Gaze-Guided Learning: Avoiding Shortcut Bias in Visual Classification

Jiahang Li
Tianjin Normal University, Tianjin, China

Shibo Xue
Tianjin Normal University, Tianjin, China

Yong Su
Tianjin Normal University, Tianjin, China

Abstract

Inspired by human visual attention, deep neural networks have widely adopted attention mechanisms to learn locally discriminative attributes for challenging visual classification tasks. However, existing approaches primarily emphasize the representation of such features while neglecting precise localization, which often leads to misclassification caused by shortcut biases. This limitation becomes more pronounced when models are evaluated on transfer or out-of-distribution datasets. In contrast, humans leverage prior object knowledge to quickly localize and compare fine-grained attributes, a capability especially crucial in complex and high-variance classification scenarios. We introduce Gaze-CIFAR-10, a human gaze time-series dataset, along with a dual-sequence gaze encoder that models the precise sequential localization of human attention on distinct local attributes. In parallel, a Vision Transformer (ViT) is employed to learn the sequential representation of image content. Through cross-modal fusion, our framework integrates human gaze priors with machine-derived visual sequences, effectively correcting inaccurate localization in image feature representations.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*