Generalizable and Interpretable Learning for Configuration Extrapolation

Yi Ding  
MIT CSAIL  
Cambridge, MA, USA  
ding1@csail.mit.edu

Michael Carbin  
MIT CSAIL  
Cambridge, MA, USA  
mcarbin@csail.mit.edu

Ahsan Pervaiz  
University of Chicago  
Chicago, IL, USA  
ahsanp@uchicago.edu

Henry Hoffmann  
University of Chicago  
Chicago, IL, USA  
hankhoffmann@cs.uchicago.edu

ABSTRACT
Modern software applications are increasingly configurable, which puts a burden on users to tune these configurations for their target hardware and workloads. To help users, machine learning techniques can model the complex relationships between software configuration parameters and performance. While powerful, these learners have two major drawbacks: (1) they rarely incorporate prior knowledge and (2) they produce outputs that are not interpretable by users. These limitations make it difficult to (1) leverage information a user has already collected (e.g., tuning for new hardware using the best configurations from old hardware) and (2) gain insights into the learner’s behavior (e.g., understanding why the learner chose different configurations on different hardware or for different workloads). To address these issues, this paper presents two configuration optimization tools, Gil and Gil+, using the proposed generalizable and interpretable learning approaches. To incorporate prior knowledge, the proposed tools (1) start from known configurations, (2) iteratively construct a new linear model, (3) extrapolate better performance configurations from that model, and (4) repeat. Since the base learners are linear models, these tools are inherently interpretable. We enhance this property with a graphical representation of how they arrived at the highest performance configuration. We evaluate Gil and Gil+ by using them to configure Apache Spark workloads on different hardware platforms and find that, compared to prior work, Gil and Gil+ produce comparable, and sometimes even better performance configurations, but with interpretable results.

CCS CONCEPTS
• Software and its engineering → Software configuration management and version control systems; • Computing methodologies → Machine learning approaches.

KEYWORDS
Configuration, machine learning, generalizability, interpretability

ACM Reference Format:

1 INTRODUCTION
The increasing configurability of modern software makes it challenging for users to tune performance due to high complexity: tremendous configuration spaces and complicated interactions between these configuration parameters [39, 45, 49, 50]. Configurations have a large influence on application performance such as latency [42], throughput [3], and energy consumption [8]. As a result, tools to help tune application configurations for high performance have become a crucial yet challenging research area.

To configure software applications efficiently, machine learning (ML) approaches have been applied to model the complex relationships between configuration parameters and performance. Most prior work on incorporating ML approaches in software performance modeling is to first randomly sample assignments of configuration parameters for the application, measure the application’s performance with the sampled parameters, train a learner on these samples, predict the performance for unsampled configurations, and then deploy the software in the configuration with the best predicted performance [1, 18, 23, 47]. The most sophisticated of these learners—e.g., neural networks [16], random forests [33], and Gaussian process regression [7, 29]—are black-box, meaning that users have no visibility into their internal workings [9].

Although black-box ML approaches are effective at software performance modeling, there are two practical limitations to be addressed for efficient, optimal configuration search:
• Difficulty incorporating prior knowledge. Prior ML-based tools have little ability to leverage information a user has previously collected. Such tools require significant data collection: they must start from scratch by generating numerous random configurations and collecting their performance. Two specific scenarios are considered as follows:

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ESEC/FSE ’21, August 23–28, 2021, Athens, Greece © 2021 Copyright held by the owner/author(s).
https://doi.org/10.1145/3468264.3468603
- Scenario (1): Workload X runs slowly due to some faulty configuration settings [50, 52]. To find a better configuration, a user can randomly sample new configurations from scratch, run X for each to collect performance data, and build a new ML model. To avoid this tedious process, a user can instead leverage the known configurations—even though they are slow—to build a model that can extrapolate to find a better configuration. This strategy saves time as it does not need new initial training data collection.

- Scenario (2): Workload Y has been tuned to work at the optimal workload of HiBench [14], which is a widely used benchmark suite with a hierarchical model connecting application-level configuration and the computer system on which it runs, extra low-level system layer in-between, or scheduling issues, A is no longer available and Y has to be deployed on hardware B. Using prior approaches, users would randomly sample configurations on B, collect performance data, and build a ML model for configuration search [13, 20, 28]. An alternative is to incorporate prior knowledge by using configurations known to run fast on A and developing a model to extrapolate the configuration space from A to B. In §5.3, we show that incorporating prior knowledge achieves better performance than starting from scratch with random sampling.

- Lack of interpretability. Prior work evaluates their black-box models only for their performance prediction accuracy; it is difficult for users to interpret the prediction results [32]. Considering the above two scenarios:
  - Scenario (1): Users hope to understand the underlying factors—from software applications to hardware resources—that cause the low performance and thus avoid the similar issues in the future [25, 43, 48].
  - Scenario (2): Users hope to interact with the learners and gain insights into why prior configurations are slow and how faster configurations are achieved. They also hope such knowledge can generalize across different hardware and workloads [47, 51].

To address these limitations, we present Gil and Gil+: configuration extrapolation tools that incorporate prior knowledge to achieve high performance and provide interpretability to advance human knowledge. To incorporate prior knowledge, Gil and Gil+ use existing information about configurations as initial training samples and then extrapolate to even higher performance configurations. To achieve interpretability, the key insight is to iteratively construct linear models that approximate the true nonlinear function such that the most important configuration parameters can be interpreted as a combination of both their weights in the linear model and the changes in those weights as the model is iteratively updated. To help understand the relationships between the application and the computer system on which it runs, Gil+ augments Gil with a hierarchical model that connects application-level configurations, low-level system metrics (e.g., context switches and cache misses), and the final application performance (see detailed definitions in §3.1). Additionally, we develop graphical tools to visualize these relationships to help users interpret learning results.

We implement Gil and Gil+, and evaluate them on Apache Spark workloads from HiBench [14] on three different hardware from a public heterogeneous cloud system—the Chameleon Cloud Research Platform [21]. We use the names that Chameleon 1 uses for different hardware with details shown in Table 2, which includes three hardware microarchitectures—Skylake, Haswell, and Storage; the first two are compute servers, the last is optimized for disk bandwidth. We determine the best configuration for each workload and specific hardware by exhaustive search among our measurements.

Gil+ develops a hierarchical model connecting application-level configurations, low-level system metrics, and performance shown in Figure 1 (see detailed definitions in §3.1). In particular, Gil+ builds a linear model between low-level system metrics and performance, and for each low-level system metric, Gil+ builds a linear model between application-level configuration parameters and it. With the extra low-level system layer in-between, Gil+ makes it easier for all learners, incorporating prior knowledge achieves 2–15% performance improvements compared to starting from scratch with random sampling for initial training (§5.3).

Gil+ enables the users to interact with the learners and interpret relationships between application-level configurations, low-level system metrics, and performance via visualization tools (§5.4).

In summary, this work makes the following contributions:

- Gil and Gil+ achieve better extrapolation performance than random sampling, neural networks [16] and genetic algorithms [53], with results comparable to Bayesian optimization [33] (§5.1, §5.2).
- For all learners, incorporating prior knowledge achieves 2–15% performance improvements compared to starting from scratch with random sampling for initial training (§5.3).
- Gil+ enables the users to interact with the learners and interpret relationships between application-level configurations, low-level system metrics, and performance via visualization tools (§5.4).
- Releasing code and data for Gil and Gil+ in https://github.com/yding/gil.

2 MOTIVATIONAL EXAMPLE

As an example of the value our tools, we study Apache Spark workloads from HiBench [14] on three different hardware from a public heterogeneous cloud system—the Chameleon Cloud Research Platform [21]. We use the names that Chameleon 1 uses for different hardware with details shown in Table 2, which includes three hardware microarchitectures—Skylake, Haswell, and Storage; the first two are compute servers, the last is optimized for disk bandwidth. We determine the best configuration for each workload and specific hardware by exhaustive search among our measurements.

Gil+ develops a hierarchical model connecting application-level configurations, low-level system metrics, and performance shown in Figure 1 (see detailed definitions in §3.1). In particular, Gil+ builds a linear model between low-level system metrics and performance, and for each low-level system metric, Gil+ builds a linear model between application-level configuration parameters and it. With the extra low-level system layer in-between, Gil+ makes it easier for all learners, incorporating prior knowledge achieves 2–15% performance improvements compared to starting from scratch with random sampling for initial training (§5.3).

Gil+ enables the users to interact with the learners and interpret relationships between application-level configurations, low-level system metrics, and performance via visualization tools (§5.4).

In summary, this work makes the following contributions:

- Gil and Gil+ achieve better extrapolation performance than random sampling, neural networks [16] and genetic algorithms [53], with results comparable to Bayesian optimization [33] (§5.1, §5.2).
- For all learners, incorporating prior knowledge achieves 2–15% performance improvements compared to starting from scratch with random sampling for initial training (§5.3).
- Gil+ enables the users to interact with the learners and interpret relationships between application-level configurations, low-level system metrics, and performance via visualization tools (§5.4).
- Releasing code and data for Gil and Gil+ in https://github.com/yding/gil.

1https://www.chameleoncloud.org/hardware/
This chart shows that IPC (instructions per cycle) changes most—that higher IPC is associated with better performance on Storage. See §3.4 to know more about how we visualize the results. The value on each axis shows how the corresponding low-level system metric influences performance when moving from Haswell (Before) to Storage (After).

Figure 2a shows the nweight workload results between low-level systems metrics and performance. The value on each axis shows how the corresponding low-level system metric influences performance when moving from Haswell (Before) to Storage (After). For users to reason about the hardware and software cross-stack interactions as well as tune application-level configurations.

To illustrate GIL’s behavior, we use the nweight workload from HiBench. Consider the case where Haswell is available first for deployment, and the high-performance configurations have been tuned for Haswell, either by human users or learners. Later, a new class of Storage machines with a small processor downgrade (slightly reduced clockspeed and last-level cache) and huge storage system upgrade (see details in Table 2) becomes available. A reasonable hypothesis would be that for the nweight workload, the upgraded storage system will outweigh the small downgradation in computing power. Thus when we deploy nweight on Storage, it is natural to start the high-performance Haswell configurations and run them directly on Storage.

Unfortunately, we find that the performance (throughput) can be more than 20% worse on Storage than Haswell, for the same configuration. Even if black-box learners from prior work are used to tune configurations using these available configurations, the best possible performance we find on Storage is still 10% worse than the best Haswell performance. Even worse, these black-box learners provide no interpretation of their results. Thus, at this point, the user might be wondering if Storage is fundamentally slower for nweight, or if the black-box learners simply need more time to search for a better configuration. With a black-box learner, there is no way for a user to gain insight into these questions. GIL, on the other hand provides an interpretable model that a user can inspect.

Figure 2b shows the lr interpretation results between application performance and low-level system metrics before and after applying GIL to tune application-level configurations on Storage. In other words, the before configurations are those that perform well on Haswell and the after configurations are those that perform well on Storage. The label for each axis in the radar chart represents a specific low-level system metric (see details in Table 3) and the value on each axis is the learned linear coefficient mapping from the low-level system metric to application-level performance. See §3.4 to know more about how we visualize the results. This chart shows that IPC (instructions per cycle) changes most—a sharp spike towards IPC—increase after applying GIL, which indicates that higher IPC is associated with better performance on Storage and thus application-level configurations should be tuned to increase IPC. These results immediately provide insights into why Haswell performs better than Storage: Haswell’s faster clock and larger last-level cache are likely beneficial for an application that wants increased computational power. Being able to interpret the model gives us certain confirmation that Haswell is really faster than Storage for this workload. Furthermore, GIL can be used to discover workloads that behave similarly and infer the possible optimal deployment for new workloads. Figure 2b shows the radar chart for lr: it behaves similarly to nweight: both having sharp spikes towards IPC, which indicates that they would likely benefit from the same hardware. In other words, it may not be worthwhile running lr on Storage, because nweight is faster on Haswell than storage and it is likely that lr will be faster on Haswell as well. In contrast, black-box learners cannot report back this type of information to human users.

Nevertheless, no configuration optimizer can directly affect IPC, instead they must tune the application-level configurations to increase nweight’s IPC and, then, its performance. Therefore, GIL builds interpretable models that map application-level configurations and low-level system metrics to tune configuration parameters. Figure 3 shows the absolute linear coefficient differences between after and before applying GIL (starting from configurations that perform well on Haswell and arriving at configurations that perform well on Storage). This figure shows that, although Storage has only a small difference in processor, the application-level configurations that help achieve this throughput change greatly (i.e., spark.broadcast.blockSize). In contrast, black-box learners are not able to provide such insights to help users understand why they get different optimal configurations for different hardware. Since GIL starts from prior knowledge—configurations that work well on Haswell—and then finds high-performance configurations on Storage, this provides users insights into the differences between
A workload is a configurable software application. We where 

\[ x_i = [x_{i1}, \ldots, x_{ij}, \ldots, x_{ip}] \] (1)

where \( x_i \) is the \( i \)-th configuration, \( x_{ij} \) is the value of the \( j \)-th configuration parameter in the \( i \)-th configuration. For instance, \( p \) is 20 corresponding to the 20 Apache Spark configuration parameters in our evaluation (§5) (shown in Table 1).

**Low-level system metric (LLSM).** LLSM is the quantifiable behavior of the underlying OS (e.g., context switches) and hardware (e.g., cache misses) that helps understand the application-level performance. These are represented as a \( d \)-dimensional vector assuming there are \( d \) metrics to measure and understand:

\[ u_i = [u_{i1}, \ldots, u_{ij}, \ldots, u_{id}] \] (2)

where \( u_i \) is the \( i \)-th low-level system metric vector, \( u_{ij} \) is the value of the \( j \)-th metric in the \( i \)-th low-level system metric vector. For example, in this work \( d \) is 5 as five low-level system metrics are studied as shown in Table 3.

**Performance.** Performance is the measured behavior for a workload on certain hardware. This work focuses on throughput, which is the rate of the total data delivered in the network [3]. Each configuration and its corresponding measurement—including low-level system metrics and performance—is a sample. Gil and Gil+ use these samples to build performance models to search the optimal configuration as an output. Specifically, Gil and Gil+ build a series of models relating application-level configurations to performance, a series of models relating the low-level system metrics to performance, and a series of models relating application-level configurations to each low-level system metric. Then, the coefficients learned from these models are passed to model interpretation to visualize the results as another output. Next, we will elaborate the performance models and interpretation visualization.

### 3.2 Gil Performance Model

To improve performance accuracy and promote search space exploration, Gil iteratively interacts with the workload by sequentially learning an unknown function \( f \)—modeling the relationships between application configuration and performance—with a sample budget. At each iteration, limited configurations are selected to update \( f \). After all iterations are completed, Gil uses the learned \( f \) to evaluate the remaining (untested) configurations to find the one with the best predicted performance. In this process, two challenges (CHs) need to be addressed:

**CH1:** Designing a model to learn the unknown function \( f \).

**CH2:** Formulating a query strategy to select the samples.

We will address these two CHs as follows.

#### 3.2.1 Learning Linear Functions.

To learn the unknown function \( f \), Gil uses a linear regression model [10]. Gil chooses linear model over other nonlinear models, such as neural networks [12] and Gaussian process regression [7, 29], due to several advantages:

- **Better extrapolation ability.** Although nonlinear models usually achieve higher overall prediction accuracy than linear models, they are more likely to overfit the training set, which leads to poor extrapolation performance when facing unseen data points [6]. In contrast, linear models are less likely to overfit and thus generalize better on unseen data.

- **Interpretability.** Linear models are naturally interpretable: the linear coefficients quantify the relationships between independent metrics and performance: a positive coefficient indicates positive correlation between a metric and performance, while a negative coefficient indicates negative correlation. The coefficients’ magnitudes represent the strength of these relationships.
For the $i$-th configuration $x_i$, assuming its measured performance is $y_i$, the linear regression model will be:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i,$$

where $\beta_1, \ldots, \beta_p$ are coefficients corresponding to $p$ configuration parameters, and $\epsilon_i$ is the noise (usually assumed to be a Gaussian). The linear model is solved via ridge regression, which is a regularized method to avoid overfitting [10]. Thus GIL addresses CH1 by iteratively estimating the unknown function with linear models.

### 3.2.2 Active Learning

GIL applies active learning to iteratively query samples over successive rounds. At each round, GIL collects samples that are intended to improve the model predictions for the best configurations. After evaluating the new samples, the model is updated using these samples and then the next round begins. As such, the key is to come up with a query function that converges quickly so that GIL can find the optimal configuration across spaces with limited samples. This is not trivial because the model needs to balance the tradeoff between exploration—avoiding local optima in training space—and exploitation—improving the predictions for the highest performance configurations in new space.

Given the base learner is linear, GIL optimizes locally by querying samples with the best predicted performance at each step: for each step $t$, assume the predictive function is $f_t$, then the optimizer searches for a subset $S$ with $k$ configurations such that:

$$S = \arg \max_{S \subseteq S, |S|=k} \sum_{i \in S} f_t(s).$$

This query strategy selects a subset of configurations with the best predicted performance at each round. Intuitively, picking configurations with the best performance is more likely to improve the prediction accuracy for the best performance configurations in the next round. With this query strategy, GIL addresses CH2 by searching the optimal configuration as rapidly as possible. Algorithm 1 summarizes the procedures for GIL performance model.

#### 3.3 GIL+ Performance Model

GIL interprets the relationships between the application-level configuration parameters and ultimate performance by quantifying the learned coefficient for each application-level configuration parameter. To further benefit from interpretation, we propose a hierarchical performance model that relates application-level configuration parameters to performance through low-level system metrics. In particular, we hope to understand not just how each application parameter influences performance, but also how that influence manifests through low-level system metrics like cache misses or context switches.

To accomplish this goal, we present GIL+, illustrated in Figure 1. We consider five low-level system metrics in this paper (see details in Table 3). In principle, GIL+ could be applied to any set of low-level system metrics. We incorporate this extra layer as follows.

Algorithm 2 summarizes the procedures for GIL+. Instead of directly building models mapping from application configurations to performance, GIL+ first builds an model that maps from low-level system metrics to performance. Then, GIL+ obtains the low-level system metric vector $[u_1, \ldots, u_6]$ with the best predicted performance from test low-level system metric set $U_{te}$, as in line 3. Next, for $i$-th low-level system metric $U^i$, $i = 1, \ldots, 5$, GIL+ builds a model that maps from application configurations to $U^i$. Then, GIL+ selects top $k$ application configurations that have the closest predicted $i$-th low-level system metric values to $u_i$, and these $k$ configurations and corresponding measurements will be queried for the next round, as in line 7-8. After all training and test sets are updated, GIL+ trains the model on the final training configuration and performance sets, and uses it to evaluate the remaining test configurations, picking the one with the best predicted performance as the output, as in line 13-14. The intuition of this hierarchical model is to select application configurations that induce the nearest low-level system metric values producing the best predicted performance.

#### 3.4 Interpretation by Visualization

We interpret results from performance models via visualization. We focus on GIL+ because of its extra low-level system metric layer, which makes it possible to investigate the cross-stack interactions between system and application levels. We visualize the learned relationships between (1) performance and low-level system metrics and (2) low-level system metrics and configuration parameters.

We use radar charts to depict the relationships between performance and low-level system metrics and how they change before and after applying GIL+. These charts visualize the differences between the initial, user-provided configurations and the final relationships that GIL+ learns. If users provide configurations that are known to work well on different hardware, then these charts will visualize the differences in the influence of these initial configurations and those GIL+ learned. We choose radar charts because they can show multiple metrics simultaneously and quickly compare different values in each axis. The label for each axis represents a specific low-level system metric and the value on each axis is the learned linear coefficient mapping from low-level system metric to performance. The changes are visualized with different colors.

We use bar charts to illustrate the relationships between low-level system metrics and application configuration parameters. We
choose bar charts because there are multiple low-level system metrics and we need to compare each with their relationships with configuration parameters. Take mweight workload in Figure 3 as an example. The y-axis represents each low-level system metric, and the x-axis represents the learned absolute linear coefficient differences before and after applying Gil+. Each bar represents different configuration parameters, and their direction and length correspond to the x-axis. With bar charts, we can interpret the learned relationship between low-level system metrics and configuration parameters, and how their influences to each low-level system metric change.

3.5 Discussion and Limitations
Gil and Gil+ use linear models as base learners to achieve interpretability, which would lose a slight amount of overall prediction accuracy compared to using more sophisticated black-box learners such as neural networks [12], gradient boosting trees [4], and Gaussian process regression [7, 35]. Moreover, these black-box learners are more likely to overfit the training data and make conservative judgment in exploring new space. To compensate for the lose of prediction accuracy, Gil and Gil+ employ an iterative learning paradigm to allow the model to correct itself every time there is an error [27]. Meanwhile, Bayesian optimization updates the model iteratively using a black-box learner—Gaussian process regression, and usually achieves high prediction accuracy [38], which is also demonstrated in our evaluation in Section 5. While powerful, it is uninterpretable and thus does not provide insights into how the highest performance configuration is reached. To sum up, Gil and Gil+ are practical trade-offs between accuracy and interpretability.

4 EXPERIMENTAL METHODOLOGY
This section describes our experimental setup and evaluation methodologies to demonstrate the effectiveness of Gil and Gil+ at configuration extrapolation in multiple workloads, hardware, and settings.

4.1 Systems
4.1.1 Software. We use Apache Spark 2.2.3 [54] as our software distributed computing framework. Each experiment has a server node and four worker nodes. We choose a wide range of configuration parameters that reflect significant Spark properties categorized by shuffle behavior, data compression and serialization, memory management, execution behavior, networking, and scheduling. Table 1 shows the total 20 parameters in detail. We use similar parameters to prior work [53] but not the same set because Spark has actually reduced the number of user visible configuration parameters—precisely because configuring them is challenging.

4.1.2 Hardware. We run experiments on a public cloud computing system: the Chameleon Cloud Research Platform [21]. We use the names that Chameleon uses for three Intel x86 processors shown in Table 2. We collect five low-level system metrics that have been shown to influence application performance [24]. Table 3 shows the five low-level system metrics in detail.

4.2 Workloads
We select ten Apache Spark workloads from the HiBench big data benchmark suite [14] with details shown in Table 4. These workloads cover various domains including microbenchmarks (micro), machine learning (ML), websearch, and graph analysis; and they exhibit a wide range of resource usage. For ten workloads with 2000 configurations per workload on three hardware platforms, it took four weeks of computing time to collect all these data for experimental evaluation in this paper.

4.3 Points of Comparisons
We compare the following approaches:
• Default: the default application-level configuration provided by the Apache Spark developers.
• RS: randomly sample configurations and select the best with a linear regression performance model.
• NN: a design space exploration method with neural network and intelligent sampling proposed in [16].
• DAC: a configuration parameter tuning approach with the ensemble tree model and genetic algorithm proposed in [53].
• BO: Bayesian optimization for configuration tuning [33].
• Gil: generalizable and interpretable learning relating application configurations and performance.
• Gil+: generalizable and interpretable learning relating application configurations, low-level system metrics, and performance.
• OPT: exhaustive search for the true optimal configuration over the entire measurements.

RS, NN, and DAC are non-iterative learning methods that train only once, while BO, Gil, and Gil+ are iterative methods that train multiple rounds until the sample budget is met. NN, DAC, and BO use black-box models, while RS, Gil, and Gil+ are interpretable. All parameters for these algorithms used in our experiments are selected via cross validations.

4.4 Evaluation Metric
For each workload, we compare the Relative Performance (RP) between the best performance (throughput) found from the learning-based configuration search methods and the true optimal performance (found through exhaustive search over our measurements):

\[
RP = 100\% \times \frac{Y_{\text{pred}} - Y_{\text{opt}}}{Y_{\text{opt}}},
\]

where \(Y_{\text{pred}}\) is the best performance from the above methods and \(Y_{\text{opt}}\) is the optimal performance. Lower RP is better.

4.5 Evaluation Methodology
We randomly generate 2000 application-level configurations for each workload, which is a commonly used size in prior work [53]. For each workload, we run it at each configuration on different hardware to record their performances. We divide the total configurations into three levels: low, modest, and high performance with corresponding ratios [0.5, 0.3, 0.2]; e.g., the high-performance configurations are the top 20% of the measured configurations. The optimal configuration for each workload on each hardware is obtained via exhaustive search over our measurements. We then use

---

3https://spark.apache.org/docs/2.2.3/configuration.html
4https://github.com/Intel-bigdata/HiBench
Table 1: Details of the 20 Apache Spark 2.2.3 configuration parameters.

<table>
<thead>
<tr>
<th>Configuration parameter</th>
<th>Range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>spark.reducer.maxSizeInFlight</td>
<td>24–128</td>
<td>Maximum size of map outputs to fetch simultaneously from each reduce task, in MB.</td>
</tr>
<tr>
<td>spark.shuffle.file.buffer</td>
<td>24–128</td>
<td>Size of the in-memory buffer for each shuffle file output stream, in KB.</td>
</tr>
<tr>
<td>spark.shuffle.sort.bypassMergeThreshold</td>
<td>100–1000</td>
<td>Avoid merge-sorting data if there is no map-side aggregation.</td>
</tr>
<tr>
<td>spark.speculation.interval</td>
<td>100–1000</td>
<td>How often Spark will check for tasks to speculate, in millisecond.</td>
</tr>
<tr>
<td>spark.speculation.multiplier</td>
<td>1–5</td>
<td>How many times slower a task is than the median to be considered for speculation.</td>
</tr>
<tr>
<td>spark.speculation.quantile</td>
<td>0–1</td>
<td>Percentage of tasks which must be complete before speculation is enabled.</td>
</tr>
<tr>
<td>spark.broadcast.blockSize</td>
<td>2–128</td>
<td>Size of each piece of a block for TorrentBroadcastFactory, in MB.</td>
</tr>
<tr>
<td>spark.io.compression.snappy.blockSize</td>
<td>24–128</td>
<td>Block size used in snappy, in KB.</td>
</tr>
<tr>
<td>spark.kryoserializer.buffer.max</td>
<td>24–128</td>
<td>Maximum allowable size of Kryo serialization buffer, in MB.</td>
</tr>
<tr>
<td>spark.kryoserializer.buffer</td>
<td>24–128</td>
<td>Initial size of Kryo's serialization buffer, in KB.</td>
</tr>
<tr>
<td>spark.driver.memory</td>
<td>6–12</td>
<td>Amount of memory to use for the driver process, in GB.</td>
</tr>
<tr>
<td>spark.executor.memory</td>
<td>8–18</td>
<td>Amount of memory to use per executor process, in GB.</td>
</tr>
<tr>
<td>spark.network.timeout</td>
<td>20–500</td>
<td>Default timeout for all network interactions, in second.</td>
</tr>
<tr>
<td>spark.locality.wait</td>
<td>1–10</td>
<td>How long to launch a data-local task before giving up, in second.</td>
</tr>
<tr>
<td>spark.task.maxFailures</td>
<td>1–8</td>
<td>Number of task failures before giving up on the job.</td>
</tr>
<tr>
<td>spark.shuffle.compress</td>
<td>false, true</td>
<td>Whether to compress map output files.</td>
</tr>
<tr>
<td>spark.memory.fraction</td>
<td>0–1</td>
<td>Fraction of (heap space - 300 MB) used for execution and storage.</td>
</tr>
<tr>
<td>spark.shuffle.spill.compress</td>
<td>false, true</td>
<td>Whether to compress data spilled during shuffles.</td>
</tr>
<tr>
<td>spark.broadcast.compress</td>
<td>false, true</td>
<td>Whether to compress broadcast variables before sending them.</td>
</tr>
<tr>
<td>spark.memory.storageFraction</td>
<td>0.5–1</td>
<td>Amount of storage memory immune to eviction.</td>
</tr>
</tbody>
</table>

Table 2: Details of the hardware platforms.

<table>
<thead>
<tr>
<th>Processor</th>
<th>Skylake</th>
<th>Haswell</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gold 6126</td>
<td>E5-2670</td>
<td>E5-2650</td>
<td></td>
</tr>
<tr>
<td>RAM size</td>
<td>192 GB</td>
<td>128 GB</td>
<td>64 GB</td>
</tr>
<tr>
<td># of Threads</td>
<td>48</td>
<td>48</td>
<td>40</td>
</tr>
<tr>
<td>Clockspeed</td>
<td>2.6 GHz</td>
<td>3.1 GHz</td>
<td>3.0 GHz</td>
</tr>
<tr>
<td>L3 cache</td>
<td>19.25 MB</td>
<td>30 MB</td>
<td>25 MB</td>
</tr>
<tr>
<td>Memory speed</td>
<td>2.666 GHz</td>
<td>2.133 GHz</td>
<td>2.133 GHz</td>
</tr>
<tr>
<td># Mem channels</td>
<td>6</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Network speed</td>
<td>10 GbE</td>
<td>0.1 GbE</td>
<td>10 GbE</td>
</tr>
<tr>
<td>Disk vendor</td>
<td>Samsung</td>
<td>Seagate</td>
<td>Seagate</td>
</tr>
<tr>
<td># Disks</td>
<td>1</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>Disk bandwidth</td>
<td>6 Gb/s</td>
<td>6 Gb/s</td>
<td>24 Gb/s</td>
</tr>
</tbody>
</table>

Table 3: Details of the five low-level system metrics.

<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Low-level metrics</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMR</td>
<td>Branch misses rate</td>
<td># branch misses / # total branch misses</td>
</tr>
<tr>
<td>CMR</td>
<td>Cache misses rate</td>
<td># cache misses / # total cache misses</td>
</tr>
<tr>
<td>CSR</td>
<td>Context switch rate</td>
<td># context switch / # cpu cycles</td>
</tr>
<tr>
<td>PFR</td>
<td>Page faults rate</td>
<td># page faults / # cpu cycles</td>
</tr>
<tr>
<td>IPC</td>
<td>Instruc. per cycle</td>
<td># instruction / # cpu cycles</td>
</tr>
</tbody>
</table>

Table 4: Details of the ten HiBench workloads.

<table>
<thead>
<tr>
<th>Workload</th>
<th>Data size</th>
<th>Workload</th>
<th>Data size</th>
</tr>
</thead>
<tbody>
<tr>
<td>wordcount</td>
<td>32 GB</td>
<td>lr</td>
<td>8 GB</td>
</tr>
<tr>
<td>terasort</td>
<td>3.2 GB</td>
<td>linear</td>
<td>48 GB</td>
</tr>
<tr>
<td>als</td>
<td>0.6 GB</td>
<td>rf</td>
<td>0.8 GB</td>
</tr>
<tr>
<td>bayes</td>
<td>19 GB</td>
<td>pagerank</td>
<td>1.5 GB</td>
</tr>
<tr>
<td>kmeans</td>
<td>20 GB</td>
<td>nweight</td>
<td>0.9 GB</td>
</tr>
</tbody>
</table>

the methods mentioned above to search for the best configuration for each workload on different hardware. Since our goal is to examine the extrapolation ability of each algorithm, we consider the following two experimental settings:

- **low2high**: training set only has configurations of low performance, and test set has configurations with a wide range of performance. This setting corresponds to scenario 1 in §1; i.e., this represents the case where users start from configurations of low performance and aim to extrapolate to high performance .
- **mod2high**: training set only has configurations of modest performance that run fast from a different hardware, and test set has configurations with a wide range of performance. For instance, if the training set has configurations that run fast on Haswell, but run modestly on Skylake, the search phase will evaluate all configurations in test set to pick the one predicted to run fastest on Skylake. Since we have 3 hardware, we have 6 training-test pairs. This setting corresponds to scenario 2 in §1; i.e., this represents the case where users have found configurations of high performance for one hardware and attempt to use those as a starting point to optimize for a different hardware.

In Algorithm 1 and 2, \( T \) and \( k \) are set as small numbers to demonstrate that Gil and Gil+ can produce comparable and better results on a small sample budget than those methods without intelligent sampling (e.g., RS). Since the collected dataset size for each hardware is 2000, we choose 20% and 10% for low2high and mod2high—i.e., 400 and 200—which are much smaller than thousands used in prior work [53]. low2high requires more labeled data for training because it extrapolates larger space than mod2high does. \( k \) is set as 20. We report results averaged over 5 runs with different seeds.
5 EXPERIMENTAL EVALUATION

We examine the following research questions (RQs):

- **RQ1**: How good are Gil and Gil+ at extrapolating from configurations of low performance to high performance (low2high)?
- **RQ2**: How good are Gil and Gil+ at extrapolating from configurations of modest performance to high performance (mod2high)?
- **RQ3**: Does incorporating prior knowledge improve extrapolation performance compared to starting from scratch?
- **RQ4**: What can we interpret from visualization in case studies?

5.1 RQ1: How good are Gil and Gil+ at extrapolation from configurations of low performance to high performance?

We examine the extrapolation results from configurations of low performance to high performance in low2high setting. Figure 5 shows the relative performance (RP) results for each method. The strip label on the right for each row of bar charts represents the hardware platform for the experiments. The x-axis represents each workload, and the y-axis represents RP. Lower is better (closer to optimal). The last column, HarMean, shows the harmonic mean results over all workloads, which are also quantified in Table 5.

Table 5: Harmonic mean results over all workloads of each hardware for low2high. The last column HarMean is the harmonic mean over three hardware.

<table>
<thead>
<tr>
<th></th>
<th>Skylake</th>
<th>Haswell</th>
<th>Storage</th>
<th>HarMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFAULT</td>
<td>43%</td>
<td>29%</td>
<td>25%</td>
<td>31%</td>
</tr>
<tr>
<td>RS</td>
<td>22%</td>
<td>11%</td>
<td>11%</td>
<td>13%</td>
</tr>
<tr>
<td>NN</td>
<td>22%</td>
<td>17%</td>
<td>13%</td>
<td>16%</td>
</tr>
<tr>
<td>DAC</td>
<td>36%</td>
<td>18%</td>
<td>19%</td>
<td>22%</td>
</tr>
<tr>
<td>BO</td>
<td>14%</td>
<td>1%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Gil</td>
<td>13%</td>
<td>7%</td>
<td>7%</td>
<td>8%</td>
</tr>
<tr>
<td>Gil+</td>
<td>10%</td>
<td>7%</td>
<td>5%</td>
<td>7%</td>
</tr>
</tbody>
</table>

For non-iterative learning methods, RS is generally better than NN and DAC, which is a bit counter-intuitive as we would think that linear models are not as powerful as black-box models. However, as discussed in §3.2.1, despite perhaps overall lower prediction accuracy, linear models are less likely to overfit and thus promote extrapolation. In contrast, black-box models such as neural networks and ensemble trees are more likely to overfit and make less accurate predictions for unseen data. For iterative learning methods, BO is slightly (5 percentage points) better than Gil and Gil+ because black-box models generally have higher prediction accuracy than linear models. Overall, iterative learning methods are better than non-iterative ones because they encourage extrapolation in the iterative process. These results demonstrate that Gil and Gil+’s iterative learning can provide performance comparable to the best black-box methods and overcome faulty (or extremely low-performance) configurations.

5.2 RQ2: How good are Gil and Gil+ at extrapolating from configurations of modest performance to high performance?

We examine the extrapolation results from configurations of modest performance to high performance in mod2high setting. Figure 6, 7, and 8 show the relative performance (RP) for each method on three target hardware platforms, respectively. In each figure, the strip label on the right for each row of bar charts represents the hardware platform where the starting configurations are from. The x-axis represents each workload, and the y-axis represents RP for the target hardware. Lower is better (closer to optimal). The last column, HarMean, shows the harmonic mean results over all workloads, which are also quantified in Table 6.

Table 6: Harmonic mean results over all workloads of each hardware for mod2high. The last column HarMean is the harmonic mean over three hardware.

<table>
<thead>
<tr>
<th></th>
<th>Skylake</th>
<th>Haswell</th>
<th>Storage</th>
<th>HarMean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEFAULT</td>
<td>43%</td>
<td>29%</td>
<td>25%</td>
<td>31%</td>
</tr>
<tr>
<td>RS</td>
<td>17%</td>
<td>10%</td>
<td>9%</td>
<td>11%</td>
</tr>
<tr>
<td>NN</td>
<td>19%</td>
<td>13%</td>
<td>9%</td>
<td>13%</td>
</tr>
<tr>
<td>DAC</td>
<td>22%</td>
<td>14%</td>
<td>10%</td>
<td>14%</td>
</tr>
<tr>
<td>BO</td>
<td>7%</td>
<td>1%</td>
<td>2%</td>
<td>2%</td>
</tr>
<tr>
<td>Gil</td>
<td>2%</td>
<td>2%</td>
<td>6%</td>
<td>2%</td>
</tr>
<tr>
<td>Gil+</td>
<td>7%</td>
<td>6%</td>
<td>3%</td>
<td>4%</td>
</tr>
</tbody>
</table>

The results analyzed in the low2high setting almost hold for the mod2high setting. RS is better than NN and DAC, and BO is almost comparable to Gil and Gil+. Overall, BO, Gil, and Gil+ are better than RS, NN, and DAC due to their iterative learning paradigm. These results demonstrate that Gil and Gil+ are comparable to the best black-box learners. Different from prior methods, Gil and Gil+ can benefit from starting from user suggested configurations, which is evidence that incorporating human knowledge is beneficial—it gets Gil and Gil+ on par with the best black-box learner.

5.3 RQ3: Does incorporating prior knowledge improve extrapolation performance compared to starting from scratch?

As further evidence of the benefits of extrapolating from prior knowledge, we compare starting from scratch with random sampling to incorporating prior knowledge for initial training. Figure 9 shows the aggregated improvements from incorporating prior knowledge over starting with random sampling across all workloads in the mod2high setting (higher is better). Overall, all learners benefit from incorporating prior knowledge with improvements from 2% to 15% in harmonic mean, where Gil and Gil+ both have 3% improvement. Starting from scratch achieves lower performance because random sampling over the whole space loses initial direction for extrapolation, while a tighter cluster of samples provides a clearer direction to pursue next, and reduces the chance of getting stuck in local optima. These results demonstrate the efficacy of Gil and Gil+ by incorporating prior knowledge for initial training.
5.4 RQ4: What can we interpret from visualization in case studies?

To demonstrate the scope of our visualization tools for interpretation, we use two case studies in different applications: cause interpretation and deployment prediction.

5.4.1 Cause Interpretation. We examine the Gil⁺’s ability to interpret the causes for low performance by visualizing the coefficients of the learned relationships mapping from low-level system metrics to performance, and mapping from application configurations to low-level system metrics. Figure 10 shows two workloads running on Skylake and how their coefficients change from low performance.
to high performance. The common thing observed in two radar charts is that both workloads need higher BMR and CMR for higher performance. This is counter-intuitive because higher BMR and CMR conventionally lead to lower performance. To interpret the radar charts correctly, we need to read every axis together. It is not that increasing BMR and CMR increases performance; rather increasing BMR and CMR together, while reducing CSR is key to achieving better performance.

Figure 11 shows the corresponding bar charts for the absolute learned coefficient differences from application configuration parameters on each low-level system metric. We can see that the ways to get similar system-level metrics changes are different for different workloads via different application-level configuration parameter changes. For linear, the influence magnitudes of BMR and CMR from configuration parameters do not change much. For lr, the largest influence changes of BMR and CMR are from spark.memory.fraction. These results demonstrate Gil+’s ability to interpret causes for low performance by discovering the relationships between application configurations, low-level system metrics, and performance.

Figure 10: Visualizing learned relationships between low-level systems metrics and performance. The value on each axis shows how the corresponding low-level system metric influences performance when moving from low performance (Before) to high performance (After) on Skylake.

5.4.2 Deployment Prediction. We show that the learned coefficients of the model mapping from low-level system metrics to performance can predict the deployment benefits across hardware. Figure 12 shows two radar charts, where for each workload, configurations running fast on Haswell and their performance on Storage are used for initial training to obtain the before coefficients. These coefficients represent how fast configurations found on Haswell behave when they are on Storage. Then, we apply Gil+ to iteratively extrapolate the configurations that can run fast on Storage until the sample budget is met. The after coefficients represent how the best configurations found on Storage behave.

Figure 12a and 12b show that the two workloads als and kmeans both have sharp spikes away from CSR, meaning that they both benefit from decreased influence from CSR. Thus, we predict that both workloads have same best hardware deployment. This prediction is validated by the true data that the hardware choice for the best performance of both workloads is Storage over Haswell. This also tells users that there is no need to tune both workloads separately on both hardware to find which hardware will be a better deployment because it is likely that they will have the same best hardware deployment due to their similar low-level system
Figure 12: Visualizing learned relationships between low-level systems metrics and performance. The value on each axis shows how the corresponding low-level system metric influences performance when moving from Haswell (Before) to Storage (After).

6 RELATED WORK

ML for configuration management. ML techniques confront the complexity of software configuration management [13, 37, 49]. ML is generally incorporated into configuration tuning via performance modeling [20, 40]; i.e., the process of learning a function mapping application-level configuration parameters to quantifiable behavior (e.g., latency, throughput, energy consumption) [3, 8, 15, 30, 31, 42, 55]. For instance, ConEx uses evolutionary Markov Chain Monte Carlo sampling to find high-performance configurations for big data systems [23]. OtterTune uses Gaussian processes to tune database configurations [42]. OpenTuner uses an ensemble of genetic algorithms to tune configuration parameters for computer programs [2]. More recently, there has been a growing interest in tuning ML systems themselves using other black-box optimization techniques [11, 26, 46].

Transfer learning. Transfer learning is a common approach to incorporate prior knowledge, which gains knowledge from one problem and applies it to a different but related problem [34]. For configurable software systems, the knowledge can be transferred across different workloads and hardware environments to reduce data collection efforts [17]. For instance, a cost-aware transfer learning method uses Gaussian process to transfer knowledge from simulators to real systems [19]. L2S uses active learning to select samples for knowledge transfer to a new domain [18]. BEETLE uses a discovery step to identify the most relevant source from multiple sources of data [22]. However, these works are fundamentally different from our proposal in that they use black-box models, in which the knowledge transferred is not interpretable by users [36]. In contrast, Gil and Gil+ not only incorporate prior knowledge to improve application performance, but produce interpretable results to help expand that knowledge.

Interpretability. There has been an emerging line of work on interpreting software systems. Complex proposes a white-box performance model based on a data-flow analysis [44], which does not incorporate prior knowledge or use learning. Valov et al. use linear models to transfer knowledge across different hardware environments [41], but they directly use the model learned from the source on the target software applications, and thus do not extrapolate to improve results. Our proposed methods use linear models to provide users an interpretation of the interactions between application-level configurations, low-level system metrics, and performance [5]. Our proposed tools Gil and Gil+ combine the advantages of transfer learning and interpretability to improve configuration extrapolation results and give feedback to users.

7 CONCLUSION

This paper introduces two configuration extrapolation tools, Gil and Gil+, that incorporate prior knowledge and produce interpretable results. These tools also produce a graphical interpretation of how they arrived at the highest performance application configuration. Our results show that Gil and Gil+ achieve application performance comparable to the best black-box learner, and outperform those starting from scratch with random sampling for initial training, which demonstrates the benefits of incorporating prior knowledge into the learning process. Additionally, the graphical visualization tools enable users to interact with the learners and interpret relationships between application-level configurations, low-level system metrics, and performance. We hope this work can inspire software engineering researchers to consider prior knowledge and model interpretability when applying machine learning techniques to performance modeling.

ACKNOWLEDGMENTS

We are grateful to Alex Renda who read the early draft of this work and provided extremely valuable feedback. We thank the anonymous reviewers for their helpful feedback to improve the final paper. Yi Ding’s work is supported by the National Science Foundation under Grant 2030859 to the Computing Research Association for the CIFellows Project. Ahsan Pervaiz and Henry Hoffmann’s work is supported by NSF (grants CCF-2028427, CNS-1956180, CCF-1837120, CNS-1764039), ARO (grant W911NF1920321), and a DOE Early Career Award (grant DESC0014195 0003). Michael Carbin’s work is supported in part by the National Science Foundation (NSF CCF-1918839). Results presented in this paper were obtained using the Chameleon testbed supported by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES


