Classificação Morfológica de Galáxias Por Meio de Redes Neurais

Matheus G. Silva[1], **Thiago M. Ventura**[1]

[1]Instituto de Computação – Universidade Federal do Mato Grosso (UFMT)
Cuiabá – MT – Brasil

matheusgambati@gmail.com, thiago@ic.ufmt.br

Abstract. *This paper proposes the development of a convolutional neural network for the morphological classification of galaxies through optical images, classifying them into six distinct classes based on the Hubble Tuning Fork model. In order to automate identification and separation of the huge volume of data generated in recent astronomical observatories, deep learning and data augmentation techniques are used to generate increased data variation and consequently improve network accuracy. Our model achieved an average precision of 88%.*

Resumo. *Esse artigo propõe o desenvolvimento de uma rede neural convolucional para a classificação morfológica de galáxias por meio de imagens ópticas, classificando-as em seis classes distintas baseadas no modelo do Hubble Tuning Fork. Com o objetivo de automatizar a identificação e separação do imenso volume de dados gerados nos observatórios astronômicos recentes, são utilizadas técnicas de deep learning e data augmentation para gerar um aumento da variação dos dados e consequentemente melhorar a precisão da rede. O modelo criado obteve uma precisão média de 88%.*

1. Introdução

Com o progresso na última década na astronomia, um enorme volume de dados passou a ser coletado. Equipamentos como telescópios, satélites e sensores têm produzido um volume de dados sem precedentes na historia da astronomia [Bell et al. 2009].

Existem diferentes métodos de classificação. Alguns focam em detectar colisões de galaxias e pertubações gravitacionais, enquanto outros focam na classificação generalista. A Classificação manual de galáxias é uma prática ainda comum. Como exemplo disso nota-se iniciativas de comunidades como Galaxy Zoo [Lintott et al. 2008] que permitiram a classificação de mais de 2,7 milhões de galáxias de forma manual, no entanto, isto ainda não é o suficiente para dar conta do gigantesco fluxo de dados.

Afim de contornar esta limitação, técnicas de classificação automatizada através do uso de inteligencia artificial passaram a ser estudadas a partir da década de 80 [O. Lahav 1996], porém, somente com os avanço no poder de processamento computacional foi que ela passou a ser adotada [Lintott 2010].

O objetivo deste trabalho consiste na classificação generalista de galáxias em 6 classes morfológicas distintas baseado em uma leve variação do modelo clássico conhecido como *Hubble Tuning Fork*, para que seja possível a sua aplicação em qualquer conjunto de dados com imagens ópticas.

2. Modelo de Classificação Morfológica

Para se classificar galáxias por seu tipo morfológico, é necessário a utilização de uma metodologia, como a proposta por Edwin Hubble que por meio da análise estatística de 400 galáxias a sequência de classificação Hubble (*Hubble Sequence*) [Hubble 1926] se tornou um modelo amplamente usado por causa sua simplicidade. Com o tempo sofreu uma série de modificações [den Bergh 1998], no qual veio a ser conhecido como *Hubble Tuning Fork*.

Nesse modelo é possível separar as galáxias em 5 tipos primários: elípticas, lenticulares, espirais, espirais barradas e irregulares. O modelo prevê uma sequência cronológica da evolução morfológica começando pelas elípticas até a formação espiral com classes intermediárias entre cada classe primária, conforme mostrada na Figura 1.

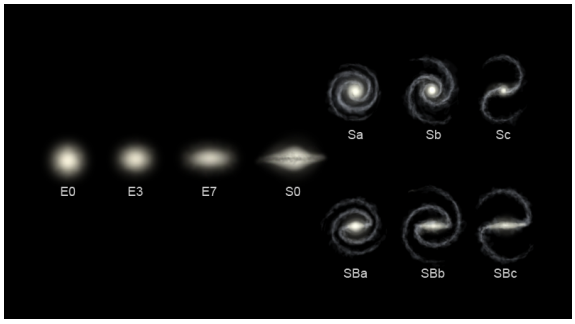

Figura 1. Exemplo de classificação baseado no Hubble Tuning Fork

Boa parte das classificações ainda são feitas de forma manual, sendo um processo lento e dependente de voluntários. Projetos como *Galaxy Zoo* tornou popular e acessível a participação de voluntários na contribuição das classificações, gerando um grande avanço no volume de dados para pesquisa. O enorme sucesso do projeto se dá, não pelas classificações manuais em si, mas sim pelo desenvolvimento de ferramentas automatizadas de classificação que somente são possíveis a partir dos dados gerados manualmente.

2.1. Trabalhos Correlatos

Devido a demanda, há trabalhos com propostas para automatização. Diversas técnicas são implementadas, algumas se utilizam de diferentes fontes de dados, utilizando infravermelho, raios-x e entre outros tipos de radiação.

O trabalho de [O. Lahav 1996] analisa e compara técnicas de classificação morfológicas via redes neurais. São comparados as técnicas *Principal Component Analysis* (PCA), *Encoder Neural Network* e *Supervised ANN*. Uma descrição detalhada dos desafios da obtenção dos conjuntos de dados que na época eram publicados com 200 à 700 imagens e a baixa capacidade computacional.

Em [Wu et al. 2018] foi desenvolvido um projeto utilizando redes neurais convolucionais para realizar a classificação de Radio-galáxias via imagens geradas pela combinação ao de sinais de rádio e radiação infravermelha. É um projeto *open source*, extremamente rápido ($<$200ms por imagem) e com uma precisão aproximada de 90%.

Em [Wu et al. 2018] foi desenvolvido um projeto para realizar a classificação de Radio-galáxias via imagens geradas pela combinação de sinais de rádio e radiação infravermelha utilizando redes neurais convolucionais.

Algo semelhante foi feito em [Zhu et al. 2019] propondo a classificação morfológica por meio de redes neurais residuais profundas (*ResNets*), utilizando imagens ópticas com conjunto de treinamento de aproximadamente 25.000 imagens classificadas manualmente pelo projeto Galaxy Zoo. O conjunto de dados foi originalmente distribuído no *The Galaxy Challenge* [Kaggle 2014]. O sistema de classificação diverge levemente no *Hubble Tuning Fork*. As precisões finais foram de: *completely round*, 96,6785%; *in-between*, 94,4238%; *cigar-shaped*, 58,6207%; *on-edge*, 94,3590% e *spiral*, 97,6953% respectivamente.

Este trabalho se aproxima do estudo de [Zhu et al. 2019], utilizando a mesma base de dados, mas divergindo da arquitetura, tentando obter melhores resultados.

3. Materiais e Métodos

Nesta seção será descrito o processo de obtenção, seleção e ampliação dos dados, além das tecnologias utilizadas para a criação da rede.

3.1. Dados

Os dados utilizados foram fornecidos pelo Galaxy Zoo, a competição The Galaxy Challenge. O conjunto é composto por 61.578 imagens classificadas manualmente pelo projeto. Cada uma possuí um GalaxyId com índices de votos recebidos para cada classe morfológica.

Como esse projeto busca classificação entre os tipos básicos de morfologias, é necessário mapear cada uma das classes dentre as 37 do conjunto original, descartando as irregulares e apenas utilizando as mais básicas.

Para selecionar as imagens que serão utilizadas para treino, ordenou-se pela porcentagem de votos de cada classe. Para as classes cujo a seleção retorne itens com índice de votos abaixo de 0.5%, aplicou-se um filtro para selecionar somente as que as classes desejadas possuíssem o maior índice de votos (Tabela 1). Neste caso foram selecionadas as N primeiras imagens e separado em treino (80%) e validação (20%).

Classe	**Seleção**	**Resultado**
completely_round	range(0, 5000)	5000
in_between	range(0, 3600)	3600
cigar_shaped	cigar_shaped >in_between cigar_shaped >on_edge range(0, 1500)	1304
on_edge	range(0, 5100)	5100
spiral	has_signs_of_spiral >0.5 range(0, 3300)	3300
spiral_barred	has_signs_of_spiral >0.5 range(0,5000)	5000

Tabela 1. Tabela de seleção

Para aumentar a quantidade e a variação dos dados, foi aplicado uma série de transformações nos dados selecionados, em um processo conhecido como *data augmen-*

tation. *Data augmentation* é um método efetivo para melhorar a generalização da rede, e consequentemente melhorar a precisão [Perez and Wang 2017]. Esse processo gerou 10.000 imagens extras. Os procedimentos realizados incluem:

- Rotação: 90º, 180º ou 270º
- Brilho: entre 0,7 e 1,8
- Contraste: entre 0,7 e 1,5
- Espelhamento: vertical ou horizontal

Adicionalmente, para todas as imagens dos conjuntos de treino e validação, aplicamos um *zoom* de 1,6x para focalizar na galáxia alvo e remover artefatos sem relevância a classificação e por fim um redimensionamento para reduzir o tempo de treino.

3.2. Tecnologias

Para a construção da rede foi escolhido a linguagem Python versão 3,6. A linguagem possui um ecossistema de bibliotecas *open source* diversificado para aplicação em aprendizado de máquina e visualização de dados, como:

- Keras: biblioteca para criação de redes neurais.
- Pandas: biblioteca para análise de dados e útil para a criação dos conjuntos de treino e validação.
- Augmentor: biblioteca para processamento de imagens, sendo utilizada para *data augmentation* do conjunto de dados para treino.
- Seaborn: fornece utilidades para visualização gráficas de dados.

4. Criação do modelo

Para a arquitetura da rede foi utilizado uma série de camadas convolucionais. A rede construída inicia com 32 filtros, aumentando gradativamente até 64 filtros na última camada de convolução (Figura 2). A função de *loss* sendo *sparse categorical crossentropy* e a última ativação sendo *softmax*. Também foi aplicado *regularizers* L2 em algumas camadas convolucionais para reduzir *overfting* da rede.

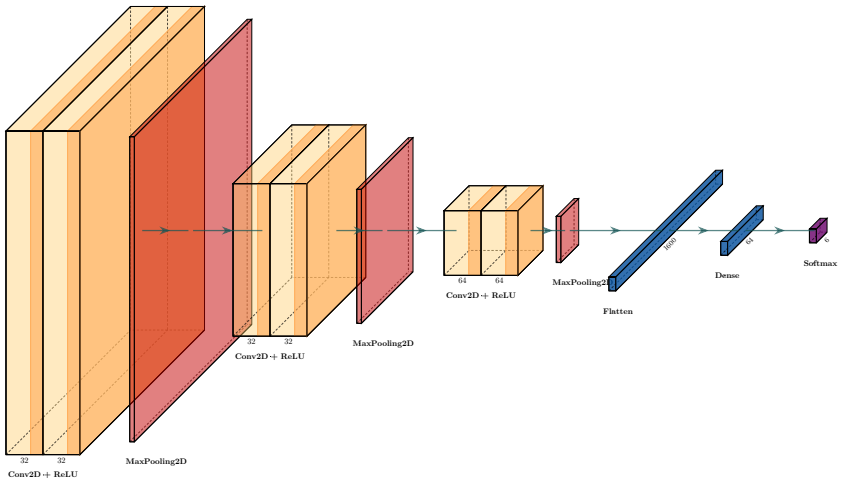

Figura 2. Arquitetura da rede

Essa arquitetura permite que a rede aprenda os atributos mantendo a generalização. Um número maior de filtros foi testado, porém a precisão no conjunto de validação caiu significativamente devido a incapacidade da rede de generalizar aspectos das imagens.

5. Resultados

Depois de realizado a preparação dos dados e o treinamento da rede com 80 épocas (2 horas de treino com uma Nvidia P100), foi possível avaliar o desempenho da rede durante a fase de treinamento. O modelo atingiu no treino uma precisão geral de 90,27%.

Por meio de uma matriz de confusão (Figura 3) pode-se analisar os dados das predições feitas. Os melhores resultados são da galáxia do tipo *Completely Round*, enquanto a *Cigar Shaped* teve maiores dificuldades de classificação.

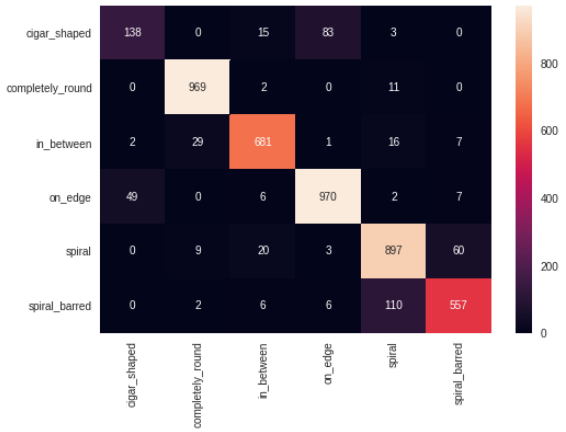

Figura 3. Precisão para classificação de cada tipo de galáxia

A Tabela 2 mostra a precisão do resultado de classificação para cada tipo de galáxia. O resultado final mostrou uma variação de 73% à 96% na classificação, tendo como média a precisão de 88%.

Classe	Precisão	F1-Score
completely_round	0,96	0,97
in_between	0,93	0,93
cigar_shaped	0,73	0,64
on_edge	0,91	0,93
spiral	0,86	0,88
spiral_barred	0,88	0,85

Tabela 2. Análise da matriz de confusão

6. Considerações finais

A rede mostrou uma boa precisão para classificar as classes de galáxias. No entando, *Cigar shaped* possui o índice de precisão mais baixo dentre as demais. Tais resultados podem ser explicados por causa de um desbalanceamento dos dados. A classe *Cigar*

shaped é a que possui um índice baixo de votos, e as pessoas podem se confundir com *On Edge*.

Como trabalhos futuros é possível melhorar o modelo com determinadas estratégias, como: aumentar o número de dados de forma a balancear o conjunto de dados por classe, refinar os hyper-parâmetros e utilizar outros algoritmos otimizadores. De forma geral, o Galaxy Zoo continua ativo e produzindo novas classificações, podendo futuramente utilizá-los para uma melhor a precisão do modelo.

Referências

Bell, C., Hey, T., and Szalay, A. (2009). Beyond the data deluge (computer science). *Science (New York, N.Y.)*, 323:1297–8.

den Bergh, V. (1998). *Galaxy Morphology and Classification*. University of Victoria, British Columbia.

Hubble, E. P. (1926). Extragalactic nebulae. *Astrophysical Journal*.

Kaggle (2014). Galaxy zoo - the galaxy challenge. https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge. Acessado: 2019-09-16.

Lintott, Schawinski, B. S. L. T. E. M. R. C. N. M. J. R. A. S. D. A. P. M. J. V. (2010). Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410:166–178.

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., and 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*, 389:1179–1189.

O. Lahav, A. Nairn, L. S. M. C. S.-L. (1996). Neural computation as a tool for galaxy classification: methods and examples. *Monthly Notices of the Royal Astronomical Society*, 283:207–221.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *CoRR*, abs/1712.04621.

Wu, C., Wong, O. I., Rudnick, L., Shabala, S. S., Alger, M. J., Banfield, J. K., Ong, C. S., White, S. V., Garon, A. F., Norris, R. P., Andernach, H., Tate, J., Lukic, V., Tang, H., Schawinski, K., and Diakogiannis, F. I. (2018). Radio Galaxy Zoo: Claran – a deep learning classifier for radio morphologies. *Monthly Notices of the Royal Astronomical Society*, 482(1):1211–1230.

Zhu, X.-P., Dai, J.-M., Bian, C.-J., Chen, Y., Chen, S., and Hu, C. (2019). Galaxy morphology classification with deep convolutional neural networks. *Astrophysics and Space Science*, 364(4):55.