Кластеризация изображений по визуальному подобию с помощью автоэнкодера

А.С. Коваленко[1], **Я.М. Демяненко**[1]

[1]Институт математики механики и компьютерных наук имени И. И. Воровича, Мильчакова 8а, Ростов-на-Дону, Россия, 344090

Аннотация. В работе рассматривается подход к решению проблемы поиска схожих изображений по визуальному подобию с использованием нейронных сетей из ранее неразмеченного набора данных. Решение сводится к использованию специальной архитектуры нейронной сети - автоэнкодера, с помощью которого осуществляется извлечение высокоуровневых признаков из изображения. Поиск ближайших элементов реализован по евклидовой метрике в сформированном пространстве признаков, после предварительной декомпозиции до двумерного пространства. Полученное множество используется для решения задачи классификации с предварительной кластеризацией.

1. Введение

На данный момент существует большое количество подходов к решению задачи классификации [7]. Но как правило, все они сводятся к использованию класса моделей с алгоритмом обучения с учителем. Всех их объединяет один существенный недостаток - требование размеченных данных для обучения. Когда возникает новая задача в сфере компьютерного зрения, как правило, размеченные данные отсутствуют, и необходимо тратить средства на их разметку.

Если рассматривать подходы, основанные на методах некотролируемого обучения, такие, как алгоритмы кластеризации, они, как правило ориентированы на работу с данными небольших размерностей. Если рассматривать в качестве обрабатываемых данных изображения, то они, как правило имеют высокую размерность. Понижение размерности пространства, к примеру методом главных компонент, все равно не дает пространства, к которому эффективно применимы алгоритмы кластеризации.

Возникает необходимость в построении отображения, действующего из пространства изображений Ω (1) в некоторое пространство признаков этих изображений, к которому можно эффективно применять методы декомпозиции и непосредственно производить кластеризацию и поиск ближайших объектов по визуальной составляющей.

2. Существующие решения

Самым часто используемым подходом к решению задачи понижения размерности является метод главных компонент. Но он приемлимо справляется только с хорошо линейно разделимыми данными. Если рассматривать объекты больших размерностей, вероятность их хорошей резделимости становится малой. Но если они составляют смеси объектов, принадлежащих к нормальным распределениям с разными параметрами, их можно

разделить с помощью алгоритма t-SNE (Laurens van der Maaten Visualizing Data using t-SNE) [2]. В большинстве случаев идет работа с данными, не удовлетворяющими этим требованиям. Возникает необходимость построения отображения из текущего пространства объектов в пространство их описательных признаков, на которое будет наложено требование их распределения по нормальному закону. Такую задачу рассматривают авторы статьи по вариационным автоэнкодерам (Doersch C. Tutorial on Variational Autoencoders) [1]. В нашей работе для решения поставленной задачи построен и обучен энкодер, выступающий необходимым отображением в пространство признаков изображений, распределенных по нормальному закону. Далее, к полученному пространству можно применять алгоритм t-SNE для дальнейшей декомпозиции пространства до размерности, где будут хорошо работать алгоритмы кластеризации.

3. Построение энкодера

Для обучения и тестирования моделей использовался набор изображений "MNIST" http://yann.lecun.com/exdb/mnist/, содержащий коллекцию изображений рукописных цифр. Пример данных изображен на рисунке 1.

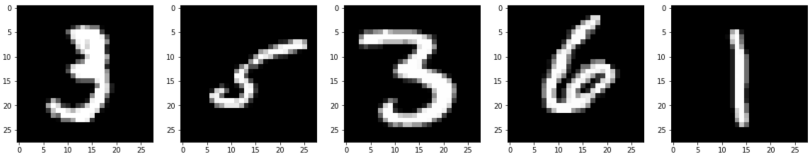

Рисунок 1. Примеры MNIST.

Построен автоэнкодер с архитектурой, изображенной на рисунке 2. Общая концепция архитектуры подробнее описана в работе [1].

В ходе экспериментов использовались следующие архитектуры нейронных сетей энкодера (Q) и декодера (P): сети, построенные только на полносвязных слоях и сверточные сети. Параметром для данных сетей выступала размерность скрытого пространства.Детальное описание архитектур данных моделей в виде блок-схем находится в репози-тории, приложенном к работе. Краткое описание моделей изображено на рисунках 3 и 4.

Обе модели обучались на указанном выше наборе данных, размером 60000. Для обучения использовался оптимизационный алгоритм Adam [6]. Количество эпох обучения: 500. Для построения и обучения использовался фреймворк keras (https://keras.io) с бекендом на tensorflow (https://www.tensorflow.org).

На выходе энкодер дает два вектора: вектор среднего значения для распределения, к которому относится объект и вектор ковариационной матрицы в диагональном виде. Для построения множества H воспользуемся средним значением, так как оно характеризует центроиду кластера в пространстве скрытых признаков, где Ω - множество исходных объектов (1), g - энкодер.

$$\Omega = \{I_m\}_{m=1}^{N} \tag{1}$$

После получения обученного энкодера построим искомое множество H следующим образом:

$$H = \{g(x)|x \in \Omega\}, \tag{2}$$

Теперь понизим размерность множества H с помощью алгоритма распределенного стохастического выделения соседей (t-SNE):

$$\hat{H} = \text{t-SNE}(H),\ \forall h \in \hat{H} \Rightarrow dim(h) = 2 \tag{3}$$

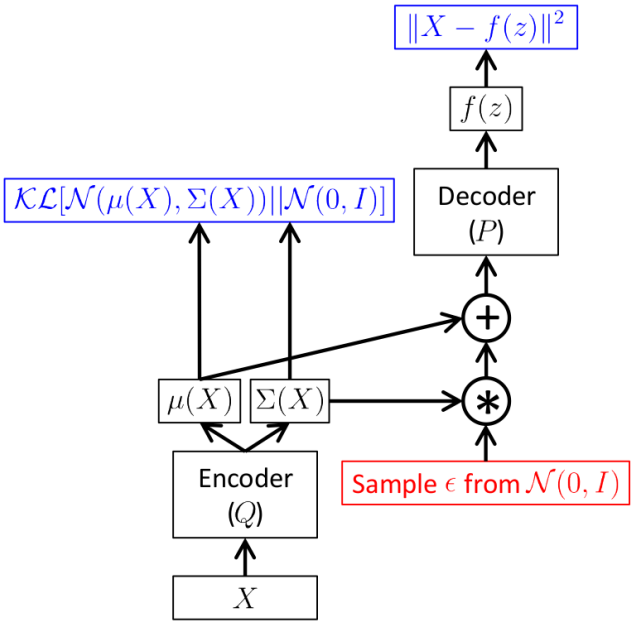

Рисунок 2. Общая блок-схема модели автоэнкодера, где синими блоками обозначены функции ошибок.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_15 (InputLayer)        (None, 28, 28, 1)         0
_________________________________________________________________
Encoder (Model)              (None, 10)                237972
_________________________________________________________________
Decoder (Model)              (None, 28, 28, 1)         237456
=================================================================
Total params: 475,428
Trainable params: 473,892
Non-trainable params: 1,536
_________________________________________________________________
```

Рисунок 3. Сводка для модели, построенной на полносвязных слоях.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_9 (InputLayer)         (None, 28, 28, 1)         0
_________________________________________________________________
Encoder (Model)              (None, 10)                25385
_________________________________________________________________
Decoder (Model)              (None, 28, 28, 1)         24924
=================================================================
Total params: 50,309
Trainable params: 50,309
Non-trainable params: 0
_________________________________________________________________
```

Рисунок 4. Сводка для модели сверточной архитектуры.

С другой стороны, можно задать размерность скрытого пространства вариационного автоэнкодерова, равным 2, тогда получим множество объектов искомой размерности. Но в этом случае возрастают потери информации при кодировании объекта энкодером и результаты декомпозиции с таким подходом получаются хуже.

4. Эксперименты

Возьмем 5000 элементов из множества данных ”MNIST” и подействуем на них энкодером, а затем алгоритмом t-SNE, получив множество $\hat{H}$ (3). Рассмотрим примеры множеств $\hat{H}$, полученных при использовании модели энкодера g с различным параметрамом размерности скрытого пространства. На рисунке 5 изображено множество $\hat{H}$ при использовании расмерности скрытого пространства, равным 2, без дальнейшего использования t-SNE.

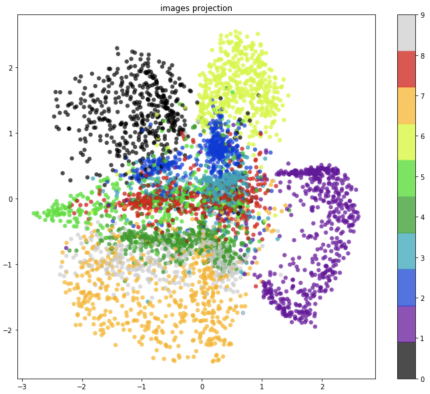

Рисунок 5. Визуализация скрытого пространства размерностью 2, полученного полносвязным энкодером.

На рисунке 6 изображено множество $\hat{H}$, построенное при использовании полносвязной модели с параметром размерности скрытого пространства, равным 10, с применением алгоритма t-SNE.

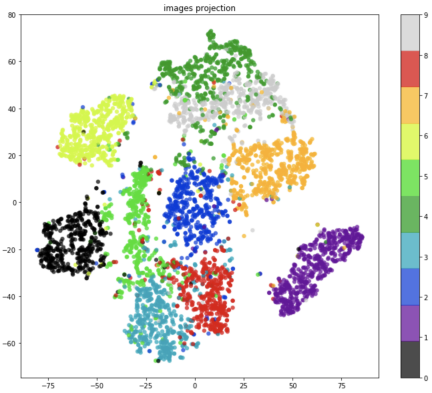

Рисунок 6. Скрытое пространство размерностью 10, полученное полносвязным энкодером, с последующим применением t-SNE.

На рисунке 7 изображено множество $\hat{H}$, построенное при использовании сверточной модели с параметром размерности скрытого пространства, равным 10, с применением алгоритма t-SNE.

Можно наблюдать, что, при использовании скрытого пространства большей размерности и последующем действии на него алгоритмом t-SNE, классы стали лучше разделимы между собой, рисунок 6. Это связано с тем, что автоэнкодер лучше восстанавливает изображение на выходе и вследствие чего пространство скрытых признаков становится более репрезентативным. При использовании сверточной архитекруты визуальная разделимость

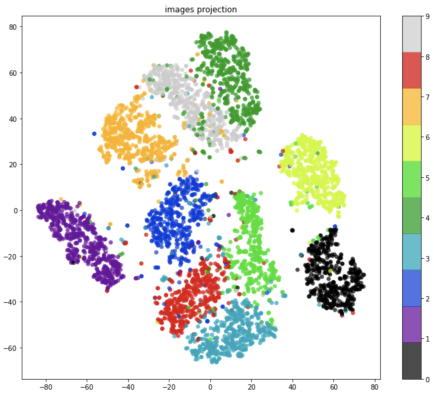

Рисунок 7. Скрытое пространство размерностью 10, полученное сверточным энкодером, с последующим применением t-SNE.

классов показывает схожие результаты, рисунок 7, причем количество параметров у данной архитектуры на порядок меньше, чем у целиком состоящей из полносвязных слоев. Так как при обучении вариационного автоэнкодера, энкодер стремится предсказывать параметры нормального распределения, к которому должен принадлежать поданный на вход объект, то при рассмотрении множества предсказанных средних значений получаем пространство нормально распределенных значений. Алгоритм t-SNE использует метрику близости объектов по нормальному распределению, что непосредственно и дает хороший результат декомпозиции. Также существует метод понижения размерности пространства, основанный на выделении глваных компонент (PCA) [4], но он показывает худшие результаты при применении на данном множестве скрытых признаков, рисунок 8.

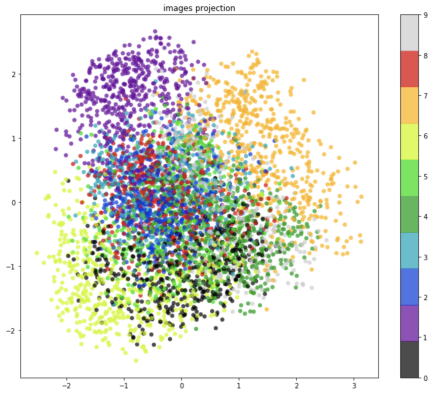

Рисунок 8. Визуализация пространства, полученного действием энкрдера, состоящего из полносвязных слоев, после декомпозиции методом главных компонент.

4.1. Поиск ближайших и кластеризация

Для примера поиска ближайших изображений по визуальному подобию, взяты случайные элементы из множества $\hat{H}(3)$ для которых найдены 5 ближайших элементов из того же множества по евклидовой метрике, рисунки 9 и 10.

Поскольку в множестве $\hat{H}(3)$ сгустки данных имеют случайную нелинейную форму, был применен алгоритм кластеризации DBSCAN [3]. После разбиения множества на кластеры,

были взяты по 10 представителей каждого класса и методом голосования выбраны метки для каждого класса из имеющейся разметки данных. После этого оценена точность классификации. Она составляет 82.2%. Если снижать пространство до размерности 2 только с помощью энкодера, точность составляет 75.9%.

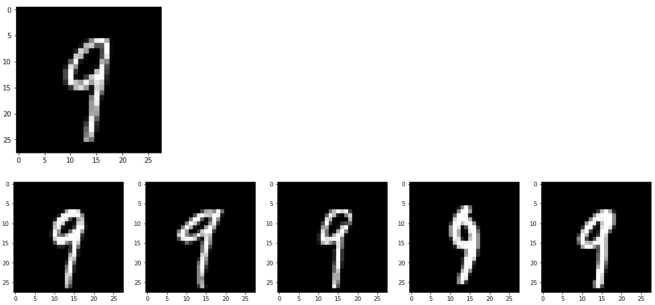

Рисунок 9. Исходное изображение (сверху) и 5 ближайших (снизу).

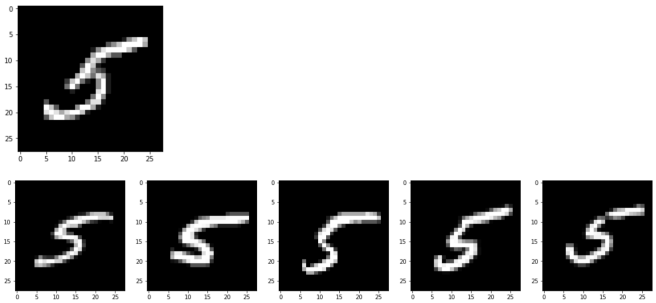

Рисунок 10. Исходное изображение (сверху) и 5 ближайших (снизу).

5. Результаты

В ходе работы была построена модель вариационного автоэнкодера, обученная на наборе данных задачи ”MNIST”. Как показывает эксперимент, описанный подход использования большей размерности скрытого пространства при его дальнейшей декомпозиции с помощью метода t-SNE дает лучшую разделимость классов, чем снижение размерности только автоэнкодером, и дает более высокую точность при кластеризации множества, 82.2% вместо 75.9%

6. Заключение

Рассмотренный подход позволяет решать задачу классификации на заранее неразмеченных данных и выполнять поиск ближайших по визуальному подобию.

Также можно выбрать ближайшие элементы к центроидам кластеров, которые будут с большей вероятностью правильно классифицированны при кластеризации, и обучить классификатор, основанный на нейронной сети [5], что, возможно позволит выполнять классифицацию на всем множестве входных данных с большей точнотью.

7. Приложения

Ссылка на репозиторий с реализацией: https://github.com/AlexeySrus/Clustering-by-VAE/.

8. Литература

[1] Doersch, C. Tutorial on Variational Autoencoders // C. Doersch, C. Mellon. – UC Berkeley, 2016.
[2] Maaten, L. Visualizing Data using t-SNE / L. Maaten, G. Hinton // Journal of Machine Learning Research. - 2008. - Vol. 9. – P. 2579-2605.
[3] Ester, M. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise / M. Ester, H.-P. Kriegel, J. Sander, X. Xu // Institute for Computer Science, University of Munic, 1998.
[4] Shlens, J. A Tutorial on Principal Component Analysis // Institute for Nonlinear Science, University of California, San Diego, 2005. – P. CA 92093-0402
[5] Wu, H. CNN-Based Recognition of Handwritten Digits in MNIST Database // Research School of Computer Science. – The Australia National University, Canberra.
[6] Kingma, D. A method for stochastic optimization / D. Kingma, B.J. Adam // ICLR, 2015.
[7] Kotsiantis, S.B. Supervised Machine Learning: A Review of Classification Techniques // Department of Computer Science and Technology University of Peloponnese, Greece, 2007.

Image clustering by autoencoders

A.S. Kovalenko[1], **Y.M. Demyanenko**[1]

[1]Institute of mathematics, mechanics and computer Sciences named after I.I. Vorovich, Milchakova street 8a, Rostov-on-Don, Russia, 344090

Abstract. The work considers the approach to solving the problem of finding similar images by visual similarity using neural networks on previously unmarked data. The solution comes down to building special architecture of the neural network - autoencoder, through which extracted high-level features from images. Search for the nearest items implemented by entered metric in the formed feature space. This metric can be applied to the classification task using pre-clustering.