F-035

データシフトによる分布外汎化性能と不確実性への影響

多田 圭吾 ♯1,a)　長沼 大樹 ♯2,3,b)
Tada Keigo[1,a)]　Naganuma Hiroki[2,3,b)]

1. はじめに

一般的な機械学習の問題設定では、訓練データを用いた経験損失最小化を行うことで、未知のデータに関する推論がうまく機能することを期待する [16]。しかしながら、実応用では、学習・推論時の環境は異なることが一般的であるため、分布外汎化問題 (Out-of-Distribution Generalization) [2] に取り組むことは、深層学習の実応用に不可欠である。また、昨今の成功によって人気を博している深層ニューラルネットワーク (DNN) は、予測のランキング性能を著しく向上させたものの、古典的な統計的モデルに比べ、その不確実性が信頼できるものでないことが問題視されている [6]。特に、オンライン広告入札システムや医療画像診断などのシナリオでは、意思決定において不確実性の信頼性が不可欠である。分布外汎化には、不確実性の較正が有効であることが示されているが [17]、扱う分布シフトの種類によってその様相が異なる可能性がある。

本研究では、教師あり分類タスクを対象とし、訓練データの分布からテストデータの分布にシフトが存在する実用的な問題設定において、複数の種類の分布のシフトによるランキング性能及び不確実性への影響の検証を行った。

2. 導入

2.1 分布外汎化

深層ニューラルネットワークの教師あり学習における一般的な問題設定においては、訓練データとテストデータが同じ分布から抽出されるという教師あり学習の一般的な仮定に基づいており、正則化付き経験損失の最小化によってテストデータの損失である汎化損失を最小化することを期待する。実応用では、テストデータの分布は訓練データの分布と異なることが一般的であり、この分布の差異、すなわち学習・推論時の環境変化は、従来の教師あり学習の前提に反するものである。

具体例を示すと、広く用いられる ImageNet-1K [7] データセットに含まれる「ラクダ」の画像は砂漠が背景であることが多く、このデータセットを訓練データとテストデータに分割して性能を評価する場合、モデルが砂漠を背景に含む画像に対して「ラクダ」という推論を行なった場合も、テストデータでの高い汎化性能を達成できる。

図 1 に示す通り、学習・推論時の環境変化が起きない場合は、先に示した Shortcut な特徴量によっても、テストデータを高い精度で推論可能であるが、分布外においては、汎化に失敗することが報告されている [4]。

2.2 不確実性

統計的モデルの評価指標として、精度として知られるランキング性能だけでなく、医療画像診断などのシナリオにおいては不確実性が重要となる [1]。例えば，同じ認識精度のモデルであっても，医療画像診断によって癌の腫瘍摘出範囲を推定する Segmentation タスクなどのシナリオでは、確信度が 99%と 50%では意思決定における意味合いが異なる。

ここで、確信度が正しい（すなわち確信度 90% のデータが 100 件あった場合、90 件は正解である）ことが意思決定には欠かせないが、昨今の深層ニューラルネットワークモデルは、確信度と精度にギャップがあることが実応用への障壁となっている [6]。この差異を定量化する指標として、Expected Calibration Error (ECE) が広く用いられている [15]。

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}\left(B_m\right) - \mathrm{conf}\left(B_m\right)\right| \tag{1}$$

ただし、$\mathrm{acc}\left(B_m\right)$, $\mathrm{conf}\left(B_m\right)$ は以下で定義される。

$$\mathrm{acc}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\left(\hat{y}_i = y_i\right) \tag{2}$$

$$\mathrm{conf}\left(B_m\right) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i \tag{3}$$

ここで、B_m は予測確信度 $\hat{p}_i$ を、ビンに区切った m 番目の集合であり、y_i は正解ラベルを、$\hat{y}_i$ は予測ラベルを示す。

3. 分布シフト

2 章で紹介した関連研究は、分布シフトを分類せずに

1 立命館大学 / Ritsumeikan University
2 モントリオール大学 / Université de Montréal
3 Mila / Montreal Institute for Learning Algorithms
a) is0463hx@ed.ritsumei.ac.jp
b) naganuma.hiroki@mila.quebec
♯ denotes equal contribution

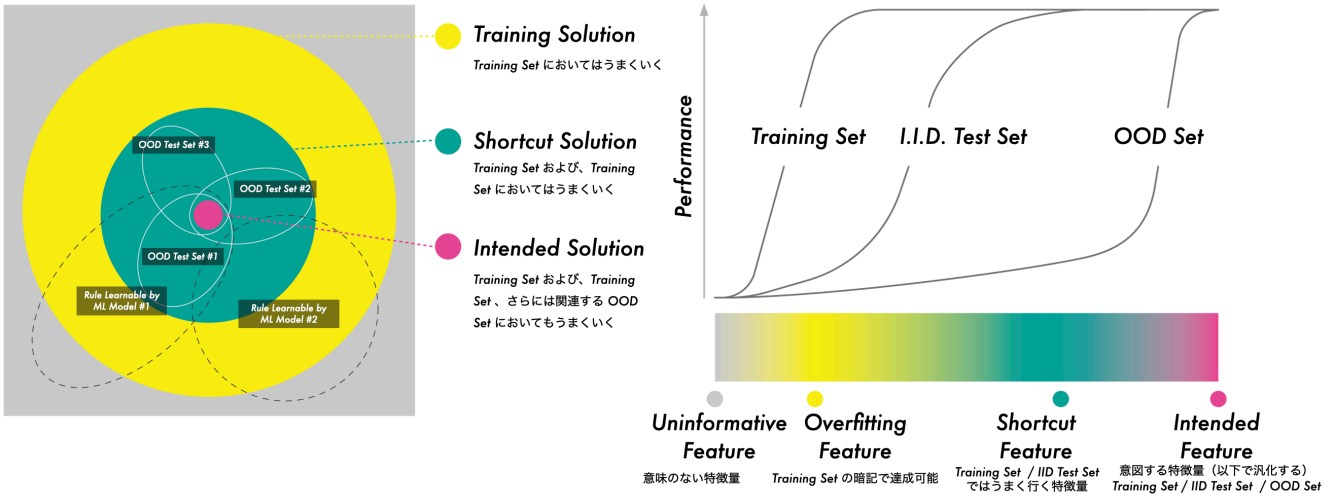

図 1: 解の分類: Training Set で性能が出る解の範囲において、Training Set のみに対して性能が出る解 (黄色)、Training Set と IID の Test Set に対しても性能が出る解 (緑)、OOD においても性能が出る解 (ピンク) に分類できる。例えば、Shortcut Deature の学習では分布内汎化には成功するが分布外汎化には失敗する。

扱っているため、どのような分布シフトによる影響であるかが定かではない。Ye ら [18] は、 Correlation Shift と Diversity Shift の二軸によって OOD のデータセットを分類することを試みているが、一般的に機械学習における文脈で、分布シフトは以下の三つに大別される（サンプリングのバイアスやその他の項目も含め、厳密な議論は [14] を参照されたい)。ただし、p を source, q を target の分布とし、y は出力、x は入力の分布である。

Covariate Shift

$$p(y|x) = q(y|x) \tag{4}$$
$$p(x) \neq q(x) \tag{5}$$

Label Shift (Prior Probability Shift)

$$p(x|y) = q(x|y) \tag{6}$$
$$p(y) \neq q(y) \tag{7}$$

Concept Shift

$$p(x|y) \neq q(x|y) \tag{8}$$
$$p(y|x) \neq q(y|x) \tag{9}$$

本研究では、画像認識における実社会のアプリケーションにおける、一般的な分布のシフトとして Covariate Shift を対象として、シフトごとの分布外汎化・不確実性の評価のため、表 1 の分類法の通り、さらに細かなシフトに区分する。

Background-Shift は画像の背景情報を他の画像で差し替えるシフトであり、Rotation-Shift は、対象のオブジェクトを回転のみさせたシフトである。画像認識で一般に用いられる Convolutional Neural Network は、特徴量の移動不変性を持つことが知られているが、回転やスケールに対しては頑健ではない [13]。Corruption-Shift 及び Perturbation-Shift、画像を観測する条件の変化 (天候やカメラレンズのの劣化) や敵対的サンプル [10] に代表される

	Background	Rotation	Pixel	Shape	Structure
Background-Shift		✓	✓	✓	✓
Rotation-Shift	✓		✓	✓	✓
Corruption-Shift				✓	✓
Perturbation-Shift					✓
Style-Shift					

表 1: 本研究で対象とする分布シフトの分類表: ✓ はシフト後も分布での画像が、元の分布から保持している視覚的特徴を示す。

ノイズによってもたらされるシフトであり画像の Pixel 情報や Shape 自体に影響を及ぼす。Style-Shift は、y としては同じ概念であるが、x の Structure 自体にも変化を及ぼすシフトである（例えば、写真の猫とイラストの猫のようなシフト)。

4. 実験

異なる分布シフトが分布外汎化性能・不確実性に及ぼす影響を調査するため、元分布として、MNIST データセットで ERM により学習を行なった DNN モデル (3 種類) に対し、分布シフト先での汎化性能 (認識精度、ランキング性能) と、不確実性の指標として 2.2 章で紹介した ECE を用いた評価を行う。本研究における実験で利用するデータセットと、分布シフトの関係を表 2 に示す。

図 2 に示すとおり、Perturbation-Shift はベースライ

Shift	Dataset
IID	MNIST [9]
Background-Shift	MNIST Background Image [8]
Rotation-Shift	Rotated MNIST [5]
Corruption-Shift	MNIST-C [11]
Perturbation-Shift (Global*1)	Morpho-MNIST [3]
Perturbation-Shift (Local*2)	Morpho-MNIST [3]
Style-Shift	SVHN [12]

表 2: 本研究で対象とするデータセットと分布シフトの対応関係

*1 thinning+thickening
*2 swelling+fractures

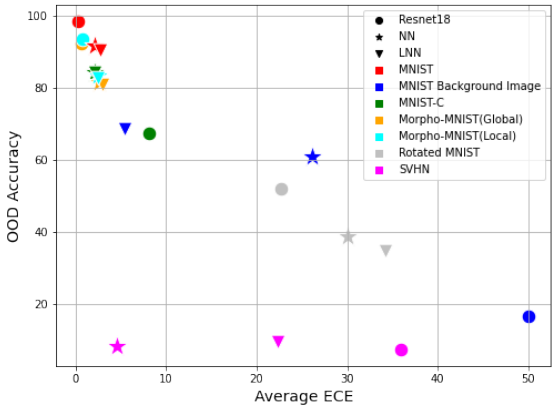

図 2: 異なるシフトにおける OOD 性能 (↑) と ECE(←) の関係: MNIST で学習した DNN を異なるシフト下において評価した。

ンの MNIST とほぼ変わらない性能を示すのに対して、Corruption-Shift や、Rotation-Shift は OOD 性能・ECE 共に劣化し、Style-Shift や Background-Shift は深刻な劣化を招く結果となった。Foreground の特徴量のシフトに関しては、OOD 性能が劣化すると、同時に ECE が劣化することが観測された他、昨今の ResNet18 などのモダンな DNN においては、Background-Shift は Style-Shift に比べ OOD 性能は劣化しないものの、ECE が大きく劣化する結果となった。

5. おわりに

本研究は、近年広く産業応用の期待される深層ニューラルネットワークの、実応用を想定した分布シフト下において、その分布外での性能と不確実性に及ぼす影響について検証した。実験の結果、分布シフトの種類によって、分布外汎化性能の劣化や不確実性への影響の度合いが異なることを明らかにした。特に、不確実性が重要なアプリケーションにおいて、ResNet18 などのモダンな DNN における Background Shift による不確実性の信頼度を必ずしも OOD 性能で評価できないことを示唆する結果となった。

最後に、本研究の限界として、次の 3 点が挙げられる。i) 実世界のアプリケーションとして Covariate Shift と同様に多くの問題設定が考えられる Label Shift などの下での検証が必要である。ii) 本研究では、MNIST のような小さなスケールのデータセット及び小さな DNN モデルでしか検証を行っていないため、ImageNet のような大きなスケールのデータセットや、巨大な DNN モデルでは様相が異なる可能性がある。iii)ERM 以外の、OOD 向けの学習アルゴリズムを利用した場合の検証を行う必要がある。

参考文献

[1] Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R. et al.: A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Information Fusion*, Vol. 76, pp. 243–297 (2021).

[2] Arjovsky, M.: Out of Distribution Generalization in Machine Learning (2021).

[3] Castro, D. C., Tan, J., Kainz, B., Konukoglu, E. and Glocker, B.: Morpho-MNIST: quantitative assessment and diagnostics for representation learning, *Journal of Machine Learning Research*, Vol. 20, No. 178, pp. 1–29 (2019).

[4] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. and Wichmann, F. A.: Shortcut learning in deep neural networks, *Nature Machine Intelligence*, Vol. 2, No. 11, pp. 665–673 (2020).

[5] Ghifary, M., Kleijn, W. B., Zhang, M. and Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders, *Proceedings of the IEEE international conference on computer vision*, pp. 2551–2559 (2015).

[6] Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q.: On calibration of modern neural networks, *International Conference on Machine Learning*, PMLR, pp. 1321–1330 (2017).

[7] Krizhevsky, A., Sutskever, I. and Hinton, G. E.: Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems*, Vol. 25 (2012).

[8] Larochelle, H., Erhan, D., Courville, A., Bergstra, J. and Bengio, Y.: An empirical evaluation of deep architectures on problems with many factors of variation, *Proceedings of the 24th international conference on Machine learning*, pp. 473–480 (2007).

[9] LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324 (1998).

[10] Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A.: Towards deep learning models resistant to adversarial attacks, *arXiv preprint arXiv:1706.06083* (2017).

[11] Mu, N. and Gilmer, J.: Mnist-c: A robustness benchmark for computer vision, *arXiv preprint arXiv:1906.02337* (2019).

[12] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B. and Ng, A. Y.: Reading digits in natural images with unsupervised feature learning (2011).

[13] Ngiam, J., Chen, Z., Chia, D., Koh, P., Le, Q. and Ng, A.: Tiled convolutional neural networks, *Advances in neural information processing systems*, Vol. 23 (2010).

[14] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A. and Lawrence, N. D.: *Dataset shift in machine learning*, Mit Press (2008).

[15] Roelofs, R., Cain, N., Shlens, J. and Mozer, M. C.: Mitigating bias in calibration error estimation (2021).

[16] Vapnik, V.: Principles of risk minimization for learning theory, *Advances in neural information processing systems*, Vol. 4 (1991).

[17] Wald, Y., Feder, A., Greenfeld, D. and Shalit, U.: On calibration and out-of-domain generalization, *Advances in Neural Information Processing Systems*, Vol. 34 (2021).

[18] Ye, N., Li, K., Hong, L., Bai, H., Chen, Y., Zhou, F. and Li, Z.: Ood-bench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms, *arXiv preprint arXiv:2106.03721* (2021).