인공신경망 압축 표현(MPEG NNR) 표준화 본격 시작

김재곤 한국항공대학교 항공전자정보공학부 교수

1. 개요

ISO/IEC JTC 1/SC 29 산하 멀티미디어부호화 국제 표준화위원회(WG 11, MPEG)는 딥러닝(Deep Learning)에 활용되는 인공신경망(Neural Network) 모델을 압축 표현하는 NNR(Compression of Neural Networks for Multimedia Content Description and Analysis, MPEG-7 Part 17) 표준화를 진행하고 있다. NNR은 2019년 1월 제125차 마라케시 MPEG 회의에서 최종 기술제안 요청서(CfP: Call for Proposal)를 공표하였고, 이번 제126차 제네바 회의에서 CfP 응답 제안 기술들을 평가하고 2021년 4월 표준완료(FDIS)를 목표로 본격적인 표준화를 시작하였다. NNR은 다양한 인공지능 프레임워크(Tensorflow, Pytorch 등) 및 HW플랫폼간의 상호호환적인 모델 표현을 제공할 뿐만 아니라 모델 압축을 통하여 모바일 환경과 실시간 응용에서의 용이한 NN 구현을 가능하게 한다. 또한 영상분류, 영상/비디오 압축 등의 여러 유스 케이스(Use Cases)들의 요구사항을 충족할 수 있는 표준 기술을 포함하도록 하여 딥러닝을 활용한 다양한 응용분야에서 활용할 수 있는 인공신경망 압축표현 표준으로 기대되고 있다.

이번 제126차 제네바 회의에서는 지난 회의에 공표한 CfP의 응답으로 제안된 8개 기관(HHI, Technicolor, Nokia, Huawei, 북경대(PKU), 저장대(ZJU), 한국항공대(KAU)/인시그널, GTI)의 9개의 응답기술들을 검토 평가하였고, 이를 바탕으로 제안 기술들의 비교평가를 위해 3개의 CE(Core Experiment)를 설정하였다. 또한, CE의 평가를 위해 CTC(Common Test Conditions)를 포함한 평가방법(Evaluation Framework)의 문서를 개정하였다. 향후 표준화는 Video 서브그룹에서 진행하기로 하였으며, 2020년 4월 CD, 2021년 4월 FDIS 발간을 목표로 하고 있다.

2. 주요 표준화 이슈 및 진행 결과

2.1 CfP 응답 기술

CfP의 응답에 대한 각 제안 기술들은 평가를 위해 미리 제시된 메모리 사용량, 실행시간, 모델 크기, 압축 성능 그리고 각 유스 케이스에서의 성능들을 보고하였다. 제안된 기술들의 성능은 유스 케이스와 압축 방법에 따라 다르지만, 원 인공신경망 모델의 크기를 평균적으로 약 7~8배 압축됨을 보였으며, 각 유스 케이스에서의 성능은 0~30%의 성능 저하를 보였다.

제안된 CfP 응답 기술(기고번호, 제안기관, 주요 제안기술)은 다음 <표 1> 과 같다.

<표 1> 제안된 CfP 응답 기술(기고번호, 제안기관, 주요 제안기술)

P1 Nokia	· (m47375) Response to the Call for Proposal on Neural Network Compression: Simultaneously Learning Architectures and Features of Deep Neural Networks
	· Pruning & Sparse coding
P2 Nokia	· (m47379) Response to the Call for Proposals on Neural Network Compression: Training Highly Compressible Neural Networks
	· Pruning & Re-training
P3 ZJU	· (m47412) Universal Compression Platform for Neural Network Compression
	· Pruning & Quantization & Huffman coding
P4 Huawei	· (m47491) Huawei's response to the Call for Proposal on Neural Network Compression
	· Quantization & Entropy coding · Palette, escape index map coding
P5 Technicolor	· (m47493) Response to the Call for Proposals on Neural Network Compression, Low Displacement Rank based compression of Deep Neural Networks
	· Low-rank approximation (dense layer)
P6 PKU	· (m47634) Response to the Call for Proposal on Neural Network Compression with Adaptive Quantization
	· Adaptive quantization
P7 HHI	· (m47698) Response to the Call for Proposals on Neural Network Compression: End-to-end processing pipeline for highly compressible neural networks
	· Sparse coding & CABAC
P8 KAU/Insignal	· (m47704) Response to the Call for Proposals on Neural Network Compression: Quantization and Low-Rank Approximation
	· Quantization · Low-rank approximation (convolution layer)
P9 GTI	· (m47929) Proposal for Coefficients Quantization of Convolutional Neural Network for ASIC Implementation
	· Quantization (fixed-point)

<표 2> 제안서의 압축 기술 요약

Organization	P1	P2	P3	P4	P5	P6	P7	P8	P9
	Nokia	Nokia	ZJU	Huawei	Technicolor	PKU	HHI	KAU	GTI
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		

<표 3> 제안서 주요 특징

Organization	P1	P2	P3	P4	P5	P6	P7	P8	P9
	Nokia	Nokia	ZJU	Huawei	Technicolor	PKU	HHI	KAU	GTI
Properties									
Data-dependent transformations as preprocessing[1] (required)	X	X							
Data-dependent transformations as preprocessing[2] (optional)							X[3]		
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		

위의 <표 2>는 각 제안기술의 세부 압축 방법을 <표 3>은 각 제안기술의 주요 특징을 요약한 것이다. <표 2>와 같이 CfP 응답 기술은 가지치기(Pruning), 가중치 근사화(양자화, Low-Rank 근사화 등) 그리고 엔트로피 부호화의 3가지로 분류되며, 이 3가지 기술 종류에 따라 CE를 설정하고 각 기술 종류의 세부 기술의 비교 검증을 진행하기로 하였다. <표 3>의 주요 특징으로는 가지치기 기법의 특징(중요도 학습 여부, 추가 검증 데이터 사용 여부), 복원 과정의 유무, 복원 모델의 크기 변화 유무 및 복수의 압축 기술을 포함한 프레임워크 제안 여부 등을 나타낸 것이다.

2.2 향후 기술 평가를 위한 CE 설정

이번 회의에서 제안된 CfP 응답 기술들을 기반으로 향후 기술 비교 검증 및 기술 표준화를 위한 CE를 설정하였으며, 이번 회의에서 결정된 CE 3개와 참여할 기관은 다음과 같다.

Core Experiments
- CE1: Pruning/Sparsification (Nokia, HHI)
- CE2: Weight approximation (ZJU, Huawei, HHI, PKU, KAU, GTI)
  - Quantization/Low-rank approximation
- CE3: Entropy coding (Huawei, HHI)

2.3 향후 NNR 계획

이번 회의에서 MPEG Video 그룹과 논의한 NNR의 일정은 다음과 같다.

Timeline on NNR
- CD : 2020년 4월
- DIS : 2020년 10월
- FDIS : 2021년 04월
- IS : 2021년 10월

3. 향후 전망

지난 2019년 1월 회의에서 공표된 인공신경망 압축표현 표준 기술에 대한 MPEG NNR의 CfP 응답 기술제안서 평가를 바탕으로 2021년을 목표로 그 표준화를 시작하였다. 이번 회의에서는 지난 회의에서 공표되었던 CfP의 응답 기술제안서 총 9개가 기고되었으며, 향후 지속적인 표준 기술 채택을 위하여 3개의 CE를 설정하고 다음 회의에 그 결과를 검토할 예정이다.

현재의 진행 결과로 미루어 보아, NNR의 표준화는 데이터 표현 및 인공신경망 압축 기술들에 대한 조사 및 구체화가 더 필요해 보이며, 세계 유수 기관 간의 표준 기술/지적권 선점을 위한 치열한 경쟁이 진행될 것으로 예견된다. 특히, 중국 및 유럽에서 NNR에 대한 기술의 관심도가 높아 적극적으로 대응하고 있다. 현재까지는 국내 기관으로는 인시그널과 한국항공대학교가 CfP 응답기술을 제안하는 등 표준화에 적극적으로 참여하고 있으며 일부 기업, 연구기관 및 대학에서 관심을 가지고 진행 현황을 주목하고 있으며, 향후 추가적으로 표준화에 참여할 것으로 예상된다.