TURING

HPC ToolkitによるSlurm環境構築と日本語LLM/VLM開発

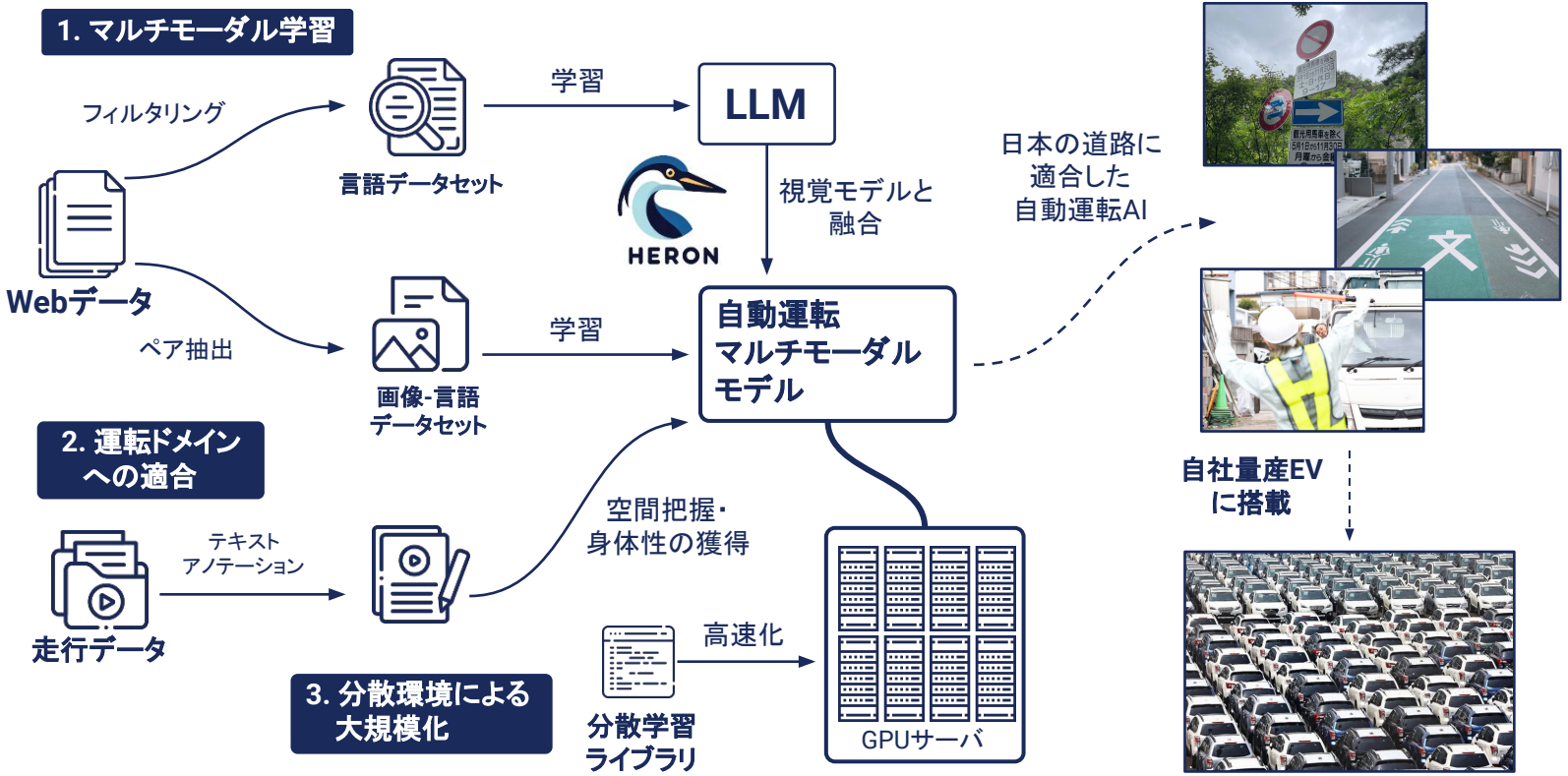
GENIAC 中間報告会 / Turing株式会社

山口 祐

2024/05/21

マルチモーダルAI x 自動運転

TURING



TURING

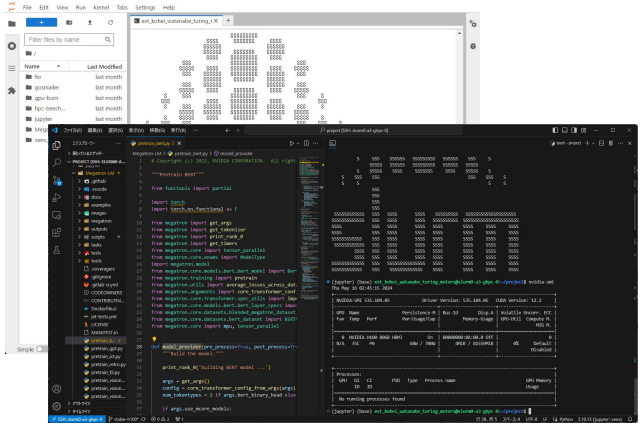
GCPでの環境構築

GCP環境構成概要 (HPC-toolkit構成)

TURING

複数人のリサーチャーが複数テーマやタスクの実行する環境ではジョブスケジューラーなどによるよる**公平なリソース管理が必要**

- **すべてのリソースはジョブスケジューラー(Slurm)管理**
  - **1GPU単位からマルチノードのリソース割り当て管理**を実現
  - Enroot/PyxisによるdockerをHPC環境で利用が可能(≠Singurarity)
- **共有ストレージ(詳細後述)**
  - /home,/dataにFilestoreを利用(現時点)
  - 大規模分散並列学習向けにLustreストレージへ移行予定
- **shell,vscode,jupyter-labでのインタラクティブ利用が可能**
  - 確保リソース上に独立したsshデーモンを起動することで実現



スケジューリングされたリソースへの
vscode jupyter-lab等での利用(GPUx1割当)

なぜ HPC-toolokit(Slurm)環境なのか

TURING

GCPにて提供されるblueprintを利用することでGCP環境に**最適化した大規模計算環境を構成**できる

- **リサーチャーの利用負担(利用習熟が少ない)環境**
  - ubuntu22.04ベースのカスタムイメージが利用(dlvm+TCPX)
  - ABCIやTSUBAME等の国内の計算基盤と同等の利用形態を提供
- **自社保有計算環境に向けた事前習熟ができる**
  - 自社保有計算環境も同一(Slurm管理)を計画
  - 利用形態や必要な環境資源を事前に洗い出せる
- **GCPのAPIを利用したノード制御機能が利用可能**
  - CPU計算ノードなどをオンデマンドでdeployできる(submitトリガで構成)

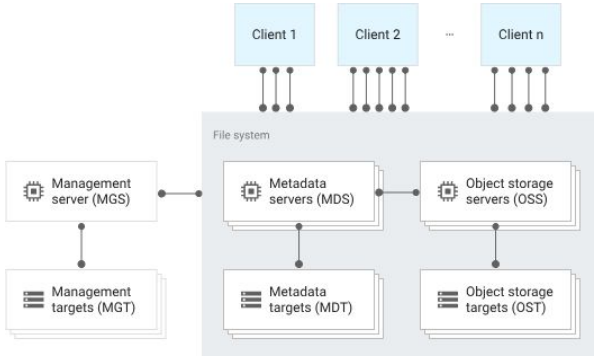
独自に各種メトリックを収集

共有Storage について

TURING

現状はGCP Filestoreを利用し、大規模ノード環境利用開始(6月中旬)からは**DIYのOpenLustre環境を利用**

- **性能の重要性**
  - **checkpoint保存スループットが重要**
  - 4時間間隔でcheckpoint保存時間20分の場合8%の計算時間ロス
- **Open Lustre環境の検証** (参考:GCP Codelab)
  - 10GB/s以上の書き込み性能は確認,最大100GB/sの性能を予定
  - ノード(MDS/OSS)スケールアウトにて容量/性能スケールが容易
- **GCPのマネージドメニューの利用について**
  - 以下のメニューは本PJで利用不可のためDIYで構成
  - DDN LustreやParallelstore(IntenDAOS)



Lustreの基本的なアーキテクチャ(GCP DDN Lustre)

その他 考慮事項

TURING

GCP環境においては以下の考慮と対応が必要となります。詳細についてはSlackにてご連絡いただければ個別に詳細を共有します。

- **GCP定期メンテナンスへの対応**
  - GPUインスタンスは停止を伴うメンテナンスが連続起動時間が2週間おきに実施される(GPUホストイベントの処理)
  - Slurmにてメンテナンス時間を予約し、ユーザー任意のタイミングで再起動することで回避
- **DIY Lustre環境について**
  - 参考に記載したGCPのCodelabの内容が古くコードの修正が必要
  - 最新のz3インスタンスを利用することで大幅なパフォーマンス向上見込み

TURING

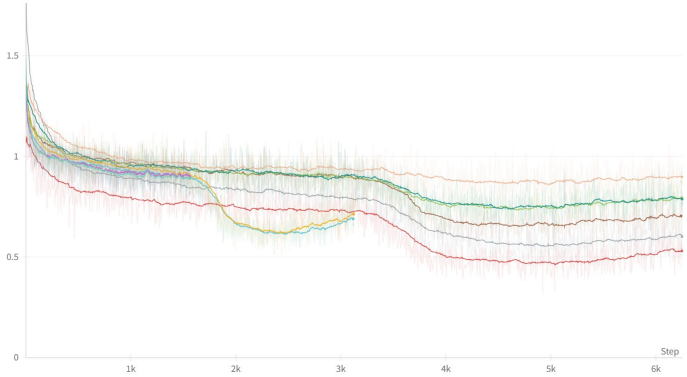
日本語LLM/VLMの学習

日本語LLMの指示チューニングの検証

TURING

前半期間は少数ノードのため、小規模モデルの事後学習を中心に検証。

- **ベースモデル**
  - Swallow-MS-7b (Mistral-MS-7b)
  - Llama3-8b（学習中）
- **データセット**
  - Google Research のFlanデータセットに回答を付与したものをベース（英語）
  - 5万件を機械翻訳で日本語の指示チューニングデータに（100M token程度）

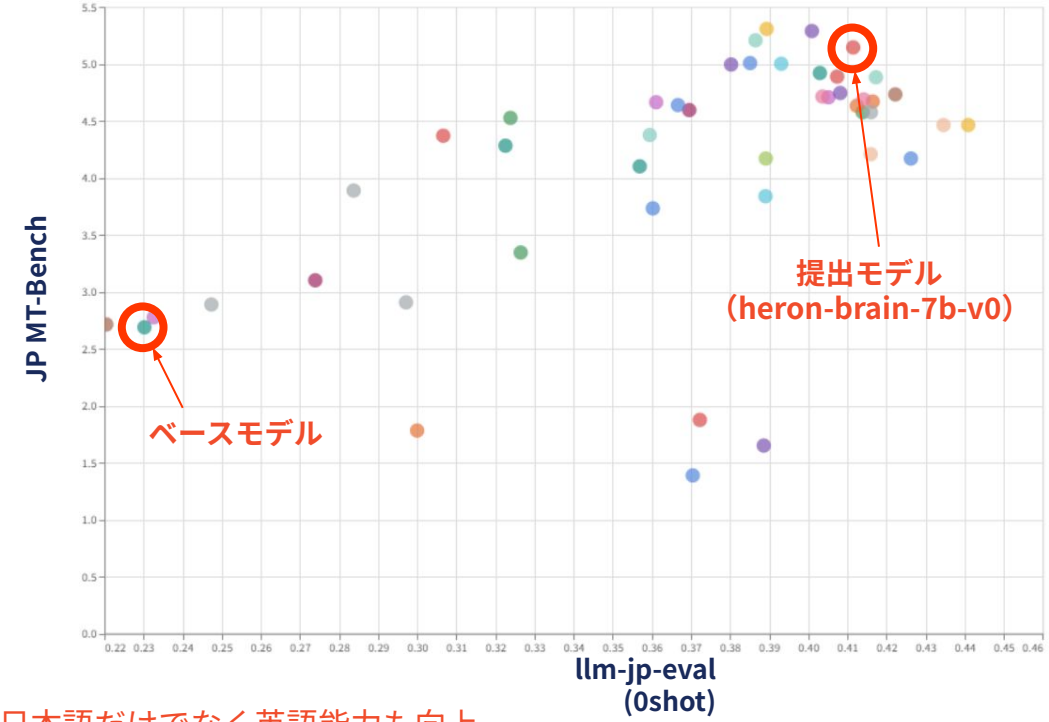


日本語LLMの指示チューニングの検証

TURING

- **日本語LLMの初期モデル（heron-brain-7b）を構築**
  - llama-recipes (pytorch-fsdp) を利用
- **ハイパーパラメータ探索**
  - データセット
  - 学習率・WeightDecay
  - バッチサイズ
  - NEFTune
  - …

→ 報告されている7Bモデルでも最高レベルの日本語スコア（llm-jp-eval/MT-Bench）



日本語だけでなく英語能力も向上

model	AVG_Overall	AVG_jaster_ja_0sh	AVG_jaster_ja_4sh	AVG_jaster_en_0s	AVG_jaster_en_4s	AVG_MTbench_ja	AVG_MTbench_en
heron-brain-7B	**0.53**	**0.43**	**0.62**	**0.39**	**0.51**	**5.57**	**5.97**
Swallow-MS-7B	0.41	0.27	0.62	0.21	0.42	4.29	4.59

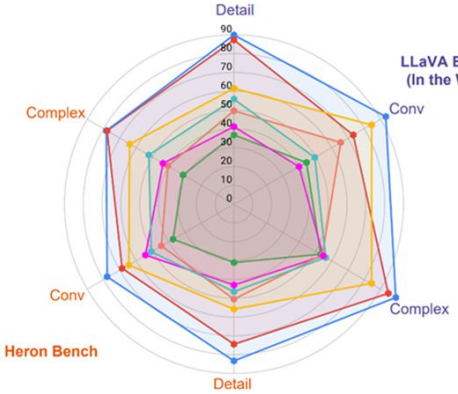
日本語VLMの評価: Heron-Bench

TURING

Heron-Bench (detail): Sports

User	力士たちは何を行っているか説明してください。
Heron GIT	画像では、2人の力士のパフォーマンスが行われており、1人はリングの上で相撲の技を披露し、もう1人の力士はリングの近くで太鼓を叩いています。このイベントは、相撲と太鼓のパフォーマンスを組み合わせたもので、観客を楽しませ、魅了するエンターテイメントとして機能します。
GPT-4V	この写真は、日本の伝統的なスポーツである相撲の試合を捉えたものです。力士たちは土俵の上で対戦しており、周囲には観客が詰めかけています。写真の中央にいる二人の力士は、試合の一環として、おそらく立ち合い直前の儀式を行っているところです。相撲では、試合前に塩をまいたり、四股を踏んだりするなど、様々な儀式が行われます。これらの儀式は、試合の公正を保ち、悪霊を払うという意味合いがあります。また、力士たちは伝統的な相撲の衣装であるまわしを身につけています。
Claude 3 Opus	画像では、力士が相撲を取っている様子が見られます。土俵の上で2人の力士が対戦しており、周りには大勢の観客が見守っています。力士は伝統的な装束である廻し姿で、激しく組み合いながら勝負を競っている最中のようです。会場の装飾からして、本場所のような大きな大会の様子だと思われます。多くの観客が見守る中、熱戦が繰り広げられているのがよくわかる1枚です。

- **日本語の視覚-言語モデルの評価ベンチマークを公開**
  - 日本特有の画像や文脈理解の質問
- **多面的な評価**
  - GPT-4VやClaude3などのマルチモーダルモデルで幅広く評価
  - WandB社と連携してリーダーボード化も検討中



Method	LLaVA Bench	Heron Bench
GPT-4V	94.05	79.70
Claude 3 Opus	85.13	73.56
Gemini Pro	76.69	61.32
LLaVA1.6 7B	44.92	33.05
Qwen-VL 7B	54.01	49.72
EvoVLM-JP	56.43	45.07
Heron GIT	45.15	46.83

今後の開発＆メッセージ

TURING

35ノードで大規模VLM学習、自動運転への適合を目指す

- **日本語LLM/VLMの学習**
  - より大規模なモデルの学習
- **運転ドメインへの適合**
  - 走行動画の独自データセット作成
  - 運転知識の学習

収集した走行データセット

- **今後開始される事業者の方へ**
  - とにかく継続的に実験を回せる堅牢な学習環境が大事
  - モデル・データは簡単にスケールしないので、まずは小さく始めて試行錯誤するのが近道

We Overtake Tesla