Toolflows for Mapping Convolutional Neural Networks on FPGAs: A Survey and Future Directions

Aayush Suresh Bansal
Anshul Suresh Bansal

IS THIS A
CAT or DOG?

~~CAT~~ DOG

OUTPUT LAYER

ACTIVATED NEURONS

INPUT LAYER

DEEP *neural* NETWORK

What is a toolflow?

Mapping of high level input to a hardware architecture.

Why do we need a toolflow?

Increases the speed of development cycle

FPGAs are a promising alternative

CNN-FPGA toolflows

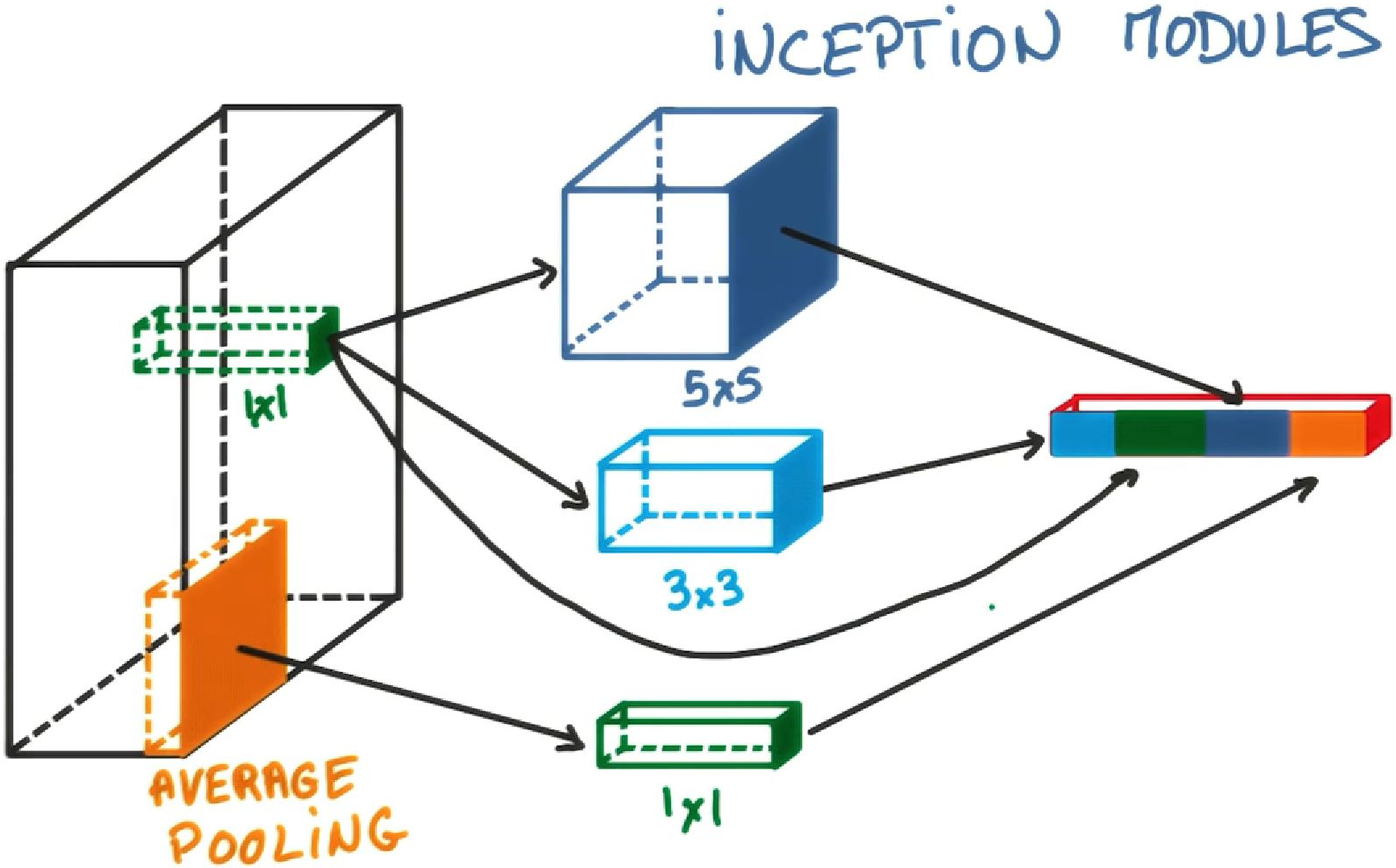
Toolflow Name	Interface	Year
fpgaConvNet [86][87][88][85]	Caffe & Torch	May 2016
DeepBurning [90]	Caffe	June 2016
Angel-Eye [68][23][24]	Caffe	July 2016
ALAMO [58][56][57][55][59]	Caffe	August 2016
Haddoc2 [1][2]	Caffe	September 2016
DnnWeaver [75][76]	Caffe	October 2016
Caffeine [98]	Caffe	November 2016
AutoCodeGen [54]	Proprietary Input Format	December 2016
Finn [84][19]	Theano	February 2017
FP-DNN [22]	TensorFlow	May 2017
Snowflake [21][10]	Torch	May 2017
SysArrayAccel [91]	C Program	June 2017
FFTCodeGen [100][97][96][95]	Proprietary Input Format	December 2017

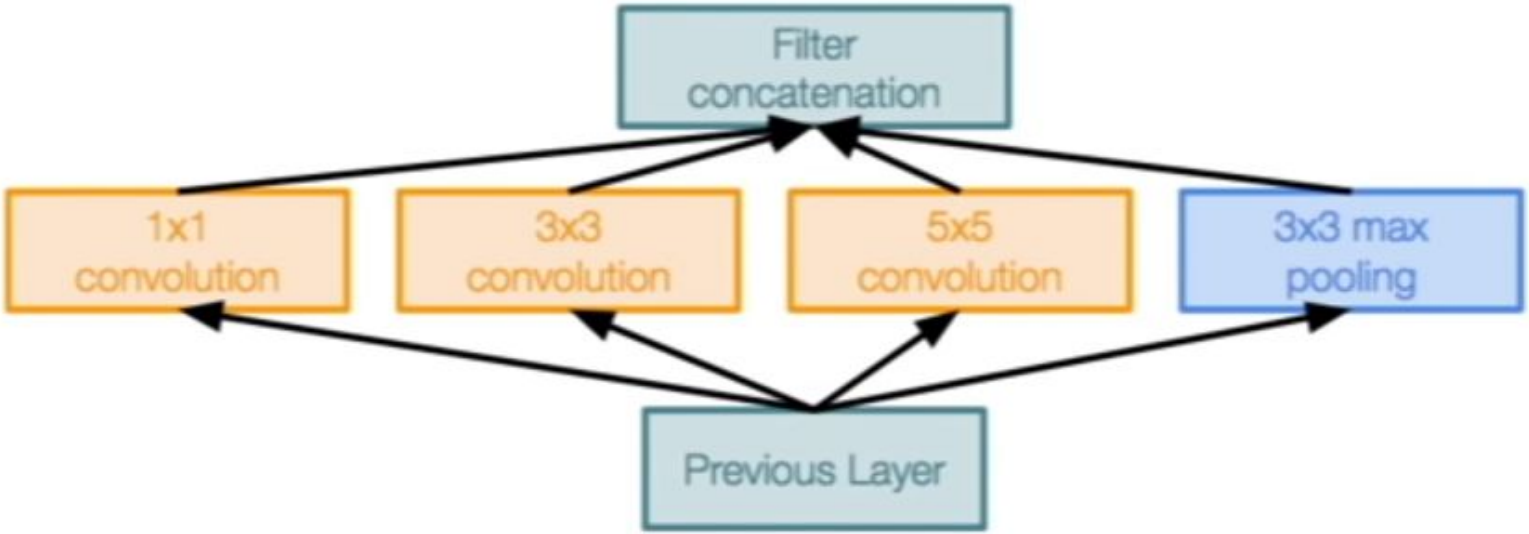
Comparison Parameters

- Supported neural networks
- Interface
- Portability
- Exploration of design space
- Arithmetic Precision
- Performance

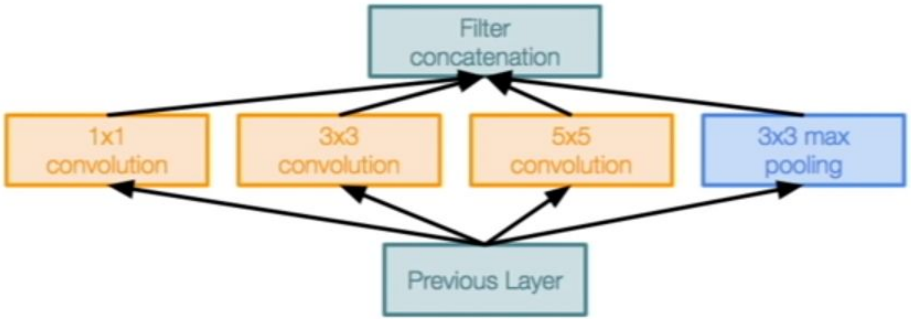
Supported Neural Networks

- CONV, POOL and FC Layer supported by all existing frameworks
- Additional layers and blocks:
  - Local Response Normalization Layer - Implements lateral Inhibition
  - Residual blocks - Used in ResNet
  - Inception Modules

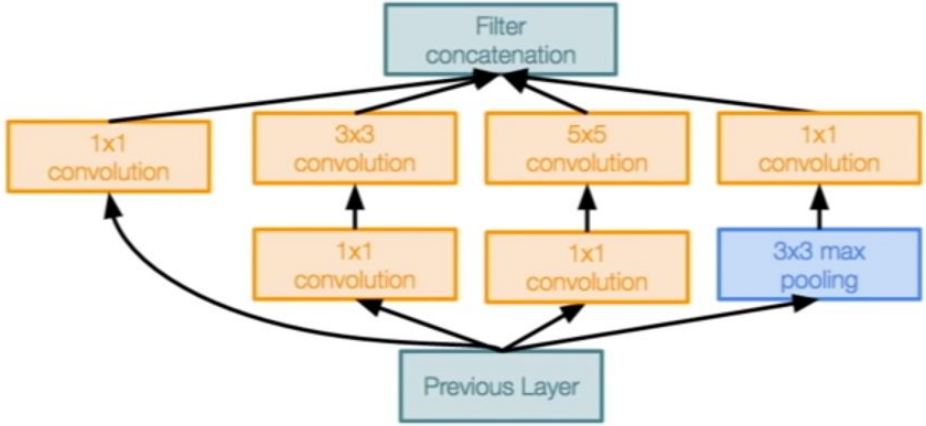

INCEPTION MODULES
1x1
5x5
3x3
1x1
AVERAGE POOLING



Naive Inception module



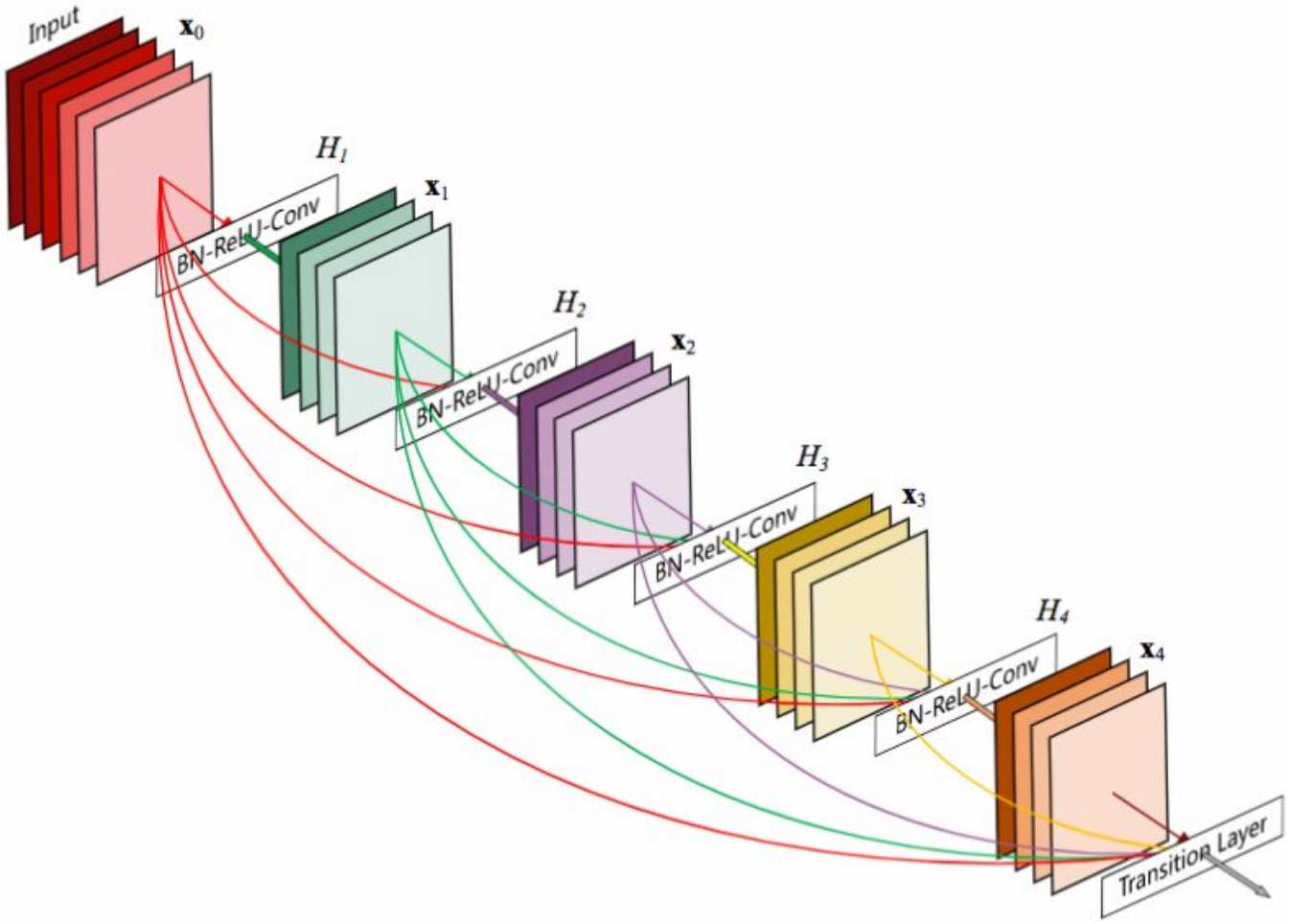
Naive Inception module



Inception module with dimension reduction

Supported Neural Networks

- CONV, POOL and FC Layer supported by all existing frameworks
- Additional layers and blocks:
  - Local Response Normalization Layer - Implements lateral Inhibition
  - Residual blocks - Used in ResNet
  - Inception Modules
  - Dense Blocks


Input
$\mathbf{x}_0$
H_1
BN-ReLU-Conv
$\mathbf{x}_1$
H_2
BN-ReLU-Conv
$\mathbf{x}_2$
H_3
BN-ReLU-Conv
$\mathbf{x}_3$
H_4
BN-ReLU-Conv
$\mathbf{x}_4$
Transition Layer

Supported Neural Networks

- CONV, POOL and FC Layer supported by all existing frameworks
- Additional layers and blocks:
  - Local Response Normalization Layer - Implements lateral Inhibition
  - Residual blocks - Used in ResNet
  - Inception Modules
  - Dense Blocks

DeepBurning and FP-DNN demonstrate the widest range of supported applications.

Interface

- Ease of access to developers
- Caffe is the most widely used front-end
- It supports 7 FPGA frameworks