

ПОИСК ПОХОЖИХ ГАЛАКТИК В ШИРОКОУГОЛЬНОМ ОБЗОРЕ PAN-STARRS1 МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Савченко Сергей Сергеевич

канд. физ.-мат. наук, старший научный сотрудник, Санкт-Петербургский государственный университет, РФ, г. Санкт-Петербург

A SEARCH OF SIMILAR GALAXIES IN THE PAN-STARRS1 WIDE-FIELD SURVEY WITH MACHINE LEARNING METHODS

Sergey Savchenko

Candidate of Science, senior researcher, Saint Petersburg state University, Russia, Saint Petersburg

Аннотация. Современные обзоры неба содержат изображения сотен тысяч галактик, что делает поиск объектов заданного типа в таком большом объеме данных трудной задачей. В работе предлагается метод, основанный на сверточной нейронной сети, позволяющей быстро находить изображения галактик, подобных заданной. Работа метода демонстрируется на примере изображений из обзора Pan-STARRS1.

Abstract. Modern sky surveys contain hundreds of thousands of galactic images, which makes a search for the objects of some specific type a very difficult task. In this article we propose a method, based on a convolutional neural network, which allows one to easily find images of galaxies that are similar to a given one. The method is applied to the images of the Pan-STARRS1 survey to demonstrate its performance.

Ключевые слова: галактики; обработка данных; машинное обучение.

Keywords: galaxies; data analysis; machine learning.

Введение. В последние два десятилетия большое значение в астрономии стали иметь широкоугольные обзоры неба. Такие обзоры выполняются на одном или нескольких телескопах и содержат изображения больших участков небесной сферы. Так, например, обзор Pan-STARRS1 содержит изображения больше, чем половины небесной сферы [1] с угловым разрешением 0.25 угловых секунд на пиксель. Всего база данных обзора содержит сотни тысяч изображений с миллионами объектов и занимает несколько десятков терабайт. Анализ такого большого объема данных представляет собой трудную задачу: ручная классификация изображений более чем миллиона галактик из другого обзора, SDSS, покрывающего треть небесной сферы потребовала почти трехлетних усилий тысяч волонтеров [2].

Одним из видов анализа, который часто требуется для больших баз астрономических данных,

является поиск изображений объектов заданного типа: как правило, отдельные исследования концентрируются вокруг галактик какой-то конкретной морфологии (эллиптических или спиральных галактик) или какой-то конкретной пространственной ориентации (дисковые галактики, видимые «плашмя» лучше подходят для анализа спирального узора, а галактики, видимые «с ребра» позволяют исследовать вертикальную структуру их звездных дисков).

«Классические» алгоритмы обработки изображений плохо справляются с работой по классификации галактик, так как галактики отличаются очень большим разнообразием и трудно запрограммировать все возможные их виды и формы. В последнее время, большую популярность в области обработки изображений набрали алгоритмы, основанные на машинном обучении, и, в частности, на искусственных нейронных сетях. Основная идея этих алгоритмов заключается в использовании большого набора данных (так называемой обучающей выборки), на основе которого алгоритм сам делает обобщающие выводы, необходимые для решения поставленной задачи [3].

В данной работе представлен алгоритм, основанный на искусственной нейронной сети, позволяющий производить поиск изображений похожих галактик.

Описание алгоритма. В основе искусственной нейронной сети лежит понятие нейрона, построенного по аналогии с биологическим нейроном мозга. Нейрон имеет несколько входов, на которые подаются значения (входные данные) и один выход. Математически нейрон может быть представлен в следующем виде:

$$y = F\left(w_0 + \sum_{i=1}^N x_i w_i\right),$$

где x_i – входные значения нейрона, w_i – так называемые веса, а F – функция активации, добавляющая нелинейность всей системе. Веса нейрона являются его свободными параметрами, и, подбирая разные значения весов, можно регулировать поведение нейрона.

Нейронная сеть состоит из большого числа (часто десятков тысяч) нейронов, организованных в слои. Выходы нейронов одного слоя подаются на вход нейронов следующего слоя, посредством чего организуется передача сигнала от входа нейронной сети до ее выхода. В процессе обучения нейронной сети веса всех ее нейронов модифицируются таким образом, чтобы входные данные, проходя через сеть, преобразовывались в требуемые выходные данные (например, изображение галактики на входе нейросети в тип галактики на ее выходе). Конкретная организация нейронов (число нейронов в каждом слое и число слоев) называется архитектурой нейронной сети и зависит от поставленной задачи.

В данной работе используется архитектура нейронной сети, которая называется автоэнкодер [4]. Схематично автоэнкодер изображен на Рис. 1. На вход автоэнкодера подаются данные (Input), которые после преобразования нейросетью считываются с выхода (Output).

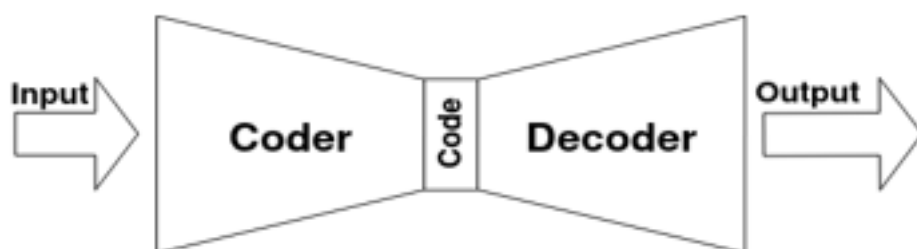


Рисунок 1: Архитектура автоэнкодера

В процессе обучения, веса автоэнкодера подбираются так, чтобы минимизировать разницу между его входом и выходом (то есть нейросеть реализует отображение идентичности). Чтобы нейросеть не сошлась к тривиальному решению, когда данные со входа просто копируются на выход, в архитектуре нейросети устраивается «бутылочное горлышко», слой в котором число нейронов значительно меньше размера входных данных (в случае изображений – числа пикселей). В результате, нейросети приходится обучаться таким образом, чтобы ее часть до бутылочного горлышка (называется кодер, *coder*) сжимала входные данные до необходимого размера с сохранением как можно большей информации об исходных изображениях. В результате такого сжатия получается вектор-код (его еще называют эмбедингом), который представляет собой сжатое входное изображение. Вторая же половина нейросети, называемая декодер (*decoder*), решает обратную задачу: восстановить оригинальное изображение из данного эмбединга.

Похожие галактики будут иметь похожие вектора-эмбединги, таким образом, сравнивая эти вектора между собой можно находить степень схожести исходных галактик. Математически расстояние между двумя векторами можно выразить в виде Евклидовой L2-меры:

$$\|\vec{x} - \vec{y}\| = \sqrt{\sum_i (x_i - y_i)^2}.$$

Пары галактик, у которых значение такой разницы между их векторами-эмбедингами будет маленьким, окажутся похожими. Таким образом, задача по поиску похожих галактик сведется к поиску разниц между их векторами-эмбедингами.

Данные. Для тренировки нейронной сети требуется обучающая выборка – набор изображений галактик, которые будут даваться ей на вход для поиска оптимальных значений весов. В качестве такой обучающей выборки были использованы изображения 50000 случайных галактик из базы данных HyperLEDA[5]. Поскольку, как правило, обучение нейронной сети происходит тем лучше, чем больше изображений в обучающей выборке, исходный массив был расширен до 256000 изображений путем случайных преобразований (зеркальное отражение, поворот, добавление шума) над оригинальными изображениями.

Результаты.

Рис. 2 показывает результаты работы автоэнкодера.

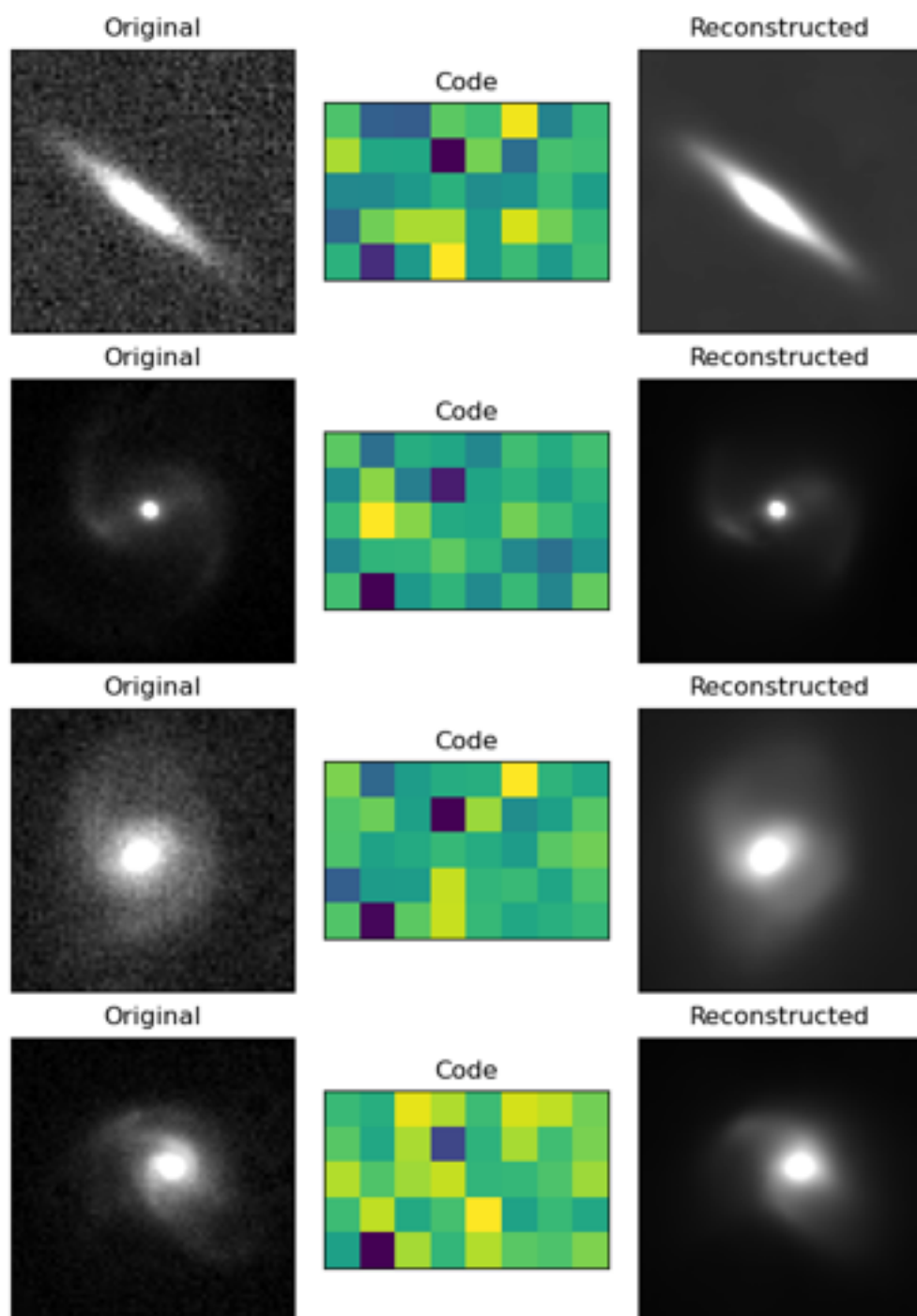


Рисунок 2. Примеры работы автоэнкодера для четырех галактик. Слева - исходное изображение, в центре вектор-эмбединг, справа - восстановленное изображение

В левом столбце показаны оригинальные изображения галактик, которые имеют размер 64x64 пикселя (то есть всего 4096 пикселей). В среднем столбце показан вектор-эмбединг, имеющий 40 значений (большим значениям вектора соответствует более яркие пиксели), то есть сжатое представление имеет ~1% от оригинального размера. В правом столбце показаны восстановленные декодером изображения галактик. Видно, что несмотря на большую степень сжатия, восстановление изображений происходит довольно качественно: все глобальные

черты галактик (ориентация дисков, форма спирального узора) воспроизведены правильно. Таким образом, можно сделать вывод о том, что предложенная модель автоэнкодера работает хорошо.

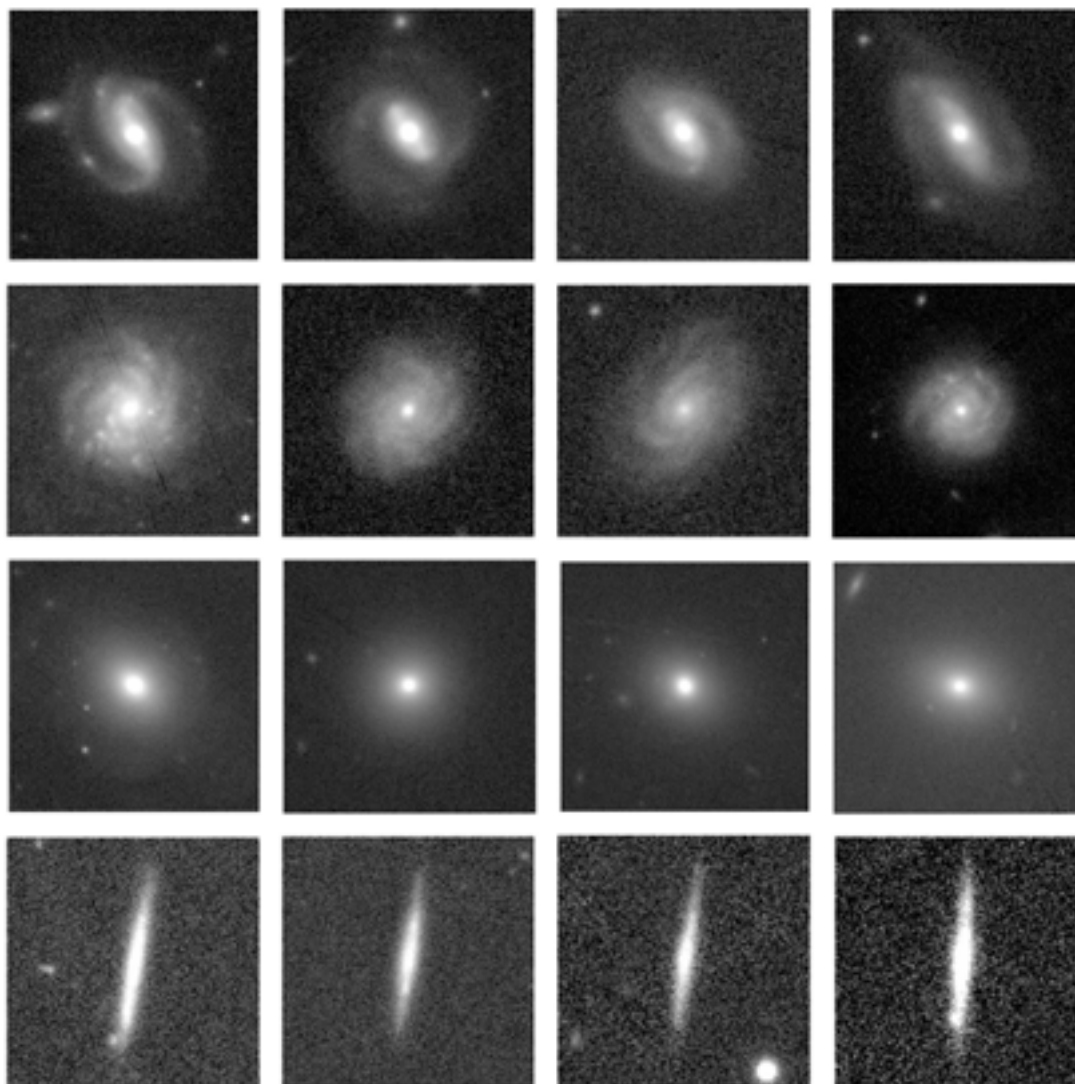


Рисунок 3. Результат поиска похожих галактик на примере четырех объектов. В левом столбце показаны целевые галактики. Во втором, третьем и четвертом - результаты поиска

Результат поиска похожих галактик приведен на Рис. 3. На этом рисунке для четырех галактик, демонстрирующих разную морфологию (показаны в левом столбце), приведены три наиболее похожие (то есть, наиболее близкие в пространстве векторов-эмбедингов) галактики. Видно, что результаты поиска вполне успешные: все галактики близкие в пространстве эмбедингов, действительно, оказались похожими, что подтверждает работоспособность метода.

Заключение. В работе решается задача о поиске похожих галактик среди изображений

астрономических баз данных. Для решения этой задачи применяется автоэнкодер – алгоритм, основанный на искусственной нейронной сети. Применение алгоритма к выборке из 50 тысяч галактик продемонстрировало, что данный метод, действительно, является мощным инструментом, позволяющим найти изображения похожих объектов в большом массиве данных.

Список литературы:

1. K. C. Chambers, et al., The Pan-STARRS1 Surveys, arXiv:1612.05560
2. Lintott, C., et al., 2011, MNRAS, 410(1):166–178.
3. LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
4. Liou, C.-Y., Cheng, W.-C., Liou, J.-W., & Liou, D.-R. 2014, Autoencoder for words. Neurocomputing, 139, 84–96.
5. Makarov D., et al., 2014, A&A, 570, A13