

Hadoop による並列処理での機械学習

Development of a QGIS plug-in for seamless integration of
ideation support and geoprocessing

安藤 祐斗

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

Teams, 9:50-10:15 Friday, December 4, 2020,
Toyama Prefectural University.

はじめに

並列分散処理

ディープラーニング

既存のサービス

サンプルプログラムの実行

まとめ

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、画像認識などにおける認識精度の向上、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。しかし、学習に長い時間を要する、学習しすぎると未知のデータに対応できなくなるといった問題点も挙げられる。

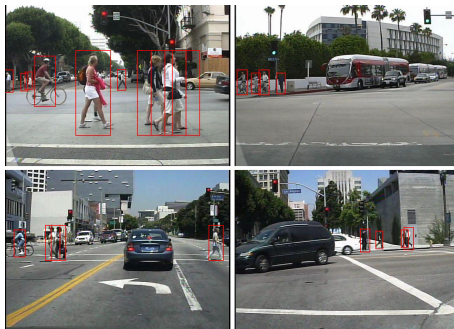


図 1: ディープラーニングの例（歩行者検知）

はじめに

並列分散処理

ディープラーニング

既存のサービス

サンプルプログラムの実行

まとめ

目的

本研究では,Hadoop 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.

はじめに

並列分散処理

ディープラーニ
ング

既存のサービス

サンプルプログラ
ムの実行

まとめ

Hadoop とは

大量のデータを複数のコンピュータで処理を行う、並列分散処理の可能としたソフトウェア。

データを複数のサーバーで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分散処理する Hadoop MapReduce というアルゴリズムで構成されている。

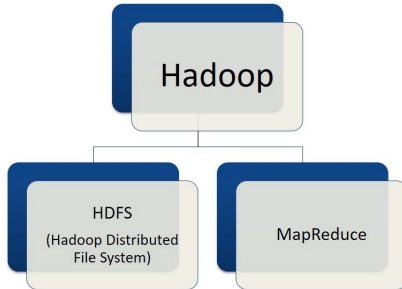


図 2: Hadoop の構成

ディープラーニングとは

人間が行うタスクをコンピュータに学習させる機械学習の手法の一つであり、ニューラルネットワークを多層に用いることが特徴である。それによって、データに含まれている特徴を深く学習することができる。

ニューラルネットワークとは

連携して電気信号のやり取りをする、神経細胞（ニューロン）の働きを模した数理モデルのことである。

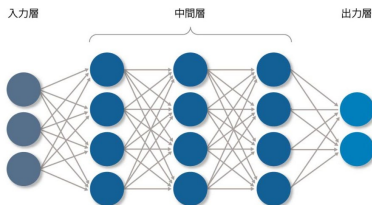


図 3: ニューラルネットワーク

Apache Submarine とは

機械学習, ディープラーニングのフレームワーク (PyTorch, TensorFlow など) を並列分散処理によって実行できるプラットフォーム.

はじめに

並列分散処理

ディープラーニング

既存のサービス

サンプルプログラムの実行

まとめ

はじめに

並列分散処理

ディープラーニング

既存のサービス

サンプルプログラムの実行

まとめ

マスターとスレイブ合わせて 5 台でサンプルプログラムを実行し、円周率を計算した。

```
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=1180
File Output Format Counters
  Bytes Written=97
Job Finished in 59.952 seconds
Estimated value of Pi is 3.1412000000000000000000
[tpu@master1 ~]$
```

図 4: 実行結果

進捗

- スレイブ PC に CentOS7 をインストールした.
- Hadoop を用い, PC5 台で円周率を求めるサンプルプログラムの起動を確認した.

今後の課題

- Apache Submarine を起動してみる.
- 他に Hadoop でディープラーニングを実行できるプログラムを探して実行し. 問題点を発見する.