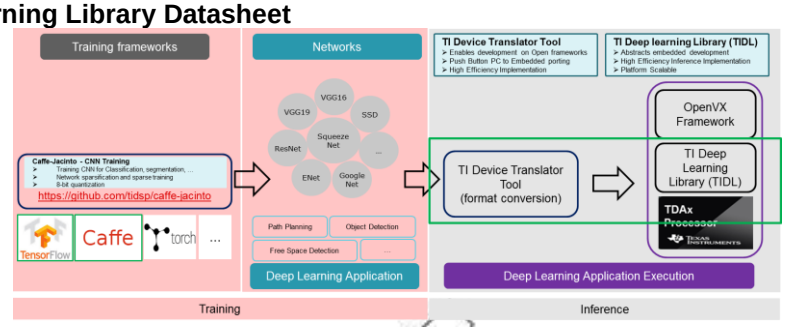
TI Deep learning Library Datasheet



- **IVISION (XDAIS) interface compliant**
- **Supports floating point Caffe Model and On the fly quantization**
- **Supports floating point Tensorflow Model (Binary protobuf) and On the fly quantization**
- **Configurable bit-depth for layer Parameters (Convolution kernel weights, Scale weights etc)**
  - 4 to 12 bits
- **Supports layers**
  - Convolution Layer
  - Pooling Layer
    - Average and Max Pooling
  - ReLU Layer
  - Element Wise Layer
    - Add, Max, Product
  - Inner Product Layer
    - Fully Connected Layer
  - Soft Max Layer
  - Bias Layer
  - Deconvolution Layer
  - Concatenate layer
  - ArgMax Layer
  - Scale Layer
  - PReLU Layer
  - Batch Normalization layer
  - ReLU6 Layer
  - Crop layer
  - Slice layer
  - Flatten layer
  - Split Layer
  - Detection Output Layer
- **Validated with a Semantic Segmentation Network and Object Classification network on TDA2x**
- **Validated with a SSD based Object detection network trained using Caffe-Jacinto on TDA2x**
- **Validated Squeezenet 1.1 caffe model**
- **Validated Inception net V1 (Googlnet) Tensorflow model**
- **Validated Mobilenet1.0 Tensorflow model.**
- **Support for dense convolution kernels (1xN and 3x3) on EVE to optimize small ROI processing.**

Note: Please refer user's guide for details on supported features in each layer.

1.1 Description

TI Deep learning Library module is TI's proprietary CNN/DNN acceleration on EVE and DSP. The performance and Memory statistic are generated using blow toolchain.

- DSP Code Generation Tools Version 7.4.2
- EVE (ARP32) Code Generation Tools Version 1.0.7

1.2 Performance Summery

This section summarizes the performance of different CNN layers on both EVE and DSP. The multidimensional data is presented in form of [Width x Height x Channels]. For example [64 x 32 x 128] represents 128 channels of 2D data of width = 64 and height = 32.

Target Platform Name: Vision28 Super ADAS Applications Processor (TDA2x)

CPU Cores: EVE and C6x DSP

OPP NOM (EVE – 535 Mhz and C6x DSP -600Mhz)

DDR3 Frequency: 532 MHz

Table 1 TIDL Convolution Layer Performance

SNo	Description	% Sparsity*	EVE		C6x DSP	
			Mega Cycles	MACs/ Cycle	Mega Cycles	MACs/ Cycle
1	[128 x 64 x 64] => ***3x3 convolution*** => [128 x 64 x 128] => ***Bias Add + RELU*** => [128 x 64 x 128] Floating point model with on the fly quantization (8 - Bit Layer Parameters)	85	8.95	12.11	39.66	2.46
2		75	12.91	12.91	57.10	2.73
3		50	23.35	13.58	101.27	3.02
4		30	31.62	13.75	135.69	3.12
5		20	35.98	13.80	153.77	3.16
6	[128 x 64 x 64] => ***3x3 convolution*** => [128 x 64 x 128] => ***Bias Add + RELU*** => [128 x 64 x 128] => ***2x2 MaxPool*** => [64 x 32 x 128] Floating point model with on the fly quantization (8 - Bit Layer Parameters)	85	8.99	12.09	39.99	2.43
7		75	12.97	12.89	57.58	2.71
8		50	23.43	13.52	101.77	3.01
9		30	31.88	13.71	136.62	3.11
10		20	36.13	13.77	154.49	3.15
11	[128 x 64 x 64] => ***5x5 convolution with 2x2 stride*** => [64 x 32 x 128] => ***Bias Add + RELU*** => [64 x 32 x 128] Floating point model with on the fly quantization (8 - Bit Layer Parameters)	85	9.50	8.00	28.74	2.43
12		75	11.98	9.70	41.63	2.64
13		50	19.25	11.48	74.79	2.87
14		30	25.41	11.95	101.09	2.94
15		20	28.44	12.14	114.32	2.97
16	[128 x 64 x 64] => ***3x3 convolution*** => [128 x 64 x 128] => ***Bias Add + RELU*** => [128 x 64 x 128] Floating point model with on the fly quantization (12 - Bit Layer Parameters)	85	9.25	11.75	39.69	2.48
17		75	13.18	12.70	57.01	2.74
18		50	23.62	13.51	101.33	3.04
19		30	31.98	13.71	135.92	3.14

20		20	36.06	13.77	153.00	3.18
21	[128 x 64 x 64] => ***3x3 convolution*** => [128 x 64 x 128] => ***Bias Add + RELU*** => [128 x 64 x 128] => ***2x2 MaxPool*** => [64 x 32 x 128] Floating point model with on the fly quantization (12 - Bit Layer Parameters)	85	9.38	11.66	40.11	2.45
22		75	13.18	12.59	57.13	2.72
23		50	23.47	13.45	100.72	3.02
24		30	31.95	13.69	136.26	3.13
25		20	36.23	13.75	153.89	3.17
26	[128 x 64 x 64] => ***5x5 convolution with 2x2 stride*** => [64 x 32 x 128] => ***Bias Add + RELU*** => [64 x 32 x 128] Floating point model with on the fly quantization (12 - Bit Layer Parameters)	85	11.89	6.43	28.57	2.44
27		75	15.56	7.50	41.36	2.66
28		50	25.01	8.82	74.14	2.89
29		30	32.63	9.32	100.28	2.97
30		20	36.37	9.49	113.29	2.99

** Sparsity indicates the % of kernel coefficients having zero value*

Table 2 TIDL Dense Convolution Layer Performance

SNo	Description	EVE	
		Mega Cycles	MACs/Cycle
1	[20 x 20 x 32] => ***3x3 convolution*** => [20 x 20 x 64] ***+ RELU*** Floating point model with on the fly quantization (8 - Bit Parameters)	0.67	11.06
2	[16 x 16 x 64] => ***3x3 convolution*** => [16 x 16 x 128] ***+ RELU*** Floating point model with on the fly quantization (8 - Bit Parameters)	1.50	12.62
3	[8 x 8 x 128] => ***3x3 convolution*** => [8 x 8 x 256] ***+ RELU*** Floating point model with on the fly quantization (8 - Bit Parameters)	1.64	11.50
4	[20 x 20 x 32] => ***3x3 convolution*** => [20 x 20 x 64] ***+ RELU*** Floating point model with on the fly quantization (8 - Bit Parameters)	0.66	11.10
5	[16 x 16 x 64] => ***3x3 convolution*** => [16 x 16 x 128] ***+ RELU*** Floating point model with on the fly quantization (8 - Bit Parameters)	1.50	11.07
6	[8 x 8 x 128] => ***3x3 convolution*** => [8 x 8 x 256] ***+ RELU*** Floating point model with on the fly quantization (8 - Bit Parameters)	1.64	11.50

Table 3 TIDL Layers Performance

SNo	Description	EVE		C6x DSP	
		Mega Cycles	Cycles/ Output	Mega Cycles	Cycles/ Output
1	[1024 x 512 x 3] => ***Bias Layer*** => [1024 x 512 x 3] with 8-bit Layer Parameters	2.13	1.34	7.32	4.65
2	[1024 x 512 x 8] => ***ArgMax Layer*** => [1024 x 512 x 8] with 8-bit Layer Parameters	2.01	0.48	4.78	1.12
3	[128 x 64 x 64] => ***Eltwise Layer*** => [128 x 64 x 64] with 8-bit Layer Parameters	0.92	1.70	1.68	3.20
4	[1024 x 512 x 8] => **4x4 Deconvolution Layer** => [1024 x 512 x 8] with 8-bit Layer Parameters	6.30	1.44	28.39	6.73
5	[128 x 64 x 128] => **2x2 Max Pooling with stride 2x2** => [64 x 32 x 128] with 8-bit Layer Parameters	1.41	5.38	1.44	5.49

6	[128 x 64 x 128] => **3x3 Max Pooling with stride 2x2** => [64 x 32 x 128] with 8-bit Layer Parameters	2.86	10.91	3.15	12.02
7	[128 x 64 x 128] => **3x3 Avg Pooling with stride 2x2** => [64 x 32 x 128] with 8-bit Layer Parameters	2.75	10.49	8.80	33.57
8	[1024 x 512 x 3] => ***Bias Layer*** => [1024 x 512 x 3] with 12-bit Layer Parameters	2.16	1.37	6.92	4.40
9	[1024 x 512 x 8] => ***ArgMax Layer*** => [1024 x 512 x 8] with 12-bit Layer Parameters	2.03	0.48	2.20	0.52
10	[128 x 64 x 64] => ***Eltwise Layer***=> [128 x 64 x 64] with 12-bit Layer Parameters	0.89	1.70	1.68	3.20
11	[1024 x 512 x 8] => **4x4 Deconvolution Layer**=> [1024 x 512 x 8] with 12-bit Layer Parameters	6.03	1.44	23.41	5.58
12	[128 x 64 x 128] => **2x2 Max Pooling with stride 2x2** => [64 x 32 x 128] with 12-bit Layer Parameters	1.41	5.38	1.44	5.49
13	[128 x 64 x 128] => **3x3 Max Pooling with stride 2x2** => [64 x 32 x 128] with 12-bit Layer Parameters	2.87	10.95	3.16	12.05
14	[128 x 64 x 128] => **3x3 Avg Pooling with stride 2x2** => [64 x 32 x 128] with 12-bit Layer Parameters	2.75	10.49	8.80	33.57

Table 4A TIDL Segmentation Network Performance[1]

SNo	Description	EVE	C6x DSP
		Mega Cycles	Mega Cycles
1	Semantic Segmentation Network - 1024x512 with 8-bit Layer Parameters	162.2	596.12
2	Semantic Segmentation Network - 1024x512 with 12-bit Layer Parameters	192.88	667.85

[1] Complete Network execution for one frame inference is measure either on EVE or DSP

Table 4B TIDL Object Detection Network Performance[2]

SNo	Description	EVE	C6x DSP
		Mega Cycles	Mega Cycles
1	SSD Based Object Detection Network - 768x320 with 8-bit Layer Parameters	96.24	5.87

[2] All layers of the network are executed / profiled on EVE except the Detection Output Layer. This Detection Output Layer is executed on C6x DSP (it is best optimal on C6x DSP). So the EVE and DSP time needs to be added together to get total runtime.

Table 4C TIDL Image classification Network Performance[3]

SNo	Description	EVE	C6x DSP
		Mega Cycles	Mega Cycles
1	MobileNet Based image classification Network - 224x224 with 10-bit Layer Parameters and dense convolutions	69.21	539.1

[3] In EVE Mega Cycles, SoftMax Layer cycles (25 Mega cycles on EVE) are not accounted. In DSP Mega Cycles, SoftMax Layer cycles (1.72 Mega cycles on DSP) are included.

Table 5 TIDL On-the-fly Quantization Accuracy

Class Id	Caffe Floating point model	TIDL dynamic Quantization to 8-bit
0	97.04%	97.06%
1	93.04%	93.08%
2	68.55%	68.64%
3	65.07%	64.35%
4	88.13%	88.15%
Mean	82.36%	82.25%

1.3 Memory Summery

Table 6 Memory Statistics

Configuration ID	Memory Statistics[1]				
	Program Memory	Data Memory			
		Stack	SCRATCH[2]		
			IALG_DARAM0	IALG_DARAM1	IALG_SARAM0
EVE	54.41	11	20 (DMEM)	8 (DMEM)	320 (OCM)
C6x DSP	91.68	12	8 (L1D)	148 (L2 SRAM)	320 (OCM)

[1] All memory requirements are expressed in kilobytes (1 K-byte = 1024 bytes) and there could be a variation of around 1-2% in the numbers.

[2] The algorithm will work fine with these memories being in DDR also. These are minimum required internal memories for optimal TIDL performance.

1.4 References

- TIDeepLearningLibrary_UserGuide.pdf

1.5 Glossary

Constants	Elements that go into .const memory section
Scratch	Memory space that can be reused across different instances of the algorithm
Shared	Sum of Constants and Scratch
Instance	Persistent-memory that contains persistent information - allocated for each instance of the algorithm

1.6 Acronyms

DMA	Direct Memory Access
EVM	Evaluation Module

IMPORTANT NOTICE

Texas Instruments Incorporated and its subsidiaries (TI) reserve the right to make corrections, enhancements, improvements and other changes to its semiconductor products and services per JESD46, latest issue, and to discontinue any product or service per JESD48, latest issue. Buyers should obtain the latest relevant information before placing orders and should verify that such information is current and complete. All semiconductor products (also referred to herein as "components") are sold subject to TI's terms and conditions of sale supplied at the time of order acknowledgment.

TI warrants performance of its components to the specifications applicable at the time of sale, in accordance with the warranty in TI's terms and conditions of sale of semiconductor products. Testing and other quality control techniques are used to the extent TI deems necessary to support this warranty. Except where mandated by applicable law, testing of all parameters of each component is not necessarily performed.

TI assumes no liability for applications assistance or the design of Buyers' products. Buyers are responsible for their products and applications using TI components. To minimize the risks associated with Buyers' products and applications, Buyers should provide adequate design and operating safeguards.

TI does not warrant or represent that any license, either express or implied, is granted under any patent right, copyright, mask work right, or other intellectual property right relating to any combination, machine, or process in which TI components or services are used. Information published by TI regarding third-party products or services does not constitute a license to use such products or services or a warranty or endorsement thereof. Use of such information may require a license from a third party under the patents or other intellectual property of the third party, or a license from TI under the patents or other intellectual property of TI.

Reproduction of significant portions of TI information in TI data books or data sheets is permissible only if reproduction is without alteration and is accompanied by all associated warranties, conditions, limitations, and notices. TI is not responsible or liable for such altered documentation. Information of third parties may be subject to additional restrictions.

Resale of TI components or services with statements different from or beyond the parameters stated by TI for that component or service voids all express and any implied warranties for the associated TI component or service and is an unfair and deceptive business practice. TI is not responsible or liable for any such statements.

Buyer acknowledges and agrees that it is solely responsible for compliance with all legal, regulatory and safety-related requirements concerning its products, and any use of TI components in its applications, notwithstanding any applications-related information or support that may be provided by TI. Buyer represents and agrees that it has all the necessary expertise to create and implement safeguards which anticipate dangerous consequences of failures, monitor failures and their consequences, lessen the likelihood of failures that might cause harm and take appropriate remedial actions. Buyer will fully indemnify TI and its representatives against any damages arising out of the use of any TI components in safety-critical applications.

In some cases, TI components may be promoted specifically to facilitate safety-related applications. With such components, TI's goal is to help enable customers to design and create their own end-product solutions that meet applicable functional safety standards and requirements. Nonetheless, such components are subject to these terms.

No TI components are authorized for use in FDA Class III (or similar life-critical medical equipment) unless authorized officers of the parties have executed a special agreement specifically governing such use.

Only those TI components which TI has specifically designated as military grade or "enhanced plastic" are designed and intended for use in military/aerospace applications or environments. Buyer acknowledges and agrees that any military or aerospace use of TI components which have ***not*** been so designated is solely at the Buyer's risk, and that Buyer is solely responsible for compliance with all legal and regulatory requirements in connection with such use.

TI has specifically designated certain components as meeting ISO/TS16949 requirements, mainly for automotive use. In any case of use of non-designated products, TI will not be responsible for any failure to meet ISO/TS16949.

Products		**Applications**	
Audio	www.ti.com/audio	Automotive & Transportation	www.ti.com/automotive
Amplifiers	amplifier.ti.com	Communications & Telecom	www.ti.com/communications
Data Converters	dataconverter.ti.com	Computers & Peripherals	www.ti.com/computers
DLP® Products	www.dlp.com	Consumer Electronics	www.ti.com/consumer-apps
DSP	dsp.ti.com	Energy and Lighting	www.ti.com/energyapps
Clocks and Timers	www.ti.com/clocks	Industrial	www.ti.com/industrial
Interface	interface.ti.com	Medical	www.ti.com/medical
Logic	logic.ti.com	Security	www.ti.com/security
Power Mgmt	power.ti.com	Space, Avionics & Defense	www.ti.com/space-avionics-defense
Microcontrollers	microcontroller.ti.com	Video & Imaging	www.ti.com/video
RFID	www.ti-rfid.com		
OMAP Applications Processors	www.ti.com/omap	**TI E2E Community**	e2e.ti.com
Wireless Connectivity	www.ti.com/wirelessconnectivity		