Automatic Pathology Detection

Ahmed Mohammed

Department of Computer Science
Norwegian University of Science and Technology
Sintef Digital
Motivation

• Colorectal cancer (CRC) is a major cause of mortality and morbidity worldwide.
• Patients with long standing ulcerative colitis are at greater risk of developing colon or rectal cancer than the general population
Motivation
Capsule Video Endoscopy (CVE): Characteristics

- Size 31x11 mm
- Images produced (~50,000)
- Requires 2 hours to review
Capsule Video Endoscopy (CVE): Characteristics

**IDEALLY**
- Vessels should be visible
- Specular reflections should be removed
- Tissues should be visible
- Color shouldn’t be distorted. (If the colors are changed, it requires retraining physician to view such images)
IDEALLY

• The phi-effect should be obtained to create apparent living pictures with natural motion portrayal.
• Flickering should be avoided when displaying image sequences.
• Save battery
Capsule Video Endoscopy (CVE): Characteristics

- Sessile
- Pedunculated
- Neoplastic vs Hyperplastic

Appearance, orientation and shape variability,

Different stages of UC and their appearances
Unstructured ➔ Structured data

Challenge:
- Medical data annotation is expensive and slow
- Privacy issues and limited access
- Diverse pathologies and long video sequence

How to deal with lack of data
- Transfer Learning, Domain Adaptation
- Weakly supervised learning
- Self-supervision
- Semi- or Unsupervised learning
Approach

- Hand-crafted features approach:
  - Shape and appearance modeling: Intensity valley detection
  - Gives high FP and low TP

- Deep learning approach
  - Hybrid: CNN + RGB + geometric features
  - CNN with RGB input, transfer learning and training from scratch
  - 3D CNN - Online and offline
Proposed: Y-Net

• Exploit the advantage of pre-trained weights
• Combines pre-trained weights with random weights
• Address the performance loss due to domain-shift
Proposed: Ensemble of encoders:
Proposed: Y-Net
Proposed: Y-Net
Proposed Method: Y-Net
Proposed Method: Y-Net
Training

$$\theta_{t+1} = \theta_t - \frac{c \cdot \eta}{\sqrt{E[g^2]} + \epsilon} \cdot g_t$$

$$\mathcal{L}(p, g) = -\frac{1}{N} \sum_{i=1}^{N} \left( \frac{\lambda}{2} \cdot g_i \cdot \log p_i \right) + \left( 1 - \frac{2 \sum_{i=1}^{N} (g_i \cdot p_i) + \epsilon}{\sum_{i=1}^{N} (p_i) + \sum_{i=1}^{N} (g_i) + \epsilon} \right)$$
Dataset:

• ASU-Mayo clinic polyp database
• 20 and 18 short segment colonoscopy videos for training and testing with pixel level annotated polyp masks
• 4278/18495 frames with polyps in the training set and 4300/17574 frames in test set.
Evaluation metrics:

- F1 and F2 Metrics

\[ P = \frac{N_{tp}}{N_{tp} + N_{fp}}, R = \frac{N_{tp}}{N_{tp} + N_{fn}} \]

\[ F1 = \frac{2PR}{P + R}, F2 = \frac{5PR}{4P + R} \]
Ablation study

<table>
<thead>
<tr>
<th>Method</th>
<th>Prec[%]</th>
<th>Rec[%]</th>
<th>F1[%]</th>
<th>F2[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net (Trained from scratch, baseline)</td>
<td>90.8</td>
<td>39.2</td>
<td>54.7</td>
<td>44.2</td>
</tr>
<tr>
<td>U-Net (Pre-trained encoder VGG19, single encoder)</td>
<td><strong>96.2</strong></td>
<td>68.2</td>
<td>79.8</td>
<td>72.4</td>
</tr>
<tr>
<td>Y-Net (Ours)</td>
<td>87.4</td>
<td><strong>84.4</strong></td>
<td><strong>85.9</strong></td>
<td><strong>85.0</strong></td>
</tr>
</tbody>
</table>
Result: Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>Prec[%]</th>
<th>Rec[%]</th>
<th>F1[%]</th>
<th>F2[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>1594</td>
<td>10103</td>
<td>2719</td>
<td>13.6</td>
<td>36.9</td>
<td>19.9</td>
<td>27.5</td>
</tr>
<tr>
<td>CVC-CLINIC</td>
<td>1578</td>
<td>3456</td>
<td>2735</td>
<td>31.3</td>
<td>36.6</td>
<td>33.8</td>
<td>35.4</td>
</tr>
<tr>
<td>OUS</td>
<td>2222</td>
<td>229</td>
<td>2091</td>
<td>90.6</td>
<td>51.5</td>
<td>65.7</td>
<td>56.4</td>
</tr>
<tr>
<td>ASU [2]</td>
<td>2636</td>
<td>184</td>
<td>1677</td>
<td>93.5</td>
<td>61.1</td>
<td>73.9</td>
<td>65.7</td>
</tr>
<tr>
<td>CUMED</td>
<td>3081</td>
<td>769</td>
<td>1232</td>
<td>80.0</td>
<td>71.4</td>
<td>75.5</td>
<td>73.0</td>
</tr>
<tr>
<td>Fusion [3]</td>
<td>3062</td>
<td>414</td>
<td>1251</td>
<td>88.1</td>
<td>71.0</td>
<td>78.6</td>
<td>73.9</td>
</tr>
<tr>
<td>Y-Net(Ours)</td>
<td>3582</td>
<td>513</td>
<td>662</td>
<td>87.4</td>
<td>84.4</td>
<td>85.9</td>
<td>85.0</td>
</tr>
</tbody>
</table>

1. Bernal et al., “WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians”.
2. Tajbakhsh, Gurudu, and Liang, “Automated polyp detection in colonoscopy videos using shape and context information”.
3. Yu et al., “Integrating online and offline three-dimensional deep learning for automated polyp detection in colonoscopy videos”.
Result:

PS-DeVCEM: Pathology-sensitive deep learning model for video capsule endoscopy based on weakly labeled data.

- Limited amount of annotated samples
- \( \approx 50,000 \) images
- Manually labeling images for pathologies simply does not scale well and is expert-intensive.
MIL: Multiple instance learning

• A type of weakly supervised learning problem where only group-level, also known as bag level annotation, is available. The instances within the bag are not labeled.
  
  E.g: the annotation could be a general statement about the category of the pathology in the video without information about the location within the video or frame labels.

• Independent samples (images):
  
  • E.g. Single-instance learning (SIL): assigns each instance the label of its bag, creating a supervised learning problem, but mislabeling negative instances in positive bags

• Temporal based (video) MIL:
  
  • E.g. The group-level prediction is given by taking the average of the instances. An objective function is introduced to encourage smoothness of inferred instance-level labels based on instance-level similarity, while at the same time respecting group-level label constraints.

Assumptions

• It is assumed that positive bag videos contain at least one instance of a given pathology while a negative bag video depicts none.

Let $V = \{f_1, f_2, f_3, \ldots, f_N\}$ be a video containing frames $f_1, f_2, f_3, \ldots, f_N$ and $N$ is the number of frames in the video. We assume individual labels are available for each video $V$ and is given by $G$ with unknown frame label $y = \{y_1, y_2, y_3, \ldots, y_N\}$. 
MIL constraints

\[ Y = \begin{cases} 
  p & \text{if } \exists n \text{ s.t. } y_n = p, p \subseteq P, n \in N \\
  0, & \text{otherwise}
\end{cases} \]

Alternative MIL constraint

\[ Y = \max_n \{ y_n \} \mid y_n = p, p \subseteq P, n \in N \]
PS-DeVCEM
PS-DeVCEM
Attention

- Provides insight into the contribution of each instance to the bag label.

\[ Z = \sum_{n=1}^{N} \alpha_n h_n \]

\[ \alpha_n = \frac{\exp\{w^T \tanh(Vh_n^T)\}}{\sum_{i=1}^{N} \exp\{w^T \tanh(Vh_i^T)\}} \]
Training

\[ Z_{bag}^+ = \sum_{b=1}^{B^+} h_b^+ | h_b^+ \subseteq h, \alpha_b > \frac{1}{N} \]

\[ Z_{bag}^- = \sum_{b=1}^{B^-} h_b^- | h_b^- \subseteq h, \alpha_b < \frac{1}{N} \]

\[ L = \frac{1}{M} \sum_{m=1}^{M} g_n \log(y_n)^+ \]

\[-\frac{\lambda}{M} \sum_{m=1}^{M} g_b \log(y_b) - (1 - g_b) \log(1 - y_b) \]
Ablation study

Train: Each convolution feature is weighted with computed value $\alpha_n$ before feeding into the LSTM network.

Input: Frame feature
Attend: Frame feature

Input: Frame feature
Attend: Hidden state of LSTM block

Input: Hidden state of LSTM block
Attend: Hidden state of LSTM block

Input: Hidden state of LSTM block
Attend: Hidden state of LSTM block + self-supervision

Input: Residual LSTM block
Attend: Residual LSTM block + self-supervision
Ablation study

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>AttenConv</td>
<td>0.229</td>
<td>0.290</td>
<td>0.246</td>
<td>0.872</td>
</tr>
<tr>
<td>AttenConvLSTM</td>
<td>0.450</td>
<td>0.461</td>
<td>0.443</td>
<td>0.939</td>
</tr>
<tr>
<td>AttenLSTM</td>
<td>0.529</td>
<td>0.478</td>
<td>0.487</td>
<td><strong>0.954</strong></td>
</tr>
<tr>
<td>GuidedLSTM</td>
<td>0.487</td>
<td>0.482</td>
<td>0.458</td>
<td>0.946</td>
</tr>
<tr>
<td>EndoscopicMIL(proposed)</td>
<td><strong>0.616</strong></td>
<td><strong>0.546</strong></td>
<td><strong>0.551</strong></td>
<td>0.951</td>
</tr>
</tbody>
</table>

Ablation study result: The values are averaged for all pathologies.
### Result

Comparison with other MIL algorithms: The values are averaged for all pathologies.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIL [1]</td>
<td>0.235</td>
<td>0.046</td>
<td>0.066</td>
<td>0.997</td>
</tr>
<tr>
<td>MissSVM [2]</td>
<td>0.130</td>
<td>0.162</td>
<td>0.123</td>
<td>0.912</td>
</tr>
<tr>
<td>Attention based deep MIL [3]</td>
<td>0.616</td>
<td>0.471</td>
<td>0.513</td>
<td>0.955</td>
</tr>
<tr>
<td>STPN [4]</td>
<td>0.592</td>
<td>0.517</td>
<td>0.536</td>
<td>0.916</td>
</tr>
<tr>
<td>W-TALC [5]</td>
<td>0.274</td>
<td>0.891</td>
<td>0.416</td>
<td>0.666</td>
</tr>
<tr>
<td>EndoscopicMIL (w/o self-supervision)</td>
<td>0.606</td>
<td>0.54</td>
<td>0.54</td>
<td>0.951</td>
</tr>
<tr>
<td>EndoscopicMIL (proposed)</td>
<td><strong>0.616</strong></td>
<td>0.546</td>
<td><strong>0.551</strong></td>
<td>0.951</td>
</tr>
</tbody>
</table>

Visual result

Visual result

Self-supervision:

• Labels for free and train supervised
• Common in NLP

Self-supervision
Conclusions and Future Work

• Improve the video frame localization through domain knowledge of the pathologies
• Weakly supervised video segmentation
• Novel self-supervision pre-text task/contrastive learning
• Explainability
• Advance to the next level and test these methods in real procedures.
Acknowledgement:

• Norwegian Colour and Visual Computing Laboratory
• “IQ-MED: Image Quality enhancement in MEDical diagnosis, monitoring and treatment, project no. 247689” and “CAPSULEAI3D Improved Pathology Detection in Wireless Capsule Endoscopy Images through Artificial Intelligence and 3D Reconstruction, project no. 300031"