DeepPhy

Exploration of Deep Learning Application to Modem UE Physical Layer Design

Nate Chizgi (nchizgi@stanford.edu)

Stanford University

CS230: Deep Learning

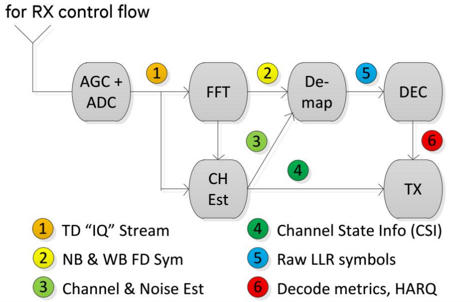
Abstract

Within the wireless physical layer realm, there has been ongoing research and application of deep learning to network side optimizations (self-organizing networks, scheduling algorithms, beamforming, etc), however the user equipment (UE) domain has largely not yet been influenced. The primary challenges in integrating deep learning into the UE physical layer are the impact on device power consumption, silicon area, latency and flexible design needed for rapid technological advancement.

This project summarizes existing research in power and complexity reduction techniques for deep neural network inference in embedded platforms, as well as discussing applications of deep learning on modem baseband physical layer design.

Modem Processing Control Flow

Brief example of input/output deep learning input options for RX control flow



Embedded Deep Learning

Memory Advancements: Up to 50x compression of inference weights in [1] through Pruning, Quantization, and Hamming coding. Dense NN [2] also giving similar performance with 2-3x fewer parameters

Processing Reduction: 3x speedup & 5x power consumption reduction in techniques discussed in [3], [4], [5] – mostly focused on exploiting zero-valued activations resulting from ReLU operation

Modem Deep Learning Exploration

Idle: (a) Choosing a technology (2G/3G/4G/5G) and frequency band to camp on, (b) Measuring and retaining a strong link over time, (c) Decoding periodic paging indicator channel from the network, (d) Classifying interfering signal

Voice: (a) SIR Target estimation, (b) Early frame decoding, (c) Voice quality (MOS) estimation and feature impact, (d) Quality of non-CRC encoded channels, (d) false pass elimination, (e) Decode metric analysis

Data: (a) Channel State Feedback (CSF) determination, (b) HARQ LLR compression, (c) Inform decoder design and 'quality of failure' analysis, (d) IC design

References

[1] S Han, H. Mao, W. Dally. (2016) "Deep Compression: Compressing Deep Neural Networks With Pruning, Trained Quantization And Huffman Coding"
[2] G. Huang, Z. Liu, L. Maaten (2018) "Densely Connected Convolutional Networks"
[3] A. Parashar, et al (2017) "SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks"
[4] S. Zhang, Z. Du, L. Zhang, H. Lan, S. Liu, L. Li, Q. Guo, T. Chen, Y. Chen. (2016) "Cambricon-X: An Accelerator for Sparse Neural Networks"
[5] J. Albericio, P. Judd, T. Hetherington, T. Aamodt, N. E. Jerger, A. Moshovos (2016) "Cnvlutin: Ineffectual-Neuron-Free Deep Neural Network Computing"

Paging Indicator (PI) Detection Example

Input: 0.67 ms of baseband IQ samples
Output: 1 if reliable PI detected
Train/Test data 1M+ generated samples from SNR sweep (-20:10 dB) + multipath fading + time/freq error distribution
Architecture: CNN with 1D input, compressed as in [1]

Layer	Layer Parameters			Output Shape		# parameters	Parameter	Compressed
	f	s	h	# chan	x		Size (kB)	Size (kB)
Input (Flat 8-bit I/Q samples)				1	10240	0	0	0
Convolution (f=1x32) + ReLU	128	1		128	10209	8192	64	4
Max Pooling (f=1x2)		2		128	5104	0	0	0
Convolution (f=1x64) + ReLU	256	1		256	5041	16384	128	8
Max Pooling (f=1x2)		2		256	2520	0	0	0
Convolution (f=1x64) + ReLU	256	1		256	2457	16384	128	8
Max Pooling (f=1x2)		2		256	1228	0	0	0
Convolution (f=1x64) + ReLU	256	1		256	1165	16384	128	8
Max Pooling (f=1x2)		2		256	582	0	0	0
Fully Connected + ReLU			512		512	76283905	595968	9312
Fully Connected + ReLU			128		128	65537	512	8
Binary Classification			1		1	129	1	0
				TOTAL (M)		76.4M	582.9MB	9.1MB

Conclusions

The modem user equipment physical layer domain appears to be an area not yet explored in the context of modern deep learning techniques. Recent advancements in neural network compression, techniques to minimize processing power, and specialized ASIC designs have made it feasible to incorporate deep learning into commercial 3G, 4G and 5G solutions.

The more promising applications appear to be within the 4G and 5G domains, for channel feedback reporting and as and tool to influence demodulation and modulation design.