SambaNova SYSTEMS

SambaNovaによるLLM-jpモデルのfine-tuningに関する取り組み

◯中野 匡彦，Reggie Lu，長尾 太介

2025-03-14 @NLP2025「大規模言語モデルのファインチューニング技術と評価」ワークショップ

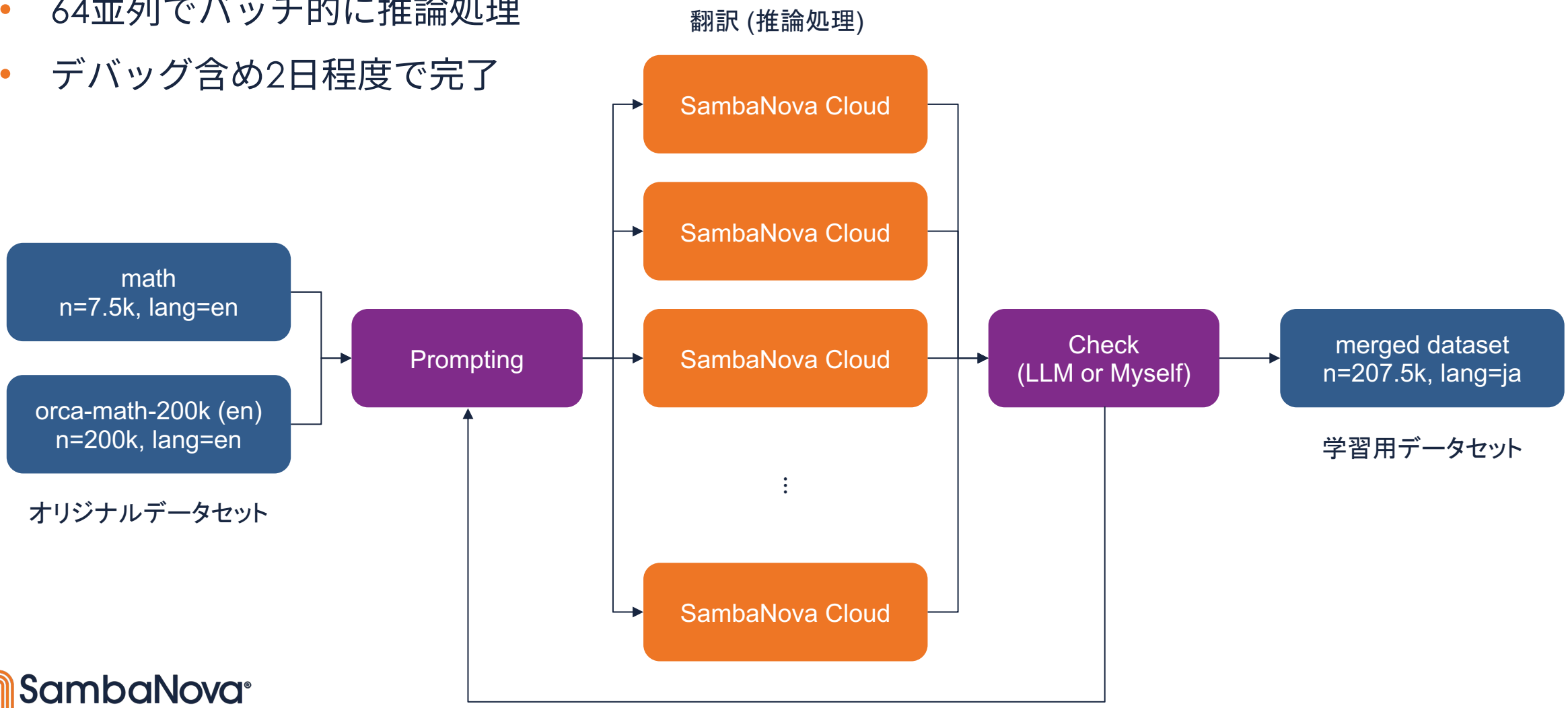
FTコンペ (数学タスク) における本チームのアプローチ

- 学習手法: シンプルなsupervised fine-tuning (SFT) を採用
- 数学系 (CoT) オープンデータセットをLLMで翻訳し、高品質な日本語データセットを構築
  - データセット
    - hendrycks/math
    - microsoft/orca-math-word-problems-200k
  - モデル
    - DeepSeek-R1-Distill-Llama-70B
    - Llama-3.1-Tulu-3-405B

SambaNova独自の超高速推論API を活用し、大規模データセットを効率的に構築

データ生成フロー

- 64並列でバッチ的に推論処理
- デバッグ含め2日程度で完了



RDU — 生成AIの学習・推論に最適なアーキテクチャ

RDU:
Reconfigurable Dataflow Unit
(再構成可能データフローユニット)

x 16

“SN40L” RDU

- **TSMC 5nm テクノロジー**
  - 1026億 トランジスタ
  - 1,040 RDUコア
  - 638 TFLOPS (bf16)
- **3層 データフローメモリ**
  - 520 MB オンチップメモリ (SRAM)
  - 64 GB 広帯域メモリ (HBM3)
  - 1.5 TB 大容量メモリ (DDR5)

Single system: **SN40L-16**

On-Chip SRAM [8.3 GB, PBs per sec]

データフローを大容量オンチップメモリで実現

25.6 TB/s

RDU 広帯域 HBM メモリ [1 TB]

超低レイテンシのモデル切り替え

1.6 TB/s

RDU 大容量 DDR メモリ [24 TB]

最大12兆個のパラメータ！

SambaNova SYSTEMS

SambaNova による超高速な推論パフォーマンス

Artificial Analysis社による推論ベンチマーク結果 (単位: token/sec/user)

	SambaNova	NVIDIA
Llama 3.2 1B 16-bit	2477	304
Llama 3.1 8B 16-bit	1066	93
Llama 3.1 70B 16-bit	460	32
Llama 3.1 405B 16-bit	200	14

GPUに対して10倍以上高速なトークン処理速度を実現

SambaNova SYSTEMS

コンペ振り返り - やりたかったこと (やりきれなかったこと)

- Training手法はもう少しバリエーションを試したかった。 (DPO含め)
- 嵩増ししたデータセット (orca-math-word-problems-200k) の問題種別が、必ずしも評価セット (math) と一致していなかった。評価セットの問題の内容や難易度とよりアラインしたデータセットを作成したかった。
- 評価セット (math) には、一部かなり高度な問題が含まれていた。(proprietaryモデルでもほとんど正解できないくらいの難易度。) コンペのスコアを上げるという観点では、平易な問題のみに着目したアプローチが取れたと思う。
- DeepSeek-R1-671Bなど、Reasoningモデルを露に活用することで、数学CoTデータセットを作成したかった。

SambaNovaの超高速推論は
データ生成にも有効です。

ぜひお試しください！

cloud.sambanova.ai

SambaNova SYSTEMS