DOI: 10.1007/978-3-658-47422-5_45

Intrinsic Correspondence of Classification Ground Truth and Image Content on the Example of Endoscopic Images

Johannes Schuiki ♦ Andreas Uhl

Department of Artificial Intelligence and Human Interfaces, University of Salzburg, Austria

Abstract

On what basis class labels ("ground truth") get assigned to images heavily depends on the application scenario, sometimes even without visual inspection of the data. Therefore, it can be of interest to evaluate whether distinguishing intrinsic structures exist within the image data. In this study, it is investigated if images from five small-scale endoscopic datasets where class labels were assigned based on domain-specific criteria can be algorithmically clustered into the desired classes. The image classification task is treated as a clustering comparison problem by comparing ground truth labels with clustering results derived from a variety of image representations.

Contents

This work has been partially supported by the Salzburg State Government WISS2025 Research and Transfer Lab on "Artificial Intelligence in BioMedical Image Analysis (AIBIA)".

1 Introduction

Image classification is one of the fundamental tasks in the domain of computer vision. It involves assigning a label or category to a given image based on its visual content. With rapid advances in deep learning techniques, image classification has become more accurate and efficient, enabling machines to perceive and interpret visual information with human-like accuracy. However, having properly labeled data points for learning and evaluation is paramount to satisfying results. Such labels are a subset of metadata widely known in the machine learning and computer science community as "ground truth" and used to train a model via supervised learning. This labeled data provide the "correct answers" that the model should learn to predict. Throughout various domains of science employing computer vision tools, the understanding of ground truth and also the process of obtaining slightly differs. In the domain of medicine (and therefore heavily influenced is the discipline of medical imaging), the term "gold standard" is used for the benchmark that is available under reasonable conditions [1]. For data annotations, experts often rely on the results from gold standard methods.

Because the process of assigning a class label to images does not necessarily rely on visual inspection, it can be unclear whether relevant structures for distinguishing actually exist within the image data. Also, some domains require experts for data labeling while it is known [2] that, especially in the medical fields, an inter-observer discrepancy exists. Furthermore, there is also the possibility that learning based approaches learn something unintended but rather forced by labeling. Thus, it is of interest to explore if a collection of images can be grouped according to the defined class labels, i.e. to assess the extent of correspondence between class label and visual content.

In this work, we investigate the correspondence between label and image content experimentally for five small-scale datasets. To evaluate this correspondence, a variety of image representations are first grouped into clusters using a number of clustering algorithms and obtained cluster labels are afterwards compared to the ground truth labels using a clustering comparison metric.

2 Materials and methods

2.1 Data

For the experiments in this work, five small-scale endoscopic image datasets are employed where data labeling happened under supervision of domain experts. Two image sets deal with celiac disease while the other three focus on colonic polyps. Although the datasets provide annotations for various stages of the respective conditions, the labels are always collapsed to a two-class scenario (healthy vs. pathologic or benign vs. malignant). In the following, every dataset is briefly introduced. A property shared across all the employed datasets is their imbalance in terms of class distribution and that classes can be further subdivided according to patient-ids. Thus, dataset statistics can be visualized using nested pie charts as depicted in Figure 1. Example patches for every dataset are shown in Figure 2.

2.1.1 Celiac disease

Celiac disease (CD) is a chronic autoimmune condition that affects approximately 1% of the global population. Currently, the gold standard for diagnosing CD is endoscopy with biopsy. Specimen are then classified in a histological analysis according to the modified Marsh classification proposed in [3] which distinguishes between classes Marsh-0 to Marsh-3, with subclasses 3A, 3B, and 3C. The used dataset includes classes Marsh-0, Marsh-3A, Marsh-3B and Marsh-3C. For this dataset, patient-level biopsy results were available rather than area-specific data. Consequently, final patch labeling incorporated both histological and visual criteria. Within this work, class 0 is used as the healthy class and classes 3A, 3B & 3C are collapsed to form the disease class. Imaging techniques within the scope of this work include the modified immersion technique (MIT) under traditional white-light (WL) illumination, as well as MIT under narrow band imaging (NBI). The green (the two left-most) pie charts in Figure 1 visualize the distribution for the celiac samples. Further information about the dataset can be found in [4].

2.1.2 Colonic polyps

Colonic polyps are growths that develop on the inner lining of the colon or rectum. For colon examination, the current gold standard is colonoscopy. Colonoscopy allows to investigate the inside of the colon using an endoscope. This work employs three

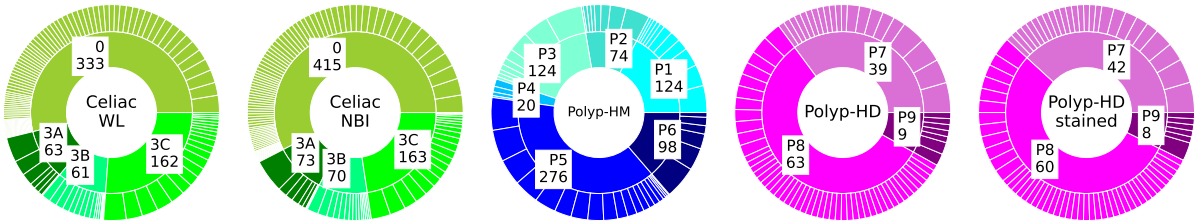

Figure 1: Distribution of the data for all datasets used in this work. The inner ring depicts the proportion of the data for each class; segments in the outer ring indicate images from the same patient. The accompanying integer value indicates the number of samples per class.

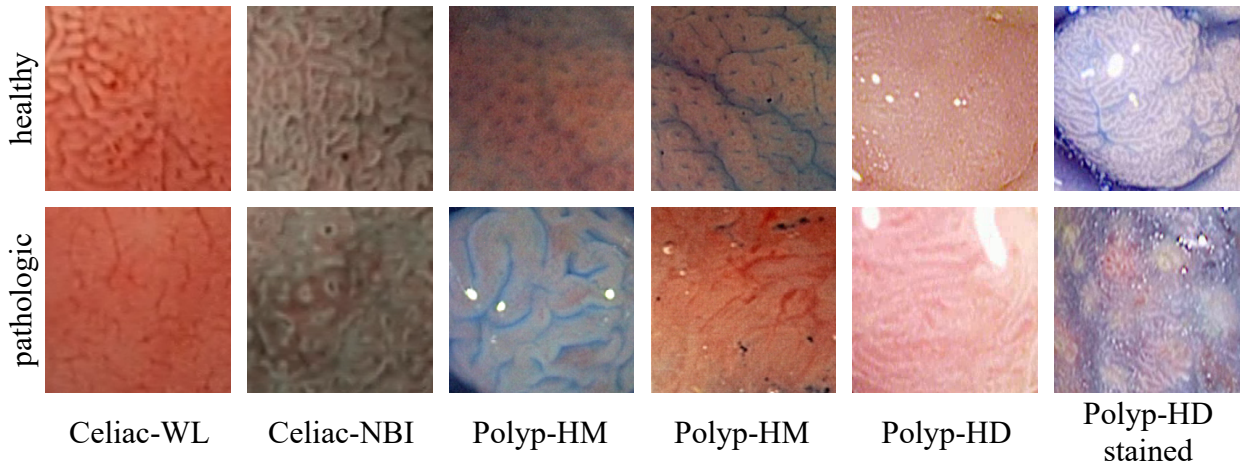

Figure 2: Example patches from all datasets included in this work.

such datasets, acquired using different imaging techniques:

- Polyp-HM: This dataset was acquired using a high magnification (HM) colonoscope together with a contrast staining. Images within this image set are classified based on six classes, known as pit patterns [5]. These class labels are assigned solely based on the visual content. The data distribution of the six classes, denoted as P1 - P6, are visualized by the blue (center) pie chart in Figure 1. They can be summarized as P1 and P2 including benign polyps and P4 - P6 including malignant polyps. Additional information about the data can be found in [6].

- Polyp-HD: Two further image datasets were acquired using a high definition (HD) endoscope with and without additional staining. For these datasets, however, images are only divided into three classes, "normal", "non-invasive" and "invasive" based on their histological diagnosis (abbreviated as P7 - P9, where P7 is treated as the healthy class and P8 & P9 together are viewed as the pathologic class). The data distribution is depicted in the magenta (the two rightmost) pie charts in Figure 1. For further information the interested reader is referred to [7].

2.2 Image representation

In our experiments, we employ a variety of image representations to evaluate the intrinsic correspondence between ground truth labels and image content. This approach is necessitated by the absence of a universally optimal image representation. By evaluating multiple representations, we aim to comprehensively assess the intrinsic relationship between image content and ground truth labels across different levels of abstraction.

- Raw pixel values: We begin with flattened raw pixel values, which provide the closest approximation to a "natural description" of the image. While this representation preserves all original information, it lacks desirable properties such as spatial invariance.

- Dimensionality reduction: We apply principal component analysis (PCA) to reduce the dimensionality of the raw pixel data. This step helps to mitigate the curse of dimensionality and potentially reveals latent structures in the data.

- Neural Networks: We utilize feature embeddings extracted from pre-trained deep learning models. These representations excel at identifying relevant structures in images due to their ability to capture complex relations within the data. However, it is important to note that these networks, trained on ImageNet, introduce biases inherent to their training data. Yet, in pursuit of finding a "natural representation" of the images and also to avoid learning something unintended by force, we explicitly refrain from fitting a model to our data. Network model architectures are taken from the PyTorch Image Models collection (https://github.com/rwightman/pytorch-image-models) which encompasses a large variety of state-of-the-art model architectures. In particular, seven different architectures are employed in this work: ResNet-18 & ResNet-50, MobileNetv3, TinyNet, EfficientNet, Vision Transformer (ViT) & Mobile-ViT. Within this work, models are utilized such that images are embedded into a feature representation using a pre-trained network where the final classification layer is removed.

2.3 Consensus of ground truth and image content

To quantify the alignment between the assigned ground truth and the image descriptors outlined in Section 2.2, these descriptors are grouped using a variety of algorithmic approaches for data clustering. Three clustering algorithms are employed for the experiments in this work: k-means clustering, hierachical (or agglomerative) clustering and spectral clustering. Implementations from the *scikit-learn* python library are used. After clustering, images have both a ground truth class label and a newly generated cluster label. For comparing both label assignments, the ground truth assignment can be treated as a different clustering result and thus the comparison be treated as a case of clustering comparison. One such comparison method is the adjusted mutual information (AMI) [8] metric, which is a version of mutual information that introduces a correction for chance by considering the expected similarity of all pairwise comparisons. For calculation of an expected similarity value, a random model must be chosen based on how the data points can be clustered. According to [8], the right random model to choose for experiments within this work is the *one-sided* (always comparing against the same basis, i.e. ground truth) comparison with a *fixed number of clusters*.

3 Results

This section presents the experimental results. Bar charts in Figure 3 depict the AMI values for comparing the clustering labels with the ground truth labels for every image representation and clustering algorithm used in this work. As mentioned in Section 2.1, image classes are always collapsed to a two-class scenario. Regardless of how many of the first principal components are used as a data representation, results for PCA embedding are quite similar. Hence, only one case is used for visualization. All experiments were carried out ten times. Error bars indicate the standard deviation.

4 Discussion

The overview presented in Figure 3 indicates that raw and PCA representations result in an overall low AMI value, suggesting their unsuitability for our analysis. Except for Polyp-HM, every dataset has a configuration achieving a metric value of 0.4 of higher. One particular combination even yields a metric value exceeding 0.5 AMI. Here, 0.5 AMI can be interpreted as roughly 90% of agreement between class label and clustering, as discussed in Section 4.1. It is interesting to observe that the Polyp-HM dataset (labels assigned solely based on visual appearance) appears to yield the lowest agreement of labels and algorithmic clustering, whereas the AMI values for other datasets (labels assigned also/solely based on histological analysis) are relatively high. We can further observe that the choice of clustering algorithm and feature embedding has a high impact on the results. Hence, we can conclude that the employed methodology appears to only provide insight into coarse trends, but seems insufficient to give a definitive answer to the correspondence between image labels and visual content.

4.1 Relationship between AMI and classification accuracy

In an attempt to establish a relationship between the AMI and an estimation of the classification accuracy, two artificial clusterings were gradually alienated and the corresponding metrics calculated. Figure 4

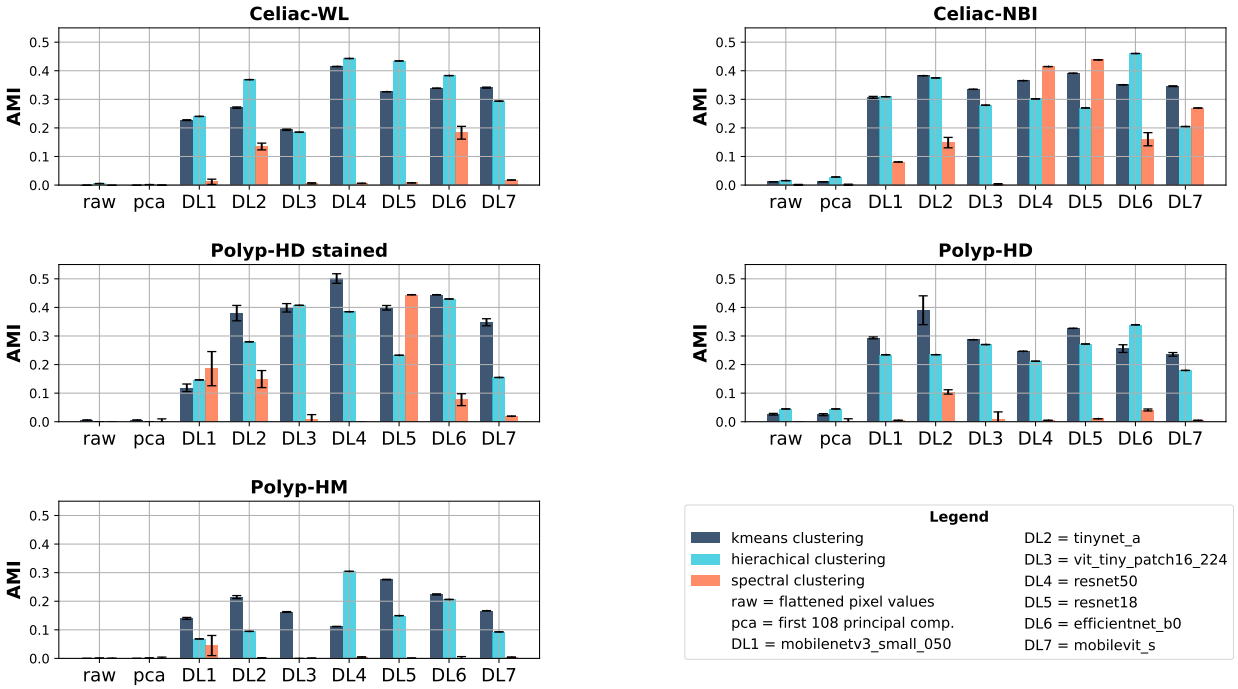

Figure 3: Results of clustering comparison using the AMI metric.

depicts the simulated relationship between AMI and accuracy for a two-class setup. Two cases for label randomization are considered: (i) permutation, where the starting point is an equal amount of elements per cluster and on every step of scrambling the labels, every cluster receives exactly one element from the other cluster. Doing so, the number of elements per cluster does not change, merely the labels get permuted. (ii) randomized, which also considers imbalanced clusters. Depending on the step, a number of elements are picked and assigned to the other cluster in a random fashion. The second case is more realistic but doing so, however, does not result in an unambiguous relationship between AMI and accuracy. The relationship can be understood as: *The clustering algorithm was able to separate the feature embeddings as if a classifier model would score x% accuracy*. Note that 0.5 accuracy corresponds to random guessing in a two-class scenario.

References

[1] J. Cardoso, L. Pereira, M. Iversen, and A. Ramos, "What is gold standard and what is ground truth?" *Dental Press J Orthod*, vol. 19, 2014.

[2] R. Corona, A. Mele, M. Amini, G. De Rosa, G. Coppola, P. Piccardi, M. Fucci, P. Pasquini, and T. Faraggiana, "Interobserver variability on the histopathologic diagnosis of cutaneous melanoma and other pigmented skin lesions.," *Journal of Clinical Oncology*, vol. 14, no. 4, 1996.

[3] G. Oberhuber, G. Granditsch, and H. Vogelsang, "The histopathology of coeliac disease: Time for a standardized report scheme for pathologists," en, *Eur. J. Gastroenterol. Hepatol.*, vol. 11, no. 10, 1999.

[4] M. Gadermayr, S. Hegenbart, R. Kwitt, and A. Uhl, "Narrow band imaging versus white-light: What is best for computer-assisted diagnosis of celiac disease?" In *2016 IEEE 13th International symposium on biomedical imaging (ISBI)*, 2016.

[5] S. Kudo, S. Hirota, T. Nakajima, S. Hosobe, H. Kusaka, T. Kobayashi, M. Himori, and A. Yagyuu, "Colorectal tumours and pit pattern," *J. Clin. Pathol.*, vol. 47, no. 10, 1994.

[6] M. Häfner, M. Liedlgruber, A. Uhl, A. Vécsei, and F. Wrba, "Delaunay triangulation-based pit density estimation for the classification of polyps in high-magnification chromo-colonoscopy," *Comput Methods Programs Biomed*, vol. 107, no. 3, 2012.

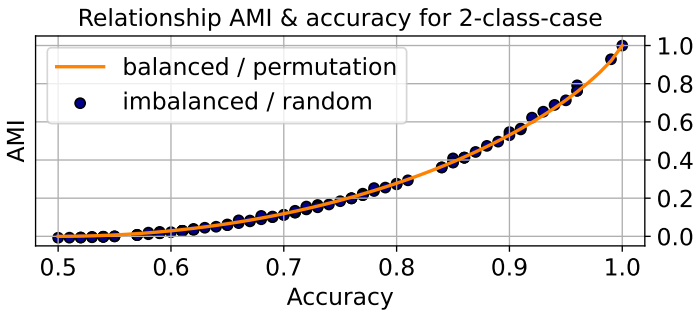

Figure 4: Simulated relation between AMI and accuracy for a two-class-scenario.

[7] E. Ribeiro, M. Häfner, G. Wimmer, T. Tamaki, J. Tischendorf, S. Yoshida, S. Tanaka, and A. Uhl, "Exploring texture transfer learning for colonic polyp classification via convolutional neural networks," in *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, 2017.

[8] A. J. Gates and Y.-Y. Ahn, "The impact of random models on clustering similarity," *J. Mach. Learn. Res.*, vol. 18, no. 1, 2017.