Kisaco Leadership Chart on AI software optimization solutions 2021: SigOpt, an Intel company

Publication Date: 25 Feb 2021. Product code KR380.4

Michael Azoff

Kisaco Research View

Motivation

The current revolution or dramatic evolution in artificial intelligence (AI) we are witnessing was sparked by the arrival of hardware accelerators onto which deep learning neural networks were ported: training times that took months ran in days or hours on Nvidia GPUs. This gave rise to the explosion in AI hardware accelerator chips competing to take a share of the large and still growing accelerator market. Now a new form of optimization, that encompasses a host of features beyond and inclusive of acceleration, has appeared in the AI market, purely software based: meaning that they operate at the software level in the machine learning (ML) technology stack. Many of the AI software optimization (AISO) products have emerged from relatively recent startups. These products can optimize ML models that run on just central processing units (CPUs) or enhance performance on standard AI accelerators: graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and digital signal processors (DSPs). AISO products also compete with the newer breed of AI chips (which we label as domain specific architectures, DSAs), making the whole AI field a lot more nuanced and competitive. For both enterprise users and product manufacturers there are now wider options in choosing the best combination of software and hardware for their AI applications and products requirements.

In this report we feature the leading players in the AISO market, compared side by side in our Kisaco Leadership Chart (KLC). We explain what this technology has to offer, reveal our analysis of the top players, and profile in-depth SigOpt, an Intel company.

Key findings

- The AISO product market is distinct from the AI hardware accelerator market: users working with AISO may choose different hardware accelerators to work with than if they had not used AISO. AISO creates new degrees of freedom.
- Working with a particular set of hardware choices, compression resulting from AISO creates a smaller AI component footprint, which may entail manufacturers having space to add more functionality/chips to a product.
- AISO products in the market operate at various levels in the ML technology stack. This means it is possible to combine AISO operations from different products in the same model.
- Using AISO can make the difference to an AI model achieving its specifications for the target application. For example, automotive applications require latency within strict tolerances, AISO can be optimized for latency reduction and make an AI model suitable for near real time response.
- Enterprises ramping up their ML applications at scale need to manage the ML lifecycle (MLOps); lack of such management, is the cause of failure to execute and deliver value. We see a good overlap between ML lifecycle management and AISO products.
- We expect to hear of more deals and partnerships formed between AISO vendors and hardware manufacturers, especially producers of off-the-shelf AI hardware accelerators: CPU, GPU, and FPGA.

AI software optimization overview

Defining AISO

For the purposes of this report, an AISO product is one that can perform at least one of the following operations on a software AI model:

- Compress the model.
- Reduce latency in the model (i.e., accelerate operation).
- Reduce power consumption by the model.
- Increase the model throughput.
- Reduce the cost of working with the model.

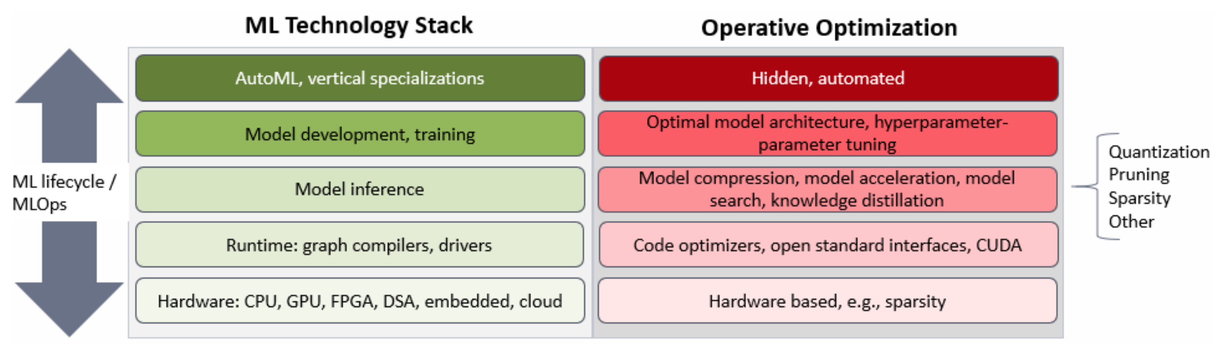
An AISO product typically achieves all of the above and does so without reducing the accuracy of the model. Thus, for inference mode AISO products, they start with a trained model that has achieved a desired accuracy, this accuracy is maintained by the AISO product as it performs optimization. However, some of the AISO products may increase accuracy. And whether the accuracy is increased or maintained, some AISO products allow this accuracy level to be reduced (typically only slightly) in order to gain a significant boost in another dimension of the optimization (e.g., reduce latency further).

AISO products may operate during training only, during inference only, or some will operate during either mode.

AISO products operate across multiple layers of the ML stack

In this section we work our way through the ML technology stack, starting at the bottom hardware layer through to the development/training level, see Figure 1, and at each layer we identify the key optimization practices that can be performed.

Figure 1: The ML technology stack and the optimizations possible at each level


ML Technology Stack
Operative Optimization
ML lifecycle / MLOps
AutoML, vertical specializations
Hidden, automated
Model development, training
Optimal model architecture, hyperparameter-parameter tuning
Model inference
Model compression, model acceleration, model search, knowledge distillation
Quantization
Pruning
Sparsity
Other
Runtime: graph compilers, drivers
Code optimizers, open standard interfaces, CUDA
Hardware: CPU, GPU, FPGA, DSA, embedded, cloud
Hardware based, e.g., sparsity

Source: Kisaco Research

Hardware layer

The AISO product marketplace does not operate at the hardware level itself, as AISO products operate solely at the software level. However, we start our coverage of optimization with the hardware level as the hardware manufacturers have introduced features designed to support some optimization operations. The hardware manufacturers are limited in what they can do as they receive a model as a given and cannot change it, nevertheless they are able to add features that will also work in tandem with the AISO operations that are performed higher up in the ML stack.

One way the AI hardware accelerator manufacturers have been fine tuning ML workloads is through modifying bit precision. Some of these techniques originated in high-performance computing and have found relevance for ML. The higher the bit precision the higher the range of numbers that can be represented. This is generally desirable during training of ML models, however, high bit precision equates with more computational resources used (due to working with longer bit word lengths), more data transfer, more memory storage, and consequently increases latency, power consumption, and costs. Hardware manufacturers have introduced multi-precision computing into their devices, whereby a device can run an application at double, single, or half precision for different parts of an application. They also introduced mixed precision, i.e., using different precisions within one computational operation. For example, the heavy computational burden of multiplying two matrices together is reduced using single precision during multiply but the result is stored in double precision.

Runtime layer

The next layer up in the stack is the runtime where graph compilers and interfaces exist. Standard interfaces are available: Nvidia has an API called Compute Unified Device Architecture (CUDA) which makes it easy to port applications onto Nvidia GPUs, Intel has Open Visual Inference and Neural network Optimization (OpenVINO) for running on Intel devices, and Xilinx has the Vitis platform for its FPGAs. All these open interfaces will connect with popular ML frameworks and development environments like TensorFlow and PyTorch, and they are also all proprietary, which means they only operate on their respective hardware.

Apache TVM operates at the runtime level and is an open source project that describes itself as “a ML compiler framework for CPUs, GPUs, and machine learning accelerators. It aims to enable machine learning engineers to optimize and run computations efficiently on any hardware backend”. TVM uses AI to optimize nested loops in code that arise in deep learning models and changes them to run optimally on the target hardware.

Also of interest are open standards such Open Computing Language (OpenCL) and Sycl from the Khronos Group, an open industry consortium. OpenCL is an open royalty-free, standards-based framework for writing cross-platform programs: write once and execute across diverse accelerators and heterogeneous platforms consisting of CPUs, GPUs, DSPs, FPGAs and other processors or hardware accelerators. Sycl is also royalty-free and is an abstraction layer siting above OpenCL, it is suitable for software developers who want to take their CUDA code and migrate it to open standards and be able to run on lots of different hardware.

Model inference

The inference stage is where a model has been trained to the required accuracy and can be used for inferencing. It is also an opportunity for applying several optimization techniques. Depending on the application, these optimizations may be essential in order to meet one or more criteria:

- Reduce size of model to fit smaller footprint edge applications.
- Reduce the latency to meet near real-time responses.
- Reduce the power consumption to within available energy sources in operating environment.
- Implement any of above to reduce cost to within product budget.

In addition, optimization can increase the performance, for example in image processing increasing the frames per sec (FPS) analyzed.

There are optimization techniques emerging continually from the academic research community, and three key ones are summarized below, however they require expertise to implement (typically at neural networks PhD level):

- **Pruning**: There are many schemes for removing connections (synapses) in neural networks. A simple scheme is to use an absolute threshold value and any weight (taking its absolute value) below the threshold is pruned.
- **Quantization**: After training weights and biases in the network are represented typically by bit precisions of length 32 (single precision) or 64 (double precision). We discussed above quantization in the hardware layer of these numbers to lower precision. This can be performed at this layer in the ML stack. Reducing the bit precision reduces the number of different weights that can be used.
- **Knowledge distillation**: In this approach the trained model is designated as teacher and smaller student models are introduced, trained against a combination of the true labels and the outputs of the teacher (for the same given input). This method distills the teacher knowledge into a smaller student model.

Model development and training

This is the top level where optimization under the control of a user may be performed. Automated deep learning hyperparameter optimization is a technique that takes a huge amount of labor out of fine-tuning hyperparameter values. This is where SigOpt, an Intel company, operates.

AutoML and vertical specific applications

We mention this as there are products on the market that perform complete model development in an easy-to-use interface designed for non-experts in ML. These solutions typical perform optimization internally and any such operations are invisible to the user, although they may appear as user interface options such minimize latency or maximize throughput.

Solution analysis: vendor comparisons

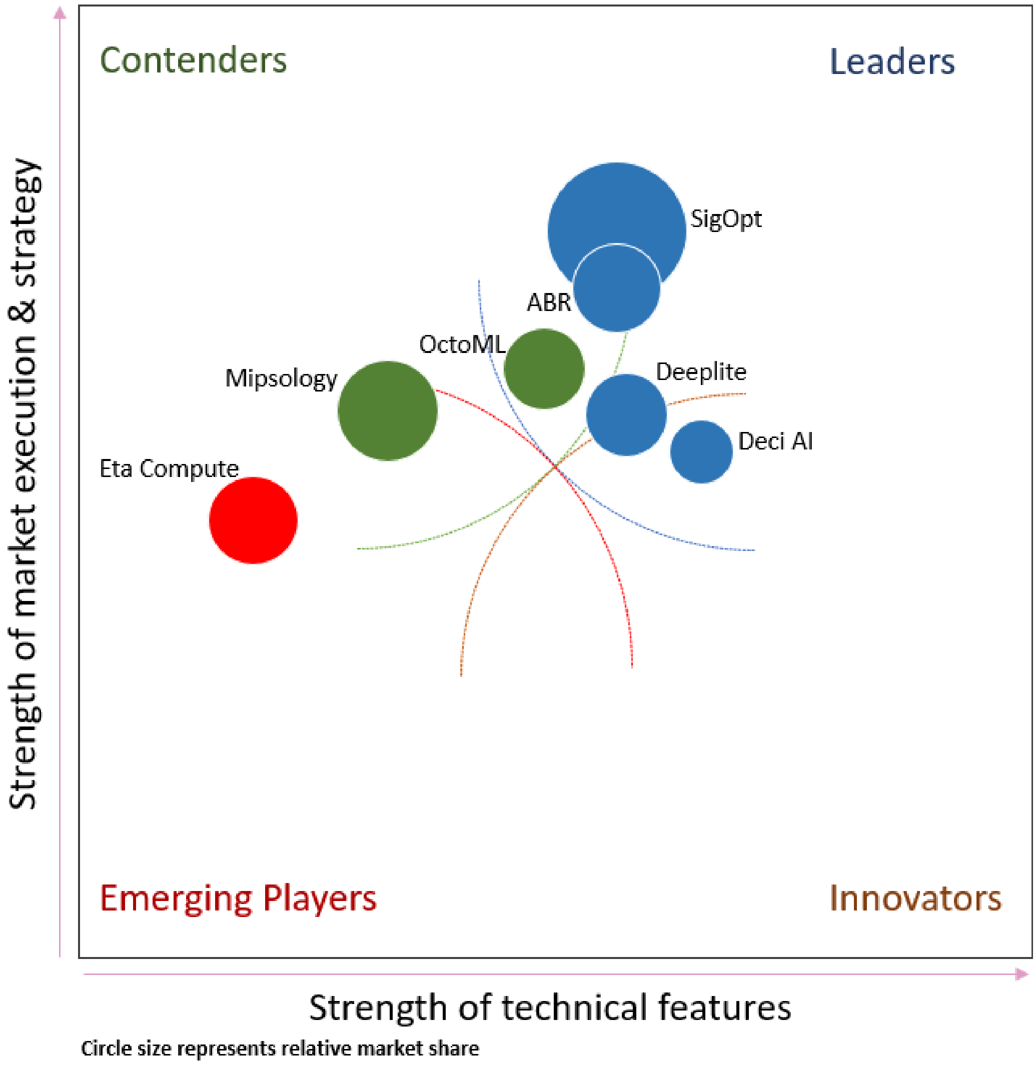
Kisaco Leadership Chart on AI software optimization solutions 2021

The KLC chart for AI software optimization solutions

For the KLC, Figure 2, the scoring was split across three dimensions: the x-axis technical features assessment is the total of all the various technical AISO features, with scores weighed and aggregated. The KLC y-axis score is the strength of market execution and strategy which comprises a number of business reach and customer engagement assessments. The third dimension is the plot circle size representative of the market revenue, we normalize (largest circle) to the highest earning vendor. We have had to use best estimates where vendors were not able to share details with us.

Our assessments have ranked our participating vendors as shown in Figure 3. We ranked SigOpt, an Intel company: as leader. SigOpt is a prominent vendor in the ML community for its automated deep learning hyperparameter optimization during training. The company has continued to innovate and grow its business and was most recently acquired by Intel.

Figure 2: Kisaco Leadership Chart on AI software optimization solutions 2021



Source: Kisaco Research. Circle size is representative of market share.

Figure 3: Kisaco Leadership Chart on AI software optimization 2021: ranking of vendors

Leader	Contender	Emerging Player
ABR	Mipsology	Eta Compute
Deci AI	OctoML	
Deeplite		
SigOpt, an Intel co.		

Source: Kisaco Research

Vendor analysis

SigOpt, an Intel Company, Kisaco evaluation: Leader

Product: SigOpt Model Experimentation & Optimization Platform, available as SaaS and on-premises.

Intel was launched in 1968 by founders Gordon Moore and Robert Noyce. Based in Silicon Valley, it will be led by new CEO Pat Gelsinger beginning 15 Feb 2021. Intel acquired SigOpt in Oct 2020. SigOpt was a private company founded in 2014 by Patrick Hayes, Scott Clark, and is based in San Francisco. SigOpt's history goes back to origins at Cornell University where Scott Clark worked on parallelizing Bayesian optimization, which led to work on hyperparameter optimization in deep neural networks. Scott then joined Yelp where he created the open source tool MOE (Metric Optimization Engine), which is an efficient way to optimize the parameters in numerical models of engineering systems. SigOpt was created to take MOE into the enterprise and grow the tool and its business to its full potential, gaining investment support from Y Combinator and Andreessen Horowitz before it was acquired by Intel.

Today, SigOpt has over 1000 academic users, over 30 papers published in leading AI events, and has 10 patents approved and over 20 pending. Its enterprise customers span seven industry verticals, such as AI based financial trading – with customers in this space who manage over $600b in assets. Others include business consulting, media & entertainment, government, enterprise technology, insurance, payments, and banking. Most of the revenue is product license based but there is also some professional services in the form of a 'white glove' services that support customers in areas ranging from applied optimization research to AI product integration, all of which are geared toward helping them make best use of its product. The largest customer footprint is in N. America but it has customers globally.

SigOpt's expertise is in algorithm and platform engineering, taking operational tasks that are time intensive for machine learning (ML) and making them easy and seamless, using AI to assist in this task. Hyperparameter optimization is a key example. The aim is to help AI developers by enabling them to create models that run faster and perform better, by creating better configurations of model parameters. The company also focuses on helping its customers with real-world ML applications, ensuring they make an impact in production. SigOpt recognizes enterprise AI as falling across a spectrum of use cases, from one end of commodity models comprising relatively straightforward implementations of usually existing or tweaked-existing ML models, to the other end of differentiated models where the goal is to unlock hard problems, and where the best ML model can make a significant difference to the business. SigOpt is applied to all of these modeling problems but drives the highest return on investment for its customers when it is used to unlock hard problems, for which teams can use a wide range of advanced experimentation features that are available out of the box in SigOpt's platform.

An example of differentiated models in the market is SigOpt customer Two Sigma Investments, a hedge fund with assets under management grown significantly and using ML as part of its trading technology. It has made the case for its benefits that include model performance gains from its algorithmic capabilities, new capacity to unlock entirely new business problems with its advanced experimentation features, and efficient utilization of compute with asynchronous parallelization of jobs. SigOpt also notices the growth in the role of AI specialists is faster than the annual growth of data scientists, and this translates to an increase in differentiated AI models within enterprises. Critically, developing these impactful AI models requires a different set of software tooling that has these types of characteristics:

- API enabled
- ML framework agnostic
- Integration with any stack
- Reliable at scale
- Designed to augment experts
- Algorithmically differentiated
- Designed to be best-in-class
- It is designed to fit into existing workflow

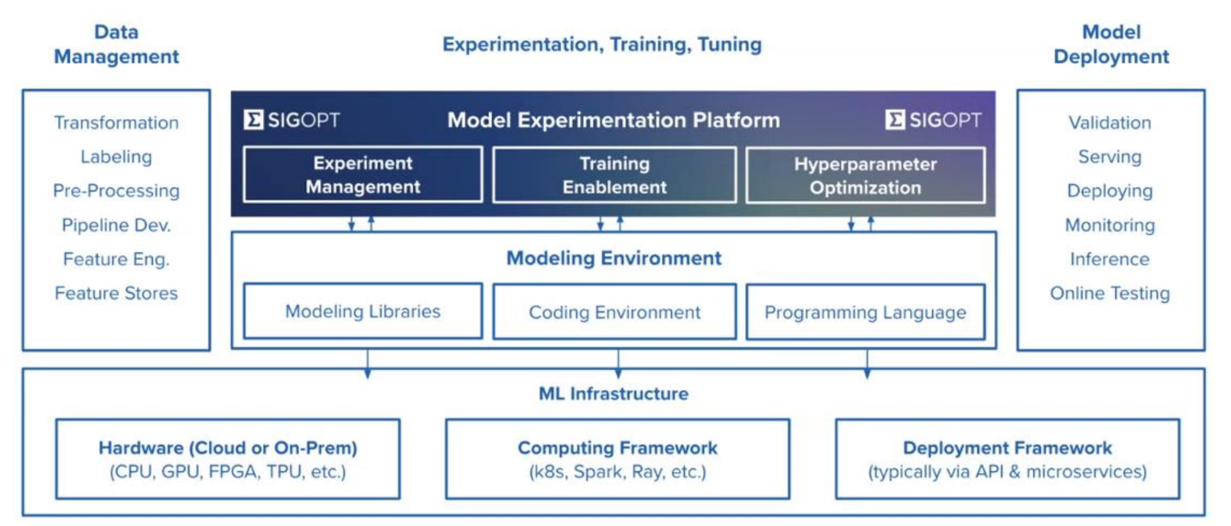
Designing this type of next-generation AI software is the big opportunity that SigOpt sees in the market and has dedicated its efforts to delivering for its customers. Considering the spectrum of enterprise AI use cases, SigOpt is focused on helping enterprises succeed with their differentiated models. SigOpt is working across every industry sector and in each of these there are the enterprises that seek to push AI to the limit and be best in class, and these are the customers SigOpt is focused on. SigOpt customers span seven industries and customers here prefer to build rather than buy their AI solutions, including one of the largest streaming services in the media space. The large global technology consulting firms also have a role to play in supporting these types of customers and their needs and are investing to provide such capability by working with SigOpt as part of their engagements.

The key functionalities offered by SigOpt have evolved, to the following:

- **Manage**: Track, analyze, reproduce modeling with just a few lines of code. Insights and automation boost ML development productivity and help modelers see the bigger picture with metric, architecture, and model comparisons.
- **Optimize**: Automate parameter and hyperparameter tuning. Patented algorithms discover higher performance models much more efficiently in fewer training runs than other methods.
- **Scale**: Seamlessly distribute and run jobs in parallel. The API is designed to work with any stack, keeps data safe, and scales to millions of calls/hour.

Figure 4 shows how SigOpt fits into the ML technology stack. It is used primarily for training ML models in development, where optimization is performed in offline training. SigOpt is agnostic to where it is used, on the cloud or on-premises, it is also agnostic in choice of ML development framework, with the major tools supported, and can also be used with a range of development languages. Integration is seamless into any workflow, including containers and Kubernetes.

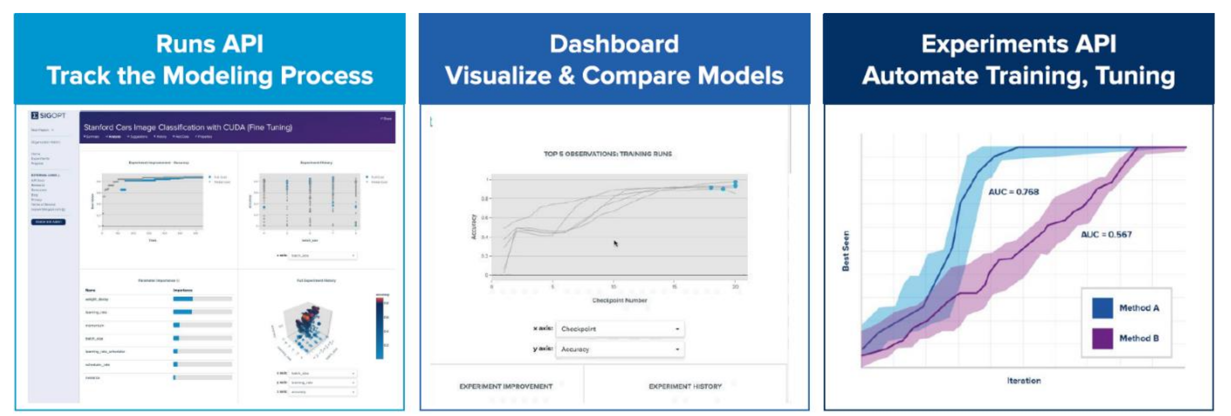
Figure 4: Where SigOpt sits in the ML technology stack



Source: SigOpt

SigOpt has made it easier to visualize and compare models and track the modeling process – See Figure 5. There is a separate API for instrumenting and logging of any attribute of the model as it is progressed through the workflow, most typically used in training runs. It can track up to 50 different metrics and the analysis is designed to be shared across a team for collaboration. Wherever SigOpt runs the data always stays at source and remains private.

Figure 5: SigOpt features for management of the training process



Source: SigOpt

Customers using SigOpt benefit from productivity gains, faster time to train, and efficient utilization of compute. In combination, this empowers modelers to bring more models into production on a faster timeline, boosting the overall return of AI for enterprises. SigOpt customers also report that SigOpt also empowers their modelers to solve entirely new business problems, and, in the process, understand their modeling problem space more deeply. SigOpt cites a customer, a large global technology consulting firm (with 3000 ML developers), that had ML groups working in siloes and suffering from a lack of standardization. SigOpt was brought in and helped automate architecture search and model optimization, which resulted in 30% productivity gain per employee and yielded nearly $50m in projected value across the firm. Finally, there are a range of case studies from SigOpt that point to its vast efficiency gains when compared to other methods like grid search and random search.

Further reading: “www.twosigma.com/articles/why-two-sigma-is-using-sigopt-for-automated-parameter-tuning”.

Kisaco Assessment

Strengths

- SigOpt, an Intel Company, achieved an impressive score across all dimensions of our KLC evaluation to be ranked as Leader. SigOpt has the largest market penetration, having been in the market for longer than its recently emerging rivals, and has significant strength in its technical offering. The acquisition by Intel offers three clear opportunities for SigOpt: 1) Intel resources can help SigOpt achieve its goals, 2) the opportunity to work deeper with Intel hardware to create new optimization features across a broad range of end user price points, and 3) working with the Intel customer base to further evolve the use cases exposed to SigOpt.
- SigOpt, which was acquired by Intel in October 2020, is a pioneer in sample-efficient, Bayesian-based optimization and this continues to be its forte in hyperparameter optimization of ML models in training. It has added functionality to offer a complete model experimentation platform, covering experiment management and training enablement. Seamless integration with any technology stack makes it easy to work with SigOpt.
- SigOpt’s go to market strategy is the most mature and strongest that we have seen in this market. It has identified the needs of global enterprises that are building world beating AI solutions and is helping them achieve their ambitions. It also has good presence within the academic research community, and this creates opportunities when the researchers move into the corporate space.

Weaknesses

- SigOpt focuses on ML training optimization but not inference optimization. This may be something that SigOpt might want to expand into. For example, a manufacturer of edge AI products may receive a trained model but then have to fit it into a device with limited power, memory, and accelerator options. Being able to optimize the model at this stage is a clear benefit.
- We find having flexibility in ingesting data from multiple sources a benefit. SigOpt relies on other tools to do this task but this could be an area worth expanding into.
- Some of the standard techniques in optimization (pruning, quantization, compact convolutional filters) may yield additional benefit and it would be useful to have these out-of-the-box in SigOpt.

Appendix

Vendor solution selection

Inclusion criteria

In general, the KLC is not designed to exhaustively cover all the players in a market but a representative set of the leading players. Kisaco also invites smaller, possibly niche vendors that have innovative solutions and are on a fast growth path. With this flexibility we consider each participant on its merits as a good fit to the KLC topic.

The criteria for inclusion of a vendor product in this report are as follows:

- Vendor has an offering fitting the topic of AISO.
- There are two categories of vendor that are considered for inclusion in this evaluation:
    - Vendor has significant market share relative to peers and is either a recognized leader in the market or has the potential to become one.
    - The vendor is a niche player or an emerging player with outstanding market leading technology.

Exclusion criteria

We exclude products that are not ready for the market and have no customers.

Methodology

- Vendors complete a comprehensive capability questionnaire in a spreadsheet, covering the three dimensions of the KLC. The resultant matrix of responses is appropriately weighted and scored, and these scores are plotted to produce the KLC.
- We hold comprehensive briefings with all participating vendors, including product demonstration.
- Supplemental information is obtained from vendor literature and publicly available information.

Definition of the KLC

The KLC spans three assessment dimensions.

Technical Features

Kisaco Research has developed a series of features and functionality that provide technology differentiation between the leading solutions in the marketplace.

Market execution and strategy

Kisaco Research reviews the capability of the solution and the vendor’s performance in executing its strategy around key areas such as vision of the business, go-to-market strategy, customer engagement, and market execution.

Market share

Market share is a metric normalized to the market leader and is based on the solution’s global revenue. Where revenue data is unavailable, Kisaco provides a representative estimate.

Kisaco Research ratings

- **Leader:** This vendor appears in the top right of the KLC chart and has established a significant market position with a product that is technologically advanced compared with peers and its market execution is strong.
- **Innovator:** This vendor appears in the bottom right of the KLC chart and has established a significant technological lead compared with peers but may be still early in its market execution.
- **Contender:** This vendor appears in the top left of the KLC chart and has established an excellent record executing on its market vision. The product is technically strong compared with peers but may be still early in its development.

- **Emerging player**: This vendor appears in the bottom left of the KLC chart and has a strong enough product to have participated in the KLC. The vendor may be still in early stages of establishing itself in the market, or it may be a niche player with a product aimed at a narrower range of customers.

Further reading

Kisaco Leadership Chart on Enterprise ML Lifecycle Solutions 2020-21, KR327, August 2020.

Kisaco Leadership Chart on AI Hardware Accelerators 2020-21 (part 1): Technology and Market Landscapes, KR301, July 2020.

Kisaco Leadership Chart on AI Hardware Accelerators 2020-21 (part 2): Data centers and HPC, KR302, July 2020.

Kisaco Leadership Chart on AI Hardware Accelerators 2020-21 (part 3): Edge and Automotive, KR303, July 2020.

Acknowledgements

I would like to thank all participating vendors for their time to provide briefings and answer many questions, as well as fill out our comprehensive questionnaire.

Author

Michael Azoff, Chief Analyst

michael.azoff@kisacoresearch.com

Kisaco Research Analysis Network

We are running a network for AI chip users, buyers, and people in AI related decision-making roles for their business. We will run surveys, members will receive free reports on the results, and we will also run unique events of interest to the network. To register interest please email: analysis@kisacoresearch.com with your contact details and “Kisaco Research Analysis Network” in the subject line.

Copyright notice and disclaimer

Whilst reasonable efforts have been made to ensure that the information and content of this product was correct as at the date of first publication, neither Kisaco Research Ltd. nor any person engaged or employed by Kisaco Research Ltd. accepts any liability for any errors, omissions or other inaccuracies. Readers should independently verify any facts and figures as no liability can be accepted in this regard - readers assume full responsibility and risk accordingly for their use of such information and content.

Any views and/or opinions expressed in this product by individual authors or contributors are their personal views and/or opinions and do not necessarily reflect the views and/or opinions of Kisaco Research Ltd.

Kisaco Research

CONTACT US

www.kisacoresearch.com

Michael.azoff@kisacoresearch.com