Dense Attention Convolutional Network for Image Classification

Han Zhang¹, Kun Qin¹*, Ye Zhang¹, Zhili Li² and Kai Xu²

¹School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, Hubei, 430079, China
²School of Geography and Information Engineering, China University of Geosciences, Wuhan, Hubei, 430074, China

*Corresponding author’s e-mail: qink@whu.edu.cn

Abstract. Convolutional neural networks (CNN) have made rapid progress for a series of visual tasks, but the bottom-up convolutional feature extraction process is not qualified to mimic the human visual perception process, which takes advantages of discriminative features. Although attention modules extract the top-down discriminative features and have been widely investigated in the past few years, current attention modules interrupt the bottom-up convolutional feature extraction process. To tackle this challenge, in this paper, we introduce dense connection structure to fuse discriminative features from attention modules and convolutional features, which we term as dense attention learning. Also, to alleviate the over-fitting problem caused by rapid feature dimension growth, we propose a channel-wise attention module to compress and refine the convolutional features. Based on these strategies, we build a dense attention convolutional neural network (DA-CNN) for visual recognition. Exhaustive experiments on four challenging datasets including CIFAR-10, CIFAR-100, SVHN and ImageNet demonstrate that our DA-CNN outperforms many state-of-the-art methods. Moreover, the effectiveness of our dense attention learning and channel-wise attention module is also validated.

1. Introduction
Convolutional neural Network (CNN) has boosted the performance of a series of visual tasks, and has drawn significant research interests[1-5].

It is often acknowledged that deeper network structure can enhance the feature representation capability and the performance is further improved. For example, AlexNet[6] has only five convolutional layers, while later VGG[7] and GoogLeNet[8] has 19 and 22 convolutional layers respectively. More impressively, the recently-developed residual connection structure[9-10] and dense connection structure[11-12] enhances the feature propagation and the convolutional network structure can be extended into hundreds or even thousand layers.

Current convolutional neural networks extract and generalize the features from low-level to high-level. However, a major weakness is that these CNN models are not qualified enough to mimic the human vision perception process[13]. This perception process is capable of capturing the discriminative features in a top-down manner. For example, for a scene where there is a duck on the lake, current CNN models tend to first extract the texture and structural features and then generalize them into semantic level, while for human they can find the duck at the first glance.
Towards this challenge, attention module has been widely utilized in the past few years[14-16]. For visual tasks, although many attention modules have been investigated, as is clearly pointed out in [17], a major drawback of current attention modules is that the top-down discriminative feature extraction process interrupts the traditional bottom-up convolutional feature extraction process.

To solve this problem, in this paper, we introduce a dense attention learning strategy (demonstrated in Figure 1) and propose a dense attention convolutional neural network (DA-CNN). In contrast with current CNN models, its dense attention block allows both top-down discriminative features and bottom-up convolutional features equally contribute to the following feature propagation. Moreover, with our dense attention learning strategy, it is easy for a CNN model to go deeper with fewer parameters.

The contribution of our work can be summarized as follows:

- We propose a dense attention learning strategy to fuse top-down discriminative features from attention modules and bottom-up features from convolutional layers. In a dense attention block, the current input is all the stacking of all the previous attention features and convolutional features, and it serves as the input for all the preceding layers.
- We propose a new convolutional neural network architecture named Dense Attention Convolutional Network (DA-CNN). It is organized by our dense attention learning strategy and its basic component is the dense attention block. Notably, it offers a stronger feature representation capability with less parameters and shallower depth.
- We propose a channel-wise attention module. It firstly highlights important channels and then compresses the convolutional feature dimensions.
- We conduct experiments on four benchmarks including CIFAR-10, CIFAR-100, SVHN and ImageNet to demonstrate that our DA-CNN outperforms many state-of-the-art methods.

2. Related Work

2.1. CNN Architecture

One of the explanations for the impressive performance of CNN is the utilization of convolutional operations in the forward propagation[6], which offers a much stronger feature representation capability compared with traditional hand-crafted features[18-19].

When the convolutional layer number increases, a CNN model is more capable of learning multi-level features and the feature representation capability is enhanced[7-8]. However, as the model goes deeper, it may become more difficult for the preceding layers to receive the information from the previous layers. This problem is often termed as gradient vanishing[20], which often leads to the phenomenon that a deeper CNN model performs worse than a shallower CNN model.

As is mentioned above, residual connection[9-10] and the recently-developed dense connection [11-12] tackled this challenge from the perspective of network architecture. Residual connection structure has served as a good backbone for many visual tasks[21-23] till now. In contrast, dense connection structure takes the output of all the previous layers as the input of all the preceding layers and the feature propagation is further enhanced[12]. In fact, it outperforms the ResNet for visual classification task and is drawing increasingly attention in the past few years[11,24,25].
2.2. Spatial Attention
Visual attention modules have been extensively investigated in the past few years[14,16,17]. Many of these approaches are based on spatial attention mechanism.

To be specific, Sun et.al. proposed a ROI attention module[27] which has more activation functions to highlight the key regions in an image. Similarly, Xu et.al. proposed a two hop attention module to further mine the spatial relation in an image[28]. It consists of two convolutional layers each followed by a softmax function. Moreover, Seo et.al proposed a hierarchical attention module to take multi-scale information into account[29].

To sum up with, recent studies indicate that the improvement of spatial attention modules usually lies in the following two aspects. The first is to introduce activation functions to enhance the non-linear feature presentation capability[24,27,28]. The second is to extract more discriminative convolutional features for attention modules by designing more complicated CNN structures [14,17,28,29].

2.3. Channel Attention
Recently, channel-wise attention modules have also been exhaustively investigated and proved effective in a series of visual tasks[14,30,31].

One of the classic channel attention modules is the squeeze and excitation network (SENet)[30]. The process of reducing feature dimensions is termed as squeeze while the process of increasing feature dimensions is termed as excitation.

Based on the former work, current approaches to develop more advanced channel attention modules lie in the following two aspects. The first is to utilize more complicated network structure to further mine the channel correlation [31-33]. The second is to learn both channel and spatial attention in a module [14,34,35].

2.4. Fusing attention features and convolutional features
One of the major drawbacks of current attention modules is that its discriminative feature extraction process interrupts the bottom-up convolutional feature extraction process[13,17].

In some applications such as visual question answering[36-37], one commonly-utilized strategy to tackle this challenge is to fuse features from both attention modules and CNN models.

However, for visual recognition, few work has tackled this challenge. Notably, a residual attention learning strategy has been proposed to fuse these two kinds of features[17]. The proposed residual attention network (RANet) outperforms many state-of-the-art methods and it has more layers with fewer parameters. Inspired by its impressive performance, later some studies also adapt this residual attention learning strategy[38].

3. Dense Attention Convolutional Network
3.1. Network Overview
Generally speaking, our DA-CNN is constructed via our dense attention learning strategy. As is demonstrated in Figure 2. It has a convolutional layer followed by pooling, three dense attention blocks, a forth dense block and finally a classification layer.

Figure 2. Demonstration of our dense attention convolutional neural network (DA-CNN).
The first convolutional layer followed by pooling aims to extract convolutional features and down-sample the images. Later, each dense attention block repeatedly extracts, refines and down-samples convolutional features. Finally, a classification layer, consisting of a global average pooling, a fully connected layer and a softmax classifier, finishes the output of predicted label.

The key of our DA-CNN is the three dense attention blocks. Each dense attention block consists of the following components, that is, a dense block, a dense attention learning module, a channel attention module and finally an average pooling operation.

### 3.2. Dense Attention Learning

Let $X$ and $H(\cdot)$ denote inputted convolutional feature map and a spatial attention module. Then, a set of attention weights $a_{ij}$ are calculated via

$$a_{ij} = H(X),$$

(1)

where weight $a_{ij}$ corresponds to the position $(i,j)$ of the inputted convolutional feature map $X$.

For current CNN models[24,27,28,29], the outputted convolutional feature map $X'$ from the attention module is calculated via

$$X' = H(X) \odot X',$$

(2)

where weight $a_{ij}$ corresponds to the position $(i,j)$ of the inputted convolutional feature map $X$.

Since $X'$ is top-down discriminative features from attention module $H(\cdot)$, it interrupts the bottom-up convolutional feature extraction process, where the features are gradually generalized from low-level to high-level.

Recently, dense connection structure is reported effective to enhance feature propagation and it outperforms residual connection structure[12]. Inspired by this, we introduce a dense attention learning strategy to fuse top-down discriminative features and bottom-up convolutional features. A basic unit to implement our dense attention learning is a dense attention module.

In a dense attention module, a one-layer spatial attention module $H(\cdot)$ follows a composite function. The composite function can be denoted as BN-ReLU-Conv(1×1)-BN-ReLU-Conv(3×3)[12], where BN, ReLU and Conv denotes the batch normalization, ReLU activation function and convolutional operation respectively.

Assume that before our targeted convolutional layer, there are already $l$ composite functions and $l$ one-layer attention module, then the input convolutional feature $X'$ for our targeted layer is organized as

$$X' = [X_1, H(X_1) \odot X_1, \cdots, X_l + H(X_l) \odot X_l],$$

(3)

where $[\cdot, \cdot]$ denote the feature staking operation.

Moreover, as is demonstrated in Figure 2, while our dense attention module refines the convolutional features, at the same time, the inputted feature is sent into a soft mask brunch $T(X)$ proposed in [17] to learn an attention mask so that the entire model converges faster.

### 3.3. Channel Attention Module

However, simply fusing densely connected convolutional features and discriminative features from spatial attention modules leads to the rapid growth of feature dimensions and it is likely to cause the over-fitting problem. To solve this problem, we propose a channel-wise attention module and insert it at the end of each dense attention module.
Assume the outputted convolutional feature $X'$ from a dense attention module has the size of $W \times H \times C$, where $W$, $H$ and $C$ denotes the width, height and channel number. Then, the convolutional feature $X'$ can be regarded as $C W \times H$ feature maps, which we denote as $X'_1, X'_2, \cdots, X'_c$.

For our channel-wise attention module, first we implement mean pooling for each channel to calculate the mean $v_i$ of feature vector $X'_i$. Let $V$ denote the output of channel-wise mean pooling of $X'$, then it can be represented as

$$V = [v_1, v_2, \cdots, v_{c-1}, v_c].$$ (4)

Later on, as is demonstrated in Figure 3, we utilize a $1 \times 1$ convolutional layer with weight $W_c$ and bias $b_c$ followed by a ReLU activation function (denoted as $relu$ below) to extract the attention weight $\beta_i$, where $i = 1, 2, \cdots, c$. It can be represented as

$$\{\beta\} = \text{softmax}(relu(W_c V + b_c)),$$ (5)

where $\text{softmax}$ denotes the softmax function.

Then, similar to equation (3), the point-wise product operation between channel attention weights $\beta$ and convolutional feature $X'$ highlights the information from those channels contributing significantly to the semantic label. Let $X''$ denote the outputted convolutional feature, then it is calculated via

$$X'' = \{\beta\} \odot X'.$$ (6)

Finally, we implement feature compression for the convolutional feature $X''$. By utilizing a $1 \times 1$ convolutional layer, its channel number $C$ is reduced into half. This effort further relieves the feature redundancy caused by staking convolutional features and top-down attention maps repeatedly.

### 3.4. Dense Attention Block

As is mentioned above, our dense attention block consists of the following components, that is, a dense block, a dense attention module, a channel-wise attention module, and finally a mean pooling operation.

#### 3.4.1 Dense Block

A dense block consists of several composite functions and the corresponding connections, which is clearly demonstrated in [12]. It extracts convolutional features for the later dense attention learning.
3.4.2 Dense Attention Module. Our dense attention module consists of two branches. In the first branch, as is demonstrated in section 3.2, composite functions and spatial attention modules are placed in turn, and all convolutional features and discriminative attention features are fused via dense connection. In the second branch, we utilize a soft mask branch in [17] for the features outputted from the aforementioned dense block to extract another attention mask. Then, this mask is fused with the features from the first branch via pot-wise product and staking operation.

3.4.3 Channel-wise attention module. After the dense attention module, it is important for us to compress and refine the features to avoid the over-fitting problem.

3.4.4 Mean pooling operation. This pooling operation down-samples feature maps for the next dense attention block.

Clearly, DA-CNN has two important hyper-parameters. The first is the composite function number in a dense block at the beginning of a dense attention module, which we term as $p$. The second is the total number of composite functions and attention modules in a dense attention module, which we term as $d$. Note that $d$ must be even since each composite function is followed by a one-layer spatial attention module. Figure 2 demonstrates the situation when \{$p=2, d=4$\}.

4. Experiments
Our proposed DA-CNN is validated on four challenging benchmarks, that is, CIFAR-10[39], CIFAR-100[39], SVHN[40] and ImageNet[5] respectively, to answer the following questions: Q1 How is the performance of our DA-CNN compared with the state-of-the-art methods? Q2 Is our dense attention learning effective to improve the classification performance? Q3 Is our channel-wise attention module effective to improve the classification performance?

4.1 CIFAR Classification

4.1.1 Implementation. Both CIFAR-10 and CIFAR-100 dataset has 50000 images for training and 10000 images for testing, while samples are equally divided into 10 and 100 categories respectively. All samples are three-channel 32×32 color images. The data normalization and data augmentation process follows the previous work[9,12,31].

4.1.2 Parameter setting. The widely utilized DenseNet[12] serves as the baseline method. For fair comparison, we do not focus on hyper-parameter tuning. Instead we keep most of the parameter settings the same with DenseNet. To be specific, the model is trained for 300 epochs with a batch size of 64. The initial learning rate is 0.1 and is divided by 10 at epoch 150 and 225. Moreover, the optimizer is the stochastic gradient descent (SGD). Following [12,41], the weight decay and the Nesterov momentum is set $10^{-4}$ and 0.9 respectively. The weight initialization follows the setting in [42]. The dropout rate is set 0.2 for all the convolutional layers.

4.1.3 Other details. We build three dense attention networks with the hyper-parameter setting of \{$p=3, d=6$\}, \{$p=6, d=8$\} and \{$p=8, d=12$\} and term them as DA-CNN-60, DA-CNN-93 and DA-CNN-145 respectively.

Following the setting in [12], each composite function and attention module in a dense attention block has the growth rate of 12, and the first convolutional layer has 16 channels.

4.1.4 Comparison with state-of-the-art methods. We compare our DA-CNN with some state-of-the-art methods including ResNet[9], WideResNet[43] (denoted as WRN), Residual Attention Network[17] and DenseNet[12] in terms of the parameter number (denoted as para) and the test error (denoted as err C10 and err C100). The results are listed in Table 1. Our DA-CNN-145 outperforms all other methods with a moderate model size. Meanwhile, the performance of our DA-CNN-60 and DA-CNN-
93 is also impressive with a relatively small model size. It suggests that our DA-CNN is capable of improving classification performance while reducing the number of parameters.

### Table 1. Comparison with state-of-the-art methods on CIFAR-10/100.

| Network         | para×106 | err C10 | err C100 |
|-----------------|----------|---------|----------|
| ResNet-164[9]   | 1.7      | 5.46    | 24.33    |
| ResNet-1001[9]  | 10.3     | 4.64    | 22.71    |
| WRN-16-8[43]    | 11.0     | 4.81    | 22.07    |
| WRN-28-10[43]   | 36.5     | 4.17    | 20.50    |
| RA-Net-92[17]   | 1.9      | 4.99    | 21.71    |
| RA-Net-236[17]  | 5.1      | 4.14    | 21.16    |
| RA-Net-452[17]  | 8.6      | 3.90    | 20.45    |
| DenseNet-40[12] | 1.0      | 5.24    | 24.42    |
| DenseNet-100[12]| 7.0      | 4.10    | 20.20    |
| DenseNet-250[12]| 15.3     | 3.62    | 17.60    |
| DA-CNN-60       | 2.0      | 4.96    | 24.03    |
| DA-CNN-93       | 5.1      | 3.87    | 19.61    |
| DA-CNN-145      | 9.8      | 3.24    | 17.15    |

4.1.5 Influence of attention module. To investigate the influence of our dense learning strategy and our channel-wise attention module, we report the test error (denoted as err) of our DA-CNN under the situations where dense attention learning strategy is removed (denoted as DA-CNN#DA) and channel-wise attention module is removed (denoted as DA-CNN#CA).

Table 2 lists the classification results on CIFAR-10/100 datasets. It can be seen that when there is no dense attention learning strategy, the performance significantly decreases. Indeed, attention modules repeatedly interrupt the bottom-up convolutional feature extraction process. Also, without dense connection structure, the feature propagation capability is not strong enough.

### Table 2. Influence of our dense attention learning strategy and channel-wise attention module on CIFAR-10/100 and SVHN datasets.

| Network         | err C10 | err C100 | err SVHN |
|-----------------|---------|----------|----------|
| DA-CNN-60#DA    | 5.17    | 24.34    | 1.88     |
| DA-CNN-60#CA    | 5.09    | 24.18    | 1.85     |
| DA-CNN-60       | 4.96    | 24.03    | 1.81     |
| DA-CNN-93#DA    | 4.21    | 20.27    | 1.82     |
| DA-CNN-93#CA    | 4.13    | 20.04    | 1.78     |
| DA-CNN-93       | 3.87    | 19.61    | 1.74     |
| DA-CNN-145#A    | 3.75    | 18.06    | 1.75     |
| DA-CNN-145#DA   | 3.56    | 17.78    | 1.73     |
| DA-CNN-145      | 3.24    | 17.15    | 1.71     |

Moreover, when there is no channel-wise attention module, the feature dimension grows too rapidly after each dense attention block. It may explain why the results of DA-CNN#CA are worse when compared with our DA-CNN.
4.2 SVHN Classification

4.2.1 Implementation. The Street View House Number (SVHN) dataset[40] contains 73257 training samples, 26032 testing samples. Each sample is a three-channel 32×32 colored image. Following the setting in [12,31], we do not augment the dataset, and all images are divided by 255 for normalization.

4.2.2 Parameter setting. The model is trained for 40 epochs with a batch size of 64. The initial learning rate is 0.1 and is divided by 10 at epoch 20 and 30. Moreover, the optimizer is also the stochastic gradient descent (SGD). Same with the experiments on CIFAR, the weight decay, Nesterov momentum and dropout rate is set 10^{-4}, 0.9 and 0.2 respectively.

4.2.3 Other details. The Network structure is the same with the structure for CIFAR.

4.2.4 Comparison with state-of-the-art methods. We compare our DA-CNN with other state-of-the-art methods including Wide Residual Network[43] (denoted as WRN), Network in Network[44] (denoted as NIN), FractalNet[45], DSNet[46] and DenseNet[12]. Table 3 lists all the related results.

| Network | err SVHN | Network | err SVHN |
|---------|----------|---------|----------|
| NIN[44] | 2.35     | FractalNet | 2.01     |
| DSNet[46] | 1.92   | WRN-16 | 1.64     |
| DenseNet40[12] | 1.79 | DA-CNN-60 | 1.81 |
| DenseNet100[12] | 1.76 | DA-CNN-93 | 1.74 |
| DenseNet250[12] | 1.74 | DA-CNN-145 | 1.71 |

It can be seen that our DA-CNN-145 outperforms all other state-of-the-art methods except the Wide Residual Network[43]. Also, our DA-CNN-60 and DA-CNN-93 achieves a good classification accuracy. Another impressive observation is that our DA-CNN-93 achieves the same performance when compared with DenseNet-250, while the parameter number of DenseNet-250 is more.

4.2.5 Influence of attention module. Similar to the experiments on CIFAR-10/100, here we also report the test error (denoted as err) of our DA-CNN under the situations where dense attention learning strategy is removed (denoted as DA-CNN#DA) or channel-wise attention module is removed (denoted as DA-CNN#CA). Table 2 lists all the related classification results on SVHN dataset. When there is no dense attention learning strategy, the performance significantly decreases. Also, without our channel-wise attention module, the performance also drops.

4.3 ImageNet Classification

4.3.1 Implementation. The ILSVRC 2012 dataset[5] has 1.2 million training images, 50000 validation images and 100000 testing images from 1000 classes. All images are three-channel 224×224 colored images. Following [9,12,31], we implement standard data augmentation for the training dataset, normalize them into [0,1] and report the single-crop error rate on the validation dataset.

4.3.2 Parameter setting. For fair comparison, we also keep most of the parameter settings the same with DenseNet[12]. To be specific, the model is trained for 90 epochs with a batch size of 256. The initial learning rate is 0.1 and is divided by 10 at epoch 30 and 60. Moreover, the optimizer is also the stochastic gradient descent (SGD). Also, the weight decay, Nesterov momentum and dropout rate is set 10^{-4}, 0.9 and 0.2 respectively.
4.3.3 Other details. The detailed network structure is listed in Table 4. The hyper-parameter setting is also \( p=3, d=6 \), \( p=6, d=8 \) and \( p=8, d=12 \), but input and output image size is different from the network structure in CIFAR and SVHN. Moreover, following the setting of DenseNet for the ImageNet\cite{12}, the initial convolutional filter number and growth rate is also 64 and 32 respectively.

4.3.4 Comparison with state-of-the-art methods. In ImageNet, we report the single-crop error rate of our DA-CNN and compare the results with state-of-the-art methods including ResNet\cite{9}, Residual Attention Network\cite{17} and DenseNet\cite{12}. Note that here both top-1 (denoted as top-1 err) and top-5 error (denoted as top-5 err) are included.

Table 4. Network structure of our DA-CNN for ImageNet. Note that Attention in each cell denotes one-layer spatial attention module, and \((1 \times 1 \text{ conv}) \times 2\) at the end of each dense attention block denotes our channel-wise attention module.

| Layers         | Output Size | DA-CNN-60 | DA-CNN-93 | DA-CNN-145 |
|----------------|-------------|-----------|-----------|------------|
| Convolutional  | 112×112     | 7×7 conv, stride 2 | 7×7 conv, stride 2 | 7×7 conv, stride 2 |
| Pooling        | 56×56       | 3×3 max pool, stride 2 | 3×3 max pool, stride 2 | 3×3 max pool, stride 2 |
| Dense Attention Block I | 56×56 | 1×1 conv \[3×3 conv\] × 3 | 1×1 conv \[3×3 conv\] × 6 | 1×1 conv \[3×3 conv\] × 8 |
|                | 28×28       | (1×1 conv) × 2 | 2×2 average pool, stride 2 | 2×2 average pool, stride 2 |
| Dense Attention Block II | 28×28 | 1×1 conv \[3×3 conv\] × 3 | 1×1 conv \[3×3 conv\] × 6 | 1×1 conv \[3×3 conv\] × 8 |
|                | 14×14       | (1×1 conv) × 2 | 2×2 average pool, stride 2 | 2×2 average pool, stride 2 |
| Dense Attention Block III | 14×14 | 1×1 conv \[3×3 conv\] × 3 | 1×1 conv \[3×3 conv\] × 6 | 1×1 conv \[3×3 conv\] × 8 |
|                | 7×7         | (1×1 conv) × 2 | 2×2 average pool, stride 2 | 2×2 average pool, stride 2 |
| Dense Block IV  | 7×7         | 1×1 conv \[3×3 conv\] × 3 | 1×1 conv \[3×3 conv\] × 6 | 1×1 conv \[3×3 conv\] × 8 |
| Classification Layer | 1×1       | 7×7 conv global average pool | 1000D fully-connected, softmax | 1000D fully-connected, softmax |

Table 5 lists all the related classification results. Our DA-CNN-145 outperforms other state-of-the-art methods. Note that in Table 1 we have demonstrated that our DA-CNN achieves similar
performance to DenseNet of similar depth but the number of parameters is much less. Although the layer number of our DA-CNN-60 and DA-CNN-93 is quite shallow when compared with models such as DenseNet-121, DenseNet-169 and ResNet-200, they still outperform these models.

These results indicate the effectiveness of our proposed DA-CNN. It may be mainly explained from its dense attention learning strategy, which fuses the bottom-up convolutional features and top-down discriminative features. It allows our model to effectively learn both features with less parameter number and shallower depth.

Table 5. Comparison with state-of-the-art methods on ImageNet.

| Network                | top-1 err | top-5 err |
|------------------------|-----------|-----------|
| ResNet152[9]           | 22.16     | 6.16      |
| Attention56[17]        | 21.76     | 5.90      |
| DenseNet121[12]        | 25.02     | 7.71      |
| DA-CNN-60              | 21.44     | 5.67      |
| ResNet200[9]           | 20.10     | 4.80      |
| Attention92[17]        | 19.50     | 4.80      |
| DenseNet169[12]        | 23.80     | 6.85      |
| DA-CNN-93              | 19.98     | 5.03      |
| DenseNet201[12]        | 22.58     | 6.34      |
| DenseNet264[12]        | 22.15     | 6.12      |
| DA-CNN-145             | 19.23     | 4.61      |

4.3.5 Discussion of feature reuse. To investigate the feature propagation and feature reuse capability, we visualize the weights of different layers inside a block[12,31]. We pick up the weights of a dense block with six composite functions (left in figure4) and the weights of our dense attention block with the hyper-parameter $d=6$ (right in figure4).

Figure 4. Visualization of the weights in the first block in pre-trained DenseNet (a) and DA-CNN (b) via calculating the average absolute value of $W_{ij}$. Note that 0 in the vertical axis denotes the input layer in this block.

It can be seen that the co relation between weights from shallower layers and from deeper layers is stronger in our dense attention block compared with the dense block.
4.3.6 Visual Results. We provide some visualized examples in 0 to demonstrate the effectiveness of our dense attention learning strategy.

It can be seen that key objects such as fish or dog is highlighted via our dense attention learning, while background information is suppressed.

5. Conclusion

In this paper, we propose a dense attention convolutional neural network (DA-CNN) for image classification. The major component of our DA-CNN is the dense attention block. In each dense attention block, we implement dense attention learning strategy to fuse top-down discriminative features and bottom-up convolutional features. It allows both features to propagate till the end of each block. Moreover, a channel-wise attention module is utilized at the end of each block to compress and refine the fused features. Exhaustive experiments demonstrate that our DA-CNN outperforms many state-of-the-art methods with shallower depth and much fewer parameters.

Acknowledgments

This research was supported by the National Key Research and Development Program of China (No.2017YBF0503600 and No.2016YFB0502600).

References

[1] Xue, N., Bai, S., Wang, F.D., Xia, G.S., Wu, T.F., Zhang, L.P. (2019) Learning Attraction Field Representation for Robust Line Segment Detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach.

[2] Hong, S.H., Oh, J.Y., Li, H.L., Han, B.Y. (2016) Learning Transferrable Knowledge for Semantic Segmentation with Deep Convolutional Neural Network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas.

[3] Li, Y.K., Ouyang, W.L., Wang, X.G., Tang, X.O. (2017) ViP-CNN: Visual Phrase Guided Convolutional Neural Network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu.

[4] Chen, T., Lu, S.J., Fan, J.Y. (2018) S-CNN: Subcategory-Aware Convolutional Networks for Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(10): 2522–2528.

[5] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Li, F.F. (2009) ImageNet: a Large-Scale Hierarchical Image Database. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Miami.

[6] Krizhevsky, A., Sutskever, I., Hinton, G. (2012) ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems (NIPS). Lake Tahoe. pp. 1097–1105.

[7] Simonyan, K., Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. In: International Conference on Learning Representations (ICLR). San Diego.

[8] Szegedy, C., Liu, W., Jia, Y.Q., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015) Going Deeper with Convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston.

[9] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2016) Deep Residual Learning for Image Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas.
[10] Xie, S.N., Gershick, R., Dollar, P., Tu, Z.W., He, K.M. (2017) Aggregated Residual Transformations for Deep Neural Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu.

[11] Huang, G., Liu, S.C., Maaten, L.V.D., Weinberger, K. (2017) CondenseNet: An Efficient DenseNet using Learned Group Convolutions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu.

[12] Huang, G., Liu, Z., Maaten, L.V.D., Weinberger, K. (2017) Densely Connected Convolutional Networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu.

[13] Pramod, R.T., Arun, S.T. (2016) Do Computational Models Differ Systematically from Human Object Perception? In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas.

[14] Chen, L., Zhang, H.W., Xiao, J., Nie, L.Q., Shao, J., Chua, T.S. (2017) SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu.

[15] Liu, S.K., Johns, E., Davison, A.J. (2019) End-to-End Multi-Task Learning with Attention. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach.

[16] Harley, A.W., Derpanis, K.G., Kokkinos, I. (2017) Segmentation-Aware Convolutional Networks Using Local Attention Masks. In: IEEE International Conference on Computer Vision (ICCV). Venice.

[17] Wang, F., Jiang, M.Q., Qian, C., Yang, S., Li, C., Zhang, H.G., Wang, X.G., Tang, X.O. (2017) Residual Attention Network for Image Classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu.

[18] Chen, X.Y., Yuan, X.T., Chen, Q., Yan, S.C., Chua, T.S. (2011) Multi-Label Visual Classification with Label Exclusive Context. In: IEEE International Conference on Computer Vision (ICCV). Barcelona.

[19] Wu, Z.F., Shen, C.H., Hengel, A.V.D. (2019) Wider or Deeper: Revisiting the ResNet Model for Visual Recognition. Pattern Recognition, 90: 249–256.

[20] Lin, T.Y., Goyal, P., Girshick, R., He, K.M., Dollar, P. (2017) Focal Loss for Dense Object Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 99: 2999–3007.

[21] Kendall, A., Gal, Y., Cipolla, R. (2018) Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City.

[22] Bi, Q., Qin, K., Li, Z.L., Zhang, H., Xu, K. (2019) Multiple Instance Dense Connected Convolution Neural Network for Aerial Image Scene Classification. In: IEEE International Conference on Image Processing (ICIP). Taipei.

[23] Zhang, Z.C., Liang, X.K., Dong, X., Xie, Y.Q., Cao, G.H. (2018) A Sparse-View CT Reconstruction Method Based on Combination of DenseNet and Deconvolution. IEEE Transactions on Medical Imaging, 37(6): 1–13.

[24] Song, T.Y., Yu, Q., Song, Y.Z., Xiang, T., Hospedales, T. (2017) Deep Spatial-Semantic Attention for Fine-Grained Sketch-Based Image Retrieval. In: IEEE International Conference on Computer Vision (ICCV). Venice.

[25] Sun, W.Y., Zhao, H.T., Jin, Z. (2018) A Visual Attention Based ROI Detection Method for Facial Expression Recognition. Neurocomputing, 296: 12–22.

[26] Xu, H.J., Saenko, K. (2016) Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In: European Conference on Computer Vision (ECCV). Amsterdam. pp. 451–466.

[27] Seo, P.H., Lin, Z., Cohen, S., Shen, X.H., Han, B.H. (2016) Hierarchical Attention Networks. in Preprint, arXiv:1606.02393v1.

[28] Hu, J., Shen, L., Albani, S., Sun, G., Wu, E.H. (2020) Squeeze-and-Excitation Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(8): 2011–2023.

[29] Glorot, X., Bengio, Y. (2010) Understanding the Difficulty of Training Deep Feedforward Neural Networks. Journal of Machine Learning Research, 9: 249–256.
[31] Yang, Y.B., Zhong, Z.S., Shen, T.C., Lin, Z.S. (2018) Convolutional Neural Networks with Alternately Updated Clique. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City.

[32] Wang, Q.L., Wu, B.G., Zhu, P.F., Li, P.H., Zuo, W.M., Hu, Q.H. (2019) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. in Preprint, arXiv: 1910.03151.

[33] Wang, Q.L., Gao, Z.L., Xie, J.T., Zuo, W.M., Li, P.H. (2019) Global Gated Mixture of Second-Order Pooling for Improving Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems (NIPS). Vancouver. pp. 1284–1293.

[34] Woo, S.H., Park, J.C., Lee, J.Y., Kweon, I.S. (2018) CBAM: Convolutional Block Attention Module. In: European Conference on Computer Vision (ECCV). Munich. pp. 3–19.

[35] Fu, J., Liu, J., Tian, H.J., Fang, Z.W., Lu, H.Q. (2019) Dual Attention Network for Scene Segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach.

[36] Gu, G.M., Kim, S.T., Ro, Y.M. (2018) Adaptive Attention Fusion Network for Visual Question Answering. In: IEEE International Conference on Multimedia and Expo (ICME). Hong Kong.

[37] Wang, C. (2018) RRA: Recurrent Residual Attention for Sequence Learning. In: AAAI Conference on Artificial Intelligence. New Orleans.

[38] Krizhevsky, A., Hinton, G. (2009) Learning Multiple Layers of Features from Tiny Images.

[39] Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y. (2011) Reading Digits in Natural Images with Unsupervised Feature Learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning. Sierra Nevada.

[40] Gross, S., Wilber, M. (2016) Training and Investigating Residual Nets. http://torch.ch/blog/2016/02/04/resnets.html.

[41] He, K.M., Zhang, X.Y., Ren, S.Q., Sun, J. (2015) Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: IEEE International Conference on Computer Vision (ICCV). Santiago.

[42] Zagoruyko, S., Komodakis, N. (2016) Wide Residual Networks. in Preprint, arXiv: 1605.07146.

[43] Lin, M., Chen, Q., Yan, S.C. (2014) Network in Network. In: International Conference on Learning Representations (ICLR). Banff.

[44] Larsson, G., Maire, M., Shakhnarovich, G. (2016) Wide Residual Networks. in Preprint, arXiv: 1605.07648.

[45] Lee, C., Xie, S., Gallagher, P., Zhang, Z., Tu, Z. (2015) Artificial Intelligence and Statistics. In: International Conference on Artificial Intelligence and Statistics (AISTATS). San Diego. pp. 562–570.