AI理論と実践講座

本日のAgenda

- GPGPU
- AIプロジェクトの勘所 (データ編)
- Pythonによる基礎データ処理

GPGPU

(General-Purpose computing on Graphics Processing Units)

GPGPUとは？

GPGPU（General-purpose computing on graphics processing units; GPUによる汎用計算）とは、GPUの演算資源を画像処理以外の目的に応用する技術のことである。

元来GPUはリアルタイム画像処理向けのデータ並列計算とパイプライン処理に特化した命令発行形態を持ち、またGPUとメインメモリ間の帯域幅は通例狭いことが多いものの、GPUと直結されるVRAM間には十分広い帯域幅を備えていることから、補助的なベクトル計算機の一種とも言える。GPGPUは、GPUが持つこの特性を活かした汎用的なストリーム・プロセッシングの一形態である。GPUを主体として計算機システムを構成した場合、専用設計のスーパーコンピュータと比較して導入・運用のコストが圧倒的に安くなることから、HPCの分野で特に多くの注目を集めている応用技術でもある。 (wikipedia)

NVIDIA

Company History

1993年： Jensen Huang (現CEO)らが創業

1999年：上場

次世代のコンピュータは画像処理が鍵を握り、このためには汎用のCPUではなく専用のデバイスが必要となるだろう。コンピュータゲームが大きな需要を産むだろう

Year[62]	Revenue in mil. USD$	Net income in mil. USD$	Total assets in mil. USD$	Price per share in USD$	Employees
2005	2,010	89	1,629	8.81	
2006	2,376	301	1,955	16.76	
2007	3,069	449	2,675	25.68	
2008	4,098	798	3,748	14.77	
2009	3,425	−30	3,351	10.97	
2010	3,326	−68	3,586	12.56	
2011	3,543	253	4,495	15.63	
2012	3,998	581	5,553	12.52	
2013	4,280	563	6,412	13.38	5,783
2014	4,130	440	7,251	17.83	6,384
2015	4,682	631	7,201	23.20	6,658
2016	5,010	614	7,370	53.33	6,566
2017	6,910	1,666	9,841	149.38	7,282
2018	9,714	3,047	11,241	245.75	11,528

Products History

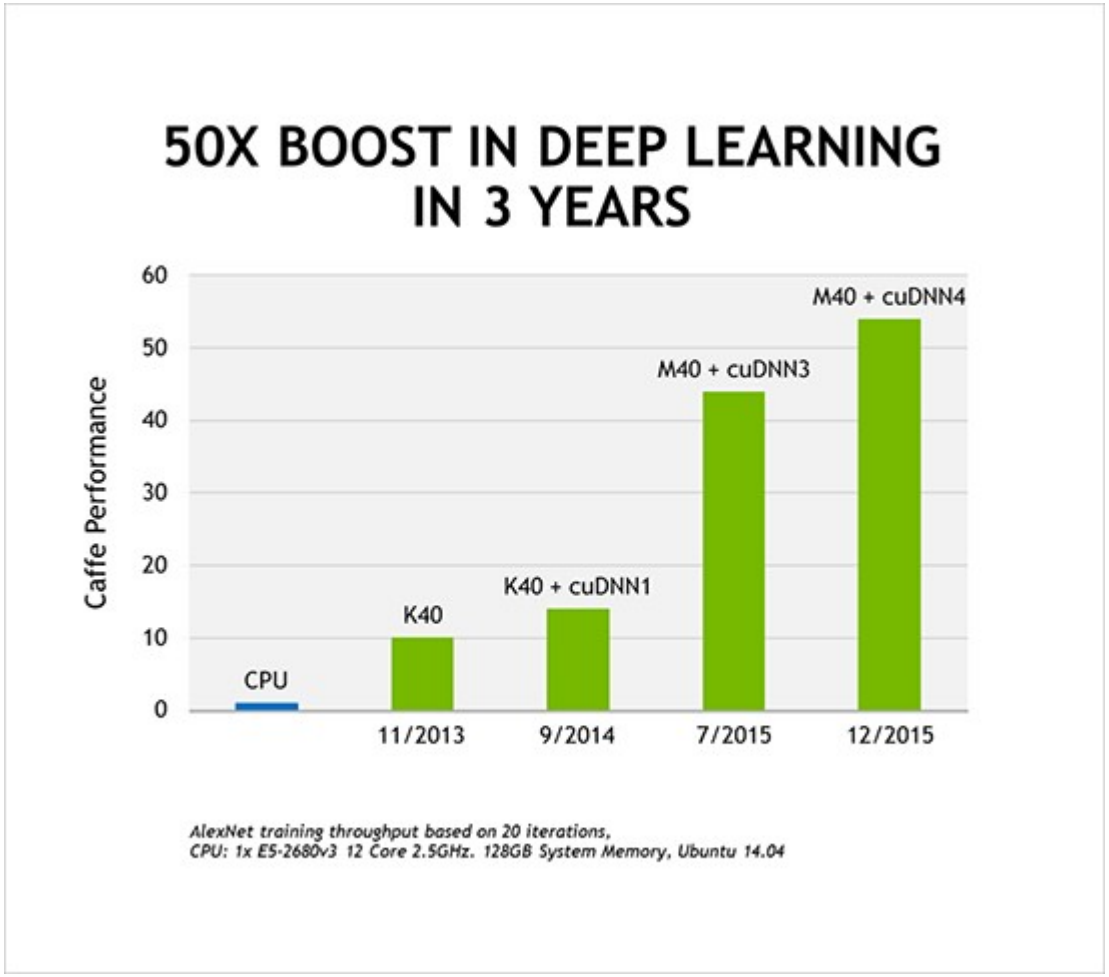
1999年 GeForce256を発売
2001年 MicroSoftとXboxを共同開発
2004年 SCEIとPlayStation3を共同開発

CPUメーカー第二位のAMDはGPUメーカー第二位のATIを買収し、GPUのCPU内蔵を開始。続いてIntelもCPUに内製のGPUをCPUに内蔵するようになる

Mobile向けのGPU SoCを開発するもイマイチ不発

Deep Learning & 仮想通貨マイニング向けのGPGPUで大躍進

Deep LearningとGPU



General
Purpose computing on
Graphics
Processing
Units
(GPGPU)

画像処理用の並列演算器を一般計算に応用

Deep Learningに使われる演算と相性がよく、学習時間を劇的に短縮

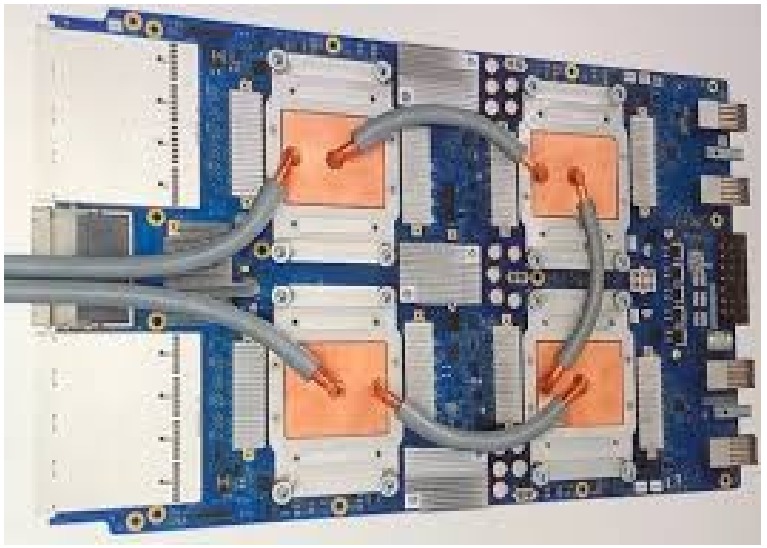
自動運転

自動車メーカーらのADAS（先進運転支援システム）開発から完全自動運転開発まで、トータルでサポートする製品を展開している

Audi, Volvo, VW, Daimler AG, Toyotaなどとパートナーシップ

自動運転車のプラットフォーマー戦略 (vs. Google?)

Competitor



Google TPU

Tensor Processing Unitは機械学習に特化したASIC
意図的に計算精度を落として高速化している
Googleの機械学習プラットフォームTensorFlowがサポートしており、AlphaGoでも使用された

License

更新されたエンドユーザー使用許諾契約書には、「データセンターでの使用禁止。データセンターの ブロックチェーン 処理が許可されている場合を除き、このソフトウェアは、データセンターでの使用にはライセンス供与していません」とある。NVIDIAの説明によれば、

「GeForceとTitan GPUは、複雑なハードウェア構造、ソフトウェアだけでなく、24時間365日稼働の熱要件があるデータセンターへの展開用に設計されたものでは決してありません。この点を明確にするために、我々は最近、GeForceに特化したエンドユーザー使用許諾契約書に、要求の厳しい大企業環境でGeForceおよびTitan製品が間違った使い方をされないように条項を追加しました」

「非商用用途やデータセンター規模では運用させていない他の研究用途に、研究者はGeForceおよびTitan製品を採用することが多いと我々も認識しています。Nvidiaはそのような使い方まで禁止しようとは思いません」

AIプロジェクトの勘所 (データ編)

AIはデータが命

AIは無計画に集めたデータを与えるだけでは機能しない

目的にあわせた収集データの設計

適切なデータサンプリング

適切なデータ分布

良質なアノテーション

収集データの分析

データ準備

目的に合わせた収集データ設計

- 特徴量の決定と収集方法の確立

天気予報

特徴量の決定：温度、湿度、気圧 …
収集方法と精度：必要なセンサは？
配置：位置、数

自動運転用画像

カメラ：解像度、視野角 …
設置：位置、数
撮影場所：地域、高速道路／一般道路
撮影環境：天気、昼間／夜間

適切なデータサンプリング

- サンプリングデータが解析しようとする対象の適切なサンプルになっているか？

  選挙予測：電話調査、インターネット調査、出口調査 …
  販売予測：地域、時期 …

- 一般に母数の適切なサンプリングになっていれば全体推定に対する信頼度を求めることが可能

適切なデータ分布

- サンプリングデータが解析しようとする対象の適切なサンプルになっているか？

タコとアルマジロの画像から２クラス分類

タコの学習画像が900枚

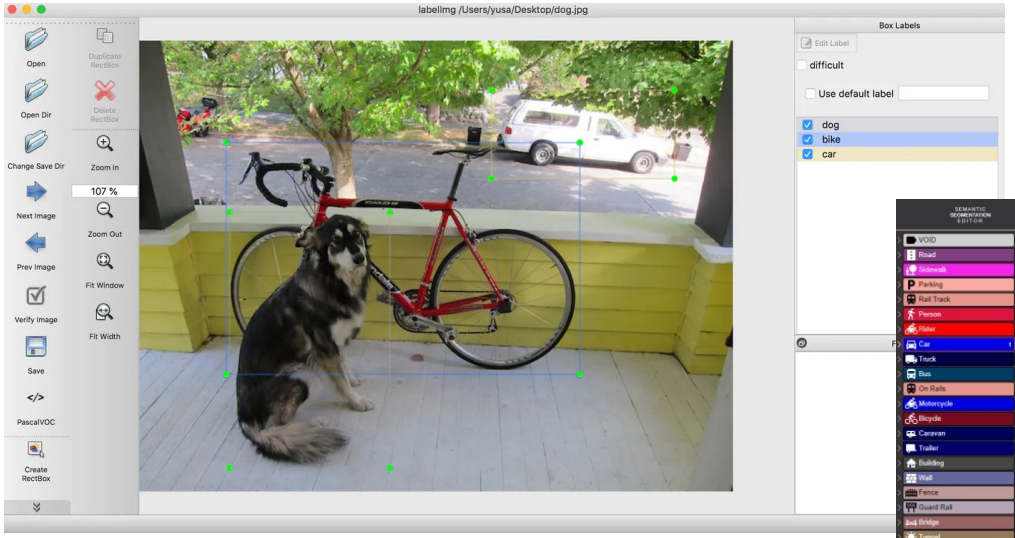
アルマジロの学習画像が100枚

機械学習はなるべく誤りを少なくするように学習する

常にタコと答えておけば正答率90%…

良質なアノテーション

教師あり学習には正解データを与える必要がある
正解データの品質は学習精度に直結
(大変な)高コスト作業



収集データの分析

収集したデータから基本統計量を算出しデータの特徴をつかむ

- 最大、最小
- 平均値
- 中央値
- **分散、標準偏差**

データのばらつきを表す。それぞれの数値と平均値の差を二乗し、平均を取ったもの。分散の平方根が標準偏差

機械学習を始めるとするAI技術は多くの業務の画期的なソリューションになる可能性がある

AIプロジェクトの勘所(データ編)

機械学習を始めるとするAI技術は多くの業務の画期的なソリューションになる可能性がある

一方でAIは良質なデータなしに動かない

さらには、データをざくっと与えてうまくやってくれるほどにAIはブラックボックス的に賢くない

データをじっと眺めて、その背景に潜む真実の気配を感じて、AIの助けを借りるような

そんな風にAIとうまくやる必要があります

だからそれが仕事の「データサイエンティスト」は高給取りなのです

Python基礎プログラミング

データの読み込みと処理

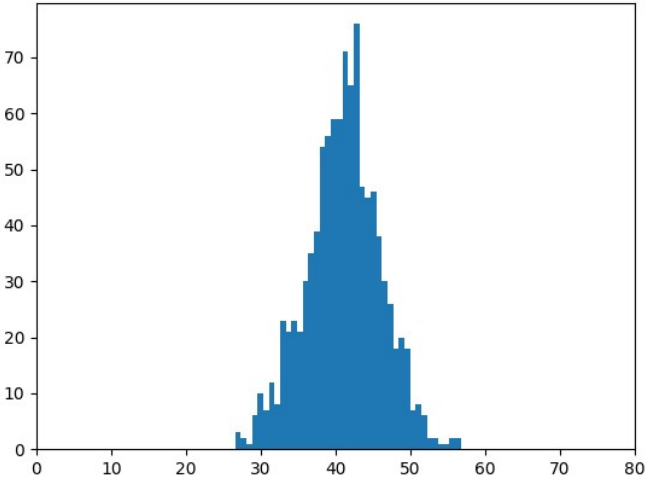
HR_0.csvおよびHR_1.csv

Emp ID,Name Prefix,First Name,Middle Initial,Last Name,...
857211,Ms.,Hermila,J,Suhr,…
514341,Mr.,Antonio,Q,Joy,…
314598,Prof.,Sebastian,J,Moores,...
...

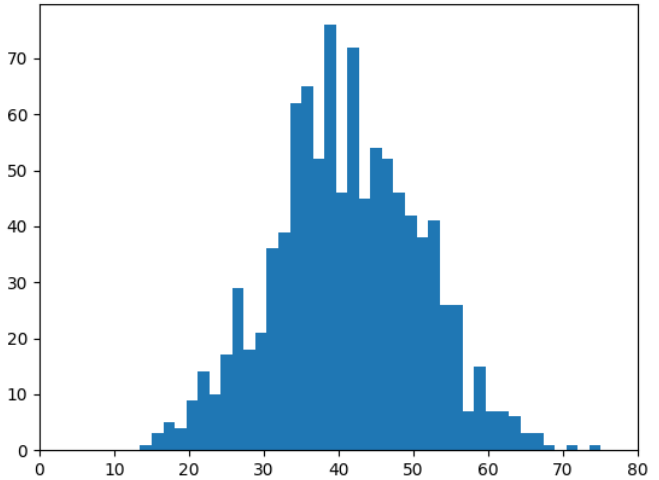
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statistics

df = pd.read_csv('HR_0.csv')

# データ数(行数)
num = len(df)

# 特定の列を取得
age = df['Age in Yrs.']

min =  min(age)
max =  max(age)
mean = statistics.mean(age)     # 平均
median = statistics.median(age)   # 中央値
stdev = statistics.stdev(age)   # 標準偏差
```

データの表示

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np

import statistics

df = pd.read_csv('HR_0.csv')

age = df['Age in Yrs.']

# ヒストグラム表示
plt.xlim([0, 80])
plt.hist(age, bins=40)
plt.show()
```



HR_0



HR_1

データの保存

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import statistics
import random

df = pd.read_csv('HR_0.csv')

for i in range(len(df)):
    df.at[i,'Age in Yrs.'] += random.randint(0, 10)

df.to_csv('out.csv', index=False)
```

保存されたout.csvを読み込んで、基本統計量を計算してみてください

演習

問1.

HR_0.csvの体重の最小値、最大値、平均値、中央値、標準偏差を求めよ

問2.

HR_0.csvで体重が平均値より20kg以上重い人の名前を表示せよ

Q&A