CASE STUDY

2nd Generation Intel® Xeon® Scalable Processors
Intel® Deep Learning Boost with Vector Neural Network Instructions (VNNI)
nGraph Compiler

Transforming Business through Deep Learning

AI technologies from Intel improve the performance of Baidu's open source deep learning framework PaddlePaddle, expanding its applications on an industrial scale.

> "Optimized with custom Intel® Xeon® Scalable processors, Baidu PaddlePaddle is able to store, process and analyze huge amounts of product image data, quickly deploy deep learning models and increase model training efficiency. With help from Intel, we have successfully enhanced PaddlePaddle's deep learning performance at the structural level, making it possible for us to continually improve customer experience, accelerate product development, and ultimately empower more businesses."
>
> **Si Cheng**
> Senior Manager
> Baidu PaddlePaddle Deep Learning Platform

A leading technology firm in China, Baidu began as an internet search provider and has continued to expand into AI, cloud computing and big data. Baidu firmly believes that AI represents the next major milestone for civilization and is of strategic importance.

Consequently, Baidu has been exploring a number of cutting-edge AI technologies, from voice recognition to image processing and machine translation. Moreover, with the mission to "Make a Complicated World Simpler Through Technology", Baidu continues to share its AI breakthroughs with the wider industry.

PaddlePaddle is a prime example of the AI solutions Baidu can provide, giving businesses and developers of all sizes the ability to implement advanced deep learning technologies in real-life situations.

Baidu's research into deep learning dates back to 2012. Today, deep learning technologies are highly modelled, standardized and automated, serving as an entry point for a range of industries to apply AI technologies on a large scale.

As an open source deep learning platform, PaddlePaddle gives users a jumpstart by linking high performance processors and computing systems with their deep learning models to solve enterprise-wide challenges. It has been widely welcomed by the developer community and has evolved into a comprehensive ecosystem.

Intel plays a vital role in this process, by providing PaddlePaddle with the compute power it needs to establish itself as a popular deep learning framework trusted by a large number of Chinese businesses.

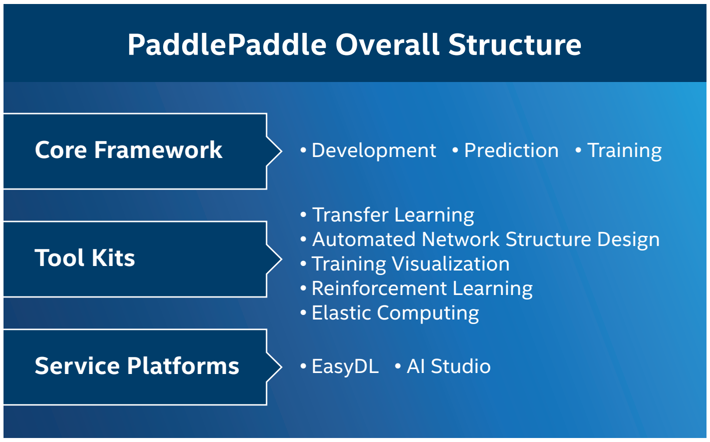
PaddlePaddle – Key Objectives and Challenges

When the PaddlePaddle design program started years ago, Baidu decided to place parallel distributed deep learning at its core (hence the name "paddle"), with a vision to enable large-scale model training for real-world applications.

In 2016, PaddlePaddle opened its source code, giving full access to developers and businesses of all sizes to support their innovation and transformation. With an aim to become a deep learning platform based on real industrial practices, PaddlePaddle focuses on application scenarios across industries. Its new Chinese name Fei Jiang, meaning Flying Paddle, was unveiled in April 2019, and speaks of its ambitions to soar.

Built from the perspective of users who value efficiency and ease of use, PaddlePaddle adopts a three-part overall structure - core framework, tool kits and service platforms.

- The core framework provides model development, training and prediction capabilities, supported by ready-made models across all areas, from image recognition to natural language processing, delivered to users in a modularized manner.
- Tool kits cover areas like transfer learning, reinforcement learning, automated network structure design, training visualization and elastic computing to meet the needs of large-scale industrial production.
- Service platforms include EasyDL*, a starter-friendly customized training and service platform, and AI Studio*, a one-stop development platform.



PaddlePaddle Overall Structure

Core Framework
- Development
- Prediction
- Training

Tool Kits
- Transfer Learning
- Automated Network Structure Design
- Training Visualization
- Reinforcement Learning
- Elastic Computing

Service Platforms
- EasyDL
- AI Studio

Figure 1. A three-part overall structure of PaddlePaddle.

Such a comprehensive structure allows users to easily take full advantage of advanced deep learning technologies. However, building a platform of such complexity and magnitude is full of challenges.

In the words of Si Cheng, Baidu PaddlePaddle Deep Learning Platform Senior Manager, "From PaddlePaddle 1.0 to 3.0, our challenges have grown bigger by the day. We need to continually push ourselves to unlock more computing power, create more efficient and friendly user experiences, and ultimately lower the barriers so businesses and developers can easily access cutting-edge AI technologies."

How can PaddlePaddle be improved so it is easier to learn and use? How can it enable more efficient development and deployment? How can it become an AI platform that has the power to support industrial applications? These are the questions PaddlePaddle's team facing each day.

Solutions with Intel® Technologies

None of these questions can be solved without tackling one fundamental task: performance optimization. To achieve this, PaddlePaddle uses a top-down approach with its infrastructure, investing heavily in its general technology framework in addition to hardware and base-level optimizations.

Intel plays a vital role in this process by supplying a robust range of hardware, in conjunction with software that enables maximum hardware performance.

Howard Chang, Intel Global Account Sales General Manager, said, "To help Baidu PaddlePaddle increase its model training performance and achieve innovative breakthroughs, we used Intel® Xeon® Scalable Processors as the starting point, while exploring optimization opportunities in computing power, memory, structure and communications. As a result, we elevated PaddlePaddle's model deployment capabilities to a new level. And Intel® Deep Learning Boost with VNNI/INT8 optimizations helped increase deep learning inference throughput significantly."

> "... we elevated PaddlePaddle's model deployment capabilities to a new level. And Intel® Deep Learning Boost with VNNI/INT8 optimizations helped increase deep learning inference throughput significantly."
>
> **Howard Chang**
> Intel Global Account Sales General Manager

2nd Generation Intel® Xeon® Scalable Processors

Intel® CPUs play a fundamental role at PaddlePaddle. With most applications, relying on CPU power is more cost effective; and when it comes to data training at scale, CPU clusters deliver impressive processing capabilities. Secondly, Intel has been committed to making CPUs optimized for deep learning inference workloads, which is exactly what PaddlePaddle requires.

With customized 2nd Gen Intel® Xeon® Scalable processors, PaddlePaddle is making big strides in performance optimization. The 2nd Gen Intel® Xeon® Scalable processors offer a 25% to 35% performance increase over the previous generation[1] and provide new features for great agility, enhanced memory, stronger security, and more.

Intel® Deep Learning Boost with Vector Neural Network Instructions (VNNI)

More importantly, the 2nd Gen Intel® Xeon® Scalable processors feature Intel® Deep Learning Boost (Intel® DL Boost) technology, which includes new Vector Neural Network Instructions, or VNNI. Based on Intel® Advanced Vector Extensions 512 (Intel® AVX-512), VNNI improves inference performance by functioning as an embedded accelerator.

This performance improvement and efficiency is delivered by using a single instruction to handle workloads which would previously require three separate AVX-512 instructions[2]. Examples of targeted applications include image classification, speech recognition, language translation, object detection and more.

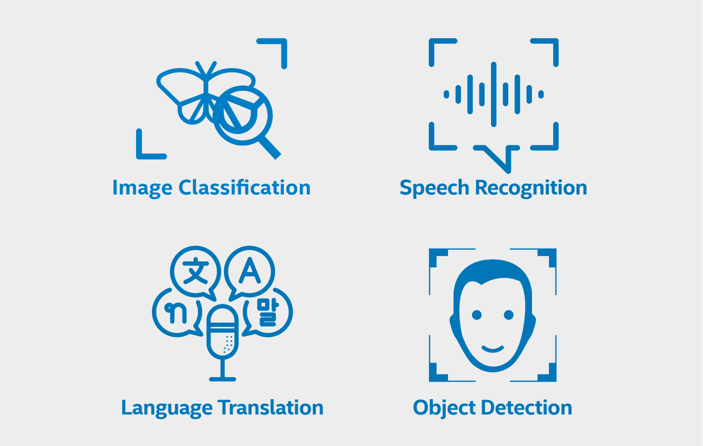


Figure 2. Examples of targeted applications.

As a deep learning platform, PaddlePaddle is particularly focused on increasing efficiency for low-precision computation (generally, deep learning models do not require precise numerical results). Intel® CPUs are well-suited for the task because they provide excellent vector processing, and VNNI enables the use of INT8 data types that can include more data in a single instruction, therefore accelerating multiplication and addition operations. Intel and PaddlePaddle's model verification results have proven VNNI's impressive acceleration power. For instance, when operating in conjunction with 2nd Gen Intel® Xeon® Scalable Processors to process a full ImageNet dataset—INT8 performance is significantly higher than FP32, with an accuracy difference of under 1%[3].

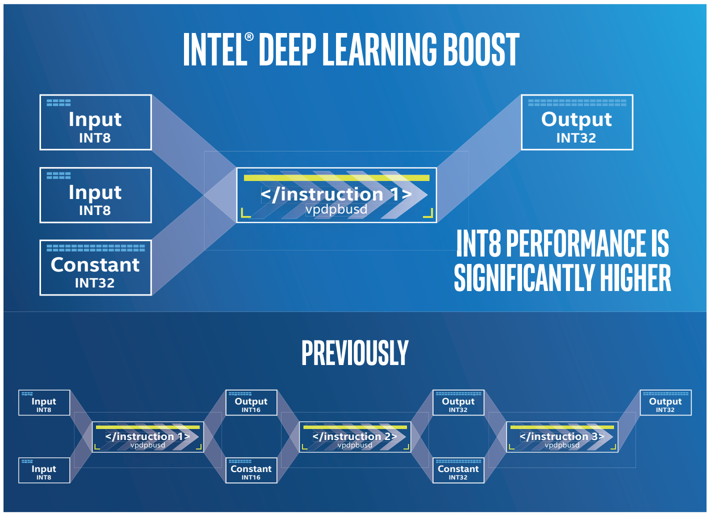

INTEL® DEEP LEARNING BOOST
Input INT8
Input INT8
Constant INT32
</instruction 1> vpdpbusd
Output INT32
INT8 PERFORMANCE IS SIGNIFICANTLY HIGHER
PREVIOUSLY
Input INT8
Input INT8
</instruction 1> vpdpbusd
Output INT16
Constant INT16
</instruction 2> vpdpbusd
Output INT32
Constant INT32
</instruction 3> vpdpbusd
Output INT32

Figure 3. INT8 performance increase enabled by Intel® Deep Learning Boost with VNNI and 2nd Gen Intel® Xeon® Scalable Processors.

Intel's Open Source nGraph Compiler

Another one of Intel's key contributions to PaddlePaddle is nGraph, an open-source C++ library and runtime/compiler suite for deep learning ecosystems. With the nGraph compiler, data scientists can use their preferred deep learning framework on any number of hardware architectures, for both training and inference.

Because nGraph is able to seamlessly interface with different hardware and frameworks, it offers great ease and flexibility when integrating with PaddlePaddle. As Dianhai Yu, Tech Lead of PaddlePaddle pointed out: "This helps PaddlePaddle use simplified bridge code to transform graph ops and deliver performance improvements. Preliminary tests carried out by Intel engineers together with the PaddlePaddle team have shown clear ResNet-50 performance gains."[4]

Team-Level Cooperation

Intel and Baidu have a long history of cooperation, and Intel has a range of solutions at its disposal to help Baidu in their quest to develop cutting-edge AI technologies.

As Howard Chang points out, "Intel and Baidu have a shared mission-empowering more businesses with AI technologies. Customized Intel® Xeon® Scalable processors are now highly integrated with Baidu's key infrastructure, dramatically boosting its AI computing power and enabling Baidu to improve the performance of their search engine, public cloud and other services. Large-scale data centers driven by Intel® Xeon® Scalable processors also help PaddlePaddle users enjoy a more convenient cross-platform experience. Across the spectrum, Baidu takes full advantage of Intel's products and expertise to help AI technologies empower businesses and benefit society."

To continue optimizing the PaddlePaddle platform as it grows and innovates, Intel goes beyond focusing on products and technologies to understand the key challenges in PaddlePaddle's core frameworks, tool kits and service platforms, then utilizing the entire Intel ecosystem to find solutions.

For example, Intel has been actively building its full-stack AI technology capabilities, with particular advantages in high-performance AI computational library instruction sets. At this specific level, Intel has helped PaddlePaddle write over 10,000 lines of code.

Tiezhu Gao, Senior Architect of Baidu Deep Learning Platform, feels the bond Intel and Baidu share: "When we work together, we are not like two companies but one team. The way we cooperate helps us achieve significant boosts in performance across many applications. Whenever a real problem arises, the Intel team works closely with us and provides valuable input. For instance, they helped us improve parallel computing efficiency, and solved the performance problem caused by CPU cache competitions. Efforts like this have cumulatively resulted in a huge increase in deep learning performance at the structural level."

Moving forward, the Intel-PaddlePaddle teams will focus on the following areas:

- Integrating Intel's model compression and quantization technologies for further performance optimization;
- Using Intel® Math Kernel Library for Deep Neural Networks (MKL-DNN) as the preferred high-performance CPU OP kernel;
- Applying nGraph in a more flexible manner for more efficiency gains on Intel® processors;
- Cooperating on a team level for further model deployment performance optimization, both internally and externally.

PaddlePaddle Applications

Internally, PaddlePaddle's comprehensive and powerful structure is enabling Baidu to carry out cutting-edge AI research, such as using PaddleMobile to achieve AI vision comparable to human eyes; and Senta, an open-source sentiment classification system based on semantics and big data.

Externally, PaddlePaddle is empowering a number of sectors from manufacturing to agriculture and education, enabling them to successfully apply deep learning technologies for a range of purposes: performing product quality inspections, optimizing steel smelting productions, remote sensing and detection on the golf course, identifying farming land slots, carrying out auscultation and more.

It is fulfilling its objectives on all frontiers, evolving into an AI framework with the power to support industrial applications.

Key Takeaways

Considered the building blocks for an intelligent world, deep learning frameworks are of paramount importance. Baidu continues to invest tremendous efforts and resources into developing deep learning technologies, including PaddlePaddle - their open source deep learning platform.

Through continuous improvement in user experience, performance and structure, PaddlePaddle is expertly-optimized for real-life applications, empowering a range of industries to take advantage of state-of-the-art AI technology.

With team-level cooperation, Intel uses its technologies, software and hardware to help PaddlePaddle meet and even exceed its optimization goals for computing power, memory, structure and communications.

Notably, the customized 2nd Gen Intel® Xeon® Scalable processors have helped PaddlePaddle achieve breakthroughs in performance optimization. By leveraging Intel® DL Boost technology, 2nd Gen Intel® Xeon® Scalable processors take deep learning inference to the next level. As well, the nGraph compiler is able to seamlessly interface with different hardware and upper frameworks, offering ease and flexibility when integrating with PaddlePaddle and delivering performance improvements.

[1] https://www.intel.com/content/www/us/en/technology-provider/products-and-solutions/xeon-scalable-family/2gen-data-centric-computing-article.html

[2] https://www.intel.com/content/www/us/en/now/your-data-on-intel/deep-learning-boost-video.html?wapkw=vnni

[3] Speech from Tiezhu Gao, Senior Architect of Baidu Deep Learning Platform at Intel® Data-Centric Innovation Product Launch Beijing 2019, http://bizwebcast.intel.cn/events/20190403/onDemandWeb.html?eid=134

[4] Speech entitled “Baidu Large-Scale Deep Learning Application Practice and Open Source AI Framework PaddlePaddle" at AIDC Beijing 2018, http://bizwebcast.intel.cn/events/20181114/onDemandWeb.html?eid=113

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No product or component can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com].
Intel, the Intel logo, Xeon, Intel Deep Learning Boost and others are trademarks of Intel Corporation or its subsidiaries in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.