intel-extension-for-transformers

Public

Watch 17 Fork 85 Star 689

main 178 branches 12 tags Go to file Add file Code

lvliang-intel Fix latest chromadb version issue (#647) 2e4e639 1 hour ago 1,282 commits

.github	Fix llm runtime ir4 multi-turn inference issue (#642)	4 hours ago
conda_meta	update conda-meta version	2 months ago
docker	update docker readme (#641)	yesterday
docs	Typos fix in multiple docs files (#598)	last week
examples	add generation inference script for deepspeed on XPU (#634)	2 hours ago
intel_extension_for_transformers	Fix latest chromadb version issue (#647)	1 hour ago
tests	[Infra] combine docker codegen and chatbot (#469)	yesterday
workflows	[LLM] support dpo on habana/gaudi. (#576)	5 days ago
.dockerignore	Add Containers and k8s Support (#998)	4 months ago
.gitignore	Upgrade Qbits backend (#294)	2 months ago
.gitmodules	Python api for cpp model (#252)	2 months ago
.pre-commit-config.yaml	[pre-commit.ci] pre-commit autoupdate (#429)	last month
LICENSE	add perf PDT reviewed data (#372)	last year
MANIFEST.in	Update python packaging (#967)	5 months ago
README.md	fix spelling (#645)	5 hours ago
SECURITY.md	add perf PDT reviewed data (#372)	last year
requirements.txt	[NeuralChat] Integrate photoai backend into restful API (#478)	last week
setup.py	fix qbits backend get wrong workspace malloc size when using Relea…	2 weeks ago
third_party_programs.txt	move neural engine to deprecated (#206)	2 months ago

README.md

Intel® Extension for Transformers

An Innovative Transformer-based Toolkit to Accelerate GenAI/LLM Everywhere

intel-extension-for-transformers Release "1.0"

🏛️Architecture | 💬NeuralChat | 🤗Inference | 🗃️Examples | 💻Documentations

🚀Latest News

- [2023/11] Published a **4-bit chatbot demo** (based on NeuralChat) available on Intel Hugging Face Space. Welcome to have a try! To setup the demo locally, please follow the instructions.
- [2023/11] Released Fast, accurate, and infinite LLM inference with improved StreamingLLM on Intel CPUs!
- [2023/11] Our paper Efficient LLM Inference on CPUs has been accepted by **NeurIPS'23** on Efficient Natural Language and Speech Processing. Thanks to all the collaborators!
- [2023/10] LLM runtime, an Intel-optimized GGML compatible runtime, demonstrates **up to 15x performance gain in 1st token generation and 1.5x in other token generation** over the default llama.cpp.
- [2023/10] LLM runtime now supports LLM inference with **infinite-length inputs up to 4 million tokens**, inspired from StreamingLLM.
- [2023/09] NeuralChat has been showcased in Intel Innovation'23 Keynote and Google Cloud Next'23 to demonstrate GenAI/LLM capabilities on Intel Xeon Scalable Processors.
- [2023/08] NeuralChat supports **custom chatbot development and deployment within minutes** on broad Intel HWs such as Xeon Scalable Processors, Gaudi2, Xeon CPU Max Series, Data Center GPU Max Series, Arc Series, and Core Processors. Check out Notebooks.
- [2023/07] LLM runtime extends Hugging Face Transformers API to provide seamless low precision inference for popular LLMs, supporting low precision data types such as INT3/INT4/FP4/NF4/INT5/INT8/FP8.

🏃Installation

Quick Install from Pypi

```
pip install intel-extension-for-transformers
```

For more installation methods, please refer to Installation Page

🌟Introduction

Intel® Extension for Transformers is an innovative toolkit to accelerate Transformer-based models on Intel platforms, in particular, effective on 4th Intel Xeon Scalable processor Sapphire Rapids (codenamed Sapphire Rapids). The toolkit provides the below key features and examples:

- Seamless user experience of model compressions on Transformer-based models by extending Hugging Face transformers APIs and leveraging Intel® Neural Compressor
- Advanced software optimizations and unique compression-aware runtime (released with NeurIPS 2022's paper Fast Distilbert on CPUs and QuaLA-MiniLM: a Quantized Length Adaptive MiniLM, and NeurIPS 2021's paper Prune Once for All: Sparse Pre-Trained Language Models)
- Optimized Transformer-based model packages such as Stable Diffusion, GPT-J-6B, GPT-NEOX, BLOOM-176B, T5, Flan-T5, and end-to-end workflows such as SetFit-based text classification and document level sentiment analysis (DLSA)
- NeuralChat, a customizable chatbot framework to create your own chatbot within minutes by leveraging a rich set of plugins Knowledge Retrieval, Speech Interaction, Query Caching, and Security Guardrail.
- Inference of Large Language Model (LLM) in pure C/C++ with weight-only quantization kernels, supporting GPT-NEOX, LLAMA, MPT, FALCON, BLOOM-7B, OPT, ChatGLM2-6B, GPT-J-6B, and Dolly-v2-3B

🌱Getting Started

Below is the sample code to enable the chatbot. See more examples.

Chatbot

```
# pip install intel-extension-for-transformers
from intel_extension_for_transformers.neural_chat import build_chatbot
chatbot = build_chatbot()
response = chatbot.predict("Tell me about Intel Xeon Scalable Processors.")
```

Below is the sample code to enable weight-only INT4/INT8 inference. See more examples.

INT4 Inference

```
from transformers import AutoTokenizer, TextStreamer
from intel_extension_for_transformers.transformers import AutoModelForCausalLM, WeightOnlyQuantCo
model_name = "Intel/neural-chat-7b-v1-1"     # Hugging Face model_id or local model
config = WeightOnlyQuantConfig(compute_dtype="int8", weight_dtype="int4")
prompt = "Once upon a time, there existed a little girl,"

tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
inputs = tokenizer(prompt, return_tensors="pt").input_ids
streamer = TextStreamer(tokenizer)

model = AutoModelForCausalLM.from_pretrained(model_name, quantization_config=config)
outputs = model.generate(inputs, streamer=streamer, max_new_tokens=300)
```

INT8 Inference

```
from transformers import AutoTokenizer, TextStreamer
from intel_extension_for_transformers.transformers import AutoModelForCausalLM, WeightOnlyQuantCo
model_name = "Intel/neural-chat-7b-v1-1"     # Hugging Face model_id or local model
config = WeightOnlyQuantConfig(compute_dtype="bf16", weight_dtype="int8")
prompt = "Once upon a time, there existed a little girl,"

tokenizer = AutoTokenizer.from_pretrained(model_name, trust_remote_code=True)
inputs = tokenizer(prompt, return_tensors="pt").input_ids
streamer = TextStreamer(tokenizer)

model = AutoModelForCausalLM.from_pretrained(model_name, quantization_config=config)
outputs = model.generate(inputs, streamer=streamer, max_new_tokens=300)
```

🎯Validated Models

You can access the latest int4 performance and accuracy at int4 blog.

Additionally, we are preparing to introduce Baichuan, Mistral, and other models into LLM Runtime (Intel Optimized llamacpp). For comprehensive accuracy and performance data, though not the most up-to-date, please refer to the Release data.

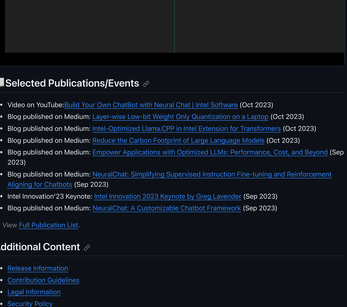
📖Documentation

OVERVIEW		
NeuralChat		LLM Runtime
NEURALCHAT		
Chatbot on Intel CPU	Chatbot on Intel GPU	Chatbot on Gaudi
Chatbot on Client		More Notebooks
LLM RUNTIME		
LLM Runtime	Streaming LLM	Low Precision Kernels	Tensor Parallelism
LLM COMPRESSION		
SmoothQuant (INT8)	Weight-only Quantization (INT4/FP4/NF4/INT8)	QLoRA on CPU
GENERAL COMPRESSION		
Quantization	Pruning	Distillation	Orchestration
Neural Architecture Search	Export	Metrics	Objectives
Pipeline	Length Adaptive	Early Exit	Data Augmentation
TUTORIALS & RESULTS		
Tutorials	LLM List	General Model List	Model Performance

🙌Demo

- Infinite inference (up to 4M tokens)

streamingLLM_v2.mp4



📃Selected Publications/Events

- Video on YouTube:Build Your Own ChatBot with Neural Chat | Intel Software (Oct 2023)
- Blog published on Medium: Layer-wise Low-bit Weight Only Quantization on a Laptop (Oct 2023)
- Blog published on Medium: Intel-Optimized Llama.CPP is Intel Extension for Transformers (Oct 2023)
- Blog published on Medium: Reduce the Carbon Footprint of Large Language Models (Oct 2023)
- Blog published on Medium: Empower Applications with Optimized LLMs: Performance, Cost, and Beyond (Sep 2023)
- Blog published on Medium: NeuralChat: Simplifying Supervised Instruction Fine-tuning and Reinforcement Aligning for Chatbots (Sep 2023)
- Intel Innovation'23 Keynote: Intel Innovation 2023 Keynote by Greg Lavender (Sep 2023)
- Blog published on Medium: NeuralChat: A Customizable Chatbot Framework (Sep 2023)

View Full Publication List.

Additional Content

- Release Information
- Contribution Guidelines
- Legal Information
- Security Policy
- Apache License

Acknowledgements

- Excellent open-source projects: bitsandbytes, FastChat, fastRAG, ggml, gptq, llama.cpp, lm-evaution-harness, peft, tpp, streamingllm and many others.
- Thanks to all the contributors.

🤷Collaborations

Welcome to raise any interesting ideas on model compression techniques and LLM-based chatbot development! Feel free to reach us, and we look forward to our collaborations on Intel Extension for Transformers!