인공신경망을 이용한 압축기술 CfP의 평가결과

천승문 인시그널 기술연구소장, NNR AHG의장

1. 머리말

2019년 3월 22일부터 31일까지 스위스 제네바에서 개최된 ISO/IEC JTC1 SC29/WG11 126차 MPEG 회의에서 인공신경망(Neural Networks)의 압축표현(Compressed Representation) 국제표준에 Call for Proposal(CfP)에 대하여 9개 기고서가 제출되었다. 참여기관은 중국의 화웨이, 독일의 HHI, 핀란드의 노키아 등이다. 한국에서는 한국항공대와 인시그널이 기고서를 제출하였다. 각 기관의 기고서를 바탕으로 공동기술에 대한 성능을 평가하고, 이를 근거로 차기 회의에서 새로운 추가기고와 Cross check를 통해 검증작업을 진행하기로 하였다.

인공신경망을 이용한 압축기술은 훈련된 인공신경망에 대해 압축되고 해석 가능하며 상호 운용 가능한 표현을 정의하는 것을 목표로 한다. 표현은 다음을 할 수 있어야 한다.

- 다양한 인공 신경망 유형(예: CNN 및 자동 인코딩과 같은 피드 포워드 네트워크, LSTM과 같은 반복 네트워크 등)
- 원래의 네트워크보다 더 빨리 추론 할 수 있도록 원래의 네트워크를 완전히 재구성하지 않고 추론 할 수 있다.
- 리소스 제한 (계산, 메모리, 전력, 대역폭) 하에서 사용 가능해야 한다.

2. 회의 주요 결과

아래와 같이 9개의 기고서를 바탕으로 각 기술의 내용을 분석한 이후에 공통적인 기술내용을 취합하여 기술분류표를 작성하였다.

- **P1**: m47375 Response to the Call for Proposal on Neural Network Compression: Simultaneously Learning Architectures and Features of Deep Neural Networks(핀란드 노키아)
- **P2**: m47379 Response to the Call for Proposals on Neural Network Compression: Training Highly Compressible Neural Networks(핀란드 노키아)
- **P3**: m47412 Universal Compression Platform for Neural Network Compression(중국 Zhejiang대)

- **P4**: m47491 Huawei's response to the Call for Proposal on Neural Network Compression(중국 화웨이)
- **P5**: m47493 Response to the Call for Proposals on Neural Network Compression, Low Displacement Rank based compression of Deep Neural Networks(미국 테크니칼라)
- **P6**: m47634 Response to the Call for Proposal on Neural Network Compression with Adaptive Quantization(중국 Peking대/홍콩시립대)
- **P7**: m47698 Response to the Call for Proposals on Neural Network Compression: End-to-end processing pipeline for highly compressible neural networks(독일 HHI)
- **P8**: m47704 Response to the Call for Proposals on Neural Network Compression: Quantization and Low-Rank Approximation(한국 인시그널/한국항공대)
- **P9**: m47929 Proposal for Coefficients Quantization of Convolutional Neural Network for ASIC Implementation(미국 gyrfalcon tec)

	P1	P2	P3	P4	P5	P6	P7	P8	P9
Methods									
Pruning structure	X		X*						
Sparsification		X		X			X		
Use loss function Supporting compressability		X					X		
Quantization (using labelled data)							X		
Quantization (using unlabeled data)				X					
Quantization (using weights)		X		X		X		X	
Quantization (data independent)		X	X	X				X	X
Quantization (rate distortion based)				X			X		
Quantization using codebook		X		X		X		X	
Uniform reconstruction				X			X		X
Matrix decomposition					X			X	
Layer partitioning				X					
Layer representation				X					
Entropy coding	X	X	X	X	X	X	X		
Properties									
Data-dependent transformations as preprocessing[1] (required)	X	X							
Data-dependent transformations as preprocessing[2] (optional)							X		
Fine-tuning between or after compression steps (optional)			X	X	X	X	X	X	X
Decompression step		X	X	X		X	X	X	
Partial decompression supported				X			X		
Reconstructed complete network smaller than original	X		X		X			(X)	
Additional features									
Framework		X	X	X			X		

3. 향후 전망

126차 MPEG 회의부터는 인공신경망 국제표준을 ISO/IEC 15938-17로 분류하였다. 표준의 제목은 'Part 17: Compression of neural networks for multimedia content description and analysis'이고 에디터로 오스트리아의 Wener와 한국의 천승문 소장이 선정되었다. 목표로 하는 국제표준화 일정은 2020년 4월 CD 단계, 2020년 10월 DIS 단계, 2021년 4월 FDIS 단계, 2021년 10월 IS 단계로 진행하기로 하였다. 또한 SC42 등 타 인공신경망 표준화 단체에 관련 기술을 소개하는 레터를 보내서 참고하도록 하였다. 127차 차기 회의부터는 추가로 새로운 기술기고서를 받고 상호관심기관에서 공동으로 검증작업을 거쳐서 최상의 압축기술을 국제표준화 할 예정이다.