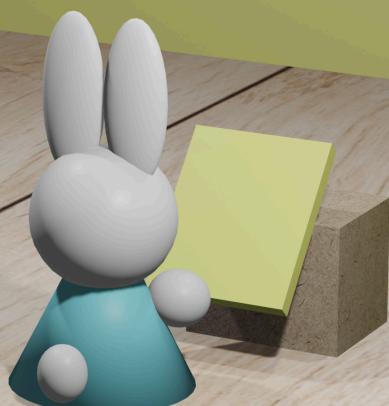




月刊

ZENKEI AI MAGAZINE

2021年5月号



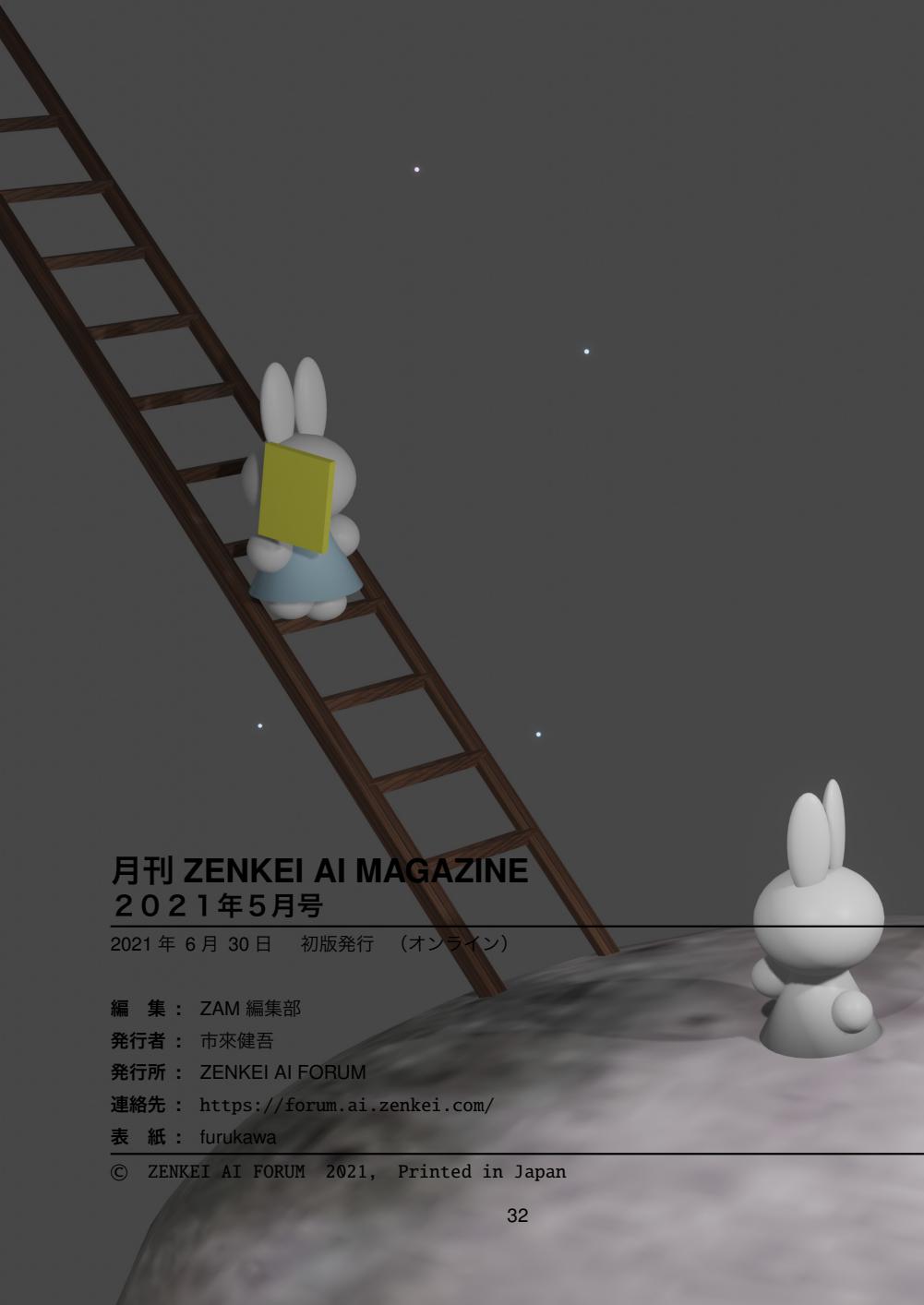
アイリスvsペギン furukawa
Kaggle奮闘記 石川達也
ZENKEI AI FORUM
への提言 米田稔

【月刊 ZENKEI AI MAGAZINE 2021年5月号】

目次

まえがき	2
第1章 当日のイベントの模様	3
第2章 アイリス VS ペンギン (furukawa)	4
第3章 Kaggle 奨闘記 (石川達也)	14
第4章 ZENKEI AI FORUMへの提言 (米田稔)	24
執筆者紹介	30
編集後記	31





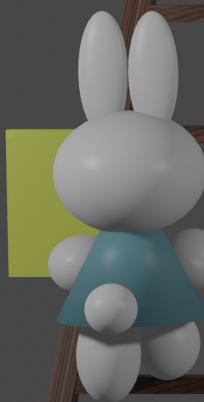
まえがき

月刊 ZENKEI AI MAGAZINE (ZAM) の5月号です。今年の2021年1月にスタートしてから本号で4冊目になります。はい、実はまだ4月号が完成していません (ZENKEI AI FORUM 4月の回は市来の一人喋りになったため、執筆者が私だけなので、発行が遅れて悲しいのも私だけ)。ということで、予定通り刊行される『月刊 ZAM 5月号』です。お楽しみください！

2021年6月30日

金沢にて

ZENKEI AI MAGAZINE 編集長
市来健吾



ZAM

編集後記

今月号（2021年5月号）の刊行は、予定通り、ZENKEI AI FORUM 6月のイベントの日（2021年6月30日開催予定）に間に合いました。

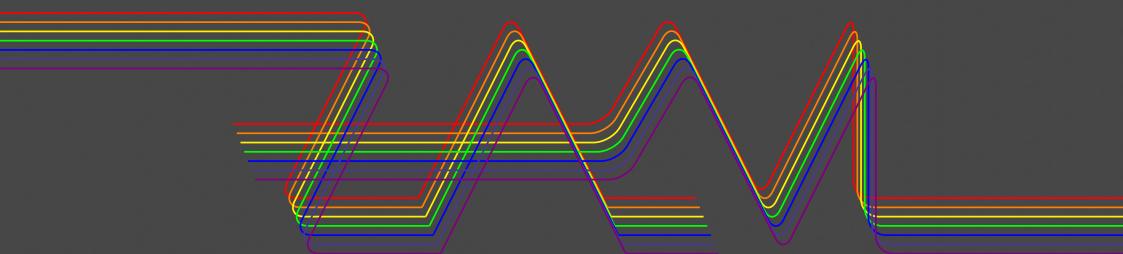
と言っても（まえがきに書いた通り）実は1つ前の『月刊 ZAM 4月号』がまだ出来上がってません。（これからがんばります。）

いつものことですが、執筆者の3名のみなさまには、きちんと原稿をいただきました。ありがとうございました。いろいろな人の書いたものが1冊の本になるのが、やっぱり雑誌の醍醐味ですね。

ところで来月の7月は ZENKEI AI FORUM もサークルとして過去2回参加した技術同人誌イベント『技術書典 11』が開催されます。私たちも今回も引き続き出典予定です。既刊の単行本 (furukawaさんの『ゼロからはじめるAI』や中野裕さんの『Jupyter Book で AI の解説本を書く方法』ほか) に加え、目玉は ZAM の有料版である『ZAM 季報』の創刊です。予定では『月刊 ZAM』の1月号から6月号までの内容をベースに、サークルメンバーによる書き下ろしコンテンツを加えて、魅力ある雑誌にしたいと考えています。みなさま是非ご参加ください！

今月号の表紙デザインおよび口絵のイラストは、『月刊 ZAM 2月号』でも表紙デザインをもった furukawa さんです。

(市来健吾)



執筆者紹介

今月の執筆者紹介です。

第1章

いちきけんご
市來健吾 (ZAM 編集長)

自分とみんなの人生が楽しくなるかなと思って、勝手に雑誌を作って、勝手に編集長をやっています。
結局いつも自分の原稿が一番遅いですね。



第2章

furukawa (全景株式会社)

職業：エンジニア、趣味：お絵描き
かわいいものが好きです。



第3章

いしかわたつや
石川達也 (全景株式会社)

1991年生まれ、愛知県出身。金沢大学大学院自然科学研究科数物科学専攻（旧 物理学部）修士課程終了。
愛知で自動車の組み込み系の開発に従事し、なんだかんだあって金沢に戻ってきました。現在は全景株式会社でプログラマをしています。AIは勉強し始めて1年くらい。



第4章

よねだみのる
米田 稔 (株式会社 COM-ONE 代表取締役社長)

石川高専卒後コマツソフト、コマツパキスタンソフト（イスラマバード）勤務を経て、2003年に株式会社 COM-ONE を設立し現在に至る。ZENKEI AI FORUM の前身である ZENKEI AI SEMINAR からのメンバー。



第1章

当日のイベントの模様

5月の ZENKEI AI FORUM (ZAF) は2021年5月26日に開催されました。当日の模様は YouTube のアーカイブ (<https://youtu.be/GwbYxcMWa7w>) でご覧になれますので、見逃した方はそちらでご覧ください。この日はゲスト3名お迎えしました。ZAM 本号の内容はこの ZAF をベースにして、以下のような構成になります。

- 【第2章】アイリス VS ペンギン (furukawa)

『ゼロからはじめるAI』シリーズの furukawa さんに、機械学習で使われるデータセットの最近のはなしを紹介していただきます。

- 【第3章】Kaggle 奮闘記 (石川達也)

全景株式会社でAIモデルなど開発している石川さんに、先日挑戦して結構いいところまで行ったという Kaggle のはなしです。

- 【第4章】ZENKEI AI FORUMへの提言 (米田稔)

前回4月の ZAF でいただいたコメントを発端に「社会に役に立つ ZENKEI AI FORUM」という視点でおはなしをしていただきます。



趣旨

- ・機械学習に使用される定番の座を狙う
PalmerPenguinsデータが2020年の夏に登場しました。
- ・Iris vs PalmerPenguins 比較をしてみました。
→ どちらもすぐ精度出て練習用として使いやすい。
Penguinsのほうが特徴量の数が多いです。
- ・プログラム言語 R を使ってみました。
→ 意外と簡単に使えました。

テーマ案：訪問介護のスケジューリング

訪問介護のスケジューリングにAIを応用できないでしょうか？

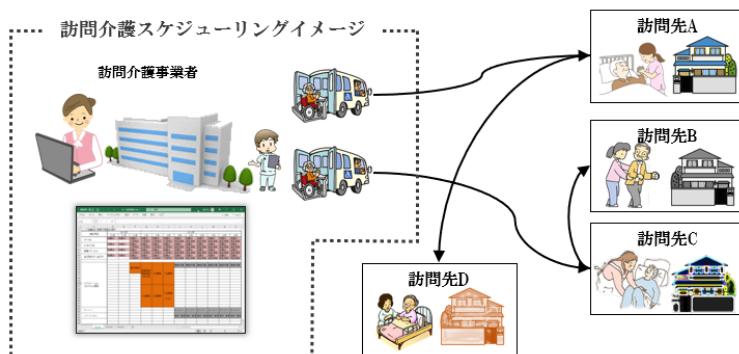
知り合いの訪問介護事業施設でヒアリングしたところ、現状はEXCELで手作業による編集でスケジューリングしているとの事でした。

→ 手間と時間を要す

→ 間違いや漏れが生じる

→ 複数の人が同時アクセス/更新

といった課題があるようです。5月のZAFでの話をしていたところ、YouTube 視聴者からの書き込みで「現状ホワイトボードでスケジューリングしているところもある」との事でした。



介護事業者のリソースと、訪問先のマッチングを的確に行えるシステムができれば、とても便利になります。学校の時間割のような機能をイメージしています。調査したところ介護スケジューリングを自動化するシステムやサービス現在はないようです。

またZAF当日、スケジューリングに加えて訪問先への移動の最適化機能も、さらに便利で面白いと考えられるという意見も頂きました。

おわりに

私が始めて劇場で見た映画は”ウェストワールド”という作品(1973年)でした。ユルブリンナーがロボット役で人間に復讐するというストーリーで、とても怖かったけどエキサイティングでした。その後”ターミネータ”がシリーズ化もされて有名な映画となっています。

ロボットやAIが人間を襲うなんて事は”SFの世界だけ”と誰もが考えていたと思いますが、第3次AIブームの先にはシンギュラリティが起こってしまうと予想する人も多いようです。

AIへのとても強い興味と感心を寄せる一方で、その脅威にも(どうにもできるものではありませんが)留意して行こうと思う今日この頃です。

機械学習データの定番。

Scikit-learnにもデータセットとして含まれている。

1936年頃から使われている。

3種類の花を、4つの特徴量（がくと花びらの長さと幅）で分類する。

各50サンプルずつ、計150データ。

PalmerPenguins

<https://allisonhorst.github.io/palmerpenguins/>

機械学習データの覇権を狙うデータセット。

2020年夏に公開された。

3種類のペンギンを、7つの特徴量（身体計測値、性別、計測年・場所）で分類する。

※地域や年齢などを含めたより詳しいバージョンも別である。

Adelie 152、Chinstrap 68、Gentoo 124、計344データ。

String型やNA（ごく少量）が含まれるので処理の練習にはなりそう。

Iris vs PalmerPenguins 印象

PalmerPenguins =

Iris + α + 分かりやすいかわいさ 

α = 文字列、NaN

α は別にほかのデータセットでもいいような。。

ちなみに

この絵文字：皇帝ペンギン

ピングー：皇帝ペンギン

タキシードサム：皇帝ペンギン

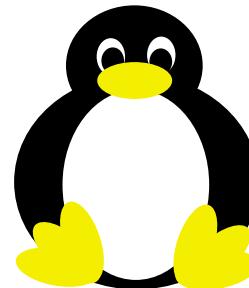
Suicaのペンギン：アデリーペンギン

Linuxのペンギン(Tux)：アデリーペンギン？

※開発者のリーナス・トーバルズが休暇中にコガタペンギンにかまれたことがマスコットにペンギンが選定されたきっかけらしいが、コガタペンギンはくちばしが黒いので、マスコットの見た目とは異なる。

バッドばつ丸：アデリーペンギン？

バッドばつ丸のパパ：ギャングのボス（いいのだろうか）



評価方法

初心者の機械学習練習用なので
簡単に一通りの練習ができるかを見る

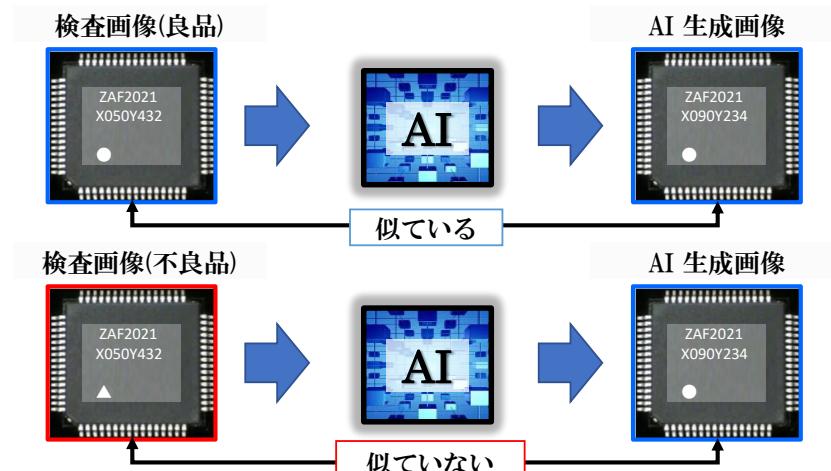
		アイリス 	
		できる	できない
ペン ギン	でき る	<ul style="list-style-type: none">簡単に精度を出す分布を見る	<ul style="list-style-type: none">NaNの処理文字列の処理コード例
	でき ない	<ul style="list-style-type: none">前処理なしで使うライブラリに入っている	-

参考事例 3：石川県工業試験場

石川県には全国で数ヵ所しかない工業試験場があります。今年3月に同場電子情報部の研究で、良品画像の学習だけで不良品の検査を実現した事例が発表されました。

製品の外観検査にAIによる画像処理が適用される例は近年多くなっていますが、学習データとして不良品の画像データを相当数準備する事が難しいという課題があります。通常の工業製品であれば、悪くても数百から数千に一つが不良品の割合でしょう。不良品の画像データを学習用に数百するのは不可能に近い訳です。

良品画像データからGANによって生成された画像と検査対象画像を比較する方式ですが、精度を実用レベルに上げるためにAEやVAEという手法を加えて3年がかりで十分な精度にたどり着いたとの事でした。



AIに入力された検査画像が良品の場合は似た画像を生成できるが
不良品の場合は似た画像を生成できない

2017年 正常画像のみ学習/異常検知 画像生成技術GANを用いた異常検知
評価 OK 80%, NG 100%
2018年 大きな画像サイズへの適用を進める
2019年 AE(AutoEncoder)を用いた
評価 OK 94%, NG 94%
2020年 AEに+ VAE(Variational AutoEncoder)
評価 OK99.3%, NG 99.0% → 実用化できる高い精度

既に石川県内の製造業企業で活用、適用を拡大してるとの事

参考事例 1：モルフォ

ディープラーニング推論エンジンを開発し、提供している会社です。
画像処理に強み、手ぶれ補正やAIレタッチのモジュール提供も。
ニュースに下記の記事発見！

モルフォAIソリューションズが最新AI技術を
活用したOCR処理プログラムの研究開発を
国立国会図書館から受託

確かZAFで「くずし文字認識」というテーマに取り組んでたような。
そう、一昨年、しかも対象はもっと古い時代の書物(もっと難読)だったはず。

例えば「古文書 → 文字認識 → 意味理解 → 現代文へ翻訳」のような事が
社会的需要として現在あれば、以前のテーマをあらためて進めるのもありではないでしょうか？

参考事例 2：エーアイ

音声合成を主なサービス事業としている企業です。「エーアイ」という会社名ですが AI 技術をどのくらい応用しているかは分かりませんでした(笑)。“音”を応用テーマの一つに考えたいので参考に挙げました。

ZAFでも市来主宰が音楽に AI を応用して「音楽と数理：才能にたよらない耳コピ」のタイトルで出版されていますね。

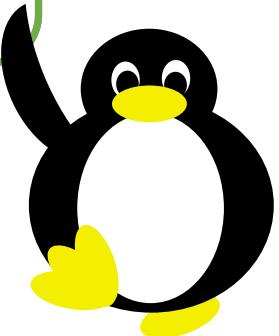
つい先日開催された高専のディープラーニングのコンテスト全国大会”D-CON2021”で、福井高専の打音検査システム”D-ON”が最優秀賞を受賞しました。企業評価額はなんと 6 億円、実投資額も 1 億円と報じられています。

マイコンを取り付けたハンマーで構造物をたたき、その打音データを専用のウェブページにアップロード。正常音と異常音を学習したAIが判定し、どちらの音に近いかを割合で示してくれるというシステム。“エッジAI”と“クラウドAI”的技術をうまく活用していて、高評価を得ています。

“音”はまだまだ AI 応用の可能性が大きい分野ではないでしょうか？

まあね。。。。

ペンギンデータはアイリスの後発なので
アイリスでできることは
ペンギンでもできるように
なってます。



項目を比較

Iris : 5項目

sepal length (cm), sepal width (cm), petal length (cm), petal width (cm), target

PalmerPenguins : 8項目

species, island, bill_length_mm, bill_depth_mm, flipper_length_mm,
body_mass_g, sex, year

※調査の元データ（17項目）も付属している

精度：ある程度のスコアが出てほしい

それぞれ2項目だけを使って

train:test=8:2 でデータを分けて LightGBMClassifier で学習させた。

Iris testscore : 0.93

Penguin testscore : 0.92 (6/69が推論間違い)

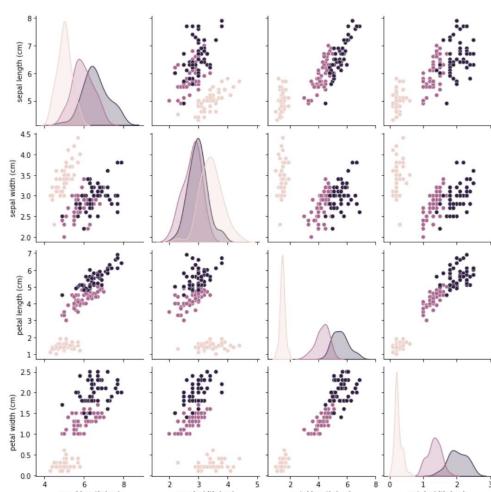
target	target_pred	TF
13	2	0 False
14	2	1 False
15	2	1 False
35	0	1 False
38	1	0 False
54	1	0 False

0.92なら十分。

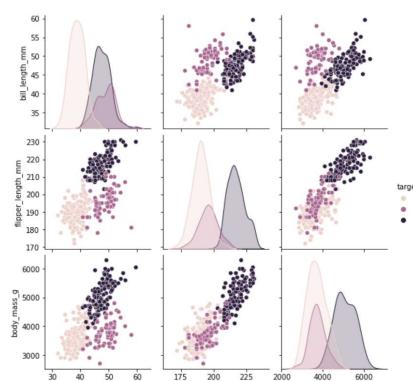
間違いには特に傾向はみられなかった

ペアプロット

Iris

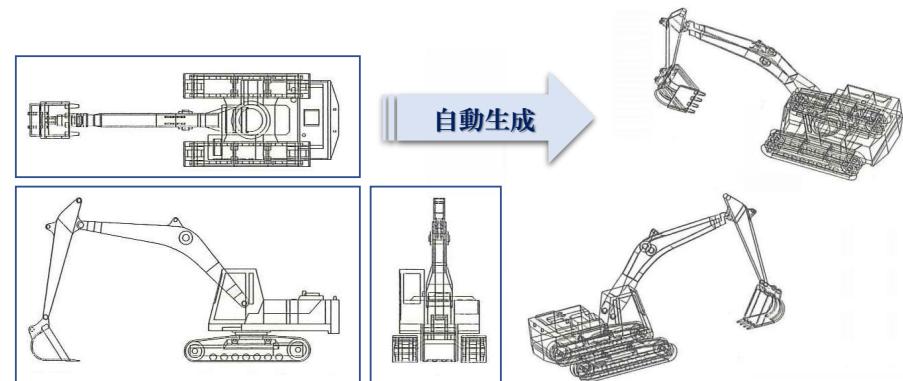


PalmerPenguins

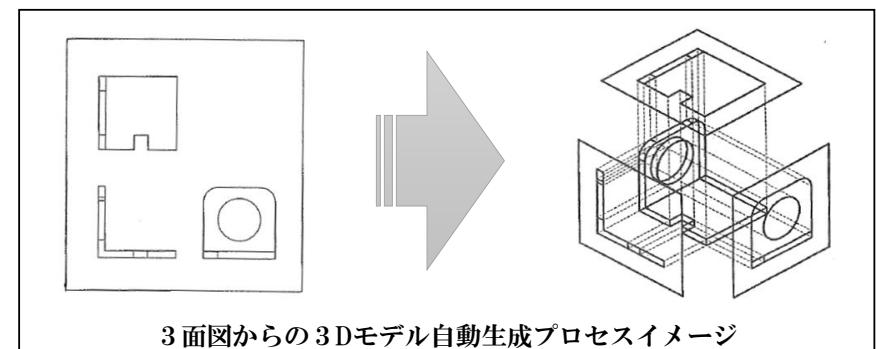


いい感じに分布している
基本的なことは一通りできそう

いきなり余談 です!! かつて取り組んだ「AI的」研究
3面図から3Dモデルを自動生成するソフトを紹介します



前職でプログラマーをしていた時代1987年頃の取り組みでした。
正面図、平面図、側面図にCADで描かれた直線や曲線の要素データを”総当たり”
させて、3D空間上で論理積を幾何学計算で求めるという原理でした。なので、要素
データの数が多くなり、例えば平均1,000だとすれば10億回の計算が必要となり、当
時大変高価(1時間で私の1月分の給料よりはるかに高額でした)なメインフレームの
CPUリソースを使い切って”クレーム”を受ける事もちょくちょくありました。



後に性能が飛躍的向上を遂げたパソコンやワークステーションにも移植して数社で
利用されました。3D CADの急速な普及によって需要が消滅して行きました。
もしかしたら、今のAI技術なら高度で的確な3D生成が可能かも知れません。土木・
建設分野などではまだ3Dモデル自動生成の需要があるのではないかと思います。

ZENKEI AI FORUMへの提言

米田 稔

“AIを適用して社会の課題を解決する”
をZAFのテーマの一つとして志向しましょう



寄生虫分類アプリ “Parasite Finder Project”
10年前の技術では識別精度が”20%”程度と低く
実用化できませんでした。
しかし、DLを適用した実験では”95%”に到達、
AI技術に大きな可能性を感じています。

ZAF を地方イノベーションの契機に!!

提言の背景

ZAFは毎回楽しみに参加していますが、もっぱらAIの技術情報などが主な話題となる事が多いです。しかし、数学や流体力学などの高度な話になると、段々追いついて行けなくなりそうに感じていました。
そう、話題をAIの理論から応用へ持って行って、ついて行けそうな領域にしなければ(笑)と思っていました。
2021年5月のZAFで話をする機会を頂き、座談会形式で意見交換する事になったので、AI応用分野の話をしようと考えました。

製造/土木/建設/農業/交通/物流/医療/介護/音楽/芸能/ etc. 適用され実用/活用されているものも多くありますが、まだまだ応用の余地は色々あるはず、そのようなケース/シーンをみつけられれば、実データで検証/評価して現場へ適用する事ができると考えています。

何よりも企業経営者としては、売上/利益につなげられ、発展できそうなテーマに取組めれば嬉しいです。

PalmerPenguinsに勝ち目はあるか？

NANの処理や文字列の処理など

PalmerPenguinsでできてIrisではできないことというのもあるが、

「それがやりたいならTitanicでいいじゃないか」という考えもあります。

※Titanic：タイタニック号の乗客データから生存者を予測する趣旨のデータセット



余談

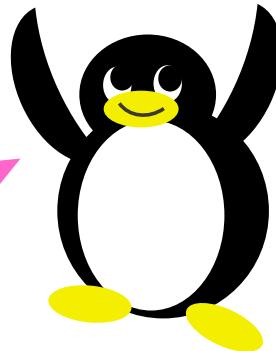
PalmerPenguins vs Titanicを比較

	いいところ	気になるところ
ペンギン	<ul style="list-style-type: none">・テーマがかわいい・簡単に精度が出る	<ul style="list-style-type: none">・精度上げについて工夫するところがあまりなさそう
タイタニック	<ul style="list-style-type: none">・「客室の等級と生存率は関係するのではないか？」など、人間の推測を分類に使うことができる。←強い魅力・工夫によって精度がだいぶ変わる。	<ul style="list-style-type: none">・テーマが重い

PalmerPenguinsを盛り上げるには

フォーラム中に出た意見で、
「このペンギンデータと対になるペンギン画像セットを用意し、
“機械学習も深層学習（画像分類・物体検出）も両方できる”
という売りにしてはどうか？」
というものがありました。

それだー！！



“機械学習”というくくりで考えると
特徴量を増やしたりする方向へいってしまうが
“初学者学習用データセット”と考えれば
機械学習も深層学習もできる、
というのはめっちゃうれしい！

PalmerPenguinsのすてきさ

個人的に感じたPalmerPenguinsのすてきさは、その提供方法です。
そもそもテーマになっているペンギン自体がかわいいです。
コード例が複数提供されていてすぐ使えます。
そして、プロットされるときの色が指定されていて、その色がおしゃれなんです。
昔の『フランニとズーイ』の表紙を思い出します。

公式ページのアートワークも
カラフルなペンギンのイラスト満載がかわいいです。

かわいくて、すぐ結果が出る、
そうすると、やっていることがうれしくなります。
そこが、いいなと思いました。

4.7. アクシデント(サプライズ)発生

妻が妊娠しコンペどころでなくなってしまいました。
基本土日に作業を考えて平日モデルの学習をしていたのですが、優先すべきことは考えるまでもないので土日にはほぼ手が出せなくなり、バタバタしているうちに期限が迫り、あまり考えずに PublicScore の精度が良いものを選んで提出してしまいました。

筆者の初めてのコンペはこれで終わりました。

5. 結果

 Makerere AI Lab
Cassava Leaf Disease Classification
Identify the type of disease present on a Cassava Leaf Image
Research · Code Competition · 3 months ago

 602/3900
Top 16%

602/3900 位(Top 16%)でした。
残念ながらブロンズメダルも取れなかった。難しいですね。精進します。

6. 反省

今回のデータセットはノイズが大きく、参加者のほとんどが Accuracy:0.9 以上を超えることが出来ませんでした。そのため、リーダーボードの Score は団子状でその中で一つ抜けることは難しく筆者には出来ませんでした。筆者のソリューションはで SEResNext50 が最も精度が良かったのでこのモデルを 5 つ学習し提出しましたが、正直モデルは深いモデルにしなくとも ResNet18 程度で Accuracy:0.9 近くにはなったので 1 つのモデルに固執せず複数モデルのアンサンブルでやったほうがよかったです。そういう視点がなかったので勉強になりました。また、悔やまれる点として筆者が提出したモデルが悪かっただけで提出しなかったモデルにブロンズメダルを貰えるモデルがあったことです。完全に期限に焦ってそこまで考えずにモデルを提出したことが原因で、ディスカッションで見かけた「PublicScore ではなく CrossValidation の Score を信頼しましょう」の言葉が後々で響きました。次は気をつけようと思います。

いろいろ考えてこのコンペに挑みましたが、実際は「気合と根性で良い結果が出るまでやった」が正直なところです(途中でやめちゃいましたが)。また次のコンペも地道に頑張ろうと思います。

7. まとめ

いろいろな方法でアプローチを見ることが出来勉強になりました。あと

CrossValidation を信じましょう

また丁度いい感じのコンペがあったら一緒にやりませんか？

話変わって、気になるのが

Cassava / seresnext50_32x4d [inference] Succeeded 0.8788 0.8813 □
(version 1/7)
7 months ago by Misyaguziya

From Notebook [Cassava / seresnext50_32x4d
[inference]]

(右 : PublicScore、左 : PrivateScore。実際コンペの開催中は PublicScore のみ見える。)

PublicScore で Accuracy:0.8813 と初めてにしては十分精度が出ました。

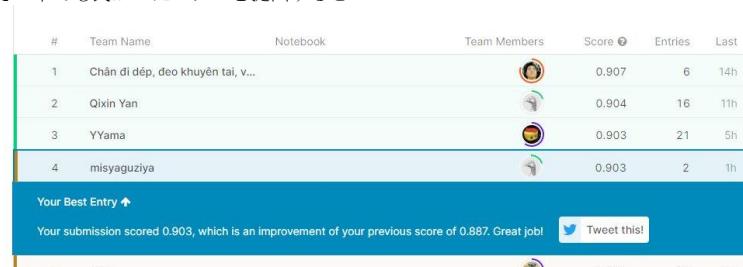
4.6. 学習の繰り返し

ここからモデルや DA、インプットの画像サイズを変更して精度を上げていきました。
自分が使用したモデルは以下になります。

- HrNet
- ResNet18
- ResNet31
- ResNet50
- SEResNext50
- EfficientNetB3
- EfficientNetB5

もっと多くのモデルを使って比較したかったのですが、一回の学習時間が半日以上掛かり、そこまで試行錯誤出来ませんでした。

その中でも良かったモデルを提出すると



Accuracy:0.903 で暫定 4 位になっていました!(misyaguziya というのが筆者)
この感じなら上位に行けると気を良くして続けて頑張っていこうと思っていたら…

Palmer Penguin のデータセットのチュートリアルは

R 前提です (pythonではない)。

Rでのコード例が入っています。

Rは統計に強いプログラム言語だそうですが、

私は使ったことありません。

なんとなく難しそうなイメージです。

使ってみました。

→ 意外に使いやすいです

R

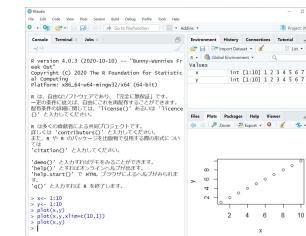
■Rとは

統計解析向けのプログラミング言語
プロットが簡単にできる

■インストール：無料でできる

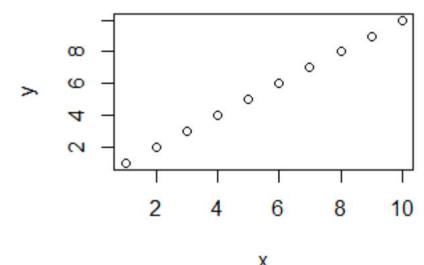
Rをインストール (v4.0.3)

R studioをインストール

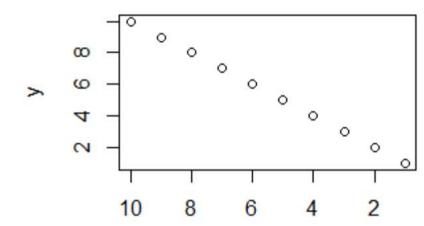


右図がコード例です。シンプル。

```
> x<- 1:10  
> y<- 1:10  
> plot(x,y)
```



```
> plot(x,y,xlim=c(10,1))
```



ColaboratoryでRを使う

上部バーより「ランタイム」→「ランタイムのタイプを変更」→R
あとは、PalmerPenguinsのチュートリアル通りに実行すれば使えます。

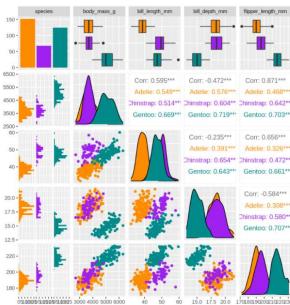
```
[2] install.packages("palmerpenguins")  
Installing package into '/usr/local/lib/R/site-library'  
(as 'lib' is unspecified)
```

```
[3] library(palmerpenguins)  
data(package = "palmerpenguins")
```

```
[4] head(penguins)
```

```
A tibble: 6 x 8  
  species   island bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex  year  
  <fct>     <fct>      <dbl>        <dbl>          <dbl>       <dbl> <fct> <int>  
1 Adelie    Torgersen     39.1         18.7          130.      3750   female 2007  
2 Adelie    Torgersen     39.5         17.4          135.      3800   female 2007  
3 Adelie    Torgersen     40.3         18.0          133.      3250   female 2007
```

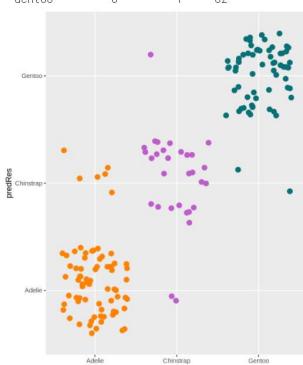
きれいです。



R で PalmerPenguins 分類を推論

linear modeling : 0.93

Adelie	Chinstrap	Gentoo
67	2	0
Chinstrap	6	27
Gentoo	0	1
		62



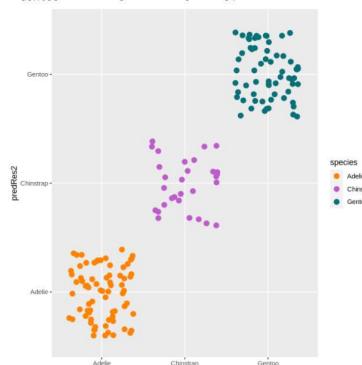
```
((67+2+6+27+2+1+62)-(6+2+1+2))/((67+2+6+27+2+1+62))
```

0.934131736526946

12

svm : 1.0

Adelie	Chinstrap	Gentoo
73	0	0
Chinstrap	0	30
Gentoo	0	0
		64



```
(73+30+64)/(73+30+64)
```

1

第2章 アイリス VS ペンギン

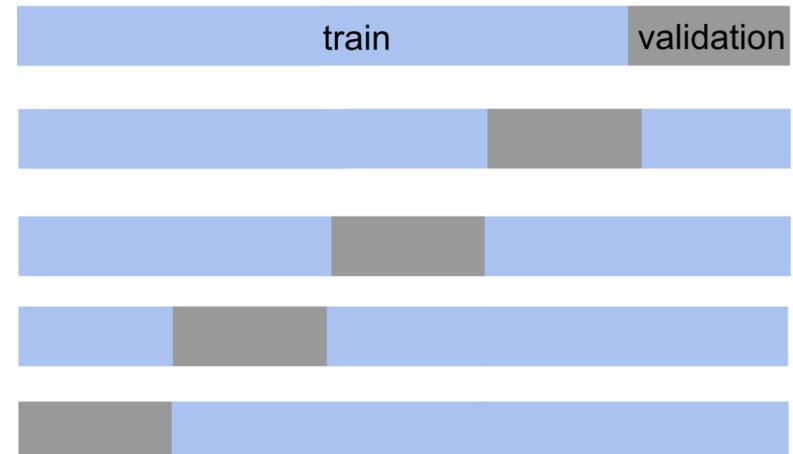
なんで精度が上がるのか

ここについては論文(<https://arxiv.org/abs/1908.02983>)を読んでもらうしかないし、筆者も完全に理解できているわけではないですが、今回の場合に限って言えばラベルノイズが大きいことが原因の一つです。疑似ラベルと元々つけられていたラベルが異なる場合がよくあり、元々のラベルでの精度を悪くともに提出したモデルの精度は良かったことが多かったです。モデルが人間のアノテーションを修正したと考えています。

4.4. CrossValidationについて

精度の検証のため、CrossValidation を使用しました。

今回は 5fold、つまり学習データを下図のように 5 セットに分割し 4 セットを学習に使用し 1 セットで検証という組み合わせで同一のモデルを 5 つ学習させる形で行いました。



5つのモデルの精度の平均でモデルの精度を評価して行きます。

4.5. ベースライン作り

兎にも角にもモデルを提出して Score を出さないとリーダーボードにも乗りません。自分で作っても良かったのですが、すでに Accuracy:0.8 を超えている Code が投稿されていましたのでこれを流用して始めました。

(<https://www.kaggle.com/vasufuminakama/cassava-resnext50-32x4d-starter-training>) どうやって提出したら良いかわからない筆者にとってはとても助かる内容でした。こういった Code を共有してくれる Kaggler の方々には感謝です。

公開されている Code から方針の設定を行い、モデルを学習させ Kaggle に投稿するとこのような結果が出ました。

石川達也

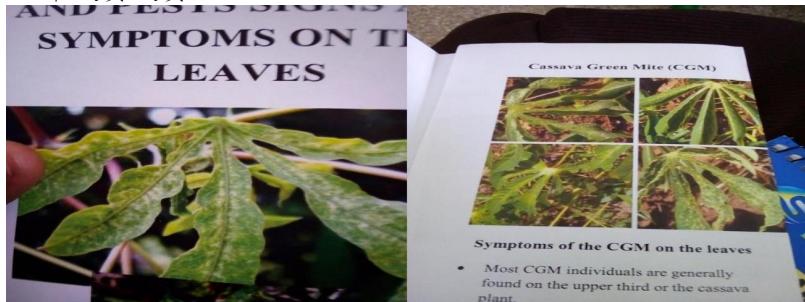
21

まとめ

- 人が写っている



- 本の写真の写真



こういったデータを取り除いていきます。

正直このデータセットを見たときは酷いと思いましたが、実際の現場で何も知らない人が撮ってきたデータはこの程度のデータでそれを何とかするのもコンペの内なのだと今は納得しています。

4.3. Pseudo-Label(疑似ラベル)について

疑似ラベルで学習することでノイズに強くしようという試みです。前回コンペの1位のソリューションや他の上位のソリューションでもやっていて効果が大きかったとあったので使いました。

学習の方法

- ① ラベル付けされたデータでモデルを学習させる(これを教師モデルという)
- ② ①で作成した教師モデルを使用してラベル付されていないデータに予測値をラベルとしてつける(これを擬似ラベルという)。今回の場合、全てのデータにラベルがついているが 1/3 のデータで教師モデルを学習し 2/3 のデータのラベルを教師モデルの予測値に置き換えた。
- ③ ②の疑似ラベル付きのデータとラベル付きデータを混ぜてモデルを学習させる(これを生徒モデルという)
- ④ 上記を繰り返し推論させ疑似ラベルを更新していく

PalmerPenguins自体は機械学習の勉強用としていいデータです。

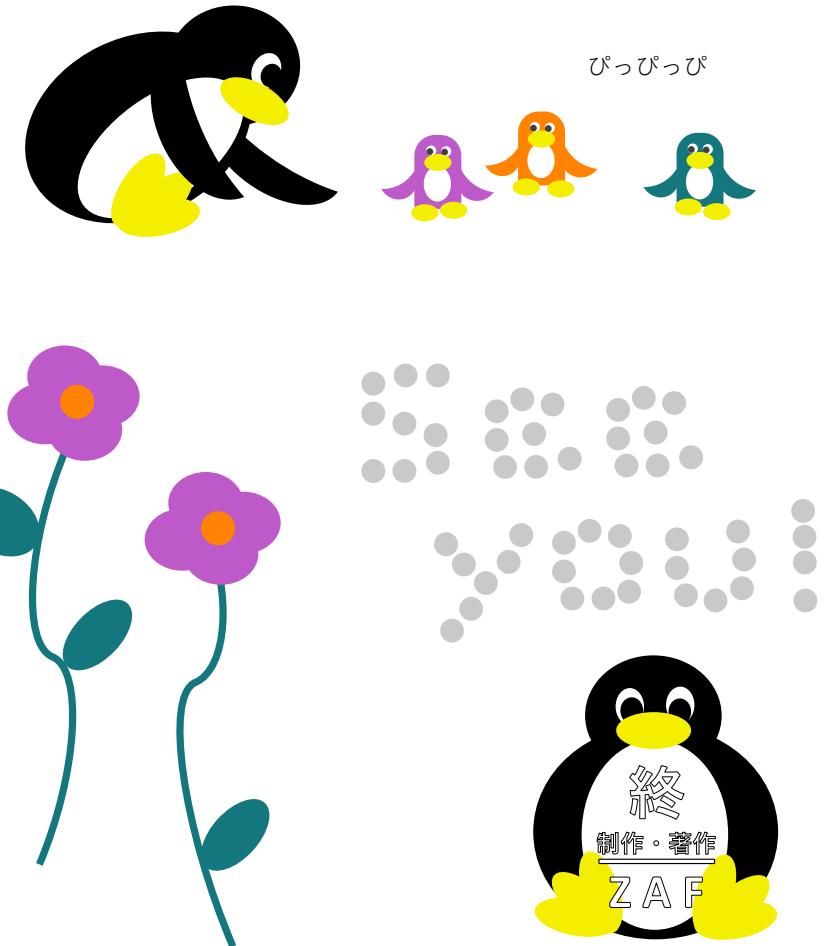
サイズも小さくて扱いやすいし、テーマがかわいくて楽しいです。

「Rを今後使いたいか?」という質問をフォーラム時にいただきました。

現在使っているpythonは

ライブラリが豊富で機械学習以外のこともできますし、特に不便はないです。

Rに乗り換えることはないと思います。



Kaggle 奮闘記

石川 達也

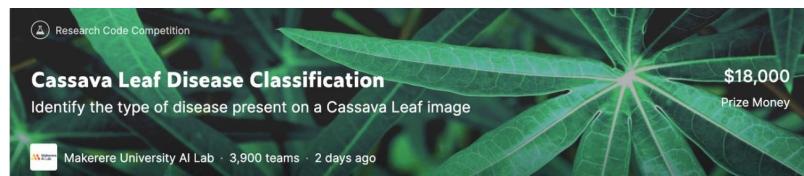
1. はじめに

筆者が DeepLearning に触れるようになったのは、ZENKEIに入社してからのことです。今から大体一年くらい前になります。前職もプログラマでしたが、組み込み系であり DeepLearning というものは無縫でした。DeepLearning について基本的なことを勉強し、numpy で ResNet や DenseNet の実装をしたり Pytorch の使い方を覚えて手を動かして画像分類なんかができるようになつたりしたのが入社から 3ヶ月くらいだったと思います。なので DeepLearning 知識 3ヶ月くらいの初心者の経験談になります。この経験が誰かの参考になれば幸いです。

2. 参加のきっかけ

この始まりは ZENKEI チームが以前(筆者の入社前に)に参加した「くずし字コンペ」を一人で挑戦してみたことでした(このコンペも勉強になることが多かったので皆さんも参加してはどうでしょうか? ちなみに最終 Score は 0.863 でした)。結局、ZENKEI チームの Score は超えられず、次なにしようかなと思っていたところに twitter で「Cassava が面白そう」という話が流れてきました。

Cassava とはなんぞや? と思いながら調べると以下のコンペを発見しました。



<https://www.kaggle.com/c/cassava-leaf-disease-classification/overview>

Cassava(タピオカの原料)の葉の画像から病気の種類を判別するコンペのようで問題設定としても物体検出とかの難しそうなものではなく単純に画像分類問題なのでとっつきやすそうだなと思いました。

コンペの概要

- 期間は 2020 年 11 月 20 日 ~ 2021 年 2 月 19 日(3ヶ月)
- 病気 4 種類 + 健康 1 種 の 5 クラスの分類問題
- データセットは Cassava の葉や茎や芋本体の画像
- 学習用のデータセット 21,367 枚
- テスト用のデータセット 15,000 枚
- 評価方法は単純な Accuracy

期間も 3ヶ月と長いく、なにか成果は出るだろうということで参加を決意しました。では初めてのコンペやってきましょう。

3.5. ディスカッションに情報はないか

事前知識がゼロなので情報収集です。このコンペは問題設定からも参加しやすく多くの人がディスカッションしていました。収集した情報は以下になります。

- 「Cassava Leaf Disease Classification」のコンペは以前にもやっていた <https://www.kaggle.com/c/cassava-leaf-disease-classification/discussion/198131>
- データセットのラベルにノイズ(ラベルの付け間違い)が多く精度が出ないあまりにも酷いので「このコンペは宝くじですか?」なんて投げかけもありました。
- PublicScore ではなく CrossValidation の Score を信頼しましょう 少し前のコンペで PublicScore と実際の結果である PrivateScore がかけ離れていたこともあり、PublicScore を信用するなど強く呼ばれていました。

4. 作業

4.1. 作業方針

EDA 等で得た情報から今回のコンペは以下の方針で進めて行きました。

- 明らかに間違っているデータの除去
- Pseudo-Label(疑似ラベル)を使用
- Cross Validation は 5fold 程度
- Data Augmentation(DA)はなんでもあり
- ラベル不均衡はダウンサンプリングで対応

大体は前回のコンペの上位のソリューションを参考にしています。

4.2. ノイズデータの除去

正しい学習データで正しくラベリングされているデータの精度を上げようという試みです。EDA を行った結果で筆者の目では病気の種類を分類することは出来ないと言いましたが、明らかに画像データとして間違っているものを除いて行きます。今回のデータはクラウドソーシングで現地農家の人に写真を撮ってもらって収集しマケレ大学の AI 研究室と共同でアノテーションしているそうなのですが、下図のように明らかに学習向きではない画像が混じっています。

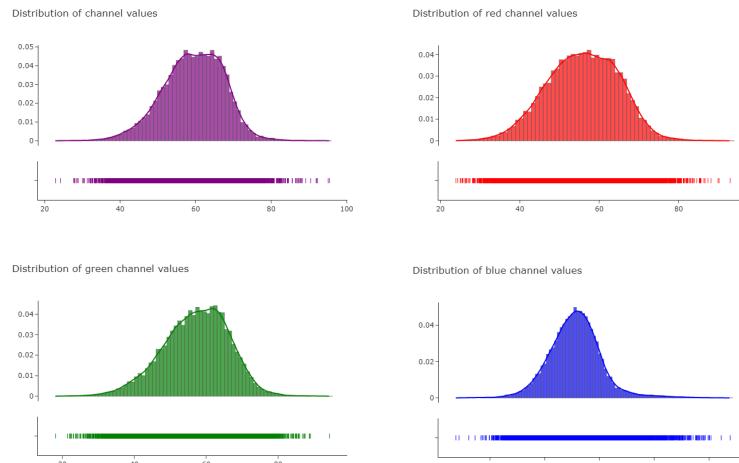
- データが壊れている



全体を見た結果、筆者の目では分類できないと判断しました。またアスペクト比を無視してリサイズされている画像もあったのでこれも厄介な問題です。

3.3. 色情報に傾向はあるのか

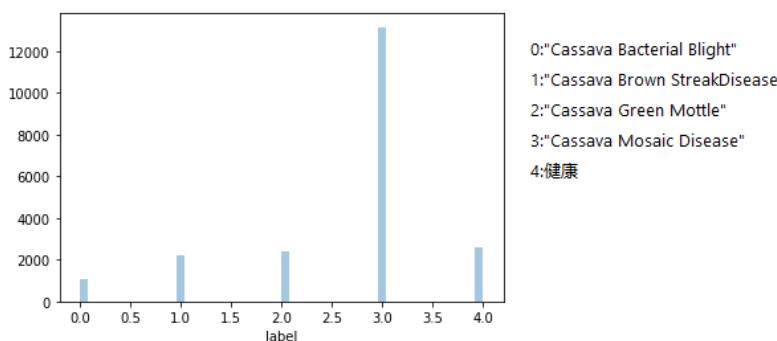
データセットの画像をラベルごとに 1000 枚ずつ抜き出して RGB の分布をグラフにしました。色全体や RGB ごとの傾向からなにか取れないかと思いましたが、中心から広がる形で極端に外れた形はないので色情報から分類していくのは難しそうでした。



3.4. ラベルの分布

データセットのラベルごとの総数をプロットしました。

4 つのラベルが 2000 枚前後なのに対し、ラベル 3 の Cassava Mosaic Disease(モザイク病)は突出して 12000 以上ありました。データセットとしてはあまりにもサンプル数に偏りがあるのでこれについて学習時に考慮する必要があります。



3. 準備

3.1. まずはどんな画像があるのかデータを見てみる

とりあえず、データセットを見てみましょう。ラベル別に 1 種類ずつ表示すると下図になります。



ぱっと見なにが病気なのかはわからないですね。ちなみに左から

- 健康
- Cassava Bacterial Blight(CBB) キャッサバ白葉枯病
- Cassava Brown Streak Disease (CBSD) キャッサバ褐色条斑病
- Cassava Green Mottle (CGM) キャッサバ緑斑紋ウイルス
- Cassava Mosaic Disease (CMD) モザイク病

違いは分かりますか？筆者は未だに分かっていません。

3.2. 病気の種類別に見る

1 種類ずつ見て比較してもよくわからなかったので種類ごとの特徴は無いか探します。

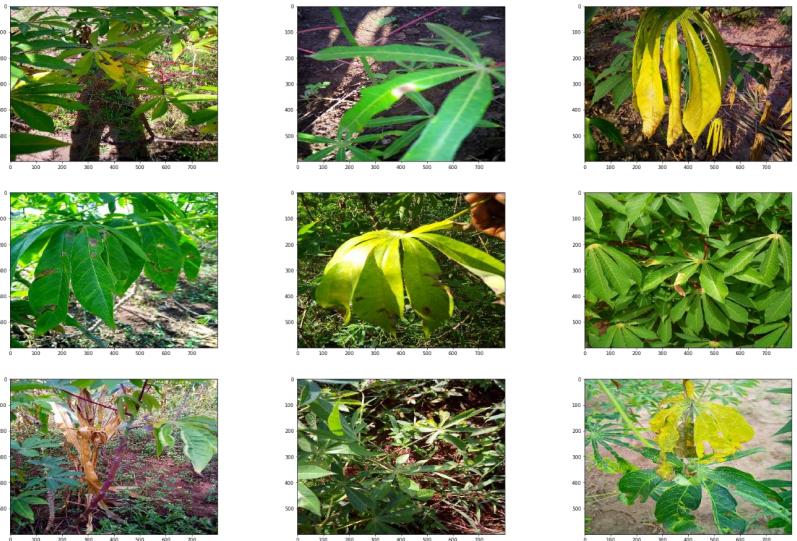
● 健康

ラベルは健康って書いてあるのに葉が枯れたり、萎れたりしている。健康なのでここ固有の特徴はないですね。



● Cassava Bacterial Blight(CBB) キャッサバ白葉枯病

正直健康と違いが目ではわからない。葉の色に黄色みがかったのが多いかな程度です。



● Cassava Brown Streak Disease (CBSD) キャッサバ褐色条斑病

この病気は芋でも判別できるらしいので葉ではなく芋単体の画像が紛れ始めました。



● Cassava Green Mottle (CGM) キャッサバ緑斑紋ウイルス

言われてみれば薄っすら斑紋があるようには見えますが、斑紋の無い葉もあります。



● Cassava Mosaic Disease (CMD) モザイク病

病名の通り、葉に黄緑のモザイク柄があるのでこの中だと分かりやすいです。

