新技術説明会
New Technology Presentation Meetings!

JST MIRAI 未来社会創造事業

事故につながる誤り多数を ねらい通りに抑制するAI修正技術

国立情報学研究所　アーキテクチャ科学研究系
准教授　石川　冬樹

JST未来社会創造事業　“Engineerable AI” プロジェクト

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

eAI

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

自己紹介

- 国立情報学研究所　准教授
  - ソフトウェア工学，特にディペンダビリティ：
    形式手法，自動テスト生成，安全性論証など

- 現在の主な研究プロジェクト
  - JST MIRAI-eAI：機械学習システムのディペンダビリティ
  - JST ERATO-MMSD：自動運転システムの安全性

- 産業界向け教育・実践研究
  - トップエスイー，日科技連SQiP，電通大AISECなど
  - 機械学習工学コミュニティ（MLSE研究会，QA4AI）

機械学習工学研究会
MACHINE LEARNING SYSTEMS ENGINEERING

AIプロダクト
品質保証
コンソーシアム

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

産学連携成果

（公開されているもののみ掲載）

- 自動運転のためのテスト自動生成，問題分析
  （マツダ株式会社）
- 自動配達ロボットにおけるサービス仕様の探索
  （パナソニックホールディングス株式会社）
- ゲームの自動テスト・自動テストの解釈説明
  （株式会社コナミデジタルエンタテインメント）
- その他AIの品質保証に関する連携多数

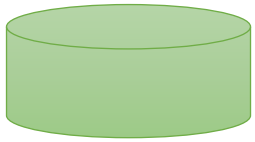
背景：深層学習技術によるAI

■現在の産業応用における一つの主流：

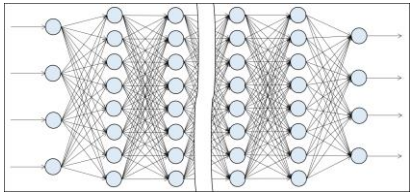
深層学習（ディープラーニング）技術を用いた教師あり学習

画像などの入力に対して正解の設定が必要（ラベリング）

数百万・数千万といった数のパラメーターをデータを基に設定



訓練データ



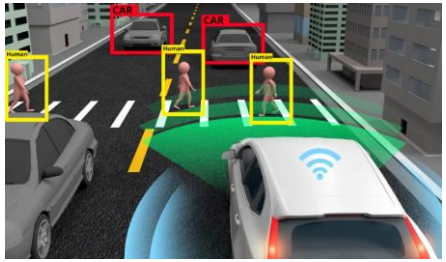
判別や予測の機能

訓練 ＝ 多数のパラメーターの設定

ニーズ：AIの安全性

- 機械学習技術を用いたAIシステムの産業応用・社会実装の盛んな追求
- ➡ 品質，特に安全性に対する懸念 [NEDO'19]
  - 50%が安全性はリスクと受け止め
  - 従来よく言われる「正解率（精度）」だけでなく

「状況Aでも安全？状況Bでは？ひどい間違いはしない？」



新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

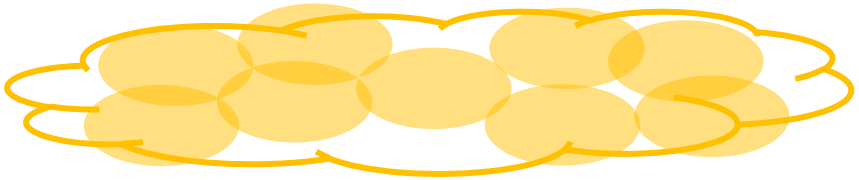
ニーズ：AIの安全性

「状況Aでも安全？状況Bでは？ひどい間違いはしない？」

- 標準やガイドラインによる要請
  - AI品質に関するガイドライン（AIQM, QA4AI）
  - 自律システム・自動運転に関する標準
    （ISO 21448, ANSI/UL 4600）

誤認識の影響による事故リスクが大きい
認識対象・環境を分析せよ

特定の状況での誤り率が
高くなっていないか？

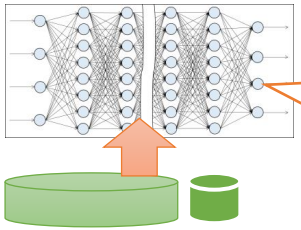


「10万件中 85% 正解」などではなく
想定する認識対象・環境
ごとの評価とリスク低減が必要！

AIの修正：従来技術とその問題

- 従来，安全性などのAI品質が十分でないとき
  - 訓練データを追加する
    - 例：「○○の誤りが多いので，○○のデータを追加収集」
  - 訓練の方法（ハイパーパラメーター）を調整する
    - 例：「○○のデータが少なめなので，重視度を0.8にしてみる」

➡多大な時間がかかる試行錯誤，結果は予測困難



“Changing anything changes everything”

訓練データ追加による再訓練
→ 数百万個以上の
パラメーターを「シャッフル」

更新結果が不確実・
制御困難！

意図せぬ性能劣化

現場の声：
「数日では済まない試行錯誤」

新技術説明会
New Technology Presentation Meetings!

JST MIRAI 未来社会創造事業

JSTプロジェクトと「AI修正」技術

- JST未来社会創造事業においてこの「AI修正」を一つの軸としたプロジェクトを実施中

プロジェクトの目標：細やかなニーズに対し
AIを仕立て上げられる“Engineerable AI”技術を創出,
医療・交通の二分野において実証

  - 「AI修正」は200名以上の実務者の要請を受けテーマ設定,
    主に国立情報学研究所，九州大学，富士通が技術開発

⇒ 自動車業界や安全性の専門家が定めたベンチマークの実施を通して技術を評価・改善してきた

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

提案AI修正技術 DistrRep ：特徴

- 深層ニューラルネットワークの誤識別の要因を分析
  数百万以上のパラメーターシャッフルを避け
  ピンポイントな変更を自動探索

- 誤識別に対するユーザーからの細かな要望に応え，
  深層学習によるAIを狙い通りに修正可能
  複数種類の誤識別に対する同時修正・調整に特化

- 安全性要求が極めて厳しい自動運転分野において，
  産業界とともに定めたベンチマークで効果を実証
  ニーズ駆動で行った技術開発の成果

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

提案AI修正技術 DistrRep ：使い方

- 入力
  - 訓練を行ったAI（深層学習モデル）
  - 「どういう誤り方を直したい・防ぎたいのか」の優先度

番号	AIの誤り種別	シーン	ハザード	リスクレベル	AI評価	許容可否
001	誤分類： 歩行者→バイク搭乗者	自車の前の 歩行者	ブレーキせず 歩行者に衝突	5	誤り 4%	○
002	誤分類： バイク搭乗者→歩行者	近距離で追随する 後方車両	不要なブレーキで 後方から衝突	3	誤り35%	×
…	…	…	…	…	…	…

（入力につながるリスク分析の例）

- 出力：修正済みAI
- *リリース前の詳細な品質調整や，*
  テストや運用で検出されたリスクへの対応に活用

提案AI修正技術 DistrRep ：性能評価

- 性能評価の例
  - 運転シーンにおいて安全性に関わる正解率・誤り率の指標12個を総合評価
  - 以下は先端モデルの一つ EfficientNet B7 での評価例

ENETB7	Arachne	Arachne_REM	DistrRep	SplitTrain_NW	SplitTrain_W	FullTrain_NW	FullTrain_W
AVG DELTA	-1,13	1,40	7,53	0,29	0,13	0,83	-0,22
MAX DELTA	0,99	2,90	8,38	1,16	1,32	1,33	0,59
MIN DELTA	-2,61	0,34	6,75	-1,12	-1,18	0,24	-1,55

類似する既存技術では
多数のリスク要因を
同時にうまく直せない

再訓練では少し直ったり
直らなかったり，制御困難

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

提案AI修正技術 DistrRep ：活用場面

- 向いている状況
  - 影響が大きい誤りとそうでないものがあり，
    優先度に応じたバランス・トレードオフの調整が重要
  - 「今までできていたことが，更新の結果できなくなる」
    デグレの悪影響（ユーザ体験や再調査コスト）が大きい
- 向いていない状況
  - そもそもデータが少なく全体の正解率の時点で低すぎる
    （これはプロジェクトの別技術が対応）
  - 当たり外れの種類は気にせず「合計点」「平均点」が
    高ければよい（投資など）

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

提案AI修正技術 DistrRep ：実用化へ

- 分類タスクについては十分に効果を確認
- ➡ 個別企業のデータ，ニーズをいただいて，技術深化と実証を行うことが重要なフェーズ
  - 類似技術を富士通の社内で効果確認済み
- 検出タスクへの技術適合が2023年度前半の目標
  - 深層学習モデルの多様性が高く，技術のアレンジが必要
- 九大，富士通の他AI修正技術との連携・統合も実施中

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

連携の進め方

（現時点では，ツールリリース前の先行技術適用）

- A．我々のプロジェクト内のベンチマークデータ等を用いた技術試行や評価
- B．企業固有の問題・データへの適用
  - こちらの場合は技術的な適合性について事前検証が必要
- いずれにしても
  - 可能でしたら，プロジェクトにご参加いただき，ニーズ調査等ご協力いただければ嬉しいです！
    （現在十数社が参加）
  - まずは個別の連携という形でも可能

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

本発表技術に関する知的財産権

■発明の名称：

モデル生成装置、モデル生成方法及びプログラム

■出願人：

大学共同利用機関法人　情報・システム研究機構

■出願番号：

特願2023-014970

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

まとめ

- 複数種別の誤りに対応したAI修正技術
  - 深層学習技術における「性能制御の困難さ」に対応
  - 安全性，すなわち誤りの影響によるリスクを考慮した識別性能の調整を可能に
- より大きなプロジェクトでの取り組み進行中
  - 「AI修正技術」は，デグレ抑制重視など他の技術バリエーションも開発
- より広く「AI工学」に関する個別企業との連携も実施中

新技術説明会
New Technology Presentation Meetings!
JST MIRAI 未来社会創造事業

問い合わせ先

国立情報学研究所
総務部 企画課 社会連携推進室
大型プロジェクト・知財チーム

TEL 03-4212-2139
FAX 03-4212-2120
EMAIL chizai@nii.ac.jp

なお現在、原則、在宅勤務勤務中のため
メールにてご連絡頂ければ幸いです。