One-shot Learning with Memory-Augmented Neural Network

...

Original work by: Adam Santoro et al.
Presented by: Ishan, Konik, Yishuo
2016.11.30

Motivation

NN based classification: Train ~10-100K epochs

- Network extracts appropriate features, associating labels
- Learning wasted when dataset changes

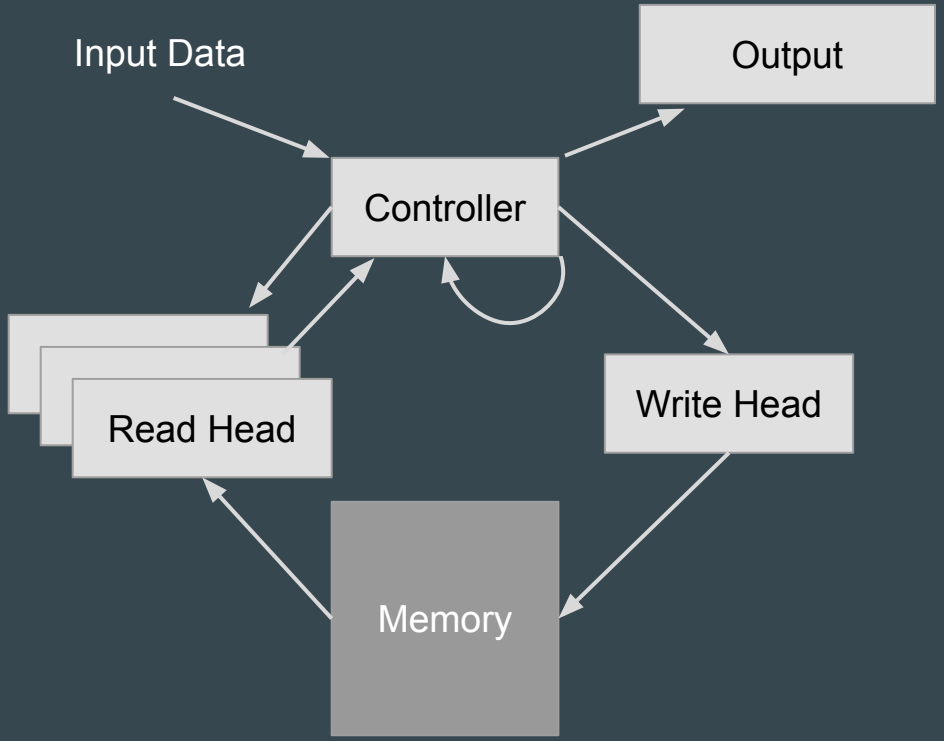
Instead want NN to ‘learn how to learn’

- Should not bind features to labels

Choose parameter set Θ to minimize $E_{D \sim p(D)}[\, L\,(D;\Theta)\,]$ instead of $E\,[\, L\,(D;\Theta)\,]$

NN should observe a few instances and learn the label for that class

Architecture/Literature Review

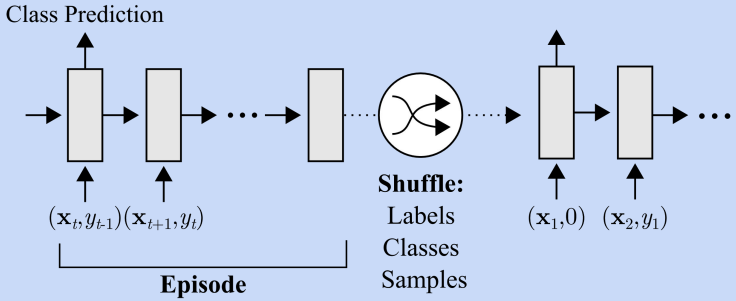


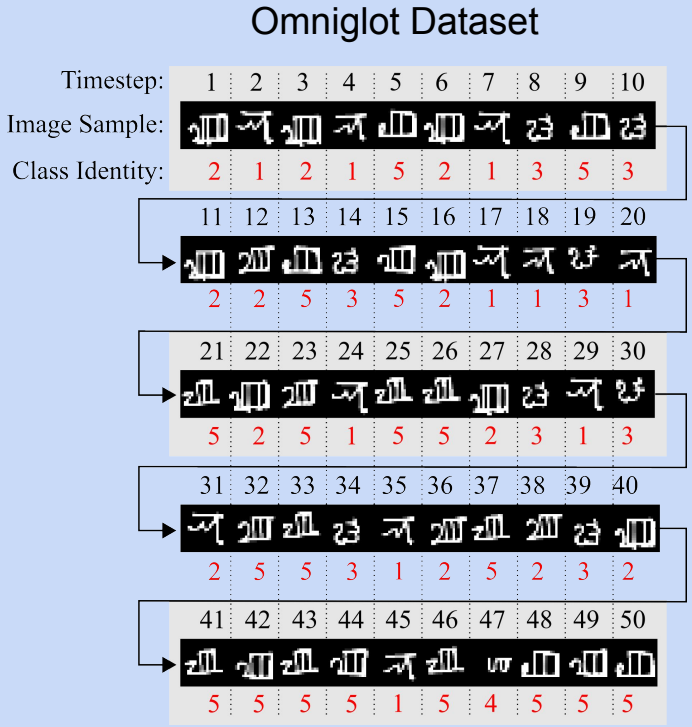
Controller network:
LSTM or Feed-Forward

Least Recent Used Access (LRUA)

- *Rows - read and write weights to generate usage weights per row*
- *Least used row erased at every step*

Architecture schematic





Courtesy: Santoro et al

Hyper-Parameter

Controller(LSTM)

units: 200; learning rate: 1e-4; # read heads: 4

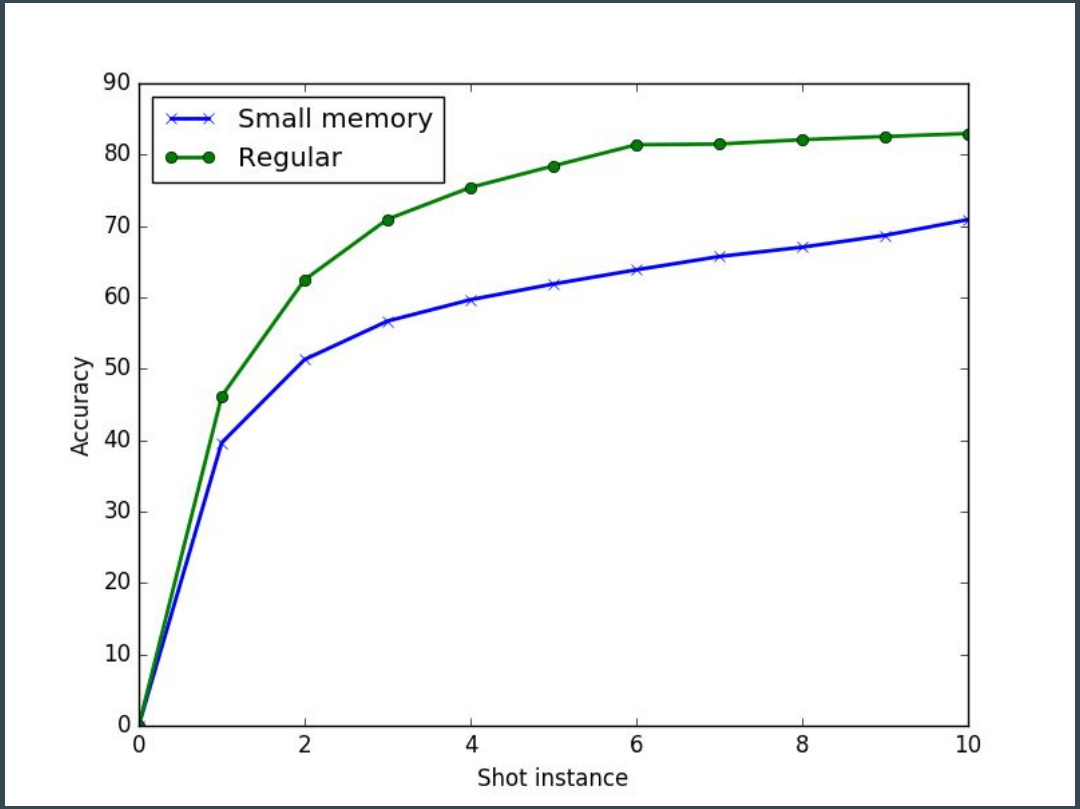
Adam: Learning rate = 1e-3, decay = 0.95

RMSprop: Learning rate: 1e-4; max learning rate: 5e-1; decay: 0.95; momentum: 0.9

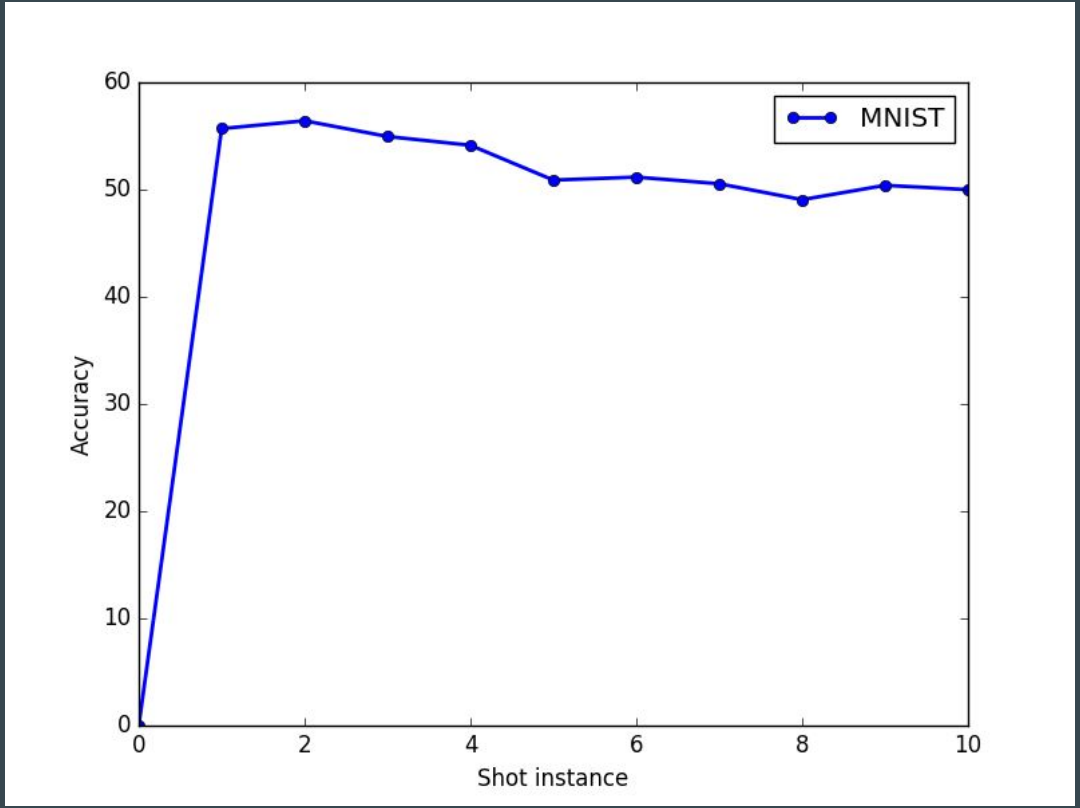
Memory size

- Regular memory: 128 rows, 40 columns
- Small memory: 2 rows, 4 columns

Omniglot-Omniglot



Omniglot-MNIST



Current drawbacks

- Error function and n-shot accuracy not linearly correlated; hence, minimizing error may not lead to direct improvement in accuracy
- (Some) feature learning inevitable in current architecture; is feature-agnostic model possible?
- Cosine similarity restrictive - learn similarity measure: cascaded memory systems

Future plans

Batch normalization:

Current training without BN - 100,000 iterations -> ~80% train accuracy;

Loss functions: Cross-entropy loss

Backprop