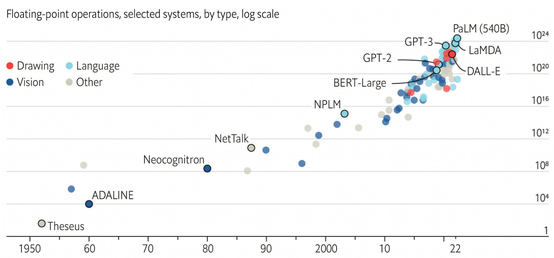
The *BitBrain* method for learning and inference at the edge

Michael Hopkins, Jakub Fil, Steve Furber
The University of Manchester
michael.hopkins@manchester.ac.uk / jakub.fil@manchester.ac.uk

Abstract

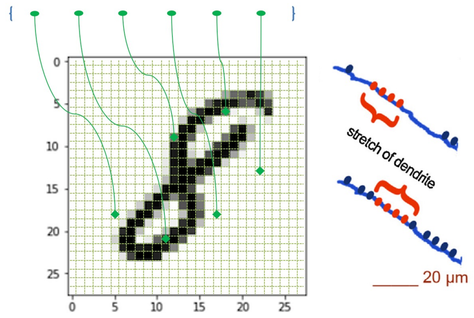
We present an innovative working mechanism (the SBC memory) and surrounding infrastructure (*BitBrain*) based upon a novel synthesis of ideas from sparse coding, computational neuroscience and information theory that enables fast and adaptive learning and accurate, robust inference. The mechanism is designed to be implemented efficiently on current and future neuromorphic devices as well as on more conventional CPU and memory architectures. It provides a unique combination of single-pass, single-shot and continuous supervised learning; following a very simple unsupervised phase. Accurate classification inference that is very robust against imperfect inputs has been demonstrated. These contributions make it uniquely well-suited for edge and IoT applications

Energy consumption



- The number of parameters of deep neural networks has been growing super-exponentially in the recent years.
- The energy usage becomes a significant concern in terms of training deep learning models.
- Training modern deep neural networks often requires thousands of GPUs running for many days e.g., some recent large language models cost millions of dollars to train.

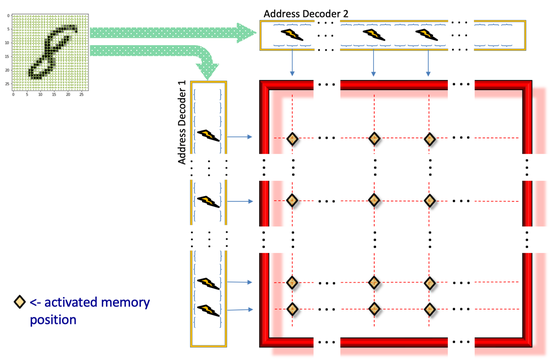
Address Decoder Elements



• Analogous to synaptic cluster/section of dendrite. Each ADE samples a small subset of the input data, e.g. the pixels of an MNIST digit but can be almost any type of data.

• When a weighted sum of the connected input values reaches a threshold (which is learned homeostatically), the ADE will 'fire' - analogous to a dendritic NMDA potential.

SBC memories

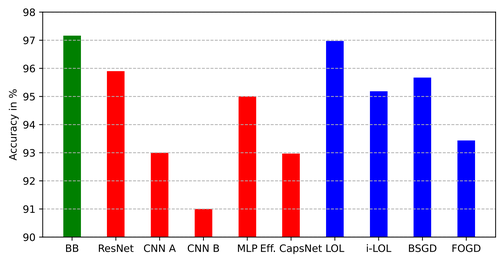


• Activation pattern is *sparse* i.e., only a small percentage of the ADEs in each AD will fire for any given input.

• Each *coincidence* of active ADEs between ADs activates a memory location that reads or writes information about the class which has activated it.

• The SBC memory stores coincidences between features detected in class examples in a training set, and infers the class of a previously unseen test example by identifying the class with which it shares the highest number of feature coincidences

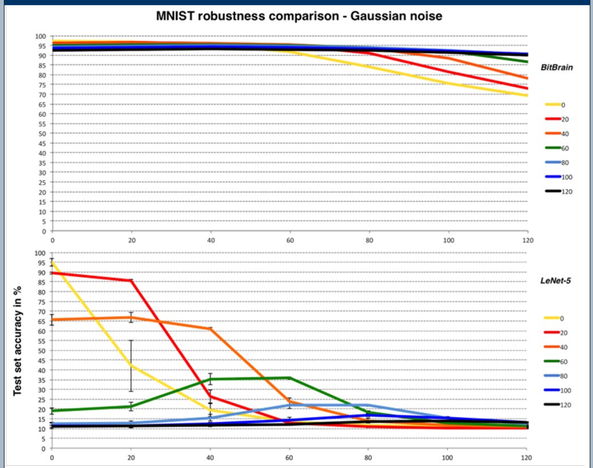
Single-pass learning



• Unlike real neuronal systems, deep learning methods require a huge number of repetitions of the inputs in order to learn.

• *BitBrain* is a state-of-the-art single-pass method - it only needs to see the data once and therefore learns very quickly and using little energy.

Robustness to noise

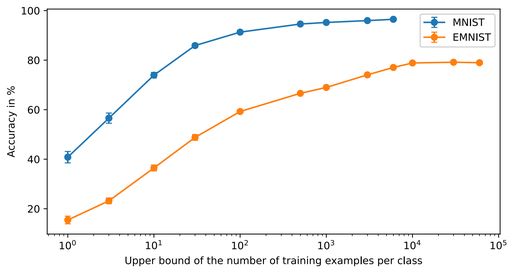


• *BitBrain* performs much better than many deep neural networks given imperfect inputs.

• It can tolerate large amounts of noise and other problems in the data required for making decisions. This makes if far less 'brittle' than other methods.

• This graceful degradation like real neural systems is a very useful feature, especially when the system is deployed e.g., in a noisy environment, with changing conditions or sensor degradation.

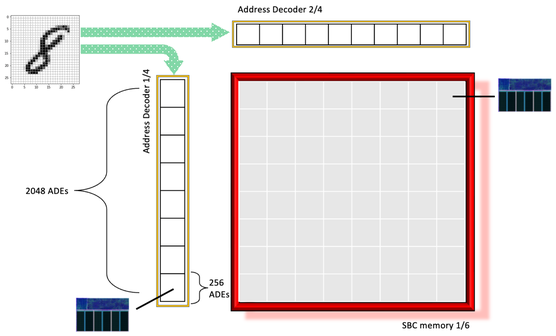
Single-shot learning



• *BitBrain* shows excellent performance on single-shot learning tasks, where the model is given only one example of a new type of input.

• This combination offers continuous and adaptive learning, which is usually either difficult or infeasible with deep neural networks.

SpiNNaker implementation



• *BitBrain* was designed to be compatible with conventional CPUs, but also with energy-efficient, distributed computers, such as SpiNNaker.

• The algorithm can be spread among arbitrarily many cores on SpiNNaker, and thus can make full use of its inherent parallelism.

• The SpiNNaker implementation offers an Improved power-efficiency: ~1W per chip.