CytoImageNet: A large-scale pretraining dataset for bioimage transfer learning

Stanley Z. Hua
Computer Science
University of Toronto
stanley.hua@mail.utoronto.ca

Alex X. Lu
Computer Science
University of Toronto
alexlu@cs.toronto.edu

Alan M. Moses
Cell and Systems Biology
Computer Science
Center of genome Evolution and Function
University of Toronto
alan.moses@utoronto.ca

Abstract

Motivation: In recent years, image-based biological assays have steadily become high-throughput, sparking a need for fast automated methods to extract biologically-meaningful information from hundreds of thousands of images. Taking inspiration from the success of ImageNet, we curate CytoImagenet, a large-scale dataset of openly-sourced and weakly-labeled microscopy images (890K images, 894 classes). We show that CytoImageNet features perform comparably to ImageNet features on downstream classification tasks.

1 Introduction

Automated microscopy has given biologists a tool for studying various cell types under varying experimental conditions with minimal assistance. In recent years, automated image analysis has proven useful for drug discovery and functional genomics research; identifying the function of novel genes and cellular response to various drugs. (Caicedo, Singh and Carpenter 2016). With the capacity to perform large-scale screens rapidly, tens of thousands of images are being collected daily, demanding computational methods to extract biology from these images (Wollman and Stuurman 2007).

Methods for automated image analysis still remains an active area of research. One strategy for automating the analysis of microscopy images is to use low-level features from classic computer vision strategies; extracting thousands of general features related to area, shape, texture and intensity of single-cells (Danuser 2011; Carpenter et al., 2006). However, these features usually require extensive engineering, parameter tuning, and effective single-cell segmentation, all of which must be customized to the images of interest. Furthermore, these features may not capture relevant biological signal and instead capture experimental noise (Caicedo et al., 2018). To reduce labor and improve the sensitivity of extracted features, some studies rely on deep learning to automatically learn features. These strategies range from supervised training with labels, autoencoders to more recently, self-supervised learning (Caicedo et al., 2018; Kraus et al., 2017; Lu et al., 2019). However, these models require users to train on their own data, which is computationally expensive, time-consuming, and requires specialized hardware like GPUs

Preprint. Under review.

Table 1: Number of labels from each metadata category

Category	Compound	Phenotype	Cell Type	Gene	Cell Visible	siRNA	Organism	**TOTAL**
Number of labels	637	93	44	43	38	36	3	894

Table 2: Number of labels contributed by each image database

Database	**Number of Labels Contributed**
Recursion	651
Image Data Resource	450
Broad Bioimage Benchmark Collection	202
Kaggle	27
Cell Image Library	1

As an alternative to training bespoke models, some studies instead rely on transfer learning. Most commonly, studies re-use features from neural networks trained to classify images from ImageNet, a large-scale natural image classification dataset (Huh, Agrawal and Efros 2016). By training models to perform well on large datasets, they can learn useful features that may be transferred to other datasets. ImageNet is a large-scale dataset of more than 14 million diverse annotated images. A defining factor in the success of ImageNet is their diversity of image labels and image sources. Models that perform the best on ImageNet become the standard for transfer learning. These models learn a diverse set of features that are able to capture complex and generalizable patterns inherent in their dataset. Surprisingly, ImageNet features have even shown good transferability on microscopy classification tasks (Pawlowski et al., 2016; Kensert, Harrison and Spjuth 2019). However, a recent study suggests that "ImageNet features are less general than previously suggested," and that pretraining on datasets with a closer domain match to the downstream task benefits transfer learning (Kornblith, Shlens and Le 2019). In fact, Caicedo et al., 2018 showed that pretraining convolutional networks (CNN) on weakly labeled microscopy images yields features that outperform ImageNet features on downstream classification tasks. We hypothesize that features from pretraining on a diverse large-scale microscopy dataset will yield more domain-relevant features than ImageNet.

We present CytoImageNet, a large-scale image dataset of weakly labeled microscopy images (890K images, 894 classes). Incorporating image data and labels from 40 distinct microscopy datasets, CytoImageNet mimicks the diverse and complex nature of ImageNet. We believe that pretraining on CytoImageNet may yield biologically-meaningful image representations useful in any downstream biological task.

We provide a framework for curating similar large-scale pretraining datasets using openly available microscopy datasets, and we show that pretraining on CytoImageNet results in features that perform comparably to ImageNet on downstream tasks.

2 Methods

CytoImageNet is a dataset of 890,737 microscopy images sourced from openly available images from 40 datasets. The selected datasets originated from 5 databases: (1) Recursion, (2) Image Data Resource (IDR), (3) Broad Bioimage Benchmark Collection (BBBC), (4) Kaggle and (5) Cell Image Library (CIL). Electron microscopy and histopathology images were excluded. Code for the methods below is available at `https://github.com/stan-hua/CytoImageNet`.

2.1 Weak label assignment & stratified downsampling

As a training target for CytoImageNet, we sought to associate each microscopy image with a class label. As expert-assigned annotations are not readily available or standardized across publicly available image datasets, we instead chose to assign images to classes based upon metadata associated with each of the image datasets we used. We considered 7 kinds of metadata: (organism, cell_type, cell_visible, phenotype, compound, gene, and sirna). Each column was searched for unique values and their counts to find potential weak labels. To create 894 labels, unique values with at least 287

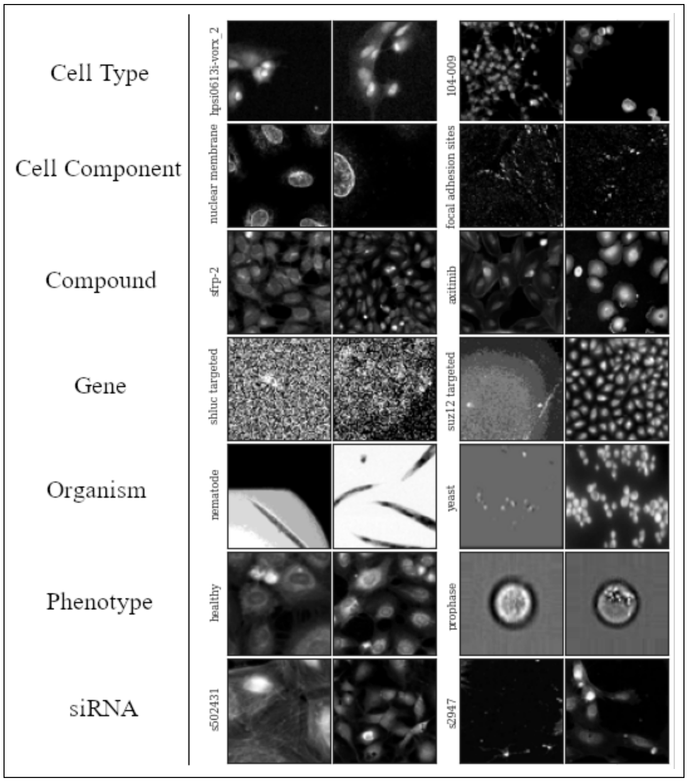

Figure 1: Examples of CytoImageNet classes from each type of category (left).

images served as potential labels. Beginning with the least counts, each potential label was iterated through, searching for images that can be assigned to the label. Images that were already assigned labels are excluded, and their unique image index is stored in a hash table pointing to their label. To improve the diversity of images selected for a label, stratified sampling was done by sampling from images that were grouped by values in the remaining potential label columns.

2.2 Image standardization & upsampling

To preprocess images, we converted all images to PNG. RGB images are converted to grayscale by averaging over channels. We normalized each channel between [0, 255] using the 0.1 and 99.9th percentile pixel intensity as the lower and upper ranges respectively, as in Kraus et al. 2017. Finally, for fluorescent images, we merged the channels.

We took 1-4 crops from each images of varying spatial scales where possible. Unless the image was smaller than 70x70 pixels, we cropped images into four quadrants. To introduce variations in spatial scale, we further cropped these quadrants by sampling crops of either full, 0.5, 0.25, or 0.125 of the quadrant (excluding any crop scales that would produce images smaller than 70x70 pixels). Finally, we filtered any crops that were all black, all white, or where the 75th percentile pixel intensity is equal to 0. Crop pixel intensities were re-normalized as we previously did for full images.

2.3 Training

CytoImageNet is split into a training and validation set with 10% used for validation. This yields roughly 900 training samples for each label. Images are fed in batches of 64 with random 0 to 360 degrees rotations. We train convolutional networks (EfficientNetB0) to classify one of the 894 labels, by minimizing the categorical cross-entropy loss of predictions to ground truth labels. Randomly-initialized models were trained for 24 epochs (2 weeks) on an NVIDIA Tesla K40C. The loss was optimized via the Adam optimizer with learning rate of 0.001.

Table 3: Results on downstream tasks

Downstream Task	Random	ImageNet	CytoImageNet
BBC021	27.18 ± 8.59%	**83.5** ± 7.17%	**83.5** ± 7.17%
CyCLOPS	53.06 ± 0.59%	**68.47** ± 0.55%	65.19 ± 0.57%
COOS7 Test Set 1	65.68 ± 0.91%	**88.87** ± 0.61%	88.58 ± 0.61%
COOS7 Test Set 2	67.81 ± 0.7%	88.93 ± 0.47%	**89.37** ± 0.46%
COOS7 Test Set 3	48.77 ± 0.54%	**75.91** ± 0.46%	65.97 ± 0.51%
COOS7 Test Set 4	51.88 ± 0.56%	**82.19** ± 0.43%	78.52 ± 0.46%

2.4 Evaluation

We compare features extracted from a CytoImageNet-pretrained model, an ImageNet-pretrained model, and a randomly-initialized model on three distinct downstream tasks: 1) BBBC021 mechanism-of-action classification, 2) Cells-Out-Of-Sample (COOS-7) protein localization classification, and 3) a CYCLoPs dataset for protein localization classification.

BBBC021 is a dataset of fully imaged human cells (Caie et al., 2010). Cells are treated with one of 113 small molecules at 8 concentrations, and fluorescent images are captured staining for nucleus, actin and microtubules. The phenotypic profiling problem is presented, where the goal is to extract features containing meaningful information about the cellular phenotype exhibited. Each of 103 unique compound-concentration treatment is labeled with a mechanism-of-action (MOA). The MOA is predicted for each unique treatment (averaging features over all treatment examples) by matching the MOA of the closest point excluding points of the same compound. We use a 1-nearest neighbors using cosine distance as the metric and report the not-same-compound (NSC) accuracy.

On the other hand, the WT2 dataset from the CYCLoPs database consists of 27,058 single-cell images of yeast cells (Koh et al., 2015). The task is to classify the subcellular localization of a fluoresced protein, given two channels staining for the protein of interest and the cytosol. An 11-nearest neighbors is used to perform leave-one-out classification, and the accuracy is reported.

Lastly, COOS-7 contains 132,209 single-cell images of mouse cells (Lu et al., 2019). Images are spread over 1 training set and 4 testing sets, where each single-cell image contains a protein and nucleus fluorescent channels. COOS-7 was curated, such that each succeeding testing set has a greater degree of covariate shift from the training set. Similar to the CYCLoPs dataset, the COOS-7 dataset presents a protein subcellular localization classification task. We use a 11-nearest neighbors fit on the training set to predict on the testing sets and report the accuracy.

We chose these datasets (1) to assess the generalizability of CytoImageNet-pretrained models on distinct biological image datasets and problems, and (2) to compare against ImageNet features that have already been shown to perform relatively well on these datasets (Pawlowski et al., 2016; Lu et al., 2019; Lu et al., 2019).

Features are extracted from the penultimate layer of the models. Predictions are generated using k-Nearest Neighbors (kNN), and accuracy is reported with a 95% confidence interval. BBBC021 kNN is implemented with scikit-learn, while kNN is implemented using the faiss library for COOS-7 and CyCLOPS datasets.

We compare the performance of features extracted following two optional preprocessing methods: normalization and channel merging. Normalization refers to channel normalization with the lower bound and upper bound of the 0.1 and 99.9th percentile pixel intensity respectively (Kraus et al. 2017). Channel merging refers to the merging of fluorescent channels into a single grayscale image before extracting features. If not, features are extracted for each fluorescent channel then concatenated.

3 Discussion

CytoImageNet features are shown to perform comparably to ImageNet features on all 3 downstream classification tasks. In some cases, ImageNet performs slightly better than our CytoImageNet features. Overall, normalizing channels and extracting features for each channel led to the best performance for all features, implying that automated methods for feature extraction still require some form of image preprocessing.

Notably, our model only achieved 13.42% accuracy on the training set and 11.32% on the validation set. Yet, it produced features competitive to ImageNet. It is surprising that features pretrained on CytoImageNet don't beat ImageNet-pretrained features. This may be that we haven't had the opportunity to optimize the model enough. Kornblith, Shlens and Le 2019 reported a strong correlation between ImageNet validation accuracy and transfer accuracy. This may be studied in future work.

In summary, we present CytoImageNet; a pretraining dataset of 890,737 microscopy images and 894 classes. We also contribute a simple framework for combining openly available microscopy datasets, in the hopes that this sparks interest in the vast amount of rich microscopy data available online and that this may inspire others to experiment on their own.

References

[1] Caicedo JC, McQuin C, Goodman A, Singh S, Carpenter AE. Weakly supervised learning of single-cell feature embeddings. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018. doi:10.1109/cvpr.2018.00970

[2] Caicedo JC, Singh S, Carpenter AE. Applications in image-based profiling of perturbations. Current Opinion in Biotechnology. 2016;39:134–142. doi:10.1016/j.copbio.2016.04.003

[3] Caie PD, Walls RE, Ingleston-Orme A, Daya S, Houslay T, Eagle R, Roberts ME, Carragher NO. High-content phenotypic profiling of drug response signatures across distinct cancer cells. Molecular Cancer Therapeutics. 2010;9(6):1913–1926. doi:10.1158/1535-7163.mct-09-1148

[4] Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang I, Friman O, Guertin DA, Chang J, Lindquist RA, Moffat J, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biology. 2006;7(10). doi:10.1186/gb-2006-7-10-r100

[5] Danuser G. Computer Vision in cell biology. Cell. 2011;147(5):973–978. doi:10.1016/j.cell.2011.11.001

[6] Huh M, Agrawal P, Efros AA. What makes ImageNet good for transfer learning? arXiv. 2016;1608.08614.

[7] Kensert A, Harrison PJ, Spjuth O. Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. SLAS DISCOVERY: Advancing the Science of Drug Discovery. 2019;24(4):466–475. doi:10.1177/2472555218818756

[8] Koh JL, Chong YT, Friesen H, Moses A, Boone C, Andrews BJ, Moffat J. Cyclops: A comprehensive database constructed from automated analysis of protein abundance and subcellular localization patterns in saccharomyces cerevisiae. G3 Genes|Genomes|Genetics. 2015;5(6):1223–1232. doi:10.1534/g3.115.017830

[9] Kornblith S, Shlens J, Le QV. Do Better Imagenet Models Transfer Better? 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019. doi:10.1109/cvpr.2019.00277

[10] Kraus OZ, Grys BT, Ba J, Chong Y, Frey BJ, Boone C, Andrews BJ. Automated Analysis of high-content microscopy data with Deep Learning. Molecular Systems Biology. 2017;13(4):924. doi:10.15252/msb.20177551

[11] Ljosa V, Sokolnicki KL, Carpenter AE. Annotated high-throughput microscopy image sets for validation. Nature Methods. 2012;9(7):637–637. doi:10.1038/nmeth.2083

[12] Lu AX, Kraus OZ, Cooper S, Moses AM. Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. PLOS Computational Biology. 2019;15(9). doi:10.1371/journal.pcbi.1007348

[13] Lu AX, Lu AX, Schormann W, Andrews DW, Moses AM. The Cells Out of Sample (COOS) dataset and benchmarks for measuring out-of-sample generalization of image classifiers. ArXiv. 2019;abs/1906.07282.

[14] Pawlowski N, Caicedo JC, Singh S, Carpenter AE, Storkey A. Automating morphological profiling with generic deep convolutional networks. bioRxiv. 2016. doi:10.1101/085118

[15] Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, Leo S, Antal B, Ferguson RK, Sarkans U, et al. Image Data Resource: A Bioimage Data Integration and publication platform. Nature Methods. 2017;14(8):775–781. doi:10.1038/nmeth.4326

[16] Wollman R, Stuurman N. High throughput microscopy: From raw images to discoveries. Journal of Cell Science. 2007;120(21):3715–3722. doi:10.1242/jcs.013623