AI研究WG EdgeTech+ 2025
組込AIの現状と
組込生成AIの可能性

2025年11月19日

応用技術調査委員会 AI研究WG
中村 仁昭

一般社団法人
組込みシステム技術協会
Japan Embedded Systems Technology Association

自己紹介

- 株式会社Bee CTO 中村仁昭
  - 新大阪の組込ソフト会社です
  - JASAパビリオンで「ネットが無くても生成AI」を展示してます
- CQ出版のInterfaceで記事を書いてます
  - 2024年9月号(7/25発売)の「OpenCVで体験　現場プロの画像処理50」特集でラズパイ+Python
  - 2024年5月号の「ラズベリーパイ5大研究」特集で生成AI(LLM)

オンデバイス・インテリジェンスが拓く未来と、直面する課題

組込みAIの現状と将来展望

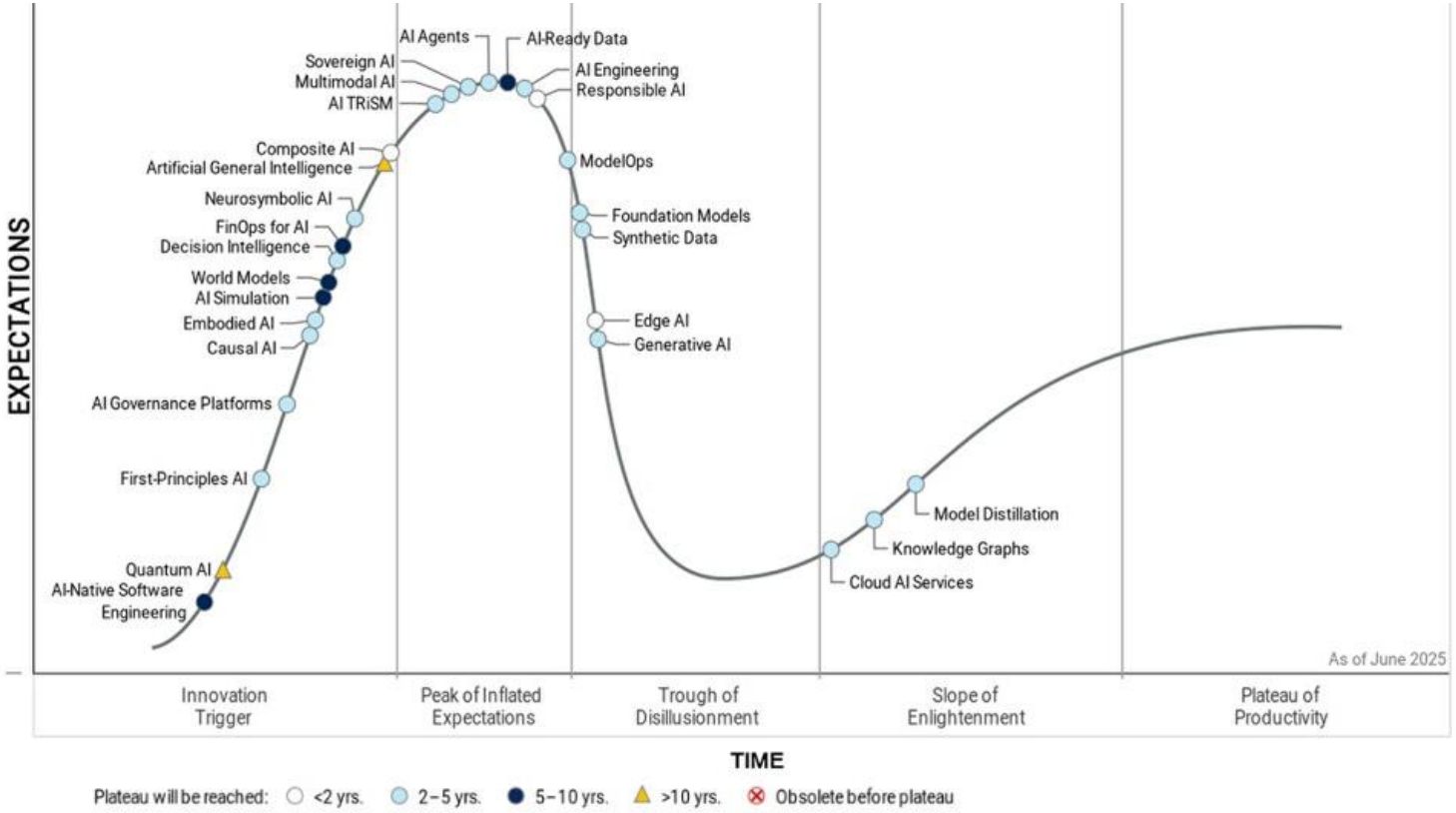
一般社団法人
組込みシステム技術協会
Japan Embedded Systems Technology Association

アジェンダ

1. 組込AIの位置付け
2. 市場概況：なぜ今、組込みAIか？
3. 主要技術トレンド：オンデバイス生成AI と TinyML
4. ハードウェア戦略：二極化する市場と主要プレイヤー
5. ソフトウェア戦略：ベンダーロックインの回避（ONNX）
6. 主要な応用分野（産業・自動車・ヘルスケア）
7. 直面する重要課題（セキュリティと開発ハードル）
8. 未来展望と戦略的提言

組込AI(EdgeAI)の位置付け

- GartnerのHype Cycle for Artificial Intelligence, 2025では引き続き幻滅期(2023～)のままで主流の採用までに時間がかかっているように見える



https://www.gartner.com/en/newsroom/press-releases/2025-08-05-gartner-hype-cycle-identifies-top-ai-innovations-in-2025

市場概況：加速する組込みAI市場

- 組込みAI市場は、単なる成長期から「加速期」へ移行
- **市場規模（2024年）**: 142億8,000万米ドル
- **市場規模（2025年）**: 156億8,000万米ドル（CAGR 9.8%）
- **中期予測（2029年）**: 259億9,000万米ドルに達する見込み 。2025年からのCAGRは13.5% ないし13.8% と予測され、成長が加速

なぜ今、組込みAIか？（クラウドAIとの違い）

- 従来のクラウドAIの課題を、デバイス側（エッジ）で解決
- **1. リアルタイム性（低遅延）:**
  - ネットワーク遅延を排除し、ミリ秒単位の応答が可能
  - 工場の異常検知 や自動運転など、即時性が求められる現場で必須
- **2. プライバシーとセキュリティ:**
  - 個人の健康情報 や工場の機密映像など、センシティブなデータをデバイス内で完結して処理
  - データ漏洩のリスクを根本的に低減
- **3. コストと安定性:**
  - クラウドへのデータ通信コストと、通信モジュールの消費電力を大幅に削減
  - ネットワーク接続が不安定な場所や、オフライン環境でもAI機能が動作

主要技術トレンド（1）オンデバイス生成AI

2025年現在、市場を牽引する最大のトレンド

■ 概要:

- 従来クラウドが必須だった大規模言語モデル（LLM）やマルチモーダルAIを、スマートフォンやPC上で直接実行

■ 最先端の事例（Qualcomm):

- スマートフォン上でリアルタイムの「画像から動画へ」の生成AIデモを実現
- 音声、テキスト、画像を統合的に理解する「マルチモーダルAIアシスタント」をプライバシーを保護しながら提供

■ 価値:

- RAG（検索拡張生成）技術と組み合わせ、デバイス内の個人データ（メール、予定）に基づいた、真にパーソナライズされた応答をオフラインで実現

主要技術トレンド (2) TinyML

AIoT (AI + IoT) を実現する中核技術

■ 概要:

- mW(ミリワット)単位の超低消費電力で動作する安価なMCU(マイクロコントローラ)上でAI推論を実行する技術

■ 仕組み:

- 「学習はパワフルな環境(クラウド)で、推論のみを省電力なデバイスで」実行

■ 必須となる「モデル軽量化」技術:

1. **量子化**: モデルのパラメータ(例: 32bit浮動小数点)を8bit整数などに変換。モデルサイズを劇的に(例: 1/4に)圧縮
2. **プルーニング**: モデル性能への寄与が低い接続を「剪定」し、計算量を削減
3. **知識蒸留**: 高性能だが重い「教師モデル」の知識を、軽量な「生徒モデル」に継承

ハードウェア戦略：二極化する市場

■ 組込みAIハードウェア市場は、性能とコストに応じて明確に二極化

	1. ハイパフォーマンス・エッジ（MPU）	2. 超低電力エッジ（MCU）
ターゲット	オンデバイス生成AI、高度な画像認識	TinyML、シンプルなセンサーAI
性能	高TOPS（Tera Operations Per Second）	GOPS / MOPS（Giga/Mega）
消費電力	数十Wクラス	mW～数Wクラス
主要プレイヤー	Qualcomm（AI Engine） NVIDIA（Jetson） Tesla（AI5/カスタム）	STMicroelectronics（STM32N6） Renesas（RZ/V）

ハードウェア主要プレイヤーの戦略

- **Qualcomm (AI Engine):**
  - NPU、GPU、CPUを最適に使い分ける「ヘテロジニアス（異種混合）コンピューティング」戦略 。オンデバイス生成AI市場を強力にリード
- **NVIDIA (Jetson):**
  - 高性能GPUを活かし、ロボティクスや高度な交通監視システム など、画像解析分野でデファクトスタンダードの地位を確立
- **STMicroelectronics (STM32N6):**
  - MCU市場の巨人が、独自NPU「Neural-ART」を搭載した初のAIマイコンを発表 。TinyML市場の本格的な開拓を進めている
- **Tesla (AI5 チップ):**
  - 究極の「垂直統合」。自動運転アルゴリズムに最適化したAIチップを自社開発
  - 目標：NVIDIA製GPU比で「消費電力1/3、製造コスト1/10」 。性能向上と劇的なコスト削減を両立

ソフトウェア戦略：ベンダーロックインの回避

■ **課題:**

- ハードウェアベンダー各社は、開発者を囲い込むため独自のソフトウェア（例: ST Edge AI Suite , Qualcomm AI Hub ）を提供
- 一度このエコシステムに依存すると、将来、他社の優れたハードウェアへ乗り換えることが困難（＝**ベンダーロックイン**）

■ **標準フレームワーク (TensorFlow Lite):**

- Googleが提供するモバイル・組込みAIの標準ツールセット 。ただし、現時点では「推論」特化であり、「オンデバイス学習」は未サポート

■ **戦略的解答 (ONNX):**

- **Open Neural Network Exchange**。AIモデルの「共通フォーマット」
- PyTorchやTensorFlowなど、どのフレームワークで学習させても、ONNX形式に変換すれば、多様なハードウェア（Qualcomm, NVIDIA, ST等）で実行可能
- **メリット**: AIモデルという最大の開発資産をハードウェアから分離・抽象化し、ベンダーロックインを回避。将来の技術選択の自由度を確保できる

応用分野（1）産業（IIoT）と自動車

■ **産業（IIoT）/ 予知保全:**

- 工場の設備に設置されたセンサー（振動、温度、異音など）のデータを、エッジAIがリアルタイムで分析
- 従来は見逃されていた複雑なパターンから「故障の予兆」を検知し、突発的なライン停止（ダウンタイム）による莫大な損失を未然に防ぐ

■ **自動車（キャビン内センシング）:**

- Tobii社がQualcommのプラットフォーム上で、**単一のカメラ**を用いて「ドライバー監視(DMS)」と「乗員監視(OMS、子供の置き去り検知など)」を同時に実現
- インパクト: EUの安全規制強化に対応しつつ、自動車メーカーのハードウェアコストを大幅に削減（2025年より量産開始）

応用分野 (2) スマートホームとヘルスケア

■ スマートホーム (エッジAI化):

- エアコン、冷蔵庫、ロボット掃除機が、クラウドを介さずデバイス本体でAI処理を実行
- 価値: ネットワーク遅延のない快適なリアルタイム応答と、家庭内の映像やデータを外部に出さないプライバシー保護を両立

■ ヘルスケア (予防医療):

- **AIパーソナルヘルスコーチ**: ウェアラブルデバイスが収集するバイタルデータ(心電図など)と、AI家電からの食事データを連携 。AIが個人の健康状態を分析し、最適な生活習慣を提案。「治療」から「予防」へのシフトを促す
- **診断支援**: 医療画像(X線、CT)のAI解析や、最適な投薬提案により、医師の診断を支援し、ヒューマンエラー(医療ミス)を大幅に削減

課題 (1) 最大のアキレス腱：セキュリティ

組込みAIの普及における最大の障害

- **問題の背景**: 自動車や産業機器は、製品ライフサイクルが10年超と非常に長い
- **脆弱性**:
  - ネットワークへの常時接続
  - 出荷後に脆弱性が発見されても、物理的・システム的にセキュリティパッチを適用することが困難
- **現実の課題 (NVIDIA Jetsonの事例)**:
  - 台湾の高速道路（遠隔地）に設置された高性能Jetsonデバイス 。導入（PoC）は成功しても、「設置後のソフトウェア更新や管理方法がない」という深刻な運用課題に直面している
- **戦略的必須事項**:
  - 「出荷したら終わり」のビジネスモデルは完全に破綻している
  - セキュアなOTA (Over-the-Air) アップデートの仕組みと、そのための「**運用コスト (OpEx)**」を、製品の初期設計とTCO（総所有コスト）に組み込むことが必須

課題 (2) 開発ハードルと倫理

■ 技術的ハードル:

- **リアルタイム処理の壁**: デバイスの処理性能が不足し、AI処理に時間がかかり遅延が発生する 。→ ハードウェア選定と、TinyMLによる徹底的なモデル軽量化が必須

■ 組織的ハードル:

- **スキルセット不足**: 「AI」「画像処理」「組込み開発」など、従来は異なっていた専門知識の融合が必要
- **ROIの欠如**: 生成AIの専門知識が不足し 、AI導入コストを正当化する明確なビジネスケース（ROI）を描けない

■ 倫理的ハードル:

- AIが人間を「監視（DMS）」「診断（ヘルスケア）」 するようになり、その判断の透明性・公正性が厳しく問われます。AIガバナンスの確立は、法令遵守（コンプライアンス）を超えた、企業の「信頼」に関わる経営課題

未来展望：次世代コンピューティング（2020年代後半）

■ エネルギー効率の高いコンピューティング：

- ITインフラの消費電力増大に伴い、AIの「持続可能性（サステナビリティ）」が最重要トレンドの一つになっている

■ ニューロモーフィック・コンピューティング（SNN）：

- 現在のAI（ANN）とは根本的に異なる、脳の神経細胞（ニューロン）の「スパイク（発火）」を模倣したアーキテクチャ
- 「イベント駆動型」で動作するため、理論上、桁違いの電力効率を達成できる可能性があり、2020年代後半の実用化が期待される

■ 人間拡張と神経系との融合：

- 空間コンピューティング（AR/VR）や、将来的にはブレイン・マシン・インターフェース（BMI）を通じ、AIが人間の認知能力や身体能力を直接的に支援・拡張

結論：戦略的提言

1. **AIは「コスト」ではなく「戦略的投資」**
   - 価値は新機能（売上増）だけでなく、「既存コストの劇的削減」（ハードウェア費 、運用費 、人為的ミス 、NVIDIA依存コスト ）にもある
2. **ハードウェア戦略の早期決断がロードマップを左右する**
   - 「ハイパフォーマンス（Qualcomm等）」か、「超低電力（STMicro等）」か、あるいは「垂直統合（Tesla型）」か。自社の戦う領域を定義する
3. **ソフトウェア戦略（ONNX）で「ロックイン」を回避**
   - AIモデルという最大の「資産」を、ハードウェアから分離・抽象化することが、将来の技術選択の自由度を確保する「戦略的防衛策」となる
4. **セキュリティは「運用コスト」として初期設計に組み込む**
   - 「出荷したら終わり」のモデルは破綻している 。セキュアなOTA（アップデート）の仕組みと、10年単位のライフサイクルを支える「運用コスト」をTCOに計上すべき

組込生成AIの可能性

一般社団法人
組込みシステム技術協会
Japan Embedded Systems Technology Association

生成AIとは？

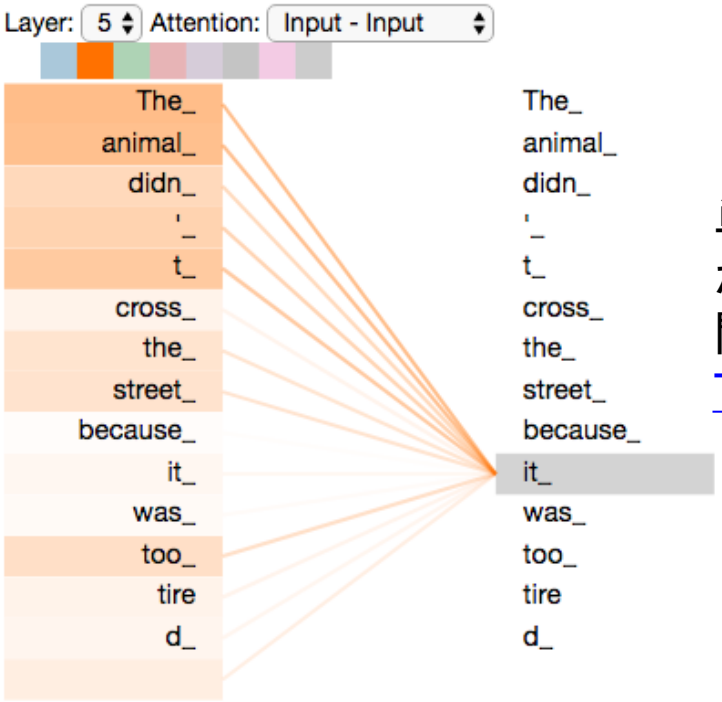
- 分類や回帰と異なり生成タスクは従来困難と言われていた
- Transformerによって言語モデルが高度な能力を獲得
  - ChatGPTなどの大規模言語モデル(LLM: Large Language Model)
  - ViT (Vision Transformer)など他タスクの応用も
- 拡散モデルによって画像、音声、動画の生成が可能に
  - Stable Diffusionなど自然な画像生成が有名
  - 化合物や制御など多くの分野で成果を出している
- それぞれの生成AIがどのように動作しているのかをみてみる

Transformer

- Transformerは2017年に発表された自然言語処理の論文”Attention is All You Need”で初めて登場したモデル
- 翻訳タスクでSeq2seq(RNNベース Encoder-Decoderモデル)より速く、精度が高い
- RNNもCNNも使わずAttentionのみ使用したEncoder-Decoderモデル
  - 並列計算が可能
- アーキテクチャのポイントは3つ
  - Encoder-Decoderモデル
  - Attention
  - 全結合層
- NLPの最近のSoTA(BERT、GPTなど)のベースとなるモデル

Attention

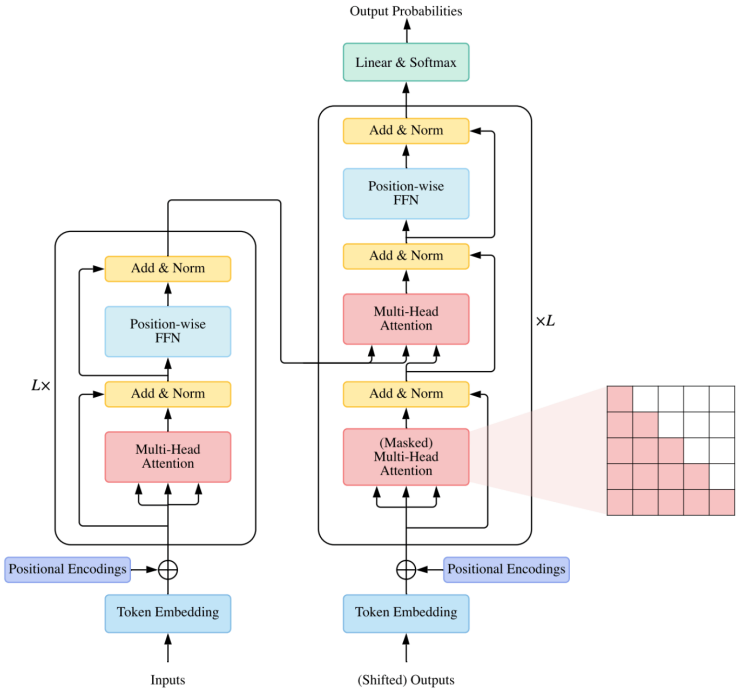
- TransformerはAttentionと呼ばれる仕組みを効率的に積層した深層学習モデル
- Attentionとはデータを検索するための鍵(Key)と実際の値(Value)のペア集合に対して、問い合わせ(Query)を投げて値を取り出す操作



単語“it”をエンコードしているときにAttention機構が”The Animal”に注目し、
関連付けている
The Illustrated Transformerより引用

Transformerの全体像

- Attention機構を積層し線形層や正規化層を適切に挟み込んだアーキテクチャとしてTransformerは提案された
- また対象入力を特徴ベクトルとして埋め込む層や、位置情報を符号化して付加する層も重要な構成要素



初期Transformerの概要
A Survey of Transformersより引用

Transformerの利活用

- 大きく3つに大別される使用法が主流に
- Encoder-Decoder
  - オリジナルのTransformerと同様の利用法。機械翻訳などある系列を異なる系列に変換するタスクに用いられる
- Encoder Only
  - 入力系列の表現学習に利用。系列分類やラベリングタスクへの転用も多い
  - BERTが代表例
- Decoder Only
  - EncoderとのCross-Attentionを除外し、自己回帰生成のデコーダ部のみを残した構造
  - ChatGPTのようなLM(Language Model：言語モデル)など、生成タスクでの利用が主

Encoder Only – BERT –

- BERTは教師なし表現学習手法で汎用的な自然言語表現能力が獲得できることを示した
  - 入力系列の一部をランダムで欠落させてその部分を予測するタスク(MLM)、2種類の結合された文章が連続するものか否かを予測するタスク(NSP)を同時に解くことで、ラベルのない大量のテキストコーパスから学習
- Decoder Onlyが注目を浴びるが検索(RAGなど)、分類(コンテンツモデレーションなど)、エンティティ抽出(プライバシーや規制遵守など)など日々発生する現実的なタスクに使用されている
- BERTは2018年に発表されたアーキテクチャだが、現代の技術で更新したModern BERTが2024年12月にリリースされるなど改善されている
- モデルサイズが小さい

Decoder Only - GPT -

- 入出力が同一のモダリティ、または異なる言語や異なるモダリティを統合して単一のトークン集合にまとめて扱えばEncoderは不要でDecoderのみでTransformerを運用できる
- 莫大な言語データを取り込むことで獲得されたFew-shot性能やZero-shot性能が高い
  - 最近のモデルでは「プロンプトエンジニアリング」によってFew-shot性能が顕著に向上している
- GPT-3など一定規模を超えたLLMには、ある種の「創発性」が発現することが観測されている
  - 小さいモデルではランダムな水準の精度しか達成できなかったタスク(四則演算など)において、ある閾値を境に不連続的にモデルがタスクに適応し精度向上を実現する現象
- モデルサイズが巨大

SLM (Small Language Model)

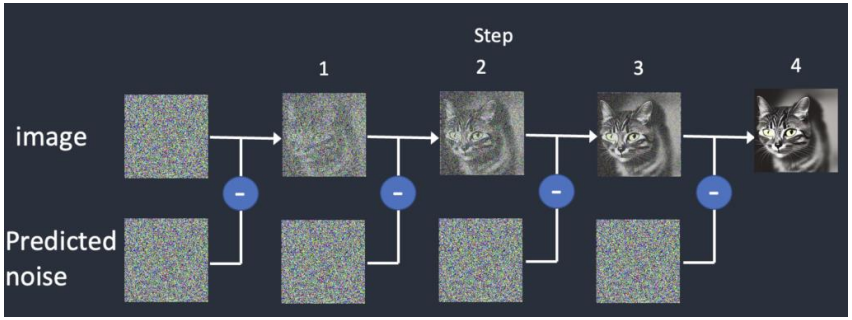
- LLMに対して近年提唱されるようになったSLM（小規模言語モデル)がある
  - LLMよりも比較的小さなパラメータ数のモデルに対する総称
- 一般常識や日常会話ができる程度の知識があり、専門性の高い情報を扱う場合はRAGやファインチューニングを使う運用が想定されている
- パラメータ数が小さいためLLMよりも小規模な計算リソースで使用可能
  - エッジデバイスで処理させることでセキュリティを担保できる
- MicroftのPhi-3 Miniや、MetaのLlama 3.2、GoogleのGemma 2などがSLMの代表例

拡散モデル

- 生成AIの礎となる技術の一つ
- 自然な静止画、動画を生成可能
- Midjourney、Stable Diffusion、OpenAIのSoraなどが有名

Stability AI Japanより引用

拡散モデルの学習



How Does Stable Diffusion Work?より引用

- トレーニング画像に段階的にノイズを加えて破壊していく拡散過程
  - ノイズは正規分布にしたがってそのスケールを段階的に大きくする
- 拡散過程を逆向きにたどって除去すべきノイズを学習しながら画像を復元する生成過程
  - 各ステップで加えられたノイズを予測するモデルをU-Netなどのニューラルネットワークで学習

拡散モデルの特徴

- 画像、音声、動画、化合物、制御など多くの分野で成果を出している
  - 2024年のノーベル化学賞を取ったたんぱく質の構造予測が有名
  - 表形式データを生成できる(tabsynなど)
    - 統計的性質を温存しつつレコード単位で全く異なるデータを生成することでプライバシーを侵害せずにデータ活用可能
  - ロボティクス分野でモデルベース強化学習や模倣学習などで成果をあげている
- 言語の生成(LM)もMasked Diffusion Modelが進展(Mercury、LLaDa)
  - Masked Diffusion: 言語のような離散データのための拡散モデル
  - 同サイズのLLaMA3と同等か上回る性能で高速
  - LLMの課題である、テキストの順序を入れ替えるだけで性能が著しく下がるReversal curseが起こりにくい
    - Reversal curse:「AはBである」で学習したときに「BはAである」を正しく予測できない

拡散モデルの課題と方向性

- 課題: 計算効率が低く計算オーバーヘッドとエネルギー消費が大きい
  - 逆拡散プロセスの反復的な性質により高品質なサンプルを生成するために数百から数千のステップが必要
- 効率的拡散モデルの進歩
  - 出力品質を維持しながらステップ数を削減するための加速サンプリング手法
  - アーキテクチャーとパラメータの最適化(モデル圧縮など)で計算量とメモリの削減
  - GPUやTPUなどハードウェアに最適化された実装

組込み環境で拡散モデル

Machine	Time
Raspberry Pi 4	32分36秒
Raspberry Pi 5	14分11秒

- Stable Diffusion v1.5で400x400pxの画像を生成
  - straczowski/raspberry-pi-stable-diffusion
- やはりCPUのみだと重い

組込み生成AIの実用性

- SLMであればある程度動作する
  - VLM(Vision and Language Model)はかなり動作が遅い
- コマンドプロンプトを工夫すればFew-shot分類機として実用可能
  - 取得したテキストに「ポジティブな内容かをYesかNoで答えて」など
- 速度的な観点でBERT(Encode Only Model)を取り入れるのは有効
- 応用範囲の広さから拡散モデルが有望だが、動作速度は遅い
  - 高速化に期待

AI研究WG紹介

一般社団法人
組込みシステム技術協会
Japan Embedded Systems Technology Association

AI研究WG

- 研究会とセミナーの2本立てで開催
- 研究会
  - 今年で7年目になるDeep Learningを既に理解して開発できるメンバーが集り、様々なテーマでAI活用研究を行なう研究会
  - メンバーは現在12社 24名
- セミナー
  - 今年で9年目になる初学者向けのDeep Learningセミナー
- AI研究WG発表会
  - 年度末に研究会/セミナー別で発表会を実施

研究会紹介

- エッジデバイス上でのDeep Learningの可能性や、様々なテーマで持続的に調査研究を行なう
- 1ヶ月に1度、定例会議を開きDeep Learning周辺の最近の動向の共有、メンバーの研究内容の進捗発表
- 全員でコンペに参加して実力を試したり
  - 個々のメンバーで興味のあるコンペに参加

2024年度研究案件

- 組込み環境で生成AI
- 低リソースデバイスAI
- LLMチューニング
- 深層教科学習の説明可能性
- Unity ML-Agentsを活用した強化学習

組込み環境で生成AI

- Transformerモデルの概要
- 拡散モデルの解説
- 組込み環境(ラズパイ)で拡散モデル
- 組込み生成AIの実用性について考察
  - SLMは実用的
  - 拡散モデルは有望だが遅い

組込み環境で生成AI

Transformer

- Transformerは2017年に発表された自然言語処理の論文”Attention is All You Need”で初めて登場したモデル
- 翻訳タスクでSeq2seq(RNNベース Encoder-Decoderモデル)より速く、精度が高い
- RNNもCNNも使わずAttentionのみ使用したEncoder-Decoderモデル
  - 並列計算が可能
- アーキテクチャのポイントは３つ
  - Encoder-Decoderモデル
  - Attention
  - 全結合層
- NLPの最近のSoTA(BERT、GPTなど)のベースとなるモデル

SLM (Small Language Model)

- LLMに対して近年提唱されるようになったSLM (小規模言語モデル)がある
  - LLMよりも比較的小さなパラメータ数のモデルに対する総称
- 一般常識や日常会話ができる程度の知識があり、専門性の高い情報を扱う場合はRAGやファインチューニングを使う運用が想定されている
- パラメータ数が小さいためLLMよりも小規模な計算リソースで使用可能
  - エッジデバイスで処理させることでセキュリティを担保できる
- MicroftのPhi-3 Miniや、MetaのLlama 3.2、GoogleのGemma 2などがSLMの代表例

組込み環境で拡散モデル

Machine	Time
Raspberry Pi 4	32分36秒
Raspberry Pi 5	14分11秒

- Stable Diffusion v1.5で400x400pxの画像を生成
  - straczowski/raspberry-pi-stable-diffusion
- やはりCPUのみだと重い

組込み生成AIの実用性

- SLMであればある程度動作する
  - VLM(Vision and Language Model)はかなり動作が遅い
- コマンドプロンプトを工夫すればFew-shot分類機として実用可能
  - 取得したテキストに「ポジティブな内容かをYesかNoで答えて」など
- 速度的な観点でBERT(Encode Only Model)を取り入れるのは有効
- 応用範囲の広さから拡散モデルが有望だが、動作速度は遅い
  - 高速化に期待

低リソースデバイスAI

- 前年度から進めていたM5Stack AtomS3(ESP32-S3)でフラッシュ暗算
- 前年度からデータ送受信を変更
  - 1px・1Byte → 1px・1bit
- 推論をアップデート
  - データ送受信に伴う推論画像変更
  - 学習部分を1から自作
- 推論高速で精度も改善

低リソースデバイスAI

AI研究WGでの取り組みについて

テーマ：低リソースデバイスでAI

ウェアラブルデバイスなどでAIが活用されるが、
こうしたデバイスでは限られたリソースの中で
AIを活用しているのが実情

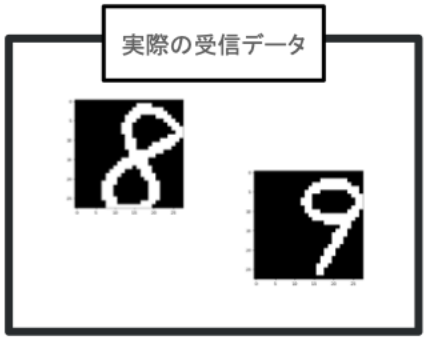
そんなリソース制限下でAIを扱うための研究

リソース制限を受けても
高速に、自由に**AIを動かしたい**！！

速度向上のための対策について

- 数字の形がはっきりとわかりさえすればMNISTで学習したモデルが使用できる
- 画像のピクセルデータを1bit単位で送信
  - 28 * 28 / 8 = **98Byte**（元々は**57600Byte**）

元のRAWデータから**0.17%**に圧縮



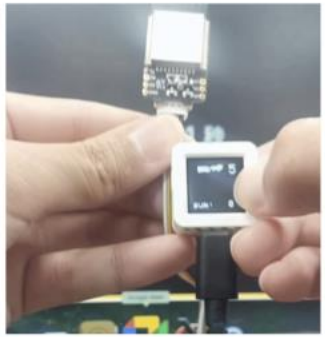
結果について

フラッシュ暗算ができる程度の精度と速度の向上が見られた

画像取得速度：10FPS程度

推論速度：秒間30回の推論程度

数字文字認識：
1→4→8→2→9が認識できている



まとめ

- 前年度から進めていたM5Stack AtomS3（ESP32-S3）でフラッシュ暗算する
  - 前年度からデータ送受信を変更
    - 1px・1Byte → 1px・1bit
  - 推論をアップデート
    - データ送受信に伴う推論画像の変更
    - 学習部分を1から自作
  - 推論は高速で精度も改善
- 次回はrv1103・rv1106による推論とESP32-S3の性能比較

LLMチューニング

- ユーザーが入力したデータを読み取り、指定した情報を抽出して出力
  - 名刺や帳票から効率よく情報を抽出するなど
- この目的のため3つの手法を模索
  - プロンプトチューニング
    - AIから期待通り出力を得られるように、AIに与える指示、命令 (プロンプト)文を最適化
  - 入力データ前処理
    - LLMが理解しやすいように入力データに前処理を行うことで精度向上を目指す
  - LLMパラメータ設定
    - Temperature(回答ランダム性を制御)
    - Frequency Penalty(生成されたテキストに特定単語やフレーズが存在することを抑制)

LLMチューニング

やってみたこと

LLMに行わせるタスク

ユーザーが入力したデータを読み取り、指定した情報を抽出して出力する

上記タスクを達成するために、以下の 3項目を実施した

- プロンプトチューニング
- 入力データの前処理
- LLMのパラメータ設定

プロンプトチューニング

プロンプトチューニング (プロンプトエンジニアリング)とは？

AIから期待通りの出力を得られるように、AIに与える指示、命令(プロンプト)文を最適化する

例えば. . .

> マーケティングについて教えて

と聞くよりも

> B2Bソフトウェア企業が導入すべき効果的なコンテンツマーケティング戦略を 5つ、
> それぞれ200字程度で説明し、各戦略の主な利点と実装する際の課題も含めて解説してください。

と聞いた方がユーザーの目的にあった詳細な回答が得られやすくなる

ただし、複雑なプロンプトにしすぎるとかえって悪化することも …

入力データの前処理

画像からデータを抽出させたい場合、
入力画像に以下のような前処理を施す

- 抽出したい部分を枠で囲む
- 枠線の太さを調整
- 枠線の色を変える
- タスクに関係のない部分はカットする

プロンプトが書きやすくなるのもメリット

例) 赤枠で囲まれた部分には宛先情報が記載されています

最後に

ChatGPTやGemini等、普段使用するようなモデルは高性能なので、工夫を凝らさずとも期待した通りの回答を得ることが出来る

ただ、次のような問題で高性能なモデルを使えない場合もあるので、そのようなケースでは精度向上に有効な手段となる

- 商用利用不可なライセンス
- ローカル環境で動作させたい場合、
  リソースの問題でパラメータ数の多いモデルを動かせない

今回得た知見を基に、来年はエッジデバイス上でLMを動かしてみたい
(注文したJetson早く届いてほしい…)

深層強化学習の説明可能性

- 強化学習における最適な方策について判断根拠を説明すること重要
  - 将棋や囲碁などボードゲームで指し手決定
  - 熟練者からノウハウ獲得、等
- 強化学習にLLMを導入してAgent行動判断根拠を説明させる
- LLMによる推論部分がネックとなり学習を完了することができなかった
- 行動判断根拠実験というより、学習データ言語化に関する実験に

深層強化学習の説明可能性

背景

強化学習における最適な方策についてその判断根拠を説明することは重要

例えば、

　将棋や囲碁などのボードゲームで指し手の決定

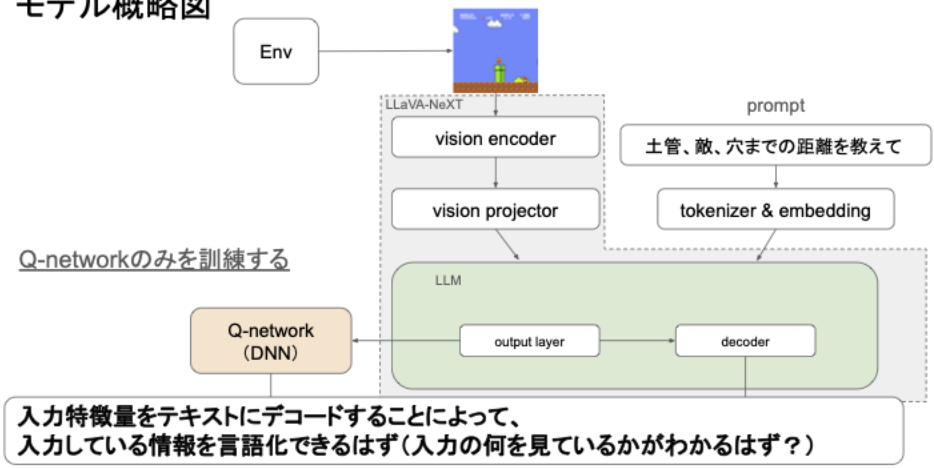
　熟練者からのノウハウ獲得　等

言語化できることに意味があるタスクが多く存在する

急成長しているLLMの高い文章生成能力を上手く使えば、

判断根拠を説明させられるのでは

モデル概略図



実験結果：VLMの推論性能

prompt:

これはスーパーマリオのゲーム画面です。

マリオから緑の土管、穴、敵までの距離を教えてください。

生成されたテキスト

提供されている画像では、マリオは画面の左側のレンガ造りのプラットフォームに立っています。
緑色の粘土パイプはマリオの右側にあり、穴はパイプの真下にあります。
敵は、グンパのように見えるが、レンガのプラットフォームに立っているマリオの右側にもある。

マリオから緑色の粘土パイプまでの距離は、スーパーマリオなどのプラットフォームゲームで通常使用される
グリッドシステムで測定されるように、約1.5〜2ユニットです。
穴がパイプの真下にあるため、マリオから穴までの距離は同じです。
マリオから敵までの距離は画像にはっきりと見えませんが、
敵はマリオの右側にあるレンガ造りのプラットフォームにも立っているため、似たような距離のようです。

実験結果：VLMの推論性能

- 相対距離を正確に予測できていない
  - 生成テキストを見ると、物体間の相対距離を正確に予測できていないように見える
  - プロンプト等で工夫が必要そう
- 推論速度が遅い
  - RTX4070Tiで推論したが、3s/frame程度かかる
  - 少なくとも10000step以上は必要になりそうなので、トライアンドエラー前提の実験はやりにくい

Unity ML-Agentsを活用した強化学習

- 目的: 3D Pinball Space Cadetで人間のプレイと同程度のスコア(1,000,000点)を取得できるモデルの作成
- 前年度の課題の解決を模索
  1. 学習に要する時間が長すぎる
  2. 序盤の学習効率の悪さ
  3. ボールを落とさない方向に学習が進む
- 今回取り組んだ内容
  - 学習時間の短縮
  - 報酬設計の見直し
  - 観測データの追加
- ML-Agentsを用いた学習の高速化(軽微…)
- 報酬設計による学習傾向の改善

Unity ML-Agentsを活用した強化学習

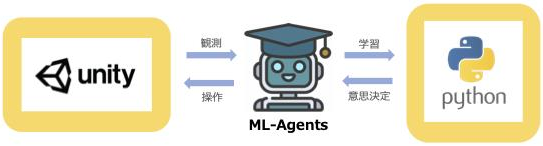
題材

操作は左右のフリッパーと発射台。(台を動かす操作は未使用)
3回ボールが落ちるとゲーム終了。

学習環境②

・学習環境はUnityのML-Agentsを使用
https://github.com/Unity-Technologies/ml-agents

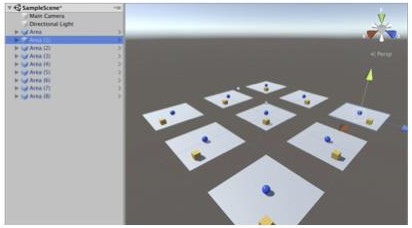
Unity 2023.2.15
ML-Agents version 21



取り組み①: 学習時間の短縮

アプローチ②: 複数のエージェントによる並列学習

1つの環境内に複数のエージェントを配置して学習

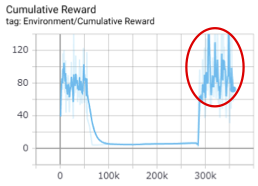


アプリをプレハブ化する対応が必要 → Unity側の改修が難航したため今回は断念

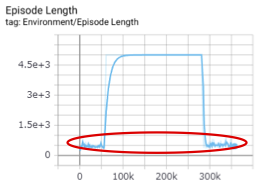
取り組み②: 学習結果の考察

学習結果としてはイマイチ

(ⅰ)報酬が不安定



(ⅱ)エピソードの長さが増加しない



→ リプレイボール
落下ボーナス によって ボール落下のペナルティが伝わっていない可能性

セミナー紹介

- 1年間で3回の座学とグループでのDeep Learningデモ作成がゴール
- 講習にはGoogle Colaboratoryを利用
  - Colaboratoryはクラウドで実行されるJupyterノートブック環境なのでお手軽
- フレームワークはTensorFlow+Keras
- グループ間の情報共有、全体連絡にSlackを活用

2024年度セミナー・成果発表

- 機械学習を用いたディスクの故障予測
- 生体認証を用いた入室管理システム

機械学習を用いたディスクの故障予測

- ディスクのS.M.A.R.T情報と故障有無、および故障発生までの時間を機械学習にて分析
- 生存時間予測にてディスクがN時間後に壊れる確率を推論
  - Cox比例ハザードモデルを用いた予測
    - ハザード回帰係数を使用した生存時間予測
  - Random Servival Forestを用いた推論
    - イベント発生までの時間データを分析するために特別に設計された高度なアンサンブル学習方法

機械学習を用いたディスクの故障予測

概要

・テーマ

機械学習を用いたディスクの故障予測

・方法

ディスクのS.M.A.R.T情報と故障有無、および故障発生までの時間を機械学習にて分析。

生存時間予測にてディスクがN時間後に壊れる確率を推論。

生存時間予測とは

・生物の死や機械システムの故障など、1つの事象が発生するまでの予想される期間を分析する、統計学の一種。

・各種説明変数と事象発生の有無、事象発生までの時間を元に学習を行い、推論を行う。

【とある疾病発生までの推論】

	尿酸値	CRP	経過時間	事象
Aさん	6.1	0.5	3年	発生
Bさん	5.2	0.2	4年	–
Cさん	7.7	0.8	2年	発生
Dさん	6.2	0.3	5年	–

作業概要

・学習環境

言語　　：Python

実行環境　：コンドプロンプト、Google Colaboratory

・利用データ

(kaggle) Backblaze Hard Drive Failure Dataset 2023

https://www.kaggle.com/datasets/priyamsaha17/backblaze-hard-drive-failure-dataset-2023

・手法

① Cox比例ハザードモデルを用いた予測

②Random Servival Forestを用いた推論

手法② Random Servival Forest 考察

・データについて

実際のデータを確認すると、1ヶ月後に壊れるデータと1年後も壊れないデータとで差異が無い場合も存在した。

	S.M.A.R.T 1	S.M.A.R.T 5	S.M.A.R.T 10	S.M.A.R.T 196	S.M.A.R.T 197	S.M.A.R.T 198	S.M.A.R.T 226
壊れる	100	100	100	100	100	100	100
壊れない	100	100	100	100	100	100	100

⇒兆候自体が無く、突然壊れる、というパターン。
実はS.M.A.R.T情報のWikipediaにも、「Googleの研究では、半数のHDDは何の兆候もなく突然死する」との記載あり。

⇒もしかすると今回利用した説明変数以外のパラメータでは兆候が取れるものがあるかもしれない。

生体認証を用いた入室管理システム

- 顔認証
  - 顔領域検出モデル(MTCNN)で顔の切り抜き
  - 顔特徴量ベクトルをInceptionResNetV1で取得
  - 事前登録した顔とコサイン類似度で比較
- 声紋認証
  - 音声から発音部分を音声認識モデル(Whisper)で抽出
  - 事前登録していた音声とベクトルDBで類似度を算出

生体認証を用いた入室管理システム

顔認証シーケンス

①認証者の顔が写った画像を撮影

②撮影した画像から顔部分切り抜き
"MTCNN"という顔領域検出モデル(学習済み)を使用

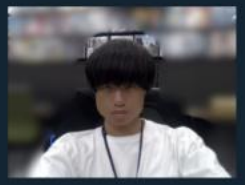
③認証者の顔の特徴量ベクトル取得
"InceptionResNetV1"という顔認識モデル(学習済み)を使用

④事前登録しておいた顔との類似度算出
コサイン角度を類似度として算出

⑤類似度が閾値を超えているか確認
超えた場合：対象の事前登録しておいた顔の人物=認証者とし、声紋認証へ
超えなかった場合：事前登録しておいた別の人物の顔画像を用いて④からやり直す
※事前登録しておいた全ての顔画像との類似度が閾値を超えなかった場合、入室を拒否

顔認証の精度

肌の色・服を変えた同一人物：同一人物だと判定可能



声紋認証シーケンス

①認証者の声が含まれる音声を録音

②録音した音声を事前登録しておいた音声と結合

③各音声から声が入っている部分のみ抽出
"Whisper"という音声認識モデル(学習済み)を使用

④認証者の声と事前登録しておいた全ての声の類似度算出
ベクトルDBを用いて類似度を算出
類似度が閾値を超えていた場合、対象の事前登録者を候補者リストに追加

⑤入室を許可できるか判定
候補者リストの中に顔認証で判定された人物が含まれる場合：入室を許可
含まれなかった場合：入室を拒否

声紋認証の精度

識別不可パターン

認証対象音声(花野)：カブトムシ

事前登録音声(木下)：おはよう

事前登録音声(森)：こんばんは

事前登録音声(花野)：よろしくお願いします

「AI研究WG EdgeTech+ 2025」

2025/11/17 発行

発行者　一般社団法人 組込みシステム技術協会
東京都 中央区 入船 1-5-11 弘報ビル5階
TEL: 03(6372)0211 FAX: 03(6372)0212
URL: https://www.jasa.or.jp/

一般社団法人
組込みシステム技術協会
Japan Embedded Systems Technology Association