РОЗДІЛ 1

МІЖНАРОДНЕ СПІВРОБІТНИЦТВО У СФЕРІ ОСВІТИ, НАУКИ І ВИРОБНИЦТВА

UDC 004.415.3:681.6

L. Becker[1], B. Moroz[2], L. Kabak[2], S. Teslenko[2]
[1]Comparus GmbH, Munich, Germany
[2]Dnipro University of Technology, Dnipro, Ukraine

ESTIMATION OF THE GEOGRAPHICAL COORDINATES OF OBJECTS ON THE IMAGE WITH MULTI-TASK CONVOLUTIONAL NEURAL NETWORKS

Abstract. Determining GPS coordinates of the objects on the image is exceptionally complex problem. Images often contain enough information such as landmarks, cloud texture, grass type, road signs or architectural features that allow suggesting the location where the photo was taken. Previously, such issue was solved with image search methods. In contrast, the problem is stated as a classification task, subdividing the Earth's surface into geographical cells using a special type of space-filling curve. Thousands of differently scaled geographical cells, used to train the model. In this paper, several deep learning methods that follow the latter approach and take advantage of multitask learning are presented. Taking into account the content of the scene of the image, i.e. inside, outside, wild or urban setting, etc. is proposed. As a result, additional information with different spatial resolutions as well as more specific features for different environments are included in the learning process of the convolutional neural network. Reported metrics demonstrate the effectiveness of our out-of-the-box approach, while using a helper network to combine two datasets combined to spread scene labels on GPS dataset and receive more robust model. This model does not rely on search methods, which require an enormous amount of computational power, and implements a probabilistic approach.

Keywords*: Deep Learning, Classification, Convolutional Neural Networks, PyTorch, Multitask learning, Space Filling Curve, Geo-location Estimation.*

Introduction. Estimation of the geographical location of the object on the image without prior knowledge is a very difficult task, as there are a huge variety of pictures taken all over the globe. Different times of the day, object location or camera characteristics makes it even harder. In addition, images are often ambiguous and therefore provide very few visual hints about the relevant recording location. For these reasons, most approaches simplify the geo-localization of the objects by limiting the problem to the city images, for example, of famous attractions and cities

or natural areas such as forests or mountains. Only a few systems have considered the general task without relying on specific imagery or any other prior assumptions. These approaches particularly benefit from advances in deep learning and the growing number of publicly available collections of geo-labelled images in such services as Instagram, Facebook, Imgur, Pinterest. Due to the complexity of the concern and the unbalanced distribution of locations from which the photo was taken across the planet, convolutional neural network (CNN)-based methods perform well on treating the geo-localization of the objects on the image as a classification task, subdividing the Earth into geographic cells with the same number of images. However, even modern CNNs are not able to memorize the visual appearance of the entire Earth and at the same time learn the scene-understanding task. The system architecture (Fig. 1) shows high-level representation of the work. Besides, geographical separation approaches entail a trade-off problem. Whereas detailed partitioning results in higher accuracy at the specific urban location scale (error is less than 1 km), wider cell partitioning increases performance at the nation scale (approx. 1000 km). On the other hand, natural scenes, such as mountains and glades or indoor scenes are more likely to be defined by features encoding plants and animals or the type of interior, respectively. Thus the geo-localization of the objects on the image can significantly benefit from additional knowledge of the surrounding scene, since the divergence in the data samples can be significantly reduced alongside with unnecessary noise.

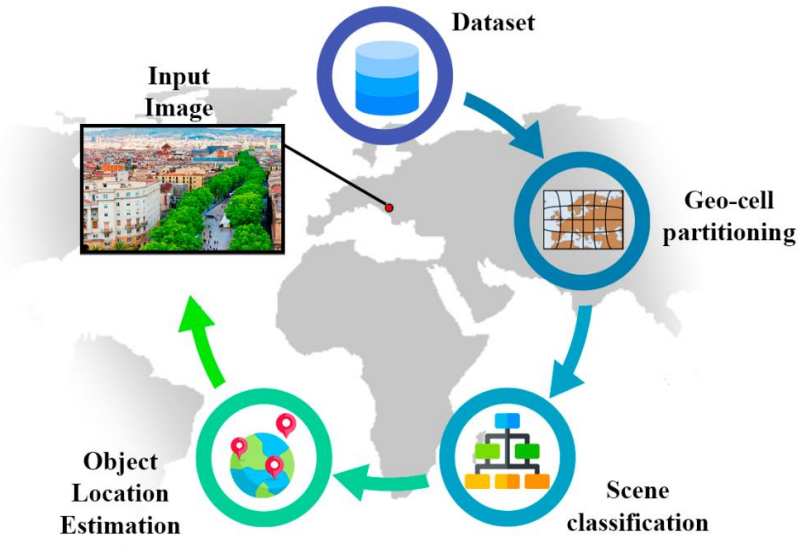

Fig. 1. High-level system architecture

This geo-estimation system can be used to solve problems in social media forensics, such as manipulation of content and metadata. Because now, thanks to the spread of social media platforms, online media can influence the millions of people in a very short period. From the other hand, attackers can easily share their photos for malicious purposes, such as creating panic or manipulating public opinion, without any particular effort.

Objective formulation: In order to achieve the objective the following tasks were introduced in this work:

- set out the principles of design and optimization of the architecture of the multitask CNN;
- implement different spatial resolutions in a multi-partition approach;
- determine the optimal way to split geo cells for classification problem;
- extract and take into account the information about the specific class of the scene;
- identify the basic, minimum necessary components of the system;
- research and apply optimal method for encoding coordinates data;
- organize training strategy for the neural network;
- draw conclusions about the feasibility of creating such a system.

Main content of the work. The following methods and tools were used during the work on this project:

- gradient descent, first-order iterative optimization algorithm for finding a local minimum of a differentiable function;
- artificial neural network (ANN) that combines biological principles with advanced statistics to solve problems in domains such as pattern recognition and game-play. ANNs adopt the basic model of neuron analogues connected to each other in a variety of ways;
- pretrained ResNet [1] architecture as a backbone for the classification;
- Python, PyTorch, Sklearn, OpenCV, PyCharm, Bash, Linux, Git.

In deep learning, a convolutional neural network is a class of artificial neural network most commonly applied to analyze visual imagery. They are also known as shift invariant or space invariant artificial neural networks (SIANN), based on the shared-weight architecture of the convolution kernels or filters that slide along input features and provide translation equivariant responses known as feature maps. Counter-intuitively, most convolutional neural networks are only equivariant, as opposed to invariant, to translation. [4-7]

The Multitask Neural Network: since the presented method of geo-location detection may not be taught efficiently due to the large variety of possible environmental representations of the area, an architecture for teaching two complementary tasks at the same time is proposed for teaching [2], [3]. More specifically, an additional fully connected layer is placed after the "global pooling" of the ResNet network (Fig. 2). The number of neurons in the final layer is the number of categories for |S| scene types. The parameters of the classifiers for both tasks are matched. As a cost function for classifying the scenes and geo-cell, the categorical cross entropy was used. Final cost function (1) L_{total} for the network training is the sum of L_{scene} and L_{geo} for both tasks.

$$L_{total} = L_{scene} + L_{geo} \quad (1)$$

These two losses have equal weight in the work. Where every part of the loss calculated separately as the categorical cross entropy (CCE), with respective number of classes (2):

$$L_{CCE} = -\sum_{i=1}^{n} y_i * \log \hat{y}_i \quad (2)$$

To reduce training time and achieve better accuracy the transfer learning (TL) was applied. This technique is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. The pretrained ResNet architecture on famous ImageNet dataset was used.

It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems.

In transfer learning, first a base network on a base dataset and task was trained, and then the learned features from ImageNet dataset were repurposed, or transferred, to a second target network to be trained to classify geo cells and scenes. This process will tend to work if the features are general, meaning suitable to both base and target tasks, instead of specific to the base task. For this research, Im2GPS and Places2 as our primary datasets were used.

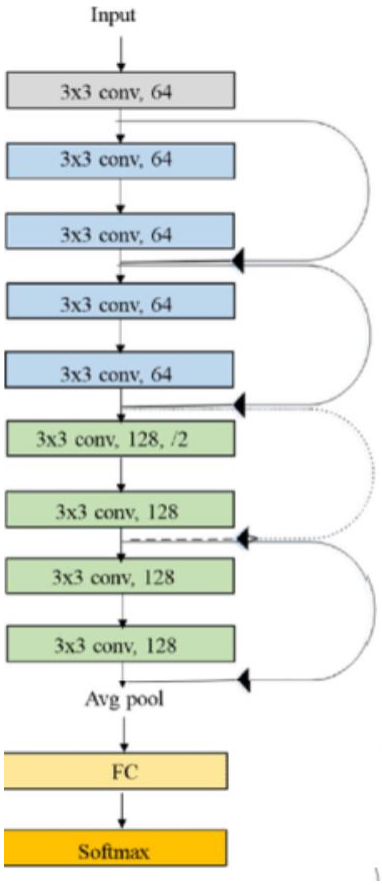

Fig. 2. CNN Architecture

The helper model was also trained to propagate labels from Places2 dataset on Im2GPS dataset by training this network to predict only class of the scene of the image. The Softmax activation function was used and the scene labels for all of the unlabeled images in Im2GPS dataset were predicted before training the MTN.

The state of art method of encoding geographical coordinates for using in a deep neural network was also introduced. The S2 geometry library is used to create a set of non-overlapping cells. In more detail, the earth's surface is projected onto the enclosing cube with six faces representing the initial cells. After that, an adaptive hierarchical division of faces based on GPS coordinates in the data set is performed, so that each cell is a node of a quad-tree. Starting from the root nodes, the corresponding quad-tree is divided recursively until all cells contain a maximum of t_{max} images. In particular, S2 cells are arranged sequentially along the space-filling curve (fractal type). The specific curve used for the S2 partition is called the S2 curve and consists of six Hilbert curves connected to form a single continuous loop across the Earth surface (Fig. 3).

Scientific novelty of the method consists in refined mathematical model for estimating of the geographical coordinates of objects on the image with multi-task convolutional neural networks and introducing helper model to deal with unlabeled parts of multi-task dataset. It is expected that this method can provide ultimate framework for media forensics and object localization.

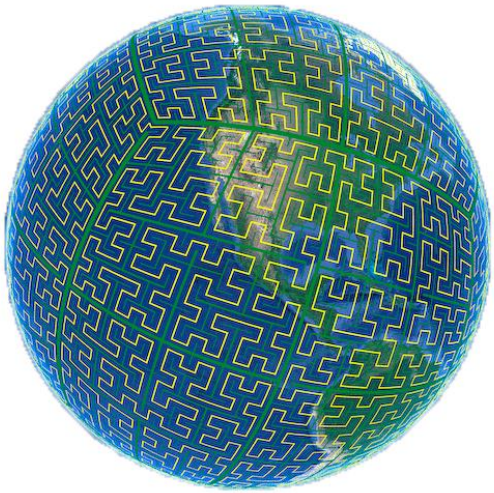

Fig. 3 Visualization of the cell division

Conclusions. In this work robust architecture for large-scale geo-location prediction of the object on specific picture by using pixel only information was developed. The teaching methods, architecture and training method demonstrated accurate result on this task. In addition, the method for treating this task as a classification with help of S2 curve geometry to represent every part of the worlds as a cell was introduced. Such a system is suitable for training on various environments and resolutions producing optimal classifier for geo-location estimation. In the future research it is planned to investigate which additional context information can help to improve the accuracy of the model.

REFERENCES

1. Kaiming H., Xiangyu Zh., Shaoqing R., Jian S. Deep Residual Learning for Image Recognition arXiv:1512.03385 (2015). [Electronic resource] - Access mode: https://arxiv.org/abs/1512.03385
2. Ruder S. An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017). [Electronic resource] - Access mode: https://arxiv.org/abs/1706.05098
3. Bishop C. M. Neural networks for pattern recognition. Oxford university press, 1995. 482p.
4. Everingham M., Van Gool L., Williams C. K., Winn J., Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. IJCV, pages 303–338, 2010.
5. Hinton G. E., N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing coadaptation of feature detectors. arXiv:1207.0580, 2012). [Electronic resource] - Access mode: https://arxiv.org/abs/1207.0580
6. Jegou H., Perronnin F., Douze M., Sanchez J., Perez P., Schmid C. Aggregating local image descriptors into compact codes. TPAMI, 2012.
7. Krizhevsky A., Sutskever I., Hinton G. Imagenet classification with deep convolutional neural networks. In NIPS, 2012.