Kisaco Research

Kisaco Leadership Chart

Kisaco Leadership Chart on AI software optimization solutions 2021: Deci AI

Publication Date: 22 Feb 2021. Product code KR374.1

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 Deci AI.

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.

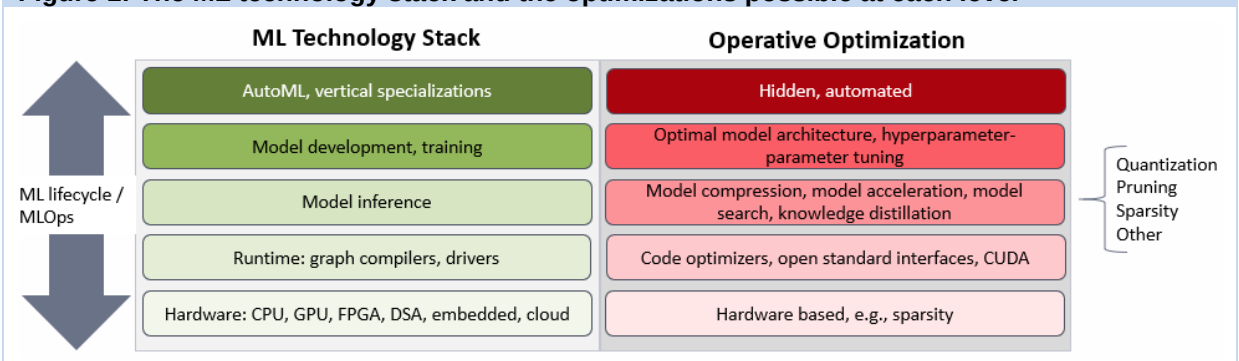
An AISO product typically achieves all of the above and does so without reducing the accuracy of the model. Thus, for inference mode AISO products, they start with a trained model that has achieved a desired accuracy, this accuracy is maintained by the AISO product as it performs optimization. However, 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



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. For example, automated deep learning hyperparameter optimization is a technique that takes a large amount of labor out of fine-tuning hyperparameter values.

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

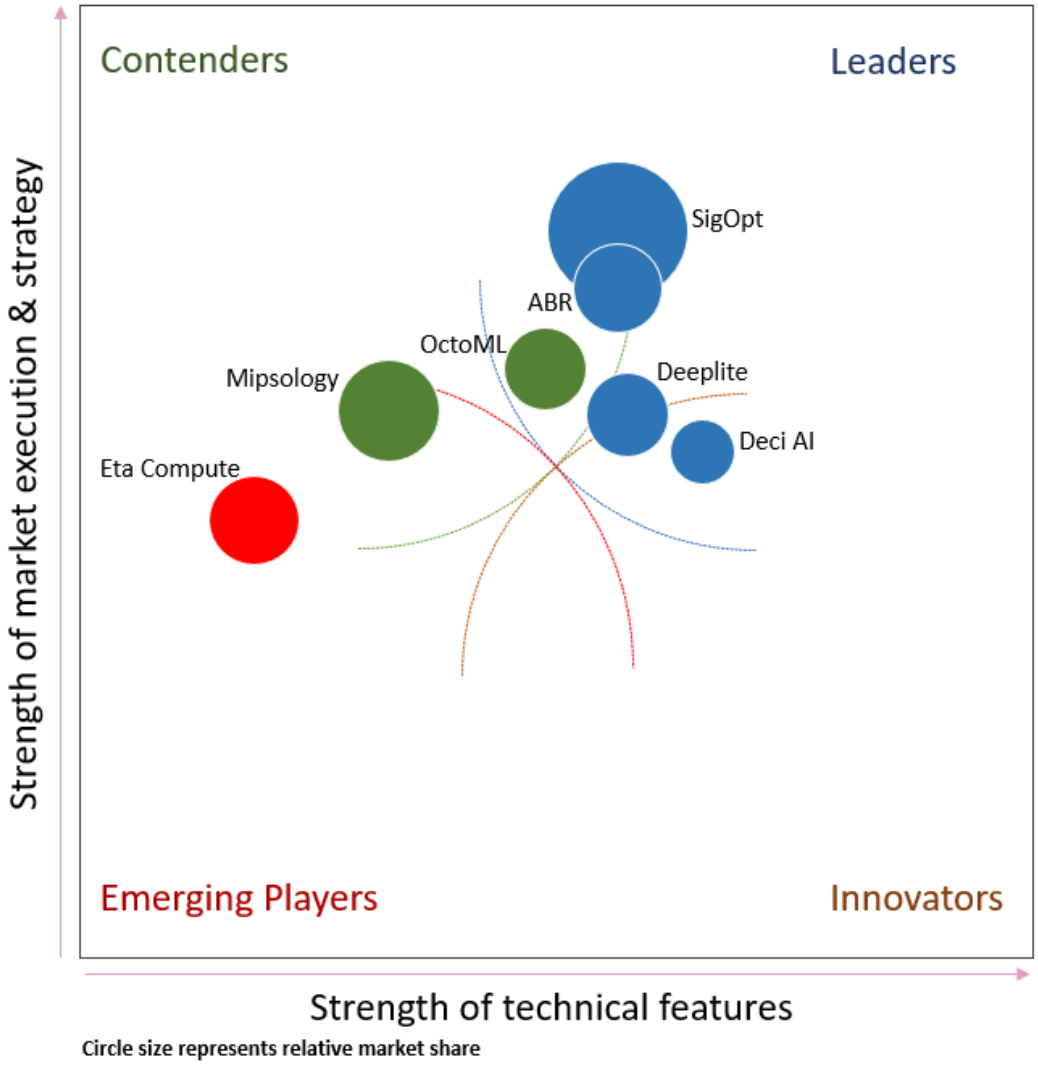
The KLC chart for AI software optimization solutions

For the KLC, Figure 2, the scoring was split across three dimensions: the x-axis technical features assessment is the total of the 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 Deci AI as leader. Deci has moved rapidly for a newcomer to the field, with a strong go-to-market strategy and is

also the top scoring KLC vendor in our technical assessment, with AISO suitable for both training and inference modes. Deci has advanced innovation in search for optimal neural network architectures.

Figure 2: Kisaco Leadership Chart on AI software optimization solutions 2020-21



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

Deci AI, Kisaco evaluation: Leader

Product: Deci Deep Learning Platform, available on-prem and hosted by Deci.

Deci AI is a private company founded in 2019 by CEO Yonatan Geifman, Chief Scientist Ran El-Yaniv, and COO Jonathan Elial, and is based in Tel Aviv, Israel. The team is 21 strong with backgrounds in PhDs from Technion (Israel Institute of Technology) and Weizmann Institute of Science and includes ex-Googlers and ex-members of the Israel Defense Unit 8200, the elite unit for intelligence and cybersecurity.

Deci recognizes that AI has a barrier to commercialization at scale due to the algorithmic complexity of deep learning. Deci's aim is to use AI to build next generation AI solutions and products and also help operationalize AI in production. The development cycle for deep learning applications is characterized over five stages: model creation, dataset collection, model training, model optimization, and deployment to production. Deci is seeing businesses take six months to get from stage one to five and at the end achieve unsatisfactory performance. For example, businesses found their edge applications had poor throughput and latency and did not meet requirements, and on the cloud operational costs and scaling costs were too high.

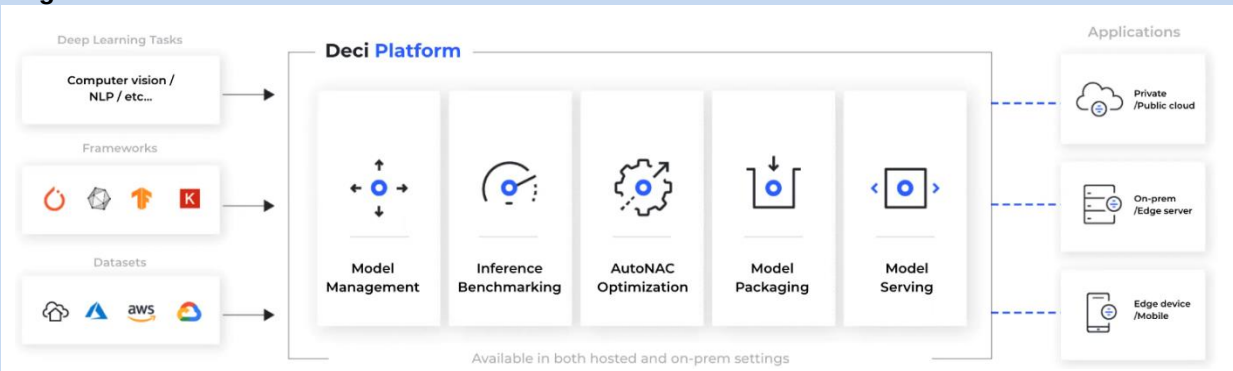
Deci meets these challenges by improving inference performance by significant factors for throughput and latency reduction, and it achieves this on any hardware – a key Deci differentiator is that Deci's technology can be applied to any hardware accelerator in theory and in practice it has a wide range supported across CPU, GPU, FPGA, ASIC, TPU, and embedded. Deci and Intel combined to submit a benchmark to MLPerf v0.7 (inference datacenter open) and achieved performance boosts of 11.8 factor in various scenarios of Intel CPU core i5 and the Intel OpenVino compiler. The results achieved were comparable to running the benchmark on an Intel Xeon processor.

These results also help reduce data center costs and cloud bills. The performance gains are achieved without compromising on accuracy – this is maintained at the same level, or at a slight reduction if the user finds this is still satisfactory for the application in order to gain even more speed and lower costs. Deep learning has traditionally had to balance performance with accuracy, however with Deci it is possible to keep accuracy unchanged while reducing the size of the AI model, which helps boost performance. Deci helps optimize the model for the target hardware and packages the model for production deployment. Deci will work with clouds, on-premises servers, and on edge and mobile devices.

The ML tools market is split between tools that optimize ML models and tools that manage the ML application lifecycle. What is different with Deci's solution is that it combines both, helping bring ML applications from development into successful deployment in production. Where enterprises are looking to scale their ML activity, we believe the use of ML lifecycle management is essential, so it is a good move by Deci to offer this as part of its optimization platform. Looking at Figure 4, the platform has five stages (the Deci Platform will be available to beta customers soon at time of writing), and it supports standard ML frameworks such as TensorFlow, Keras, and ONNX.

Deci has a runtime inference container, called RTiC, which is a containerized ML runtime engine that turns any model into a server (RTiC runs on Docker and Kubernetes). It makes deployment of Deci solutions easy to integrate into a variety of production environments, and easy to scale out.

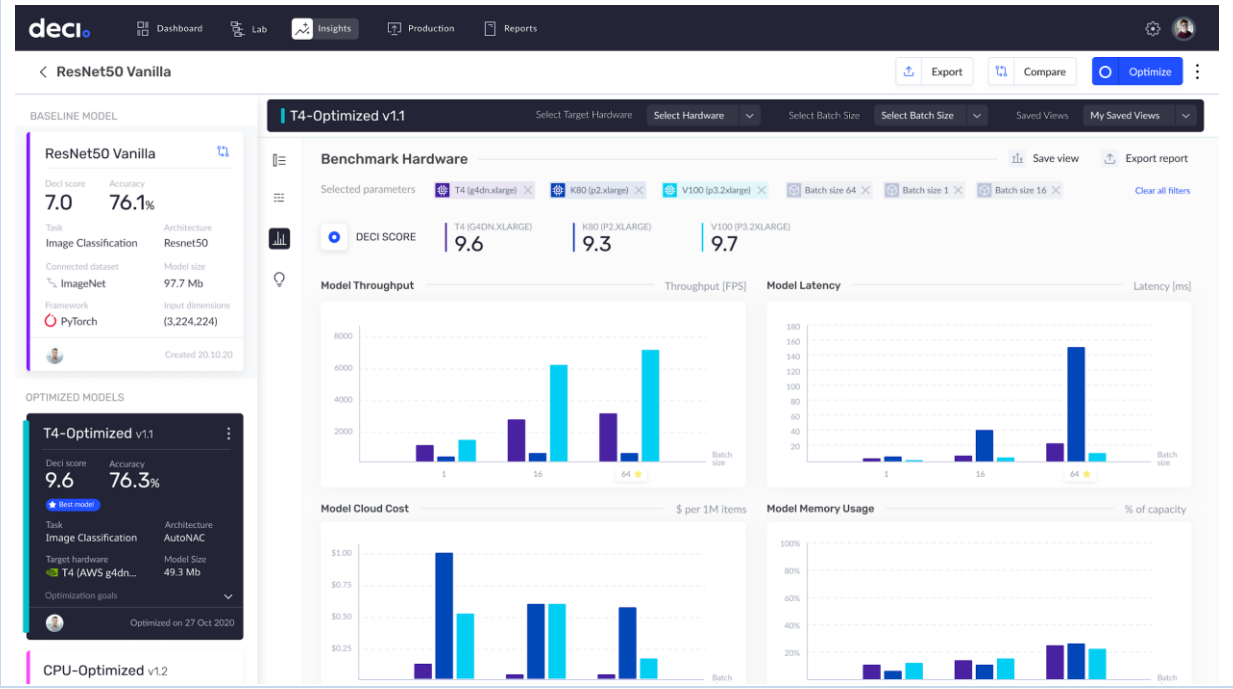
Figure 4: The Deci Platform



Source: Deci

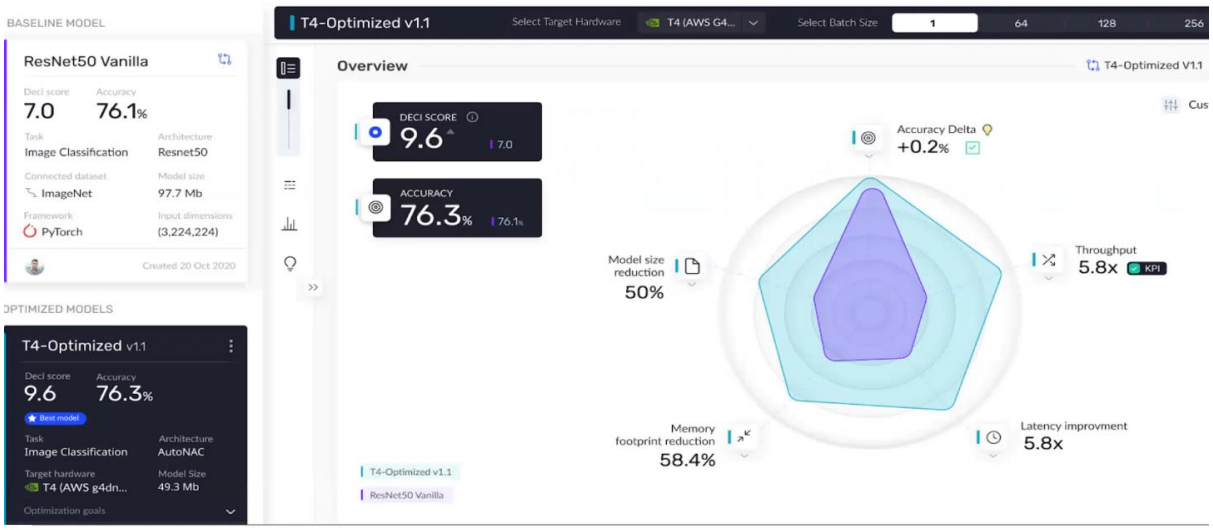
Figure 5 show the Deci platform optimizing a model named ResNet50 Vanilla. There are two Deci versions of this model, one being optimized for Nvidia GPU T4 and one just for running on a CPU. The accuracy of the models is shown and can be compared with the accuracy of the original model. Next to this is a score named “Deci” which is a metric that encapsulates all the individual metrics for this model. If the user selects/clicks the model this opens out to see the individual metrics being optimized – shown in Figure 6. The original model in purple is compared with the optimized one in green. As one can see the optimized version has improved: throughput, latency, memory footprint reduction, and model size reduction. The Deci scores provides a convenient single score to represent all these metrics.

Figure 5: The Deci Platform



Source: Deci

Figure 6: The Deci Platform Insights screen



Source: Deci

The platform allows these models to be benchmarked running on a variety of hardware and select the best optimized model and hardware accelerator combination. There is a lot of fine control over metrics, for example the model throughput is measured by different batch sizes, e.g., 64, 16, and 1. One of the most important lessons Deci learned is that optimizing models toward specific target hardware devices is key to achieving outstanding speedup. Avoiding fine hardware awareness even within the same hardware group can be detrimental. For instance, an excellent model optimized by Deci for core-i-7 CPUs turned out to be inferior to the baseline when inferencing on core i-5 CPUs, and vice versa. Such cases are frequently encountered and emphasize the need in fine hardware awareness.

Some of Deci's competitor technologies operate at the graph compilation level (such as Nvidia TensorRT, Apache TVM, Intel OpenVino, Nvidia CUDA, and Apple CoreML). Such technology, however, is somewhat limited because the computation graph must preserve the same input output functionality. On the other hand, these tools do preserve the accuracy of the original model. Before using such tools, or in conjunction with them, it is possible to perform standard quantization and pruning to compress the model and achieve gains from use of a smaller model with reduced numerical representation. Such operations, for example pruning 50% of the weights, can maintain the accuracy but the speedup is not realized in a linear way and this relates to standard (CPU, GPU, FPGA) hardware accelerators not being able to fully leverage sparse networks.

At the highest level of the technology stack and the point where the greatest inference speedup can be performed is where Deci introduces its key innovation: this is the model construction level where neural architecture search can be performed. Deci's technology is called automated neural architecture construction (AutoNAC) and it performs a search for the optimum model running on a given hardware. It uses the baseline model to set and preserve the accuracy level, and then performs optimizations across all the levels: given the data and hardware it performs model search at the top level, while adding acceleration on top of the speedup gains of lower levels, such as quantization and compilation.

AutoNAC will often retain the stem of the neural network model (the first few layers) and the predictor layers (the last few layers at the output), and in the middle it replaces the baseline network with a router/mixer control network, followed by parallel options of small, uncorrelated expert networks that can be selected by the controller network to run singly or in a parallel mixture of experts.

AutoNAC then performs a search for the optimum router/mixed setting and expert selection to replace the innards of the baseline model. Finally, AutoNAC performs fine tuning of all the model layers, including stem and predictor parts. This approach allows AutoNAC to rapidly search in a very expressive yet manageable space of candidate architectures.

Sometimes Deci customers have already performed lower-level optimizations and AutoNAC is able to improve these with the model search. The mixture of experts may include pre-built library components that Deci has found are optimal for certain applications, however it finds that even small differences in the data and hardware (such as different CPU) will affect the optimum model and so a new search is typically performed.

Currently Deci focuses on all types of machine vision as a mature market with strong opportunities for inference optimization. Deci is also beginning to look at natural language processing and conversational AI. Looking further ahead, current deep neural networks cannot handle tabular data but Deci believes that in time everything will be handled by neural networks and the range of applications will increase (largely due to the memory capacity of devices like mobile phones continuing to increasing and so be able to hold more data in memory that these applications require).

Finally, Deci has started with a focus on the inference market as this has the largest market opportunity (correlated to the size of the population at large and the amount of data in the world), but on its roadmap it has plans to optimize deep neural networks during training (correlated to the number of data scientists in the world). Also, on its roadmap is adding more end-to-end lifecycle support features, such as monitoring the quality of model predictions in production.

Kisaco Assessment

Strengths

- Deci AI achieved the highest score in our technical features assessment and also impressed us with its go to market strategy and the early partnerships it has secured, despite being a young company, and we have positioned it as Leader in our KLC. The solution excels in every area of our assessment.
- The scope for Deci AI covers both training and inference and by sitting at the top of the ML technology stack it is able to have the highest degree of influence in optimizing the ML model. The key technology offered is the automated neural architecture search (AutoNAC) which is able to both compress and accelerate the original model, without reducing the accuracy.
- We were also impressed by Deci AI's inclusion of ML lifecycle management capabilities in its solution (sometimes called MLOps): in our assessment Deci AI scored highest for this function. Enterprises typically struggle to move ML models from the lab to production, and lifecycle functionality helps reduce the friction in deployment, manages change, and monitors performance in production. Deci's decision to add lifecycle support features we believe was a smart move that ensures users are able to exploit the development of ML models to best effect.

Weaknesses

- Deci AI addressed all the main features we were looking for so there is little that is weak. The main challenge for the company is that it is fresh in the market and needs to remain above the lure of an exclusive deal with a hardware accelerator supplier that wants a differentiated offering. So far Deci AI has navigated such waters well and we expect it to grow its customer base.
- We do not see it as a problem weakness, but Deci AI does not apply its technology to neuromorphic or spiking neural networks (SNN). We are seeing an upsurge in SNN hardware startup activity and there will be an opportunity and a gap in the market for optimizing SNN – something for Deci to consider on its roadmap.
- For a startup covering two major areas, optimization and lifecycle, would normally be a challenge. The ML lifecycle management tools market is an active and distinct one but still relatively small, which means that Deci can be effective in its activities in this space because it is more likely customers will not have such a dedicated solution in place. For this reason, it is currently a smart move as we said, but it is something to watch and ensure the main target of optimization is kept in focus.

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

CONTACT US

www.kisacoresearch.com
Michael.azoff@kisacoresearch.com

