低ビットの量子化されたオープン LLM リーダーボード

クライアント向けに高品質のモデルを見つける新しいツール

Kaokao Lv、Wenjiao Yue、Wenhua Cheng、Jun Lin、Hanwen Chang、Tai Huang、Haihao Shen
インテル コーポレーション

Hugging Face* にはすでにリーダーボード（英語）がありますが、なぜ新しいリーダーボードを作成したのでしょうか？ それは、量子化の結果の比較は簡単ではないからです。主な問題は、ほとんどの量子化モデルで精度の結果が不足していることで、もう 1 つの問題は、Hugging Face* リーダーボードで特定のモデル名を検索すると多数のモデルが見つかることです。そのため、それらがマージされているか、ファインチューニングされているか、FP16 で量子化されているか、混合されているかを判断するには、多くの手作業が必要です。これは、モデルのデプロイメントに課題をもたらします。

そこで、量子化 LLM モデルに注目し、検索エンジンを強化することでこれらの問題に対処する、低精度量子化リーダーボード（英語）を開発しました。ユーザーは、アルゴリズム（AutoRound（英語）、GPTQ（英語）、AWQ（英語）、BitsAndBytes（英語）、および GGUF（英語））、計算のデータ型（int8、fp16、bf16 など）、重みのデータ型（fp4、int4、nf4 など）、モデルサイズ、二重量子化が有効かどうかで、量子化 LLM を即座に検索できます。低ビットの量子化されたオープン LLM リーダーボードは、特定のクライアントに効率良くデプロイできる高品質のモデルを見つけられる貴重なツールです。

量子化アプローチ

さまざまな量子化方法と重みおよび計算のデータ型にわたって LLM モデルのベンチマークを効果的に行うには、堅牢な量子化ツールが必要です。インテルのリーダーボードは、LLM 量子化サポートに Transformers 向けインテル® エクステンション（英語）を利用しています。このソリューションは、GPTQ（英語）や AWQ（英語）などのよく知られている重みのみの量子化方法をシームレスに統合するインターフェイスを備えた Transformers のような API を提供します。さらに、このツールには、低ビット LLM 推論用のインテルの AutoRound 量子化アルゴリズム（英語）が組み込まれています。Transformers 向けインテル® エクステンションの量子化機能は、オープンソースのモデル圧縮ツールであるインテル® ニューラル・コンプレッサー（英語）をベースに構築されています。

低ビットの量子化されたオープン LLM リーダーボード

このリーダーボードには 10 の異なるベンチマーク（ARC-c、ARC-e、Boolq、HellaSwag、Lambada_openai、MMLU、Openbookqa、Piqa、Truthfulqa_mc1、Winogrande）が含まれています。ランキングは、これらのベンチマークの平均スコアによって決定され、再ランキング時に特定のベンチマークを優先するオプションがあります。

私たちの評価では、AutoRound は一般的なさまざまなモデルで準可逆圧縮に近づいており、GPTQ や AWQ などのほかの方法よりも一貫して優れており、GGUF よりも優れた精度を示しています。Llama* 2 7b-chat と Mistral*-7b-instruct の精度を**図 1** と**図 2** に示します。

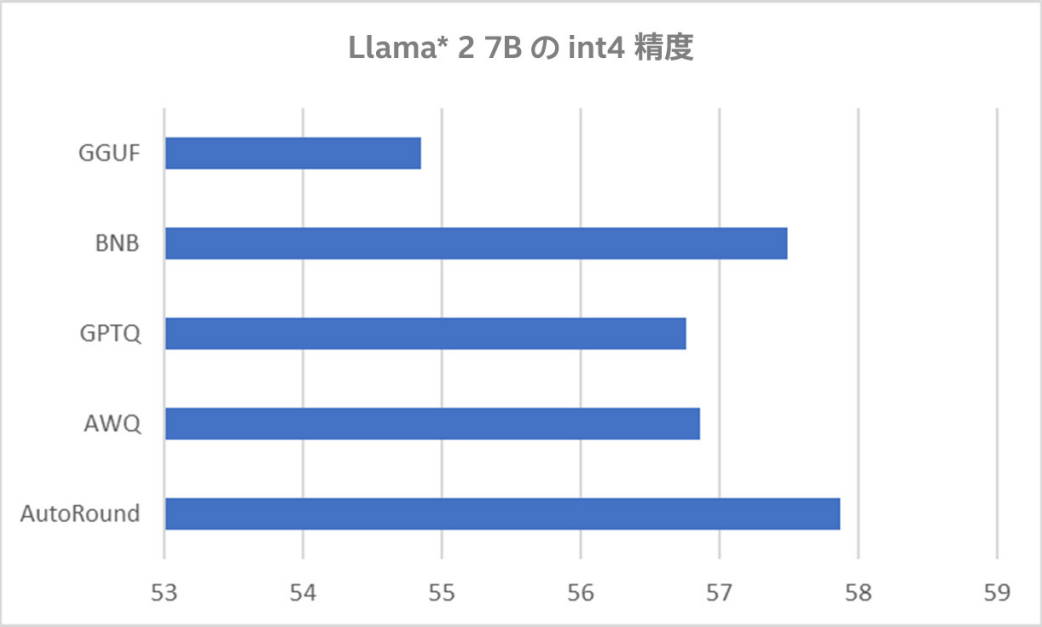

図 1. int4 Llama* 2 7B-chat の平均精度（値が大きいほうが良い）

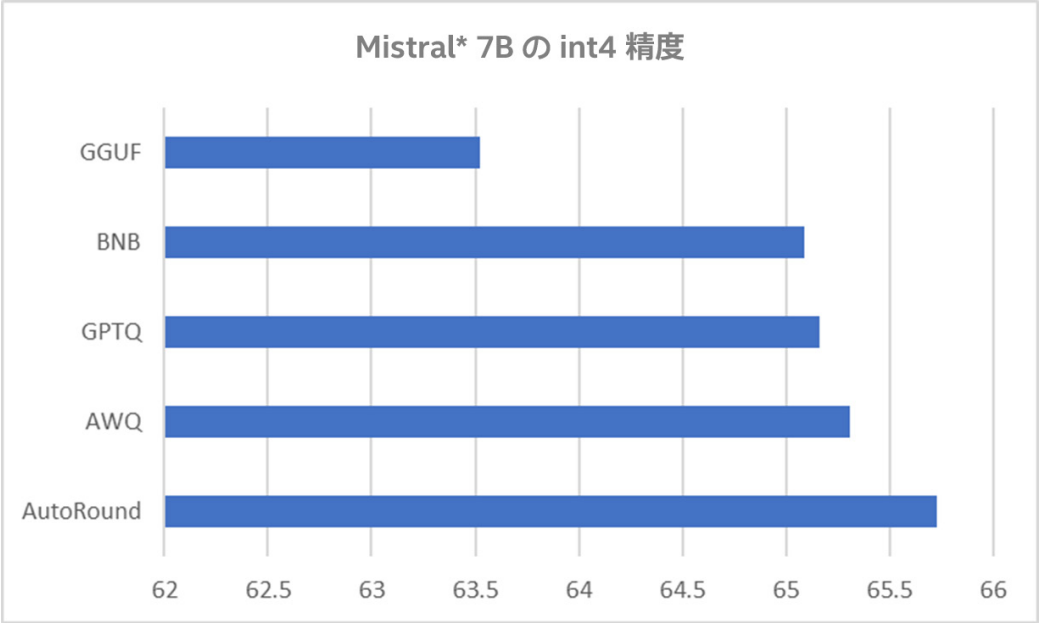

図 2. int4 Mistral*-7B-instruct-v0.2 の平均精度（値が大きいほうが良い）

また、13B AutoRound モデルと fp16 7B モデルを比較したところ、AutoRound はすべてのメトリックと平均で一貫して fp16 7B を上回っていることが分かりました（**表 1**）。これにより、低ビットの量子化された中規模 LLM を利用することで、モデルサイズが小さくても、半精度の小規模 LLM よりも優れたパフォーマンスを発揮できます。

	Average	Arc-c	Arc-e	boolq	lambada	mmlu	openbookqa	piqa	truthfulqa	winogrande
Baichuan2-7B-Chat	57.1	41.64	72.72	79.45	67.63	50.82	30.4	74.48	31.21	68.98
Baichuan2-13B-Chat AutoRound int4	60.46	47.35	74.92	82.11	71.32	55.78	31.2	75.84	36.35	73.01

表 1. fp16 7B モデルと 13B AutoRound モデルの精度（値が大きいほうが良い）

さらに、AutoRound は一般的なモデルのほとんどに対応できますが、ほかの量子化アルゴリズムには制限があります。これが、AutoRound モデルが利用しやすい理由です（**表 2**）。

T	Model	Average	ARC-c	ARC-e	Boolq	HellaSwag	Lambada	MM
	Intel/SOLAR-10.7B-Instruct-v1.0-int4-inc	68.49	60.49	82.66	88.29	68.29	73.36	62
	TheBloke/SOLAR-10.7B-Instruct-v1.0-GPTQ	68.3	60.92	83.33	87.89	67.65	72.95	62
	TheBloke/SOLAR-10.7B-Instruct-v1.0-AWQ	68.19	59.81	83.08	87.98	68.06	72.79	62
	TheBloke/SOLAR-10.7B-Instruct-v1.0-GGUF	66.6	60.41	83.38	88.29	67.73	52.42	62
	Intel/Mistral-7B-Instruct-v0.2-int4-inc	65.73	55.38	81.44	85.26	65.67	70.89	58
	TheBloke/Mistral-7B-Instruct-v0.2-AWQ	65.31	53.75	80.43	85.11	65.59	70.99	58
	unsloth/mistral-7b-instruct-v0.2-bnb-4bit	65.09	54.44	81.99	85.14	65.56	71.36	58
	Intel/Phi-3-mini-4k-instruct-int4-inc	65.09	57.08	83.33	86.18	59.45	68.14	66
	leliuga/Phi-3-mini-4k-instruct-bnb-4bit	64.66	56.91	83.08	86.02	59.71	67.46	66
	kaitchup/Phi-3-mini-4k-instruct-gptq-4bit	64.2	55.2	81.78	85.5	59.36	66.97	66
	TheBloke/Mistral-7B-Instruct-v0.2-GGUF	63.52	53.5	77.9	85.44	66.9	50.11	58

表 2. 低ビットの量子化されたオープン LLM リーダーボード（2024 年 5 月 11 日現在）

AutoRound は、AWQ などのアルゴリズムにはない lm_head の量子化もサポートしています。GPTQ や AWQ とは異なり、AutoRound は新しいモデルを自動的に受け入れることができます。最後に、AutoRound はさまざまなデータセットにわたるキャリブレーション機能を提供します。

量子化サンプルコード

次のサンプルコードは、Transformers 向けインテル® エクステンション（英語）の Transformer のような API を使用して AutoRound を適用し、LLM を量子化する方法を示します。

```
from transformers import AutoTokenizer
from intel_extension_for_transformers.transformers import AutoModelForCausalLM, AutoRoundConfig

model_name_or_path = "Intel/neural-chat-7b-v3-3"
prompt = "Once upon a time, a little girl"
tokenizer = AutoTokenizer.from_pretrained(model_name_or_path, trust_remote_code=True)
inputs = tokenizer(prompt, return_tensors="pt").input_ids

q_config = AutoRoundConfig(bits=4, tokenizer=tokenizer)
int4_model = AutoModelForCausalLM.from_pretrained(
                                        model_name_or_path,
                                        quantization_config=q_config,
                                         )
output = int4_model.generate(inputs, max_new_tokens=100, do_sample=True)
```

詳細については、Transformers 向けインテル® エクステンションの AutoRound サンプル（英語）を参照してください。

コラボレーションと今後の取り組み

ぜひ、リーダーボードを試して、量子化モデルをアップロードしてください。皆さんからのフィードバック、質問、コメントをお待ちしています。独自のモデルで試してみたい場合は、チケットを作成して、インテルのチームからどのようなサポートが得られるかをご確認ください。

将来的には、リーダーボードを拡張して、超低ビットの量子化されたオープン LLM をサポートする予定です。このトピックに興味がある場合は、お気軽にメールでお問い合わせください。繰り返しになりますが、皆さんからのフィードバックと貢献をお待ちしています。