Computer Methods and Programs in Biomedicine 187 (2020) 105236

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Celiac disease diagnosis from videocapsule endoscopy images with residual learning and deep feature extraction

Xinle Wang[a], Haiyang Qian[a], Edward J. Ciaccio[b], Suzanne K. Lewis[b], Govind Bhagat[b,d], Peter H. Green[b], Shenghao Xu[c], Liang Huang[a], Rongke Gao[a,*], Yu Liu[a,*]

[a] School of Instrument Science and Opto-electronic Engineering, Hefei University of Technology, Hefei 230009, China
[b] Columbia University Medical Center, Department of Medicine – Celiac Disease Center, New York, USA
[c] Shandong Key Laboratory of Biochemical Analysis, College of Chemistry and Molecular Engineering, Qingdao University of Science and Technology, Qingdao 266042, China
[d] Columbia University Medical Center, Department of Pathology and Cell Biology, New York, USA

ARTICLE INFO

Article history:
Received 30 August 2019
Revised 14 November 2019
Accepted 19 November 2019

Keywords:
Block-wise channel squeeze and excitation component
Residual network
Videocapsule endoscopy
Machine learning
Celiac disease

ABSTRACT

Background and Objective: Videocapsule endoscopy (VCE) is a relatively new technique for evaluating the presence of villous atrophy in celiac disease patients. The diagnostic analysis of video frames is currently time-consuming and tedious. Recently, computer-aided diagnosis (CAD) systems have become an attractive research area for diagnosing celiac disease. However, the images captured from VCE are susceptible to alterations in light illumination, rotation direction, and intestinal secretions. Moreover, textural features of the mucosal villi obtained by VCE are difficult to characterize and extract. This work aims to find a novel deep learning feature learning module to assist in the diagnosis of celiac disease.

Methods: In this manuscript, we propose a novel deep learning recalibration module which shows significant gain in diagnosing celiac disease. In this recalibration module, the block-wise recalibration component is newly employed to capture the most salient feature in the local channel feature map. This learning module was embedded into ResNet50, Inception-v3 to diagnose celiac disease using a 10-time 10-fold cross-validation based upon analysis of VCE images. In addition, we employed model weights to extract feature points from training and test samples before the last fully connected layer, and then input to a support vector machine (SVM), k-nearest neighbor (KNN), and linear discriminant analysis (LDA) for differentiating celiac disease images from heathy controls.

Results: Overall, the accuracy, sensitivity and specificity of the 10-time 10-fold cross-validation were 95.94%, 97.20% and 95.63%, respectively.

Conclusions: A novel deep learning recalibration module, with global response and local salient factors is proposed, and it has a high potential for utilizing deep learning networks to diagnose celiac disease using VCE images.

1. Introduction

Celiac disease (CD) is a common immune-based disease that often affects the small intestine, which is also known as gluten-sensitive enteropathy. It is a genetically determined autoimmune disease in which the environmental precipitant, gluten, is known [1,2]. This is a long-term chronic disease, prevalent in about 1% of the worldwide population, and it is not currently curable. Gliadin/gluten is resistant to endopeptidases present in the intestinal mucosa, with a result that large gluten peptides up to 33 amino acids in length remain. In CD patients, an immune response to these immunodominant peptides develops. The gliadin peptides trigger an immune response that damages the small intestinal mucosa, which leads to many of the manifestations of CD [1,3]. CD is only partially clinically distinguishable from other small intestinal diseases with malabsorption, because patients with these disorders can present with similar signs and symptoms, including diarrhea, abdominal pain, weight loss, fatigue, and edema. Hence, the diagnosis of CD may be missed and therapy might be delayed, which may cause numerous medical complications, including malignancy, osteoporosis, and infertility [4]. The small intestinal mu-

* Correspondence authors.
E-mail addresses: rkgao@hfut.edu.cn (R. Gao), yuliu@hfut.edu.cn (Y. Liu).

https://doi.org/10.1016/j.cmpb.2019.105236
0169-2607/

Fig.1. VCE images of healthy controls (upper row) and celiac disease patients (lower row) .

cosa of healthy individuals contains finger-like projections known as villi (Fig. 1, upper row). However, as shown in Fig. 1 (lower row), the intestinal mucosa of CD patients with villous atrophy may have a mosaic appearance and fissuring, which leads to a lowered ability to absorb nutrients.

Diagnostic steps to detect CD usually include serological testing for tissue transglutaminase antibodies, upper gastrointestinal endoscopy, and duodenal biopsy [1,5,6]. However, these examination methods are somewhat invasive. More recently, VCE has been developed to promote a visual confirmation of suspected villous atrophy in CD [5]. Compared with conventional endoscopy, VCE is noninvasive and is able to provide evidence for subtle pathology throughout the entire small intestine [5,6]. Physicians evaluate the mucosal images manually according to videoclips captured by VCE. However, it is a time-consuming and tedious process to retrospectively analyze approximately 50,000 videoclips per six-hour video. As a consequence, computer-aided algorithms have been suggested, and rudimentary types have been introduced, to assist the diagnosis of CD in recent years.

Previous work focused on the visual quantitative analysis of villous atrophy to assess the severity of mucosal abnormalities in CD [7,8]. Koh et al. [9] applied the discrete wavelet transform (DWT) to extract significant textural and nonlinear features according to wavelet decomposition coefficients, and employed particle swarm optimization to select discriminative features for CD classification. Although the results obtained by the DWT method were promising, the method is limited to evaluating the presence of lesions based on grayscale images, which may ignore subtle color information. The performance of image modeling has been shown to be effective in patch-wise methods, because selective image patches are more resistive to noise and artifacts as compared to the whole image. Furthermore, the image patch is a tradeoff between pixel-wise spatial texture information and image-wise structural information, which can effectively capture the local image description and provide informative edge representation. Some of the satisfactory patch-wise CD detection methods have included threshold and incremental learning [7], dominant period analysis [10], shape-from-shading models [11], and color masking method [12]. In practice, these methods were shown to be useful to detect subtle abnormalities and presence of mucosal atrophy in the small bowel. However, high-level feature representation has not been addressed as yet.

With the availability of immense computing power, deep learning has emerged as a mainstream and state-of-the-art method in high-level computer vision tasks [13–17], and these could potentially become useful techniques to detect salient and subtle features of villous atrophy during automated analysis of videocapsule images. Convolutional neural networks (CNN) are a trainable end-to-end architecture that can be used for hierarchical feature representation. Zhou et al. [18], revealed that the sensitivity and specificity of CD analysis can be significantly improved by introducing the pre-rotation scheme in a GoogLeNet network. In Ref [19]., the multi-layer perceptron (MLP) model was developed based on conventional endoscopic image data, wherein the results indicated deep MLP architectures that can be highly suited for the classification of CD. Nevertheless, a multi-layer neural network architecture would suffer from the well-known nuisance of the vanishing gradient.

The deep learning method is a pool of the CNN-based network. Several approaches were applied to elevate network performance, such as increasing network depth [20], extending the width of the network (e.g. strict constraints on computational budget to achieve higher performance using inception architecture) [21], and attention mechanism [22]. However, more parameters were introduced as increasing network depth, which is not applicable in several sensitive scenes (e.g. embedded system development and mobile phone application). The channel attention mechanism is an amazing component, which sufficiently utilizes the connectivity of the feature channel to adaptively boost useful information and suppress weak information.

It is well known that attention mechanism plays an important role in human perception [23,24] . The squeeze and excitation (SE) attention mechanism models the channel dependence in a global receptive domain, since local spatial information in celiac disease is prone to strong correlation. We got the idea about the application of the block-wise channel squeeze and excitation (BCSE) attention module from human visual perception, which can selectively focus on important factors. For example, in order to recognize a new subject, one important property of our visual system is to give a whole view of scene, and the other is to exploit the most salient

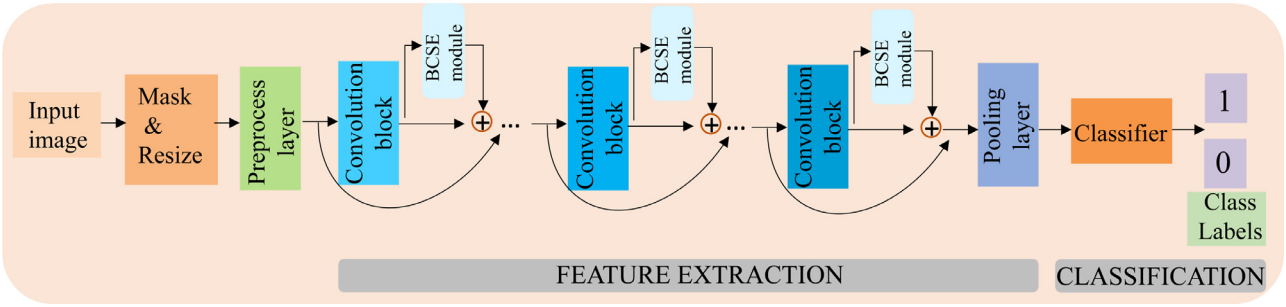

Fig. 2. Schematic diagram of the BCSE learning module embedded in residual network.

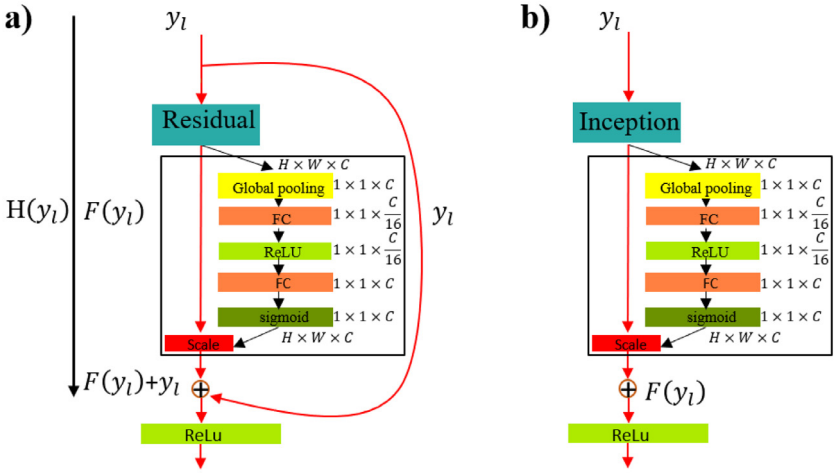

Fig. 3. The structure and deployment of SE embedded in residual and inception unit.

parts in the local spatial domain. The BCSE attention module not only retained the importance of the channel-wise feature map, but also improved the representation capacity in the block-wise spatial location.

In this work, we explored performance gains on SE [22] and squeeze spatial information after channel recalibration (SCSE) [25] and proposed BCSE learning module. Three recalibrated blocks can then be seamlessly integrated into ResNet50 and Inception-v3 by combining the optimal channel and space-wise information. As depicted in Fig. 2, the BCSE learning module was embedded in ResNet50 to obtain a refined network. The original images were masked, resized and delivered to a series of stacked convolution blocks for feature extraction. The BCSE module was embedded after convolution blocks to recalibrate the importance of the block-wise channel. The detailed information concerning the preprocessing layer and convolution block was reported by Hu et al. [22]. Implementation details concerning the BCSE module can be found in Fig. 3(c). The outer arc represented skip connections.

In addition, a different auxiliary classifier, such as a support vector machine (SVM), K-nearest neighbor (KNN), and linear discriminant analysis (LDA) can be used to validate the availability of our proposed BCSE learning module in adaptively detecting villous atrophy in the small intestinal mucosa, which is evidence of CD. The novelty of our work is threefold:

(1) We proposed a novel BCSE learning module which addressed the importance of local salient features into consideration; it can be merged into CNN-based networks to promote CD recognition. Control experiments with Inception-v3 and ResNet50 networks indicated significant productivity of BCSE in the diagnosis of CD.
(2) The SE block with channel recalibration is firstly applied to adaptively recalibrate the features of CD images. This can boost the useful pathology information and suppress less salient content.
(3) The combination of ResNet50 with the SVM classifier can be useful to measure discriminative and subtle villous atrophy in CD.

2. Methods

2.1. SE learning module embedded in residual and inception unit

Residual mapping is inherently important for training extremely deep networks, and it can be performed by residual units. As shown in Fig. 3(a), the residual unit can be realized by attaching a skip connection of stacked convolutional layers, rectified linear unit layers, and batch-normalization layers. It can be formulated as:

$$y_{l+1} = Relu(y_l + F(y_l,\ w_l)) \tag{1}$$

Here y_l and y_{l+1} denote the input and output feature maps of the l-th residual unit, $F(\bullet)$ is the residual unit mapping function, and $Relu(\bullet)$ represents the activation function of the rectified linear unit (Relu) [26].

The original plain network needs to learn the mapping function $H(y_l)=F(y_l)+y_l$, but it has to undergo sharp gradient increases. To avoid this problem, a deep residual network, which inherits the advantage of a normalization operation, can only fit the other mapping function of $F(y_l) = H(y_l) - y_l$. The residual unit can be deployed in a sequential or skipped layer to realize identity mapping, and then directly deliver the feature from the former layer to the subsequent one. Furthermore, the introduction of a residual is more sensitive to propagate gradients and rapidly converges without additional parameters.

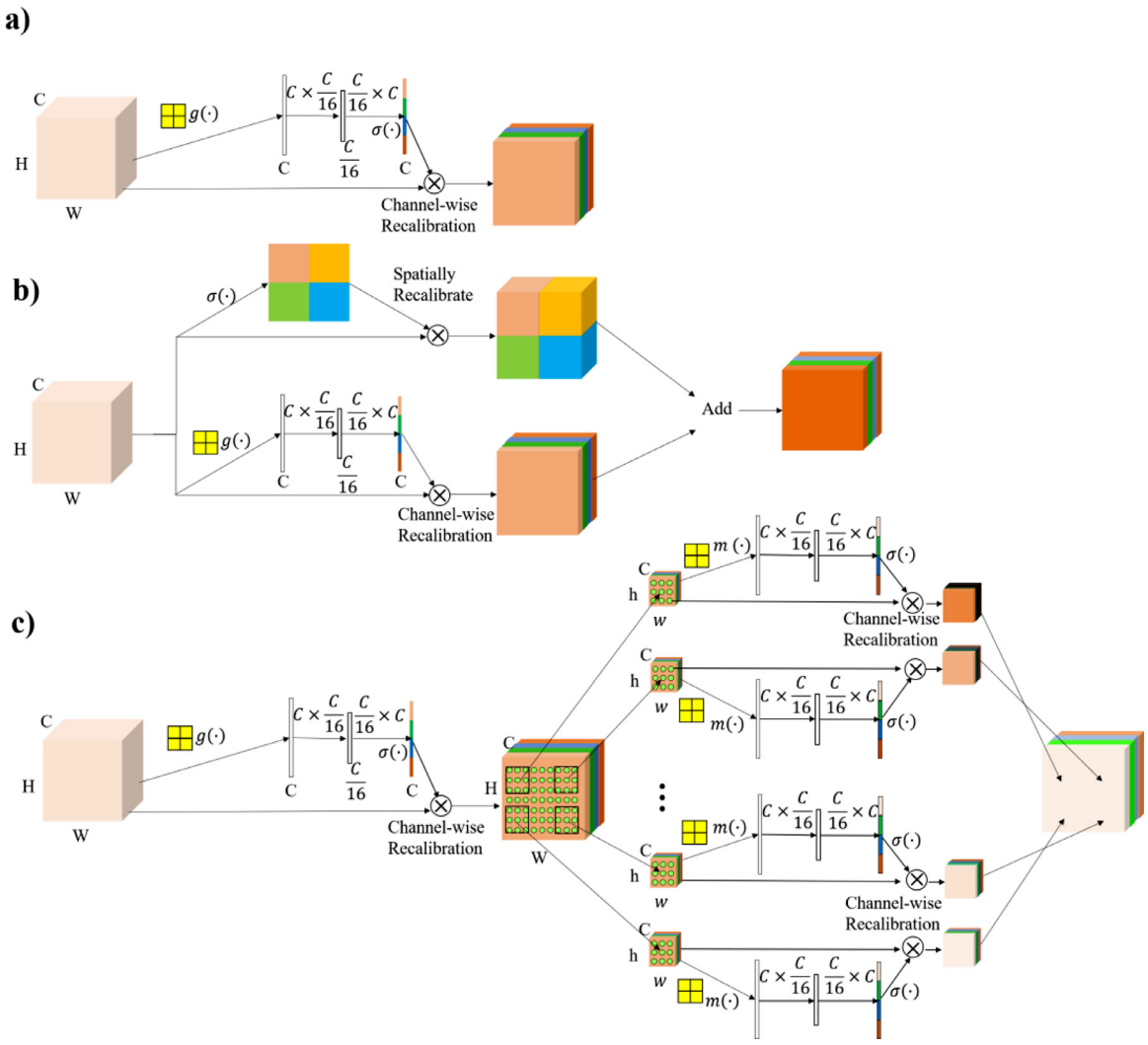

Fig. 4. Illustration of network encoding component with SE and SCSE module (a and b). The proposed BCSE module (c).

As can be seen from the black box of Fig. 3, the SE learning module is embedded after the residual and inception unit to complete the recalibration process. The dimension of the feature map is assumed as $R^{H \times W \times C}$, it can pass through a global average pooling operation to aggregate global response in a channel-wise manner. Then, a shared multi-layer perceptron (MLP) is used for attention inference to complete the excitation of channel dependence. Finally, the feature maps are reweighted by sigmoid activation to generate a new version that can be delivered to subsequent layers. SE was seamlessly integrated into ResNet50 and Inception-v3 networks. Fig. 4(c) shows the process of sequentially inferring a finer attention map though SE and subsequent block-wise recalibration. Here, the max-pooled features can compensate the average-pooled features.

Inception-v3 is a CNN-based network [21]. It inherits a special incarnation of inception architectures, and the idea of factorization into small convolutions was adopted in the inception mode. The introduction of an inception mode has proven to be effective in overcoming high computational costs [21]. Detailed information about network architecture can be found in [21].

The derivation process of BCSE is as follows. Fig. 4(a) shows the 3-dimensional recalibration process of SE, where SE was firstly reported in [22]. The output feature map was $U \in R^{H \times W \times C}$ and calculated through the residual network, which contains valuable channel and spatial information. Here, the height of the feature map was defined as H and the width as W, and C denoted the channel size. For each residual unit, the SE learning module was embedded to complete the recalibration process (Fig. 3(a) black box). SE divided the feature map U into a vectorization version $[U_1, U_2, \ldots U_C]$, here $U_z \in R^{H \times W \times 1}$, where U_z is the z-th element of C. We acquired the global response of the feature channel through average pooling to complete the spatial squeeze:

$$g_z = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_z(i, j) \tag{2}$$

Through the above spatial squeeze operation, the spatial dimension of the feature map was compressed, and each two-dimensional feature channel U_z was transformed into a scalar. This scalar can capture global spatial information and distribution on the feature channel. Then, the feature vector g_z was computed as: $\hat{g} = W_1(\delta(W_2\ (g)))$ with $W_1 \in R^{C \times \frac{C}{16}}$ and $W_2 \in R^{\frac{C}{16} \times C}$, where W_1 and W_2 denote the weighting of two fully connected (FC) layers, and $\delta(\bullet)$ represents the rectified linear unit layer [26]. Then, the sigmoid activation operation is added to the output $\sigma(\hat{g})$ vector range in the interval [0, 1] to recalibrate the importance of each global feature channel:

$$\hat{U} = \left[U_1\ \sigma\left(\hat{g}_1\right),\ U_2\ \sigma\left(\hat{g}_2\right), \cdots,\ U_C\ \sigma\left(\hat{g}_C\right)\right] \tag{3}$$

In the SE attention module, the global average pooling operation serves as a global information container; it can capture the

global response of an image. Nevertheless, it has low sensitivity in a local spatial region and may miss the most salient feature. In this regard, we proposed a novel hybrid BCSE learning module. Fig. 4(c) depicts that the feature map computed after SE was assumed as $\hat{U} \in R^{H\times W\times C}$. BCSE divides the feature map $\hat{U}$ into a vectorization version $[\hat{u}_1, \hat{u}_2, \bullet\bullet\bullet, \hat{u}_N]$, here $\hat{u}_N \in R^{(h\times w\times C)}$, $N \in R^{(H/h\times W/w)}$, $\hat{u}_k = [\hat{u}_k^1, \hat{u}_k^2, \bullet\bullet\bullet, \hat{u}_k^C], k = 1, 2, \bullet\bullet\bullet, N$ where h is the height and w is the width of the block-wise feature map. The key operations of the block-wise squeeze and excitation exhibited spatial max-pooling and channel excite. We treated each feature map in an iterative manner to acquire the block-wise salient response through spatial max-pooling:

$$m_k = \arg\max \hat{u}_k(i,\ j),\ i = 1,\ \cdots,\ \text{h and } j = 1,\ \cdots,\ \text{w} \qquad (4)$$

where $m_k \in R^{1\times 1\times C}$, and each block-wise channel feature map can acquire the most salient feature. Then, the feature vector m_k was computed as: $\hat{m}_k = W_1^{'}(\delta(W_2^{'}(m_k)))$ with $W_1^{'} \in R^{C\times \frac{C}{16}}$ and $W_2^{'} \in R^{\frac{C}{16}\times C}$. Thereafter the sigmoid activation operation was added to the output $\sigma(\hat{m}_k)$ vector range in the interval [0, 1] to recalibrate each block-wise feature channel:

$$\hat{u}_k^{'} = \left[\hat{u}_k^1\ \sigma\left(\hat{m}_k^1\right),\ \hat{u}_k^2\ \sigma\left(\hat{m}_k^2\right), \cdots,\ \hat{u}_k^C\ \sigma\left(\hat{m}_k^C\right)\right] \qquad (5)$$

Then, the block-wise recalibration was obtained by $\hat{U}' = [\hat{u}_1', \hat{u}_2', \cdots, \hat{u}_N']$. Through these operations, each block-wise feature channel assigned a weight to capture max-pooled channel-wise dependencies.

2.2. SCSE learning module embedded in residual and inception unit

Spatial information is an important component in medical image analysis. As can be observed from Fig. 3(b), Roy et al. proposed SCSE attention module for image analysis [25]. The channel recalibration of the residual mapping was supposed as $\hat{U}$, where $\hat{U} \in R^{H\times W\times C}$. If we consider pixel-wise spatial information, $\hat{U}$ can be written as $\hat{U} = [\hat{u}^{1,1}, \hat{u}^{1,2}, \cdots \hat{u}^{i,j} \cdots \hat{u}^{H,W}]$, here $\hat{u}^{i,j} \in R^{1\times 1\times C}$ Each $\hat{u}^{i,j}$ in feature map $\hat{U}$ represented the location of each spatial element. The spatial squeeze was computed through a convolution operation, where feature channel, kernel size, and stride are 1, 1, and 1, respectively. Therefore, the original feature map $\hat{U}$ was converted to $\hat{U}^{'} \in R^{(H\times W\times 1)}$. In this regard, each $\hat{U}_{i,j}^{'}$ can capture the global dependency in point-wise spatial information and assign a weight to each spatial location. The sigmoid function $\sigma(\bullet)$ projected each spatial location to a scalar version. Then, the spatial recalibration can be realized as follows:

$$U_{out} = \left[\sigma\left(\hat{U}_{1,1}^{'}\right)\hat{u}^{1,1},\ \sigma\left(\hat{U}_{1,2}^{'}\right)\hat{u}^{1,2}, \cdots \sigma\left(\hat{U}_{i,j}^{'}\right)\hat{u}^{i,j}, \cdots \sigma\left(\hat{U}_{H,W}^{'}\right)\hat{u}^{H,W}\right]$$

3. Experiments

3.1. Data acquisition and preprocessing

All experimental data were acquired from two PillCam imaging systems (Given Imaging, Yokneam, Israel and Medtronics)—SB2 and SB3, which were interpreted by two experienced gastroenterologists at Columbia University Medical Center, New York, USA. The SB2 dataset contained eight patients' data with villous atrophy and data from eight healthy controls, and the recorded videoclips included regions of the duodenal bulb, distal duodenum, proximal jejunum, distal jejunum, and ileum. The SB3 dataset contained four celiac patients' data and that of five healthy controls, including the regions of duodenum, ileum and jejunum. Both celiac patients and healthy controls swallowed a PillCam capsule after fasting. Then the capsule, which contained a mini camera, travelled through the small intestine by peristalsis. The raw data obtained from the external receiving sensor was a series of videoclips with a frame rate of two per second. Finally, the total dataset contained 52 CD videoclips and 55 healthy clips.

The videoclips were divided into frames with dimension 576 × 576 pixels. Then, 20 images of each videoclip were selected with evident villous atrophy. Those images included a mosaic appearance, fissuring, and scalloping of mucosal folds, mostly excluding degraded images with opaque extraluminal fluid, low-contrast, air bubbles, and overexposure. Finally, the entire database of 2140 color images with 576 × 576 pixels were gathered by means of VCE. Among them, 1100 images belong to the healthy mucosa class, and the remaining 1040 images belong to the damaged mucosa class, as likely affected by CD. Then, the original 576 × 576 pixel images were masked and cropped into 512 × 512 dimensions, to remove the influence of boundary font and the black box.

3.2. Celiac disease classification

In this section, we discuss several persuasive comparative experiments that were performed in the study. SE, BCSE and SCSE in conjunction with ResNet50 and the Inception-v3 network are presented. All of the classification performance measures were implemented in the Python 2.7 interpreter, based on the PyTorch version 0.4.1 framework. They ran on an Inter(R) Xeon(R) CPU 2.40 GHz and 64.0 GB RAM accompanied by two NVidia GTX1080Ti graphics cards with 22 GB memory. The residual network structure of ResNet50 was reported in previous studies [22,27]. Briefly, the bottleneck structures of 1 × 1, 3 × 3 and 1 × 1 were embedded into each residual block to form different convolutional blocks, which differed in the output channel size. The first and last pooling layers were not counted. The remaining 3, 4, 6, and 3 convolution blocks, the preprocessed convolutional layer, and the last fully connected layer, formed the 50 layers structure of ResNet50. As can be seen from Fig. 2, SE, SCSE and BCSE learning modules were embedded after each convolution block to construct different variants. This enabled the ResNet50 to reap the benefits of the SE, SCSE and BCSE units. Similarly, SE, SCSE and BCSE learning modules were embedded after each inception structure to form different variants of Inception-v3.

A 10-time 10-fold cross-validation strategy was performed in this study. All data were randomly shuffled and split into 10 folds. We took turns to let 9 folds be involved in training, while the other participated in test. The mean and standard deviation were given from the 10 times results. All data were processed as a tensor, and each channel was normalized with mean = [0.46, 0.456, 0.406] and standard deviation = [0.229, 0.224, 0.225], as described in [22]. The weight initialization was based on He initialization [28]. CD images were assigned with label 0, and control subjects with label 1, respectively. The learning rate was initialized with a value of 0.01. If the present training loss was higher than the previous one, the learning rate was decayed by 0.5. The batch-size was 16 in all networks, due to the limitation of computer memory. In addition, we adopted the L2 regularization strategy to punish the part which took excessive weight in the loss function with a weight decay λ=0.0005, to avoid overfitting. The learning rate and weight decay were referred to the original paper of the baseline SE and SCSE modules for better optimization and validation of the proposed BCSE module [22,25]. The dense feature of CD was constructed by the models that adaptively optimized through the Adam algorithm. The cross-entropy loss function was employed to calculate the network output with ground-truth. After one shot training, two aspects were tested. One was the network output provided by its end-to-end Softmax classifier. Secondly, the feature points were extracted by a parametric model before the last fully connected layer in the training and test data. After extraction, training and test data points were obtained with dimension of the

Table 1
celiac disease classification with ResNet50 embedded with different attention modules.

Network	Acc.(%)	Sen.(%)	Spe.(%)	F1-score(%)	Recall(%)	Param.(M)	GFLOPs
ResNet50+None	90.73 (0.07)	92.58 (0.08)	88.70 (0.11)	90.15 (0.09)	88.70 (0.11)	22.42	4.136
ResNet50+SE	93.63 (0.06)	93.62 (0.08)	93.30 (0.05)	93.72 (0.05)	93.30 (0.05)	24.84	4.142
ResNet50+SCSE	87.44 (0.07)	90.71 (0.09)	84.37 (0.12)	86.54 (0.07)	84.37 (0.12)	24.85	4.145
ResNet50+BCSE	95.85 (0.04)	96.94 (0.05)	94.57 (0.05)	95.75 (0.04)	94.57 (0.05)	27.66	4.146

*mean (standard deviation) of 10-time statistical results.
**None represent the original network, with accuracy(Acc.), sensitivity(Sen.), specificity(Spe.), F1-score, Recall, parameters(Param.) and floating point operations(FLOPs) as evaluation matrix.

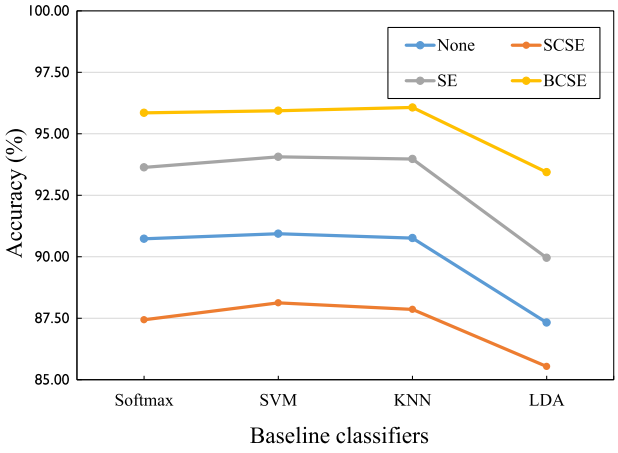

Fig. 5. The mean accuracy plot of 10-time 10-fold using ResNet50 embedded with SCSE, SE and BCSE.

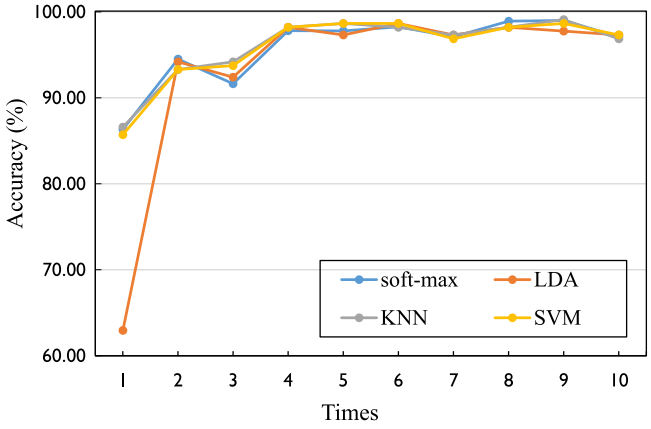

Fig. 6. Celiac disease classification rate in 10-time statistics with Softmax, SVM, KNN and LDA classifiers using 10-fold cross-validation.

sample size*2048. Those dense feature points can be used to fit SVM, KNN, and LDA to obtain the corresponding ten times statistics. SVM was configured with a radial basis function (rbf) kernel, penalty coefficient (gamma) of 1, and cost factor (C) of 1. The non-parametric KNN was a distance-based classifier with a neighbor of 10 in accordance with a previous study [9].

4. Results

4.1. Evaluation of the BCSE learning module in ResNet50

To validate the contribution of the proposed BCSE learning module in CD recognition, we reimplemented ResNet50 and embedded with several baseline learning modules. All images were resized from 512 × 512 to 224 × 224. In this study, the Grid Search was used to enumerate all candidate h and w with a step size of 2. The optimal block-size in BCSE was 14 × 14. As displayed in Fig. 5, performance gains of BCSE in Softmax, SVM, KNN, and LDA classifiers all outperformed SE and SCSE with respect to average accuracy statistics. Table 1 lists the performance matrix from average accuracy, sensitivity, specificity, F1-score, recall rate and model parameters, where its definition can be found in [29]. From the analysis of standard deviation, the proposed BCSE is more stable than SE and SCSE. Furthermore, it was found that the BCSE module demonstrated significant improvement in performance as compared with state-of-the-art baseline attention modules, with minimal additional parameters and computational complexity. ResNet50 embedded with BCSE achieved an average accuracy, sensitivity, specificity, F1-score, and recall rate of 95.85%, 96.94%, 94.57%, 95.75% and 94.57%, respectively. Those results exceed by 2.22%, 3.32%,1.27%, 2.03% and 1.27%, as compared to ResNet50 with SE. It also can be found that SCSE suffered from a low productivity in CD recognition. This may result from the SCSE was generally employed for image segmentation.

Fig. 6 depicted the 10-fold average recognition accuracy of BCSE embedded in ResNet50 from 10-times statistics. The feature points were further sent to SVM, KNN, and LDA for a second training process. It can potentially promote gains versus its end-to-end Softmax classifier. In particular, retraining SVM was superior to Softmax, KNN, and LDA classifiers each time. Table 2 shows the mean and standard deviation for 10-time 10-fold. Here, BCSE demonstrated a substantial performance increase as compared with SE. Fig. 7 depicts the training and validation loss of ResNet50. It was found that BCSE consistently enjoys performance improvements during training. After 100-epochs, the models were evolved until training and validation loss converged.

4.2. Evaluation of BCSE learning module in Inception-v3

To assess the shared benefit gains in other CNN based networks, a benchmark Inception-v3 network was selected. We resized the images from 512 × 512 to 299 × 299 pixels, and the auxiliary classifiers were closed as parameters regularization [21]. The protocol of hyper-parameter and cross validation were consistent with ResNet50. As shown in Fig. 8, the same phenomenon was observed that Inception-v3 embedded with BCSE was superior to SCSE and SE in CD recognition. Table 3 lists the performance comparison of the end- to-end Softmax model. BCSE can potentially outperform SCSE and SE, which has an average accuracy, sensitivity, specificity, F1-score, and Recall rate of 87.30%, 88.09%, 86.45%, 87.05% and 86.45%, respectively. They exceeded Inception-v3 embedded with SE by 0.53%, −2.07%, 3.05%, 0.76% and 3.05%. This demonstrates that the BCS E learning mechanism is more effective than SE in the Inception-v3 network for CD recognition. Fig. 9 depicts the training curves of SE-Inception-v3 and BCSE-Inception-v3 within 100 iterations, which showed the robustness of our proposed BCSE learning module.

Table 2
celiac disease classification with different classifiers.

Network	SVM(rbf)			KNN		
	Acc.(%)	Sen.(%)	Spe.(%)	Acc.(%)	Sen.(%)	Spe.(%)
ResNet50+CSE	94.06 (0.05)	94.17 (0.06)	94.04 (0.063)	93.97 (0.05)	93.26 (0.05)	94.73 (0.04)
	95.94 (0.04)	97.20 (0.03)	94.53 (0.03)	95.07 (0.04)	96.23 (0.04)	95.63 (0.04)
Network	LDA			soft-max		
	Acc.(%)	Sen.(%)	Spe.(%)	Acc.(%)	Sen.(%)	Spe.(%)
ResNet50+CSE	89.96 (0.10)	91.39 (0.10)	88.64 (0.10)	93.63 (0.06)	93.62 (0.08)	93.30 (0.05)
ResNet50+BCSE	93.44 (0.10)	95.53 (0.07)	91.27 (0.13)	95.85 (0.04)	96.94 (0.05)	94.57 (0.05)

** where accuracy (Acc.), sensitivity (Sen.), specificity (Spe.) as evaluation matrix.

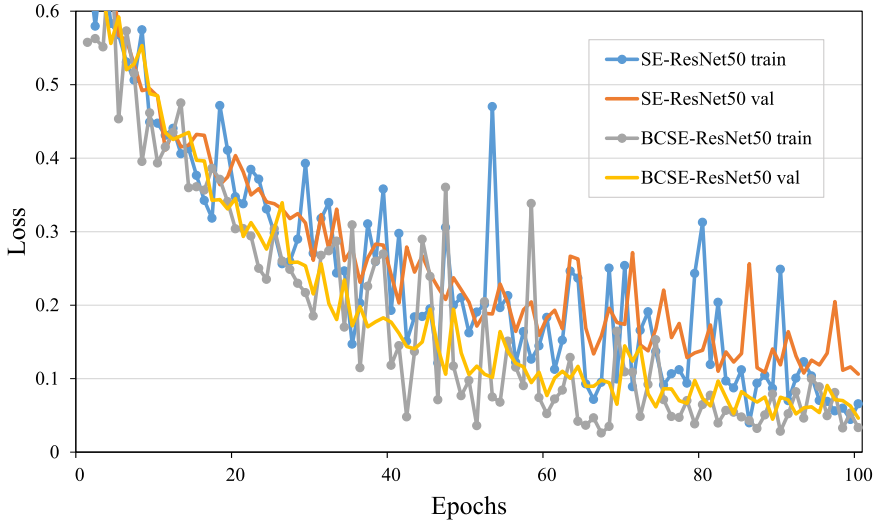

Fig. 7. Training curves of SE-ResNet50 and BCSE-ResNet50 in 10-time 10-fold.

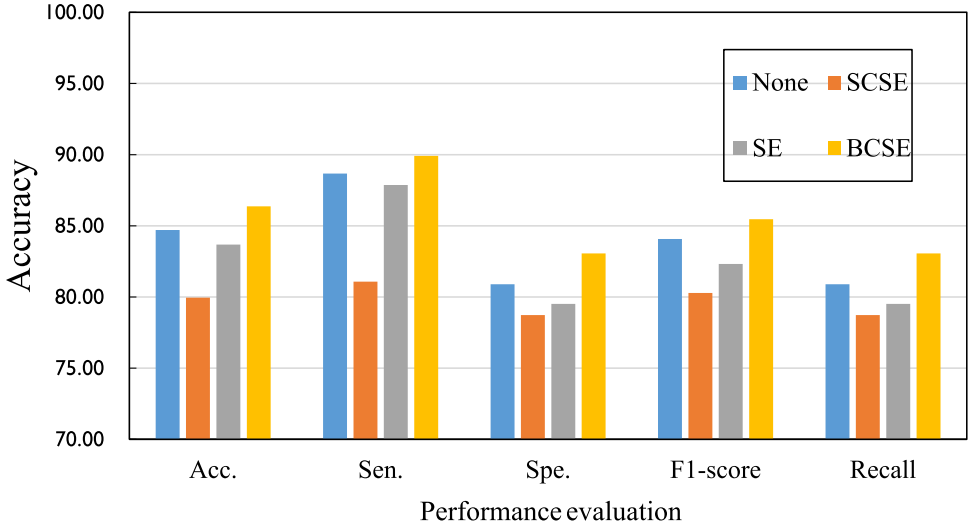

Fig. 8. The mean performance plot of 10-time 10-fold using Inception-v3 embedded with SCSE, SE and BCSE.

Table 3
Celiac disease classification with Inception-v3 embedded with different learning modules.

Network	Acc.(%)	Sen.(%)	Spe.(%)	F1-score(%)	Recall(%)	Param.(M)	GFLOPs
Inception-v3+None	84.71 (0.09)	88.67 (0.10)	80.89 (0.11)	84.08 (0.09)	80.89 (0.11)	20.78	5.747
Inception-v3+SE	86.77 (0.06)	90.16 (0.08)	83.40 (0.12)	86.29 (0.07)	83.40 (0.12)	21.87	5.747
Inception-v3+SCSE	79.96 (0.09)	81.08 (0.20)	78.73 (0.10)	80.28 (0.07)	78.73 (0.10)	21.88	5.747
Inception-v3+BCSE	87.30 (0.06)	88.09 (0.07)	86.45 (0.09)	87.05 (0.07)	86.45 (0.09)	23.40	5.747

*mean (standard deviation) of 10-time statistical results.
**None represent the original network, with accuracy(Acc.), sensitivity(Sen.), specificity(Spe.), F1-score, Recall, parameters(Param.) and floating point operations(FLOPs) as evaluation matrix.

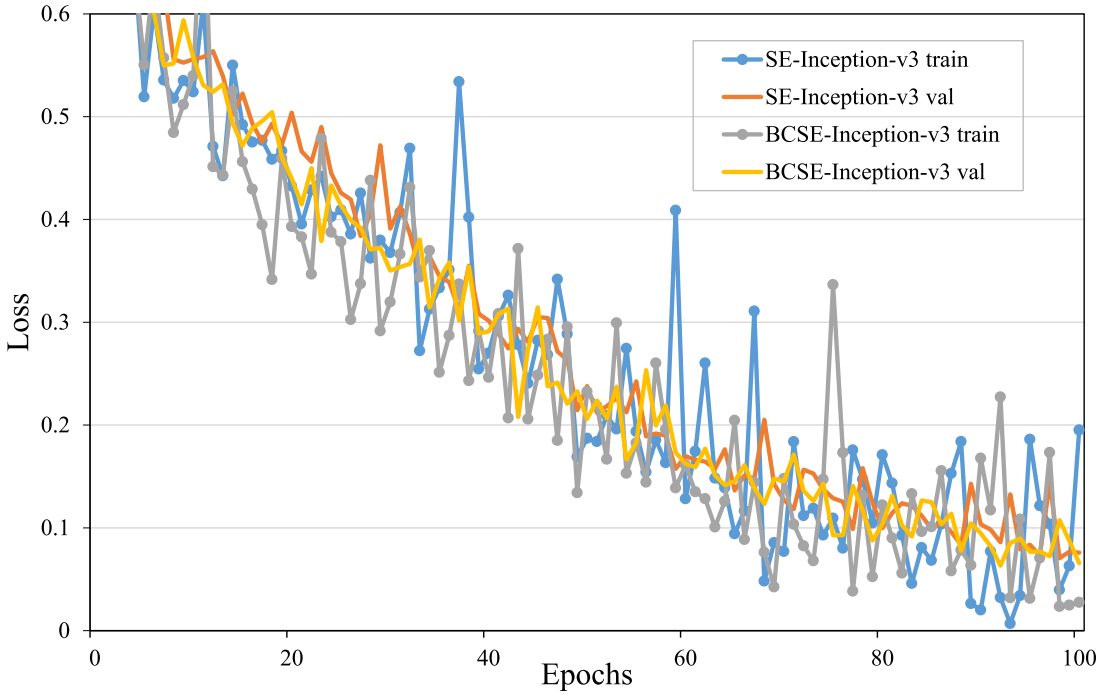

Fig. 9. Training curves of SE-Inception-v3 and BCSE-Inception-v3 in 10-time 10-fold.

5. Discussion

CD often manifests as scalloped duodenal folds, mucosal fissures, and submucosal vessels in the small intestinal mucosa. Owing to its special feature alignment patterns, we proposed a novel BCSE recalibration module to capture its salient features, where it can be more sensitive to a local feature region. The comparative results showed that BCSE achieved better feature representation capacity as compared with SE and SCSE.

The purpose of this work was to find a novel deep learning attention module for CD recognition of VCE images. However, current intelligent deep learning solutions usually are equipped with many parameters, resulting in serious time consumption of the computing resource. Simultaneously, greater computational complexity would occur with a larger data dimension. It is difficult to make a satisfactory balance between computer memory and real-time processing. Hence, most computer vision tasks (e.g., CIFAR-10, ImageNet and FERET) are downsampled so that the image size is smaller than 224 $\times$ 224 pixels [30–32].

The traditional methods were based on a prior feature selection and classifier training, which can distinguish villous atrophy from normal small intestinal mucosa [33]. However, this may lack sensitivity for subtle changes in villous atrophy and high-dimensional features of the image. The deep learning method is a hierarchical dense representation mechanism. In this work, we integrated a global SE learning module into the CD recognition process. In particular, the importance of local salient textural features were taken into consideration. The proposed BCSE learning module, serving as a special attention and gating mechanism, was used in each feature channel to enhance the expressive ability of each residual and inception unit.

Computational time plays an important role in medical image analysis. Each feature map was divided block-wise by the BCSE, which adopted a new cascaded recalibrate mode. Fig. 10 displayed the forward propagation running time versus different baseline attention module in ResNet50 and Inception-v3 networks, which was input with batch-size VCE images. Regarding the ResNet50 network, the running time was increased as the complexity of the baseline attention module increased. It also showed a similar trend in the Inception-v3 network. Notably, BCSE blocks showed a time extension in ResNet50 as compared to other attention modules. It only increased a very small fraction in the Inception-v3 network. It can be found that the original ResNet50 (no additional attention module) has a small time consumption, while the original Inception-v3 network requires a larger time consumption. Hence, the proportion of original network time consumption strongly affected the running time of the attention module embedded network. In addition, the time extension of BCSE in ResNet50 was higher than BCSE in the Inception-v3 network, because the ResNet50 network had 16 residual units while only 11 inception units were presented in Inception-v3 network. The time extension occurred at the network training phase. As the network became stable, there was no need to train again for inference. Hence, the time cost was negligible in the context of significant performance gains, which ResNet50 embedded with BCSE exceeded by 8.55%, 8.85%, 8.12%, 8.7% and 8.12%, as compared to Inception-v3 with BCSE in average accuracy, sensitivity, specificity, F1-score, and recall rate.

It should be noted that the proposed method still has several limitations. It is comparatively new to study celiac disease by using video capsule endoscopy. The benchmark dataset is small from patient-wise division (only comprised of 12 patients and 13 controls). In addition, with the limited numbers of volunteers, the patient-wise results may not help to statistically evaluate the performance of different learning modules (BCSE, SE and SCSE). We would need to recruit more volunteers to collect data for patient-wise division in future studies. The purpose of this study was to implement a novel joint learning module for celiac disease analysis. The parameter setting was referred to the reported baseline attention module [22,25]. It could be finely adjusted and further optimized for better classification performance. Another limitation is that the BCSE module requires a longer computation time for

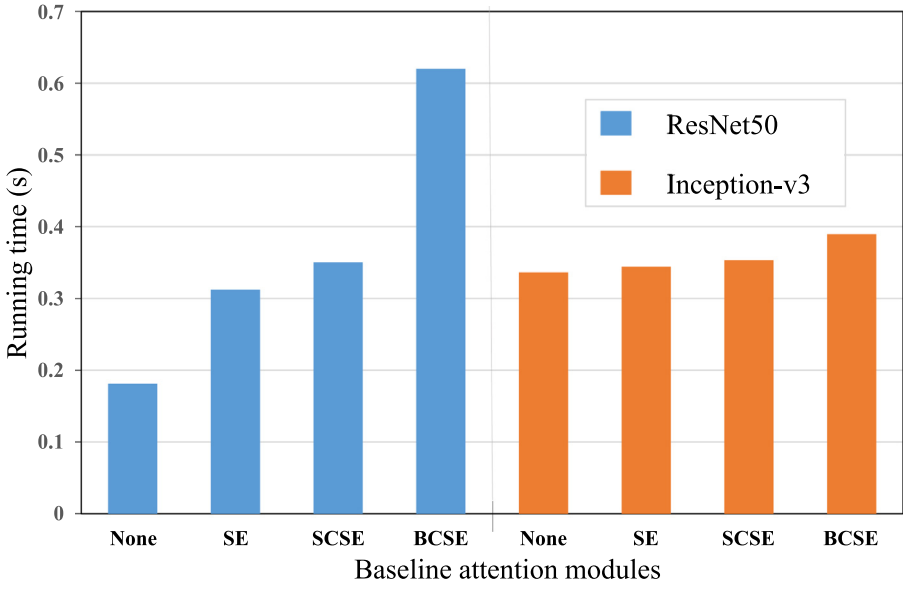

Fig. 10. The forward propagation time versus different baseline attention modules in the ResNet50 and Inception-v3 networks.

Table 4
A list of previous publications on the diagnosis of CD.

Authors	Year	Techniques	Imaging system	Conclusion
Ciaccio et al.	2010a	• Mean and SD in brightness	SB2	Threshold classifier: SEN: 80% SPE: 96% Incremental classifier: SEN: 88% SPE: 80%
Ciaccio et al.	2010b	• Mean and SD in brightness	SB2	SEN: 92.7% SPE: 93.5%
Ciaccio et al.	2011	• Shape-from-shading principles • Image transformation	SB2	SEN:83.9 SPE:92.9 ACC:88.1
Ciaccio et al.	2014	• Histogram level	SB2	SEN: 76.9%–84.6% SPE: 69.2%–92.3%
Zhou et al.	2017	• CNN (22-layer GoogLeNet)	SB2	SEN: 100% SPE: 100%
Koh et al.	2018	• DWT • Nonlinear, textural features • PSO • SVM classifier	SB2 and SB3	SEN: 88.43% SPE: 84.60% ACC: 86.47%
Ciaccio et al.	2019	• color masking • linear discriminant function	SB2 and SB3	ACC:80%
Vicnesh et al.	2019	• Daisy descriptors • PSO • Decision Tree, kNN,PNN, SVM • PNN, SVM	SB2 and SB3	SEN: 94.35% SPE: 83.20% ACC: 89.82%
Present work	2019	• BCSE learning module • ResNet50, Inception-v3 • SVM, KNN, LDA	SB2 and SB3	SEN: 97.20% SPE: 95.63% ACC: 95.94%

network inference. However, the current study, including the joint global channel and local salient spatial recalibration can serve as a proof-of-concept for celiac disease diagnosis.

As can be seen from Table 4, compared to the state-of-the-art methods, our proposed methods sufficiently utilized the RGB color information to extract high-level and specific dense features of villous atrophy in CD without any image selection.

6. Conclusions

A novel deep learning feature learning module was developed to boost the useful local pathology information while suppressing less meaningful components, to enable the development of a tool assistive for CD diagnosis. The CNN based method is a pool of models, which does not involve the process of feature

selection, while the proposed BCSE learning module offers a new choice for utilizing deep learning methods over the conventional machine learning algorithm to improve the diagnosis of CD. We extracted dense feature vectors from ResNet50 and Inception-v3 embedded with SE, SCSE and BCSE. Then, classical SVM (rbf), KNN, and the LDA classifiers model were utilized to validate the utility of the proposed methods. The results demonstrated that block-wise channel squeeze and excitation successfully distinguishes villous atrophy in CD from controls, with an accuracy of 95.94%, and sensitivity and specificity of 97.20% and 95.63%, respectively.

Recently, we became more focused on computer assisted diagnosis for CD. Due to the special imaging modality, the Scale-invariant feature transform (SIFT) is of interest and can be integrated into deep learning methods in future work. In addition, deep learning techniques have shown significant advantages in the field of medical image analysis. More powerful algorithms and datasets of larger size and structure can be considered for integration of technologies for the diagnosis of CD, to achieve rapid diagnosis and treatment.

Declaration of Competing Interest

There are no conflicts to declare.

Acknowledgments

The National Natural Science Foundation of China supported this work (Nos. 61601165, 61571176). We also acknowledge financial support from the Fundamental Research Funds for the Central Universities (No. JZ2019HGTB0088), the China Postdoctoral Science Foundation (No. 2018T110613), the Anhui Key Project of Research and Development Plan (No. 1704d0802188). We also wish to thank Mr. and Mrs. Daniel Wallen for supporting this research. This work was also partially supported by the Open Project of Faculty of Chemistry of Qingdao University of Science and Technology (QUSTHX201805) .

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2019.105236.

References

[1] P.H.R. Green, C. Cellier, Celiac disease, N. Engl. J. Med. 357 (17) (2007) 1731–1743.
[2] V. Fuchs, K. Kurppa, H. Huhtala, M. Maki, L. Kekkonen, K. Kaukinen, Delayed celiac disease diagnosis predisposes to reduced quality of life and incremental use of health care services and medicines: a prospective nationwide study, United Eur. Gastroenterol. J. 6 (4) (May 2018) 567–575.
[3] V. Rotondi Aufiero, A. Fasano, G. Mazzarella, Non-Celiac gluten sensitivity: how its gut immune activation and potential dietary management differ from celiac disease, Mol. Nutr. Food Res. 62 (May (9)) (2018) e1700854.
[4] P.H.R. Green, A.T. Fleischauer, G. Bhagat, R. Goyal, B. Jabri, A.I. Neugut, Risk of malignancy in patients with celiac disease, Am. J. Med. 115 (3) (2003) 191–195.
[5] S.K. Lewis, C.E. Semrad, Capsule Endoscopy and Enteroscopy in Celiac Disease, Gastroenterology Clinics of North America, 2018.
[6] M.E. Robert, S.E. Crowe, L. Burgart, R.K. Yantiss, B. Lebwohl, J.K. Greenson, S. Guandalini, J.A. Murray, Statement on best practices in the use of pathology as a diagnostic tool for celiac disease: a guide for clinicians and pathologists, Am. J. Surg. Pathol. 42 (Sep (9)) (2018) e44–e58.
[7] E.J. Ciaccio, C.A. Tennyson, G. Bhagat, S.K. Lewis, P.H. Green, Classification of videocapsule endoscopy image patterns: comparative analysis between patients with celiac disease and normal individuals, Biomed. Eng. Online 9 (September (1)) (2010) 44.
[8] E.J. Ciaccio, C.A. Tennyson, G. Bhagat, S.K. Lewis, P.H. Green, Use of basis images for detection and classification of celiac disease, Biomed. Mater. Eng. 24 (6) (2014) 1913–1923.
[9] J.E.W. Koh, Y. Hagiwara, S.L. Oh, J.H. Tan, E.J. Ciaccio, P.H. Green, S.K. Lewis, U. Rajendra Acharya, Automated diagnosis of celiac disease using DWT and nonlinear features with video capsule endoscopy images, Future Gener. Comput. Syst. 90 (2019) 86–93.
[10] E.J. Ciaccio, C.A. Tennyson, S.K. Lewis, S. Krishnareddy, G. Bhagat, P.H. Green, Distinguishing patients with celiac disease by quantitative analysis of videocapsule endoscopy images, Comput. Methods Programs Biomed. 100 (Oct (1)) (2010) 39–48.
[11] E.J. Ciaccio, C.A. Tennyson, G. Bhagat, S.K. Lewis, P.H. Green, Implementation of a polling protocol for predicting celiac disease in videocapsule analysis, World J. Gastrointest. Endosc. 5 (Jul (7)) (2013) 313–322.
[12] E.J. Ciaccio, S.K. Lewis, G. Bhagat, P.H. Green, Color masking improves classification of celiac disease in videocapsule endoscopy images, Comput. Biol. Med. 106 (Mar) (2019) 150–156.
[13] M. Salvi, F. Molinari, N. Dogliani, M. Bosco, Automatic discrimination of neoplastic epithelium and stromal response in breast carcinoma, Comput. Biol. Med. (2019).
[14] F. Molinari, G. Barbato, and N. Michielli, “Analisi della performance neurocognitiva usando un singolo canale EEG,” 2019.
[15] N. Michielli, U.R. Acharya, F. Molinari, Cascaded LSTM recurrent neural network for automated sleep stage classification using single-channel EEG signals, Comput. Biol. Med. 106 (2019) 71–81.
[16] A. Shahin, Y. Guo, K. Amin, A.A. Sharawi, White blood cells identification system based on convolutional deep neural learning networks, Comput. Methods Programs Biomed. (2017).
[17] U.R. Acharya, U. Raghavendra, J.E. Koh, K.M. Meiburger, E.J. Ciaccio, Y. Hagiwara, F. Molinari, W.L. Leong, A. Vijayananthan, N.A. Yaakup, Automated detection and classification of liver fibrosis stages using contourlet transform and nonlinear features, Comput. Methods Programs Biomed. 166 (2018) 91–98.
[18] T. Zhou, G. Han, B.N. Li, Z. Lin, E.J. Ciaccio, P.H. Green, J. Qin, Quantitative analysis of patients with celiac disease by video capsule endoscopy: a deep learning method, Comput. Biol. Med. 85 (Jun) (2017) 1–6.
[19] G. Wimmer, S. Hegenbart, A. Vécsei, and A. Uhl, "Convolutional neural network architectures for the automated diagnosis of celiac disease." pp. 104–113.
[20] N.N. Sultana, B. Mandal, N. Puhan, Deep residual network with regularised fisher framework for detection of melanoma, IET Comput. Vis. 12 (8) (2018) 1096–1104.
[21] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision."
[22] J. Hu, L. Shen, and G. Sun, “Squeeze-and-Excitation networks.”
[23] R.A. Rensink, The dynamic representation of scenes, Vis. Cognit. 7 (1–3) (2000) 17–42.
[24] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (Nov (11)) (1998) 1254–1259.
[25] A.G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel ‘squeeze & excitation’in fully convolutional networks." pp. 421–429.
[26] V. Nair, and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines." pp. 807–814.
[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." pp. 770–778.
[28] K. He, X. Zhang, S. Ren, J. Sunleee, Delving deep into rectifiers: surpassing human-level performance on imagenet classification, in: IEEE International Conference on Computer Vision, IEEE International Conference on Computer Vision, 2015, pp. 1026–1034. 2015.
[29] S. Raschka, An overview of general performance metrics of binary classifier systems, Comput. Sci. (2014).
[30] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” 2016.
[31] M.C. Kruithof, H. Bouma, N.M. Fischer, and K. Schutte, “Object recognition using deep convolutional neural networks with complete transfer and partial frozen layers,” 2016.
[32] B.N. Li, Q. Yu, R. Wang, K. Xiang, M. Wang, X. Li, Block principal component analysis with nongreedy L1-Norm maximization, IEEE Trans. Cybern. 46 (Nov (11)) (2016) 2543–2547.
[33] E.J. Ciaccio, C.A. Tennyson, G. Bhagat, S.K. Lewis, P.H. Green, Transformation of videocapsule images to detect small bowel mucosal differences in celiac versus control patients, Comput. Methods Programs Biomed. 108 (Oct (1)) (2012) 28–37.