

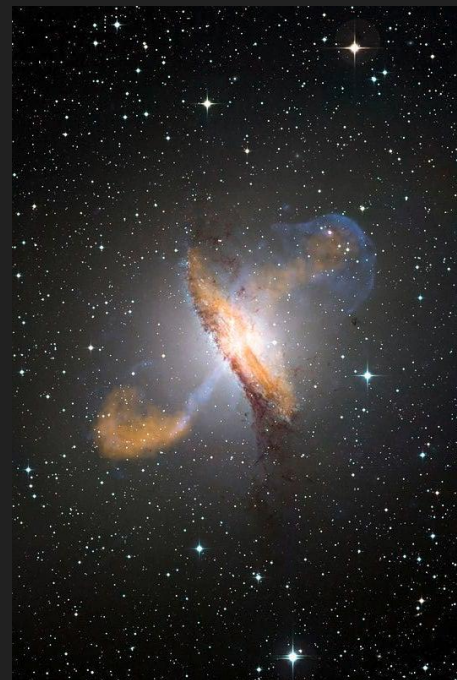
Galaxy Classification



Hanna Born, George Sangiolo, Yiming Yu
ANLY 590 Fall 2021

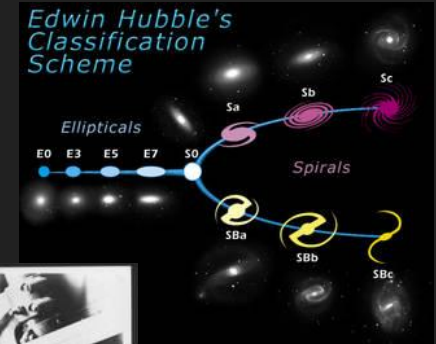
Overview

- Motivation
- Literature Review
- Data
- Models
 - Baseline Model: Logistic Regression
 - DFF
 - CNN
- Web Application
- Questions?



Motivation

- Andromeda “Nebula” → Galaxy
- Galaxies beyond our own
- Hubble Classification Scheme
- Galaxy classification important for studying and making sense of galaxy formation and evolution of the universe
 - Reimagined our place in the universe
 - Universe expanding as well



Literature Review

Deep Galaxy: Classification of Galaxies based on Deep Convolutional Neural Networks

Nour Eldeen M. Khalifa^{1,*}, Mohamed Hamed N. Taha^{1,*}, Aboul Ella Hassanien^{1,*}, I. M. Selim^{2,3}

- Presented a deep convolutional neural network architecture for galaxy classification
 - 8 layers
 - 1 main convolutional layer for features extraction with 96 filters
- 97% test accuracy
- Only using 3 labels
 - Elliptical, Spiral, Irregular

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 <http://arxiv.org/abs/1709.02245>

3 classes

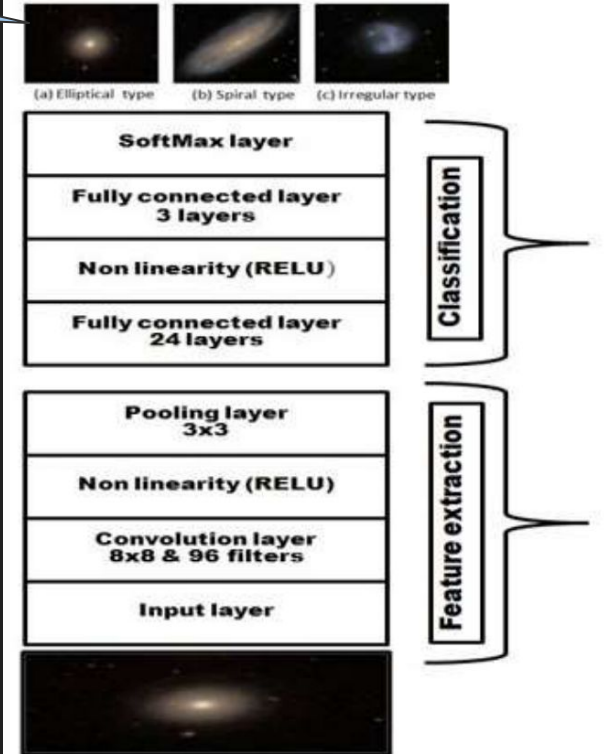


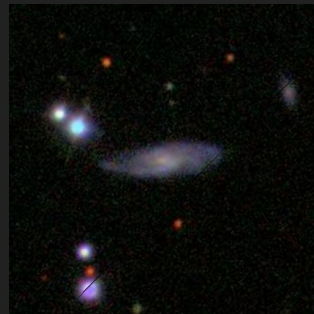
Fig. 2. Detailed layers description for the proposed deep CNN architecture

The Data

Kaggle: Galaxy Zoo - The Galaxy Challenge

Classify the morphologies of distant galaxies in our Universe

- **images_training:**
 - JPG images of **61578** galaxies
 - Files named according to GalaxyId
- **solutions_training:**
 - Probability distributions for the classifications for each of the training images
 - 37 dimensions
- **images_test:**
 - JPG images of **79975** galaxies
 - Files are name according to their GalaxyId



Data Labeling

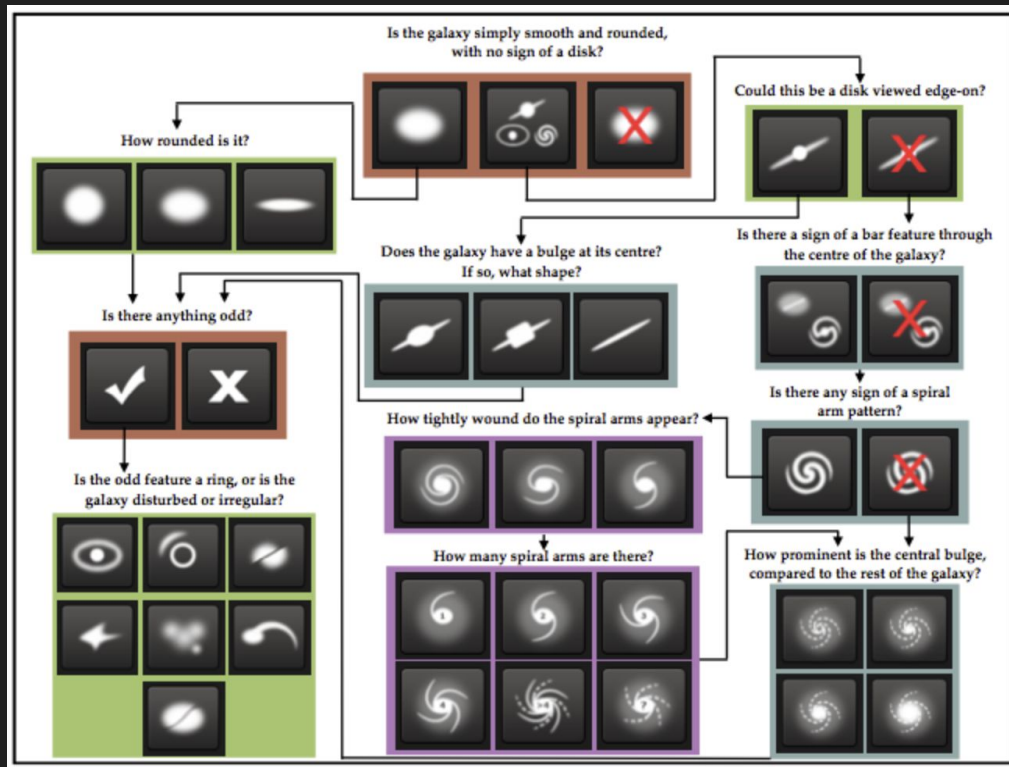
Initial
Question

Task	Question	Responses	Next
01	<i>Is the galaxy simply smooth and rounded, with no sign of a disk?</i>	smooth features or disk star or artifact	07 02 end
02	<i>Could this be a disk viewed edge-on?</i>	yes no	09 03
03	<i>Is there a sign of a bar feature through the centre of the galaxy?</i>	yes no	04 04
04	<i>Is there any sign of a spiral arm pattern?</i>	yes no	10 05
05	<i>How prominent is the central bulge, compared with the rest of the galaxy?</i>	no bulge just noticeable obvious dominant	06 06 06 06
06	<i>Is there anything odd?</i>	yes no	08 end
07	<i>How rounded is it?</i>	completely round in between cigar-shaped	06 06 06
08	<i>Is the odd feature a ring, or is the galaxy disturbed or irregular?</i>	ring lens or arc disturbed irregular other merger dust lane	end end end end end end end
09	<i>Does the galaxy have a bulge at its centre? If so, what shape?</i>	rounded boxy no bulge	06 06 06
10	<i>How tightly wound do the spiral arms appear?</i>	tight medium loose	11 11 11
11	<i>How many spiral arms are there?</i>	1 2 3 4 more than four can't tell	05 05 05 05 05 05

Non-linear path
through questions
depending on
answer to current
question

Data Classification

- 37 classes
- at each node (question), total initial probability of a classification will sum to 1.0
 - initial probabilities are then weighted
 - weighting emphasizes that a good solution must get the broad categories correct but best solutions also perform well for detailed categories further down decision tree



Baseline Model: Logistic Regression

- Single-class problem (using `argmax()` of training labels)
 - Not all 37 classes present in single-class problem
 - 5 distinct classes in training data
 - Represent most definitive/distinguishable characteristics in image decision tree
- Score on test data = **0.562**
 - LogReg performs better than random guess (>0.20)

Feed Forward Network

- Trained with KerasTuner
 - Automatically explores different combinations of hyperparameters to find the best-performing model
- Trained a Multilabel Classification with Threshold = 0.5
- Used ImageGenerator to increase training data by ~20%
- Test Accuracy: 93.6%
- Unfortunately, history plot got lost in space...a downside to KerasTuner is that it eats a lot of disk space!

```
Epoch 10/10
762/762 [=====] - 4s 5ms/sample - loss: 0.6913 - acc: 0.9214 - val_loss: 0.6912 - val_acc: 0.9210
Trial 30 Complete [00h 00m 37s]
acc: 0.9213662147521973
```

```
Best acc So Far: 0.9690713286399841
Total elapsed time: 00h 13m 31s
```

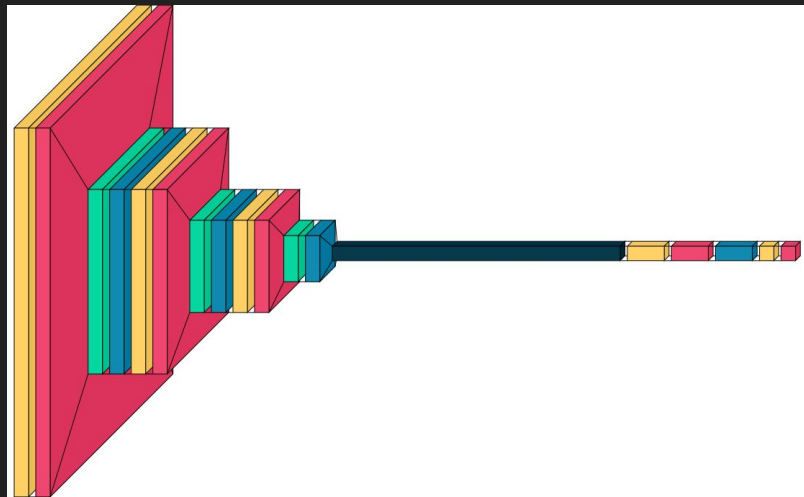
```
The hyperparameter search is complete. The optimal number of units in the first densely-connected layer is 288 and the optimal learning rate for the optimizer is 0.001.
```

```
Epoch 45/45
32/762 [>.....] - ETA: 5s - loss: 0.0000e+00 - acc: 1.0
64/762 [=>.....] - ETA: 4s - loss: 0.0307 - acc: 0.9996
762/762 [=====] - 6s 8ms/sample - loss: 0.8964 - acc: 0.9975 - val_loss: 132.4939 - val_acc: 0.9328
160/239 [=====>.....] - ETA: 0s - loss: 107.5755 - acc: 0.935
192/239 [=====>.....] - ETA: 0s - loss: 112.5924 - acc: 0.933
224/239 [=====>.....] - ETA: 0s - loss: 110.2219 - acc: 0.935
239/239 [=====] - 1s 4ms/sample - loss: 108.3694 - acc: 0.9358
[test loss, test accuracy]: [108.36942093342418, 0.9357684]
```

Convolutional Neural Network

CNN Baseline:

- 3 Convolutional layers

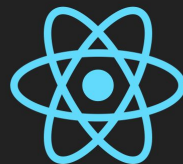


Xception (Extreme Inception)

- Outperform Inception V3 with the same amount of parameters
- Backbone of MobileNet for device deployment
- Same preprocessing and tuning setup as the feed forward network

Web Application

- Models are best when they can be seen!
- Deployment Considerations:
 - Ease – only a few weeks to do the project!
 - Accessibility – several “canned” user functions for limited access
 - Size – ML model files tend to be too big for Lambda functions
 - Cost – cloud resources are \$\$\$
- Stack
 - AWS S3 Stores Models and Images
 - FastAPI Backend
 - React.js Frontend
- Demo!



References

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 <http://arxiv.org/abs/1709.02245>
<https://www.nasa.gov/content/about-story-edwin-hubble>