Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem

Ridge-i inc.

Masanari Kimura (mkimura@ridge-i.com)

About

Education & Career

- 筑波大学卒 (2018)
- 株式会社Ridge-iエンジニア (2018 ~)
- 産総研特専研究員 (2019 ~)

Twitterやってます

@machinery81

Researches ater joining Ridge-i

- Interpretation of Feature Space using Multi-Channel Attentional Sub-Networks (CVPRW2019)
- Intentional Attention Mask Transformation for Robust CNN Classification (MIRU2019)
- PNUNet: Anomaly Detection using Positive-and-Negative Noise based on Self-Training Procedure (MIRU2019)
- Progressive Data Increasing as the Neural Network Initializer (JSAI2019)
- Anomaly Detection Using GANs for Visual Inspection in Noisy Training Data (ACCVW2018)
- Analyzing Centralities of Embedded Nodes (ICDMW2018)

今回紹介する論文

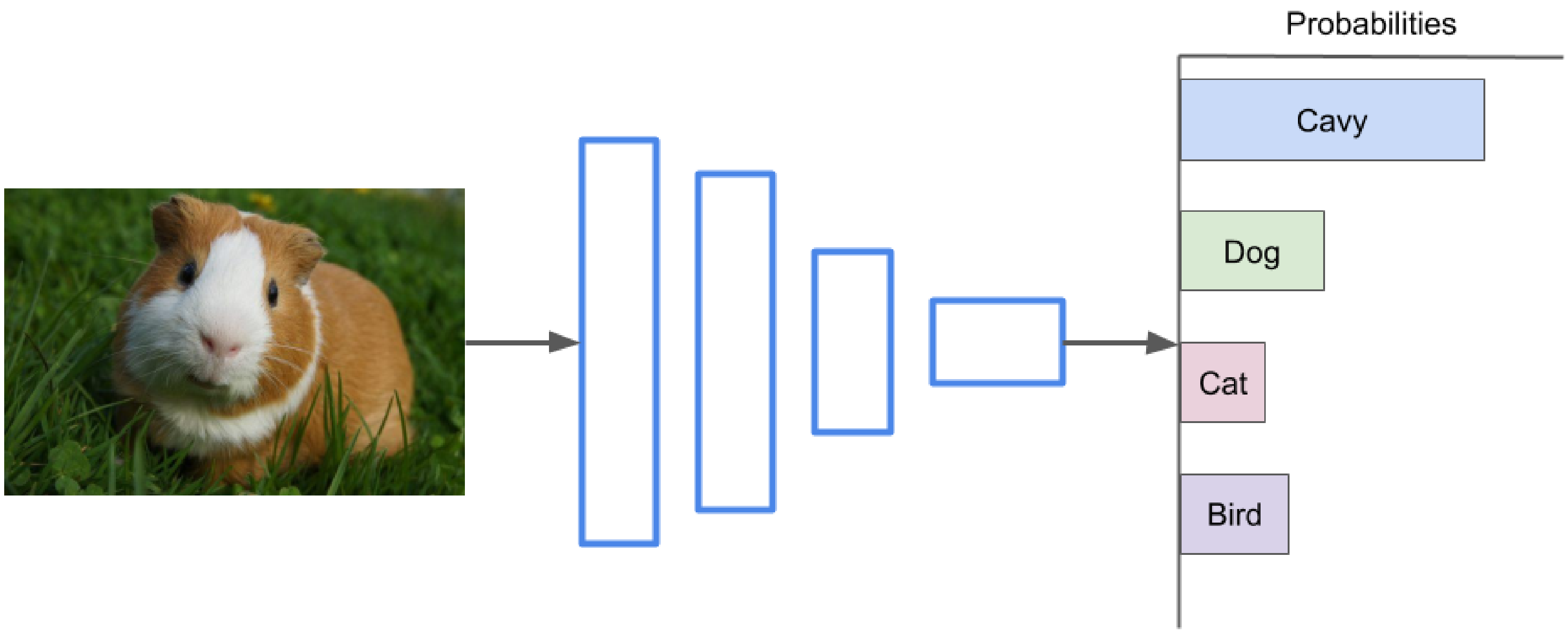
Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem

Abstract

- CVPR2019採択論文 [1]
- DNNsの出力の信頼度に関する論文

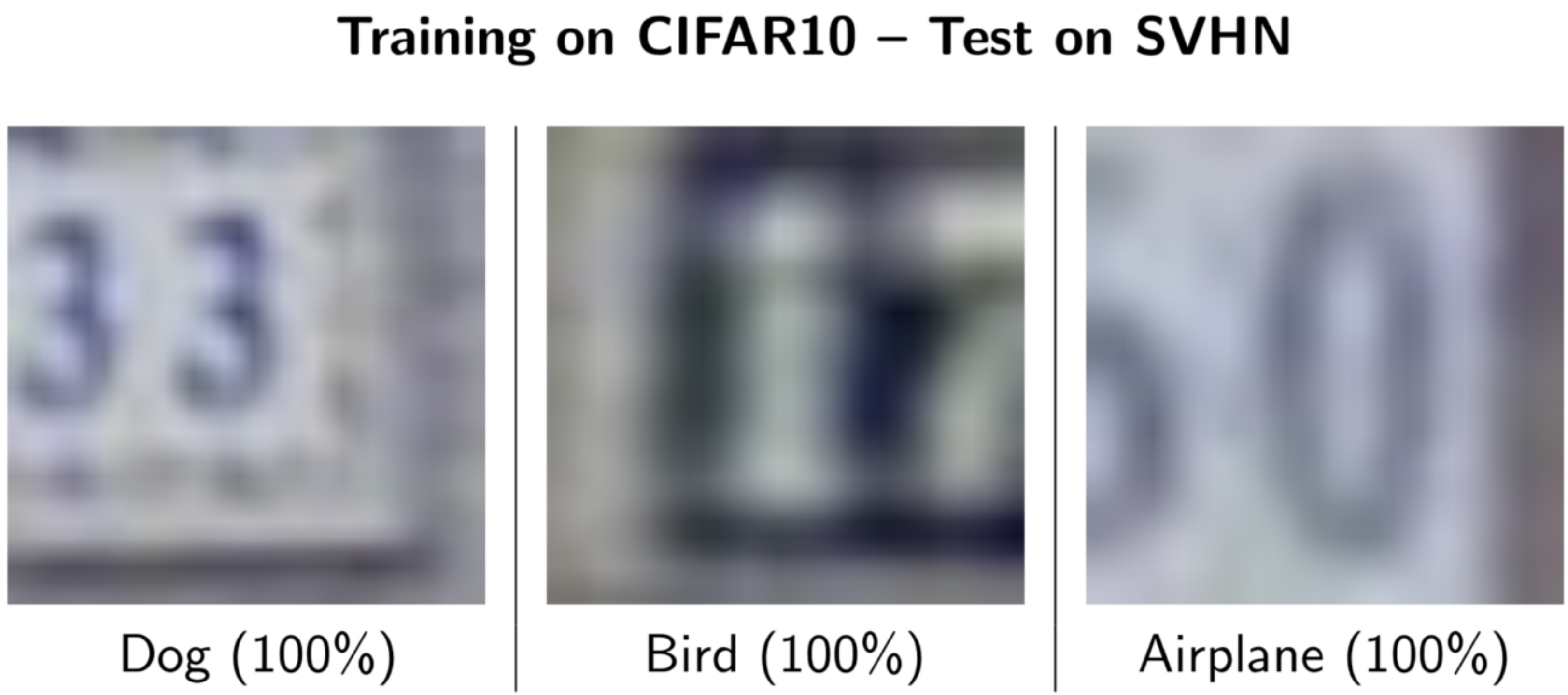
Confidence of DNNs outputs

- 一般的なsoftmaxを用いたDNNsの出力例

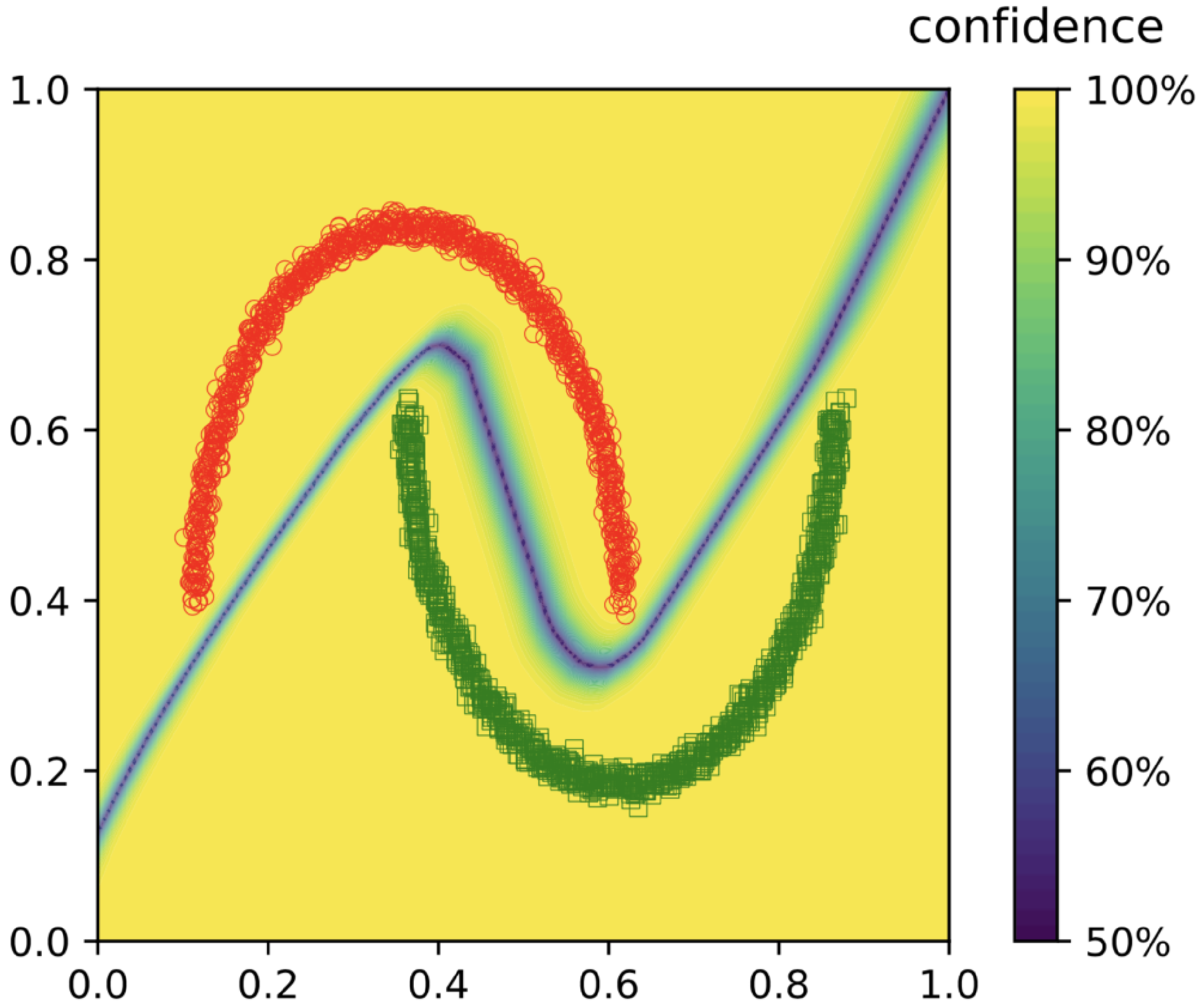


Problem of Overconfident Predictions

- 学習データと全く関係ないタスクのデータを入力
  - * DNNsは非常に高い確信度で適当なクラスに分類してしまう
  - * 本当はどのクラスの予測確率も低くあってほしい



Problem of Overconfident Predictions



Why ReLU Networks lead Overconfident ?

ReLU networks produce piecewise affine functions

Definition 2.1. *A function* $f : \mathbb{R}^d \to \mathbb{R}$ *is called* piecewise affine *if there exists a finite set of polytopes* $\{Q_r\}_{r=1}^{M}$ *(referred to as* linear regions *of* f*) such that* $\cup_{r=1}^{M} Q_r = \mathbb{R}^d$ *and* f *is an affine function when restricted to every* Q_r*.*

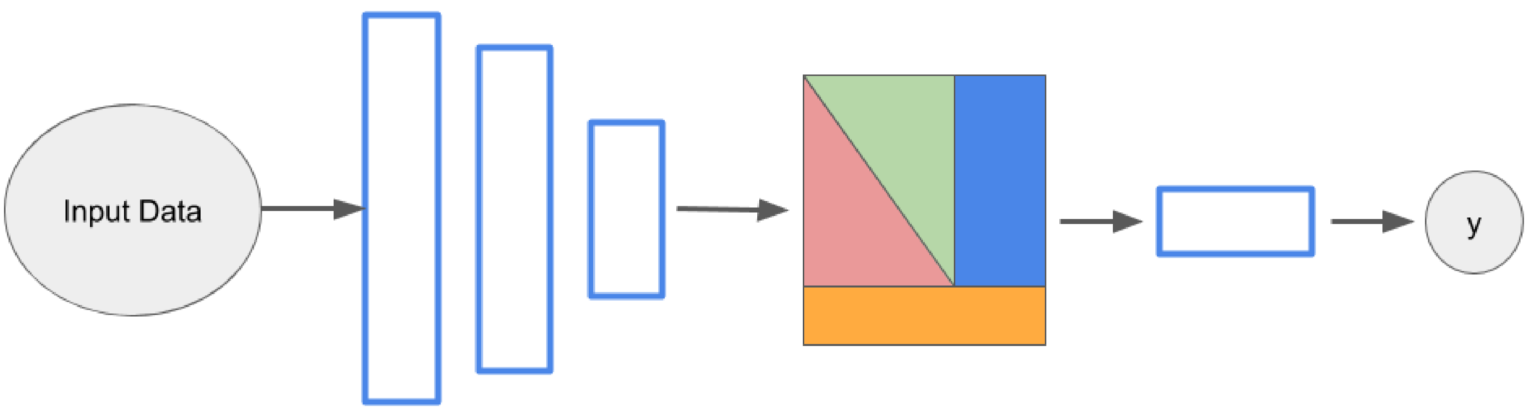
→ NN with ReLU = piecewise affine function ○ classifier

ReLU networks produce piecewise affine functions

ReLUを用いた，最終層が全結合層であるようなネットワークは,

1. 入力を有限の超多面体に分割
2. 全結合層で分類

と解釈できる[2].



Why ReLU Networks lead Overconfident ?

Lemma 3.1. *Let* $\{Q_i\}_{l=1}^{R}$ *be the set of linear regions associated to the ReLU-classifier* $f : \mathbb{R}^d \to \mathbb{R}^K$. *For any* $x \in \mathbb{R}^d$ *there exists* $\alpha \in \mathbb{R}$ *with* $\alpha > 0$ *and* $t \in \{1, \ldots, R\}$ *such that* $\beta x \in Q_t$ *for all* $\beta \geq \alpha$.

All the proofs can be found in the supplementary material. Using Lemma 3.1 we can now state our first main result.

Theorem 3.1. *Let* $\mathbb{R}^d = \cup_{l=1}^{R} Q_l$ *and* $f(x) = V^l x + a^l$ *be the piecewise affine representation of the output of a ReLU network on* Q_l. *Suppose that* V^l *does not contain identical rows for all* $l = 1, \ldots, R$, *then for almost any* $x \in \mathbb{R}^d$ *and* $\epsilon > 0$ *there exists an* $\alpha > 0$ *and a class* $k \in \{1, \ldots, K\}$ *such that for* $z = \alpha x$ *it holds*

$$\frac{e^{f_k(z)}}{\sum_{r=1}^{K} e^{f_r(z)}} \geq 1 - \epsilon.$$

Moreover, $\lim_{\alpha \to \infty} \frac{e^{f_k(\alpha x)}}{\sum_{r=1}^{K} e^{f_r(\alpha x)}} = 1.$

Why ReLU Networks lead Overconfident ?

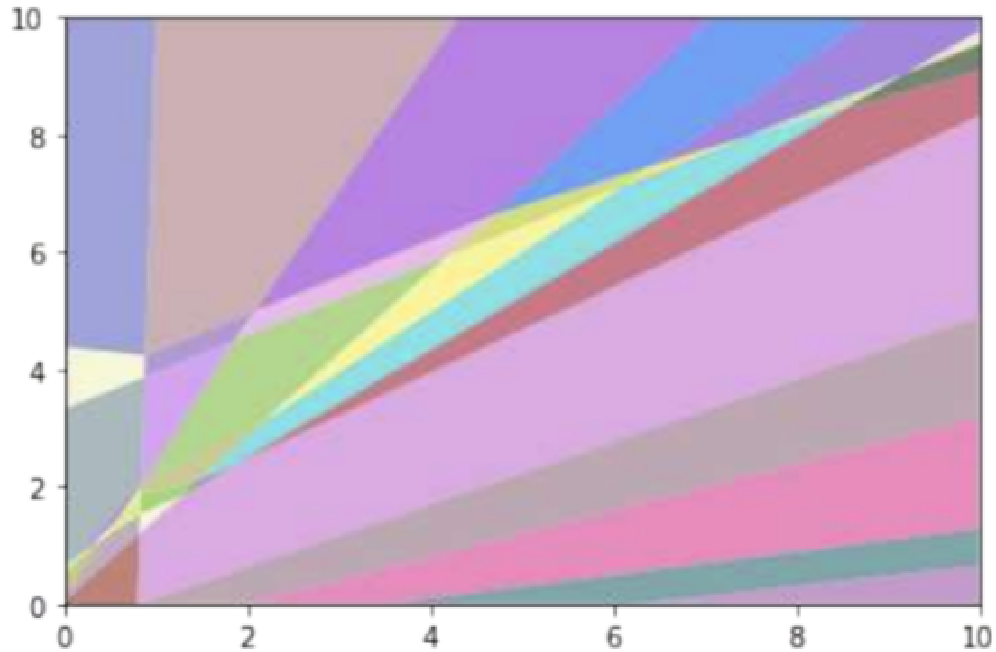
$$\lim_{\alpha \to \infty} \frac{e^{f_k(\alpha x)}}{\sum_{r=1}^{K} e^{f_r(\alpha x)}} = 1.$$

ReLUを使ったネットワークはαxの予測確率の極限が1になる

- αは定数
- αxとは？

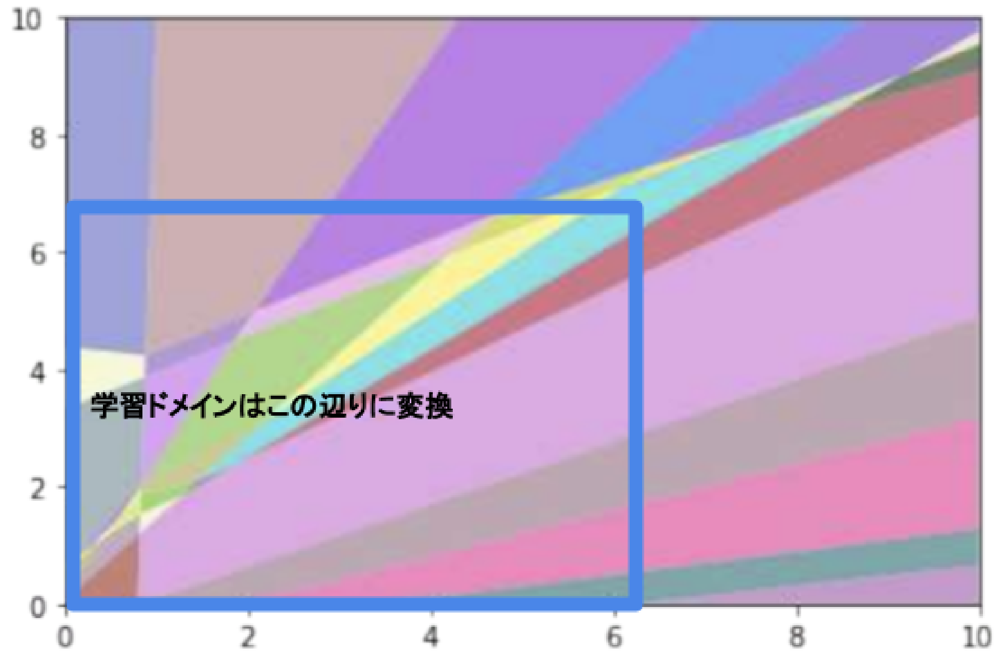
Intuitive Understanding

- 以下のように入力を変換



Intuitive Understanding

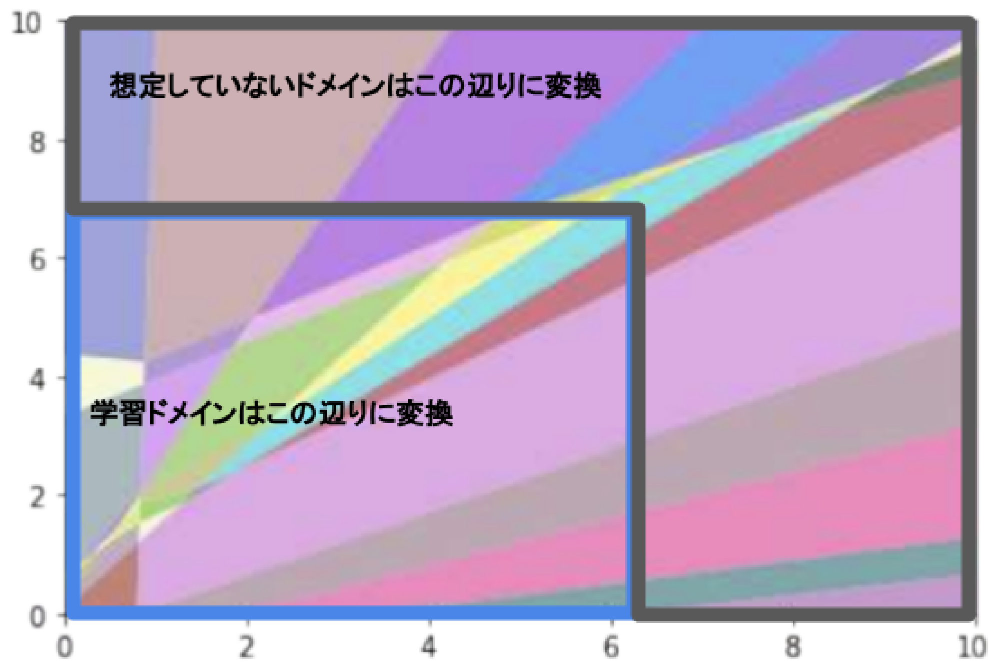
- 学習データ＆想定した入力データは以下のエリアに変換



Intuitive Understanding

- 想定していない入力データは以下のエリアに変換
- 定数$\alpha > 0$を掛ける→変換後の空間で右上に．．．

→　αx = 想定していない入力



（再活）

$$\lim_{\alpha \to \infty} \frac{e^{f_k(\alpha x)}}{\sum_{r=1}^{K} e^{f_r(\alpha x)}} = 1.$$

つまり，想定していない入力に対する予測確率の極限が1になる

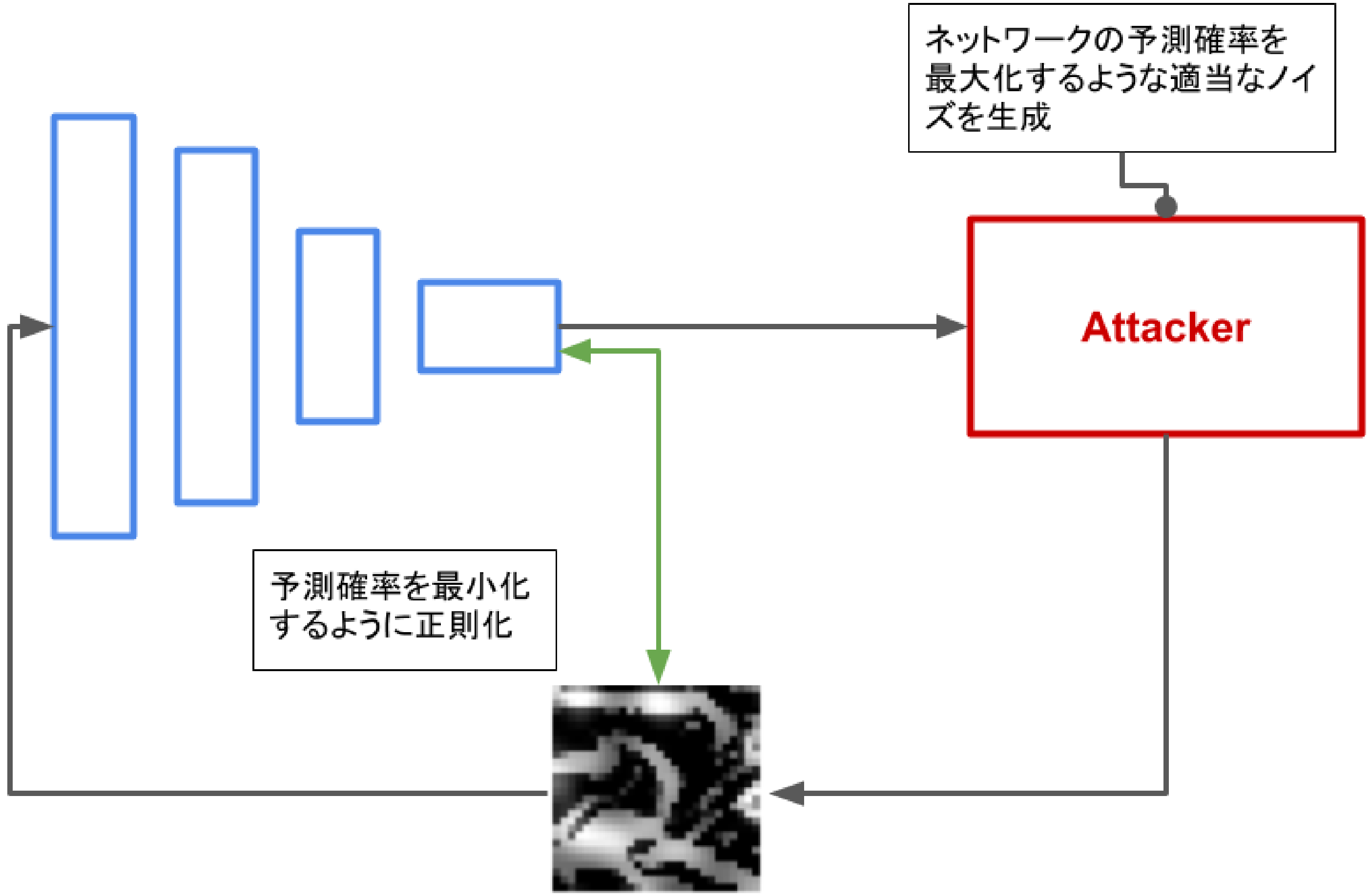
Adversarial Confidence Enhanced Training

- 想定していない入力に対する予測確率を低くするような正則化

$$\frac{1}{N}\sum_{i=1}^{N} L_{CE}(y_i, f(x_i)) + \lambda \mathbb{E}\Big[\max_{\|u-Z\|_p \leq \epsilon} L_{p_{\mathrm{out}}}(f, u)\Big],$$

Adversarial Confidence Enhanced Training

- 想定していない入力に対する予測確率を低くするような正則化

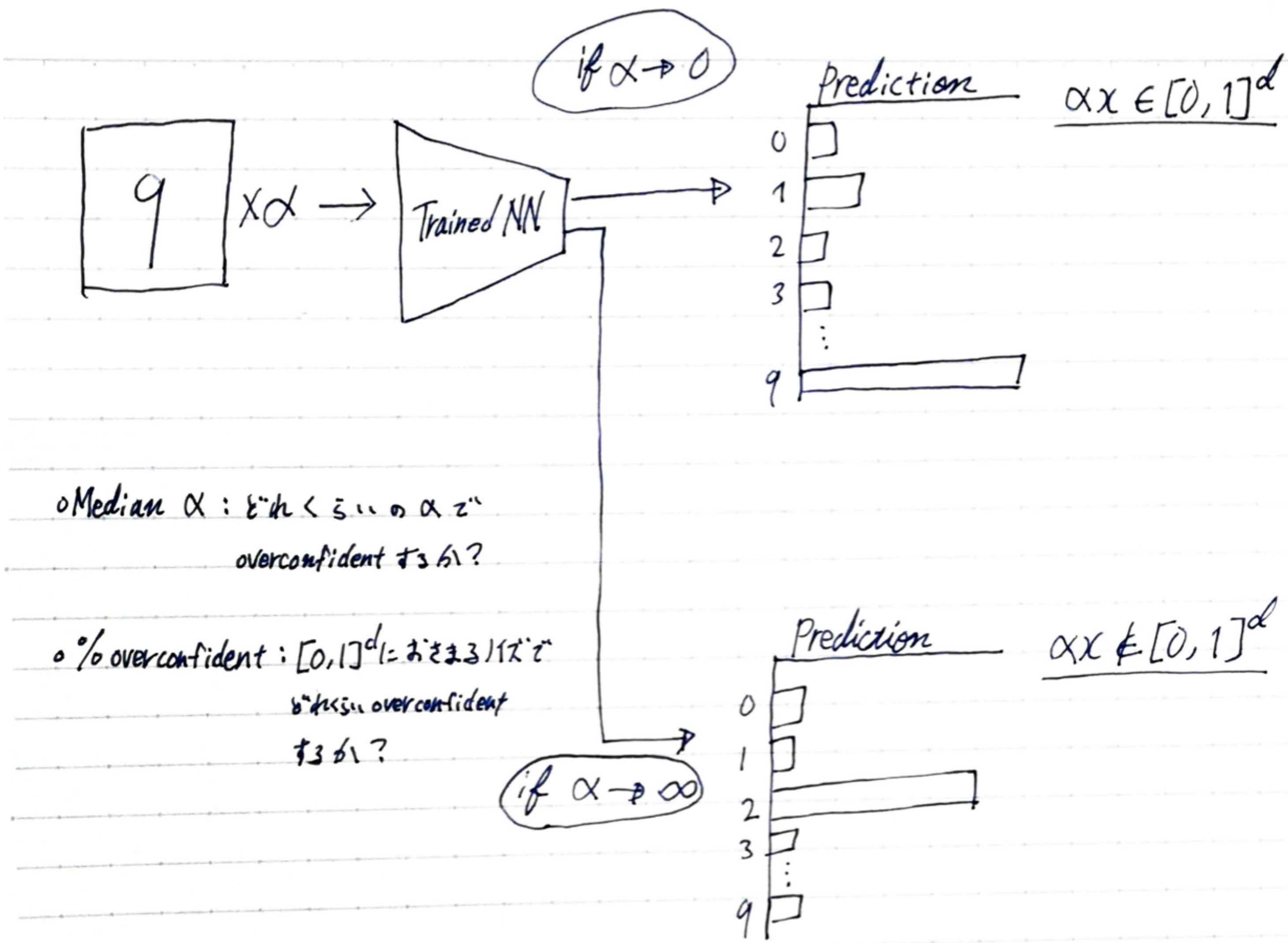


Experimental Results

- 各データセットに対するOverconfidentの実験結果

	Plain				ACET			
	MNIST	SVHN	CIFAR-10	CIFAR-100	MNIST	SVHN	CIFAR-10	CIFAR-100
Median α	1.5	28.1	8.1	9.9	$> 10^6$	**49.8**	**45.3**	**9.9**
% overconfident	98.7%	99.9%	99.9%	99.8%	**0.0%**	**50.2%**	**3.4%**	**0.0%**

	Plain				ACET			
	MNIST	SVHN	CIFAR-10	CIFAR-100	MNIST	SVHN	CIFAR-10	CIFAR-100
Median α	1.5	28.1	8.1	9.9	$> 10^6$	**49.8**	**45.3**	**9.9**
% overconfident	98.7%	99.9%	99.9%	99.8%	**0.0%**	**50.2%**	**3.4%**	**0.0%**



Conclusion & 疑問

- ReLUを用いたネットワークのoverconfident問題について理由づけ
- それを解決する正則化手法を提案
- そもそもsoftmaxの出力を信頼度と捉えてしまっていいのか
    - 不確実性を取り扱う手法を検討した方がいい気もする
        - Bayesian NNs etc.

References

- [1] Hein, et al. "Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem" The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019.
- [2] R. Arora, A. Basuy, P. Mianjyz, and A. "Mukherjee.Understanding deep neural networks with rectified linear unit" International Conference on Learning Representations (ICLR).2018.