Deep learning approach for compound similarity analysis

Dariusz Brzeziński, DSc PUT, Maciej Filanowicz, PUT, Rafał Stachowiak, PUT, Kajetan Kubik, PUT, Bartosz Maślanka, PUT, Franciszek Olejnik, PUT, Mikołaj Kruś, PUT, Piotr Masłowski, PUT,

***Abstract*—In recent years, neural network-based approaches for image processing spread throughout both the industry and academia. However, high-content screening research still typically relies on analysis and comparison of various statistical markers derived from measured samples. This study aims to determine the viability of employing convolutional neural networks for high-content screening applications. Specifically, it investigates whether these modern deep learning approaches can successfully examine the effects induced by various chemical compounds on the cells. It can be important for the production of drugs especially speeding up the analysis of various statistics in the image data which in the classical approach took a lot of time. Notably, we use residual neural networks as the base architecture. Their parameters are mostly pretrained, with certain layers adjusted to our problem definition. This way the model can perform image classification and ultimately assess the similarity of the scanned sample to a previously known substance. As opposed to a supervised learning approach, we wanted to try an unsupervised method: clustering, and see if we can find anything interesting. We have determined that ResNets can indeed be successfully used for the devised task while achieving high classification accuracy - usually above 95%. They can be trained within minutes and don't require substantial power for evaluation - a regular, consumer-grade CPU can easily perform this task. Our results suggest that convolutional neural networks are a promising route for future high-content screening. Currently, they can provide researchers with interesting data points and after further polishing our tools, they could conceivably replace some of the previous, statistical methods.**

***Index Terms*—IEEE, High Content Screening, ResNet, Cell-Painting, Clustering**

I. Introduction

HIGH-CONTENT SCREENING (HCS), also known as High-Content Analysis (HCA), has been a tremendously strong tool utilized in biological activity research at the model organism, cellular, pathway, or molecular level. The standard HCS [1] process outlines the planning and execution of investigations utilizing cell painting, a morphological profiling assay that multiplexes six fluorescent dyes and images them in five channels to identify eight generally relevant cellular components or organelles. Cells are grown on multiwell plates, disturbed with the treatments being investigated, stained, fixed, and photographed using a high-throughput microscope. Subsequently, an automated image analysis software recognizes individual cells and assesses 1,500 morphological characteristics (different measurements of size, shape, texture, intensity, and so on) to provide a rich profile ideal for detecting subtle phenotypes. Profiles of cell populations treated with various experimental perturbations can be compared to achieve a variety of goals, including determining the phenotypic impact of chemical or genetic perturbations, grouping compounds and/or genes into functional pathways, and identifying disease signatures. [2]

However, this traditional technique is both extremely laborious and time consuming since, in addition to biological experiments, image processing and quantification need supervised effort and significant computational resources, after which an extra statistical analysis must be conducted. Both image and data analysis are recognized as considerable bottlenecks in HCS. Furthermore, employing existing image analysis software and statistical tools for analysis requires a certain level of experience and training.

To address these shortcomings, we propose a deep learning approach based on a convolutional neural network (CNN) for evaluating biological activity of tested perturbations encoded in images directly in acquired images. Furthermore, our technique eliminates the phase of obtaining numerical characteristics from images and instead identifies phenotypic similarity or morphological grouping directly from the images.

Deep learning in image classification is a process that involves training a convolutional neural network with sample images, after which the trained model can be used to recognize different images. It has been demonstrated to be highly effective for descriptor extraction and classification in visual recognition tasks. Deep learning's capability and efficiency have recently surpassed human recognition, and it has been applied to a wide range of related fields in cutting-edge science and engineering. Object classification is a very common task in image analysis. The classification can be used to select the desired objects after initial segmentation, as well as to categorize the final images into defined state categories. [3]

In this report we developed a software for compound similarity analysis based on its phenotypic impact on cellular level. We established a base set of eight well-known compounds, namely Berberine Chloride, Brefeldin A, Fluphenazine, Latrunculin B, Nocodazole, Rapamycin, Rotenone, and Tetrandrine, as well as a control group with no compound used (DMSO). Our solution enables rapid cross-comparison (clustering and similarity estimation) of new, previously unknown compounds with our base set.

II. Related work

Image-based profiling [4] is a maturation technique in which a multidimensional profile, or a collection of extracted image-based attributes, is created using the rich information inherent in biological pictures. These profiles may be analysed for

patterns that show unexpected biological activity, which is valuable for various stages of the drug development process. Identifying illness-associated screenable phenotypes, understanding disease processes, and forecasting a drug's efficacy, toxicity, or mode of action are examples of such uses. Within academia and the pharmaceutical sector, some of these applications have recently been verified and put into production mode. Some of them have provided poor results in reality, but improved machine-learning algorithms that better use image-based information has reignited interest in them.

Image-based screening [5] may be used to assess a wide range of phenotypes in cells and entire organisms. When combined with perturbations like RNA interference, small chemicals, and mutations, such screens are a valuable tool for learning about biological processes systematically. Various processes, such as protein localization changes, cancer cell vulnerabilities, and complex organismal phenotypes, have been studied using screens. Large-scale perturbation screens have recently been expedited thanks to recent developments in imaging and image-analysis technologies. The current state of the art for image-based screening tests is described in the article, along with experimental and image-analysis methodologies, as well as obstacles and future possibilities, such as utilising CRISPR/Cas9-mediated genome modification.

High-dimensional data collected from microscope pictures for a large number of single cells opens up a lot of possibilities for data processing. It's an easy activity for a skilled human, but it's extremely tough to automate on robots. In this [6] article, a 50-layer neural network Resnet is used on a large quantity of data from microscope pictures of cells, reaching classification accuracy of 95% for cell localization and 99% for protein localization on images. This article establishes and shows that low-level network properties correlate to basic visual qualities, whereas deeper layers distinguish between classes.

III. Methods/Approach

A. Data set preprocessing

A high-content screening microscope equipped us with high-quality 16-bit images in the size of 1080 x 1080 px. From eleven distinct substance concentrations, we focused on the two highest ones – 50% and 25%. To construct one data set sample with all the organelles we utilised four images, later referred to as the 'channels', out of which each was captured using a different light wavelength, hence every file depicts various cell organelles. The division of organelles per channel is expressed by the below mapping (channel number, organelles)

1) cell nuclei
2) endoplasmic reticulum and nucleoli
3) actin F (cytoskeleton protein), Golgi apparatus and cytoplasmic membrane
4) mitochondria

All the channels were then properly resized using the Lanczos interpolation algorithm [7] to the dimensions of 224 x 224, merged together into one picture and saved to the drive. (Fig. 1) For samples to be used in the deep learning models we standardized them using the mean and standard deviation calculated from the entire data set.

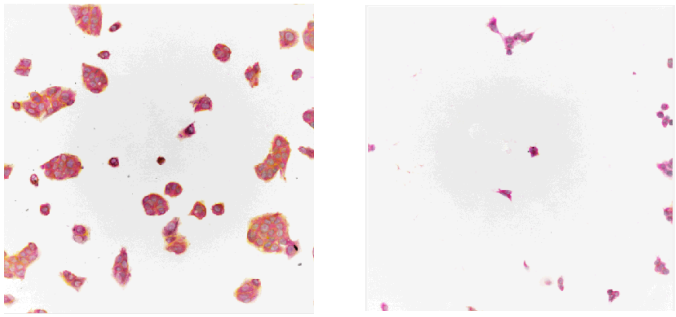

Fig. 1. Images can capture various cell sizes, locations and groups.

B. Data set split into test / train samples

The testing data set consists of one whole plate. The remaining plates were separated into training and validation sets using the StratifiedGroupKFold function from Scikit-learn. The reason we preferred this particular split algorithm is twofold. Firstly, we wanted it to be stratified, so that the ratios of different classes remain unaffected between our data sets. Secondly, since several different images can depict a single well and can thus be dependent on each other, it was crucial that they are grouped during data set split. This was not essential for the test set split because any wells are already contained within a plate.

C. Classification model

To conduct a compound classification task we propose a pre-trained version of the ResNet model. Since it was designed for three-channel images on the input, one of our modifications was to replace its initial layer with an extended copy. We have sourced the additional weights from one of the existing layers. By convention the pre-trained model on the ImageNet data set has one thousand classes at the output [8], we altered that number to the number of substances we scrutinised.

D. Clustering model

Unlike in the standard approach, instead of using HCS tools to extract features from images, we take our classification model and we remove the last layer. Thanks to this operation, we get a vector of 4096 features, which we can apply next to map each image (merged all channels to one image) of cells to the Cartesian coordinate system to allow an objective assessment of the distribution of photos from the entire study. To do this, we use Uniform Manifold Approximation and Projection [9]. With UMAP, we reduced the vector of features to a 2-dimensional point that we can visualise in the coordinate system. Taking into account the UMAP parameters, we can modify the results in real time.

IV. Results/Experiments

A. Comparison of ResNets

We tested three sizes of ResNet models. The models in which we trained only the last layer - performing the classification task - delivered surprisingly good results. We proved that training all the layers can give even better results. Our primary goal was to develop a system capable of operating on both the CPU and the GPU with a short inference time. Since the difference in performance was not significant we picked the least ResNet (due to the lowest training time).

We are calculating the accuracy score in two ways. Our data set consists of batches of images describing the same well, each including nine distinct scans. Based on 9 judgements of the model, we derive a single result. 'Majority' score represents the simple voting procedure and 'Average' score - the label with the highest mean probability.

Per-well accuracy (Dense layer)			
ResNet	Data set	Majority	Average
ResNet18	Training	98.214%	98.810%
	Validation	95.833%	95.833%
	Testing	97.917%	97.917%
ResNet34	Training	97.619%	98.214%
	Validation	95.833%	95.833%
	Testing	97.917%	98.958%
ResNet50	Training	99.405%	99.405%
	Validation	95.833%	95.833%
	Testing	48.958%	97.917%

TABLE I
EACH MODEL WAS TRAINED FOR 20 EPOCHS WITH A LEARNING RATE OF 0.01.

Per-well accuracy (All layers)			
ResNet	Data set	Majority	Average
ResNet18	Training	98.810%	98.810%
	Validation	95.833%	95.833%
	Testing	73.958%	71.875%
ResNet34	Training	99.405%	99.405%
	Validation	95.833%	95.833%
	Testing	97.917%	100.000%
ResNet50	Training	99.405%	99.405%
	Validation	95.833%	95.833%
	Testing	100.000%	100.000%

TABLE II
EACH MODEL WAS TRAINED FOR 20 EPOCHS WITH A LEARNING RATE OF 0.001.

B. Clustering models

During our experiments, we tested several algorithms for mapping the feature vector PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbour Embedding), UMAP (Uniform Manifold Approximation and Projection) and the combinations of these algorithms. The results were very similar, and after performing comparison analysis and consulting this [10] article, we came to the conclusion that UMAP gives the most satisfying results.

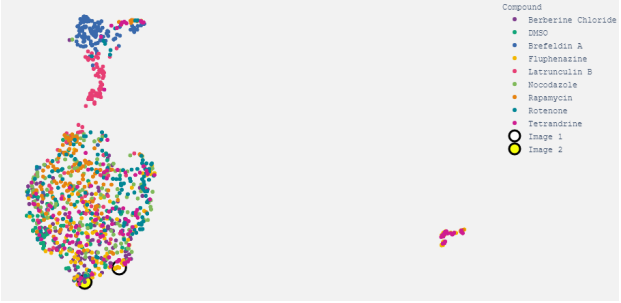

Fig. 2. How UMAP map your data into 2d space. Different colors corresponding different compounds.

V. Discussion/Conclusion

Our study aimed to evaluate the viability of replacing classic statistical methods in High Content Screening data analysis with end-to-end convolutional neural networks. We have tested transfer learning using residual neural networks and verified its fitness for this purpose. Furthermore, we found out that significantly downscaling the scans did not prevent the model from producing highly correct results. On the other hand, neither gamma correction of pixel values nor clipping outlier points to a given percentile value have yielded any noticeable improvement but instead turned out detrimental to the model.

In the future, we could devise a semantic segmentation algorithm or check why the model chose a given class (by performing a thorough analysis using e. g. SHAP).

In terms of clustering, we could add something called latent space adaptation, for better separability in the final graph [11], but this (probably) would force us to train some GAN(Generative adversarial network). The other big idea is to use the first channel of the images (which contains the cell nucleus), and on these images perform instant segmentation. This allows us to know the amount of cells in the image and their average size.

Acknowledgment

The authors would like to thank PhD Magdalena Otrocka Head of Laboratory of Molecular Assays at Institute of Bioorganic Chemistry, Polish Academy of Sciences and PhD Jacek Kolanowski Head of Molecular Probes and Prodrug Laboratory at Institute of Bioorganic Chemistry, Polish Academy of Sciences for their knowledge and help in the process of understanding the problem.

References

[1] J. Bellis, Q. Guichard, B. Chilian, K. Kottig, P. Rochard, and G. Mondesert, "Cell painting for compounds clustering and mechanism of actions characterization," evotec. [Online]. Available: https://www.evotec.com/f/7db58faf46fc51e7fc0c5c815ab1be37.pdf

[2] M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, M. Kost-Alimova, S. M. Gustafsdottir, C. C. Gibson, and A. E. Carpenter, "Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes," *Nature Protocols*, vol. 11, no. 9, pp. 1757–1774, Sep 2016. [Online]. Available: https://doi.org/10.1038/nprot.2016.105

[3] I.-H. Lee, S. Passaro, S. Ozturk, J. Ureña, and W. Wang, "Intelligent fluorescence image analysis of giant unilamellar vesicles using convolutional neural network," *BMC Bioinformatics*, vol. 23, no. 1, p. 48, Jan 2022. [Online]. Available: https://doi.org/10.1186/s12859-022-04577-2

[4] S. N. Chandrasekaran, H. Ceulemans, J. D. Boyd, and A. E. Carpenter, "Image-based profiling for drug discovery: due for a machine-learning upgrade?" *Nature reviews. Drug discovery*, vol. 20, no. 2, pp. 145–159, Feb 2021, 33353986[pmid]. [Online]. Available: https://doi.org/10.1038/s41573-020-00117-w
[5] M. Boutros, F. Heigwer, and C. Laufer, "Microscopy-Based High-Content screening," *Cell*, vol. 163, no. 6, pp. 1314–1325, Dec. 2015.
[6] H. Elnashar and I. Abd, "Deep learning: Protein cells classifications using resnet-50 model," 12 2021.
[7] C. Lanczos, "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators," *Journal of research of the National Bureau of Standards*, vol. 45, pp. 255–282, 1950.
[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385
[9] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.
[10] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, "Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," 2021.
[11] D. Zhang, Y. Sun, B. Eriksson, and L. Balzano, "Deep unsupervised clustering using mixture of autoencoders," 12 2017.