

# 畳み込みニューラルネットワークを用いたすばる望遠鏡による Lyman-alpha Emitter 観測データの自動分類

大倉 真一希† 天笠 俊之†† 伊藤 凌平††† 大内正己†††† 北川博之††

† 筑波大学情報学群情報科学類

†† 筑波大学計算科学研究センター

††† 東京大学大学院 理学系研究科物理学専攻

†††† 東京大学 宇宙線研究所

## 1 はじめに

2012年よりすばる望遠鏡に搭載された超広視野 CCD カメラ Hyper Suprime-Cam(HSC)は一度の観測で $10^8$ に及ぶ膨大な量の超遠方宇宙のデータを得ることを可能とした。観測画像より重要な天体を見つけ出す際は、画像中の光源の等級を測定し、設定された基準を超えたものを取り出すという処理を行う。しかし、その処理の結果として得られる天体の候補には、目的外の対象を誤検出したもの (spurious sources) が数多く含まれる。真に重要な天体を見つけ出すには、天体の候補から spurious sources を取り除くことが必要となる。従来の天体の観測では HSC と同程度の規模の観測データを得られるものは存在せず、人手でのスクリーニングによって、目的の天体と spurious sources を分類していた。しかし、莫大なデータを人手分類することは困難である。そのため、人間と同様の基準を持って自動で分類を行えるような手法の開発が望まれている。

本研究では、HSC によって得られる観測データのうち、特に遠方銀河 Lyman-alpha Emitter(LAE) の候補について、それを真の LAE と spurious sources に自動で分類するモデルを設計し、実データに対して適用することを目的とする。現在までに Convolutional Neural Networks(CNN) を用いた分類器の設計及びその性能評価を行い、実環境下での有効性を検証した。

## 2 前提知識

### 2.1 Lyman-alpha Emitter

LAE は宇宙形成初期の銀河の一種である。LAE のような初期の銀河はダストをほとんど含まないため、通常の銀河からは観測されない狭い帯域に光が観測される。よって、特定の狭帯域を観測し、光の有無を確認することで LAE の同定が行える。

HSC での遠方銀河の観測では、主に5つの広帯域フィルタ (g, r, i, z, y) を決め撮影を行う。結果として、1つの銀河に対して、異なる5つの波長で撮影された画像が得られる。LAE を検出する際は5つの広帯域フィルタに加えて、観測する赤方偏移に応じた狭帯域フィルタを選択する。この時、広帯域フィルタの範囲に光源が現れておらず、狭帯域には現れているならば、LAE と同定することができる。今回は比較対象の広域フィルタとして、光源の有無が明確な g バンドを利用する。図1に LAE の例を示す。

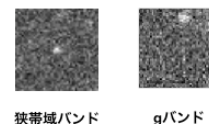


図1: LAE の例。狭帯域には中心に光源があるが、g バンドにはない。

LAE の候補には図2に示すような本来検出されるべきでない spurious sources が多く含まれ、全候補の98%を占める。真の LAE を得るため、候補からこれらを取り除く必要がある。

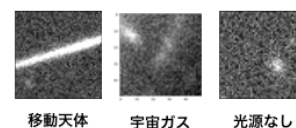


図2: spurious sources の例。

### 2.2 Convolutional Neural Network

Convolutional Neural Network(CNN)[1]は近年の画像認識の分野で大きな成果を成し遂げている手法であり、様々な分野に応用がされている。CNN はニューラルネットワークを発展させた手法であり、データから特徴を抽出する層と、抽出した特徴を元に分類を行う層から成っている。CNN の特徴量抽出部は任意の数のフィルタの畳み込みにより画像から特徴マップを取り出す Convolution 層と、特徴マップを正方領域に分割し統計値に変換することでダウンサンプリングを行う Pooling 層を積み重ねた構造を持っている。特徴量抽出部のパラメータを誤差逆伝搬方によって最適化することで、特

Automatic classification of Subaru-telescope's Lyman-alpha Emitter observation data with convolutional neural networks

†Maiki OKURA ††Toshiyuki AMAGASA †††Ryohei ITOH

††††Masami OUCHI ††Hiroyuki KITAGAWA

†College of Information Science, University of Tsukuba

††Center for Computational Science, University of Tsukuba

†††Graduate School of Science, University of Tokyo

††††Institute for Cosmic Ray Research, University of Tokyo

微量の設計を必要としない End-to-End な学習を可能としている。

### 3 CNNによるLAEの分類

#### 3.1 画像への前処理

観測画像は fits 形式として得られる。fits データには観測の際の情報なども含まれるが、今回は画像のみを利用する。fits 中の画像はピクセル値に上下限がないため、学習に利用できるような正規化を行う必要がある。今回は実際に天体画像を見る際によく用いられる zscale を利用する。zscale は画像全体のピクセルの輝度を中央値付近から線形的に分布させるようなアルゴリズムであり、極端な輝度分布を画像に対して用いられる。画像の本来のサイズは 77x77 であるが、光源の有無の判定は中心点のみを見るため、中央部分 50x50 のみを切り出す。学習の際は画像のピクセル値を [0, 1] の範囲に納める。

#### 3.2 CNNの構成

本研究ではデータセットの性質変化に対応しやすいようシンプルな CNN アーキテクチャを用いることとした。図 3 に実験に利用した CNN の構成を示す。

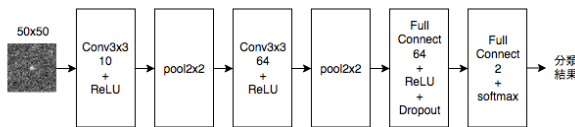


図 3: 本研究で利用した CNN アーキテクチャの構成。

## 4 実験

本来の LAE はマルチバンドを用いて検出されるが、学習データセットとしては十分なものが用意できない。そこで本研究では、本来の分類タスクを、中心に光源がある天体画像と spurious sources の分類に簡単化し、性能評価を行なった。

#### 4.1 データセット

CNN の学習とパラメータ調整に利用する訓練用データセットとして、36000 個の spurious sources と 48000 枚の天体画像を用意した。訓練用データセットのうち、spurious sources は実際の観測から得られたものであり、天体画像は設定した条件の下でデータを自動生成できる SynPipe[3] を用いて作成したものである。また、最終的な性能評価のためのデータセットとして 62914 個の spurious sources と実際の観測より得られた 776 枚の天体画像を用意した。

#### 4.2 実験手順

まず、訓練用データセットをランダムに 5 分割し、交差検定を行う。その際、validation には訓練に利用しな

い残り 1 分割の中の spurious sources とテスト用から固定で取り出した 200 個の LAE を利用する。評価指標としては Recall と Specificity を用いる。5 回の validation の平均結果を評価し、結果が最良となるようなパラメータを決定する。最終的に全ての訓練データと最適なパラメータを用いて CNN の訓練を行い、テストデータでモデルの実環境での性能を測る。

#### 4.3 実験結果と考察

実験より、最適化手法として確率的勾配法を利用し、学習率を 0.001、学習エポック数を 220 と設定することとした。最適なパラメータで訓練を行なったモデルでのテストデータに対する分類結果を表 1 にまとめる。

表 1: 実観測データに対する CNN による分類結果

真のラベル \ 分類結果	天体	spurious sources
	天体	550
spurious sources	671	62243

結果として、spurious sources 検出率を表す Specificity は 98.8% と目標を達成している。天体検出率を表す Recall は 95.5% と目標を超えているが、今後この分類結果を元に LAE の検出を行うと考えると不十分である。この原因として、実観測された天体の画像を訓練に用いておらず、そのデータ分布を学習できていないことが考えられる。また、spurious sources のデータの中にも天体と思われるものが残っていたため、それらを取り除くことが性能向上に繋がると思われる。

## 5 まとめと今後の課題

本研究では CNN を用いた LAE の分類器の設計及び、実環境データでの性能評価を行なった。その結果、spurious sources は十分な精度で取り除けるものの、天体はマルチバンドでの LAE 検出を考慮すると精度が不十分であることがわかった。今後は実データでの精度を向上させるとともに、マルチバンドでの LAE 検出の性能を測る必要がある。

### 参考文献

- [1] Alex Krizhevsky et al, “ImageNet Classification with Deep Convolution Neural Networks,” NIPS, 2012.
- [2] Mikio MORII et al, “Machine-learning Selection of Optical Transients in Subaru/Hyper Suprime-Cam Survey,” Publ. Astron. Soc. Japan, 2014.
- [3] Song Huang et al, “Characterization and Photometric Performance of the Hyper Suprime-Cam Software Pipeline,” Publ. Astron. Soc. Japan, 2014.