THE EFFECT OF BACKGROUND ON A DEEP LEARNING MODEL IN IDENTIFYING IMAGES OF BUTTERFLY SPECIES

Tianyu Xi[1,2], Jiangning Wang[2], Yan Han[2], Tianshan Wang [2] and Liqiang Ji[2]

[1]University of Chinese Academy of Science, beijing, China

[2]Institude of Zoology,Chinese Academy of Science, Beijing, China

ABSTRACT

The biodiversity of Lepidoptera plays an vital role in ecological systems. Identification and recognition are the first steps in assessing biodiversity status. Automatic image-based identification tools for butterflies (Lepidoptera) can help taxonomists identify species effectively. In this paper, we propose a deep learning method with improved model accuracy and target object detection; the method was improved by the removal of image backgrounds. We compiled our dataset from images of Chinese butterfly specimens and replaced the images' original backgrounds with a green background in Photoshop. We then trained two models based on the original image set and the green-background set. Our results achieved training model accuracy of 98% and showed that the image background removal enhanced model generalizability, providing better results for test sets with different kinds of image data. Additionally, the experiments proved that it is feasible to use a small amount of data to train deep learning models, which could have wide applicability in the image recognition field.

KEYWORDS

Butterfly classification; Biodiversity; Background remove; Deep learning; Alexnet

1. INTRODUCTION

Butterflies are important in biodiversity; there are approximately 16000 recorded species around the world (Y. Zhou, 1994). Butterfly species diversity responds to climate change, usually in the form of northward or elevation range shifts (Walther et al., 2002). They also interact with plants in coevolution; not only can plant diversity increase the diversity of animals, but also vice versa (Ehrlich & Raven, 1964). Currently, for the sake of further reveal the evolution of these species' ecological status, scientists are focused on maintaining the diversity of species in each ecosystem, as their numbers have dramatically decreased. Therefore, the classification of species is crucial but complicated and difficult. The identification of Lepidoptera species requires information on their morphology. Traditional methods of insect taxonomy differentiate species of butterflies by analysing the colour and size of the wing spot, the wing veins and other anatomical features, including the labial palpus, tarsus and external genitalia (M. Q. Zhou, Geng, & Huang, 2006). Many automatic methods have been developed to help entomologists with identification (Martineau et al., 2017). The use of automatic identification systems can not only reduce damage to specimens but also facilitate the integration of massive amounts of data.

Over the past few decades, automatic insect recognition and classification research has given extra attention to image capturing and processing tools, which have proven useful to non-specialists and ecologists aiming in rapid and objective recognition of butterflies at the family or order level .

DOI: 10.14810/elelij.2019.8101

2. MATERIAL AND METHODS

2.1. Image sources

All images for the training set were collected from two sources: images of specimens scanned from a monograph on Chinese butterflies (Y. Zhou, 1994) and a Taiwanese butterfly website (Hong, Chen, & Hsiang, 2000).

2.2. Pre-processing and Datasets

There are 4494 scanned images in total covering six families of butterfly. Ten percent of these images were extracted as a validation set. To analyse the model's generalizability, we produced two kinds of test sets. One kind consists of 5-10 images randomly selected from the original collection of scanned images, and the other kind uses Taiwanese butterfly photos.

In this paper, before training the model, we removed the original image backgrounds using the Photoshop process tools, replacing them with a solid green colour (R=0, G=255, B=0). We henceforth refer to this dataset by "MG" and to the dataset of images as originally scanned by "MO". "TO" represents the original dataset of Taiwanese butterfly images; "TG" is this dataset with green backgrounds. Table 2 shows the distribution of images in each dataset by butterfly family. Figs. 1-2 are sample images.

2.3 Classification methods

The experiment was carried out using a finetuned version of the AlexNet model (Krizhevsky, Sutskever, & Hinton, 2012) in the deep learning framework Caffe (Jia et al., 2014). We trained the model on the butterfly image datasets and used cross validation to test the model's generalizability. Then, we analysed the results.Convolutional neural networks (CNNs) designed to process two-dimensional images are variants of multi-layer perceptrons (MLPs), which are inspired by animal neural structures. The input space of these filters is local, so it is beneficial for exploiting the robust spatially local correlation appear in natural images (Hubel & Wiesel, 1968; Lecun, Bottou, Bengio, & Haffner, 1998). In this paper, we applied the CNN model AlexNet.

Image classification on the CNN architecture known as AlexNet was proposed by Alex Krizhevsky and won the 2012 ImageNet Large Scale Visual Recognition Challenge (Krizhevsky et al., 2012). There are eight layers in the network, containing five convolutional layers and three fully connected layers (Fig. 3). All feature extractors were initialized with white Gaussian noise and learned from data, and the resulting feature maps were passed through rectified linear units (ReLUs), a type of non-linearity unit. The model was trained using stochastic gradient descent and the backpropagation algorithm included in Caffe (Jia et al., 2014), with the learning rate policy set to "step". The learning rate set to 10-2 for all layers initially for the reason of accept the newly defined last fully connected layer set to 10-2. The learning rate (lr) was decreased by a factor of 10 every 69 iterations and training was stopped after 6900 iterations. The number of units in the third fully connected layer (fc3) was changed according to the number of classes of training data. We set the batch size to 69 and momentum to 0.9 and applied L2 weight decay with penalty multiplier set to 5 × 10-4, dropout ratio set to 0.5, CPU mode

3.RESULTS

We trained two classification models on the MG and MO image datasets, then tested each model's performance on several test sets. Table 3 shows the test results.

3.1. MG model vs MO model

The experiment shows that we can get a well-trained deep learning model based on the AlexNet framework using only small datasets and fine-tuned parameters. This model can classify images of butterflies at family level

3.2.Test Accuracy

Table 2 shows that the original image model and background-removed model have almost the same validation performance (to two decimal places), reaching 98% accuracy, while the MG model performed better on the test sets, which means that training with the MG dataset leads to better ability to generalize to these test sets.

3.3. Figures and Tables

Table 1. Butterfly datasets

Dataset	ID	number of images							pre-processing	image source
		Lycaenidae	Nymphalidae	Pieridae	Papilionidae	Satyridae	Hesperiidae	Total		
Training set	MO	823	1147	406	347	702	560	3985	none	(Zhou, 1994)
	MG	828	1152	413	351	706	565	4015	Green background	
Validation set	MO	92	128	46	39	78	63	446	none	
	MG	92	129	46	40	79	63	449	Green background	
Test set	MO	5	5	5	5	5	5	30	none	
	MG	5	5	5	5	5	5	30	Green background	
	TO	202	100	56	100	30	89	577	none	(Hong et al., 2000)
	TG	202	100	56	100	30	89	577	Green background	

Figure 1. The left is original image, the right is pre-processed image used for training set

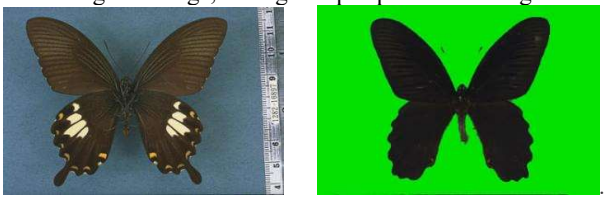

Figure 2. Testing image

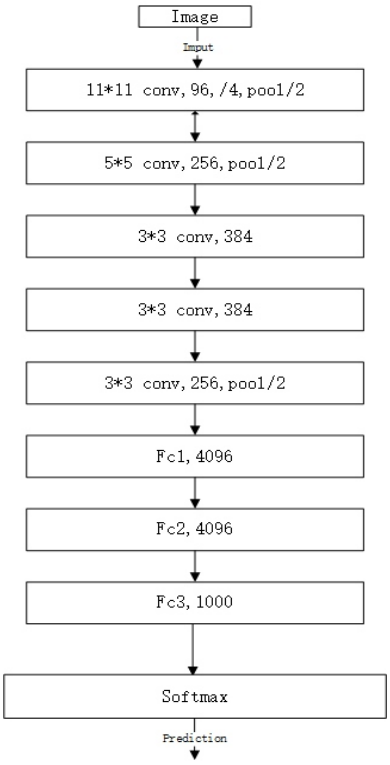

Fig. 3: Flowchart for AlexNet model.

Notes: Convs1 padding: 2
All other Convs padding: 1
The first Overlapping Pooling layer: (Kernel size: 5, Stride: 4)
The rest all Overlapping Pooling layers: (Kernel size: 3, Stride: 2)

Table 2. Model test result

Dataset	ID	MO model	MG model
		Accuracy	
Validation set	MG	0.98	
	MO		0.98
Testing set	MG	0.98	0.43
	MO	0.36	0.98
	TO	0.24	0.27
	TG	0.16	0.75

Table 3. Test results for MG model by Family

Class	Papilionidae	Pieridae	Lycaenidae	Nymphalidae	Satyridae	Hesperiidae
Balanced images	380	380	380	380	380	380
Recognition rate	0.86	0.97	0.84	0.67	0.79	0.3
Original images	351	413	828	1152	706	565
Recognition rate	0.94	0.92	0.82	0.79	0.76	0.04

Figure 4. Processed by Photoshop

Figure 5. Bad samples processed by automatic tools.

4. DISCUSSION

4.1. Background removal improves results

The experiments were performed on a small-size dataset without data augmentation. The model results demonstrate that it is feasible to use small dataset to train deep learning models as long as the parameters are chosen properly. The MG and MO models can both get good training results and they perform well in their own validation set, proving that deep learning is a powerful method for classification model training. In this case, the butterfly specimen data sets were standardized, and determining whether the models can achieve the same good results with other various complex images or not will require more experiments after collecting more images from different sources.

It is widely known that learning high-level features always leads to overfitting problems. In our case, the results show that the original model is as good as the background-removed model, indicating the noise caused by the scan can reduce overfitting to some extent, although it lacked generalizability.

Our experiment suggests that for identification of specimen images, it is better to remove the original background when using deep learning methods to train a classification model. Pre-processing the image to be identified (consistently with the images in the training set) can greatly improve classification accuracy.

4.2. Factors affecting model performance

4.2.1 Data source

There are some factors that affect classification accuracy with scanned images may be introduced by the scanning process. These factors may include: scanner quality, scan resolution, type of printed documents (laser-printed or photocopied), paper quality and complicated background (Alginahi, 2010). We are considering a new method to avoid these problems, that is, by pre-processing the scanned images in some way. Some existing methods can be used for processing and removing backgrounds, for example, edge detection, which greatly reduces the amount of data and eliminates information that can be considered irrelevant, while preserving the important structural attributes of the image (Pala & Pala, 1993). However, such methods cannot remove backgrounds well. Therefore, we used Photoshop, which is easy to use and helps clear background (Figure 4). Figure 5 shows some bad samples.

4.2.2 Data distribution

In the machine learning literature, much research has focused on problems with imbalanced data (He & Garcia, 2009). Previous studies have noted that imbalanced data would make the significant negative impact on CNN models (J. N. Wang et al., 2017). To compare balanced data performance with imbalanced data, we used data resampling in training a new model. The new model covered six classes, where the classes are evenly distributed. Validation accuracy was only 94% because the data size has decreased (Masko & Hensman, 2015). The recognition rate for the family Hesperiidae increased, and the other rates have changed at the same time (Table 3).

We used images of six butterfly families to make our dataset because the image size and number are similar. The biological classification of these butterflies is recognized by taxonomists (M. Q. Zhou et al., 2006), so we can only use these known categories to make labels.

4.3 Butterfly morphological classification based on images

In the classification of butterflies, the criteria used to distinguish species are their morphological characteristics, including internal structure and external structure. Wing veins, color and external genitalia were used to make a detail description of species (Frost, 1942). It is currently feasible to use supervised learning to identify species according to known classifications and to assist scientists in preliminary identification of species (Martineau et al., 2017). However, because of differences in the division of tax a of the world, there could be misjudgment due to the controversial nature of the species itself (M. Q. Zhou et al., 2006). In addition, the non-synchronous evolution of butterflies' molecular and morphological characteristics (Nice & Shapiro, 1999) and wing patterns mimicking other closely related species (Joron et al., 2011) can make classification ambiguous; as a result, some species are still worth refining when making labels.

ACKNOWLEDGEMENTS

It's glad to thank all the authors and collaborators who have contributed to the paper. First, I am deeply grateful to Prof Liqiang Ji and Dr Jiangning Wang, who gave me detailed research advice. Yan Han helped collect the images and Tianshan Wang assisted with coding. This project was supported by the National Natural Science Foundation of China (31501841).

REFERENCES

[1] Alginahi, Y. (2010). Preprocessing Techniques in Character Recognition: InTech.

[2] Arbuckle, T. (2002). Automatic identification of bees' species from images of their wings. Paper presented at the Proc. 9th Int. Workshop on Systems, Signals and Image Processing, Manchester.

[3] Chen, B. C. (2000). Content-based image retrieval of butterflies. (Master), National Taiwan University, Taibei.

[4] Chen, Y.-S., Kao, T.-C., Yu, G.-J., & Sheu, J.-P. (2004, 2004). A mobile butterfly-watching learning system for supporting independent learning. Paper presented at the Wireless and Mobile Technologies in Education, 2004. Proceedings. The 2nd IEEE International Workshop on.

[5] Ehrlich, P. R., & Raven, P. H. (1964). BUTTERFLIES AND PLANTS: A STUDY IN COEVOLUTION. Evolution, 18(4), 586-608. doi: 10.1111/j.1558-5646.1964.tb01674.x

[6] Faruk, E. Ö., Kaya Yılmaz, K. L., & Ramazan, T. (2015). Identification of Butterfly Species by Similarity Indexes Based on Prototypes. Journal ISSN: TBA, 1.

[7] Faruk, E. Ö., Yılmaz, K., Lokman, K., & Ramazan, T. (2015). A Vision System for Classifying Butterfly Species by using Law's Texture Energy Measures. International Journal on Computer Vision, Machine Learning and DataMining, 1, 16-24.

[8] Feng, L., Bhanu, B., & Heraty, J. (2016). A software system for automated identification and retrieval of moth images based on wing attributes. Pattern Recognition, 51, 225-241.

[9] Frost, S. W. (1942). General entomology. General Entomology.

[10] Grajales-Múnera, J. E., & Martínez, A. R. (2013). Butterfly Classification by HSI and RGB Color Models Using Neural Networks. [Clasificación de Mariposas por Modelos de Color HSI y RGB Usando Redes Neuronales]. Revista Tecno Lógicas, Edición Especial, 669-679.

[11] He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

[12] Hong, J., Chen, H., & Hsiang, J. (2000). A digital museum of Taiwanese butterflies. Paper presented at the Proceedings of the fifth ACM conference on Digital libraries, San Antonio, Texas, United States.

[13] Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology, 195(1), 215-243.

[14] Institute of Zoology. (2010). China Patent No. 2010SR000756.

[15] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., . . . Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. Paper presented at the Proceedings of the 22nd ACM international conference on Multimedia.

[16] Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., . . . Ferguson, L. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. nature, 477(7363), 203-206.

[17] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Paper presented at the International Conference on Neural Information Processing Systems.

[18] Lecun, Y. (2013). Deep Learning Tutorial.

[19] Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[20] Liu, F., Shen, Z.-R., Zhang, J.-W., & Yang, H.-Z. (2008). Automatic insect identification based on color characters. Chinese Bulletin of Entomology, 45(1), 150-153.

[21] Martineau, M., Conte, D., Raveaux, R., Arnault, I., Munier, D., & Venturini, G. (2017). A survey on image-based insect classification. Pattern Recognition, 65, 273-284. doi: https://doi.org/10.1016/j.patcog.2016.12.020

[22] Masko, D., & Hensman, P. (2015). The Impact of Imbalanced Training Data for Convolutional Neural Networks. (Independent thesis Basic level (degree of Bachelor) Student thesis). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-166451 DiVA database.

[23] Nice, C. C& Shapiro, A. M. (1999). Molecular and morphological divergence in the butterfly genus Lycaeides (Lepidoptera: Lycaenidae) in North America: evidence of recent speciation. Journal of evolutionary biology, 12(5), 936–950.

[24] O'Neill, M. A., Gauld, I. D., Gaston, K. J., & Weeks, P. J. D. (2000). Daisy: an automated invertebrate identification system using holistic vision techniques. Paper presented at the Inaugural Meeting of the BioNET-International Group for Computer-aided Taxonomy, Cardiff.

[25] Pala, N. R., & Pala, S. K. (1993). A review on image segmentation techniques. Pattern Recognition, 26(9), 1277-1294.

[26] Walther, G., Post, E., Convey, P., Menzel, A., Parmesan, C., Beebee, T. J. C., . . . Bairlein, F. (2002). Ecological responses to recent climate change. nature, 416(6879), 389-395.

[27] Wang, J., Ji, L., Liang, A., & Yuan, D. (2012). The identification of butterfly families using content-based image retrieval. Biosystems Engineering, 111(1), 24-32. doi: doi:10.1016/j.biosystemseng.2011.10.003

[28] Wang, J. N., Chen, X. L., Hou, X. W., Zhou, L. B., Zhu, C. D., & Ji, L. Q. (2017). Construction, implementation and testing of an image identification system using computer vision methods for fruit flies with economic importance (Diptera: Tephritidae). Pest management science, 73(7), 1511.

[29] Yeh, R. B., Liao, C., Klemmer, S. R., Guimbretiere, F., Lee, B., Kakaradov, B., . . . Paepcke, A. (2006). ButterflyNet: A Mobile Capture and Access System for Field Biology Research. Paper presented at the Conference on Human Factors in Computing Systems (CHI 2006), Montreal, Quebec, Canada. http://ilpubs.stanford.edu:8090/853/

[30] Zhou, M. Q., Geng, G. H., & Huang, S. G. (2006). Ontology Development for Insect Morphology and Taxonomy System. Paper presented at the International Conference on Web Intelligence and Intelligent Agent Technology Workshops.

[31] Zhou, Y. (1994). Monographia Rhopalocerorum Sinensium. Zhengzhou: Henan Science and Technology Press.