Characterizing the Performance of Deep Neural Networks for Eye-Tracking

Arnab Biswas
arnab@nevada.unr.edu
University of Nevada, Reno
Reno, Nevada, USA

Kamran Binaee
kbinaee@unr.edu
University of Nevada, Reno
Reno, Nevada, USA

Kaylie Capurro
capurro.kaylie@gmail.com
University of Nevada, Reno
Reno, Nevada, USA

Mark D. Lescroart
mlescroart@unr.edu
University of Nevada, Reno
Reno, Nevada, USA

ABSTRACT

Deep neural networks (DNNs) provide powerful tools to identify and track features of interest, and have recently come into use for eye-tracking. Here, we test the ability of a DNN to predict keypoints localizing the eyelid and pupil under the types of challenging image variability that occur in mobile eye-tracking. We simulate varying degrees of perturbation for five common sources of image variation in mobile eye-tracking: rotations, blur, exposure, reflection, and compression artifacts. To compare the relative performance decrease across domains in a common space of image variation, we used features derived from a DNN (ResNet50) to compute the distance of each perturbed video from the videos used to train our DNN. We found that increasing cosine distance from the training distribution was associated with monotonic decreases in model performance in all domains. These results suggest ways to optimize the selection of diverse images for model training.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Video segmentation**.

KEYWORDS

deep eye tracking, pupil tracking, eyelid tracking

ACM Reference Format:
Arnab Biswas, Kamran Binaee, Kaylie Capurro, and Mark D. Lescroart. 2021. Characterizing the Performance of Deep Neural Networks for Eye-Tracking. In *2021 Symposium on Eye Tracking Research and Applications (ETRA '21 Adjunct), May 25–27, 2021, Virtual Event, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3450341.3458491

1 INTRODUCTION

Video-based eye-tracking methods have evolved from classical methods which tracked the corneal reflection [Carmichael and Dearborn 1947; Cornsweet 1958] to recent methods using deep neural networks (DNNs) that use the full anatomical shape of the eye for gaze estimation [Chaudhary et al. 2019; Kothari et al. 2020a]. DNNs have been successful in dealing with noisy data, but the robustness of DNN solutions to pupil estimation under different sources of image perturbation encountered during eye-tracking have not yet been well characterized.

Part of the reason for this is that many commonly used data sets have been collected under constrained circumstances, e.g. in a few types of indoor environments, in order to assure high-quality data. Such data is useful for testing multiple models on the same videos but cannot help to determine the robustness of DNN algorithms under more challenging circumstances such as outdoor freely moving mobile eye-tracking. Generalization is a well-known problem for DNNs. A DNN trained on a certain dataset may not perform well on new held-out data. The ability of a DNN to generalize may depend upon the choice of the DNN architecture or the data used to train the DNN. Selecting training data is a notoriously difficult problem. It is very difficult to sample the long tails of data in the wild [Salakhutdinov et al. 2011; Van Horn and Perona 2017]. Among all possible rare events to sample, it is difficult to know a priori whether the inclusion of particular data will improve model performance.

A crucial step in making DNNs increasingly robust is to characterize the sources of image variation that cause them to fail. Understanding failure cases can lead to insights about how to improve model performance, e.g., by strategically expanding the training set. Here, we study the degree to which different image perturbations cause a deep neural network (in our case DeepLabCut [Mathis et al. 2018]) to fail. We further define a method to characterize the performance of our DNN model on these image perturbations based on their distance from the training distribution. This opens up the possibility of predicting model performance on new data points.

2 METHODS

For this study, we used eye videos from 13 participants from the Gaze in Wild (GiW) dataset [Kothari et al. 2020b] that used a head-mounted pupil labs eye tracker [Kassner et al. 2014] to track the gaze position of participants while they performed different tasks indoors.

ETRA '21 Adjunct, May 25–27, 2021, Virtual Event, Germany

ACM ISBN 978-1-4503-8357-8/21/05…$15.00
https://doi.org/10.1145/3450341.3458491

2.1 Adding perturbations to eye videos

One of our goals was to investigate potential failure cases for our eye-tracking neural network model under the types of challenging conditions encountered outdoors and in datasets with many participants. For example, variable lighting conditions, individual differences in eye anatomy, and hardware and software choices all affect eye tracking data.

We evaluated our DNN model on eye videos with increasing exposure, reflection, defocus blur, eye rotation, and JPEG compression. These perturbations were added artificially using image processing to eye videos from two participants from the GiW dataset. We selected these two participants (one train and one test participant) because they had the smallest difference in DNN model performance between them. This minimizes the effects of individual differences on model performance across the test and train participant. For each participant, we created four videos with uniformly increasing levels of perturbation (Figure 1A) and then tested the performance of our neural network model on these videos.

2.1.1 Exposure. To simulate the effect of an increase in exposure and decrease in the contrast between the pupil/eyelashes and other regions in the eye video in bright sunlight, we added four steps of luminance increments (each 35 units) to all pixels in each frame. After each increment, pixel values were clipped to a maximum of 255.

2.1.2 Rotation. To simulate the effect of different camera angles and facial anatomy across participants, we rotated the eye videos in four five-degree increments followed by scaling and cropping to ensure uniform frame size. This rotation resulted in the eye going partially out of the frame for the 15 and 20-degree rotation conditions.

2.1.3 Reflection. Corneal reflection and shadows on the eye present a challenge while recording eye videos outdoors. We used the method presented in [Eivazi et al. 2019] to add reflections and shadows to the eye images. We modified the blending factor for images superimposed on the eye video in four steps. For every frame, we randomly selected the reflected image from the Driving Events Camera Dataset [Rebecq et al. 5555] which contains videos from dashboard cameras of cars driving through highways and cityscapes.

2.1.4 JPEG artifacts. Compressed video formats are desirable when storing eye videos as they take up less space. Thus, we tested the robustness of our DNN to compression artifacts by altering the video frames with JPEG compression. We varied the JPEG quality parameter (which varies from 100 to 0, denoting best to worst quality) from 32 to 8 in four steps of 8.

2.1.5 Defocus Blur. Finally, to mimic the defocus blur from a camera we used the *imgaug* image augmentation library [Jung et al. 2020] and iteratively increased the severity parameter from 1 to 4 to create an incremental loss of focus in the eye videos.

2.2 DeepLabCut for pupil and eyelid detection

For detecting the pupil position and the eyelid shape, we used DeepLabCut (DLC) [Mathis et al. 2018; Nath* et al. 2019], a markerless pose estimation library. DLC uses a convolutional neural network based on ResNet50 architecture [He et al. 2016], which has been pre-trained on the ImageNet dataset [Deng et al. 2009]. DLC uses k-means in pixel space for each eye video to select 40 distinct frames from each participant for hand labeling. Our DLC model was trained using 240 frames (40 frames from each of 6 participants) which were hand-labeled with the 48 key points: 32 localizing the eyelids, and 16 encircling the pupil. The trained DLC model uses eye video frames as its input and outputs the coordinates of the 48 keypoint locations (Figure 1A) along with the likelihood for each of these key points. The likelihood, which varies from 0 to 1, is a metric of how confident the model is about its prediction for the given keypoint and is generally correlated with annotation error.

2.3 Distance from training distribution

Neural Net feature spaces have become a popular choice for feature extraction. For example, the Alexnet feature space has been used successfully as a space for representing and classifying images [Kiros et al. 2014; Krizhevsky et al. 2012; Ponce et al. 2019; Venugopalan et al. 2015]. Since DLC is based on the ResNet50 architecture, we used the ResNet50 feature space. To map an eye video frame into ResNet50 space, we ran it through a ResNet50 neural network pre-trained on ImageNet and looked at the output of the final convolutional layer before the fully connected layers. This gives us a 100,352-dimensional feature vector for each frame, which is considered the representation of the frame in ResNet50 space. To calculate the distance of new eye video frames from the training distribution, first, we mapped all the 240 training images into the ResNet50 space, took the mean of these 240 vectors, and then calculated the cosine distance of the new frames to this mean training distribution vector.

3 RESULTS AND DISCUSSION

To validate that DLC captures pupils accurately, we estimated the pixel error between the DLC model's predicted keypoints and human-labeled keypoints in 160 test frames from 4 participants that were not part of the training set. The root mean squared error (RMSE) between DLC and the human labels was 10.78 pixels. For context, the RMSE for the same frames between two human labelers was 12.91 pixels, suggesting that DLCs accuracy lies within the variance of human labelers.

We evaluated DLC performance on each perturbed video by measuring the change in likelihood with an increase in the perturbation intensity for each domain (Figure 1B). As expected, keypoint annotation confidence drops with an increase in perturbation intensity. We attempted to scale the amount of perturbation in each domain to a range likely to be seen in real data. However, it would be useful to have a single metric for "image change" to investigate the relative drop in model performance between perturbations.

To this end, we investigated the use of ResNet50 feature space as a representation of image similarity to the training distribution to explain model performance across perturbations.Figure 1C shows the DLC likelihood as a function of cosine distance from the training distribution across perturbations in ResNet50 feature space. We calculated the correlation and the rate of change (*ROC*), the average slope, between the likelihood and the cosine distance. The correlation tells us how good a metric the distance from the

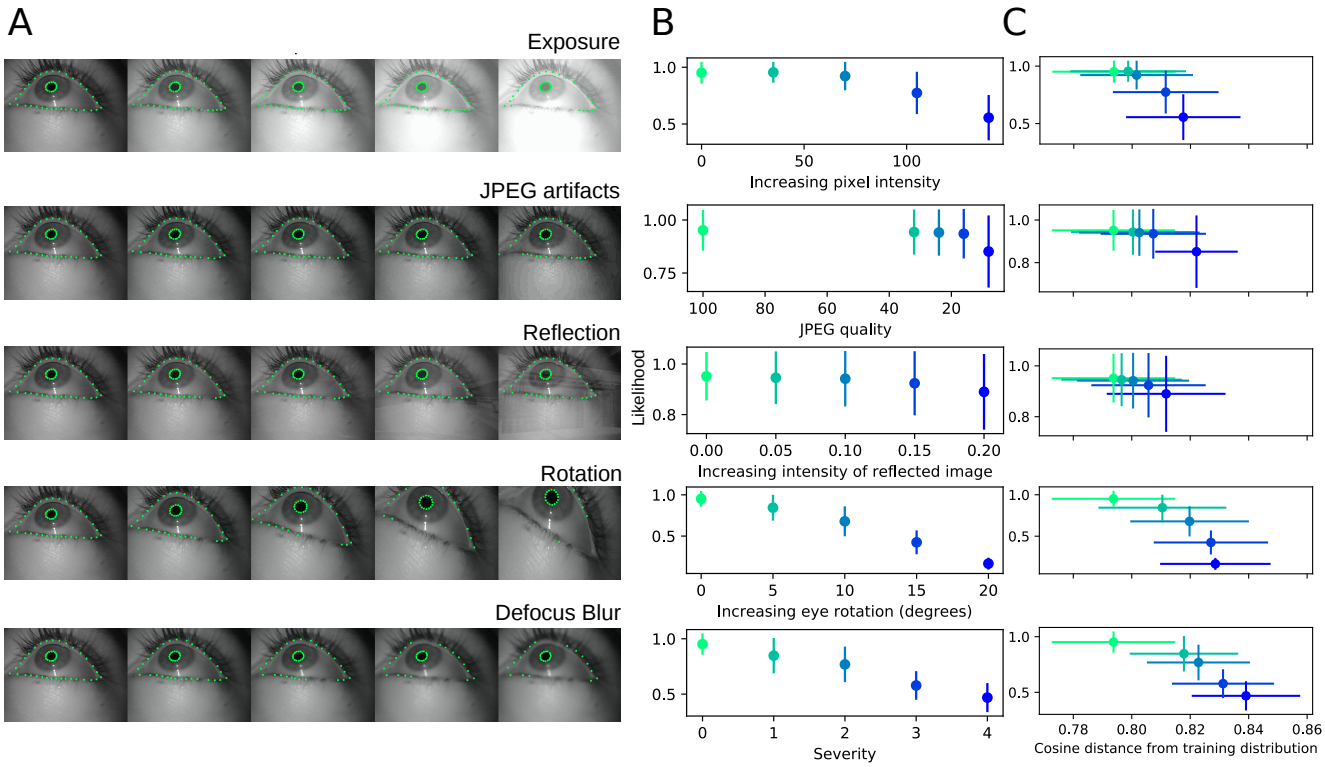

Figure 1: Frames illustrating perturbed eye videos and the keypoint annotation performance of our neural network model. Note the increasing number of missing keypoints as the intensity of the perturbations increases from left to right. (B) Mean model likelihood as a function of increasing perturbation intensity for each domain. (C) Likelihood as a function of cosine distance of the perturbed video from the training distribution mean in ResNet50 feature space. The likelihood drops monotonically with an increase in distance from the mean of the training distribution.

training distribution is in explaining model performance. The slope gives us the drop in likelihood per unit distance: the more negative the slope the more detrimental is the effect of the perturbation on model performance. Across all domains, the likelihood drops monotonically with an increase in distance from the training distribution ($r = -0.4$). We see that even a small increase in rotation($r = -0.37, ROC = -22.5$), exposure($r = -0.53, ROC = -16.6$) or blur($r = -0.22, ROC = -10.65$) results in a large increase in the distance from the training distribution. This suggests that our model may not be robust to these perturbations as evident from the larger drop in likelihood with an increase in distance. On the other hand, we see that the videos with even the strongest perturbation intensity for compression($r = -0.16, ROC = -3.5$) and reflection($r = -0.03, ROC = -3.4$) domains are relatively nearer to the training distribution. Our model is correspondingly robust to these perturbations.

Ideally, the drop in likelihood with distance should be predictable using a single function for all perturbations. The differences between the slopes and correlations in Figure 1C suggest that we have not yet found an ideal generic image space for such a function. However, these correlations and slopes can be used as a reasonable approximation of model performance decline. Our approach is a step in the direction to find a feature space which explains model performance as a function of image variability from the training distribution. Also, image spaces are very high dimensional. Thus, a single metric (e.g. distance from the training distribution) which is linearly correlated with a decrease in model likelihood may not exist.

In addition to the ResNet50 feature space we evaluated Euclidean and cosine distance in pixel space to calculate distance from the training distribution. The performance drop due to blur ($r = 0.08$), compression ($r = 0.15$), and reflection ($r = 0.04$) was not reliably related to an increase in distance from the training distribution in pixel space. This suggests that using k-means in ResNet50 feature space (instead of pixel space) to uniformly sample training data may be more effective in maximizing the variance of the dataset. This would reduce any redundancy in the training dataset by including only those frames which add new information based on the distance metric. This is important when a DNN is trained on a limited number of frames and labeling new frames is costly.

Our work also suggests productive directions for training data augmentation: addition of variation in exposure, rotation, and defocus blur to the training data seems more likely to improve model performance than addition of reflection and JPEG artifacts.The current study guides the training regime that one would use in order to retrain a single model to generalize across conditions.

Another possible use of a metric for distance from the training set is as a data quality filter. Currently, the correlation between distance from the training set and model likelihood is too low to be used as a frame-by-frame quality metric. However, computing distance from a training set could be useful to determine whether a whole new data session is likely to be problematic.

Each neural network is different and is trained on a different dataset, thus each of them is differently susceptible to these perturbations. We plan to further investigate if these methods are applicable for other neural networks. We are also looking into Mahalanobis distance and mutual information as alternate distance metrics between the training dataset and image perturbations.

ACKNOWLEDGMENTS

This research was supported by NSF EPSCoR #1920896 to MDL

REFERENCES

Leonard Carmichael and Walter F Dearborn. 1947. *Reading and visual fatigue.* Houghton Mifflin.

A. K. Chaudhary, R. Kothari, M. Acharya, S. Dangi, N. Nair, R. Bailey, C. Kanan, G. Diaz, and J. B. Pelz. 2019. RITnet: Real-time Semantic Segmentation of the Eye for Gaze Tracking. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 3698–3702. https://doi.org/10.1109/ICCVW.2019.00568

Tom N Cornsweet. 1958. New technique for the measurement of small eye movements. *JOSA* 48, 11 (1958), 808–811.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Shaharam Eivazi, Thiago Santini, Alireza Keshavarzi, Thomas Kübler, and Andrea Mazzei. 2019. Improving Real-Time CNN-Based Pupil Detection through Domain-Specific Data Augmentation. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 40, 6 pages. https://doi.org/10.1145/3314111.3319914

K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. 2020. imgaug. https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-Based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (Seattle, Washington) *(UbiComp '14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1151–1160. https://doi.org/10.1145/2638728.2641695

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *CoRR* abs/1411.2539 (2014). arXiv:1411.2539 http://arxiv.org/abs/1411.2539

Rakshit Kothari, Zhizhuo Yang, Christopher Kanan, Reynold Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020b. Gaze-in-wild: A dataset for studying eye and head coordination in everyday activities. *Scientific reports* 10, 1 (2020), 1–18.

Rakshit S Kothari, Aayush K Chaudhary, Reynold J Bailey, Jeff B Pelz, and Gabriel J Diaz. 2020a. EllSeg: An Ellipse Segmentation Framework for Robust Gaze Tracking. *arXiv preprint arXiv:2007.09600* (2020).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

Alexander Mathis, Pranav Mamidanna, Kevin M Cury, Taiga Abe, Venkatesh N Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. 2018. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience* 21, 9 (2018), 1281–1289.

Tanmay Nath*, Alexander Mathis*, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie W Mathis. 2019. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols* 14, 7 (2019), 2152–2176. https://doi.org/10.1038/s41596-019-0176-0

Carlos R Ponce, Will Xiao, Peter F Schade, Till S Hartmann, Gabriel Kreiman, and Margaret S Livingstone. 2019. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences. *Cell* 177, 4 (2019), 999–1009.

H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza. 5555. High Speed and High Dynamic Range Video with an Event Camera. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 01 (dec 5555), 1–1. https://doi.org/10.1109/TPAMI.2019.2963386

Ruslan Salakhutdinov, Antonio Torralba, and Josh Tenenbaum. 2011. Learning to share visual appearance for multiclass object detection. In *CVPR 2011*. IEEE, 1481–1488. https://doi.org/10.1109/CVPR.2011.5995720

Grant Van Horn and Pietro Perona. 2017. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450* (2017).

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. (May–June 2015), 1494–1504. https://doi.org/10.3115/v1/N15-1173