Probing PredNet's Internal Representations using Feature Visualization

Aniekan Umoren[1,3], Nikolas McNeal[2,4], Dr. Tai-Sing Lee[3,4]

[1]Massachusetts Institute of Technology, [2]The Ohio State University, [3]Carnegie Mellon University,
[4]Center for Neural Basis of Cognition

CNBC
CENTER FOR THE NEURAL BASIS OF COGNITION

Carnegie Mellon University

Motivation

PredNet[1,2] is a generative neural network trained via self-supervised learning to perform next-frame prediction.

Self-supervised learning is an unsupervised framework which utilises unlabelled data to automatically generate supervision signals. PredNet uses the difference between its predictions and the ground-truth frames in its objective function.

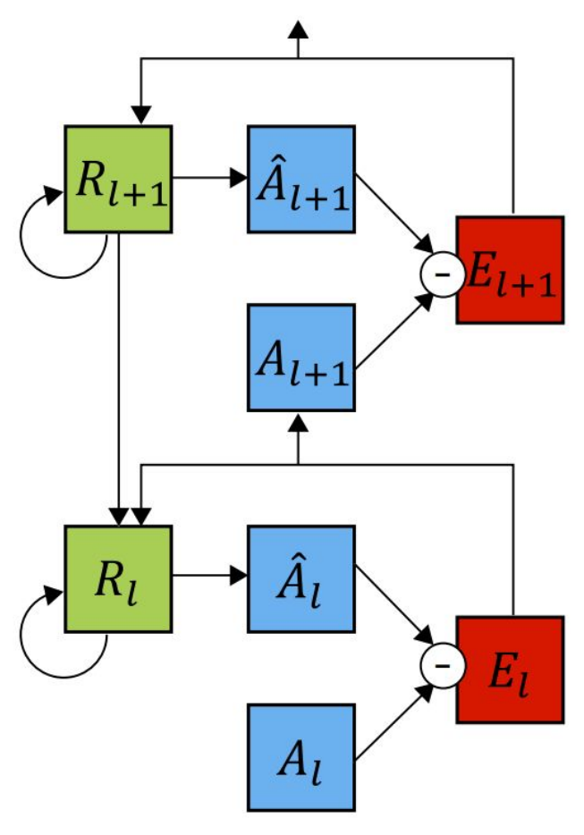
PredNet was inspired by the predictive coding principle; the brain creates a hierarchical internal model of the world to explain away sensory inputs. The network implements this idea by continuously generating predictions of future sensory input via a top-down path and sends prediction errors via its bottom-up path.

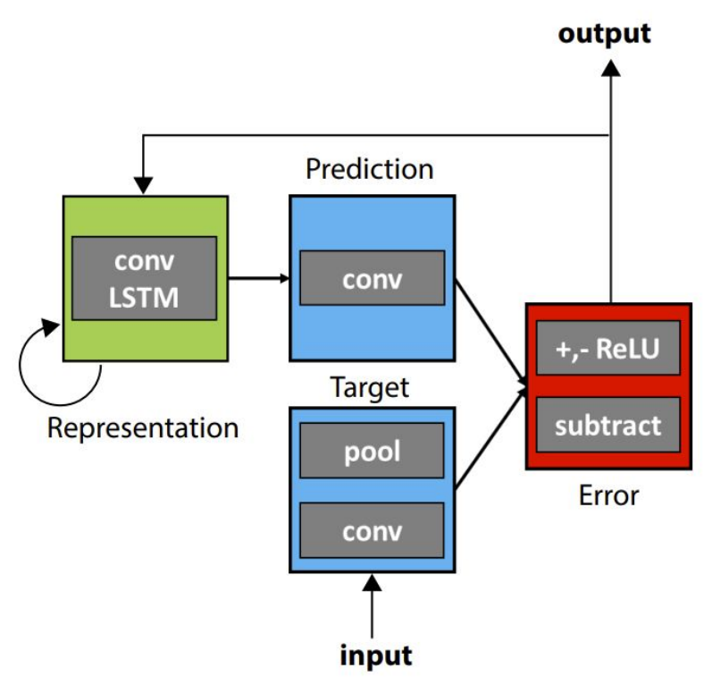
This principle allows the model to reproduce salient phenomena in the visual cortex such as illusory percepts and single-unit dynamics[2].

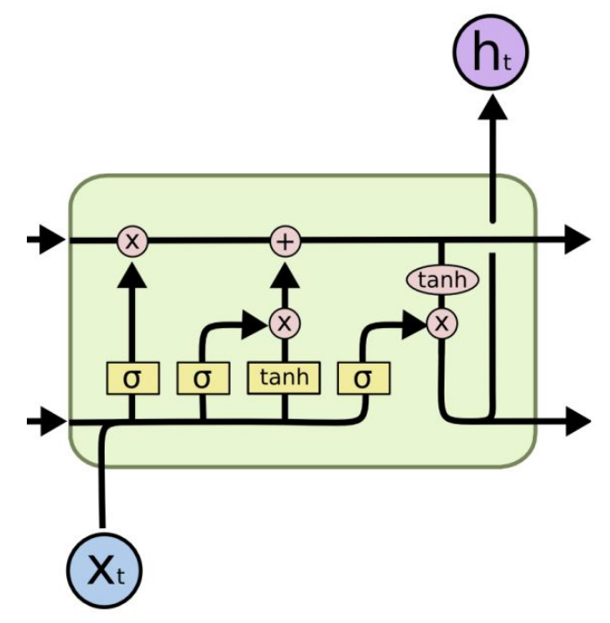
The representations of neurons in PredNet are not known.

In so far as PredNet is a useful model of the primate visual cortex, learning its internal representations can provide insights into the functioning of the visual system.

Visualizing the stimuli that elicit a strong or weak response gives us insight into PredNet's internal representations.







Methodology

PredNet was originally implemented in TensorFlow, so the first step was to reimplement it in PyTorch and reproduce its performance on the KITTI dataset.

After acquiring a trained model, the optimal stimulus for A and R units was visualized via the image optimization technique[3,4]:

1. Initialize a random image
2. Perform a forward pass and collect the activity of the unit(s) under investigation
3. Keeping weights fixed, backpropagate the loss to the image
4. Update image pixels. Smooth the image to constrain spatial smoothness
5. Perform 500 iterations

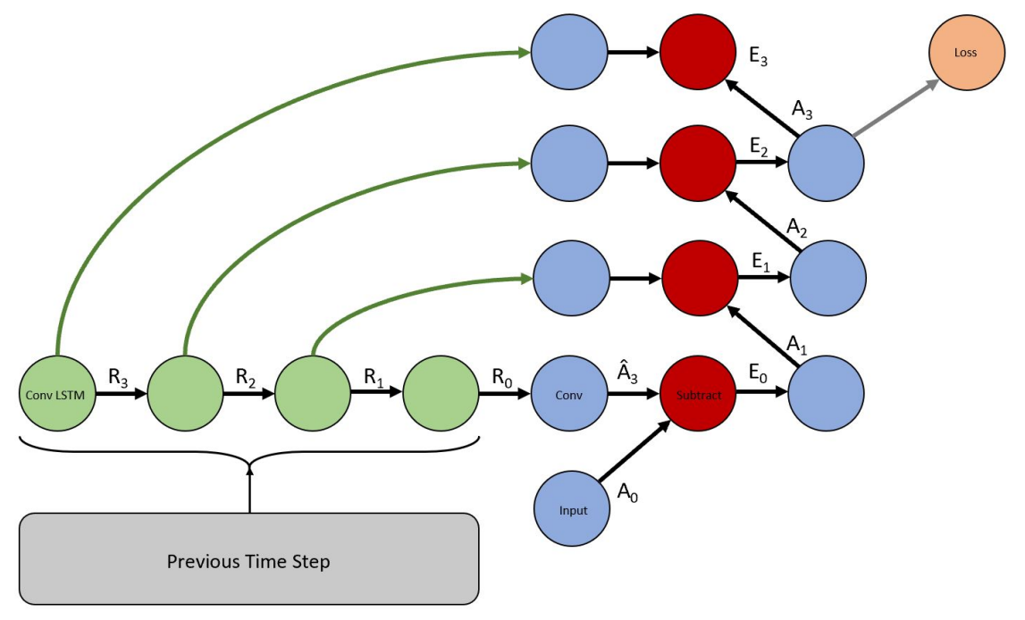


Figure 1. The gradient flows through layers and time allowing the activation to update the current frame and past frames.

The loss function depended on the type of response to be elicited:

- ***Strong sustained response (SSR)*** stimuli minimize the negative of neuron's time-averaged activity (ignoring the first frame)
- ***Weak sustained response (WSR)*** stimuli minimize the neurons time-averaged activity (ignoring the first frame).
- ***Strong impulse response (SIR)*** stimuli minimize the activity of the neuron's activity *after the last frame.*

Exemplary Visualizations

Figure 2. Sizes of the receptive fields for neurons at each layer
(Top) A1-A3 **(Middle)** R0-R3 **(Bottom)** A3 stimulus frames

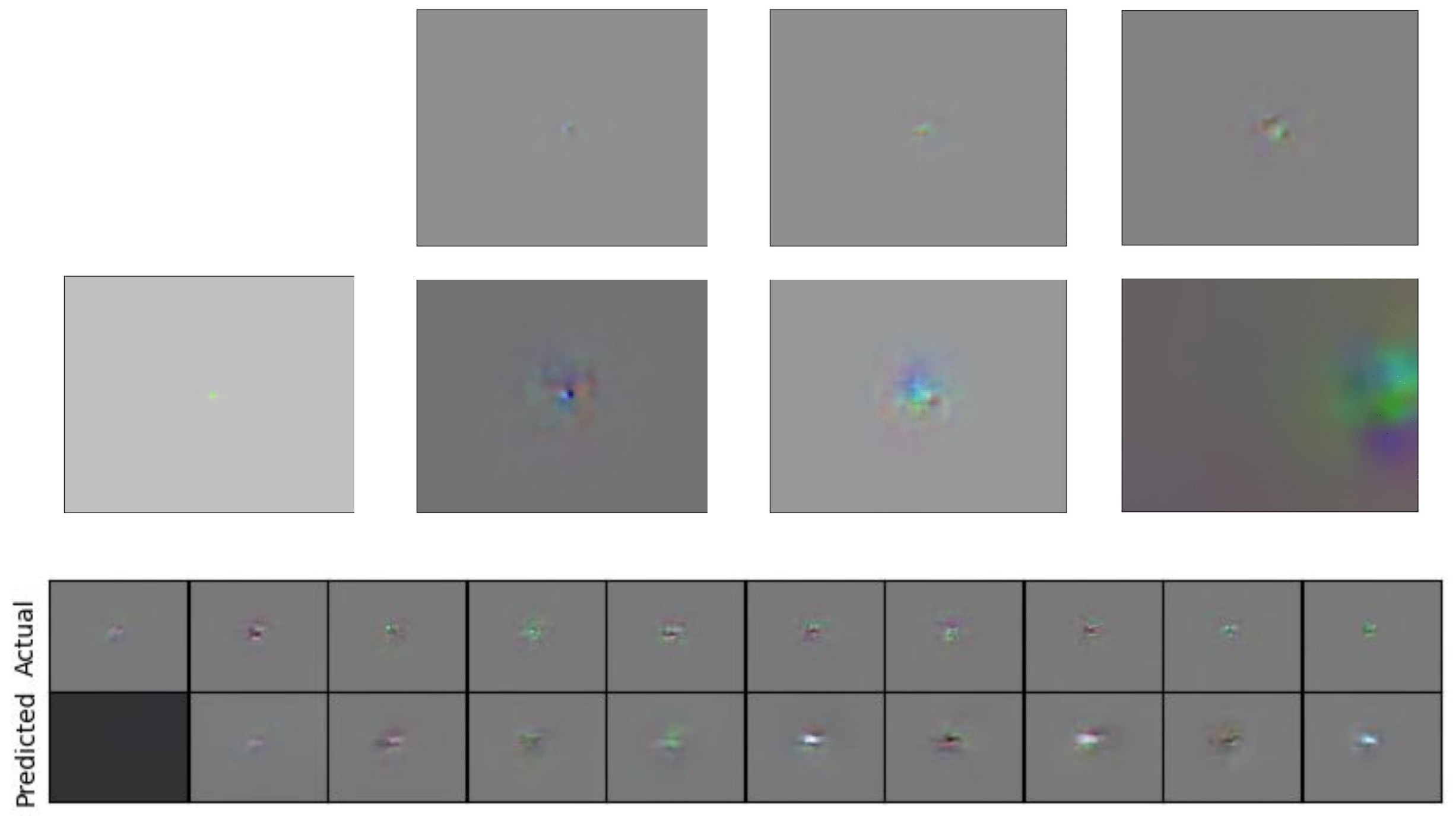


Figure 3a. Visualizing stimuli that elicit a strong/weak sustained response or a strong impulse response for example neurons in A3 and R3.
(Col 1) strong sustained response; **(Col 2)** weak sustained response; **(Col 3)** strong impulse response
(Row 1) A3; **(Row 2)** R3

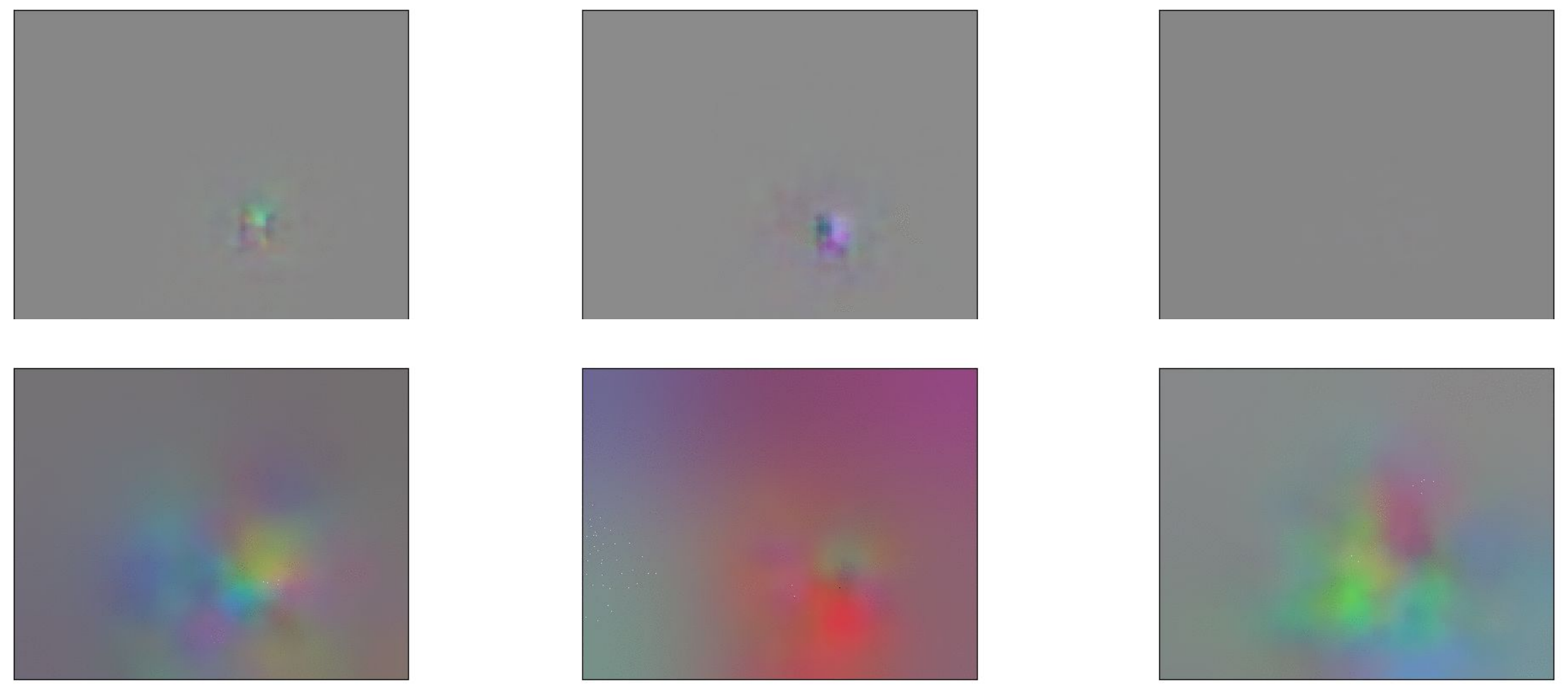


Figure 3b. Still frames for SSR and SIR stimuli
(Row 1) A3 SSR; **(Row 2)** R3 SSR; **(Row 3)** A3 SIR; **(Row 4)** R3 SIR;

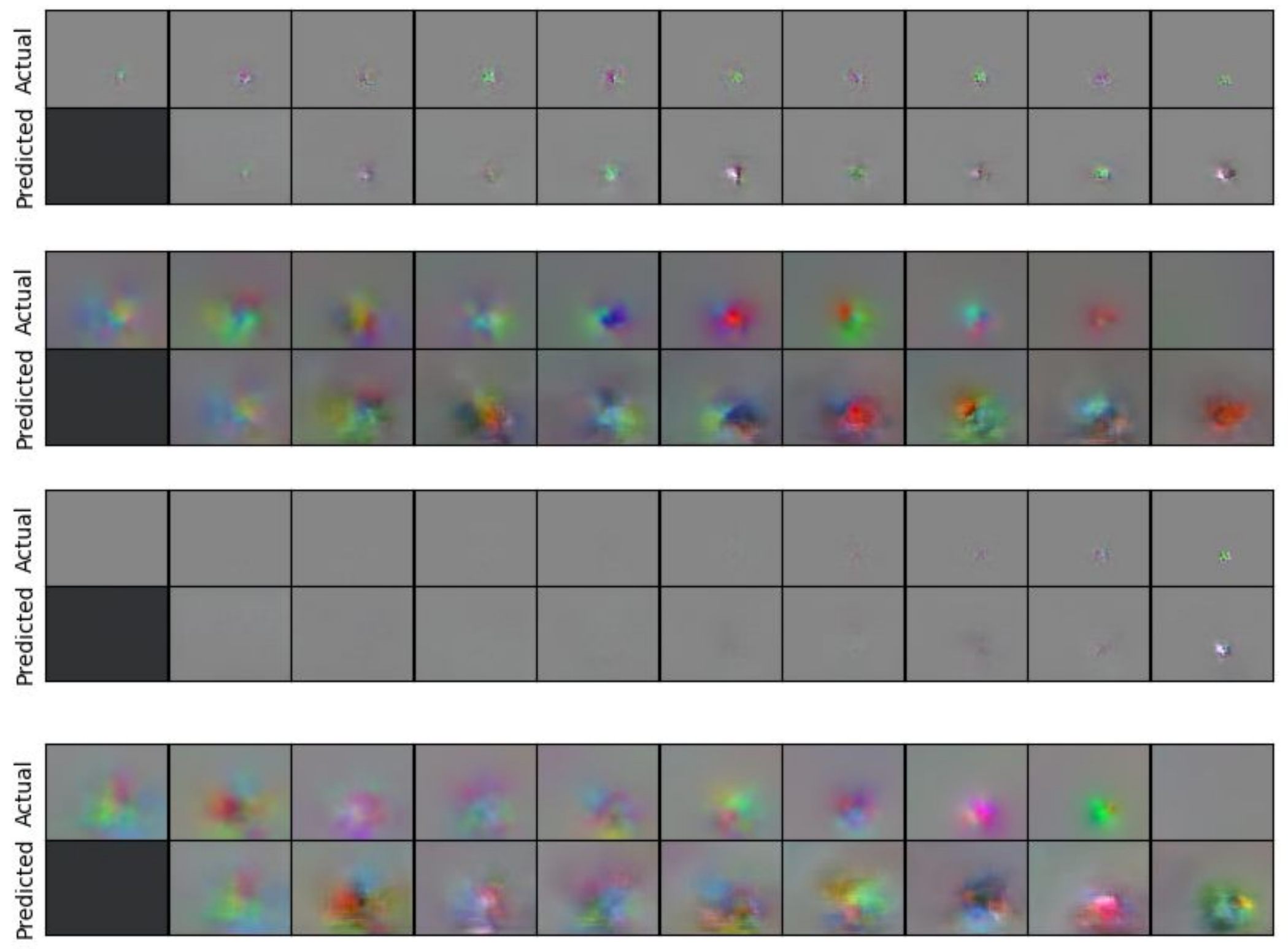
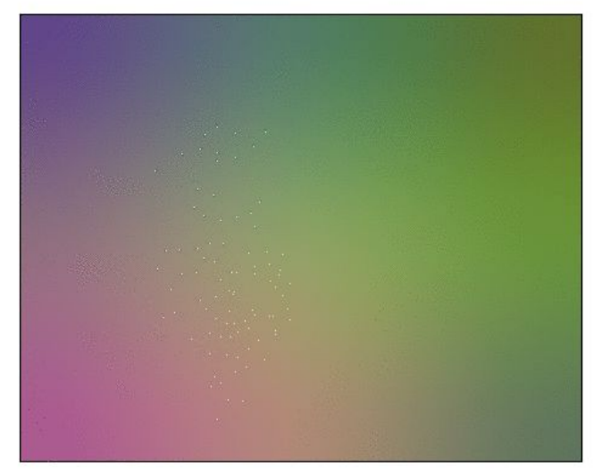


Figure 4. Optimized stimuli for R3 neuron which failed to elicit an SSR



Results

Inspection was focused on the highest layers units A3 and R3. The neurons this deepest layer have the largest receptive fields allowing for easier discernment of salient features

The following features were observed in Figures 2 and 3:

- The size of receptive fields shrinks over time. This is analogous to the shrinking of receptive fields as you move from higher to lower layers (*Fig. 2*). Convolution through time gives neurons a larger spatial context of past frames.
- SSR stimuli for A3 neurons display object motion while the WSR stimuli are stationary (*Fig. 3a*). This suggests that A3 neurons show sensitivity to object translation.
- R3 neurons do not display any coherent motion. However, many neurons do display an enhanced sensitivity to color (as compared to A3 neurons).
- The features in R3 SIR stimuli span further into the past than A3 SIR stimuli. This demonstrated the memory capabilities of the Convolutional LSTM units.
- There were R3 neurons for which feature visualization failed to elicit an SSR/ SIR. Failed R3 feature visualizations result in a full-image color display without a clear receptive field. (*Fig. 4*).

Summary and Discussion

A distinction between the bottom-up and top-down pathways emerges.

1. The bottom-up pathway shows sensitivity to motion while the top-down pathway has access to longer time-dependencies.
2. A more thorough survey is needed to cement these preliminary results and perhaps discover neurons that are sensitive to other types of motion such as radial expansion or rotation. Optical flow analysis can be used to quantify these results and detect various motion and form detectors.
3. The optimization may be underconstrained. Adding regularizers to enforce both spatial temporal smoothness may result in more coherent and natural stimuli.
4. Attempt hyperparameter tuning (learning rate and spatial/temporal smoothing) for R3 neurons to avoid failed visualizations.
5. The receptive fields of A3 and (to a lesser extent) R3 are still relatively local. Making the network deeper may allow neurons to be more sensitive to a global context.

References

[1] Lotter, William, Gabriel Kreiman, and David Cox (Mar. 2017). "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning". In: URL: https://arxiv. org/pdf/1605.08104.pdf.

[2] Lotter, W., Kreiman, G. & Cox, D. A neural network trained for prediction mimics diverse features of biological neurons and perception. Nat Mach Intell 2, 210–219 (2020). https://doi.org/10.1038/s42256-020-0170-9

[3] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. CoRR, abs/1312.6034.

[4] Olah, et al., "Feature Visualization", Distill, 2017.

[5] Aravindh Mahendran, & Andrea Vedaldi. (2016). Visualizing Deep Convolutional Neural Networks Using Natural Pre-Images.