Interpreting Capsule Networks for Image Classification by Routing Path Visualization (Abstract Reprint)

Amanjot Bhullar[1], **Michael Czomko**[1], **R. Ayesha Ali**[1], **Douglas L. Welch**[2]

[1]Department of Mathematics and Statistics, University of Guelph, Canada
[2]Department of Physics and Astronomy, McMaster University, Canada

Abstract Reprint. This is an abstract reprint of the journal article by Bhullar, Czomko, Ali, and Welch (2025).

Abstract

Artificial neural networks are popular for computer vision as they often give state-of-the-art performance, but are difficult to interpret because of their complexity. This black box modeling is especially troubling when the application concerns human well-being such as in medical image analysis or autonomous driving. In this work, we propose a technique called routing path visualization for capsule networks, which reveals how much of each region in an image is routed to each capsule. In turn, this technique can be used to interpret the entity that a given capsule detects, and speculate how the network makes a prediction. We demonstrate our new visualization technique on several real world datasets. Experimental results suggest that routing path visualization can precisely localize the predicted class from an image, even though the capsule networks are trained using just images and their respective class labels, without additional information defining the location of the class in the image.

References

Bhullar, A.; Czomko, M.; Ali, R. A.; and Welch, D. L. 2025. Interpreting capsule networks for image classification by routing path visualization. *Artificial Intelligence*, 348: 104395.