ACHIEVE NEXT-GENERATION NATURAL LANGUAGE PROCESSING

How to keep up with the latest innovations in NLP

SambaNova® SYSTEMS

CONTENTS

Natural Language Processing is at the Center of Innovation

We interact with natural language processing (NLP) models every day. From sentiment analysis to fraud detection, NLP technology is often at the center of our day-to-day interactions with AI. NLP is one of the most actively pursued areas of machine learning among researchers, engineers, and data scientists who count on its ubiquitous automated applications, such as predictive text, virtual assistants, chatbots, semantic search, and social media listening. These applications reliably accelerate innovation, establish market differentiation, and ultimately save a lot of time and effort.

As critical as they are to the future of AI, developing state-of-the-art NLP models is very complex, expensive, and time-consuming. How can organizations accelerate time to value?

A Driving Force in AI

NLP adoption is increasing across industries. In fact, IDC recognizes AI software spend supporting NLP-powered solutions in five key industries: banking, retail, manufacturing, healthcare, and securities and investments.[1] Given its applicability and rising popularity, it's no surprise that NLP use cases are growing faster than organizations can keep up.

> Worldwide AI systems spend has moved beyond the early adopters to mainstream industrywide use-case implementation.[2]
>
> IDC

1 IDC, NLP by AI Industry and Use Case - Nov 2019
2 IDC, NLP by Artificial Intelligence Industry and Use Case, Nov 2019

NLP Technology is Well-known for Contributing to the Modern Innovations We Use Every Day

Sentiment analysis

Fraud detection

Personal assistants

Automated service agents

Search engines

Ad content suggestions

Trading signals

Document analysis

Chatbots

SECTION 1

CONVENTIONAL TECHNOLOGIES

Conventional Technologies Can't Keep Up

NLP is evolving faster than the technology used for training and inference in language models. But as demand for fast, more accurate NLP models grows, compute performance progressively lags. The two things model owners say matter most are accuracy and time to results, outputs that require adequate compute resources. Today, it's only a matter of time before NLP's requirements overwhelm the hardware and software capabilities of most organizations. In fact, we're already starting to see it happen.

From an evolutionary standpoint, NLP models are trending away from simpler models, such as LSTM, to computationally intensive, large-capacity transformer models, such as BERT and GPT-3. To perform, these larger NLP models require machine learning expertise and a powerful computational infrastructure—both of which are difficult resources to come by for most organizations.

SECTION 2

COMMON NLP MODELS

SECTION 2

Common NLP Models

Organizations use NLP to increase business efficiency by automating labor-intensive or repetitive tasks traditionally performed by people. To successfully leverage NLP for competitive advantage, they need the ability to more quickly develop highly accurate models to drive better business outcomes, justify NLP investments, and deliver innovative applications.

While the NLP model-to-application pipeline is ripe for optimization and innovation, obstacles loom large between model owners and cutting-edge breakthroughs. Training models still takes too much time in most scenarios. And the pipeline itself, with its multiple stages, presents challenges as model owners work to incorporate new efficiencies into workflows.

Models Matter

In the world of NLP modeling, these models dominate:

BERT

In late 2018, Google researchers released BERT (Bidirectional Encoder Representations from Transformers). In addition to capturing semantic meanings of words statically, BERT can pick up on the contextual meaning of words depending on how they appear in a sentence. BERT models can also capture the intention behind words, understanding subtle nuances in language—even when expressed improperly. A properly trained BERT model can enable virtually unlimited possibilities in language applications. Facebook, for instance, developed their own version of BERT called RoBERTa that tackled the social network's challenges in content moderation by having it learn multiple languages simultaneously. Unfortunately, outside of these hyperscalers, performing domain-specific pretraining and fine-tuning requires expertise and resources that are hard to come by for many organizations.

Dataflow-as-a-Service
for Language

GPT-3

Generative Pretrained Transformer-3 (GPT-3) is a deep learning language model that's capable of generating state-of-the-art, human-like text. Achieving remarkable results on a number of benchmarks, the full version of GPT-3 features 175 billion parameters—making it one of the largest natural language models in existence. The expansiveness of this powerful model comes at a price, with estimates suggesting that it costs up to a staggering $12 million in compute resources to train. Despite all of the promising capabilities that it demonstrates, this places GPT-3 firmly out of reach for most organizations. While OpenAI offers GPT-3's API for beta trials, access requests are waitlisted, and results are not guaranteed.

SECTION 3

PROCESS DEMANDS

Process Demands, Constraints, and Challenges

Ideally, remedies to deep learning training compute problems should deliver great model accuracy and be flexible enough to support memory-heavy workloads of any size when needed. Whether performing sentiment analysis on large amounts of data, or distilling large models into small models to be deployed quickly—compute capacity matters—and so does efficiently using that capacity.

Enabling state-of-the-art machine intelligence is crucial for text and language understanding in many domains. Users are pushing for larger and larger models to advance machine intelligence for NLP. But training large NLP models requires massive compute resources, and the process is constrained by the limitations of conventional technologies (i.e., CPUs and GPUs).

From training to inference, NLP requires technical expertise that is in short supply and expensive. Where once building and deploying large deep learning models was limited to hyperscale enterprises that can afford to hire hundreds of skilled technical developers, today, organizations of all sizes are looking for the means to take advantage of machine learning.

Step-by-step, Token-by-token

The standard model to application pipeline follows these primary steps:

1. Extract Information

Take in data and clean it up

2. Tokenization

Define, identify, and sequence data

3. Model Processing

Operate, compute, and perform tasks

4. Post-process Output

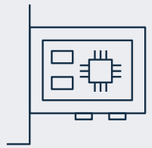
Evaluate and refine

SECTION 4

COMPUTE BARRIERS

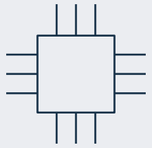
Current Compute Solutions and Small Models

Highly-detailed, smaller models require intense compute capacity and can waste compute resource.



GPUs

- Underutilized capacity
- Kernel-by-kernel execution mode
- High overhead
- Low utilization
- One operation at a time

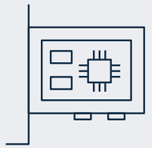


CPUs

- Capacity is too low
- Not designed for heavy linear algebra compute

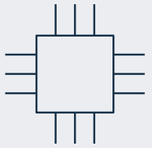
Compute Barriers for Larger Language Models

Larger language models require more compute and capacity than today's solutions deliver.



GPUs

- Design does not support enough memory
- Hundreds or thousands of GPUs needed



CPUs

- Capacity is too low
- Subject to many constraints

As NLP continues on its trajectory to evolve faster than its supporting technology, conventional processors will face more elaborate and complex bottlenecks that will further clog the NLP pipeline at the model processing stage and limit the pace of innovation due to the lack of performance and available capacity.

SECTION 5

AN EXTENSIBLE ML SERVICES PLATFORM FOR NLP

An Extensible ML Services Platform for NLP

Enterprises working toward the cutting edge of innovation with NLP need solutions that overcome compute limitations and remove obstacles to developing more advanced machine learning models. There is a better way to meet the model training needs of NLP applications.

SambaNova Systems has developed an optimal product that enables users to lower the NLP model training time, from months to days, and keep models up to date efficiently—Dataflow-as-a-Service™ for Language.

Quickly Ramp NLP Model Training From Zero to State-of-the-Art

Organizations can now rapidly train NLP applications and scale on-demand. SambaNova's Dataflow-as-a-Service for Language accelerates AI workloads and provides comprehensive services, models, and a platform—empowering you to deploy customized NLP solutions with confidence.

Dataflow-as-a-Service for Language is designed to help organizations quickly get up and running with strategic NLP initiatives. This solution enables users to overcome obstacles that hinder their organization's ability to leverage NLP for next-generation AI innovation—and competitive advantage. These challenges include:

- NLP training is time-prohibitive, and experts are hard to find and costly to hire.
- Deployment requires hard trade-offs that sacrifice accuracy for expediency—adding complexity to the NLP production pipeline.
- Exponentially growing datasets and models compound the difficulty, hindering the advancement of AI applications.

SECTION 5

Quickly Ramp NLP Model Training From Zero to State-of-the-Art

By providing an end-to-end platform for training, fine-tuning, and inference, SambaNova's Dataflow-as-a-Service for Language™ offers a flexible, subscription-based NLP solution that effectively addresses these challenges and more.

The fastest path to AI

- Accelerate AI Innovation. Use your data to deploy customized NLP solutions at best-in-class accuracy.

The easiest path to AI

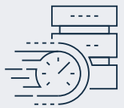
- Augment your organization's ML expertise. You don't need to be a tech giant to deploy world-class machine learning

The guaranteed path to AI

- Stay current with the latest advancements in NLP seamlessly. Get the most up-to-date NLP models and algorithmic techniques.

A Flexible New Design Architecture to Support NLP Applications

1.



A Core Enabling Platform

Optimize dataflow from algorithms to silicon using SambaNova Systems DataScale®. Build quickly and deploy next-generation AI technologies at scale using DataScale as a core enabling platform. Built on SambaNova's Reconfigurable Dataflow Architecture™ (RDA), DataScale enables organizations to achieve unparalleled model training efficiency and performance across NLP applications

2.

The Next-Generation Processor

SambaNova's Reconfigurable Dataflow Unit™—RDU—enables access to native dataflow processing and programmable acceleration. RDUs with high compute and memory capacity can process huge language models while still offering the flexibility to optimize for smaller detailed models, high accuracy, cost-efficiency, and greater scalability. Easy to use, the new RDU solutions provide NLP model owners cost-effective, democratized access to high accuracy and scalability.

3.

A Complete Software Stack

The complete software stack—SambaFlow™—is designed to take input from standard machine-learning frameworks such as PyTorch and TensorFlow. It automatically extracts, optimizes, and maps dataflow graphs onto RDUs, allowing high performance to be obtained without the need for low-level kernel tuning.

CONCLUSION

ACHIEVE BREAKTHROUGH EFFICIENCY

Achieve Breakthrough Efficiency in NLP Model Training

With optimized dataflow from the algorithms to the silicon, SambaNova is always rethinking how modern compute solutions can more efficiently and accurately process machine learning models to better support next-generation AI.

Dataflow-as-a-Service for Language provides a flexible platform solution that accelerates the entire NLP pipeline across model training, fine-tuning, distillation, and deployment. With state-of-the-art accuracy, performance, scale, and ease of use, SambaNova's NLP solution opens up a new realm of attainable AI innovation for organizations of all sizes.

DISCOVER HOW YOUR ORGANIZATION CAN DEPLOY ACCURATE AND POWERFUL NLP SOLUTIONS UP TO 18 MONTHS FASTER COMPARED TO DO-IT-YOURSELF SOLUTIONS.

Visit us on the Web at SambaNova.AI

SambaNova SYSTEMS

About SambaNova Systems

SambaNova Systems is an AI innovation company that empowers organizations to deploy best-in-class solutions for computer vision, natural language processing, recommendation, and AI for science with confidence. SambaNova's flagship offering, Dataflow-as-a-Service™, helps organizations rapidly deploy AI in days, unlocking new revenue and boosting operational efficiency. SambaNova's DataScale® is an integrated software and hardware system using Reconfigurable Dataflow Architecture™, along with open standards and user interfaces. Headquartered in Palo Alto, California, SambaNova Systems was founded in 2017 by industry luminaries, and hardware and software design experts from Sun/Oracle and Stanford University. Investors include SoftBank Vision Fund 2, funds and accounts managed by BlackRock, Intel Capital, GV, Walden International, Temasek, GIC, Redline Capital, Atlantic Bridge Ventures, Celesta, and several others. For more information please visit us at sambanova.ai or contact us at info@sambanova.ai. Follow SambaNova Systems on LinkedIn.