A Novel Global Spatial Attention Mechanism in Convolutional Neural Network for Medical Image Classification

Linchuan Xu\textsuperscript{1}, Jun Huang\textsuperscript{2}, Atsushi Nitanda\textsuperscript{2}, Ryo Asaoka\textsuperscript{3}, Kenji Yamanishi\textsuperscript{2}

\textsuperscript{1}PQ813, Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
linch.xu@polyu.edu.hk

\textsuperscript{2}Graduate School of Information Science and Technology
The University of Tokyo, Tokyo, Japan
\{jun.huang, nitanda, yamanishi\} @mist.i.u-tokyo.ac.jp

\textsuperscript{3}Department of Ophthalmology
The University of Tokyo, Tokyo, Japan
rasaoka-tky@umin.ac.jp

Abstract

Spatial attention has been introduced to convolutional neural networks (CNNs) for improving both their performance and interpretability in visual tasks including image classification. The essence of the spatial attention is to learn a weight map which represents the relative importance of activations within the same layer or channel. All existing attention mechanisms are local attentions in the sense that weight maps are image-specific. However, in the medical field, there are cases that all the images should share the same weight map because the set of images record the same kind of symptom related to the same object and thereby share the same structural content. In this paper, we thus propose a novel global spatial attention mechanism in CNNs mainly for medical image classification. The global weight map is instantiated by a decision boundary between important pixels and unimportant pixels. And we propose to realize the decision boundary by a binary classifier in which the intensities of all images at a pixel are the features of the pixel. The binary classification is integrated into an image classification CNN and is to be optimized together with the CNN. Experiments on two medical image datasets and one facial expression dataset showed that with the proposed attention, not only the performance of four powerful CNNs which are GoogleNet, VGG, ResNet, and DenseNet can be improved, but also meaningful attended regions can be obtained, which is beneficial for understanding the content of images of a domain.

Index terms— Attention, convolutional neural networks, medical image classification
1 Introduction

1.1 Background and Motivation

Medical imaging, such as computed tomography (CT), magnetic resonance (MR), and fundus photography, has played a crucial role in the detection, diagnosis, and treatment of diseases [1]. Due to variations in pathology and increased size of images, researchers and doctors have been resorting to artificial intelligence (AI) techniques for automated image analysis. Among all the AI techniques, deep learning, particularly convolutional neural network (CNN), has brought the most promising performance [1, 2, 3]. The successful application of CNNs is due to their automatic extraction of good feature representations which are the key to success of AI techniques [4].

Recently, spatial attention has been introduced to CNN to improve both its performance and interpretability [5, 6, 7, 8, 9]. The essence of spatial attention is to learn a weight map which represents relative importance of features within the same layer or channel. By promoting important features and suppressing unimportant features, the performance can be reasonably improved. In terms of interpretability, by looking at the regions to which large weights are assigned, i.e., attended regions, we may get to understand what features are mostly utilized by the CNN to produce a particular result.

All existing attention mechanisms are local attentions in the sense the weight maps are image-specific. This is because the weight map for each image is mainly determined by itself and each image is unique in the intensities of pixels. But in the medical field, even though individual images within a set are also unique, there are cases that the set of images should share the same weight map, i.e., one global weight map. The global weight map may be needed when the set of images are took on one kind of object for the same kind of disease by the same imaging technique at the same angle from which the object faces the image-taking device. We can easily tell that there exist not a few medical image datasets fit in with the procedure of the image generation mentioned above. We give a name, structured images, to these kinds of image sets. As the name suggests, it is expected that the structured images have organized information which conforms to a certain format. Therefore, images in other domains with organized information may be structured images as well.

When applying existing local attention mechanisms to the structured medical images, both the performance and the interpretability of a CNN may even be deteriorated. Local weight maps may cause additional overfitting since the maps are additional parameters, especially when the dataset is small-scale, which is often the case in the medical field. Local weight maps may also make the interpretation confusing because they may assign different weights to the same structured regions due to small variations in the pixel intensities.

In this paper, we thus propose a novel global spatial attention mechanism to learn a global weight map for structured medical images. A single weight map may be ignored in terms of the number of parameters, but can still enjoy the improvement in both the performance and the interpretability.

1.2 Novelty and Significance

The novelty and significance of this paper is summarized as follows.

(1) Proposal of a global spatial attention mechanism for the first time. To the best of our knowledge,
all existing attention mechanisms are local. Moreover, existing ones learn local weight maps in the hidden layers. In the proposed global attention mechanism, there is only one global weight map, and the global weight map is directly learned for the input layer, i.e., the input raw images. In brief, the global weight map is instantiated by a decision boundary between important pixels and unimportant pixels. And we propose to realize the decision boundary by a binary classifier. Note that we in fact do not know which pixels are important or unimportant. Therefore, the binary classifier is integrated into the image classification CNN and to be optimized together with the CNN.

(2) Applications to medical image classification. The proposed global attention mechanism is a generic approach and can be straightforwardly integrated into any CNN architectures. We study four powerful CNNs which are GoogleNet, VGG, ResNet and DenseNet, and apply them to two medical datasets. Moreover, an additional facial expression dataset is studied because the proposed method is essentially designed for structured images. The results showed that with the proposed attention, not only the performance of GoogleNet, VGG, ResNet, and DenseNet can be improved, but also meaningful attended regions can be obtained, which is beneficial for understanding the content of images of a domain.

1.3 Organization of This Paper

The rest of this paper is organized as follows. Section 2 presents notations and definitions used in this paper. The proposed method is developed in section 3. Section 4 presents results of empirical evaluation. Section 5 presents related work. In section 6, conclusions are drawn from this study and future directions are introduced.

2 Notations and Definitions

This paper deals with structured images defined as follows:

**DEFINITION 1 (STRUCTURED IMAGES)** Structured images have the same sizes in all dimensions, and are taken on one kind of objects under similar circumstances which include similar image-taking devices, similar context within which the objects locate, and similar angles from which the objects face the devices.

Although the proposed method is mainly for medical image classification, it essentially works for structured images defined above. It happens that structured images are mainly produced in the medical field because of the circumstances required for the images.

The way the structured images are generated guarantees that different images have similar characteristics at the same pixels, which gives the opportunity to learn a single weight map shared by all images within a set. Three images from each of the three studied structured image datasets are presented in Fig. 1 for illustration. The first two datasets contain medical images taken on human eyes. We studied an additional dataset about facial expression images which focus on human faces only to demonstrate the potential applications to fields other than the medical field. Details about the description of the three datasets can be found in the following section of experiments.
The classification of images is based on their characteristics which are pixel intensities. Similarly, for the global attention weight map, we propose to learn the importance of pixels based on pixels’ characteristics. In our problem setting, we only have images but have no additional information about each pixel. We thus obtain pixel characteristics by utilizing images themselves. We have learned that to measure the importance of features in the filter-style feature selection methods [10, 11, 12], the characteristics of each feature can be obtained as the values of all data instances at the feature. Since the studied images are structured images, we can also obtain the characteristics of a pixel as the intensities of all images at the pixel. And it is expected that there are different distributions of intensities among different pixels.

To facilitate the image classification and the pixel classification, we define an image representation and a pixel representation of a structured image dataset as follows.

**DEFINITION 2 (IMAGE REPRESENTATION)** Images are represented by $\mathcal{X} \in \mathbb{R}^{N \times C \times W \times H}$, which is a four-dimensional tensor where $N$ is the number of images, $C$ is the number of image channels, $W$ and $H$ are the width and height of each image, respectively.

**DEFINITION 3 (PIXEL REPRESENTATION)** Pixels are represented by $P \in \mathbb{R}^{NC \times W \times H}$, which is a three-dimensional tensor obtained by reshaping $\mathcal{X}$. The number of pixels is $WH$, and the dimension of the representation of each pixel is $NC$.

Note that we do not have ground-truths to perform the classification of pixels. Since the pixel importance learning is to serve the task of image classification, we thus propose to integrate the pixel classifier into the image classification CNN, and to optimize the pixel classifier together with the CNN. We define the studied problem as follows:

**DEFINITION 4 (THE STUDIED PROBLEM)** Given image representation $\mathcal{X}$ and pixel representation $P$, and image labels $Y \in \mathbb{R}^N$, the objective to learn a function mapping $\mathcal{X}$ to $Y$ where the function is parameterized by a conventional neural network and a pixel classifier.
3 Methodology

3.1 Overview

The proposed model is illustrated in Fig. 2. The model consists of two convolutional neural networks, one for the classification of images and another for the classification of pixels. In the rest of the paper, the two CNNs are referred to as image CNN and pixel CNN, respectively. There exist many powerful image CNNs such as VGG [13] and ResNet [14], and hence we directly employ an existing one. The novelty of the proposed model lies in the integration of the pixel CNN into the image CNN as a global spatial attention. The pixel CNN is carefully designed and described in the following subsection. The model takes two representations from a structured image dataset as input, i.e., the image representation $X$ and the pixel representation $P$. The pixel representation is fed to the pixel CNN to obtain the attention weight map. In particular, the pixel CNN produces the importance values of pixels which range from zero to one. The value of one and zero denote the highest importance and the lowest importance, respectively. The image representation multiplied by the importance values of pixels is fed to the image CNN as input.

3.2 Pixel CNN

The pixel CNN functions as a binary classifier. Since the features of a pixel are well-defined, many existing binary classifiers are applicable, such as the logistic regression and the multilayer perceptron. But...
instead of only the features of the pixel itself, we propose to further consider the features of surrounding pixels as its auxiliary features because nearby pixels have dependencies. Moreover, considering surrounding pixels allows for small transformations of the objects in the images and the small transformations actually may be inevitable even in the structured images. Therefore, we propose to employ a CNN which is designed to capture spatial dependencies among pixels.

But different from the image CNN which can take a batch of images as input at each time, the pixel CNN should take all the pixels simultaneously as input because the importance of all the pixels should be simultaneously considered in the image CNN. The architecture of the proposed pixel CNN is illustrated in Fig. 3. To accord with the convention of the convolution operation, the dimension of the input is denoted as $1 \times NC \times W \times H$. The pixel CNN only has two convolutional layers, and works well in the experiments. Moreover, the experiments showed that the increase of hidden layers does not consistently bring improvement in performance.

In the first convolutional layer, the design of the kernel size is to take into account surrounding pixels as auxiliary features. The stride and padding are designed to keep the number of pixels the same in the output of the convolutional layer. As a result, the first convolutional layer makes abstract features of each pixel out of its raw features and surrounding pixels’ raw features. The second convolutional layer is designed to generate the importance value of each pixel by only looking at its own abstract features, which is realized by a kernel of size $1 \times 1$.

### 3.3 Cost Function

The cost function can be obtained as follows:

$$
\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \text{loss}(F(\mathcal{X}_i \ast M(P; \theta_M); \theta_F), Y_i) + \lambda |M(P; \theta_M)|, \quad (1)
$$

where loss($\cdot$) is the cross entropy loss function, $F(\cdot; \theta_F)$ and $M(\cdot; \theta_M)$ are classification functions parameterized by the image CNN and the pixel CNN, respectively, $\ast$ realizes the element-wise multiplication with broadcasting of image pixels and corresponding pixel importance values, $\theta_F$ and $\theta_M$ are parameters of the image CNN and the pixel CNN, respectively. $\mathcal{X}_i$ is the $i$-th image, and $Y_i$ is the corresponding class label. The L1 regularization on the output of the pixel CNN is to enforce pixel selection, and works well as demonstrated in the experiments. $\lambda \in \mathbb{R}$ is the coefficient.
Algorithm 1: The optimization procedure

\[ \text{Input : } X, Y, P, K, \lambda \text{ and } E \]

1. Initialize a pre-trained CNN which is existing for image classification;
2. for epoch = 1, 2, 3, ..., E do
3. \hspace{1em} Optimize the image CNN and the pixel CNN;
4. for epoch = E+1, E+2, E+3, ... do
5. \hspace{1em} Fix the pixel CNN, and optimize the image CNN only;

3.4 Optimization

Both the image CNN and the pixel CNN can be solved by backpropagation. And by regarding the pixel CNN as just another layer of the image CNN, all the optimization algorithms developed for CNNs can be directly utilized to solve the proposed model. We present the optimization procedure in Algorithm 1.

There are mainly three hyper-parameters introduced by the proposed algorithm. \( K \) is the number of channels in the hidden layer, \( \lambda \in \mathbb{R} \) is the coefficient of the L1 regularization on the output of the pixel CNN, and \( E \) is a cut-off training epoch after which the pixel CNN is fixed. The insight behind the cut-off epoch is that after important pixels are selected, the image CNN should be fine-tuned on the important pixels. In fact, the idea of cut-off epoch is borrowed from pruning low-weight connections [15] in deep neural networks to obtain efficient networks. In particular, the optimization procedure in the connection pruning starts with normal training, and then prunes low-weight connections among neurons, and finally fine-tunes the network on the remaining connections. Here, we firstly learn the important pixels, and fine-tune the image CNN on the learned important pixels. Note that we do not set the small importance values of pixels to zeros because it is non-trivial to determine the threshold. The cut-off epoch may depend on datasets, but works well with a small number, e.g., 60, in the experiments.

4 Empirical Evaluation

4.1 Datasets

We studied three datasets including two medical datasets and one facial expression dataset, which are described as follows:

- **Glaucoma** [16]: Glaucoma is an eye disease, and is the second leading cause of blindness over the world. We studied a dataset consisting of 86 eyes of 43 normative subjects and 505 eyes of 304 patients with open glaucoma (OAG). Each of the eyes belongs to one of three severity levels of the disease according the mean deviation (MD) of visual field sensitivity (Humphrey 10-2 test), i.e., early phase (MD > -6 dB), moderate phase (-12 dB < MD < -6 db), and serious phase (MD < -12 dB). The images are retinal layers thickness produced by optical coherence tomography (OCT). The images were obtained from Tokyo University Hospital, Osaka University Hospital, Hospital of Kyoto Prefectural University of Medicine, Oike-Ikeda Eye Clinic, Shimane University Hospital and Hiroshima Memorial Hospital.
Table 1: Statistics of datasets.

| Name   | # Data Points | # Classes          |
|--------|---------------|--------------------|
| Glaucoma | 591           | 3                  |
| IDRiD   | 516           | 3 (diabetic retinopathy) and 5 (macular edema) |
| JAFFE   | 213           | 6                  |

• **IDRiD** [17]: This dataset consists of retinal fundus photographs that may have diabetic retinopathy lesions. There are two tasks corresponding to the severity level of diabetic retinopathy and diabetic macular edema, respectively. The images have $4288 \times 2848$ pixels, and have background pixels on both the left side and the right side. After removing the background pixels, we resize them into $224 \times 224$. Moreover, we horizontally flipped the images of the right eyes to reconcile the horizontal symmetry. This dataset is publicly available at [https://idrid.grand-challenge.org/Data/](https://idrid.grand-challenge.org/Data/).

• **The Japanese Female Facial Expression (JAFFE) Database** [18]: The dataset contains 213 images of 6 facial expressions posed by 10 Japanese female models. Each image has been rated on 6 emotion adjectives by 60 Japanese subjects such that each image has an averaged value for each kind of expression. We assign the emotion with the largest value as the label to each image. Original images are $256 \times 256$ gray level and were resized into $224 \times 224$ in the experiments. The images are publicly available at [https://zenodo.org/record/3451524#.XyI_6fj7Ts0](https://zenodo.org/record/3451524#.XyI_6fj7Ts0).

The statistics of the three datasets are summarized in Table 1, and we have presented three samples from each of the datasets in Fig. 1. Images from either of the first two medical datasets were taken on human eyes with the same kind of devices such that they fit in well with the definition of structured images. For the images from the facial expression dataset, even though they may not strictly accord with the definition, it is not difficult to see that different images share similar structural content at most pixels.

The number of images of each dataset is not large, which is common in the medical field. It is obvious that for the facial images, there are many pixels irrelevant to the classification task because the task is about facial expressions. For the medical images, after learning from the medical field, we figured out that not all the pixels are relevant to the diseases as well. But the knowledge about the explicit boundary between relevant pixels and irrelevant pixels for each dataset was not provided. Therefore, CNNs may get overfitted on these image datasets with irrelevant pixels, and the proposed global attention method could be useful for alleviating the overfitting problem.

### 4.2 CNN Models for Image Classification

We studied four popular CNN models which are **VGG-16** [13], **GoogleNet** [19], **ResNet-152** [14], and **DenseNet-161** [20]. The four models mark four different types of powerful CNN architectures. The success of VGG is based on very small (3x3) convolution filters. Unlike VGG in which there is only one convolution with a fixed kernel size from one layer to its consecutive layer, GoogleNet utilizes multiple convolutions with different sizes. Before ResNet, many very deep neural networks suffer the problem of degradation of performance as the number of layers increases. To solve this problem, ResNet has a novel
deep residual learning framework which creates an identity connection from one layer to the next layer. As a result, ResNet-152 can have as many as 152 layers and enjoys performance boosting. DenseNet further encourages the connections from early layers to later layers to strengthen feature propagation and to promote feature reuse. These CNN models are now popular building blocks for many visual tasks. We studied these different CNN models to demonstrate that the proposed global attention mechanism is a generic solution, and is beneficial for different CNN architectures in the image classification tasks.

4.3 Baselines

The experiments were mainly conducted to compare the performance of the CNN models with and without the proposed global attention mechanism. Since local attention mechanisms in CNNs are also available, we employed VGG with attention mechanism at the last three levels (VGG-att3) proposed in [8] and residual attention network with 92 trunk layers (ResAttNet-92) proposed in [7] as baselines.

When the attention is regarded as a pixel selection method, it is similar to the embedded category of feature selection methods [10, 11, 12]. Therefore, we further compare the performance of a CNN with the attention and the same CNN with the L1 regularization. In particular, we multiply each pixel by one as the initialization weight, and enforce L1 regularization on the pixel weights.

Moreover, because of the nature of the structured images, we studied the performance of two non-deep models, which are logistic regression (LR) and support vector machine (SVM). We did not compare other methods making hand-crafted features because feature engineering is not needed for deep learning, e.g., the direct application of CNN on the raw retinal fundus photographs has been widely recognized [21].

4.4 Implementation

For the LR and SVM, we employed the functions provided by scikit-learn. For the CNNs, we used pre-trained CNNs provided by Pytorch. For the hyper-parameters of all baseline models, we performed grid-search on their appropriate ranges. For optimizing the CNNs, we employed the Adam optimizer, and set the mini-batch size as 32, the learning rate as 0.0001, performed the search for the weight decay within the range \{0.001, 0.0001, 0.00001\}. For the CNNs with L1 regularization, the coefficient was searched over \{100000, 10000, 1000, 100, 10, 1\} which contains large values because the regularization loss on each pixel was averaged by the number of pixels. For the proposed models, the number of channels \(K\) was searched over \{32, 62, 128, 256\}, and the coefficient of the regularization \(\lambda\) on the outputs of the pixel CNN was searched over \{0.1, 0.01, 0.001, 0.0001\}, and the cut-off epoch \(E\) was set as 60. The total number of training epochs for all CNNs was set as 250. An example of implementation is available at https://drive.google.com/drive/folders/1aNgm1XP-Xu4gsS92fmsKTxx814BoAx-j?usp=sharing. The algorithm was implemented with Pytorch V1.3.1 and was executed on a GPU of GeForce RTX 2080 Ti.

\footnote{https://scikit-learn.org/stable/}
Table 2: Results of the image classification tasks. Bold numbers for each CNN indicate the existence of statistically significant difference between the vanilla CNN and it with global attention, and * is for the difference between L1 and global attention.

| Accuracy(standard deviation) | Glaucoma | IDRI D | Retinopathy | Macular Edema | JAFFE |
|-----------------------------|----------|--------|-------------|---------------|-------|
| LR                          | 74.28    | 39.81  | 66.99       | 74.06         |       |
| SVM                         | 74.45    | 45.63  | 64.08       | 75.00         |       |
| VGG-att3                    | 79.63(3.09) | 47.25(4.45) | 78.86(2.10) | 74.26(1.92) |       |
| ResAttNet-92                | 77.17(2.33) | 41.81(5.07) | 67.15(3.98) | 70.78(6.04) |       |
| GoogleNet                   | 81.06(1.18) | 52.46(4.38) | 80.49(1.90) | 78.94(2.56) |       |
| GoogleNet (with L1)         | 81.93(2.27) | 54.72(2.66) | 78.60(2.14) | 82.02(2.66) |       |
| GoogleNet (with global attention) | *83.61(2.43) | *57.21(3.20) | *81.36(3.05) | 82.64(2.57) |       |
| VGG-16                      | 82.38(1.30) | 56.96(4.63) | 80.23(2.32) | 81.78(8.44) |       |
| VGG-16 (with L1)            | 82.71(3.02) | 57.99(5.41) | 78.93(1.02) | 84.88(2.00) |       |
| VGG-16 (with global attention) | *84.22(1.26) | *66.41(1.36) | *83.59(0.47) | *84.26(6.19) |       |
| ResNet-152                  | 82.16(1.87) | 56.78(4.98) | 79.06(1.97) | 77.83(4.35) |       |
| ResNet-152 (with L1)        | 83.41(1.10) | 60.91(2.31) | 79.54(1.39) | 80.00(3.79) |       |
| ResNet-152 (with global attention) | *84.37(0.98) | *66.17(2.16) | *82.13(0.70) | *80.39(3.85) |       |
| DenseNet-161                | 81.06(1.44) | 59.84(3.80) | 80.12(1.27) | 79.39(3.42) |       |
| DenseNet-161 (with L1)      | 80.08(3.18) | 60.58(1.80) | 81.07(2.56) | 82.95(3.53) |       |
| DenseNet-161 (with global attention) | *82.27(1.88) | *62.88(4.40) | *82.07(2.71) | *84.88(3.04) |       |

4.5 Evaluation Metric

For evaluating the classification performance, we employed the metric accuracy defined as follows:

\[
\text{Accuracy} = \frac{\text{Number of correctly predicted images}}{\text{Total number of images}} \times 100\% \quad (2)
\]

4.6 Experimental Results

The two medical datasets were provided with 80% training instances and 20% test instances. Similarly, we separated the facial expression dataset into 80% and 20% by random. The classification accuracy is presented in Table 2. Since early stopping was adopted for training all the CNNs models, we designed a method based on an empirical observation to choose the epoch for measuring the performance. In particular, we performed five-fold cross validation on the training instances only, and obtained the 30 epochs which were associated with the highest accuracy as the epochs for the measurement of performance on...
the test instances. As a result, the resulting accuracy on the test instances by a CNN is the mean of 30 measurements with standard deviation.

From Table 2, we can see that all the CNNs with the proposed global attention consistently outperformed the corresponding vanilla CNNs. We conducted the t-test between each CNN with the global attention and the vanilla CNN, and obtained the p-values less than 0.01. The reason behind the superior performance was because larger importance values were assigned to relevant pixels than to irrelevant pixels. We visualize the weight map in the following subsection. The global attention significantly outperformed the L1 regularization at most times. The exceptions happened only on the JAFFE dataset. Note that it is not difficult to distinguish the relevant pixels from the irrelevant pixels as illustrated in the face images in Fig. 1. Therefore, it may not be difficult to identify the relevant pixels even for the L1 regularization.

The poorer performance of VGG-att3 and ResAttNet compared with the vanilla VGG and ResNet (we used the pretrained version provided by Pytorch) may be due to two reasons. First, VGG-att3 and ResAttNet cannot enjoy the benefits brought by pre-training since their attention layers should be data-dependent. Second, we examined that the important pixels learned by VGG-att3 and ResAttNet for one image are always different from that of another, and figured out that the difference is due to the difference in the pixel intensities of raw images despite being structured. But in fact, all the images of a dataset should share important pixels. Therefore, the additional local attention layers may further aggravate overfitting.

4.7 Attention Visualization

In this subsection, we visualized the attention weight map, i.e., the output of the pixel CNN for each dataset, and only present the representative result of one CNN for each dataset. The pixel importance values in the weight map range from zero to one. In the following visualizations, darker colors denote larger values. In fact, the smallest importance values were not always zero and the largest values were not always one. In the visualizations, the values were normalized into the range from zero to one.

For the JAFFE dataset, we can easily tell that the ground-truth attention should be potentially paid on human faces since it is about facial expressions. The initial attention by a random initialization of the pixel CNN and the final attention after training are presented in Fig. 4. It is very surprising that the initial
attention appear to be consistent with the potential attention, which may be because the pixel CNN is effective at capturing spatial dependencies and common characteristics among pixels. After the learning, the attention got closer to the potential ground-truth attention.

For the IDRiD dataset, studies [22] showed that the potential attention may be a circular region centered at fovea and with one optic disc diameter as illustrated in Fig. 5 (b). By carefully examining the learned attention in Fig. 5 (a), we are able to see that there exists such a circular region out of which the color is usually lighter.

For the glaucoma dataset, studies [23, 24] showed that the potential attention locates in the region of black dots as illustrated in Fig. 6 (b). The dots in black are visual field (VF) test points, and the mean deviation of the test points were employed to determine the categories of the images, i.e., the severity level of glaucoma. In Fig. 6 (a), we are able to see that attended regions locate in the VF test points.

As a conclusion, the proposed method can learn meaningful attended regions, which is beneficial for understanding the content of images and may be relied on to guide the development of the domain knowledge.

4.8 Complexity Analysis

The complexity of the proposed model is mainly determined by the number of parameters. The proposed model introduces an additional pixel CNN. The number of parameters in the pixel CNN is $K \times (3 \times$
3 \times N \times C + 1 + (K + 1) \) where \( K \) is the number of channels in the hidden layer, \( N \) is the number of images, and \( C \) is the number of channels of the images. The number of parameters depends on the number of images. For the three studied datasets, the number of images is not large. As a result, the number of parameters is very small compared with that of the image CNN. Even for very large datasets, such as millions of images, the number of parameters is at the same level as that of VGG, ResNet and DenseNet.

4.9 Parameter Sensitivity

We studied the sensitivity of the proposed model with respect to the hyper-parameters introduced by the proposed global attention mechanism, i.e., the number of channels \( K \), the coefficient of the L1 regularization on the output of the pixel CNN \( \lambda \), and the cut-off epoch \( E \). We present the results for DenseNet in Fig. 7. In the experiments, the value of a hyper-parameter was varied within an appropriate range while the other two hyper-parameters were fixed. The results show that the proposed model is not much sensitive to the three hyper-parameters within the studied range. But the coefficient for the L1 regularization should not be large (e.g., 10).

4.10 Additional Observations on the Pixel CNN

There actually exist many choices for designing the architecture of the pixel CNN thanks to the various architectures developed for the CNNs, and we have tried some of them. But all the other designs did not consistently bring improvement compared with the design proposed in this paper. Here, we list the studied choices as follows:

- Kernel size: The kernel size in convolution determines how many surrounding pixels are considered as auxiliary features of a pixel. To keep the dimension of an image the same after the convolution, the kernel size should be odd numbers and we tried 5 \( \times \) 5 and 7 \( \times \) 7.

- Number of hidden layers: In a CNN, hidden layers are important to learn abstraction of raw features. Moreover, if the kernel size is larger than 1 \( \times \) 1, the range of surrounding pixels as auxiliary features would be expanded as the number of hidden layers increases. We tried two, three, and four.
• Fully convolutional network: Since the last layer is implemented as a convolution layer instead of a fully connected layer, the CNN can regarded as a fully convolutional network [?]. As a result, the kernel size of the last layer can be larger than one, and we tried $3 \times 3$ and $5 \times 5$.

• Dense connections: Dense connections from early layers to later layers were introduced by DenseNet.

We conducted experiments to compare the proposed design with an alternative design with respect to each individual of the four choices. For instance, while comparing the kernel with size of $3 \times 3$ (the proposed) and that with $5 \times 5$, we kept all the other aspects of the architecture as the same as that in the proposed architecture. The results showed that there were no statistical differences at most times between the performance obtained by the proposed design and by the alternative design in terms of each choice. As a result, we recommend the proposed architecture as the first trial if one would like to apply the proposed method. Note that the results may be data-dependent. Moreover, the combination of multiple choices may be beneficial, which is left as future studies.

5 Related Work

5.1 Medical Image Classification with CNNs

The state-of-the-art fashion of medical image classification is to employ deep learning which can perform automatic feature extraction because good features are crucial [2]. There have been many successful applications of deep learning, especially convolutional neural networks (CNNs), and several reviews are available [25, 3, 2]. The typical way of application of CNNs is to utilize existing CNN architectures which are pre-trained on natural images and then to fine-tune the pre-trained CNNs on the images at hand. The pre-training is important because powerful CNNs have excessive number of parameters and medical image datasets are usually small-scale.

However, CNNs may still get overfitted on medical images especially when there exist pixels irrelevant to the classification task, e.g., the task is only related to a particular region within the images. In order to avoid irrelevant pixels, many methods [26, 27, 28, 29, 30, 31] rely on domain knowledge to obtain relevant pixels, and then somehow enforce the CNNs to pay attention to the pixels of interest. This kind of methods works at the data preprocessing stage. But the domain knowledge is not always available, and may be expensive to obtain in some scenarios. Other methods [32, 33] obtain relevant pixels according to characteristics of pixel values, but still work as data preprocessing techniques.

The proposed attention mechanism can be regarded as an end-to-end data-driven approach to learning relevant pixels by estimating their importance values, and training image classification CNNs simultaneously.

5.2 Local Attention in CNNs

Existing local attention mechanisms [5, 6, 7, 8, 9] are also end-to-end data-driven approaches to learning regions of interest, but the regions of interest are image-specific. Instead, in the proposed method, the regions of interest are shared by all images within a dataset. Moreover, existing local attentions mainly work at hidden layers while the proposed global attention mechanism only works at the input layer.
5.3 Feature Selection

The proposed global weight map learning method can be regarded as a pixel selection method, which is similar to the concept of feature selection. Feature selection has been extensively studied to avoid the curse of dimensionality, and has been demonstrated successful on non-deep learning models. The following paragraph briefly introduces existing feature selection methods, and explains why pixel selection for CNNs has been much less studied.

Feature selection methods mainly fall into three categories, filter methods, wrapper methods, and embedded methods [10, 11, 12]. Filter methods are data pre-processing methods, and mainly conduct feature selection by analyzing the importance of features based on their own inherent characteristics. Since the importance of each pixel to the classification task is unknown and CNNs are able to automatically learn high-level abstraction of features, filter methods may not be preferable. Wrapper methods select features based on evaluating the performance of models on subsets of features, and multiple runs of model training are required. Since the training of CNNs is usually computationally expensive, wrapper methods are not feasible in practice. Embedded methods, as the name suggests, feature selection mechanisms are components of the model building. The commonly used embedded methods apply regularization on model parameters, e.g., L1 and L2 regularization. L1 and L2 regularization have already been employed in CNNs. But because of the huge number of parameters in CNNs, CNNs may still easily get overfitted, which explains why other regularization techniques are especially designed for deep learning models, such as dropout [34].

The proposed pixel selection method belongs to the embedded method, but is not another kind of regularization. Moreover, the proposed method is based on the importance of pixels to tasks where the importance is determined by inherent characteristics of a pixel. In this sense, it is similar to filter methods. But filter methods usually employ pre-defined measurements of importance, such as mutual information [35] and Fisher score [36]. The proposed method instead measures the importance via a pure data-driven approach.

6 Conclusion and Future Directions

This paper for the first time proposes a global spatial attention mechanism in convolutional neural networks (CNNs) for structured medical image classification. The attention learning is realized by a binary classifier where the intensities of all images at a pixel are employed as the features of the pixel. The attention is a generic solution and can be integrated into any CNN architectures. With the global attention, the overfitting problem of CNNs can be alleviated. As a result, the experiments showed that all the studied CNNs with the proposed attention significantly outperformed the vanilla CNNs. Moreover, the attended regions are useful for understanding the structural content of the images.

The proposed method is a pure data-driven approach. In fact, for each of the studied medical datasets, the medical field has accumulated years of domain knowledge. In future, we plan to integrate the domain knowledge into the attention mechanism to further improve the performance for a particular medical dataset.
Acknowledgments

This work was partially supported by JST KAKENHI 19140000190 and JST-AIP JPMJCR19U4. Atsushi Nitanda was partially supported by JSPS Kakenhi (19K20337) and JST-PRESTO.

Declarations of interest

The authors declare no competing interests.

References

[1] Brody, H., 2013. Medical imaging. Nature, 502(7473), pp.S81-S81.
[2] Litjens, G. et al. 2017. A survey on deep learning in medical image analysis. Medical image analysis, 42, pp.60-88.
[3] Shen, D., Wu, G. and Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering, 19, pp.221-248.
[4] Bengio, Y., Courville, A. and Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8), pp.1798-1828.
[5] Jaderberg, M., Simonyan, K. and Zisserman, A., 2015. Spatial transformer networks. In Advances in neural information processing systems (pp. 2017-2025).
[6] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A., 2016. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2921-2929).
[7] Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X. and Tang, X., 2017. Residual attention network for image classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
[8] Jetley, S., Lord, N.A., Lee, N. and Torr, P.H., 2018, February. Learn to Pay Attention. In International Conference on Learning Representations.
[9] Schlemper, J. et al., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53, pp.197-207.
[10] Molina, L.C., Belanche, L. and Nebot, ., 2002, December. Feature selection algorithms: A survey and experimental evaluation. In 2002 IEEE International Conference on Data Mining, 2002. Proceedings. (pp. 306-313). IEEE.
[11] Chandrashekar, G. and Sahin, F., 2014. A survey on feature selection methods. Computers & Electrical Engineering, 40(1), pp.16-28.
[12] Tang, J., Alelyani, S. and Liu, H., 2014. Feature selection for classification: A review. Data classification: Algorithms and applications, p.37.
[13] Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[14] He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[15] Han, S., Pool, J., Tran, J. and Dally, W., 2015. Learning both weights and connections for efficient neural network. In Advances in neural information processing systems (pp. 1135-1143).

[16] Uesaka, T. et al., 2017, August. Multi-view Learning over Retinal Thickness and Visual Sensitivity on Glaucomatous Eyes. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 2041-2050).

[17] Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V. and Meriaudeau, F., 2018. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data, 3(3), p.25.

[18] Lyons, M., Akamatsu, S., Kamachi, M. and Gyoba, J., 1998, April. Coding facial expressions with gabor wavelets. In Proceedings Third IEEE international conference on automatic face and gesture recognition (pp. 200-205). IEEE.

[19] Szegedy, C., et al., 2015. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[20] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., 2017. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4700-4708).

[21] Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. and Kim, R., 2016. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. Jama, 316(22), pp.2402-2410.

[22] Deepak, K.S. and Sivaswamy, J., 2011. Automatic assessment of macular edema from color retinal images. IEEE Transactions on medical imaging, 31(3), pp.766-776.

[23] Hood, D.C., Raza, A.S., de Moraes, C.G.V., Liebmann, J.M. and Ritch, R., 2013. Glaucomatous damage of the macula. Progress in retinal and eye research, 32, pp.1-21.

[24] Fujino, Y., et al., 2018. Mapping the Central 10 degree Visual Field to the Optic Nerve Head Using the Structure-Function Relationship. Investigative ophthalmology & visual science, 59(7), pp.2801-2807.

[25] Ravi, D. et al. 2016. Deep learning for health informatics. IEEE journal of biomedical and health informatics, 21(1), pp.4-21.

[26] Chen, X., Xu, Y., Wong, D.W.K., Wong, T.Y. and Liu, J., 2015, August. Glaucoma detection based on deep convolutional neural network. In 2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 715-718). IEEE.

[27] Nie, D., Zhang, H., Adeli, E., Liu, L. and Shen, D., 2016, October. 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. In International conference on medical image computing and computer-assisted intervention (pp. 212-220). Springer, Cham.

[28] Eppel, S., 2017. Setting an attention region for convolutional neural networks using region selective features, for recognition of materials within glass vessels. arXiv preprint arXiv:1708.08711.

[29] Paeng, K., Hwang, S., Park, S. and Kim, M., 2017. A unified framework for tumor proliferation score prediction in breast histopathology. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support (pp. 231-239). Springer, Cham.
[30] Ni, X., Yan, Z., Wu, T., Fan, J. and Chen, C., 2018, September. A region-of-interest-reweight 3D convolutional neural network for the analytics of brain information processing. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 302-310). Springer, Cham.

[31] Zhang, M., Li, W. and Du, Q., 2018. Diverse region-based CNN for hyperspectral image classification. IEEE Transactions on Image Processing, 27(6), pp.2623-2634.

[32] Gao, X., Lin, S. and Wong, T.Y., 2015. Automatic feature learning to grade nuclear cataracts based on deep learning. IEEE Transactions on Biomedical Engineering, 62(11), pp.2693-2701.

[33] Mahapatra, D., Roy, P.K., Sedai, S. and Garnavi, R., 2016, October. Retinal image quality classification using saliency maps and CNNs. In International Workshop on Machine Learning in Medical Imaging (pp. 172-179). Springer, Cham.

[34] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), pp.1929-1958.

[35] Estvez, P.A., Tesmer, M., Perez, C.A. and Zurada, J.M., 2009. Normalized mutual information feature selection. IEEE Transactions on neural networks, 20(2), pp.189-201.

[36] Gu, Q., Li, Z. and Han, J., 2011, September. Generalized fisher score for feature selection. In 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011.
Supplementary Information
A Novel Global Spatial Attention Mechanism in Convolutional Neural Network for Medical Image Classification

Linchuan Xu\textsuperscript{1}, Jun Huang\textsuperscript{2}, Atsushi Nitanda\textsuperscript{2}, Ryo Asaoka\textsuperscript{3}, Kenji Yamanishi\textsuperscript{2}

\textsuperscript{1}PQ813, Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
linch.xu@polyu.edu.hk

\textsuperscript{2}Graduate School of Information Science and Technology
The University of Tokyo, Tokyo, Japan
\{jun.huang, nitanda, yamanishi\}@mist.i.u-tokyo.ac.jp

\textsuperscript{3}Department of Ophthalmology
The University of Tokyo, Tokyo, Japan
rasaoka-ty@umin.ac.jp

Section 1 presents data preprocessing. Section 2 presents the implementation.

1 Preprocessing of Datasets

1.1 IDRiD

This dataset is publicly available at https://idrid.grand-challenge.org/Data/. For each image, background pixels on both the left side (pixel index on the horizontal axis less than 260) and the right side (pixel index larger than 3685) were removed. Afterwards, we resized them into $224 \times 224$ using interpolation = \texttt{cv2.INTERAREA} in Python. Moreover, we horizontally flipped the images of the right eyes to reconcile the horizontal symmetry. For information, the set of flipped images in the training data by index is $\{0, 6, 8, 9, 12, 13, 16, 17, 19, 21, 22, 24, 26, 28, 30, 32, 34, 36, 38, 40, 45, 46, 48, 51, 54, 56, 58, 60, 61, 63, 64, 65, 66, 68, 70, 73, 75, 77, 79, 81, 84, 86, 89, 90, 93, 94, 98, 99, 100, 101, 102, 104, 106, 109, 111, 115, 117, 118, 119, 123, 124, 127, 128, 131, 134, 135, 137, 139, 141, 144, 145, 147, 149, 151, 152, 156, 157, 158, 159, 160, 163, 166, 168, 170, 171, 172, 174, 176, 177, 181, 183, 184, 186, 187, 188, 190, 192, 193, 194, 195, 197, 200, 201, 202, 204, 206, 209, 211, 212, 214, 216, 217, 220, 222, 226, 230, 234, 235, 238, 241, 244, 246, 248, 251, 252, 253, 256, 258, 259, 261, 264, 268, 270, 272, 273, 274, 275, 276, 279, 282, 284, 286, 288, 289, 291, 292, 299, 300, 302, 304, 305, 306, 307, 309, 312, 314, 315, 318, 320, 321, 322, 325, 329, 331, 333, 334, 337, 340, 341, 343, 345, 349, 351, 356, 357, 359, 361, 362, 367, 369, 370, 372, 373, 375, 377, 378, 381, 383, 384, 387, 389, 392, 394, 396, 398, 402,
403, 405, 406, 407, 408, 409, 411, 412}, and the set in the test data is {1, 3, 4, 5, 8, 9, 14, 15, 16, 19, 21, 24, 26, 27, 31, 34, 35, 36, 38, 39, 41, 42, 45, 48, 50, 53, 55, 57, 60, 62, 63, 64, 66, 67, 69, 71, 76, 81, 82, 84, 86, 87, 90, 94, 96, 100, 102}. Finally, the pixel values were normalized into range from 0 to 1, and then were standardized using means and standard deviations in the set {0.485 and 0.229 for the first channel, 0.456 and 0.224 for the second channel, 0.406 and 0.225 for the third channel}.

1.2 JAFFE

The images are publicly available at https://zenodo.org/record/3451524#.XyI_6fj7Ts0. There are six values corresponding to six emotions for each image, and we assigned the emotion with the largest value as the label to each image. Each image was resized into 224 × 224 using interpolation = cv2.INTER_AREA in Python. The normalization and standardization of the pixel values were performed in the same way as that for the IDRiD data.

1.3 Glaucoma

The preprocessing for the Glaucoma data is basically similar to that for the other two datasets. More details are omitted due to review policies and privacy issues.

2 Implementation

2.1 The Proposed Model

An example of the implementation of the proposed model is provided through Google Drive at https://drive.google.com/drive/folders/1aNgmlXP-Xu4gsS92fmsKTxx8l4BoAx-j?usp=sharing.

2.2 Image CNNs

The image CNNs, e.g., GoogleNet, VGG-16, ResNet-152, and DenseNet-161, are pretrained models provided by Pytorch, which were instantiated from torchvision.models.

2.3 CNNs with Attention

For VGG-att3, we used the implementation publicly available at https://github.com/SaoYan/LearnToPayAttention. For ResAttNet-92, we used the implementation publicly available at https://github.com/osmr/imgclsmob.

2.4 SVM and LR

We used the models implemented in sklearn https://scikit-learn.org/stable/.