Do We Need Annotation Experts? A Case Study in Celiac Disease Classification

Roland Kwitt[1], Sebastian Hegenbart[1], Nikhil Rasiwasia[3], Andreas Vécsei[2], and Andreas Uhl[1]

[1] Department of Computer Science, University of Salzburg, Austria
[2] St. Anna Children's Hospital, Medical University Vienna, Austria
[3] Yahoo Labs! Bangalore, India

Abstract. Inference of clinically-relevant findings from the visual appearance of images has become an essential part of processing pipelines for many problems in medical imaging. Typically, a sufficient amount labeled training data is assumed to be available, provided by domain experts. However, acquisition of this data is usually a time-consuming and expensive endeavor. In this work, we ask the question if, for certain problems, expert knowledge is actually required. In fact, we investigate the impact of letting non-expert volunteers annotate a database of endoscopy images which are then used to assess the absence/presence of celiac disease. Contrary to previous approaches, we are not interested in algorithms that can handle the *label noise*. Instead, we present compelling empirical evidence that label noise can be compensated by a sufficiently large corpus of training data, labeled by the non-experts.

1 Motivation

Many problems in medical imaging involve some sort of decision-making process based on the visual appearance of images acquired by some modality. Typical examples include, but are not limited to, computer-aided assessment of various types of cancer, or the classification of tissue types for subsequent segmentation. The prevalent paradigm of these approaches is to assume the existence of *expert-annotated* data to train a classification system which is then used to make predictions for new data instances. For segmentation tasks, predictions are typically made on a pixel level, whereas for computer-aided diagnosis, predictions are made on suitable representations of images regions or even the full images.

While many approaches demonstrate fairly good performance for the respective task, classifier training inherently depends on the pristine expert annotations. In practice, though, such annotations are typically hard to obtain, since the annotation task is often time-consuming and thus expensive. Consequently, the amount of available training data tends to be rather limited which can lead to non-conclusive statements about the generalization ability of a system. This is in contrast to many computer vision problems, where annotation tasks can typically be "crowd-sourced" easily.

P. Golland et al. (Eds.): MICCAI 2014, Part II, LNCS 8674, pp. 455–462, 2014.

In this work, we ask the question whether we can circumvent the requirement for expert annotations by using a substantially larger corpus of training data labeled by non-experts. This is an interesting and potentially far-reaching question, since most works in the literature that assume a certain amount of label noise either rely on a separate pre-processing step to remove suspect samples [3], or on learning algorithms that can handle the noise implicitly. Examples include multiple instance learning [7], robust variants of logistic regression [2], or SVM formulations that incorporate the possibility for label flipping [13]. Either way, a change in the classification architecture is required to handle label noise.

In contrast, we do not propose a novel learning algorithm to handle label noise, but instead study the fundamental question if, in certain cases, the potentially negative impact of noisy training data can be alleviated by simply increasing the number of available training instances. While typical crowd-sourcing scenarios are impractical for medical data due to privacy issues, it is still relatively easy to obtain non-expert annotations from supporting personnel for instance. In Leung et al. [7], similar advantages have been highlighted when using "amateur" raters for video annotation tasks. It is important to note, though, that such strategies will only be suitable for certain visual recognition problems where little or no domain-specific knowledge is required to achieve reasonable annotation performance with moderate training effort. Finally, we highlight the difference to weakly-supervised segmentation problems, such as in [10]. In these problems, labels are given at the image-level (not the pixel-level) and indicate the presence/absence of some object of interest (e.g., Crohn's disease [10]). Nevertheless, labels are assumed to be correct which corresponds to 100% sensitivity at the pixel-level. In our case, with noisy image-level labels this is not guaranteed.

Contribution. The contribution of this work is an experimental study on the impact of noisy, non-expert image labels on the performance of a classification system to assess the presence/absence of celiac disease in endoscopy imagery. This is a clinically relevant problem, since it's relatively easy to acquire images but expert labels for a large corpus of data are hard to obtain, not least since consistency with the histopathological diagnosis is required. Based on a study with eight volunteers, we first establish a basis of what error is to be expected. By relying on a standard classification architecture and three state-of-the-art image representations, we then present empirical evidence that a large corpus of non-expert labeled data can in fact compensate for the potentially negative impact of label noise.

2 Experimental Study

In our experimental study, we consider the problem of automated assessment of endoscopy imagery for the presence/absence of celiac disease, i.e., a complex autoimmune disorder caused by the introduction of materials containing gluten such as wheat, rye and barley. During the course of the disease, hyperplasia of the enteric crypts occurs and the mucosa eventually looses its absorptive villi. This leads to a diminished ability to absorb nutrients. Visible celiac-specific

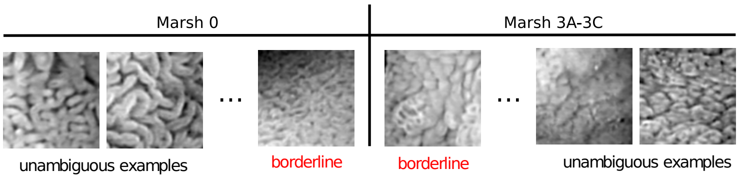

Fig. 1. Typical and borderline examples of images showing non-celiac (Marsh 0) vs. celiac disease (Marsh 3A-3C)

markers that are reported [4] to be characteristic for the pathologic changes of the mucosa include mosaic mucosal patterns, nodular mucosa, scalloping of the duodenal folds, visualization of underlying blood vessels and villous atrophy.

Dataset. Our dataset consists of images acquired during duodenoscopies using pediatric gastroscopes without magnification. The main indications for endoscopy were the diagnostic evaluation of dyspeptic symptoms, positive celiac serology, anemia, malabsorption syndromes, inflammatory bowel disease, and gastrointestinal bleeding. Images were recorded by using the modified immersion technique. The condition of the mucosal areas covered by the images was determined by histological examination of biopsies from the corresponding regions. Severity of villous atrophy was classified according to the modified Marsh classification [11]. This histological classification scheme identifies six stages of severity of celiac disease, ranging from class Marsh 0 (i.e., no visible change of villi structure) up to class Marsh 3C (i.e., absence of villi). A medical expert assisted in extracting suitable sub-images with a dimension of 128×128 pixels from the images captured during endoscopy. Each image shows specific markers for either the absence or presence of celiac disease. While the expert-guided extraction process slightly biases the results, no domain-specific knowledge is needed in practice, since the selection of sub-images is only guided by inspection with respect to certain quality criteria (e.g., sharpness or distortions). Experts were only involved to establish a ground-truth for evaluation purposes. Our dataset consists of 1050 images from 320 patients with 592 images (240 patients) categorized (consistent with the histopathology) as normal (Marsh 0) and 458 images (80 patients) categorized as containing evidence for celiac disease (Marsh 3A-3C). All images from a single patient have a consistent label. We focus on this, most clinically-relevant binary categorization, as the distinction between all six classes is difficult during endoscopy, even for specialists. This is due to the non-distinct visual appearance of certain classes. A reliable, fine-grained categorization can only be done using histopathology. Some typical and borderline cases for Marsh 0 vs. Marsh 3A-3C cases are shown in Fig. 1.

2.1 Performance of Non-expert Annotators

To assess the performance of human, non-expert annotators, we randomly selected 100 images with a roughly equal split of *non-celiac* vs. *celiac* instances

Table 1. Performance (in %) of eight human “non-expert” annotators

	Accuracy	**Specificity**	**Sensitivity**
AVERAGE	83.8 ± 3.9	81.3 ± 6.7	87.5 ± 9.3
MIN.	79.0	68.3	70.0
MAX.	92.0	90.0	97.5

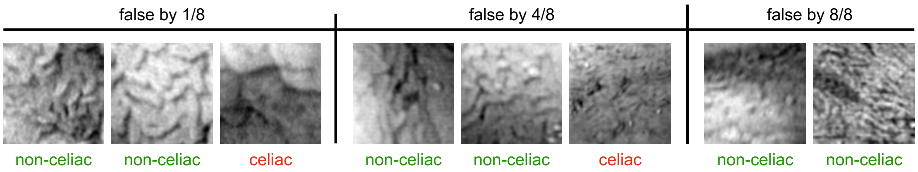

Fig. 2. Labeling erros of human ”non-expert“ annotators. Images are grouped by errors made by single individuals (left) to errors made by all annotators (right). Images are annotated by their actual ground-truth label.

(60/40). These images were then shown to eight non-expert volunteers, after a 10 minute introduction to the annotation problem, where the differences between celiac vs. non-celiac disease were illustrated on an example of 10 typical images per category. This introduction was intended to quickly outline (1) how the disease manifests in architectural changes of the villi and (2) how this affects the visual appearance. Each person then labeled all 100 images individually, *without* knowledge of the class distribution. In addition to the assigned labels, we also recorded the time spent to annotate each image. Interestingly, the mean annotation time per image was only 1.9 seconds with a low standard deviation of ± 0.3 seconds. Table 1 lists the accuracy, sensitivity and specificity, averaged over all eight annotators. In our setup, sensitivity corresponds to the percentage of true celiac images actually identified as celiac images by the annotators.

A closer examination of the annotation errors (with respect to the ground truth) revealed that, out of 100 images, only two images were consistently assigned a false label by all annotators, see Fig. 2 (right). Although, the ground truth label is *non-celiac* in these cases, the presence of the villi is not clearly pronounced making these images hard to categorize without substantial domain knowledge. Some images, falsely labeled by half of the annotators are shown in Fig. 2 (middle). For the *non-celiac* cases, the situation is similar as before in the sense that villi are less pronounced; for the *celiac* image, the non-typical scaling is deceiving and suggests the presence of villi. The left-hand side of Fig. 2 shows images which were falsely labeled only by *single* individuals each. For these images, the appearance is relatively typical for the respective category.

2.2 Classification Architecture and Evaluation Protocol

We implement a standard classification architecture with a linear support vector machine as a discriminant classifier at the end of the pipeline. Three variants of image representations are used: (1) the state-of-the-art Fisher vector encoding of [12], computed from SIFT descriptors extracted on a dense 6×6 pixel grid; (2) a standard local-binary pattern (LBP) texture representation [9] with 3 scales, 8 neighbors and uniform patterns [8]; (3) a transform-domain based approach to statistical texture characterization that uses the mean and standard deviation of complex (dual-tree) wavelet-subband coefficients (at 6 scales) for image representation [6]. Fisher vectors represent a generative-discriminative approach, whereas approaches (2)-(3) are purely discriminative approaches. We remark that the focus of this paper is *not* on designing an optimal classification architecture, but to study the impact of label noise and an increasing amount of training data within established frameworks.

Implementation. All three approaches are implemented in MATLAB using `vlfeat` [14], the linear C-SVM implementation of `LIBLINEAR` [5] and custom implementations of [9] and [6]. Gaussian mixture model estimation for Fisher vectors is done via the standard EM algorithm using diagonal covariance matrices. Component weights, mean vectors and covariances are initialized using k-means++. In all experiments, we use 8 mixture components. While this is a relatively low setting, we remark that our problem is only binary. In vision problems, the number of categories, and consequently the appearance variability, is often much higher, thus requiring a larger number of components.

Evaluation Protocol. *All* reported results are averaged over 50 cross-validation (CV) runs. In each CV run, a random split between training and evaluation image data is selected such that 90% of all patients are used for training and the remaining patients are used for testing. Although we will restrict the amount of training data in some experiments, this restriction applies only to the 90% of the training portion, whereas the evaluation portion remains unchanged. We will refer to the number of patients used for training by N. Further, we ensure a balanced class distribution. The SVM cost factor C is cross-validated on the training splits using an additional 5-fold CV and $C \in \{0.5, 1, 2, 4, 8, 16, 32\}$.

2.3 Results

Impact of Training Set Size. In our first experiment, we investigate the impact of increasing the amount of training data using expert labels. We start by randomly selecting 10% of the patients in the training portion of each 90/10 CV split and evaluate the classification performance. We then successively increase the amount of data by increments of 10% until all patients of the original training split are used. As already mentioned, the size of the testing set is unaffected by these changes. Fig. 3 shows the average CV accuracy as a function of the fraction of patients used for training.

For all three image representations, it appears that performance starts to level off as 50% of the patient data is used for training. Given our experimental setup, this is equivalent to ≈ 144 patients. Interestingly, a Wilcoxn rank-sum test at $\alpha = 0.001$ reveals that at $3N$ to $4N$ results start to become significantly different from the results of using all training data (i.e., at $10N$). In setups with more

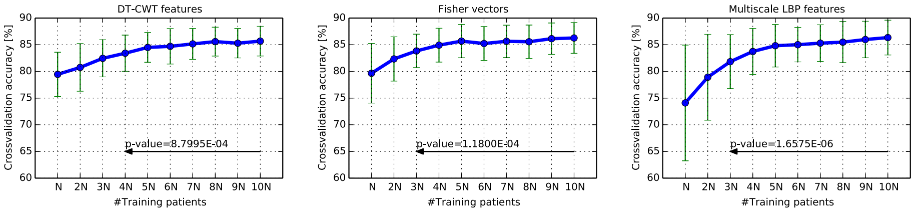

Fig. 3. Impact of increasing the training set size (w.r.t. the # patients), starting from 10% (corresponds to N) of all patients available in the training portion of the data.

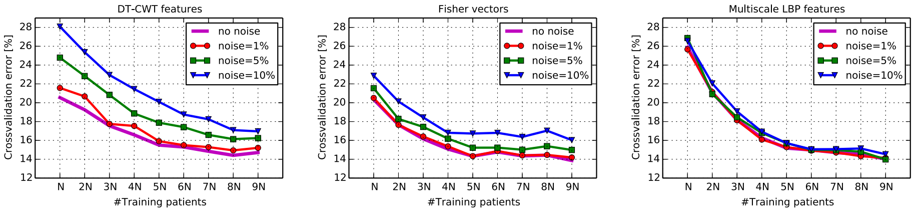

Fig. 4. Impact of different amounts of label noise as a function of the training set size w.r.t. the #patients (best-viewed in color)

than two classes, we expect the "level-off effect" to occur substantially later, due to the increased complexity of the problem. In practice, this means that expert labels are required for at least ≈ 480 images (since we have 3 images/patient on avg.) to achieve stable performance.

Impact of Artificial Label Noise. In our second experiment, we study (1) the impact of increasing label noise to simulate non-expert annotators with gradually decreasing performance, and (2) the impact of increasing the number of training samples at the same time. This allows to assess *if*, and to what extent, compensation of label noise can be achieved. We remark, though, that the scenario of *random* label flipping is an unrealistic worst case. In fact, as we have seen in §2.1, labeling errors tend to happen for borderline cases and do not occur totally random. For better illustration, we select the CV error instead of accuracy as an evaluation measure in Fig. 4.

We observe that the positive effect of increasing the training corpus is *not* mitigated by the introduction of label noise. In fact, up to a certain size of pristine training labels, we can achieve equal rates by simple increasing the size of the noisy training corpus. On the example of *DT-CWT features* for instance, $9N$ noisy labels suffice to achieve an error that is comparable to the error achieved by using $5N$ pristine labels (in fact, the null-hypothesis of equal population median cannot be rejected at $\alpha = 0.001$ with a p-value of 0.09). While the magnitude of the impact of label noise seems to be dependent on the image representation, the general behavior remains the same. For the other image representations the compensation effect is even more pronounced.

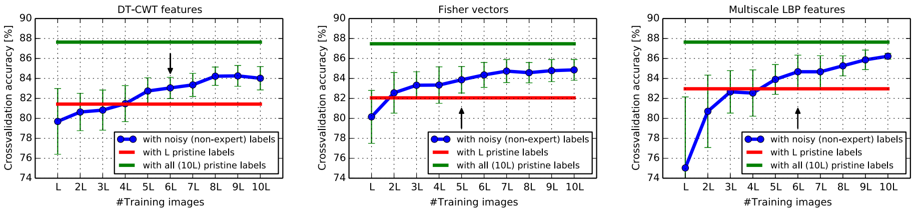

Fig. 5. CV performance as a function of the number of images labeled by non-experts. The baseline result at L is obtained by taking 50/500 images but training is performed using the pristine labels; the performance when *all* 500 pristine training labels are used is indicated by the top line (best-viewed in color).

Training with Non-expert Labels. We consider the actual practical scenario where 500 (randomly chosen) images are labeled by a non-expert, i.e., one volunteer from the experiment in §2.1. The "top annotator" is selected and reaches 89% accuracy on this set. We then compare the classification performance of a system trained with $L = 50$ (10%) of the 500 images using pristine labels vs. systems trained on an increasing number of images with non-expert labels. All remaining $1050 - 500 = 550$ images are used for testing. Results are shown in Fig. 5. We performed a left-tailed Wilcoxn rank-sum test at $\alpha = 0.001$ to assess the null-hypothesis that the population median of the CV results obtained with L pristine labels is *less than* the median of the results obtained with non-expert labels (for each training set size)[1]. The position at which the null-hypothesis cannot be rejected is marked by an arrow. For all three representations this occurs at $6L$ or earlier (with different p-values).

3 Discussion

Given the presented results, several points are worth discussing. First, as we have shown in Fig. 3, a small training corpus with pristine labels does not suffice to achieve stable performance, at least not for the considered problem of celiac disease assessment. In fact, a substantial amount of data is needed until results stabilize and improvements level-off. This effectively shows that limited availability of expert data is an actual problem.

Second, we have presented empirical evidence that a large corpus of non-expert labeled (i.e., noisy) training data can in fact be used to build a classification system that performs equally well as a system trained solely on a limited number of pristine labels. Further, in our particular problem, the relatively good performance of the non-experts reduces the amount of training data required to compensate for the noisy labels. Nevertheless, the question obviously arises how this behavior generalizes to other problems. In our case, visual categories have

[1] We correct for multiple comparisons using the Benjamini-Hochberg [1] procedure to control for FDR.

relatively distinct appearances which renders the problem appropriate for non-experts. In situations with more categories or less distinct visual characteristics, non-experts are likely to perform worse and the amount of data needed to compensate for errors might be larger. However, on difficult problems, the probability of expert errors is expected to be higher as as well.

Finally, our results indicate that the architecture of existing systems does not necessarily need to be changed if label noise introduced by non-experts annotators is expected, as long as enough data is available. In problems where the task of acquiring images is not the limiting factor, this could substantially broaden the use of computer-aided diagnosis or decision support systems, due to the sudden availability of large training corpora.

Acknowledgements. This work has been supported, in part, by the Austrian Science Fund (project no. 24366).

References

1. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B 57(1), 289–300 (1995)
2. Bootkrajang, J., Kabán, A.: Label-noise robust logistic regression and its applications. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part I. LNCS, vol. 7523, pp. 143–158. Springer, Heidelberg (2012)
3. Brodley, C., Friedl, M.: Identifying mislabeled training data. J. Artif. Intell. Res. 11, 131–167 (1999)
4. Dickey, W., Hughes, D.: Prevalence of celiac disease and its endoscopic markers among patients having routine upper gastrointestinal endoscopy. Am. J. Gastroenterol. 94, 2182–2186 (1999)
5. Fan, R., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. JMLR 9, 1871–1874 (2008)
6. Kwitt, R., Uhl, A.: Modeling the marginal distributions of complex wavelet coefficient magnitudes for the classification of zoom-endoscopy images. In: MMBIA (2007)
7. Leung, T., Song, Y., Zhang, J.: Handling label noise in video classification via multiple instance learning. In: ICCV (2011)
8. Mäenpää, T., Ojala, T., Pietikäinen, M., Soriano, M.: Robust texture classification by subsets of local binary patterns. In: ICPR (2000)
9. Mäenpää, T., Pietikäinen, M.: Multi-scale binary patterns for texture analysis. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 885–892. Springer, Heidelberg (2003)
10. Mahapatra, D., Vezhnevets, A., Schüffler, P., Tielbeek, J., Franciscus, M., Buhmann, J.: Weakly supervised semantic segmentation of Crohn's disease tissues from abdominal MRI. In: ISBI (2013)
11. Oberhuber, G., Granditsch, G., Vogelsang, H.: The histopathology of coeliac disease: time for a standardized report scheme for pathologists. Eur. J. Gastroen. Hepat. 11, 1185–1194 (1999)
12. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR (2007)
13. Vahdat, A., Mori, G.: Handling uncertain tags in visual recognition. In: ICCV (2013)
14. Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms (2008), `http://www.vlfeat.org/`