MNIST と CIFAR10 の推論の弱点について

村上 茉耶　(指導教員：粕川 正充)

1　はじめに

機械学習用のソフトウェアライブラリである Tensorflow を使用した研究で、 MNIST という手書き数字の 28 × 28 ピクセルのグレースケール画像のデータセットを識別する研究と、 CIFAR-10 という 10 のクラ スで分けられている 32 × 32 ピクセルのカラー画像のデータセットを識別する研究がある [1],[3]。

本研究では、MNIST や CIFAR-10 の推論において、推論結果が曖昧という理由から「わからない」と判断されるデータについて考察した。MNIST では推論の最大の確率が 0.5 未満であるデータを抽出し、CIFAR-10 では、その前段階として、不正解であったものの推論におけるロジットを抽出した。

2　先行研究

CIFAR-10 に対して「天然物か人工物かどうか」という弱いラベルを与えて訓練させることで収束が良くなるという報告がある [4]。

3　実験環境

実験は以下の環境で行った。

- ◯ CPU : AMD ryzen7 1700 3.4GHz
- ◯ Memory DDR4 : 2666MHz，16GB
- ◯ OS : Ubuntu16.04LTS
- ◯ GPU : NVIDIA Geforce1080ti(Memory :11GB)
- ◯ Software : CUDA 10.0，python3.5.2，Tensorflow1.12.1

4　実験 1

MNIST は「0」〜「9」の手書き数字の画像を正しく分類するものであり、正解率は約 92%である。MNIST のテストデータセットの推論における確率の最大値が 0.5 未満のものの各数字の確率のグラフとその画像データを表示させた。MNIST のトレーニングデータセットで 1,000 回学習させ、MNIST のテストデータセットに対する正解率が 96.7%を示すプログラムで実験した。1,000 回学習させて、グラフとデータを表示させるのに、約 1 分半であった。

4.1　結果 1

テストデータセット 10,000 個に対し、推論における確率の最大値が 0.5 未満のものは 90 個得られた。得られた各数字の確率のグラフとその画像データ例を図 1 に示す。また、そのデータにおいて、正解と不正解であったものの個数を図 2 に示した。また、最大の確率が 2 つの数字で同率 1 位のものが 1 つだけあった。また、得られた 90 個のデータのうち、最大の確率 と 2 番目に大きい確率の値の差が 0.1 未満のものは 34 個あった。

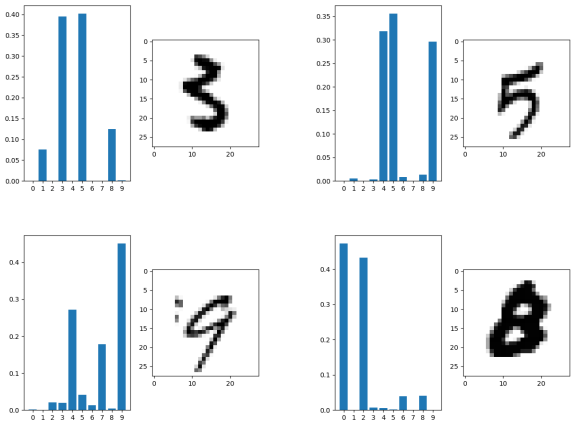

図 1: 例

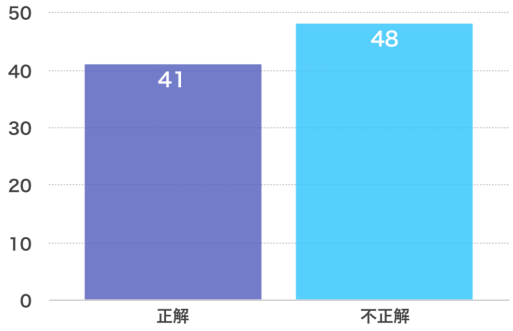

図 2: 正解と不正解の個数

4.2　考察 1

全体の正解率が 96.7%であるのに対し、確率の最大値が 0.5 未満のものの正解率は 45.6%であったので、確率の最大値が 0.5 未満であると明らかに正解率が落ちることから、上手く推論ができていないものが多い。また最大の確率と 2 番目の確率の差が 0.1 未満のものが 37.8%あることから、2 つの数字に判別されそうな曖昧なものが一定数ある。

5　実験 2

CIFAR-10 は「飛行機」、「自動車」、「鳥」、「猫」、「鹿」、「犬」、「蛙」、「馬」、「船」、「トラック」という 10 種類のカラー写真画像をカテゴリに分類するものである。CIFAR-10 のテストデータセットの推論において不正解であったもののロジットを正解のクラス名と間違えて予想したクラス名とともに表示させた。CIFAR-10 のトレーニングデータセットで 100,000 回学習させ、CIFAR-10 のテストデータセットに対する精度が 86.2%を示すプログラムで実験した。

ここで、ロジットとは、0 から 1 の値をとる p に対し

$$\mathrm{logit}(p) = \log(\frac{p}{1-p}) = \log(p) - log(1-p)$$

で表される値をいい、ここでは推論の関数 $inference$

がこれを返している (対数の底は 1 より大きければ何でもよい)。

5.1 結果 2

テストデータセットから関数 $inference$ に渡される際の 128 個のバッチに対し、不正解のものは 15 個得られた。不正解のもののロジットの最大値は、最大で 5.728、最小で 1.996 で、それらの平均値は 4.048 であった。その中で、最大のロジットと 2 番目に大きいロジットの差が 1 未満のものは 4 個あった。また、クラスを動物と乗り物に分けて考えたときの、「正解が動物のものを他の動物として予想したもの」、「正解が動物のものを乗り物として予想したもの」、「正解が乗り物のものを動物として予想したもの」、「正解が乗り物のものを他の乗り物として予想したもの」、の各個数を図 3 に示した。

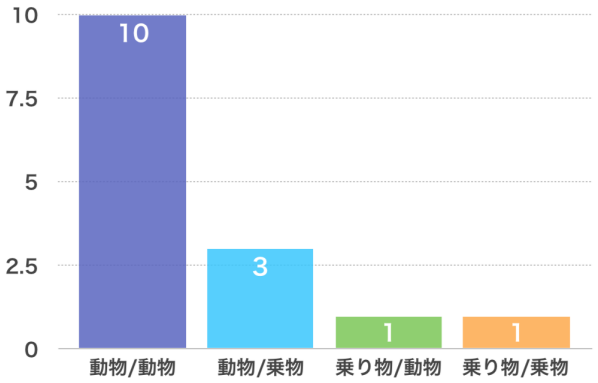

図 3: 正解のクラス/予想したクラス

5.2 考察 2

最大のロジットと 2 番目に大きいロジットの値の差が 1 未満のものが 26.7%あることから、2 つのクラスに判別されそうな曖昧なものが一定数存在する。また、動物のデータの推論においての不正解が圧倒的に多いことから、乗り物と比べ、動物の方が推論が難しい。また、その中でも動物を他の動物と推論してしまう間違いが圧倒的に多いことから、推論において動物同士の判断が難しい。

6 まとめ

実験 1 より、MNIST の推論における確率の最大値が 0.5 未満のものは、全体の正解率と比較して正解率が半分くらいに落ち、さらに推論から得られる結果が曖昧になっているものがそれなりに存在していることがわかった。

実験 2 より、CIFAR-10 においても、不正解のものの中に、推論から得られる結果が曖昧になっているものがあることがわかった。また、CIFAR-10 においては、動物の推論の方が乗り物の推論より圧倒的に間違いが多く、動物と乗り物のデータで精度が異なっていることもわかった。

7 今後の課題

実験 1 では、MNIST の推論における確率の最大値について、今回は 0.5 未満で区切ったが、推論が曖昧になるものを取り出すにはどこで区切るのが適切か実験して求める。

実験 2 では、1 つのバッチでしか不正解であったもののロジットと正解のクラス名と間違えて予想したクラス名のデータを得られなかったので、複数のバッチからデータを得られるようにコードを改変して、得られるデータを増やす必要がある。

また、CIFAR-10 においても、適切なロジットの値で区切り、推論が曖昧になるものを取り出す実験をする。

最終的には、MNIST や CIFAR-10 において、推論しても曖昧な結果が得られるものを「わからない」ものとして分類し、正解率や精度を高める。

また、CIFAR-10 においては、動物と乗り物で精度が異なる要因を調べ、それを推論に反映させることで動物の推論の精度を高める。

参考文献

[1] 中井悦司 : Tensorflow で学ぶディープラーニング入門 畳み込みニューラルネットワーク徹底解説, 株式会社マイナビ出版, 2016, ISBN978483996088

[2] 有山圭二 : Tensorflow はじめました 実践!最 新 Google マシンラーニング, 株式会社インブレ ス R&D , 2016, ISBN9784802090889

[3] Github, https://github.com/tensorflow/models/tree/master/tutorials/image/cifar10

[4] CIFAR-10 で「天然物か人工物かどうか」のフラグを入れて分類すると精度が上がるか？- Qiita, https://qiita.com/koshian2/items/f1e86fb5fe979f525c23

[5] TensorFlow の CIFAR-10 で実際に予測してみる - SuprSonicJetBoy's blog, http://blog.suprsonicjetboy.com/entry/2017/04/30/204951

[6] TensorFlow cifar10_eval.py のオリジナル画像で試す。, http://www.netosa.com/blog/2018/01/tensorflow-cifar10-evalpy.html

[7] ロジスティック回帰分析 (5) — ロジスティック変換の逆変換, https://bellcurve.jp/statistics/blog/14099.html

[8] TensorFlow による推論 — 画像を分類する CIFAR-10 の基礎 - Build Insider, https://www.buildinsider.net/small/booktensorflow/0202