組込みの生成AIの可能性

第12回HEPTコンソーシアムフォーラム

中村 仁昭

自己紹介

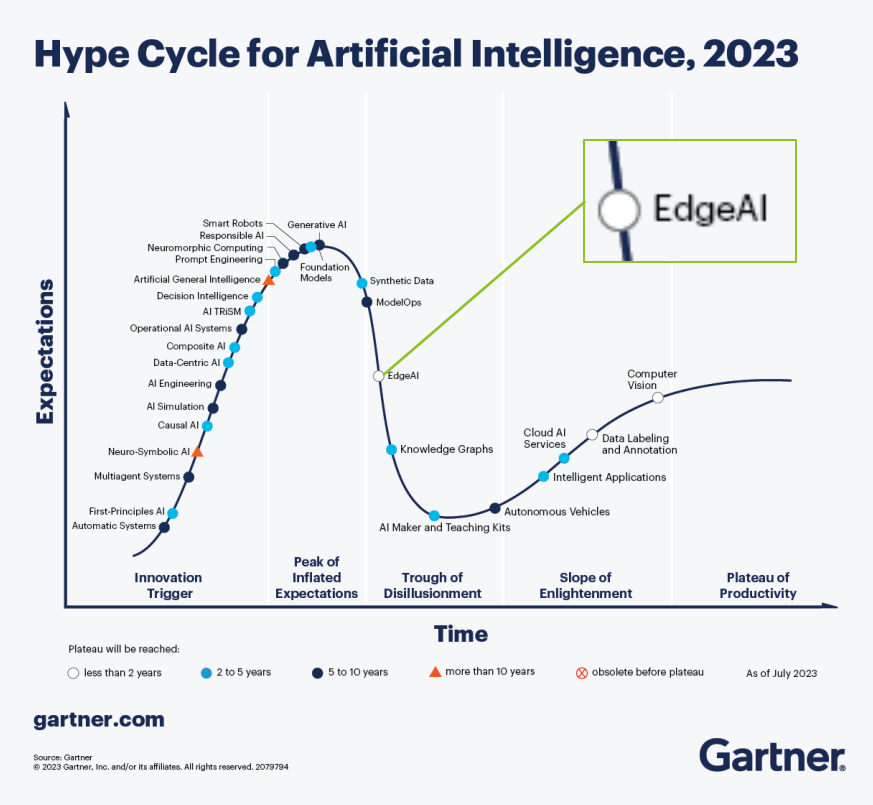
- 株式会社Bee CTO、JASA AI研究WG主査 中村 仁昭
  - 社内AIチームPM、車載プラットフォームPMを担当
  - 2018年からJASAでAIセミナー、研究会を実施
- CQ出版 Interface誌でAI含めた様々な記事を執筆
  - 2023年10月号から2024年3月号まで「Rustプログラミング問題集」の連載
  - 2024年9月号に「第4部の顔検出/異常検知/病害検知/テンプレート・マッチングを移植 ラズベリー・パイで画像処理」を執筆

アウトライン

- 組込みAIの現状
- 組込みの生成AIの可能性
- JASA AI研究WGの紹介

組込みAIの現状

組込みAI(EdgeAI)の位置付け



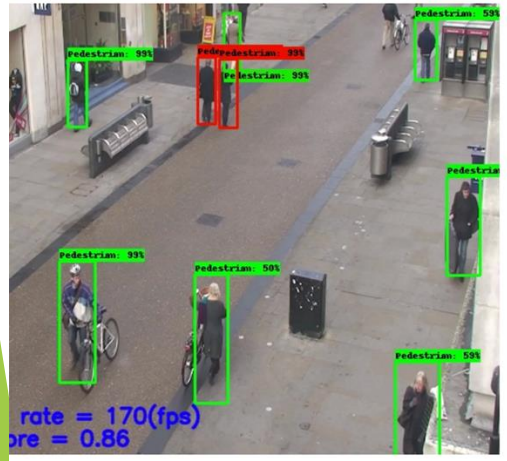
- GartnerのHype Cycle for Artificial Intelligence，2023ではようやく幻滅期に入り、ますます実装例は手に入りにくい状況に
  - 主流の採用までに要する年数は2年未満とされている

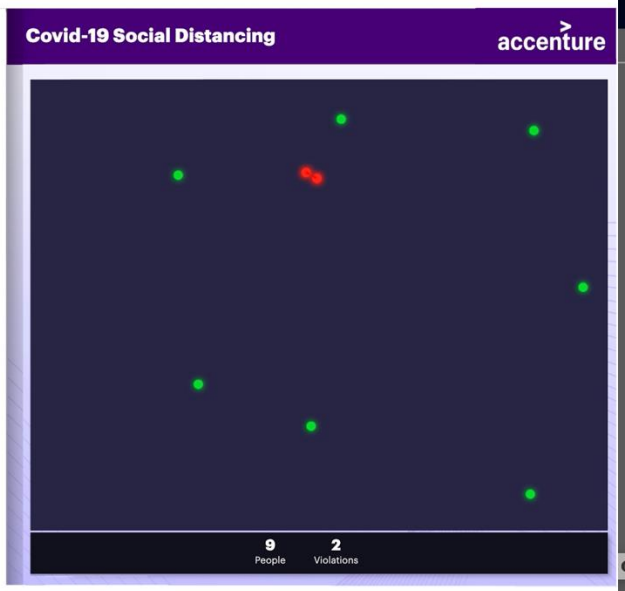
組込みAIの実装例

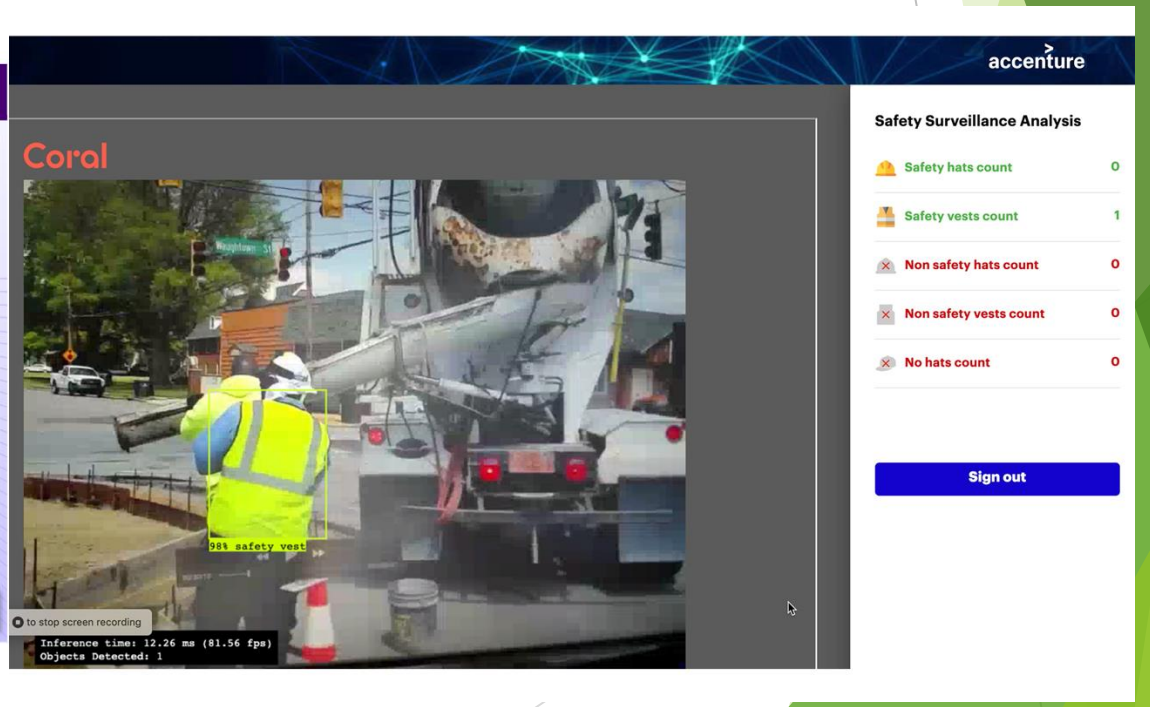
- スマホの音声認識や、カメラ画像処理は採用されて久しい
- Intel CoralのCustomer Storiesが数少ないながらも実際に使用されている実例が更新されている
  - コンサルティング会社AccentureのAIを活用した外観検査の取り組み
  - くら寿司の皿を数える装置
  - ノルウェーの配電会社Pratexoの電力グリッドの変圧器の異常検査
  - など

Accentureの外観検査の取り組み

- 知的財産と労働者を保護するためにプライバシーとセキュリティを強化する必要があり、ネットワークを使用しないローカルの分析にCoral EdgeAIを選択

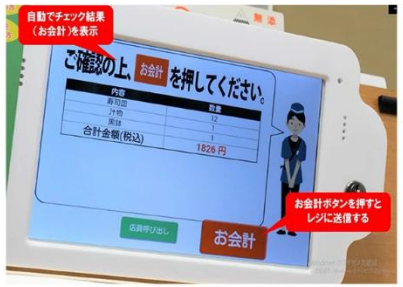






くら寿司の皿を数える装置

- 回転レーンから取った皿の数をRaspberry　Pi4でQRコードの識別とTensorFlow(Coral　USB　Accelarator)を使った画像認識で皿の種類と数をカウント



Pratexoの電力グリッドの変圧器の異常検知

- Coral M.2アクセラレーターで各変圧器が発する音から機械学習モデルで問題が発生するかを予測し電力グリッドの信頼性を確保

組込みAI実装の普及に向けて

- 実装例が表に出てこない
  - 調査は継続する必要はありそう
- もう幻滅期に入ってメディアに注目されないため、自ら実装を進めないと情報が集まらない可能性が高い
  - Computer Visionなど啓発期に入っている技術を組込みに導入するのであれば、低リスクで実装を進められる

組込みの生成AIの可能性

そもそも生成AIとは？

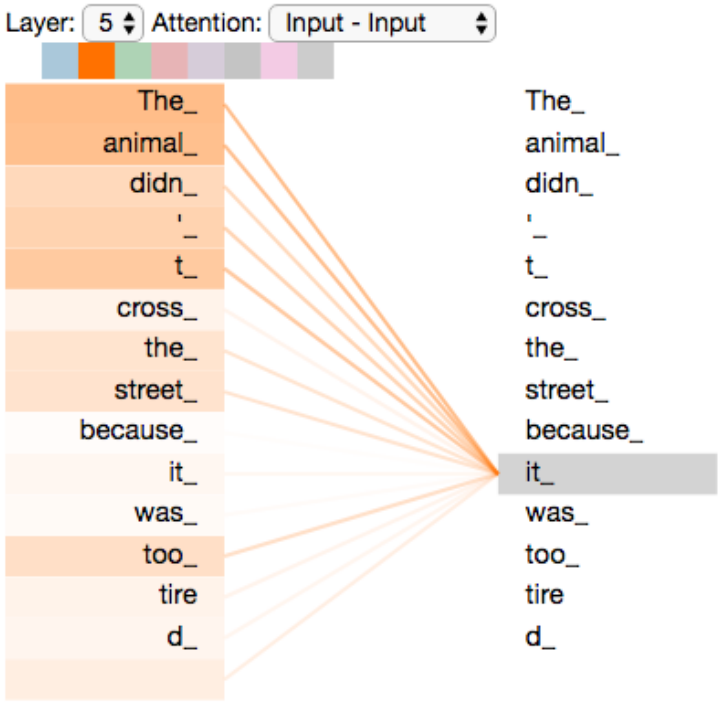
- 分類や回帰と異なり生成タスクは従来困難と言われていた
- Transformerによって言語モデルが高度な能力を獲得
  - ChatGPTなどの大規模言語モデル(LLM: Large Language Model)
  - ViT (Vision Transformer)など他タスクの応用も
- 拡散モデルによって画像、音声、動画の生成が可能に
  - Stable Diffusionなど自然な画像生成が有名
  - 化合物や制御など多くの分野で成果を出している
- それぞれの生成AIがどのように動作しているのかをみてみる

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

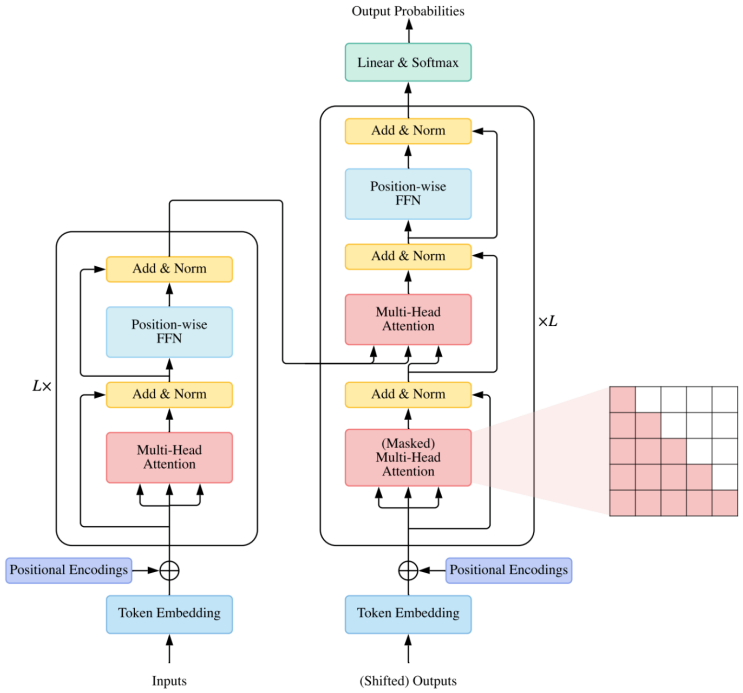
- TransformerはAttentionと呼ばれる仕組みを効率的に積層した深層学習モデル
- Attentionとはデータを検索するための鍵(Key)と実際の値(Value)のペア集合に対して、問い合わせ(Query)を投げて値を取り出す操作



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

RAG (Retrieval Augmented Generation)

- LLMに入力をそのまま与えずに、入力に関係する情報を検索して入力に情報を追加した内容をLLMに与えることで回答精度を上げる技術
- 最新情報や社内情報などを回答できるように
- インターネット検索や、入力内容の特徴ベクトル使ったベクトルDB検索などが使われる

SLM (Small Language Model)

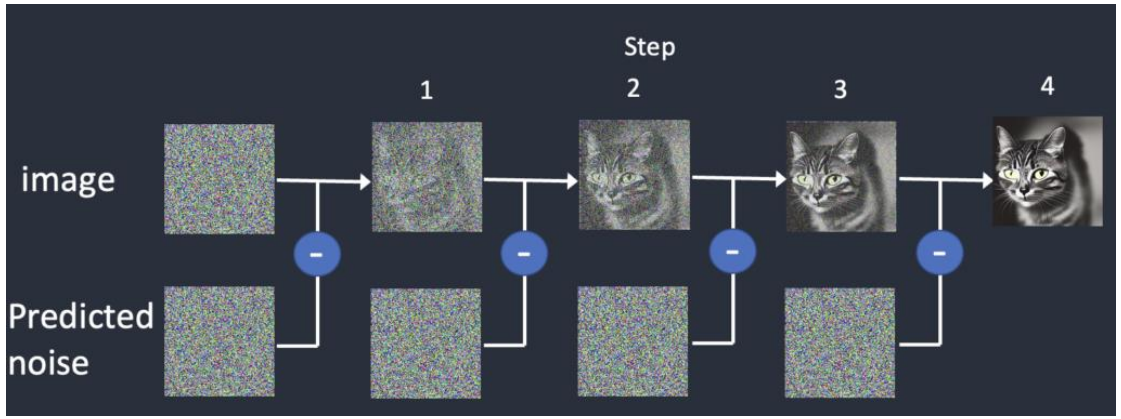
- LLMに対して近年提唱されるようになったSLM (小規模言語モデル)がある
  - LLMよりも比較的小さなパラメータ数のモデルに対する総称
- 一般常識や日常会話ができる程度の知識があり、専門性の高い情報を扱う場合はRAGやファインチューニングを使う運用が想定されている
- パラメータ数が小さいためLLMよりも小規模な計算リソースで使用可能
  - エッジデバイスで処理させることでセキュリティを担保できる
- MicoroftのPhi-3 Miniや、MetaのLlama 3.2、GoogleのGemma 2などがSLMの代表例

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
  - Excelのような表形式データを生成できる
    - 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の応用例も
- 言語の生成(LM)はまだ十分な成果は挙げていない
  - 言語が離散的なデータであるためと予想されいてる
  - もし言語モデルが実現した場合、生成候補が多様性を持ち効率的に列挙できる
    - 多峰性と呼ばれる拡散モデルの特徴

組込み機器でllama2を動作させてみた

model	Tokens/sec
Raspberry Pi 4	0.96
Raspberry Pi 5	2.97
Jetson Xavier NX	10.97
GeForce RTX 4070Ti(12GB) *1	12.01

- Raspberry Pi 4はさすがに遅い
- Raspberry Pi 5はある程度実用的
- Jetson Xavier NXは快適に動作
  - Web版ChatGPTより速い印象
  - デスクトップPCとほぼ同レベルの速度

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も期待できるが、動作速度は遅いと思われる
  - 類似のフローマッチングモデルでは速度的に期待できるか？

JASA AI研究WGの紹介

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4年目になるDeep Learningをすでに理解して開発できるメンバーが集まり、様々なテーマでAI活用研究を行う研究会
  - メンバーは現在8社17名
- セミナー
  - 今年で7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を行なっている
- 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

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Jupyter Notebook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

研究会の個々の研究案件の紹介

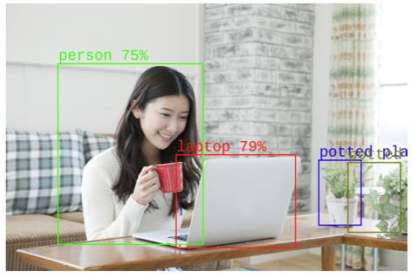
- 推論時の消費電力調査
- 競馬AI予測研究
- 低リソースデバイスAI
  - Edge TPUで推論
  - リザバーコンピューティング(ESN)の調査
- 異常音検出
- FPGA上での学習
- 強化学習
- JetBotで自動運転

推論時の消費電力調査

- 組込みAIの実用化に向け、Raspberry Pi4やEdge TPUで推論させた時の実消費電力を計測
  - Raspberry Pi4(人物検出)：4.2～6.0W
  - Edge TPU(物体検出)：5.2～6.0W

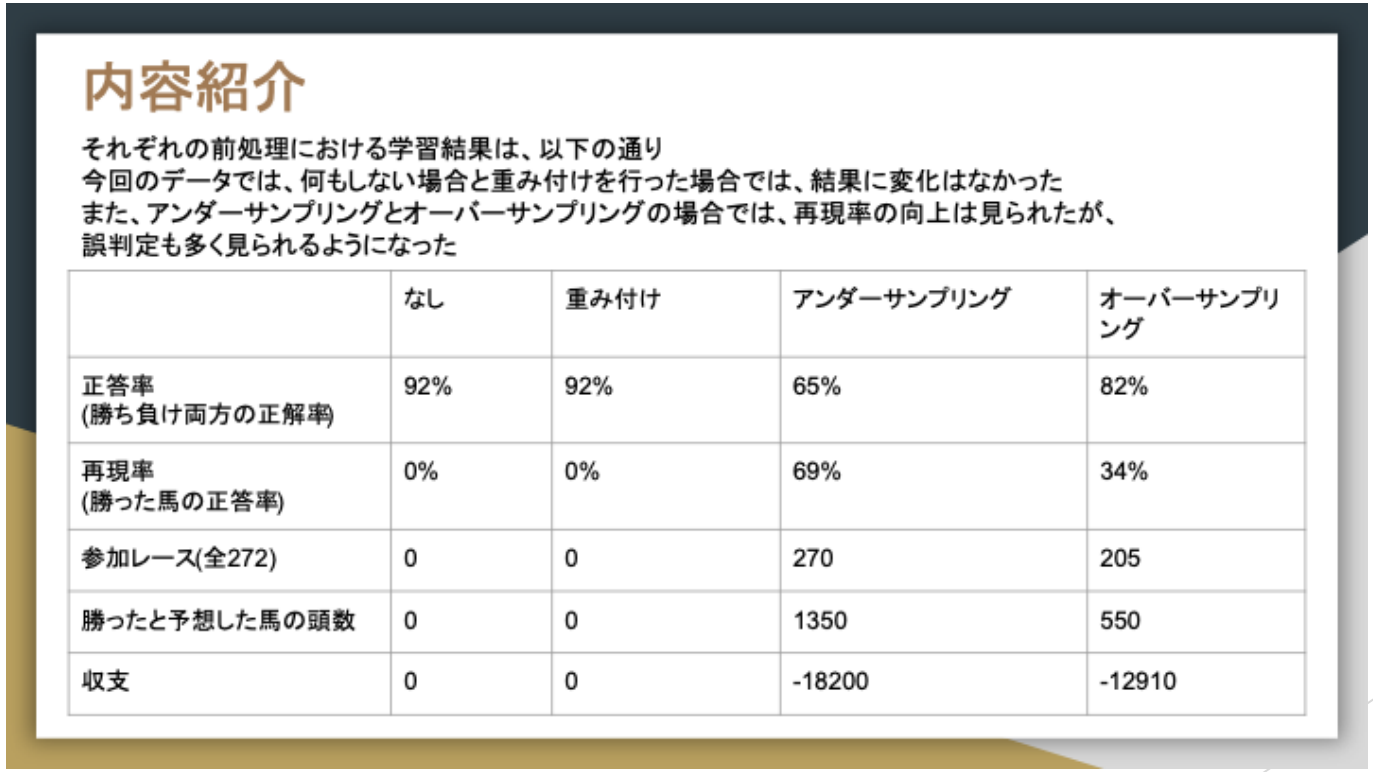
EdgeTPUでの物体検出

- 前掲の人物検出モデルがpycoralベースで動作しなかったため、EdgeTPUのexampleにあったMobileNetV2ベースの物体検出SSDをpycoralで動作
  - IDLE(EdgeTPU接続): 3.5W (接続のみで1.4W程度消費)
    - EdgeTPU単体: 0.3W
  - 推論: 5.2～6.0W 20fps
    - EdgeTPU単体: 1.0W～0.3W
  - 検出精度は低いが速度は速い
  - 思ったより消費電力は高くない



競馬AI予測研究

- 前年の研究から、前処理の導入による予測の改善を実施
  - 重みづけ・アンダーサンプリン・オーバーサンプリングを比較



内容紹介

それぞれの前処理における学習結果は、以下の通り
今回のデータでは、何もしない場合と重み付けを行った場合では、結果に変化はなかった
また、アンダーサンプリングとオーバーサンプリングの場合では、再現率の向上は見られたが、
誤判定も多く見られるようになった

	なし	重み付け	アンダーサンプリング	オーバーサンプリング
正答率 (勝ち負け両方の正解率)	92%	92%	65%	82%
再現率 (勝った馬の正答率)	0%	0%	69%	34%
参加レース(全272)	0	0	270	205
勝ったと予想した馬の頭数	0	0	1350	550
収支	0	0	-18200	-12910

低リソースデバイスAI

- インテルのEdge TPUを試してみる
  - 性能が高くもっと掘り下げてみる価値ありと判断

Edge TPU 推論性能比較

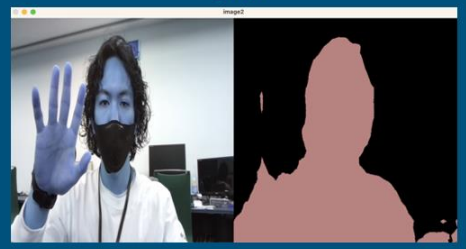
model: Semantic Segmentation

Mac CPU
- FPS: 1

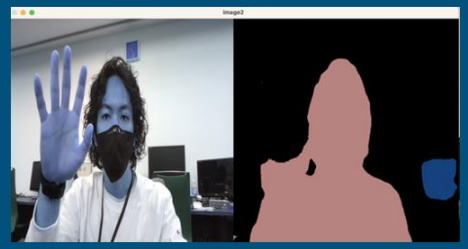
Mac Radeon Pro 555X 4 GB
- FPS: 3.70

EdgeTPU:
- FPS: 17.48

GPU



TPU

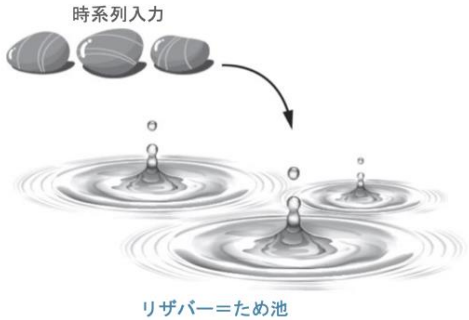


低リソースデバイスAI

- リザバーコンピューティング(ESN)の調査
  - リカレントニューラルネットワークの特殊なモデルを一般化した概念で、時系列情報処理に適している

概要

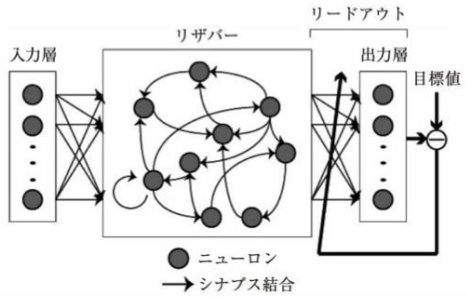
リザバーコンピューティング



石（入力） を入れると波紋が生じる
波紋は、石の大きさや形、投げ入れる順番によって異なる
波紋を見ればそこから入力情報について分析可能

代表例

Echo State Network



2001年にJaeger（イェーガー）によって提案
リカレントニューラルネットワークの特殊なモデル
出力層の重みのみを学習の対象（他は固定）
→低リソース環境で活用可能性あり

https://www.skillupai.com/blog/tech/reservoir-computing/

FPGA上での学習

- FPGAでの学習の可能性、限界、課題の調査を行う
  - 去年度はまず推論をSignateのエッジAIコンテストに参加して試してみた

AIエッジコンテスト取組内容

DeepLabV3の量子化モデルをFPGAで動作

- 前述の内容を考慮し、推論用のpythonスクリプトを作成して動作させた
- FPGAボード：KV260
- 入力：192x120x3
- 実行速度：5fps

入力画像

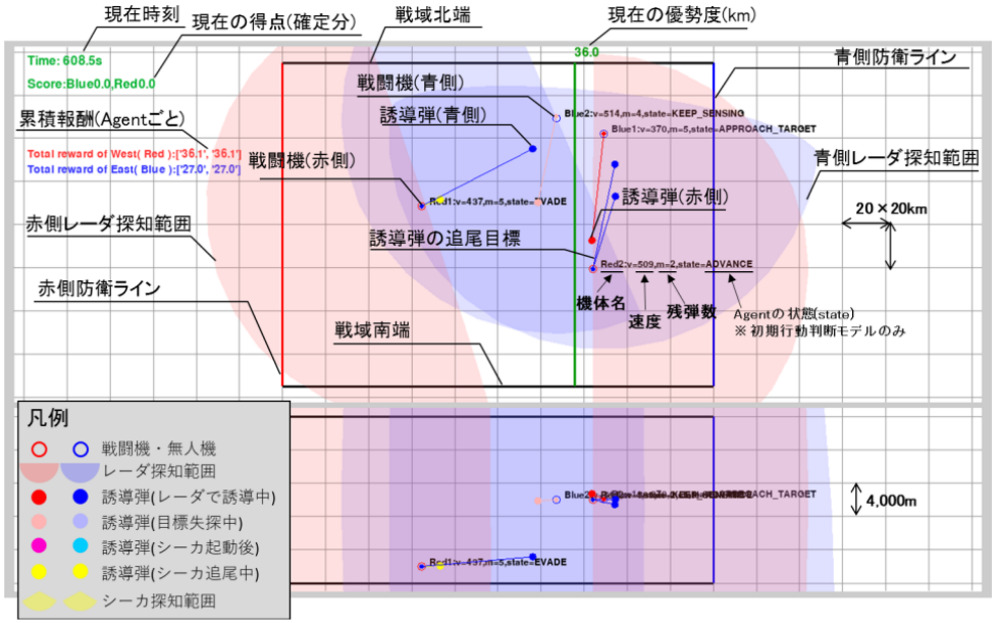
可視化画像

PointPainting全体を動かすところまではいけなかった

強化学習

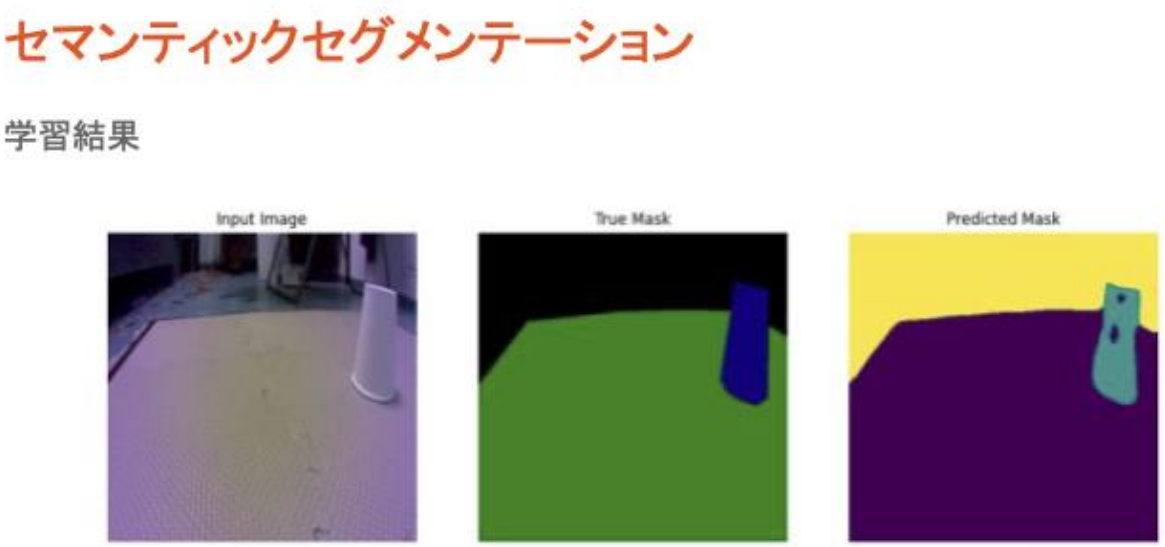
- 強化学習を主題としたシミュレーションコンペへの参加
  - 空戦AIチャレンジに再挑戦

コンペの概要③



JetBotで自動運転

- ラインをトレースして、ライン上の障害物を検知し停止するのを目標とした
  - セマンティック・セグメンテーション モデルの学習と、ライントレースのお試しまで完了



入力画像(左)、正解画像(中央)、予測画像(右)