UDK 004.93.1

Borys I. Tymchenko[1], Institute of Computer System, E-mail: tim4bor@gmail.com, ORCID: 0000-0002-2678-7556
Anhelina A. Hramatik[1]**,** Institute of Computer System, E-mail: angelinagramatik08@gmail.com, ORCID: 0000-0003-4569-1637
Heorhii P. Tulchiy[1], Institute of Computer System, E-mail: gearty@rambler.ru
ORCID: 0000-0001-8396-1672
Svitlana G. Antoshchuk[1], Doctor of Technical Sciences, Professor, Director of Institute of Computer System, E-mail: asg@opu.ua, ORCID: 0000-0002-9346-145X
[1]Odessa National Polytechnic University, Shevchenko Avenue, 1, Odessa, Ukraine, 65044

CLASSIFYING MIXED PATTERNS OF PROTEINS IN MICROSCOPIC IMAGES WITH DEEP NEURAL NETWORKS

Abstract. *Nowadays, accurate diagnosis of diseases, their treatment and prognosis is a very acute problem of modern medicine. By studying information about human proteins, you can identify differentially expressed proteins. These proteins are potentially interesting biomarkers that can be used for an accurate diagnosis, prognosis, or selection of individual treatments, especially for cancer. A surprising finding from this research is that we have relatively few proteins that are tissue specific. Almost half of all proteins are categorized as housekeeping proteins, expressed in all cells. Only 2,300 proteins in the human body have been identified as tissue enriched, meaning they have elevated expression levels in certain tissues. Thanks to advances in high-throughput microscopy, images are generated too quickly for manual evaluation. Consequently, the need for automating the analysis of biomedical images is as great as ever to speed up the understanding of human cells and diseases. Historically, the classification of proteins was limited to individual patterns in one or more cell types, but in order to fully understand the complexity of a human cell, models must classify mixed patterns according to a number of different human cells. The article formulates the problem of image classification in medical research. In this area, classification methods using deep convolutional neural networks are actively used. Presented article gives a brief overview of the various approaches and methods of similar research. As a dataset was taken "The Human Protein Atlas", that presents a tissue-based map of the human proteome, completed in 2014 after 11 years of research. All protein expression profiling data is publicly accessible in an interactive database, enabling tissue-based exploration of the human proteome. It was done an analysis of the work and the methods that were used during the research. To solve this problem, the deep neural network model is proposed taking into account the characteristics of the domain and the sample under study. The neural network model is based on Inception-v3 architecture. Optimization procedure contains combination of several tweaks for fast convergence: stochastic gradient descent with warm restarts (learning rate schedule for exploring different local minima), progressive image resizing (training starts from small resolution and sequentially increases each cycle of SGDR). We propose new method for threshold selection for F1 measure. Developed model can be used to create an instrument integrated into the medical system of intellectual microscopy to determine the location of the protein from a high-performance image.*

Keywords: *Human Protein Atlas; deep learning; neural networks; classification; pattern recognition, stochastic gradient descent*

Introduction

Proteins are the doers in the human cell, executing many functions that together enable life. This includes maintaining the structure (cell shape), chemical catalysis, and motor function (muscle contraction, for example), and transport (say, hemoglobin protein transports oxygen from the lungs to the tissue and carbon dioxide in the opposite direction) and complex regulatory functions to maintain consistency internal environment (e.g. protein hormones and all intracellular regulatory systems) and many others.

Historically, classification of proteins has been limited to single patterns in one or a few cell types, but in order to fully understand the complexity of the human cell, models must classify mixed patterns across a range of different human cells.

Images visualizing proteins in cells are commonly used for biomedical research, and these cells could hold the key for the next breakthrough in medicine.

However, thanks to advances in high-throughput microscopy, these images are generated at a far greater pace than what can be manually evaluated. Therefore, the need is greater than ever for automating biomedical image analysis to accelerate the understanding of human cells and disease [1], which leads to accurate diagnosis, and selection of individual treatments, especially in cancer treatment.

As making diagnosis of diseases, prescribing their treatment and prognosis more accurate is a very acute problem of modern medicine, studying information about human proteins is a very important task.

The Human Protein Atlas (HPA) provides high-resolution insights into the spatio-temporal distribution of proteins within human cells. The protein localization data is derived from antibody-based profiling by immunofluorescence confocal microscopy, using a panel of 64 cell lines to represent various

cell populations in different organs and tissues of the human body [11].

The highly promising technology that is nowadays used to localize proteins is high-throughput fluorescence microscopy imaging (HTI). This imaging technology allows a selected protein to be stained with fluorescent antibodies. Then a microscopic image of the whole cell is taken.

Together with the information of other staining, such as the Hoechst staining [12] of the cell nucleus, and the actin staining of the cytoskeleton [12], these images provide a rich source of information on the protein location. Which is very important and can be used for studying the structure of different types of proteins and therefore solving problems that need a lot of resources.

The problem is that high amount of images, their big size and complexity require high performance in localizing proteins from deep learning methods since imaging data together with the annotation by the HPA project are an exhaustive source of training data. Due to lack of resources another problem is to optimize neural network model for using as few re-sources as possible.

The Aim of this work is to create the method and implement the model that is capable of localization of target protein of interest within various cell organelles in different cell types and lines from a high-performance image using modern techniques in the field of neural networks and deep learning.

Recent research

Big leap in performance caused by convolutional neural networks (CNNs), aroused comparisons of computational methods with humans or even human experts. Esteva et al. [3] have compared CNNs with expert performance at detecting melanoma in images of skin lesions. Swamidoss et al. [14] have conducted a comparison with two human experts of human proteins classification. They worked with microscopy data on tissue level belonging to four classes. In this work, we aim at a more challenging task in which proteins have to be localized within 28 classes with multiple possible locations per sample.

Sullivan et al. proposed to combine two approaches for large-scale classification of fluorescence microscopy images. First, using the publicly available data set from the Cell Atlas of the HPA, authors integrated an image-classification task into a mainstream video game (EVE Online) as a minigame, named Project Discovery. This data was then integrated into a tool for automatic protein labeling [16].

The M-CNN model [8] was designed for phenotype classification of human cell data. The main idea of the architecture is to combine features extracted from the input at several spatial resolutions. This is achieved by scaling the original image dimensions to widths and heights. These scaled versions of the input are processed by different convolutional layers and the feature maps of the last layer, are downscaled via pooling to the smallest resolution. Then, the feature maps are concatenated, combined via 1x1 convolutions, and passed on to a fully connected layer and the output layer.

Another prominent work, designed specifically for high-throughput microscopic imagery is Convolutional Multiple Instance Learning. Authors focus on the problem of weak labels, i.e. that microscopy images not only contain cells of the target or labeled class but also cells that do not comply with the label. The authors propose to tackle this problem with multiple instance learning , where cells belonging to the class label of an image are identified automatically while the influence of other cells on the result of the model is down-weighted by using a special pooling function called noisyAND [13].

In the field of HPA classification with deep neural networks, one of the most recent state-of-the-art works is GapNet [7].

Authors propose architecture designed specifically to process high-throughput microscopy images. Authors achieve the possibility to learn from fine structures within images, as they do not have to be downscaled via a two-step approach. In a first step, an encoder consisting of several convolution layers interspersed with max-pooling layers to learn abstract features on different spatial resolutions is used.

In the second step, they reduce the feature maps from three different layers via global average pooling to a size of one pixel and concatenate the resulting feature vectors. The resulting features, representing different spatial resolutions, are then passed on to a fully connected network with two hidden layers for the final prediction.

Huang et al. proposed [9] the densely connected convolutional architecture (DenseNet). The basic idea of DenseNet is to reuse features learned on early layers of a network contain fine-grained localized information, on higher layers which have a more abstract representation of the input. This is achieved by passing feature maps of a layer to all consecutive layers. A stated benefit of this architecture is that it does not have to re-learn features several times throughout the network. Hence, the individual convolutional layers have a relatively small number of learned filters.

Human Protein Atlas Dataset

We conducted all experiments on the dataset released for the Kaggle Human Protein Atlas Image Classification Challenge [1] by the Human Protein Atlas. Challenge dataset contains more than 30,000 images taken from the Cell Atlas, which is the part of the Human Protein Atlas.

Every sample consists of four high-resolution images corresponding to the different fluorescent channels (Fig. 1).

These four channels correspond to four different filters:

- green filter for the target protein structure of interest;
- blue landmark filter for the *nucleus;*
- red landmark filter for *microtubules;*
- yellow landmark filter for the *endoplasmatic reticulum.*

The distinct patterns in the images together with the reference markers make it possible to precisely classify the spatial distribution of a protein within the cell.

The dataset has different limitations:

- The staining of target proteins in the green channel is not equally successful and differs in intensity.
- The images differ in their intensities and the target proteins are not always located the same way.
- There are morphological differences because cells in the dataset are of different types.

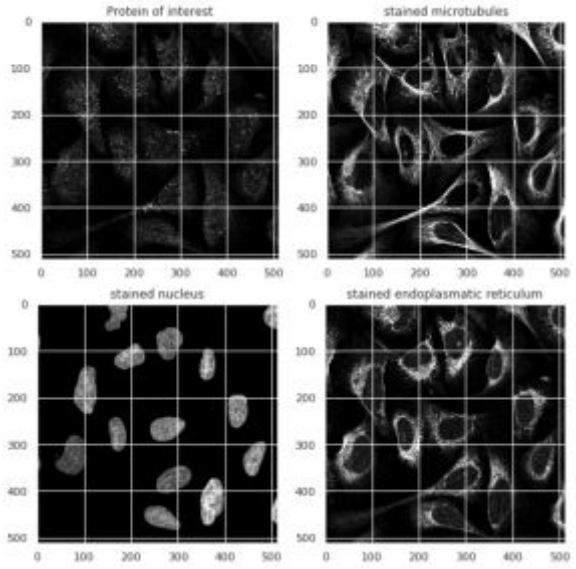

Fig. 1. The fluorescent channels of a single sample from HPA dataset

The immunofluorescence-based approach used in the Cell Atlas allows a simultaneous analysis of the protein distribution in all organelles. This enables the possibility to study the spatial distribution of proteins in their cellular context and the identification of all proteins that are located to more than one organelle, which can be called “multilocalizing proteins” [18].

For each sample, the task is to determine in which of the 28 organelles the protein of interest appears, where multiple organelles are possible as same proteins can come together in different cell organelles due to cell functions.

Dataset has highly imbalanced classes’ distribution with two orders of magnitude between most- and least-counted classes (Fig. 2).

Consequently, accuracy only is not the right score to measure the performance and fine validation strategy is needed. Metrics for this validation will be discussed later here.

We can see that most common protein structures belong to coarse grained cellular components like the plasma membrane, the cytosol and the nucleus. In contrast small components like the lipid droplets, paroxysms, endoscopes, lysosomes, microtubule ends, rods and rings are very seldom in the train data. We explore correlations of different targets in order to find common patterns in classes’ distribution. Targets have small correlations, except of lysosomes and endosomes, that occur together frequently in cell operation (Fig. 3). We can note high correlation between endosomes and lysosomes. They both contain important substances for the functioning of the cell. Furthermore, endosomes store internalized materials until their digestion while lysosomes fuse with endosomes, aiding the digestion of materials inside the endosome. Thus, protein of interest is frequently contained in both of them [21].

For these classes the prediction will be very difficult as we have only a few examples that may not cover all variability’s and model probably will be confused during learning process by the major classes. Due to this confusion, the model will make less accurate predictions on the minor classes.

Too little sample numbers in rare classes required to combine oversampling with under sampling to achieve more uniform distribution in training phase. We oversampled classes that have less than 100 samples by the factor of five. To reduce influence of dominant classes, we undersampled them with the factor of two.

We spitted preprocessed data into train (85 %) and validation (15 %) sets using iterative approach for multilabel stratification [15] in order to prevent dominance of rare classes in the train set due to its bigger percentage. As test set, we used test data and evaluation procedure that was providing by Kaggle [1] platform.

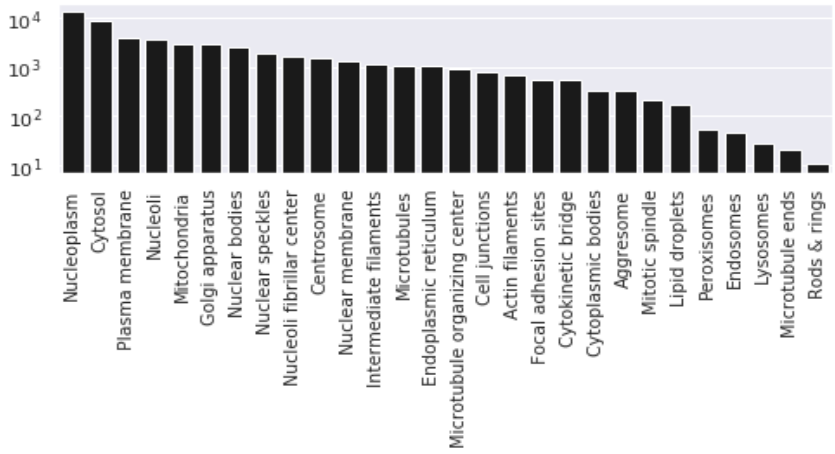

Fig. 2. Distribution of labels in dataset

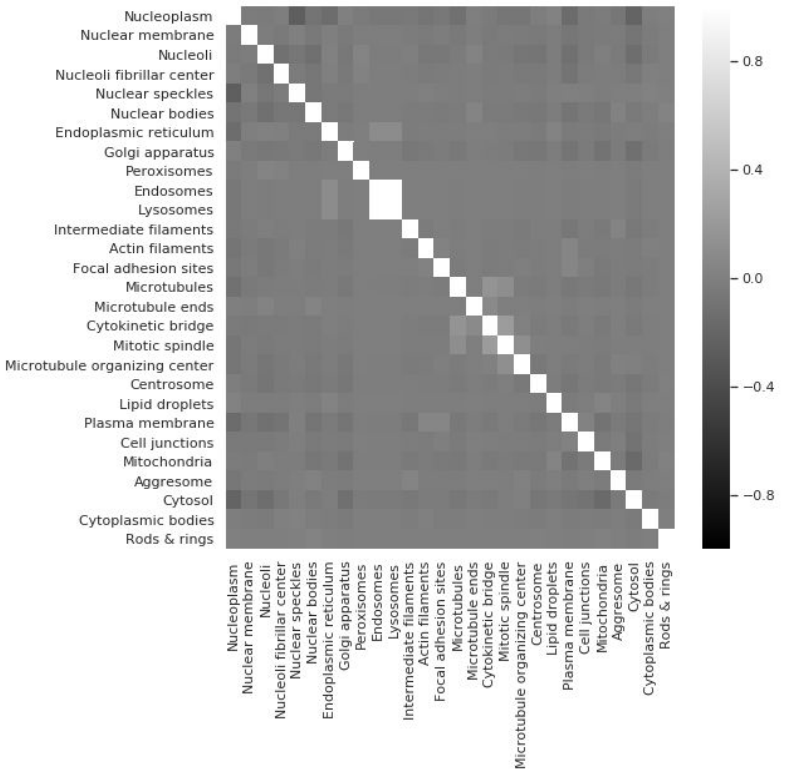

Fig. 3. Target location correlation matrix

Methods

Neural network structure

Cells in dataset can be of different sizes. To be able to deal with multi-scale recognition, we needed the appropriate network structure. Through experiments with different neural network architectures, we found Inception-v3 [4] to be most efficient in terms of capacity and computational resources. It consists of blocks that can utilize features of different scales (Fig. 4). Each block represents convolution operation with kernel size written on block. Thus, each column of operations can have different receptive fields and can grasp different scales. In addition, its balanced width/depth change in each module aid reduction of information loss in the flow, what leads to greater network capacity w.r.t. parameters number?

To make our deep network invariant to image size (which is needed for progressive training), we use global average pooling as a final layer before dense layers. It reduces spatial dimensions of the feature map of any size to one pixel.

We used the modified version of Inception-v3 modules that also includes Batch Normalization after every inception block. This addition allows much faster convergence than traditional Inception model [19].

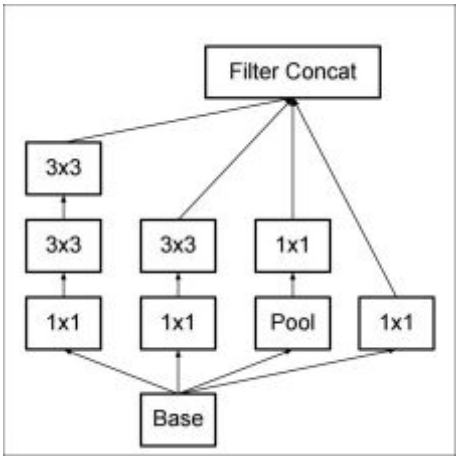

Fig. 4. Inception-v3 module

Objective function

Imbalanced dataset imposes limitations on objective functions that we could use. As the evaluation measure is F1, so we considered to be optimal to use soft-F1 loss function, which is represented by differentiable lower bound for Fbeta score [6] at fixed beta value of one.

$$F_\beta = \left(1+\beta^2\right)\frac{tp}{\left(1+\beta^2\right)\cdot tp + \beta^2 \cdot fn + fp}\ , \quad (1)$$

where: *tp*, *fn* and *fp* are true positives, false negatives and false positives respectively.

Here, these values are calculated in differentiable manner as multiplications of one-hot encoded targets and raw predictions after sigmoid output.

Progressive image resizing

To save resources while training, training starts from small resolution and sequentially increases it when validation loss plateaus. This heuristic acts as regularization at early stages forcing neural network to learn coarse features first, and then refining details only when it is really needed. In addition, it reduces training wall time significantly due to bigger batch size with lower image resolution. The batch size for model depend on memory consumption and was chosen as large as possible to fit in the GPU with 11GB memory given desired image size (Table 1).

Table 1. Progressive image size sequence

Step	Image size	Batch size
1	221x221	48
2	256x256	36
3	296x296	27
4	320x320	20
5	384x384	15
6	440x440	11
7	512x512	8

Warm restarts

Ensembles of neural networks are known to be much more robust and accurate than individual networks. To increase performance we used Snapshot Ensemble technique to create ensemble. As in original paper [10], we used cosine annealing learning rate (Fig. 5). As an optimizer, we used Stochastic Gradient Descent with momentum.

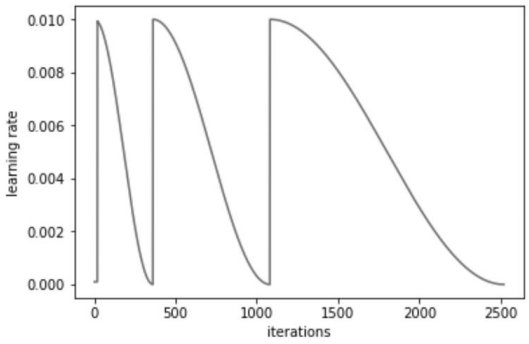

Fig. 5. Cosine annealing learning rate schedule

To stabilize training at high learning rate part of the cycle, gradient clipping was used. The gradients normalized, if their L1 norm exceeds threshold (which was set to 10.0 in experiments).

We incorporated warm restarts with progressive image resizing by increasing image size at each cycle of learning rate; such multi-scale ensemble performed better, than ensemble trained only with highest image resolution possible. We suppose this happening due to different size of different cells and different morphology of different cell lines in the dataset. Thus, neural network could learn useful features at different scales.

Threshold selection

As the problem is multi-label with severe class imbalance and correlations between several classes, we decided to select thresholds for each class individually. We chose threshold with linear search for each class while keeping thresholds for other classes fixed. This optimization could increase F1 score for the same network up to 22 %.

Other researchers mentioned that such approach could lead to poor generalization [2]. To increase generalization, we searched for thresholds that not only maximize F1 score, but also keep precision and recall close to each other to increase stability (Fig. 6).

We added additional penalty δ to linear search, equal to:

$$\delta = -\alpha \cdot P - R\ , \quad (2)$$

where: α is a small number, *P* is precision and *R* is recall. It allows more robust threshold selection with optimal precision-recall tradeoff.

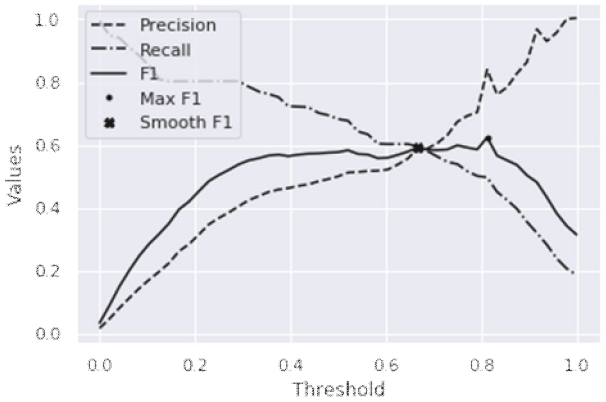

Fig. 6. Dependency of Precision, Recall and F1 scores from threshold value

Results

In this work, we introduced a method and practical experimental notes for multi-target protein localization. It shows high predictive performance and can perform even better with more data of higher resolution or different staining images. For it, we used modified version of Inception-v3 neural network architecture. In order to ensure fast and robust training, we used combination of stochastic gradient descent with restarts and progressive image size. We show the method of robust threshold selection for imbalanced dataset with noisy labels for F1-score metric. As part of this work, we took part in Kaggle Human Protein Atlas Image Classification Challenge. Presented solution took 231-th place among 2172 teams with final F1 value of 0.495 on a separate test set.

Future research

The main future aim is to improve model performance for classes that have small amount of samples overall. For this purpose, existing approach can be supplemented with metric learning [20]. This will allow matching new unknown samples to known samples using learned distance metric. This approach can sufficiently decrease overfitting for rare classes.

Another direction of the research can be focused of increasing the generalization properties of existing neural network architectures. Domain knowledge augmentations or AutoAugment can be used [5].

In addition, approach with progressive resizing combined with warm restarts needs future investigation. It can be useful for heterogeneous cell data that was captured at different magnification. In addition, this combination in can be successful in other different problems, e.g. it can be useful in tasks with strong influence of perspective.

Conclusions

We presented an approach for protein localization using deep neural network. Tests show robustness of the presented solution and high quality of predicted labels. It is a generic and robust method which can process input images of arbitrary size capable of learning images from various heterogeneous cell types. Future advances of this technique can show substantial potential in other localization tasks with future research of miniaturizing the deep neural network as well.

References

1. (2018). “Human Protein Atlas Image Classification” [Electronic source] – Available at : https://www.kaggle.com/c/human-protein-atlas-image-classification. – Active link : (11.03.2019).

2. (2018). “Protein Atlas Image Classification”. 12-th place solution [Electronic source]. – Available at : https://www.kaggle.com/c/human-protein-atlas-image-classification/discussion/77325. – Active link : (11.03.2019).

3. Andre, Esteva, Brett, Kuprel, Rob, Novoa, Justin, Ko, Susan, Swetter, Helen, Blau, & Sebastian, Thrun. (2017). “Dermatologist-level Classification of Skin Cancer with Deep Neural Networks”. *Nature Publ.*, pp. 115-118.

4. Christian, Szegedy, Vincent, Vanhoucke, Sergey, Ioffe, Jon, Shlens & Zbigniew, Wojna. (2016). “Rethinking the Inception Architecture for Computer Vision”, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR*), pp. 2818-2826.

5. (2018). Cubuk. Ekin, Barret. Zoph et al. “Auto Augment: Learning Augmentation Policies from Data – in proceedings of Computer Vision and Pattern Recognition” – site.

6. Elad. Eban, Mariano. Schain, Alan. Mackey, Ariel. Gordon, Rif A. Saurous & Gal. Elidan. (2017). “Scalable Learning of Non-Decomposable Objectives”. In arXiv preprint arXiv: 1608.04802v2, pp. 1-7 site.

7. Elisabeth. Rumetshofer, Markus, Hofmarcher, Clemens, Röhrl, Sepp, Hochreiter & Günter, Klambauer. (2018). „Human-level Protein Localization with Convolutional Neural Networks”. *ICLR 2019 Conference Blind Submission*, pp. 221-232.

8. Godinez, W. J., Hossain, I., Lazic, S. E., Davies, J. W., & Zhang, X. (2019). “A multi-scale convolutional neural network for phenotyping high-content cellular images”. *Bioinformatics*, 33(13), pp. 2010-2019. DOI:10.1093/bioinformatics/btx069.

9. Huang, Gao & Zhuang, Liu. (2017). “Densely Connected Convolutional Networks - In proceeding of Computer Vision and Pattern Recognition”, pp. 1-77. DOI: 10.1109/CVPR.2017.243.

10. Ilya, Loshchilov & Frank, Hutter. (2017). “SGDR: Stochastic Gradient Descent with Warm Restarts”. In ICLR 2017 proceedings, https://arxiv.org/abs/1608.03983.

11. Kraus, O. Z., Ba, J. L., & Frey, B. J. (2017). “Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics”, 32(12), pp. 52-59. DOI: 10.1093/bioinformatics/btw252.

12. Latt, S. A., Stetten, G., Juergens, L. A., Willard, H. F., & Scher, C.D (July 1975). “Recent developments in the detection of deoxyribonucleic acid synthesis by 33258 Hoechst fluorescence”. *Journal of Histochemistry and Cytochemistry*. 23 (7): pp. 493-505, doi:10.1177/23.7.1095650.

13. Maximilian, Ilse, Jakub, M. Tomczak & Max, Welling. (2018). “Attention-based Deep Multiple Instance Learning”. In arXiv preprint arXiv:1802.04712v4, https://arxiv.org/abs/1802.04712.

14. (2013). Niwas, S, Issac, Kårsnäs, er al. „Automated classification of immunostaining patterns in breast tissue from the human protein Atlas”. *Journal of pathology informatics.* 4. S14. DOI: 10.4103/2153-3539.109881.

15. Sechidis, K., Tsoumakas, G. & Vlahavas, I. (2011). “On the Stratification of Multi-Label Data. Machine Learning and Knowledge Discovery in Databases”. ECML PKDD 2011. Lecture Notes in Computer Science, Vol. 6913. Springer, Berlin, Heidelberg, pp. 145-158. DOI: 10.1007/978-3-642-23808-6_10.

16. (2018). Sullivan, D, Casper, F. Winsnes, et al. “Deep learning is combined with massive-scale citizen science to improve large-scale image classification”, *Nature Biotechnology Publ*., Vol. 36, pp. 820-828. DOI: 10.1038/nbt.4225.

17. (2017). Thul, P. J, Åkesson, L er al. „A subcellular map of the human proteome”. Science. 356(6340): eaal3321 PubMed: 28495876. DOI: 10.1126/science.aal3321.

18. (2019). “The multilocalizing proteome” [Electronic source]. – Available at : – https://www.proteinatlas.org/humanproteome/cell/multilocalizing. – Active link : 11.03.2019.

19. Vincent, Vanhoucke, Christian, Szegedy, Sergey, Ioffe, Jonathon, Shlens & Zbigniew, Wojna. (2015). “Rethinking the Inception Architecture for Computer Vision”. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9.

20. (2016). Wenbin, Li, Jing, Huo, Yinghuan, Shi et al. “Online Deep Metric Learning”, *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11-19.

21. Luzio, J. P. (2001). “Relationship between Endoscopes and Lysosomes”. *Biochemical Society Transactions*, Vol. 29, No. 4, pp. 476-480, doi: 10.1042/bst0290476.

Received 17.01.2019

УДК 004.93.1

[1]**Тимченко, Борис Ігорович**, Інститут комп’ютерних систем, каф. інформаційних систем, E-mail: tim4bor@gmail.com, ORCID: 0000-0002-2678-7556
[1]**Граматік, Ангеліна Антоліївна**, Інститут комп’ютерних систем, каф. інформаційних систем, E-mail: angelinagramatik08@gmail.com, ORCID: 0000-0003-4569-1637
[1]**Тульчий, Георгій Петрович**, Інститут комп’ютерних систем, каф. інформаційних систем, E-mail: gearty@rambler.ru, ORCID: 0000-0001-8396-1672
[1]**Антощук, Світлана Григорівна,** д-р техн. наук, професор, директор Інституту комп’ютерних систем, E-mail: asg@opu.ua, ORCID: 0000-0002-9346-145X
[1]Одеський національний політехнічний університет, пр. Шевченка, 1, Одеса, Україна, 65044

КЛАСИФІКАЦІЯ ПАТЕРНІВ БІЛКІВ НА МІКРОСКОПІЧНИХ ЗОБРАЖЕННЯХ З ВИКОРИСТАННЯМ ГЛИБОКИХ НЕЙРОННИХ МЕРЕЖ

Анотація. *В наш час точна діагностика захворювань, їх лікування та прогноз є гострою проблемою сучасної медицини. Вивчаючи інформацію про людські протеїни, можливо ідентифікувати диференційно експресовані білки. Ці протеїни є потенційно цікавими біомаркерами, які слід використовувати для точного діагнозу, прогнозу або вибору індивідуального лікування, особливо в разі онкологічних захворювань. Результати досліджень показують, що відносно мало білків мало білків в людському тілі є тканеспецифічнимі. Майже половина всіх білків класифікується як допоміжні білки, що експресуються в усіх клітинах. Тільки 2300 білків в організмі людини були ідентифіковані як тканеспецифічні, що означає, що вони*

мають підвищені рівні експресії в певних тканинах. Завдяки досягненням в області високопродуктивної мікроскопії зображення генеруються занадто швидко для ручної оцінки. Отже, потреба в автоматизації аналізу біомедичних зображень як ніколи велика, щоб прискорити розуміння людських клітин і захворювань. Історично класифікація білків обмежувалася індивідуальними паттернами в одному або декількох типах клітин, але для повного розуміння складності людської клітини моделі повинні класифікувати змішані па терни відповідно до кількості різних типів людських клітин. У статті сформульована проблема класифікації зображень в медичних дослідженнях. У цій області активно використовуються методи класифікації з використанням глибоких загортальних нейронних мереж. Представлена стаття дає короткий огляд різних підходів і методів подібного дослідження. Як набору даних було взято «Human Protein Atlas», що представляє тканинну карту протеома людини, складену в 2014 році після 11 років досліджень. Всі дані профілювання експресії протеїнів загальнодоступні в інтерактивній базі даних, що дозволяє досліджувати протеом людини на тканинній основі. Було проведено аналіз робіт і методів, які були використані в ході дослідження. Для вирішення цієї завдані запропонована модель глибокої нейронної мережі з урахуванням характеристик домену і досліджуваної вибірки. Модель нейронної мережі заснована на архітектурі Inception-v3. Процедура оптимізації містить комбінацію декількох методів для швидкої збіжності: стохастичний градієнтний спуск з перезапуском (зміна швидкості навчання для вивчення різних локальних мінімумів), прогресивне зміна розміру зображення (навчання починається з невеликої роздільної здатності і послідовно збільшує її кожен цикл SGDR). Ми пропонуємо новий метод вибору порогу для заходи F1. Розроблена модель може бути використана для створення приладу, інтегрованого в медичну систему інтелектуальної мікроскопії, для визначення місця розташування білка по високоефективному зображенню.

Ключові слова: *Human Protein Atlas;, глибоке навчання; нейронні мережі; класифікація; розпізнавання образів*

УДК 004.93.1

[1]**Тимченко, Борис Игоревич**, Институт компьютерных систем, кафедра информационных систем, E-mail: tim4bor@gmail.com, ORCID: 0000-0002-2678-7556, г. Одесса, Украина
[1]**Граматик, Ангелина Анатольевна**, Институт компьютерных систем, кафедра информационных систем, E-mail: angelinagramatik08@gmail.com, ORCID: 0000-0003-4569-1637, г. Одесса, Украина
[1]**Тульчий, Георгий Петрович**, Институт компьютерных систем, кафедра информационных систем, E-mail: gearty@rambler.ru, ORCID: 0000-0001-8396-1672, г. Одесса, Украина
[1]**Антощук, Светлана Григорьевна,** д-р техн. наук, профессор, директор Института компьютерных систем, E-mail: asg@opu.ua ORCID: /0000-0002-9346-145X, г. Одесса, Украина
[1]Одесский национальный политехнический университет, пр. Шевченко, 1, Одесса, Украина, 65044

КЛАССИФИКАЦИЯ ПАТТЕРНОВ БЕЛКОВ НА МИКРОСКОПИЧЕСКИХ ИЗОБРАЖЕНИЯХ С ИСПОЛЬЗОВАНИЕМ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ

Аннотация. *В настоящее время точная диагностика заболеваний, их лечение и прогноз являются острой проблемой современной медицины. Изучая информацию о человеческих белках, возможно идентифицировать дифференциально экспрессируемые белки. Эти белки являются потенциально интересными биомаркерами, которые следует использовать для точного диагноза, прогноза или выбора индивидуального лечения, особенно в случае онкологических заболеваний. Результаты исследований показывают, что относительно мало белков мало белков в человеческом теле являются тканеспецифичными. Почти половина всех белков классифицируется как вспомогательные белки, экспрессируемые во всех клетках. Только 2300 белков в организме человека были идентифицированы как тканеспецифичные, что означает, что они имеют повышенные уровни экспрессии в определенных тканях. Благодаря достижениям в области высокопроизводительной микроскопии изображения генерируются слишком быстро для ручной оценки. Следовательно, потребность в автоматизации анализа биомедицинских изображений как никогда велика, чтобы ускорить понимание человеческих клеток и заболеваний. Исторически классификация белков ограничивалась индивидуальными паттернами в одном или нескольких типах клеток, но для полного понимания сложности человеческой клетки модели должны классифицировать смешанные паттерны в соответствии с количеством различных типов человеческих клеток. В статье сформулирована проблема классификации изображений в медицинских исследованиях. В этой области активно используются методы классификации с использованием глубоких сверточных нейронных сетей. Представленная статья дает краткий обзор различных подходов и методов подобного исследования. В качестве набора данных был взят «Human Protein Atlas», представляющий тканевую карту протеома человека, составленную в 2014 году после 11 лет исследований. Все данные профилирования экспрессии белка общедоступны в интерактивной базе данных, что позволяет исследовать протеом человека на тканевой основе. Был произведен анализ работ и методов, которые были использованы в ходе исследования. Для решения этой задачи предложена модель глубокой нейронной сети с учетом характеристик домена и исследуемой выборки. Модель нейронной сети основана на архитектуре Inception-v3. Процедура оптимизации содержит комбинацию нескольких методов для быстрой сходимости: стохастический градиентный спуск с перезапусками (изменение скорости обучения для изучения различных локальных минимумов), прогрессивное изменение размера изображения (обучение начинается с небольшого разрешения и последовательно увеличивает каждый цикл SGDR). Мы предлагаем новый метод выбора порога для меры F1. Разработанная модель может быть использована для создания прибора, интегрированного в медицинскую систему интеллектуальной микроскопии, для определения местоположения белка по высокоэффективному изображению.*

Ключевые слова: *Human Protein Atlas; глубокое обучение; нейронные сети; классификация; распознавание образов*