DOI:http://dx.doi.org/10.18524/1810-4215.2019.32.182092

DEEP LEARNING FOR MORPHOLOGICAL CLASSIFICATION OF GALAXIES FROM SDSS

V. Khramtsov[1]*, D. V. Dobrycheva[2,3], M. Yu. Vasylenko[2,4], V. S. Akhmetov[1]

[1] Institute of Astronomy, V.N. Karazin Kharkiv National University, 35 Sumska Str., Kharkiv, 61022, Ukraine
[2] Main Astronomical Observatory of the National Academy of Sciences of Ukraine, 27 Akademika Zabolotnoho Str., 03143, Kyiv, Ukraine
[3] Bogolyubov Institute for Theoretical Physics of the National Academy of Sciences of Ukraine, 14-b Metrolohichna Str., 03143, Kiyv, Ukraine
[4] Institute of Physics of the National Academy of Sciences of Ukraine, 46, Nauka avenu, 03028, Kyiv, Ukraine

ABSTRACT. We present the results of applying deep convolutional neural network to the images of redshift-limited ($z < 0.1$) sample of $\sim 300\,000$ galaxies from the SDSS DR9. We aimed to classify galaxies into the two classes: Elliptical and Spiral. To create the training sample, we used a set of $\sim 6\,000$ galaxies from our previous work with visually inspected morphological types, and also added 80 000 well-confirmed galaxies from Galaxy Zoo 2 dataset, that were also classified visually. With a given sample of $\sim 86\,000$ galaxies, we used the deep neural network, namely Xception, to provide a classification of *g-r-i* composite images (25 arcsec in each axis in size) of galaxies. Keeping in the mind a relatively small training dataset, we provided the data augmentation (horizontal and vertical flips, random shifts on $\pm 10\,pixels$, and rotations within 180 degrees), that was randomly applied to the images during learning. The data augmentation is a key technique within our algorithm to display the variative nature of the observed galaxies, and avoid overfitting problem. We compared our classification result with the Support Vector Machine (SVM) classification performed on the SDSS photometric data (absolute magnitudes, colour indices, inverse concentration index, ratios of semiaxes, etc.), and proposed a method to learn the benefits from both approaches (Deep Learning and photometric classification). We show the common mistakes of both algorithms, and propose to stack these two approaches to block these mistakes, with a main goal to increase the overall classification quality of SDSS galaxies.

Keywords: galaxies, morphological classification, machine learning.

АНОТАЦІЯ. В цій роботі ми представляємо результати застосування згорткової нейронної мережі до зображень галактик вибірки, яка обмежена по червоному зміщенню ($z < 0.1$). Вибірка містить $\sim 300\,000$ зображень галактик з цифрового огляду SDSSDR9. Ми мали на меті класифікувати галактики на два класи: Еліптичні та Спіральні. Тренувальна вибірка містить $\sim 6\,000$ зображень галактик з наших попередніх робіт, де візуально було визначено морфологічний тип кожної галактики. До тренувальної вибірки ми додали 80 000 зображень галактик з вибірки даних GalaxyZoo2, що також були класифіковані візуально. На основі даних тренувальної вибірки галактик $\sim 86\,000$ ми застосували згорткову нейронну мережу, а саме Xception, щоб зробити морфологічну класифікацію галактик використовуючи *g-r-i* складені зображення галактик (розміром 25 на 25 кутових секунд). Для навчання моделей ми використовуємо відносно маленьку вибірку зображень, що накладає обмеження на оцінку якості моделей, та їх подальшу експлуатацію. Для покращення результатів ми використали декілька відомих прийомів (горизонтальні та вертикальні перевороти, випадкові зсуви на ± 10 пікселів та обертання в межах 180 градусів), які були випадковим чином застосовані до зображень галактик під час навчання. Збільшення кількості даних є ключовою технікою в нашому алгоритмі для відображення варіативної природи спостережуваних галактик та уникнення проблеми перенавчання моделі. Ми порівняли отримані результати класифікації з результатами нашої попередньої роботи, де використовувались фотометричні данні (абсолютні зоряні величини, показники кольору, зворотній індекс концентрації, співвідношення

*E-mail: vld.khramtsov@gmail.com

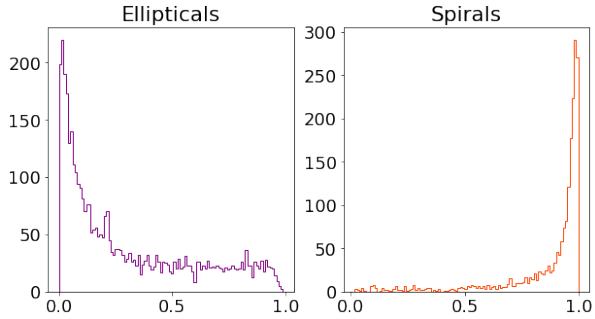

Figure 1: Probability distribution for validation sample of 6 000 galaxies, classified visually and with DL. Probabilities for Ellipticals (E) (*left*) and for Spirals (S) (*right*) of being E- and S- types, respectively.

півосей та ін.) та метод опорних векторів. В результаті, ми запропонували метод для вивчення переваг з обох підходів (глибоке навчання та фотометрична класифікація). Ми показуємо загальні помилки обох алгоритмів і пропонуємо скласти ці два підходи для уникнення помилок з метою підвищити точність морфологічної класифікації галактик SDSS.

Ключові слова: морфологічна класифіквція галактик, машинне навчання.

1. Introduction

Morphological classification of galaxies can provide insights into the processes that form the evolution of Universe. The modern wide-field surveys (like the SDSS [Blanton et al. 2017]) include $\sim 10^5$ of resolved galaxies and require the machine learning application. Deep Learning (DL) methods (namely, Convolutional Neural Networks) mimics visual inspection of images by expert but with a much higher speed-performance. In this study we used Deep Convolutional Neural Network to classify 300 000 galaxies ($z < 0.1$) from the SDSS DR9 [Ahn et al. 2012, Dobrycheva et al. 2017] and compared it with photometric classification approach [Dobrycheva et al. (2018), Dobrycheva et al. (2015)]. We note that binary classification (Elliptical and Spiral galaxies) is a primary method for searching gravitational lenses [Sergeyev et al. 2018, Khramtsov et al. 2019] and locating gravitational waves' host galaxies [Dalya et al. 2018] as well as important for Zone of Avoidance identification [Vavilova et al. 2018].

2. Data and methods

We used the data from crowd-source Galaxy Zoo 2 project [Willett et al. 2013] providing positional cross-matching of this catalogue with inference sample [Dobrycheva 2013] and obtained 170 000 common sources. Then, we selected reliable classification with using flags (from Galaxy Zoo 2 catalogue) for the Elliptical (`t01_smooth_or_features_a01_smooth_flag`) and Spiral (`t01_smooth_or_features_a02_features_or_disk_flag`) galaxies, that returned 15 264 Ellipticals and 64 441 Spirals. Besides this training sample we used a list of 6 163 visually inspected galaxies (4 148 Ellipticals and 2 015 Spirals) as a validation sample. We applied the Deep Convolutional Neural Network called Xception [Chollet 2016] that gives the state-of-art performance in classifying images. We trained the Xception network on 2/3 of training sample from Galaxy Zoo data and validated the results on remaining 1/3 fraction and on visually inspected galaxies. The distribution of final probabilities for visually inspected 6 000 galaxies is shown in Fig.1.

3. Results

We compare the obtained probabilities returned by our Deep Learning (DL) model with the corresponding probabilities obtained by Support Vector Machine (SVM, [Vapnik 1979]) method. One can see in Fig.1, that our DL model performed well on Spirals, when some Ellipticals from visually inspected sample flowed to the class of Spiral galaxies ($p > 0.5$). We inspected the galaxies, which have rival probabilities (i.e., were classified differently with two methods, see Fig.2, top). Galaxies classified with DL as Ellipticals and as Spirals with SVM look like smooth rounded sources, but in most of cases, they are the starforming galaxies (as we directly checked with SDSS spectra) despite the lacking of resolved spiral structure on their images. This result indicates that DL method can classify rounded sources as Ellipticals but it can not catch the spectral energy distribution properties of galaxies more clearly than SVM, trained on photometric features of galaxies. Also, the galaxies classified as Spirals with DL and as Ellipticals with SVM, are, mostly, the edge-on or face-on Spirals. This confirms that SVM method could not deal with such galaxy images and one should include the additional information (e.g., semi-axis ration) in classification [Vasylenko et al. 2019]. The total amount of differently classified galaxies is 60 000. So, an overall classification of galaxies with DL is satisfactory (Fig.2, bottom) and can be joined with SVM classification to obtain more confident result. We propose to use the Stacking algorithm, when some meta-classifier learns to select sources with using probabilities returned by some basic classifiers as input features. We expect that this method will improve the final classification adopting the best from both methods.

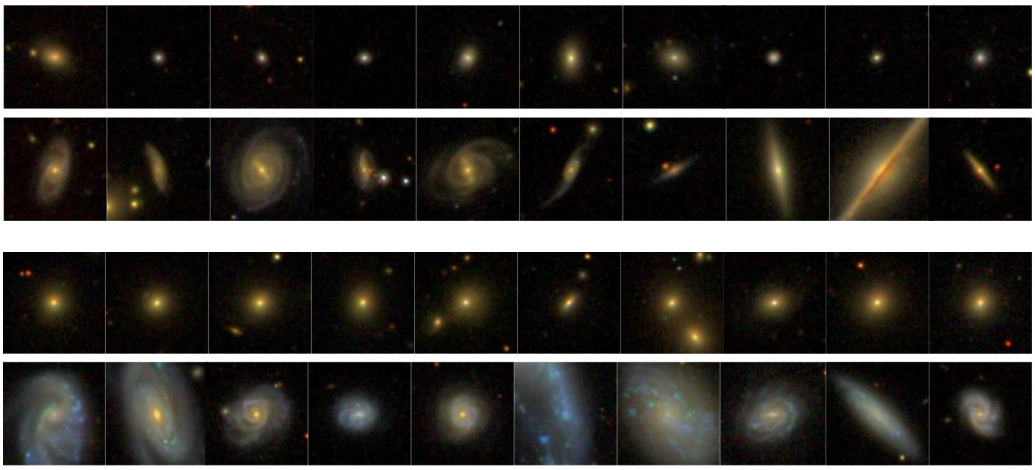

Figure 2: (Top figure) Representative galaxy sample with rival probabilities; top row: Ellipticals with DL, Spirals with SVM; bottom row: Ellipticals with SVM, Spirals with DL. (Bottom figure) Representative galaxy sample with reliable DL classification; top row: Ellipticals; bottom row: Spirals.

Acknowledgement. The work was partially supported by the grant for Young Scientists Research Laboratories (2018-2019, Dobrycheva D.V.) and the Youth Scientific Project (2019-2020, Dobrycheva D.V., Vasylenko M.Yu.) of the NAS of Ukraine.

References

Ahn C. P., Alexandroff R., Allende P. C. et al.: 2012, *ApJS*, **203**, 21.

Blanton M. R., Bershady M. A., Abolfathi B. et al.: 2017, *Astron. J.*, **154**, 35.

Chollet F.: 2016, arXiv:1610.02357.

Dalya G., Galgoczi G. et al.: 2018, *MNRAS*, **479**, 2374.

Dobrycheva D. V.: 2013, *OAP*, **26**, 187.

Dobrycheva D. V., Melnyk O. V., Vavilova I. B. et al.: 2015, *Astrophysics*, **58**, 168.

Dobrycheva D. V. et al.: 2017, arXiv:1712.08955.

Dobrycheva D.V., Vavilova I. B., Melnyk O. V. et al.: 2018, *Kinemat. Phys. Celest. Bodies*, **34**, 290.

Khramtsov V., Sergeyev A., Spiniello C. et al.: 2019, *arXiv:1906.01638*.

Sergeyev A., Spiniello C., Khramtsov V. et al.: 2018, *AAS*, **2**, 189.

Vapnik V.: 1979, Estimation of Dependences Based on Empirical Data [in Russian].

Vavilova I. B., Elyiv A. A., Vasylenko M. Yu.: 2018, *Radio Phys. Radio Astron.*, **23**, 244.

Vasylenko M. Yu., Dobrycheva D. V., Vavilova I. B. et al.: 2019, *OAP*, this issue.

Willett K. W., Lintott C. J., Bamford S. P. et al.: 2013, *MNRAS*, **435**, 2835.