Case Study

Service Providers
Artificial Intelligence

Scaling the Global Customer Experience

Lilt's adaptive neural machine translation (NMT) platform, running Google Cloud N2 instances and optimized for Intel® architecture, enables enterprises to deliver first-class customer experiences to global audiences no matter what language they speak.

Currently, global enterprises spend millions refining their customer experience in one language, usually English. But, when they attempt to transfer this experience to other countries and regions, they typically don't invest comparable resources or effort. The same high level of customer experience isn't translated across borders.

Lilt encourages enterprises to rethink this approach by focusing on the global customer experience. It works with localization, global growth, and customer experience teams to ensure they reach every customer at every touchpoint in their preferred language. Lilt achieves this goal by providing faster and better translation, all enabled by artificial intelligence (AI).

Challenge

- To provide enterprises with localization services at scale, Lilt augments human translation with adaptive neural machine translation (NMT) suggestions and technologies.
- Returning accurate AI-supported translations as fast as possible requires massive computing power.
- Since Lilt serves translation requests from customized models via the CPU, CPU efficiency is often as crucial as GPU efficiency.

Solution

- As a member of Intel® AI Builders, Lilt has early access to Intel's technology roadmap, enabling it to optimize its platform to run inference on Intel® processors in the cloud.
- Using TensorFlow 2.6 and Python high-level programming language, Lilt and Intel optimized Lilt's NMT platform to run on Google Cloud N2 instances with 2nd Generation Intel® Xeon® Platinum 8280L processors. Lilt's NMT platform now delivers 2.53x better latency and increases inference throughput by 355.6 tokens per second, compared with Baseline TensorFlow[1].
- On bare metal instances, 3rd Gen Intel® Xeon® Platinum 8380 processors optimized with TensorFlow 2.4 improve performance by 3.84x, compared with Baseline TensorFlow. And up to 2.6x, compared with previous-generation Intel® Xeon® Scalable processors[2].

Results

- Working with Intel® AI Builders and Intel Capital has provided Lilt with the foundation to launch its differentiated, adaptive NMT platform.
- With Intel optimizations, Lilt can deliver faster, more accurate, scalable AI-powered localization for a lower total cost of ownership (TCO).
- In turn, faster, more accurate AI-powered translations enable enterprises to cost-effectively manage global customer experiences at scale. Something that wasn't previously possible using human translators alone.
- Lilt's customers can reduce translation costs by up to 60 percent year on year for the same volume of content[3].

> "Intel has helped us a lot with benchmarking our translation system. They have put in CPU-specific optimizations so it's really efficient, making a meaningful difference in the live production setting."
>
> **Spence Green,**
> co-founder and CEO at Lilt

Fast neural machine translation requires low latency inference

Lilt's innovative solution combines adaptive neural machine translation (NMT) technology, an enterprise global experience management system, and world-class professional translators. It enables organizations to scale their translation programs, speed time to market, and improve the global customer experience.

The Lilt platform's core technology is an interactive, adaptive NMT system that learns in real-time from human feedback and existing translation memory data. Adaptation allows the system to provide progressively better suggestions to human translators, and higher quality for fully automated translation. This patented solution reduces translation errors by nearly 70 percent and helps linguists become three to five times more productive[3].

Critical to the success of the Lilt platform is achieving low latency on inference instances to return accurate AI-supported translations in the fastest time possible. But this level of performance requires a massive amount of computing power. Spence Green, co-founder, and CEO at Lilt, explains: "Our system trains machine learning models in production to adapt to the people using it and the workflows it's being used to translate. As a result, we're continuously swapping in and out personalized models in our production system. And that means it's not easy for us to accelerate our production system using GPUs, because we serve every request from a different model. We need to be very efficient on the CPU. That's where Intel® processors can help."

Optimizing Lilt's platform to run on Intel® architecture

Since 2015, Lilt and Intel have collaborated to improve the performance of Lilt's adaptive NMT platform. As a member of Intel® AI Builders, Lilt benefits from having early access to Intel's technology roadmap, enabling it to test, integrate, and optimize its platform to run inference on Intel® processors in the cloud.

Currently, Lilt's NMT platform runs on Google Cloud N2 highmem-16 instances with Intel® Xeon® Platinum 8280L processors. The platform is optimized with the TensorFlow 2.6 open-source software library for machine learning and AI and Python high-level programming language. This solution delivers 2.53x better latency and an inference throughput increase of 355.6 tokens per second, compared to Baseline TensorFlow[1].

As a result of these performance gains, Intel has now moved virtually all of its localization workload to Lilt. See box out case study.

To improve performance further, Lilt is investigating the benefits of upgrading, when available, to Google Cloud instances running on 3rd Gen Intel® Xeon® Scalable processors. On bare metal instances, running normalized inference workloads, Lilt has already experienced performance gains of 3.84x on the Intel® Xeon® Platinum 8380 processor at 2.3 GHz. These gains were made using Intel® Advanced Vector Extensions 512 (Intel® AVX-512), the Intel® oneAPI Deep Neural Network Library (oneDNN), and Intel-optimized TensorFlow 2.4 with native format. Compared with previous-generation Intel Xeon Scalable processors, this configuration delivers a performance improvement of 2.60x[2].

Commenting on Lilt's collaboration with Intel, Green says: "Intel has helped us a lot with benchmarking our translation system. They have put in CPU-specific optimizations so it's really efficient, making a meaningful difference in the live production setting."

Intel also introduced Lilt to Intel Capital, which invested in Lilt by leading its Series B funding round. The funds have been used to expand development of Lilt's AI-powered enterprise translation software and increase adoption amongst enterprises.

Case study: Maximizing return on investment (ROI) of localization spend[4]

Intel's localization team was frequently asked to do more with less.

- As Intel expanded into new verticals and locales, the localization team had to serve new internal business partners with flat or even reduced budgets.
- With each additional business partner came new subject matter and content types, meaning that the team's ability to use existing linguistic assets like translation memories (TMs) was limited.

To improve ROI on localization, Intel needed a solution that incorporated potentially transformational technologies like adaptive neural machine translation (NMT) and intelligent automation.

- Intel conducted an initial proof of concept. It explored Lilt's ability to scale rapidly and work with Intel's existing systems.
- After carefully analyzing the performance, costs and benefits of Lilt's solution, Intel moved forward with deployment.

Adding Lilt to the Intel IT translation portfolio enables the localization team to do more, without sacrificing quality:

- Lilt's adaptive NMT technology enables Intel to translate the same amount of content while reducing costs by 40 percent year-over-year.
- The impact of that reduction is significant, enabling one of Intel's business units to double its volume of translated content with only a marginal increase in budget.

Faster, more accurate localization, for less

Optimizing Lilt's neural machine translation to run on Intel architecture has helped to improve translation speed. Lilt can now translate more text with fewer linguists, lowering its total cost of ownership (TCO).

Since Intel-based Google Cloud N2 instances quickly load translation models simultaneously, Lilt can easily update the translation model for each customer with every new translated sentence, enabling it to improve translation accuracy. Fast, accurate, and scalable translation appeals to a much broader range of enterprise customers and provides Lilt with a competitive edge in the localization and global experience market.

Lilt's customers also benefit from faster, more accurate translations for reduced cost. Lilt only charges customers for localizing 'new' words that haven't previously been translated. Lilt's customers can, over time, translate the same amount of content for a reduced cost, or translate ever more content for the same budget.

Businesses can deliver better experiences by engaging employees and customers in their language of choice, helping them expand their business across the globe. They also benefit from having a single platform that allows them to scale translation across the business, improving time to market for digital content and services.

Finally, a single point of contact allows enterprise customers to spend more time developing their global customer experience strategy, rather than dealing with the day-to-day overhead and complexities of managing multiple vendors.

> "The productivity gains from Lilt's AI-powered translation services have enabled us to reduce translation costs by 40 percent year on year for the same volume of content, and we expect those cost savings to continue to increase. Because translation makes up a significant share of our overall localization costs, this was very material to the business[3]."
>
> **Loïc Dufresne de Virel,**
> head of localization at Intel

Improving future customer experiences

Working with Intel® AI Builders and Intel Capital has provided Lilt with the foundation to launch its differentiated, adaptive NMT platform, transforming the way enterprises approach and consume localization services.

In turn, enterprises can now manage global customer experiences at scale. The Lilt AI-powered platform enables them to speak to customers cost-effectively at every touchpoint in their preferred language. With human translators alone, this achievement wouldn't be possible.

Lilt has several plans to improve further the performance of its NMT platform:

- Adding more context into the predictions made.
- Building models that adapt to the domain they're trained on even more rapidly.
- Introducing ways to ensure the quality of its machine translation services.

"All these systems require more compute and ever more expressive models. That's where the hardware collaboration with Intel continues to deliver an advantage," concludes Green.

About Lilt

Headquartered in San Francisco, Lilt is the modern language service and technology provider enabling the global customer experience. Lilt's mission is to make the world's information accessible to everyone regardless of where they were born or which language they speak. Lilt brings human-powered, technology-assisted translations to global enterprises, empowering product, marketing, support, e-commerce, and localization teams to deliver exceptional customer experiences to global audiences. For further information, visit: https://lilt.com/.

Learn More

You may find the following resources helpful:

- **Intel® Xeon® Scalable Processors**
- **Intel® AI Builders**

Find the solution that is right for your organization. Contact your Intel representative or visit **intel.com/ai**.

[1] Testing conducted comparing inference using Baseline TensorFlow compared to Intel-optimized TensorFlow Testing carried out by ISV/Intel on November 14, 2020. CONFIGURATIONS: CPU: Intel® Xeon® Platinum 8280L processor at 2.70 GHz; OS: Ubuntu 18.04.4 LTS; BIOS: SE5C620.86B.02.01.0011.032620200659

[2] Testing conducted comparing inference using Baseline TensorFlow compared to Intel-optimized TensorFlow. Testing carried out by Intel on April 6, 2021. CONFIGURATIONS: CPU: Intel® Xeon® Platinum 8380 processor at 2.3 GHz; OS: Ubuntu 18.04.5 LTS; BIOS: SE5C6200.86B.0022.D08.2103221623; compared with CPU: Intel® Xeon® Platinum 8280L processor at 2.70 GHz; OS: Ubuntu 18.04.4 LTS; BIOS: SE5C620.86B.02.01.0011.032620200659

[3] https://lilt.com/customer-stories

[4] https://lilt.com/customer-stories/intel