Machine learning for timing estimation
Abderaouf Nassim Amalou, Elisa Fromont, Isabelle Puaut

To cite this version:
Abderaouf Nassim Amalou, Elisa Fromont, Isabelle Puaut. Machine learning for timing estimation. D3 - Architecture séminaire: PhD days / Journées doctorants – 2022, Nov 2022, Rennes, France. hal-04260161

HAL Id: hal-04260161
https://hal.science/hal-04260161
Submitted on 26 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
Machine learning for timing estimation

1. CONTEXT
The execution time is calculated in specific cases depending on the needed application domain, (e.g. worst case execution for realtime applications, average case for compiler optimization).

Challenges of using Machine Learning for timing estimation:
- Context-awareness (Cache and Pipeline).
- Representativeness of training data.
- Reliability of the estimations.

Figure 1 Execution time estimation classification

Figure 2 Challenges to estimate the execution time of modern processor

3. TIMING AVERAGE CASE

- ITHEMAL [3] uses 2 stacked LSTMs layers to estimate the execution time for a BB in isolation.
- CATREEN [2] improves ITHEMAL to estimate the execution time of a BB in its context.
  - Instruction == Word.
  - Basic block == Sentence.
  - Sequence of basic blocks == Paragraph.

Figure 3 Training/Estimation workflow of WE-HML

4. RESULTS

Figure 4 Unfolding of a LSTM cell of the instruction layer on a machine instruction

Figure 5 Architecture of CATREEN

Figure 6 Estimated WCET WE-HML versus Maximum Observed Execution Time (MOET)

Figure 7 Absolute Percentage Error on MiBench for CATREEN, ITHEMAL and MLP regressor

5. PERSPECTIVE

- Transformers [4] are able to handle sequential data, but this is done in parallel, which drastically reduces both training and inference times.

Abderaouf Nassim AMALOU
PACAP/LACODAM

Pr. Isabelle PUAUT
PACAP

Pr. Elisa FROMONT
LACODAM

Bibliography