Visual Interpretability of Capsule Network for Medical Image analysis

Mighty Abra Ayidzoe¹,², Yongbin Yu¹, Patrick Kwabena Mensah ², Jingye Cai¹, Faiza Umar Bawah²
¹School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu, P.R. China
²Department of Computer Science and Informatics, University of Energy and Natural Resources, Sunyani, Ghana

Abstract: Deep learning (DL) models are currently not widely deployed for critical tasks such as in health. This is attributable to the "black box," making it difficult to gain the trust of practitioners. This paper proposes the use of visualizations to enhance performance verification, improve monitoring, ensure understandability, and improve interpretability needed to gain practitioners’ confidence. These are demonstrated through the development of a CapsNet model for the recognition of gastrointestinal tract infection. The gastrointestinal tract comprises several organs joined in a long tube from the mouth to the anus. It is susceptible to diseases that are difficult for medics to diagnose since it is not easy to have physical access to the sick regions. Consequently, manual access and analysis of images of the unhealthy parts requires the skills of an expert, as it is tedious, prone to errors, and costly. Experimental results show that visualizations in the form of post-hoc interpretability can demonstrate the reliability and interpretability of the CapsNet model applied to gastrointestinal tract diseases. The outputs can also be explained to gain practitioners’ confidence in paving the way for its adoption in critical areas of society.

Key words: Capsule Network, Explainable Artificial Intelligence (XAI), Convolutional Neural Network

1. Introduction

Deep learning has proven efficient and indispensable in several computer vision tasks such as speech recognition, image classification, textual analysis, and many more. However, its adoption in sensitive areas such as health, banking, etc., is yet to reach its full potential due to challenges resulting from the “black box” concept [1]. The “black box” makes it extremely difficult for data practitioners and industry players to build the trust and confidence needed to practically deploy DL models for such critical tasks. Health is a sensitive field with lots of uncertainties. Dataset imbalance, unavailability of large datasets, issues of privacy, and regulatory policies [2] are additional problems that make the field challenging for DL. Therefore, the adoption of DL models is a prudent decision when many uncertainties exist, especially if the model outputs cannot be trusted partly because the operation of the model is not understandable. Consequently, explainability and interpretability of the outcomes of DL algorithms applied in health is a pressing need for improved confidence in model results. These may not be strict requirements in other domains such as online translation services, computer electronics, and some business sectors, as long as the models perform up to expectation [3].

The sole aim of understandability and explainability is to enable humans to understand every decision made by the model to establish trust and transparency. Understanding another individual’s decision is a key factor in human intelligence; understanding the outputs of a DL model will be of great benefit to practitioners and will serve as a motivation to perform rigorous performance evaluation to determine their suitability for

*Correspondence: ybyu@uestc.edu.cn
adoption in critical applications. This will prepare the grounds for (i) system performance verification and consequent improvement, (ii) humans to learn from the system, (iii) making legislation compliance check easier [4] and (iv) the developer to understand their system for improved debugging. Due to its potential to achieve the advantages mentioned above, model interpretability has recently received a great boost of interest, especially the development of methodologies to make DL models explainable. Some of the methods are visual, example-based, or text-based [5]. Some visual methods in the literature involve visualization of the feature vectors using nearest neighbors [7] and dimensionality reduction techniques. Other methods analyze the contribution of each pixel using Sensitivity Analysis (SA) [9], layer-wise relevance propagation (LRP) [10], gradient-weight class activation mapping and local interpretable model-agnostic explanation (LIME) approach [11]; with LRP and SA being the most popular. However, SA suffers from gradient shattering, and explanation discontinuities [3] whereas LRP is unsatisfying when applied on nested or complex architectures [13].

Capsule Networks (CapsNet) [14] were proposed to overcome the limitations of Convolutional Neural Network such as invariance. CapsNets have proven to be effective and highly appropriate for medical image classification [15] [16] [17]. It can model the spatial variability of an object in an image and can infer pose parameters from the image [17]. This is an essential requirement of any medical diagnosis algorithm since sick parts are located in varied image regions. In addition, CapsNets have a decoder layer that acts as a regularizer and utilizes the learned features to reconstruct the input images. The advantage is that CapsNets rarely overfit on smaller datasets (like those found in medical images) [30] as well as producing visual outputs in the form of reconstructed images that can be used for comparison with the inputs. However, CapsNet’s mode of operation is slightly different from traditional neural networks or convolutional neural networks (CNNs), necessitating the need for additional methods of explainability for this black-box model.

This paper, therefore, explores existing visualization methods that can enhance CapsNet interpretability and goes further to propose the use of other post-hoc visualization analyses aimed at explaining the inner workings and outputs of dynamic routing-based CapsNets. The paper adopts visualizations taking cognizance of the ease with which both experts and novices will understand the reasons behind model decisions as interpreted by visual inspection of deep learning model outputs can shed light on the activities of the neuron [19][20]. Experimental results demonstrate that CapsNets are explainable [13] and their results interpretable to obtain the confidence needed for practical adoption in medical diagnosis.

This paper is organized as follows; section 2 presents a brief review of capsules and the methods for capsule network explainability. Section 3 details the experimental setup adopted for this study, leading to Section 4, where the results of the experiments are presented and discussed concerning post-hoc visual interpretability. The paper is concluded in section 5.

2. Capsule Network Explainability

Sabour et al [14] proposed an implementation of Capsule Network with dynamic routing showing promise in many application domains such as agriculture [21][22], health [17][15], facial analysis [23], manufacturing [24], sentiment analysis, emotion classification [25], speech recognition [26], meteorological event detection [27], stock prediction [28], internet of things [29] and many more.

A capsule refers to a group of neurons forming a vector of features of an entity or parts of an object. Long vectors signify the probability that an entity exists, with short vectors representing the opposite. The iterating dynamic routing algorithm exists between two capsule layers to transfer outputs to the next layer. In the first
capsule layer, each capsule is transformed using a transformation matrix $W_{ij}$ which aids in the computation of the prediction vectors $u_{j|i}$:

$$\hat{u}_{j|i} = W_{ij}u_i$$ \hspace{1cm} (1)

The prediction vectors vote for capsules in the upper capsule layer. The voting approach enables the connection of similar capsules and vice versa. The total input to a capsule $s_j$ in the upper capsule layer is a weighted sum of all the prediction vectors. This is computed as follows:

$$s_j = \sum_i c_{ij} \hat{u}_{j|i}$$ \hspace{1cm} (2)

where $c_{ij}$ represents coupling coefficients; updated by a routing SoftMax. The coupling coefficients between a capsule in the lower layer and all the capsules in the upper layer sum up to one. The initial inputs to the routing SoftMax are log prior probabilities $b_{ij}$ which are initially set to 0 and are updated by the addition of the previous $b_{ij}$ and the vote.

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}$$ \hspace{1cm} (3)

The length of the vectors in the capsule layers are squashed using a squash function computed as;

$$v_j = \frac{||s_j||^2}{1 + ||s_j||^2} \frac{s_j}{||s_j||}$$ \hspace{1cm} (4)

The vote to each capsule in the upper layer is calculated as follows;

$$a_j = v_j \cdot \hat{u}_{j|i}$$ \hspace{1cm} (5)

The capsule network architecture consists of a decoder layer which is made up of several fully connected layers that are responsible for reconstructing the input images using the learned features. The output of the decoder gives an idea of the features learned by the encoder model.

Some works on the explainability of CapsNets with applications in health exist in the literature. Afshar et al. [30] explained the features of their proposed model using two approaches. Firstly, an investigation of the ability of the model to distinguish between tumor types was carried out. The study involved the visualization of the learned features via T-distributed stochastic neighbor embedding (T-SNE). In the second aspect, the learned features were used to perform reconstruction, activation maximization, and comparison with handcrafted radiomics features. Peer et al. [31] proposed $\gamma$-capsule network and scaled distance agreement routing, which they hypothesized to be more transparent. The $\gamma$-capsule network uses an inductive bias to introduce robustness and withstand adversarial attacks. LaLonde et al. [32] proposed the use of high-level visual attributes in capsule networks to enable the diagnosis of lung cancer at the radiologist level. Shahroudnejad et al. [13] investigated the underlying components and learning mechanisms of CapsNet algorithms. Shahroudnejad et al. [13] argued that CapsNet has inbuilt explainability properties. In explainability and interpretability, the need to attain good accuracy and improve model robustness is not a significant concern. Instead, the objective is to adopt suitable methods and techniques to enhance the understandability of the model’s output. Research in the area of Capsule explainability and interpretability is still open and not fully explored. Further research in this field
will increase the potential of CapsNets for practical adoption in critical areas such as health, banking, and many more. This paper contributes to the field by employing visualization of capsule layer outputs and features as an additional means of enhancing the explainability and understandability required to earn the trust of industry practitioners for adoption.

3. Experimental Details and Dataset Description

All experiments in this study were carried out on a 64-bit windows PC with NVIDIA GeForce GTX 1060 graphic processing unit (GPU) running CUDA 10.1 and an 8GB dedicated memory. The Keras (TensorFlow backend) code at https://github.com/XifengGuo/CapsNet-Keras was adopted and modified for this study. The model was trained using 1, 2, 3, 4 and 5 routing iterations and for 200 epochs each. Other hyperparameters used include the learning rate = 0.001, learning rate decay =0.9 and a batch size=100. The margin loss function in [14] was adopted and is computed as:

\[ L_k = T_k \max(0, m^+ - ||v_k||)^2 + \lambda(1 - T_k)\max(0, ||v_k||, m^-)^2 \]  

where \( T_k = \)

\[ \begin{cases} 
1, & \text{if class } K \text{ is active} \\
0, & \text{otherwise}, \lambda=0.5, \ m^+=0.9, \ m^-=0.1 
\end{cases} \]

The kvasir-v2 [33] dataset used for this study consists of 8,000 RGB images with 8 classes; namely; z-line, pylorus, cecum, esophagitis, polyps, ulcerative colitis, dyed & lifted and dyed & resection. Esophagitis, ulcerative colitis and polyps are the pathological finding class whereas dyed & lifted and dyed & resection are polyp lesions removal class. Z-line, pylorus and cecum are the anatomical landmarks class. The dataset was divided using the ratio 8:2 hold-out approach with images resized to 28 × 28.

3.1. CapsNet Model

Seven convolutional layers were implemented to enable the extraction of the best features, with one primary capsule layer and one class capsule layer (Figure 1).

![Figure 1. The CapsNet architecture used for the experiments.](image-url)
The first, second, and fifth convolutional layers consist of $3 \times 3$, 96 filters. The third and sixth convolutional layers are made up of $3 \times 3$, 48 filters. The fourth and seventh convolutional layers have $3 \times 3$, 64 filters. All the convolutional layers have both the stride and padding set to one with ReLU activation. The Primary capsule layer comprises 16 component capsules, with each vector having a dimension of 8. The dimension of each class capsule is 16. The decoder layer consists of three fully connected layers with 512, 1024, and 2352 neurons.

4. Post-hoc Visual Interpretability of Results

In post-hoc analysis, the outputs of a model are used to understand the intuition behind the model’s decision. Several visualizations were considered. For example, plots of training and test, routing iteration graphs, the cluster of capsules, receiver’s operating characteristics graph (AUC-ROC), precision-recall curve (PR), feature maps, and model attention maps.

To give a deeper understanding of some of the visualizations, such as the confusion matrix, a thorough analysis of the values was carried out to determine the most identified class in the kvasir-v2 dataset.

Training and test accuracy plots are very popular visualizations used to report the performance of deep learning models. Aside from the advantage of reading the accuracy and loss from the graph, it is possible to understand the reasons behind some outputs. For example, during the first few epochs, the proposed model and baseline model are unable to pick the best features from the input images. This can be observed from the fluctuations [21] in test accuracy during training (see Figure 2a).

However, for the proposed model, the accuracy stabilizes as training progresses and the network approaches convergence unlike in Figure 2b where the instability of the baseline model continues until the epoch.
The inability of the models to extract the best features can be attributed to the complexity of the images and the use of random weights during the initial stages. The confusion matrix is a popular tool that is used to provide information on the dataset and performance of a model. From the confusion matrix in Figure 3a, it is observed that the model performs poorly on the classes that are considered to be complex (that is, polyps and esophagitis).

![Confusion Matrix](image)

**Figure 3.** Confusion Matrix for (a) the experimental model and (b) the baseline model.

<table>
<thead>
<tr>
<th>Class</th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
<th>TN</th>
<th>Precision</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Accuracy per class (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dyed &amp; lifted</td>
<td>161</td>
<td>39</td>
<td>40</td>
<td>1366</td>
<td>0.805</td>
<td>0.801</td>
<td>0.972</td>
<td>95.08</td>
</tr>
<tr>
<td>Dyed &amp; resection</td>
<td>160</td>
<td>40</td>
<td>32</td>
<td>1374</td>
<td>0.800</td>
<td>0.833</td>
<td>0.977</td>
<td>95.52</td>
</tr>
<tr>
<td>Esophagitis</td>
<td>142</td>
<td>58</td>
<td>51</td>
<td>1355</td>
<td>0.710</td>
<td>0.736</td>
<td>0.964</td>
<td>93.21</td>
</tr>
<tr>
<td>Cecum</td>
<td>185</td>
<td>16</td>
<td>30</td>
<td>1375</td>
<td>0.920</td>
<td>0.860</td>
<td>0.979</td>
<td>97.14</td>
</tr>
<tr>
<td>Pylorus</td>
<td>197</td>
<td>4</td>
<td>12</td>
<td>1393</td>
<td>0.980</td>
<td>0.943</td>
<td>0.991</td>
<td>99.00</td>
</tr>
<tr>
<td>z-line</td>
<td>149</td>
<td>51</td>
<td>58</td>
<td>1348</td>
<td>0.745</td>
<td>0.720</td>
<td>0.959</td>
<td>93.21</td>
</tr>
<tr>
<td>Polyps</td>
<td>133</td>
<td>67</td>
<td>35</td>
<td>1371</td>
<td>0.665</td>
<td>0.792</td>
<td>0.975</td>
<td>93.65</td>
</tr>
<tr>
<td>Ulcerative colitis</td>
<td>167</td>
<td>37</td>
<td>54</td>
<td>1348</td>
<td>0.819</td>
<td>0.756</td>
<td>0.961</td>
<td>94.33</td>
</tr>
<tr>
<td>Average Accuracy</td>
<td>80.57%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area under curve</td>
<td>81.23%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average Precision</td>
<td>82.07%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10-Fold Mean</td>
<td>81.74%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For instance, the polyps class has many misclassifications showing that it is a challenging class that may require additional preprocessing. The confusion matrix of the baseline model is presented in Figure 3b. Consistent with the proposed model’s observations, the polyps, dyed & lifted, and esophagitis classes have many misclassified images. However, comparing the number of correctly identified polyps, dyed & lifted and esophagitis images of the two models, it could be said that the proposed model has better recognition abilities than the baseline model. This observation explains the need for model modifications. To derive more benefit from a model, the practitioner may consider enhancing the feature extraction abilities of the model. Additionally, from the confusion matrices, metrics such as precision, sensitivity, specificity, and recall can be obtained as shown in Table 1. These are needed to provide further insights into the performance and behavior of the models.
Figure 4. Experimental model: ((a) Receiver’s operating characteristics curves and (c) Precision-Recall curve) and Baseline model: ((b) Receiver’s operating characteristics curve (d) Precision-Recall curve.) Class 0: dyed & lifted, Class 1: dyed & resection, Class 2: esophagitis, Class 3: cecum, Class 4: pylorus, Class 5: z-line, Class 6: polyps and Class 7: Ulcerative Colitis.

of the model relative to the complexity of individual classes in the dataset. Table 2 represents a summary of the 10-fold results of the experimental model.

The ROC and PR curves are famous for evaluating the performance of balanced and imbalanced datasets, respectively. The area under the curve (AUC) is preferred to accuracy as it is not biased in class imbalance. The presence of large areas under the curves (ROC and PR) is an indication that the model is confident about its prediction. Since the presence of class imbalance makes accuracy the wrong choice of metric, a good AUC of the PR curve is preferred, as can be observed in Figure 4(a).

From the ROC and PR curves, practitioners would understand the tradeoff between clinical sensitivity and precision for a mixture of the test (classes) in a graphical manner with any potential cutoff. In a given clinical situation, the AUC is used to indicate the usefulness of the tests and determine the model’s discriminative potential. For example, a practitioner may realize an overlap between classes when the ROC AUC is less than or equal to 0.5. This phenomenon may help the practitioner to obtain an interpretation of the cluelessness that
Table 2. 10-Fold Cross Validation results

<table>
<thead>
<tr>
<th>K-folds</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fold-1</td>
<td>82.86</td>
</tr>
<tr>
<td>Fold-2</td>
<td>80.52</td>
</tr>
<tr>
<td>Fold-3</td>
<td>82.40</td>
</tr>
<tr>
<td>Fold-4</td>
<td>83.72</td>
</tr>
<tr>
<td>Fold-5</td>
<td>81.46</td>
</tr>
<tr>
<td>Fold-6</td>
<td>81.17</td>
</tr>
<tr>
<td>Fold-7</td>
<td>80.92</td>
</tr>
<tr>
<td>Fold-8</td>
<td>82.82</td>
</tr>
<tr>
<td>Fold-9</td>
<td>81.14</td>
</tr>
<tr>
<td>Fold-10</td>
<td>80.36</td>
</tr>
<tr>
<td>10-Fold Mean</td>
<td>81.74%</td>
</tr>
</tbody>
</table>

the model will show. On the other hand, the PR curve allows practitioners to experience the actual effect of class imbalance on the performance of the algorithm. The majority class whose predictions overshadow the other classes and the minority class whose predictions weigh down the model’s performance can be identified, and corrective measures are taken if necessary. Consistent with the data in Table 1, the AUC for the ROC and PR values (Figure 4) of the pylorus class are high. This means that this particular CapsNet model is very good at identifying pylorus followed by cecum and others in that order. Figure 4b and d presents the ROC and PR curves of the baseline model.

4.1. Effect of Routing Iterations

To test the ability of the model to scale without performance degradation, the model is evaluated using several routing iterations. One main objective is to determine the amount of change in accuracy as the network capacity increases from 1 to 5 routing iterations. The preferred result is for the model not to overfit. According to [14], three routing iterations are best for training a CapsNet model; however, in this experiment, four routing iterations (Figure 5) produced the highest accuracy after 200 epochs.

![Figure 5](image_url)  
**Figure 5.** Effect of the number of Routing Iterations on Accuracy. The highest accuracy was obtained with 4 routing iterations.
This can be attributed to the fact that the number of routing iterations is partially dependent on the dataset [21]. As the model is scaled up and the size of the data remains the same, the threat of overfitting increases. As overfitting begins to occur during scale-up, the practitioner’s first point of call is to consider the dataset and decide whether to perform data augmentation or not. This model, therefore, is expected to produce an optimal performance at four routing iterations when used in clinical image analysis, making it easier for practitioners to adopt and trust.

4.2. Ablation Study with the corresponding Clusters, and Feature maps

The two techniques carried out during ablation are the removal of trainable weight from the trained model and pruning of layers. The pruning technique is the popularly used method for the assessments of structurally damaged artificial neural networks. Furthermore, it helps to identify the contribution of each layer to the performance of the classification algorithm. It identifies components that can stand in for damaged parts and contribute significantly to network recovery during failure. In other words, this technique is employed to measure the model’s ability to degrade gracefully. Graceful degradation prevents failed components from grinding the network to a halt. Several experiments were carried out to investigate the robustness of the proposed model. We analyzed the impact of the ablation studies procedure on the clusters form at the class capsule layer. We compared the compactness and separability of the clusters to determine the effect pruning technique.

![Figure 6. Clusters formed in the class capsule layer for each ablation test as seen in Table 2.](image)

From Table 3, we observe that the presence of the conv1 layer is critical to the model’s performance in terms of accuracy. The removal of this layer (SNo. 1 and 10 in Table 3) drastically reduces the model’s overall performance. The next influential layer is the conv 4 layer (SNo. 4 and 12), where the model’s performance
is negatively affected by removing the layer. The corresponding capsule clusters at the class capsule layer for each of the tests in Table 3 can be seen in Figure 6. The separable and compact a cluster is, the higher the performance.

Table 3. Result of ablation studies. “-” before a layer means the removal of that layer.

<table>
<thead>
<tr>
<th>SNo.</th>
<th>Layer</th>
<th>Overall recognition accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>conv1 layer</td>
<td>71.23</td>
</tr>
<tr>
<td>2</td>
<td>conv2 layer</td>
<td>71.21</td>
</tr>
<tr>
<td>3</td>
<td>conv3 layer</td>
<td>76.52</td>
</tr>
<tr>
<td>4</td>
<td>conv4 layer</td>
<td>72.41</td>
</tr>
<tr>
<td>5</td>
<td>conv5 layer</td>
<td>72.73</td>
</tr>
<tr>
<td>6</td>
<td>conv6 layer</td>
<td>74.12</td>
</tr>
<tr>
<td>7</td>
<td>conv7 layer</td>
<td>75.28</td>
</tr>
<tr>
<td>8</td>
<td>conv3 and 4 layers</td>
<td>74.16</td>
</tr>
<tr>
<td>9</td>
<td>conv5 and 7 layers</td>
<td>70.91</td>
</tr>
<tr>
<td>10</td>
<td>conv1 and 6 layers</td>
<td>71.06</td>
</tr>
<tr>
<td>11</td>
<td>conv2 and 4 layers</td>
<td>72.10</td>
</tr>
<tr>
<td>12</td>
<td>conv4 and 7 layers</td>
<td>72.18</td>
</tr>
<tr>
<td>13</td>
<td>conv1 and 5 layers</td>
<td>70.03</td>
</tr>
<tr>
<td>14</td>
<td>conv3 and 6 layers</td>
<td>74.96</td>
</tr>
</tbody>
</table>

4.3. Regions of Interest and Activation maps

Visualization of feature maps can be used to identify layers in the model that gets activated by regions of an input image to influence the model’s final decision-making.

Figure 7. Demonstration of the model’s areas of interest for (a) esophagitis, (b) polyp and (c) ulcerative colitis.

It is a technique that can show whether the model layers tend to focus on the shape or texture of the input image. This is helpful to practitioners in determining the type of applications suitable for the model.
For instance, the model in this study recognizes gastrointestinal diseases based on the color, texture, and shape of the sick area. To demonstrate this idea, three images, one from each of esophagitis, polyps, and ulcerative colitis classes, were used to perform the feature visualizations shown in Figure 7. It is observed that the model’s focus is on the part commonly identified by gastroenterologists as the sick region. Unaffected parts on the walls of the colon are ignored.

Feature visualization has seen heightened research interest [34] due to its potential to make deep learning explainable. In this study, activation maps of network layers are visualized to identify extracted features and regions of interest. To have a clear view of the active pixels, the activation maps are displayed in the form of heatmaps shown in Figure 8 for both the proposed model and baseline model, respectively. Models may have the same classification performance but differ in terms of what features they use to make decisions [4]. Therefore, the identification of the most appropriate model requires the knowledge of the regions based on which a model makes its decisions. For instance, in health applications where wrong predictions can be costly, the model’s reliance on the right features (regions) for prediction must be guaranteed and identifiable by the practitioner. It is observed that the first convolutional layer tries to learn everything in the dataset, unlike...
the second convolutional layer, which learns edges. The third convolutional layer learns shape in addition to edges. Interestingly, the primary capsule layer focuses on the active pixels, and at this stage, the parent-child relationship is built with the capsules in the class capsule layer. From the activation maps, it can be inferred that no single layer can extract good features. It takes the effort of all the layers to extract enough features for effective classification. It points out that, for this model, it is essential to have at least three convolutional layers before the primary capsule layer. This enables the CapsNet model to learn low-level part descriptions.

4.4. Prediction and Reconstruction

Individual predictions and reconstructed images obtained by using the instantiation parameters of the actual class are illustrated in Figure 9. Out of the ten unseen images tested, eight were correctly identified from the pylorus and cecum classes with high prediction probabilities signifying the increased certainty of the model. In Figure 8, four samples of these images are displayed. It can be observed that each class capsule exactly replicates nine of the images, with each object in the image positioned at the exact spatial location as the original image. This proves that CapsNets are good at learning spatial information.

![Figure 9](image.png)

**Figure 9.** Reconstruction and prediction of random images in the kvasir-v2 dataset.

“Explanation” under these circumstances is a list of prediction probabilities for each class after the original image. The likelihood of the prediction is a means by which the model shows its level of confidence in its outputs. Additionally, the reconstructed output image of the decoder is shown to enable the practitioner to qualitatively evaluate the output. Practitioners have prior domain knowledge based on which they can accept or reject the predictions if only the model’s decisions are understandable. The practitioner may use the reconstructed images to verify whether the network is capturing the required visual attributes of the input images. Poor reconstruction can be blamed on poor feature extraction, and the routing process since relevant features from the initial layers of the model are still subject to the routing process. Additionally, the reconstruction results can be used to verify the parts of the image that the network pays attention to as it initially analyses the whole image, after which it gradually focuses on the relevant features of the image.

4.5. Clusters

The network learned features (the instantiation parameters) could be visualized to understand the level to which the model successfully categorized them into the different classes. Primary capsules with similar features form clusters around secondary capsules with which the agreement $a_{ij}$ is high. The separability of the clusters into the respective classes can measure the performance of the routing algorithm. Starting with the raw test set in Figure 10(a), the t-distributed stochastic neighbor embedding (TSNE) [8] converts its high dimensional feature space into a lower one and retains the embedded information. As the dynamic routing successfully performs coupling, visible clusters begin to form at the class capsule layer, as depicted in Figure 10(b). The
compactness and separability of the clusters is a measure of the discriminative power or effectiveness of the routing process. Again, difficult classes can easily be identified to serve as a basis for further action by the practitioner or developer. Figure 10c presents the cluster formed at the class capsule layer of the baseline model. The clusters of the respective classes are not separable signifying the inefficient abilities of the model to distinguish between the different classes.

4.6. Comparison with related works

The main objective of this study is to use visualizations of layer outputs to enhance performance verification, improve monitoring, ensure understandability and improve the interpretability of the inner workings of the model. Notwithstanding, a comparison of the results of our model to related works in the literature is presented in Table 4. These related works include state-of-the-art Capsules and Convolutional neural network models trained with slight variations of the gastrointestinal datasets.

Table 4. Comparison of results to other works.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN [12]</td>
<td>90.96</td>
</tr>
<tr>
<td>VGG16 [36]</td>
<td>96.33</td>
</tr>
<tr>
<td>GaborCapsNet [6]</td>
<td>93.10</td>
</tr>
<tr>
<td>Deep CNN [35]</td>
<td>90.42</td>
</tr>
<tr>
<td>D-caps [32]</td>
<td>82.00</td>
</tr>
<tr>
<td>Our model</td>
<td>80.57</td>
</tr>
</tbody>
</table>

5. Conclusion and future work

This study explored post-hoc visual interpretability with the dynamic routing algorithm to encourage its adoption in medical image analysis and gastrointestinal disease classifications. Initially, the model’s regions of interest were visualized, and the results show that the model focused on sick parts of the images similar to what the gastroenterologists will pay attention to. The effectiveness of the routing algorithm was also visualized via a cluster of features formed at the class capsule layer. Through ablation studies, the most essential layers in terms of performance were identified, serving as an explanation for the different accuracies generated by the model. Prediction and reconstruction of images demonstrated the model’s ability to learn spatial information and engender trust. It is believed that the untapped abilities of CapsNets can be fully explored only if they are interpretable, a requirement to earn the confidence of practitioners in critical sectors such as health.
the future, ad-hoc and ante-hoc interpretability of capsules will be investigated to further champion their acceptability in critical real-life applications.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (NSFC Grant No. 61550110248); Research on Sino-Tibetan multi-source information acquisition, fusion, data mining and its application (Grant No. H04W170186) and Sichuan Science and Technology Program (Grant No. 2019YFG0190).

References


