Classifying Galaxies Using A CNN

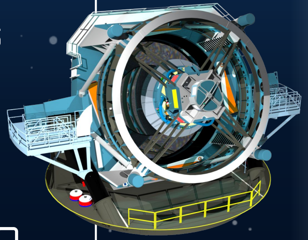
By: Steve Young

01

Scientific Motivation

The research question, personal motivation, and importance of it.

Scientific Motivation



- Research Question: Can Convolutional Neural Networks classify the different types of Galaxies?
- I chose this research question because with the introduction of new telescopes such as the Vera C. Rubin Observatory, it will capture billions of images of the sky and within these images will be new galaxies.
- However, if we want to classify these newly discovered galaxies, we can not do it manually because there are just so many and it will take too much time.
- With the use of CNN's this process of classifying galaxies can be done automatically.

02

Data

Data used, where and how it was obtained.

Classifications of Galaxies

Elliptical Galaxy

- Ovular in shape
- Minimal star formation
- Redder due to old star populations

Spiral Galaxy

- Circular in shape
- Contain spiral arms
- Lots of star formation
- Has a stellar disk
- Bluer due to new stars forming

Irregular Galaxy

- Does not have any particular shape
- Some of the youngest galaxies
- Lots of star formation

Data Used

- The datasets that I used for this project are Galaxy Zoo 1 and EFIGI (Extraction de Formes Idealisées de Galaxies en Imagerie)
- Galaxy Zoo 1 is a project where volunteers help researchers classify galaxies through their morphological features
- EFIGI was a project where the goal was to measure galaxy morphologies for a variety of galaxies
- Within the database of Galaxy Zoo 1, they have the classification of around 900,000 galaxies
- In the EFIGI database, they have morphological information and images for 4,458 galaxies
- I then cross referenced the EFIGI dataset with Galaxy Zoo 1's database and this left me with 2053 Spiral Galaxies, 230 Elliptical and 33 Irregular Galaxies

03

Methods

Procedure and techniques used.

Methods

- The methods I used is creating a Convolutional Neural Network using the Keras library. This is the architecture I used.

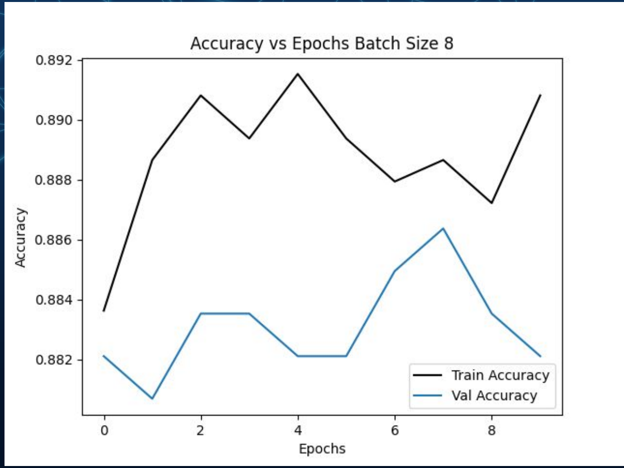
```
model1 = keras.Sequential()
model1.add(keras.layers.Input(shape=(256, 256, 1)))
model1.add(keras.layers.Conv2D(96, (8, 8), activation='relu'))
model1.add(keras.layers.Dropout(0.25))
model1.add(keras.layers.Activation('relu'))
model1.add(keras.layers.MaxPooling2D((3,3)))
model1.add(keras.layers.Flatten())
model1.add(keras.layers.Dense(24))
model1.add(keras.layers.Activation('relu'))
model1.add(keras.layers.Dropout(0.5))
model1.add(keras.layers.Dense(3, activation='softmax'))
model1.summary()
```

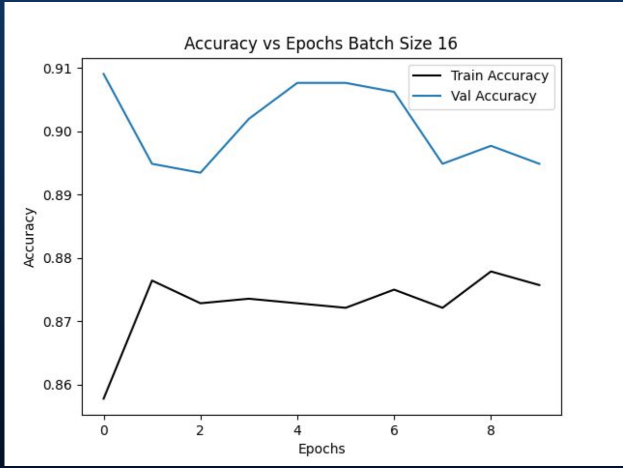
- This architecture that I used is based on the article Deep Galaxy: Classification of Galaxies based on Deep Convolutional Neural Networks by Khalifa, et al. (2017).

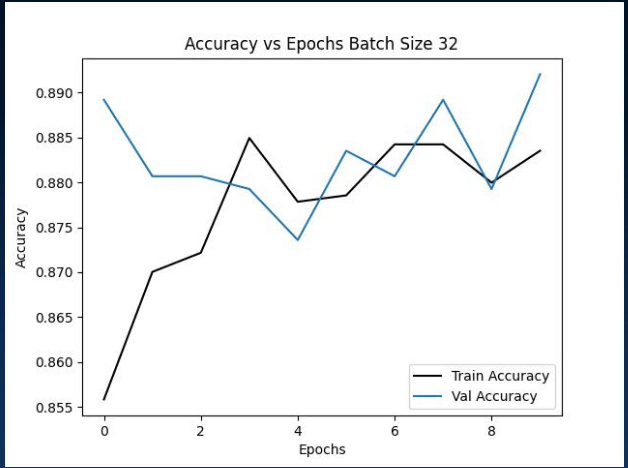
04

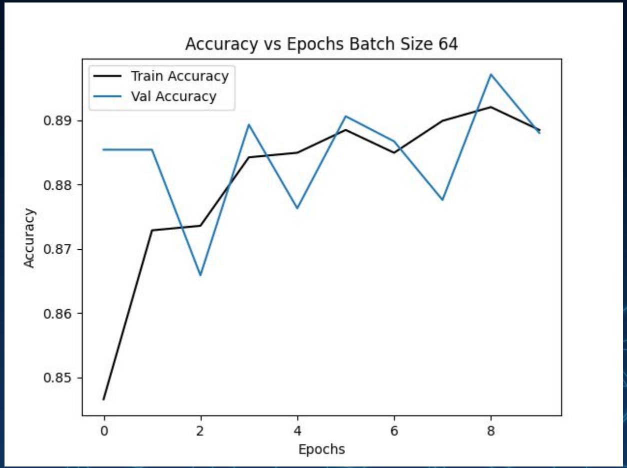
Results

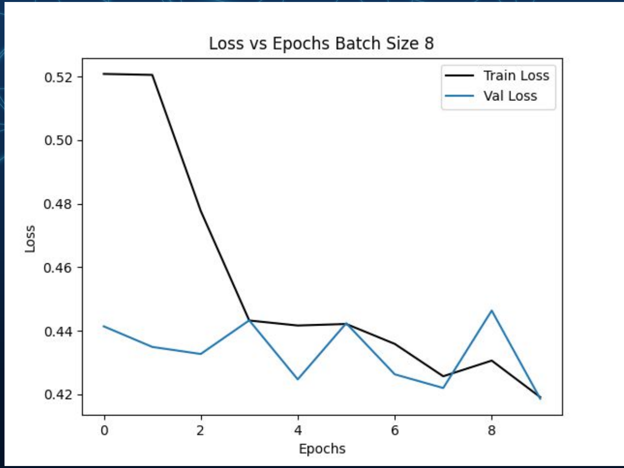
Findings about the research question and significance of it.

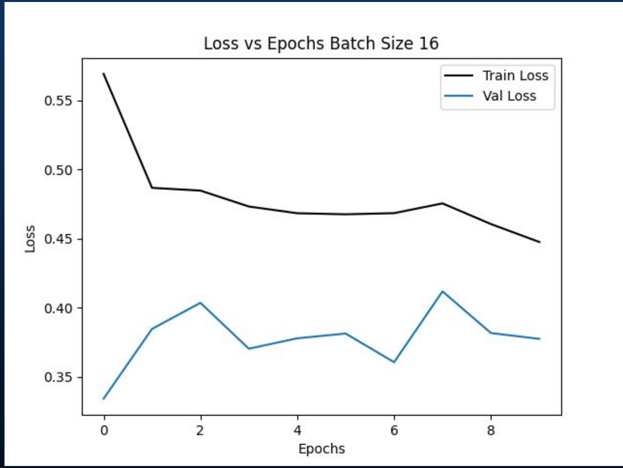

Accuracy vs Epochs Batch Size 8
Accuracy
Epochs
Train Accuracy
Val Accuracy
0.892
0.890
0.888
0.886
0.884
0.882
0
2
4
6
8

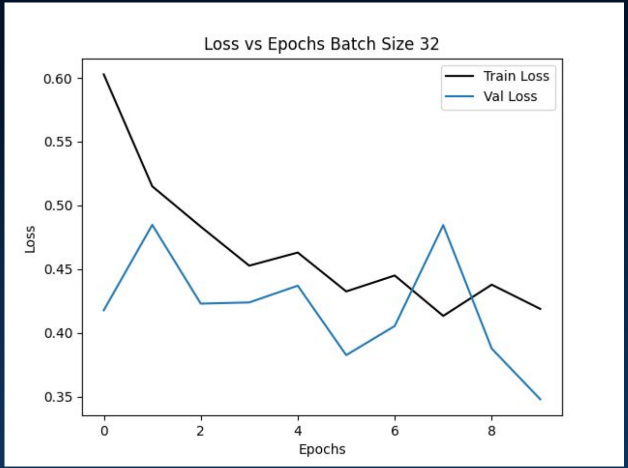

Accuracy vs Epochs Batch Size 16
Accuracy
Epochs
Train Accuracy
Val Accuracy
0.91
0.90
0.89
0.88
0.87
0.86
0
2
4
6
8

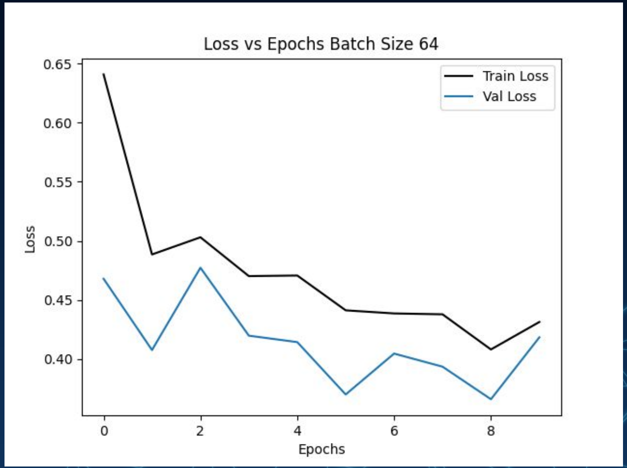

Accuracy vs Epochs Batch Size 32
Accuracy
Epochs
Train Accuracy
Val Accuracy
0.890
0.885
0.880
0.875
0.870
0.865
0.860
0.855
0
2
4
6
8

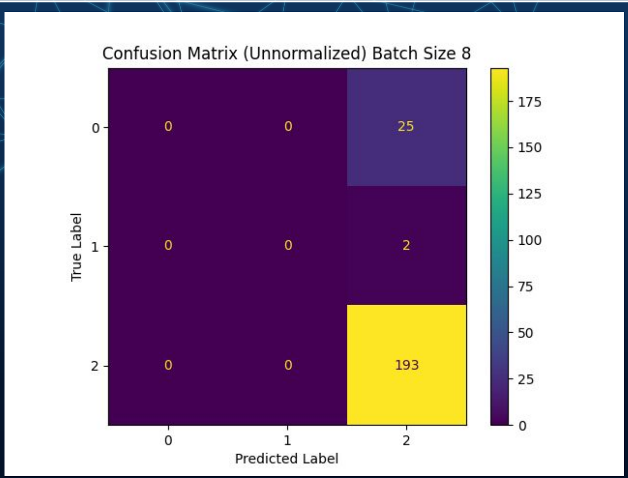

Accuracy vs Epochs Batch Size 64
Accuracy
Epochs
Train Accuracy
Val Accuracy
0.89
0.88
0.87
0.86
0.85
0
2
4
6
8

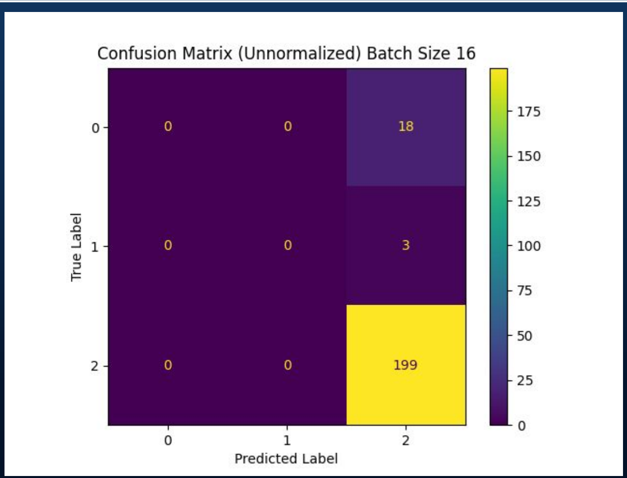

Loss vs Epochs Batch Size 8
Train Loss
Val Loss
Loss
Epochs
0.52
0.50
0.48
0.46
0.44
0.42
0
2
4
6
8

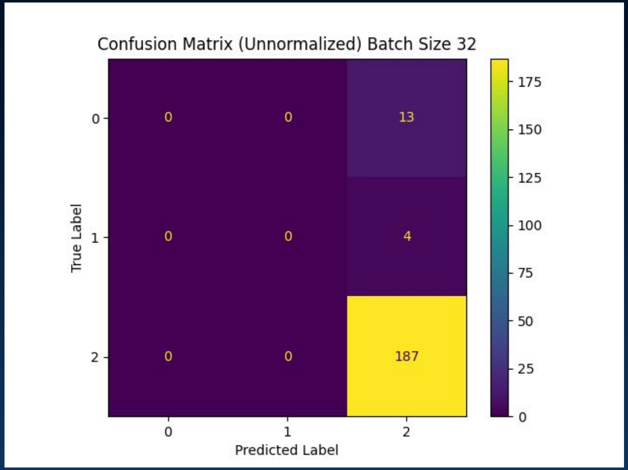

Loss vs Epochs Batch Size 16
Train Loss
Val Loss
Loss
Epochs
0.55
0.50
0.45
0.40
0.35
0
2
4
6
8

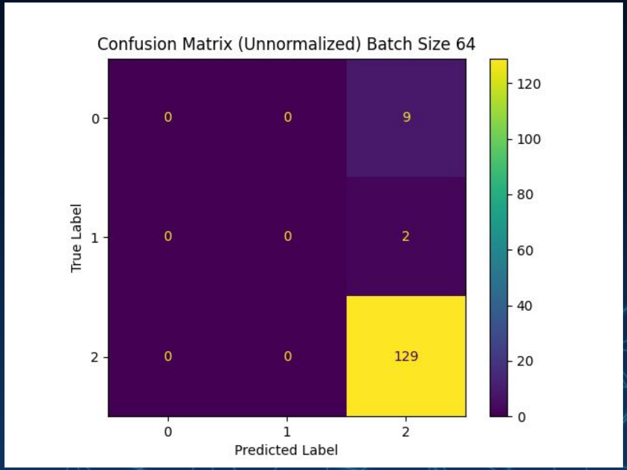

Loss vs Epochs Batch Size 32
Train Loss
Val Loss
Loss
Epochs
0.60
0.55
0.50
0.45
0.40
0.35
0
2
4
6
8


Loss vs Epochs Batch Size 64
Train Loss
Val Loss
Loss
Epochs
0.65
0.60
0.55
0.50
0.45
0.40
0
2
4
6
8


Confusion Matrix (Unnormalized) Batch Size 8
True Label
Predicted Label
0 0 25
0 0 2
0 0 193
0 1 2
0 25 50 75 100 125 150 175


Confusion Matrix (Unnormalized) Batch Size 16
True Label
Predicted Label
0 0 18
0 0 3
0 0 199
0 1 2
0 25 50 75 100 125 150 175


Confusion Matrix (Unnormalized) Batch Size 32
True Label
Predicted Label
0 0 13
0 0 4
0 0 187
0 1 2
0 25 50 75 100 125 150 175


Confusion Matrix (Unnormalized) Batch Size 64
True Label
Predicted Label
0 0 9
0 0 2
0 0 129
0 1 2
0 20 40 60 80 100 120

Results

- The accuracy of the CNN for the different batch sizes are around 90% for the test data
- However, when looking at the confusion matrix of the different batch sizes, the CNN classified all the galaxies it saw as Spiral Galaxies
- This means that the CNN either overfitted the data or it was due to not having an even distribution of each galaxy class
- To answer my research question: Yes, CNN's can classify galaxies, however, it needs an even distribution of galaxy types for it to accurately give results.

05

Next Steps & Discussion

Future research, strengths/weaknesses and improvements.

Next Steps & Discussion

- Some problems that I ran into when training my data was it took alot of computational power from my PC's CPU
- I tried to use Nvidia's Cuda toolkit to try and use my GPU instead but this didn't work because the program wouldn't open
- Improvements for my dataset could be adding in more known elliptical and irregular galaxies so that there is an even distribution of them
- Another Improvement could be creating a more complex Convolutional Neural Network with multiple convolution layers

Literature review

- Baillard, A., Bertin, E., de Lapparent, V., Fouqué, P., Arnouts, S., Mellier, Y., Pelló, R., Leborgne, J.-F., Prugniel, P., Makarov, D., Makarova, L., McCracken, H. J., Bijaoui, A., & Tasca, L. (2011). *The EFIGI catalogue of 4458 nearby galaxies with detailed morphology*. https://doi.org/10.48550/ARXIV.1103.5734
- Khalifa, N. E. M., Taha, M. H. N., Hassanien, A. E., & Selim, I. M. (2017). *Deep galaxy: Classification of galaxies based on deep convolutional neural networks*. arXiv. https://doi.org/10.48550/ARXIV.1709.02245
- Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. (2008). Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey★. *Monthly Notices of the Royal Astronomical Society, 389*(3), 1179–1189. https://doi.org/10.1111/j.1365-2966.2008.13689.x

Thanks You

```
174/174 [==============================] - 15s 85ms/step - loss: 0.4360 - sparse_categorical_accuracy: 0.8800
29/29 [==============================] - 3s 82ms/step - loss: 0.4067 - sparse_categorical_accuracy: 0.9123
Train accuracy = 88.0%
Test accuracy = 91.2%
```

```
87/87 [==============================] - 13s 150ms/step - loss: 0.4156 - sparse_categorical_accuracy: 0.8757
14/14 [==============================] - 2s 138ms/step - loss: 0.3743 - sparse_categorical_accuracy: 0.9045
Train accuracy = 87.6%
Test accuracy = 90.5%
```

```
44/44 [==============================] - 13s 292ms/step - loss: 0.3557 - sparse_categorical_accuracy: 0.8857
7/7 [==============================] - 2s 257ms/step - loss: 0.3123 - sparse_categorical_accuracy: 0.9069
Train accuracy = 88.6%
Test accuracy = 90.7%
```

```
loss_train, acc_train = model1.evaluate(train)
loss_test, acc_test = model1.evaluate(test)

print(f'Train accuracy = {acc_train:.1%}')
print(f'Test accuracy = {acc_test:.1%}')
```

```
Total params: 15,878,595
Trainable params: 15,878,595
Non-trainable params: 0
```