Paper Review: Hallucidoctor: Mitigating Toxicity in Visual Instruction Data

By: Lam Nguyen

1. SUMMARY

Multi-modal LLMs (MLLM) tuned on machine generated instruction following data perform very well on various understanding and generation tasks. However, hallucinations are still a regular occurrence. Hallucinations are defined as semantically correct outputs that are still incorrect in relation to the input question being asked.

To address Hallucinations, a process known as Hallucidoctor was developed. One of the factors that contribute to hallucinations are spurious correlations arising from long-tail data… In simple terms, what this means is that hallucinations occur because in every sort of domain from engineering, art, psychology, etc… There are freak occurrences that happen very rarely and are extremely unlikely… But they do happen. These unlikely events re then accounted for by the LLM, but this ends up causing hallucinations.

This kind of long-tail data is especially prevalent from artificially/machine generated data that is not completely based in the limits of reality.

Hallucidoctor executes a counterfactual visual instruction balance data distribution to essentially balance out these rare occurrences that contribute to hallucinations. This methodology reduces hallucinations by 44.6% and still maintains competitive performance with LLava.

Hallucidoctor takes place in these 3 steps:

1. Answer Chunks extraction
2. Answer-Based Question Generation
3. Consistency Cross Checking: Involves taking answers from multiple MLLMs and cross-checking them together.

2. STRENGTHS

- Other methods of mitigating hallucinations involve collecting remedial data for fine-tuning the model and adding an additional trained reward model during the inferencing phase. The problem with these kinds of methods is that they significantly increase training cost, training time, financial cost and increase inferencing time. They also don't address the underlying issues which is the toxicity of the dataset. What Hallucidoctor does is to fix the inherent toxicity within the machine-generated-dataset itself. And it does this in a very efficient and low-resource manner.

3. WEAKNESSES

- Methodology was not tested on other MLLMs besides LLAVA, LLAVA++, and GPT-4. Need to test on other Language Models.
- To further test the method of removing tail-end data in machine generated data… Test on other media formats

4. TECHNICAL EXTENSIONS

There are other things that can be done to mitigate Hallucinations that can be combined with Hallucidoctor without increasing computational cost. One addition that comes to mind is the addition of Visual Contrastive Decoding (VCD). VCD doesn't require extra training and just creates a correction factor between undistorted vs distorted images that can then be applied to calculations to mitigate hallucinations.

Another way to expand on this work is to move beyond image data input. Other forms of machine-generated training data such as video, 3D models, sensor input data, texts, etc… can be tested with this methodology.

5. OVERALL REVIEW

Hallucidoctor is able to effectively reduce hallucinations by removing tail end data from a dataset. It doesn't require extra GPU training time and the power usage that it entails. Further ways to reduce hallucinations would be to combine it with (VCD) which is a method that is also computationally efficient. Extensions to this project would be to use Hallucidoctor on other media formats besides images.