A Whole-Slide is Greater Than the Sum of Its . . . Patches

Anon

Abstract

Muscular-invasive bladder cancer (MIBC) is a common form of cancer which can necessitate complex treatment decisions. Different methods involving machine learning have been developed with the goal of improving and making MIBC diagnosis more specific, and thus limiting the amount of invasive testing needed for MIBC patients. A particularly fruitful direction of research involves the use of tissue images and the application of deep learning. In order to deal with extremely large whole slide images (WSIs), the state of the art methods approach the problem by using a patch-based convolutional neural network which takes small patches (often 256×256 pixels) of WSIs as input and provides a classification of cancerous or not-cancerous as output. Patch-to-slide classification is then often achieved by classifying a WSI as cancerous if and only if the majority of its patches are classified as cancerous. In this work we compare different approaches to the integration of local, patch based decisions, as a means of arriving at a robust global, WSI based classification. Our results suggest that an absolute, positive patch count based decision-making, with an appropriately learnt threshold, achieves the best results.

Introduction

Bladder cancer is a common form of cancer, with about 10,000 diagnoses each year in the UK alone. The survivability of a bladder cancer diagnosis dramatically decreases if the cancer has grown into the muscle of the bladder wall (Kennelly et al. 2017). Though it is very common, especially with people who have a strong history of smoking, the prognosis of outcomes of muscular invasive bladder cancer has not improved in many years. Patient survival and treatment could potentially be improved by the use of technologies for a better prognosis of bladder cancer in muscle tissue as well as eliminating cost and time of human analysis.

One of the most promising fields of technology that can be utilized to improve bladder cancer diagnoses is deep learning. Deep learning is a subset of the Artificial Intelligence field of machine learning. Machine learning is a form of data analysis by which models are trained with large amounts of data in order to *learn*, by identifying patterns within the data, to make *predictions* about novel instances of the same type of data. Deep learning, which has proven extremely effective in a wide range of medical (Dimitriou, Arandjelović, and Caie 2019a) and non-medical applications (Cooper and Arandjelović 2020), refers to a subset of machine learning models in the form of artificial neural networks with many layers.

In the context of to bladder cancer, deep learning has been employed on a variety of data types, such as summary statistics of cell morphology in a sample (Tun, Arandjelovic, and Caie 2018), patient meta-data, images themselves (Dimitriou, Arandjelović, and Caie 2019a), and others, in order to improve diagnosis. In this work our focus is on what is both the most promising and the most challenging data modality, in terms of data volume and its complexity – images of bladder muscle tissue. In particular, since whole-slide images of bladder tissue are far too large to be used as direct input to deep learning models, we compare different ways of integrating local deep learning based decisions to make the best whole slide level decision.

Previous Work

Bladder Cancer

Bladder cancer can be divided into two types of cases: non-muscle invasive (NIBC) and muscle invasive (MIBC). NIBC is cancer found in the inner tissue of the bladder which has not penetrated to the bladder muscle. In this form the cancer does not spread outside the bladder. Once cancer has penetrated the bladder muscle it is known as muscular invasive. MIBC has a far greater chance of leading to a fatal outcome. Once cancer has spread to muscle tissue it can spread to nearby lymph nodes and other nearby structures (Kennelly et al. 2017). In traditional staging, MIBC would fit in tumours stages T2 – T4. In T2, the tumours have grown into the muscle layer of the bladder. In T3, the tumours have surpassed the muscle layer and entered tissues around the bladder. T4 describes a stage where the tumours has grown to nearby lymph nodes and other organs.

MIBC is ultimately diagnosed through imaging which can take on the form of computed tomography (CT), magnetic resonance imaging (MRI), and ultrasound (US) (Kirkali et al. 2005). Currently, imaging for MIBC is considered the

best way to determine prognosis and most appropriate treatment for patients (Witjes et al. 2014). Different staged tumours within MIBC have not improved much in terms of conveying prognosis for several years (Kirkali et al. 2005). As such, multiple invasive tests must be conducted on patients to determine a path for treatment. Otherwise, comorbidity factors are also taken into account yet these factors such as chronological age have shown little evidence of improving prognosis.

The separation of MIBC and NIBC bladder cancers as well as low-grade and high-grade tumours is of extreme importance for cancer diagnosis methods. Because of the effects on prognosis these aspects have, a combination of the following diagnostic techniques is now being recommended to provide a full prognosis of a patient's condition:

- **Pathological evaluation:** Morphological assessment of primary tumours and adjacent structures
- **Urinary cytology:** A test for abnormal cells in urine
- **Imaging studies:** Primarily MRI (magnetic resonance imaging) and CT (computed tomography) images of muscle slides.

The variability of staging accuracy is one of the most important issues with imaging studies. Currently, accuracy of image-based staging ranges from 40% to 98% and when imaging occurs after an operation on the bladder, staging accuracy decreases to a range of 32% to 55% because of the post-operation inflammation which appears similar to tumour. Therefore there is a practical need for more accurate measures based on imaging.

Machine Learning for Bladder Cancer Diagnosis

One of the most influential machine learning methods for early and late bladder cancer diagnosis was developed at the University of California, San Diego (Kouznetsova et al. 2019). The proposed model was able to classify metabolites at 82.54% accuracy, performing better on early stage metabolites rather than late stage ones (Kouznetsova et al. 2019). At the crux of the model is a Multilayer Perceptron, which is trained with Stochastic Gradient Descent, and a logistic regression loss function (Kouznetsova et al. 2019). This approach has been reported to produce promising results for identifying important biomarkers from urine samples, thus facilitating the development of non-invasive methods for cancer staging.

Wu et al. (Wu et al. 2019) also adopt a deep learning approach, in this case for classifying bladder cancer. Their convolutional neural network (CNN), trained on CT scans of over 100 patients, predicts whether a cancer case will be responsive to chemotherapy or not (Wu et al. 2019). In experiments, the area under the receiver operating characteristic curve (AUROC) achieved by the model was 0.73, which suggests comparable performance to that of human radiologists, whose AUROC score was found to be approximately 0.76. Interestingly, Wu et al. also reported that a pre-trained CNN performed better than a one trained from scratch (Wu et al. 2019).

A novel measurement combining imaging biomarkers and quantitative features is called radiomics (Ge et al. 2019). The combination of both image and quantitative data could provide a powerful basis for machine learning algorithms for diagnosis of bladder cancer (Ge et al. 2019). Features extracted post-modelling have also been shown to contribute to better understanding of bladder cancer prognosis (Ge et al. 2019). For example, Garapati et al. (Garapati et al. 2017) found that morphological and textural features were particularly helpful to stage bladder cancer slides.

Deep Learning & Whole Slide Images

Whole Slide Images (WSI) analysis presents several major challenges to deep learning based methods. Many of these stem from the fact that WSIs are often extremely large (billions of pixels is common) (Dimitriou, Arandjelovic, and Caie 2019b). This means that they cannot be presented in their whole at once to deep networks. Rather, most methods in the literature, and indeed the state of the art, approaches the problem bottom-up, by breaking up WSIs into image patches, analysing these patches using manageable deep networks, and finally using patch based decisions to arrive at a slide level one. This process, and in particular its last step, is made even more challenging by the coarseness of annotation available. In particular, labelling WSIs requires expertize and is a laborious process, which is why in most cases only a slide level label is available, but more nuanced annotation is lacking.

Pathology slide corpora available to machine learning researchers are typically annotated in one of three ways. The first of these, patch level labelling, is most fine grained and indeed most useful, but requires the most labour. With annotations at the patch level, strong supervision is possible, which generally improves performance (Dimitriou, Arandjelovic, and Caie 2019b). Considering the high labelling burden of patch level annotation, it is unsurprising that slide level annotation is far more frequent. Most deep learning based research in the published literature uses slide level annotations, with promising results reported by numerous authors (Yue, Dimitriou, and Arandjelovic 2019). Lastly, patient level annotation is sometimes used, again for time and cost saving reasons. In this case, multiple WSIs come from the same patient and are all labelled using the same label (Dimitriou, Arandjelovic, and Caie 2019b). This annotation protocol is uncommon.

Technical Approach

In this section we describe each of the steps in the data processing pipeline in detail, starting from the design of the neural network used to label patches, to the different methods which are examined in their ability to integrate patch level predictions and arrive at a slide level prediction.

Patch Classifier

We used Keras and Tensorflow, and adopted the VGG16 model owing to its favourable performance reports in the published literature (Yue, Dimitriou, and Arandjelovic 2019). Recall that the VGG16 is a very deep convolutional network with small sized kernels, which comprises 16 layers

with, in our implementation, 135 million parameters. It was implemented as a Keras Sequential model.

As suggested earlier, the network was trained on patch level with the target prediction being the label of the corresponding WSI. On average approximately 250 patches were extracted from each slide. Observe the consequence of this approach: the correct network output for some patches was the cancerous label even if the patch itself contained no actual tumours. The model should thus be seen not as classifying the patch as cancerous or not, but rather as predicting whether the patch comes from a cancerous slide. Class imbalance, resulting from there being many more non-cancerous patches than cancerous ones, was taken into account during training. Instead of random sampling, the relative sampling rates for the two classes were adjusted to as to result in a balanced cancerous and non-cancerous sample sets. Class weights were also used to mitigate the unbalanced classes.

From Patch Level to Slide Level Predictions

Recall that the due to the large size of whole slide images, direct deep learning decision-making is done on the patch level – in order words, locally in the context of the image. Hence, what needs to be done next is the integration of these local decisions into a single decision pertaining to the slide as a whole. The current standard, in that most work adopts it, employs a majority vote based consensus, that is, if most patch predictions are that of a cancerous slide, the slide is deemed cancerous, and if most are non-cancerous, the slide is deemed non-cancerous as well.

Here, we sought to investigate two alternatives to this process. The first of these is based on the thresholding of the absolute vote count. It is motivated by the simple observation that adding further health patch samples to a set of patches extracted from a slide does not change the nature (cancerous or not) of the original set; more negative votes should not affect the prediction in the way that they do in the majority vote approach. The second alternative is that of thresholding not the absolute but the relative number of cancerous votes. In other words, a slide is deemed cancerous if and only if the proportion of cancerous patches exceeds a certain value. The different patch-to-slide methods are summarized for the reader's convenience in Table 1. Note that majority vote is a special case of the proportion thresholding approach (Prop<N> in the table), with the proportion threshold equal to 0.50.

Combining Local and Global Information

As we explained in the previous sections, one of the major challenges (though by no means the only one) in the analysis of WSIs stems from their large size. It is for this reason that the state of the art methods in the literature apply deep learning not on WSIs themselves, but on their patches (sub-images), and then integrate these local decisions to arrive at a whole slide one. It is clear that valuable information is lost in this process. In particular, any geometric relationship between different localities, i.e. patches, is lost as is therefore the holistic view of the original slide. Hence, herein we also explore the possibility of combining bottom-up patch based reasoning which focuses on fine, local detail, and global reasoning which focuses on slide level features. Considering the inherent challenge noted earlier, for the latter stage, rather than attempting to feed the entire original slide into a network, and bearing in mind that the complementary part of the method (patch based) deals with detail, we use severely down-sampled WSI instead to capture slide level appearance and (implicitly) patch relationships. As this has not been attempted before, we adopt a simple decision level fusion whereby a WSI is deemed cancerous if either the aggregated patch prediction or the down-sampled slide prediction are positive.

Method	Description
MajorityVote	Classifies the WSI as cancerous iff most of its patches (i.e. $> 50\%$) are deemed cancerous.
Thresh<N>	Classifies the WSI as cancerous iff at least N of its patches are deemed cancerous.
Prop<P>	Classifies the WSI as cancerous iff the number of cancerous patches is at least P times the number of non-cancerous ones.

Table 1: Summary of the three different approaches for patch label to WSI label inference compared.

Experimental Analysis

Prior to any actual experimental analysis, we tuned and trained the adopted VGG16 network, using the standard training-validation paradigm. The final model achieved patch level precision of 0.48 and recall of 0.80.

Patch to Slide Decisions

We started the main part of our analysis by comparing different approaches at integrating patch level decisions into a unified whole slide level decision. Recall that two of the approaches we described, namely the absolute and relative thresholding based ones, have a free parameter – the respective threshold. The determination of these thresholds was also done prior to any comparison across methods, and again using the standard training-validation paradigm. In summary, we found the optimal absolute threshold in our case to be $N = 110$, and the relative one $P = 2$.

The results of our comparison are summarized by the plots in Figure 1. Firstly, note that the precision values obtained with the three methods are rather similar, with the proportion based thresholding approach being somewhat superior to the other two. Differences in recall are more stark, though, with the absolute thresholding method being significantly better both than the standard MajorityVote approach and the proportion based thresholding one. The most important conclusions that can be drawn from our results is that the absolute thresholding approach is best, and certainly much better than the MajorityVote used in the literature. Considering the theoretical arguments put forward in the previous

section for it, this is not surprising. Moreover, MajorityVote can also not be said to be absolutely preferable to the proportional thresholding method either, but rather that the preference for one or the other depends on the clinical decision regarding the trade-off between false positives (and thus the unnecessary upset and inconvenience caused to some patients) and false negatives (and thus the increased rate of mortality and additional complications).

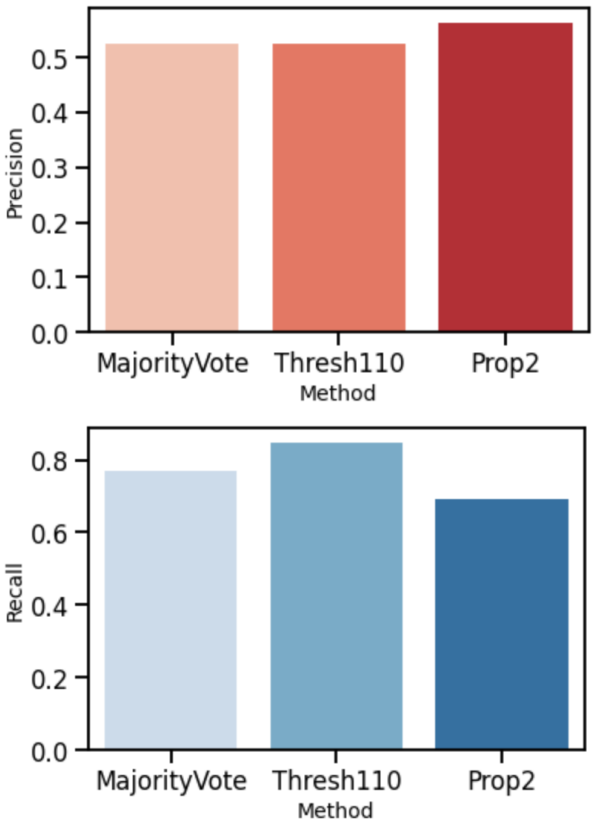

Figure 1: Comparison of precision and recall for different approaches.

Combining Local and Global Information

Lastly, we turned our attention proposed in the previous section and which attempts to ameliorate the loss of information effected by breaking up a WSI into patches, by combining aggregated patch based learning and learning done on the level of the whole slide, down-sampled for computational reasons. Considering the findings we already discussed and the superiority of the absolute thresholding method for the aggregation of patch level predictions, this approach was adopted for the patch based learning part of the method (with the previously determined optimum free parameter value, that is the threshold, of $N = 110$). Promisingly, the performance was indeed far superior to what was achieved using patches only. In particular, while the precision remained unchanged, the recall was improved dramatically; quite in fact, recall was found to be perfect, i.e. equal to 1.0.

Summary and conclusions

In digital pathology, the state of the art methods for whole slide image analysis approach the problem by breaking up a slide into a large number of constituent sub-images, patches, performing deep learning on the said patches, and then integrating patch level predictions into a slide level one. Hence, the first goal of this paper was to examine whether the commonly used approach for patch level to slide level inference, that is majority vote, can be improved upon. In particular, we proposed two alternatives in the form of absolute and relative thresholding of patch level decisions and on a real-world corpus of bladder whole slide images demonstrated that a tuned absolute thresholding approach indeed achieves superior results. All three methods, the original majority vote based one and the two newly proposed herein, attained similar precision scores, but significant differences were observed in recall, with absolute thresholding outperforming others by a significant margin.

Furthermore, motivated by the most significant limitation of the aforementioned patch to slide level analysis, namely the loss of geometric relationship between different patches, we proposed a new fusion based approach which combines aggregated patch level predictions (local information, lacking in global awareness) with the prediction made directly on the whole slide (global information, lacking in detailed information), down-sampled for the sake of computational tractability. This approach was shown to be successful too, improving performance yet further and indeed achieving perfect recall.

Our results should have both effects on the immediate practice in machine learning in digital pathology as well as on future work. As regards the former, the presented experiments suggest that researchers should adopt, or at the very least investigate the use of, more sophisticated methods for making slide level predictions from patches. For future work, the promising results of our approach for the utilization of both local information, in the form of patches, and global information, in the form of down-sampled slides, calls for more research effort into how local and global information may be combined. Our immediate plan is to apply more sophisticated decision level fusion approaches (Arandjelovic 2016).

References

Arandjelovic, O. 2016. Weighted linear fusion of multimodal data: a reasonable baseline? In *Proceedings of the ACM international conference on Multimedia*, 851–857.

Cooper, J., and Arandjelović, O. 2020. Learning to describe: A new approach to computer vision based ancient coin analysis. *Sci* 2(2):27.

Dimitriou, N.; Arandjelović, O.; and Caie, P. D. 2019a. Deep learning for whole slide image analysis: an overview. *Frontiers in medicine* 6:264.

Dimitriou, N.; Arandjelovic, O.; and Caie, P. D. 2019b. Deep Learning for Whole Slide Image Analysis: An Overview. *Frontiers in Medicine* 6.

Garapati, S. S.; Hadjiiski, L.; Cha, K. H.; Chan, H.-P.;

Caoili, E. M.; Cohan, R. H.; Weizer, A.; Alva, A.; Paramagul, C.; and Wei, J. e. a. 2017. Urinary Bladder Cancer Staging in CT Urography Using Machine Learning. *Medical Physics* 44(11):5814–5823.

Ge, L.; Chen, Y.; Yan, C.; Zhao, P.; Zhang, P.; A, R.; and Liu, J. 2019. Study Progress of Radiomics With Machine Learning for Precision Medicine in Bladder Cancer Management. *Frontiers in Oncology* 9:1296.

Kennelly, M. J.; Smith, A. M.; Meeks, J. J.; and Quale, D. Z. 2017. *Muscle Invasive Bladder Cancer Patient Guide*. Urology Care Foundation, 1000 Corporate Boulevard, Linthicum, MD 21090, USA, 1 edition.

Kirkali, Z.; Chan, T.; Manoharan, M.; Algaba, F.; Busch, C.; Cheng, L.; Kiemeney, L.; Kriegmair, M.; Montironi, R.; and Murphy, W. M. e. a. 2005. Bladder cancer: Epidemiology, Staging and Grading, and Diagnosis. *Urology* 66(6, Supplement 1):4–34. International Consultation on Bladder Tumors.

Kouznetsova, V. L.; Kim, E.; Romm, E. L.; Zhu, A.; and Tsigelny, I. F. 2019. Recognition of Early and Late Stages of Bladder Cancer Using Metabolites and Machine Learning. *Metabolomics* 15(7).

Tun, W.; Arandjelovic, O.; and Caie, P. D. 2018. Using machine learning and urine cytology for bladder cancer prescreening and patient stratification. In *AAAI Workshops*.

Witjes, J. A.; Compérat, E.; Cowan, N. C.; De Santis, M.; Gakis, G.; Lebret, T.; Ribal, M. J.; Van der Heijden, A. G.; and Sherif, A. 2014. EAU Guidelines on Muscle-invasive and Metastatic Bladder Cancer: Summary of the 2013 Guidelines. *European Urology* 65(4):778–792.

Wu, E.; Hadjiiski, L. M.; Samala, R. K.; Chan, H. P.; Cha, K. H.; Richter, C.; Cohan, R. H.; Caoili, E. M.; Paramagul, C.; Alva, A.; and Weizer, A. Z. 2019. Deep Learning Approach for Assessment of Bladder Cancer Treatment Response. *Tomography* 5(1):201–208.

Yue, X.; Dimitriou, N.; and Arandjelovic, O. 2019. Colorectal cancer outcome prediction from h&e whole slide images using machine learning and automatically inferred phenotype profiles. *arXiv preprint arXiv:1902.03582*.