Datasheet

Building Transformer-Based Natural Language Processing Applications

NVIDIA Deep Learning Institute

Workshop Overview

Applications for natural language processing (NLP) and generative AI have exploded in the past decade. With the proliferation of applications like chatbots and intelligent virtual assistants, organizations are infusing their businesses with more interactive human-machine experiences. Understanding how transformer-based large language models (LLMs) can be used to manipulate, analyze, and generate text-based data is essential.

Modern pretrained LLMs can encapsulate the nuance, context, and sophistication of language, just as humans do. When fine-tuned and deployed correctly, developers can use these LLMs to build powerful NLP applications that provide natural and seamless human-computer interactions within chatbots, AI voice agents, and more.

Transformer-based LLMs, such as Bidirectional Encoder Representations from Transformers (BERT), have revolutionized NLP by offering accuracy comparable to human baselines on benchmarks like SQuAD for question answering, entity recognition, intent recognition, sentiment analysis, and more.

Learning Objectives

By participating in this workshop, you'll learn:

- How transformers are used as the basic building blocks of modern LLMs for NLP applications
- How self-supervision improves upon the transformer architecture in BERT, Megatron, and other LLM variants for superior NLP results
- How to leverage pretrained, modern LLM models to solve multiple NLP tasks such as text classification, named-entity recognition (NER), and question answering
- How to manage inference challenges and deploy refined models for live applications

Overview	
Duration	8 hours
Price	**Contact us for pricing**.
Prerequisites	> Experience with Python coding and use of library functions and parameters > Fundamental understanding of a deep learning framework such as TensorFlow, PyTorch, or Keras > Basic understanding of neural networks Suggested materials to satisfy prerequisites: **Python tutorial**, **Overview of Deep Learning Frameworks**, **PyTorch tutorial**, **Deep Learning in a Nutshell**, **Deep Learning Demystified**
Tools, libraries, and frameworks	PyTorch, pandas, NVIDIA NeMo™, NVIDIA Triton™ Inference Server
Assessment type	Skills-based coding assessments evaluate learners' ability to train deep learning models on multiple GPUs.
Certificate	Upon successful completion of the assessment, participants will receive an NVIDIA DLI certificate to recognize their subject matter competency and support professional career growth.
Hardware Requirements	Desktop or laptop computer capable of running the latest version of Chrome or Firefox. Each participant will be provided with dedicated access to a fully configured, GPU-accelerated workstation in the cloud.
Language	English

Workshop Outline	
Introduction (15 minutes)	Meet the instructor. > Create an account at **courses.nvidia.com/join**
Introduction to Transformers (120 minutes)	**Explore how the transformer architecture works in detail:** > Build the transformer architecture in PyTorch. > Calculate the self-attention matrix. > Translate English to German with a pretrained transformer model.
Break (60 minutes)	
Self-Supervision, BERT, and Beyond (120 minutes)	**Learn how to apply self-supervised transformer-based models to concrete NLP tasks using NVIDIA NeMo::** > Build a text classification project to classify abstracts. > Build a named-entity recognition (NER) project to identify disease names in text. > Improve project accuracy with domain-specific models.
Break (15 minutes)	
Maintaining Model Accuracy When Scaling to Multiple GPUs (90 minutes)	**Understand and apply key algorithmic considerations to retain accuracy when training on multiple GPUs:** > Understand what might cause accuracy to decrease when parallelizing training on multiple GPUs. > Learn and understand techniques for maintaining accuracy when scaling training to multiple GPUs.
Inference and Deployment for NLP (120 minutes)	**Learn how to deploy an NLP project for live inference on NVIDIA Triton:** > Prepare the model for deployment. > Optimize the model with NVIDIA® TensorRT™. > Deploy the model and test it.
Final Review (15 minutes)	> Review key learnings and answer questions. > Complete the assessment and earn a certificate. > Take the workshop survey. > Learn how to set up your own environment and discuss additional resources and training.

Why Choose NVIDIA Deep Learning Institute for Hands-On Training?

> Access workshops from anywhere with just your desktop/laptop and an internet connection. Each participant will have access to a fully configured, GPU-accelerated server in the cloud.

> Obtain hands-on experience with the most widely used, industry-standard software, tools, and frameworks.

> Learn to build deep learning and accelerated computing applications for industries, such as healthcare, robotics, manufacturing, accelerated computing, and more.

> Gain real-world expertise through content designed by NVIDIA and industry experts.

> Earn an NVIDIA DLI certificate to demonstrate your subject matter competency and support your career growth.

Ready to Get Started?

For the latest DLI workshops and trainings, visit
www.nvidia.com/dli

For questions, contact us at **nvdli@nvidia.com**

NVIDIA.