Few-Shot Learning with Ferroelectric Analog Content-Addressable Memory

Keshava Katti (SEAS 2022) | Mentors: Dr. Deep Jariwala (Faculty, SEAS, ESE) , Xiwen Liu (PhD Student, SEAS, ESE) | Funding Programs: JFJ, GfFMUR

Motivation

- Humans learn new concepts with very little supervision
- A child can generalize the concept of "giraffe" from a few pictures in a book
- But our best deep learning systems need hundreds or thousands of examples

What is Meta-Learning?

- "Learning to learn" — machine learning (ML) models that can learn new skills, adapt to new environments rapidly with few training examples
- More closely emulates human intelligence

Few-Shot Learning — A Type of Meta-Learning

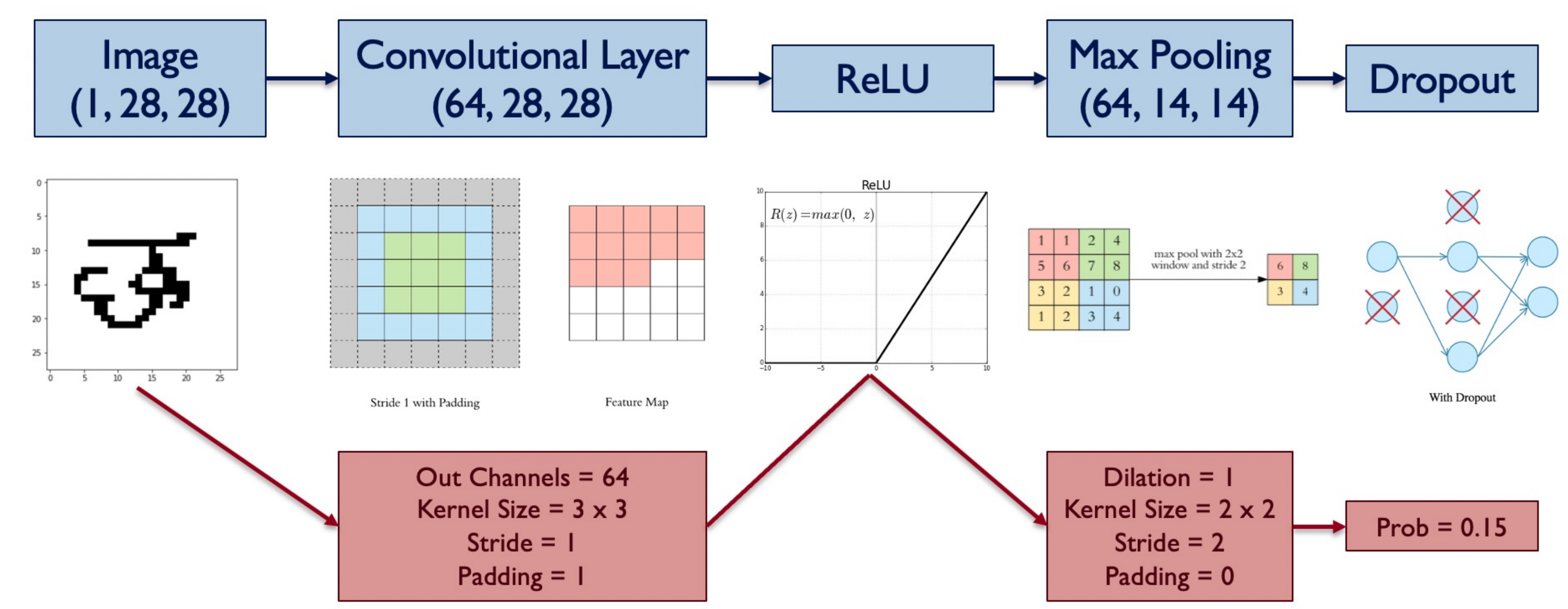
- Model learns a class from few (< 10) labeled examples
- "Lifelong learning" models — continuously learn from small episodes of data containing various unseen classes

Accelerating Few-Shot Learning Via Hardware

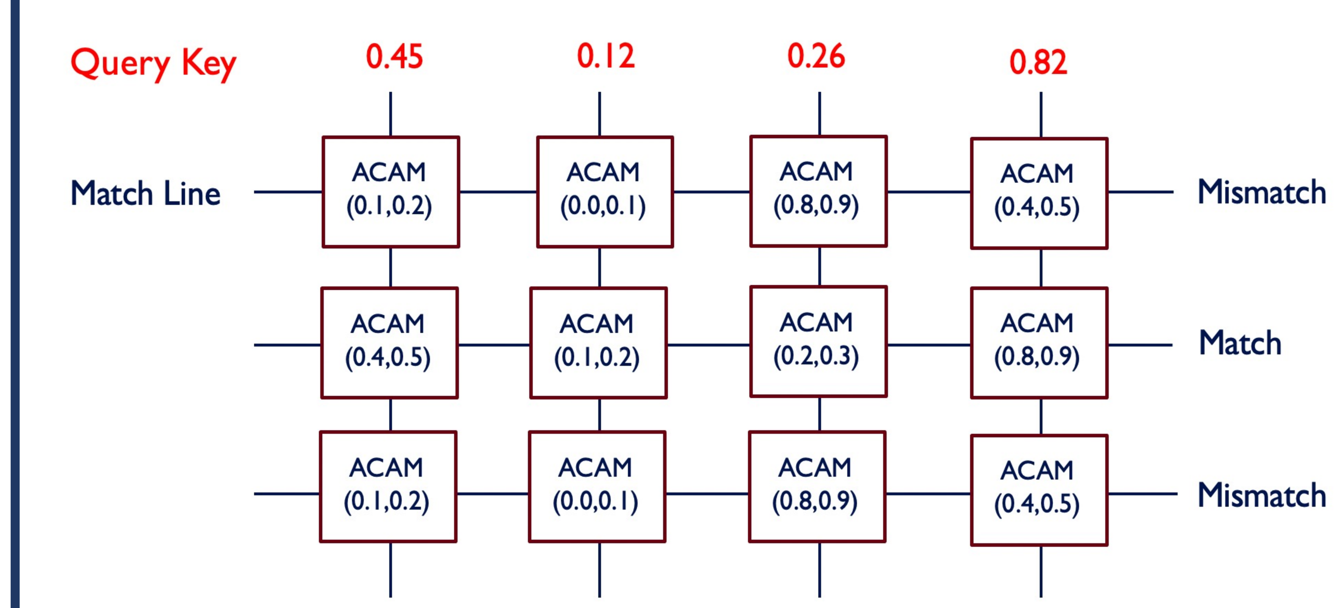
- Goal is to improve energy efficiency, space requirements, and runtime without compromising inference accuracy

Methodology

Convolutional Neural Network (CNN)

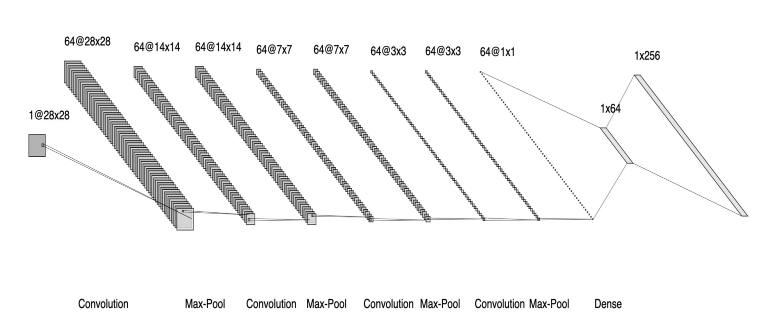


Analog Content-Addressable Memory (ACAM)



Embedding Function

- 4 convolutional layers
- 64-dimensional, real-valued output embedding



Matching Network Algorithm

1. **Given:** Support set $S = ((x_1, y_1), \ldots, (x_n, y_n))$, Query image $Q = (x_q, y_q)$, Embedding function f
2. **Compute:** Attention kernel $a(x_q, x_i)$ using:

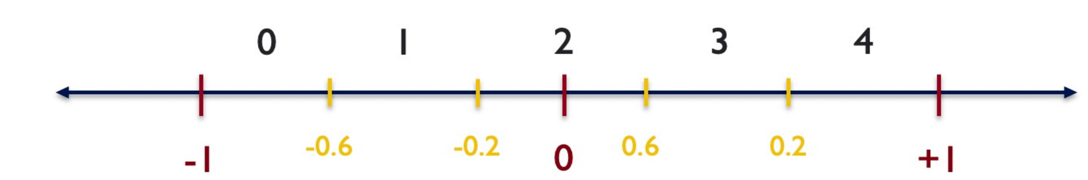
$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

For all $i = 1, \ldots, n$:

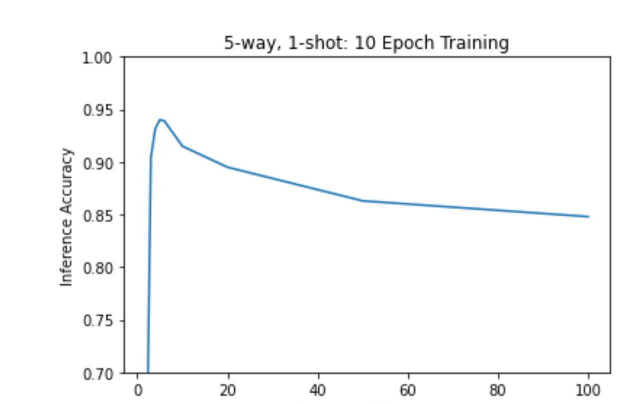
$\widehat{y_q} = argmax_{y_i}\, a(x_q, x_i)$

Backpropagate $cross\_entropy\_loss(y_q, \widehat{y_q})$

Embedding Quantization

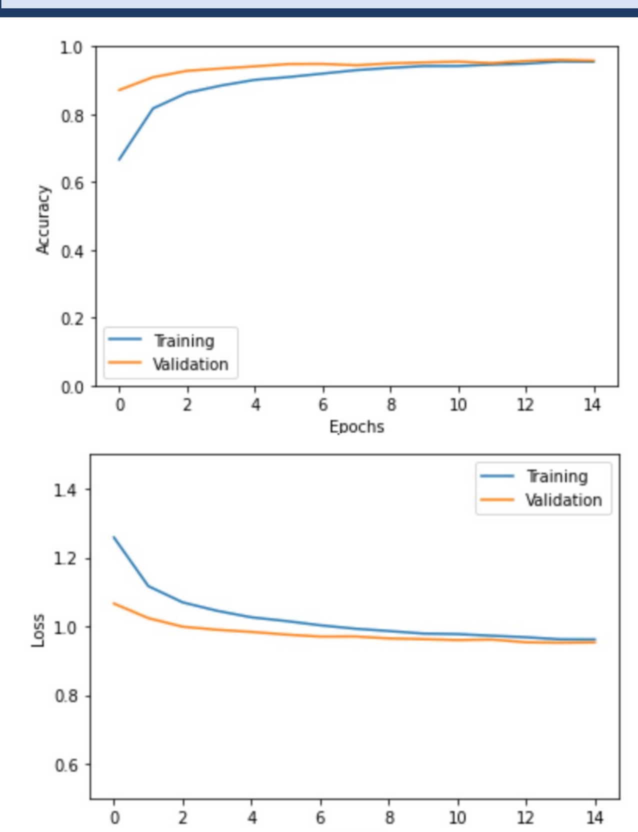


- ACAM stores real values in intervals
- 5 quantization intervals yields optimal inference accuracy



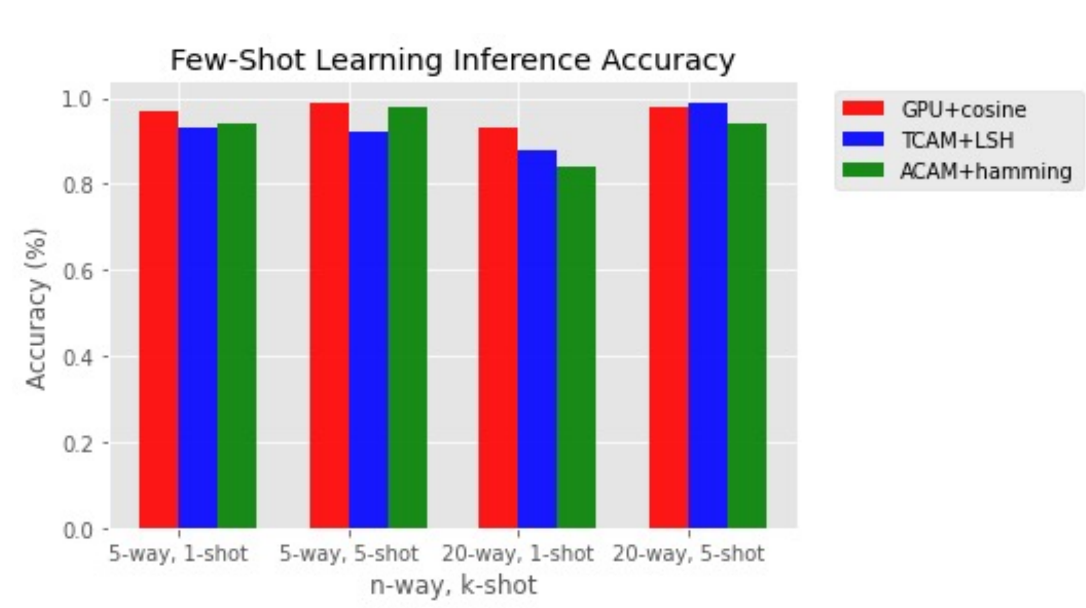
Results

Training / Validation



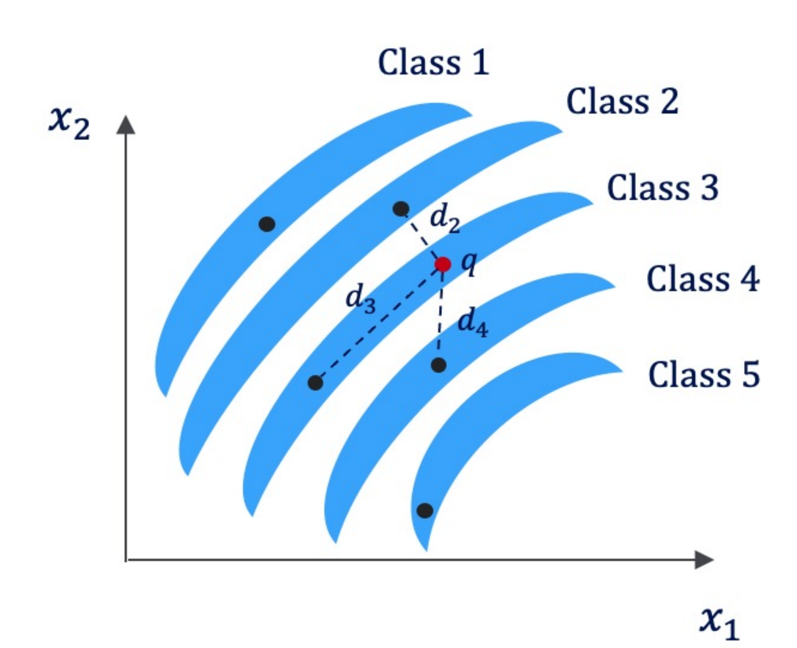
ACAM Accuracy vs GPU, TCAM

- Inference accuracy for GPU+cosine, TCAM+LSH, and ACAM+Hamming are very similar



Conclusion

- Few-shot learning with ACAM requires less energy, space, and search time than alternative (GPU, TCAM) implementations with negligible compromise on inference accuracy
- Future steps — matching network algorithm is not optimal for few-shot learning, implement improved approach



$d_2 < d_4 < d_3 \Rightarrow q \in$ Class 2

References

1. Vinyals, et. al. (2016), "Matching Networks for One-Shot Learning," NeurIPS
2. Dhillon, Chaudhari, et. al. (2020), "A Baseline for Few-Shot Image Classification," ICLR
3. Ni, et. al. (2019), "Ferroelectric ternary content-addressable memory for one-shot learning," Nature Electronics
4. Raginsky & Lazebnik (2009), "Locality-sensitive binary codes for shift-invariant kernels," NeurIPS
5. Andoni & Indyk (2006), "Near-Optimal Hashing Algorithms for Approximate Nearest-Neighbor in High Dimensions," IEEE FOCS
6. Shinde, et. al. (2010), "Similarity search and locality-sensitive hashing using TCAMs," ACM SIGMOD