Left-Right Skip-DenseNets for Coarse-to-Fine Object Categorization

Changmao Cheng[1], Yanwei Fu*[1], Wenlian Lu[2], Yu-Gang Jiang[1], Jianfeng Feng[2] and Xiangyang Xue[1]

[1]School of Computer Science, Fudan University, China
[2]Center for Computational Systems Biology, Fudan University, China
{*cmcheng16,yanweifu,wenlian,ygj*}*@fudan.edu.cn* *jianfeng64@gmail.com* *xyxue@fudan.edu.cn*

Abstract

Inspired by the recent neuroscience studies on the left-right asymmetry of the brains in the low and high spatial frequency processing, we introduce a novel type of network – the left-right skip-densenets for coarse-to-fine object categorization. This network can enable both coarse and fine-grained classification in a single framework. We also for the first time propose the layer-skipping mechanism which learns a gating network to predict whether skip some layers in the testing stage. This layer-skipping mechanism assigns more flexibility and capability to our network for the categorization tasks. Our network is evaluated on three widely used datasets; the results show that our network is more promising in solving the coarse-to-fine object categorization than the competitors.

Introduction

Recent success of deep neural networks (NNs), such as convolutional neural networks (CNNs) and Recurrent Neural Networks (RNNs) has inspired the researchers to further explore the mechanisms and functions of human brains. Even though the engineering-tailored NNs may surpass human performance on several specific vision tasks, simulating the flexibility and the complexity of human brain system is still very hard for the current level NNs; and the Artificial General Intelligence (AGI) is also a mission impossible in the near future. However, the researchers (Hassabis et al. 2017) do believe that biological plausibility can be a guide to design intelligence systems that achieve neuroscience-level understanding; and particularly, we focus on visual understanding in this paper.

Though there are still lots of arguments towards that the exactly how and where visual analysis is processed within the brain, there is considerable evidence showing that visual analysis take place in a predominately and default coarse-to-fine sequence (Kauffmann, Ramanoël, and Peyrin 2014) as shown in Fig. 1(1). Interestingly, the coarse-to-fine perception is also proportional to the length of the cerebral circuit path, i.e. the time. For example, when we are presenting the image of Fig. 1(1) in a very short time, only very coarse visual stimuli can be perceived, such as the sand, and umbrella, which is usually of low spatial frequencies. Nevertheless, if we are given a longer time, more fine-grained details

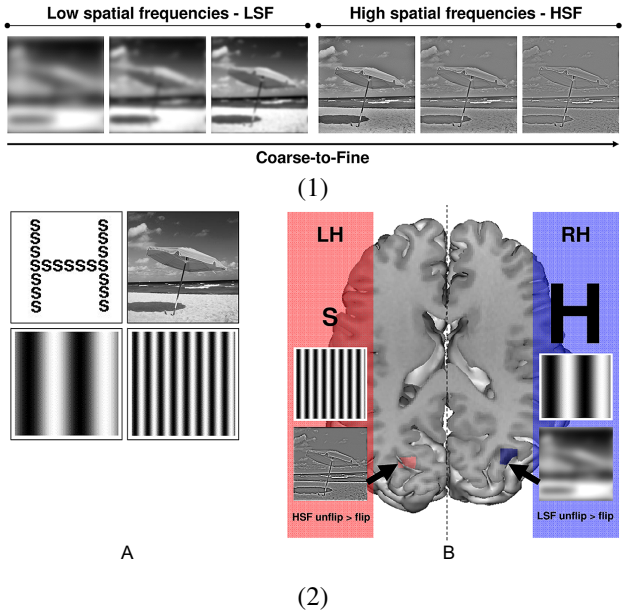

Figure 1: (2-A) Example images to assess cerebral asymmetries for spatial frequencies. (2-B) Hemispheric specialization: the left hemisphere (LH)/the right hemisphere (RH) are predominantly involved in the local/global letter identification. Figures from (Kauffmann, Ramanoël, and Peyrin 2014).

of high spatial frequencies can be perceived. It is natural to ask whether our network has such mechanism of predicting coarse information with fewer layers, and fine visual stimuli with more layers.

Another question is how the coarse-to-fine sequence being processed in human brains? Recent biological studies (Kauffmann, Ramanoël, and Peyrin 2014) (R. et al. 2003; Y. et al. 2008) postulate that the two cerebral hemispheres may be differently involved in spatial frequency processing, i.e., left-right asymmetry of the brain. The left hemisphere (LH)/the right hemisphere (RH) are predominantly involved in the high/low spatial frequency processing respectively. As illustrated in Fig. 1(2-A), the top-left figure is a large global letter made up of small local letters; and the top-right figure

*Corresponding Author

is a scene image. The bottom-left and bottom-right are the corresponding sinusoidal grating of corresponding images above. In Fig. 1(2-B), given the same input visual stimuli, the high activated regions in LH and RH are corresponding to high (in red) and low (in blue) spatial frequencies. Additionally, the left-right asymmetry of the brain was argued as a result of evolution of human beings (Macneilage, Rogers, and Vallortigara 2009): the key initial functional configurations of LH and RH should be roughly the same; and the cerebral asymmetries may be represented on different synapsis connections and distinctive ability of processing high and low spatial frequency come from the evolution process.

In the light of this understanding and to mimic the hemispheric specialization, we for the first time propose left-right skip-densenets for coarse-to-fine object categorization tasks. Our network can enable the coarse-to-fine object categorization in a single framework. The whole network is structured in Fig. 2. Our network has two branches by referring to LH and RH respectively. Both branches have roughly the same initialized layers and structures. The networks are built upon the Skip-Dense blocks and transition layers. The former one is an abstraction block for visual concept abstraction; and the latter one aims at manipulating the size of feature maps learned in the previous layer. The unique connections are built by learning varying knowledge or abstraction. The functionality of each branch is "memorized" from the given input and supervised information in the learning stage. The "Guide" arrow refers to the information flows from coarse branch to fine branch; this indicates that the fine-grained categorization tasks can also use the information learned from the coarse categorization tasks. The same mechanism is also observed in human brain (Ge, Kendrick, and Feng 2009).

We for the first time introduce the *layer-skipping mechanism* for single input to utilize only a part of layers in the deep model for the purpose of computation sparsity and flexibility. The organisms like humans tend to use their energy "wisely" for the task given visual stimuli (Macneilage, Rogers, and Vallortigara 2009). Some recent study (AW and PM 2003) in neuroscience also showed that the synaptic cross-layer connectivity is common in human neural system, especially in the same abstraction level. In contrast, most state-of-the-art deep convolutional neural networks (e.g. AlexNet (Deng et al. 2009)) do not have this mechanism and have to process the input data with the entire network layers when doing the inference. This is not only computational intensive; but also recent study (Bolukbasi et al. 2017) found that most of examples are easy to be correctly classified without the utilization of very deep networks in the testing stage. In particular, we propose the gating network learns to predict whether skipping one convolutional layer in the testing stage. Our networks are evaluated on three datasets in the coarse-to-fine object recognition tasks. The experimental results strongly support our claims and thus can be used to validate our proposed network.

Contributions. In this paper, inspired by the left-right asymmetry of the brain, we propose a left-right skip-densenets for coarse-to-fine object classification tasks. The main novelties of our networks come from several parts.

1. The left and right branch network structure to solve coarse-to-fine classification is firstly proposed. Our network is inspired by the recent theory in neuroscience (Kauffmann, Ramanoël, and Peyrin 2014).
2. A novel *layer-skipping mechanism* is for the first time proposed to skip some layers at the testing stage.
3. Additionally, we create a novel dataset named small-big MNIST (sb-MNIST) dataset for the related research.

Related Work

Hemispheric Specialization. Left-right asymmetry of the brain has been widely studied in through psychological examination and functional imaging in primates (AW and PM 2003). Recent research showed that both shape of neurons and density of neurotransmitter receptor expression depend on the laterality of presynaptic origin (R. et al. 2003; Y. et al. 2008). There is also a proof in terms of information transfer that indicates the connectivity changes between and within left and right inferotemporal cortexes as a result of recognition learning (Ge, Kendrick, and Feng 2009). Learning also differs in both local and population as well as theta-nested gamma frequency oscillations in both hemispheres and there is greater synchronization of theta across electrodes in the right IT than the left IT (Kendrick et al. 2009). It has recently been argued that the left hemisphere specializes in controlling routine and tends to focus on local aspects of the stimulus while the right hemisphere specializes in responding to unexpected stimuli and tends to deal with the global environment (M. 1994). For more information, we refer to one recent survey paper (Kauffmann, Ramanoël, and Peyrin 2014).

Deep Architectures. Starting with the notable victory of AlexNet (Krizhevsky, Sutskever, and Hinton 2012), ImageNet (Deng et al. 2009) classification contest has boomed the exploration of deep CNN architectures. Later, (He et al. 2016a) proposed deep Residual Networks (ResNets) which mapped lower-layer features into deeper layers by shortcut connection and element-wise addition, making training up to hundreds or even thousands of layers feasible. Prior to this work, Highway Networks (Srivastava, Greff, and Schmidhuber 2015) devised shortcut connection with input-dependent gating units.

Recently, (Huang, Liu, and Weinberger 2017) proposed a compact architecture called DenseNet that further integrated shortcut connections to make early layers concatenated to later layers. The simple dense connectivity pattern surprisedly achieves the state-of-the-art accuracy with fewer parameters. Different from the shortcut connections used in DenseNets, ResNets, our layer-skipping mechanism is learnt to predict whether skipping one particular layer in the testing stage.

Feedback Networks (Zamir et al. 2017) developed a coarse-to-fine representation via recurrent convolutional operations, so that the previous iteration's output gives a feedback to the prediction at the next iteration. With different

emphasis on biological mechanisms, our design is also a coarse-to-fine formulation but in a feedforward fashion. The coarse-level categorization will guide the fine-level task.

Additionally, our network is also different from the previous two branch networks such as Siamese Net (Bromley et al. 1993) and two-stream networks (Simonyan and Zisserman 2014) which usually have different input data to each branch and can not solve the coarse-to-fine object categorization problems.

Conditional Computation. The evaluation of deep models are still time-consuming and resource-intensive. Some efforts have been made to bypass this problem, including the network compression (Zhang et al. 2017), and conditional computation (CC) (Bengio, Léonard, and Courville 2013). The CC often refers to input-dependent activation for neurons or unit blocks, resulting in partial involvement fashion for the network. The CC learns to drop some data points or data blocks in the feature map and thus it can be taken as a adaptive variant of dropout. (Bengio, Léonard, and Courville 2013) introduced Stochastic Times Smooth neurons as binary gates within a deep neural network and termed straight-through estimator whose gradient is learned by heuristically back-propagating through the threshold function. (Ba and Frey 2013) proposed a 'standout' technique which uses an auxiliary binary belief network to compute the dropout probability for each node. (Bengio et al. 2015) tackle the problem of selectively activating blocks of units via reinforcement learning. Later, (Odena, Lawson, and Olah 2017) used a RNN controller trained by REINFORCE to examine and constrain intermediate activations of a network at test-time. (Figurnov et al. 2017) incorporates attention into ResNets for learning image-dependent early-stop policy in residual units, both in layer level and feature block level.

Compared with conditional computation, our layer-skipping mechanism is different in two points: (1) CC learns to drop out some data points or data block of feature map, whilst our gating network learns to predict whether skipping the layers. (2) In most cases, CC usually employs the reinforcement learning methods which has in-differentiable loss function and needs huge computational cost for the policy search algorithm; in contrast, our gating network is a differentiable function which can be used for any layer in our network.

Model

To simulate the left-right hemispheric specialization, we propose the Left-Right DenseNets architecture as shown in Fig. 2. The input image is firstly processed by the first convolutional layer shared with two subnets, representing low-level visual signal processing module[1].

Each subnet is stacked mainly by abstraction blocks and transition layers iteratively. As in Fig. 2 (right), we define the Skip-Dense block which can be viewed as a level of visual concept abstraction and the information flows can be adaptively changed by the gating networks. Each transition layer can do convolutional and pooling operations on feature maps to change channel number and spatial size. The guide link in Fig. 2 (left) indicates that the coarse/global branch can help guide the fine/local one, also inspired by the coarse-to-fine cortical model (Kauffmann, Ramanoël, and Peyrin 2014; Peyrin et al. 2010; Bullier 2001). Finally, pooling and linear classifier are added to each branch for prediction.

Both the left and right subnets are equivalent in the general structure but will perform differently due to the different grained level of supervised information. One branch is fed into coarse (or global) level object class information and another branch is fed into fine (or local) level object class information. Given an input, the coarse/global path is commonly shorter and faster than the fine/local one since it is learned from rougher information.

We explain the details of our proposed network below.

Hidden Paths

Residual networks containing shortcut connections behave like ensembling numerous shallower paths (Veit, Wilber, and Belongie 2016). Thus, removing few layers in these networks would not degrade the performance seriously. Similarly, DenseNets containing dense connectivity pattern allow any layer to access all the preceding layers, which can make individual layers additionally supervised by shorter connections (Huang, Liu, and Weinberger 2017). We may conclude that both ResNets and DenseNets can improve the data flow via short paths throughout the networks. By allowing layer selection mechanism in a DenseNet (ResNet) containing L dense layers (residual units), we can obtain 2^L hidden paths. Considering that ensembling so many hidden paths is intuitively redundant for per-input basis, we believe that introducing path selection can speedup the inference process and even enlarge the capacity of the network as flexible dynamic path selection may make the evaluation more targeted for single input. Besides, this mechanism increases the complexity and diversity of nested convolutions, which is still unexplored. Note that some experimental evidence (Bullier 2001; Peyrin et al. 2010) also indicate that the path optimization may exist in the parvocellular pathways of the left and right hemispheres when the low and high-pass signals are processed.

Skip-Dense Block

We discuss the structure of a Skip-Dense block in Fig. 2 (right). For the input pattern $\mathbf{x}$ from a shallower abstraction level, shortcut connection is applied after each dense layer. And the option of flowing through each dense layer is conditionally checked by a cheap gating network which is input-dependent. Then the output pattern $\mathbf{y}$ is processed into the next transition layer and entered into the next abstraction level.

Gating Gating network is introduced to predict whether or not skipping the convolutional layer in terms of input data pattern. It can also be taken as one special type of regularizations: if the input data flow is too complex, the gating network should be inclined not to skip too many layers; and vice versa. Here, we utilize a $N \times 1$ fully-connected layer for

[1]In general, this convoluational layer is biologically corresponding to the primary visual cortex, V1 (Naselaris 2015).

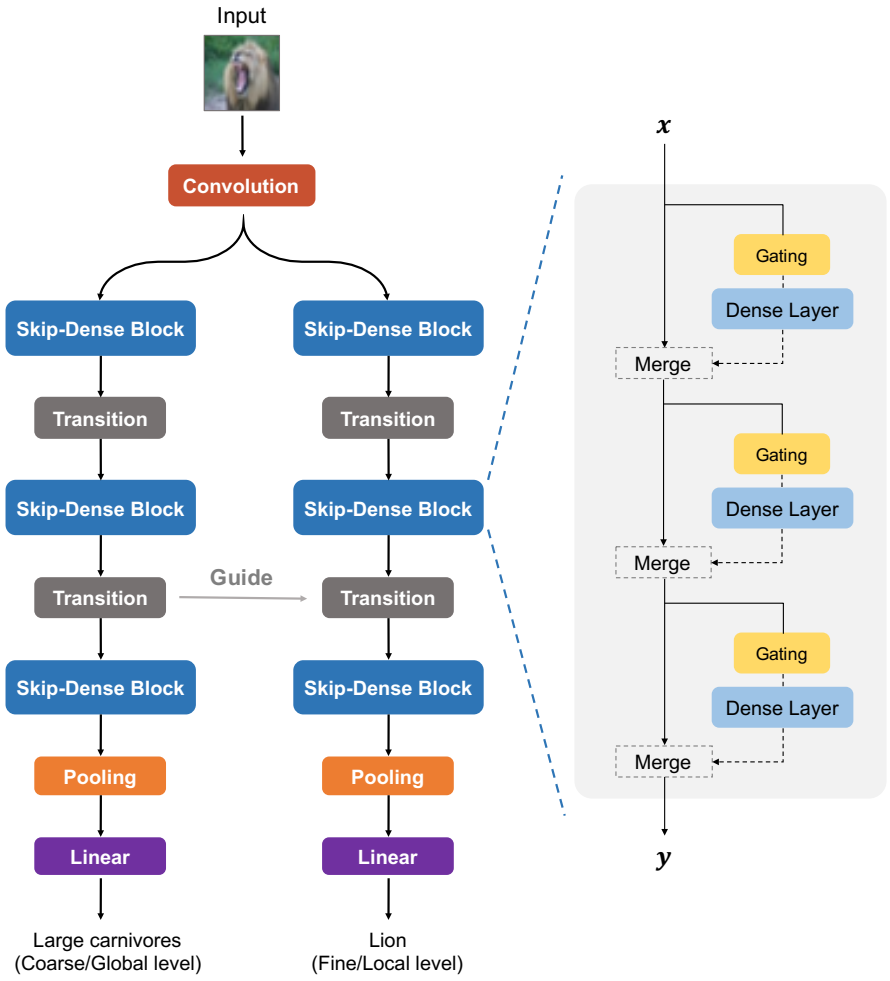

Figure 2: **Left-Right Skip-DenseNets architecture.** *Left:* Two homogeneous subnets derived from DenseNets process input visual information asymmetrically. One branch is to predict coarse/global level object classes, the other is to predict fine/local level object classes. They share the same prelimilary visual processing module, namely the first convolutional layer here. Each subnet is stacked mainly by abstraction blocks and transition layers iteratively. The guide link delivers global context information from the coarse/global subnet to the fine/local one. *Right:* Each Skip-Dense block contains several skipable densely connected convolutional layers controlled by a cheap affiliated gating network.

the N dimensional input feature and then a threshold function is applied on the scalar output. We preprocess the input feature by average pooling in practice.

The policy of designing and training the threshold function as an estimator is very critical for the success of skipping mechanism. Luckily, the magnitude of an unit often determines its importance for the categorization task in CNNs. Based on this observation, we derive a simple end-to-end training scheme for the entire network. Specifically, the output of threshold function is multiplied with each unit of the convolutional layer output, which affects the layer importance for categorization. As long as the threshold function is differentiable, gating network training can be incorporated into back-propagation naturally and smoothly. The mutual adjustment and regularization of gatings and dense layers abbreviate the problem of training difficulty and hard convergence of reinforcement learning (Bengio et al. 2015).

We exploit two kinds of threshold functions. One is adjustable *soft sigmoid function*

$$\text{soft sigm}(x) = \frac{1}{1+e^{kx}} \tag{1}$$

the other is adjustable *hard sigmoid function*,

$$\text{hard sigm}(x) = \max\left(0, \min\left(kx + \frac{1}{2}, 1\right)\right) \tag{2}$$

the slope variable k of a threshold function is regulated by the data distribution statistics of the previous outputs of the function when training. By gradually adjusting the value of k, the proportion of 'unsettled' output values lying in $(\epsilon, 1-\epsilon)$ is appropriately scheduled for training, where ϵ is the margin of error. The slope variable for each dense layer

is adjusted independently. At the ending of training, the adjusted sloped variables of gating functions are large enough and the learned gating networks can do approximate binary decisions. When evaluation, we compute the dense layer if its gating output approaches to 1 and skip it if its gating output approaches to 0. That is, gating networks make binary decisions for inference to achieve computational sparsity.

Merge The merge of preceding layers includes two types of operation: *channel concatenation* and *element-wise addition*. The former is used in DenseNets, which is claimed to improve feature reuse and efficiency. And an instance of the later can be Identity Mappings (He et al. 2016b) in ResNets. The different merge operations heavily lead to different behaviors of the two deep architectures. The pre-activation unit (He et al. 2016b) facilitates both of the merge operations from our observation, which is also used in our model. We replace skipped convolutions with the feature maps filled with zero values skipped convolutions for merge when evaluation, which make both addition and concatenation feasible. And the merged feature maps will be rescaled properly according to gating results both for training and test. Different from standard dropout (Srivastava et al. 2014), the layer dropping and feature rescaling in our model are dynamic and data-dependent.

Transition Layer

A transition layer contains 1×1 convolution for the interaction of cross channel information and pooling for downsampling feature maps if needed. When merge operations in Skip-Dense blocks are the concatenation type, the convolution at transition can reduce the number of feature channels for compression. When merge operations are addition type, the convolution at transition should increase the number of feature channels for higher level information distillation.

Guide

The faster coarse/global subnet can guide the slower fine/local subnet with global context information of objects in higher levels. In here, we select the last transition layer for delivering guiding information. Specifically, the output feature of the penultimate skip-dense block in the global subnet is copied and then concatenated into the output feature of the penultimate skip-dense block in the local subnet. Intuitively, the injection of feedback information from coarse level can be beneficial for the fine-level object categorization.

Experiments

Datasets. We conduct experiments on three datasets, namely sb-MNIST, Cifar-100 (Krizhevsky 2009) and Stanford Cars (Krause et al. 2013).

I. Inspired by the experiments in (Kauffmann, Ramanoël, and Peyrin 2014), we build the sb-MNIST dataset by randomly selecting two images from MNIST figures and using the first one as the local figure to construct the second one. We generate 12000 training images and 20000 testing images for building sb-MNIST dataset. Some examples are illustrated in Fig. 3.

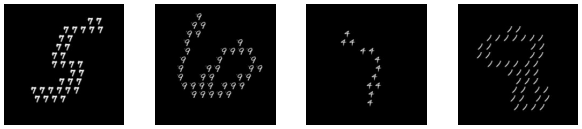

Figure 3: Illustrated examples of sb-MNIST dataset. Each "big" number is made up by the "small" number.

II. Cifar-100 dataset has 60000 images from 100 fine-grained classes, which are further divided into 20 coarse-level classes. The image size is 32×32. We use the standard training/testing splits (Krizhevsky 2009)[2].

III. Stanford Cars dataset (Krause et al. 2013) is another fine-grained classification dataset. It contains 16,185 car images of 196 fine classes (e.g. Tesla Model S Sedan 2012 or Audi S5 Coupe 2012) which describes the properties such as Maker, Model, Year of the car. In terms of the basic car types defined in (Zamir et al. 2017), it can be categorized into 7 coarse classes including Sedan, SUV, Coupe, Convertible, Pickup, Hatchback, Wagon. All the images are resized into 96×96; and top-1 mean accuracy is reported for both fine-grained and coarse-level classification.

Competitors. We compare the following several baselines.

(1) *Feedback Net* (Zamir et al. 2017): instantiated using existing Recurrent Neural Networks (RNNs), it is a feedback based learning architecture in which each representation is formed in an iterative manner based on feedback received from previous iteration's output. Thus Feedback Net has better classification performance than the standard CNNs.

(2) *DenseNet* (Huang, Liu, and Weinberger 2017): DenseNet connects each layer to all of its preceding layers in a feed-forward fashion; and the design of our Skip-Dense block is derived from DenseNet. Thus comparing with DenseNet, our network has two new components: the gating network to skip some unnecessary layers; and the two branch structures solving the coarse-to-fine classification.

(3) *ResNet* (He et al. 2016a): it is an extension of traditional CNNs by learning the residual of each layer to enable the network of being trained substantially deeper than previous CNNs.

Implementation details. We mainly report our configuration for Cifar-100. The model structures for other smaller datasets are similar. For the concat-type network, we start with 24 feature maps by the shared convolutional layer. The two branches are built based on DenseNet-40, namely, with growth rate of 12 and 3 skip-dense blocks. No compression and 2×2 max pooling are applied at each transition layer. For the addition-type network, we start with 16 feature maps by the shared convolutional layer. Both branches separately contain 6 skip-dense blocks with 4 dense layers each. The

[2] `https://www.cs.toronto.edu/~kriz/cifar.html`.

Methods	Accuracy(%)	
	Local	Global
LeNet (LeCun 1998)	90.3	70.2
DenseNet	98.3	97.2
ResNet	98.1	96.8
Ours	**99.1**	**99.1**

Table 1: Results of sb-MNIST dataset. Our network has 26 layers and averagely $\sim 50\%$ layers can be skipped in the testing step. DenseNet and ResNet have 40 and 18 layers respectively.

feature maps are doubled by 1×1 convolutions at transition layers and 2×2 max pooling is only applied after every two skip-dense blocks. For networks of both merge types, no transition layer is after the last skip-dense block and 8×8 average pooling follows instead. All the pooling operations in gating networks are 8×8 average pooling.

Data preprocessing and training procedure follow (Huang, Liu, and Weinberger 2017). The dropout rate for fully connected layers is set to 0.5. By default, we use the *element-wise addition* and *soft sigmoid function* for merge and gating network respectively.

Main Results and Discussion

sb-MNIST The results on this dataset are compared in Tab. 1. The "Local" and "Global" indicates the classification of "small" number and "big" number. On this dataset, DenseNet, ResNet and LeNet are running separately for both "Local" and "Global" tasks. This toy dataset gives us a general understanding on this task. The classification of "small" number and "big" number in the image corresponds to the identification task of high and low spatial frequency processing in spatial frequency processing (Kauffmann, Ramanoël, and Peyrin 2014). Comparing with the other baselines, our method can jointly solve the two tasks, i.e., the coarse-to-fine processing.

Judging from the results in Tab. 1, our method greatly outperforms all the other methods by a clear margin. It is largely due to two reasons. Firstly, the skip-dense block can efficiently and gradually learn the visual concepts; secondly, the "Guide" network transfers the learned information of coarse branch to the finer branch. The performance of original version LeNet is greatly suffered from the low number of layers and convolutional filters.

Cifar-100 We compare the results in Tab. 2. We compare Feedback Net, DenseNet and ResNet as well as other previous works. The "Local" and "Global" indicate the classification tasks at fine-grained and coarse-level individually. DenseNet and ResNet are running separately for the "Local" and "Global" tasks respectively.

In Tab. 2, again our network greatly outperforms the other baselines. Note that for the "Local" and "Global" classification tasks, the DenseNet and ResNet are running separately whilst Feedback Net and our network can do them jointly. Comparing the Feedback Net with ours, the accuracy margins on "Local" and "Global" are 5.0 and 2.9 individually.

Methods	Accuracy(%)	
	Local	Global
(Snoek et al. 2014)	72.6	–
(Mishkin and Matas 2016)	72.3	–
Feedback Net	71.1	80.8
DenseNet	75.6*	80.9
ResNet	72.8*	–
Ours	**76.1**	**83.8**

Table 2: Results on Cifar-100 dataset. The depths of ResNet, DenseNet and Feedback Net are 101, 40 and 48 respectively. Note that the virtual depth of Feedback Net is reported from (Zamir et al. 2017). Our network uses 31 layers and averagely $\sim 40\%$ layers can be skipped in the testing step. *: reported from (Huang, Liu, and Weinberger 2017).

This implies that our network has better capability of learning the coarse-to-fine information by the two branches.

Interestingly, even on the "Global" task, our network can still have around 3% improvement over Feedback Net and DenseNet. One possible explanation is that since the "Global" task is relatively easy, the networks with standard layers may be tend to overfit the training data; in contrast, our "Gating Network" can skip some layers and avoid the problem of overfitting in some degree.

Stanford Cars We show the results in Tab. 3. Still, the three baselines are compared in this dataset. The "Local" and "Global" tasks still refer to the classification of fine-grained and coarse-level classification.

Firstly, the Feedback Net hits the top accuracy 80.7% on the "Global" task which is 4.6% higher than our results. One possible reason of good "Global" performance of Feedback Net comes from its early prediction stage (Zamir et al. 2017), which can effectively avoid the overfitting in the "Global" task.

On the other hand, the gating network in the skip-dense block of our network can also prevent the overfitting by skipping some layers when the "Global" task is relatively easy. Thus our network still outperforms DenseNet and ResNet on the "Global" task. Finally, with the guiding information from coarse branch to fine branch, our network can achieve the best performance on "Local" task, outperforming the Feedback Net by 16.5%. This implies that our network has better coarse-to-fine processing ability than Feedback Net and this also further prove the effectiveness of our proposed framework.

Ablation Study on Cifar-100 dataset

We also conduct some ablation study to further explain/evaluate the impact towards the performance in choosing some key parameters on Cifar-100 dataset.

Merge types and gating functions We compare the different choices on merge types and gating functions. The results are compared in Tab. 4. Two evaluation metrics have been used here, namely, accuracy and skip ratio. The skip

Methods	Accuracy(%)	
	Local	Global
Feedback Net	53.4	**80.7**
DenseNet	68.1	74.8
ResNet	66.3	75.2
Ours	**69.9**	76.1

Table 3: Results on Stanford Car dataset. The depths of ResNet, DenseNet and Feedback Net are 34, 40 and 24 (virtual) respectively. Our network uses 26 layers and averagely $\sim 40\%$ layers can be skipped in the testing step.

Merge	Gating	Accuracy (%)		Skip Ratio (%)	
		Local	Global	Local	Global
concat	hard	65.4	75.8	36.2	55.3
	soft	69.0	80.8	18.2	24.1
addition	hard	75.0	82.5	44.6	43.7
	soft	76.1	83.8	42.2	59.5

Table 4: Results on Cifar-100 dataset. 'concat', 'addition','hard','soft' indicate the channel concatenation, element-wise addition, soft sigmoid function and hard sigmoid function.

ratio is computed by number of skipped layers dividing the total number of dense layers in each individual branch.

Judging from the results in Tab. 4, The combination of element-wise addition for merge and gating via soft sigmoid is our best network configuration. Specifically, we notice that soft sigmoid is generally better than hard sigmoid for both merge types on most cases. We postulate that this is probably due to the good differential property of soft sigmoid function, which can be better stabilized and optimized in the back-propagation step.

We compare the skip ratio on "Local" and "Global" tasks. In testing stages, the coarse branch of our network will skip much more layers on the "Global" tasks than the skipped layers in its sibling fine branch[3]. This is largely due to the fact that the "Global" task is easy, and using relative a few layers of coarse branch are good enough to grasp the coarse-level information. Interestingly, this point also follows the coarse-to-fine perception in the recent neural science study (Kauffmann, Ramanoël, and Peyrin 2014) that low spatial frequency information is predominantly processed by right hemisphere with relatively shorter paths in the cerebral system.

Skip ratio vs. accuracy To further verify the flexibility of our model, we vary the threshold of the gating network and it can affect the accuracy of "Local" and "Global" tasks as reported in Fig. 4.

From the results we can conclude that both branches are capable of producing acceptable accuracy under large skip

[3]One exception is addition-hard combination, which has almost the same skip ratio on both branch.

[4]We make the skip ratio controllable by reducing each learned slope variable by factor of 10 for plotting the curves.

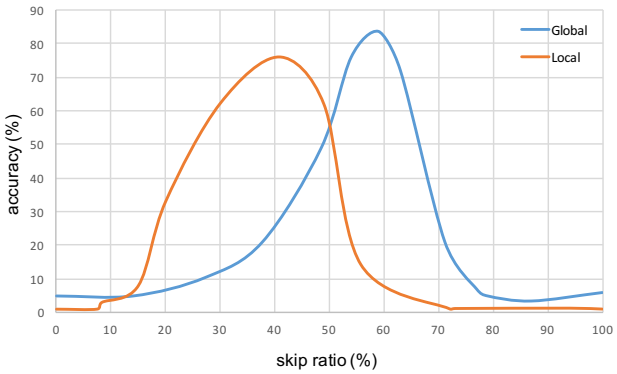

Figure 4: Local and Global accuracy under different skip ratios of dense layers in each individual branch.[4]

ratios. As expected, the optimal range of skip ratio for two branches are different and even not overlapped due to the different grained level of recognition tasks. For the Global branch, $50\% - 65\%$ is the optimal skip ratio range. For the Local branch, $25\% - 50\%$ is the optimal skip ratio range, which is lower and wider than the Global branch. The good thing is that the learned gating networks at each dense layer make the performance of two branches keep consistent under various user-defined thresholds. And we also notice that the performance of two trained branches without layer skipping is as inferior as the case of skipping all dense layers, which is also significantly different from standard dropout (Srivastava et al. 2014) and tranditional conditional computation. This intriguing phenomenon suggests that individual dense layer in our model learns specialized visual abstraction information. In other words, the skipping control of gating networks is critical for the success of recognition. Without skipping any layer, the network may become disordered.

Conclusion

Inspired by recent study on left and right hemisphere specialization, and the coarse-to-fine perception, we proposed a novel left-right Skip-DenseNets for coarse-to-fine object categorization. We use a new design philosophy to make this network simultaneously classify coarse and fine-grained classes. Additionally, the layer-skipping behaviour of densely connected convolutional layers controlled by auxillary gating networks is for the first time proposed. The experiments on three datasets validate the performance and show the promising results of our network.

References

AW, T., and PM, T. 2003. Mapping brain asymmetry. *Nat Rev Neurosci*.

Ba, J., and Frey, B. J. 2013. Adaptive dropout for training deep neural networks. In *NIPS*.

Bengio, E.; Bacon, P.-L.; Pineau, J.; and Precup, D. 2015. Conditional computation in neural networks for faster models. *CoRR* abs/1511.06297.

Bengio, Y.; Léonard, N.; and Courville, A. C. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR* abs/1308.3432.

Bolukbasi, T.; Wang, J.; Dekel, O.; and Saligrama, V. 2017. Adaptive neural networks for efficient inference. In *ICML*.

Bromley, J.; Bentz, J. W.; Bottou, L.; Guyon, I.; LeCun, Y.; Moore, C.; Säckinger, E.; and Shah, R. 1993. Signature verification using a "siamese" time delay neural network. *IJPRAI* 7:669–688.

Bullier, J. 2001. Integrated model of visual processing. *Brain Res. Rev.*

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Figurnov, M.; Collins, M. D.; Zhu, Y.; Zhang, L.; Huang, J.; Vetrov, D. P.; and Salakhutdinov, R. 2017. Spatially adaptive computation time for residual networks. In *CVPR*.

Ge, T.; Kendrick, K. M.; and Feng, J. 2009. A novel extended granger causal model approach demonstrates brain hemispheric differences during face recognition learning. *PLoS Comput Biol*.

Hassabis, D.; Kumaran, D.; Summerfield, C.; and Botvinick, M. 2017. Neuroscience-inspired artificial intelligence. *Neuron* 95(2):245–258.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Identity mappings in deep residual networks. In *ECCV*.

Huang, G.; Liu, Z.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*.

Kauffmann, L.; Ramanoël, S.; and Peyrin, C. 2014. The neural bases of spatial frequency processing during scene perception. *Frontiers in integrative neuroscience* 8.

Kendrick, K. M.; Zhan, Y.; Fischer, H.; Nicol, A. U.; Zhang, X.; and Feng, J. 2009. Learning alters theta-nested gamma oscillations in inferotemporal cortex. *Nature*.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. *IEEE International Conference on Computer Vision Workshops* 554–561.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Krizhevsky, A. 2009. Learning multiple layers of features from tiny images.

LeCun, Y. 1998. Gradient-based learning applied to document recognition.

M., T. 1994. *Right-brain left-brain reflexology: a self-help approach to balancing life energies with color, sound, and pressure point techniques*. RochesterVT: Healing Arts Press.

Macneilage, P.; Rogers, L.; and Vallortigara, G. 2009. Origins of the left & right brain. *Scientific American*.

Mishkin, D., and Matas, J. 2016. All you need is a good init. In *ICLR*.

Naselaris. 2015. A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*.

Odena, A.; Lawson, D.; and Olah, C. 2017. Changing model behavior at test-time using reinforcement learning. *CoRR* abs/1702.07780.

Peyrin, C.; Michel, C. M.; Schwartz, S.; Thut, G.; Seghier, M.; and Landis, T. 2010. The neural substrates and timing of top-down processes during coarseto- fine categorization of visual scenes: a combined fmri and erp study. *J. Cogn. Neurosci*.

R., K.; Y., S.; Y., K.; H., S.; R., S.; and I., I. 2003. Asymmetrical allocation of nmda receptor epsilon2 subunits in hippocampal circuitry. *Science*.

Simonyan, K., and Zisserman, A. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*.

Snoek, J.; Rippel, O.; Swersky, K.; Kiros, R.; Satish, N.; Sundaram, N.; Patwary, M. M. A.; Prabhat; and Adams, R. P. 2014. Scalable bayesian optimization using deep neural networks. In *ICML*.

Srivastava, N.; Hinton, G. E.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15:1929–1958.

Srivastava, R. K.; Greff, K.; and Schmidhuber, J. 2015. Highway networks. *arXiv preprint arXiv:1505.00387*.

Veit, A.; Wilber, M. J.; and Belongie, S. J. 2016. Residual networks behave like ensembles of relatively shallow networks. In *NIPS*.

Y., S.; H., H.; M., W.; M., I.; M., T.; and R., S. 2008. Left-right asymmetry of the hippocampal synapses with differential subunit allocation of glutamate receptors. *PNSA*.

Zamir, A. R.; Wu, T.-L.; Sun, L.; Shen, W.; Malik, J.; and Savarese, S. 2017. Feedback networks. In *CVPR*.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2017. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *CoRR* abs/1707.01083.