SagaNet: A Small Sample Gated Network for Pediatric Cancer Diagnosis

Yuhan Liu [1] **Shiliang Sun** [1]

Abstract

The scarcity of available samples and the high annotation cost of medical data cause a bottleneck in many digital diagnosis tasks based on deep learning. This problem is especially severe in pediatric tumor tasks, due to the small population base of children and high sample diversity caused by the high metastasis rate of related tumors. Targeted research on pediatric tumors is urgently needed but lacks sufficient attention. In this work, we propose a novel model to solve the diagnosis task of small round blue cell tumors (SRBCTs). To solve the problem of high noise and high diversity in the small sample scenario, the model is constrained to pay attention to the valid areas in the pathological image with a masking mechanism, and a length-aware loss is proposed to improve the tolerance to feature diversity. We evaluate this framework on a challenging small sample SRBCTs dataset, whose classification is difficult even for professional pathologists. The proposed model shows the best performance compared with state-of-the-art deep models and generalization on another pathological dataset, which illustrates the potentiality of deep learning applications in difficult small sample medical tasks.

1. Introduction

Deep learning-based methods have enjoyed marvelous success over a variety of tasks in recent years, and their application in the medical field also illustrates excellent potentiality. The outstanding performance of deep models heavily depends on the availability of vast training databases. Nevertheless, this is a considerable obstacle when tackling medical data due to several reasons. First, different from labeling natural images or language with the help of ordinary volunteers, annotating medical data requires experts to devote a lot of time and endeavors during busy work. Second, for the protection of data privacy, many hospitals or institutions are unwilling or not allowed to share data about specific diseases. Third, although some diseases have high mortality and are worthy of attention, the affected patients have a small population base and incidence. Under such difficulties, designing effective deep models becomes a challenge that researchers focusing on medical data processing have to face.

Strong dependence on large datasets makes deep learning challenging to be applied to diseases whose cases are scarce. As the experimental analysis in (Campanella et al., 2019) concludes, when the number of pathological slices is less than 10,000, the performance of deep models is often unsatisfactory. This problem is especially severe in many pediatric diseases. The small round blue cell tumors (SRBCTs) (Sharma et al., 2017) we studied in this work are one kind of such diseases due to their small number and diverse shapes.

The reason for the common small sample problem in pediatric pathology research is that pediatric cancer data are very difficult to collect. Compared with adults, the cancer incidence is lower in children. For example, the most common tumor in adults is hepatocellular carcinoma, which has an incidence rate of about 40 per million (Di Bisceglie et al., 1988), while the hepatoblastoma in children, the most common pediatric liver malignancy, is only 1.5 per million on incidence (Tang et al., 2011). Coupled with the considerable difference in the population base, the number of pathological tumor slices in children is significantly lower than the number of adult tumor slices, which is a huge challenge for deep learning. At the same time, the severe shortage of pediatric pathologists and expensive time cost bring more obstacles.

Faced with such a pediatric cancer research problem with small data size, a simple idea is to directly use a model from adult cancer pathology research for transfer learning. However, compared with adult tumors, children tumors are very different in terms of tumor origin, cell source, frequent location, and disease spectrum (Ma et al., 2018)(Marshall et al., 2014), which in turn leads to significant differences in cell morphology on pathological slices. In most cases, models suitable for adult pathology tasks cannot be directly

[1]School of Computer Science and Technology, East China Normal University, Shanghai, China. Correspondence to: Shiliang Sun <slsun@cs.ecnu.edu.cn>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

applied to pediatric cancers.

In this work, we present a novel small sample gated network, named SagaNet, for the pathological image classification task of SRBCTs. The problem we face is that challenging small datasets usually have much noise and high diversity, and are difficult to distinguish visually even for experts. In terms of probability, 50% diagnostic accuracy is an average performance of pathologists. In the actual diagnosis process, ancillary techniques such as immunohistochemical markers are often used (Shimada et al., 1999) (Barden & Lewey, 1949). However, when performing immunohistochemistry, a definite diagnosis can still be impossible due to insufficient available slices, considering that making slice samples from living children via surgeries is a costly and risky task. Thus, we hope that the classification task can be performed directly from the pathological slides through deep learning, and a reasonable reference can be provided before the pathologist makes the immunohistochemistry.

When data are insufficient, the first problem to be solved is removing image noise caused by various cell tissues or impurities. Therefore, we propose to generate a mask for the input pathological image to roughly shield the noise area and design the gating mechanism cooperating with a partial reconstruction loss to force the network to focus on the meaningful tissue regions. Another problem is the diversity of feature patterns themselves. If massive training data are available, the model can be gradually taught to assimilate various feature patterns of the same category. For the current small sample problem, we redesigned the final classification loss, utilizing the length of features as the basis for classification. Feature patterns of the same length can be very different, which improves the model's tolerance for output feature diversity.

Our model shows outstanding performance, which is much higher than human attempts, and illustrates dramatic superiority compared with state-of-the-art deep models. In addition, we also conducted training and verification on another small-scale pathology dataset. The experimental results far exceeded the performance of baseline models on this dataset, confirming the generalization of SagaNet. This work provides a promising prospect for the auxiliary diagnosis of pediatric tumors and starts a good exploration for further research on small sample pathological problems with deep learning.

2. Overview of Related Work

At present, most medical research based on deep learning has focused on some mainstream adult diseases, and childhood diseases have not received much attention. These studies generally use neural networks to detect or segment the corresponding objects in the disease or directly make classification diagnoses.

In some research problems of high incidence adult diseases, many studies have utilized deep learning to make relevant explorations and obtained good achievements. Chen et al. (2017) proposed a deep contour-aware network to automatically detect and segment histological objects by combining multi-layer contextual features. They added three weighted auxiliary classifiers into the network to alleviate gradient vanishing, and the experimental results demonstrated the superior performance of their method. In the same year, Harrison et al. (2017) applied a fully connected convolutional network in the task of pathological lung segmentation through merging output maps from different layers, finally obtaining a detailed mask as the result.

In the work of Titano et al. (2018), a 3D convolutional neural network architecture was demonstrated to detect acute neurologic events on head CT images by performing weakly-supervised classification. Their method shows excellent performance and greatly reduces diagnosis time. A pathological descriptor was proposed in (Niu et al., 2019) to describe the position and quantity of lesions in diabetic retinopathy. They trained a generative adversarial network (GAN) to synthesize the corresponding retinal image of the given descriptor and a binary vessel segmentation, and this method improved the interpretability of deep learning on medical tasks. Campanella et al. (2019) presented a deep learning model based on multi-instance learning through weakly-supervised training on a massive amount of whole slide images from several different cancers and obtained accurate results on several common tumors. These research efforts are mainly focused on mainstream diseases because it is easier to obtain a large amount of medical data on these disease tasks.

Compared with numerous research works in the field of adult medical care, there are far fewer works related to child medical care (Shu et al., 2019). Larson et al. (2017) estimates the skeletal maturity of children through deep neural networks, achieving similar performance compared with experts. In the work of Lakhani & Sundaram (2017), an ensemble of AlexNet (Krizhevsky et al., 2012) and GoogLeNet (Szegedy et al., 2015) was utilized to detecting tuberculosis in chest radiographs. Tabrizi et al. (2018) combined a deep neural network and a weighted fuzzy active shape model in pediatric 3DUS images to automatically segment kidneys with various shapes, sizes, and texture characteristics. Through training on images of entire hands and specific parts of hands with deep learning, Iglovikov et al. (2018) successfully automatically estimated the pediatric skeletal bone age with high accuracy.

In pediatric cancer research works discussed above, datasets are small but with a relatively stable positional relationship with images, which is different from highly diverse patho-

logical images in our dataset. The SRBCTs classification task in our work is more challenging to solve.

3. SagaNet

When a dataset is small, the two problems of high noise and high diversity will be particularly prominent, which will greatly reduce the performance of the model on the dataset. For these two problems, we proposed two sets of schemes: mask filtering and diversity tolerance respectively, and integrated them into the same network model, named SagaNet.

In this section, we will elaborate on the ideas and implementation strategies of the two sets of schemes and show the details of each module of the corresponding scheme in subsections.

3.1. Noise Filtering

Massive image noises are inevitable in pathological images of SRBCTs due to several factors. The first one is SRBCTs can happen in many parts of a child's body, such as the abdomen, neck, chest, and pelvis, and spread to other parts of the body with high probability (Janoueix-Lerosey et al., 2010) (Brereton et al., 1975). This makes the form of tissue in slices highly varies with different sample positions, sometimes accompanied by muscles, red blood cells, or vascular cavity, and sometimes not. The second one is that the blade may deform a small part of the tissue when the doctor cuts the sample to make slices, and some operations may cause tissue breaks or air bubbles in the slice. Last but not least, because the samples of SRBCTs are rare, the sample collection covers a long-time span (usually 3 to 10 years). Some slices may discolor due to improper or long-time storage. As the number of available training samples is limited, the problem of noise may significantly degrade model performance.

In order to shield this disturbing classification information, we propose a set of architecture to force the network to ignore the noise area in the pathological image. This mechanism is composed of image masks, gating mechanism, and a partial reconstruction loss. First, we unsupervisely generate masks for pathological images to shield areas such as cavities, blood clots, etc. These masks are used as the input of the network together with pathological images. Then, we equip the network with gating layers (Dauphin et al., 2017), which can dynamically determine whether to spread the information to the next layer with some probability, and thus selectively filter out the information shielded by the mask. The objective function which guides this information filtering process is a partial reconstruction loss with the mask. Under such a design, the network is forced to pay attention to the unmasked image regions and focus on useful potential discriminant information, leading to the improvement of the classification performance with less noise.

3.1.1. Mask Generating

We consider constructing masks of these pathological images by exploiting the visual attributes of stained slices themselves. The mask results are gained via an unsupervised segmentation algorithm inspired by (Kanezaki, 2018). Different from their work, we propose a recognition mechanism for the valid area after the initial training process and change the architecture of convolutional networks to make the optimization process more smooth. Non-linear layers are placed in front of batch normalization layers in the original work, which is found to make optimization unstable in our experiments. Therefore, we adjust the structure and further improve the architecture to be four convolutional layers with the squeeze-and-excitation mechanism as SENet (Hu et al., 2018).

First, an initial segmentation result of the input image $X = \{x_n\}^N$ with N pixel values is gained via traditional clustering like SLIC (Achanta et al., 2010). We record a pixel location set of every unique segmentation label, namely $A = \{A_{y_k}\}^K$, where y_k is the kth label. Then, during each epoch, a convolutional network accepts the image as an input and output probabilities for segmentation labels at each pixel, namely $P = \{\mathbf{p}_n\}^N$. A target image $T = \{t_n\}^N$ is constructed by marking each pixel with the most probable label, namely $l_k = argmax(|l_n|_{n \in A_{y_k}})$, where $|l_n|$ is the count number of the label l_n. Afterwards, each area in T corresponding to the segmented area in A will be marked to the most frequently occurring label. The backpropagation process is performed by minimizing the cross-entropy loss between P and T, until the number of segmented areas S reaches the minimum label number u. The pseudo code of the specific process is shown in Algorithm 1.

After obtaining a coarse result by applying the steps above, we need to merge valid parts and mark the collection of remaining parts as the mask of this image by a special judgment mechanism. To allow pathologists to clearly see where the cells are, slices are often stained by H&E (Titford, 2005). Thus, we can utilize color and texture characteristics to detect valid areas. We use T as an abbreviation for threshold. The first one is that the average value of (R_m, G_m, B_m) must be within $[T_{low}, T_{high}]$, where m means the mean value of the corresponding area, such that dark points and blank areas can be excluded. The second one is the $(R_m + B_m)/2 - G_m \leq T_{red}$, removing red regions. The last one is $mean(R_s, G_s, B_s) \geq T_{smooth}$, where s is the standard variance of the color channel in that field so that only rough-textured areas (cell regions) will be considered. Some examples of generated masks are shown

Algorithm 1 Mask Construction

Input: image $X = \{x_n\}^N$
Output: mask M

$\{A_{y_k}\}^K = Slic(X)$
SENet.parameters = Init()

repeat
  $\{p_n\}^N = SENet(X)$
  $\{t_n\}^N = \{argmax(\mathbf{p}_n)\}^N$
  for $k = 1$ **to** K **do**
    $l_k = argmax(|l_n|_{n \in A_{y_k}})$
    $t_n^{new} = l_k$ for $n \in A_{y_k}$
    $Loss = Cross\text{-}entropy(x_n, t_n)$
    SENet.SGD(Loss)
  end for
  $S = unique(\{l_k\})$
until $S \leq u$

for $s = 1$ **to** S **do**
  $mean_s = mean(RGB_{A_s})$
  $std_s = std(RGB_{A_s})$
  if $ValidOrNot(mean_s, std_s)$ is $true$ **then**
    Merge V with A_s
    Update M
  end if
end for

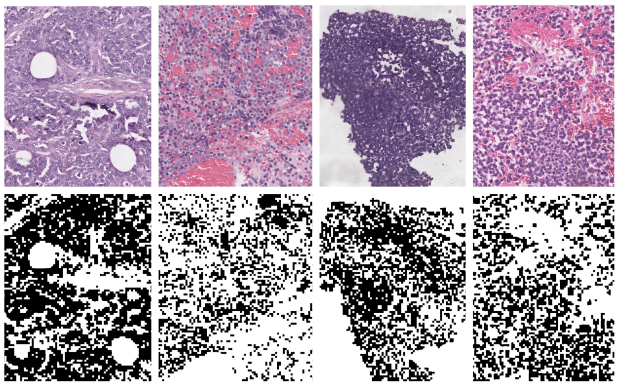
in Figure 1. In order to make the proportion of the cell tissue area in the image as large as possible, we rotate and randomly shift this area within a specific range to fill masked regions, and update the mask at the same time. In most cases, this direct method will not completely fill the blank area but can provide the model more useful information.

3.1.2. Feature Gating

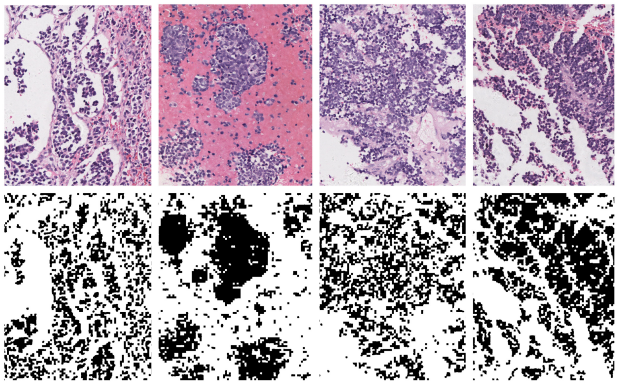
Gated layers are proven to be good at handling masked input in previous works (Yu et al., 2019) (Chang et al., 2019) to achieve image inpainting tasks. In these tasks, when networks based on vanilla convolutional layers are ill-fitted with large holes existing in image inputs, non-signal information will cause a bad effect on gradient updates and lead to blurriness, color discrepancy, and obvious edge responses surrounding holes of filled parts. That is because the model treats the pixel at any location of the whole image as equally valid. By applying a gating mechanism, trainable dynamic feature selection can be achieved at each local spatial region across layers. It is formulated as:

$$Gating_{x,y} = W_g I,$$
$$Feature_{x,y} = W_f I,$$
$$O_{x,y} = \sigma(Gating_{x,y}) \odot \phi(Feature_{x,y}),$$

where I represents the input feature and $O_{x,y}$ is the output feature at the location (x, y). W_g and W_f are the convolutional filters for gating and feature representation, respectively. $\sigma(\cdot)$ is the sigmoid non-linear function such that the gating value can be restricted within $(0, 1)$ and $\phi(\cdot)$ is an activation function such as ReLU and ELU.



(a) EWS



(b) NBUD

Figure 1: Some examples of masks for EWS and NBUD categories, where the black parts in the second row of each image represent valid areas.

3.1.3. Recontruction Regularization

During the optimization, it is undesired for the classification model to know any original information that should be masked. To let the network aware of signal-containing regions, we introduce mask into the loss by only calculating the least-squares loss between unmasked pixels of the input and output, written in a formula: $L_{reconst.}(x, y) = ||(I_{x,y} - O_{x,y}) \odot M||^2$, where $I_{x,y}$ and $O_{x,y}$ means the input and output of the auto-encoder at the location (x, y) and M means the corresponding mask.

This reconstruction loss can impose a straining effect on the dynamic feature filtering mechanism of gated layers and lead the network to discover available visual regions.

3.2. Diversity Tolerance

The high diversity of samples from the same category is caused by the various possibilities of sample appearances. For the SRBCTs studied in this paper, the tissue morphology on pathological slices is ever-changing due to the diversity of its frequent sites and easy metastasis. The high diversity of samples brings a great challenge when training data are insufficient. When a large dataset is available, the network adjusts the parameters during the training process to obtain a uniform and reasonable probability output after the feature goes through a series of linear and nonlinear transformations. However, when a small sample problem occurs, the high diversity of features cannot be accurately digested by the network. Therefore, we consider relaxing the probability calculation step. Instead of using the precise linear and non-linear calculations, we utilize the vector length of the feature to compute the probability. The advantage of this method is that the features of many different patterns can correspond to the same length. For instance, the different patterns $(0, 0, 1, 1)$ and $(1, 0, 1, 0)$ have the same value $\sqrt{2}$. This explicitly models the loss function with the relaxation condition, which we call a length-aware hinge loss.

3.2.1. Length-aware Hinge Loss

The original cross-entropy function is undoubtedly the most commonly utilized loss in the field of deep learning. However, it often does not work well on small sample datasets because it tends to consider the entire feature pattern, which reduces the flexibility of the model. For a small sample dataset, we utilize a length-aware hinge loss, which can treat different feature vectors extracted by the network as the same length value. This loss was also used in (Sabour et al., 2017), in conjunction with the dynamic routing mechanism, to solve a problem of structuring features. We use it based on a new starting point, i.e., the vector length can correspond to many different vector patterns. Our approach potentially allows the features coming from the same category to have higher diversity. In a multinomial classification problem with C categories, we can separately write the cross-entropy loss CE_c and the length-aware hinge loss LH_c as follows:

$$\begin{aligned} CE_c &= -t_c \log(\phi_c(W\mathbf{f})), \\ LH_c &= t_c(\max(0, T_{max} - |\mathbf{f}_c|))^2 + \\ &\quad \alpha(1 - t_c)(\max(0, |\mathbf{f}_c| - T_{min}))^2, \end{aligned}$$

where $\mathbf{f}$ is the feature vector to be input into the last layer and $\mathbf{f}_c$ is the feature for class c, which is obtained by performing matrix transformation on the convolved features $\mathbf{u}$, namely $\mathbf{f}_c = \sum_j W_{c,j}\mathbf{u}_j$. t_c equals 1 if the sample belongs to class c and 0, otherwise. $\phi(\cdot)$ is the softmax layer and W is the weights of the last fully connected layer. T_{max} and T_{min} are respectively the upper and lower threshold of the hinge loss and α is the loss down-weighting for absent classes.

3.3. Model Backbone

The architecture of SagaNet for easier understanding is displayed in Figure 2 and the final loss of the model is:

$$Loss = \sum_c^C LH_c + L_{reconst.}.$$

One important difference between natural images and pathological ones is that the latter's bottom features are more important. Thus, we choose a relatively light DenseNet (Huang et al., 2017) architecture as the base of our model, whose densely connected residual structure prevents the network from losing low-level representations. Upsampling followed by convolution is applied with kernel size 1×1 to replace deconvolution, which can lead to more stable optimization. To preserve the bottom signal of input, we also add skip-connections (Ronneberger et al., 2015) into the auto-encoder architecture. These connections help store and spread low-level information, which is essential to pathological images, throughout the whole network.

4. Experiments

In this section, we first introduce the pediatric cancer dataset which we mainly focus on. Then the experimental configurations are detailed and specific comparisons of results are reported. Furthermore, to show the generalization of the model, the comparison results on another dataset BreakHist (Spanhol et al., 2015) are also provided.

4.1. SRBCT2 Dataset

We train and evaluate our model on a small sample SRBCTs dataset with two tumor categories: neuroblastoma, undifferentiated subtype (NBUD) and Ewing sarcoma (EWS), named SRBCT2. The two tumor categories are usually considered impossible to visually diagnose without auxiliary techniques (immunohistochemistry, electron microscopy, and/or cytogenetics) (Shimada et al., 1999) (Barden & Lewey, 1949). There is a high degree of morphological diversity within either single category but a high degree of similarity between the two, which makes them challenging to classify. All the pathological images were collected from a national pediatric hospital. The record time of patient cases varies from 2010 to 2015, and 17 patients are involved in total, in which eight patients suffer NBUD and nine patients suffer EWS. There are 25 whole slide images in total, and each patient has 1 to 5 slices. Through taking screenshots with the size of 768×768 from them without overlapping, smaller patches were organized and labeled by a professional pathologist, and the patch number of each patient varies from 10 to 141. As a consequence, we got

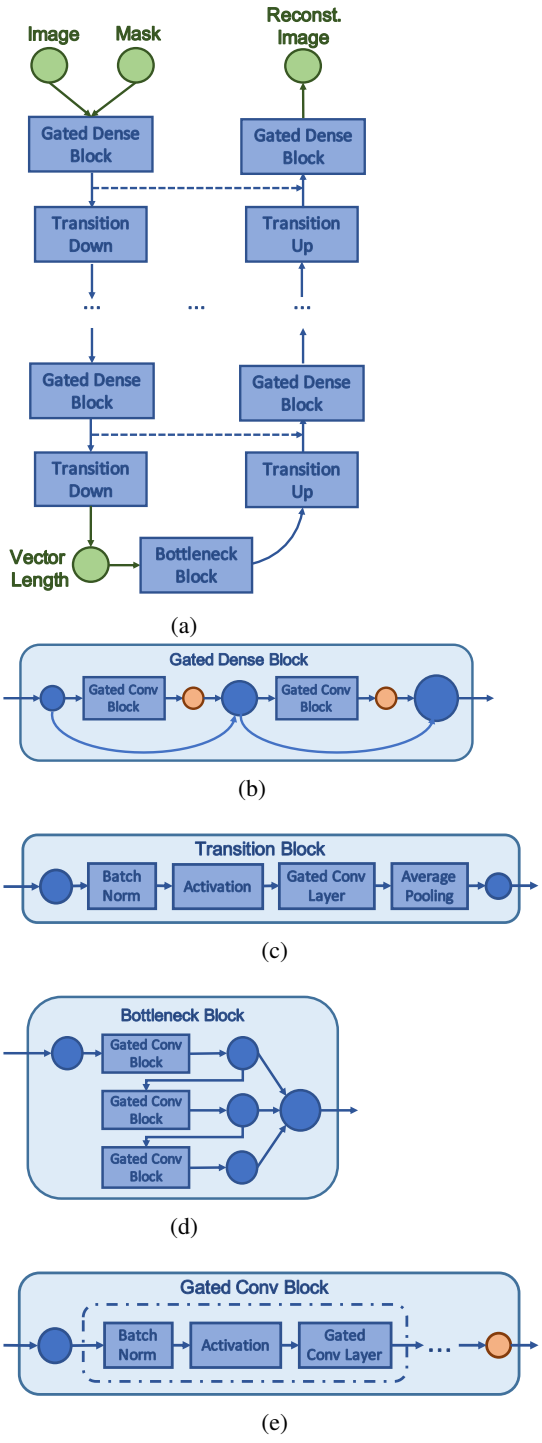


Figure 2: The schematic structure of SagaNet and the internal structure of major modules. The green circles represent input and output data, and the blue circles mean the features between layers. In gated dense layers, there is the growth of feature channel number marked with orange, which is one of the characteristics of DenseNets.

300 annotated patches for either of the two categories. After obtaining the SRBCT2 dataset, we performed data augmentation with rotating and flipping, and finally expanded the number of images by eight times. Then we split the dataset for model evaluation, in accordance with the practical and challenging situation that patches for training and evaluation should not come from the same patients. The patch number of each patient is different, and so the dataset is unable to be split evenly. Thus, we split the dataset with the constraint that the patients of the training set, the validation set, and the test set cannot overlap each other and the number of patches in the last two parts is within the range $[45, 55]$. Finally, we randomly take 20 splits.

4.2. Configurations

In this paper, we train our model from scratch and consider comparison deep models with and without pre-trained weights. When constructing masks for input images, we use the SGD optimizer with the learning rate of 0.05 when training the SENet network and set the number of minimum segmented areas as 7. In the process of merging segmented areas, we set $T_{low} = 78$, $T_{high} = 158$, $T_{red} = 37$ and $T_{smooth} = 35$.

Considering that different comparison models may have different levels of adaptability to inputs with masks, we perform experiments on two forms of the image input for fairness: one is the raw image without any mask, and the other is the image multiplied with its mask. In the proposed model, four gated dense blocks are used, and the number of gated convolutional blocks is set to one and two for the first two and the last two blocks, respectively. The growth rate of each dense block is 32, and the first convolutional layer contains 64 filters. We use ELU (Clevert et al., 2015) as the nonlinear function in the proposed model. When training neural networks, we configure the initial learning rate to $5e^{-6}$ with the Adam optimizer and multiply it by 0.7 at the end of each epoch. The whole number of training epochs is set to 300, and an early stopping mechanism is applied with the patience of 20 epochs. After the training process stops, the best-saved model is used to evaluate the performance on the test set. One GPU (NVIDIA GeForce RTX 2080 Ti) is used.

4.3. Comparison with Deep Models

The state-of-the-art deep models used for comparison are DenseNet121, DenseNet169, DenseNet201 (Huang et al., 2017), Inception-v3 (Szegedy et al., 2016), ResNet50 (He et al., 2016), Xception (Chollet, 2017), Mobilenets (Howard et al., 2017) and NASnet (Zoph et al., 2018).

A new top classifier with two fully connected layers and one dropout layer with the probability of 0.5 is applied to each model for the SRBCTs classification task. The

Table 1: Classification accuracies, p values of paired t-test and parameter numbers (Param. Num.) for state-of-the-art deep models and SagaNet.

	MASK INPUT		RAW INPUT		
MODEL	ACCURACY	$P_{t\text{-}test}$	ACCURACY	$P_{t\text{-}test}$	PARAM. NUM.
DENSENET121	0.6150± 0.1055	$5.1107e^{-6}$	0.6007± 0.1121	$5.4310e^{-4}$	$3.27e^{7}$
DENSENET121$_{pre}$	0.7568± 0.0576	$2.1010e^{-4}$	0.6793± 0.0832	$9.5404e^{-1}$	–
DENSENET169	0.5613± 0.0956	$7.1318e^{-12}$	0.6082± 0.1323	$5.0798e^{-6}$	$5.44e^{7}$
DENSENET169$_{pre}$	0.7799± 0.0658	$1.9576e^{-2}$	0.6720± 0.0810	$9.6755e^{-1}$	–
DENSENET201	0.5440± 0.0977	$1.9573e^{-12}$	0.5159± 0.0561	$1.1007e^{-8}$	$6.65e^{7}$
DENSENET201$_{pre}$	0.7847± 0.0606	$1.0021e^{-2}$	0.6925± 0.0709	$5.9305e^{-1}$	–
INCEPTION-V3	0.5265± 0.1039	$1.5035e^{-12}$	0.6175± 0.1549	$4.5514e^{-4}$	$4.80e^{7}$
INCEPTION-V3$_{pre}$	0.6565± 0.0657	$1.3520e^{-15}$	0.6611± 0.0847	$5.7810e^{-1}$	–
RESNET50	0.4898± 0.0153	$2.1705e^{-17}$	0.6512± 0.1111	$9.5022e^{-4}$	$7.50e^{7}$
RESNET50$_{pre}$	0.7268± 0.0826	$1.1542e^{-3}$	0.6396± 0.0970	$2.4182e^{-1}$	–
XCEPTION	0.6004± 0.1459	$1.1690e^{-7}$	0.5216± 0.0671	$2.2246e^{-9}$	$7.22e^{7}$
XCEPTION$_{pre}$	0.7257± 0.0614	$1.8365e^{-7}$	0.6766± 0.1042	$8.7516e^{-1}$	–
MOBILENETS	0.5798± 0.0918	$8.5734e^{-11}$	0.6402± 0.1325	$2.2565e^{-3}$	$2.11e^{7}$
MOBILENETS$_{pre}$	0.7550± 0.0523	$5.2889e^{-4}$	0.6715± 0.0865	$6.7780e^{-1}$	–
NASNET	0.4943± 0.0162	$5.4818e^{-18}$	0.4848± 0.0132	$2.1779e^{-10}$	$1.86e^{8}$
NASNET$_{pre}$	0.6867± 0.0547	$1.1731e^{-9}$	0.6848± 0.0819	$8.0931e^{-1}$	–
SAGANET	**0.8031± 0.0578**	–	**0.7457± 0.0763**	–	$1.96e^{6}$

performance results are shown in Table 1. Considering that these comparison models may be disadvantaged on small datasets due to a large number of parameters, which may lead to unfair comparisons, we also consider loading pre-trained weights for them and making the last 4 layers of parameters trainable. In the table, we use the subscript pre to indicate that the model has been loaded with pre-trained weights. Moreover, paired t-test experiments were also performed on the results of 20 data splits, and when the p-value is less than 0.05, the difference between results will be considered significant. The parameter numbers of each model are shown in the last column.

As is shown in Table 1, the proposed model with masked inputs achieves the highest accuracy, and its superiority is also evident as the p-value is far lower than 0.05 for any other deep model. Considering that this dataset is very small and the accuracy of visual classification alone in professional pathologists' recognition is only 50%, the current performance improvement is actually very remarkable. The performance of the comparison model has increased after using the pre-trained weights, and uniformly shows better performance when using masks. This phenomenon shows that when with good initialization settings, using masks can indeed help reduce noises. SagaNet surpassed these models with high significance, indicating that the gating and partial reconstruction designs for masks are necessary and effective. These designs inside the network may not be fully functional without masks. What can prove this is that when only raw images are used as input, the significance of SagaNet performance advantage compared with the pre-trained comparison model is reduced. All in all, the mask input and the corresponding designs inside the model are closely matched and indispensable.

For SagaNet, the performance of the model has taken a qualitative leap. This is because the combination of gated layers and reconstruction loss enables the information in the valid area to be exploited more fully without interference from other noises. When we use raw inputs, there is no explicit noise-filtering mechanism, but the model still shows a certain degree of advantage, showing that the design of length-aware hinge loss is more suitable for small datasets than the cross-entropy loss.

These results may illustrate that when the dataset is small and highly diverse within each class, popular deep models cannot be directly applied to the task, and some targeted designs for data as in SagaNet are necessary. Thus, it is very meaningful to build such a delicate model to solve real problems. We also perform comparison experiments on traditional classifiers and results are shown in Appendix A.

4.4. Ablation Experiments

To evaluate the influence of each novel part in SagaNet, we performed ablation experiments in the same condition as the proposed model. The parts that we take into consideration are the gated layer, the reconstruction regularization term, the length-aware hinge loss, and the masked input. Models that lack one of these parts are respectively named as $Masked^C$, $Gated^C$, $Reconst^C$, and $Hinge^C$, which means the complement set of the corresponding design part. We use raw images as inputs for $Masked^C$, use vanilla convolutional layers to replace gated layers for $Gated^C$, remove the reconstruction loss term from the total loss

for $Reconst^C$, and use a cross-entropy loss to replace the length-aware hinge loss for $Hinge^C$. The baseline model has the same mainframe as SagaNet but without all of the parts mentioned above. Experimental results are shown in Table 2.

Table 2: Classification accuracies and p values of paired t-test in the ablation experiments of SagaNet.

MODEL	ACCURACY	$P_{t\text{-}test}$
$Masked^C$	0.7457± 0.0763	$2.6499e^{-4}$
$Gated^C$	0.7058± 0.0904	$4.8692e^{-5}$
$Reconst^C$	0.6268± 0.1221	$1.4353e^{-6}$
$Hinge^C$	0.6086± 0.1039	$3.9894e^{-7}$
BASELINE	0.5550± 0.1146	$4.2072e^{-12}$
SAGANET	**0.8031± 0.0578**	–

It can be found that the performance of the baseline is dramatically improved by applying our thoughts. Furthermore, the results show that loss design plays a crucial role in the optimization of the model. Both the hinge loss and reconstruction regularization term are necessary for the current design. The former provides the model the outstanding ability to tolerate high sample diversity, and the latter guides gated layers to filter features dynamically. Masks and gated layers also help improve accuracies significantly, which shows either of them cannot be ignored for the aim of filtering image noises. No matter which part of SagaNet's design is removed, noticeable performance degradation will occur, which shows the rationality and integrity of our model design.

4.5. Experiements on BreakHist

In order to verify the generalization of SagaNet in other small-scale pathological images, we conducted experiments on another dataset named BreakHist (Spanhol et al., 2015). This dataset contains pathological images of the same case set at four magnifications, $40\times$, $100\times$, $200\times$, and $400\times$. All images are of size 700×460. Each set of magnification has two types of tissue images, benign and malignant, and the data distribution is shown in Table 3. We perform 5 data splits according to the requirements in (Spanhol et al., 2015), and make the test set account for about 30% of the total data while ensuring that the test set and training set cases do not overlap. In order to use the same hyperparameters when generating the mask, we use the image hue in the SRBCTs dataset as the reference hue for BreakHist color normalization (Vahadane et al., 2016). The final result is shown in Table 4. We also evaluate the model according to a recognization rate defined in that work. If N_{reg} images are correctly recognized from N_P cancer images of a patient, the formulation of a patient score N_{ps} can be written as $N_{ps} = \frac{N_{reg}}{N_P}$. The recognition rate P_{reg} of N_{pat} patients is $P_{reg} = \frac{\sum N_{ps}}{N_{pat}}$. The results shown in the table illustrate that the proposed method has very high superiority and stability, and combining mask-based noise filtering and a length-aware loss is very reasonable and effective, which can better deal with small-scale pathological datasets.

Table 3: Data distribution of BreakHist dataset

MAGNIFICATION	BENIGN	MALIGNANT	TOTAL
$40\times$	652	1,370	1,995
$100\times$	644	1,437	2,081
$200\times$	623	1,390	1,995
$400\times$	588	1,232	1,820
$Total$	2,480	5,429	7,909

Table 4: Recognition rates for comparison classifiers and SagaNet.

	MAGNIFICATION FACTORS			
MODEL	$40\times$	$100\times$	$200\times$	$400\times$
1-NN	80.9± 2.0	80.7± 2.4	81.5± 2.7	79.4± 3.9
QDA	83.8± 4.1	82.1± 4.9	84.2± 4.1	82.0± 5.9
RF	81.8± 2.0	81.3± 2.8	83.5± 2.3	81.0± 3.8
SVM	81.6± 3.0	79.9± 5.4	85.1± 3.1	82.3± 3.8
SAGANET	**96.2± 0.6**	**96.0± 1.5**	**94.4± 1.2**	**92.7± 1.3**

5. Conclusion

Artificial intelligence assisted diagnosis is an important direction that has attracted much attention in the medical diagnosis field. The pediatrics field faces a more serious data shortage than adult medicine, which makes it very challenging for deep models to provide excellent and stable performance. In this work, we start by solving a difficult pediatric tumor pathological problem and propose a general deep network model for small sample classification, named SagaNet.

Due to the high similarity of the two considered cancer types, even professional pathologists have difficulty in distinguishing them via observing the slices with microscopes. In this model, we propose to extract masks based on the characteristics of pathological images and use them as the input of the network together with the corresponding images. In order to make the network focus on non-masked areas, we use gating layers to filter the information and use a partial reconstruction loss to guide the training process. To further improve the network's tolerance to feature diversity, we propose a length-aware hinge loss. In experiments, SagaNet largely outperforms other state-of-the-art deep models and shows good generalization.

Acknowledgements

This work was supported by NSFC Project 62076096, Shanghai Municipal Project 20511100900, the Open Research Fund of KLATASDS-MOE, and the Fundamental Research Funds for the Central Universities.

References

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. Slic superpixels. Technical report, 2010.

Barden, R. P. and Lewey, F. H. Metastasizing cerebellar tumors: the difficulty in distinguishing between medulloblastoma and neuroblastoma. *Journal of Neurosurgery*, 6(6):439–449, 1949.

Brereton, H. D., Simon, R., and Pomeroy, T. C. Pretreatment serum lactate dehydrogenase predicting metastatic spread in ewing's sarcoma. *Annals of Internal Medicine*, 83(3): 352–354, 1975.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8): 1301–1309, 2019.

Chang, Y.-L., Liu, Z. Y., Lee, K.-Y., and Hsu, W. Free-form video inpainting with 3D gated convolution and temporal patchgan. In *IEEE International Conference on Computer Vision*, pp. 9066–9075, 2019.

Chen, H., Qi, X., Yu, L., Dou, Q., Qin, J., and Heng, P.-A. Dcan: Deep contour-aware networks for object instance segmentation from histology images. *Medical Image Analysis*, 36:135–146, 2017.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258, 2017.

Clevert, D. A., Unterthiner, T., and Hochreiter, S. Fast and accurate deep network learning by exponential linear units (Elus). *arXiv preprint arXiv:1511.07289*, 2015.

Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, pp. 933–941. JMLR. org, 2017.

Di Bisceglie, A. M., Rustgi, V. K., HOOFNAGLE, J. H., DUSHEIKO, G. M., and LOTZE, M. T. Hepatocellular carcinoma. *Annals of Internal Medicine*, 108(3):390–401, 1988.

Harrison, A. P., Xu, Z., George, K., Lu, L., Summers, R. M., and Mollura, D. J. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 621–629. Springer, 2017.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Iglovikov, V. I., Rakhlin, A., Kalinin, A. A., and Shvets, A. A. Paediatric bone age assessment using deep convolutional neural networks. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 300–308. Springer, 2018.

Janoueix-Lerosey, I., Schleiermacher, G., and Delattre, O. Molecular pathogenesis of peripheral neuroblastic tumors. *Oncogene*, 29(11):1566, 2010.

Kanezaki, A. Unsupervised image segmentation by backpropagation. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1543–1547. IEEE, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

Lakhani, P. and Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 284(2):574–582, 2017.

Larson, D. B., Chen, M. C., Lungren, M. P., Halabi, S. S., Stence, N. V., and Langlotz, C. P. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology*, 287 (1):313–322, 2017.

Ma, X., Liu, Y., Liu, Y., Alexandrov, L. B., Edmonson, M. N., Gawad, C., Zhou, X., Li, Y., Rusch, M. C., Easton, J., et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature*, 555(7696):371–376, 2018.

Marshall, G. M., Carter, D. R., Cheung, B. B., Liu, T., Mateos, M. K., Meyerowitz, J. G., and Weiss, W. A. The prenatal origins of cancer. *Nature Reviews Cancer*, 14(4): 277–289, 2014.

Niu, Y., Gu, L., Lu, F., Lv, F., Wang, Z., Sato, I., Zhang, Z., Xiao, Y., Dai, X., and Cheng, T. Pathological evidence exploration in deep retinal image diagnosis. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 1093–1101, 2019.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pp. 234–241. Springer, 2015.

Sabour, S., Frosst, N., and Hinton, G. E. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pp. 3856–3866, 2017.

Sharma, S., Kamala, R., Nair, D., Ragavendra, T. R., Mhatre, S., Sabharwal, R., Choudhury, B. K., and Rana, V. Round cell tumors: classification and immunohistochemistry. *Indian Journal of Medical and Paediatric Oncology: Official Journal of Indian Society of Medical & Paediatric Oncology*, 38(3):349, 2017.

Shimada, H., Ambros, I. M., Dehner, L. P., Hata, J., Joshi, V. V., and Roald, B. Terminology and morphologic criteria of neuroblastic tumors: Recommendations by the international neuroblastoma pathology committee. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 86(2):349–363, 1999.

Shu, L., Sun, Y., Tan, L., Shu, Q., and Chang, A. C. Application of artificial intelligence in pediatrics: past, present and future. *World Journal of Pediatrics*, 15(2):105–108, 2019.

Spanhol, F. A., Oliveira, L. S., Petitjean, C., and Heutte, L. A dataset for breast cancer histopathological image classification. *Transactions on Biomedical Engineering*, 63(7):1455–1462, 2015.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, 2015.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Tabrizi, P. R., Mansoor, A., Cerrolaza, J. J., Jago, J., and Linguraru, M. G. Automatic kidney segmentation in 3d pediatric ultrasound images using deep neural networks and weighted fuzzy active shape model. In *International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 1170–1173. IEEE, 2018.

Tang, J., Li, Z., Chen, J., et al. Pediatric oncology, 2011.

Titano, J. J., Badgeley, M., Schefflein, J., Pain, M., Su, A., Cai, M., Swinburne, N., Zech, J., Kim, J., Bederson, J., et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nature Medicine*, 24(9):1337–1341, 2018.

Titford, M. The long history of hematoxylin. *Biotechnic & Histochemistry*, 80(2):73–78, 2005.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A. M., Esposito, I., and Navab, N. Structure-preserving color normalization and sparse stain separation for histological images. *Transactions on Medical Imaging*, 35(8):1962–1971, 2016.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., and Huang, T. S. Free-form image inpainting with gated convolution. In *IEEE International Conference on Computer Vision*, pp. 4471–4480, 2019.

Zoph, B., Vasudevan, V., Shlens, J., and Le, Q. V. Learning transferable architectures for scalable image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8697–8710, 2018.