Kaizen
The Crowd Pathologist

Darius Lam
Phillips Academy Andover
180 Main St.
Andover, MA 01810
dlam@andover.edu

Genevieve Patterson
Microsoft Research New England
1 Memorial Drive
Cambridge, MA 02142
gen@microsoft.com

Abstract

This paper introduces Kaizen, a system to bootstrap classifiers for pathology detection in whole-slide images. Kaizen contains automatic patch creation as well as deep convolutional neural network feature extraction and uses an active-learning pipeline seeded by a small number of annotated examples. We examine the efficacy of the crowd to identify visually similar but complex pathologies with little prior knowledge. We run several experiments on the Camelyon16 dataset released by the International Symposium on Biomedical Imaging (ISBI) grand challenge in order to evaluate the utility of classifiers seeded by a varying number of starting examples.

1 Introduction

A pathologist's diagnosis has significant implications for patient treatment and overall health. Human pathologists have the time consuming task of sorting through thousands of stained whole-slide images. In this paper, we examine methods for the automated pathological diagnoses. Training classifiers to detect pathologies from whole-slide images is a daunting task, not least because of the immense size of the images. Accurate classifiers also require significant amounts of training data and, most importantly, labels. Collecting this medical training data is often difficult on two fronts. Firstly, using and aggregating data can pose privacy concerns and require the machine learning specialist to navigate through a bureaucratic maze. Secondly, accurate global labels (and segmentations) are incredibly time-consuming and require the effort of pathologists, costing significant sums of money. Labeling also requires a high level of expertise because of the sometimes minute differences that differentiate healthy from unhealthy tissue. In this paper, we seek to cost-efficiently label pathological whole-slide image patches. We will use a crowd-in-the-loop active learning pipeline to bootstrap labeling and classifier creation.

The dataset used in this paper is Camelyon16, provided by the ISBI. Camelyon16 is a set of whole-slide images of sentinel lymph node biopsies. 41% of the whole-slide images contain metastatic breast cancer complete with annotations. The manual review of sentinel lymph nodes took around 1 hour per whole-slide image for individual pathologists. The Camelyon16 dataset is unprecedented in scale for metastatic breast cancer in sentinel lymph nodes.

This paper introduces Kaizen, an online active learning system that queries the crowd in order to create classifiers using only a small number of labeled examples. Kaizen has similar functionality to Tropel, the system designed by (Patterson et al. 2015), but is more extensible, allows deep-feature extraction, and contains GPU support . Unlike Patterson et al., who asked the crowd to identify common objects like birds or clothing, we prompt crowd workers with visual events they have never seen before, namely tissue samples from whole slide images. We use the Camelyon16 dataset because of its large number of labeled examples that can be used for future baseline testing. Our goal is to answer the question: Can a relatively unskilled crowd train a detector for high-expertise pathologies?

2 Related Works

Using machine learning to identify pathology in whole-slide images is an expanding field. (Veta et al. 2014) describe various algorithms for the automatic counting of mitotic figures in breast tumors from teams competing in the AMIDA13 challenge (Assessment of Mitosis Detection Algorithms 2013). Many teams use shallow convolutional neural networks (five or less layers) as classifiers. Common image processing techniques include thresholding methods for image segmentation or manipulating color channels. (Cruz-Roa et al. 2014) use a similar approach to detect invasive ductal carcinoma, a common form of breast cancer, from whole-slide images. Cruz-Roa et al. crop each WSI into non-overlapping 100x100 pixel sections, and then change each patch from the RGB space to the YUV color space. The CNN for classification contains two convolutional layers with pooling. Cruz-Roa et al. show that even a shallow CNN performs better than classifiers trained on handcrafted features (RBG-hist, Fuzzy color hist, Haralick, etc). (Wang et al. 2016) use the same CAMELYON16 dataset as this paper, using deep convolutional nets to detect metastatic regions in lymph node whole-slide images. They use Otsu thresholding to process images and select patches that contain tissue as well as a patch-based classification technique. State-of-the-art deep learning models including GoogLeNet, AlexNet, VGG16, and FaceNet are trained from scratch on a variety of patch magnification levels. Wang et al. show that a deep-learning based approach is effective at solving complex classification questions in whole-slide images. How-

ever, the papers above are completely dependent on access to large, fully labeled datasets. Without access to datasets like CAMELYON16 or AMIDA13, supervised classifiers for specialized pathologies cannot be trained.

Crowdsourcing for medical purposes is also an expanding field. Both (Nguyen and Patrick 2014) and (Cocos et al. 2015) describe methods of employing the crowd for annotations of radiology reports. These reports contain textual information filled out by radiologists as opposed to visual image information. (Irshad et al. 2015) use the crowd for nucleus detection and segmentation in whole-slide images. They use the CrowdFlower platform to obtain crowdworkers and prompt them with multiple images of kidney renal clear cell carcinoma, asking them to perform both a detection task (select nuclei) and segmentation task (use polygon tool to isolate nuclei). Irshad et al. then compare the crowd results to an expert as well as automated methods. The crowd, while having a higher sensitivity than the automated method, had a lower sensitivity and precision than an expert. (Park et al. 2016) use the Amazon Mechanical Turk crowd to detect polyps in virtual colonoscopy (VC) videos. After uploading the VC videos, they present the crowd users with a binary-classification task. Park et al.'s work shows that the crowd is able to detect colon polyps with nearly the same specificity as an expert radiologist. However, the above studies require exhaustively labeled datasets. We investigate crowd-in-the-loop active learning to bootstrap classifier training and reduce labeling cost.

Active-learning has proved useful in past experiments for bootstrapping classifiers from large unlabeled datasets (Patterson et al. 2015; Collins et al. 2008; Hoi et al. 2006). The authors believe that Kaizen is the first system to explore the potential of crowd-bootstrapped classifiers for breast-cancer detection in lymph node whole-slide images.

3 The Camelyon16 Dataset

Experiments in this paper use the Camelyon16 dataset of sentinel lymph node whole-slide images. The dataset was released for the Camelyon Grand Challenge 2016 by the International Symposium on Biomedical Imaging (ISBI) for the automatic detection of metastatic breast cancer. Camelyon16 consists of 400 whole slide images (WSIs) from two independent datasets split into 270 WSIs for training and 130 WSIs for testing. The datasets were collected from the Radboud University Medical Center and the University Medical Center Utrecht. Each of the 400 whole slide images had 10 downsample levels, with level 0 being around 20 Gigapixels in size. Ground truths for the train set images were stored in both XML and binary mask format. For binary masks, values of 0 were considered negative and values of 255 were considered positive. The 130 WSIs labeled for testing did not come with ground truth labels.

4 Kaizen

Kaizen is designed to give pathologists the ability to train classifiers using a very small number of labeled images and a large unlabeled dataset. It is inspired by the Tropel active-learning system for general visual classifier creation (Patterson et al. 2015). The user is allowed to upload a dataset along with pre-labeled seed examples.

Once the dataset is loaded, the user is given full control of patch cropping and feature extraction. Users can either select pre-defined patch specifications (PatchSpecs) or create their own, setting height, width, overlap percentage, and scale percentage. Upon submission the Kaizen system performs patch extraction. Because of the large dimensions of the whole-slide images, patching takes a significant amount of memory and time. After patching, the user is given the option to add a feature extractor. Aside from commonly used feature extractors like HoG and RGB Histogram, Kaizen includes a deep-network feature extractor using off-the-shelf CNNs. Using deep convolutional neural networks for feature extraction has been shown to be effective for a variety of image classification problems [13]. We implement GPU processing, allowing Kaizen to utilize deep networks including VGG16, Caffenet, and GoogLeNet. After feature extraction, the user is then able to create a classifier with the previously uploaded pre-labeled examples. To begin, Kaizen uses the seed positive examples and randomly selects negative events from the dataset. Because the visual events are assumed to be sparse, a random selection of a single patch from the training set is likely to be negative. Kaizen then trains a linear SVM, which is used to find the top 200 most confidently scored patches from the training set. These 200 patches are then shown on the active-labeling UI.

At this point, Kaizen uses the selected images to seed the classifier, using a algorithm of the user's choice (SVM, Linear Regression, KNeighbors Regressor). Once a type is selected, the classifier is trained on the crowd selected examples. The top 200 classifications are then presented on the active UI. Kaizen continues to update the classifier after each active-learning iteration.

5 Experiments on the Camelyon Dataset

The Camelyon dataset serves as a baseline for our classifier experiments. We use 110 labeled whole slide images for the experiments, with each WSI large enough to generate hundreds of thousands of patches. Experiments ran on an Amazon EC2 g2.2xlarge instance, with a single K20 NVIDIA GPU, 8 CPU cores, and 16 GB RAM. Currently, Kaizen does not support multiple-GPUs.

5.1 Image preprocessing

We first identify which downsample level provides optimal results. Having lower downsample levels gives workers closer views of the actual cells, whereas higher levels gives workers a better overview of the shape of the lymph node tissue (see Figure 1). Lower downsample levels are useful for identifying pathologies that are more easily recognizable at the cellular level, as opposed to pathologies that are more easily detectable as large bodies.

After examining ground truth patches, it becomes clear that classifying metastasis in high-level croppings is more difficult than classification in low-level croppings. The histological appearance of cancerous tumor cells often differs

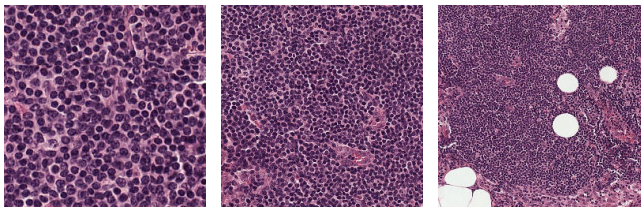
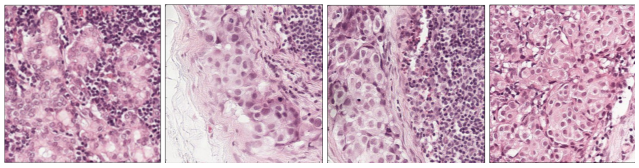
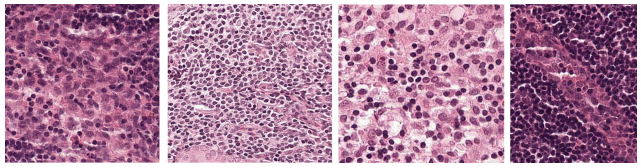


Figure 1: Downsample levels 1, 2, 3 left to right, respectively.



(a) *Metastatic cell samples*



(b) *Non-cancerous cell samples*

Figure 2: Types of cells considered.

from normal cells in three main ways starting with the cellular nucleus, which has changes in nuclear size, shape, and transparency (often reflected in stain intensity). Cancerous regions also have the tendency to form tubular structures and have a greater mitotic rate, thus giving a highly clumped appearance. Figure 2 shows various cancerous and non-cancerous patches from the whole-slide images.

We programmatically extract 256x256 cancerous patches. Patches are considered cancer-positive if they overlap with the binary mask label by more than 80%. We then create a train and test set with a 70/30 split.

5.2 Evaluation and Comparison

Once the whole-slide images and ground-truth patches are loaded into Kaizen, we create a PatchSpec that crops each whole-slide image into thousands of 256x256 subpatches. We use a deep convolutional neural network feature extractor, running off of the AlexNet architecture trained on ImageNet using model weights from Caffe's ModelZoo. Kaizen performs feature subsampling with a 200-dimensional output feature vector. Subsampling saves significant RAM and computational time, while sacrificing little in terms of image representation. For these experiments, the authors act as the crowd. Future classifier training can be completed by crowd workers with Amazon Mechanical Turk integration. The ground truth patches are sampled from a single whole-slide image because annotating a single whole-slide image is the bare minimum work required to seed the classifier. We create multiple classifiers using different numbers of starting examples: 1, 5, 10, 20, 50, and 100. Training is stopped after 3 iterations.

Classifier Starting Examples	Average Precision
1	.73
5	.61
10	.54
20	.65
50	.66
100	.47

Table 1: Average precision at iteration 3 for experimental classifiers using different numbers of seed examples.

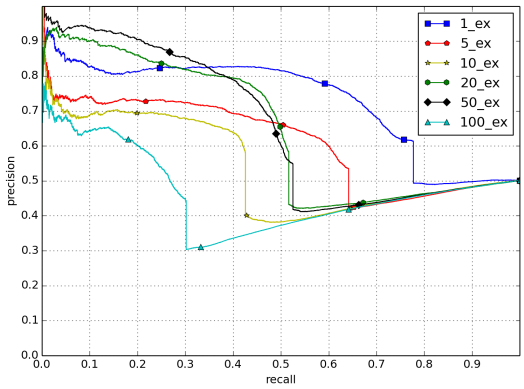


Figure 3: Precision-recall curves for all classifiers at iteration 3.

Table 1 shows average precision scores and Figure 3 shows precision-recall graphs for each of the 6 classifiers we created.

Overall, the average precision for each classifier decreased as the number of classifier starting examples increased. This could be due to using a single whole-slide image for patch extraction. The classifier learned the features of the metastatic tissue in the seed whole-slide image better as the number of starting examples increased, but failed to identify cancerous regions of other WSIs. Similarly, Kaizen, which returns the top 200 classifier results as candidate images for the next iteration, makes it easy for the classifier lock into a specific mode of cancerous tissue, returning images that look nearly identical to the starting examples (see Figure 4).

Because of the large number of unlabeled patches, successive training rounds can very well contain the same 200 patches as the previous iteration, thus preventing the classifier from learning from different examples.

From Figure 3 we observe that at a certain varying threshold, each classifier has a steep trade-off between precision and recall. In classifier "100_ex", for example, the trade-off occurs as recall reached .3, while the value is .5 for classifiers "50_ex" and "20_ex". Secondly, all of the classifiers converge to the same precision value of .5 as recall goes to 1.

Qualitatively, as the iterations progress, the returned patches with highest classifier output probability become

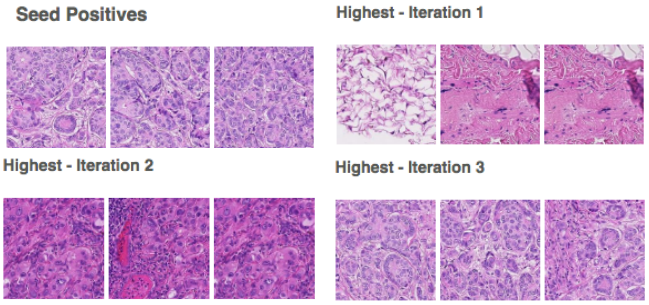


Figure 4: Ground truth patches from Classifier 100_ex and retrieved unlabeled images from successive iterations.

more and more visually similar to the ground truth. This is an expected output. However, not all cancerous patches within the test and even train dataset look similar to the ground truth images provided. By using only a single whole-slide image for seed patch extraction, we train the classifier to detect only cancerous patches that are highly visually similar to the ground truth, explaining the decline in average precision as we seed the classifier with more examples from the same WSI. Kaizen can retrieve results from specific modalities very well but that it performs poorly in retrieving results from a variety of modes.

6 Conclusion and Future Work

In this paper we introduce Kaizen, a system for bootstrapping classifiers for pathology identification in WSIs with a small number of training examples. Kaizen includes patching and feature extraction functionality so users can upload annotated patches from WSIs and use them to create more labeled examples and seed classifiers. Our experiment on the Camelyon16 dataset shows that seeding Kaizen with examples from only a single whole slide image is not enough to create a classifier able to perform well on a variety of lymph node tissues. The classifier trained using only a single seed example performed better than all others. Increasing the number of starting examples from a single WSI does not improve the classifier, rather, it decreases average precision. A future experiment would include examples from multiple different whole-slide images in order to test whether Kaizen can learn the multimodal nature of Camelyon16. Similarly, training a standard classifier (SVM or the like) on the Camelyon dataset would provide a good baseline for comparing our classifiers. Future work also includes exploring iteration cutoff procedures and testing autoencoder feature extraction. While the off-the-shelf AlexNet model functioned properly, it would be interesting to see whether an autoencoder trained on the whole-slide images will improve performance.

7 References

Dung H M Nguyen and Jon D Patrick. “Supervised machine learning and active learning in classification of radiology reports.” Journal of the American Medical Informatics Association. 21.5 (2014): 893-901.

M Veta, P J van Diest, S M Willems, H Wang, A Madabhushi, A Cruz-Roa, F Gonzales, A Larsen, J Vestergard, A Dahl, D Ciresan, J Schmidhuber, A Giusti, L Gambardella, F B Tek, T Walter, CW Wang, S Kondo, B J Matuszewski, F Precioso, V Snell, J Kittler, T E de Campos, A M Khan, N M Rajpoot, E Arkoumani, M M Lacle, M A Viergever, J Plium. “Assessment of Algorithms for Mitosis Detection in Breast Cancer Histopathology Images.” arXiv. Cornell University Library, 21 Nov. 2014. Web.

Cocos, Anne, Aaron Masino, Ting Qian, Ellie Pavlick, and Chris Callison-Burch. “Effectively Crowdsourcing Radiology Report Annotations.” Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis (2015): 109-14. Web.

Cruz-Roa, Angel, Ajay Basvanhally, Fabio Gonzalez, Hannah Gilmore, Michael Feldman, Shridar Ganesan, Natalie Shih, John Tomaszewski, and Anant Madabhushi. “Automatic Detection of Invasive Ductal Carcinoma in Whole Slide Images with Convolutional Neural Networks.” Medical Imaging 2014: Digital Pathology (2014): n. pag. SPIE. Web.

Wang, Dayong, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew Beck. “Deep Learning for Identifying Metastatic Breast Cancer.” arXiv. 18 Jun. 2016. Web.

Gecer, Baris, Ozge Yalcinkaya, Onur Tasar, and Selim Aksoy. “Evaluation of Joint Multi-Instance Multi-Label Learning For Breast Cancer Diagnosis.” arXiv. 10 Oct. 2015. Web.

Hou, Le, Dimitris Samaras, Tahsin Kurc, Yi Gao, James Davis, and Joel Saltz. “Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification.”arXiv. 9 Mar. 2016. Web.

Patterson, Genevieve, Grant Van Horn, Serge Belongie, Pietro Perona, and James Hays. “Tropel: Crowdsourcing Detectors with Minimal Training.” HCOMP (2015): n. pag.Third AAAI Conference on Human Computation and Crowdsourcing. Web.

Irshad, H., L. Montaser-Kouhsari, G. Waltz, O. Bucur, J.a. Nowak, F. Dong, N.w. Knoblauch, and A.h. Beck. “Crowdsourcing Image Annotation For Nucleus Detection And Segmentation In Computational Pathology: Evaluating Experts, Automated Methods, And The Crowd.” Pacific Symposium on Biocomputing 2015 (2015): n. pag. Web.

Park, Ji Hwan, Seyedkoosha Mirhosseini, Saad Nadeem, Joseph Marino, Arie Kaufman, Kevin Baker, and Matthew Barish. “Crowdsourcing for Identification of Polyp-Free Segments in Virtual Colonoscopy Videos.” ArXiv. N.p., 27 June 2016. Web.

Collins, Brendan, Jia Deng, Kai Li, and Li Fei-Fei. “Towards Scalable Dataset Construction: An Active Learning Approach.” ECCV (2008): 86-98. Web.

Hoi, Steven C. H., Rong Jin, Jianke Zhu, and Michael R. Lyu. “Batch Mode Active Learning and Its Application to Medical Image Classification.” The 23rd International Conference on Machine Learning (2006): n. pag. Web.

Razavian, Ali Sharif, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. “CNN Features Off-the-Shelf: An Astounding Baseline for Recognition.” IEEE Conference on Computer Vision and Pattern Recognition Workshops (2014): n. pag. Web.