DMAGNet: an Interpretable Convolutional Neural Network for Galaxy Morphology Classification using Deep Dream

Chessa Giovanni
giovanni.chessa.1@studenti.unipd.it

Murru Alessandro
alessandro.murru.1@studenti.unipd.it

Lazzari Tommaso
tommaso.lazzari.1@studenti.unipd.it

Abstract

The detection, measurement and classification of galaxy morphological parameters is a key element to study their formation and evolution. Large public surveys, such as the Sloan Digital Sky Survey (SDSS), have made extensive galaxy image datasets available to the scientific community, enabling the development of computer vision approaches for automatic galaxy morphology classification, a task traditionally performed through expert visual inspection.
In this work, we present a convolutional deep neural network trained to distinguish between ten morphological classes from the public Galaxy10 DECaLS dataset. To investigate the representations learned by the model, we apply the Google Deep Dream algorithm to different network layers, providing visual interpretations of the features driving the final classification. The resulting visualizations revealed class-specific morphological structures, such as spiral arms or bulges, as well as finer patterns that may be difficult to discern through direct visual inspection. These results suggest that the network captures meaningful and potentially astronomically relevant features, offering both competitive classification performance and improved model interpretability.

1. Introduction

Galaxies exhibit a wide variety of shapes, colours and sizes. These properties play a fundamental role in astronomical studies, as they provide valuable information about galaxy age, formation history, and interactions with other galaxies over their lifetimes. The study of galaxy formation and evolution relies heavily on morphological analysis to infer the physical processes that drive these phenomena. In particular, large astronomical surveys, are essential for uncovering complex relationships between galaxy properties, such as metallicity, age, environment, and morphology.

Such studies require vast numbers of observations and reliable morphological classifications. Large-scale surveys, such as the Sloan Digital Sky Survey (SDSS), have made millions of galaxy images publicly available. However, manual inspection and classification of such datasets are impractical when performed solely by domain experts. Early attempts to develop automated or semi-automated classification systems struggled to meet scientific accuracy requirements, motivating the creation of the Galaxy Zoo project. Galaxy Zoo enabled the morphological classification of nearly 900,000 galaxies by leveraging public participation through an online platform.

Following the success of Galaxy Zoo, additional datasets were released, incorporating images acquired by new-generation ground-based and space-based telescopes. The increasing availability of large, high-quality datasets has fostered the development of machine learning and deep learning approaches, which have achieved excellent performance in galaxy morphology classification tasks.

In this work, we propose a Convolutional Neural Network, DMAGNet, designed for galaxy morphology classification. The model follows a hierarchical architecture that progressively extracts low-level visual features and higher-level structural patterns. Multi-scale convolutional operations allow the network to capture morphological structures of varying spatial extent such as extended spiral arms, smaller central bulges as well as barely noticeable galactic tides or halos. In addition, attention mechanisms are employed to emphasize the most informative regions of the images, improving the discrimination between visually similar morphological classes.

Furthermore, we investigate the interpretability of the model through the application of the Google Deep Dream algorithm. Deep Dream is a computer vision program that uses a convolutional neural network to find and enhance pat-

terns. The optimization resembles backpropagation; however, instead of adjusting the network weights, the weights are held fixed and the input is adjusted. Thanks to this method, once trained, a neural network can also be run in reverse, being asked to adjust the original image slightly so that a given output neuron (e.g. the one for bulges or spiral arms) yields a higher confidence score. In this work, Deep Dream is employed as a qualitative tool to analyze the morphological structures that drive the network's classification decisions.

2. Related Work

One of the first systematic galaxy classification schemes was introduced by Hubble (1926 [6], 1936 [7]), dividing the galaxies into two broad types: galaxies with a dominant bulge component (also known as early-type galaxies, ETGs) and galaxies with a significant disc component (late-type or spiral galaxies). The spiral galaxies are further divided into barred (with the presence of a bar shaped central structure) or unbarred and ordered according to their spiral arms strength. The intermediate class between elliptical and spiral galaxies is referred to as S0, while there is also a population of galaxies with irregular or distorted shapes. According to this visual classification, a number can be assigned to each type of galaxy, which is known as the T-type (de Vaucouleurs 1963 [3]).

When extremely large datasets became available, an effective approach to address this limitation of visual classification for large amounts of data was the Galaxy Zoo project (Lintott et al. 2008 [12]), where 'science citizens' volunteered to classify galaxies through a user-friendly web interface. The first approach to the task of galaxy morphology classification was a very simple classification into three types (ETGs, spirals or mergers) but, given the success of the project, a more complex classification system, GalaxyZoo2, was proposed in Willett et al. (2013) [15]. However, galaxy classifications performed by amateur astronomers, a task that is challenging even for professionals, present several limitations. For example, features such as bars are only selected when the bar is obvious and the volunteers tend to choose intermediate options when available. There are also a large number of galaxies with uncertain classifications caused by the disagreement between classifiers.

Automated classifications using a set of parameters that correlate with morphologies (e.g. concentrations, clumpiness, asymmetries, Gini coefficients, etc.) have also been attempted (Abraham et al. 1996 [1]; Conselice, Bershady & Jangren 2000 [2]; Lotz et al. 2008 [13]). A generalization of that approach, using an n-dimensional classification with optimal non-linear boundaries in the parameter space, was proposed in Huertas-Company et al. (2011) [8].

A natural step forward is to take advantage of the recently popular Deep Learning algorithms. CNNs have been proven very successful in the last years for many different image recognition purposes. CNNs have also been used for morphological classification of galaxies, with a high success rate. For example, Huertas Company et al. (2015) [9] applied CNNs to classify 50,000 CANDELS (Grogin et al. 2011 [5]; Koekemoer et al. 2011 [10]) galaxies into five groups (spheroid, disc, irregular, point source, and unclassifiable). They obtained zero bias, approximately 10% scatter and less than 1% misclassification. The CNN model presented in Dieleman, Willett & Dambre (2015)[4], was able to reproduce the GZ2 classification with large accuracy for galaxies with certain classifications.

3. Dataset

The Galaxy10 dataset was originally derived from Galaxy Zoo Data Release 2, in which volunteers visually classified approximately 270,000 RGB images of galaxies from the Sloan Digital Sky Survey (SDSS). From this collection, around 22,000 images were selected and grouped into ten broad morphological classes based on consensus volunteer votes. Subsequently, Galaxy Zoo incorporated higher-resolution imaging from the DESI Legacy Imaging Surveys (DECaLS), significantly improving image quality and spatial resolution.

The resulting Galaxy10 DECaLS [11] dataset combines Galaxy Zoo morphological labels with DECaLS images from multiple observing campaigns, yielding approximately 441,000 unique galaxies covered by DECaLS. From this larger sample, approximately 18,000 RGB images were selected and assigned to ten refined morphological classes using more stringent filtering criteria. Compared to the original Galaxy10 dataset, the class definitions were adjusted to improve inter-class separability, and the "Edge-on Disk with Boxy Bulge" category, originally containing only 17 samples, was excluded.

The final dataset consists of 17,736 images distributed across the following classes: Disturbed (1,081), Merging (1,853), Round Smooth (2,645), In-between Round Smooth (2,027), Cigar-Shaped Smooth (334), Barred Spiral (2,043), Unbarred Tight Spiral (1,829), Unbarred Loose Spiral (2,628), Edge-on without Bulge (1,423), and Edge-on with Bulge (1,873).

For training, all images were kept at their original size of 256×256 pixels and normalized. Data augmentation was applied through random rotations sampled uniformly

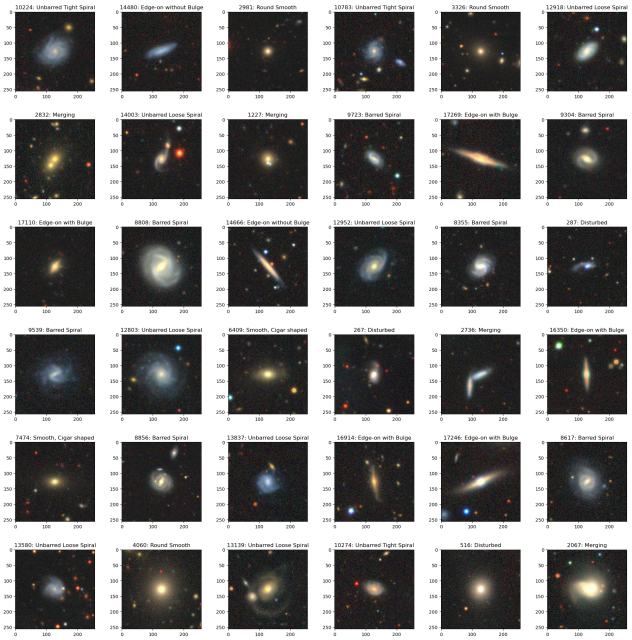

Figure 1. Random galaxy images from the Galaxy10 DECaLS dataset.

in $[-180°, 180°]$, random horizontal and vertical flips with probability 0.5, and mild random perturbations of brightness and contrast (±10 %). These augmentations promote invariance to orientation and reflection and improve robustness to photometric variations commonly encountered in astronomical imaging. A random sample of the Galaxy10 DECaLS images is provided in Figure 1.

4. DMAGNet

DMAGNet is a deep convolutional network specifically designed to classify galaxies extracting rich multi-scale spatial features from astronomical images. The architecture (Figure 2) is made of four main components, arranged into nine layers. An initial STEM stage for early spatial reduction, useful to contain the computational cost of the network training, followed by three layers of residual DMA blocks, a local attention refinement module and a final light classification head. Starting from RGB images of size 256×256, the network progressively reduces spatial resolution ($256 \times 256 \rightarrow 64 \times 64 \rightarrow 32 \times 32 \rightarrow 16 \times 16 \rightarrow 8 \times 8$) while increasing channel capacity ($3 \rightarrow 64 \rightarrow 128 \rightarrow 256 \rightarrow 512 \rightarrow 1024$).

The main unit, the DMA blocks, integrates three complementary mechanisms, able to generate multi-receptive-field features, enriches spatial and channel interactions and finally decides which channel are actually useful:

- a Dilated Large Kernel (DLK) module that adaptively combines small, medium, and large receptive fields. Extracts feature at different effective receptive fields (small/medium/large), then let the network learn how much to use each per channel.
- a Multi-Scale Feed-Forward Network (MS-FFN) that enhances feature representations by expanding channel capacity, performing parallel depthwise convolutions at multiple spatial scales, and fusing the resulting information through a bottleneck projection with residual learning.
- Attention Feature Fusion (AFF) that is a channel-wise attention module that learns which feature channels are important and reweights them without changing spatial resolution, while preserving information via a residual connection.

Stage	Description	Output Shape	Params
Stem	Initial conv + pooling	$64 \times 128 \times 128$	9,536
DMA Layer 1	$2 \times$ DMA Block	$128 \times 64 \times 64$	768,704
DMA Layer 2	$1 \times$ DMA Block	$256 \times 32 \times 32$	1,375,040
DMA Layer 3	$2 \times$ DMA Block	$512 \times 16 \times 16$	11,922,176
LAM	Local Attention Module	$1024 \times 16 \times 16$	7,343,104
Head	Conv + Att. Pool + FC	10	2,402,251
Total			**23,820,811**

Table 1. Architectural summary of DMAGNet.

After the three DMA blocks, the Local Attention Module (LAM) performs spatially adaptive feature refinement by learning a dense attention map over local regions of the feature tensor, enabling the network to emphasize morphology-relevant structures while suppressing background noise prior to global aggregation.

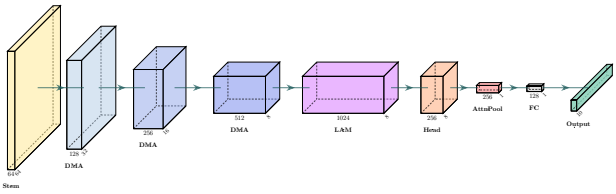

Figure 2. DMAGNet backbone structure

As the last block, the classification head consists of a channel reduction convolution followed by an attention-based global pooling mechanism and a lightweight fully connected classifier, enabling robust aggregation of spatially localized morphology cues into a compact representation suitable for final galaxy classification.

4.1. Galaxy Morphology Classification

The performance of the proposed DMAGNet architecture was evaluated on the Galaxy10 DECaLS dataset, which was partitioned into 11,487 augmented training images, 1,276 clean validation images, and a held-out test set of 3,194 images. On the test set, DMAGNet

achieves an overall classification accuracy of 87%, with a weighted F1-score of 0.86, indicating robust performance across morphological classes despite the pronounced class imbalance present in the dataset.

High precision and recall are obtained for well-defined galaxy morphologies, such as In-between Round Smooth, Cigar-shaped Smooth, and Unbarred Tight Spiral galaxies, all of which reach F1-scores above 0.92 (Table 3). This suggests that the network effectively captures both global structural properties and fine-grained morphological features. In contrast, more ambiguous classes, including Merging and Round Smooth galaxies, exhibit lower recall values, reflecting intrinsic visual overlap with disturbed or edge-on systems.

True \ Pred	D	M	RS	IRS	CS	BS	UTS	ULS	EWB	EWB$^+$
Disturbed	313	3	5	2	1	7	1	1	23	12
Merging	0	51	1	0	3	0	1	2	1	1
Round Smooth	10	3	112	2	2	2	7	14	38	5
In-between Round Smooth	2	0	2	310	13	2	1	1	5	1
Cigar-shaped Smooth	2	2	1	3	241	2	0	0	3	2
Barred Spiral	5	1	2	2	4	304	6	1	8	1
Unbarred Tight Spiral	4	0	4	0	0	4	451	0	0	13
Unbarred Loose Spiral	1	5	3	1	0	2	0	349	3	1
Edge-on without Bulge	22	1	20	5	3	11	6	4	363	38
Edge-on with Bulge	7	1	0	0	1	2	7	9	34	269

Table 2. Confusion matrix of DMAGNet predictions on the Galaxy10 DECaLS test set.

The confusion matrix (Table 2) reveals that misclassifications predominantly occur between morphologically related classes, most notably between edge-on disks with and without bulges and between round smooth galaxies and edge-on configurations. These confusion patterns are astrophysically plausible and consistent with known degeneracies in visual galaxy classification. Overall, the results indicate that DMAGNet learns discriminative and physically meaningful representations, providing reliable performance on a challenging multi-class galaxy morphology classification task.

Class	**Prec.**	**Rec.**	**F1**	**Sup.**
Disturbed	0.86	0.85	0.85	368
Merging	0.76	0.85	0.80	60
Round Smooth	0.75	0.57	0.65	195
In-between Round Smooth	0.95	0.92	0.94	337
Cigar-shaped Smooth	0.90	0.94	0.92	256
Barred Spiral	0.90	0.91	0.91	334
Unbarred Tight Spiral	0.94	0.95	0.94	476
Unbarred Loose Spiral	0.92	0.96	0.94	365
Edge-on without Bulge	0.76	0.77	0.76	473
Edge-on with Bulge	0.78	0.82	0.80	330
Accuracy			**0.87**	3194
Macro avg.	0.85	0.85	0.85	3194
Weighted avg.	0.86	0.87	0.86	3194

Table 3. Classification performance of DMAGNet on the Galaxy10 DECaLS test set.

5. Google Deep Dream

One of the challenges of neural networks is understanding what exactly goes on at each layer. We know that after training, each layer progressively extracts higher and higher-level features of the image, until the final layer essentially makes a decision on what the image shows. For example, the first layer maybe looks for edges or corners. Intermediate layers interpret the basic features to look for overall shapes or components. The final few layers assemble those into complete interpretations.

Google Deep Dream [14] is a visualization technique designed to probe and interpret the internal representations learned by convolutional neural networks. The key idea is that the image is no longer just data being analyzed. It becomes the thing to be optimized. The image is treated as a variable whose pixel values can change. When the image is passed through network, each layer produces internal features responses that represent *what the network sees* at that level. Google Deep Dream algorithm aims to answer the following question:

How would the pixels of a given image need to change so that this layer responds more strongly?

The method operates by fixing the network parameters and iteratively modifying the input image. Measuring how intense the activations of the chosen layer are and computing how sensitive those activations are to each individual pixel, a gradient is obtained for each pixel, telling, via gradient ascent, whether to increase or decrease its value in order to maximize the activation of the chosen layer. When this process is repeated many times, small structures that excite the network start to appear. This is the so called *"dreaming"* effect.

To prevent raw gradients to be unstable and noisy, they are smoothed spatially, spreading their influence over neighboring pixels. The gradients are also normalized so that each update has a controlled magnitude, preventing the image from blowing up or collapsing.

If the image were dreamed at a single resolution, the results would tend to be either overly repetitive or overly local. To avoid this, the algorithm works across multiple spatial scales. It starts by dreaming on a smaller version of the image. At low resolution, each pixel represents a larger area of the image, so any change affects broader regions. This encourages the emergence of large, global structure and repeating motifs that span wide ares. The image is then gradually increased in resolution and dreamed again. At higher resolutions, the algorithm can add finer details on top of the existing structure.

5.1. DMAGNet: Feature Visualization

In this work, Google Deep Dream is employed as a qualitative interpretability tool to analyze the representations learned by DMAGNet at different depths. In our framework we obtained "dreamed" images for each one of the main layers of DMAGNet: Stem, DMA 1, DMA 2, DMA 3 and LAM.

From left to right, the visualizations displayed in Figure 3 reveal a clear evolution from low-level, high-frequency responses, dominated by point-like sources and local contrast variations in the Stem, to more structured, spatially coherent patterns in the DMA blocks. These intermediate stages emphasize repetitive, cell-like or ring-like motifs that align with salient galactic structures such as spiral arms, bulges, and elongated bars. In the final column, corresponding to the LAM, the visualizations reflect the effect of a large-scale attention mechanism that aggregates information over extended spatial regions. The activations are strongly concentrated on the central galaxy structure and its dominant components, with background details largely suppressed, indicating that the model is explicitly weighting globally relevant morphological cues rather than responding to localized or fragmented patterns.

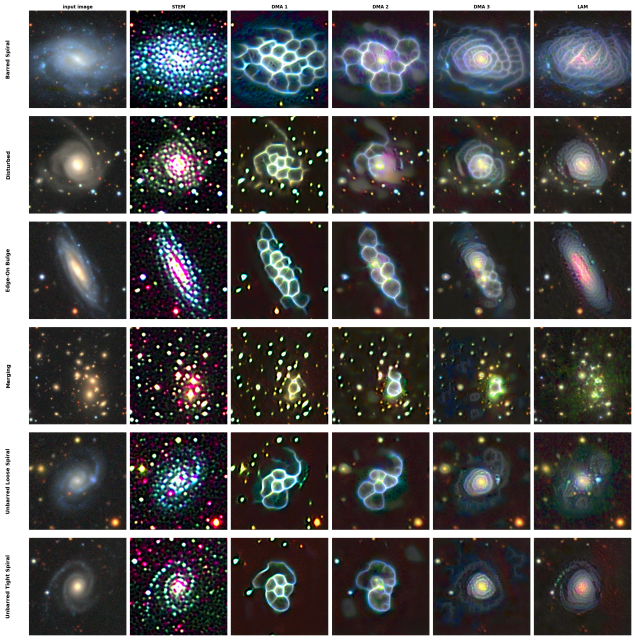

Figure 3. DeepDream visualizations for galaxy morphology classification. Each row corresponds to an input galaxy image from a different morphological class (leftmost column). The remaining columns show the enhanced patterns obtained by applying DeepDream to progressively deeper components of the DMAGNet (STEM, DMA1, DMA2, DMA3, LAM), highlighting class-discriminative features learned at each stage.

In our framework, we generated DeepDream visualizations for each of the principal stages of DMAGNet, namely the Stem, DMA 1, DMA 2, DMA 3, and the LAM module, in order to probe the internal representations learned at different depths of the network. To facilitate a clearer and more interpretable visual inspection of these internal features, we adopt a multi-resolution re-dreaming pipeline designed to produce sharper and more detailed activation visualizations. Specifically, the pipeline operates sequentially at three increasing image resolutions ($256 \rightarrow 512 \rightarrow 1024$), where at each resolution a complete DeepDream optimization is performed starting from the upscaled output of the previous stage. Importantly, the optimization is reinitialized at each resolution rather than being directly continued, and although each stage internally makes use of a two-level spatial pyramid, no gradients are propagated or shared across different resolutions.

6. Conclusion

In this work, we introduced DMAGNet, a convolutional neural network tailored for galaxy morphology classification that combines multi-scale feature extraction with attention mechanisms to effectively capture complex astrophysical structures. Evaluated on the Galaxy10 DECaLS dataset, the model achieves competitive classification performance across ten morphological classes, with errors predominantly arising between visually and physically related categories, reflecting intrinsic ambiguities in galaxy morphology rather than model deficiencies.

Beyond quantitative performance, a central contribution of this study lies in the interpretability analysis conducted through Google Deep Dream. The resulting visualizations provide qualitative evidence that DMAGNet learns a hierarchical and progressively more abstract representation of galaxy morphology, evolving from low-level photometric patterns in early layers to globally coherent, class-discriminative structures in deeper attention-based modules. In particular, the large-scale attention mechanism in the LAM stage appears to focus selectively on the dominant morphological components of galaxies, suppressing background noise and emphasizing astrophysically meaningful features.

These findings suggest that DMAGNet does not rely solely on superficial texture cues, but instead encodes structurally relevant information aligned with established morphological concepts. Future work may extend this same pipeline to different research domains, where it could provide to domain experts visual insights that are difficult to detect through direct human inspection.

References

[1] Roberto G. Abraham, Sidney van den Bergh, Karl Glazebrook, Richard S. Ellis, Basil X. Santiago, Peter Surma, and Richard E. Griffiths. The morphologies of distant galaxies. i. an automated classification system. *The Astrophysical Journal Supplement Series*, 107:1–17, 1996.

[2] Christopher J. Conselice, Matthew A. Bershady, and Andrew Jangren. The asymmetry of galaxies: Physical morphology for nearby and high-redshift galaxies. *The Astrophysical Journal*, 529:886–910, 2000.

[3] Gérard de Vaucouleurs. Classification and morphology of external galaxies. *The Astrophysical Journal Supplement Series*, 8:31–91, 1963.

[4] Sander Dieleman, Kyle W. Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

[5] Norman A. Grogin et al. Candels: The cosmic assembly near-infrared deep extragalactic legacy survey. *The Astrophysical Journal Supplement Series*, 197(2):35, 2011.

[6] Edwin P. Hubble. Extragalactic nebulae. *The Astrophysical Journal*, 64:321–369, 1926.

[7] Edwin P. Hubble. *The Realm of the Nebulae*. Yale University Press, New Haven, 1936.

[8] Marc Huertas-Company, Jesús A. L. Aguerri, Mariangela Bernardi, Simona Mei, and Jorge Sánchez Almeida. A morphological classification of galaxies based on a support vector machine. *Astronomy and Astrophysics*, 525:A157, 2011.

[9] Marc Huertas-Company et al. A catalog of visual-like morphologies in the 5 candels fields using deep learning. *The Astrophysical Journal Supplement Series*, 221(1):8, 2015.

[10] Anton M. Koekemoer et al. Candels: The hubble space telescope observations, imaging data products, and mosaics. *The Astrophysical Journal Supplement Series*, 197(2):36, 2011.

[11] Henry Leung and Jo Bovy. Galaxy10 decals: A labeled galaxy image dataset, 2019. Based on Galaxy Zoo classifications and DECaLS imaging.

[12] Chris J. Lintott et al. 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, 2008.

[13] Jennifer M. Lotz et al. A new nonparametric approach to galaxy morphological classification. *The Astrophysical Journal*, 672:177–197, 2008.

[14] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks, 2015. Google Research Blog.

[15] Kyle W. Willett et al. Galaxy zoo 2: detailed morphological classifications for 304,122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, 2013.