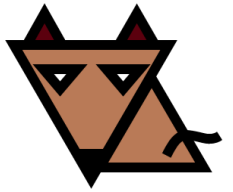
論文Summary

Entropy Penalty: Towards Generalization Beyond the IID Assumption

今回の論文



- Entropy Penalty: Towards Generalization Beyond the IID Assumption.
  - Arpit, Devansh, Caiming Xiong, and Richard Socher. 2019.
  - ArXivとOpenReviewでしか検索が引っかからない、ICLRに落ちた？

- Robustな機械学習
  - Domain adaptationに近いが少し違う
  - **Target domainの知識を使わない**で、
    Training Set (Source domain)の分布から外れたサンプルに対応する
  - **分類問題に特化**した手法だが、
    検出問題の特徴抽出部や、セグメンテーションなどでも使えるかも

偽の特徴量を獲得してしまう問題

- 人の直観とは違うものを学習してしまう

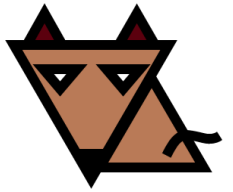
Training Set

らくだと自然風景は相関が高いものとして学習

Test Set

街中だとうまく認識できない

何故か？



- ニューラルネットは
  全ての手がかりを総動員して入力と出力を関連付けようとする

- 極端に言えば・・・**茶色い動物＋自然風景＝らくだ**

- **背景を学習してしまっている**ので、街にいるらくだを認識できない

意図する特徴だけを抜き出すには？

- Training Setを通して一貫した特徴だけを採用する

Training Set

らくだ本体の特徴は全てのサンプルに一貫している

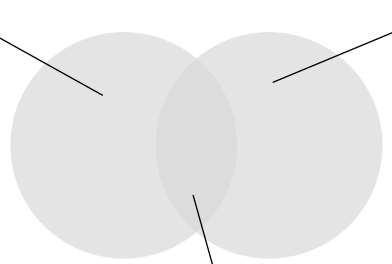
情報ボトルネック法

- もともとは情報圧縮のレート歪みの説明のための手法
- 一貫した特徴の抜き出しに応用

相互情報量とは

サイコロ
||
Aの平均情報量= H(A)

偶数・奇数
||
Bの平均情報量=H(B)



AとBの相互情報量=I(A,B)

={Aにより得られる情報量} –{Bが分かっている時、Aにより得られる情報量}
={Aの情報量のうち、Bを知ることで得られる情報量}
={偶数か奇数かを知ることでサイコロの出目が説明される度合い}

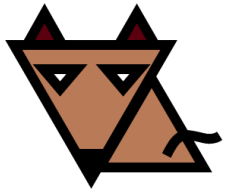
ロス関数

予測量　教師ラベル　特徴量　入力

$$\mathcal{L}_{IB}(\theta) = \underbrace{-\mathcal{I}(f_\theta(\mathbf{X}), \mathbf{Y})}_{\text{予測量が教師ラベルをよく説明するように学習}} + \underbrace{\beta\mathcal{I}(f_\theta(\mathbf{X}), \mathbf{X})}_{\text{特徴量を見ても、入力のことが分からないように学習}}$$

⇒入力の違いに拠らない特徴量を抽出

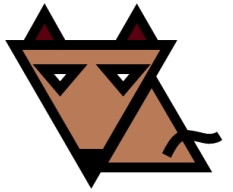
第何層の特徴量に適応すればよいか？



- より浅い層のほうがエントロピーが高い
- 第1層への適応が最も効果が高い
  ※この点は、ちゃんとした理論的な考察はできていない

$$\mathcal{H}(\mathbf{h}_1) \geq \mathcal{H}(\mathbf{h}_2) \geq \ldots \geq \mathcal{H}(\mathbf{h}_L)$$

Entropy Loss



- ガウス分布を仮定すれば、
  情報ボトルネックによるLoss関数は以下のように書ける

$$\mathcal{L}_{IB}(\theta) = -\mathcal{I}(f_\theta(\mathbf{X}), \mathbf{Y}) + \beta\mathcal{I}(f_\theta(\mathbf{X}), \mathbf{X})$$

f(x)が決定的な関数と仮定

$$\hat{\mathcal{L}}_{IB}(\theta) = (1-\beta)\mathcal{H}(\mathbf{Y}|f_\theta(\mathbf{X})) + \beta\mathcal{H}(f_\theta(\mathbf{X})|\mathbf{Y}) + C$$

従来のラベルを説明するためのLoss

f(x)がガウス分布に従うと仮定

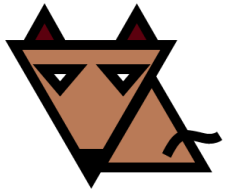
Entropy Loss

$$\mathcal{R}_{EP} = \sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}(\mathbf{x}|y=k)}[(\mathbf{h}(\mathbf{x}) - \mu_k)^2]$$

$$\text{where } \mu_k := \mathbb{E}_{\mathbf{x}\sim\mathcal{D}(\mathbf{x}|y=k)}[(\mathbf{h}(\mathbf{x})]$$

h(x): 第1層特徴量（非線形変換の前）

単純な例で効果を確認(1)



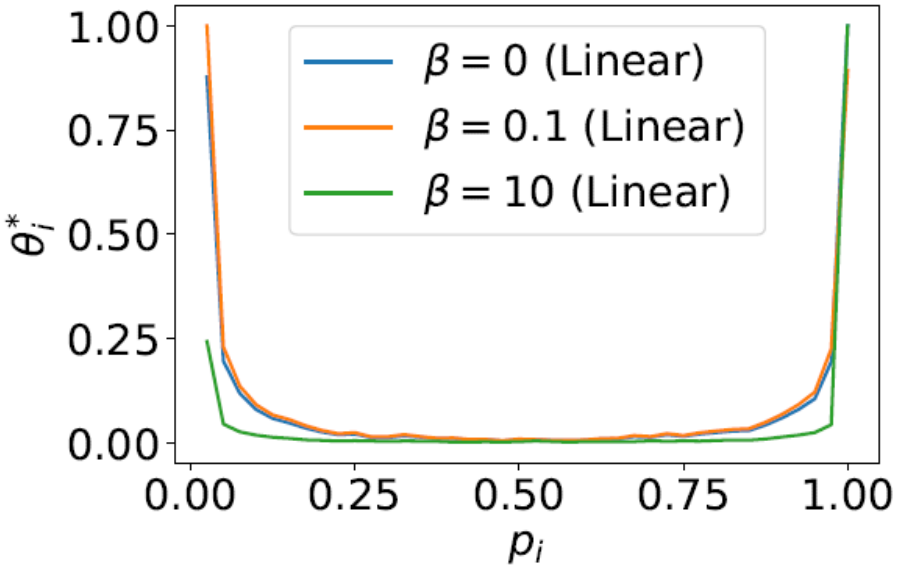
データセット $y \sim \{-1, 1\}$ $x_i \sim \begin{cases} \mathcal{N}(y, \sigma^2) & \text{with probability } p_i \\ \mathcal{N}(-y, \sigma^2) & \text{with probability } 1 - p_i \end{cases}$

モデル $f_\theta(\mathbf{x}) := \theta^T \mathbf{x}$ と 3層MLP

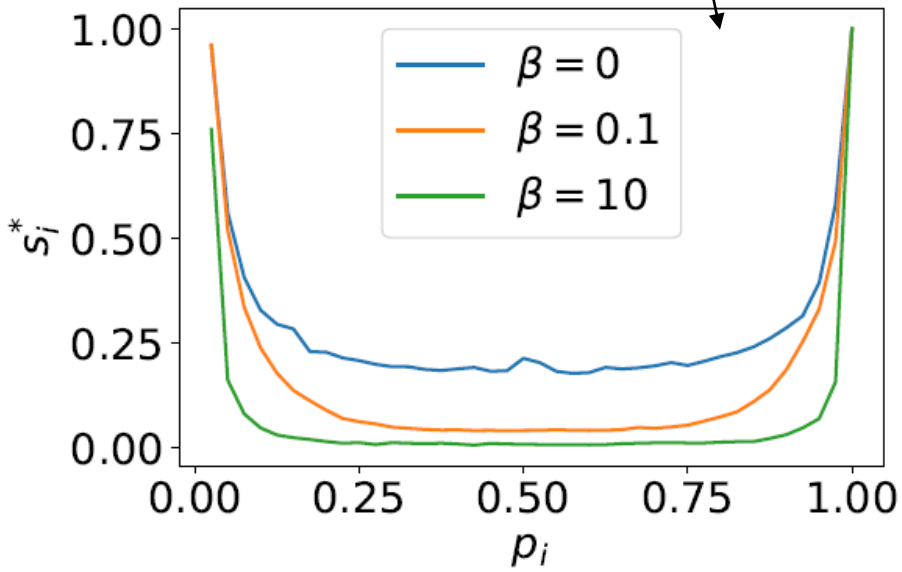
Loss $J(\theta) = \mathbb{E}[(f_\theta(\mathbf{x}) - y)^2] + \lambda \|\theta\|^2 + \frac{\beta}{2\pi e} e^{H(f_\theta(\mathbf{x})|y)}$

$s_i^* := \mathbb{E}_{\mathbf{x}} \left[\left| \frac{\partial f_{\theta^*}(\mathbf{x})}{\partial x_i} \right| \right]$

Linear Model



MLP



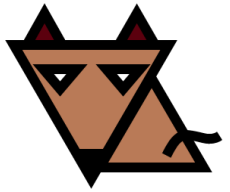
正規化された最適解

pi=0.5の特徴が選ばれていない

入出力感度 θ と同じ意味合い

pi=0.5の特徴は感度が低い

単純な例で効果を確認(2)



データセット $y \sim \{-1, 1\}$ $x_i \sim \begin{cases} \mathcal{N}(y, \sigma^2) & \text{with probability } p_i \\ \mathcal{N}(y, k\sigma^2) & \text{with probability } 1 - p_i \end{cases}$

モデル $f_\theta(\mathbf{x}) := \theta^T \mathbf{x}$ と 3層MLP

Loss $J(\theta) = \mathbb{E}[(f_\theta(\mathbf{x}) - y)^2] + \lambda \|\theta\|^2 + \frac{\beta}{2\pi e} e^{H(f_\theta(\mathbf{x})|y)}$

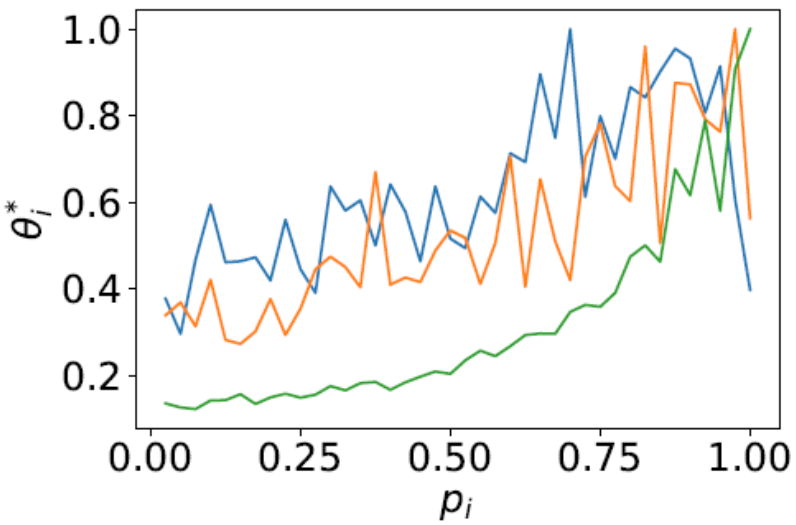
$s_i^* := \mathbb{E}_{\mathbf{x}} \left[\left| \frac{\partial f_{\theta^*}(\mathbf{x})}{\partial x_i} \right| \right]$

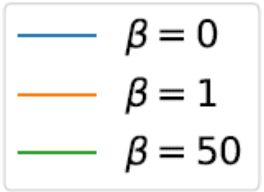
入出力感度
θ と同じ意味合い

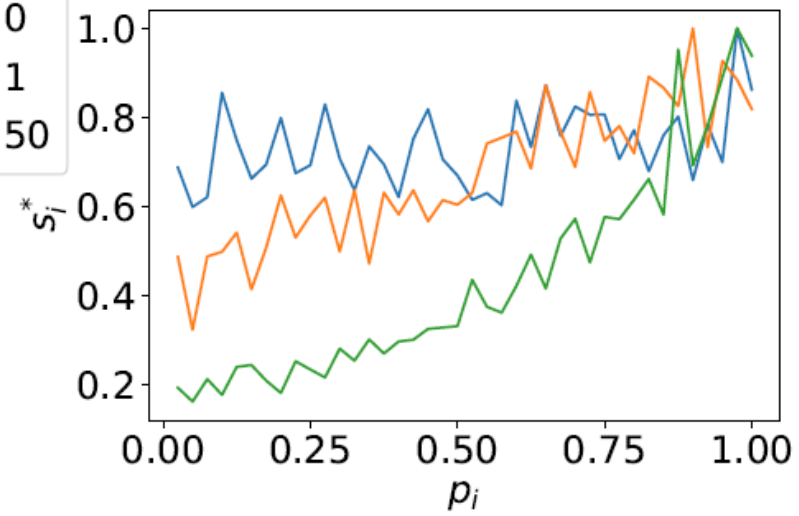
Linear Model

MLP

正規化された最適解







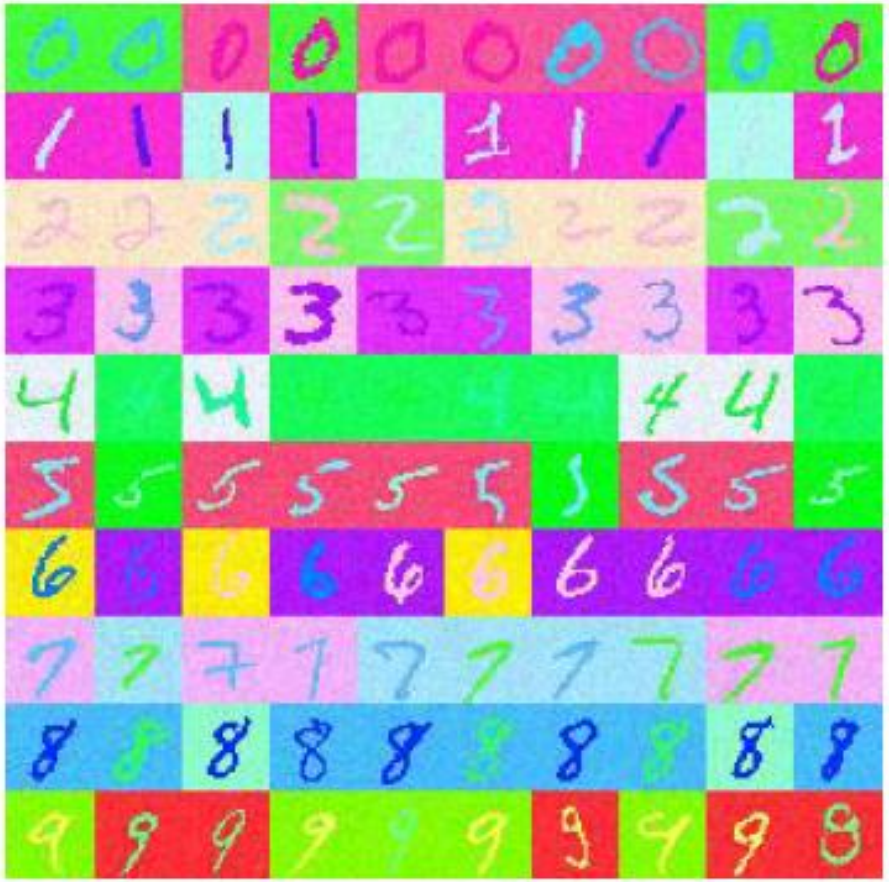
pi=0の特徴が選ばれていない

pi=0の特徴は感度が低い

画像認識で確認

- Color MNIST (C-MNIST)

Training Set



Test Set

画像認識で確認

- C-MNISTのTrain/Test

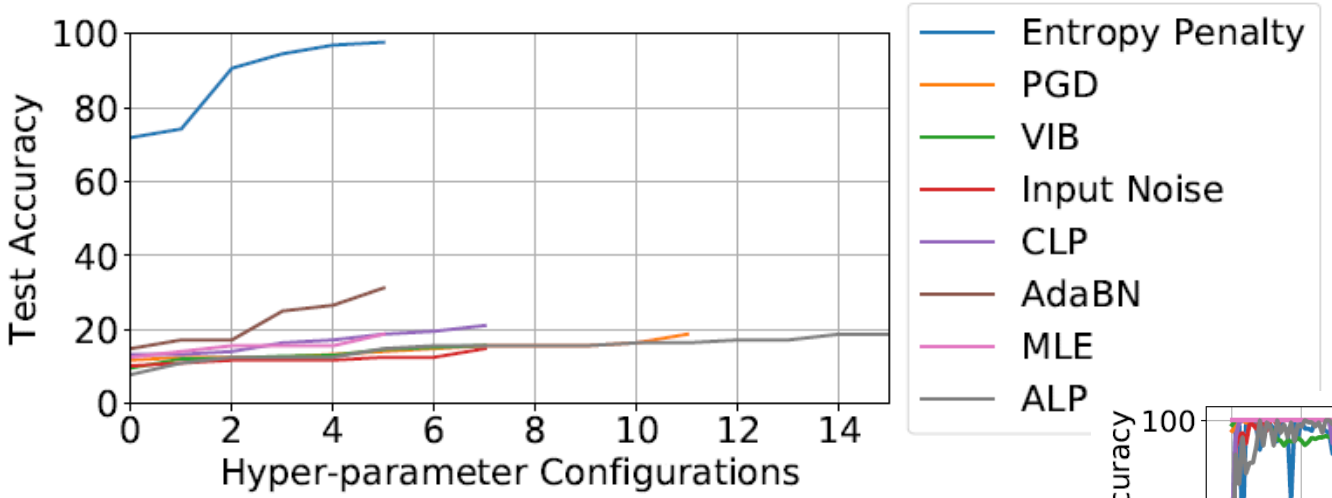


Figure 3: Performance on the distribution shifted te
MNIST for various methods trained on C-MNIST t

PGD, ALP: Adversarial training系の手法
VIB, CLP: 提案手法に近い手法
Input Noise: ノイズ付加
AdaBN: BN使用のシンプルなDA手法

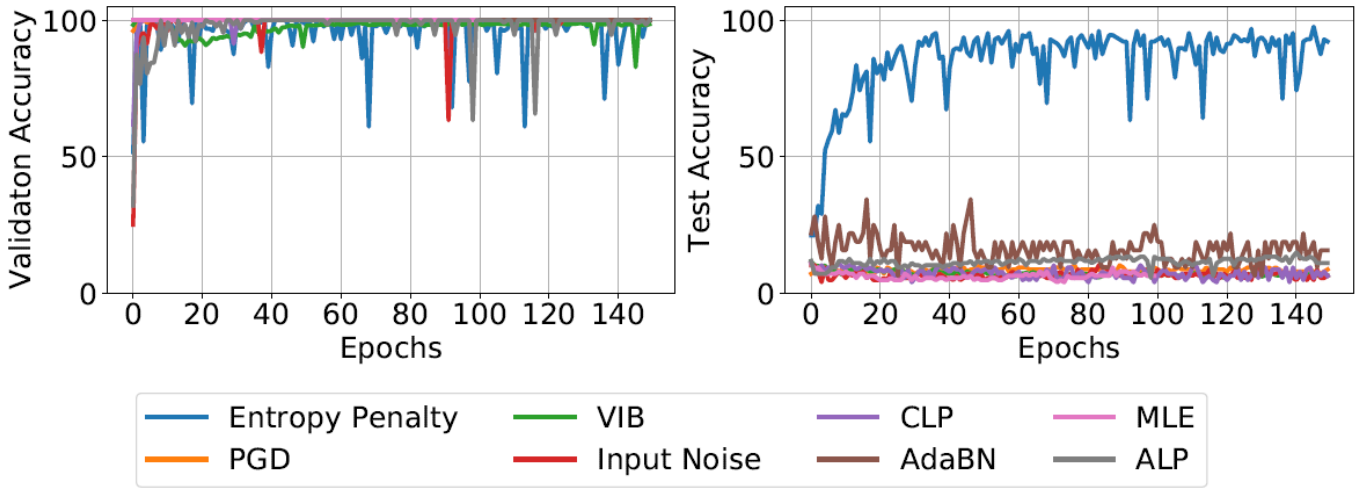


Figure 4: Baseline methods severely overfit color features in the C-MNIST training set leading to near 100% accuracy on C-MNIST validation set but close to chance performance on the distribution shifted C-MNIST test set.

モデルは、ResNet56

画像認識で確認

- その他のデータセットへの適応度

Dataset	Accuracy
C-MNIST	96.88
MNIST	93.75
MNIST-M	85.94
SVHN	60.94

Table 1: Out of distribution performance on test sets using a model trained with Entropy Penalty on C-MNIST dataset.

Training

Test

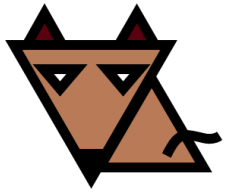
C-MNIST

MNIST

SVHN

MNIST-M

議論すべき点



- 論文でふれられている点
  - 特徴量の分布がGaussianという仮定はいつでも成り立つか？
  - 第1層に適用するのが本当に最適か？本論文では理論的な考察が不足

この手法の弱点は？

- （個人的な考察）背景も含めて完全に偏っている場合、やはり背景も学習してしまう？

Training Set

⇨ 砂漠も学習してしまう