Text Scene Detection with Transfer Learning in Price Detection Task

Vladimir Fomenko, Dmitry Botov, and Julius Klenin

Chelyabinsk State University, Chelyabinsk, Russia

ironyship@gmail.com

Abstract. This paper discusses the use of the transfer learning method in a text scene detection task. The transfer learning is an effective method in image analysis tasks, in particular, classification and object detection. Nevertheless, the application of this approach in combination with various methods for detecting text scenes almost is not described. The experiment is conducted to transfer knowledge about text detection to the detection for price tags using Fully Convolutional Network. COCO-Text, based on the MS COCO dataset, ImageNet dataset and dataset made up of a large number of images of the prices of various stores obtained during monitoring are taken as base datasets for the transfer learning. The target dataset is compiled from photographs of price monitoring with price tags and prices marked on them. The results of the experiment show how the application of transfer learning affects the training speed of a FCN in the task of detection price tags and prices for each of the base datasets.

Keywords: Text Scene Detection, Transfer Learning, Convolutional Neural Networks, Semantic Segmentation

1 Introduction

Monitoring of prices allows for effective pricing based on the prices of competitors. At the moment, many retail chains conduct the price monitoring process in the following way: low-qualified personnel visits competitors' shops, photographs the goods and price tags of the monitored goods, and writes down information about the goods and prices to the database, and then other employees verify the correctness of the information entering into the database. Every day more than one million photographs are received for monitoring, and therefore this process does not allow to quickly analyze the prices of competitors and conduct pricing based on this information.

The task of price tag recognition aims to significantly speed up the process of monitoring the prices of competitors by reducing the number of photos being checked. The whole task is performed in two stages: the localization of the price tag and the price and the recognition of the name of the product and its price [15]. Localization can be performed by methods based on object detection and

semantic segmentation. Recognition of the name of the products can be carried out by means of the classification of sentences or words or methods of OCR [5]. Price recognition can only be performed by OCR methods. There are also end-to-end recognition methods [12], but they require significant computational resources.

The collection and labelling of the dataset for the detection of price tags and prices is quite complex and expensive, and therefore this article proposes the use of the method of transfer learning to solve the problem of finding price tags.

In the transfer learning, two approaches can be distinguished: the transfer learning from a similar task and the transfer learning from a similar domain [1]. We can assume that the task of finding the price tag is similar to the task of finding the text and dataset can be used to search for text, such as COCO-Text [11]. Considering that retail chains collect more than a million photographs every day with labeling of product names, we can take a model trained to classify goods by photo and use it as a base model for finding price tags.

2 Related works

There are two approaches to finding text in an image. The first approach involves methods based on region proposals. State-of-the-art region proposals methods such as R-CNN in its pure form are not suitable for text searching because the design of the anchor box is not suitable for a large aspect ratio of text strings [13]. There are methods that solve this problem by using long anchor boxes [6] or the Region Proposal Network [10]. At the moment, such methods are working well only with horizontal text.

The second approach includes methods based on segmentation, such as Text-Block FCN [14], which is a modification Fully Convolutional Networks [7]. These methods allow you to do per-pixel prediction of finding text in the image and do not have problems with the text of irregular shape, but they require time-consuming process of separation to get the result on the words.

Moreover, hybrid methods are emerging that combat the shortcomings of both approaches [2].

In recent years, the transfer learning method has been applied to most image analysis tasks, such as the classification [8] and the objects detection [1]. The transfer learning makes it possible to accelerate the process of training computer vision models at times. There are two approaches to the transfer learning: based on a similar problem and based on a similar domain. Many deep learning methods for imaging tasks are used as a base architecture methods that show high results in the ImageNet competition [4].

3 Method

For the experiment, the Fully Convolutional Network method was chosen. This method solves the problem of semantic segmentation. The key feature of this

method is that the architectures implementing this method do not contain any fully connected layers (see Fig. 1) [7].

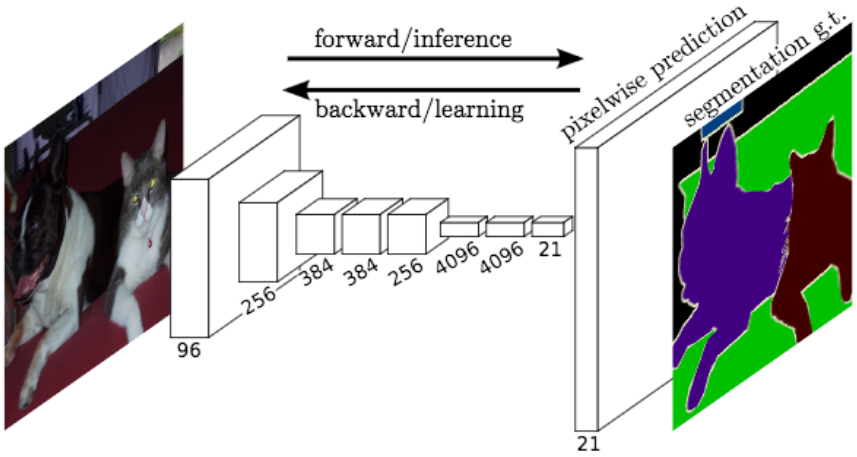

Fig. 1. Example of a Fully Convolutional Network architecture

At the input, such a neural network receives an image and an output image is created with the number of channels equal to the number of predicted classes, where each channel is a binary mask showing where the object of the corresponding class is on the image. In the original article [7], the best results were shown by an architecture based on the architecture of VGG16 [9], which showed the best results in the ImageNet contest in 2014.

To convert VGG16 architecture into FCN, the following operations were performed: fully connected layers, flatten layer and last max pooling layer have been removed; convolutional layer with a kernel size of 1x1 , the number of filters equal to 128 and the relu activation function, upsampling layer increases the output of the last convolutional layer 16 times and convolution layer with a kernel size of 3x3, a number of filters equal to 2, and a sigmoid activation function have been added. The Adam optimizer was used with the parameters: learning rate is 10^{-3}, β_1 is 0.9, β_2 is 0.999, ϵ is 10^{-7}.

In this article, in addition to VGG16, the use of the MobileNet [3] architecture is proposed. To convert MobileNet architecture into FCN, the following operations were performed: a fully connected layer and a global average pooling layer have been removed; were then added convolutional layer with a kernel size of 1x1 , the number of filters equal to 128 and the relu activation function, upsampling layer increases the output of the last convolutional layer 32 times and convolution layer with a kernel size of 3x3, a number of filters equal to 2, and a sigmoid activation function have been added. The Adam optimizer was used with the parameters: learning rate is 10^{-4}, β_1 is 0.9, β_2 is 0.999, ϵ is 10^{-7}.

The transfer learning was carried out as follows: the weights of the target network were initialized by the weights of the networks trained on other tasks, then the target network was trained on the target dataset.

The training was terminated by the early stopping method: the training stopped when there was no improvement in the metric on the validation set for five epochs.

The following data augmentations were used: image rotation from -10 to 10 degrees, shift from 0 to 0.1 in any direction, zoom from 0.9 to 1.1.

4 Dataset

4.1 Base datasets

The transfer learning was conducted by the following datasets.

1. ImageNet. Dataset used in the annual competition for pattern recognition. At the moment it contains 14,197,122 images containing 21,841 categories.

2. COCO-Text. On this dataset, the problem of text search was solved. It is made up of pictures included in the MS COCO dataset of photos containing text. It contains 63,686 images with 145,859 text instances.

3. Dataset collected from the price monitoring pictures. This dataset was used to solve the problem of classification of products from the photo. It contains about 500,000 photos of goods divided into 150 classes.

4.2 Target dataset

The target dataset consists of 10,642 price monitoring photographs of more than 10 stores, on which there is one of six types of goods and a price tag. For each photo, the areas in which the price tag and price are located are labeled. In the photo, there may be more than one price tag and more than one price on the price tag and in such cases a price tag that corresponds to the product in the photo was chosen, and the price was selected based on the expert's opinion (see Fig. 2).

5 Evaluation

5.1 Metric

The Jaccard coefficient metric, also known as Intersection over Union, was chosen, which is calculated as the intersection of the predicted and true areas of the object's location by the image divided by their union.

Fig. 2. Example of markup of the target dataset

5.2 Results

As shown in Table 1, using the transfer learning, it was possible to increase the speed of training and it was not possible to improve the quality of models. The greatest increase in the speed of training was due to the transfer learning based on one subject area. The transfer learning based on the same task did not give a gain in speed in comparison with the transfer learning based on the ImageNet dataset.

Table 1. The results of the experiment for Fully Convolutional Network models based on the VGG16 and MobileNet architectures.

	MobileNet		VGG16	
Base dataset	Mean IOU	Number of epochs	Mean IOU	Number of epochs
Without a base dataset	0.9355	42	0.9603	43
ImageNet	0.9356	26	0.9598	20
COCO-Text	0.9354	23	0.9595	24
Price monitorings	0.9354	9	0.9595	16

6 Conclusion and Future Work

Based on the results of the experiment, it can be concluded that the use of the transfer learning method makes it possible to make learning process several times faster. The best approach was the approach of transferring learning from the tasks of same subject area. In our experiment, one epoch of training took 5 minutes and we managed to reduce the training of models from 3.5 hours to 45 minutes. Acceleration of training models will allow us to more effectively test models with different parameters.

In this paper, the application of the transfer learning to the Fully Convolutional Network method, which is based on segmentation, was considered. In the future, it is planned to explore the effectiveness of the transfer of learning for methods based on region proposals. In addition, in the future, the domain's dataset will grow to more than a million images divided into more than 1000 classes, which, perhaps, will further accelerate the training process for finding price tags and prices.

References

1. Aytar, Y.: Transfer Learning for Object Category Detection. `http://www.robots.ox.ac.uk/~vgg/publications/2014/Aytar14a/` (2014)
2. He, W., Zhang, X., Yin, F., Liu, C.: Deep Direct Regression for Multi-Oriented Scene Text Detection. arXiv preprint arXiv:1703.08289 (2017)
3. Howard, A., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H.: MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. aeXiv preprint arXiv:1704.04861 (2017)
4. Huh, M., Agrawal, P., Efros, A.: What makes ImageNet good for transfer learning? arXiv preprint arXiv:1608.08614 (2016)
5. Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading Text in the Wild with Convolutional Neural Networks. arXiv preprint arXiv:1412.1842 (2014)
6. Liao, M., Shi, B., Bai, X., Wang, X., Liu, W.: TextBoxes: A Fast Text Detector with a Single Deep Neural Network. arXiv preprint arXiv:1611.06779 (2017)
7. Long, J., Shelhamer, E., Darrell, T.: Fully Convolutional Networks for Semantic Segmentation. arXiv preprint arXiv:1605.06211 (2014)
8. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. CVPR'14 Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1717-1724 (2014)
9. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv preprint arXiv:1409.1556 (2014)
10. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting Text in Natural Image with Connectionist Text Proposal Network. arXiv preprint arXiv:1609.03605 (2016)
11. Veit, A., Matera, T., Neumann, L., Matas, J., Belongie, S.: COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. arXiv preprint arXiv:1601.07140 (2016)
12. Wojna, Z., Gorban, A., Lee, D., Murphy, K., Yu, Q., Li, Y., Ibarz, J.: Attention-based Extraction of Structured Information from Street View Imagery. arXiv preprint arXiv:1704.03549 (2017)
13. Xing, D., Li, Z., Chen, X., Fang, Y.: ArbiText: Arbitrary-Oriented Text Detection in Unconstrained Scene. arXiv preprint arXiv:1711.11249 (2017)
14. Zhang, Z., Zhang, C., Shen, W., Yao, C., Liu, W., Bai, X.: Multi-Oriented Text Detection with Fully Convolutional Networks. arXiv preprint arXiv:1604.04018 (2016)
15. Zhu, Y., Yao, C., Bai, X.: Scene text detection and recognition: recent advances and future trends. Frontiers of Computer Scienc February 2016, Volume 10, Issue 1, pp. 19-36 (2016)