Multi-attention Guided Activation Propagation in CNNs

Xiangteng He and Yuxin Peng(✉)

Institute of Computer Science and Technology, Peking University, Beijing, China
pengyuxin@pku.edu.cn

Abstract. CNNs compute the activations of feature maps and propagate them through the networks. Activations carry various information with different impacts on the prediction, thus should be handled with different degrees. However, existing CNNs usually process them identically. Visual attention mechanism focuses on the selection of regions of interest and the control of information flow through the network. Therefore, we propose a multi-attention guided activation propagation approach (MAAP), which can be applied into existing CNNs to promote their performance. Attention maps are first computed based on the activations of feature maps, vary as the propagation goes deeper and focus on different regions of interest in the feature maps. Then multi-level attention is utilized to guide the activation propagation, giving CNNs the ability to adaptively highlight pivotal information and weaken uncorrelated information. Experimental results on fine-grained image classification benchmark demonstrate that the applications of MAAP achieve better performance than state-of-the-art CNNs.

Keywords: Multiple attention · Activation propagation
Convolutional Neural Networks

1 Introduction

Neural networks have advanced the state of the art in many domains, such as computer vision, speech recognition and natural language processing. Convolutional Neural Networks (CNNs) [1], one type of the popular and classical neural networks, have been widely used in computer vision due to its strong power in feature learning, and have achieved state-of-the-art performance in image classification [2], object detection [3], semantic segmentation [4] and so on.

Recent advances of CNNs focus on designing deeper neural network structure, which promote the performance of image classification. In 2012, Krizhevsky et al. designed an 8-layer convolutional neural network, called AlexNet [5], which contains 5 convolutional layers and 3 fully-connected layers. In 2014, VGGNet [2] was designed and its depth was increased to 16/19 layers by using an architecture with very small (3×3) convolutional filters, which achieved significant

J.-H. Lai et al. (Eds.): PRCV 2018, LNCS 11257, pp. 16–27, 2018.
https://doi.org/10.1007/978-3-030-03335-4_2

improvement on image classification. In 2016, He et al. designed a residual network with the depth of up to 152 layers, 8 times deeper than VGGNet, called ResNet [6], which also had a 1000-layer version.

These popular CNNs take images as inputs, conduct convolutional operation on each pixel, compute the activations of feature maps and propagate the activations through the networks layer by layer. Activations carry various information with different impacts on prediction, thus should be handled with different degrees of attention. However, existing CNNs usually process the activations identically in the propagation process, leading to the fact that the pivotal information is not highlighted and the uncorrelated information is not weakened, which is contradictory with visual attention mechanism that pays high attention to pivotal information, such as regions of interest. For addressing above problems, an intuitive idea is to adaptively highlight or weaken the activations based on their importance degrees for final prediction. The importance degree can be defined as attention.

Attention is a behavioral and cognitive process of selectively concentrating on a discrete aspect of information [7]. Tsotsos et al. state that visual attention mechanism seems to involve at least the following basic components [8]: (1) the selection of regions of interest in the visual field, (2) the selection of feature dimensions and values of interest and (3) the control of information flow through the network. Therefore, we apply visual attention mechanism to guide the activation propagation in CNNs, selecting the activations of interest and feature values of interest, as well as controlling the activation propagation through the network based on the attention. Karklin et al. indicate that neurons in primary visual cortex (V1) respond to the edge over a range of positions, and neurons in higher visual areas, such as V2 and V4, are more invariant to image properties and might encode shape [9]. According to the studies on visual attention mechanism, different level attentions focus on different attributes of objects.

Inspired by these discoveries about visual attention mechanism, we propose a multi-attention guided activation propagation approach (MAAP), which can be applied into existing CNNs to improve the performance, and give CNNs the ability to adaptively highlight pivotal information and weaken uncorrelated information. The main contributions of the proposed approach can be summarized as follows:

(i) **Low-level Attention Guided Activation Propagation (LAAP).** Neurons in primary visual cortex (V1) respond to the edge over a range of positions, which is significant for discovering the shape of the object. Inspired by this discovery, we first extract the attention map based on the activations of feature maps output from the first convolutional layer as the low-level attention, and then guide the activation propagation based on the low-level attention, enhancing the pivotal activations that carry key information such as the edge of the object. Low-level attention guided activation propagation feeds such key information forward to the high-level convolutional layer, which helps to localize the object as well as learn discriminative features.

(ii) **High-level Attention Guided Activation Propagation (HAAP).** Neurons in higher visual areas (V2 and V4) might encode shape, which is significant for recognition. Inspired by this discovery, we first extract the attention map output from high-level convolutional layer, and then apply the high-level attention to guide the activation propagation, preserving the pivotal activations that carry key information such as the object and removing the uncorrelated activations that carry less significant information such as background noise. Then we feedback the activations to the input data to eliminate the background noise, which carries uncorrelated information. High-level attention guide activation propagation to feed activations backward to the input data, which finds the region of interest and boosts discriminative feature learning.

(iii) **Multi-level Attention Activation Propagation (MAAP).** Low-level attention and high-level attention jointly guide activation propagation in CNNs to promote the discriminative feature learning, and enhance their mutual promotions to achieve better performance. The two activation propagation strategies have different but complementary focuses: LAAP focuses on enhancing the pivotal activations, while HAAP focuses on reducing the uncorrelated activations. With the guide of multi-level attention, activations are propagated through CNNs with different weights, where the key information is enhanced through the forward propagation and the uncorrelated information is removed through the backward propagation, which boost the performance of CNNs.

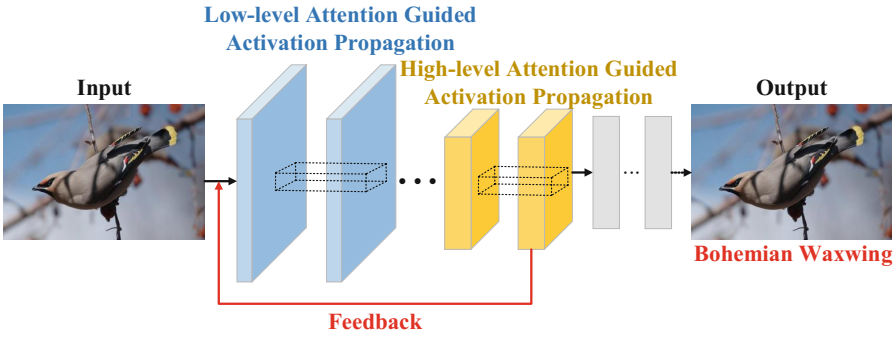

Fig. 1. Overview of the multi-attention guided activation propagation approach.

2 Multi-attention Guided Activation Propagation

Researchers state discovery of visual attention mechanism: different level visual areas of cerebral cortex concentrate on different aspects of the visual information [10]. Like neurons in V1 and V2 respond to the edge and shape of the object

respectively [9], neurons in convolutional layers have similar functions. For example, neurons in low-level convolutional layers focus on the edge of the object and neurons in high-level convolutional layers pay attention to the shape of the object. Inspired by this discovery, we propose a multi-attention guided activation propagation approach (MAAP), applying the low-level and high-level visual attention into activation propagation, which can be inserted into the CNNs, and its overview is shown in Fig. 1.

For a CNN, in the training phase: (1) We adopt the low-level attention guided activation propagation (LAAP) after the first convolutional layer to give the activations variant weights based on their attention values. (2) We employ the high-level attention guided activation propagation (HAAP) after the last convolutional layer, and feedback the activations to the input data, which is to drop the background noise and preserve the region of interest at the same time. (3) We utilize alternative training strategy to train the CNNs with LAAP and HAAP. The three components are presented in the following paragraphs. The high-level attention is performed on the input data, and frequent data modification is time-consuming and not sensible, so only LAAP is adopted in the testing phase.

2.1 Low-Level Attention Guided Activation Propagation

Neurons in convolutional layers have higher activation to some specific spatial positions of the input data, and have the pattern that focusing on the significant and discriminative features which is help for recognizing the image. We extract the feature maps from some specific convolutional layers in the widely-used CNN, e.g. VGGNet [2], and visualize their average feature map in Fig. 2. We can observe that: (1) Low-level convolutional layer focuses on the edge of the object just like neurons in primary visual cortex (V1). (2) Average feature map generated from middle-level convolutional layers has some noises, which are not helpful for recognition.

Therefore, we consider enhancing the significance of the key information, such as the edge shown in the sub figure of "Conv1_1" in Fig. 2. An intuitive idea is to give the pivotal activations with higher weights. We propose low-level

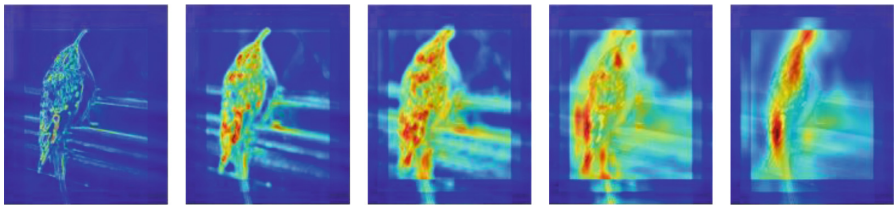

Fig. 2. Visualization of average feature maps in convolutional layers. "Conv1_1" to "Conv5_1" indicate the name of the convolutional layers in VGGNet [2].

attention activation propagation approach, which consists of attention extraction and activation enhancement. The detailed processing is shown in Fig. 3.

Attention Extraction. For a given image I, we first extract its feature maps $F = \{F_1, F_2, \ldots, F_n\}$ from the first convolutional layer in the CNN, such as AlexNet [5], VGGNet [2] and ResNet [6]. n indicates the number of neurons in this convolutional layers, and F_n indicates the feature map extracted from the first convolutional layer responding to the n-th neuron. Each feature map is a 2D matrix with the size of $mh \times mw$. Then we calculate their average feature map, denoted as FA, its definition is:

$$FA = \frac{1}{n}\sum_{1}^{n} F_i \tag{1}$$

For each element in the average feature map FA, we perform sigmoid function to normalize it to the range of $[0, 1]$. Then we get the attention map, where the element indicates the importance of each activation in the feature maps to the recognition. Each element in attention map A is calculated as follows:

$$A_{i,j} = \frac{1}{1 + e^{-FA_{i,j}}} \tag{2}$$

where i and j indicate the spatial position of element in the attention map A.

Activation Enhancement. After generating the attention map A, we apply it to guide the activation propagation. For each feature map $F_i \in F$, we calculate the new feature map F_i' based on the attention map as follows:

$$F_i' = (1 + A) * F_i \tag{3}$$

where $*$ denotes element wise product. Through the activation enhancement manipulation, we infuse the attention information to the feature outputs of convolutional layer, in order to guide the feature learning processing by highlighting the pivotal activations.

2.2 High-Level Attention Guided Activation Propagation

From Fig. 2 we can observe that the high-level convolutional layer (as "Conv5_1" shown in Fig. 2) concentrates on the shape of the whole-object, just like the higher visual areas of cerebral cortex (V2 and V4). Inspired by this, we propose high-level attention guided activation propagation, dropping the uncorrelated information of the input data, such as background noise, and preserving the region of interest of the input data at the same time. We implement this through attention extraction and activation elimination, which are presented in the following paragraphs.

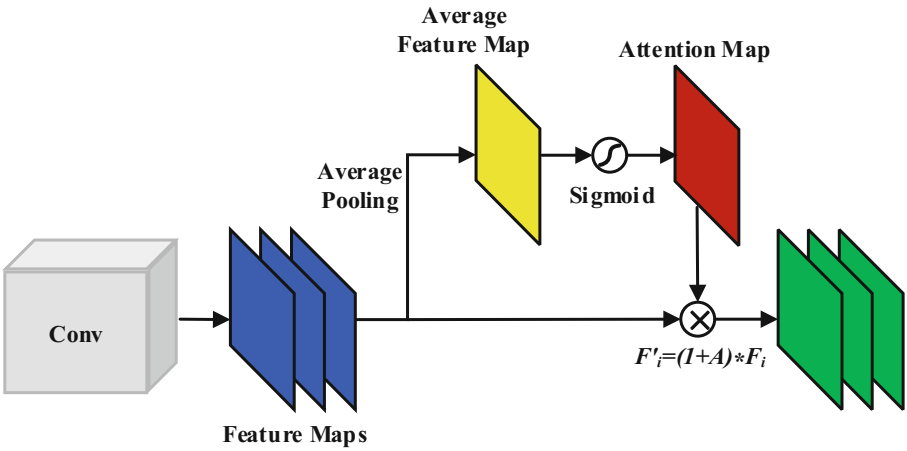

Fig. 3. Overview of low-level attention guided activation propagation approach.

Attention Extraction. The attention extraction is the same with the process in LAAP. First, we extract the attention map A from the last convolutional layer. Second, we perform binarization operation on the attention map with an adaptive threshold, which is obtained by OTSU algorithm [11], and take the bounding box that covers the largest connected region as the discriminative region R.

Activation Elimination. We propagate the attention activation backward to modify the input data D to the new data D^*as follows:

$$D_i^* = A^* * D_i \tag{4}$$

where i indicates the i-th channel of the input data. We experiment with 6 definitions of A^*:

(i) *A-RoI*: Retrain the region of R and remove the uncorrelated region.
(ii) *A-uncorrelated*: Set values of pixels outside the region of R as 0. Modify A as follow:

$$A_{i,j} = \begin{cases} 1 & , \ pixel\ (i,j)\ inside\ R \\ 0 & , \ pixel\ (i,j)\ outside\ R \end{cases} \tag{5}$$

(iii) *A-enhance*: Enhance the region of R in the input data and set values of pixels outside R as 0. Modify A as follow:

$$A_{i,j} = \begin{cases} A_{i,j} + 1 & , \ pixel\ (i,j)\ inside\ R \\ 0 & , \ pixel\ (i,j)\ outside\ R \end{cases} \tag{6}$$

(iv) *A-reduce*: Preserve the region of R in the input data and reduce the value of pixels outside R based on the attention activations. Modify A as follow:

$$A_{i,j} = \begin{cases} 1 & , \ pixel\ (i,j)\ inside\ R \\ A_{i,j} & , \ pixel\ (i,j)\ outside\ R \end{cases} \tag{7}$$

(v) *A-allsoft*: Directly adopt the extracted attention map A as A^*.

(vi) *A-allsoft*+1: Inspired by residual learning [6], we plus the original input data with the new data. Modify A as follow:

$$A = A + 1 \tag{8}$$

2.3 Alternative Training of MAAP

Considering that the high-level attention is performed on the input data, frequent data modification is time-consuming and not sensible, we design an alternative training strategy for the application of MAAP in CNNs, which is described as Algorithm 1.

Algorithm 1. Alternative Training

Input: Training data D, maximal iterative epoch it_m.
Output: Trained CNN model N.

```
Initialize N, such as pre-training on the large scale dataset, ImageNet [12].
for epoch = 1, ..., it_m do
    Compute attention map A and feature maps F of first convolutional layer.
    Modify feature maps F as formula (3).
    Perform a feed-forward pass.
    Compute the loss and perform back-propagate manipulation.
    if loss is converged then
        Stop this phase of training.
    end if
end for
Perform feed-forward pass to compute attention map A of last convolutional layer.
Modify input data D as formula (4).
Repeat 2 to 10.
return N.
```

3 Experiments

Fine-grained image classification task aims to recognize hundreds of subcategories belonging to the same basic-level category, such as 200 subcategories belonging to the category of bird. It is a challenging task due to the large variances in the same subcategory and small variances among different similar subcategories. It covers a lot of domains, such as animal species [13], plant breeds

Table 1. Classification results on CUB-200-2011 dataset.

Method	Accuracy (%)		
	AlexNet	VGGNet	ResNet
Baseline	59.0	72.2	76.0
+LAAP	59.7	72.9	76.4
+HAAP	62.2	78.0	78.1
+MAAP	**63.0**	**78.2**	**78.7**

[14], car types [15] and aircraft models [16]. We choose fine-grained image classification to evaluate the effectiveness of our MAAP approach. We conduct experiments on the widely-used CUB-200-2011 [13] dataset for fine-grained image classification. Accuracy is adopted to evaluate the effectiveness of our proposed approach, which is widely used in fine-grained image classification [17,18].

CUB-200-2011 dataset [13] is the most widely-used dataset in fine-grained image classification task, which contains 11788 images of 200 subcategories belonging to the same basic-level category of bird. It is split into training and test sets, with 5994 images and 5794 images respectively. For each image, detailed annotations are provided: an image-level subcategory label, a bounding box of object, and 15 part locations. In our experiments, only image-level subcategory label is utilized to train the CNNs.

3.1 Implementation

We implement our MAAP approach as two layers: *enhancement layer* and *elimination layer*, which are corresponding to low-level attention guided activation propagation and high-level attention guided activation propagation respectively. We implement the two layers based on the open source framework Caffe[1] [19].

Table 2. Results of adopting dropout in different convolutional layers of AlexNet.

Net	**AlexNet**	conv1	conv2	conv3	conv4	conv5
Accuracy (%)	**59.0**	57.8	58.7	58.7	57.9	58.7

To verify the effectiveness of our proposed MAAP approach, we insert MAAP into the state-of-the-art CNNs: AlexNet with 8 layers [5], VGGNet with 19 layers [2] and ResNet with 152 layers [6]. Following Algorithm 1, it consists of 3 steps in the training phase. (1) Each of these CNNs is pre-trained on the 1.3M training data of ImageNet [12]. (2) We make some modifications for each CNN. In general, for each CNN, we follow the original settings, only incorporate it

[1] http://caffe.berkeleyvision.org/.

Table 3. Comparisons of different definitions of A^* in high-level attention guided activation propagation.

Net	AlexNet	**A-RoI**	A-uncorrelated	A-enhance	A-reduce	A-allsoft	A-allsoft+1
Accuracy (%)	59.0	**62.2**	58.2	52.6	59.5	55.3	55.3

with our MAAP approach. Specifically, we make the following modifications: For AlexNet, we insert our implemented enhancement layer after "relu1", resulting in a mapping resolution of 55 × 55. For VGGNet, we insert enhancement layer after "relu1_1", resulting in a mapping resolution of 224 × 224. For ResNet, we insert enhancement layer after "conv1_relu", resulting in a mapping resolution of 112 × 112. And then we fine-tune each CNN on CUB-200-2011 dataset, obtaining the first CNN with enhancement layer. (3) We further insert our implemented elimination layer. For AlexNet, we insert elimination layer after "relu5", resulting in a mapping resolution of 13 × 13. For VGGNet, we insert elimination layer after "relu5_4", resulting in a mapping resolution of 14 × 14. For ResNet, we insert elimination layer after "res4b2_branch2b_relu", resulting in a mapping resolution of 14 × 14. And then we fine-tune each CNN on CUB-200-2011 dataset. Finally, we obtain the final CNNs.

3.2 Effectiveness of MAAP in State-of-the-Art CNNs

This subsection presents the experimental results and analyses of adopting our MAAP in 3 state-of-the-art CNNs, and analyzes the effectivenesses of the components in our MAAP. Table 1 shows the results of MAAP incorporated with AlexNet, VGGNet and ResNet respectively on CUB-200-2011 dataset. From the experimental results, we can observe:

(i) The application of low-level attention guided activation propagation (LAAP) improves the classification accuracy via enhancing the pivotal information in the forward propagation to help the high-level convolutional layers learn the shape of the object, which boosts the discriminative feature learning. Compared with the results of CNNs themselves, without adopting our proposed MAAP approach, LAAP improves 0.7%, 0.7%, and 0.4% respectively.

(ii) The application of high-level attention guided activation propagation (HAAP) improves the classification accuracy, via retaining the region of interest and removing the background noise of the input data at the same time. Comparing with the results of CNNs themselves, HAAP improves 3.2%, 5.8%, and 2.1% respectively.

(iii) Combination of LAAP and HAAP via alternative training achieves more accurate results than only one of them is adopted, e.g. 63.0% vs. 59.7% and 62.2% of AlexNet, which shows the complementarity of LAAP and HAAP. The two activation propagations have different but complementary focuses: LAAP focuses on enhancing the discriminative features, while

HAAP focuses on dropping the background noise. Both of them are jointly employed to boost the discriminative feature learning, and enhance their mutual promotion to achieve the better performance.

3.3 Comparison with Dropout

Low-level attention guided activation propagation can be regarded as weighting activations of feature maps. Dropout [20,21] randomly drops units from the neural networks during training. It can be regarded as weighting activations, which is a special case of low-level attention guided activation propagation with weights equal to 0 or 1. So we present the results of adopting traditional dropout in different convolutional layers of AlexNet in Table 2. For AlexNet, we add dropout layer after "relu1", "relu2", "relu3", "relu4", "relu5" respectively, which are denoted as "conv1" to "conv5" respectively in Table 2. We can observe that no matter where to add dropout layer, the classification accuracy is not improved. It is because that dropout is performed randomly on units, which may lead to that key information is lost in a large probability in convolutional layers. The experimental results of comparison with dropout show that our proposed MAAP is highly useful for improving the performance of CNNs.

3.4 Effectivenesses of A^* in HAAP

In high-level attention guided activation propagation, we conduct experiments with 6 definitions of A^*. From Table 3, we can see that *A-RoI* and *A-reduce* bring improvements for classification performance. It is because that they all focus on reducing the impact of background noise and preserving the key information simultaneously. The other definitions retain negative impact of background noise more or less.

4 Conclusion

In this paper, the multi-attention guided activation propagation approach (MAAP) has been proposed to improve the performance of CNNs, which explicitly allows CNNs to adaptively highlight or weaken activations of feature maps in the propagation process. The activation propagation can be inserted into state-of-the-art CNNs, enhancing the key information and dropping the less significant information. Experimental results show that the application of MAAP approach achieves better performance on fine-grained image classification benchmarks than the state-of-the-art CNNs.

The future work lies in two aspects: First, we will adopt the low-level attention to guide the feature learning of higher convolutional layers and vice versa. Second, we will also attempt to apply the attention mechanism to compress the neural networks. Both of them will be employed to further improve the effectiveness and efficiency of CNNs.

Acknowledgments. This work was supported by National Natural Science Foundation of China under Grant 61771025 and Grant 61532005.

References

1. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: Proceedings of the IEEE, vol. 86, pp. 2278–2324. IEEE (1998)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems (NIPS), pp. 91–99 (2015)
4. Liu, X., Xia, T., Wang, J., Lin, Y.: Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition. arXiv:1603.06765 (2016)
5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Neural Information Processing Systems (NIPS), pp. 1097–1105 (2012)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
7. Anderson, J.R.: Cognitive Psychology and Its Implications. WH Freeman/Times Books/Henry Holt & Co., New York (1990)
8. Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artif. Intell. **78**(1–2), 507–545 (1995)
9. Karklin, Y., Lewicki, M.S.: Emergence of complex cell properties by learning to generalize in natural scenes. Nature **457**(7225), 83–86 (2009)
10. Zhang, X., Zhaoping, L., Zhou, T., Fang, F.: Neural activities in V1 create a bottom-up saliency map. Neuron **73**(1), 183–192 (2012)
11. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
12. Deng, J., Dong, W., Socher, R., Li, L-J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255 (2009)
13. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset (2011)
14. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Sixth Indian Conference on Computer Vision, Graphics & Image Processing, pp. 722–729 (2008)
15. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: International Conference of Computer Vision Workshop (ICCV), pp. 554–561 (2013)
16. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv:1306.5151 (2013)
17. He, X., Peng, Y.: Fine-grained image classification via combining vision and language. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
18. Fu, J., Zheng, H., Mei, T.: Look closer to see better: recurrent attention convolutional neural network for fine-grained image recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
19. Jia, Y., et al.: Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678. ACM (2014)

20. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
21. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)