Learning Convolutional Features and Text Information to Draw Image

Wenxin Yu
School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China

Chang Liu
School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China

Shiyu Chen
School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China

Kang Xu
School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China

Zhiqiang Zhang*
School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China
zzq.zhangzhiqiang2018@gmail.com

ABSTRACT

In this paper, a more effective and general joint exploration method (JEM) is proposed to synthesize images. By combining the technology of image segmentation, feature extraction, and image synthesis, high-quality images can be generated based on the text description and the convolutional segmentation information. Experiments on the Oxford-102 dataset show that our method is more effective than the GAN-CLS-INT method proposed recently. It also shows that in the training process, using VGG for feature extraction has a faster convergence speed than using AlexNet. Simultaneously, we demonstrate that the segmentation image's background information plays an active role in the training process.

CCS CONCEPTS

• **Computing methodologies** → Modeling and simulation; Model development and analysis; Modeling methodologies.

KEYWORDS

Deep Learning, Image Synthesis, Computer Vision, Image Segmentation, Feature Extraction

ACM Reference Format:
Wenxin Yu, School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, China, Shiyu Chen, Chang Liu, Kang Xu, and Zhiqiang Zhang. 2021. Learning Convolutional Features and Text Information to Draw Image. In *The 2nd International Conference on Computing and Data Science (CONFCDS 2021), January 28–30, 2021, Stanford, CA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3448734.3450462

1 INTRODUCTION

Generating realistic images have always been an important research topic in the field of computer vision. In recent years, the introduction of deep learning technology has made many breakthroughs in this subject, especially the emergence of Generative Adversarial Networks (GAN) makes it possible to generate realistic images that are difficult to distinguish between true and false.

The input of traditional GAN [1, 2] is only noise vector from Gauss distribution or normal distribution, which makes it difficult for traditional GAN to control the generation of image types. Based on this, Mirza et al. [3] proposed conditional GAN that can control image generation by introducing constraints such as image labels and other attributes. Although the introduced conditional constraints have been successfully to control image generation, they often require professional domain knowledge, which makes them less flexible in practical applications.

Recently, Reed et al. [4] proposed to generate images based on natural language. Compared with image attributes, natural language is more flexible and more consistent with people's habits. They proved the feasibility of synthesizing images from natural language and obtained encouraging results. However, on the one hand, the quality of the synthesized images needs to be improved. On the other hand, the text description is global information. Only use text description to synthesize images will bring multi-modality problems——a text description can have more than one corresponding image result. In this situation, the process of image synthesis becomes uncontrollable, which makes the practicability very poor.

To further improve the quality of synthesized images and ensure the practicability of the model, the joint exploration method (JEM) is proposed. The method first obtains additional annotations by preprocessing operation and then uses the obtained annotations to guide the process of image synthesis. In most cases, datasets only

[*] * Corresponding author

CONFCDS 2021, January 28–30, 2021, Stanford, CA, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8957-0/21/01…$15.00
https://doi.org/10.1145/3448734.3450462

contain images and corresponding text descriptions, so the method first gets additional annotations from the images directly, which uses mask R-CNN [5] to segment the images. Then it uses AlexNet [6] or VGG [7] structure to convolute the extracted segmentation image, and the obtained convolution features are used as additional annotations. Finally, the convolution features guide the text description to synthesize the image. Experiments show that the results obtained by our method are better than those of GAN-CLS-INT [4]. Simultaneously, the model trained by our method is more practical because it can control the direction of image synthesis. Besides, experiments also show that the results obtained using background segmentation images are better than those obtained by using only foreground segmentation images, which proves that background information plays a positive role in the guidance process.

Our main contributions are as follows: (1) A image synthesis method with better effectiveness and practicability is proposed; (2) Experiments demonstrate that using the VGG model in the training process has faster convergence; (3) The background information of the segmentation image has played a positive role in the whole generation process is proved through experiments.

The following contents of the paper are as follows: Section II presents the related basic technology. Section III introduces our methods in detail. The experimental results are shown in Section IV and the conclusion of our work is presented in Section V.

2 BACKGROUND

2.1 Conditioned Generative Adversarial Networks

The game theory is used for generative adversarial networks. A generator G and a discriminator D are used to carry out antagonistic games continuously. Finally, an excellent generating ability G is obtained, while a good discriminant ability D is also obtained. A particular game process is shown in Equation 1:

$$\min_{G} \max_{D} V(D,G) = \sum\nolimits_{x\sim p_{data}} \log D(x) + \sum\nolimits_{z\sim p_z} \log(1 - D(G(z))) \tag{1}$$

where p_{data} and p_z represent the original data distribution and noise distribution, respectively.

The goal of generator G is to fit the original data distribution as much as possible so that D can not distinguish true or false. The goal of discriminator D is to distinguish the original data and the fake data generated by G as much as possible. The ability of G and D has been improved in the continuous game of the confrontation.

Although outstanding achievements have been made in GAN, the general GAN can not control the generation of images. In this case, Mirza et al. [3] proposed cGAN. Based on GAN's basic architecture, conditional control is introduced into both generator and discriminator to control the generation of image types. The concrete implementation of cGAN is shown in Equation 2:

$$\min_{G} \max_{D} V(D,G) = \sum\nolimits_{x\sim p_{data}} \log D(x|c) + \sum\nolimits_{z\sim p_z} \log(1 - D(G(z|c))) \tag{2}$$

where c is conditional information, such as image labels, text descriptions, or other data information forms. By introducing c, the process of image generation is controlled.

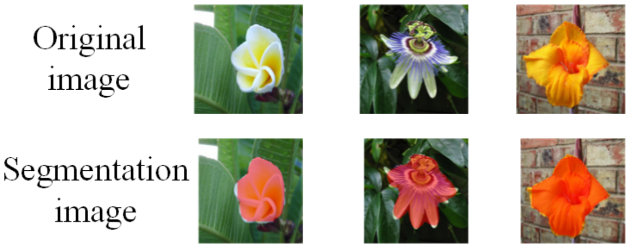

Figure 1: The results of the segmentation image corresponding to the original image are shown above.

2.2 Instance Segmentation

Mask R-CNN [5] is the most effective image segmentation method proposed in recent years. After the rise of deep learning, the convolutional neural network (CNN) has been widely used in image segmentation. R-CNN [8] combines CNN and support vector machine (SVM) to achieve effective instance segmentation. Based on R-CNN's structure, [9] and [10] proposed fast R-CNN and faster R-CNN, which hoisted the whole training speed and achieved a better instance segmentation effect.

In order to further improve the segmentation effect, He et al. [5] proposed mask R-CNN architecture. By introducing the RoIAlign mechanism, the problem of losing part convolution features in the RoIPooling mechanism in faster R-CNN is overcome. It retained all convolution features, thus achieving more satisfactory segmentation results.

2.3 Feature Extraction

The rise of deep learning is largely due to the proposed structure of AlexNet [6]. AlexNet combines ReLU [11] activation function, dropout [12] and max-pooling method. Furthermore, it introduces the local response normalization layer (LRN) to successfully extract features suitable for deep learning structure, making deep learning architecture brilliant. In order to further improve the ability of network expression and extract more effective features, Simonyan et al. [7] proposed the VGG architecture. By deepening the network structure, VGG extracts more excellent convolution features.

3 OUR METHOD

3.1 Joint Exploration Method

The joint exploration method (JEM) uses convolutional features and text information to explore a more versatile and useful image synthesis architecture. The method is mainly divided into three modules: image segmentation module, feature extraction module, and image synthesis module.

In the image segmentation module, the mask R-CNN [5] is used for image segmentation because it is the most widely used and most effective image segmentation method. The flower image segmentation results are shown in Figure 1, which also reflects the practicability and effectiveness of mask R-CNN.

In the feature extraction module, the AlexNet or VGG model is used to extract the segmentation image's features. The output of the seventh layer of AlexNet and the fifteenth layer of VGG will be used as guidance information.

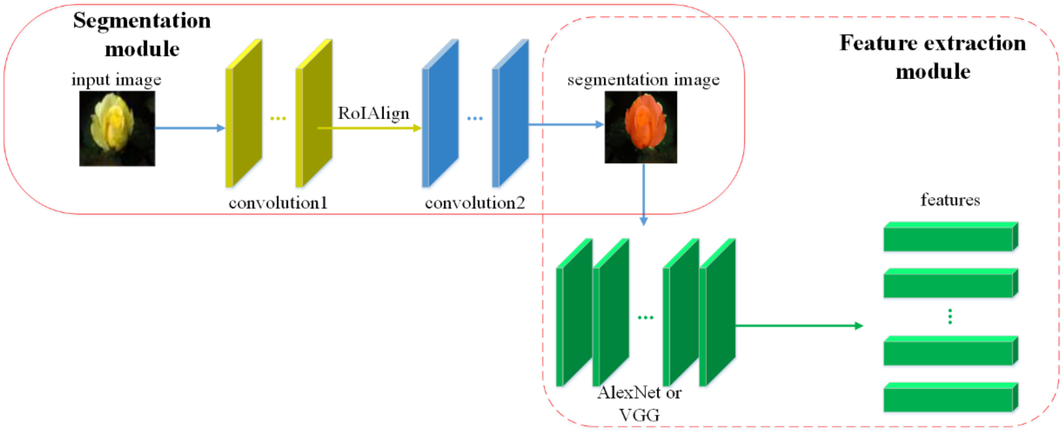

Figure 2: The structure diagram of the preconditioning part. It includes segmentation module and feature extraction module.

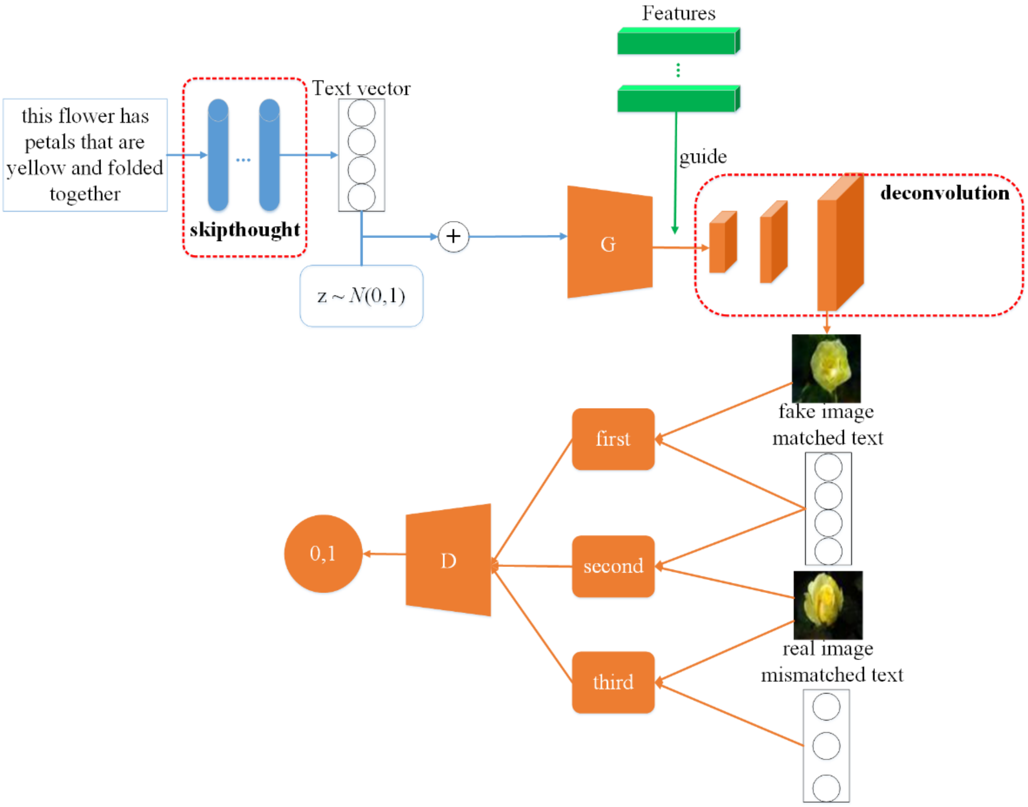

Figure 3: The structure diagram of the image synthesis part. It includes text processing, image synthesis and feature guidance.

In the image synthesis module, the extracted features will be combined with text information to synthesize images. Text information is encoded into text vectors by the skip-thought model [13]. Text vectors are combined with noise vectors as the generator's input. The deconvolution [14] operation is performed in the generator to synthesize images. In the process of generation, the features extracted before are used as guidance information to guide the whole generation process. In the discriminator, the generated image and the original real image are discriminated. In order to improve the matching degree between text information and image, two cases of matched text and mismatched text are added to the discriminator. Through continuous iterative training, the generator can finally synthesize high-quality images corresponding to the text information.

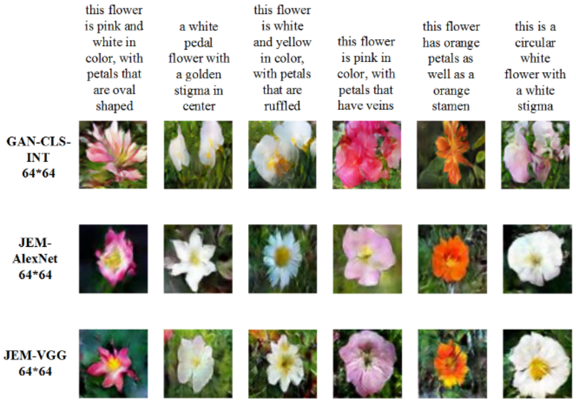
this flower is pink and white in color, with petals that are oval shaped
a white pedal flower with a golden stigma in center
this flower is white and yellow in color, with petals that are ruffled
this flower is pink in color, with petals that have veins
this flower has orange petals as well as a orange stamen
this is a circular white flower with a white stigma
GAN-CLS-INT 64*64
JEM-AlexNet 64*64
JEM-VGG 64*64

Figure 4: The comparison results of GAN-CLS-INT, JEM-AlexNet, and JEM-VGG methods are shown above.

3.2 Network Structure

The network structure is mainly divided into two parts. One is the preprocessing structure, and the other is the image synthesis structure. The preprocessing structure is shown in Figure 2. The segmentation image can be produced by the original image firstly. Then AlexNet or VGG model is used to extract convolution features from the segmentation images.

The structure of image synthesis is shown in Figure 3. First, the skip-thought model is used to encode the text, and then the noise vector is input into the generator, which synthesizes the image by deconvolution. The discriminator discriminates three situations: fake image combined with matched text, original image combined with matched text, original image combined with mismatched text.

Simultaneously, the direction of synthesis is guided by the convolution feature extracted during the generation process.

3.3 Specific details

In the process of image synthesis, the dimension of the noise vector is 100, the text vector is 256, the size of Batch Normalization (BN) [15] is 64, and Adam [16] is used as the optimizer. Two different sizes of results, 64*64 and 128*128, are synthesized. The dimensions of the corresponding generator and discriminator are 64 and 128, respectively.

4 EXPERIMENTAL RESULTS

We conducted experiments on the Oxford-102 [17] dataset. The Oxford-102 dataset contains 8189 flower images of 102 classes (80 are used as the training set, and 20 are used as the test set). Each image corresponds to 10 text descriptions.

4.1 Comparison results

In flower image synthesis, the 64*64 image results are generated based on AlexNet and VGG models. It is compared with the state-of-the-art method (GAN-CLS-INT). The comparison results are shown in Figure 4

The figure shows that the results obtained by our method (JEM-AlexNet and JEM-VGG) are better than those generated by GAN-CLS-INT in detail processing, which shows the effectiveness of our method. By comparing the results of JEM-AlexNet and JEM-VGG, it can be found that there is no significant difference between them. Therefore, we observed the two methods' training process in detail and found that feature extraction based on the VGG model made the image synthesis training have a faster convergence speed. Table 1 shows the specific comparison result. The table shows that JEM-AlexNet converges at about 600th training epochs, and JEM-VGG converges at 300th.

Table 1: The comparison of the convergence speed of JEM-AlexNet and JEM-VGG.

Method	JEM-AlexNet	JEM-VGG
Convergence time	600 epochs	300 epochs

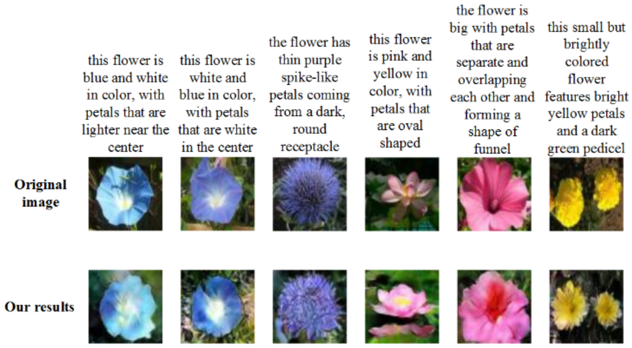
this flower is blue and white in color, with petals that are lighter near the center
this flower is white and blue in color, with petals that are white in the center
the flower has thin purple spike-like petals coming from a dark, round receptacle
this flower is pink and yellow in color, with petals that are oval shaped
the flower is big with petals that are separate and overlapping each other and forming a shape of funnel
this small but brightly colored flower features bright yellow petals and a dark green pedicel
Original image
Our results

Figure 5: The comparison of the original image and our results are shown above.

In view of the existing multi-modal problems mentioned earlier, our method also achieves a good solution. As shown in Fig. 5, our results are compared with the original image. From the comparison results, it can be seen that the results generated by our method are basically consistent with the original image in shape, direction, and quantity, which shows that our method can control image synthesis very well.

4.2 Extended experiments

In order to verify the role of background in the guiding process, an extended experiment using only the foreground image for the guide is conducted. This experiment is conducted in the Oxford-102 dataset. The segmentation images of the flower with a pure blue background on the official website are downloaded firstly. Then the background color of the segmentation images is converted to white to obtain the desired foreground flower image. The foreground results obtained are shown in Figure 6. The results generated with foreground images are compared with those using the whole segmentation image are shown in Figure 7. From the comparison, the results using the segmentation image have better detail processing and better continuity than those using only the foreground image. This shows that the background information in the segmentation image plays an active role in the guidance process.

The image synthesis with higher pixels is also explored with our JEM method. The 128*128 pixels of images are generated and

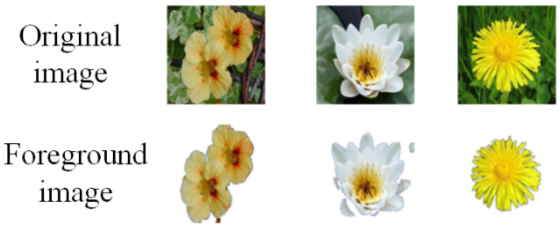

Figure 6: The results of the foreground image are shown above.

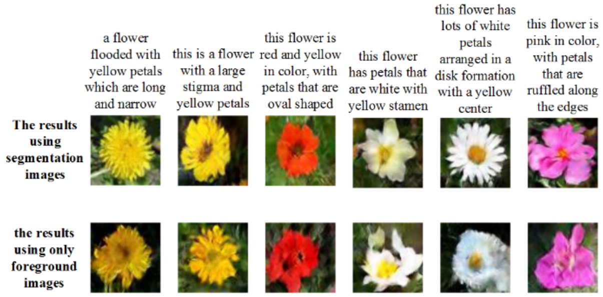
a flower flooded with yellow petals which are long and narrow
this is a flower with a large stigma and yellow petals
this flower is red and yellow in color, with petals that are oval shaped
this flower has petals that are white with yellow stamen
this flower has lots of white petals arranged in a disk formation with a yellow center
this flower is pink in color, with petals that are ruffled along the edges
The results using segmentation images
the results using only foreground images

Figure 7: The comparison results of using only foreground images and using the whole segmentation image are shown above.

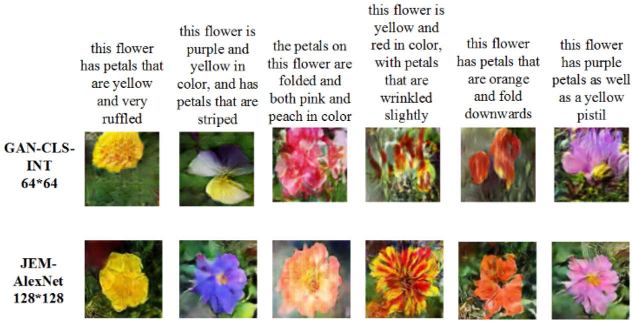
this flower has petals that are yellow and very ruffled
this flower is purple and yellow in color, and has petals that are striped
the petals on this flower are folded and both pink and peach in color
this flower is yellow and red in color, with petals that are wrinkled slightly
this flower has petals that are orange and fold downwards
this flower has purple petals as well as a yellow pistil
GAN-CLS-INT 64*64
JEM-AlexNet 128*128

Figure 8: The comparison results of GAN-CLS-INT (64*64) and JEM-AlexNet (128*128) are shown above.

compared with GAN-CLS-INT, shown in Figure 8. It can be seen that our high-pixel results are still better than GAN-CLS-INT.

5 CONCLUSION

This paper proposes a more general and effective image synthesis method (Joint Exploration Method). On the one hand, this method solves the existing multimodal problem and realizes controllable image synthesis. On the other hand, by combining convolution features and text information, higher quality images are synthesized. Experiments show the effectiveness of our method. At the same time, it also proves that the background information of the segmentation image can play a promoting effect in the whole training process.

ACKNOWLEDGMENTS

This research is supported by Sichuan Science and Technology Program (No. 2020YFS0307, 2020YFG0430, 2019YFS0146, 2019YFS0155).

REFERENCES

[1] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," Montreal, Quebec, Canada, In: Neural Information Processing Systems 27, pp. 2672–2680, December 2014.
[2] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," San Juan, Puerto Rico, In: International Conference on Learning Representations, May, 2016.
[3] M. Mirza, and S. Osindero, "Conditional generative adversarial nets," In: arXiv:1411.1784, 2014.
[4] S. E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," Venice, Italy, In: International Conference on Computer Vision, pp. 2242–2251, October, 2017.
[5] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "mask R-CNN," Venice, Italy, In: International Conference on Computer Vision, pp. 2980–2988, October, 2017.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Lake Tahoe, Nevada, United States, In: Advances in Neural Information Processing Systems, pp. 1106–1114, December, 2012.
[7] K.Simonyan, and A. Zisserman., "Very Deep Convolutional Networks for Large-Scale Image Recognition," San Diego, CA, USA, In: International Conference on Learning Representations, May, 2015.
[8] R.B. Girshick, J. Donahue, T. Darrell, and J. Malik., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," Columbus, OH, USA, In: Computer Vision and Pattern Recognition, pp. 580–587, June, 2014.
[9] R.B. Girshich., "Fast R-CNN," Santiago, Chile, In: International Conference on Computer Vision, pp. 1440–1448, December, 2015.
[10] S. Ren, K. He, R.B. Girshick, and J. Sun., "Faster R-CNN: Towards Real-Time Objects Detection with Region Proposal Networks," Montreal, Quebec, Canada, In: Advances in Neural Information Processing Systems, pp. 91–99, December, 2015.
[11] V. Nair, and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," Haifa, Israel, In: International Conference on Machine Learning, pp. 807–814, June, 2010.
[12] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," In: J. Mach. Learn. Res., vol 15, no 1, pp. 1929–1958, 2014.
[13] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler, "Skip-Thought Vectors," Montreal, Quebec, Canada, In: Advances in Neural Information Processing Systems, pp. 3294–3302, December, 2015.
[14] M. D. Zeiler, and R. Fergus, "Visualizing and understanding convolutional networks," Zurich, Switzerland, In: European Conference on Computer Vision, pp. 828–833, September, 2014.
[15] S. Ioffe, and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," Lille, France, In: International Conference on Machine Learning, 448–456, July, 2015.
[16] D. P. Kingma, and J. Ba, "Adam: a method for stochastic optimization," San Diego, CA, USA, In: International Conference on Learning Representations, May, 2015.
[17] M.-E. Nilsback, and A. Zisserman., "Automated flower classification over a large number of classes," Bhubaneswar, India, In: Indian Conference on Computer Vision, Graphics and Image Processing, pp. 722–729, December, 2008.