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

機械学習のテスト・検証 研究論文アップデート

超簡易版・2018秋

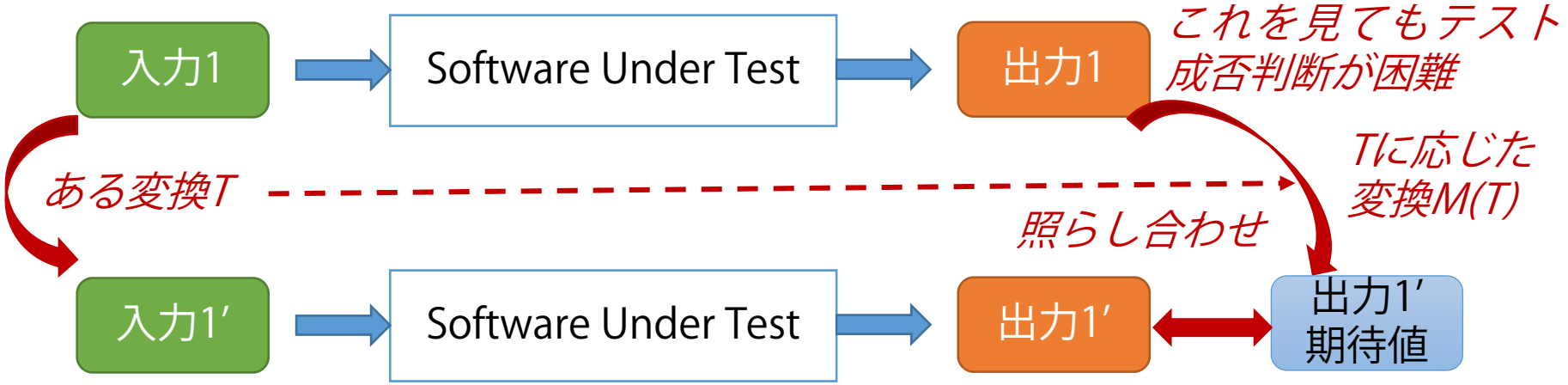
国立情報学研究所　石川　冬樹
f-ishikawa@nii.ac.jp / @fyufyu
http://research.nii.ac.jp/~f-ishikawa/

ＡＩプロダクト
品質保証
コンソーシアム

固有のテスト技術の一例（１）

- メタモルフィックテスティング
  - 多数の入力を用いてテストをしたくても，各出力の正しさ（テスト成否）を判定するのが困難または高コスト

➡ 「入力を変えると出力はこう変わるはず」という関係を検証，既存テストケースから多数のテストケースを生成



例：オススメ商品ランキングを，1位商品を抜いたデータから出し直してみると？

[Segura et al., A Survey on Metamorphic Testing, 2016]

メタモルフィックテスティングの適用

- 訓練アルゴリズム，訓練済みモデル，システム全体のテストへの適用事例
  （ある入力で出力を出してみた後に）
  - ランキング生成： 入力購買データから1位商品を抜く
  - 時系列分析： 入力信号をすべて定数値ずらす
  - ドローン探索プランニング： 地図を回転させる
  - 画像処理： 画像のRGBを入れ替える
  - ・・・

NEW

[Murphy et al., Improving the Dependability of Machine Learning Applications, 2008]
[Jarman et al., Metamorphic Testing for Adobe Data Analytics Software, 2017]
[Lindvall et al., Metamorphic Model-based Testing of Autonomous Systems, 2017]
[中島,データセット多様性のソフトウェア・テスティング, 2017]
[Dwarakanath, Identifying Implementation Bugs in Machine Learning, 2018]

ISSTA'18

Identifying Implementation Bugs in Machine Learning Based Image Classifiers using Metamorphic Testing (Accenture India)

- SVMでの文字認識
  - 例：訓練データの並び替えや定数値付加
    （分類結果だけでなくスコアが変わらない）
- CNNでの画像認識
  - 例：RGB入れ替え，回転，正規化，定数値付加
    （分類結果だけでなくloss, accuracyが変わらない）
- ミューテーション分析により，71%の実装バグを見つけられることを確認

関連して ISSRE'18

DeepMutation: Mutation Testing of Deep Learning Systems (Lei Ma, 九大など)

- その名の通り，ミューテーション分析（テストの評価）
- ソース（訓練データと訓練アルゴリズムコード）をいじると再学習がいるので，訓練済みモデルをいじる手も
- 偏ったテストデータではちゃんと悪いスコアが出る

TABLE I: Source-level mutation testing operators for DL systems.

Fault Type	Level	Target	Operation Description
Data Repetition (DR)	Global Local	Data	Duplicates training data Duplicates specific type of data
Label Error (LE)	Global Local	Data	Falsify results (e.g., labels) of data Falsify specific results of data
Data Missing (DM)	Global Local	Data	Remove selected data Remove specific types of data
Data Shuffle (DF)	Global Local	Data	Shuffle selected training data Shuffle specific types of data
Noise Perturb. (NP)	Global Local	Data	Add noise to training data Add noise to specific type of data
Layer Removal (LR)	Global	Prog.	Remove a layer
Layer Addition (LA_s)	Global	Prog.	Add a layer
Act. Fun. Remov. (AFR_s)	Global	Prog.	Remove activation functions

TABLE II: Model-level mutation testing operators for DL systems.

Mutation Operator	Level	Description
Gaussian Fuzzing (GF)	Weight	Fuzz weight by Gaussian Distribution
Weight Shuffling (WS)	Neuron	Shuffle selected weights
Neuron Effect Block. (NEB)	Neuron	Block a neuron effect on following layers
Neuron Activation Inverse (NAI)	Neuron	Invert the activation status of a neuron
Neuron Switch (NS)	Neuron	Switch two neurons of the same layer
Layer Deactivation (LD)	Layer	Deactivate the effects of a layer
Layer Addition (LA_m)	Layer	Add a layer in neuron network
Act. Fun. Remov. (AFR_m)	Layer	Remove activation functions

固有のテスト技術の一例（２）

- 機械学習モデル（画像識別器や進路判断器等）の「あら探し」テスティング（敵対的サンプル等）
  - 既存画像に雨やノイズを加えることで様々な入力を作る
  - 「ニューロンカバレッジ」を最大化するテストスイートを出す（「作ったものを一通り試す」新指標を利用）
  - これによりテストスイートの多様性・網羅性を上げつつ，「悪いケース」を探す
    - 同じ入力を別バージョンの実装に入れたときと比べて，より大きく出力結果が変わるケース
    - 前述のメタモルフィック関係が満たされない不具合ケース

1.1 original　　1.2 with added rain

[Pei et al., DeepXplore: Automated Whitebox Testing of Deep Learning Systems, 2017]

画像元：[Tian et al., DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars, 2018]

ASE'18

DeepGauge: Multi-Granularity Testing Criteria for Deep Learning Systems（これもLei Ma, 九大など）

- 初期のニューロンカバレッジは雑（全ニューロンが活性化すればOK）
- より粒度の細かい指標
  - 「ニューロン値が様々な範囲の値をとったか」
  - 「各レイヤでTop-kに入ったニューロン組が多様か」
- より効果的なテストの指標

ASE'18

Automated Directed Fairness Testing

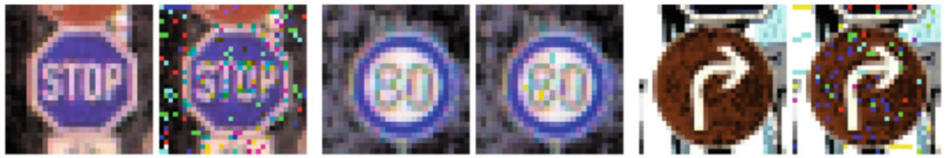
- 給与判断であれば，「性別が違うだけで給与が違う」といった出力をモデルが出すような入力を生成可能

（人種，性別など「これで差別していてはいけない」という特徴を指示すると）

- 全体の中からそういう入力を探した後，近辺を探していく（そういう入力に関する確率的な見積も）
- 再訓練まで
- 速い

形式検証技術の適用

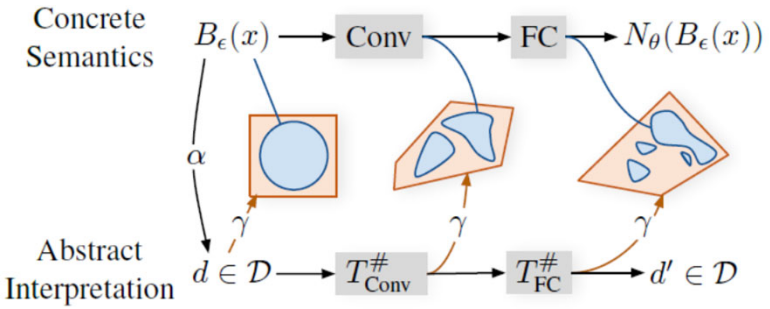
- 形式検証技術による頑健性検証
  - 画像認識において，「一定範囲の画像操作」を行っても認識結果が変わらないかどうかを網羅的に検証
  - SMTソルバーや抽象解釈を応用



"stop" to "30m speed limit"　"80m speed limit" to "30m speed limit"　"go right" to "go straight"

画像元：[Huang et al., Safety Verification of Deep Neural Networks, 2017]

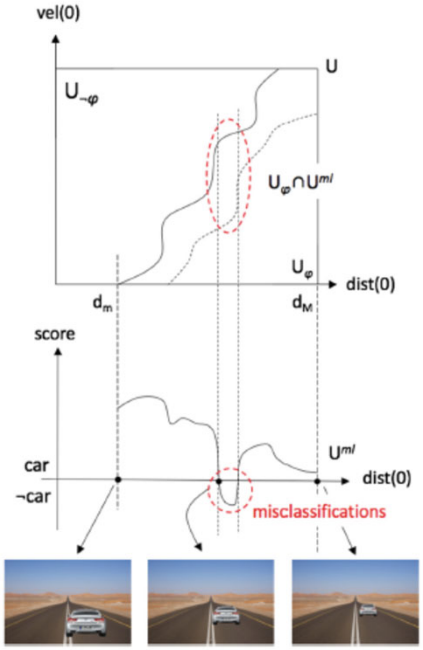
NEW



画像元：[Mirman et al., Differentiable Abstract Interpretation for Provably Robust Neural Networks, 2018]

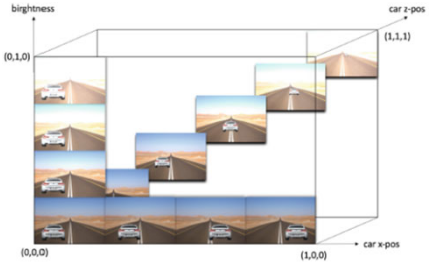
システムレベルの要求に基づくテストの例

- システム全体の要求を踏まえるのが重要？
  - 機械学習部品（生成されたモデル）の精度だけむやみに突き詰めてもしょうがない
  - 例： 遠くの物体を誤認識しても衝突には至らないかも



例題：自動ブレーキシステム（人の操作や先行車の位置が入力，「先行車との距離が一定以上ある」ことが要求）

1．完璧な機械学習部品を用いると要求を満たすが，常に失敗する機械学習部品を使うとそうならないような入力パラメータ領域を絞り込む



2．その領域に限定して機械学習部品が誤認識するような画像を探す

3．その画像を用い要求を満たさないケースを探す

[Dreossi et al., Compositional Falsification of Cyber-Physical Systems with Machine Learning Components, 2017]

HSCC'18

Sim-ATAV: Simulation-Based Adversarial Testing Framework for Autonomous Vehicles
（Toyota US）

- Closed-loopでのシステム全体のテスト