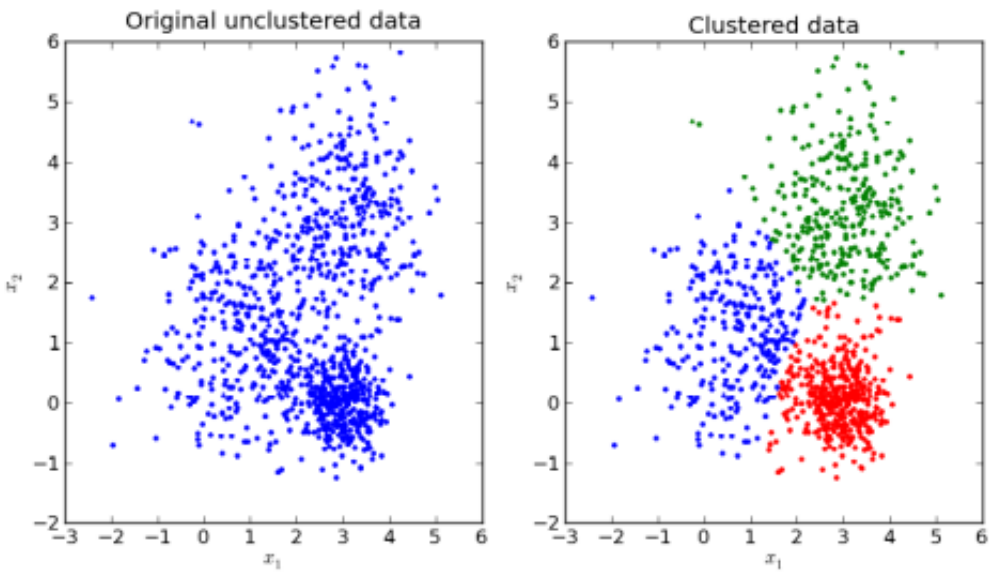
ASYNC TASKING FOR DEEP LEARNING

Stephen Jones, NVIDIA, 18th April 2017

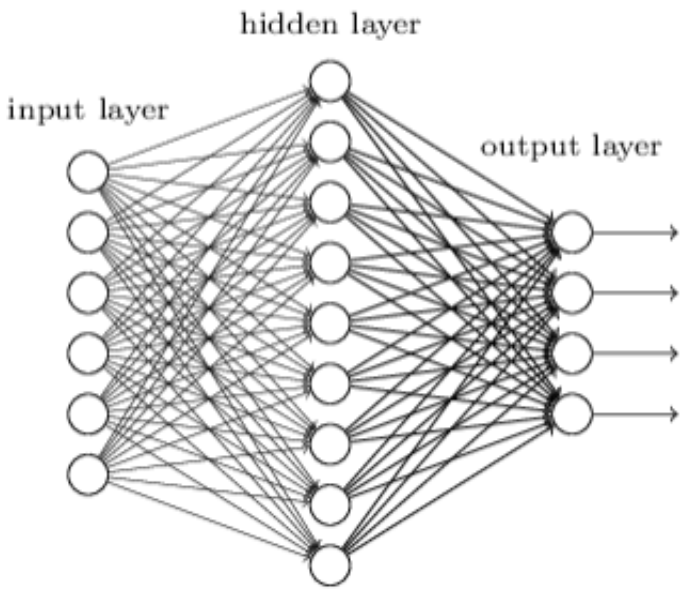
WHAT IS DEEP LEARNING?

Model-based machine learning



Fit model to training data
Test real data against model

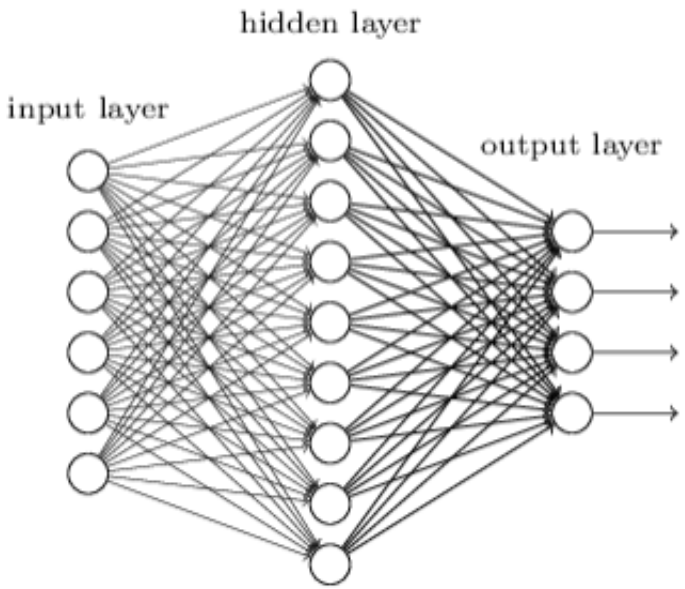
"Non-deep" feedforward neural network



Output is combination of linear operations & filters on input

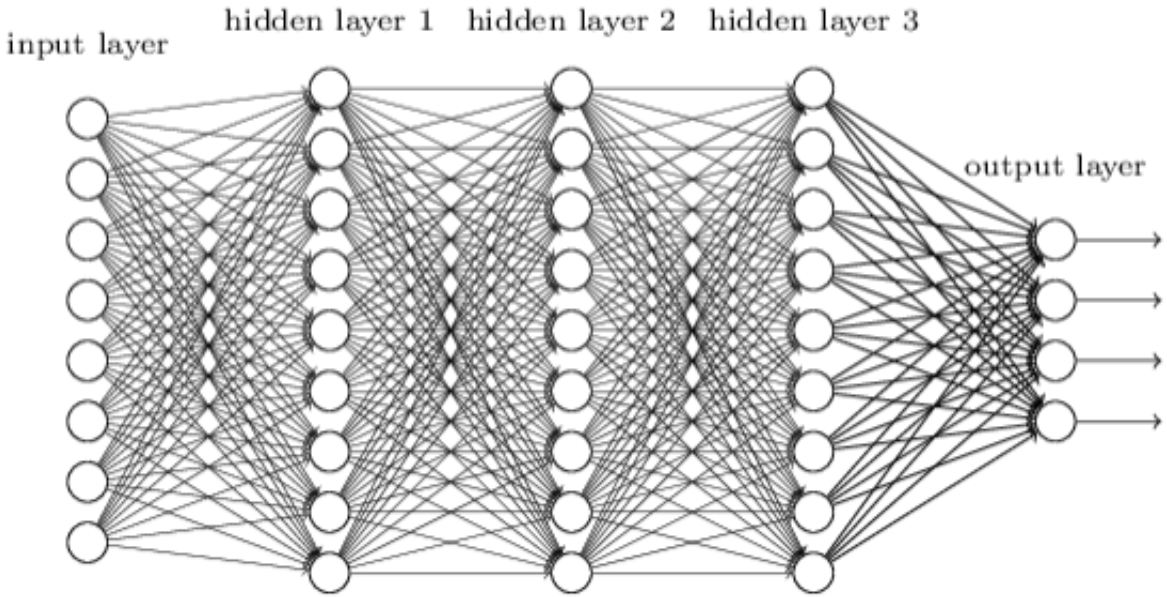
DEEP NEURAL NETWORKS

"Non-deep" feedforward neural network



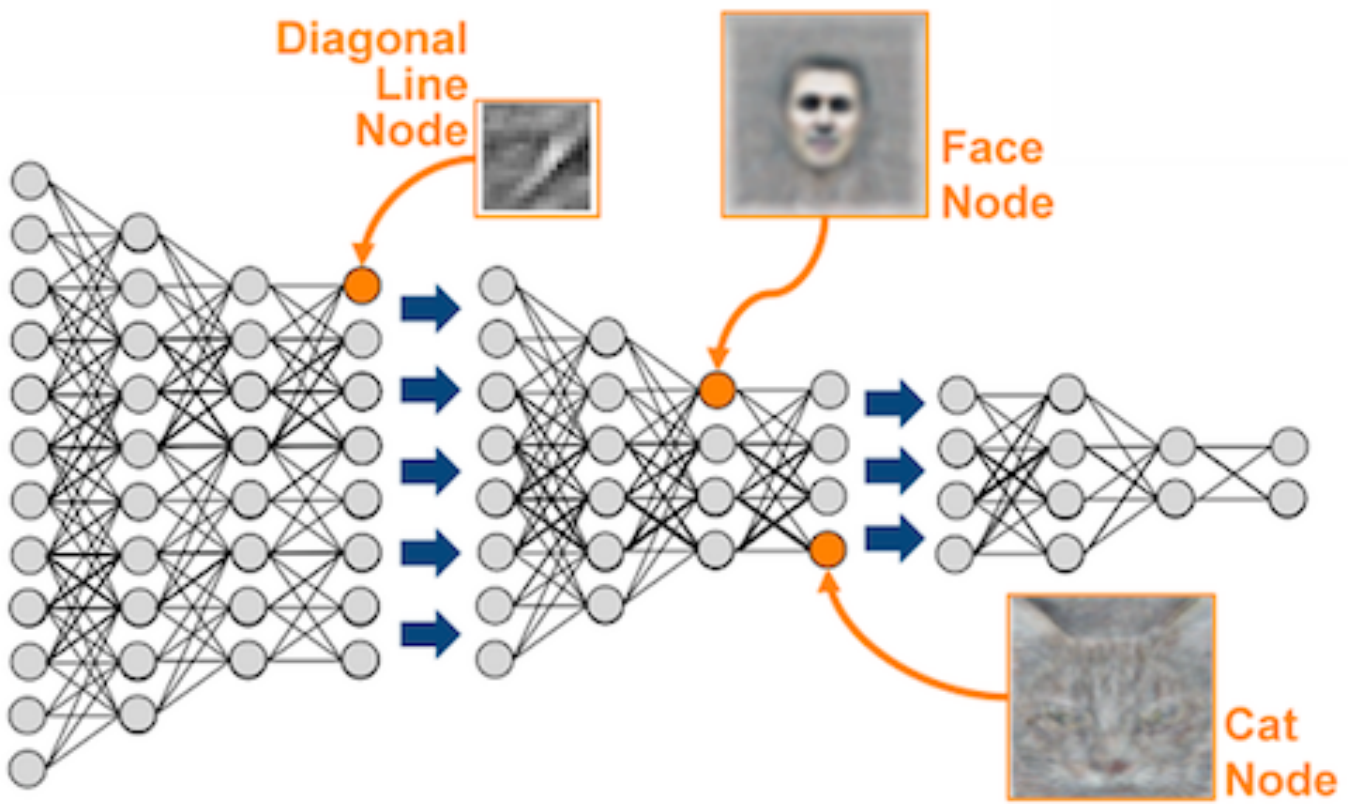
Output is combination of linear operations & filters on input

Deep neural network



Multiple layers each extract different characteristics from input

FEED-FORWARD NEURAL NETWORKS



Data flow is uni-directional: work graph is acyclic

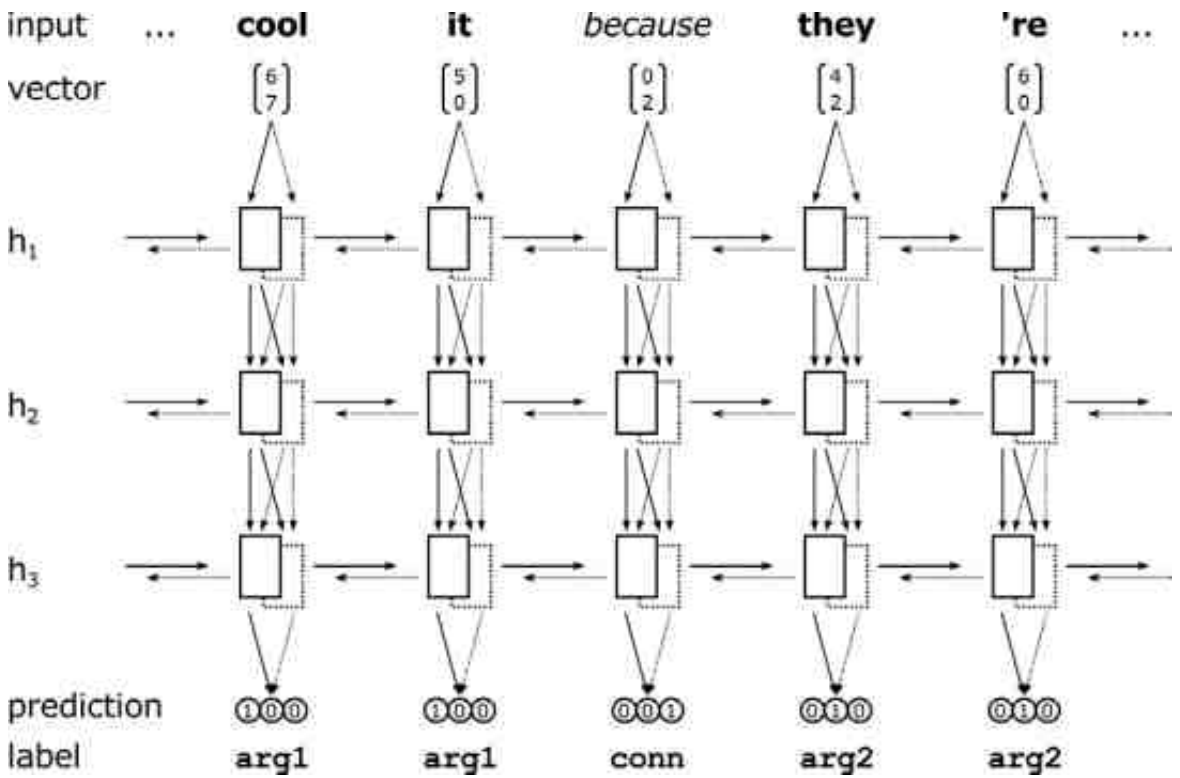
HOW A DEEP NEURAL NETWORK SEES

Image

"Audi A7"

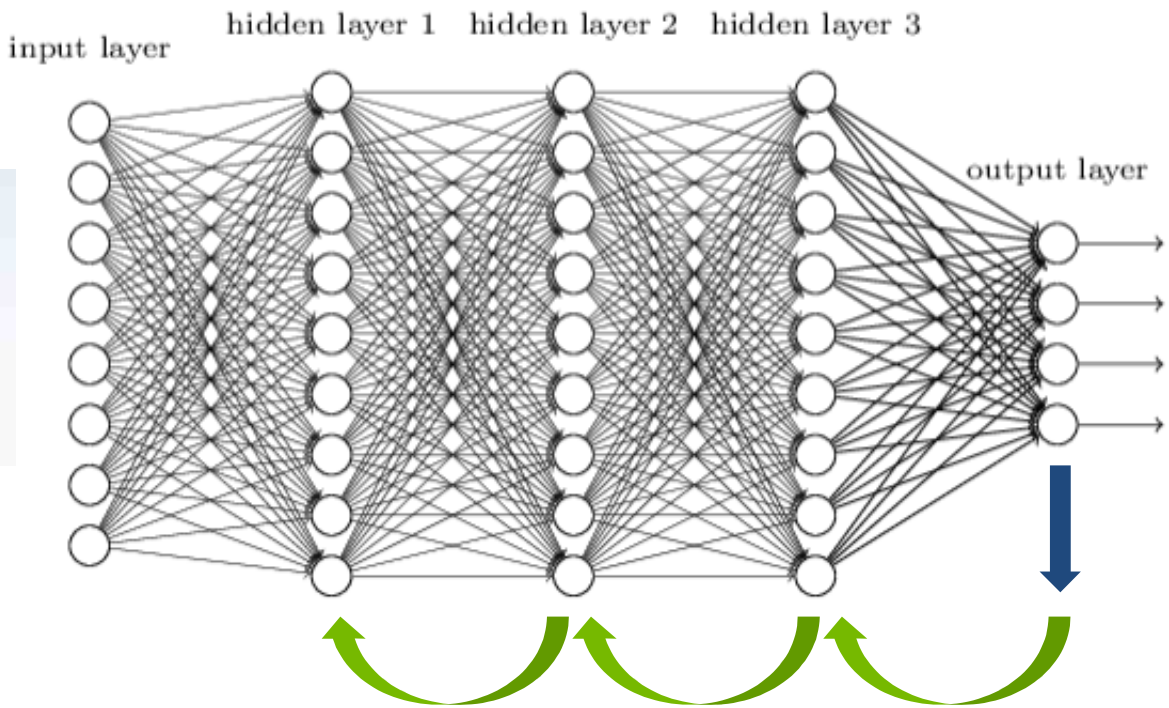
Image source: "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks" ICML 2009 & Comm. ACM 2011. Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng.

RECURRENT NEURAL NETWORKS



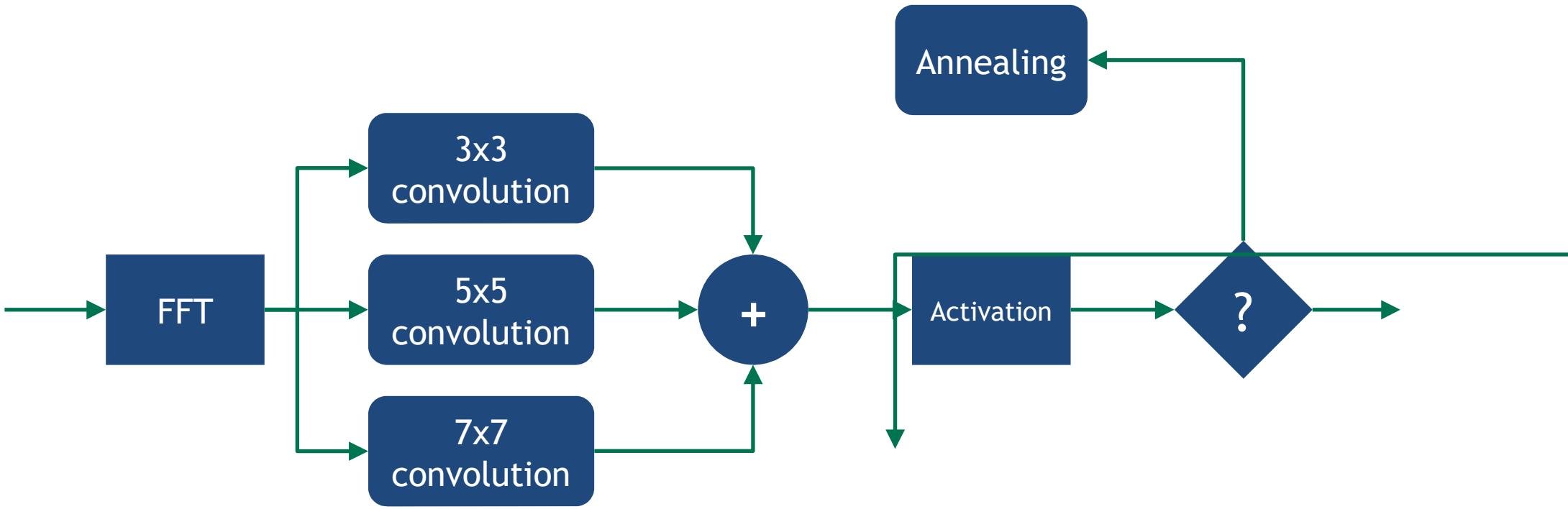
Feedback is fundamental to training: work graph is cyclic

TRAINING DEEP NEURAL NETWORKS



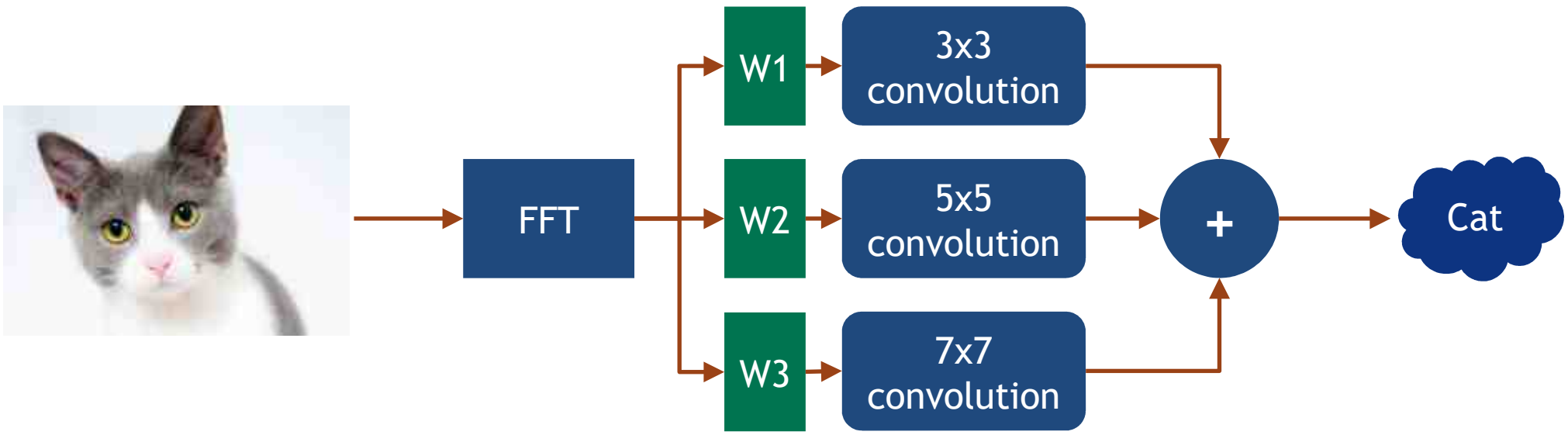
Back-propagate result to adjust layers,
then repeat forward process

EXAMPLE DNN TRAINING WORK GRAPH



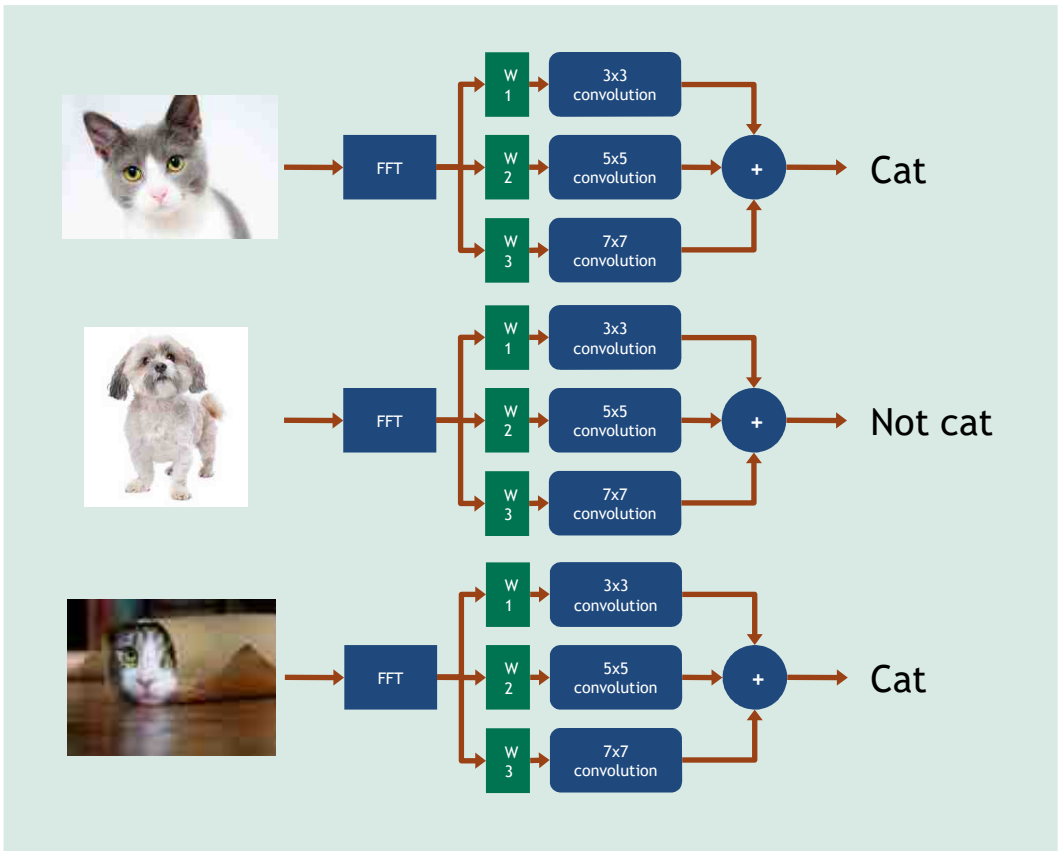
Iterate millions of times of very large input data

USING DNNs: INFERENCE



Inference network is weighted, acyclic version of training network sub-graph.
Network is optimized for size and performance.
Hardware is highly heterogeneous.

INFERENCE BATCHING



Each inference step is small and fast

Aim to process as many inputs as possible to maximize resource use

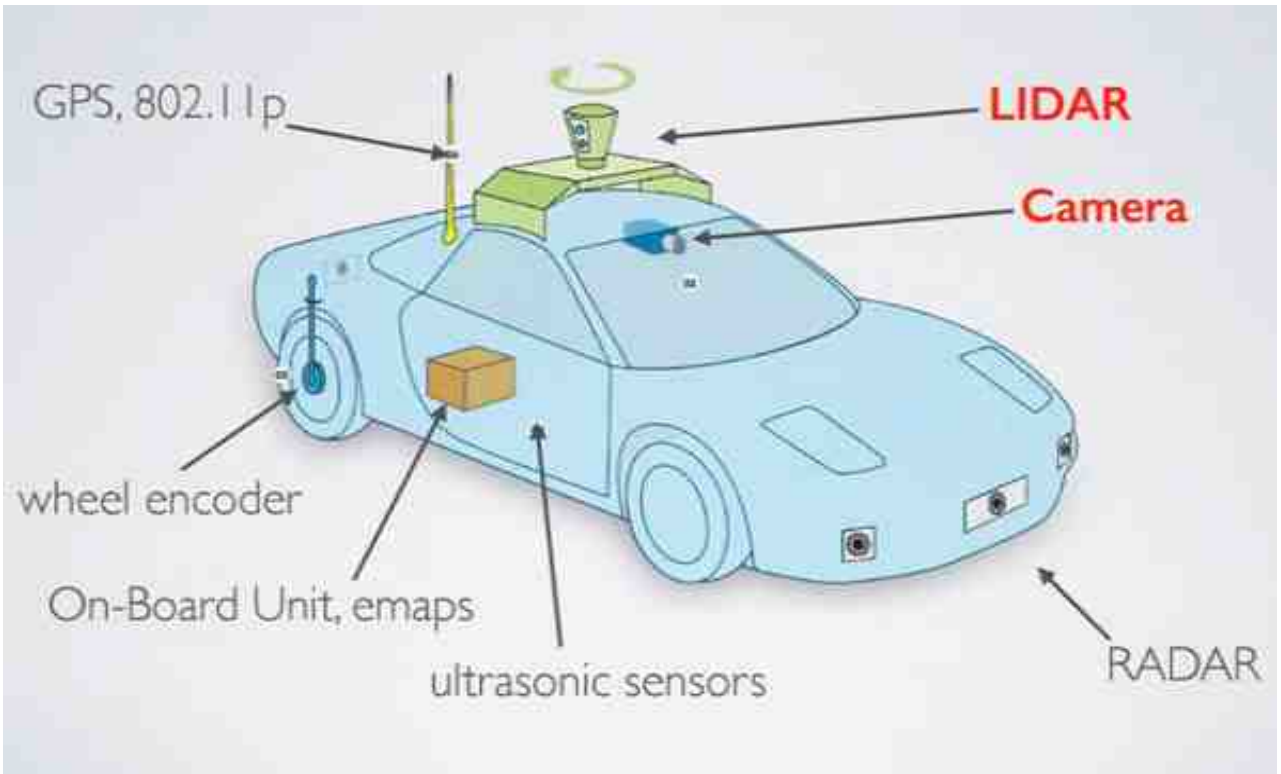
Issue work rapidly to minimize idle time

Requires high degrees of concurrency

Fine-grained scheduling decisions are critical

Extremely latency-sensitive

HETEROGENEOUS INFERENCE SYSTEMS



Multiple co-operating hardware types (CPUs, FPGAs, GPUs)

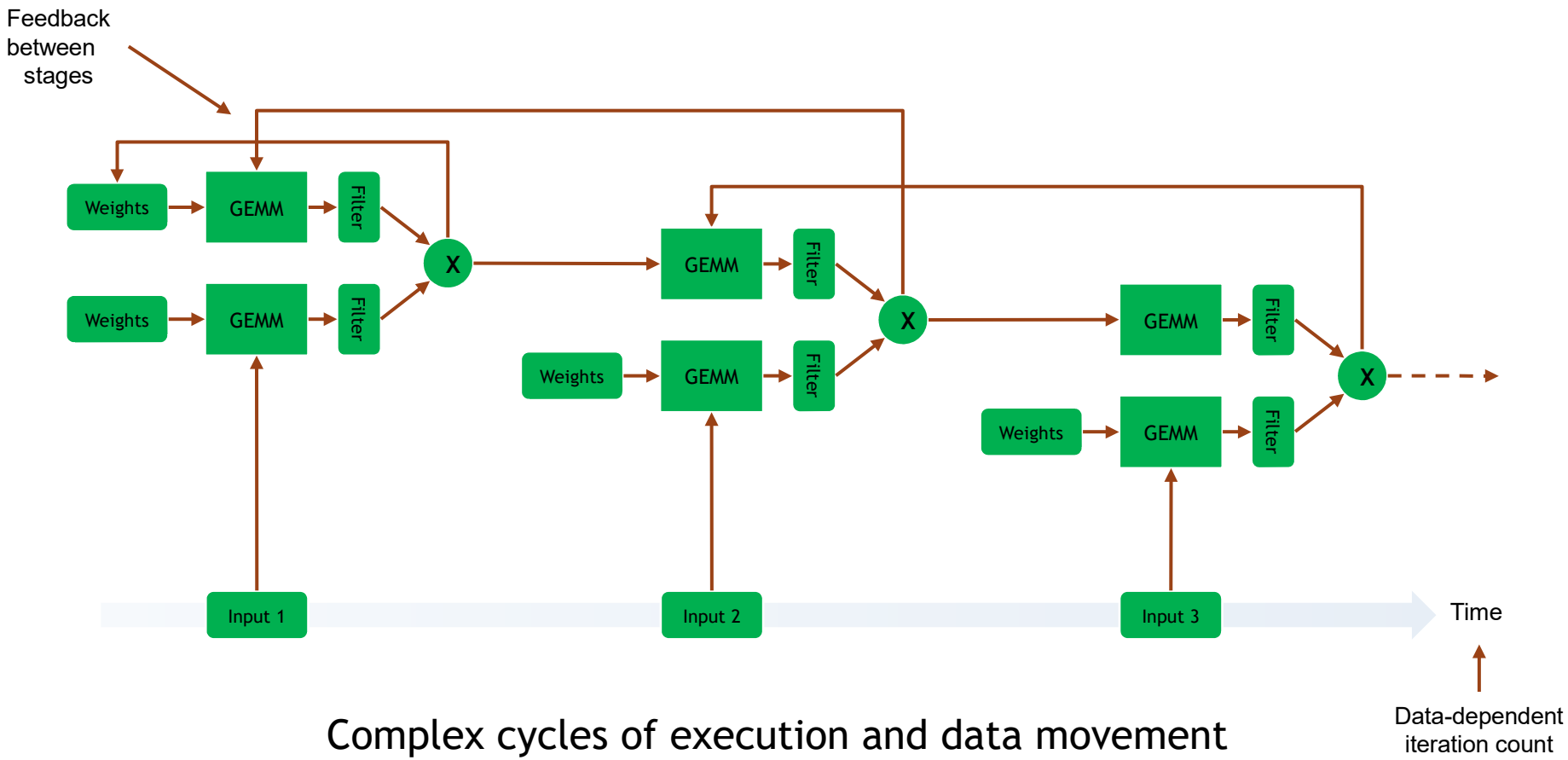
Tight integration between units

Execution control switches between hardware at fine granularity

Input signals arrivinig at widely varying rates

Real-time system constraints

RECURRENT NETWORKS



Complex cycles of execution and data movement
Resource management is critical

DNN REQUESTS FROM HiHAT

Highly asymmetric workloads present very different requirements:

Training

- Large amounts of data require sophisticated communication & memory management
- Large compute loads span multi-node clusters
- Load balancing and resource management important

Inference

- Small, fast kernels are extremely latency-sensitive
- Seeking high degree of concurrency from fine-grained scheduling
- Extremely heterogeneous platforms


NVIDIA