Apache Sparkによるディープラーニングの分散処理

1815008　安藤　祐斗

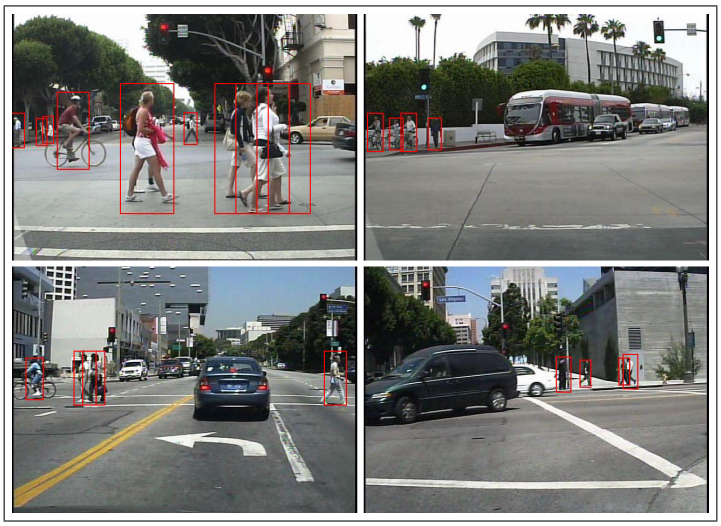

図 1: ディープラーニングの例（歩行者探知）

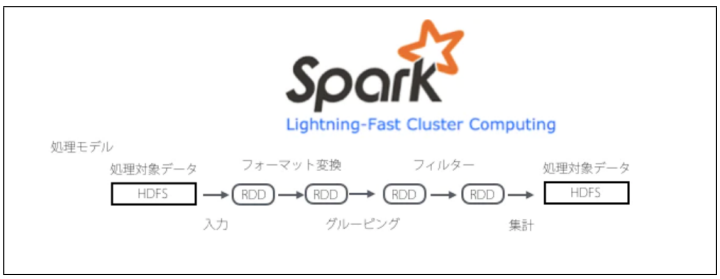

図 2: Spark の構成

1. はじめに

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 はディープラーニングの例のひとつである.

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.

2. Apache Spark による並列分散処理

Aparch 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である.HDFS（Hadoop Distributed File System）と呼ばれる複数のサーバーでデータを格納するファイルシステムと, 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 に主な Spark の構成を示す.

3. 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, 図 3 のように 3 層以上の

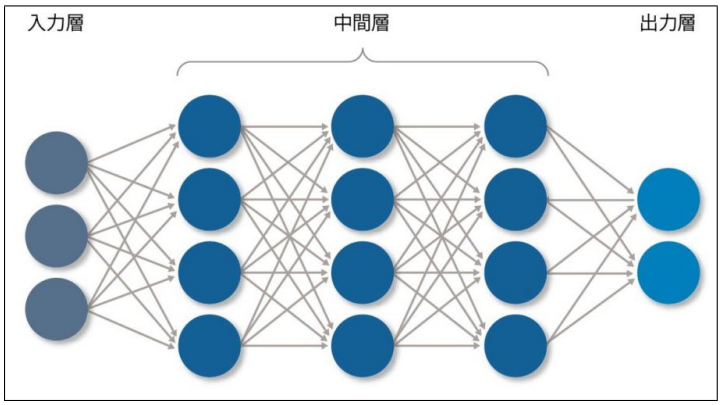

図 3: ニューラルネットワーク

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.

4. ディープラーニングを分散処理させるライブラリ

本研究では, ディープラーニングの分散処理化ライブラリの一つである,BIGDL を使用する.

5. 実験結果ならびに考察

6. 進捗状況と課題

現在,BIGDL と Spark のダウンロード, インストールとクラスタの環境構築を 3 台の PC で行っている.2 台以上でプログラムを実行する際にエラーが発生しているため, 解決を急ぐ. また,Deeplearningforjava といった他のライブラリもできたら試すことと. 発展してどんなことができるかを考える.

参考文献