IBM IEEE CAS/EDS
AI Compute Symposium 2021
October 13-14

Characterizing Neuro-Symbolic Workloads

Zachary Susskind, Bryce Arden, Lizy K. John
The University of Texas at Austin

Patrick Stockton, Eugene B. John
The University of Texas at San Antonio

Introduction

- Recent innovations in machine learning have transformed the field of artificial intelligence
  - The “state-of-the-art” is continually and rapidly changing

- An understanding of the performance of these workloads is needed to propose software and hardware solutions

- In this work, we analyze three workloads in the emerging domain of *Neuro-Symbolic Artificial Intelligence* - NSAI

Background

- Neuro-Symbolic Artificial Intelligence (NSAI) is a hybrid approach to machine learning:
  - Modern, DL-based approaches are good at extracting features from data
  - Traditional, rules-based approaches (“expert systems”) are good at operating on these features
  - Idea: Combine both approaches into a single model

- NSAI is a novel field, and the performance characteristics of models are not well-understood in literature

Model Overview

- We analyzed three separate NSAI models

- Neuro-Symbolic Concept Learner / NSCL (MIT/IBM):
  - NS model for image question answering
  - Takes scene and question as inputs; outputs prediction

- Neuro-Symbolic Dynamic Reasoning / NS-DR (MIT/IBM)
  - *Video* question answering - objects move and collide

- Neural Logic Machines / NLM (Google)
  - Framework for neuro-symbolic models
  - Can be trained for a variety of tasks

Methodology

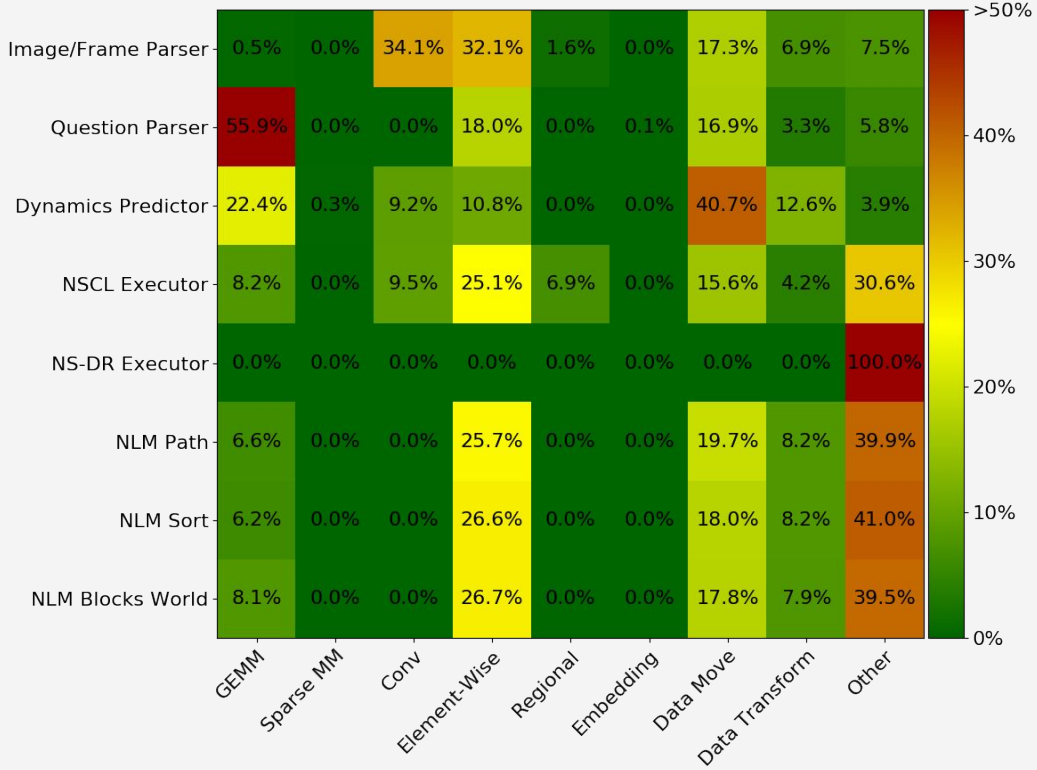
- Analyze performance characteristics of the three models:
  - NSCL and NS-DR have independent submodels
    - Some submodels are neural; other are symbolic
    - Analyze these individually
  - NLM is a framework which can be specialized for different tasks
    - Looked at three distinct tasks

- Performed profiling at a function level
  - Analyze where the CPU and GPU are spending the most time
  - Gives a high-level overview of performance behavior
  - Most models use the PyTorch library, which has built-in tools for doing this

Activity Characterization

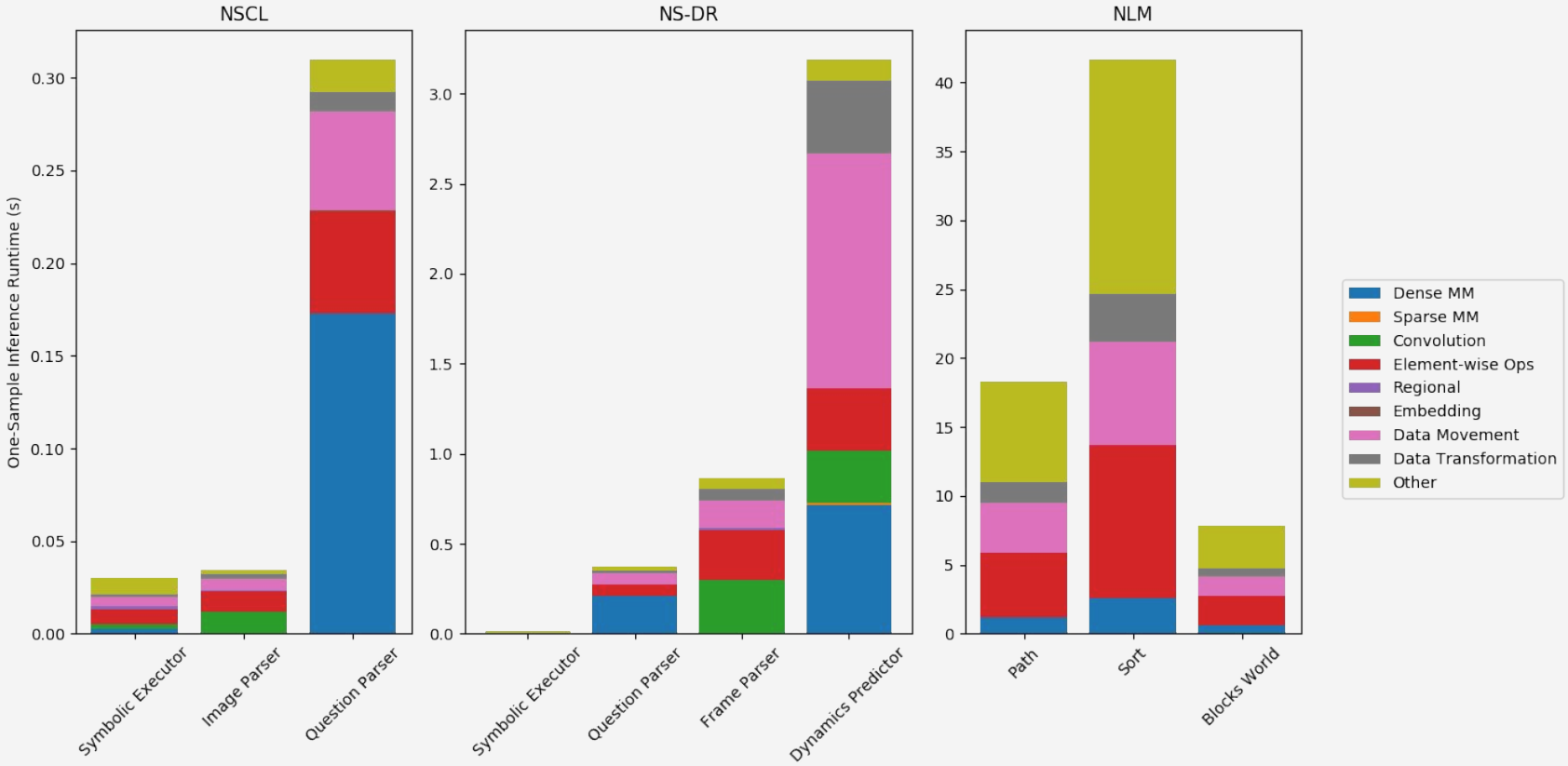
- These models do not have clearly dominant functions
  - However, many functions are similar
    - e.g. tensor addition and subtraction
  - Developed a categorization scheme to report results
    - Dense matrix multiplication
    - Sparse matrix multiplication
    - Convolution
    - Element-wise Tensor Operations
    - Regional Tensor Operations
    - Embedding Lookups
    - Data Movement
    - Data Transformation

Results: Work Breakdown



- Values normalized to execution time of workload
- Takeaways:
  - NS workloads spend proportionately more time on “element-wise” ops
  - Low operational intensity, “streaming” access patterns

Results: Runtime Breakdown



Analysis

- NSCL and NS-DR:
  - Symbolic execution is comparatively fast
    - This is a claimed strength of symbolic AI
  - Dynamics prediction in NS-DR incurs substantial overhead
    - Image and question parser are well-optimized models
    - Dynamics prediction is a relatively novel area of DL
    - Should explore ways to reduce predictor's data movement

- NLM:
  - Element-wise operation and data movement dominate
    - Ratios are similar between tasks, despite different run times
    - These ops are easy to parallelize (SIMD-like), but have poor operational intensity

Conclusion

- Conclusion
  - DL models and frameworks are fast-moving - this presents both opportunities and challenges
  - Analysis shows bottlenecks and opportunities for parallelization

- Future Work
  - Contrast against DL-only models for the same tasks
  - Profile on different platforms and at microarchitectural level

- Acknowledgement
  - This research was supported in part by Semiconductor Research Corporation (SRC) Task 3015.001/3016.001 and National Science Foundation grant number 1763848. Any opinions, findings, conclusions or recommendations are those of the authors and not of the funding agencies.