Viewpoint Diversity Improves Convolutional Neural Network Generalization and Robustness

Yifan Luo (yifan.luo@student.uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

Niklas Müller (n.muller@uva.nl)
Psychology Research Institute, University of Amsterdam, The Netherlands

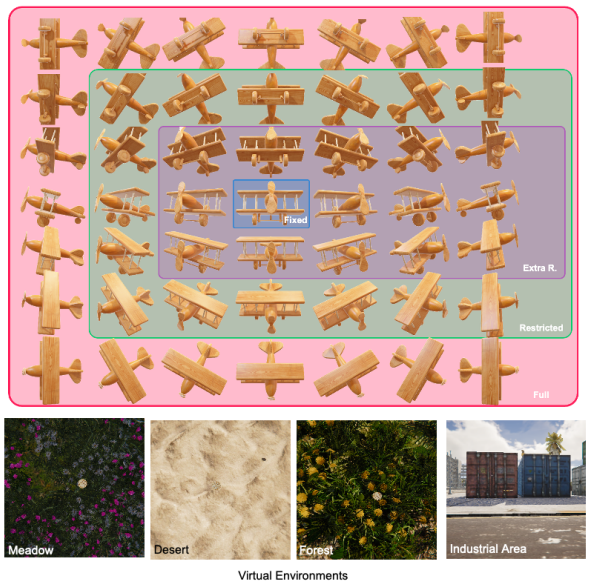

Figure 1: We curated four equally-sized image datasets varying in viewpoint diversity. Four virtual environments were used as image background.

Abstract

Although convolutional neural networks (CNNs) reach human-level accuracy on standard object recognition tasks, they perform poorly when faced with changes in viewpoint or corrupted images. In this study, we demonstrate that these two distinct failure modes can be addressed using a single strategy: training on diverse viewpoints. To investigate this, we created artificial image datasets that systematically vary in viewpoint diversity while keeping the dataset size constant, to train and evaluate CNN object recognition performance. Our results reveal a core trade-off between learning speed and generalization performance. On the one hand, models trained on restricted viewpoints exhibit fast learning and achieve near-perfect *in-distribution* accuracy, but they overfit to specific views, resulting in dramatic performance drops on unfamiliar viewpoints. On the other hand, training with diverse viewpoints slows learning but significantly improves *out-of-distribution* performance. Notably, exposure to diverse viewpoints also greatly enhances robustness to common image corruptions. These results point to a shared mechanism for achieving robustness to both viewpoint variation and image corruption, and further alignment with human performance.

Keywords: Convolutional Neural Networks; Viewpoint Diversity; Out-of-Distribution Generalization

Introduction

Convolutional neural networks (CNNs) are among the top performing models for object recognition, and are widely used for studying visual recognition and human-model alignment (Kriegeskorte, 2015). While recognizing objects from novel, uncommon viewpoints or with perturbation to the images are tractable problems to humans, they pose a notable challenge to CNNs (Geirhos et al., 2018; Madan et al., 2021; Sakai et al., 2022). One explanation for the CNN generalization limitation could be a lack of viewpoint diversity in the training sets. Several large computer vision datasets such as Caltech 101 (Fei-Fei, Fergus, & Perona, 2004) and ImageNet (Deng et al., 2009) do not feature dedicated variation of viewpoints, but are instead designed to capture a large variety of object categories. In comparison, exposure to multiple viewpoints plays a critical role in human and non-human animal object recognition development (Ayzenberg & Behrmann, 2024; Kraebel & Gerhardstein, 2006; Milivojevic, 2012; Okamura, Yamaguchi, Honda, Wang, & Tanaka, 2014; Tarr, 1995; Hayward & Williams, 2000; Wood & Wood, 2016). While CNNs learn from restricted viewpoint sets, humans learn from a dynamic environment where they can actively interact with objects and see them from multiple viewpoints (Meijer & Van der Lubbe, 2011; Sasaoka, Asakura, & Kawahara, 2010). To better understand the influence of viewpoint diversity on CNN object recognition behavior, we leverage the high-maneuverability of 3D objects to curate image datasets that systematically vary in viewpoint diversity. We evaluate CNN object recognition performance as a function of viewpoint diversity. Furthermore, we test CNNs on corrupted images to investigate how viewpoint diversity affects model robustness to perturbations.

Methods

We created four datasets with varying viewpoint diversity by progressively restricting the camera placement range (Figure 1) which is defined by the polar (θ) and azimuth angle (ϕ) of the camera origin. The viewpoint diversity levels are fixed ($\theta = \pi, \phi = 0.5\pi$), extra restricted ($\theta \sim U(1.4\pi, 2\pi)$, $\phi \sim U(0.4\pi, 0.8\pi)$), restricted ($\theta \sim U(0.8\pi, 2\pi)$, $\phi \sim U(0.2\pi, 0.8\pi)$), and full ($\theta \sim U(0, 2\pi)$, $\phi \sim U(0, \pi)$). We used 3D object from Objaverse 1.0 dataset (Deitke et al., 2022), selecting 1544 instances across 32 diverse categories. Objects were imported into Unity (Unity Technologies, 2023) and rendered in four virtual scenes, with each object rendered equally across scenes. Cameras were placed 1 unit from the object, offset slightly (0–0.1 units) from the center, and 30 images were rendered per object per scene at 256×256 resolution. Each training dataset contained 185K images. In this study, in-distribution (ID) and out-of-distribution (OOD) are defined solely in terms of viewpoint distribution. Therefore, we generated ID test sets (61K each) using the same pipeline. An additional 61K-image test set sampled from unseen viewpoints served as an OOD set for all but the full-view model, which by design has no true OOD counterpart. We trained four ResNet18 instances (He, Zhang, Ren, & Sun, 2015) from scratch, each on one of the four viewpoint diversity datasets for 30 epochs. To test the robustness to image corruption, we applied 19 kinds of image corruption from Hendrycks and Di-

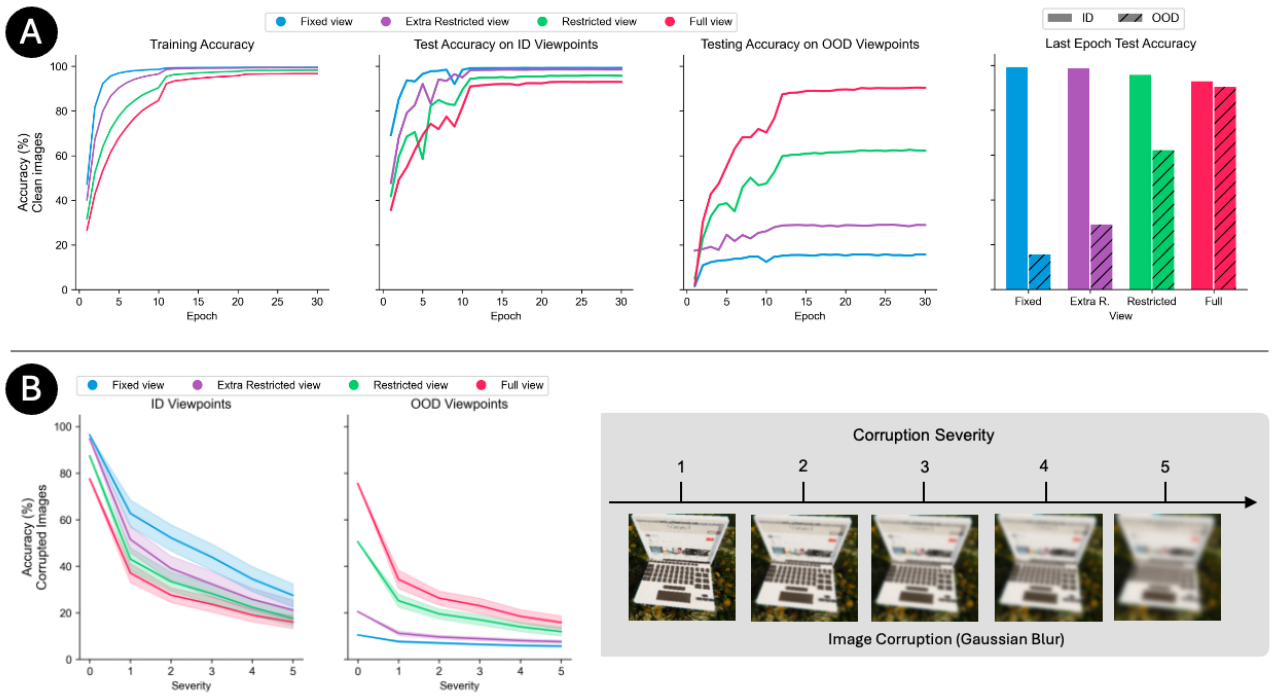

Figure 2: **a)** Accuracy of ResNet18 evaluated on training set, ID validation set, OOD viewpoint dataset. **b)** Accuracy of ResNet18 evaluated on ID and OOD viewpoint datasets with corrupted images of increasing severity. Line and bar colors indicate the viewpoint dataset (fixed, extra restricted, restricted, and full) used for training. Gaussian blur is shown as one exemplar of 19 kinds of image corruption.

etterich (2019) to test set images. We randomly sampled 640 pictures from each of the five test datasets balancing the object categories and backgrounds. Finally, we evaluated each model on the corresponding ID (corrupted) dataset and OOD (corrupted) dataset.

Results

We observed a trade-off between rapid learning and generalization (Figure 2a): as viewpoint diversity increased, training and ID accuracy declined slightly (e.g., training accuracy: fixed 99.6% $\rightarrow$ full 96.7%; ID accuracy: fixed 99.3% $\rightarrow$ full 93.0%). However, viewpoint-restricted models failed under OOD viewpoints (e.g., fixed: 15.6%, extra-restricted: 28.9%), whereas viewpoint-diverse models maintained high OOD performance (restricted: 62.2%, full: 90.3%).

Our image corruption experiment (Figure 2b) shows that with mild severity (level 1) condition, fixed-view models performed well on ID data (62.7%) but performed poorly on OOD views (7.6%). In contrast, full-view models showed better performance, achieving lower ID accuracy under corruption (37.1%) but sustaining strong OOD performance (34.4%). These differences between viewpoint-diverse and viewpoint-restricted models persisted under more severe image corruption as well.

Discussion

Our data reveals how viewpoint diversity influences CNN object recognition performance. Viewpoint-restricted models memorized specific features fast, while learning speed of viewpoint-diverse models was more gradually, however leading to better generalization performance. Image corruption experiments shows the viewpoint-diverse models staying robust to corruptions on OOD images. These findings point toward a pathway for improving human-CNN alignment in object recognition. By emulating the way humans interact with real-world objects, CNNs can be trained to perform in a more robust and generalizable manner. This improved generalizability and robustness makes them better suited for real-world tasks, where variability in viewpoint is common.

References

Ayzenberg, V., & Behrmann, M. (2024). Development of visual object recognition [Journal Article]. *Nature Reviews Psychology*, *3*(2), 73-90. doi: 10.1038/s44159-023-00266-w

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., ... Farhadi, A. (2022, December 01, 2022). *Objaverse: A universe of annotated 3d objects* [Electronic Article]. doi: 10.48550/arXiv.2212.08051

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Fei-Fei, L., Fergus, R., & Perona, P. (2004). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop* (p. 178-178). doi: 10.1109/CVPR.2004.383

Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, *31*.

Hayward, W. G., & Williams, P. (2000). Viewpoint dependence and object discriminability. *Psychological Science*, *11*(1), 7–12.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.*

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International conference on learning representations.*

Kraebel, K. S., & Gerhardstein, P. C. (2006). Three-month-old infants' object recognition across changes in viewpoint using an operant learning procedure. *Infant Behavior and Development*, *29*(1), 11-23. doi: https://doi.org/10.1016/j.infbeh.2005.10.002

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing [Journal Article]. *Annual Review of Vision Science*, *1*(Volume 1, 2015), 417-446. doi: https://doi.org/10.1146/annurev-vision-082114-035447

Madan, S., Henry, T., Dozier, J., Ho, H., Bhandari, N., Sasaki, T., ... Boix, X. (2021). *When and how cnns generalize to out-of-distribution category-viewpoint combinations.*

Meijer, F., & Van der Lubbe, R. H. (2011). Active exploration improves perceptual sensitivity for virtual 3d objects in visual recognition tasks. *Vision Research*, *51*(23), 2431-2439. doi: https://doi.org/10.1016/j.visres.2011.09.013

Milivojevic, B. (2012). Object recognition can be viewpoint dependent or invariant – it's just a matter of time and task. *Frontiers in Computational Neuroscience*, *6*. doi: 10.3389/fncom.2012.00027

Okamura, J.-y., Yamaguchi, R., Honda, K., Wang, G., & Tanaka, K. (2014). Neural substrates of view-invariant object recognition developed without experiencing rotations of the objects. *Journal of Neuroscience*, *34*(45), 15047–15059. doi: 10.1523/JNEUROSCI.1898-14.2014

Sakai, A., Sunagawa, T., Madan, S., Suzuki, K., Katoh, T., Kobashi, H., ... Sasaki, T. (2022). *Three approaches to facilitate dnn generalization to objects in out-of-distribution orientations and illuminations.*

Sasaoka, T., Asakura, N., & Kawahara, T. (2010). Effect of active exploration of 3-d object views on the view-matching process in object recognition. *Perception*, *39*(3), 289-308. (PMID: 20465167) doi: 10.1068/p5721

Tarr, M. J. (1995). Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects [Journal Article]. *Psychonomic Bulletin Review*, *2*(1), 55-82. doi: 10.3758/BF03214412

Unity Technologies. (2023). *Unity.* (Game development platform)

Wood, J. N., & Wood, S. M. W. (2016). The development of newborn object recognition in fast and slow visual worlds. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1829), 20160166. doi: 10.1098/rspb.2016.0166