

# 深層学習における不確かさ評価の重要性

## Importance of Uncertainty Estimation in Deep Learning

前田 巍<sup>\*1</sup>  
Iwao Maeda

松島 裕康<sup>\*1</sup>  
Hiroyasu Matsushima

坂地 泰紀<sup>\*1</sup>  
Hiroki Sakaaji

和泉 潔<sup>\*1</sup>  
Kiyoshi Izumi

ディグロー デビット<sup>\*2</sup>  
David deGraw

富岡 博和<sup>\*2</sup>  
Hirokazu Tomioka

加藤 慎雄<sup>\*3</sup>  
Atsuo Kato

北野 道春<sup>\*3</sup>  
Michiharu Kitano

<sup>\*1</sup>東京大学  
The University of Tokyo

<sup>\*2</sup>大和証券株式会社  
Daiwa Securities Co. Ltd.

<sup>\*3</sup>株式会社大和総研  
Daiwa Institute of Research Ltd.

In recent years, predictions by machine learning and deep learning methods are utilized in various scenes of society. A model trained with deep learning methods can predict the target with high accuracy, but can not consider the predictive confidence sufficiently, and may predict high confident for extrapolated data which is hard to predict. In this study, we applied ordinary deep learning methods and methods considering predictive uncertainty, proposed in recent years, to an image classification task, and verified the robustness of trained models against extrapolated data. Models trained with the ordinary deep learning methods predicted high confidence values for data having characteristics not existing in the training data, but models trained with the methods considering uncertainty predicted low confidence values for such data. By using methods considering uncertainty, it is possible to avoid mispredictions for extrapolated data. Experimental results suggest the importance of uncertainty estimation in deep learning.

### 1. はじめに

近年の機械学習研究の発展に伴い、社会の様々な場面で機械学習モデルによる予測が活用されている。特に深層学習(Deep Learning)を用いた予測は画像識別や機械翻訳、強化学習等の分野において目覚ましい成果を上げている。深層学習は多数の線形変換および非線形変換を組み合わせることにより高い汎化性能を実現し、対象データの持つ潜在的な特徴を学習することができる。しかし、予測信頼性という観点においては、現状の深層学習手法による予測では不十分な部分が多い。

予測信頼性は予測精度と異なる概念である。予測精度はモデル検証用データに対する予測結果が正解にどれだけ一致しているかで評価され、クラス分類問題の場合 Accuracy や F1 score といった指標で定量的に評価される。ここで重要なのは、モデルの学習および評価が内挿データのみによってなされており、外挿データ(予測の困難なデータ)に対する予測が考慮されていないことである。

図 1. に外挿データに対する予測の例を示す。図 1. の実験では手書き数字データセットである MNIST[Lecun 1998] を用いて学習した畳み込みニューラルネットワーク(CNN)モデルで、数字以外の画像データを予測している。各画像下の棒グラフが MNIST データセットに含まれるそれぞれのクラスに対する予測の確信度(confidence)を示している。MNIST データセットを用いて学習したモデルは 0 から 9 までの数字以外の画像を識別できないため、そのような画像に対しては予測の不確かさが大きく(confidence が小さく)出力されることが望ましい。しかし、図 1. に示す全ての画像に対し、CNN モデルは高い confidence で予測を行ってしまっている。CNN モデルは一般的に画像識別の性能が非常に高く、図 1. で用いた CNN モデルもモデル検証用データについてほぼ完璧に正解を予測することができる。一方でどの数字にも属さないような外

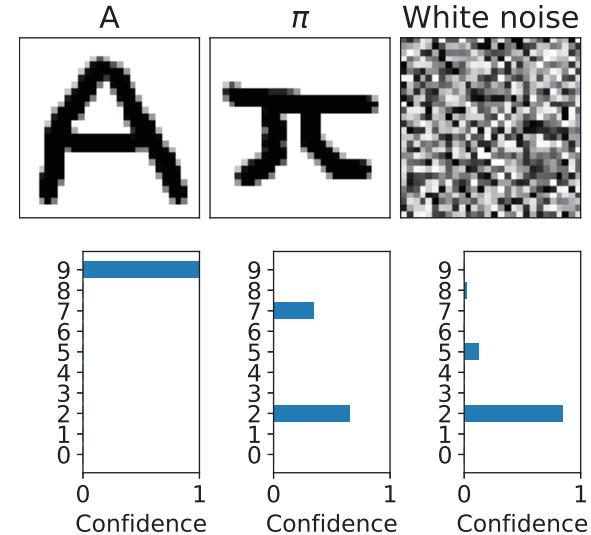


図 1: MNIST データセットで学習した畳み込みニューラルネットワークモデルによる、数字以外の画像に対する予測結果。

挿データの入力に対し脆弱であり、誤った予測を大きな確信を持って行ってしまう危険がある。

今後、深層学習手法がより広い領域の課題に応用されることが期待される。その際に、深層学習モデルが人間の手に頼らずに入力データに対する予測の確実性、あるいは予測の不確かさを見積もることが重要となる。深層学習モデルが適切に不確かさ評価を行うことができれば、不確かさの大きい領域では意思決定を行わない等の対処が可能となり、深層学習モデルの予測がより信頼のおけるものとなる。

本研究では、同じネットワーク構造を持つ深層学習モデルに対し、一般的な学習方法および不確かさを考慮した学習方法を

適用し、内挿データおよび外挿データに対する予測結果を比較した。不確かさを考慮した学習方法を適用することで外挿データに対し頑健になり、より信頼性の高い予測が行えることを確認した。

## 2. 手法

### 2.1 置み込みニューラルネットワーク

学習を行う深層学習モデルとしては、置み込みニューラルネットワーク (Convolutional Neural Networks, CNN)[Krizhevsky 2012] を採用した。CNN は画像データや系列データといったパターン情報を持つ入力に対し適用され、入力データ内のそれぞれの領域のパターンをもとに予測を行う。予測が入力データに対し位置普遍性を有するため特に画像認識に優れ、pooling や batch normalization といった手法と組み合わせることで高い予測精度を実現できる。

本研究では表 1 に示すネットワーク構成のモデルを用いた。ここで表の各列は層の種類、フィルターの大きさ、ストライドおよび出力の次元を表し、convolution, max pool, dense はそれぞれ置み込み層、max pooling 層、全結合層を表している。各入力データに対し最終層の出力は 10 次元のベクトルとして得られ、その各次元が 0 から 9 までの各クラスに対応している。

表 1: 実験で用いた深層学習モデルのネットワーク構成

| type        | filter size  | stride | output size              |
|-------------|--------------|--------|--------------------------|
| convolution | $3 \times 3$ | 1      | $28 \times 28 \times 64$ |
| convolution | $3 \times 3$ | 1      | $28 \times 28 \times 64$ |
| max pool    | $3 \times 3$ | 2      | $14 \times 14 \times 64$ |
| convolution | $3 \times 3$ | 1      | $14 \times 14 \times 64$ |
| convolution | $3 \times 3$ | 1      | $14 \times 14 \times 64$ |
| max pool    | $3 \times 3$ | 2      | $7 \times 7 \times 64$   |
| dense       |              |        | 10                       |

多くの場合、最終層の出力は正規化を経て最終的なモデルの出力値となる。最も一般的な正規化は以下に示す softmax function である。

$$\text{softmax}(y_i) = \frac{\exp(y_i)}{\sum_j \exp(y_j)} \quad (1)$$

ここで、 $\sum_j \exp(y_j)$  は全てのクラスに対する出力値の指数の和を表している。Softmax function を通すことで、各クラスに対する出力値の和が 1 に正規化され、出力値を各クラスに対する予測確率 (predicted probability) あるいは予測信頼性 (confidence) とみなすことが可能となる。

Softmax function による正規化は、入力データが必ず出力クラスのいずれかに属することを前提としている。したがって図 1. で示したような外挿と考えられる入力データに対しても、高い confidence で予測を行ってしまう。外挿データがモデルに入力されることが想定される場合には、softmax function 以外の正規化を用いることを検討すべきである。以下に代替として利用可能な正規化の 1 つとして sigmoid function を示す。

$$\text{sigmoid}(y_i) = \frac{1}{1 + \exp(-y_i)} \quad (2)$$

Sigmoid function を通すことでの出力値はそれぞれ  $(0, 1)$  の区間に正規化され、各クラスに対する出力はデータが各クラ

スに属する確率と対応する。Softmax function の出力が他のクラスを含めた相対的な値であるのに対し、sigmoid function の出力は他のクラスに影響されない値であり、入力データに対し全てのクラスの出力値が小さくなる（すなわち、どのクラスにも属さないと予測する）ことが可能である。ただし、各クラスに対する出力値の和が 1 を超えてしまい、予測確率として解釈できなくなることも起こりうる。

深層学習を用いた識別モデルの学習には、以下に示すクロスエントロピー  $H$  を損失関数として用いることが一般的である。

$$H(X, y) = - \sum_i^N y_i \log(f(X_i)) \quad (3)$$

ここで、 $X, y$  はそれぞれ入力データおよび one-hot 形式の正解ラベル、 $f(\cdot)$  はニューラルネットワークを表している。クロスエントロピーは予測精度の向上に対しては大きく寄与する一方、予測信頼性を考慮した学習を行うことは難しい。

### 2.2 不確かさ学習手法

予測の不確かさあるいは予測信頼性を考慮したニューラルネットワークの学習手法は近年多く提案されている。

通常のニューラルネットワークに対し、学習および推論の調整を施すことにより、予測信頼性評価の妥当性を向上させる手法が提案されている。学習済みのニューラルネットワークモデルに対し、[Guo 2017] では、出力値の較正 (calibration) を行うことにより、モデルの不確かさ評価がより妥当になることが報告されている。[Szegedy 2016] では、学習時に label smoothing(正解ラベルの線形変換) を行うことにより、過剰に予測信頼性が大きくなることを抑制することができると述べられている。

モデルが予測できないデータを想定し、ニューラルネットワークの学習を行う手法も提案されている。[Thulasidasan 2019] では、出力に「予測不可能」クラスを追加し、入力に対し「予測不可能」クラスが予測されすぎないように制約をかけながらモデルを学習する手法 (deep abstaining classifiers) が提案されている。[Sensoy 2018] では、ニューラルネットワークにより入力データの各クラスに対する evidence を予測し、予測した evidence を元に最終的な予測および不確かさの計算を行う手法 (evidential deep learning) が提案されている。

さらに近年では、ニューラルネットワークでベイズ推論を行う Bayesian Neural Networks (BNN) [Blundell 2015] が提案されている。BNN により近似した出力値の事後分布を用いて、確率的に出力値の予測を行うことができる。BNN の推論および学習手法としては、reparameterization gradient[Kingma 2014] や、variational dropout[Kingma 2015] およびその発展形である sparse variational dropout[Molchanov 2017] が提案されている。

## 3. 実験

手書き数字データセットである MNIST を用いて実験を行った。データセット内の各画像は 28 pixel × 28 pixel のモノクロ画像であり、各画像には 0 から 9 の整数の内いずれか 1 つのラベルが割り当てられている。モデル学習用データ 60000 件、モデル検証用データ 10000 件の全 70000 データを用いた。

MNIST データセットに対し、通常の CNN モデル、出力値の較正 (calibration) を行った CNN モデル、label smoothing を行った CNN モデル、evidential deep learning モデル、sparse variational dropout CNN(VDCNN) モデル、calibration を

行った VDCNN モデル、および label smoothing を行った VDCNN モデルで学習を行い、結果を比較した。evidential deep learning モデル以外については、出力の正規化に softmax function および sigmoid function それぞれを用いた場合について実験を行った。全てのモデルは表 1 に示す同一のネットワーク構造を持つ。

### 3.1 予測精度

モデル検証用データに対する各モデルの予測精度を表 2 に示す。手法 (Method) の列はそれぞれ、CNN が通常の畳み込みニューラルネットワークモデル、CNN + LS が label smoothing を行った CNN モデル、Evidential CNN が evidential deep learning を用いて学習した CNN モデル、VDCNN が sparse variational dropout CNN モデルを表す。Calibration に関しては、予測クラスが calibration を行わないモデルと同一であるため省略している。Activation は出力層の正規化に用いる関数 (softmax function あるいは sigmoid function) を示している。F1 score はモデル検証用データに対する F-measure の値を表している。ECE は Expected Calibration Error [Naeini 2015] の略であり、モデル検証用データについて、モデルが予測した信頼性と実際の予測精度との間の解離度を表す指標である。表 2 の結果を見ると、F1 score においては evidential CNN 以外の手法でほとんど差がなく、ECE においては label smoothing を施さない CNN および sparse variational dropout CNN モデルが特に優秀な値を示している。MNIST データセットは学習が比較的容易であり、モデル検証用データについてもほぼ完璧に正解を予測できるため、過剰に大きな confidence の値を出力するモデルが良い結果となっていると解釈できる。

表 2: MNIST データセットにおける、モデル検証用データに対する各モデルの予測精度

| Method         | Activation | F1 score | ECE    |
|----------------|------------|----------|--------|
| CNN            | Softmax    | 0.9947   | 0.0042 |
| CNN            | Sigmoid    | 0.9947   | 0.0042 |
| CNN + LS       | Softmax    | 0.9946   | 0.0625 |
| CNN + LS       | Sigmoid    | 0.9948   | 0.0374 |
| Evidential CNN |            | 0.9642   | 0.0208 |
| VDCNN          | Softmax    | 0.9948   | 0.0035 |
| VDCNN          | Sigmoid    | 0.9951   | 0.0043 |
| VDCNN + LS     | Softmax    | 0.9957   | 0.0656 |
| VDCNN + LS     | Sigmoid    | 0.9954   | 0.0422 |

F1 score および Expected Calibration Error はいずれもモデル検証用データに対し計算されるものであり、外挿データに対する予測の頑健性は評価できていない。

図 2 に、MNIST データセット内の 1 つの数字画像を角度を変えて回転させ、各モデルに入力した場合の confidence の予測結果を示す。1 列目に通常の CNN モデルを元にした手法の予測結果を、2 列目に sparse variational dropout CNN を元にした手法の予測結果を、3 列目に evidential CNN の予測結果を、4 列目に回転角度に対応する入力画像を表示している。今回入力に用いた 6 の画像について、角度 0 rad および  $\pi$  rad 付近ではそれぞれ目視でも 6 および 9 の数字と認識でき、モデルによる confidence の出力値も大きいことが望ましい。図 2 の 1 列目および 2 列目を見ると、calibration および label smoothing を行わないモデルは、ほぼ全ての角度において confidence の値が 1 に近くなっていることがわかる。Calibration および label smoothing をすることで confidence

の出力値は角度に対し滑らかに変化するようになり、予測の不確かさをより妥当に学習できていることがわかる。3 列目に示す evidential CNN では、角度 0 rad および  $\pi$  rad 付近を除く多くの領域で confidence の値が 0 になっており、信頼性の高い予測が行えないことを適切に判断できている。

### 4. おわりに

本研究では、深層学習モデルを用いた予測における不確かさ評価の重要性について、画像識別データセットに対する実験を通して検証を行った。通常の手法により学習した深層学習モデルは内挿データに対する予測精度のみしか考慮できておらず、予測の困難な外挿データに対し確信度の大きい誤った予測を行ってしまう危険性がある。予測の不確かさを考慮した深層学習モデルの学習により、外挿データに対しても予測信頼性を適切に評価でき、より信頼のおける予測が行えると期待される。

### 5. 免責事項

本稿は、著者の個人見解を表すものであり、大和証券株式会社および株式会社大和総研の公式見解を表すものではありません。

### 参考文献

- [Blundell 2015] Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight Uncertainty in Neural Networks, *Proceedings of the 32nd International Conference on Machine Learning*, 1613–1622 (2015)
- [Guo 2017] Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q.: On Calibration of Modern Neural Networks, *Proceedings of the 34th International Conference on Machine Learning* 1321–1330 (2017)
- [Kingma 2014] Kingma, D. P., Welling, M.: Auto-Encoding Variational Bayes, *Proceedings of the 2nd International Conference on Learning Representations* (2014)
- [Kingma 2015] Kingma, D. P., Salimans, T., Welling, M.: Variational Dropout and the Local Reparameterization Trick, *Advances in Neural Information Processing Systems* 28, 2575–2583 (2015)
- [Krizhevsky 2012] Krizhevsky, A., Sutskever, I., Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems* 25, 1097–1105 (2012)
- [Lecun 1998] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, 2278–232 (1998)
- [Molchanov 2017] Molchanov, D., Ashukha, A., Vetrov, D.: Variational Dropout Sparsifies Deep Neural Networks, *Proceedings of the 34th International Conference on Machine Learning*, 2498–2507 (2017)
- [Naeini 2015] Naeini, M. P., Cooper, G. F., Hauskrecht, M.: Obtaining Well Calibrated Probabilities Using Bayesian Binning, *Proceedings of the Twenty-Ninth*

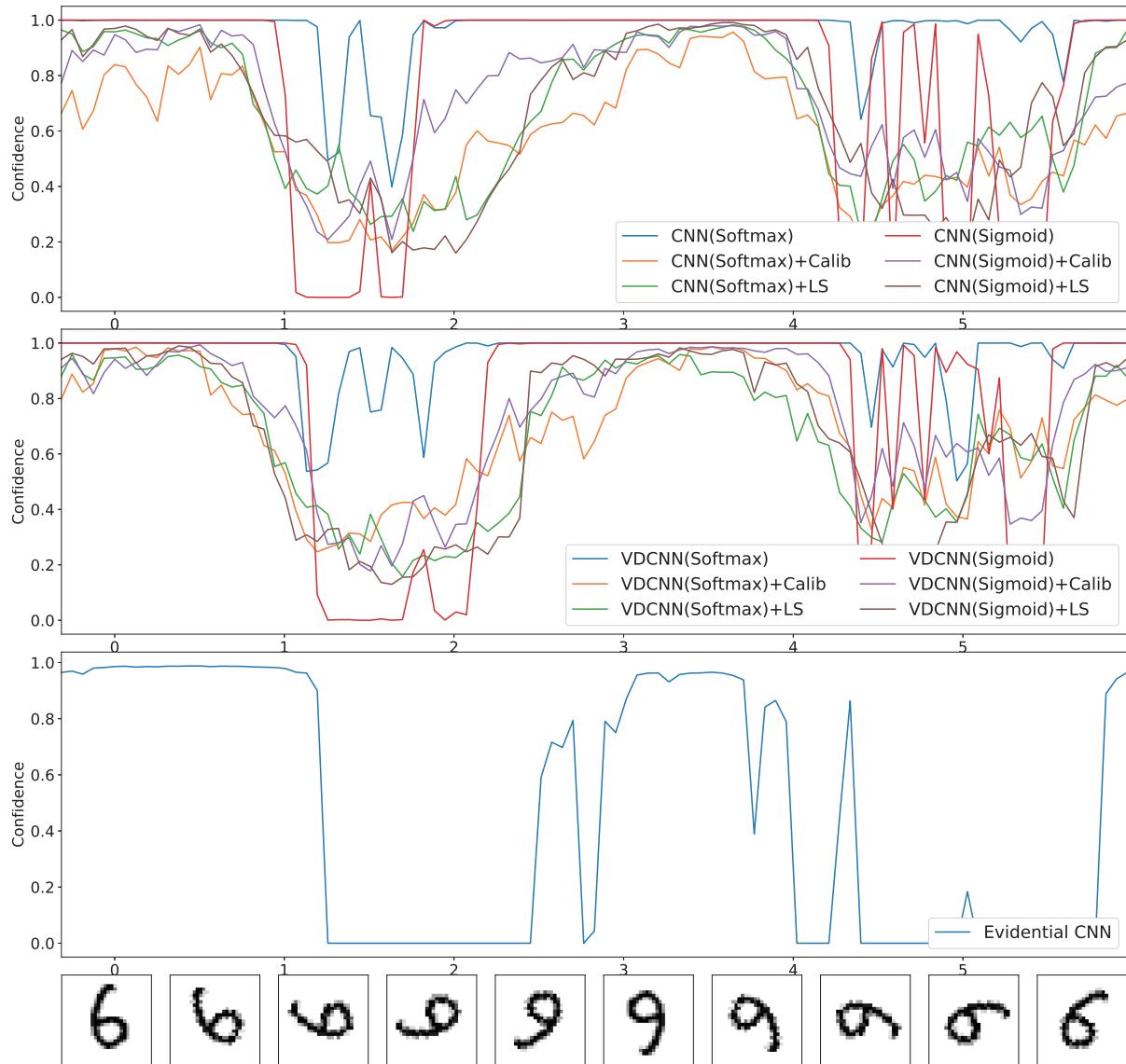


図 2: 各手法で学習したモデルを用いた、手書き数字の回転画像に対する予測結果。

AAAI Conference on Artificial Intelligence 2901–2907  
(2015)

[Sensoy 2018] Sensoy, M., Kaplan, L., Kandemir, M.: Evidential Deep Learning to Quantify Classification Uncertainty, *Advances in Neural Information Processing Systems 31*, 3183–3193 (2018)

[Szegedy 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision, *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826 (2016)

[Thulasidasan 2019] Thulasidasan, S., Bhattacharya, T., Bilmes, J., Channupati, G., Mohd-Yusof, J.: Knows When it Doesn't Know: Deep Abstaining Classifiers (2019)