RETAIL

Solution Brief

Revision 0.1

habana®

I.

AI in Retail

The Artificial Intelligence in the Retail market was estimated to be worth roughly $2.9 billion in 2021 and is expected to reach $17.1 billion by 2028 at a compound annual growth rate (CAGR) of 34.10% [1]. Retail seems to be the fastest growing sector adopting AI-based technologies.

AI has been widely used to process business-related information in order to reach deeper business insights and market forecasting. However, the ubiquitous growth in surveillance and monitoring has also opened a new door for artificial intelligence to help analyze the business activity in physical stores, which benefits store management. Store occupancy tracking and queue monitoring are just a couple of such valuable applications.

The advancement in AI has been so promising that the analysis can be done at the level of individual items in the stores, bringing checkout-free technologies to smart stores. This is a booming new market with a lot of competition between large companies and startups to build smart stores or retrofit solutions to existing stores. AI is also used in automatic inventory management and to reduce shrinkage in stores.

Habana offers hardware and software solutions to increase the pace of research and development in the field by shortening the time of AI experiment cycles. We are also proud that our solutions reduce the development and validation costs letting the researchers explore more within their budget.

II.

Applications of AI in Retail

Most retail stores are already equipped with surveillance cameras. These cameras are usually installed for security purposes. However, with the help of AI, they can be transformed into powerful monitoring and analytics devices. that help optimize operations. This can help business owners make more informed decisions that improve customer satisfaction, staffing, and sales.

Computer vision models can be used to detect people and extract their demographics. This has many benefits for store management. A store manager can automatically count the number of customers in the store. They can market their products better by knowing the demographics of their customers throughout the day and where the members of each gender/age group spend the most time in the store. They can manage their staff efficiently by monitoring the register queue or detecting crowd gatherings.

Inventory management is another application of AI in retail. Forecasting demand using predictive analysis helps business owners have the right products in the right store at the right time. On the other hand, some companies have built AI robots that circle in the stores and watch the shelves to detect misplaced, mispriced, and out-of-stock items. This technology arms retailers with stronger insights into shelf availability, and ensures that items are more quickly restocked and corrected, improving the customer experience. AI can also be helpful in reducing shrinkage by providing tools to monitor the stores and warehouses actively and automatically.

Applications of AI in Retail
CONTINUED

Maybe the most advanced application of AI in retail is in checkout-free stores. The information from various sensors is processed in real-time to monitor which item is picked up by which customer and charge the customer automatically when they leave the store without having a checkout register. CV-based solutions for smart stores need iterative targeted model retraining as new SKUs are frequently added to the store inventory.

Key Use Cases	Task	Popular Model Architectures
Checkout-free technology	object detection activity recognition	YOLOv5 YOLOX Mask R-CNN U-Net 3D DenseNet TCN
Business activity analytics	occupancy management occupancy tracking people counting heat mapping crowd gathering detection queue monitoring	
Inventory management	low-stock and out-of-stock shelf recognition demand forecasting reduce shrinkage	

In addition to the CV-related use cases, Natural Language Processing (NLP) techniques have also found use in the retail domain specifically in e-commerce applications. The main applications of NLP models in retail include customer feedback tracking, product categorization in online e-commerce sites, and buyer/seller fraud detection. Customer feedback tracking [3] includes mining product reviews for sentiment analysis towards brands/products, identifying fake reviews, identifying actionable recommendations from product reviews, etc. Product categorization [4] involves automatically assigning the correct product category based on product description and title. NLP applications for e-commerce also include product search [5]. In addition, the e-commerce domain also uses NLP models for conversational AI agents to engage with customers and chatbots. State-of-the-art transformer-based NLP models have been used in these applications.

Challenges

Stores love it when customers take home way more than they planned for. So, they want to carry more items. According to the Food Marketing Institute [2], a traditional supermarket has, depending on its size, anywhere from 15,000 to 60,000 SKUs.

Now, imagine a company that provides checkout-free technology to grocery stores. If it's a vision-based system, the object detection model needs to know all these items. On the other hand, new products come to the market every day. These models need be frequently retrained and updated to be able to generalize to more stores and cover more items. The cameras that are installed for this purpose look at dense scenes with many small objects. Therefore, large input images are favorable to reach the desired detection accuracy. This means heavier computing and need for more memory.

III.

Why is Gaudi a good fit for retail use cases?

Training deep object detection and tracking models for checkout-free stores require a large amount of processing that can be parallelized and thus accelerated. They can benefit specifically from accelerators that can handle data parallelism when the training dataset is huge or distributed and model parallelism when the models are large.

The two primary considerations that come into play in employing AI processing - whether for computer vision or NLP applications - are time-to-train models to the desired level of accuracy and cost-to-train. Habana's Gaudi Training Processors are expressly designed—in both hardware and software—to deliver high-efficiency cost- and time-to-train, making AI training more accessible to more organizations and for more applications. This helps to reduce development and validation costs and to enable rapid innovation and faster time to market.

Training with Gaudi clusters is available both in the cloud with AWS EC2 DL1 instances consisting of 8 Gaudis and on-premises with the Supermicro X12 Gaudi Training Server, also featuring 8 Gaudis.

The ideal equation for end users is to achieve desired AI price-performance, meaning that the cost and time to train each image or language sequence meets cost and time investment criteria. In other words, enabling more training at a low cost is the objective for data scientists and IT infrastructure management.

First-generation Gaudi, in fact, has proven delivery of up to 40% better price-performance than with comparable GPU-based solutions—for both the EC2 DL2 instance as well as for on-premise systems. And, there are customer cases that have proven even greater cost savings, which will be shared in part 4 of this brief.

In addition, Gaudi2, which launched in May 2022, offers substantial performance advances that enable significantly faster training of models, while preserving cost-efficiency. Gaudi2 systems will be available in 2H 2022 for on-premises implementation.

IV.

News and customer testimonials

"Our experiences with Gaudi give us confidence that we will be able to lower our training costs while improving model quality and translate that into even more powerful tools for our end-user."

Maxime Bergeron
R&D Director
Riskfuel

"On our own models the increase in price performance met and even exceeded the published 40% mark."

Chaim Rand
Machine Learning Algorithm Developer
Mobileye

"We are consistently seeing cost-savings compared to existing GPU-based instances across model types, enabling us to achieve much better Time-to-Market for existing models or training much larger and complex models."

David Peer
DevOps Tech Lead & Specialist
Mobileye

V.

References

[1] Vantage Market Research, "Artificial Intelligence (AI) in Retail Market Size to Reach 17,086.54 USD Million by 2028| Increasing AI-powered Application and Chatbots in the Retail Industry Flourish the Market", March 2022.

[2] Food Marketing Institute: Supermarket Facts, available at https://www.fmi.org/our-research/supermarket-facts

[3] Semantics between customers and providers: The relation between product descriptions, reviews, and customer satisfaction in E-commerce, available at https://arxiv.org/abs/2203.16489

[4] DeepCAT: Deep Category Representation for Query Understanding in E-commerce Search, available at https://arxiv.org/abs/2104.11760

[5] Generating Rich Product Descriptions for Conversational E-commerce Systems, available at https://arxiv.org/pdf/2111.15298.pdf