細やかな要求に応じて安全なAIを仕立て上げる

eAIプロジェクト

自動運転の安全性に向けたAI修正技術

石川　冬樹（アーキテクチャ科学研究系）・ eAI プロジェクト
吉岡　信和, Paolo ARCAINI, Thomas LAURENT

どんな研究？

機械学習・深層学習技術を用いたAIソフトウェアに対して，細やかな要求やリスクを考慮し対応するための工学技術に取り組んでいます．

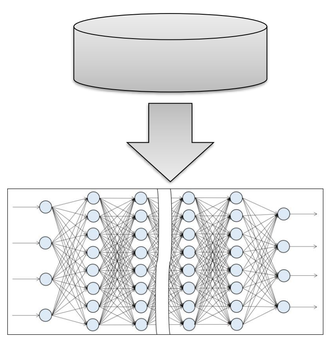
何ができる？

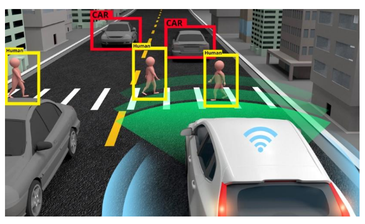
プログラムに対する欠陥箇所推定や自動修正の技術を適合させ，AIソフトウェアに対しても，要求・リスクに応じた「狙った修正」を行えます．

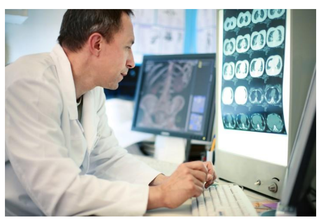
eAI プロジェクトの概要

機械学習・深層学習技術を用いたAIソフトウェア：
データを用いた訓練によるパラメータ調整を通して
機能を構築・修正
（深層学習では数百万以上のパラメータ）

1．大量データが必要となる
2．避けたい誤りの考慮など細かい調整ができない



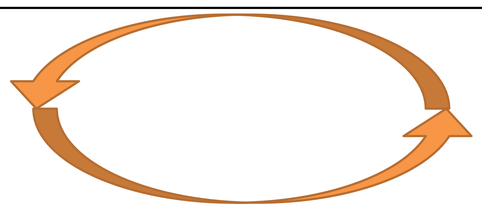




安全・信頼が重要な分野におけるAI活用の障害

要求やリスクに応じAIを仕立て上げるフレームワーク → 医療・交通の二分野で実証

フレームワーク技術
要求分析・リスク分析技術



早大

モジュール型AI構築技術
構造保存型AI構築技術
・・・

東工大　　阪大

データが少ない希少な症例に対しAI構築

リスク考慮型AI修正技術
ノイズ適応型AI修正技術
・・・

NII　　九大・東大

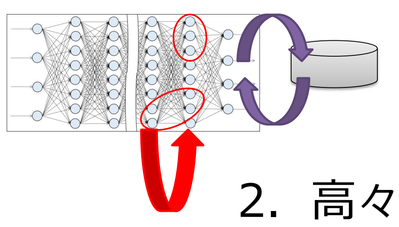
安全性要求を踏まえてAI修正

AI修正技術へのアプローチ

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	近距離の後方車両	不要なブレーキで後方から衝突	3	誤り 35%	×
…	…	…	…	…	…	…

**自動車企業や安全性の専門家とともに
AI誤りの影響を分析してベンチマークを設定**

データ追加や訓練ハイパーパラメータの調整では特定の誤りを減らすような修正になかなかたどり着かず，意図せず他の誤りが増えてしまうことも



**従来ソフトウェアプログラム向けの
欠陥箇所推定・自動修正技術の
考え方を活用！**

1．訓練データや運用データから致命的な誤りの要因を分析

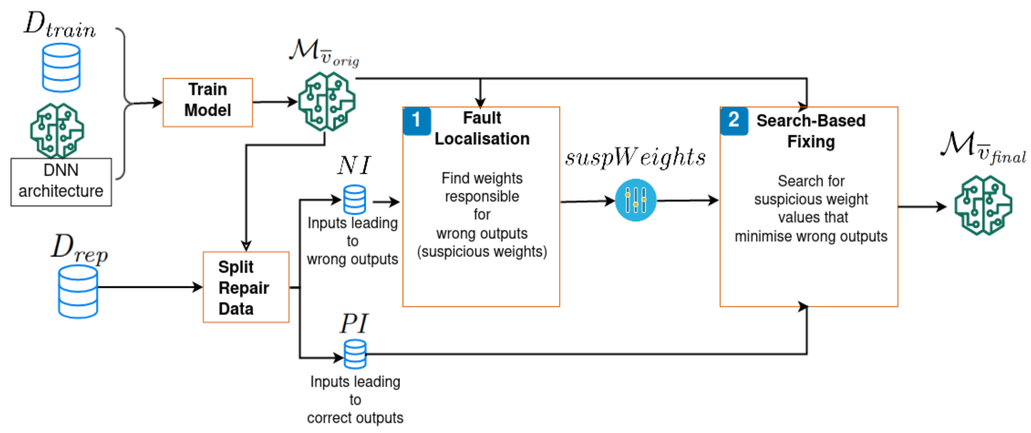
2．高々数十個程度に絞ってパラメータを調整

連絡先：国立情報学研究所　アーキテクチャ科学研究系　石川　冬樹
URL：https://research.nii.ac.jp/~f-ishikawa/ Email：f-ishikawa@nii.ac.jp

Baseline AI Repair

2 steps:
- Fault Localisation: find "suspicious weights"
- Search-based repair: find optimal suspicious weights values

Limitations:
- Considers all faults (wrong outputs) together
- Takes a static view of the system/faults

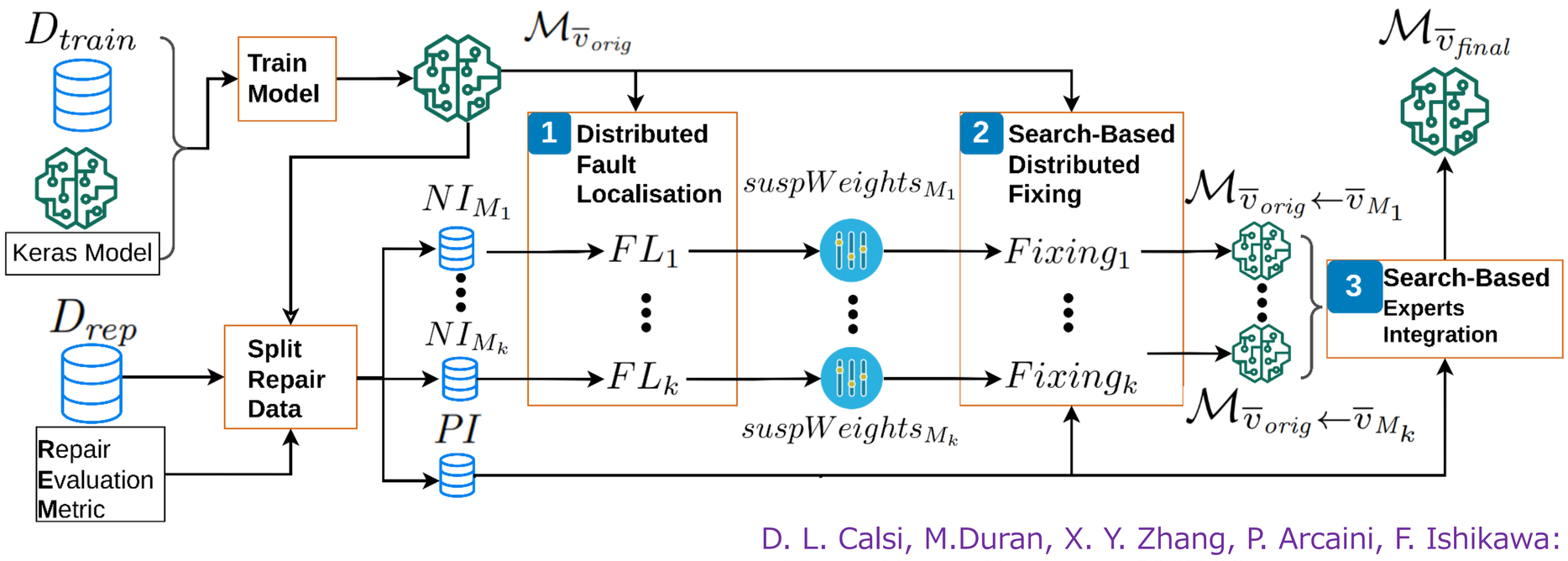


Distributed Repair Technique

Motivation

Deep Neural Networks have different types of faults, how to consider them all when repairing?
Current approaches consider all the wrong model outputs at once. They do not distinguish by fault type.
For example, “car -> truck” fault vs. “car -> pedestrian” fault

Method: Perform multiple specialised repairs and merge them



D. L. Calsi, M.Duran, X. Y. Zhang, P. Arcaini, F. Ishikawa: Distributed Repair of Deep Neural Networks. ICST’23

Results

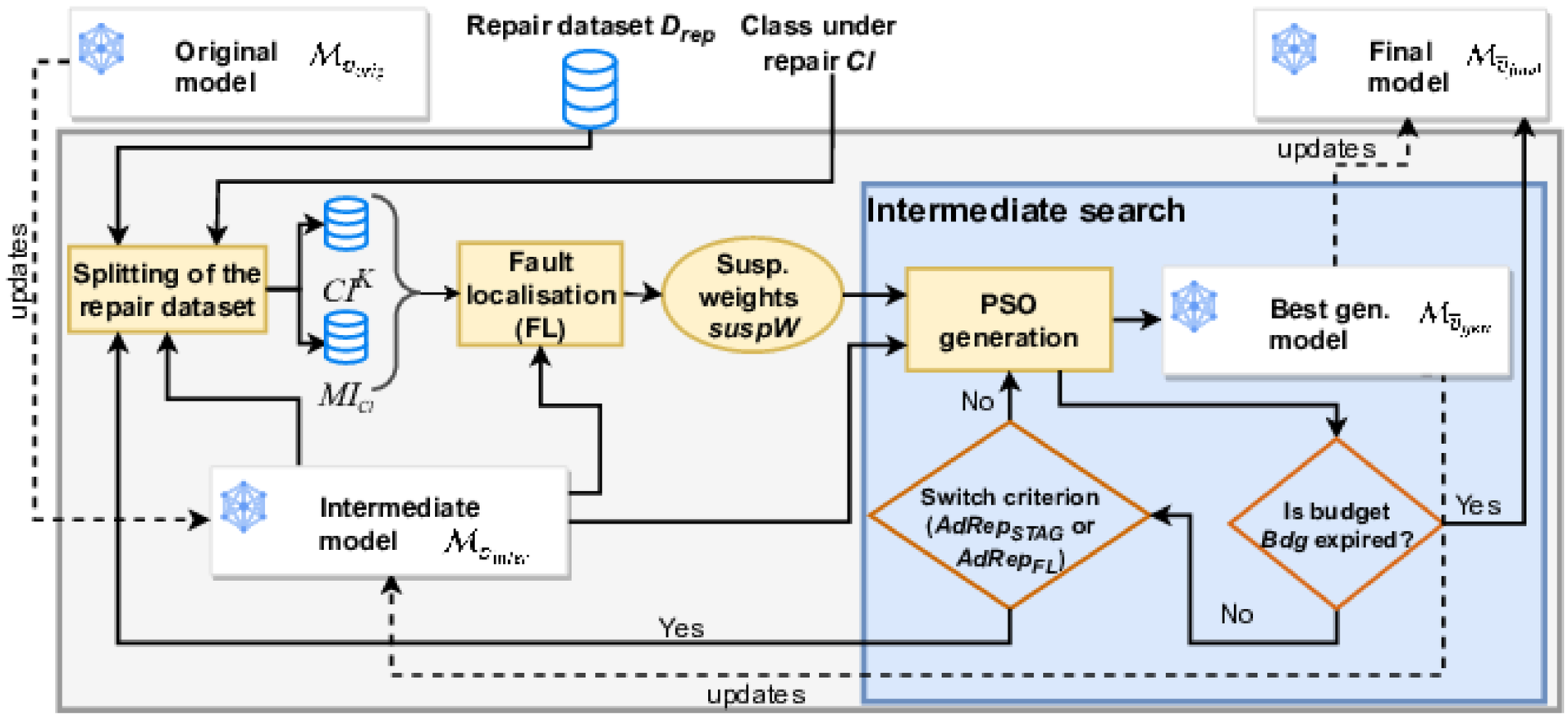
- Each repair is specific and performs well
- The merged repaired model outperforms models repaired with a global (no fault-specificity) repair
- When merging the different fault-specific repairs, we can give them different priorities

Distributed Repair Technique

Motivation

As we repair the system, it changes and the fault(s) can "shift". We might not be repairing the right weights anymore.

Method: Continuously perform fault localisation to make sure we still repair the right weights



Results

- Multiple intermediate searches are performed, i.e., the faults shift
- Adaptive Repair performs better than the baseline approach

D. L. Calsi, M. Duran, T. Laurent, X. i. Zhang, P. Arcaini, F. Ishikawa: Adaptive Search-based Repair of Deep Neural Networks. GECCO 2023

連絡先：国立情報学研究所　アーキテクチャ科学研究系　石川　冬樹
URL：https://research.nii.ac.jp/~f-ishikawa/ Email：f-ishikawa@nii.ac.jp