Kaggle 奮闘記

石川 達也

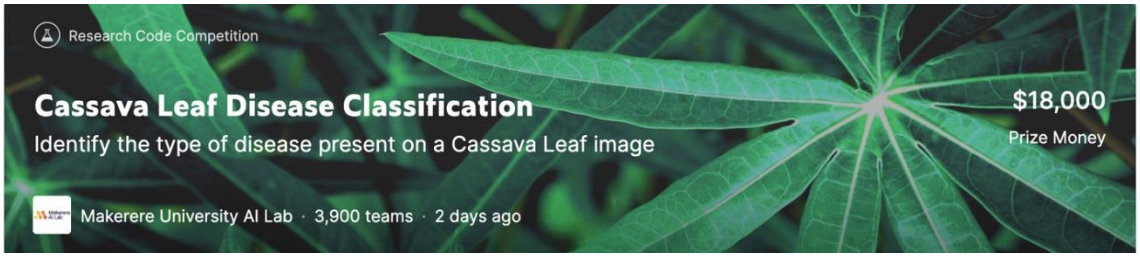
1. はじめに

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 というものは無縁でした。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っ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

ことの始まりは ZENKEI チームが以前(筆者の入社前)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. 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ていたり、萎れていたりしている。健康なのでここ固有の特徴はないですね。

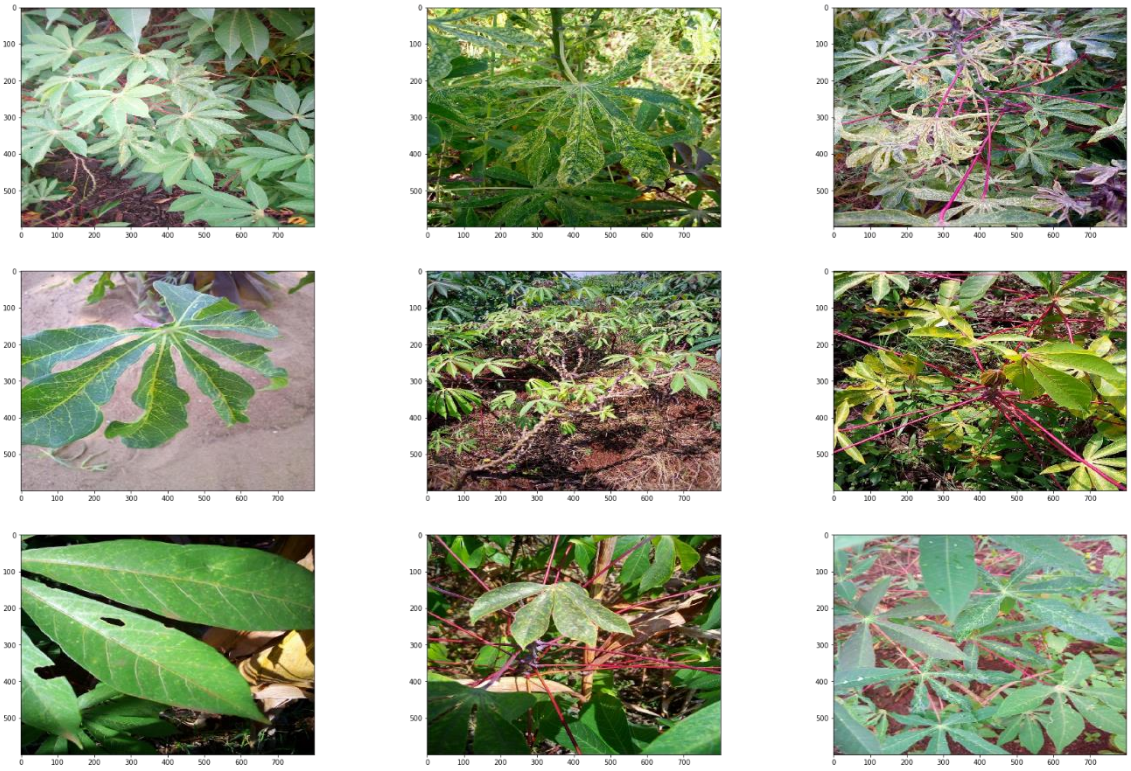
- **Cassava Bacterial Blight(CBB) キャッサバ白葉枯病**

  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。

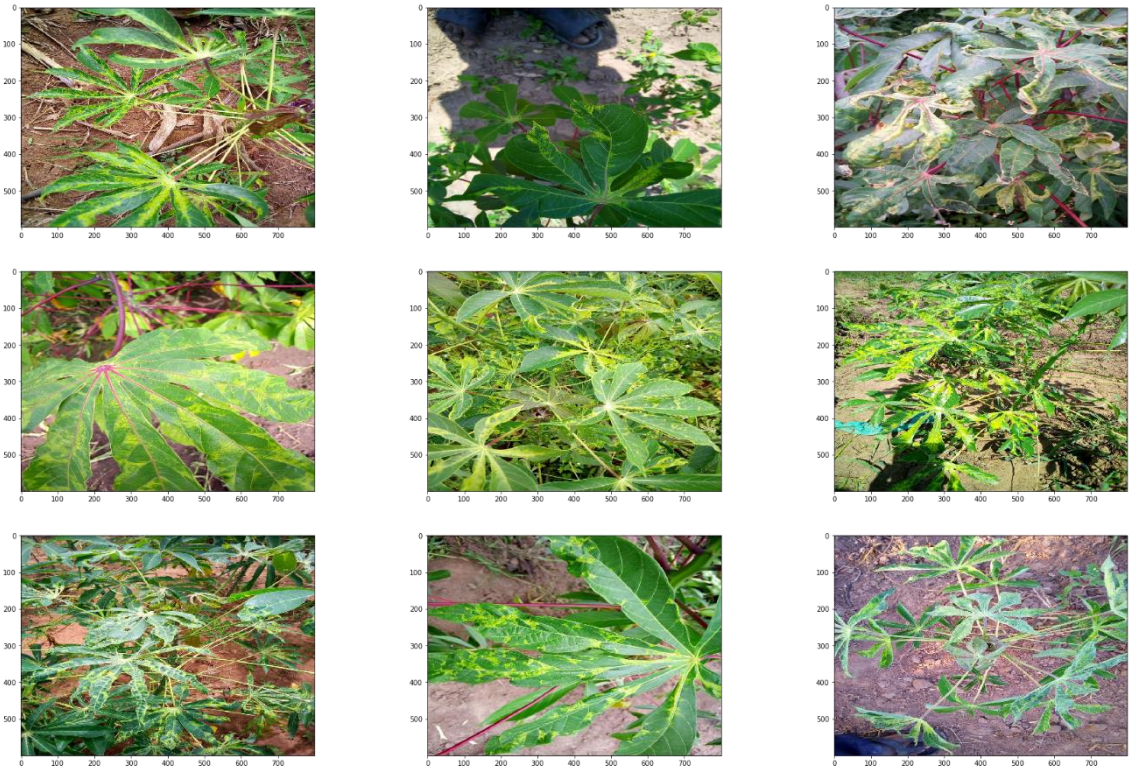
- **Cassava Brown Streak Disease (CBSD) キャッサバ褐色条斑病**

  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。



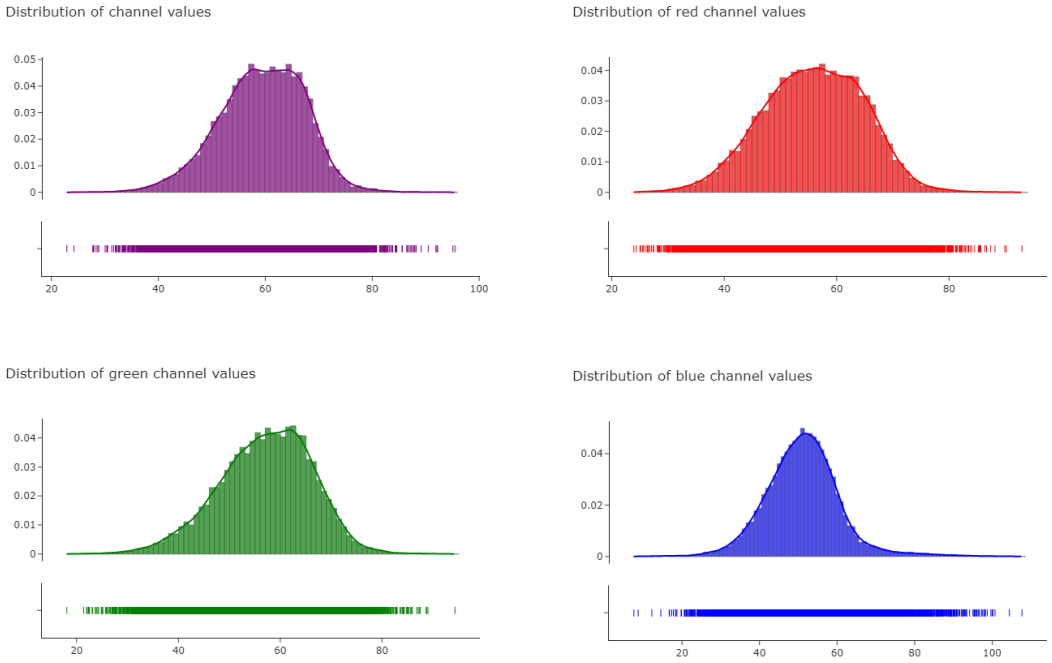
- **Cassava Mosaic Disease (CMD) モザイク病**
  病名の通り、葉に黄緑のモザイク柄があるのでこの中だと分かりやすいです。



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。

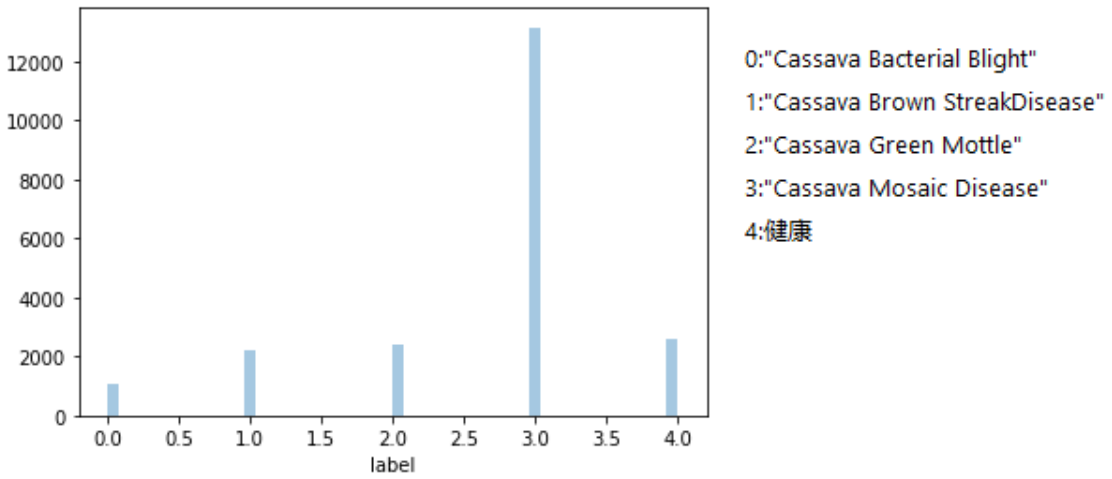
3.3. 色情報に傾向はあるのか

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に持っ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。



3.5. 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 を信用するなと強く叫ばれていました。

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。

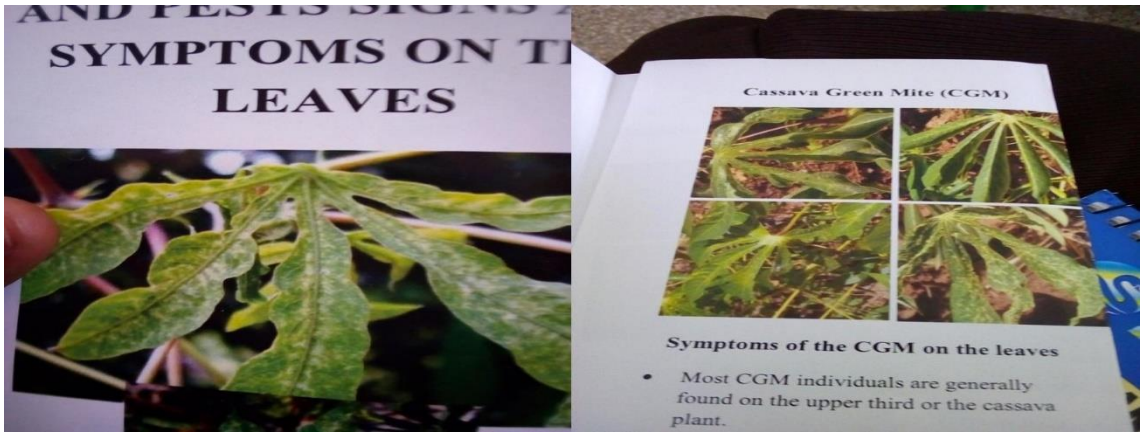
4.2. ノイズデータの除去

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

- 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な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

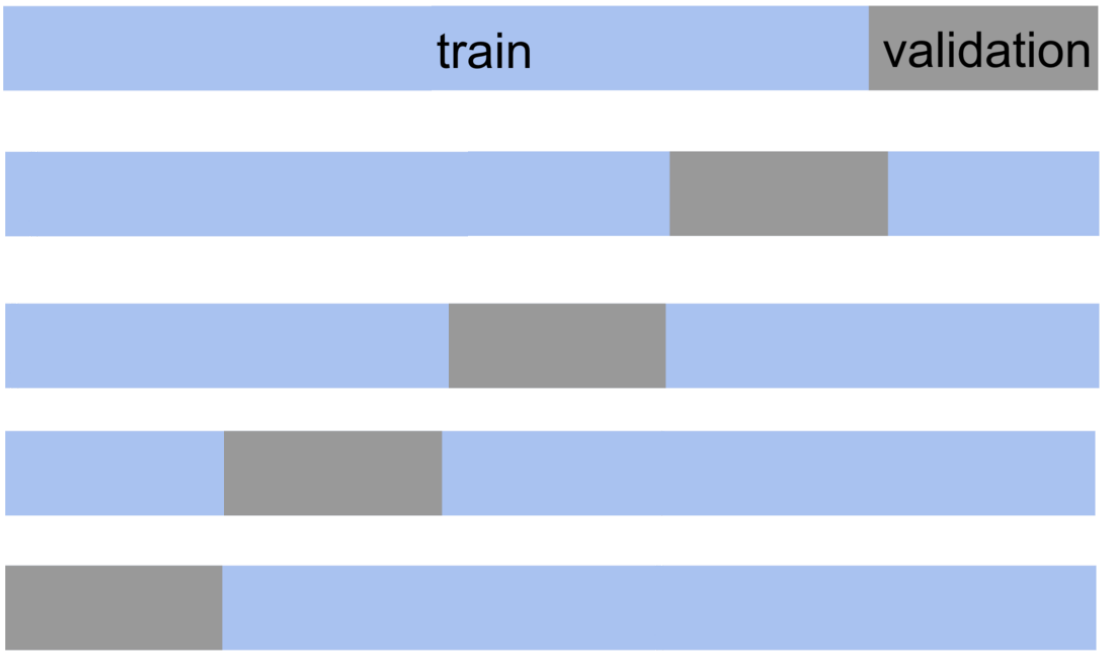
① ラベル付けされたデータでモデルを学習させる(これを教師モデルという)
② ①で作成した教師モデルを使用してラベル付されていないデータに予測値をラベルとしてつける(これを擬似ラベルという)。今回の場合、全てのデータにラベルがついているが 1/3 のデータで教師モデルを学習し 2/3 のデータのラベルを教師モデルの予測値に置き換えた。
③ ②の疑似ラベル付きのデータとラベル付きデータを混ぜてモデルを学習させる(これを生徒モデルという)
④ 上記を繰り返し推論させ疑似ラベルを更新していく

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のですが、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。

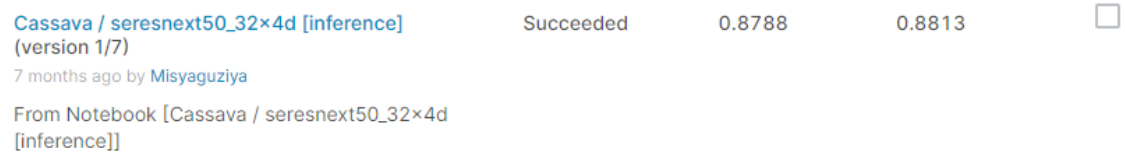


5 つのモデルの精度の平均でモデルの精度を評価して行きます。

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 が投稿されていたのでこれを流用して始めました。
(https://www.kaggle.com/yasufuminakama/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。



(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。

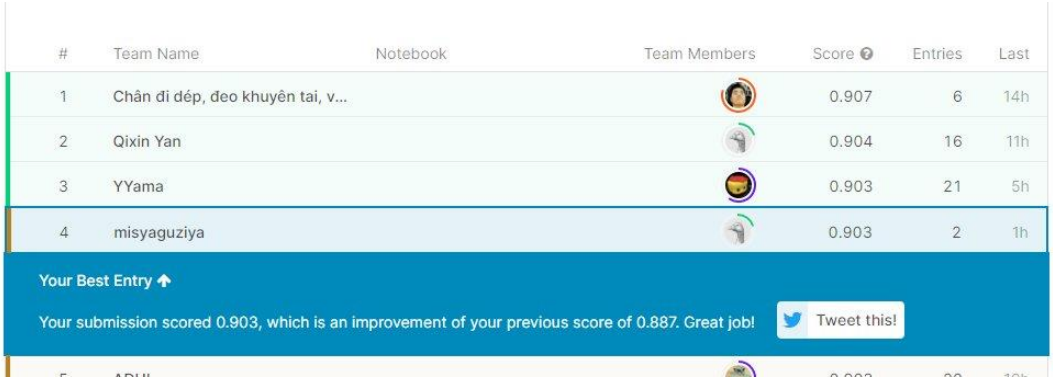
4.6. 学習の繰り返し

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



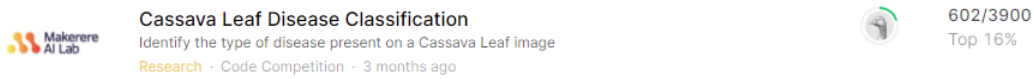
Accuracy:0.903 で暫定 4 位になっていました!(misyaguziya というのが筆者)
この感じなら上位に行けると気を良くして続けて頑張っていこうと思っていたら…

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



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と思います。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？