embedded VISION SUMMIT 2018

Neural Network Compiler: Enabling Rapid Deployment of DNNs on Low-Cost, Low-Power Processors

cadence®

Megha Daga

05/23/2018

Agenda

- Industry Trend
- Software Challenges
- Software Solution:
  - Static Tools
  - Dynamic Tools
- Tensilica® Xtensa® Neural Network Compiler (XNNC): Static Tool Support
- End-to-End deployment example
- Android Neural Network API (ANN): Dynamic Tool Support
- Conclusion

Neural Networks Everywhere

“Alexa, Add a 2pm meeting to my calendar”

Where is Processing Happening?

- AI in cloud
  - Adds latency
  - Undependable connectivity
  - Breaches privacy
  - High cost
- Trend: AI On-Device
  - Face detection
  - Seeing AI
  - Speech to Text

AI On-Device

- Requirements
    - Low Power
    - Low Area
    - Programmable
    - High Performance
    - Efficient Memory Management
    - Low Latency
    - Scalable

AI Specialized Embedded Device
Example:
Tensilica® Vision P6,
Tensilica® Vision Q6,
Tensilica® Vision C5

Embedded Platform: NN Implementation Challenges

Specialized Engines

- Optimum performance needs utilization of intrinsics
- Optimum performance needs most appropriate selection of ISA

Market Needs

- Accelerate algorithm implementation time to keep up with the speed of product announcements
- Need for dynamic platform to adapt to rapidly changing neural network algorithms

Optimization

- Need for optimization due to huge compute and bandwidth needs from neural networks
- Need for optimum memory management for bandwidth and power reduction

Solution: Automated Code Generation Tools for NN

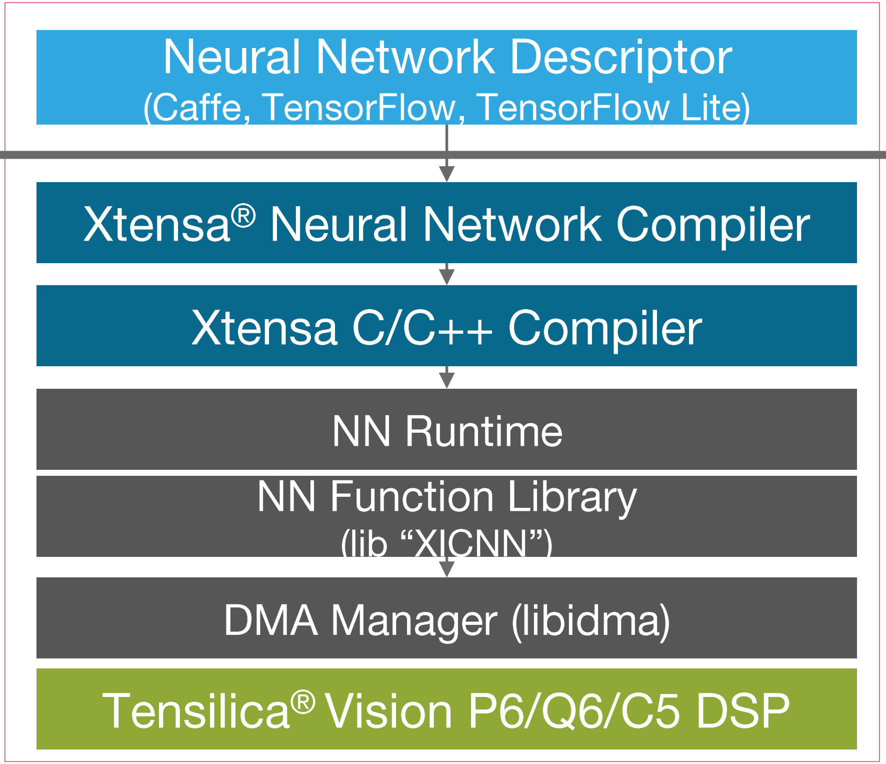
Offline

- Varied framework support
- Pre-defined NN models
- Need most optimum solution
- Example: Always-on face recognition
- Need for custom offline NN compilers
- Example: Tensilica® Xtensa Neural Network Compiler

Online

- Varied framework support
- Dynamic application development
- Need most convenient porting solution
- Example: Custom NN app on mobile phone
- Need for dynamic NN compilers
- Example: Android Neural Network API

Offline Software Solution: Tensilica® Xtensa Neural Network Compiler (XNNC)

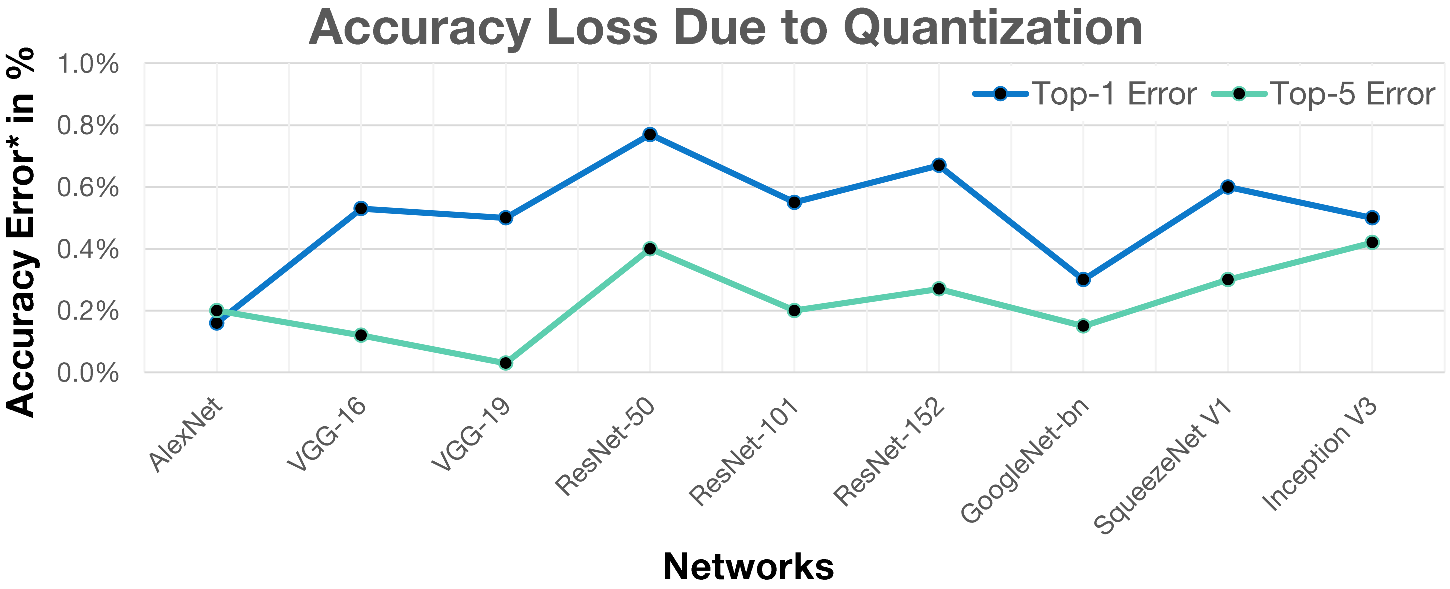

Neural Network Descriptor
(Caffe, TensorFlow, TensorFlow Lite)
Xtensa® Neural Network Compiler
Xtensa C/C++ Compiler
NN Runtime
NN Function Library
(lib “XICNN”)
DMA Manager (libidma)
Tensilica® Vision P6/Q6/C5 DSP

Compute & Bandwidth Optimized Code Generator

- Custom Quantization to 8b data & weights
- Use of target specific optimized NN Library with convolution and non-convolution layers
- Inclusion of performance enhancement features like selection of most optimum library function, kernel fusion, kernel rejection, DMA & tile management

🟦 User Code
■ Tensilica® Compiler / Tool
⬛ Tensilica® SW Library / Runtime
🟩 Tensilica® Vision DSPs

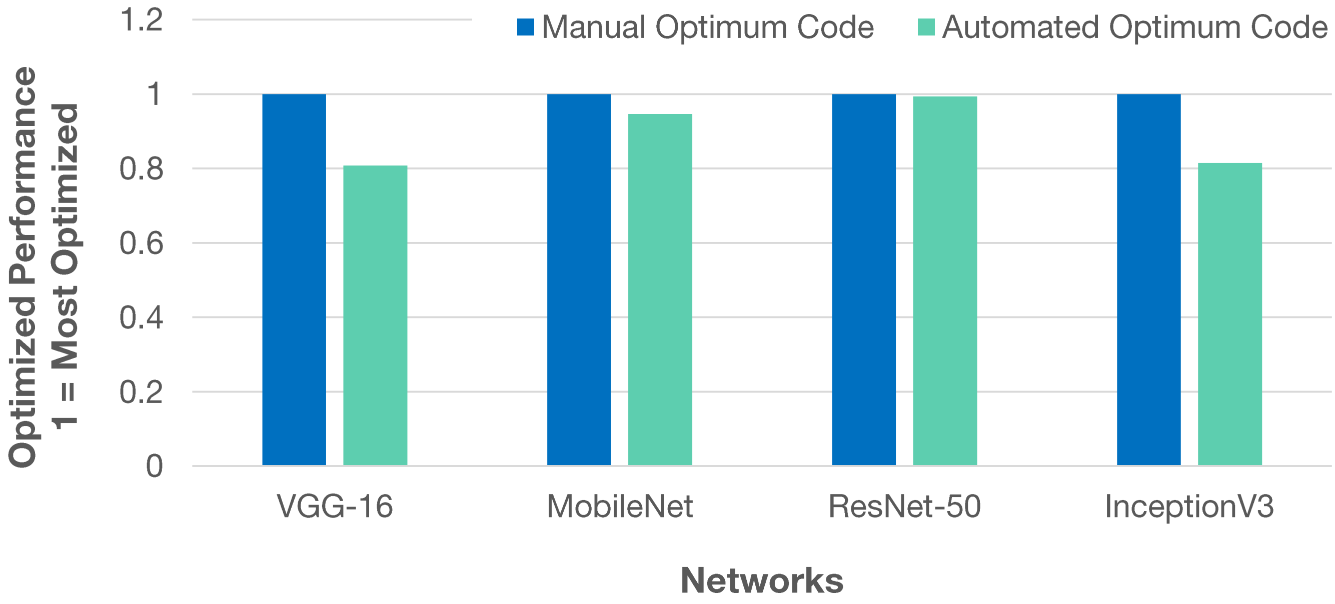
XNNC Highlights: Quantization



Vision DSP’s use of 8b fixed point for AI processing has negligible accuracy impact

*Code Generated through XNNC and tested over 50K Images

XNNC Highlight: Optimum Code Generation

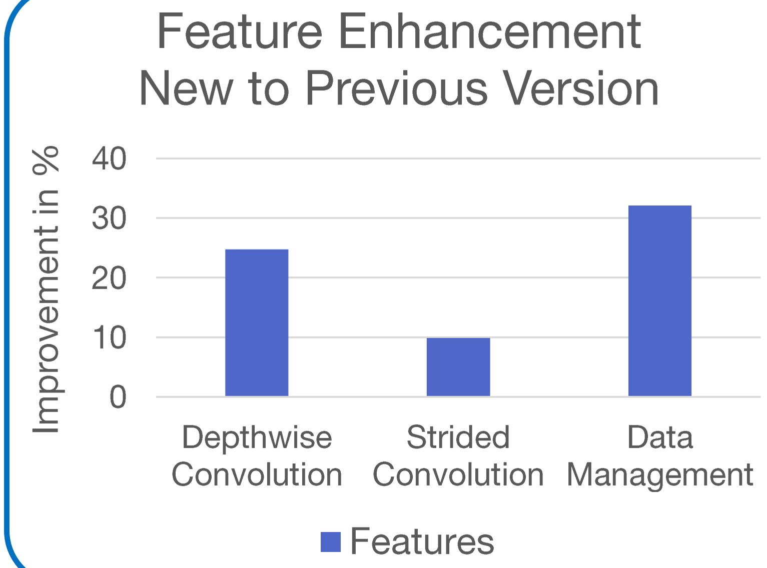


Automated generated code is close to optimum implementation resulting in speedy time to market

*Code Generated through XNNC for Vision P6 DSP

XNNC Highlights: Continuous Feature Update

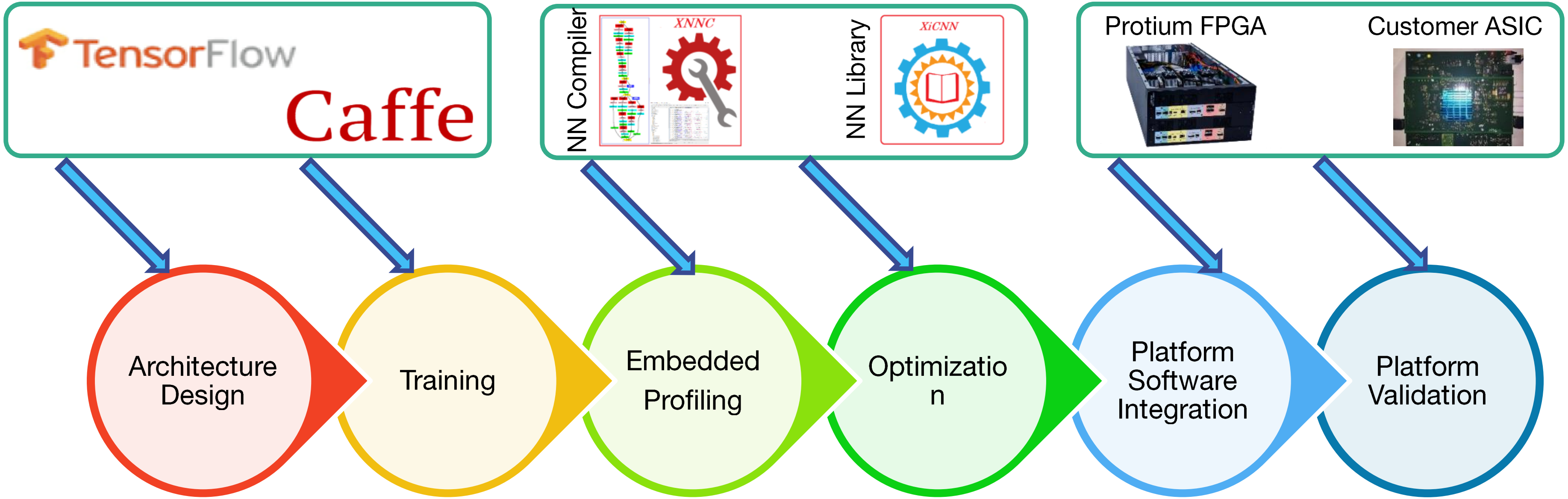
Continuous Investment in Optimization



Continuous Addition of New Layers, Framework and Features

- Wide support for framework descriptors: Caffe, TensorFlow, TensorFlow Lite
- Wide support of network types: classification, object detection, segmentation, recurrent, regression
- Custom layer
- Sparsity utilization for bandwidth and compute optimizations

XNNC with Dynamic Deployment: Time to Market Process



TensorFlow

Caffe

NN Compiler

XNNC

NN Library

XiCNN

Protium FPGA

Customer ASIC

Architecture Design

Training

Embedded Profiling

Optimizatio n

Platform Software Integration

Platform Validation

- Network Type
- Network Topology
- Network Depth

- Performance Enhancement
- Layer Pruning

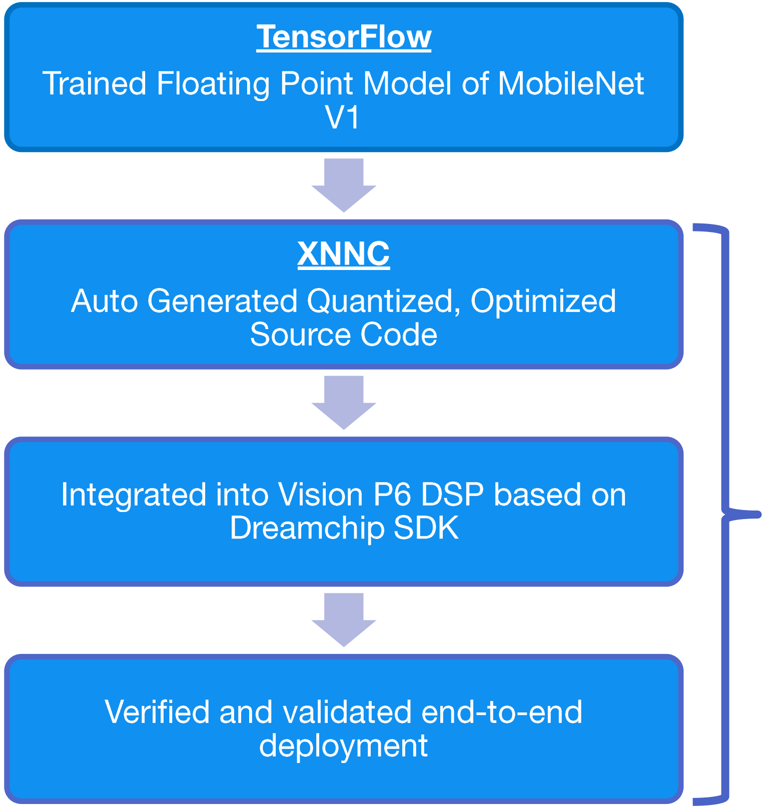
- Quantization
- Layer Merging
- Layer Organization

- Tiling
- Sequencing
- DMA
- Vectorization

- OS/Driver
- Video framework
- Vision/NN framework
- NN Application

- E2E Application
- Performance Validation
- In-field Validation

XNNC with Dynamic Deployment: Time to Market Example



TensorFlow

Trained Floating Point Model of MobileNet V1

XNNC

Auto Generated Quantized, Optimized Source Code

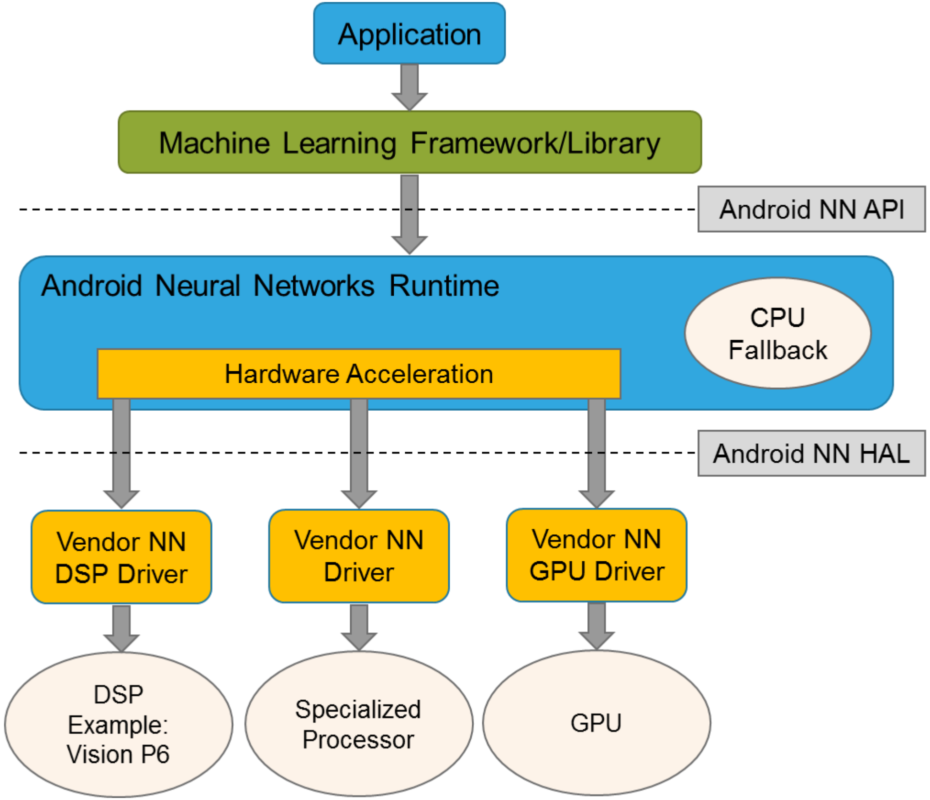
Integrated into Vision P6 DSP based on Dreamchip SDK

Verified and validated end-to-end deployment

Several man-months reduced to less than a couple of days

ON-DEVICE AI FOR
IMAGE CLASSIFICATION
TENSILICA VISION DSP
cādence
Cadence Xtensa® Neural Network Compiler (XNNC)
Caffe and TensorFlow NN Framework to Tensilica® AI DSP
Framework Definition
Caffe based Trained Floating point model of Inception V3 and MobileNet Neural Networks
Cadence XNNC
NN Function Library (lib "XI-CNN")
Cadence Automated Code Generator
Fixed point 8-bit NN code for both Inception V3 and MobileNet neural networks from XNNC to Tensilica AI DSP
Tensilica® AI DSP
Cadence Or Customer HW
Generated Optimized neural network code deployed on an SoC with Vision P6* and Vision C5 DSPs
shutterstock · 96990671

Online Software Solution: Android Neural Network API



Courtesy: https://developer.android.com/ndk/guides/neuralnetworks/

Easy Deployment of NN on Android Devices

- Called by ML Libraries or frameworks
- Efficiently distributes workload across available on-device processors

Android NN Highlights: Dynamic APP Development

Enabling App Development on Android Devices

- Apps directly use higher level frameworks to deploy trained model on device
- Specialized inferencing engines enables highest performance at lowest power penalty

Real-Time Optimized Execution on Tensilica® Vision P6 DSP

- Executes graph/sub-graph/layer
- Gets best runtime optimization using tile, DMA management, data rearrangement
- Use of hand-optimized ML library

Conclusion

- **Trend**: See, hear and speak more clearly with On-Device AI
- **Software**: Need for static and dynamic ecosystem for development and deployment of trained networks
- **Example**:
  - Tensilica® Xtensa Neural Network Compiler
  - Android Neural Network API

Demo

ON-DEVICE AI FOR IMAGE CLASSIFICATION
TENSILICA VISION DSP
cadence
Cadence Xtensa® Neural Network Compiler (XNNC)
Caffe and TensorFlow NN Framework to Tensilica® AI DSP
Framework Definition
Caffe based Trained Floating point model of Inception V3 and MobileNet Neural Networks
Cadence XNNC
NN Function Library
Cadence Automated Code Generator
Tensilica® AI DSP
Cadence Or Customer HW
shutterstock · 96990671

ON-DEVICE AI FOR PEOPLE DETECTION
TENSILICA VISION DSP
cadence
On-Device AI for People Detection
Tensilica® Vision DSP
You Only Look Once (YOLO) based neural network trained for people detection
Extremely lightweight – ideal for applications requiring fast, power-efficient object detection (with localization) for multiple object categories
Can be re-trained for up to 20 object classes
Custom network optimizations, including pruning and quantization
Reduced memory footprint and bandwidth requirements
Developed in partnership with BDTi
Contact info@BDTI.com for info about licensing, customization, and integration services

See our AI demos at booth 200. Visit our webpage ip.cadence.com/vision.

Resources

- Cadence Web Page
  - https://ip.cadence.com
- Android NN
  - https://developer.android.com/ndk/guides/neuralnetworks/index.html