Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning

David Mascharka, Philip Tran, Ryan Soklaski, Arjun Majumdar | MIT Lincoln Laboratory

Overview

Visual Question Answering involves determining the correct answer for a given question-image pair

Unlike existing methods, TbD-nets leverage attention masks that are explicitly grounded in visual primitives.

Related Work

- Early MMNs [7] produced interpretable outputs using visual attention masks, but struggled to achieve good performance.
- By improving the natural-language parser and developing modules that process high-dimensional features rather than attention, Johnson et al. [5] significantly improved performance at the cost of interpretability.

Transparency by Design Networks

- Transparency by Design networks (TbD-nets) are built to achieve the performance of black-box models while surpassing the interpretability of initial MMNs by specializing each module type.
- Our approach reuses the program generator from [5] and focuses on improving the visual reasoning component to yield highly performant and interpretable modules.
- The visual reasoning component is comprised of modules which operate on and produce visual attentions.
- Each module is designed to perform spatial transformations on visual attention to suit its specific task.

TbD Visual Reasoning Component

Results on Main Task

- We evaluate our model on the CLEVR dataset [4], a visual reasoning benchmark comprised of synthetic scenes containing 3D shapes.
- We achieve state-of-the-art 99.1% accuracy on CLEVR with c=0.07.

Quantifying Interpretability

- Adding regularization and increasing the spatial resolution reduces the noise in and improves localization of the attentions.
- Specifically, we measure the center-of-mass overlap of the attentions with the ground-truth regions.

Results on Generalization Task

- The Compositional Generalization Test (CoGenTest) evaluates generalizability to new color/shape combinations.
- While our model learns entangled representations of color and shape (Train A), we quickly recover performance fine-tuning on a small amount of data (Fine-tune B).

Quantifying Entanglement

- We find that our model’s representation of shape is entangled with color (Predict Shape A), but its color representation is not entangled with shape (Predict Color A).
- Fine-tuning on a small amount of data rectifies the entanglement (Fine-tune B).

Code available at github.com/davidmascharka/tbd-nets

References