Performance improvements of neural networks on image classification tasks with depth image

Zizui CHEN, M.Eng.; Stephen Czarnuch, P.Eng., PhD,

***Abstract*—Computer vision, especially object classification, has been significantly enhanced by the advancement of machine learning [1]. Sophisticated artificial neural networks, such as LeNet [2], AlexNet [3] and YOLO [4], archive at least 80% accuracy [3], [5], [6] in general classification. However, existing deep learning algorithms have only been evaluated on RGB images. Commonly, approaches for deep learning networks process RGB images to extract features from the entire image [7]. During the process, the networks generally first detect low-level features (e.g., edges, corners) and move to high-level features (e.g., scale variations), ignoring other information, such as 3D object details. Depth images include 3-D details of the objects and offer more information to the classifier than intensity edges. We introduce a novel classification approach which uses depth images, instead of RGB images, to train deep learning algorithms. We conducted four preliminary experiments using 6,387 frames of common household objects, a Memorial University of Newfoundland's multi-cameras dataset, and a in-house human dataset. We observed mixed results on accuracy from a traditional CNN (three convolution layers) by training this network with depth images and colour images. Further, the sizes of training data are reduced by two-thirds due to the smaller size of the depth images, and our approach is more robust against adversarial attacks. Our approach significantly reduced the dataset size and training time, while showing acceptable results in certain indoor scenarios, when depth images are easy to capture with commercially available, inexpensive IR sensors, such as Microsoft Kinect.**

***Index Terms*—Depth images, RGB-D, image classification, depth image processing, security, machine learning.**

I. Introduction

Computer vision, especially object classification, has been significantly enhanced by the advancement of machine learning [1]. Sophisticated artificial neural networks, such as LeNet [2], AlexNet [3] and YOLO [4] archive at least 80% accuracy in general classification [3], [4], [6]. These approaches feature on processing RGB images to automatically extract features from the original image without pre-processing [7]. During the process, the networks generally first detect low-level features (e.g. edges, corners) and move to high-level features (e.g., scale variations) [8]. However, the information utilised by these approaches include only 2D features, since RGB images contain no 3D information. In addition, the traditional approaches are vulnerable to adversarial attacks. Furthermore, the attacks are relative easily to be carried out since the attacker only needs to apply special patterns or features to objects.

In contrast, depth images include 3D details of the objects and offer more information to the classifier than only intensity edges. We believe that artificial neural networks trained with depth images have the advantages of better performance, smaller data size, and are more robust against traditional adversarial attacks.

II. Literature review

Existing works on image classification with artificial neural networks are mainly focus on RGB images for both supervised [4], [3], [2] and unsupervised learning [9]. Recent work [10], [11] in supervised training for RGB image classification tasks significantly improved the performance accuracy [12], [13]. However, these approaches ignore (or at least, fail to exploit) 3D details of the object due to the limitation of RGB images, and most of them are weak to adversarial attacks [14], [15], [16], [17], [18].

Depth images reveal 3D details of objects. Special equipment, including LiDAR [19], [20] and active infrared cameras, are capable of generating depth images. With the commercialisation of inexpensive IR cameras (e.g., Microsoft Kinect [21]), depth images are now widely accessible [22]. However, existing research on applying depth images to object classification tasks are limited [23], [24]. We focus on using depth images as the only source to train a neural network for object classification tasks.

III. Methodology

A. Dataset

We perform our experiments on four datasets:

1) Dataset 1: a sub-dataset from University of Washington's RGB-D Object Dataset [25]. This dataset is pre-segmented and includes common household objects. We resized/normalised the images in this dataset to 256x256 for both colour and depth images in order to train our network. Our dataset included 10 household objects[1]. The size of the dataset is 4.67GB for colour images and 1.55GB for depth images, in the format of *npy*[2]. This dataset includes 6,387 frames of images. Sample images are shown in fig.1
2) Dataset 2: a sub-dataset from Memorial University of Newfoundland's multi-camera dataset [26]. This dataset is not segmented and includes four different human participants. The size of each image is 1080x1920, 3-channels for colour images and 320x288, 16-bit for depth images. We include data from three participants

[1]The following objects are used in this research: apple_1, banana_1, bell_pepper_1, calculator_1, coffee_mug_1, dry_battery_1, hand_towel_1, potato_1, shampoo_1, water_bottle_1

[2]Standard binary file format in NumPy

in our current project. The size of the dataset is 8.22GB for colour images and 124MB for depth images in the format of *npy*. This dataset includes 355 frames of images. A sample image is shown in fig.2

3) Dataset 3: a in-house collected dataset with three participants. This dataset is background removed and includes only depth images. Data are collected with Microsoft Kinect 2 [21]. All data in the dataset are included in this research. The size of each image is 424x512 with 16-bit depth channel. The dataset includes 2,996 images. A sample image is shown in fig.3
4) Dataset 4: similar to dataset 3, this dataset includes 1,642 frames full depth images (without background removal) for the same participants. A sample image is shown in fig.4

(a) Apple (b) Calculator (c) Dry battery

Fig. 1: Sample images from dataset 1

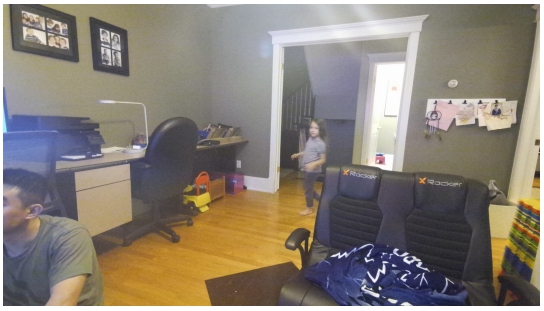

Fig. 2: Sample image from dataset 2

Fig. 3: Sample image from dataset 3

Fig. 4: Sample image from dataset 4

TABLE I: CNN layers summary

Layer type	Filter	Kernel size	Units
Reshape	-	-	-
Conv2D	32	3x3	-
MaxPooling2D	-	-	-
Conv2D	64	3x3	-
MaxPooling2D	-	-	-
Conv2D	64	3x3	-
MaxPooling2D	-	-	-
Flatten	-	-	-
Dense	-	-	64
Dense	-	-	# classes

B. CNN architecture

In this paper, we present a comparison of the performances of a simple CNN trained by colour images and depth images. The network architecture is illustrated in fig.5. We repeat the feature extraction, three times, and our network includes 10 layers, where three convolutional layers have an activation function of ReLU [27], [28]. The parameters of the convolutional layers are listed in tab.I.

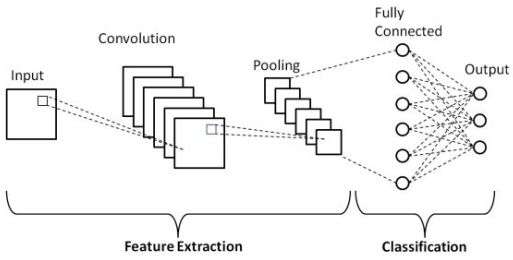

Fig. 5: CNN architecture[29]

IV. Experiment and Results

We run standard k-fold with $k = 3$ to cross-validate the network performance in dataset 1, 2, 3 and 4. Validation accuracies are listed in tab.II and tab.III.

Although the performance of the CNN trained with depth images (tab.III) is worse than the CNN trained with RGB images (tab.II) on dataset 1 and 2, the accuracy for classification

TABLE II: k-fold accuracies for CNN trained with RGB image

Iterations	Dataset 1	Dataset 2
1	1.000	0.588
2	0.999	0.567
3	0.991	0.508
Average	0.997	0.555

TABLE III: k-fold accuracies for CNN trained with depth image

Iterations	Dataset 1	Dataset 2	Dataset 3	Dataset 4
1	0.769	0.521	0.994	0.897
2	0.752	0.457	1.000	0.917
3	0.345	0.508	0.997	0.890
Average	0.622	0.495	0.997	0.901

tasks (dataset 3 and 4) is high, averaging 0.997 and 0.901 respectively. In dataset 1, the overall performance of the CNN trained by depth images is significant lower (by 38%) than the network trained by RGB images. In dataset 2, the performance discrepancy is reduced to 6%. However, in dataset 3 and dataset 4, the network achieves high classification accuracies, at 99.7% and 90.1%, respectively.

Depth and colour images in dataset 2 have different resolutions. In order to compare the ratio of sizes of colour images and depth images, we estimate the size of dataset 2 based on *size per pixel*.

Assume S_{d1} and S_{d2} represents the size of depth images in dataset 1 and dataset 2, respectively; S_{c1} and S_{c2} represents the size of colour images in dataset 1 and dataset 2, respectively.

The ratio between the size of colour image and depth images in dataset 2 is

$$R_2 = \frac{S_{c2}/(1080 \times 1920) \times (320 \times 288)}{S_{d2}} \tag{1}$$

$$\approx 3 \tag{2}$$

The ratio of dataset 1 can be directly computed with

$$R_1 = \frac{S_{c1}}{S_{d1}} \tag{3}$$

$$\approx 3 \tag{4}$$

It is obvious that the size of depth dataset is $2/3$ times smaller than the size of RGB dataset.

V. Conclusion and future works

Our research shows that training a simple CNN with depth images has mixed performance compared to training a CNN network with the exact same parameters with RGB images. The performance could be significantly different depending on the input dataset. The overall performance for a CNN trained by depth images is still acceptable for classification tasks. Furthermore, the approach we proposed in this paper shows the advantage of smaller dataset size, faster training time and the potential of higher robustness against adversarial attacks.

As a preliminary research project, our findings reveal questions about the huge performance discrepancies between dataset 1, 2 and dataset 3, 4 unresolved. Future work should include investigating the performance discrepancies, evaluating performance on other public datasets and extending the research to other more sophisticated artificial neural networks. Theoretical research on the behavioural differences between depth images and RGB images in convolution operations may reveal the fundamental reason for the performance discrepancies observed.

References

[1] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, vol. 2018, 2018.

[2] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[5] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] F. Sultana, A. Sufian, and P. Dutta, "Advancements in image classification using convolutional neural network," in *2018 Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*. IEEE, 2018, pp. 122–129.

[7] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai *et al.*, "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.

[8] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[9] Q. V. Le, "Building high-level features using large scale unsupervised learning," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 8595–8598.

[10] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," *arXiv preprint arXiv:1602.07261*, 2016.

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[12] H. Touvron, A. Vedaldi, M. Douze, and H. Jégou, "Fixing the train-test resolution discrepancy," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[13] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[14] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[15] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.

[16] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.

[17] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1625–1634.

[18] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[19] T. S. Taylor, *Introduction to Laser Science and Engineering*. CRC Press, 2019.

[20] N. O. US Department of Commerce and A. Administration, "What is lidar," Oct 2012. [Online]. Available: https://oceanservice.noaa.gov/facts/lidar.html

[21] M. Corp, "Kinect for windows." [Online]. Available: https://developer.microsoft.com/en-us/windows/kinect
[22] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, no. 2, pp. 4–10, 2012.
[23] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust rgb-d object recognition," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 681–687.
[24] Jie Tang, S. Miller, A. Singh, and P. Abbeel, "A textured object recognition pipeline for color and depth image data," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 3467–3474.
[25] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 1817–1824.
[26] S. C. Chengsi Zhang, Zizui Chen, "Toward synchronized, interference-free, distributed 3d multi-sensor real-time image capture," submitted to NECEC 2020.
[27] R. H. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, 2000.
[28] R. H. Hahnloser and H. S. Seung, "Permitted and forbidden sets in symmetric threshold-linear networks," in *Advances in neural information processing systems*, 2001, pp. 217–223.
[29] V. H. Phung, E. J. Rhee *et al.*, "A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets," *Applied Sciences*, vol. 9, no. 21, p. 4500, 2019.