Learning long-range spatial dependencies with horizontal gated-recurrent units

Drew Linsley (drew_linsley@brown.edu)

Junkyung Kim (junkyung_kim@brown.edu)

Vijay Veerabadran (vijay_veerabadran@brown.edu)

Thomas Serre (thomas_serre@brown.edu)
Brown University Department of CLPS
Providence, RI 02912, USA

Abstract

Progress in deep learning has spawned great successes in many engineering applications. As a prime example, convolutional neural networks, a type of feedforward neural networks, are now approaching – and sometimes even surpassing – human accuracy on a variety of visual recognition tasks. Here, however, we show that these neural networks and their recent extensions struggle in recognition tasks where co-dependent visual features must be detected over long spatial ranges. We introduce the horizontal gated-recurrent unit (hGRU) to learn intrinsic horizontal connections – both within and across feature columns. We demonstrate that a single hGRU layer matches or outperforms all tested feedforward hierarchical baselines including state-of-the-art architectures which have orders of magnitude more free parameters. We further discuss the biological plausibility of the hGRU in comparison to anatomical data from the visual cortex as well as human behavioral data on a classic contour detection task.

Keywords: Biological vision; recurrent networks; horizontal connections; deep learning.

Consider the images in Fig. 1a: A sample image from the Berkeley Segmentation Data Set (BSDS500) is shown on top and the corresponding contour map produced by a state-of-the-art deep convolutional neural network (CNN) (Lee et al., 2015) underneath it. Although this task has long been considered challenging because of the need to integrate global contextual information with inherently ambiguous local edge information, CNNs now rival humans at detecting contours in natural scenes. Now, consider the image in Fig. 1b, which depicts a variant of a visual psychology task called "Pathfinder" (Houtkamp & Roelfsema, 2010). Reminiscent of the everyday task of reading a subway map to plan a commute (Fig. 1c), the goal of Pathfinder is to determine if two white circles in an image are connected by a path. Compared to natural images such as the one shown in Fig. 1a, these images are visually simple, and the task is indeed easy for human observers to solve (Houtkamp & Roelfsema, 2010). Nonetheless, as detailed below, we find that modern CNNs struggle to solve this task. Why is it that a CNN can accurately detect contours in a natural scene like Fig. 1a but also struggle to integrate paths in the stimuli shown in Fig. 1b? In principle, the

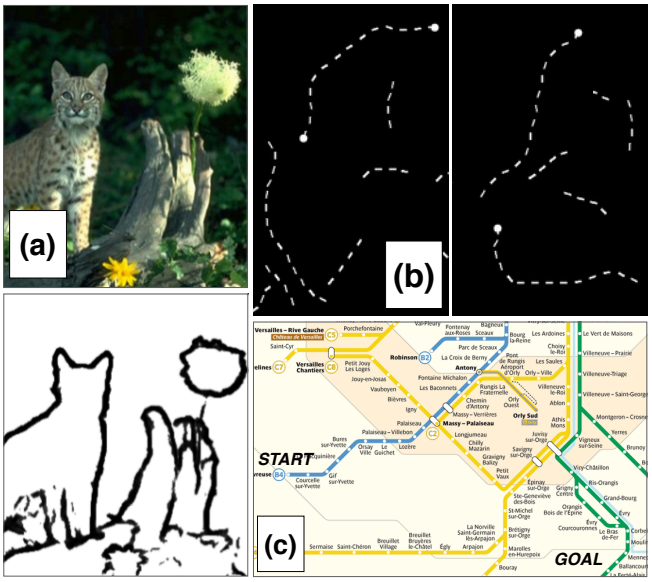

Figure 1: State-of-the-art CNNs excel at natural image contour detection benchmarks, but are strained by a task that depends on detecting long-range spatial dependencies. **(a)** Representative contour detection performance of a leading model. **(b)** Exemplars from the Pathfinder challenge: a task consisting of synthetic images which are parametrically controlled for long-range dependencies. **(c)** Long-range dependencies similar to those in the Pathfinder challenge are critical for everyday behaviors, such as reading a subway map to navigate a city.

ability of CNNs to learn such long-range spatial dependencies is limited by their localized receptive fields (RFs). This is typically addressed by building deeper networks, which increases the size and complexity of network RFs.

An alternative solution to problems that stress long-range spatial dependencies is provided by biology. The visual cortex contains abundant horizontal connections which mediate nonlinear interactions between neurons across distal regions of the visual field (Field et al., 1993). These intrinsic connections, popularly called "association fields", are thought to form the main substrate for mechanisms of contour grouping according to Gestalt principles, by mutually exciting colinear elements while also suppressing clutter elements that do not form an extended contour (Field et al., 1993). Such "extra-classical receptive field" mechanisms, mediated by horizontal

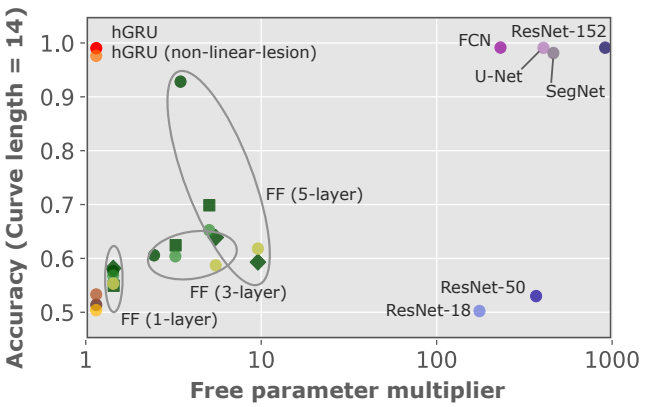

Figure 2: The hGRU efficiently learns long-range spatial dependencies that otherwise strain feedforward architectures. Its nearest competing feedforward models need at least 200× its number of parameters to match its performance on Pathfinder tasks. The x-axis shows the number of parameters in each model versus the hGRU (as a multiple of the latter). The y-axis depicts model accuracy on a version of Pathfinder featuring paths made up of 14 paddles (see 1b for examples).

connections, allow receptive fields to adaptively "grow" without additional processing depth. Several computational neuroscience models of these neural circuits have been proposed to account for an array of phenomena from perceptual grouping to contextual illusions (e.g., Mely & Serre, 2016). However, because these models are fit to data by solving sets of differential equations using numerical integration, they have so far not been amenable to computer vision. We implement the core ideas of these models in an end-to-end trainable extension of the popular gated recurrent unit (GRU) (Cho et al., 2014), which we call the horizontal GRU (hGRU).

We compared feedforward and recurrent approaches to capturing long-range spatial dependencies in a large-scale analysis of model performance on Pathfinder. This revealed a striking trend: feedforward models struggle at solving pathfinder, with only state-of-the-art feedforward models featuring millions of parameters across many processing layers succeeding (Fig. 2). An hGRU, on the other hand, efficiently solves Pathfinder with just *one layer* and a fraction of the number of parameters and training samples as feedforward models. The hGRU also outperforms all other tested recurrent models, including versions with lesions to its various mechanisms (such as linear nonlinear forms of excitation and inhibition), versions with less processing time, and a standard convolutional GRU (Fig. 2; yellow and brown dots).

We further investigated the nature of the horizontal connections learned by the hGRU by training it for contour detection in natural images (BSDS500 dataset). The learned kernels capture many of the canonical horizontal connectivity patterns found in cortex, including antagonistic near-excitatory vs. far-inhibitory surrounds, and the association field. We additionally find that fine-tuning this natural-image trained hGRU to a contour detection task yields a pattern of behavior in response to manipulations of contour salience that strongly correlates with

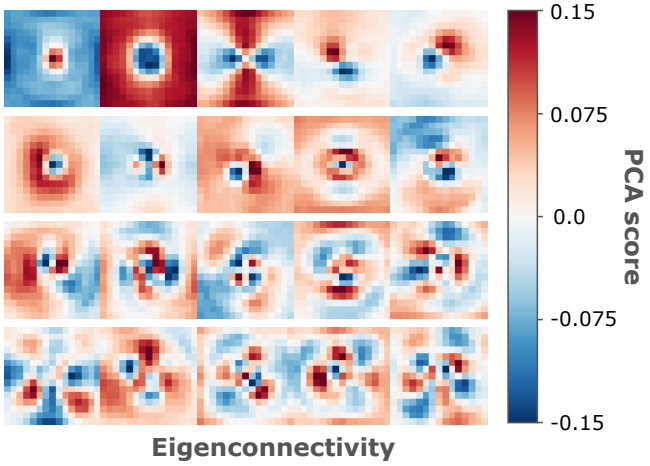

Figure 3: The hGRU learns horizontal connections that resemble cortical patterns of connectivity. Processed kernels from the hGRU depict polar center-surround interactions and association fields.

human observers (Li & Gilbert, 2002).

This work diagnoses a computational deficiency of feedforward networks, and introduces a biologically-inspired solution that can be easily incorporated into existing deep learning architectures. Beyond its effectiveness in computer vision, the weights learned by the hGRU and its corresponding behaviors are consistent with those associated with visual cortex, demonstrating its potential for establishing novel connections between machine learning, cognitive science, and neuroscience.

Acknowledgments

This work was supported by the Carney Institute for Brain Science, NSF early career award (IIS-1252951), and DARPA young faculty award (N66001-14-1-4037).

References

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014, June). Learning phrase representations using RNN Encoder-Decoder for statistical machine translation.

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: Evidence for a local "association field". *Vision Res.*, *33*(2), 173–193.

Houtkamp, R., & Roelfsema, P. R. (2010, December). Parallel and serial grouping of image elements in visual perception. *J. Exp. Psychol. Hum. Percept. Perform.*, *36*(6), 1443–1459.

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2015, February). Deeply-Supervised nets. In *Artificial intelligence and statistics* (pp. 562–570).

Li, W., & Gilbert, C. D. (2002, November). Global contour saliency and local colinear interactions. *J. Neurophysiol.*, *88*(5), 2846–2856.

Mely, D. A., & Serre, T. (2016, August). Opponent surrounds explain diversity of contextual phenomena across visual modalities. *bioRxiv*, 070821.