iverse Vision Transformers align with humans and one another in representing textures

udovica de Paolis

Current research in vision neuroscience relies on object recognition, the leading paradigm in biological vision, which in turn influenced computational models of vision. In fact, Convolutional Neural Networks (CNNs) trained on object classification are regarded as the better model of human biological vision thanks to their plausibility and interpretability. In this study we challenge the current state-of-the-art by introducing an extensive study of texture representations. We consider textures of di erent complexity generated by three synthesis algorithms; using a rank-based statistics we quantify the information encoded in the internal representations in a CNN and in three Vision Transformers (ViTs) trained on di erent tasks with object images; finally, we compare the similarity between these representations to those inferred from human psychophysics data. We find that textures with increasingly complex attributes elicit activations compatible with the representation of objects in all models; that ViT representations are more similar to one another than those of the CNN; and that humans performance can be predicted better from ViTs than from the CNN. Taken together, these results suggest that the representational geometry of texture stimuli may be driven more by the network architecture, and that ViTs may capture more faithfully how texture patterns are visually processed by humans due to attention. We conclude that object recognition is not a sufficient framework to model texture representations, and that ViTs overcome CNNs despite the fact that they still hold a canonical role as models of the ventral visual stream.