Сегментация неструктурированного текста на изображениях книжных обложек с помощью сверточной сети, основанной на архитектуре U-Net

П.Л. Николаев

***Аннотация* — В рамках данной статьи рассмотрена сверточная нейронная сеть для сегментации изображений с книжными обложками. Приведена структура сети с указанием всех составляющих ее блоков и слоев, а также их параметров, и подробно описан принцип работы каждой части. В качестве основы сети используется модель U-Net. Архитектура данной модели выделяется среди других своей энкодерно-декодерной структурой, позволяющей генерировать новые изображения. При этом энкодерная часть сети ответственна за распознавание изображения, а декодерная – за генерацию нового изображения. Предлагаемая нейронная сеть способна создавать бинарные (черно-белые) маски, на которых текст выделяется одним цветом, а все остальные элементы – другим. Таким образом, осуществляется отделение текста от других элементов на изображении. Для обучения и проверки сверточной нейросети используется самостоятельно собранный и размеченный датасет из 200 примеров. Несмотря на малый объем данных, сеть на основе U-Net неплохо обучается и показывает приемлемые результаты работы, что подтверждается результатами тестирования. Обученную сеть можно применять на практике. В частности, ее предполагается использовать для улучшения точности распознавания текста на книжных обложках.**

***Ключевые слова* — Искусственные нейронные сети, машинное обучение, глубокое обучение, сверточные нейронные сети, сегментация изображений, бинаризация изображений, распознавание текста, U-Net.**

I. Введение

Обработка изображений является сложной и трудоемкой задачей. К одной из таких относится распознавание неструктурированного текста на изображении – когда текст не имеет определенной структуры и может располагаться в произвольных местах. Также при решении подобных задач могут возникнуть сложности из-за множества посторонних элементов на изображении (различные артефакты), различных геометрических фигур или неважных деталей. Для того чтобы упростить процесс распознавания текста можно произвести предварительную обработку исходного изображения путем сегментации объектов. Сегментация изображения – разбиение изображения на множество покрывающих его областей [1]. В данном случае текстовые элементы можно выделить одним цветом (например, черным), а все остальное – другим (белым). В итоге, при сегментации должна получаться бинарная маска.

Цель данной работы состоит в разработке нейросетевого решения для сегментации неструктурированного текста на изображениях, в частности, на книжных обложках. Для этого необходимо выполнить следующие задачи:

- создать модель глубокой нейронной сети для сегментации текста на изображении;
- создать набор данных – собрать изображения книжных обложек и сделать для них бинарные маски;
- обучить нейросеть и проверить эффективность ее работы.

II. Обзор существующих работ

В области распознавания изображений для сегментации используют различные методы: классические алгоритмы наращивания областей, алгоритмы кластеризации и определения прямых дуг и окружностей [1], классические методы машинного обучения. Что касается выделения текста, то в [2] рассматривается использование вейвлет-преобразования с методом k-средних для сегментации текста на изображении. В [3] также используются вейвлет-преобразования, но с методом опорных векторов. В [4] используется адаптивный бустинг с деревьями решений. Но с помощью подобных методов не всегда можно добиться необходимой точности. К тому же, в большинстве работ речь идет о выделении структурированного текста.

В связи с развитием нейронных сетей появились новые способы сегментации изображений, позволяющие получить более качественные результаты. К примеру, в работах [6-7] рассматривается сегментация текстовых элементов с помощью одного из видов глубоких нейронных сетей – сверточных сетей. Однако в обоих случаях речь идет о рукописном структурированном тексте.

Исходя из вышеизложенного, и было принято решение о разработке собственного метода для сегментации неструктурированного текста на книжных обложках.

Статья получена 18 июля 2023.

П.Л. Николаев – старший преподаватель Московского авиационного института (национального исследовательского университета) (e-mail: npavel89@gmail.com).

III. Модель глубокой сети для сегментации изображений с неструктурированным текстом

Сверточные нейросети показывают довольно хорошие результаты работы при решении задач, связанных с обработкой изображений. Среди моделей, предназначенных для сегментации изображений, выделяется сеть U-Net [8]. Изначально она была создана для сегментации биомедицинских изображений, но нашла свое применение при сегментации и обычных изображений. Ключевая особенность сети U-Net состоит в том, что для ее обучения достаточно небольшого датасета.

Что касается архитектуры сети U-Net, то она состоит из двух частей – энкодера и декодера [9]. На вход сети подается цветное изображение, которое проходит через сверточные слои, а после по полученным признакам строится черно-белое изображение (маска).

В рамках работы была создана модифицированная модель сети U-Net, структура которой представлена в таблице I.

Таблица I. Структура сверточной нейронной сети на основе модели U-Net.

№	Слой (тип)	Выходная форма
1	Входной слой	(None, 320, 256, 3)
2	Conv2D(32, (3,3))+ReLU	(None, 320, 256, 32)
3	Conv2D(32, (3,3))+ReLU	(None, 320, 256, 32)
4	MaxPooling2D(2,2)	(None, 160, 128, 32)
5	Dropout(0.2)	(None, 160, 128, 32)
6	Conv2D(32, (3,3))+ReLU	(None, 160, 128, 64)
7	Conv2D(32, (3,3))+ReLU	(None, 160, 128, 64)
8	MaxPooling2D(2,2)	(None, 80, 64, 64)
9	Dropout(0.2)	(None, 80, 64, 64)
10	Conv2D(128, (3,3))+ReLU	(None, 80, 64, 128)
11	Conv2D(128, (3,3))+ReLU	(None, 80, 64, 128)
12	MaxPooling2D(2,2)	(None, 40, 32, 128)
13	Dropout(0.2)	(None, 40, 32, 128)
14	Conv2D(256, (3,3))+ReLU	(None, 40, 32, 256)
15	Conv2D(256, (3,3))+ReLU	(None, 40, 32, 256)
16	MaxPooling2D(2,2)	(None, 20, 16, 256)
17	Dropout(0.2)	(None, 20, 16, 256)
18	Conv2D(512, (3,3))+ReLU	(None, 20, 16, 512)
19	Conv2D(512, (3,3))+ReLU	(None, 20, 16, 512)
20	Conv2DTranspose(256, (3,3), (2,2))	(None, 40, 32, 256)
21	Объединение слоев 20 и 16	(None, 40, 32, 512)
22	Dropout(0.2)	(None, 40, 32, 512)
23	Conv2D(256, (3,3))+ReLU	(None, 40, 32, 256)
24	Conv2D(256, (3,3))+ReLU	(None, 40, 32, 256)
25	Conv2DTranspose(128, (3,3), (2,2))	(None, 80, 64, 128)
26	Объединение слоев 25 и 12	(None, 80, 64, 256)
27	Dropout(0.2)	(None, 80, 64, 256)
28	Conv2D(128, (3,3))+ReLU	(None, 80, 64, 256)
29	Conv2D(128, (3,3))+ReLU	(None, 80, 64, 128)
30	Conv2DTranspose(64, (3,3), (2,2))	(None, 160, 128, 64)
31	Объединение слоев 30 и 8	(None, 160, 128, 128)
32	Dropout(0.2)	(None, 160, 128, 128)
33	Conv2D(64, (3,3))+ReLU	(None, 160, 128, 64)
34	Conv2D(64, (3,3))+ReLU	(None, 160, 128, 64)
35	Conv2DTranspose(32, (3,3), (2,2))	(None, 320, 256, 32)
36	Объединение слоев 35 и 4	(None, 320, 256, 64)
37	Dropout(0.2)	(None, 320, 256, 64)
38	Conv2D(64, (3,3))+ReLU	(None, 320, 256, 32)
39	Conv2D(64, (3,3))+ReLU	(None, 320, 256, 32)
40	Conv2D(1, (3,3))+Sigmoid	(None, 320, 256, 1)

На входной слой подается цветное изображение следующей формы (высота, ширина, количество каналов): 320x256x3. На выходе сети получается маска – черно-белое изображение формы 320x256x1. На этой маске текст должен быть выделен черным цветом, а все остальное – белым.

После входного слоя следует энкодерная часть, состоящая из повторяющихся блоков, которые отличаются параметрами.

В одном блоке идут двумерные сверточные слои с одинаковым количеством карт признаков. Везде используется одинаковая функция активации ReLU. После сверточных слоев следует слой субдискретизации (MaxPooling2D), а затем – слой прореживания (Dropout). Прореживание необходимо для уменьшения переобучения сети, оно означает, что определенная доля нейронов принимает значение ноль. В данном случае изменению подвергается 0,2 всех нейронов.

После идут повторные блоки из таких же слоев, но уже с другими значениями карт признаков.

После энкодерной части следует пара сверточных слоев с 512 картами признаков. А уже после них начинается декодерная часть. Она также состоит из повторяющихся блоков с различными параметрами.

Каждый блок декодера включает в себя пару сверточных слоев с одинаковым количеством карт признаков. Далее следует транспонированный слой свертки – деконволюции (Conv2DTranspose). Он предназначен для создания выходной карты признаков большего размера. А после этого происходит объединение конца блока с соответствующему ему по параметрам концу блока из энкодерной части.

В конце идет последний сверточный слой с одной картой признаков и сигмоидальной функцией активации. Выход этого слоя и будет являться выходом сети – бинарной маской.

IV. Набор данных

Для того чтобы обучить нейросеть, а потом проверить точность ее работы, был собран набор данных из 200 изображений с книжными обложками различного размера. Для каждой обложки вручную была создана черно-белая маска. При этом текст выделялся черным цветом, а все остальное – белым. Таким образом, получились бинарные маски. По сути, задача обучения нейросети сводится к бинарной классификации каждого пикселя.

Из собранного набора были составлены 3 выборки: обучающая (120 изображений), валидационная (20 изображений) и тестовая (60 изображений).

V. Обучение сети

Разработка кода для обработки датасета, программная реализация сети и ее непосредственное обучение осуществлялись в среде Google Colab [9], которая предоставляет возможность использования различных библиотек, а также видеокарт для ускорения обучения нейронных сетей.

U-Net подобная сверточная сеть была реализована на языке программирования Python 3.10 с помощью глубокой нейросетевой библиотеки TensorFlow 2.12.0 с API Keras.

Для рассматриваемой сети были заданы следующие функции и параметры обучения:

- оптимизатор – Adam с коэффициентом обучения, равным 0,001;
- размер мини-выборки – 8 образцов;
- функция ошибки (потерь) – бинарная энтропия;
- метрика – точность (accuracy).

Обучение сети длилось 50 эпох. Для оценки результативности использовалось значение функции потерь на валидационной выборке – наименьшее значение, равное 0.09, было получено на 45 эпохе.

На рисунках 1 и 2 показаны изменения значений функции потерь и точности во время обучения и валидации по каждой эпохе.

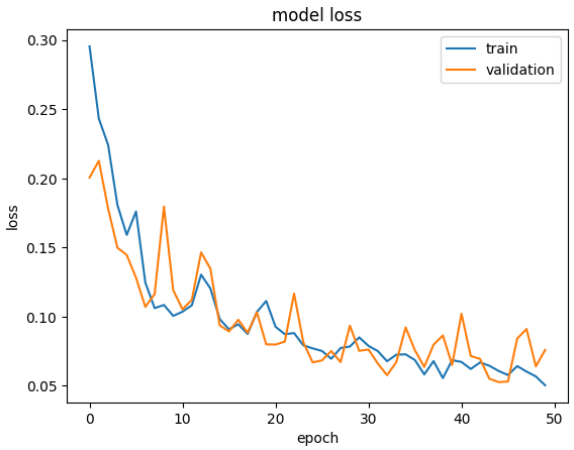


Рис. 1. Изменение значений функции потерь по эпохам.

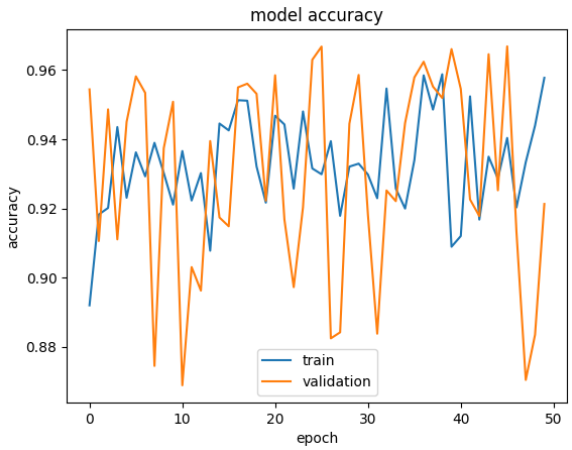


Рис. 2. Изменение значений точности по эпохам.

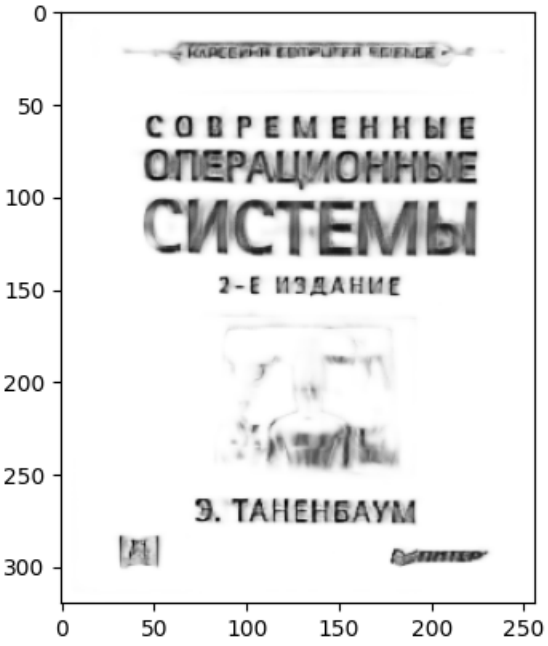
В таблице II показаны результаты обучения и проверки нашей сети с лучшими значениями весовых коэффициентов, полученными на 45 эпохе.

Таблица II. Результаты обучения и проверки сверточной сети для сегментации текста на изображении.

Выборка	Точность (accuracy)	Потери (loss)
Обучающая	95.64%	0.05
Валидационная	90.78%	0.09
Тестовая	90.35%	0.10

На рисунке 3 приведен пример сегментации произвольной книжной обложки. Следует подчеркнуть, что данное изображение не входило в собранный набор данных.

(а) Исходное изображение



(б) Сегментированное изображение

Рис. 3. Пример сегментации книжной обложки.

VI. Заключение

Таким образом, было получено решение для сегментации изображений с неструктурированным текстом на примере книжных обложек. Сегментация осуществляется с помощью сверточной нейронной сети, построенной по модели U-Net. При этом для обучения было достаточно небольшого набора данных.

Полученные в результате тестирования показатели (значение функции потерь 0,10 и точность около 90%) говорят о том, что сеть показывает вполне приемлемое качество работы.

Возможно улучшить работоспособность сети путем увеличения набора данных, что планируется сделать в дальнейшем. Что касается использования обученной сети, то она может найти применение при распознавании неструктурированных текстовых данных на изображениях.

Библиография

[1] Шапиро Л. Компьютерное зрение / Л. Шапиро, Дж. Стокман. – 2-е изд. — М. : БИНОМ. Лаборатория знаний, 2013. – 752 с.

[2] Guatam A. Segmentation of Text from Image Document / Ankush Guatam // International Journal of Computer Science and Information Technologies. – 2013. – Vol. 4. – № 3. – PP. 538-540.

[3] Grzegorzek M. Texture-Based text detection in digital images with wavelet features and Support Vector Machines / M. Grzegorzek, C. Li, J. Raskatow, D. Paulus, N. Vassilieva // Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013. – 2013. – P. 857-866. DOI:10.1007/978-3-319-00969-8_84.

[4] Le V.P. Textand non-text segmentation based on connected component features / V.P. Le, N. Nayef, M. Visani, J. M. Ogier, C. D. Tran // Proceedings of the 2013 13th International Conference on Document Analysis and Recognition(ICDAR). – 2015. – P. 1096–1100. DOI:10.1109/ICDAR.2015.7333930.

[5] Nair R.R. Segmentation of highly unstructured handwritten documents using a neural network technique / R.R. Nair, B.U. Kota, I. Nwogu, V. Govindaraju // Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR) . – 2016. – P. 1291–1296. DOI:10.1109/ICPR.2016.7899815.

[6] Bezmaternykh P.V. U-Net-bin: hacking the document image binarization contest / P.V. Bezmaternykh, D.A. Ilin, D.P. Nikolaev // Компьютерная оптика. – 2019. – №43 (5). – С. 825-832. DOI: 10.18287/2412-6179-2019-43-5-825-832.

[7] Ronneberger O. U-net: Convolutional networks for biomedical image segmentation /O. Ronneberger, P. Fischer P, T. Brox // MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. – 2015. – P. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[8] Николаев П.Л. Анализ деятельности человека посредством глубокого обучения / П. Л. Николаев // Системный администратор. – 2018. – №12 (193). – С. 80-83.

[9] Google Colab [Электронный ресурс]. – URL: https://colab.research.google.com

Segmentation of unstructured text on the book cover images using the convolutional network based on the U-Net architecture

P.L. Nikolaev

***Abstract* — This paper discusses the convolutional neural network for image segmentation with book covers. The structure of the network is given, indicating all its constituent blocks and layers, as well as their parameters, and the operating principle of each part is described in detail. The U-Net model is used as the basis of the network. The architecture of this model stands out among others with its encoder-decoder structure, which allows generating new images. In this case, the encoder part of the network is responsible for image recognition, and the decoder part is responsible for generating a new image. The proposed neural network is capable of creating binary (black and white) masks, on which the text is highlighted in one color, and all other elements in another. Thus, the text is separated from other elements in the image. To train and test the convolutional neural network, the self-assembled and labeled dataset of 200 examples is used. Despite the small amount of data, the U-Net-based network trains well and shows acceptable performance results, which is confirmed by the test results. The trained network can be used in practice. In particular, it is supposed to be used to improve the accuracy of text recognition on book covers.**

***Keywords* — Artificial neural networks, machine learning, deep learning, convolutional neural networks, image segmentation, image binarization, text recognition, U-net.**

REFERENCES

[1] L. Shapiro, G. Stockman, Kompyuternoe zrenie in Moscow, Russia: BINOM. Laboratoriya znanij (In Russian), 2013.

[2] A. Guatam, "Segmentation of Text from Image Document", International Journal of Computer Science and Information Technologies, vol. 4, no. 3, pp. 538-540, 2013.

[3] M. Grzegorzek, C. Li, J. Raskatow, D. Paulus, N. Vassilieva, “Texture-Based text detection in digital images with wavelet features and Support Vector Machines,” in Proceedings of the 8th International Conference on Computer Recognition Systems CORES 2013, pp. 857-866, 2013. DOI: 10.1007/978-3-319-00969-8_84.

[4] V.P. Le, N. Nayef, M. Visani, J. M. Ogier, C.D. Tran, “Textand non-text segmentation based on connected component features,” in Proceedings of the 2015 13th International Conference on Document Analysis and Recognition(ICDAR), pp. 1096–1100, 2015. DOI:10.1109/ICDAR.2015.7333930.

[5] R.R. Nair, B. U. Kota, I. Nwogu, and V. Govindaraju, “Segmentation of highly unstructured handwritten documents using a neural network technique,” in Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 1291–1296, 2016. DOI:10.1109/ICPR.2016.7899815.

[6] P.V. Bezmaternykh, D.A. Ilin, D.P. Nikolaev, “U-Net-bin: hacking the document image binarization contest”, Computer Optics, vol. 43(5), pp. 825–832, 2019. DOI: 10.18287/2412-6179-2019-43-5-825-832.

[7] O. Ronneberger, P. Fischer,T. Brox, “U-net: Convolutional networks for biomedicalimage segmentation,” in Proceedings. of MICCAI 2015: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241, 2015. DOI: 10.1007/978-3-319-24574-4_28.

[8] P.L. Nikolaev, “Analysis of human activity by deep learning,” in Sistemnyj administrator, vol. 12 (193), pp. 80-83, 2018. (In Russian)

[9] Google Colab, Available at: https://colab.research.google.com