Introduction

1 in 8 women are affected by breast cancer and over 40,000 women are expected to die from it in 2018 alone [1]. Early detection using mammograms is very important as it can open up treatment options and improve survival rate to 93% [2]. State-of-the-art CNNs have been used previously in breast cancer detection. We aim to extend previous work by applying transfer learning and multiple-instance learning techniques to CNNs.

Data

**DDSM Dataset [3]**
- 2555 cases, 4 mammogram images each, different views of same patient
- Labeled as normal (688 cases), benign (814), benign without callback (140), cancer (913 cases)
- Images came as Grayscale in various sizes up to 7000x7000
- 255 cases (10%) held out for validation

Data Preprocessing
- Reshaped to 299x299 grayscale for stacked MIL, RGB otherwise
- Normalized each image
- Data augmentation: rotation, flip, and zoom on each view

Models and Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shallow CNN</td>
<td>16.4%</td>
<td>37.5%</td>
<td>21.9%</td>
</tr>
<tr>
<td>Transfer (Inception)</td>
<td>24.9%</td>
<td>50.3%</td>
<td>31.4%</td>
</tr>
<tr>
<td>Transfer (Inception trained)</td>
<td>23.8%</td>
<td>55.1%</td>
<td>32.1%</td>
</tr>
</tbody>
</table>

Multiple-Instance Learning (MIL)

- Combine the views in a case to try and get a better prediction
- Can stitch images together in 2x2 grid or stack each view as a channel
- Can also send 1-by-1 and take a vote with max or mean of output or feed through NN

Discussion

- Our models are unable to balance recall and precision. The high recall suggests that we catch most of the cancer, but low precision indicates that we predict cancer too often for normal or benign (high false positive rate).
- As expected, transfer learning shows improvement compared to our baseline shallow network.
- The combined MIL + transfer learning model did not perform the best. This was unexpected because we thought having both more info per input and the low-level feature extraction of Inception would lead to better performance. However, we need to train it for more epochs or change the FC architecture to confirm this.
- Low precision could be due to class imbalance. We tried to use an error rate multiplier to weight the loss of a true positive and noticed some improvements. We could similarly weight the loss of false positives.

Future Work

- Immediate: train each model for longer
- Immediate: standardize evaluations between models that use MIL and do not use MIL to give a better idea of how much MIL helps
- Visualize what the network focuses on when predicting
- Use attention or object segmentation to focus on breast/tumor portion of image

References