M. C. Tsai, and J. S. Wang, "A 160kGates 4.5KB SRAM H.264 Video Decoder for HDTV Applications," 2006 IEEE International Solid-State Circuits Conference, Digest of Technical Papers, Session 22.3, pages 9-11, Feb. 2006.
[15] Chih-Da Chien, Chien-Chang Lin, Yi-Hung Shih, He-Chun Chen, Chia-Jui Huang, Cheng-Yen Yu, Chih-Liang Chen, Ching-Hwa Cheng, and Jiun-In Guo, "A 252K Gates/71mW Multi-Standard Multi-Channel Video Decoder for High Definition Video Applications," Proc. 2007 IEEE International Solid-State Circuits Conference, Feb. 2007.
[16] Hsiu-Cheng Chang, Jia-Wei Chen, Ching-Lung Su, Yao-Chang Yang, Yao Li, Chun-Hao Chang, Ze-Min Chen, Wei-Sen Yang, Chien-Chang Lin, Ching-Wen Chen, Jinn-Shan Wang, and Jiun-In Guo, "A 7mW~183mW Dynamic Quality-Scalable H.264 Video Encoder Chip," Proc. 2007 IEEE International Solid-State Circuits Conference, Feb. 2007.
[17] Tay-Jyi Lin, Cheng-An Chien, Pei-Yao Chang, Ching-Wen Chen, Po-Hao Wang, Ting-Yu Shyu, Chien-Yung Chou, Shien-Chun Luo, Jiun-In Guo, Tien-Fu Chen, Yuan-Hua Chu, Liang-Chia Cheng, Hong-Men Su, Chewnpu Jou, Meikei Ieong, Cheng-Wen Wu, Gene C.H. Chuang, Jinn-Shyan Wang, "A 0.48V 0.57nJ/Pixel Video Recording SoC in 65nm CMOS," Proc. 2013 IEEE International Solid-State Circuits Conference, Digest of Technical Papers, Session 9.3, Feb. 2013.
[18] Sheng-Wei Hsu, Guan-Yu Chen, Po-Chun Shen, and Jiun-In Guo, "Dynamic Local Contrast Enhancement for Advanced Driver Assistance System in Harsh Environments," Proc. IEEE International Conference on Consumer Electronics-Taiwan (IEEE ICCE-TW 2014), May 26-28, 2014, Taipei, Taiwan, 2014.
[19] Guan-Yu Chen, Po-Chun Shen, Chao-Yi Cho, Vinay M.S, and Jiun-In Guo," A Forward Collision Avoidance System Adopting Multi-feature Vehicle Detection," Proc. IEEE International Conference on Consumer Electronics-Taiwan (IEEE ICCE-TW 2014), May 26-28, 2014, Taipei, Taiwan, 2014.
[20] Che-Cheng Li, Sheng-Wei Hsu, Po-Chun Shen and Jiun-In Guo," A Single-Camera High Dynamic Range Technique by Using Contrast Enhancement and Exposure Control," Proc. APISPA ASC 2014, Siem Reap, city of Angkor Wat, Cambodia, Dec. 9-12, 2014.
[21] Po-Hsiang Huang, Yuan-Hsiang Maio and Jiun-In Guo," High Dynamic Range Imaging Technology for Micro Camera Array," Proc. APISPA ASC 2014, Siem Reap, city of Angkor Wat, Cambodia, Dec. 9-12, 2014.
[22] Yi-Ting Lin, Ting Chou, Vinay M.S, and Jiun-In Guo," Algorithm derivation and its embedded system realization of speed limit detection for multiple countries," Proc. 2016 IEEE Int'l Symposium on Circuits & Systems, Montreal, Canada, May 22-26, 2016.
[23] Chun-Yu Chung, Yi-Ting Lai, and Jiun-In Guo," Design and Implementation of a Dangerous Driving Behavior Analysis System", Proc. 2016 International Symposium on VLSI Design, Automation & Test (VLSI-DAT), April 25-27, Hsinchu, Taiwan, 2016.
[24] Chia-Chi Tsai, Yi-Ting Lai, Yuan-Fu Li, and Jiun-In Guo, "A Vision Radar System for Car Safety Driving Applications," Proc. 2017 International Symposium on VLSI Design, Automation & Test (VLSI-DAT), April 24-27, Hsinchu, Taiwan, 2017.
[25] Ting Chou, Ssu-Yuan Chang, Vinay M.S. and Jiun-In Guo, "Triangular Road Signs Detection and Recognition Algorithm and its Embedded System Implementation," Proc. The 21th International Conference on Image Processing, Computer Vision & Pattern Recognition (IPCV'17), July 17-20, 2017, Las Vegas, USA.
[26] Yuan-Fu Li, Chia-Chi Tsai, Yi-Ting Lai, and Jiun-In Guo," A Multiple-Lane Vehicle Tracking Method for Forward Collision Warning System Applications," Proc. APSIPA ASC 2017, Kuala Lumpur, Malaysia, 12-15 Dec. 2017.
[27] Po-Chun Shen, Kuan-Hung Chen, Jui-Sheng Lee, Guan-Yu Chen, Yi-Ting Lin, Bing-Yang Cheng, Guo-An Jian, Hsiu-Cheng Chang, Wei-Ming Lu, and Jiun-In Guo, "Intelligent Vision Processing Technology for Advanced Driver Assistance Systems," A book chapter in Smart Sensors and Systems: Innovations for Medical, Environmental, and IoT Applications, ISBN 978-3-319-33200-0, pp.175-206, Springer, 2016.

[28] Jiun-In Guo, Chia-Chi Tsai, Jian-Lin Zeng, Shao-Wei Peng, and En-Chih Chang," Hybrid Fixed Point/Binary Deep Neural Network Design Methodology for Low Power Object Detection," IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS), vol. 10, no. 3. pp. 388-400, September 2020.
[29] Guan-Ting Lin, Vinay Malligere Shivanna, and Jiun-In Guo, "A Deep Learning Model with Task-Specific Bounding Box Regressors and Conditional Back-Propagation for Moving Object Detection in ADAS Applications," Sensors, special issue on "Sensor and Communication Systems Enabling Autonomous Vehicles", vol. 20, no. 18, pp. 1-21, Sept. 2nd, 2020.
[30] Wen-Chia Tsai, Jhih-Sheng Lai, Kuan-Chou Chen, Vinay M. Shivanna and Jiun-In Guo, "A Lightweight Motional Objects Behavior Prediction System Harnessing Deep Learning Technology for Embedded ADAS Applications," Electronics, Electrical and Autonomous Vehicles session, special issue Autonomous Vehicles Technology, vol. 10, issue 6, pp. 692, 2021.
[31] Chun-Yu Lai, Bo-Xun Wu, Vinay M. Shivanna, and Jiun-In Guo," MTSAN: Multi-Task Semantic Attention Network for ADAS Applications," IEEE Access, vol. 9, pp. 50700-50714, 2021.
[32] Chia-Chi Tsai and Jiun-In Guo," IVS-Caffe – Hardware-Oriented Neural Network Model Development," IEEE Transactions on Neural Networks and Learning Systems, Vol. 33, Issue 10, pp. 5978-5992, Oct. 2022.
[33] Tzu-Hsien Sang, Feng-Tsun Chien, Chia-Chih Chang, Kuan-Yu Tseng, Bo-Sheng Wang, and Jiun-In Guo," DoA Estimation for FMCW Radar by 3D-CNN," Sensors, vol. 21, no. 16, Aug. 2021.
[34] Yu-Shu Ni, Vinay Malligere Shivanna, Jiun-In Guo," iVS Dataset and ezLabel: A Dataset and a Data Annotation Tool for Deep Learning based ADAS Applications," Remote Sensing, 2022, 14, 833. Feb. 10, 2022.
[35] Tzu-Hsien Sang, Kuan-Yu Tseng, Feng-Tsun Chien, Chia-Chih Chang, Yi-Hsin Peng, and Jiun-In Guo, "Deep Learning-based Velocity Estimation for FMCW Radar with Random Pulse Position Modulation," IEEE Sensors Letters, Vol. 6, Issue 3, March 2022.
[36] Hung-Wei Lin, Vinay M. Shivanna, Hsiu Chi Chang, and Jiun-In Guo," Real-Time Multiple Pedestrian Tracking with Joint Detection and Embedding Deep Learning Model for Embedded Systems," IEEE Access, May 9th, 2022.
[37] Bo-Xun Wu, Vinay M. Shivanna, Hsiang-Hsuan Hung, Jiun-In Guo," ConcentrateNet: Multi-Scale Object Detection Model for Advanced Driving Assistance System Using Real-Time Distant Region Locating Technique," Sensors 2022, 22, 7371.
[38] Chun-Yu Lai, Bo-Xun Wu, Tsung-Han Lee, Vinay, and Jiun-In Guo," A Light Weight Multi-Head SSD Model for ADAS Applications," Proc. The International Conference on Pervasive Artificial Intelligence (ICPAI2020), Taipei, Taiwan, Dec. 3-5, 2020.
[39] Bo-Xun Wu, Jia-Jheng Lin, Hsien-Kai Kuo, Po-Yu Chen, and Jiun-In Guo, "Radar and Camera Fusion for Vacant Parking Space Detection," Proc. IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS2022), Songdo Convensia, Incheon, Korea, June 13-15, 2022.
[40] An-Tai Hsiao, Chun-Hsien Liu, Po-Hsuan Chen, Yao-Lun Liu, Wei-Chi Wang, Tzu-Hsien Sang, Chia-Ming Tsai, Gray Lin, Jiun-In Guo, and Sheng-Di Lin," Real-time LiDAR module with 64x128-pixel CMOS SPAD array and 940-nm PCSEL," Proc. 2022 IEEE Sensors Applications Symposium (SAS), August 1-3, 2022, Sundsvall, Sweden.
[41] Po-Yuan Chen, Hung-Che Lin, Jiun-In Guo," "Multi-Scale Dynamic Fixed-Point Quantization and Training for Deep Neural Networks", Proc. 2023 IEEE International Symposium on Circuits & Systems (ISCAS), Monterey, May 21 - 25, 2023, California, USA.
[42] Qualcomm AI Model Efficiency Toolkit (AIMET), https://developer.qualcomm.com/software/ai-model-efficiency-toolkit
[43] Qualcomm AMIET AI Model Efficiency Toolkit (AIMET) - GitHub, https://github.com/quic/aimet
[44] Jiun-In Guo, Chia-Chi Tsai, Yong-Hsiang Yang, Hung-Wei Lin, Bo-Xun Wu, Ted T. Kuo, and Li-Jen Wang," Summary-Embedded Deep Learning Object Detection Model Competition in IEEE MMSP 2019," Proc. IEEE 21st International Workshop on Multimedia Signal Processing (MMSP 2019), Kuala Lumpur, Malaysia, Sept. 27-29, 2019.
[45] Chia-Chi Tsai, Yong-Hsiang Yang, Hung-Wei Lin, Bo-Xun Wu, En Chih Chang , Hung. Yu Liu, Jhih-Sheng Lai, Po Yuan Chen, Jia-Jheng Lin, Jen Shuo Chang, Li-Jen Wang, Ted T. Kuo, Jenq-Neng Hwang, and Jiun-In Guo, "The 2020 Embedded Deep Learning Object Detection Model Compression Competition for Traffic in Asian Countries," Proc. 2020 IEEE International Conference on Multimedia and Expo (ICME2020), July 6-10, 2020, London, United Kingdom.
[46] Yu-Shu Ni, Chia-Chi Tsai, Bo-Xun Wu, Po-Chi Hu, Ted T. Kuo, Jenq-Neng Hwang, Po-Yu Chen, Hsienkai Kuo, and Jiun-In Guo, "SUMMARY ON THE 2021 EMBEDDED DEEP LEARNING OBJECT DETECTION MODEL COMPRESSION COMPETITION FOR TRAFFIC IN ASIAN COUNTRIES," Proc. 2021 ACM International Conference on Multimedia Retrieval (ICMR2021), Taipei, Taiwan, August 21-24, 2021.
[47] Yu-Shu Ni, Chia-Chi Tsai, Chih-Cheng Chen, Po-Yu Chen, Hsien-Kai Kuo, Man-Yu Lee, Chin-Chuan, Kuo, Zhe-Ln Hu, Po-Chi Hu, Ted T. Kuo, Jenq-Neng Hwang, and Jiun-In Guo," Summary of the 2022 Low-Power Deep Learning Semantic Segmentation Model Compression Competition for Traffic Scene in Asian Countries," Proc. 2022 IEEE International Conference on Multimedia and Expo (ICME2022), July 18-22, 2022, Taipei, Taiwan.

MULTIMEDIA COMPACT REPRESENTATIONS: APPROACHES, APPLICATIONS AND CHALLENGES

Jingkuan Song
jingkuan.song@gmail.com
Jingkuan Song is a full professor with University of Electronic Science and Technology of China (UESTC) from 2017. He was the winner of ACM SIGMM Rising Star Award 2021 for his continuous contributions to Multimedia Compact Representation and Analysis.

Xiaosu Zhu
xiaosu.zhu@outlook.com
Xiaosu Zhu is a Ph.D student at University of Electronic Science and Technology of China (UESTC) from 2020. His research interests include compact representation learning, multimedia retrieval and compression.

A. Introduction

Nowadays, a rapid increase on contents of images, videos and other media raises challenges on multimedia processing. Although modern computers or mobile devices have much more computational and storage resources than before, a system for handling billion-scale medias still requires excessive power beyond them. An efficient and performant processing pipeline is needed to meet such conditions.

Compact representations and corresponding technology are designed for such a large-scale multimedia processing, including image and video compression [1,2,3], retrieval [4,5,6], synthesis [7,8,9,27], etc. Other than full-precision data, compact representations use binary or low-bit values to formulate feature vectors, operators, and neural network weights. The storage is reduced by 10-100x while processing speed is accelerated via XOR, bit-count, lookup-table and other operators with hardware support [10]. By applying it to the above multimedia systems, the severe challenge of storage space and power consumption is largely alleviated.

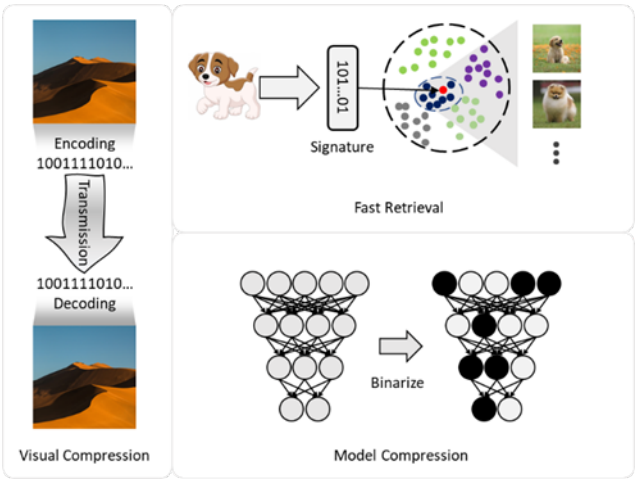


Figure 1. Typical applications of compact representations.

Today's compact representation algorithms, e.g., quantization and hashing are applied in a significant

wide space with heavy development. For visual compression, we could observe a steady improvement on compression ratio and quality from the classical JPEG image compression codec [1] to the latest neural image compression networks [2]. Similarly, a lot of vector libraries are emerged for large-scale data indexing and retrieval, e.g., faiss [6], milvus [11]. Thanks to the support of compact representations, a search in the billion-scale database is reduced to a few milliseconds and is conducted completely in memory. The recently popular tokenized image synthesis models VQ-GAN [8] and variants have initiated the new fashion of Artificial Intelligence Generated Contents (AIGC). The compact image tokens are obtained by a learnt vector quantizer, which hold abundant visual and semantic information to generate amazing arts, photographs and cartoons. Note that all the above techniques involve compact representations and make a great contribution to the current multimedia society in terms of power consumption, storage, etc. Next, a brief introduction of the fundamental techniques for them is brought.

B. Approaches

A regular vector held in computer is represented as an array of floating-point numbers, 32 bits for each. The goal of generating compact representation is to compress it into short binaries via hashing or quantization and keeps key information preserved.

B.1. Hashing

A hash function reduces high-dimensional data to a few fixed binary values, where similar data are allocated into similar hash values to keep the affinities. Therefore, the data distribution could be partially preserved by the converted binaries. It is commonly used for data clustering and nearest neighbor search. The famous Locality Sensitive Hashing (LSH) family [5] takes a concept to assign data into “buckets” while data collisions are maximized. Similarly, MinHash [12], Random Projection [13], etc. are designed in a data-independent way to achieve the goal, while Locality-Preserving Hashing performs in a data-dependent way. Furthermore, by utilizing semantic information to help data hashing, we could map data inputs according to their semantic similarity.

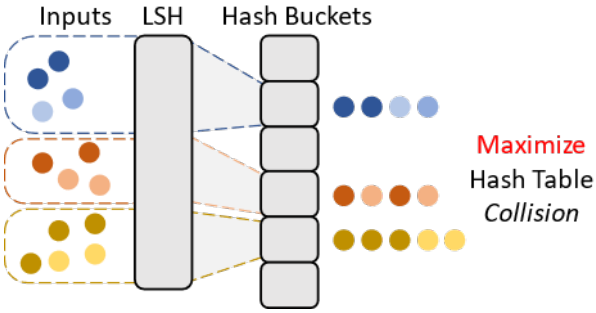


Figure 2. Demonstration of LSH family. They try to maximize collision of similar inputs in the hash table.

There is another approach to obtain hashing function that by several typical machine learning algorithms, called learning to hash [4]. Other than hand-crafted hashing rules, this kind of method uses mapping model, such as linear projection, to reduce feature dimension and binarize output by sign function. Neighborhood structure from input to output is preserved by defining objective function and performing optimization on mapping model. Then, the implicit hashing function is learnt and held in model.

Optimizing hashing models are not easy since finding an appropriate hashing output involves discrete and combinatorial optimization. Thus, a few heuristic algorithms are employed for it, e.g., coordinate descent in supervised discrete hashing, or a smooth relaxation over sign function to make it differentiable.

B.2. Quantization

Quantization acts like hashing since it also converts the raw full-precision data into a series of binary codes. While hashing directly makes comparison over hashing outputs, i.e., Hamming distance between different hashing codes, quantization uses a codebook to reconstruct original data while successive computations are happened in the reconstructed feature space.

Specifically, a scalar quantization is simple yet efficient since the implementation is rather naïve. It directly performs a rounding on floating-point numbers. Extended from this, a normal vector

quantization maps vectors to its nearest centroid, which is obtained by k-means [14]. Firstly, data are clustered by a few codewords, which are the representative vector on behalf of all neighborhoods. Combining all the codewords we obtain a codebook. Then, each input data is replaced by its nearest codeword and is stored by the index of codeword. Therefore, the quantization result of each vector is a single value, whose upper-bound is the total amount of codewords. It represents as fixed-length binaries in the storage.

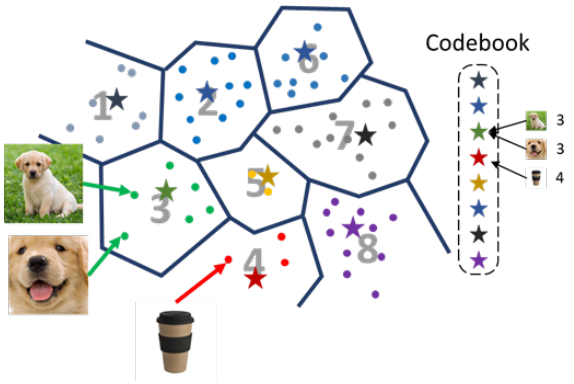


Figure 3. Demonstration of vector quantization. By clustering over input data, centers are collected as codebook. Every input is assigned to its nearest center.

The classical vector quantization is effective and has been used for a long time, but it has disadvantages in computational and space complexity, which is linearly related to the codebook size. The quantization precision is also limited with the total training data points since codebook size could not be larger than training set size. Therefore, product quantization is developed to fix the above issues by splitting raw data into a few orthogonal feature subspaces and quantizing separately by independent codebooks [15]. Therefore, the above complexity is reduced to logarithm-scale and the training becomes flexible. Further variants on vector quantization include composite quantization [16], additive quantization [17], etc. Note that the latter would be the super set of the former for the aforementioned algorithms with an improved performance.

C. Applications

As above demonstrates, we have two powerful tools for generating compact representations. Both of them have a long-time attraction and are under heavy development. Correspondingly, there are a lot of applications to integrate them into novel multimedia systems.

C.1. Visual Compression

Compact representations play the fundamental role in the pipeline of image compression, video compression and other widespread data compression techniques, since redundancies naturally exist in visual data and could be compressed by compact representations. For instance, a large area of blue sky contains low-frequency information mostly and could be represented by very few bits of a single color.

Conventional codecs for visual compression consist of transform coding, quantization, entropy coding, etc. Such practices have been confirmed as standards in modern image/video coding standards such as JPEG, H.264, VVC, etc [3].

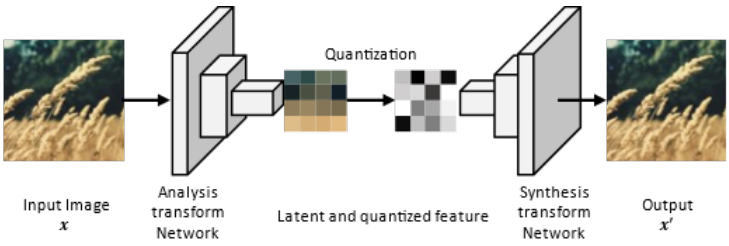


Figure 4. Typical operation diagram of a neural image compressor.

With the development of deep learning, compressing visual data by a neural network becomes a new fashion. It replaces the pre- and post-transform coding by the neural network to convert and reconstruct images to/from quantized latent features in a variational auto-encoder style. Meanwhile, a Straight-Through Estimator (STE) is applied on the quantization operation to enable end-to-end training via stochastic gradient descent. With the capacity of weights in neural networks and design of entropy models, deep image compression achieves superior performance than conventional ones. Recent works also extend quantization operations to multi-codebook vector quantization to significantly enhance coding efficiency [18]. With these methods, one could obtain 100-200x compression ratio compared to raw images without perceptible artifacts.

C.2. Fast Retrieval

Adopting hashing and quantization to perform fast retrieval is suitable since they could find approximate nearest neighbors in an efficient way with power of hardware acceleration. Moreover, the inverted file index could be built upon them to further increase search speed with a non-exhaustive search. Therefore, for million-scale and larger databases, several vector search software toolkits are developed for production, e.g., faiss, milvus. A huge number of applications are boosted by them, including but not limited to search engines, e-commerce, database indexing.

We would introduce a typical application: fast image retrieval, which digests images into short binary codes to formulate a database and provides ways to find similar images of a query. To achieve this, we could directly take image descriptors such as SIFT, GIST to extract features of images and obtain compact representations with the above methods. On the other hand, with the help of deep neural networks, an end-to-end fast image retrieval model could be made up with a network as backbone and a hashing/quantization layer on top of it [19]. Objective functions are designed to increase intra-class similarity and decrease inter-class similarity based on training image labels. Then, the images of same category are clustered together in order to organize the database. During retrieval, query image is firstly transformed into binary code in the same way and search by distance in the compact feature space, which could be accelerated by XOR (hashing) or look-up tables (quantization).

C.3. Tokenized Image Synthesis and Recognition

Another important application that involves compact representations is tokenized image synthesis. Formally, it utilizes and quantizes feature vectors extracted from a visual model, typically neural networks, and use the quantized results for further generation or recognition. Such an operation is dubbed tokenization.

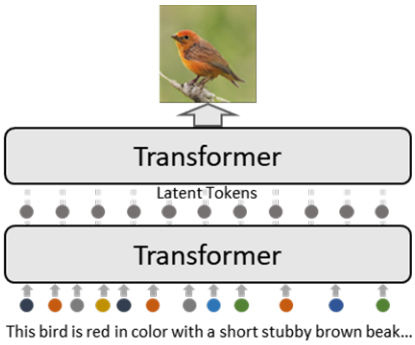


Figure 5. Demonstration of text to image synthesis by using tokens and transformers.

Specifically, VQ-VAE [7] constructs a VAE model while latent features produced by encoder are quantized by a codebook. By controlling the quantization value, the decoder could generate fake images. VQ-GAN uses transformer to enhance the generation ability upon the former. It also bridges multi-modal visual generation by unifying them into codewords.

With guide of quantized latent features of images, we would also perform visual recognitions. DALL-E [9], BeiT [20], CLIP [21], etc. bring the idea to produce classification results by formulating image tokens as sequence and performing sequence-to-sequence translation by transformers. They achieve satisfying performance and outperform convolutional networks of similar model size significantly.

C.4. Model Compression

The above applications utilize deep neural networks to obtain powerful and abundant latent representations and achieve better performance than conventional machine learning algorithms, thus become the current fashion. However, the model size and inference time of deep networks block them from running on devices with limited resources, such as mobile phones. Therefore, such demands make model compression become an important topic which benefits for deployment in real world.

Involved in model compression, binarizing or quantizing the model weights and intermediate activations are basic ways for reducing model size. It works since such values are also floating-point

vectors or matrices that could be converted into compact representations. By using hashing and scalar quantization techniques, the compression procedure is approximated by STE or relaxing for optimization. After training, a model with low-bit weights and activations would be obtained, which has a small size and the inference could be accelerated by XOR and bit-count operations [22]. To achieve good performance, knowledge distillation and weights regularization are adopted. The former forces small model to have similar activations with large models, while the latter prevents model to output trivial results [23].

D. Challenges

As a technique that receives attention in a long time, compact representations are now come with several powerful toolkits. Models or hand-crafted algorithms are well developed and could produce desired performance. However, there still has a few scenarios that previous works rarely study or could not handle well, especially for the realistic tasks. Next, we would explain them in detail.

Since discrete optimization over binaries is involved in compact representations, which is generally NP-hard, a lot of works focus on finding a solid solution to give near-optimal results. There are several remaining issues. For hand-crafted approaches, they either add constraints on it, which blocks them from the global-optima, or take a high time or space complexity for solving the problem. For deep learning based approaches, they would be trapped in local-optima where hashing layer produces trivial results or quantization falls into the “codebook collapse” problem. A few works try to tackle the above issues such as UNQ [24] and SQ-VAE [25], but a theoretical study on them remains rarely explored.

Due to the sensitivity of compact representations, i.e., data would be assigned with wrong binary value when inputs have distribution shift, current works result in bad performance especially when they meet novel inputs that are from other domains or categories. Note that a fine-tuning on models is not feasible since the new model would have gaps on representations with the old one, especially for tasks like fast retrieval, meaning that the retrieval database has to be re-built. Unfortunately, there are few works to study this problem [26].

This article makes a brief introduction to current advances in compact representations. Typical approaches such as hashing and quantization and corresponding optimization algorithms are explained. Applications that involve compact representations are introduced, while possible challenges are demonstrated. We could confirm that compact representations have wide use cases and potential improvements exist as the current research goes on.

References

[1] Gregory K. Wallace: The JPEG Still Picture Compression Standard. Commun. ACM 34(4): 30-44 (1991).
[2] Johannes Ballé, Valero Laparra, Eero P. Simoncelli: End-to-end Optimized Image Compression. ICLR 2017.
[3] Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, Shanshe Wang: Image and Video Compression With Neural Networks: A Review. IEEE Trans. Circuits Syst. Video Technol. 30(6): 1683-1698 (2020).
[4] Jingdong Wang, Ting Zhang, Jingkuan Song, Nicu Sebe, Heng Tao Shen: A Survey on Learning to Hash. IEEE Trans. Pattern Anal. Mach. Intell. 40(4): 769-790 (2018).
[5] Omid Jafari, Preeti Maurya, Parth Nagarkar, Khandker Mushfiqul Islam, Chidambaram Crushev: A Survey on Locality Sensitive Hashing Algorithms and their Applications. ArXiv preprint. 2102.08942 (2021).
[6] Jeff Johnson, Matthijs Douze, Hervé Jégou: Billion-Scale Similarity Search with GPUs. IEEE Trans. Big Data 7(3): 535-547 (2021).
[7] Aäron van den Oord, Oriol Vinyals, Koray Kavukcuoglu: Neural Discrete Representation Learning. NIPS 2017: 6306-6315.
[8] Patrick Esser, Robin Rombach, Björn Ommer: Taming Transformers for High-Resolution Image Synthesis. CVPR 2021: 12873-12883.
[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever: Zero-Shot Text-to-Image Generation. ICML 2021: 8821-8831.
[10] H. Naseri and S. Timarchi, Low-power and fast full adder by exploring new XOR and XNOR gates. IEEE Trans. Very Large Scale Integr. Syst. 26(8): 1481-1493, 2018.
[11] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, Kun Yu, Yuxing Yuan, Yinghao Zou, Jiquan Long, Yudong Cai, Zhenxiang Li, Zhifeng Zhang, Yihua Mo, Jun Gu, Ruiyi Jiang, Yi Wei, Charles Xie: Milvus: A Purpose-Built Vector Data Management System. SIGMOD Conference 2021: 2614-2627.
[12] Andrei Z. Broder, Moses Charikar, Alan M. Frieze, Michael Mitzenmacher: Min-Wise Independent Permutations (Extended Abstract). STOC 1998: 327-336.
[13] Ella Bingham, Heikki Mannila: Random projection in dimensionality reduction: applications to image and text data. KDD 2001: 245-250.
[14] Dana H. Ballard: An introduction to natural computation. Complex adaptive systems, MIT Press 2000, pp. I-XXII, 1-306.

[15] Hervé Jégou, Matthijs Douze, Cordelia Schmid: Product Quantization for Nearest Neighbor Search. IEEE Trans. Pattern Anal. Mach. Intell. 33(1): 117-128 (2011).
[16] Ting Zhang, Chao Du, Jingdong Wang: Composite Quantization for Approximate Nearest Neighbor Search. ICML 2014: 838-846.
[17] Artem Babenko, Victor S. Lempitsky: Additive Quantization for Extreme Vector Compression. CVPR 2014: 931-938.
[18] Xiaosu Zhu, Jingkuan Song, Lianli Gao, Feng Zheng, Heng Tao Shen: Unified Multivariate Gaussian Mixture for Efficient Neural Image Compression. CVPR 2022: 17591-17600.
[19] Tan Yu, Junsong Yuan, Chen Fang, Hailin Jin: Product Quantization Network for Fast Image Retrieval. ECCV 2018: 191-206.
[20] Hangbo Bao, Li Dong, Songhao Piao, Furu Wei: BEiT: BERT Pre-Training of Image Transformers. ICLR 2022.
[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021: 8748-8763.
[22] Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, Nicu Sebe: Binary neural networks: A survey. Pattern Recognit. 105: 107281 (2020).
[23] Haotong Qin, Ruihao Gong, Xianglong Liu, Mingzhu Shen, Ziran Wei, Fengwei Yu, Jingkuan Song: Forward and Backward Information Retention for Accurate Binary Neural Networks. CVPR 2020: 2247-2256.
[24] Stanislav Morozov, Artem Babenko: Unsupervised Neural Quantization for Compressed-Domain Similarity Search. ICCV 2019: 3036-3045.
[25] Yuhta Takida, Takashi Shibuya, Wei-Hsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, Yuki Mitsufuji: SQ-VAE: Variational Bayes on Discrete Representation with Self-annealed Stochastic Quantization. ICML 2022: 20987-21012.
[26] Soumava Paul, Titir Dutta, Soma Biswas: Universal Cross-Domain Retrieval: Generalizing Across Classes and Domains. ICCV 2021: 12036-12044.
[27] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, Baining Guo: Learning texture transformer network for image super-resolution. CVPR 2020: 5791-5800.