DIANNAO: Only forward pass!

→ Backward, uses the values in forward pass to calculate gradients.

↑ memory usage.

DATA REUSE:

Classifier:

Weights are reused for different inputs. (Tiling is needed)

Convolution:

Inputs/Outputs:

→ Multiple filters/convolutions use same input feature maps to generate output layers (occurs naturally)

Output feature maps

→ Same filter slides → input data reused. (Tiling)

Partial sums stored.

Synapses/Filter weights:

→ Same in a filter layer. One layer of output. (Tiling)

Pooling:

Only sliding and that too 2D. So already can be stored. Tiling not much of a use!

PROCESSING:

→ Staggered Pipeline:

multiply, add all products, sigmoid

→ 16-bit fixed point arithmetic.

Accuracy doesn't vary much even if precision is less!

MEMORY:

→ Scratchpads

→ Separate for input/output/weights

→ Width optimized

→ Conflicts handled as we can decide which input/output or synapses to be loaded based on locality.