ARI2129 Principles of Computer Vision for AI

Part A Question 1

David Cachia Enriquez

david.cachia.enriquez.21@um.edu.mt

Jan Formosa

jan.lawrence.formosa.21@um.edu.mt

Matthew Kenely

matthew.kenely.21@um.edu.mt

Question 1

1.1 Introduction

Data augmentation is a technique used in machine learning [1] to artificially inflate the quantity of training data, with the aim of improving the generalisability of learning models [2]. In the field of computer vision, this is done by performing a number of geometric or colour-based transformations on a subset of the images in the training data set [3]. Data augmentation is of particular importance due to the large amount of manual annotation required (this challenge is highlighted with respect to object detection in [4]), as well as the need for computer vision models to detect and extract abstract features within images.

1.2 Historical Context

1960s: The history of computer vision spans several decades. A work from this time period is Larry Roberts ' "Machine Perception of Three-Dimensional Solids" [5] wherein he created a computer vision method to recognise three-dimensional objects, their orientation, and position, within two-dimensional images.

1970s: A greater focus on the understanding of entire scenes within images, achieved through edge extraction and line labelling [6] [7].

1980s: The 1980s saw the development of object contour detection [8] as well as optical flow (estimation of the motion of objects in video) [9].

1990s: The emergence of new machine learning techniques (e.g. SVM [10]) and the adoption of convolutional neural networks as a foundation for more sophisticated computer vision models. Significant progress was made in face recognition algorithms [11] and video surveillance [12].

2000s: Characterised by large image datasets such as ImageNet [13] and the development of convolutional neural network architectures such as AlexNet [14], VGG-19 [15] and ResNet [16]. These architectures have been shown to be applicable to a wide range of computer vision tasks with great accuracy [17] [18] [19].

1.3 Data Augmentation Techniques

Flipping

The image is mirrored horizontally or vertically [20] – useful in situations where images are mirrored in the real world (e.g. mirrors).

Rotation

The image is rotated around its centre by a degree in the range $[0, 360]$ [20] [21] – useful in situations where objects are viewed from different angles.

Scaling

The image is resized by a factor $s > 0$ [20] [21] – helps in making models invariant to resolution changes, viewing objects from different distances and capturing features at multiple scales.

Cropping

A subset (usually square) of the image is kept [20] – useful when convolutional neural networks require images to have specific dimensions [22].

Colour Jittering

The colours of the image are changed using either a random [23] or set [24] combination and permutation of hue, saturation, and brightness adjustments – useful in situations where images are taken under different lighting conditions and to accommodate for variance in camera settings such as white balance.

1.4 Recent Advancements

Cutout

Random or specific parts of the image are intentionally occluded (set to black) – this imitates real world situations where objects are partially observable and helps models to learn to extract more context from different areas in the image [25].

Mixup

New images are generated by linearly interpolating (blending) the pixel values of prior two images – this helps to generalise the model as it is trained to behave linearly (smoothing out of decision boundaries) in the "gaps" in information between training images [26].

AutoAugment

This technique approximates the best data augmentation policies for a given dataset and task using a machine learning procedure. E. D. Cubuk et al. [27] used reinforcement learning to obtain effective combinations of data augmentation techniques very efficiently as the need for manual tuning was eliminated.

1.5 Application in Computer Vision

Data augmentation has always been a significant asset in the diversification and generalisation of training data sets (particularly when training CNNs). One of the earliest examples of this is the LeNet-5 architecture developed by Y. LeCun et al. [28] in 1998 which artificially distorted training images, and showed this technique to reduce test error rate when carrying out handwritten digit recognition. The following are examples of recent state-of-the-art computer vision models which have used and introduced data augmentation techniques:

YOLOv4

A. Bochkovskiy et al. [29] coined the term "bag of freebies" to describe methods which improve the performance of a model without significant training and resource overhead. In the case of YOLOv4, whose task is object detection, the bag of freebies consisted of data augmentation techniques, specifically, CutOut, MixUp and CutMix (a combination of CutOut and MixUp) with the authors introducing two new techniques: Mosaic (like CutMix but with 4 prior images) and Self-Adversarial Training (the neural network intentionally sabotages images in one stage and is trained on the sabotaged image in the second stage).

AlexNet

A. Krizhevsky et al. [14] employed two forms of data augmentation when training AlexNet: image translation and horizontal reflection, and multiplication of the image RGB values proportional to the corresponding eigenvalues found using Principal Component Analysis, multiplied by a random number generated using a gaussian distribution.

References

[1] T. M. Mitchell, *Machine learning*. McGraw-hill New York, 2007, vol. 1.

[2] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[3] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*. Springer, 2020, pp. 566–583.

[4] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1301–1310.

[5] L. G. Roberts, "Machine perception of three-dimensional solids," Ph.D. dissertation, Massachusetts Institute of Technology, 1963.

[6] M. Clowes, "On seeing things," *Artificial Intelligence*, vol. 2, no. 1, pp. 79–116, 1971. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0004370271900051

[7] A. Rosenfeld, R. A. Hummel, and S. W. Zucker, "Scene labeling by relaxation operations," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 6, pp. 420–433, 1976.

[8] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[9] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.

[10] E. Osuna, R. Freund, and F. Girosit, "Training support vector machines: an application to face detection," in *Proceedings of IEEE computer society conference on computer vision and pattern recognition*. IEEE, 1997, pp. 130–136.

[11] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[12] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," *Transportation Research Part C: Emerging Technologies*, vol. 6, no. 4, pp. 271–288, 1998.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 248–255.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[19] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3128–3137.

[20] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *2018 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 2018, pp. 1542–1547.

[21] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *Monthly notices of the royal astronomical society*, vol. 450, no. 2, pp. 1441–1459, 2015.

[22] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[23] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep image: Scaling up image recognition," *arXiv preprint arXiv:1501.02876*, vol. 7, no. 8, p. 4, 2015.

[24] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2014, pp. 806–813.

[25] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[26] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[27] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 113–123.

[28] 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.

[29] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.