CDLE HACKATHON2020

予測性能部門

「画像データに基づく気象予測」

氏名：鈴木淳哉

CDLE

自己紹介

- 氏名：　鈴木　淳哉
- G検定：JDLA Deep Learning for GENERAL 2019#1　取得
- E資格：JDLA Deep Learning for ENGINEER 2019#2　取得
- 現在、本業の傍ら夜間・週末を利用して、SIGNATE、kaggleなど各種データサイエンスプラットフォームにて、機械学習中の修行中。

採用した方針・手法について

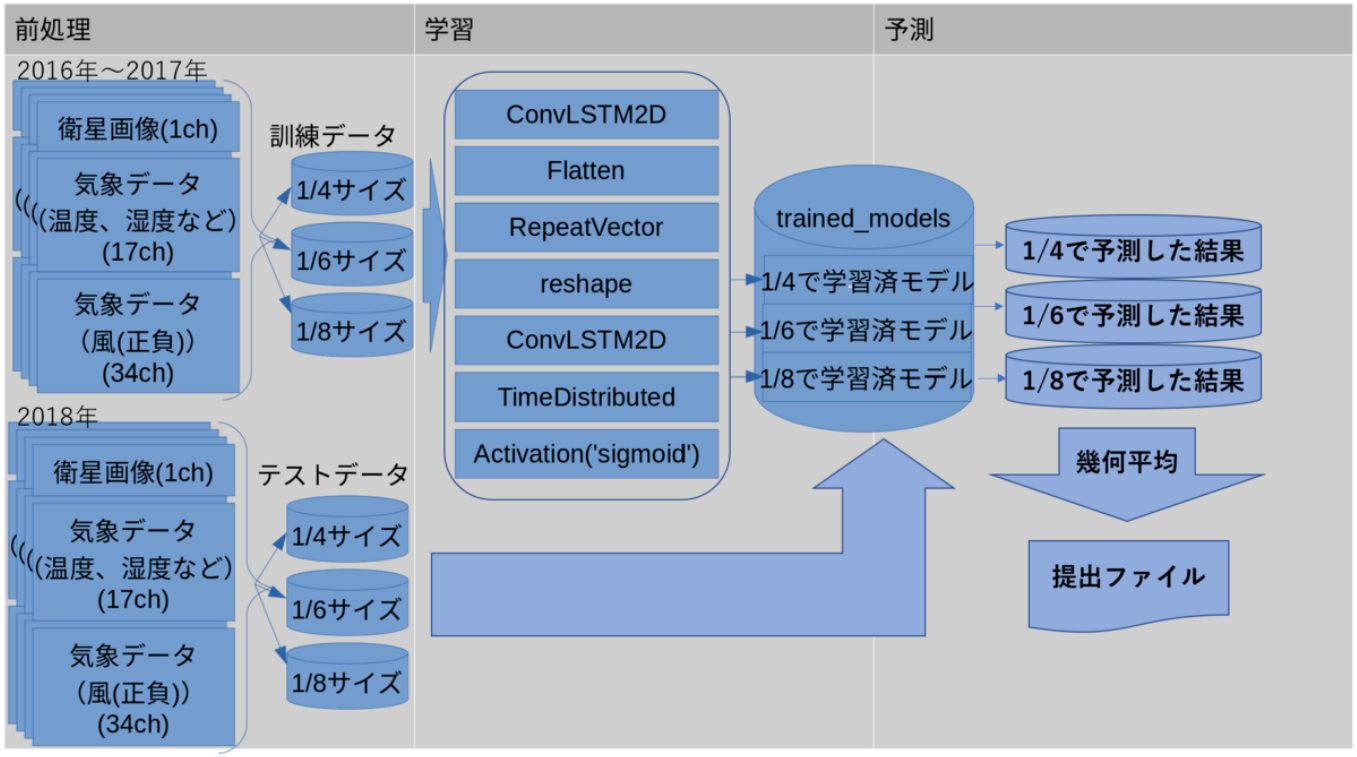
- 昨年開催時のフォーラムページと入賞者レポートなどを参考にさせていただきながら、以下の方針・手法を採用いたしました。
- 特に工夫した点は、損失関数にSSIMを採用した点と、アンサンブルに幾何平均を用いた点です。

Deep Learningフレームワーク	Keras(TensorFlow)
アルゴリズム	Convolution LSTM
損失関数	1-SSIM(Structural Similarity)
最適化手法	Adam
各データの正規化手法	最小0、最大1に正規化
入力の縮小サイズ	1/4,1/6,1/8
アンサンブル方法	幾何平均

CDLE

全体構成図

- 全体構成としては、以下の通り、前処理で気象データもchとして追加し、複数サイズで学習させたあと、予測結果を幾何平均でアンサンブルしました。



CDLE

前処理（衛星画像データ）

- 衛星画像データの前処理は、以下のような欠測補完と縮小処理を実施いたしました。

分類	処理内容
欠測補完	衛星画像の欠測には２パターンありました。１つは、特定の時間帯の衛星画像がないパターンで、もう１つは、衛星画像はあるけれども、画像の一部がないパターンでした。 これらの欠測補完は、単純に前後の画像の値の平均値で補完しました。
縮小	画像サイズを１/4、1/6、1/8に縮小して、24時間分を１日分のデータとして保存しました。 1/10、1/12、1/15の縮小サイズも作成しましたが、アンサンブルした際にMAEは向上したものの、SSIMが下がってしまったので、採用しませんでした。

CDLE

前処理（気象データ）

- 気象データの前処理は、以下の通り区分ごとに正規化と画像データのchに追加するための分割・補完処理を実施いたしました。

分類	処理内容
気温 湿度 海面気温	区分ごとの最大値・最小値を取得して、最小値０、最大値１となるように正規化。 気象データは、３時間ごとのデータしかなかったので、前後のデータから線形に変化した前提で、間の時間のデータを補完しました。
東西風 南北風 鉛直風	区分ごとの最大値・最小値を取得して、最小値０、最大値１となるように正規化。 気温とは異なり、正負があるので、正と負に分けて保存。気象データは、３時間ごとのデータしかなかったので、前後のデータから線形に変化した前提で、間の時間のデータを補完しました。

CDLE

モデル構築

- モデルには、Convolution LSTM層を含むニューラルネットワークを採用して、損失関数をMAE、1-SSIMの比較を行った結果、Public Scoreが向上した1-SSIMを採用しました。

サイズ	損失関数	CV	Public Score	Private Score
1/4	MAE	(0.9062)	0.5565107	0.5332547
1/4	1－SSIM	0.5065	**0.5665850**	0.5436109

損失関数に1-SSIM を使用することで、Scoreが向上

CDLE

アンサンブル（入力縮小サイズ）

- アンサンブルについて、最適な【入力縮小サイズの組み合わせ】の検証と、最適な【アンサンブル手法】の検証を行いました。
- 入力縮小サイズについては以下組み合わせが最適となりました。

入力	1/4	1/6	1/8	1/10	1/12	1/15	Public Score
1個	○	-	-	-	-	-	0.5665
2個	○	○	-	-	-	-	0.5683
3個	○	○	○	-	-	-	**0.5688**
4個	○	○	○	○	-	-	0.5682
5個	○	○	○	○	-	○	0.5668
6個	○	○	○	○	○	○	0.5676

各入力数毎に最もPublic Scoreが良かった組み合わせを抽出し、さらに、その中で最もPublic Scoreが良かった個数を採用。

CDLE

アンサンブル（アンサンブル手法）

- アンサンブル手法については、算術平均、幾何平均、加重平均を比較しました。
- 加重平均は訓練時に検証用データの予測値を使って最適な荷重の探索を行いました。
- 検証の結果、Public Scoreが高かった幾何平均を採用しました。

方法	Public Score
算術平均	0.5688
幾何平均	0.5691
加重平均	0.5665

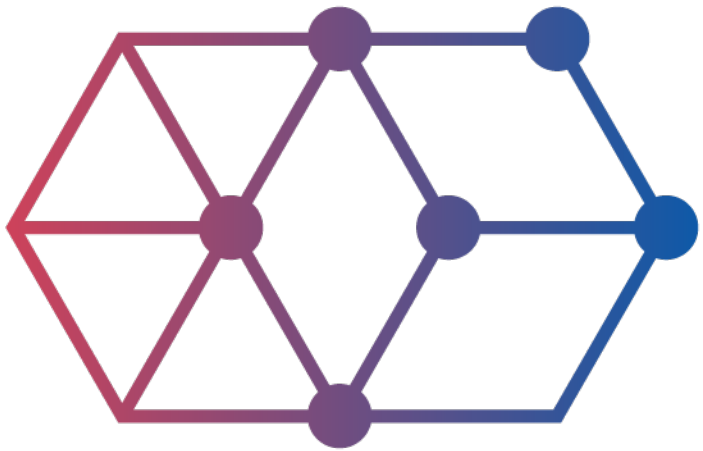
CDLE

分析のポイントについて

- **分析時に苦労したこと**
  - とにかく実行に時間がかかる点が苦労しました。
  - 前処理・学習・アンサンブルの各段階での中間出力を細かく取得し、複数の組み合わせが実行できるよう工夫しました。
- **試したが上手く行かなかったこと**
  - ２段目をLightGBMとするスタッキングに挑戦しましたが、うまく精度を出すモデルを構築できませんでした。
- **特に予測精度の向上に寄与したポイント**
  - 損失関数にSSIMを採用したことが精度向上に寄与しました。

感想・まとめ

- 今年の夏は、なかなか外出することもできず、自宅で過ごす時間が多く、CDLE HACKATHON2020のような学習機会に恵まれたことは、非常に有意義でした。
- 特に、いままで経験してこなかった課題に取り組めたことや、新たな知識を得られたことで、今後に向けた意欲が向上できたことは、今後の私の人生に大きな一歩となりました。
- G検定、E資格を取得した大きなメリットは、資格自体の評価というよりも、こういった学習機会や同じ志をもつ皆さんと交流できる機会を得られることだと改めて感じました。
- このような機会をご提供いただき誠にありがとうございました。



Community of
Deep Learning Evangelists

CDLE