Apache Sparkによるディープラーニングの並列分散処理

安藤 祐斗

富山県立大学　情報基盤工学講座
t815008@st.pu-toyama.ac.jp

April 9, 2021

本研究の背景

背景

機械学習の手法の一つであるディープラーニングは, 近年の進歩により, 画像認識などにおいての認識精度の向上, 自動運転, 医療研究などの幅広い分野での活用がされている.

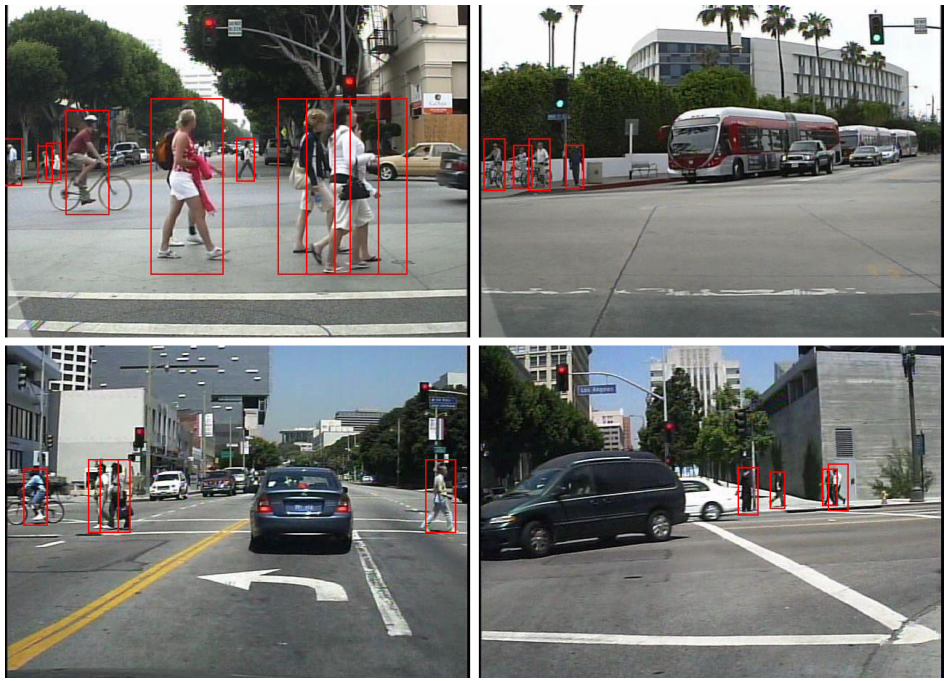

図 1: ディープラーニングの例（歩行者検知）

研究の目的

目的

本研究では,Apache Spark の並列分散処理機能を使いディープラーニングを実行する.
次に, この二つの組み合わせによって得られる優位性や, 既存のプログラムにはない新規性を確認する.

Apache Spark とは

大量のデータを複数のコンピュータで処理を行う, 並列分散処理を可能としたソフトウェア.
複数のサーバーでデータを格納するファイルシステムである HDFS（Hadoop Distributed File System）と, 格納されたデータを繰り返し加工し処理する RDD という分散データセットによって構成されている.

図 2: Spark の構成

ディープラーニング

ニューラルネットワークとディープラーニング

ニューラルネットワークとは, 神経細胞 (ニューロン) と神経回路網 (シナプス) で構成された, 人間の脳神経を模倣した数理モデルである. ニューラルネットワークは入力層, 中間層, 出力層の 3 つの層に分けられ, この中のさまざまな計算を行う中間層が,3 層以上のニューラルネットワークを用いた手法をディープラーニングと呼ぶ. 中間層を多く用いることによってより複雑な分析ができ, データの特徴を抽出することができる.

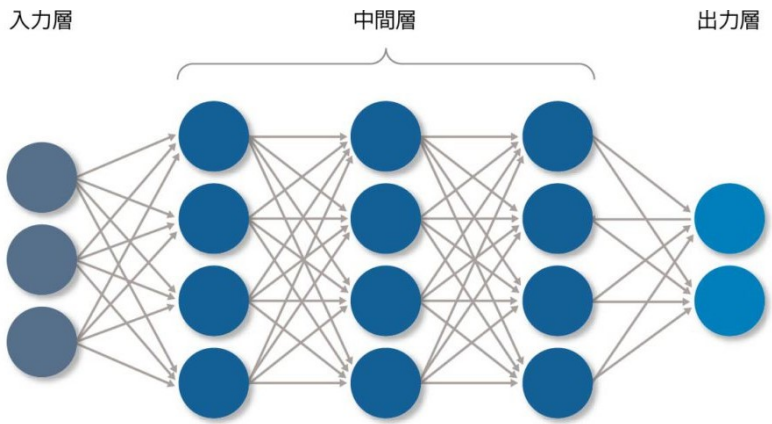

図 3: ニューラルネットワーク

BIGDL

BIGDL とは

Spark によるディープラーニングの分散処理を容易にするライブラリである.

現在,BIGDL の公式サイトに則り, 使い方を勉強中です.

サンプルプログラムの概要

最初に, 画像からパターンや物体の認識に最も利用されている, 畳み込みニューラルネットワークの一つである LeNet5 をベースに構築し,MNIST と呼ばれる手書き画像のデータセットを用いて学習をさせる. 次に, 学習で作成したモデルのテストを行い, 正確性を確認する.Spark を使い, これらを分散処理させる.

結果

現在,2 台以上で実行するとエラーが出るため,1 台でモデルの学習をしテストした時の, 学習にかかった時間と結果を示します.

```
2021-04-05 21:41:38 INFO  DistriOptimizer$:180 - [Epoch 15 60000/60000][Iteration 900000]
[Wall Clock 18621.790747666s] Top5Accuracy is Accuracy(correct: 9970, count: 10000, accur
acy: 0.997)
2021-04-05 21:41:38 INFO  DistriOptimizer$:180 - [Epoch 15 60000/60000][Iteration 900000]
[Wall Clock 18621.790747666s] Loss is (Loss: 4808.387, count: 10000, Average Loss: 0.4808
3872)
2021-04-05 21:41:38 INFO  DistriOptimizer$:220 - [Wall Clock 18628.989826929s] Save model
 to ./model/20210405_163107
2021-04-05 21:41:38 INFO  DistriOptimizer$:226 - [Wall Clock 18628.989826929s] Save optim
Method com.intel.analytics.bigdl.optim.SGD$mcF$sp@ae2db25 to ./model/20210405_163107
```

図 4: かかった時間

```
Top1Accuracy is Accuracy(correct: 9054, count: 10000, accuracy: 0.9054)
```

図 5: テスト結果

まとめ

進捗

- BIGDL と Spark の環境構築を 3 台の PC で行った.

今後の課題

- BIGDL の公式サイトにあるディープラーニング× Spark の例を 2 台以上で実行する. エラーの解決
- Deeplearningforjava といったライブラリもできたら試す.
- 発展してどんなことができるかを考える.