ДОДАТОК А

Слайди презентації

Презентація атестаційної роботи
магістра на тему:

Дослідження методів комп’ютерного зору для адаптації домену попередньо натренованих моделей

Студента групи ІПЗм-18-3 Юсіфов Р. Г.
Науковий керівник: доц. каф. ПІ Турута О. П.

Актуальність роботи

Існує ряд задач комп`ютерного зору, для яких відсутні дані для навчання:
Дані важко отримати (дорожні аварії), тоді застосовують синтетичні дані.
Дані не розмічені (відео з регістраторів), тоді дані розмічуються власноруч.

Обидва рішення потребують ресурсів і часу, застосовувати їх для кожної нової задачі неефективно.

Альтернатива – використати досвід моделі при вирішенні однієї задачі для іншої схожої задачі.

2

Семантична сегментація

Людина

Велосипед

Фон

Семантична сегментація в реальних умовах

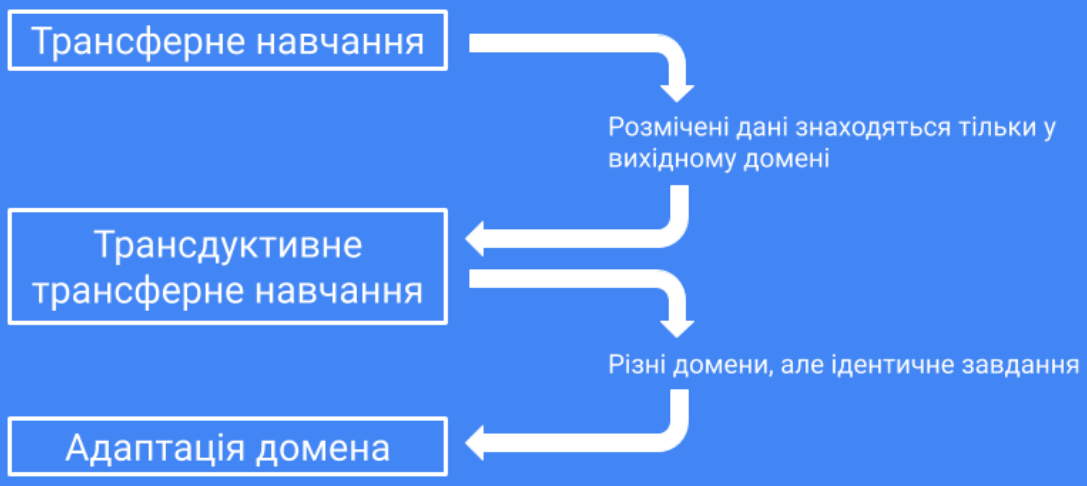
No More Discrimination: Cross City Adaptation of Road Scene Segmenters

В реальних умовах модель демонструє результат гірший на 25-30%

Створення нової моделі займає багато часу

Вихід: адаптація моделі до нового домену

Адаптація домену



Постановка задачі

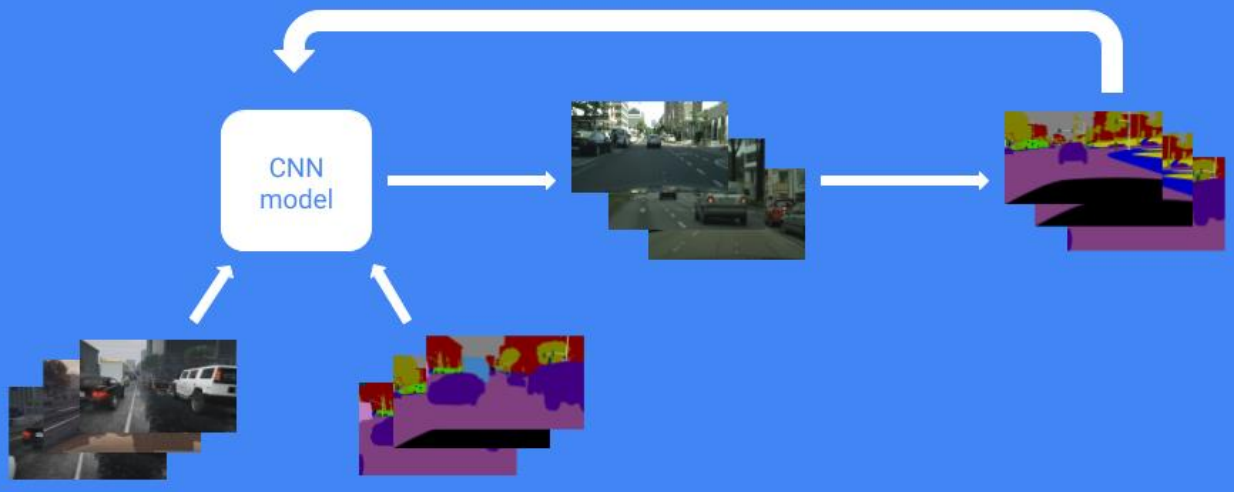
- розглянути методи доменної адаптації для семантичної сегментації, які виникають при цьому проблеми, як ці проблеми вирішують, шляхи вдосконалення;
- запропонувати власний метод або удосконалення існуючого, експериментально перевірити ефективність запропонованого метода/вдосконалення при адаптації домену між різними датасетами (GTA, Cityscapes, SYNTHIA).

Датасети

	SYNTHIA	GTA	Cityscapes
Тип даних	синтетичні	синтетичні	реальні
Класів	13	18	30
Зображень	200 000	24 966	25 000
Якість анотацій	висока	висока	висока / низька

8

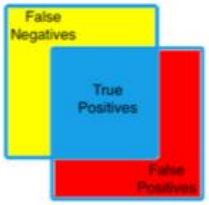
Адаптація домену з генерацією псевдоміток



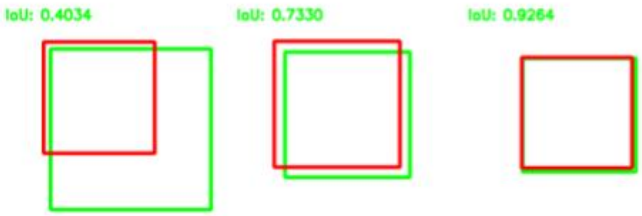
9

Метрика для семантичної сегментації

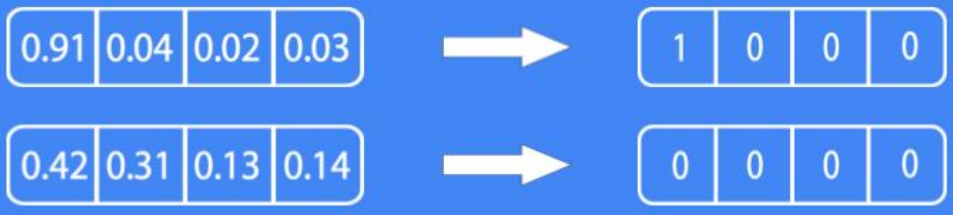
$$IoU = \frac{TP}{(TP + FP + FN)}$$



де *TP* – кількість правильно розпізнаних пікселей,
FP – кількість пікселей помилково віднесених до об`єкту,
FN – кількість пікселей помилково вилучених з об`єкта



10



$$\hat{y}_{t,n}^{(c)*} = \begin{cases} 1, if\ c = \arg\max p_n(c|w, I_t), \\ p_n(c|w, I_t) > \exp(-k) \\ 0, \quad otherwise \end{cases}$$

Значення менше exp(-k) не приймають участі у адаптації моделі

11

Адаптація домену для SYNTHIA -> Cityscapes

	Дорога	Тротуар	Будівля	Стіна	Паркан	Стовп	Світлофор	Дорожній знак
Без адаптації	0.172	0.197	0.473	0.011	0.0	0.191	0.03	0.091
З адаптацією	0.002	0.145	0.538	0.016	0.0	0.189	0.009	0.078

	Рослини	Небо	Людина	Їздок	Машина	Автобус	Мотоцикл	Велосипед
Без адаптації	0.718	0.783	0.376	0.047	0.422	0.09	0.001	0.009
З адаптацією	0.722	0.803	0.481	0.063	0.677	0.047	0.002	0.045

12

Отримання псевдоміток з балансуванням класів

Проблема: при навчанні за псевдомітками модель “схиляється” до певних класів і ігнорує інші.

$$\hat{y}_{t,n}^{(c)*} = \begin{cases} 1, if\ c = \arg\max \dfrac{p_n(c|w, I_t)}{\exp(-k_c)}, \\ p_n(c|w, I_t) > \exp(-k_c) \\ 0, \quad otherwise \end{cases}$$

13

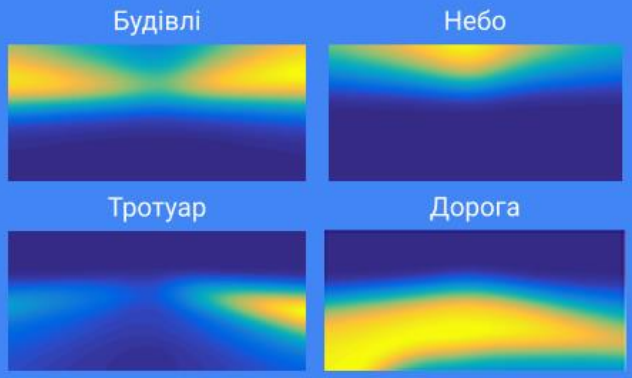
Адаптація домену SYNTHIA -> Cityscapes з урахуванням балансування класів

	Дорога	Тротуар	Будівля	Стіна	Паркан	Стовп	Світлофор	Дорожній знак
Без адаптації	0.172	0.197	0.473	0.011	0.0	0.191	0.03	0.091
З адаптацією	0.002	0.145	0.538	0.016	0.0	0.189	0.009	0.078
БК	0.696	0.287	0.695	0.121	0.001	0.254	0.119	0.136

	Рослини	Небо	Людина	Їздок	Машина	Автобус	Мотоцикл	Велосипед	mIoU
Без адаптації	0.718	0.783	0.376	0.047	0.422	0.09	0.001	0.009	0.262
З адаптацією	0.722	0.803	0.481	0.063	0.677	0.047	0.002	0.045	0.278
БК	0.82	0.819	0.491	0.145	0.66	0.066	0.037	0.324	0.331

14

Використання подібностей між датасетами



Використаємо особливості зображень (дорога внизу, небо зверху, будівлі по краях, тощо...)

15

Адаптація домену для GTA -> Cityscapes з картами класів

	Дорога	Тротуар	Будівля	Стіна	Паркан	Стовп	Світлофор	Дорожній знак	Вантажівка
Без адаптації	0.44	0.121	0.386	0.133	0.087	0.199	0.155	0.059	0.061
Без карт	0.567	0.168	0.437	0.148	0.095	0.283	0.259	0.101	0.037
З картами	0.581	0.308	0.441	0.183	0.095	0.272	0.286	0.141	0.059

	Рослини	Небо	Людина	Їздок	Машина	Автобус	Мотоцикл	Велосипед	Поїзд	mIoU
Без адаптації	0.449	0.37	0.371	0.103	0.312	0.012	0.108	0.029	0.018	0.209
Без карт	0.455	0.416	0.372	0.062	0.319	0.022	0.108	0.324	0.054	0.253
З картами	0.464	0.481	0.326	0.145	0.414	0.125	0.189	0.286	0.012	0.281

16

Висновки

Запропонований метод застосовує карти частот класів для балансування класів. Має просту архітектуру і порівняно швидко навчається. Це означає, що його можна обирати у якості “baseline” при вирішенні задач доменної адаптації для семантичної сегментації.

Подальші дослідження:

- застосувати метод для інших задач: виявлення об’єктів, сегментація екземплярів;
- дослідження гіперпараметрів для балансування класів;
- дослідження застосувань карт частот класів;
- провести дослідження запропонованого методу з рядом моделей нейронних мереж, датасетів і метрик.

Дякую за увагу!

ДОДАТОК Б

Апробація результатів роботи

Short significant image description

Oleksii Turuta
Computer Science Faculty
Kharkiv National University of Radio Electronics (NURE)
Kharkiv, Ukraine
oleksii.turuta@nure.ua

Ramal Yusifov, Yana Daniiel
Department of software engineering
Kharkiv National University of Radio Electronics (NURE)
Kharkiv, Ukraine
{ramal.yusifov, yana.daniiel}@nure.ua

Abstract— **The goal is to capture a short description of the image, to detect essential objects and speech or to react on them. For instance, to support blind people / visually impaired people, an autonomous car has to recognize environment, and robot has to do some actions. For example, to assist local navigation for blind or visually impaired people by providing (1) generic feedback about essential objects and obstacles and (2) description of identified useful objects. This appears to be a very challenging problem because the existing methods retrieve more information than we need at the moment.**

We offer a method of describing essential objects that are received in a real time. The existing methods of segmentation and recognition of objects allocate more information than necessary (find the right sign, see the traffic light signal, bypass the obstacle, pay attention). It is proposed to generate activity profiles and to form training datasets for each profile. The dataset is planned to be based on the results of segmentation and image recognition using pre-trained networks and additional information.

Keywords—object recognition, image capturing, computer vision, deep CNN

I. Introduction

In this paper, we focus on the problem of retrieval of short capturing of image, short scene of video, that contains essential objects. The general problem is that there is a need to extract objects sorted in order of importance from the video or to classify the objects in order of importance. A short description of the image has to be received to read the names of objects visible in the image. If a large number of objects is recognized, it will be impossible to read aloud them online. Therefore, it is necessary to determine the list of essential objects.

For object recognition we are using the existing pretrained networks, so we are focused more on classification than on detection. We consider small video fragments in which pretrained networks recognize objects, and then automatically generated metadata is added. The dataset is temporarily annotated at the frame level of 4 possible classes. Finally, we have to prepare online descriptions of essential objects, which can be read aloud online.

We introduce the dataset based on combination of two datasets - COCO[1] and a custom dataset. The dataset consists of 1,301 frames annotated with 4 classes. We compare different methods to retrieve the objects. As a result, our model allows to select essential objects for a particular profile. We create a short description of the image based on the importance of the selected objects. If necessary, the user receives warnings or sound messages about the collected information.

In the next section, we discuss related papers and work, in section 3 the challenging factors, dataset formation features and our assumptions are described, the section 4 is devoted to recognition of essential objects, formation of custom profiles, testing of the network and its adaptation to real world conditions.

II. Related work

Object detection. Computer vision can be used to enable purposeful navigation and object identification. In the paper [2] author suggests using QR-codes to denote objects. This method is efficient and allows easily and accurately identify the object. However, it is impossible to apply QR-codes to all objects in the real world.

To recognize objects in the image as B-bounding boxes, R-CNN[3], Fast-CNN[4], YOLO[5] are used. Segmentation methods are based on datasets with pixel-based markup and get other models that segment the object[6]. These models also use synthetic data for learning CNN, for example, ROAD[7].

Image to Speech converter. In the paper [8] it is proposed to choose one object, generate its name or its corresponding sound and then use stereo sounds to create the effect of sounding of an object from its natural environment. However, if you playback the sounds of all objects one after one, the user faces an information overload. This is caused by the fact that system tries to notify the user about a big number of recognized objects.

III. System description

A. Challenge

The existing methods of object recognition allow to generate descriptions that consist of large numbers of objects. Due to this fact, there is a need to determine only important objects and notify the user about them. For example, 5-7 objects were recognized in the frame. These objects will be relevant about 4-5 seconds of video, so during this time it will be necessary to read aloud their names. For this time in total, the list has replenished to 15 objects. So the user will definitely

face the "information overload" caused by the fact, that the system tries to inform the user about all the recognized objects at the same time.

Obviously, reading the names of 15 objects within 5 seconds is impossible. Therefore, it is necessary to determine and choose essential objects to be pronounced.

B. Dataset

In this work, we used the COCO dataset[1] and prepared our dataset annotated 1,301 frames with 4 classes. We define 4 actions. Annotations were done manually and afterwards. The video is ripped at a frame rate of 25fps and has a resolution of 240 by 320 pixels. About 20 hours of video are annotated in our dataset. The concept of a hierarchy of objects is introduced, for example, traffic light and signals, road sign-indicator and text inside of it. Classes of objects with the status are determined - the color of signal of the traffic light, the content of the pointer. As a result, an instance map of recognized objects with automatically generated metadata is formed. We divide the frames over training, validation and testing sets. Every set contains at least 10% frames for each class.

C. Assumptions

It is assumed, that there is an indoor environment, which main scene surfaces follow the Manhattan world assumption [9]. All the objects of interest are adjacent to the ground plane. These objects can be recognized using pretrained models on the basis of their distinct geometry. The recognized objects are divided into several classes and include objects like “chair,” “table,” “stair up,” “stair down,” and “wall”. We suppose that the ground plane can be initially observed and that the height and rotation of the sensor remain constant when the ground plane cannot be seen. To successfully perform the computations, a board with real-time detection and feedback is used.

IV. FUNDAMENTAL

A. ESSENTIAL OBJECT RECOGNITION

The selection of the essential object is performed in two stages. At the first stage, an excessive number of objects is allocated using the YOLO algorithm[5]. Then we increase the value of threshold, which leads to an increase in the list of certain objects. The next step is to select the essential object. For this we use two types of classifiers. The first one is based on GoogLeNet [10], the second one is made on the basis of the SVM method.

We have developed three profiles to evaluate the performance of classifiers. The results of comparing performance of classifiers are shown in Table 1. Each profile indicates a specific task. Profile 1 is designed to identify dynamic obstacles. Profile 2 aims to highlight the tablets and determine the priority of reproducing the inscriptions. Profile 3 is made to determine a free chair. The results in Table I allow us to draw conclusions about the effectiveness of using the classifier based on GoogLeNet to determine dynamic obstacles. At the same time, SVM effectively determines the essential objects for playback.

TABLE I. EVALUATION OF THE EFFECTIVENESS OF SPECIFIC TASKS

Methods	Profiles		
	Profile 1	*Profile 2*	*Profile 3*
CNN	0.78	0.72	0.70
SVM	0.65	0.77	0.69

As a result, we receive profiles that allow users to solve special tasks.

B. Adaptation of the network to real weather conditions

To increase the number of items in the datasets, it is allowed to add synthetic data based on video games[11]. In this case, the first stage uses a pre-trained model, and then the model trains on the real and synthetic data. The approach that allows the Conv Layers to train on the real data is proposed in the paper[6]. Then the real data is frozen and used to train Conv Layers on the basis of synthetic data. Figure 1 shows a diagram of such a network.

Formally, for a real image, let us denote $x_{i,j}$ as the activation at the position (i, j) of the feature map from the essential recognition model, and also denote $z_{i,j}$ as an activation at the same location of the feature map from the pretrained model, then the loss for target guided distillation can be written as,

$$L_{dist} = \frac{1}{N}\sum_{i,j}\left\|x_{i,j} - z_{i,j}\right\|_2 \quad (1)$$

where N is the number of activation at the feature map, $\|\cdot\|_2$ is the Euclidian distance.

The results of object recognition based on computer vision methods essentially depend on the lighting, and for outdoor navigation, the environmental state has an additional influence. To improve generalization, we performed data augmentation and dropout. 10% of the total dataset of the videos were shot in the same places, but under different

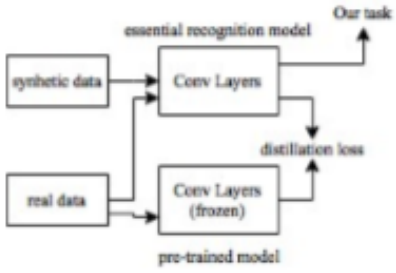


Fig. 1. Illustration of learning convolution filters on synthetic data, taking into account the pre-trained model

TABLE II. TESTING THE MODEL FOR DIFFERENT DATASETS

Method	Datasets		
	Test set	*New weather old place*	*New weather new place*
CNN	0.75	0.61	0.57

environmental conditions and at different times of the day. To solve the problem of identifying obstacles, we use the end-to-end learning approach based on imitation learning. This approach allows you to use fewer items in the dataset than in reinforcement learning. For our data set, we split the 90% of the train set and approximately 50-90 frames as test samples for each class. We train the model and perform three experiments with different types of test set (Table II).

1. Same weather same place - the test set is collected in the same weather and place, but at distinct times of the day

2. New weather old place - the test set contains items with entries made in the same places, but with other weather conditions

3. New weather new place - the test set is made elsewhere and with a different weather conditions.

As a result, we obtain an evaluation of the quality of object recognition depending on the change in location and weather.

V. CONCLUSIONS

Object recognition approaches often retrieve a huge amount of objects to create an image description. However, such a number of objects can not be read aloud online. Reading the object list takes more time than the time it is changed or supplemented. This fact makes this method of object recognition ineffective to be used to help people with visual impairment.

In this paper we have presented the approach that allows us to select essential objects to be read aloud online. For this purpose, the markup of the dataset was made with an assessment of the importance of the object. In addition, the problem of determining the obstacle and the direction of the obstacle bypass was solved.

This method was tested for urban scenes in various environmental and lighting conditions. The experiments were taken on the combination of two datasets COCO and custom dataset. We faced the problem of recognizing and classifying the same objects in the center of focus and at the edges. The created prototype successfully recognizes and sorts essential objects.

However, the prototype has several limitations. Testees are used to focusing on a certain object from afar and navigation to location close to the object, so the objects that are outside of the center of attention are often unrecognized or misclassified.

Solution of this problem has practical application. Blind people can use this approach as an assistance for navigation to obtain information about obstacles and descriptions of objects.

REFERENCES

[1] T.-Y. Lin *et al.*, 'Microsoft coco: Common objects in context', in *European conference on computer vision*, 2014, pp. 740–755.

[2] A. S. M. Yasin, M. M. Haque, S. B. Anwar, and M. S. A. Shohag, 'Computer vision techniques for supporting blind or vision impaired people: An overview', Int. J. Sci. Res. Eng. Technol., vol. 2, no. 8, pp. 498-503, 2013.

[3] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, 'Rich feature hierarchies for accurate object detection and semantic segmentation', *CoRR*, vol. abs/1311.2524, 2013.

[4] R. Girshick, 'Fast R-CNN', in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, 'You only look once: Unified, real-time object detection', in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[6] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, 'One-shot video object segmentation', in *CVPR 2017*, 2017.

[7] Y. Chen, W. Li, and L. Van Gool, 'ROAD: Reality Oriented Adaptation for Semantic Segmentation of Urban Scenes', *arXiv preprint arXiv:1711.11556*, 2017.

[8] S. Wu and L. A. Adamic, 'Visually impaired users on an online social network', in *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, 2014, pp. 3133–3142.

[9] J. M. Coughlan and A. L. Yuille, 'The Manhattan world assumption: Regularities in scene statistics which enable Bayesian inference', in *Advances in Neural Information Processing Systems*, 2001, pp. 845–851.

[10] C. Szegedy *et al.*, 'Going Deeper with Convolutions', in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, 'CARLA: An Open Urban Driving Simulator', in *Proceedings of the 1st Annual Conference on Robot Learning*, 2017, pp. 1–16.