Application of Deep Neural Networks to Model Omnidirectional Gaze Behavior in Immersive VR

Thomas L. Botch[1], Erica L. Busch[1], Caroline E. Robertson[1]

[1]Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

robertsonlab.com

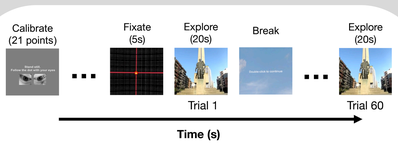
BACKGROUND

- Vision in naturalistic, 360° environments is an active process, influencing every aspect of gaze behavior [1].
- Convolutional neural networks (CNNs) are powerful computational models traditionally used to model visual images representing a single field-of-view [2]
- CNNs are good predictors of eye-movement patterns across images [3].
- Here, we illustrate a method of leveraging existing, pre-trained CNNs to provide information about 360° naturalistic scenes and demonstrate comparisons of CNN activation maps with 360° gaze behavior.

VR / Eye-tracking: Measuring Active Vision

- Wide visual field (100°)
- 360° real-world, complex scenes
- Active, first-person viewing

METHODS – VR FREE-VIEWING TASK



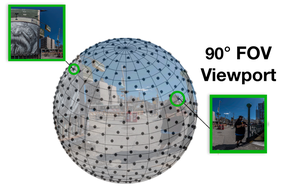
Gaze behavior in complex, real-world immersive scenes was measured in a head-mounted display (HMD). 12 adults observed 60 scenes (20s each).

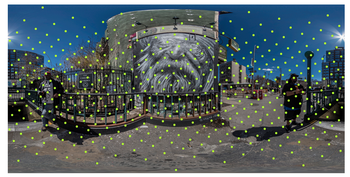
- HMD specs: Oculus DK2: resolution: 960 x 1080; field-of-view: ~100 DVA; 75 Hz refresh).
- Eyetracker: (Pupil Labs. 120 Hz refresh; 5.7ms latency; 0.5 visual degrees accuracy). 1.51° mean accuracy and 0.434° mean precision measured in n=11 ideal observers.

Gaze heat maps were created by weighting fixations by duration and smoothing with a gaussian filter for all scenes for each participant.

METHODS – 360° CNN PIPELINE

Sample Spherical Space

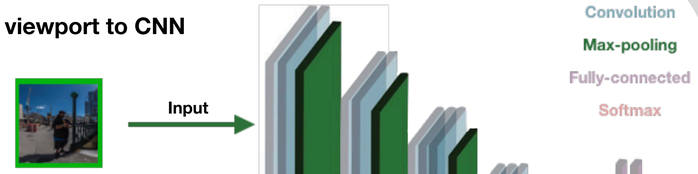




Points are uniformly sampled from each spherical image. The image is decomposed into square viewports centered on each point. The field-of-view (FOV) of each viewport is similar to the head-mounted display (90°).

Generate CNN Activation Maps

1. Input viewport to CNN



2. Extract layer activations

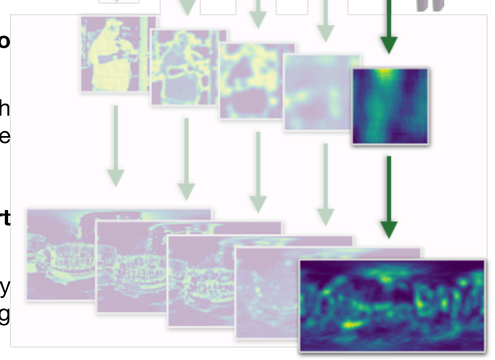
- Here, we extracted activations from the final pooling layer

3. Project activations into panoramic space

- Repeat for each viewport. Here we use 500 viewports.

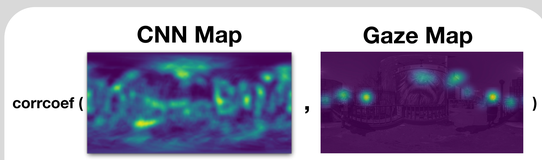
4. Aggregate viewport activation maps

- Average each pixel by the pixel's sampling density



Correlate Maps

Activation maps are normalized and histogram matched to individual gaze maps [4]. Points are sampled from each map and correlated.

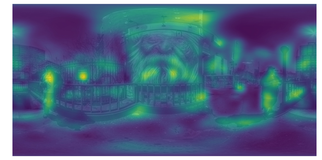


EXAMPLE APPLICATION

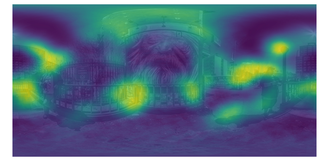
What guides attention during free-viewing of naturalistic, 360° scenes?

Here, we show the final layer activation maps from CNNs pretrained on different tasks: image memorability (AMNet) and scene recognition (VGG16) [5,6].

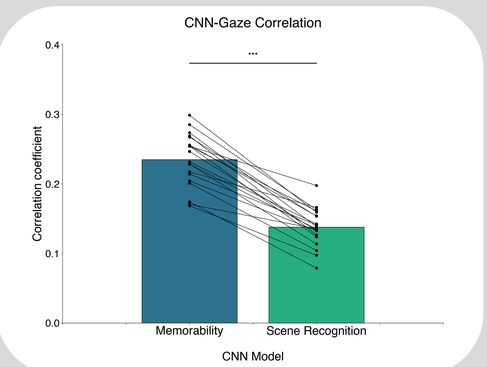
Memorability CNN



Scene Recognition CNN



RESULTS



Individual gaze behavior during 360° scene exploration is better predicted by CNN pretraining on memorability, compared to CNN pretraining on scene recognition ($t=9.34$, $p < 0.0001$).

CONCLUSIONS

- We demonstrate a method for modeling 360° real-world scenes using CNNs pretrained on different tasks.
- By comparing individual gaze behavior with CNN activation maps, we gain insight into how people actively process naturalistic, real-world environments.

REFERENCES

[1] Haskins, A.J., Mentch, J., Botch, T. L., Robertson, C. E. (2020) Active vision in immersive, 360 real-world environments. bioRxiv:10.1101/2020.03.05.976712.

[2] Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446.

[3] Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv:1610.01563 [Cs, q-Bio, Stat]*.

[4] Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10.

[5] Fajtl, J., Argyriou, V., Monekosso, D., & Remagnino, P. (2018). AMNet: Memorability Estimation with Attention. *ArXiv:1804.03115 [Cs]*.

[6] Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*.