Neural Network Exchange Format
May 2017

Peter McGuinness
Chairman, NNEF Working Group
Visualisetheworld.com
peter.mcguinness@gobrach.com

What Is NNEF?

- **It's a kind of pdf for neural networks**
  - There are many frameworks, many inference engines, all with different input and output formats.
  - That's a recipe for fragmentation, which spells disaster for embedded applications.

- **Training in general is OK**
  - Training is usually offline and handled by highly scalable machines with full programmability and universal operand availability

- **Inference is a different story**
  - While there will be a substantial mount of inference done in data centers, edge devices with dominate this market
  - To succeed, embedded devices need a universal deployment platform. NNEF, along with OpenVX and Khronos' low level and more abstracted programming languages, provide this.

Caffe

torch

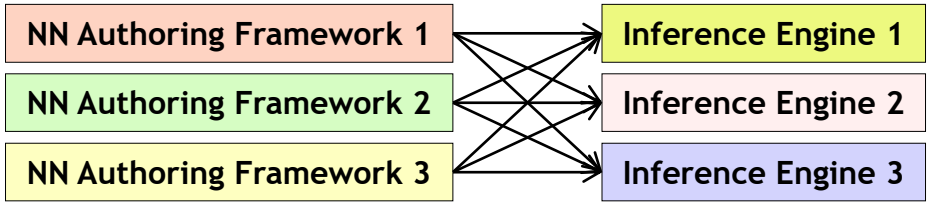
TensorFlow

Microsoft CNTK

theano

KHRONOS GROUP

Objective and use model

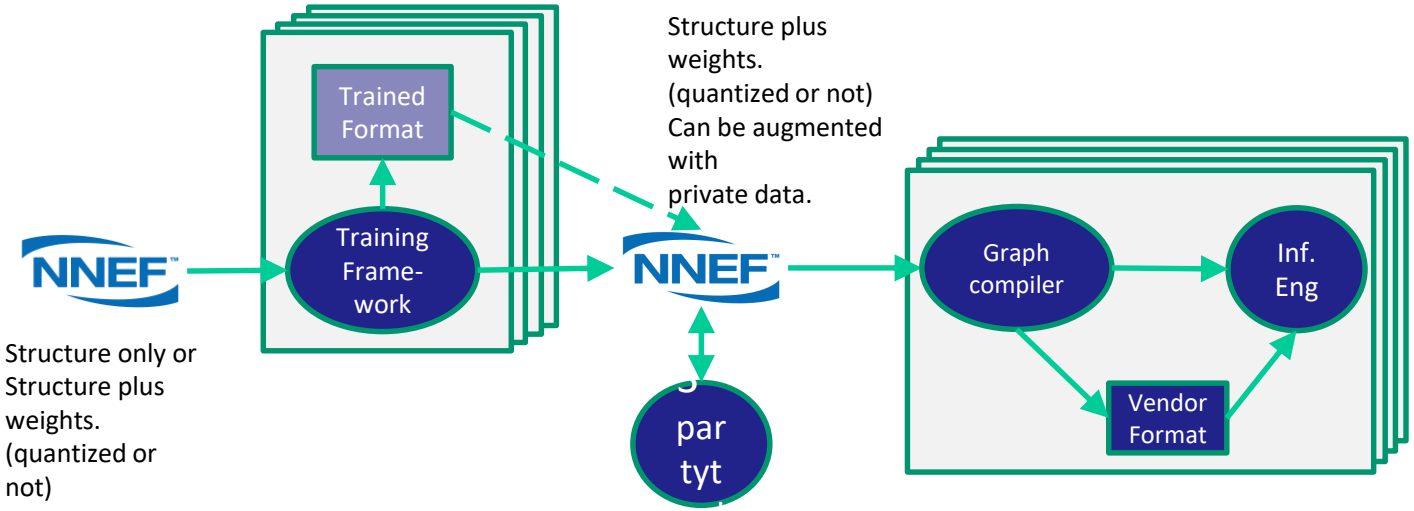


Replaces many : many mapping
with many-one : one-many.

Enables third party tools industry.

Eases deployment onto embedded systems.

A sort of pdf for neural nets.



NNEF encapsulates neural network structure, data formats, commonly used operations (such as convolution, pooling, normalization, etc.) and formal network semantics

Status and roadmap

- **V1.0 is under development, will soon start to solicit industry comments**
  - NNEF will form an advisory panel, you are invited today to participate.
- **First version will focus on embedded inference (but will allow training)**
- **'First cut' range of network types.**
  - Field is moving very fast but we aim to keep up with developments,

- **Future versions:**
  - will track development of new network types.
  - Will address a wider range of applications (outside vision apps).
  - Will aim to increase the expressive power of the format.