Deep Learning based Endoscopy Video Classification Intermediary Report

Oussama Fadil
Stanford University
Stanford, CA
fadil@stanford.edu

Ismael Lemhadri
Stanford University
Stanford, CA
lemhadri@stanford.edu

Crystal Zheng
Stanford University
Stanford, CA
czheng11@stanford.edu

Abstract

1. Introduction

Bladder cancer is the most common urinary tract disease, and early detection is vital in reducing the risk of cancer recurrence and progression. Although white light cystoscopy is currently the primary strategy for bladder cancer management, it has various limitations regarding the visualization and detection of cancer tumors. Recent advancements in endoscopic imaging has made it a promising alternative [4] in addressing these shortcomings. Furthermore, artificial intelligence has been successfully used in developing better computer-aided detection and diagnostic tools for gastrointestinal cancer from real-time endoscopic videos [1]. Since endoscopy has not been commonly used for bladder cancer particularly, there is a need to develop algorithms to improve detection and characterization of cancer in bladder endoscopic videos. Our goal is to develop a classification model between "cancer" and "non-cancer" patients that can accurately identify the most informative image frames from bladder endoscopy videos.

To approach this problem, our overall plan was first to examine the study by Lucas et al. [5] who developed a computer-aided classification of endomicroscopic videos of bladder lesions through a feature extractor combined with an LSTM network. We reviewed technical papers to help inform us on how to best develop a suitable model, including Data Shapley [3] and weakly supervised methods [2]. We collaborated with Liangqiong Qu, a Stanford researcher who provided mentoring and guidance in addition to a labelled endoscopy data.

After project setup and initial experiments, our strategy involves a supervised weighted model to deal with the presence of uninformative image frames while simultaneously learning the most accurate features for classification task. The project will undergo two-fold evaluation. First, a small sample of selected frames will be provided to physicians in training at Stanford Hospital, to obtain expert feedback on the frame's relevance to the classification task. Second, a dataset including only the selected relevant frames will be used to train a new classification model, and performance evaluated according to:

- 1) the reduction in input size achieved, while maintaining similar prediction accuracy;
- 2) the subsequent reduction in training time;
- 3) the agreement between this method's selected frames and competing methods [4].

We further elaborate on our dataset, technical approach, and results and evaluation in the statement section.

2. Problem Statement

Modern urologic endoscopy is the result of continuous innovations since early 19th century. Landmark innovations over the last two centuries have shaped modern urologic endoscopy and particularly endoscopic management of bladder tumors. Based on the well-established principle of fluorescence confocal microscopy, CLE is an optical biopsy technology that enables in vivo high resolution, subsurface imaging. It is of great current interest to devise an automated method for bladder cancer diagnostic taking as input a patient recording from confocal endoscopy. A significant challenge is that many of the frames from the said recording are uninformative. Many frames are unrelated to the bladder tissue at hand, because the camera must pass through multiple other stages before reaching the bladder; or, they might focus on the bladder but on a non-cancerous region.

Our problem statements are therefore:

1) Can we use a weakly supervised model to detect uninformative image frames?

Using the selected frames to train a new classification model, can our new model:

2) With the subsequent reduction in input size, maintain similar prediction accuracy?

3) reduce training time?

4) have an agreement between this model's selected frames and the performance of competing methods?

2.1. Dataset

From our collaboration with Liangqiong Qu, we obtained a dataset containing 129 confocal endoscopy videos, which was securely transferred since it contained protected health information. This data only had video-level labels, _normal denoting normal and _LG denoting cancerous, but not image-level labels.

The training dataset contained 44,983 confocal endoscopic image frames from 129 different videos (11.94 GB of data). There were a total of 31281 image frames in 92 videos labelled normal (8.21GB) and 13702 image frames from 37 videos labelled cancer (3.73GB).

For the validation set, we received 11,329 image frames from 30 videos, 24 normal and 3293 image frames were from 6 cancerous (video-level labels) videos. Out of them, 576 image frames were labelled as 0-normal, 1- cancerous.

2.2. Pre-processing

We used a script to perform pre-processing on confocal endoscopy frame images. We first did [0,1] normalization of the images. In other words, we subtracted the image's minimum intensity from the original image. Then, divided by the difference between the image's max intensity and image's min intensity. The formula is:

$$normalized = \frac{image - min(image)}{max(image) - min(image)}$$

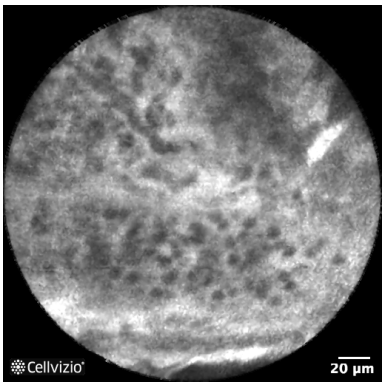
Then, we downsampled the normalized pixel array by a downsample factor of 2. Files were saved in either a folder named "training," which we will later split into train and test, or a folder names "validation."

The training and validation sets remained the same size as preprocessing.

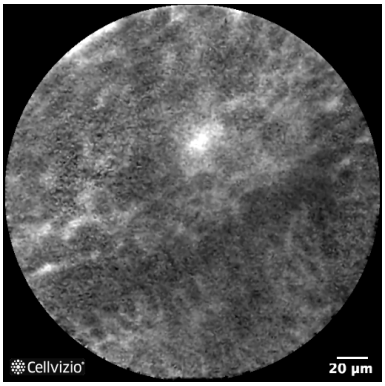
Afterwards, for development purposes, we randomly sampled the data to select 5 videos, 3 normal and 2 cancer (video-level labelled). For this development set of images, we had 69 image frames from the normal patients and 56 image frames from the cancer patients, for a total of 125 image frames. We also sampled for an even smaller dataset, with 27 image frames from 1 normal patient and 22 image frames from 1 cancer patient (video-level labelled). These sets of images allowed us to quickly prototype and debug our models/code before training on GCP.

3. Technical Approach

A patient endoscopy video is represented as a collection of m frames $X = x_1, ..., x_m$, where each frame x_i is a 224×224 image with 1 greyscale channel. Each frame is modeled as an labeled data point x_i with label $Y \in [-1, 1]$ corresponding to the patient's diagnosed condition. Therefore the labels are identical across all frames for a given patient.



(a) Normal (video-level label) patient image frame



(b) Cancer (video-level label) patient image frame

Figure 1: Sample pre and post-processed confocal endoscopic image frames from the Stanford Hospital provided dataset. On this simple example, the difficulty of the task becomes apparent: it is unclear which image frames truly discriminate cancer from normal patients; in addition, the signal-to-noise is very low and the task is fundamentally difficult. Even expert physicians may disagree on the diagnosis, and ground truth labels may require post-endoscopy biopsy to confirm the result, an extensive and heavy medical operation.

3.1. Methods

We propose to study the following two methods:

- A pre-trained ResNet18 model fine-tuned on our dataset;
- A frame-importance model that simultaneously per-

forms classification and frame importance.

Baseline: fine-tuned ResNet18

The first method shall not be considered a simple baseline; in fact, it is a standalone classification method that can be used for cancer classification. It will also be embedded as a building block in the second method, which we describe in more detail below.

Main method: CNN-LSTM

The method involves a discriminative model. The final output is a classification into normal/cancer status. The model uses t contiguous MRI frames ($5 \leq t \leq 30$, where t is a hyperparameter) and a single label per patient.

Conceptually, the model is inspired from [2], who originally applied it to MRI sequence data, and consists of two parts: a frame encoder and a sequence encoder.

- The frame encoder learns frame-level features and a sequence encoder for combining individual frames into a single feature vector. For the frame encoder, we use a Residual Network (ResNet18) pretrained on ImageNet. We also plan to test other common pretrained image neural networks such VGG16 and ResNet50 (with the aim of capturing the best possible features, although conceptually all networks have a similar role). The ResNet architecture takes advantage of low-level feature maps at all layers, making it well-suited for medical imaging applications where low-level features (e.g., edge detectors) often carry substantial explanatory power.

- The sequence encoder consists of a Bidirectional Long Short-term Memory (LSTM) sequence model with soft attention to combine all endoscopy frame features. The soft attention layer optimizes the weighted mean of frame features, allowing the network to automatically give more weight to the most informative frames in an MRI sequence.

The CNN-LSTM model is trained using noise-aware binary cross-entropy loss: the standard bce loss is modified into a *weighted* loss, where the weights are learned through usual gradient descent so as to satisfy

$$\hat{w} = \text{argmin} \frac{1}{N} \sum_{i=1}^{N} L(w, x_i, y_i).$$

thereby providing an indication of frame informativeness. Therefore, training frames with higher weights have more impact on the learned model and may be considered more valuable.

In summary, each MRI frame is encoded by the ResNet into a feature vector. These frame features are fed sequentially to the LSTM sequence encoder, which uses a soft-attention layer to learn a weighted mean embedding of all frames. This forms the final feature vector used for binary classification.

Remark: The sequence encoder could be simplified into a mean or a max pooling architecture instead of an LSTM. We preferred using the latter in order to align with recent methods for state-of-the-art video classification and medical imaging. It would be interesting to explore these simpler architectures as additional baselines.

4. Preliminary Results

Model	Blind guess	ResNet18	CNN-LSTM
Test accuracy	50%	67.5%	TBD

Due to space constraints, we only report the test accuracy metric, leaving a more comprehensive report to the final stage, as described in the Future Steps section.

5. Future Steps

Our final report will include a much deeper analysis of the model, comprising:

- a visualization of the features produced by the frame encoder;

- simpler sequence encoder architectures as additional baselines;

- a grid search trainer for hyperparameter tuning based on the validation set;

- an effective early stopping metric monitored on the validation set.

It would also be interesting to consider the following simple baseline. By flipping the frames's labels between "normal" and "cancer", and monitoring the impact on model performance, we can obtain an indication of frame importance. We intuitively expect that a large number of frames will be found deemed unimportant, just like a human labeller would.

References

[1] Muthuraman Alagappan, Jeremy R Glissen Brown, Yuichi Mori, and Tyler M Berzin. Artificial intelligence in gastrointestinal endoscopy: The future is almost here. *World journal of gastrointestinal endoscopy*, 10(10):239, 2018.

[2] Jason A Fries, Paroma Varma, Vincent S Chen, Ke Xiao, Heliodoro Tejeda, Priyanka Saha, Jared Dunnmon, Henry Chubb, Shiraz Maskatia, Madalina Fiterau, et al. Weakly

supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. *Nature communications*, 10(1):1–10, 2019.

[3] Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. *arXiv preprint arXiv:1904.02868*, 2019.

[4] Aristeo Lopez and Joseph C Liao. Emerging endoscopic imaging technologies for bladder cancer detection. *Current urology reports*, 15(5):406, 2014.

[5] Marit Lucas, Esmee IML Liem, C Dilara Savci-Heijink, Jan Erik Freund, Henk A Marquering, Ton G van Leeuwen, and Daniel M de Bruin. Toward automated in vivo bladder tumor stratification using confocal laser endomicroscopy. *Journal of endourology*, 33(11):930–937, 2019.