Global-and-local attention networks for visual recognition

Drew Linsley (drew_linsley@brown.edu)

Dan Shiebler (danshiebler@gmail.com)

Sven Eberhardt (sven2@brown.edu)

Thomas Serre (thomas_serre@brown.edu)
Cognitive Linguistic & Psychological Sciences Department
Brown University
Providence, RI 02912, USA

Abstract

Most recent gains in machine vision have originated from the development of network architectures which incorporate some form of attention. While biology is sometimes mentioned as a source of inspiration, the attentional mechanisms that have been considered by the computer vision community remain limited in comparison to the richness and diversity of the processes used by our visual system. Here, we describe a biologically-motivated "global-and-local attention" (GALA) module which is shown to yield state-of-the-art object recognition accuracy when embedded in a modern deep neural network. We further describe `ClickMe.ai`, a large-scale online experiment designed for human participants to identify diagnostic image regions for visual recognition in order to co-train a GALA network. Adding humans-in-the-loop is shown to significantly improve network accuracy, while also yielding visual representations that are more interpretable and more similar to those used by human observers.

Keywords: Object recognition; deep learning; biological vision; human-in-the-loop machine learning; visual features.

Decades of work on human attention have shown that our visual system uses at least two pathways to quickly guide attention to regions of interest (Torralba et al., 2006): A "global pathway" rapidly extracts a statistical summary of the whole scene in as little as a glance while a complementary "local pathway" leverages local features to extract salience cues (Itti & Koch, 2001). Most state-of-the-art networks including last year's squeeze-and-excitation ILSVRC winner, implement a global pathway (Hu et al., 2017). Here, we explore the role of the complementary local pathway and its interplay with a global pathway within a modern residual network architecture.

We designed the *global-and-local attention* (GALA) block as a circuit for learning complex combinations of local saliency and global contextual modulations in feedforward neural networks. GALA modulates input feature maps with an attention mask of the same dimension as the input. Importantly, this attention can be supervised by a human-in-the-loop to focus on the visual features that humans favor for object recognition.

We validated the effectiveness of GALA by embedding it within six mid- to high-level visual layers in ResNet-50 (layers 24, 27, 30, 33, 36, 39; each belonging to the same ResNet-50 processing block). This GALA-ResNet-50 slightly outperforms (6.35 top-1 error) both the standard ResNet-50 (6.86 top-1 error) and the SE-ResNet-50 (6.55 top-1 error) on ILSVRC.

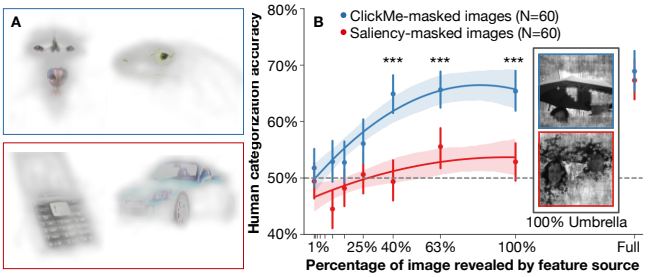

Figure 1: **(A)** A representative selection of ILSVRC'12 images and their ClickMe maps. The transparency channel of select images reflect the fraction of clicks for that location. Image features consistently deemed important for recognition are opaque and unimportant ones are transparent. Animals are outlined in blue and non-animals in red. **(B)** Features identified in ClickMe maps are more diagnostic for object recognition than those identified in saliency maps. A rapid visual categorization experiment compared human performance in discriminating animals vs. vehicles when features were revealed according to ClickMe maps (blue curve) or saliency maps (red curve). ClickMe- and Saliency-masked image exemplars are depicted for the condition in which 100% of important features are visible, demonstrating how Saliency is not relevant to the task. For clarity, we omitted data between 1-10% of features visible from this plot where accuracy was chance for of both groups. Error bars are S.E.M. ***: $p < 0.001$.

We further developed `ClickMe.ai`, a large-scale dataset of nearly 500,000 visual feature importance maps for images in the ILSVRC12 challenge. The game had a player "bubble" image regions deemed important for recognition. At the same time, a DCN tried to recognize a version of the image where only these bubbled parts were visible. The round ended as soon as the DCN recognized the image, with the player receiving points based on how quickly this occurred. Feature importance maps from this game were validated using human psychophysics by demonstrating the sufficiency of ClickMe fea-

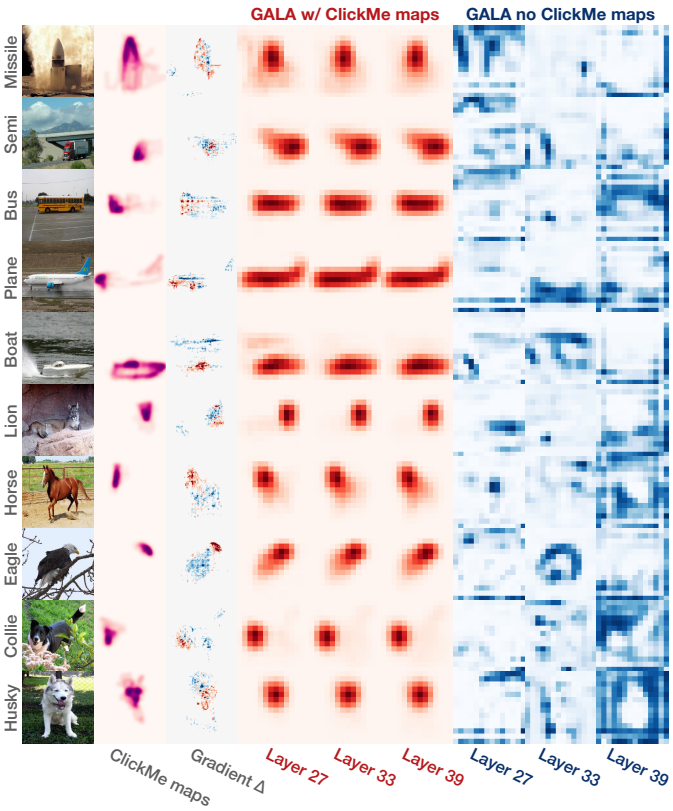

Figure 2: A GALA-ResNet-50 cued with ClickMe maps during training for object recognition uses visual features that are more similar to those used by human observers than a GALA-ResNet-50 trained without such cueing. ClickMe maps were gathered for ILSVRC12 validation set images that were held out of training, and highlight object parts that were deemed important for recognition. The difference between normalized smoothed gradient images from each network highlights regions that were more important to each network (Gradient Δ). Image pixels that were more important to a ClickMe GALA-ResNet-50 are colored in red, and those more important to a vanilla GALA-ResNet-50 are colored in blue. The column-wise L_2 norm of each network's GALA module reveals highly interpretable object and part-based attention for the GALA-ResNet-50 trained with ClickMe maps (in red) vs. less interpretable attention for the vanilla GALA-ResNet-50 (in blue).

tures for rapid visual categorization: on average, human observers were able to recognize objects pre-attentively with as little as 40% of the most important pixels derived from ClickMe maps (6% of all image pixels) 1. In comparison, when pixels were revealed according to their saliency-derived importance (from SALICON), observers only reached this performance when the whole image was visible. These results indicate that `ClickMe.ai` may provide insights into human vision with a measure of feature diagnosticity that goes beyond classic saliency measures.

We incorporated ClickMe maps into GALA-ResNet-50 training by optimizing for object classification and the similarity between the model's attention with visual features highlighted in ClickMe maps. After splitting the ClickMe dataset into separate train, validation, and test splits, we used the train and validation splits to investigate the trade-off of optimizing a model for either of these objectives. We found that with the proper trade-off, a GALA-ResNet-50 could learn more interpretable visual representations while yielding a robust improvement in classification accuracy. This result generalized to the held-out test: this GALA-ResNet-50 trained with ClickMe maps was more accurate than the same model trained without ClickMe maps, a SE-ResNet-50, and a standard ResNet-50 (49.29% 53.90%, 66.17%, and 63.68% top-1 error) while also having visual representations that explained a far greater fraction of human variance than any other model (88.56%** 64.21%**, 64.36%**, 43.61%; **: $p < 0.01$.).

The benefits of pairing the GALA-ResNet-50 with ClickMe also bear out in visualizations of its features. Strikingly, attention in the GALA-ResNet-50 trained with ClickMe maps, virtually without exception, focuses either on a single important visual feature of the target object class, or segments the figural object from the background. This effect persists in the presence of occlusion (Fig. 2, second row of GALA w/ ClickMe maps) and clutter (Fig. 2, fifth row of GALA w/ ClickMe maps). In comparison, some object features can be made out in the attention maps of a GALA-ResNet-50 trained without ClickMe maps, but there is no such localization, and the maps themselves are significantly more difficult to interpret.

In summary, our contributions are three-fold: **(i)** We extended the leading squeeze-and-excitation (SE) module with a novel global-and-local attention (GALA) module which combines global contextual guidance with local saliency to achieve state-of-the-art recognition performance on ILSVRC. **(ii)** We further described a large-scale online experiment to supplement ImageNet with a half-million image feature importance maps derived from human participants. These maps were psychophysically validated and used to co-train a 50-layer GALA residual network. **(iii)** Adding humans-in-the-loop is shown to significantly improve recognition accuracy while also creating visual representations that are more interpretable and more similar to those derived from human observers.

Acknowledgments

We are indebted to Vijay Veerabadran and Andreas Karagounis for their help. 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

Hu, J., Shen, L., & Sun, G. (2017, September). Squeeze-and-Excitation networks.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci.*, *2*(3), 194–203.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006, October). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. Rev.*, *113*(4), 766–786.