深層ニューラルネットワークに対する分散型修正技術 DistrRep

石川 冬樹　Paolo Arcaini

自動運転など安全性や信頼性が重要な領域においては，リスクが大きいような誤りの発生率を個別に評価するなど，粒度が細かい予測性能評価が求められる．一方，深層ニューラルネットワークなど複雑なモデルに対しては，訓練データを追加して再訓練の調整を行っても，特定の誤り率を下げるような修正は難しい．本発表においては，優先度やリスクレベルを踏まえ，複数の誤り種別に対して深層ニューラルネットワークの修正を行う技術 DistrRep [1] について述べる．DistrRep では，異なる誤り種別ごとに，欠陥局所化による修正パラメーターの絞り込みと最適化を行った上で，誤り種別ごとの修正案をマージする．これにより，予測性能を細かな粒度で評価する場合においても，従来手法よりもその評価に応じた修正を効果的に行うことができる．

1　はじめに

機械学習，特に深層学習技術を用いた画像認識は，自動運転など安全性が重要な領域も含めて多様な応用領域にて活用されるようになっている．安全性を重視する場合，評価データセット全体に対して正解率などの予測性能を考えるだけでは不十分で，リスクが高い事故につながるような誤りの発生率などより，特定の対象や状況に注目した細粒度の指標を用いた評価を行っていく必要がある [2] [3]．また多様な対象・状況に対して安全であるということを論じるために，考えるべき誤りの種別や対応する予測性能の評価指標は複数存在することになる．

深層学習技術を用いた場合，予測モデルである深層ニューラルネットワーク（以降 DNN）は何百万，あるいは何千万というパラメーターから構成され，データを用いた訓練を通してそれらのパラメーターの値を設定していくことになる．このような訓練により，細粒度の評価指標すべてに対して十分な予測性能を達成するのは困難であり，特定の誤り率を下げるなどの調整が必要となる．一方で，訓練データを追加しての再訓練においても，大量のパラメーターを再設定することになり，意図した予測性能の向上が得られるかは不確かであるとともに，他の種別の誤りが増加するような可能性もある．

これらに対し，プログラムに対して行われてきた自動修正（repair）技術のように，誤りの要因となる箇所を局所化し，その限られた箇所の修正を試みるという DNN 修正技術 [4] [5] が取り組まれてきた．このアプローチでは，DNN 内に含まれ一般的には訓練により設定されるパラメーターのうち，致命的な出力の誤りなど修正したい挙動に影響しているものを同定し，それらに絞って最適化を行うことで DNN の局所的な修正を行う．しかし既存技術では，複数種別の誤りを同時に修正することは，技術としても評価としても考えていない．

本発表では，以上に対して著者らが取り組んだ DistrRep 技術 [1] について述べる．DistrRep では，異なる誤り種別ごとに，欠陥局所化による修正パラメーターの絞り込みと最適化を行った上で，誤り種別ごとの修正案をマージする．これにより，予測性能を細かな粒度で評価する場合においても，従来手法よりもその評価に応じた修正を効果的に行うことができる．

This talk abstract is based on the published paper “Distributed Repair of Deep Neural Networks” at ICST 2023 [1].

Fuyuki ISHIKAWA and Paolo Arcaini, 国立情報学研究所, National Institute of Informatics.

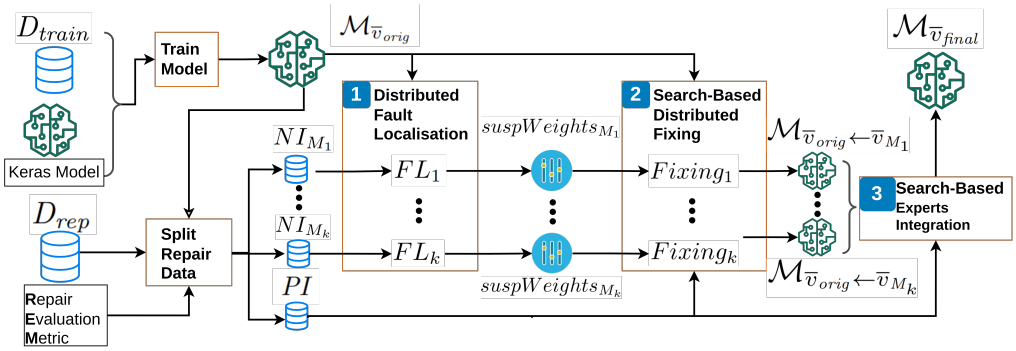

図 1　DistrRep 技術のワークフロー（[1] より引用）

2　DistrRep 技術

本稿では DistrRep の技術および評価について概要を抜粋して述べる．形式的な定義や詳細については文献 [1] を参照いただきたい．

2.1　技術の概要

図 1 に DistrRep における処理ワークフローを示す．図左上では一般的な訓練によりモデル $M_{\bar{v}_{orig}}$ が得られている．これに対して図左下では，修正を行うための追加データ D_{rep} と，修正の評価基準となる REM (Repair Evaluation Metric) が修正処理の入力として与えられている．REM としては例えば，運転シーンにおける物体の分類タスクにおいて，歩行者をバイク搭乗者と誤分類する誤り率，他の物体と誤分類する誤り率などの複数指標を，リスクの大きさを加味して一つの指標として統合したものを想定している．

これらに対して，まず準備工程となる “Split Repair Data” という部分では，REM に含まれる異なる種別の誤りが発生するような入力データを追加データ D_{rep} から選び，$NI_{M_1}, \cdots, NI_{M_k}$ へと分割している．また正しく分類することができた入力データは PI として保存しておく．

ステップ 1 では，各誤り種別に対し，その要因となっているパラメーター群を分析する．ステップででは，それらを修正する最適化を行うことで，それぞれの誤りに特化した修正モデル群 $M_{\bar{v}_{orig} \rightarrow \bar{v}_{M_1}}, \cdots, M_{\bar{v}_{orig} \rightarrow \bar{v}_{M_k}}$ を得ている．各誤り種別に対してそれぞれ行われるこのステップ 1 および 2 の処理は，既存の DNN 修正技術に該当する．ステップ 3 では，REM を踏まえて得られた修正モデル群の統合を最適化を用いて行う．

2.2　評価の概要

DistrRep に対する評価においては，運転シーンに関する既存のデータセットから，物体部分の部分画像を抽出することで，誤分類の種別によるリスクが異なる物体の分類を行うデータセットを構築した．分類モデルとしては 3 つのモデルを用いたが，ここでは EfficientNetB7 によるものを表 1 に抜粋する．

他の比較対象に対して，DistrRep のみが REM の値を向上できていることが示されている．一方で，すべてのサンプルを同等に評価するような，平均的な精度 (Accuracy) については，DistrRep は悪化させている．DistrRep は，安全性を重視するような場合に有効であることが見てとれる．

比較対象の詳細としては，再訓練について四種類用意している．NW と W という下添字があるものは，REM を踏まえて重み付きの学習目標（損失関数）を用いたかどうかを表している（non weighted および weighted の意）．上添字の rep と rep+tr については，追加データを用いて追加訓練する場合と，追加データと元の訓練データをあわせて再度訓練を行う場合とを表している．また Arachne は，従来の DNN 修正技術として比較対象に含めている．こちらについても，既存技術をそのまま用いるものと，REM を目標にするように変更したものを用いている．

Approach	REM			Accuracy (%)		
	min	max	avg	min	max	avg
$\text{ReTr}^{rep}_{\text{NW}}$	-1.09	1.20	0.33	-0.44	0.25	0.13
$\text{ReTr}^{rep}_{\text{W}}$	-1.14	1.36	0.17	-0.22	0.16	-0.05
$\text{ReTr}^{rep+tr}_{\text{NW}}$	0.28	1.37	0.87	0.13	0.31	0.22
$\text{ReTr}^{rep+tr}_{\text{W}}$	-1.51	0.63	-0.18	-1.42	-0.06	-0.35
Arachne	-2.57	1.03	-1.09	-19.92	-6.30	-11.77
$\text{Arachne}_{\texttt{REM}}$	0.38	2.94	1.44	-20.38	-15.24	-18.30
DistrRep	6.79	8.42	7.57	-7.82	-4.18	-5.80

表 1　DistrRep の評価抜粋（[1] より引用）

3　おわりに

本発表では，優先度やリスクレベルを踏まえ，複数の誤り種別に対して深層ニューラルネットワークの修正を行う技術 DistrRep [1] について述べた．今後，産業界のニーズや課題により踏み込んで技術の発展とさらなる実践的な評価を行っていきたい．

謝辞

本研究は，JST 未来社会創造事業 JPMJMI20B8（通称 Engineerable AI プロジェクト）の支援により実施された．

参 考 文 献

[1] Calsi, D. L., Duran, M., Zhang, X.-Y., Arcaini, P., and Ishikawa, F.: Distributed Repair of Deep Neural Networks, *2023 IEEE Conference on Software Testing, Verification and Validation (ICST)*, 2023, pp. 83–94.

[2] Salay, R., Angus, M., and Czarnecki, K.: A Safety Analysis Method for Perceptual Components in Automated Driving, *2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE)*, 2019, pp. 24–34.

[3] Salay, R., Czarnecki, K., Kuwajima, H., Yasuoka, H., Abdelzad, V., Huang, C., Kahn, M., Nguyen, V. D., and Nakae, T.: The Missing Link: Developing a Safety Case for Perception Components in Automated Driving, Technical report, SAE, 2022.

[4] Sohn, J., Kang, S., and Yoo, S.: Arachne: Search-Based Repair of Deep Neural Networks, *ACM Trans. Softw. Eng. Methodol.*, Vol. 32, No. 4(2023).

[5] Tokui, S., Tokumoto, S., Yoshii, A., Ishikawa, F., Nakagawa, T., Munakata, K., and Kikuchi, S.: NeuRecover: Regression-Controlled Repair of Deep Neural Networks with Training History, *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, Los Alamitos, CA, USA, IEEE Computer Society, mar 2022, pp. 1111–1121.