NIIGATA UNIVERSITY

新潟大学の紹介 / 太陽データの機械学習実習

新潟大学
大学院自然科学研究科 情報工学コース
/ 工学部 知能情報システムプログラム

飯田 佑輔

March 17th, 2022

太陽研究最前線ツアー

自己紹介

NIIGATA UNIVERSITY

名前：飯田 佑輔 (イイダ ユウスケ)

研究分野：
太陽物理学，**天文情報学**

研究キーワード：
磁気対流，**画像認識**，**ビッグデータ**

職歴：

1985	三重県 鈴鹿市 で生まれる
2003 - 2007	東京大学 理科I類 / 地球惑星物理学科
2007 - 2012	東京大学 大学院理学系研究科 地球惑星科学専攻
2012 - 2013	東京大学 JSPS特別研究員(PD)
2013 - 2016	JAXA/ISAS　プロジェクト研究員
2016 - 2019	関西学院大学 助手
2019 -	新潟大学　　准教授

新潟大学 大学院情報工学コース / 知能情報システムプログラム

NIIGATA UNIVERSITY

情報工学コース分野

- コンピュータサイエンス
- 情報ネットワーク
- マルチメディア
- 情報セキュリティ

March 17th, 2022

太陽研究最前線ツアー

研究室活動の基本方針

NIIGATA UNIVERSITY

宇宙情報学研究室として，**新しい宇宙科学の知見獲得**と**情報科学技術の発展**を目指しています．

- ✓ **データ科学技術**（深層学習，画像認識，… etc.）
- ✓ **ビッグデータ**（画像・動画，スペクトル，… etc.）
- ✓ **宇宙科学**（太陽活動，銀河，… etc.）

2022年度 研究テーマ

NIIGATA UNIVERSITY

研究生 Jargalmaa Batmunkh, 高次元データ効率圧縮による太陽スペクトルの異常検知手法の開発

大学院（10名）

大沼 伊織，太陽活動領域の成長予測モデル構築
小松 耀人，太陽全球磁気画像からの end-to-end な太陽フレア予測モデル開発
本間 裕也，銀河形態分類モデルの学習データと異なる空間解像度データへの適用

安藤 秀一，機械学習を利用した植物スペクトルからの病理判断
杉原 侑剛，Blind Deconvolutionによる衛星光学系劣化診断
竹部 良　，深層学習を用いた太陽磁場画像でのコロナホール検出
田所 拓馬，オートエンコーダを用いた太陽フレアの新規予測手法の開発
津田 和輝，High-z領域における銀河画像からのphoto-z推定手法の開発
前澤 健一，深層学習を用いた人肌画像からのパラメータ診断手法の開発
渡邊 健斗，GANを用いた黒点スケッチからの磁場データ復元

太陽：9件
銀河：3件
情報：6件

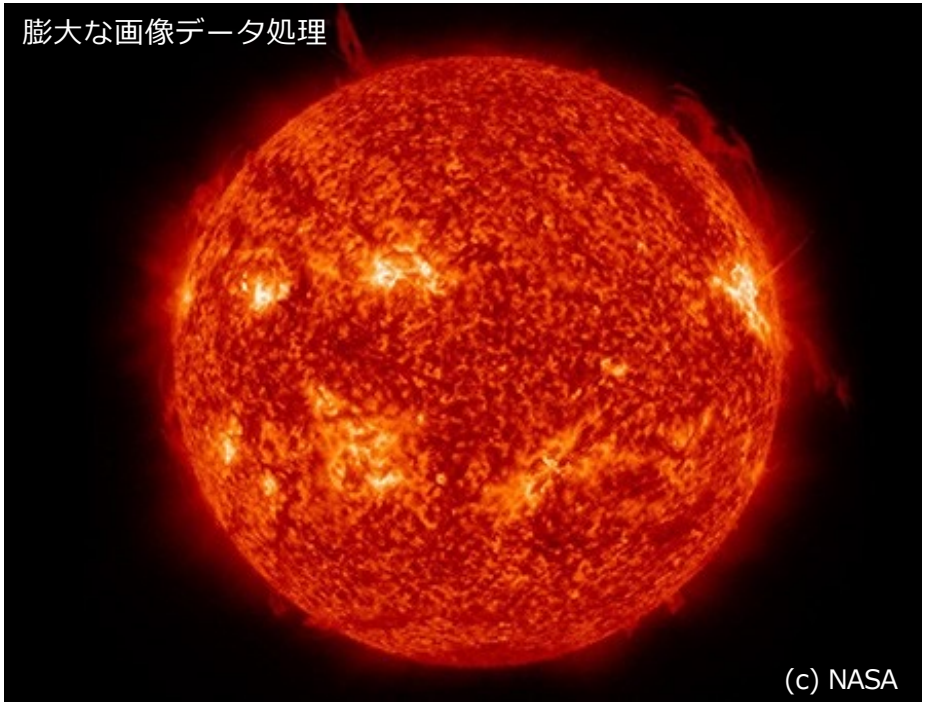
学部生（7名）

佐々木 明良，Mask R-CNNを用いた太陽フィラメント検出の精度向上
佐藤 智哉　，圃場画像からのSPAD値推定手法開発
長谷川 幸大，Giant Cell検出のための太陽対流構造追跡の高速化検討
本多 飛翔　，太陽黒点の出現予測モデル構築への挑戦
町田 瑞樹　，教師なし学習による銀河画像の特徴抽出手法の開発
村田 実広　，教師ラベルエラーに対するMLPとCNNの挙動差に関する研究
横山 光輝　，太陽活動領域の画像予測手法の開発

他に…
関学情報M1　太陽シグモイド自動検出
新潟大環境M1　短命氷河湖の自動検出

「宇宙 x 情報」の現在

近年，宇宙科学と情報科学の融合が急速に進む．**天文情報学(AstroInformatics)**という呼称も使われるように．



「宇宙 x 情報」の現在

NIIGATA UNIVERSITY

2010年代から天文情報学関連の学会・研究会が開催.

IEEE Bigdata 2019

Challenges with Extreme Class-Imbalance and Temporal Coherence: A Study on Solar Flare Data

Azim Ahmadzadeh
dept. Computer Science
Georgia State University
Atlanta, GA, USA
aahmadzadeh1@cs.gsu.edu

Maxwell Hostetter
dept. Computer Science
Georgia State University
Atlanta, GA, USA
mhostetter1@cs.gsu.edu

Berkay Aydin
dept. Computer Science
Georgia State University
Atlanta, GA, USA
baydin2@cs.gsu.edu

Manolis K. Georgoulis
RCAAM of the Academy of Athens,
Athens, Greece
manolis.georgoulis@phy-astr.gsu.edu

Dustin J. Kempton
dept. Computer Science
Georgia State University
Atlanta, GA, USA
dkempton1@cs.gsu.edu

Sushant S. Mahajan
dept. Physics & Astronomy
Georgia State University
Atlanta, GA, USA
mahajan@astro.gsu.edu

Rafal A. Angryk
dept. Computer Science
Georgia State University
Atlanta, GA, USA
angryk@cs.gsu.edu

***Abstract*—In analyses of rare-events, regardless of the domain of application, class-imbalance issue is intrinsic. Although the challenges are known to data experts, their explicit impact on the analytic and the decisions made based on the findings are often overlooked. This is in particular prevalent in interdisciplinary research where the theoretical aspects are sometimes overshadowed by the challenges of the application. To show-case these undesirable impacts, we conduct a series of experiments on a recently created benchmark data, named Space Weather ANalytics for Solar Flares (SWAN-SF). This is a multivariate time series dataset of magnetic parameters of active regions. As a remedy for the imbalance issue, we study the impact of data manipulation (undersampling and oversampling) and model manipulation (using class weights). Furthermore, we bring to focus the auto-correlation of time series that is inherited from the use of sliding window for monitoring flares' history. Temporal coherence, as we call this phenomenon, invalidates the randomness assumption, thus impacting all sampling practices including different cross-validation techniques. We illustrate how failing to notice this concept could give an artificial boost in the forecast performance and result in misleading findings. Throughout this study we utilized Support Vector Machine as a classifier, and True Skill Statistics as a verification metric for comparison of experiments. We conclude our work by specifying the correct practice in each case, and we hope that this study could benefit researchers in other domains where time series of rare events of are interest.**

***Index Terms*—class imbalance, sampling, time series, flare forecast**

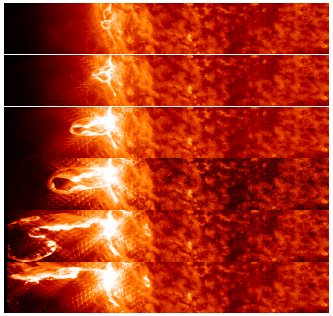


Fig. 1. Snapshots of an X-class flare, peaking at 7:49 p.m. EST on Feb. 24, 2014, observed by NASA's Solar Dynamics Observatory, in the 304Å wavelength channel. (Images source: https://helioviewer.org/)

I. Introduction

To gain valuable insights or robust predictive performance from data, we must first ensure the integrity of our data. Beyond data collection, this involves data-cleaning. It requires a thorough investigation by the experts of the domain and data scientists to produce a reliable dataset. Nonetheless, there are some challenges which are inherited from the subject under study due to unique characteristics of the data which should be identified, understood and dealt with appropriately. Class-imbalance issue is one of the main problems of this kind, which is present in many natural or other nonlinear dynamical systems. This is often due to the nature of the events, not the data collection process.

Class-imbalance is a common problem, with many potential remedies. Some of these remedies are common and well-known, but can still be misapplied. This is particularly true when the primary objective is not machine learning per se but the testing and scrutiny of domain-specific theories. The complexity of the problem at hand and the absence of data experts very often underestimate the needed level of care,

概要

これからの天文学・宇宙物理学分野では，データの量が爆発的に増え，まさにビッグデータとなっていくでしょう．この流れは今後さらに拡大していきます．一方で，計算機の進歩，そして機械学習や統計学といった応用数学の進歩によって，新しいデータ科学の手法が急速に発達しています．今後，良い成果をあげていくためにはデータ科学の方法を積極的に天文学・宇宙物理学分野に導入する必要があるでしょう．本研究会では関連分野からデータ科学を共通項として幅広く話題をあつめます．現在行われている取り組みをお互いに理解し，必要となる問題を洗い出し，今後の天文分野において必要となるデータ科学の手法に関して議論をすすめたいと考えています．本研究会は2015年, 2017年に続き，三回目の開催となります．

詳細

共催：統計数理研究所 統計的機械学習研究センター

JST CREST 「広域撮像探査観測のビッグデータ分析による統計計算宇宙物理学」

日時：2019年5月27, 28, 29日

場所：統計数理研究所 (〒190-8562 東京都立川市 緑町10-3),

大会議室（2F, B201）

参加：参加費無料，こちらから参加の登録，講演の申し込みをお願いします．(登録情報は本研究会の連絡と主催機関への開催報告のみに使用します．)

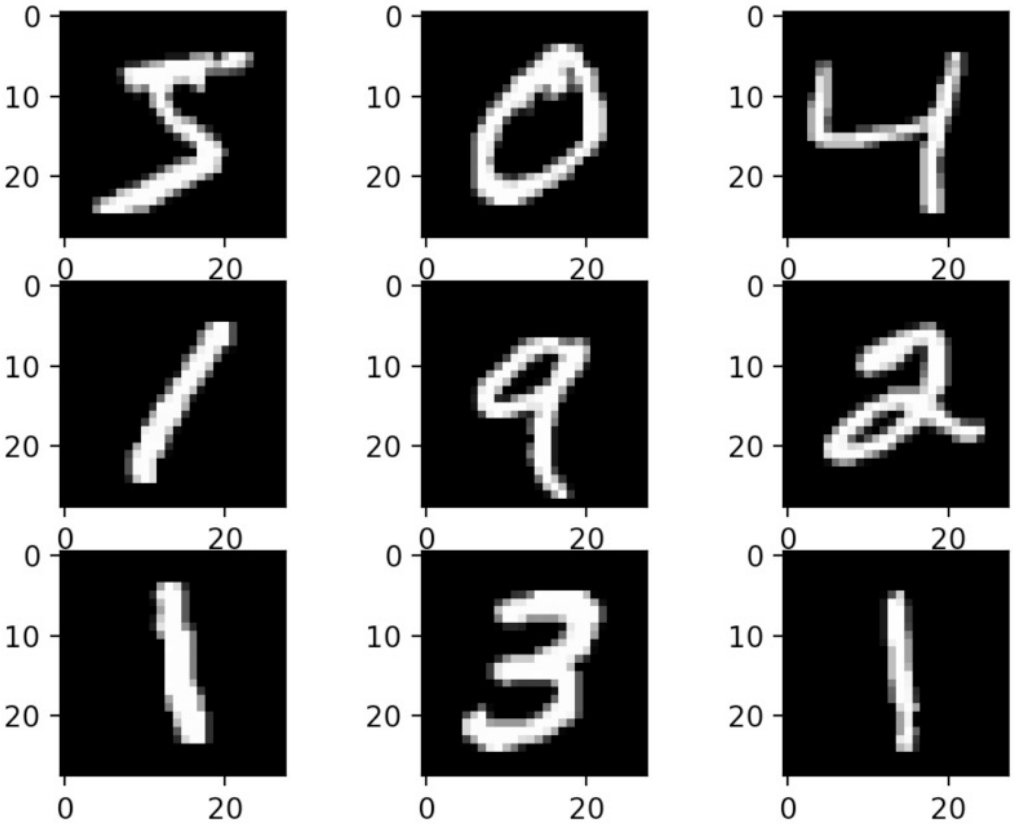
March 17th, 2022

太陽研究最前線ツアー

機械学習

NIIGATA UNIVERSITY

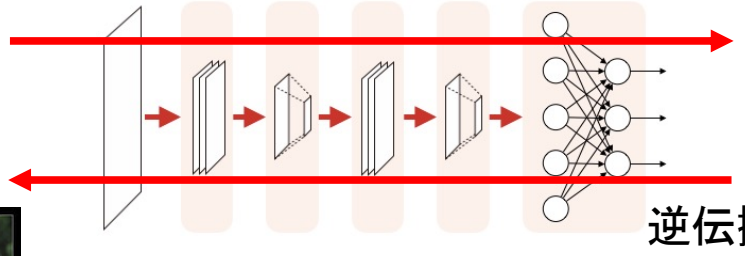
分類や回帰の入出力パターンを学習する，AIのコア技術．特に深層学習は，過パラメータでありつつ汎化性を



機械学習における学習と予測

大量のデータ＝ビッグデータ

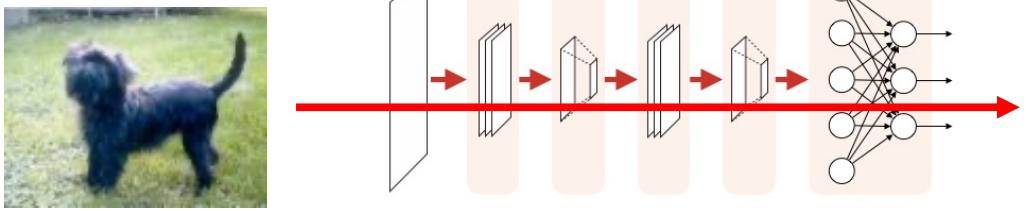
モデルの学習
＝重みと域値の最適化



教師データ

誤差

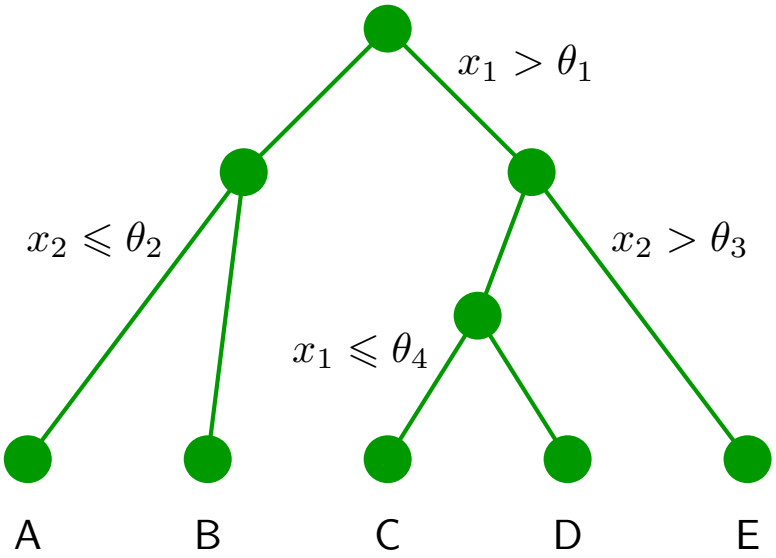
モデルの予測



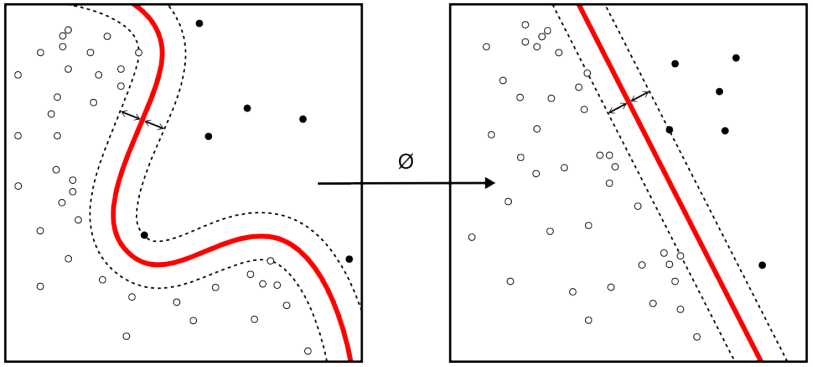
モデルによる
解答

代表的な機械学習手法

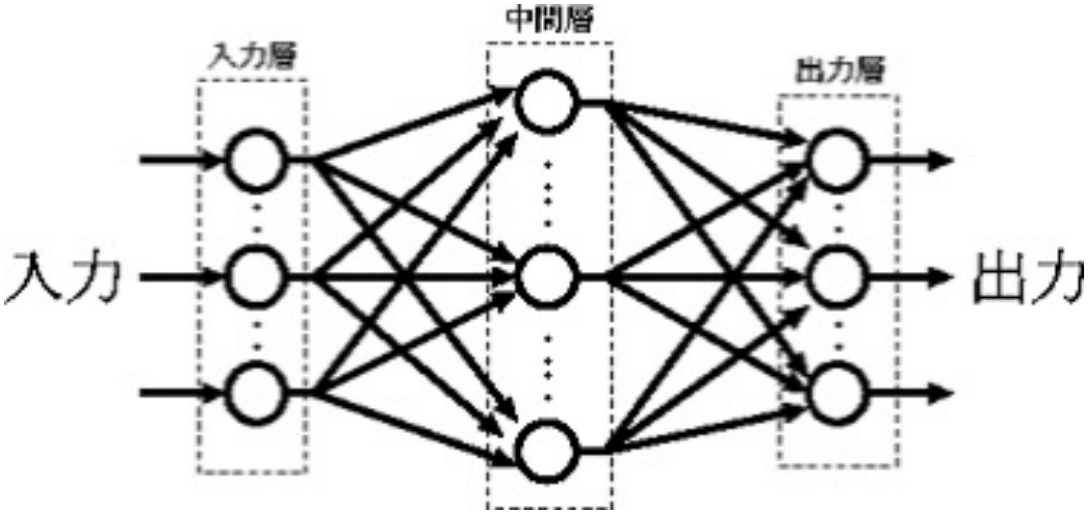
Random forest



Support vector machine

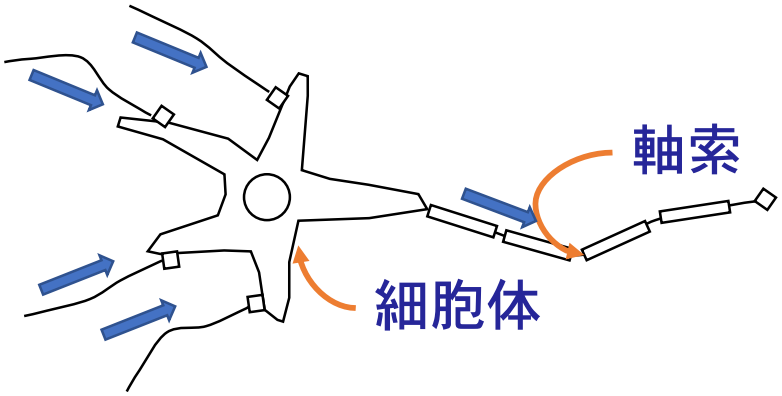


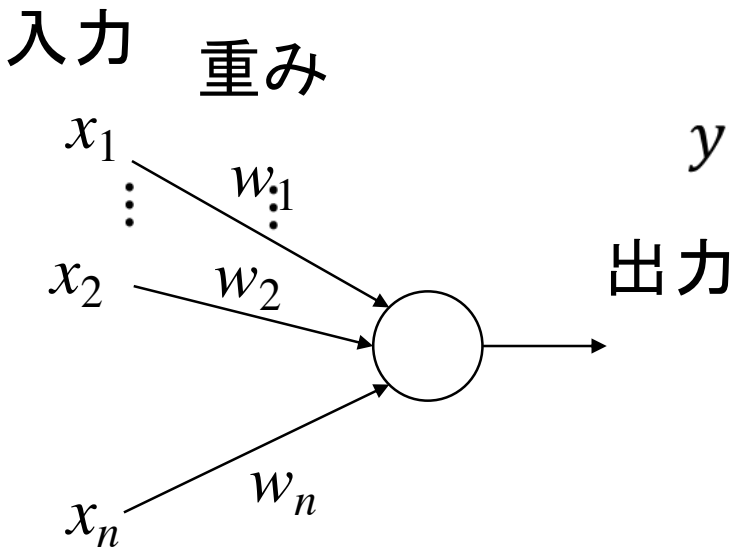
Neural Network



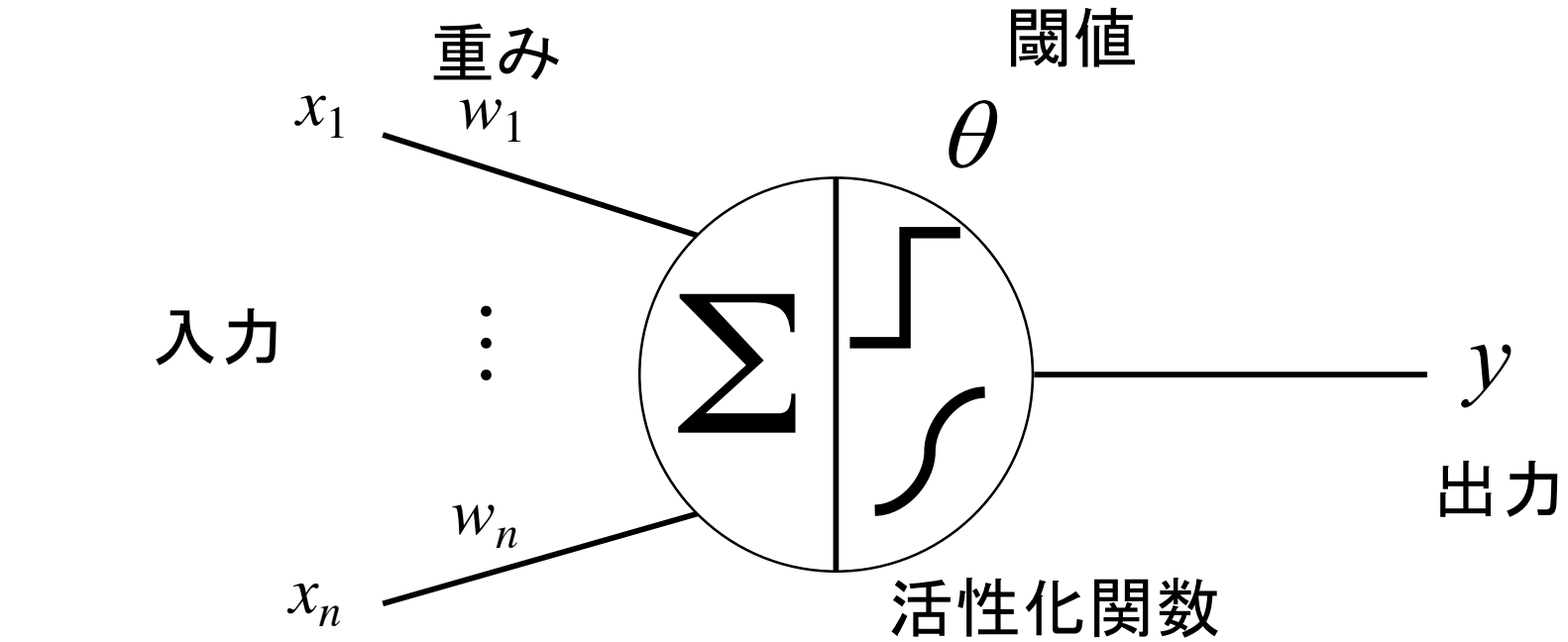
マカロック-ピッツの素子モデル

- 神経細胞を模倣した最初のモデル
- 複数の入力に対して１つの値を出力





神経細胞のモデル化



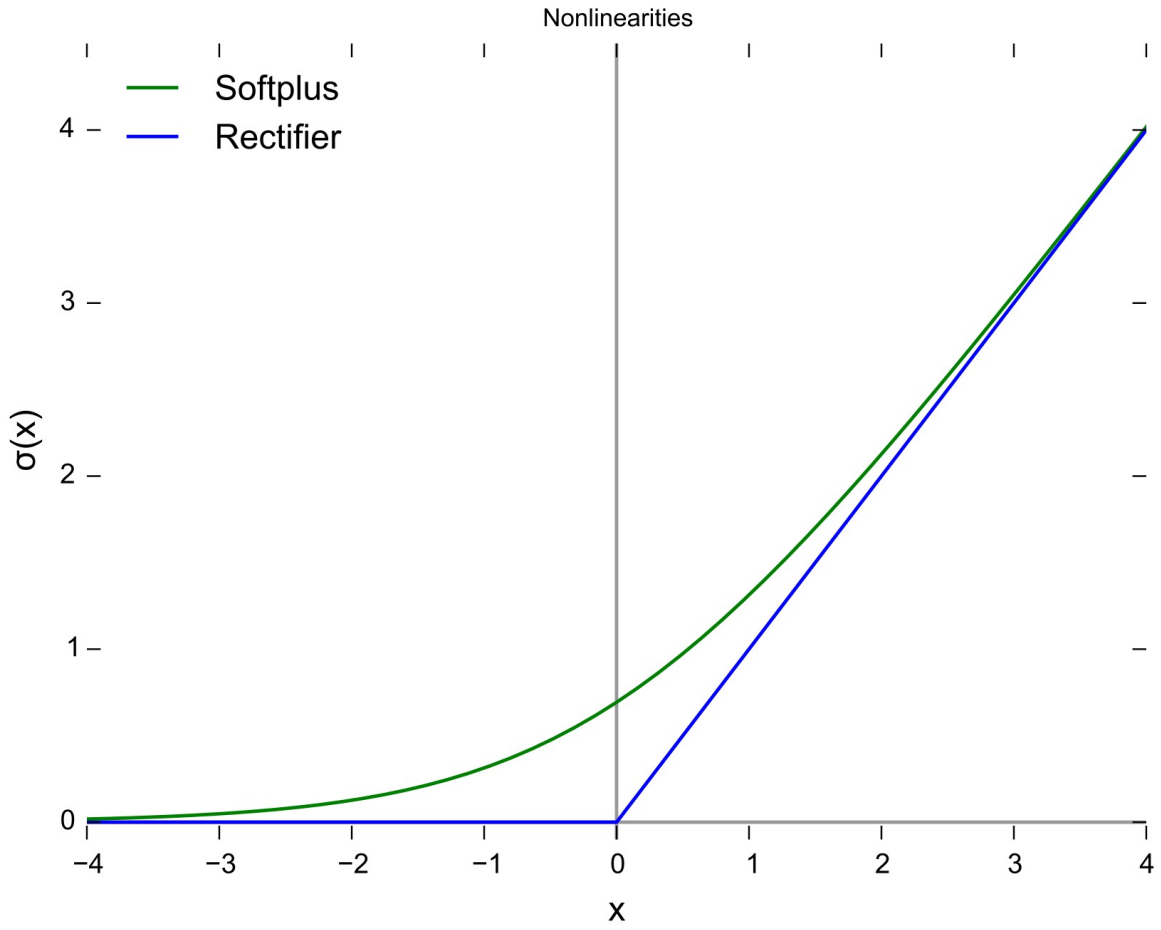
活性化関数

$$y = g\left(\sum_{i=0}^{n} w_i x_i - \theta\right)$$

内部ポテンシャル

活性化関数：ReLU

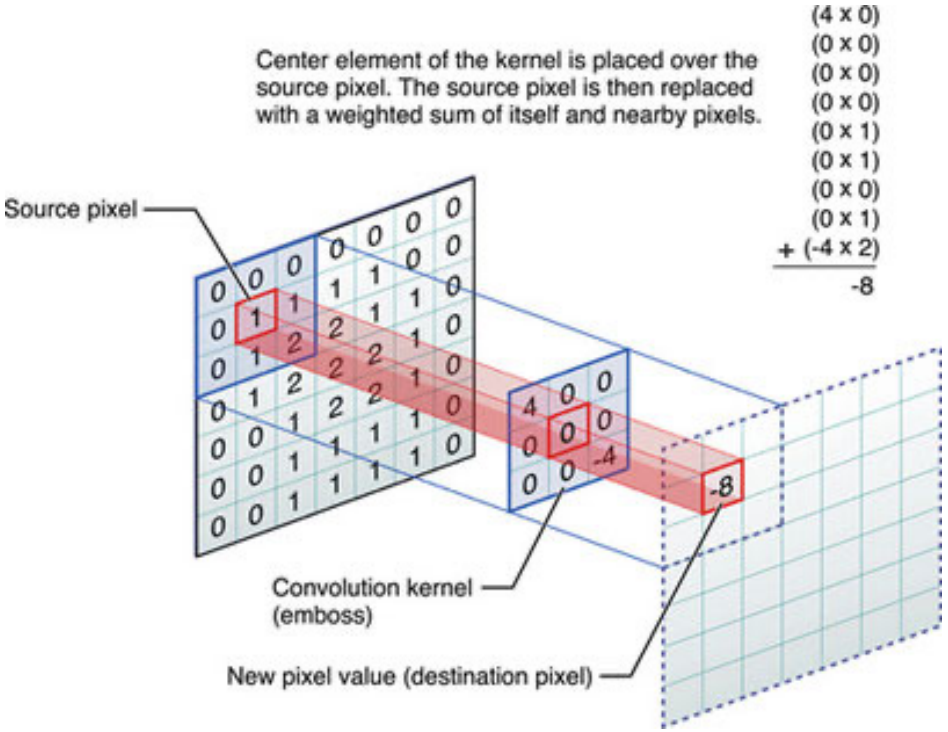
傾きが１という性質から，優れた性質を持つ.



畳み込みニューラルネットワーク

画像データ(空間構造)を学習するにはどうすればよいだろうか？

1. 画像から特徴量を作成し，入力変数とする
2. 特徴量の作成（畳み込み計算）を学習する
   ＝ 畳み込みニューラルネットワーク（CNN）



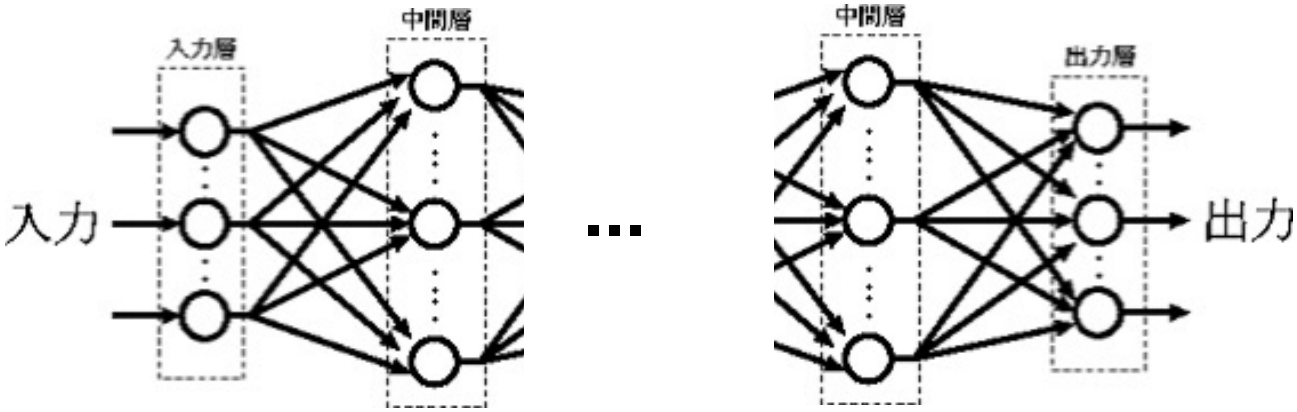
$$O_{i,j} = \Sigma_{m,n} I_{i+m,j+n} K_{m,n}$$

$$K = \begin{bmatrix} K_{1,1} & K_{1,2} & K_{1,3} \\ K_{2,1} & K_{2,2} & K_{2,3} \\ K_{3,1} & K_{3,2} & K_{3,3} \end{bmatrix}$$

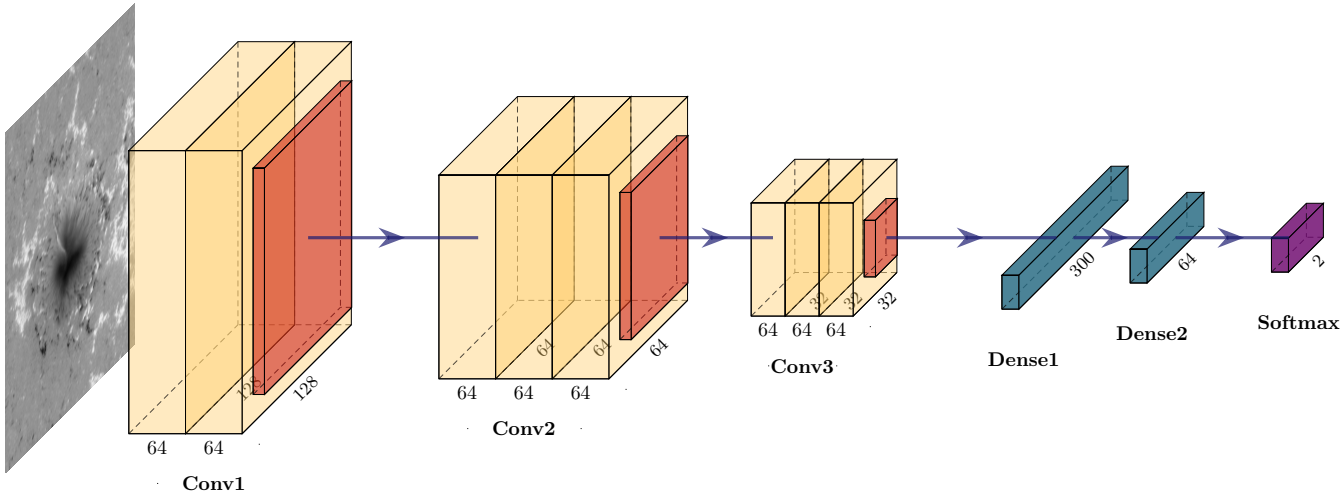
MLPとCNN

NIIGATA UNIVERSITY

Multi Layer Parceptron (MLP)



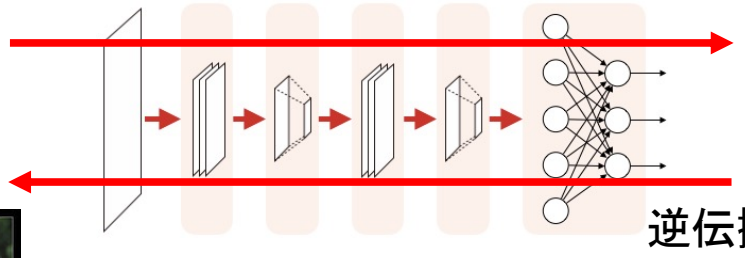
Convolutional Neural Network (CNN)



機械学習における学習と予測

大量のデータ＝ビッグデータ

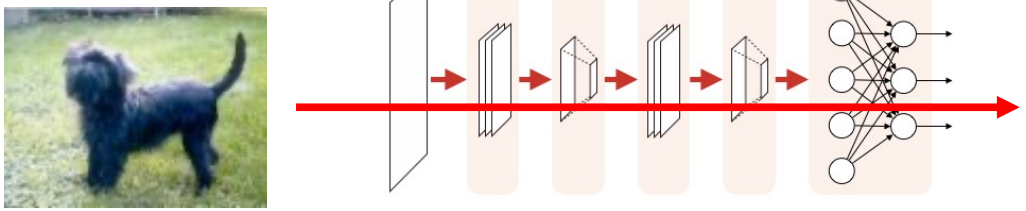
モデルの学習
＝重みと域値の最適化



教師データ

誤差

モデルの予測

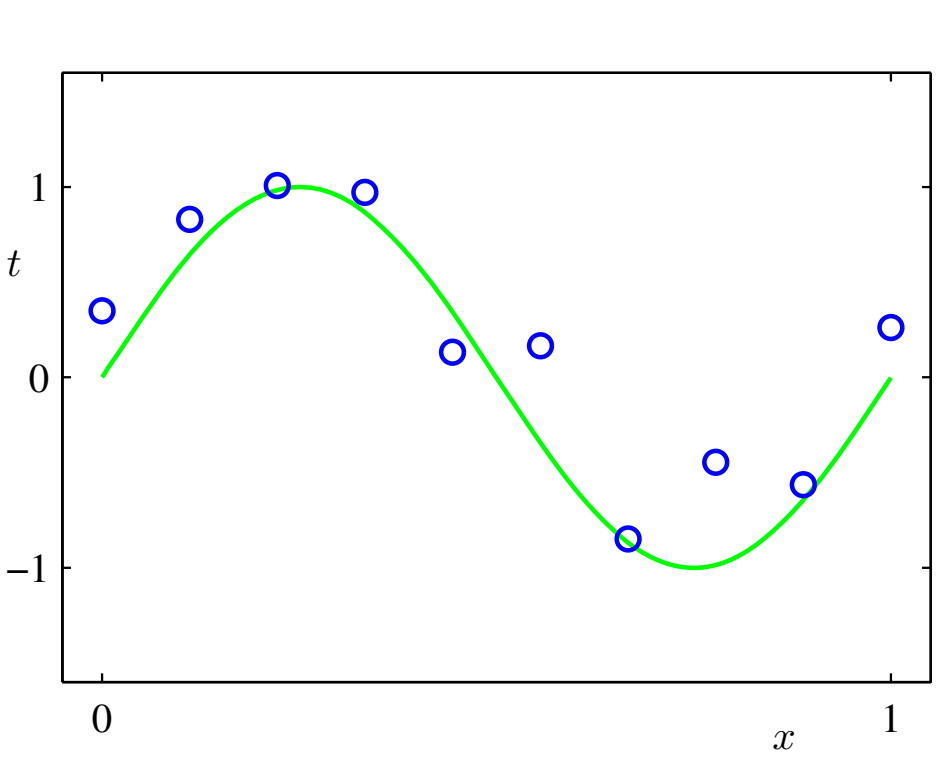


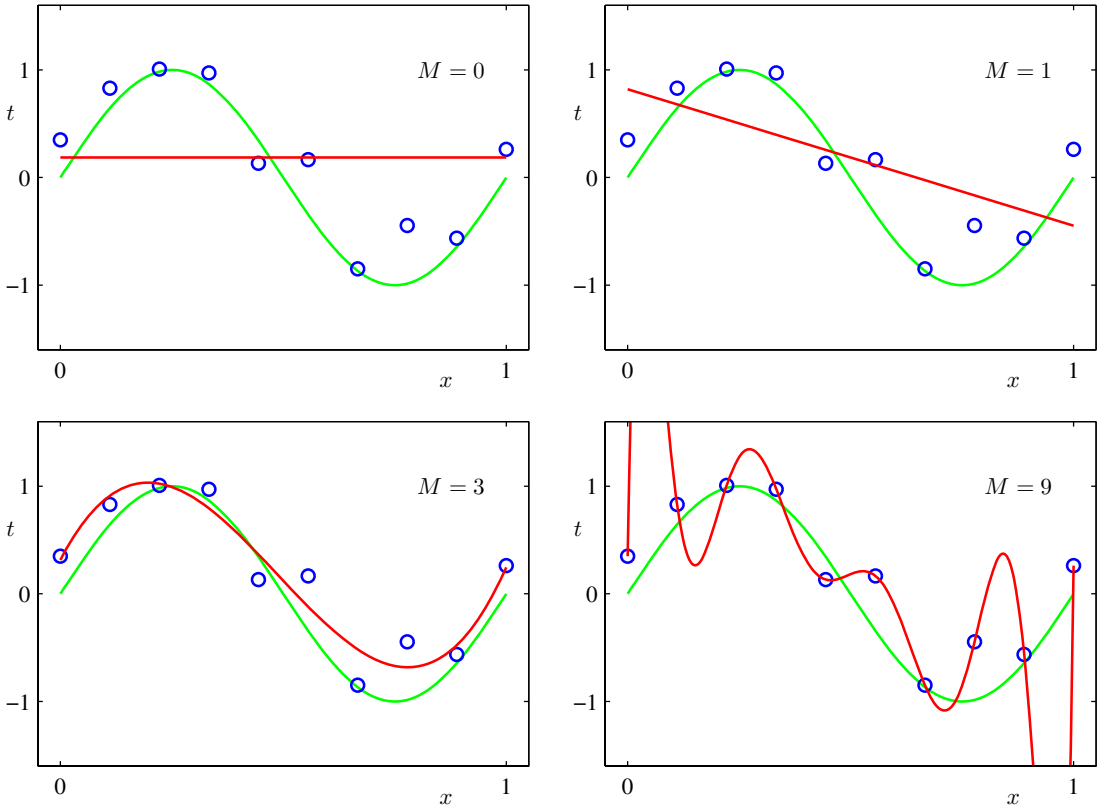
モデルによる
解答

機械学習モデルのパラメータ数の影響

パラメータ数が増えると，表現力は上がるが汎化性能は下がる．適切なパラメータ数を決める必要がある．

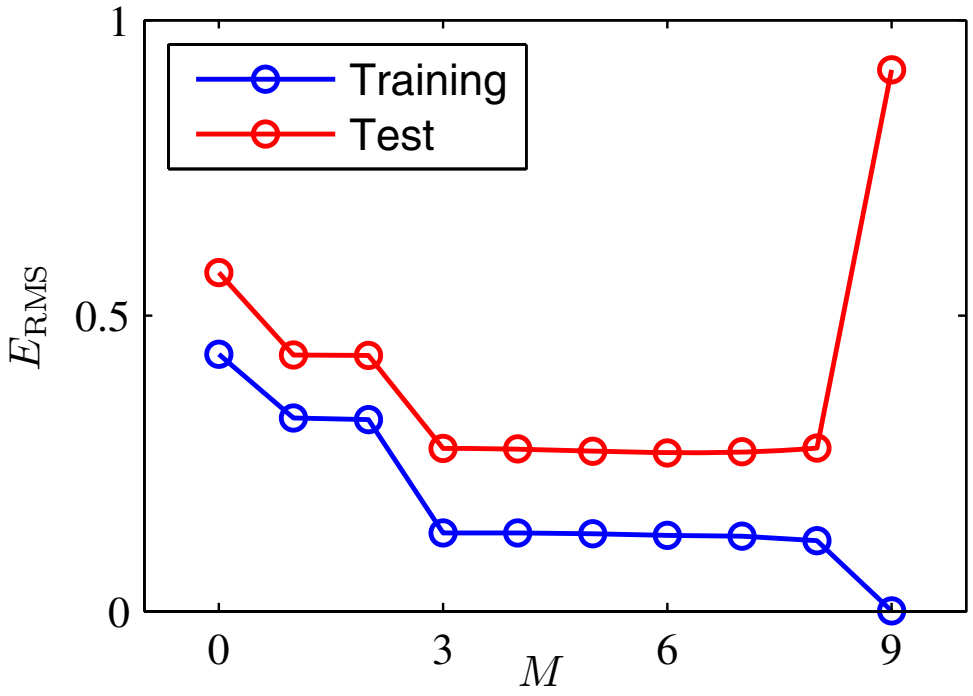
NNは過パラメータであるが，汎化性能が下がらないのかも？





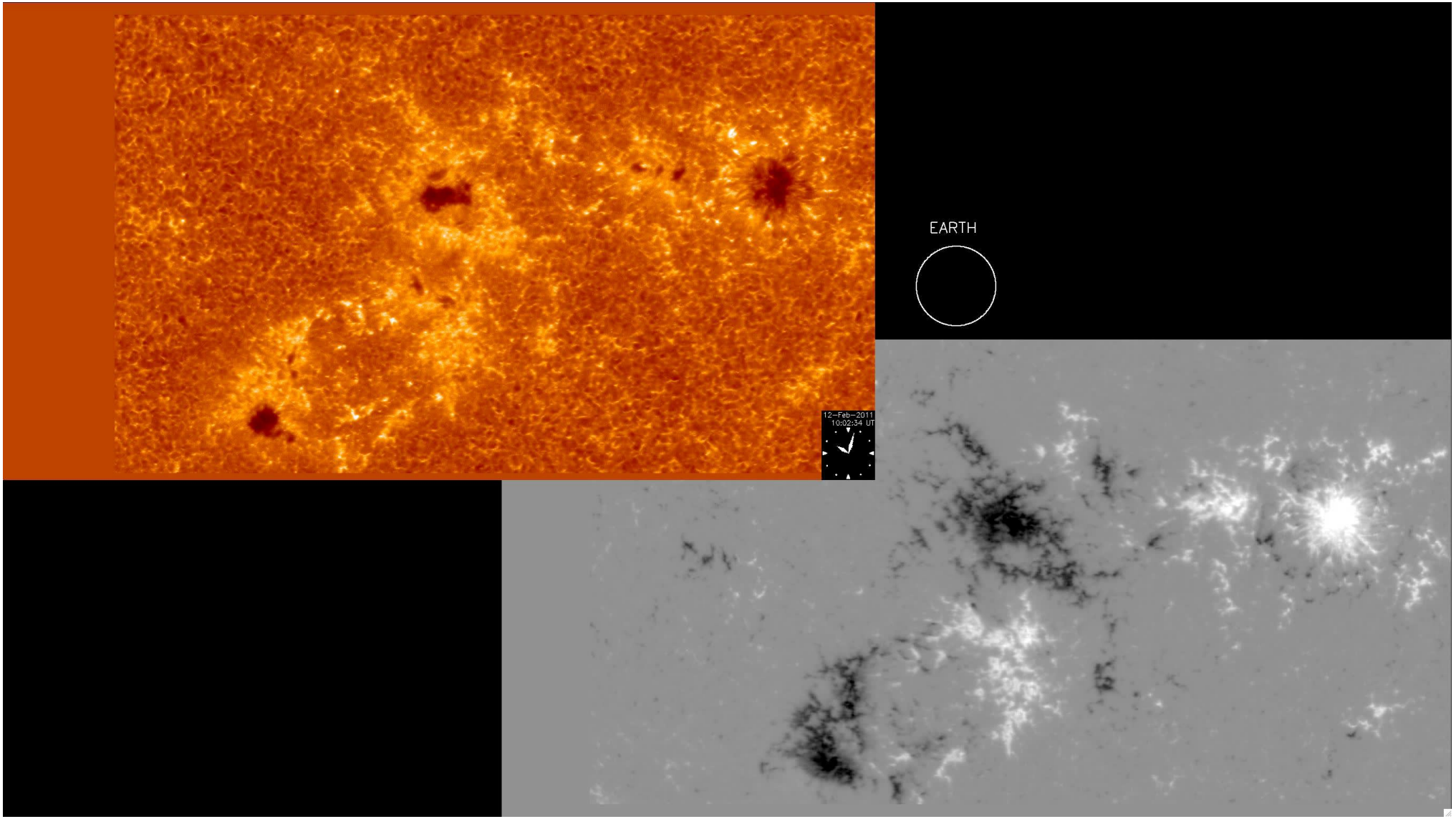
機械学習モデルの評価

汎化性能を調べるためには，訓練に使用していないデータ（検証/テストデータ）で評価が必要！



太陽フレアの予測

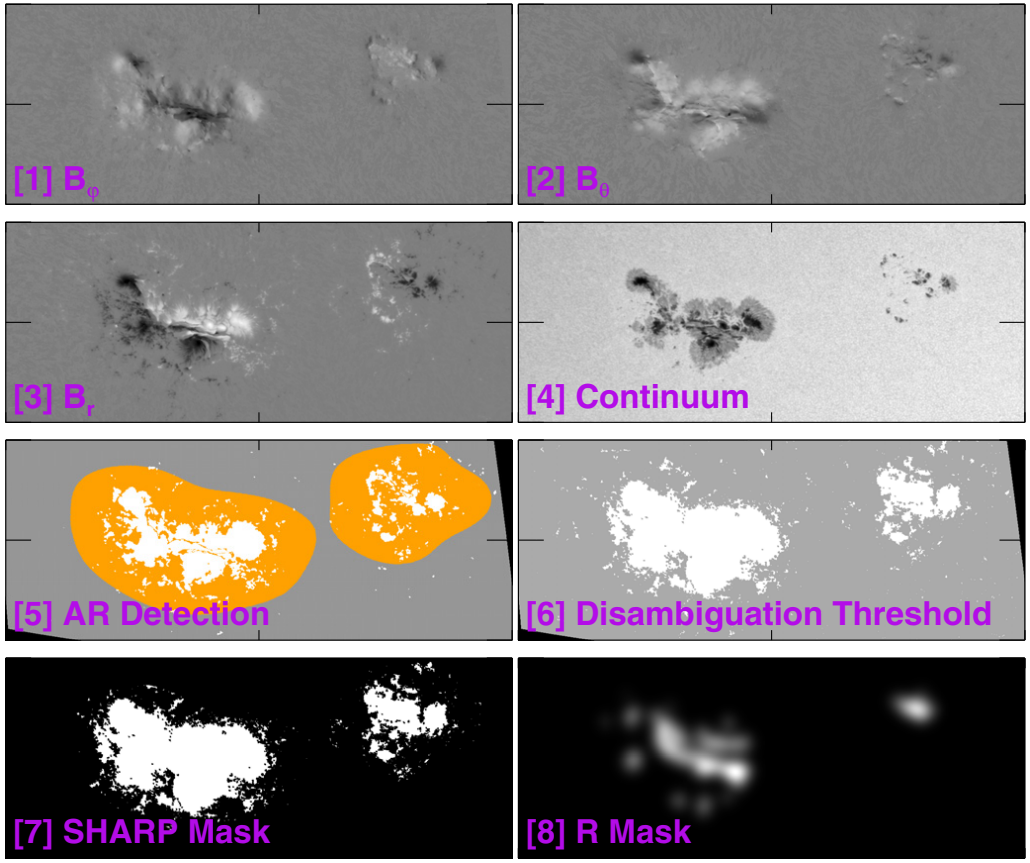
NIIGATA UNIVERSITY



機械学習によるフレア予測 (Bobra+, 2015)

NIIGATA UNIVERSITY

- 約150万枚の活動領域画像から予測モデルを構築
- 人を超えたTSS=0.76を達成　　cf) TSS～0.5？@研究者



$$TSS = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$$

Prediction	Flare	No flare
Flare	TP	FP
No flare	FN	TN

Flare　　No flare

Ground truth

これまでの予測スコア

NIIGATA UNIVERSITY

DNNよりCNNの方がスコアが低い？？

予測手法	TSS
線形分類器 (時系列+画像データ) (Jonas et al. 2018)	0.810±0.030
DNN (Nishizuka et al. 2018)	0.800
LSTM (Liu et al. 2019)	0.792±0.008
SVM (Bobra et al. 2015)	0.761±0.039
CNN (Li et al. 2020)	0.749±0.079
Random Forest (Florios et al. 2018)	0.740±0.020
CNN (Huang et al. 2018)	0.662

CNNによるフレア予測モデル構築

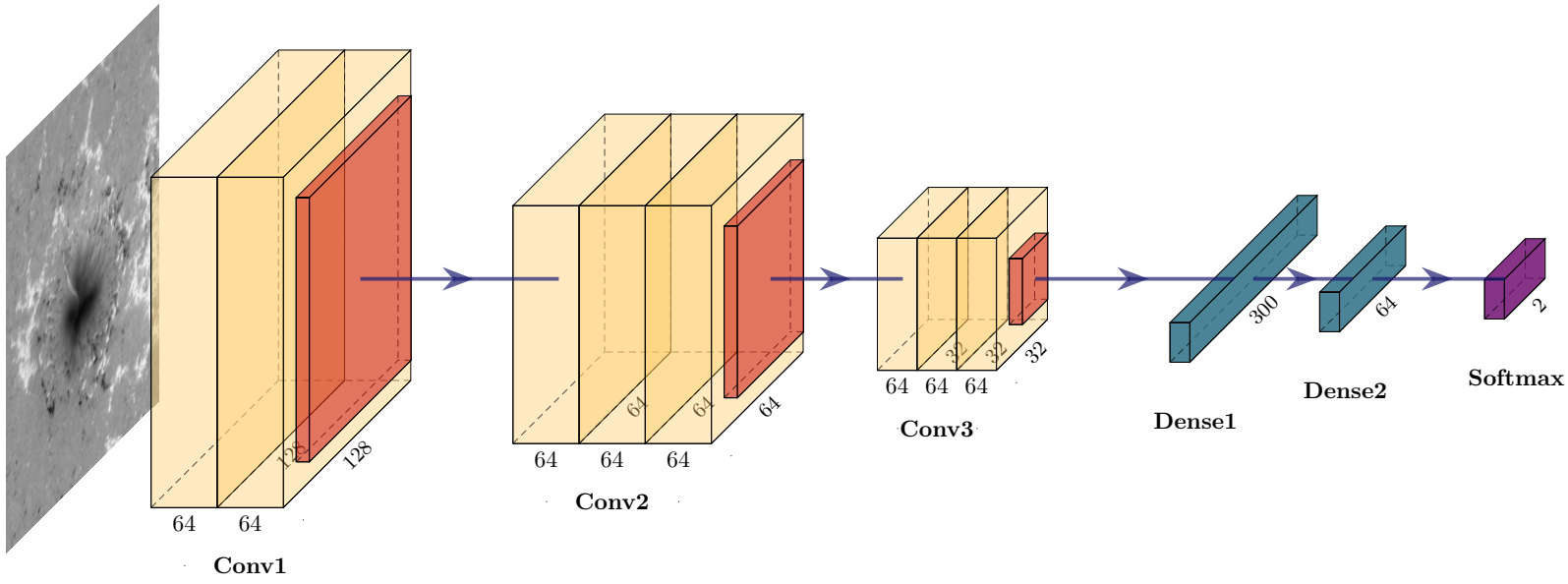
NIIGATA UNIVERSITY

調べてみたところ，作成されたデータのミスが原因だったよう…

- 全手法で予測精度が上昇
- CNNの予測精度が一番に

学習手法	TSS
CNN	0.928±0.009
DNN	0.911±0.005
Random Forest	0.901±0.006
k 近傍法	0.891±0.005
SVM	0.777±0.011
ロジスティック回帰	0.762±0.011

CNNの方が間違いデータに敏感？



End-to-endなフレア予測手法の開発

NIIGATA UNIVERSITY

従来手法

全球画像 > ルールベースで検出 + 人手による補正 > フレア予測

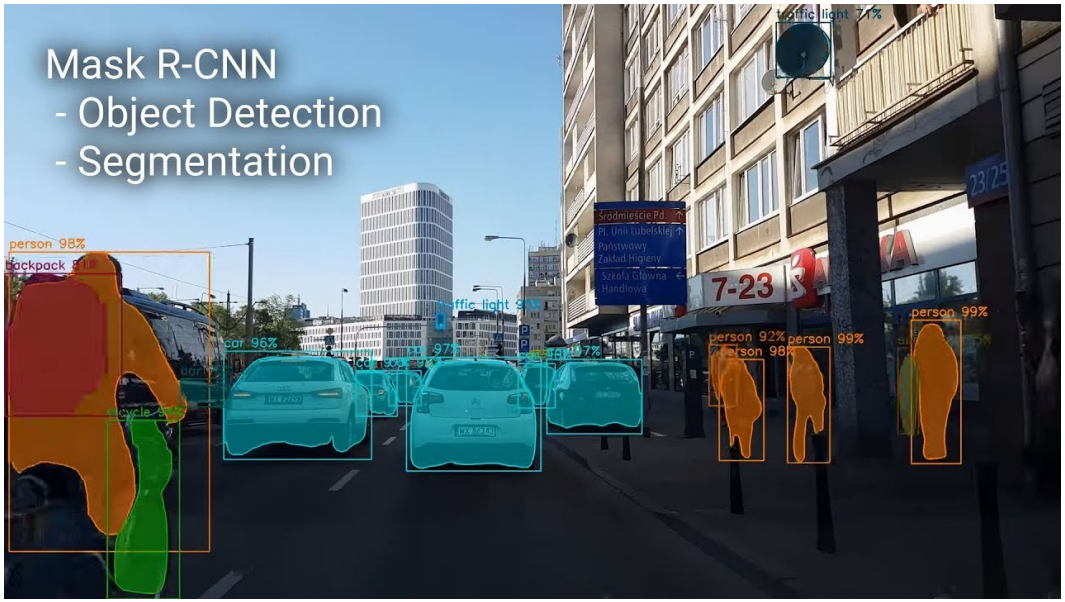
提案手法

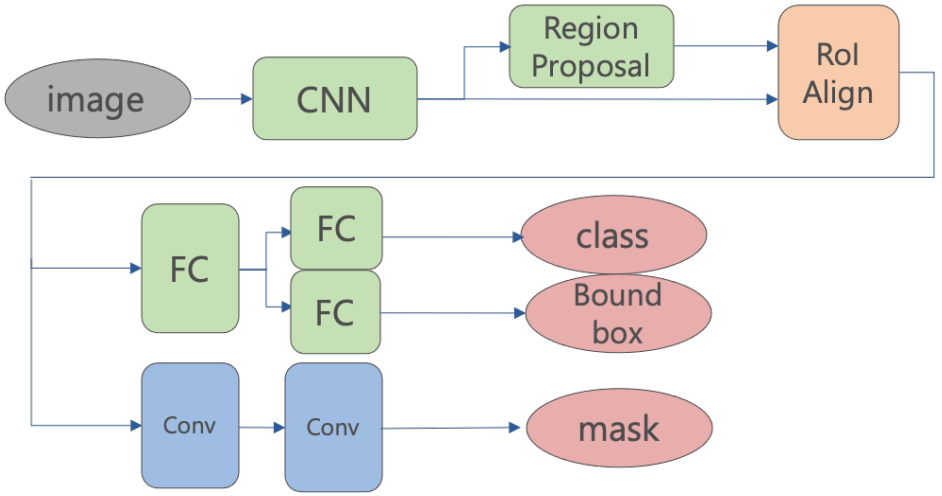
全球画像 > 深層学習で領域検出 > フレア予測

Mask R-CNN (He et al., 2017)

NIIGATA UNIVERSITY

- **マスク学習ブランチ**を持つ，物体検出モデル.
- ３つの損失関数(BB, MSK, CLS)を最適化する.
- COCO分類問題において最高スコアを達成.





$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{box}} + \mathcal{L}_{\text{mask}}$$

$$\mathcal{L}(\{p_i\},\{t_i\}) = \frac{1}{N_{\text{cls}}}\sum_i \mathcal{L}_{\text{cls}}(p_i, p_i^*) + \frac{\lambda}{N_{\text{box}}}\sum_i p_i^* \cdot L_1^{\text{smooth}}(t_i - t_i^*)$$

$$\mathcal{L}_{\text{cls}}(p_i, p_i^*) = -p_i^* \log p_i - (1 - p_i^*)\log(1 - p_i)$$

$$\mathcal{L}_{\text{mask}} = -\frac{1}{m^2}\sum_{1 \le i,j \le m} [y_{ij} \log \hat{y}_{ij}^k + (1 - y_{ij}) \log(1 - \hat{y}_{ij}^k)]$$

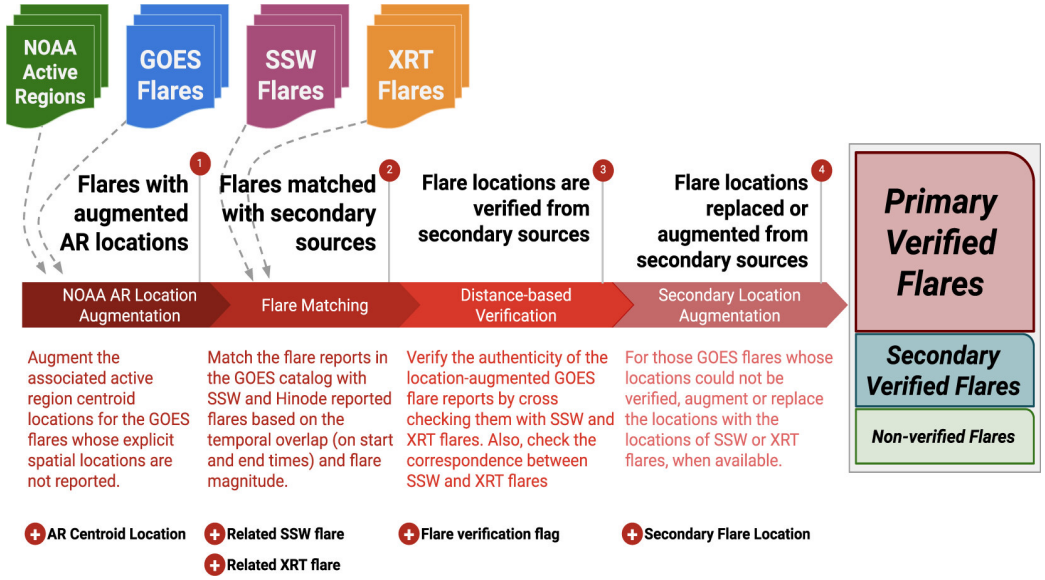
データセット

SHARP (Bobra et al., 2017)

- ルールベース+人手による活動領域検出
- 磁場画像 + 座標，形状，物理量
- 領域Annotationに使用

SHARP (Angryk et al., 2020)

- 複数データを統合したフレアデータベース
- これまでのものよりも正確にラベル付け
- フレア発生ラベルに使用



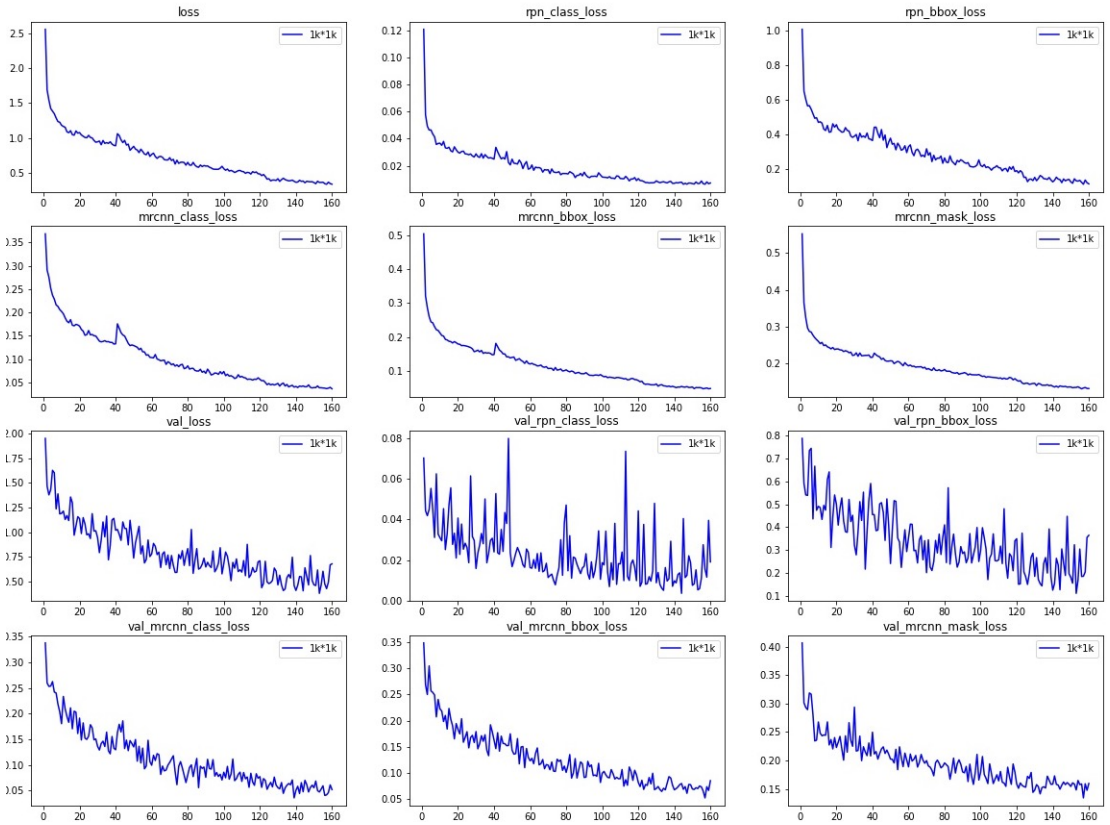
データセット

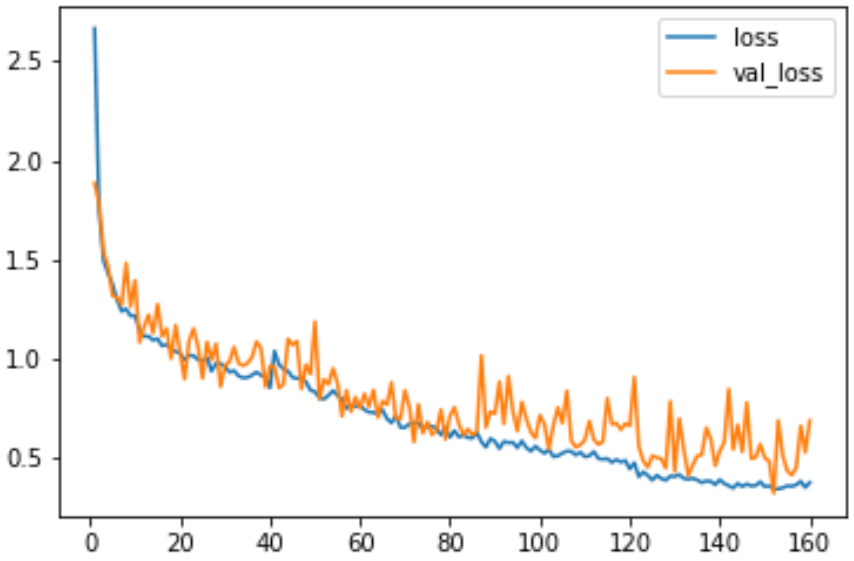
2010年5月～2011年4月で以下のデータを使用

入力画像　　- SDOの視線方向磁場画像
領域Annotation　- SHARPを利用
フレア発生ラベル - SWANから作成．24時間以内発生で2クラス

	学習用（8割）			検証用（2割）		
期間	画像 [枚]	フレア未発生[例]	フレア発生[例]	画像[枚]	フレア未発生[例]	フレア発生[例]
201005-201104	4,999	32,988	2,976	1,290	8,284	750

学習の推移



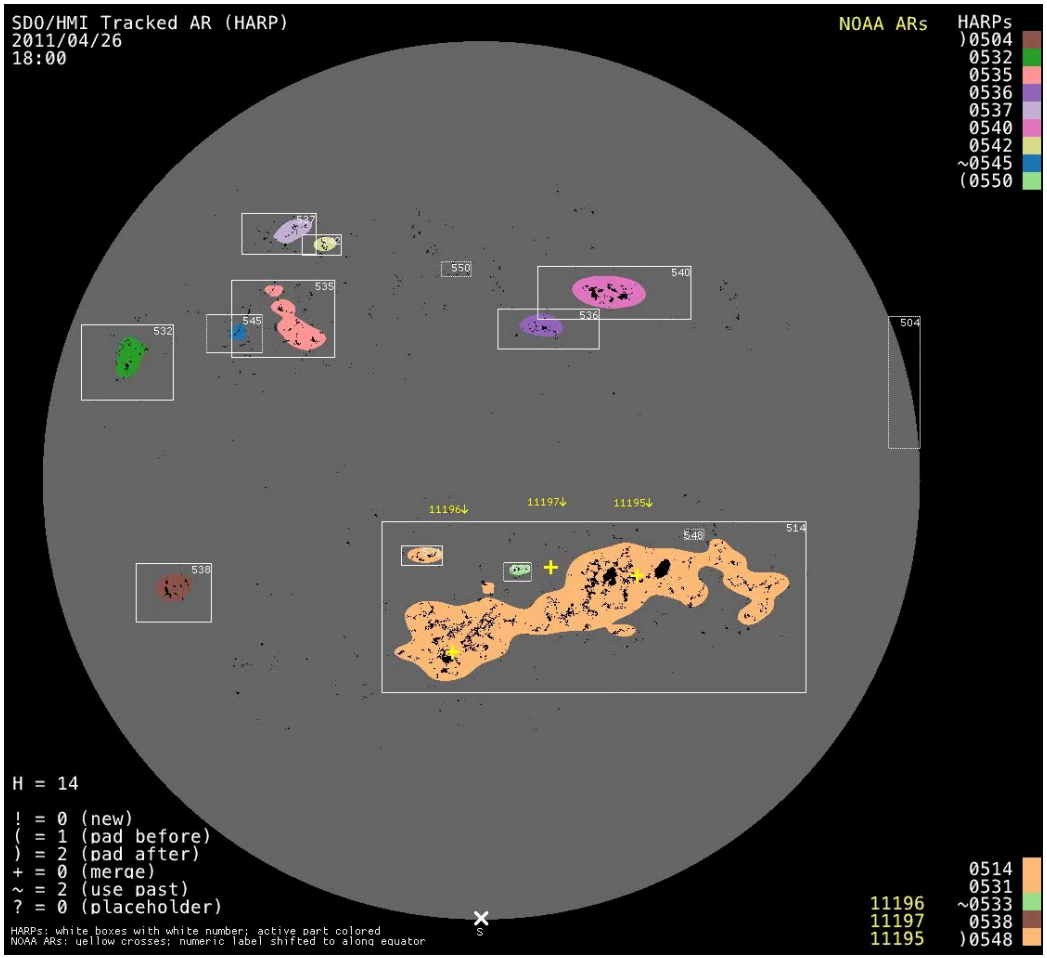


汎化性能を維持しながら推移

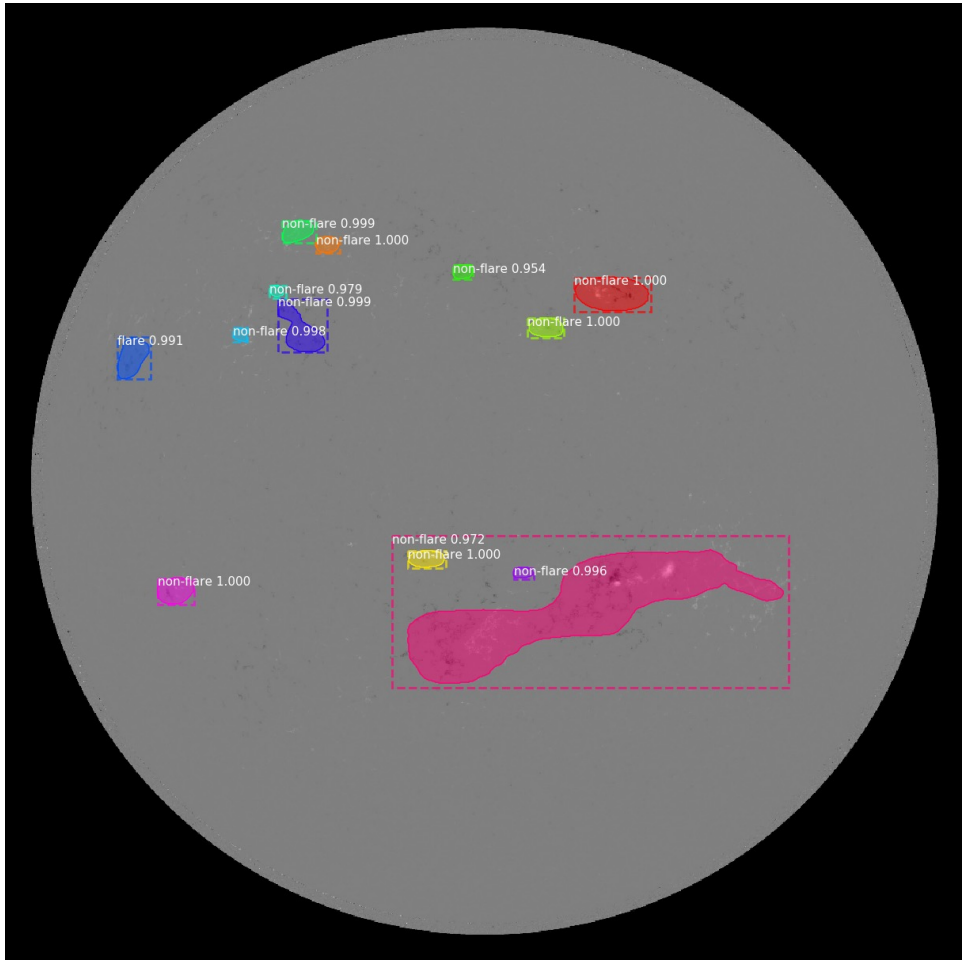
検出結果

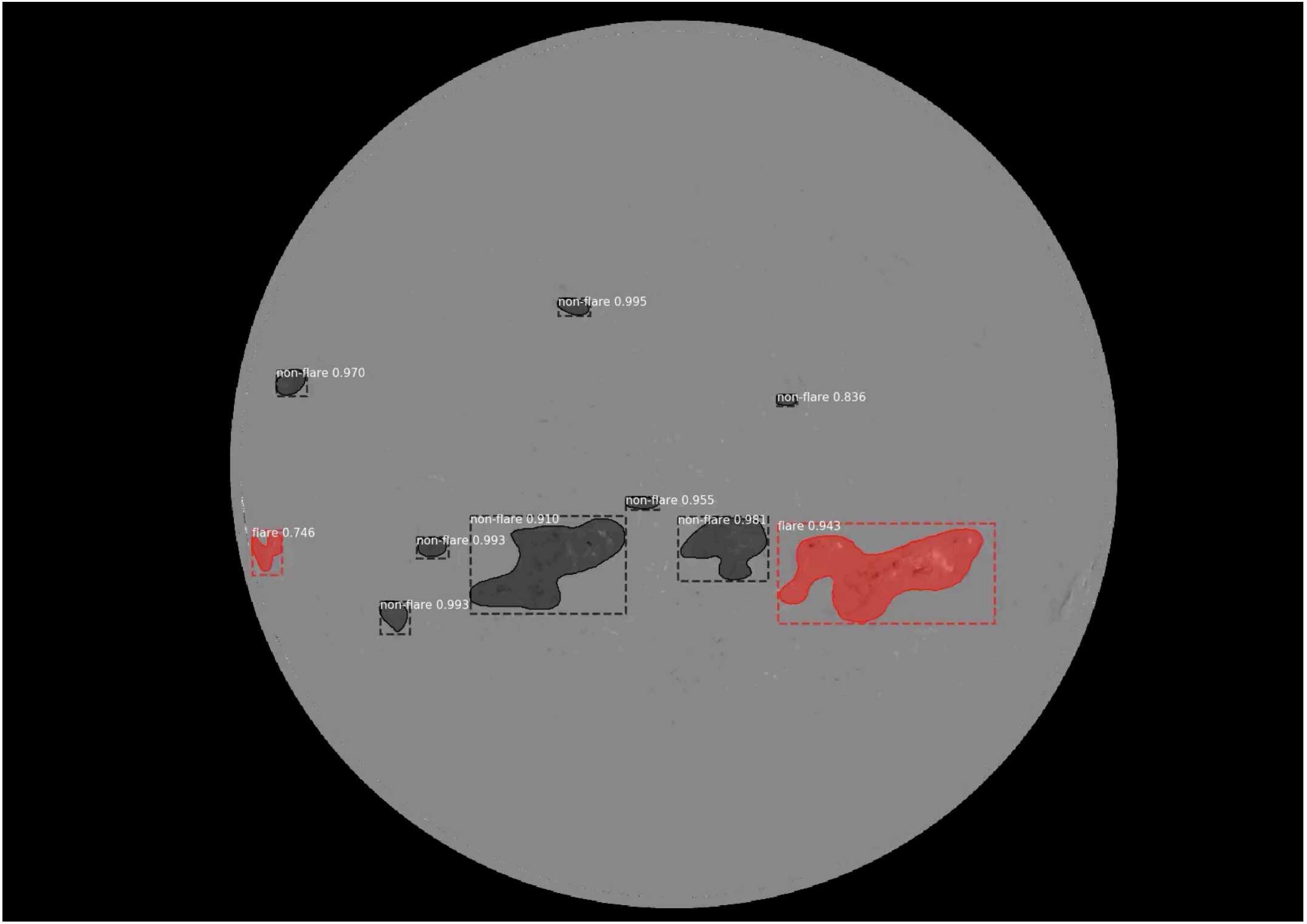
NIIGATA UNIVERSITY

Ground Truth



検出結果



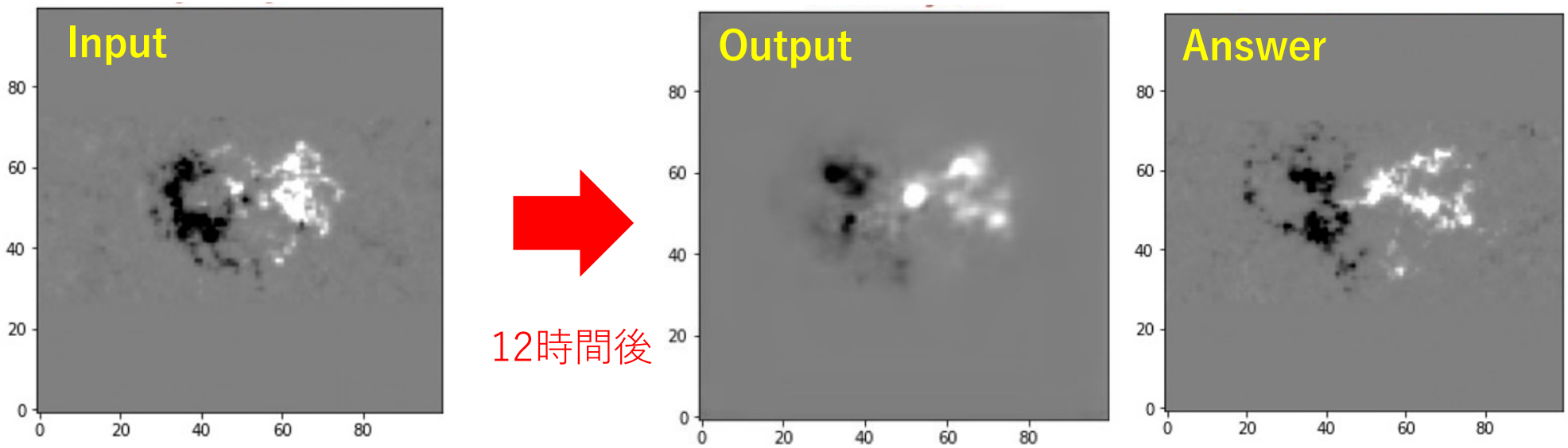

non-flare 0.995
non-flare 0.970
non-flare 0.836
non-flare 0.955
non-flare 0.910
non-flare 0.981
flare 0.943
flare 0.746
non-flare 0.993
non-flare 0.993

黒点成長の画像予測

NIIGATA UNIVERSITY

エンコーダ-デコーダモデルにより画像予測を行う.
→ それなりに特徴をとらえていそう.
→ MHD方程式が学習できている？

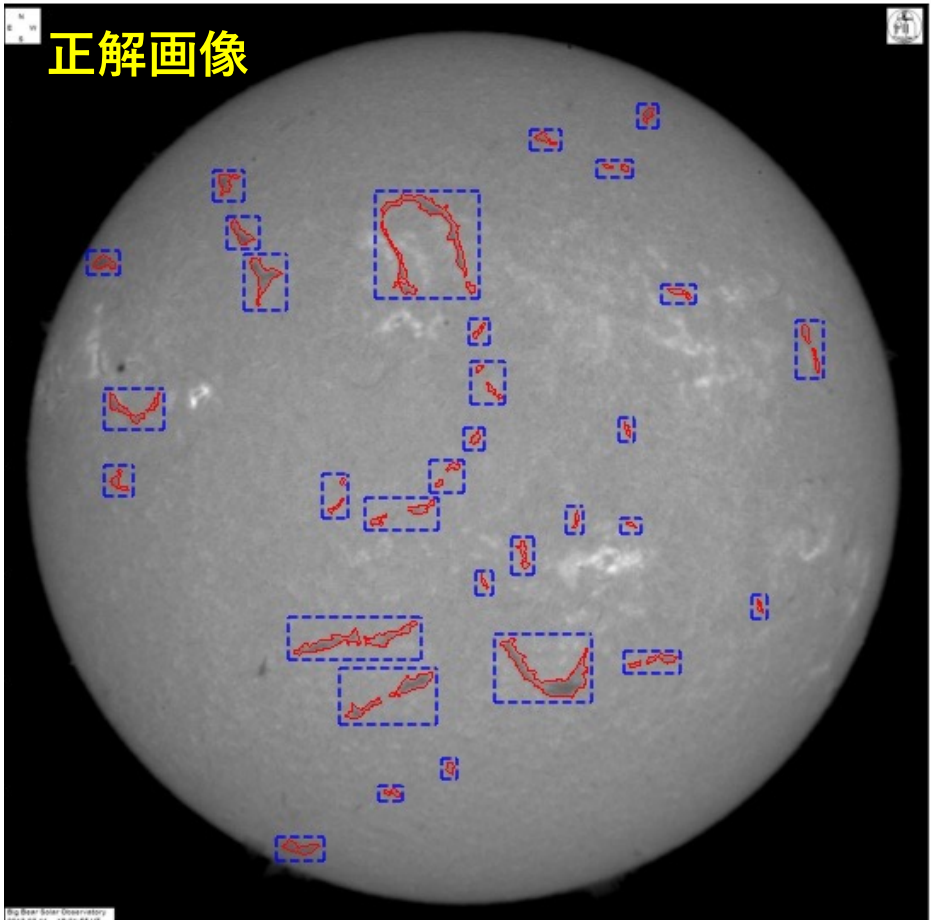
高精度な予測を目指してGANを導入中

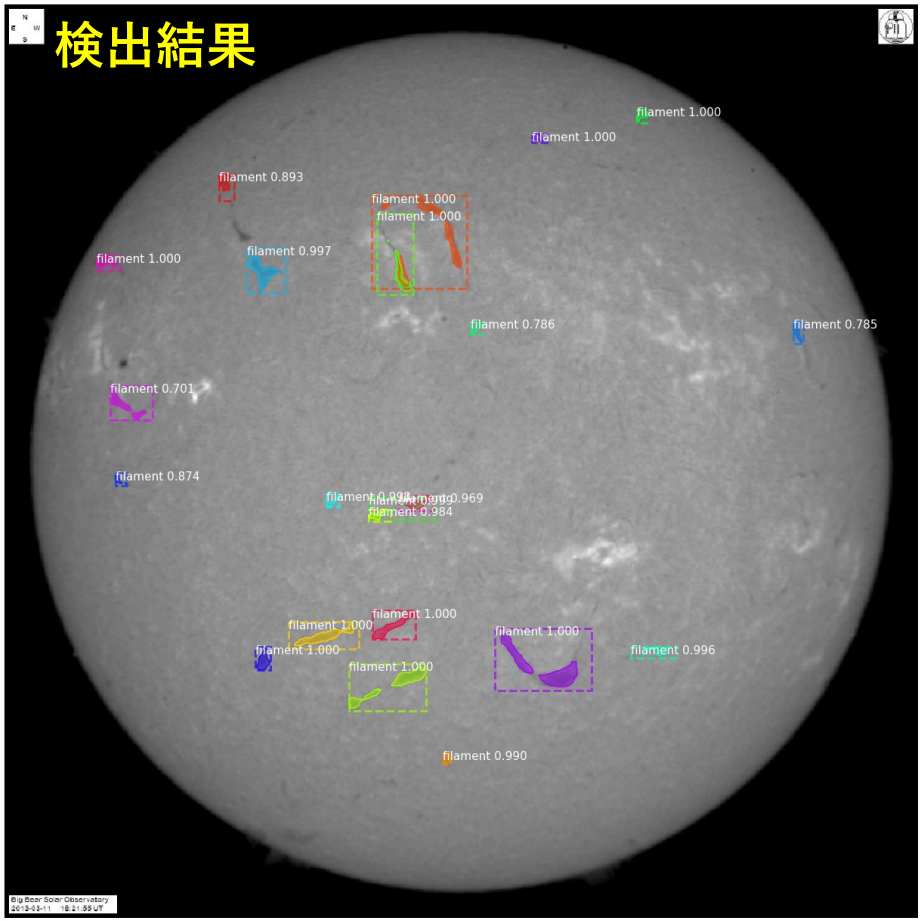


太陽フィラメント検出

NIIGATA UNIVERSITY

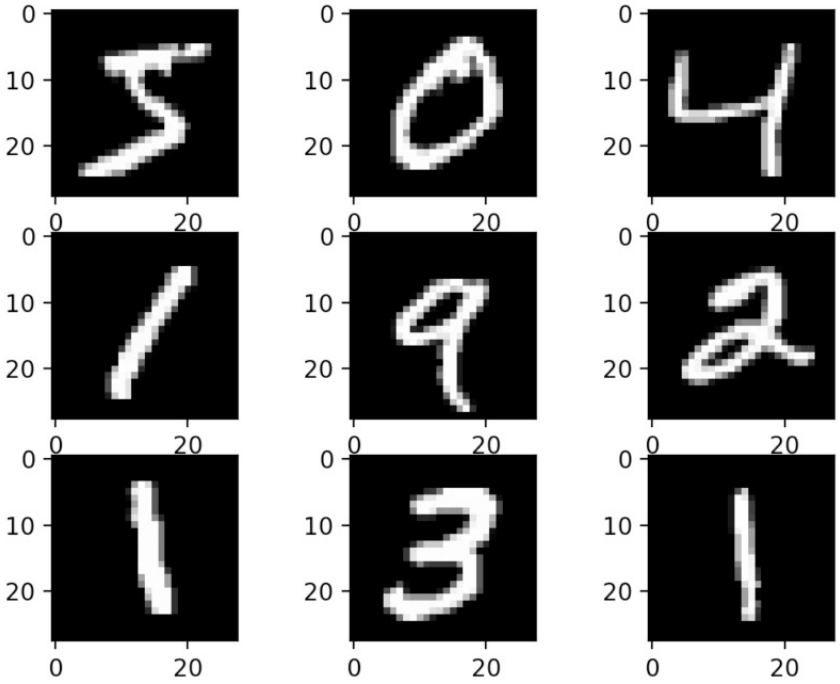
“空間相似性を持つ物体では部分構造を抽出”してしまう，現DLアルゴリズムの弱点が明らかに... → RPNを改良中！

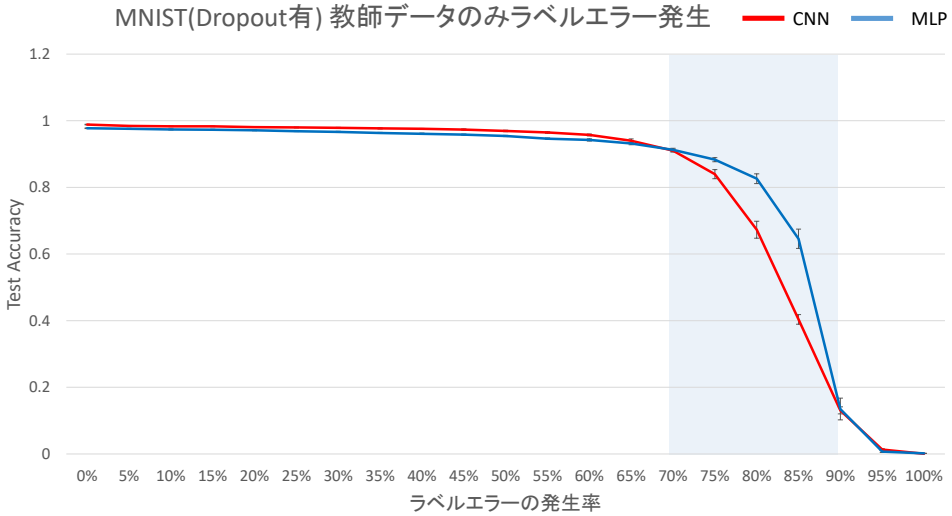


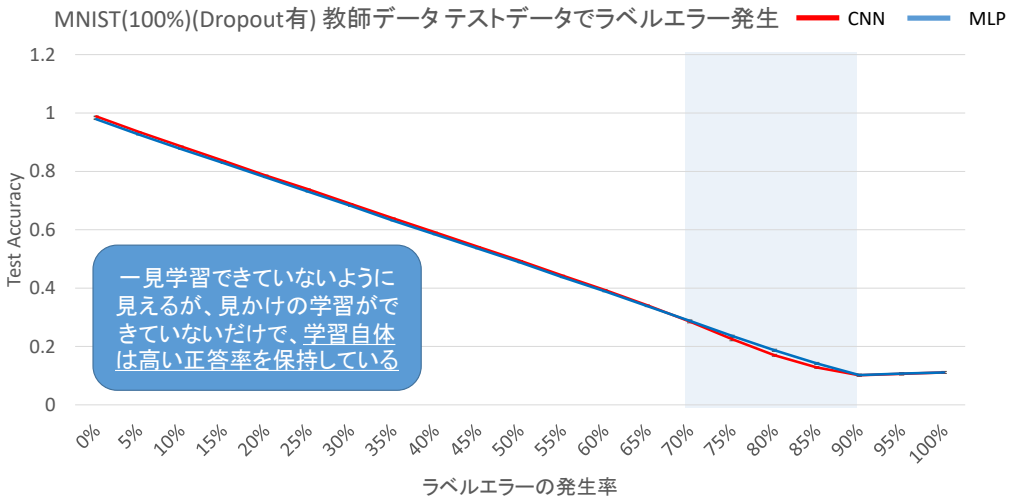


教師ラベルエラーの影響探査

実は，深層学習モデルは教師ラベルエラーにかなり強いかも？？







座学の最後に…

NIIGATA UNIVERSITY

宇宙情報学研究室として，**新しい宇宙科学の知見獲得**と**情報科学技術の発展**を目指しています．興味がある方は，ぜひ一緒に研究しましょう．