Boosted Regression Trees for Data-Driven Auto-Tuning

First Last
Affiliation line 1
Affiliation line 2
anon@mail.com

First Last
Affiliation line 1
Affiliation line 2
anon@mail.com

First Last
Affiliation line 1
Affiliation line 2
anon@mail.com

ABSTRACT

Auto-tuning is a widely used and effective technique for optimizing a parametrized GPU code template for a particular computation on particular hardware. Its drawback is that thorough or exhaustive auto-tuning requires compiling many kernels and calling each one many times; this process is slow. Furthermore, library abstraction boundaries provide operations such as image filtering and matrix multiplication, which actually correspond to a large set of potential problem configurations with a wide variety of memory access patterns and computational bottlenecks. How can we draw on data from previous auto-tuning of related problems on related hardware to make a just-in-time implementation decision for a novel problem? This paper presents a machine learning approach to auto-tuning, in which features of (a) the current hardware platform, (b) the kernel configuration and (c) the problem instance are passed to a regression model (boosted regression trees) which predicts how much faster this kernel will be than a reference baseline. Combinatorial optimization strategies that would normally implement auto-tuning by evaluating kernel configurations on the real hardware are orders of magnitude faster when evaluating the regression model instead. We validate our approach using the filterbank correlation kernel described in [anon], where we find that 0.1 seconds of hill climbing on the regression model can achieve an average of 90% of the speed brought by 120 seconds of real auto-tuning. Our approach is not specific to filterbank correlation, or even to GPU auto-tuning: the approach of using a non-linear regression model on top of simple features applies to a variety of problem types, kernel types, and platforms.