Distinguishing HII Regions from Galaxies using Machine Learning

Annabel To, Allen Papp
Department of Computer Science, College of Natural Sciences, The University of Texas

The University of Texas at Austin
College of Natural Sciences
The University of Texas at Austin
Freshman Research Initiative
College of Natural Sciences

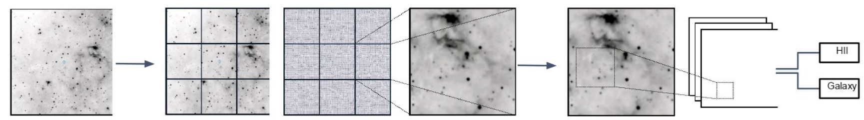
Introduction

HII regions are pockets of hydrogen gas that surround and are ionized by O and B type stars, producing unique emission lines known as forbidden lines. The principal goal of our project is to create an astronomical pipeline that is able to locate areas of space where there are appreciably large HII regions present. The pipeline uses neural nets to distinguish HII regions from other objects, most notably galaxies, which are often confused with HII regions by modern classification pipelines due to their similarity in structure. Our source of data for HII regions is the Wide-Field Infrared Survey Explorer (WISE) available through IRSA as this survey contains an all-sky view in the 21 μM range, which is coincident with the forbidden line spectra of Hydrogen. For galaxy coordinates, we use the Galaxy10 DECals dataset, sourced from Galaxy Zoo's 2nd data release. This project sits at the intersection of Computer Science and Astronomy and is a nice showcase of interesting concepts from both fields.

Research Goal

Using our pipeline, we hope to create a catalog of HII regions, as current repositories are limited on the scale of several thousand artefacts. For each object identified, we cross reference with existing catalogs to find potentially missed or misidentified HII regions. We also aim for an efficient and compact pipeline, for future general use in classifying other objects such as nebulae, stars, and different types of galaxies.

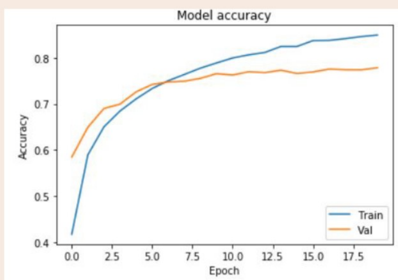
Classification Pipeline

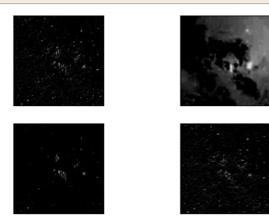


Original image   Recursive grid division/focus using gradient   CNN classification, label weighted by zoom

Figures

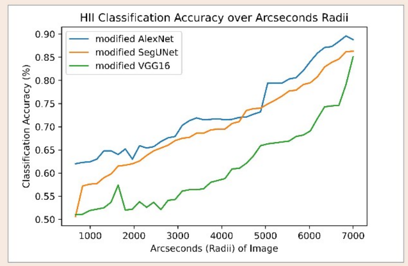
We measured the classification accuracy of our pipeline at various snapshot radii (measured as arcseconds through IPAC's telescope), to have a better confidence level of classification based on the scope of each image. We plotted the general progression of training our pipeline's base neural net (using the pipeline illustrated above). The pipeline's architecture is an adaptation of AlexNet, and classified using two labels (HII region or galaxy).





Notable feature extraction at 2000 (left) vs 6000 arcseconds (right)

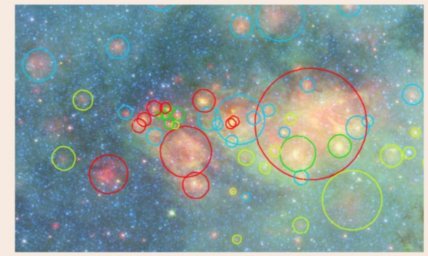
Results



We found that larger images (capped at around 6000 arc seconds of radii) were much more accurately identifiable as HII regions or galaxies by our pipeline. Because our neural net uses pixel-based classification, we handled the extra computational cost of larger images using a gradient filter based on pixel brightness. If an image contained appreciably high gradient values, then the image was included in the testing dataset. This filter reduced the number of images that we needed to examine on our test patch of sky by nearly a third.

Furthermore, these optimal classification arc seconds signify that visual indicators of an HII region, notably the shifting pattern of gasses clustered around stars, only become recognizable by the neural net at a particular zoom level. We extracted the features identified after the first layer of convolution, and found that these cloud-like gas formations were sometimes misidentified in images with arc second radii below 3000. Thus, when deciding whether a particular grid unit of a larger image held objects of interest, we weighted confidence in proportion to the neural net's test accuracy at the corresponding arc seconds.

Conclusion



With image-based analysis on forbidden line spectra, we were able to achieve decent classification accuracy, and hope to increase this accuracy using instance segmentation, allowing for greater flexibility in labeling images with multiple components (eg. an HII region image may also contain point galaxies). We believe that being able to identify HII regions from other structures within the same image would further improve accuracy and allow for more complex (rather than gridlike) divisions over large patches of space.

Acknowledgments

We express sincere gratitude for the image data from the WISE catalog through IPAC at the California Institute of Technology, the image data from the Galaxy10 DECals dataset, and the assistance of other members of the Geometry of Space research group, namely Jose Ordonez and Rik Ghosh, as well as Dr. Karl Gebhardt and Dr. Shyamal Mitra.

References

- Mannucci, F., Belfiore, F., Curti, M., et al. 2021, Monthly Notices of the Royal Astronomical Society
- Borgman, C. L., Wofford, M. F. 2021, Harvard Data Science Review
- Bowen, I. S. 1928, The Astrophysical Journal, 67, 1
- Xie, T., Mundy, L. G., Vogel, S. N., amp; Hofner, P. 1996, The Astrophysical Journal, 473
- Russell, S. J., Norvig, P. 2010, Artificial Intelligence: A modern approach (Upper Saddle River: Prentice-Hall)
- Geman, S., Bienenstock, E., Doursat, R. 1992, Neural Computation, 4, 1
- Brodley, C. E., Friedl, M. A., IGARSS '96 1996 International Geoscience and Remote Sensing Symposium
- Hausen, R., Robertson, B. E. 2020, The Astrophysical Journal Supplement Series, 248, 20