Neurocomputing 511 (2022) 105–116

Contents lists available at ScienceDirect

Neurocomputing

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

Combating spatial redundancy with spectral norm attention in convolutional learners

Jiansheng Fang [a], Dan Zeng [b], Xiao Yan [b], Yubing Zhang [c], Hongbo Liu [c], Bo Tang [b], Ming Yang [c], Jiang Liu [b,*]

[a] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[b] Research Institute of Trustworthy Autonomous Systems and Department of Computer Science and Engineering, Southern University of Science and Technology, ShenZhen, China
[c] CVTE Research, Guangzhou, China

ARTICLE INFO

Article history:
Received 1 June 2022
Revised 14 July 2022
Accepted 4 September 2022
Available online 9 September 2022

Keywords:
Spatial redundancy
Spectral norm
Attention
Convolutional Neural Networks
Singular value decomposition

ABSTRACT

There is an inherent and longstanding challenge for vision learners to exploit informative features from digital images with spatial redundancy. Given pre-processing image methods require task-specific customization and may rise unanticipated poor performance due to redundancy removal, we explore improving vision learners to combat spatial redundancy during vision learning, a task-agnostic and robust solution. Among popular vision learners, vision transformers with self-attention can mitigate pixel redundancy by capturing global dependencies, while convolutional learners fall into locality via a limited receptive field. To this end, based on investigating inter-pixel spatial redundancy of images, in this work, we propose spectral norm attention (SNA), a novel yet efficient attention block to help convolutional neural networks (CNNs) highlight informative features. We can seamlessly plug SNA into off-the-shelf CNNs to suppress the contributions of redundant features by globally differentiating and weighting. In particular, SNA performs singular value decomposition (SVD) on intermediate features of each image within a mini-batch to obtain its spectral norm. The features in the direction of the spectral norm are most informative, while the discriminative features in other directions leave less. Hence, we apply the rank-one approximation of the spectral norm direction as attention weights to enhance informative features. Besides, we adopt the power iteration algorithm to approximate the spectral norm to significantly reduce the matrix computation overhead during training, thus keeping inference speed on par with vanilla CNNs. We extensively evaluate our SNA on four mainstream natural datasets to demonstrate the effectiveness and favourability of our SNA against its counterparts. In addition, the experimental results of image classification and object detection show our SNA can bring more gains to medical images with heavy redundancy than other state-of-the-art attention modules.

1. Introduction

Inter-pixel spatial redundancy is an intrinsic characteristic of digital images [1]. It refers to that neighboring pixels are not statistically independent and the value of any given pixel can be predicated from the value of its neighbors that is they are highly correlated. Fig. 1 can demonstrate that inter-pixel spatial redundancy is inherent and varies in degree for different imaging ways. By performing singular value decomposition (SVD) on two images of column (a), we can calculate explained variance ratios (EVR) of their singular values. EVR measures the percentage occupying information in a matrix for each singular value. The information carried by individual pixels is relatively small. We can group pixels into two categories: informative pixels and redundant pixels. Those pixels in the directions of singular values with non-zero EVR are informative while others are redundant. The qualitative schematics of Columns (b) and (c) can disclose that the low-rank approximation only using top-30 singular values can almost recover original images, and the top-1 singular value (*i.e.*, the spectral norm) contributes the most information. By observing column (d), we can find that the number of singular values with non-zero EVR is less than 30 (red dot in the first row) for the natural image while the medical image is less than 10 (red dot in the second row). By combining the dimensions of the two images, it can clearly be seen that the medical image is more redundant. The statistical result in Column (e) can also prove such a claim, where the averages of maximal explained variances of medical images on varying numbers are larger than natural images.

* Corresponding author.
E-mail address: liuj@sustech.edu.cn (J. Liu).

https://doi.org/10.1016/j.neucom.2022.09.075
0925-2312/

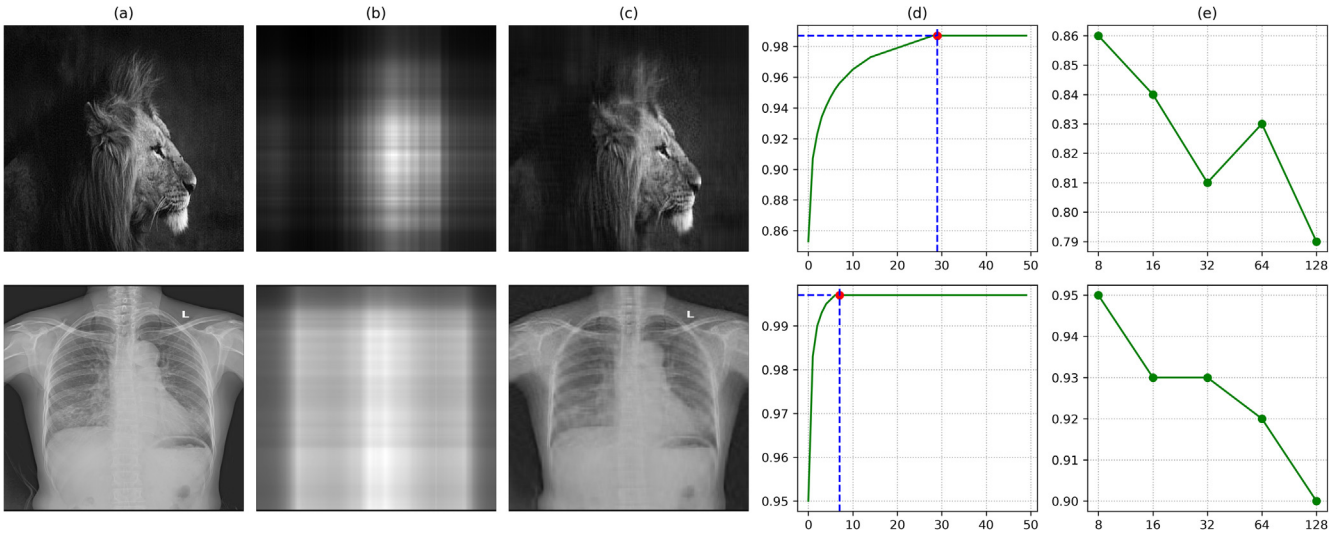

Fig. 1. Illustration of inter-pixel spatial redundancy in images. Column (a) includes a natural image with dimensions of (280, 415) and a chest X-ray image with dimensions of (3072, 2540). Column (b) shows the two rank-one approximation images of the spectral norm (top-1 singular value). And the two low-rank approximation images of top-30 singular values lie in columns (c). Column (d) is their cumulative sum (y-axis) of explained variance ratios of top-50 singular values (x-axis). Column (e) counts the mean (y-axis) of maximal explained variances on varying image numbers (x-axis).

With the above investigation, it is inevitable to consider inter-pixel spatial redundancy during image processing. Inter-pixel spatial redundancy leads to low efficiency and repetitive features occurring when inferring vision learners [2,3]. The redundant features have a limited contribution to vision learners and may even bring over-fitting, thus diminishing the generalization quality [4]. Hence, when the highly informative features are drowned in the considerable redundant features, it is challenging to extract the discriminative features to increase learning accuracy [5]. The intuitive strategy of combating spatial redundancy is to apply redundancy removal approaches to pre-process images before being input into the networks, including PCA [6], DCT [7], *etc*. Recently, to encourage learning useful features from images with spatial redundancy, MAE [8] presents a simple strategy to reduce redundancy by masking a very high portion of random patches. Although pre-processing image methods can effectively address data redundancy, they require task-specific customization and may rise to unanticipated inferior performance. Hence, encouraged by a pioneer work [3] reducing spatial redundancy by reparameterizing convolutional layers, we consider improving vision learners to combat spatial redundancy, a task-agnostic and robust solution.

Among vision learners, vision transformers (ViTs) [9] enjoy the ability to capture intra-image long-range dependencies and exhibit impressive performances, as MAE [8] uses ViTs as encoders. The self-attention mechanism in ViTs favors learning global discriminative features [10], thus addressing spatial redundancy to some extent. In contrast to ViTs, convolutional neural networks (CNNs) provide locality via a limited receptive field [11], establishing prior for image processing. CNNs naturally capture local patterns, not global context, so it is impractical to screen out redundant features. With the above comparison between ViTs and CNNs, we intuitively consider introducing a global attention mechanism to improve CNN networks, like self-attention for ViTs. Attention-based models offer an adaptive aggregation mechanism, where the aggregation scheme itself is input-dependent or spatially dynamic [10]. Through the proposed global attention mechanism, we aim to identify informative and redundant features and then generate differentiated attention weights to play their respective contributions to the final discriminative decision.

With the above intention of our proposed attention mechanism, we first make a brief retrospective discussion on attention mechanisms. Both global and local attention mechanisms[12–16] can help advance the state-of-the-art performance of CNNs by selectively emphasizing salient features and suppressing less useful ones. A representative method of local attention [17] is squeeze-and-excitation (SE), which learns channel-wise attention for each convolution block, showing encouraging performance for various CNNs. On the other hand, global attention [18–21] has emerged as a recent advance due to its advantage of capturing long-range interdependencies. Due to quadratic memory and computational requirements involving high-resolution images, global attention is intractable. Instead, local attention is the current *de facto* choice in CNNs for better vision backbones. However, local attention can not identify informative and redundant features. To this end, we propose a novel attention mechanism for CNNs, termed spectral norm attention (SNA), to leverage its globally expressive power while retaining the locality and shift-invariance of convolutional learners. The global SNA aims to help convolutional learners effectively learn informative features from images with large redundant pixels without paying dear computation.

Specifically, given a matrix obtained from feature aggregation on an image, SNA performs SVD on this matrix to obtain the spectral norm. The highly informative features lie in the direction of the spectral norm, and the features in the other directions of singular values are redundant. We intend to enhance informative features of the direction of the spectral norm and penalize redundant features of other directions. After screening out redundant features of the matrix, we employ the rank-one approximation of the spectral norm direction as attention weights to boost the discriminative contributions of informative features. In addition, to efficiently solve the spectral norm, we apply the power iteration algorithm to approximate the spectral norm. Compared to vanilla CNNs, the computation overhead slightly increase after integrating SNA, almost negligible. Concretely, we make the following contributions:

- A longstanding challenge for CNNs is learning discriminative features from digital images with spatial redundancy. We propose a novel and efficient attention block to address the challenge, named spectral norm attention (SNA), which can elaborately designate attentive scores for features in terms of their global informativity.

- SNA performs SVD on an aggregated feature matrix for each image within a mini-batch to obtain its spectral norm. We further generate distinguishing attention weights for informative and redundant features by utilizing the rank-one approximation of spectral norm direction. By laying emphasis on the highly informative features in the direction of the spectral norm, SNA can help convolutional learners improve generalization.
- SNA can be seamlessly integrated into the off-the-shelf CNNs to help improve performance without paying dearly to model efficiency. We demonstrate the effectiveness of our proposed SNA by conducting experiments on five image datasets. Experimental results also show that our SNA can gain more when encountering heavily spatial redundancy, such as medical images.

The remainder of this paper includes related works reviewing in Section 2, our method describing in detail in Section 3, experiment analysis in Section 4, discussion in Section 5, and the conclusion stating in Section 6.

2. Related works

2.1. Vision learners

The inherent inductive biases make CNNs achieve currently state-of-the-art in computer vision and therefore widely used in different image recognition tasks [11]. Recently, ViTs have emerged as a competitive vision learner to CNNs by using multi-head self-attention without requiring the image-specific biases [9,22]. The self-attention mechanism favors ViTs for capturing global dependencies but also arises intractable with higher-resolution inputs due to a quadratic complexity with respect to the input size [23]. Without the CNN inductive biases, a vanilla ViT model faces many challenges in being adopted as a generic vision backbone [24]. Swin Transformer employs a hybrid strategy to unify the intrinsic superiorities of ViT and CNN, thus bringing this gap [25]. Inspired by reintroducing attention within local windows to ViT [25], which is similar to the essence of convolutions, we consider introducing global attention to CNN with local nature to address spatial redundancy in visual learning. We intend to unify the advantages of the convolution locality and the attention globality in vision learners to effectively extract discriminative features.

2.2. Attention mechanisms

Attention mechanisms [18,26,27] can help focus on salient features by explicitly modeling local or global feature interdependencies, showing encouraging potential in performance improvement. Recently, incorporation of attention modules into CNNs has demonstrated their utility across many computer vision tasks [28,13,29,19,30–32], including image classification, object detection, and semantic segmentation. From the range of capturing feature interdependencies, the development of current attention methods relying on feature aggregation can roughly be divided into two directions: local attention and global attention.

SE [17] is the first local attention to learn channel interactions and attracts the exploration of attention methods for CNNs because of its promising performance. Another pioneering attention method is the convolutional block attention module (CBAM) [12], which combines spatial- and channel-wise attention to boost the representation power of CNNs by employing both average and max pooling. Subsequently, efficient channel attention (ECA) [33] generates channel weights by performing a fast 1D convolution on the aggregated features obtained by global average pooling. Generally, local attention captures cross-channel or -spatial dependencies by performing convolution or fully connected layers operations on aggregated features.

On the other hand, global attention capture global-range interactions by modeling all feature relationships. Using global attention in cooperation with convolutions [34–36] has attracted a lot of interest due to its tremendous success in natural language processing. Generally, the modeling strategy employs dot-product along channel-wise to aggregate features and build a relationship matrix for pairwise features [19]. Unlike the pooling or convolutional operator used in local attention, the attentive scores of global attention are produced dynamically via a similarity function between features. For example, attention augmentation (AA) [11] develops a two-dimensional relative global attention to maintains translation equivariance while being infused with relative position information, making it well suited for images. Unfortunately, self-attention-based global attentions need heavy computation consumption. Concerning this, some rising methods explore capturing long-range dependency indirectly. EMNet [37] applies the expectation–maximization manner to compute the correlation between pixels. And CTNet [32] interactively explores the spatial contextual dependency between pixels and the semantic dependency between channels for semantic segmentation tasks. The informativeness of features can be measured in the global context. Hence, we rely on global attention to identify informative and redundant features, then assign their varying attentive scores. Inspired by global information blocks [38] modeling the global context issue as a low-rank recovery problem by matrix decomposition, our SNA applies SVD to identity informative features, then boosts their contributions to the discriminative decision of CNNs.

2.3. Application of spectral norm

SVD is very helpful for analyzing properties of a matrix and has been applied in a broad of range tasks, including recommender systems [39,40] and signal processing [41,42]. SVD-based matrix norms, including nuclear norm, Frobenius norm, and spectral norm, are usually used as the additional regularization term of the objective function to help obtain better generalization. Classically, it is thought that Frobenius norm helps models achieve weight decay while nuclear norm enforcing model sparsity. And the spectral norm has been claimed to be related to generalizability [43]. There are two regularizers related to the spectral norm: one penalizes the spectral norm by adding an explicit regularization term to the objective function [44], and the other normalizes weight matrices by imposing the spectral norm, this is, an implicit regularization term which is independent of the objective function [45]. Both regularizers adopt the power iteration method to compute the spectral norm due to the expensive computation overhead. Another application of the spectral norm is spectral initialization [46], which initializes weights of factorized neural layers using SVD so that their product approximates the target un-factorized weight matrix. Unlike the current application of the spectral norm imposing on weight matrix, our SNA utilizes SVD-based spectral norm to differentiate informative features and redundant features, playing the effect of feature selection. We further apply the low-rank approximation to give differentiated attention weights. Moreover, considering the expensive overhead of SVD computation, we utilize the power iteration method to obtain an approximate solution of the spectral norm to achieve a comparable efficiency.

3. Methodology

This section will elaborate on our proposed SNA method, including SNA block, spectral norm solution, and vision learners.

3.1. SNA block

Consider the input feature maps with a batch size of $B, \boldsymbol{X} \in \mathbb{R}^{B\times C\times H\times W}$, where H and W are the height and width of each feature map, C denote the number of channels. Through our SNA block, the output feature maps is $\widetilde{\boldsymbol{X}}$ with the size of $B \times C \times H \times W$. Fig. 2 illustrates the overview of our SNA block. After channel reduction by using global average pooling (GAP) on channel-wise, our SNA performs SVD on each feature matrix $\boldsymbol{M}$ with the size of $H \times W$. Based on SVD, our SNA obtains the largest singular value (σ_{max}), and its left and right singular vectors ($\boldsymbol{U}$ and $\boldsymbol{V}$), as follows:

$$\sigma_{max}, \boldsymbol{U}, \boldsymbol{V} = SVD(\boldsymbol{M}), \tag{1}$$

where the sizes of $\boldsymbol{U}$ and $\boldsymbol{V}$ are $1 \times H$ and $1 \times W$, respectively.

The spectral norm of a matrix refers to its largest singular value and its direction occupies the most informative features of the matrix. We apply rank-one approximation of the spectral norm to get a spatial matrix $\widehat{\boldsymbol{M}}$, as follows:

$$\widehat{\boldsymbol{M}} = \sigma_{max} \times \boldsymbol{U}^T \times \boldsymbol{V}, \tag{2}$$

where the size of $\widehat{\boldsymbol{M}}$ is $H \times W$.

In fact, the value of $\widehat{\boldsymbol{M}}$ is the approximation of the matrix $\boldsymbol{M}$ only considering the direction of the largest singular value. A higher value of the spectral norm implies better approximation. About the spectral norm and the Frobenius norm, there is an important inequality:

$$\|\boldsymbol{M}\|_S = \sigma_{max}(\boldsymbol{M}) \leqslant \|\boldsymbol{M}\|_F, \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This inequality is derived from the trace of a matrix that equals the sum of its eigenvalues, and equality holds if and only if the matrix $\boldsymbol{M}$ is a rank-one matrix or a zero matrix. We can hypothesize that the value of $\widehat{\boldsymbol{M}}$ is infinitely close to the matrix $\boldsymbol{M}$ when the largest singular value covers most of the important information. The higher attentive scores in $\widehat{\boldsymbol{M}}$ correspond to the informative features in $\boldsymbol{M}$, and the redundant features in $\boldsymbol{M}$ are assigned lower scores. Based on the attention matrix $\widehat{\boldsymbol{M}}$, we can enhance informative features.

Each spatial feature of each input has a value in this matrix. The values of corresponding positions in the matrix can be used to denote whether informative feature or redundant feature. Further, our SNA generates spatial weights as attentive scores by performing a Softmax function on the matrix $\widehat{\boldsymbol{M}}$. Then, this attentive matrix is reshaped as the size of $B \times 1 \times H \times W$ and is used to element-wise product with the input feature maps of $B \times C \times H \times W$ to output the feature maps $\widetilde{\boldsymbol{X}}$.

In this work, we identify informative and redundant features by performing SVD on feature matrices. The direction of the spectral norm covers highly informative features. Those directions of non-zero singular values also contain limited informative features. For simplicity, we categorize features of each matrix according to whether they lie in the direction of the spectral norm. After SVD, we generate a matrix of $H \times W$ by using the rank-one approximation of the spectral norm. This matrix can be used as attentive scores corresponding to feature informativeness, thus enhancing informative features of the direction of the spectral norm.

The features in the direction of non-zero singular values are informative. Especially, the direction of the largest singular value enjoys dominant eigenvectors. The intention of leveraging SVD is to get the largest singular value. In particular, we utilize the power iteration algorithm to approximate the largest singular value (*i.e.*, the spectral norm), then apply the rank-one approximation matrix as attentive scores to build our SNA. The rank-one approximation matrix on the spectral norm can help enhance the most informative features in global feature matrices, thus helping enforce globally expressive power implicitly. Hence, the contribution of matrix decomposition in this work lies in helping locate the most informative features in global feature matrices, then boosting their contributions to the final discriminative decision through SNA.

3.2. Spectral norm solution

.

Algorithm 1 Power iteration method for spectral norm

1: Input a batch matrix M with the size of $H \times W$.
2: $v \leftarrow$ a random unit vector with the size of $1 \times W$.
3: $e \leftarrow$ a small floating-point value close to zero.
4: $\overline{\boldsymbol{M}} = \boldsymbol{M}^{\mathrm{T}} \times \boldsymbol{M}$.
5: **while** True **do**
6: $\quad \hat{v} = v$
7: $\quad v = \overline{\boldsymbol{M}} \times v^T$
8: $\quad v = \frac{v}{\|v\|}$
9: $\quad$ **if** $v \cdot \hat{v} > (1 - e)$ **do**
10: $\quad\quad$ Converge and exit the loop.
11: $\quad$ **end if**
12: $\quad$ **end while**
13: $u = \boldsymbol{M} \times v^T$
14: $\sigma_{max} = \|u\|$
15: $\hat{\boldsymbol{M}} = \sigma_{max} \times u^T \times v$
16: Output $\hat{\boldsymbol{M}}$.

For the proposed SNA block design to be of practical use, it must offer a good trade-off between improved performance and increased model complexity. The computational burden associated with our SNA mainly includes an operation: the computation of the spectral norm. It is expensive to compute the spectral norm by SVD during model training. Hence, we approximate the spectral norm by a greedy algorithm. This algorithm is what we called power iteration. The pseudocode of power iteration for spectral decay is provided in Algorithm 1.

Assuming the first singular value of a matrix is sufficiently larger than the second one and in turn all the other singular values, we start with a random unit vector v as the right singular vector of the spectral norm for the matrix $\boldsymbol{M}$, and let $\overline{\boldsymbol{M}} = \boldsymbol{M}^T \times \boldsymbol{M}$. $\overline{\boldsymbol{M}}$ is a square matrix with $W \times W$ dimensions. Then we loop computing $v = \overline{\boldsymbol{M}} \times v^T$ and normalizing at each step. The condition for stopping is that the magnitude of the dot product between $\hat{v}$ and v is very close to 1. Since both are unit vectors, this is the cosine of the angle between them. After a sufficient number of iteration times, we get the right singular vector v of the largest singular value. Then we can compute the corresponding left singular vector $u = \boldsymbol{M} \times v^T$ and normalize it. σ_{max} is the largest singular value, u is the left vector, and v is the right vector. The rank-one approximation to the matrix $\boldsymbol{M}$ is $\hat{\boldsymbol{M}}$ by using the spectral norm.

Generally, global attention methods rely on matrix multiplication in capturing long-range feature interactions. Our SNA conducts matrix calculations on global feature maps. The computation overhead is extremely costive if we apply SVD to solve the spectral norm. Hence, we utilize the power iteration algo-

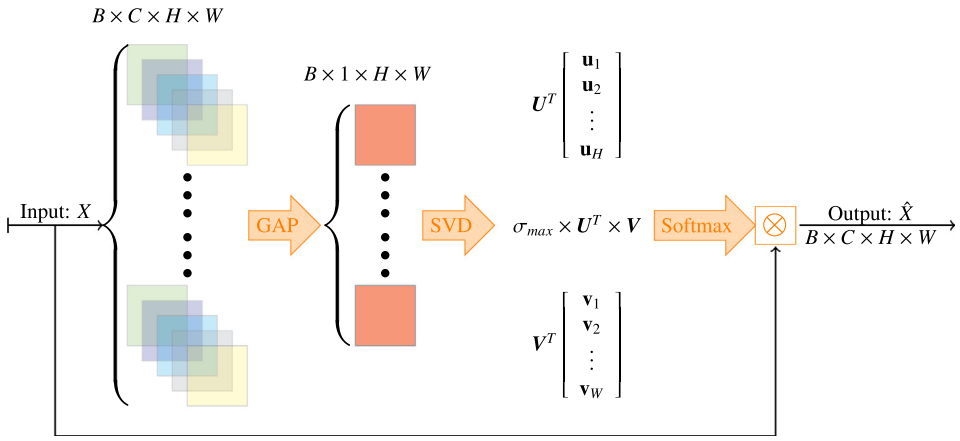

Fig. 2. Diagram of our spectral norm attention (SNA) block. Given the feature matrix obtained by global average pooling (GAP) along channel-wise, SNA generates spatial weights by using SVD to solve the spectral norm (σ_{max}) and its left and right singular vector ($\boldsymbol{U}, \boldsymbol{V}$).

rithm to solve the spectral norm and only perform multiplication on vectors to generate global attention. Generally, the time complexity of SVD in factorizing a matrix with $H \times W$ is O ($W^2 * H + W * H^2$), which is O(W^3). The power iteration algorithm starts with a random vector v, which may be an approximation to the dominant eigenvector. The most time-consuming operation of the algorithm is the multiplication of matrix $\overline{\boldsymbol{M}}$ by this vector. Based on repeatedly applying the matrix to an arbitrary starting vector and renormalizing, the time complexity is O(W^2). However, the power iteration algorithm may converge slowly. Hence, one premise of using power iteration for approximating the spectral norm in SVD is that the matrix has a dominant eigenvalue with a strictly greater magnitude than other eigenvalues. This premise guarantee that the power iteration algorithm converges to a reasonable result. The smaller is difference between the dominant eigenvalue and second eigenvalue, the longer it might take to converge. Especially, image matrices with spatial redundancy conform to this. In our experiments, the iteration number for converging is usually less than 10, and the average is 2. Hence, the power iteration algorithm in SNA can reduce the time complexity to O(W), just linear overhead, thus avoiding the dear overhead of computation.

3.3. Convolutional learners

Our SNA block can be seamlessly integrated into classical architectures of convolutional learners to discriminate feature weight. Fig. 3 shows that our SNA can be flexibly incorporated into ResNet [47] and DenseNet [48]. Both learners are the CNN cornerstone and the first choice paradigm of computer vision tasks. For high-resolution images with inter-pixel spatial redundancy, we first devote a convolutional layer and a max-pooling layer to extract high-level features from pixel-level attributes. We deliver the spatial redundancy to locally high-level features by an equivalent translation on a limited receptive field. After lots of convolutional or pooling layers, we can not preserve the spatial redundancy maximally. Hence, we impose SNA to enforce vision learners feature selection and weighting before residual and dense blocks. Following the max-pooling layer, vision learners still under-develop channel-wise features. Redundancy in spatial-wise still retains. Then, SNA performs GAP along the channel-wise to obtain feature matrices along the spatial-wise. Further, SNA decomposes feature matrices to get the spectral norm and the attention masks.

On the other hand, we consider the construction of optional SNS integration within residual and dense blocks. As the above analysis for the redundancy efficacy, how much spatial redundancy remains is immeasurable after considerable convolution layers. And the effectiveness of SNA is synchronized with redundancy. Hence, SNA is optional within residual and dense blocks. When facing heavily spatial redundancy, it is desirable to add SNA. It is worth mentioning that SNA consumes computing resources. If not necessary or bring negligible performance improvement, we suggest plugging in SNA only once before residual and dense blocks. Apart from the two popular convolutional learners, there are several viable ways for our SNA to be integrated into existing architectures. The flexibility of our SNA module entails that it can be directly grafted on any layer to remove spatial redundancy.

Our SNA aims to explicitly differentiate the importance of the informative and redundant features by SVD. Taking into account that SNA only requires one call to play the effect of feature classification and importance differentiating, our SNA is free from increasing model complexity. Compared to local attention blocks, such as SE, ECA, and so on, it is generally required to embed multiple times, thus bringing dear computing costs. According to specific tasks, embedding SNA in residual and dense blocks may be necessary. In this situation, we have to make a trade-off between performance and efficiency. With channel-wise features extending, the size of the spatial matrix also declines. Hence, even if adding SNA in residual and dense blocks, the training and inference computation complexity can not be prohibitive. Our SNA performs SVD on global feature maps, and its costs of matrix calculation are equal to other global attention blocks theoretically. However, considering SNA no need to acquire a precise solution for the spectral norm, we leverage the power iteration method to solve the approximation of the spectral norm. This algorithm significantly reduces the requirements of matrix calculation in SNA, thus helping control the computational consumption.

4. Experiments

4.1. Experimental settings

.

- **Datasets.** We conduct experiments on five image datasets, including CIFAR-100 [49], ImageNet [50,51], COCO2017 [52], VOC2012 [53] using the validation set for evaluation, and VIN-CXR [54] which localizes and classifies 14 types of thoracic abnormalities. We convert the data format of VOC2012 to the format of COCO2017, including generating mask labels, so we can perform experiments on VOC2012 using Mask R-CNN [55]. The input images of the CIFAR-100 dataset are randomly cropped to 32 × 32 and the input images of other natural data-

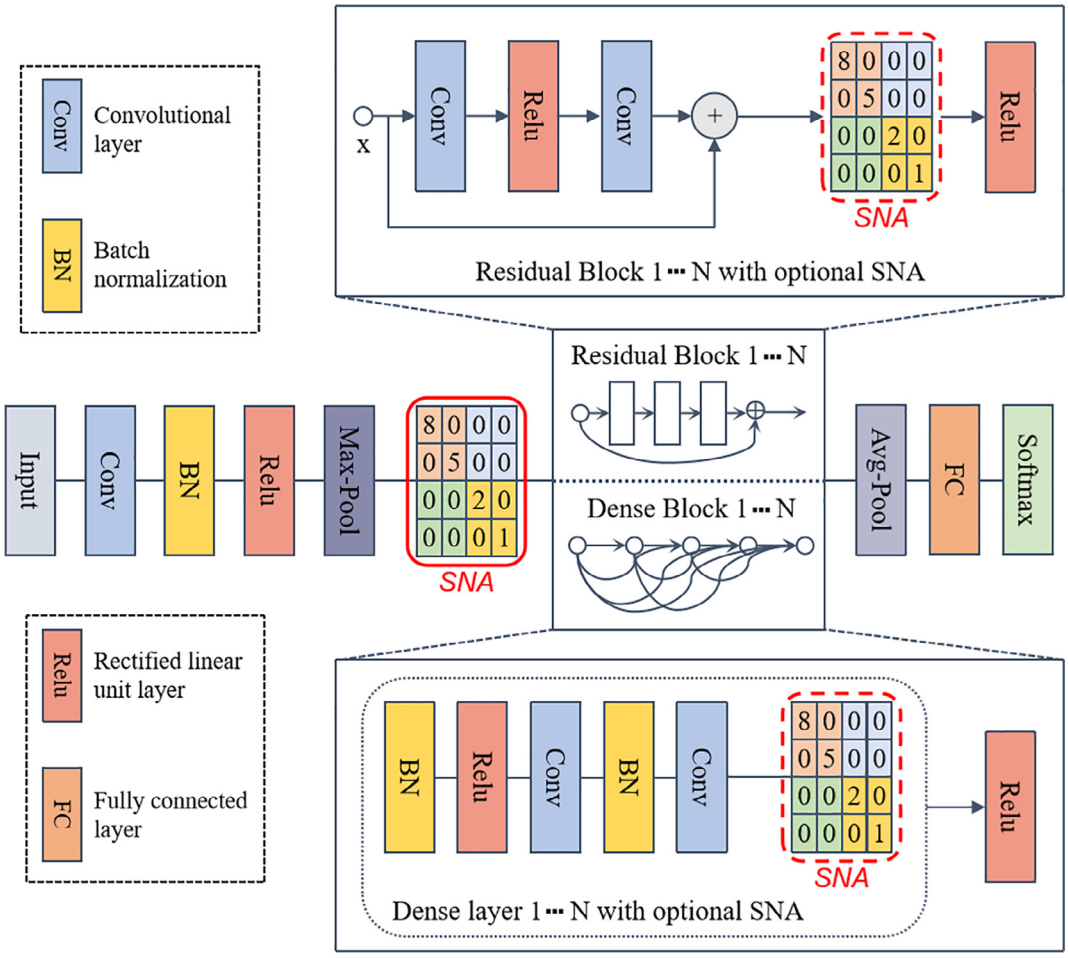

Fig. 3. Schema of ResNet and DenseNet integrating our SNA.

sets are randomly cropped to 224 × 224, both with random horizontal flipping and normalization. For medical images, we resize each image as 224 × 224 and normalize them to 0–1 range.

- **Methods**.We use ResNet [47] and DenseNet [48] as the backbones and Faster R-CNN [56] and Mask R-CNN [55] for the abnormalities detectors. Our spectral norm attention (SNA) is compared against three state-of-the-art attention methods, including two local attention modules (SE [17] and ECA [33]) and two global attention modules (SA [19] and AA [11]). If not specified, we lay our SNA into the residual and dense blocks for the VIN-CXR dataset and not for other natural datasets. Due to the large computation overhead of AA, we can not train detectors within our computation resources even if one image per GPU (1 GPU with 11 GB of memory). Hence, we do not report the performance of AA for object detection tasks.
- **Evaluation Metrics.** For the task of image classification, we use the Top-1/Top-5 accuracy to evaluate the performance on the natural datasets and the area under the receiver operating characteristic (AUROC) to compare the performance of different diseases of the medical datasets. For the task of object detection, we use average precision (AP) to evaluate the performance of the natural and medical datasets on different size of intersection over union (IoU). For example, we report AP for the VIN-CXR dataset at IoU $>$ 0.4.
- **Optimizer**. All models are trained from scratch by using Adam optimizer with a learning rate of 0.1 for natural images and 0.01 for medical images. The classifiers are trained within 100 epochs while the detectors are 20 epochs. Specifically, the learning rate is decay scheduled by setting a gamma of 1 and a step size of 10. We set the batch size as a multiple of 8 for all tasks, *i.e.*, 16 (8 GPUs with 2 input images per GPU). The parameters of networks are optimized with weight decay of 5e-4, momentum of 0.9. In particular, we use Detectron2 framework [57] to train COCO2017 and VOC2012 by following the same setting.

4.2. Image Classification

4.2.1. CIFAR-100 dataset

We first investigate how the complexity and performance of our SNA on the CIFAR-100 dataset, a standard benchmark for low-resolution imagery, by using ResNet-18 and DenseNet-121 backbones. According to the reports in Table 1, our SNA achieves the best Top-1 and Top-5 accuracy on the two backbones. And the second-highest performances are obtained by ECA and AA (underline) respectively. Although the global attention method (AA) exhibit better performance, their computation overhead is extremely large due to the global modeling of feature interdependencies. The metrics values of AA, including #.Param., FLOPs, and FPS, can demonstrate such a result by comparing to ECA and SE. The performance of another global attention (SA) is on par with the backbones while increasing model complexity. However, our SNA overcomes the paradox of performance and complexity trade-off. Compared to the backbones, our SNA does not add additional network parameters and has the same FLOPs. And the inference speed

Table 1
Comparison of different attention methods on the CIFAR-100 dataset using the ResNet-18 and DenseNet-121 backbones in terms of network parameters (#.Param., in M), floating point operations per second (FLOPs, in G), inference speed (frame per second, in FPS), and Top-1/Top-5 accuracy (in %).

Metrics	ResNet-18						DenseNet-121					
	-	SE	ECA	SA	AA	SNA (Ours)	-	SE	ECA	SA	AA	SNA (Ours)
#.Param.	10.71	10.80	10.71	12.03	13.07	**10.71**	6.74	6.75	6.74	6.86	14.25	**6.74**
FLOPs	32.64	32.64	32.64	39.51	52.25	**32.64**	52.45	52.45	52.45	61.24	87.82	**52.45**
Inference	**225**	186	202	153	76	217	**56**	31	33	24	13	51
Top-1	77.85	78.04	78.23	77.86	78.37	**79.17**	79.31	79.84	79.93	78.97	80.06	**80.81**
Top-5	93.69	93.65	93.91	93.58	93.76	**94.13**	94.83	94.79	94.91	94.72	95.03	**95.20**

is almost the same with the backbones without using any attention block. The lower model complexity of our SNA benefits from generating attention weights by using the method of redundancy removal.

Another research question is how the performance of our SNA on rank-k (k = 1,2,4,8,16) singular value. Since the direction of the spectral norm contains most of the important information. In this work, we only use the rank-1 singular value, *i.e.*, the spectral norm, to approximate the feature matrix. Then the approximated matrix is used as attentive scores. Here, as Fig. 4 shows, we compare the performance of the approximated matrix generated by different ranks of singular value. We can observe that the attention matrix using rank-2 singular value can bring the most gain. And the performance declines as the number of singular values increases. The reports illustrate that rank-2 singular values can best identify informative and redundant features. The attention matrix approximated by more singular values can almost recover the feature matrix, thus failing to impose attention. On the other hand, we employ a rank-1 singular value to approximate the feature matrix for balancing performance and efficiency. The computational complexity of rank-2 singular values is twice of rank-1.

4.2.2. ImageNet dataset

We next examine how the complexity and performance of our SNA on the ImageNet dataset, a standard large-scale dataset for high-resolution imagery, by using ResNet-50. Table 2 benchmarks our SNA against SE, ECA, and AA on the ResNet-50 backbone. We can again find that our SNA has comparable FLOPs with ResNet-50 and local attention modules and without increasing network parameters. As for performance, our SNA achieves the best Top-1 accuracy and second-highest Top-5 accuracy. In particular, our method achieves a 1.54% Top-1 accuracy improvement on ImageNet classification over a ResNet50 baseline and outperforms other attention mechanisms for images. In terms of efficiency, global attention blocks (SA and AA) inevitably increase model complexity, thus helping improve model expressivity. Conditioned on a dot-product similarity matrix as attentive scores, SA brings equal weights for informative and redundant features, thus only slightly outperforming the backbone. Further, by introducing relative positional encodings, AA can alleviate the even weighting, hence is superior to the backbone and local attention blocks (SE and ECA). Considering the self-attention mechanism brings performance improvement with sacrificing efficiency, our SNA can circumvent this issue by directly identifying and boosting the informative features. By combining reports of Table 1 and Table 2, we can demonstrate the effectiveness of our SNA on multi-class classification tasks for natural images. Our SNA enforces feature selection and weighting by SVD for convolutional learners, showing great potential in improving the performance without increasing model complexity.

4.2.3. VIN-CXR dataset

We last observe the performance of our SNA on the VIN-CXR dataset, a multi-label classification task. Each medical image is generated by the same imaging way in the same body region. Hence, compared to natural images, medical images vary heavy-

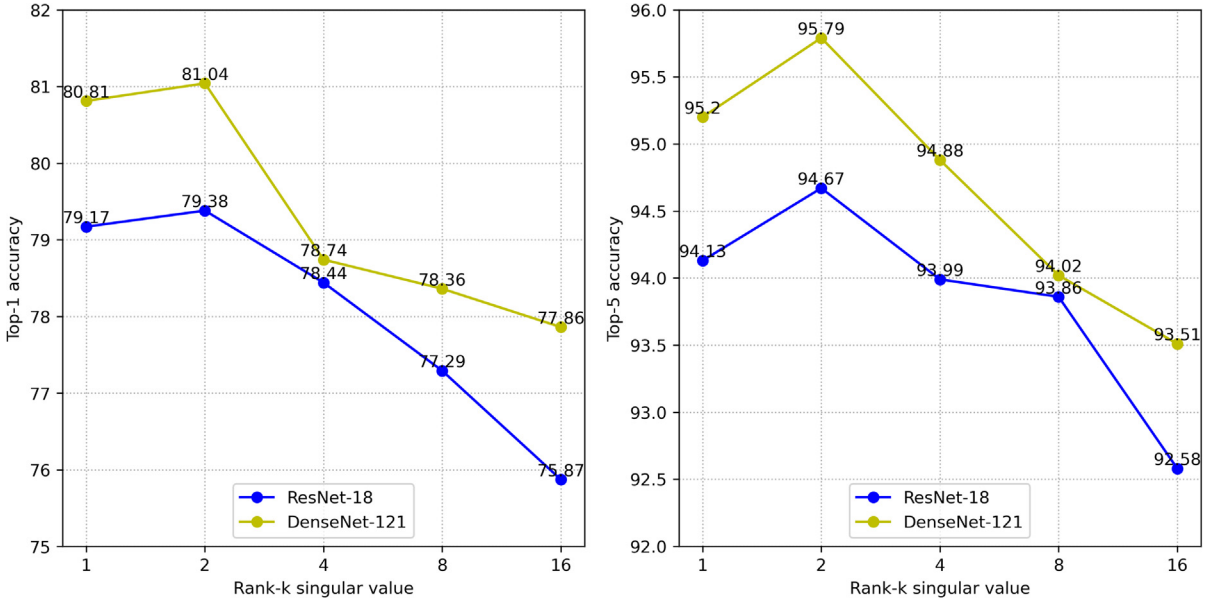

Fig. 4. Comparison of our SNA on rank-k (k = 1,2,4,8,16) singular value for the CIFAR-100 dataset using the ResNet-18 and DenseNet-121 backbones in terms of Top-1/Top-5 accuracy (in %).

Table 2
Comparison of different attention methods on the ImageNet dataset using the ResNet-50 backbone in terms of network parameters (#.Param., in M), floating point operations per second (FLOPs, in G), and Top-1/Top-5 accuracy (in %).

Method	#.Param.	FLOPs	Top-1	Top-5
ResNet-50	24.38	5.19	76.10	92.35
+ SE	26.79	5.19	76.69	92.97
+ ECA	24.38	5.19	76.75	93.24
+ SA	27.51	6.21	76.34	92.41
+ AA	28.43	7.17	76.94	**93.75**
+ SNA (Ours)	**24.38**	**5.19**	**77.27**	93.61

redundant. Such a property can better plays the effect of our SNA using the method of redundancy removal. As Fig. 5 shown, our SNA brings clear performance gain on 7 classes over other methods. It is worth mentioning that AA obtains the lowest average AUROC of 85.62 while DenseNet-121 is 87.41. The highest performances of 15 classes are obtained by our SNA, SE, ECA, and DenseNet-121, respectively, but no SA and AA. We argue that the lower information richness of medical images prevents AA from effectively capturing long-range interactions. In other words, informative features are drowned in redundant features. It is not very effective to capture long-range feature interdependencies for global attention in this situation. Differentiating from other global attention, our SNA can precisely suppress more redundant features while enhancing less informative features by redundancy removal approaches. The mechanism of our SNA performs SVD on the global feature matrix to locate useful features. The superiority of this mechanism can be proven by the above observations.

4.3. Object detection

4.3.1. COCO2017 and VOC2012 datasets

Based on the results of image classification, we can observe that our method has strong competitiveness on complexity and performance against other attention methods by experimenting on different datasets and different backbones. It is highly desirable to show its usefulness for other tasks of image processing. Here, we first validate the performance of our SNA on COCO2017 and VOC2012, two mainstream datasets for object detection. Using Faster R-CNN and Mask R-CNN detectors, we employ ResNet-50 along with FPN [58] as the backbone. As shown in Table 3, integration of attention blocks, including SE, ECA, SA, and our SNA, can improve the performance of object detection by a clear margin. Meanwhile, our SNA outperforms the ECA block by 4.91% and 4.07% in terms of *AP* using Faster R-CNN and Mask R-CNN on the COCO2017 dataset, respectively. And on the VOC2012 dataset, our SNA achieves the best performance in terms of *AP* and AP_{50} while ECA obtains the highest in terms of AP_{75}. The reports in Table 3 show that our SNA also has great potential to further improve the performance of the compared attention methods.

4.3.2. VIN-CXR dataset

We further explore the VIN-CXR dataset to verify the effectiveness of our SNA on the object detection task by using Faster R-CNN detector with ResNet-18 and DenseNet-121 backbones. As shown in Table 4, our SNA is superior to the ResNet-18 by 12.45% and the DenseNet-121 by 15.53% in terms of total *AP*. And the best performances of half of the 14 classes are obtained by our SNA. Although the classes of the best results obtained by our SNA on the two backbones are different, there is one commonality, which is that our SNA tends to achieve good results in small sample classes. For example, the incorporation of our SNA with the two backbones achieves the highest performance for the pneumothorax class which only has 81 samples for training and 15 samples for test. Such results demonstrate that the informative features of small sample classes are effectively exploited by incorporating our SAN with convolutional learners. Imbalanced data is a long-standing challenge for multi-class classification tasks. A distinct property of global attention is class-specific by capturing global dependencies, which can circumvent the class imbalance problem, while local attention can not. Unlike superfluous features, redundant features also make contributions to the global dependencies. Hence, current global attention methods (SA) can not reflect the advantage of addressing the class imbalance. But, our SNA can

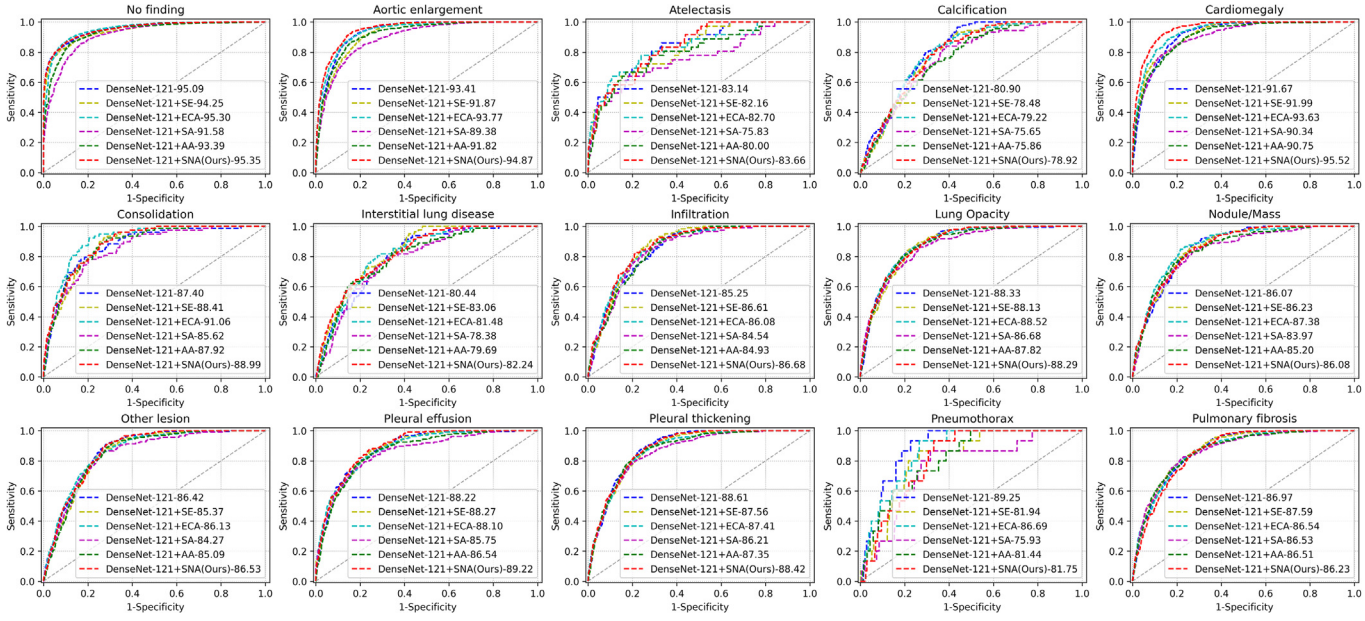

Fig. 5. Comparison of different attention methods on the VIN-CXR dataset using the DenseNet-121 backbone in terms of area under the receiver operating characteristic (AUROC, in %). The best result in 15 classes, our SNA occupies 7 (No findings, Aortic enlargement, Atelectasis, Cardiomegaly, Infiltration, Other lesion, Pleural effusion), remaining 7 classes scatter on different methods.

Table 3
Comparison of different attention methods on the COCO2017 and VOC2012 datasets using the Faster R-CNN [56] and the MASK R-CNN [55] detectors with the ResNet-50 backbone in terms of AP (in %) on different IoU values.

Backbones	Detectors	COCO2017						VOC2012		
		AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	*AP*	AP_{50}	AP_{75}
ResNet-50	**Faster R-CNN**	36.61	56.36	39.42	21.20	39.35	48.29	61.48	90.35	71.42
+ SE		37.45	57.45	40.33	21.21	40.11	49.21	62.44	90.77	71.29
+ ECA		38.11	58.24	41.03	22.41	40.95	50.29	62.80	90.84	**72.18**
+ SA		37.74	57.76	40.59	21.36	40.24	49.81	62.51	90.78	71.41
+ SNA (Ours)		**39.98**	**60.66**	**43.31**	**23.88**	**43.12**	**51.75**	**62.95**	**91.14**	71.63
ResNet-50	**Mask R-CNN**	37.28	57.17	40.50	21.35	39.73	50.02	62.21	90.83	71.95
+ SE		38.03	58.20	41.25	21.78	40.95	50.46	62.72	91.35	72.14
+ ECA		38.54	58.82	41.56	22.42	41.32	51.09	63.04	91.96	**72.35**
+ SA		37.95	58.31	41.32	21.85	41.12	50.49	62.78	90.41	72.28
+ SNA (Ours)		**40.11**	**60.45**	**43.69**	**23.83**	**43.10**	**52.36**	**63.52**	**92.20**	72.31

effectively stay away from redundant features. Hence, our SNA can help alleviate sample imbalance by enhancing highly informative features with global SVD.

We impose SNA on residual and dense blocks for medical images with heavily spatial redundancy to assemble convolutional learners. If not, the average AP of ResNet-18 is 62.31, and DenseNet-121 is 61.67. Combining the reports in Table 4, residual blocks with SNA can further prompt performance to 62.69, and dense blocks with SNA also facilitate implementation to 62.05. SNA in residual and dense blocks helps improve performance as redundancy remains active in the convolutional process. From the perspective of the redundancy degree, our SNA can gain more benefits for medical image processing. In essence, SNA before residual and dense blocks is most effective, and SNA in residual and dense blocks is also beneficial but accompanied by dear computation overhead. Another observation is that ResNet is superior to DenseNet, although the latter has more power model complexity and expressivity.

Another observation is that ResNet is superior to DenseNet, although the latter has more power model complexity and expressivity. This result is opposite to performances on natural images in Table 1. We have explored combating the spatial redundancy of medical images by factorizing convolutions in [59]. We prove that moderate model complexity can be better coped with medical images with heavy spatial redundancy. Thus, ResNet-18 can achieve better AP than DenseNet-121. This claim is the experience gained in the long-term practice of medical image processing. SNA is an exploration of combating spatial redundancy from the standpoint of features. By above comparison and analysis, we again demonstrate the superiority of our SNA, which can bring better performance without increasing computation complexity.

5. Discussions

Our SNA is a dedicated global attention block for convolutional learners with locality and shift-invariance to combat spatial redundancy. SNA globally identifies and weights informative and redundant features for each intermediate feature map by SVD. Based on the above experimental reports, we would like to discuss the pros and cons of our SNA.

From the viewpoint of eliminating inter-pixel redundancy, SNA is a unique exploration of combating spatial redundancy during vision learning for convolutional learners. Compared to preprocessing image methods executing redundancy removal before inputs, SNA is a task-agnostic and robust solution. The extensive experiments can demonstrate the effectiveness and availability of our SNA. By comparing against popular local and global attention methods, our SNA can achieve competitive performance. First, SNA lays emphasis on global attention for convolutional learners without sacrificing model efficiency. CNNs with our SNA can hold comparable training and inference speed to vanilla CNNs. By contrast, SA and AA bring better accuracy at cost of efficiency. We have witnessed the power expressivity of ViTs capturing global dependencies. Experiments show global attention (AA) can outperform local attention for convolutional learners. But the performance of SA only keeps pace with local attention blocks (SE and ECA) due to the self-attention mechanism can not precisely distinguish between informative and redundant features. Second, global atten-

Table 4
Comparison of different attention methods on the VIN-CXR dataset using the Faster R-CNN [56] detector with the ResNet-18 and DenseNet-121 backbones in terms of AP (in %) at IoU $>$ 0.4. ILD denotes interstitial lung disease, AE is aortic enlargement, PT is pleural Thickening, and PF is pulmonary fibrosis.

Diseases	Cases		ResNet-18					DenseNet-121				
	train	test	–	SE	ECA	SA	SNA	–	SE	ECA	SA	SNA
AE	2,433	634	98.42	**98.76**	97.66	98.69	98.42	98.21	98.90	**98.97**	98.59	97.46
Atelectasis	150	36	31.67	38.33	45.00	26.67	**55.00**	33.33	**50.00**	40.00	23.33	43.33
Calcification	364	88	20.88	19.23	14.29	14.29	**26.37**	22.72	19.32	21.59	16.09	**25.27**
Cardiomegaly	1,855	445	99.14	99.52	98.95	**99.71**	98.85	99.14	**99.81**	99.71	99.57	99.43
Consolidation	275	78	56.30	63.87	68.07	65.63	**75.63**	56.30	71.43	68.91	**74.79**	70.59
ILD	304	82	58.22	61.03	**79.34**	68.54	71.36	77.46	75.12	**83.10**	71.03	68.54
Infiltration	492	121	48.28	51.72	64.75	**67.01**	54.02	52.49	64.75	65.13	**70.11**	62.84
Lung opacity	1,053	269	52.33	54.26	56.59	60.74	**61.43**	57.56	59.88	59.11	60.55	**61.24**
Nodule/Mass	655	171	**38.88**	35.70	33.27	27.29	36.07	16.26	22.24	21.68	30.28	**37.38**
Other lesion	903	231	30.30	35.38	32.42	36.02	**39.62**	25.64	21.82	19.92	32.84	**34.96**
Pleural effusion	818	214	78.35	80.12	83.86	82.45	**84.45**	75.59	80.31	78.54	80.12	**84.65**
PT	1,590	391	63.36	**70.88**	65.14	67.56	67.85	64.30	72.03	**73.90**	63.88	68.06
Pneumothorax	81	15	55.56	40.00	62.22	42.22	**62.22**	33.33	51.11	42.22	37.78	**62.22**
PF	1,316	301	48.74	46.81	**53.91**	46.68	46.33	39.59	48.13	50.90	45.35	**52.71**
Total	12,289	3,076	55.75	56.82	61.11	57.39	**62.69**	53.71	59.63	58.83	57.45	**62.05**

tion is class-specific and can alleviate sample imbalance. We have discussed this benefit according to the reports in Table 4.

We make an ablation study on our SNA to observe the performance of rank-k (k = 1,2,4,8,16) singular values used for approximating the feature matrix. Considering the trade-off between performance and efficiency, we utilize rank-1 singular value, *i.e.*, the spectral norm, to recover the feature matrix and use it as an attention matrix. We also observe the performances of residual and dense blocks with and without SNA. The experimental results prove our claim that SNA is optional during the long-range convolutional process.

Generally, natural images have rich information used to drive vision learners while medical images are relatively poor. From the experimental results of object detection tasks, we can find that our SNA can bring more promotion to medical images than natural images. Such observation can demonstrate that our SNA can better help convolutional learners effectively exploit informative features for medical images with heavy redundancy. Benefiting from using the method of redundancy removal, our SNA can identify informative features and give them higher attention weights. On the other hand, the pixel-level features of medical images are highly homogenous. Such a property leads to the directions of informative features more concentrating, thus helping our SNA better capture global interactions.

There are two limitations to our SNA. First, in terms of performance, there is no significant improvement for natural images and a shortage of evidence for addressing the class imbalance. Second, from the perspective of interpretability, due to the attention mechanism using SVD implicitly calculating attentions, we can not visualize and explain the attention maps. SVD is used to implicitly locate dominant eigenvectors or informative features from global feature matrices, then SNA is dedicated to enhancing their contributions. It is intractable to visualize informative features. In the future, we intend to combat inter-pixel spatial redundancy by building attention blocks similar to SNA based on other matrix decomposition methods to explain performance improvement better.

6. Conclusions

Attention mechanisms have demonstrated their impressive progress in improving the performance of convolutional learners. In this work, we study how to exploit informative features by introducing a novel attention mechanism using the method of redundancy removal. We utilize the spectral norm of a feature matrix to calculate the attentive scores to give different weights for informative features and redundant features. We extensively evaluate our SNA on the tasks of image classification and object detection to demonstrate the effectiveness of our SNA. Our SNA can help CNNs further improve performance on four mainstream natural datasets and a medical dataset compared to the state-of-the-art attention modules. The most valuable is that we try to explore capturing global interaction by SVD during visual learning.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Subramanya, Image compression technique, IEEE Potentials 20 (1) (2001) 19–23.

[2] Y. Wang, K. Lv, R. Huang, S. Song, L. Yang, G. Huang, Glance and focus: a dynamic approach to reducing spatial redundancy in image classification, Adv. Neural Inform. Process. Syst. 33 (2020) 2432–2444.

[3] Z. Xie, Z. Zhang, X. Zhu, G. Huang, S. Lin, Spatially adaptive inference with stochastic feature sampling and interpolation, in: European conference on computer vision, Springer, 2020, pp. 531–548.

[4] L. Wolf, A. Shashua, D. Geman, Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weight-based approach, J. Mach. Learn. Res. 6 (11) (2005).

[5] Z. Li, J. Tang, Unsupervised feature selection via nonnegative spectral analysis and redundancy control, IEEE Trans. Image Process. 24 (12) (2015) 5343–5355.

[6] W. Chen, M.J. Er, S. Wu, Pca and lda in dct domain, Pattern Recogn. Lett. 26 (15) (2005) 2474–2482.

[7] Z. Pan, A.G. Rust, H. Bolouri, Image redundancy reduction for neural network classification using discrete cosine transforms, in: Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, Vol. 3, IEEE, 2000, pp. 149–154.

[8] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, arXiv preprint arXiv:2111.06377 (2021).

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[10] M. Arar, A. Shamir, A.H. Bermano, Learned queries for efficient local attention, 2021, arXiv preprint arXiv:2112.11435.

[11] I. Bello, B. Zoph, A. Vaswani, J. Shlens, Q.V. Le, Attention augmented convolutional networks, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3286–3295.

[12] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

[13] Y. Chen, Y. Kalantidis, J. Li, S. Yan, J. Feng, A 2-nets: Double attention networks, Adv. Neural Inform. Process. Syst. 31 (2018) 352–361.

[14] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.

[15] H. Zhao, J. Jia, V. Koltun, Exploring self-attention for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10076–10085.

[16] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, et al., Resnest: Split-attention networks, arXiv preprint arXiv:2004.08955 (2020).

[17] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[19] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.

[20] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, J. Shlens, Stand-alone self-attention in vision models, Adv. Neural Inform. Process. Syst. 32 (2019).

[21] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, A. Vaswani, Bottleneck transformers for visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16519–16529.

[22] C.-F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.

[23] F. Babiloni, I. Marras, G. Slabaugh, S. Zafeiriou, Tesa: Tensor element self-attention via matricization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13945–13954.

[24] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, arXiv preprint arXiv:2201.03545 (2022).

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[26] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: Advances in neural information processing systems, 2014, pp. 2204–2212.

[27] G. Cheng, p. lai, D. Gao, J. Han, Class attention network for image recognition, Sci. China Inform. Sci. (2022). doi:https://doi.org/10.1007/s11432-021-3493-7.

[28] J. Hu, L. Shen, S. Albanie, G. Sun, A. Vedaldi, Gather-excite: exploiting feature context in convolutional neural networks, in: Proceedings of the 32nd International Conference on Neural Information Processing Systems, 2018, pp. 9423–9433.

[29] J. Park, S. Woo, J.-Y. Lee, I.S. Kweon, Bam: Bottleneck attention module, arXiv preprint arXiv:1807.06514 (2018).

[30] Z. Gao, J. Xie, Q. Wang, P. Li, Global second-order pooling convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3024–3033.

[31] H. Li, Channel locality block: A variant of squeeze-and-excitation, arXiv preprint arXiv:1901.01493, 2019.

[32] Z. Li, Y. Sun, L. Zhang, J. Tang, Ctnet: Context-based tandem network for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2021).

[33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: efficient channel attention for deep convolutional neural networks, in: 2020 ieee, in: CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2020.
[34] B. Yang, L. Wang, D. Wong, L.S. Chao, Z. Tu, Convolutional self-attention networks, arXiv preprint arXiv:1904.03107 (2019).
[35] A.W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi, Q.V. Le, Qanet: Combining local convolution with global self-attention for reading comprehension, in: International Conference on Learning Representations, 2018.
[36] D. So, Q. Le, C. Liang, The evolved transformer, in: International Conference on Machine Learning, PMLR, 2019, pp. 5877–5886.
[37] X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, H. Liu, Expectation-maximization attention networks for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9167–9176.
[38] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, Z. Lin, Is attention better than matrix decomposition?, arXiv preprint arXiv:2109.04553 (2021).
[39] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2008, pp. 426–434.
[40] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer (8) (2009) 30–37.
[41] A. Singh, P. Rajan, A. Bhavsar, Svd-based redundancy removal in 1-d cnns for acoustic scene classification, Pattern Recogn. Lett. 131 (2020) 383–389.
[42] L. Zhu, H. Song, X. Zhang, M. Yan, T. Zhang, X. Wang, J. Xu, A robust meaningful image encryption scheme based on block compressive sensing and svd embedding, Signal Processing 175 (2020) 107629.
[43] N.S. Keskar, J. Nocedal, P.T.P. Tang, D. Mudigere, M. Smelyanskiy, On large-batch training for deep learning: Generalization gap and sharp minima, in: 5th International Conference on Learning Representations, ICLR 2017, 2017.
[44] Y. Yoshida, T. Miyato, Spectral norm regularization for improving the generalizability of deep learning, arXiv preprint arXiv:1705.10941 (2017).
[45] T. Miyato, T. Kataoka, M. Koyama, Y. Yoshida, Spectral normalization for generative adversarial networks, in: International Conference on Learning Representations, 2018.
[46] M. Khodak, N. Tenenholtz, L. Mackey, N. Fusi, Initialization and regularization of factorized neural layers, arXiv preprint arXiv:2105.01029 (2021).
[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[48] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.
[49] A. Krizhevsky, G. Hinton, et al., Learning multiple layers of features from tiny images (2009).
[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, 2009 IEEE conference on computer vision and pattern recognition, Ieee 2009 (2009) 248–255.
[51] S. Kornblith, J. Shlens, Q.V. Le, Do better imagenet models transfer better?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp 2661–2671.
[52] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, Springer, 2014, pp. 740–755.
[53] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, Int. J. Comput. Vision 111 (1) (2015) 98–136.
[54] H.Q. Nguyen, K. Lam, L.T. Le, H.H. Pham, D.Q. Tran, D.B. Nguyen, D.D. Le, C.M. Pham, H.T. Tong, D.H. Dinh, et al., Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations, arXiv preprint arXiv:2012.15029 (2020).
[55] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
[56] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Informa. Process. Syst. 28 (2015) 91–99.
[57] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, URL: https://github.com/facebookresearch/detectron2 (2019).
[58] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.
[59] M. Zeng, N. Zeng, J. Fang, J. Liu, Factorized convolution with spectral normalization for fundus screening, 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), IEEE 2022 (2022) 1–5.

Jiansheng Fang is a Ph.D. candidate in the School of Computer Science and Technology at Harbin Institute of Technology and a researcher at CVTE. His main research interests include computer vision, medical image processing, and image retrieval.

Dan Zeng received B.E. and Ph.D. in computer science and technology from Sichuan University in 2013 and 2018. From 2018 to 2020, she worked as a post-doc research fellow in the Data Management and Biometrics Group at the University of Twente, the Netherlands. She is currently a research assistant professor at the Southern University of Science and Technology. Her main research interests include image processing, biometrics, and deep learning.

Xiao Yan obtained his Ph.D. in 2020 from the Chinese University of Hong Kong and is currently a research assistant professor in the Department of Computer Science and Engineering at the Southern University of Science and Technology. His research interests include large-scale machine learning, algorithms and systems for database, and especially large-scale vector search.

Yubing Zhang is a researcher manager and technical expert at the Machine Vision Institute of CVTE Research. He won the 20th China Invention Patent Excellence Award and the 2011 U.S. College Student Mathematical Modeling Outstanding Award (top 1%). Besides, he also got the Second Prize for the CVTE Invention Patent Quality Gold Award and the CVTE Founder's Innovation Award. His research interests include face recognition, digital human, and metaverse.

Hongbo Liu is the Director of Machine Vision Institute at CVTE Research. Before joining CVTE, he worked as the PM of the System Architecture Department at Hisilicon. His research interests include machine learning-based ISP image processing, algorithms of video codec, and hardware-software architecture for computer vision systems.

Bo Tang received his Ph.D. in computer science from The Hong Kong Polytechnic University in 2017. He is currently an assistant professor at the Southern University of Science and Technology. He won ACM SIGMOD China Rising star 2021. His research interests include query optimization and data-intensive system.

Ming Yang is the CTO with CVTE (002841.SZ) and serves as the Director of CVTE Research. Before joining CVTE, he received a B.E. (2009) and a Ph.D. (2014) in Computer Science from Sun Yat-sen University. His research interests include machine learning and computer vision.

Jiang Liu obtained his Ph.D. in 2004 from the Department of Computer Science of the National University of Singapore and is currently a full professor in the Department of Computer Science and Engineering at the Southern University of Science and Technology. His main research interests include medical image processing and artificial intelligence.