

ラベルノイズ量とクラス尤度分布を用いたラベルノイズ検出手法

林佑介¹ 田口賢佑¹

概要：近年深層学習を用いた画像認識は高い認識精度を達成しており、大きな成功を収めている。しかし、画像と一致しないラベル（ラベルノイズ）が学習データに含まれていると、それらのデータに過剰にフィッティングしてしまい、汎化性を損なうという問題がある。これを解決するためには、正しくラベルを付けたデータセットを生成する必要があるが、正確にアノテーションを行うには時間もコストもかかってしまう。本稿では、このような問題を解決するために、ラベルと異なる推論結果を示すデータのクラス尤度分布を用いて効率的にラベルノイズを検出する手法を提案する。検証として画像分類問題のタスクにおいて、提案手法によるラベルノイズの検出精度の評価を行った。

キーワード：Label Noise, Classification, Convolutional Neural Network

1. はじめに

Convolutional Neural Network(以下CNN)[1]は、コンピュータビジョンにおいて様々な分野に応用可能な技術である。CNNは十分な量の学習データを用いて学習を行うことで汎化性の高い推論が可能となる。CNNが適用されるタスクの一つに、画像分類問題のタスクがある。画像分類問題とは1枚の画像全体に対してクラス分類を行う問題である。画像分類問題において、CNNの学習を行うためには学習データセットに対して1対1でラベルを与える必要がある。また、先述の通りCNNの学習に汎化性をもたせるためには十分な品質かつ大量の学習データが必要となる。一方、画像分類タスクにおいて、基本的には学習データに対するラベルは人間が付与する必要がある。人間が数万枚もの画像に対してラベルを1対1で対応させた場合、画像に対する誤認識や主観的な要素が介在する等、様々な要因によって画像と一致しないラベルノイズが混在してしまう[2]。ラベルノイズとは、あるデータに対する誤ったラベルを指す。学習データにラベルノイズが含まれる場合、CNNは誤った学習を行い認識精度が低下してしまう。そのため、学習データからラベルノイズを除去することが必要になるが、ラベルノイズの混入したデータセットをすべて人間が確認するためには、多くの時間的コストが必要となる。一方、人間が確認する代わりにCNNを用いた様々なラベル検出手法が提案されているが、モデルが複雑になり甚大な学習時間を必要とするという問題がある。

本研究では、ラベルノイズ候補を検出することでこれらの時間的コストの問題を緩和する。ラベルノイズの含まれるデータセットを極めて少ない学習回数でモデルに学習させて推論した後、ラベルと異なる推論結果となるデータのうち、高いクラス尤度を示すデータを抽出してラベルノイズを検出する。

2. 関連研究

近年、ラベルノイズへの対策として多くの手法[2]が提案されているが、これらは特徴によって次の2種類に分類される。

- ・ラベルノイズにロバストなモデルを学習する方法
- ・ラベルノイズを検出してデータを浄化する方法

2.1 ノイズにロバストなモデル

まずノイズにロバストな損失関数を設定する試みが挙げられる。Ghoshら[3]は平均絶対誤差(MAE: Mean Absolute Error)がラベルノイズに対してロバスト性を向上させることを理論的に示した。しかしながら、MAEをCNNに適用すると精度が劣化してしまう問題がある。その他の手法として、ラベルの信頼度に応じてそれぞれの学習データに対する重みを付加する[4]ことが提案されている。しかし、効果的な重みを推定することは簡単ではない。Renら[5]はメタラーニングを利用して重みを直接的に最適化することによって汎化性を向上した。また、[6]では、メタラーニングによってラベルを人工的に付け替えてラベルにロバストな学習を行う手法が提案されている。しかし、これらの手法はラベルノイズの含まれない一定量のデータセットが必要とされ、多くの場合、このようなラベルノイズのないデータセットを得ることは難しい。

上述したラベルに対するロバストなモデルや学習方法は、汎化性の向上が期待できる一方でラベルノイズの影響を全く受けないわけではなく、少なからずモデルの識別制度に影響を及ぼしてしまう。そのため、本論文ではラベルノイズの浄化を目指す。

2.2 ノイズを浄化するアプローチ

Kohら[10]はモデルのトレーニングに対して害のあるデータであることを測る影響度に関する関数を提案した。しかし、バリデーションのデータセットに対して各学習データの影響度を計算する必要があるため、実用的な解決策にはなりにくい。一方、Kimら[7]は相補的なラベルを学習す

¹ 京セラ (株)
Kyocera Corp.

る Negative Learning を用いてラベルノイズを検出する手法を提案している。相補的なラベルで学習を行うため、ラベルノイズを学習して過学習が起こることを防ぐことができる一方、学習時間が増えてしまうという問題がある。また、通常の学習よりも多くのデータが必要となる。

2.1 および 2.2 で挙げた従来法では、ノイズに関する先行知識を利用する[4,5]必要がある。また、メタラーニング[5,6]を用いた手法は複雑なモデルとなり、学習時間に多くの時間を費やす必要がある。さらに、学習データに対する影響度[10]や Negative Learning を利用[7]してラベルノイズを浄化する手法があるが、これらも甚大な計算コストを必要とするという問題が発生する。本研究は、ラベルノイズへの過剰なフィッティングを避けるため学習回数を1回もしくは数回に止める。したがって、従来手法のように学習に時間をかける必要がない。また、オーバーフィッティングを避けることにより、ラベルと異なる推論結果となるデータのうち、高いクラス尤度を示すデータに含まれるラベルノイズを多くすることができる。そのため、確認すべきデータの量を格段に減らすことによって効率的にラベルノイズを検出できる手法を提案する。

3. タスク概要

本稿では提案手法の有効性の検証のため、タスクとして2クラスの画像分類問題を用いる。ノイズ浄化の対象がマルチクラスのデータセットであっても、浄化を対象とするクラスを2クラスの中に入れて選択することですべてのラベルノイズを浄化することができるため、本手法の有効性は失われない。以降に学習のために用いるデータセット、ネットワーク構造、ラベルノイズの付加方法について示す。

3.1 データセット

車載シーンを対象とした物体検知用データセットを用いる。物体検知用データセットは車両の後方に装着したカメラで撮影されている。図1に画像例を示す。

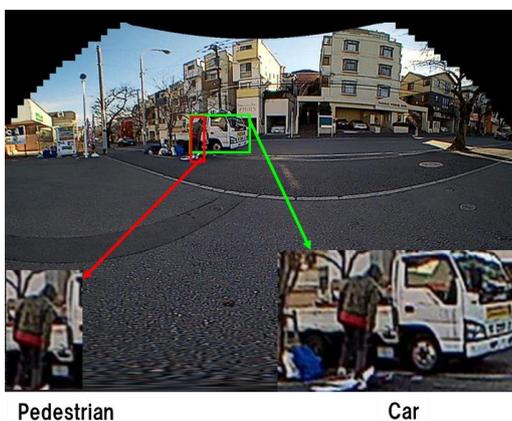


図1: 物体検知用データセットに含まれる画像の一例

本研究では、ラベルが付けられたバウンディングボックス毎に画像を切り出す。そのため、画像サイズはそれぞれ異なる。学習用に10000枚、評価用に10000枚を用いる。対象クラスは前述の通り Car と Pedestrian の2クラスで、学習用と評価用に各クラス5000枚で構成されている。

3.2 ネットワーク構造

識別器として用いるモデルのベースは ResNetV20V[11]とした。前述の通り画像サイズが異なるため Batch Size を1とし、最終層を全結合層ではなく Global Average Pooling 層に置き換えている。

3.3 ノイズの付加

学習データへのノイズの付加は次のように行う。まず、ラベルノイズ比率を設定する。次に、対象となるデータセットからラベルノイズ比率に合わせてラベルを入れ替える。ただし、ラベルノイズ比率とはデータの総数に対するラベルノイズの数の比を表す。

4. 提案手法

本研究では、ラベルと異なる推論結果を示したデータセットの中からクラス尤度の高いデータを抽出してラベルノイズを検出する手法を提案する。以下に提案手法の詳細を述べる。

4.1 用語定義

以下に提案手法に用いる各用語の定義を示す。

- **ラベルノイズ**
あるデータに対する誤ったラベル
- **クリーンラベル**
あるデータに対して正しいラベル
- **NG データセット**
ラベルと異なる推論結果を示したデータセット
- **ラベルノイズ比率**
データの総数に対するラベルノイズの数の比

4.2 概念

ラベルノイズが混在したデータセットを学習した CNN は、ラベルノイズに過剰にフィッティングしてしまい、汎化性を損なう[8]ことが知られている。一方、[9]では学習率が高ければ、CNN はある程度の汎化性を獲得し、ラベルノイズの影響を受けにくくなることが報告されている。また、勾配法による最適化手法を用いた CNN は、フィッティングが“easy”なデータから知識を獲得する傾向を持つ[12]。反対に、フィッティングが“hard”なデータは学習が進んだ後に獲得される。そこで、学習率を高く、学習回数を少なく設定すると、比較的“easy”なデータについては、ラ

ベルによらず画像の内容を正しく推論することが可能となる。つまり、ラベルノイズを推論するとラベルと異なる推論結果を示すとともにクラス尤度も高くなることが期待される。

図2に、ラベルノイズが混在するデータセットを学習したCNNにおいて、NGデータセットのクラス尤度に対するヒストグラムを示す。3.1で説明したデータセットを用いて、Pedestrianのみ250枚のラベルノイズを付加した。また、学習率を0.0001、学習回数を1に設定した。図2で示すように、ラベルノイズを含むPedestrianクラスは、ラベルノイズが1付近の階級でクラス尤度の最頻値を持つ。一方、クリーンラベルは0.5付近に最頻値を持つ。1の階級でのラベルノイズ比率を求めると97%と高い数値になる。また、ラベルノイズを含まないCarクラスでは、クラス尤度が1付近と0.5付近の両方で高くなっていることが分かる。

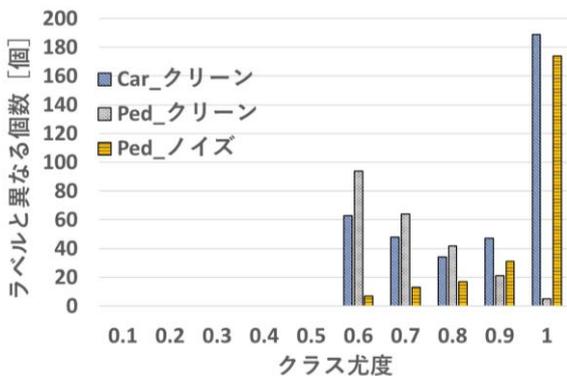


図2: ラベルノイズが混在するデータセットを学習したCNNでのNGデータセットのヒストグラム

図3に、ラベルノイズを含まないデータセットを学習したCNNにおいて、NGデータセットのクラス尤度に対するヒストグラムを示す。図2のラベルノイズを含まないCarクラスの尤度の傾向と同様に、図3ではCarクラス、Pedestrianクラスともに尤度が1付近と0.5付近で高くなっている。

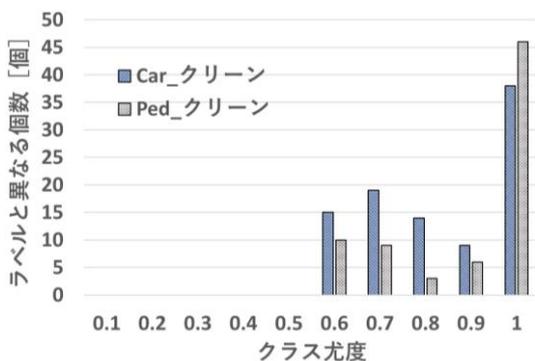


図3: ラベルノイズを含まないデータセットを学習したCNNでのNGデータセットのヒストグラム

これらの結果から、ラベルノイズが混在したデータセットを学習したCNNを用いて、NGデータセットのうちクラス尤度の高いデータを抽出すれば、ラベルノイズを高い確率で検出することができる。逆に、NGデータセットのうちクラス尤度の高いデータを抽出してノイズが含まれていなければラベルノイズが含まれていないことが分かる。

4.3 ラベルノイズ検出手順

まず、ラベルノイズが混在したデータセットでCNNを学習する。次に、CNNの推論結果のうち、ラベルと異なる結果を示すデータを選択し、クラスごとに尤度を階級としてヒストグラムを算出する。そして、例えば、クラス尤度に閾値を設定して閾値以上のデータを抽出し、ラベルノイズを正しいラベルに付け直す。ラベルと異なる結果を示すデータが少なければ、クラス尤度の高いデータから上位一定数のデータを抽出しても構わない。その後、浄化したラベルを追加して再度CNNで学習し、正しいラベルを付け直すまでの手順をラベルノイズが検出できなくなるまで繰り返す。全体の検出手順アルゴリズムをAlgorithm1に示す。

Algorithm 1 The Algorithm of Our Noisy label Cleansing

```

INPUT:  $D$ : Dataset with noisy labels
          $n$ : The total number of  $D$ 
1: WHILE  $noisy\ label \neq 0$ 
2:   Initialize  $noisy\ label (=0)$ 
3:   Train a CNN on dataset  $D$ 
4:   Set a threshold  $T$ 
5:   FOR  $i = 1:n$ 
6:     Extract the false data whose class confidence is
       over  $T$ 
7:     IF The label is wrong
8:       Correct the label and Put it into  $D$ 
9:        $noisy\ label = 1$ 
10:    ENDIF
11:  ENDFOR
12: End WHILE
OUTPUT: Clean dataset  $D$ 
    
```

5. 評価実験

提案手法の有効性を確認するため、以下3点の実験を行った。最初の実験として、提案手法を用いてラベルノイズを浄化することが可能か検証する。また、学習データに含まれるラベルノイズ量を変化させた場合の浄化可否および浄化に必要な検出の繰り返し回数についても確認する。次に、学習回数を変化させてオーバーフィッティングの程度を変えた場合のクリーンラベルとラベルノイズの尤度分布を検証する。特にクラス尤度の高いデータに含まれるラベルノイズの比率に注目する。最後に、損失関数の違いによるラベルノイズ量の推移の差について確認する。以下に、各々の実験方法、実験結果を述べる。

5.1 実験1：提案手法によるラベルノイズの浄化検証

5.1.1 実験方法

実験1では、提案手法を用いてラベルノイズが浄化できることを検証する。また、ラベルノイズ比率によるラベルノイズ量の推移を比較する。ラベルノイズ比率は5%、40%、50%の3つのパターンで確認する。ラベルノイズが検出できなくなるまで4.3で述べた検出手順を繰り返し、ラベルノイズ量の推移を評価する。表1に実験条件を示す。学習率は、[15]に提案されている手順にしたがって、学習率を徐々に増加させながらもっとも損失が落ちる値で決定した。表1に含まれない条件として、学習回数を1、損失関数を交差エントロピーと設定する。

表1: 実験条件

学習データ	物体検知用データセット
学習データ数	10000
クラス数	2
クラスごとのデータ数	5000
バッチサイズ	1
学習率	0.0001
識別モデル	ResNet20 v1

5.1.2 実験結果

本実験によって得られたラベルノイズ量の推移を図4に示す。図4で示すように、いずれのラベルノイズ比率の場合でもノイズを浄化できることが分かる。また、ラベルノイズ比率が高いほど繰り返しの回数が必要となる。なお、ラベルノイズ比率50%では学習が収束しなかった。

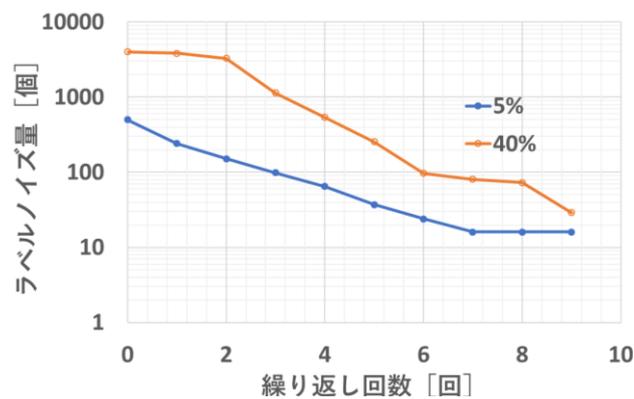


図4: ラベルノイズ量の推移

5.2 実験2：オーバーフィッティングと尤度分布の関係性

5.2.1 実験方法

4.2で述べたように、提案手法はラベルノイズへの過剰なフィッティングを避けることによって、NGデータセットのうちクラス尤度の高いデータに含まれるラベルノイズ

量を多くできることを利用している。そこで、実験2では学習回数を変化させてオーバーフィッティングの程度を変えた場合のクリーンラベルとラベルノイズの尤度分布を検証する。表1に実験条件を示す。表1に含まれない条件として、ラベルノイズ比率を5%、損失関数を交差エントロピー、学習回数を1、20回と設定する。評価は、1の階級に含まれるラベルノイズ比率で行う。

5.2.2 実験結果

学習回数1のNGデータセットのクラス尤度に対するヒストグラムを図5、6に示す。図5はCarクラス、図6はPedestrianクラスのヒストグラムを表す。学習回数1では、いずれのクラスでも、ラベルノイズのクラス尤度が高く、クリーンラベルのクラス尤度は低い。

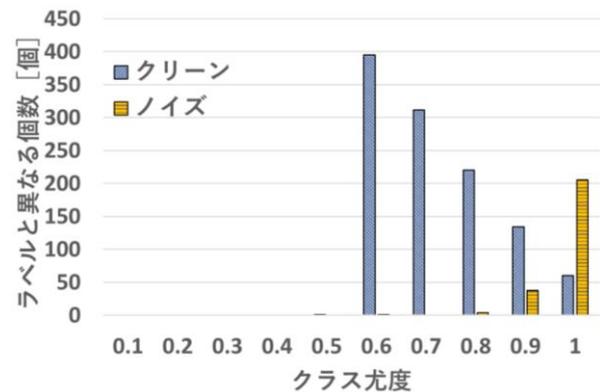


図5: NGデータセットのCarクラスの尤度：学習回数1

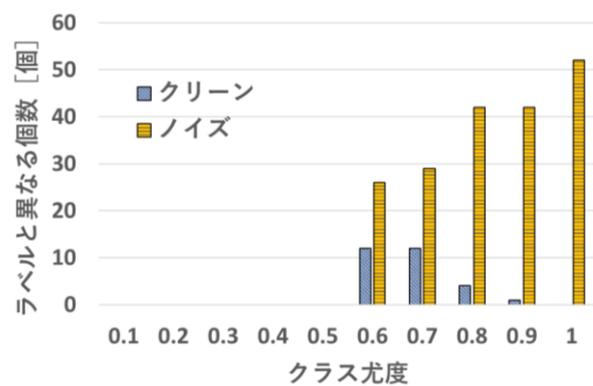


図6: NGデータセットのPedestrianクラスの尤度：学習回数1

学習回数20のNGデータセットのクラス尤度に対するヒストグラムを図7、8に示す。図7はCarクラス、図8はPedestrianクラスのヒストグラムを表す。学習回数20では、ラベルノイズ、クリーンラベルともにクラス尤度が高い。

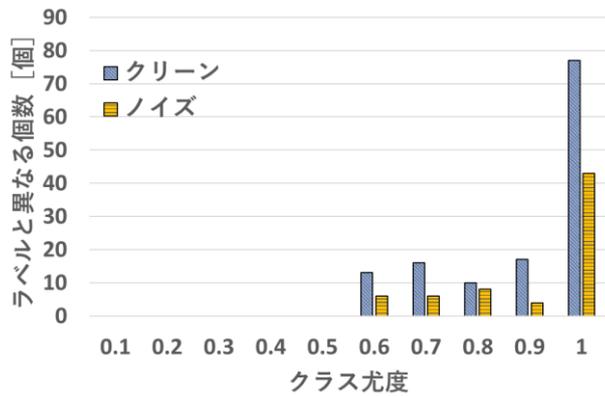


図 7: NG データセットの Car クラスの尤度：
 学習回数 20

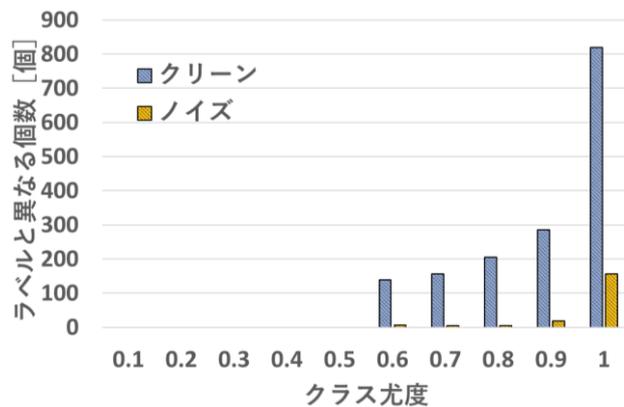


図 8: NG データセットの Pedestrian クラスの尤度：
 学習回数 20

図 9 は学習回数に対する損失の推移を示す。図 9 で示すように、学習回数 20 ではオーバーフィッティングしていることが分かる。

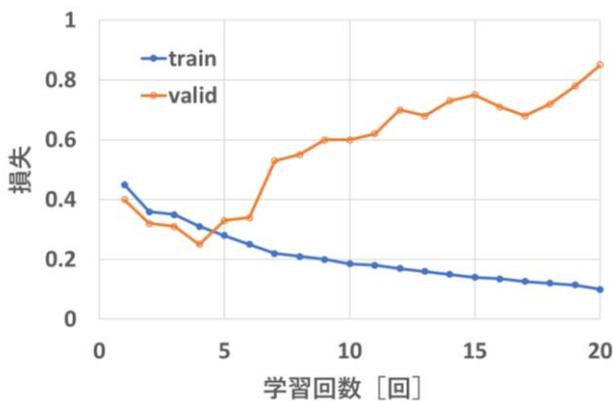


図 9: 学習回数に対する損失の推移

学習回数ごとの 1 の階級のラベルノイズ量とノイズ比率を表 2 に示す。ただし、ノイズ比率は 1 の階級でのラベルノイズ比率を表す。表 2 で示すように、学習回数 1 はいずれのクラスも学習回数 20 よりもノイズ比率が高い。つまり、

ラベルノイズに過剰にフィッティングしないよう学習することでクラス尤度の高いデータのノイズ比率を高くすることができる。これにより、確認すべきデータの量を削減し、効率的にラベルノイズを検出することが可能となる。

表 2: 学習回数別のラベルノイズ量

学習回数	1		20	
	Car	Pedestrian	Car	Pedestrian
クラス				
クリーン	60	0	77	820
ノイズ	206	52	43	157
ノイズ比率	77.4%	100.0%	35.8%	16.1%

5.3 実験 3: 損失関数の違いによるノイズ浄化の差

5.3.1 実験方法

関連研究[3]では、ラベルノイズにロバストな損失関数として MAE が挙げられている。また、[13]では、MAE を損失関数として用いると収束が遅くなることが示されている。そこで、実験 3 では、実験 1, 2 で用いた交差エントロピーとラベルノイズにロバストな MAE の 2 種類の損失関数についてラベルノイズ量の推移を比較する。表 1 に実験条件を示す。表 1 に含まれない条件として、学習回数を 1、ラベルノイズ比率を 40% と設定する。

5.3.2 実験結果

本実験によって得られた損失関数の違いによるラベルノイズ量の推移を図 10 に示す。図 10 で示すように、いずれの損失関数を用いてもラベルノイズを浄化できることが分かる。また、MAE は交差エントロピーよりも繰り返し回数を少なくすることができる。これは、MAE が収束に学習回数を多く必要とするとともにラベルノイズにロバストであるため、ラベルへの過剰なフィッティングを抑制することによる。そのため、一回の浄化でラベルノイズ量を多く削減することができる。



図 10: ラベルノイズ量の推移

6. おわりに

本論文では、ラベルノイズが含まれるデータセットをCNNに学習させて推論した結果、ラベルと異なる推論結果を示すデータのうち、クラス尤度の高いデータを抽出してラベルノイズを検出する手法を提案した。学習回数を少なくしてラベルノイズに過剰にフィッティングすることを避けることで、ラベルと異なる推論結果を示すデータセットにおいて、クラス尤度の高いデータにラベルノイズが多く含まれることを示した。また、損失関数としてMAEを用いることで交差エントロピーよりも繰り返し回数を減らして効率的にラベルノイズを絞り込むことができた。

今後の課題として、汎用的なデータセットでの検証や、マルチクラスにおけるラベルノイズ検出可能性の調査が必要である。

参考文献

- [1] Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp.2278-2324, 1998.
- [2] B. Frénay and M. Verleysen, "Classification in the presence of label noise: a survey," IEEE Transactions on Neural Networks and Learning Systems, pp.845–869, 2014.
- [3] A. Ghosh, H. Kumar and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in AAAI, pp.1919–1925, 2017.
- [4] T. Liu and D. Tao, "Classification with noisy labels by importance reweighing," IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(3), pp. 447–461, 2015.
- [5] M. Ren, W. Zeng, B. Yang and R. Urtasun, "Learning to reweight examples for robust deep learning," in ICML, pp. 4334–4343, 2018.
- [6] J. Li, Y. Wong, Q. Zhao and M. S. Kankanhalli, "Learning to learn from noisy labeled data," in CVPR, pp. 5046-5054, 2019.
- [7] Y. Kim, J. Yim, J. Yum and J. Kim, "NLNL: negative learning for noisy labels," in ICCV, pp. 101-110, 2019.
- [8] C. Zhang, S. Bengio, M. Hardt, B. Recht and O. Vinyals, "Understanding deep learning requires rethinking generalization," in ICLR, 2017.
- [9] D. Tanaka, D. Ikami, T. Yamazaki and K. Aizawa, "Joint optimization framework for learning with noisy labels," in CVPR, pp. 5552-5560, 2018.
- [10] P. W. Koh and G. Hinton, "Understanding black-box predictions via influence functions," in ICML, pp. 1885-1894, 2018.
- [11] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in CVPR, pp. 770-778, 2015.
- [12] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio and S. Lacoste-Julien, "A closer look at memorization in deep networks," in ICML, pp. 233-242, 2017.
- [13] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," in CoRR, 2017.
- [14] K. Lee, H. Lee, K. Lee and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in ICLR, 2018.
- [15] L. N. Smith, "Cyclical learning rates for training neural networks," in WACV, pp. 464–472, 2017.