Evidence for Configural Superiority Effects in Convolutional Neural Networks
Shaiyan Keshvari & Ruth Rosenholtz

INTRODUCTION
• Configural superiority effect (CSE) – combinations of parts are perceived more quickly and accurately than the parts alone\textsuperscript{1,2}
• CSEs thought to be driven by “emergent” feature (EF) differences between target and distractors\textsuperscript{1}
• EFs may result from the visual system learning abstract representations to support complex tasks, like object recognition, at the expense of simpler but less ecologically relevant tasks
• Convolutional Neural Nets (CNNs) excel at object recognition, as well as tasks for which they are not trained. Feature vectors at different layers correlate with responses of various brain areas\textsuperscript{3}

Research question:
• Do the higher levels in a CNN show CSEs?

Approach:
• Use the VGG-16 network\textsuperscript{4} pre-trained on ImageNet as a stand-in for the visual hierarchy
• Train a classifier to do an oddball localization task using layer activations as the input features

METHODS
• Base (no EF) and composite (EF) stimuli
• Noisy, translated, rotated, resized, and contrast-adjusted images to promote generalizability
• For each EF x (base, composite), trained a multi-class linear SVM on the last fully connected layer (fc7, 4096 “neurons”) to locate the “odd quad”
• Compute cross-validated performance
• Also tested a network with random weights

RESULTS
• Average cross-validated performance
• Composite better than base (CSE!): - orthogonality (Δ33 percentage points) and roundness (53 pp)
• No effect: - closure (< 1 pp)
• Base better than composite: - parallelism (23 pp) and 3D (21 pp)
• Random CNN weights: no effect for any EFs
• Pilot behavioral experiment (N=2) confirmed CSE (44 +/- 0.06 pp)

CONCLUSIONS
• Some evidence that later CNN features compute “emergent” features, but not consistently
• Need to test other layers and networks
• Other factors like “false pop-out”, may explain some effects not modeled by CNN


Special thanks to Yrvine Thelusma, Carl Vondrick, Shaoxiong Wang, and Ben Wolfe

--
Cross-validated SVM performance for all EFs
average human performance (N=2) vs average SVM performance
0.5
0.7
0.9
0.1
0.3
0.5
0.7
0.0
0.2
0.4
0.6
0.8
1

fraction of data left out during training

---

\( \text{noise}, \text{translations}, \text{etc.} = 28,797 \text{ stimuli in total} \)

---

\text{example input stimuli}