Detection of Hallucinations in Multimodal Models Based on Internal Representations

Kseniia Petrushina

Moscow Institute of Physics and Technology,
Skolkovo Institute of Science and Technology

Scientific supervisor: Dr. Alexander Panchenko

2024

Purpose of the study

Problem

As AI-generated images become more convincing, distinguishing realism from fiction is increasingly challenging.

Goal

Develop quantifiable realism measure to detect contextual and commonsense inconsistencies in visual content.

Tasks

1. Explore existing approaches in detecting image manipulation and realism.
2. Develop a method for obtaining reality score using language model reasoning.
3. Validate the method on a dataset of pairs of real and *weird* images.
4. Analyze the explanation of the *weirdness* of the images.

Problem Statement

Develop a **reality-check function** $f_{\text{reality}} : \mathbb{R}^{n \times n} \to \mathbb{R}$, which assigns a *reality score* to images such that:

- $f_{\text{reality}}(\mathbf{x}_{\text{r}}) > f_{\text{reality}}(\mathbf{x}_{\text{w}})$ for **real** images $\mathbf{x}_{\text{r}}$ and **weird** images $\mathbf{x}_{\text{w}}$, with high probability.
- Given a threshold τ, the function f_{reality} can be used to classify images as:

$$\text{Real: } f_{\text{reality}}(\mathbf{x}) \geq \tau, \quad \text{Weird: } f_{\text{reality}}(\mathbf{x}) < \tau.$$

Assumption: We are provided two datasets:

$$\mathcal{D}_{\text{r}} = \{\mathbf{x}_{\text{r}}^i\}_{i=1}^N, \quad \mathcal{D}_{\text{w}} = \{\mathbf{x}_{\text{w}}^i\}_{i=1}^N,$$

where $\mathbf{x}_{\text{r}}^i, \mathbf{x}_{\text{w}}^i \in \mathbb{R}^{n \times n}$.

Existing methods

1. **Probability**
   Realistic objects have high probability under distribution P.
2. **Adversarial losses** f-divergence between densities p, q

$$D_f[q||p] \geq \mathbb{E}_q[T(\mathbf{x})] - \mathbb{E}_p[f^*(T(\mathbf{x}))]$$

   Real-valued function T acts as a *critic* and produces larger values for samples from q and smaller for samples from p.
3. **Maximum mean discrepancy** Given two sets of iid examples $\mathbf{x}^M$, $\hat{\mathbf{x}}^N$

$$MMD^2(\mathbf{x}^M, \hat{\mathbf{x}}^N) = \|\frac{1}{M}\sum_m \Phi(x_m) - \frac{1}{N}\sum_n \Phi(\hat{x}_n)\|^2$$

   Φ is high dimensional feature space.

Base method

1. Extracting atomic facts
Using multi-modal model $f_{\text{cap}} : \mathbb{R}^{n \times n} \to \mathrm{T}^{m \times L}$ we obtain m sequences of language tokens of length L, which describe the details about the image:

$$f_{\text{cap}}(\mathbf{x}) = F_{\text{A}} = \{[t_1^i, \ldots, t_L^i] \mid i \in \overline{1, m}\}$$

2. Pairwise natural language inference
For each ordered pair of facts $(f_i, f_j) \in F_{\text{A}} \times \mathrm{F}_{\text{A}}$ we calculate entailment score via $f_{\text{nli}} : \mathrm{T}^L \times \mathrm{T}^L \to [-1, 1]$.

3. Aggregating pairwise scores
We form a set of sums for reordered facts

$$S_{\text{nli}} = \left\{ s_{\text{nli}}(f_i, f_j) + s_{\text{nli}}(f_j, f_i) \mid i, j \in \{1, \ldots, N\},\ i \neq j \right\}$$

Final score is calculated using aggregation (min, abs max, clustering).

2. Obtaining internal representations
For fact f_i we compute internal representations using text encoder $f_{\mathrm{enc}} : \mathrm{T}^L \to \mathbb{R}^d$.

$$X = f_{\mathrm{enc}}(F_{\mathrm{A}}) \in \mathbb{R}^{m \times d}$$

3. Attention-pooling layer Learnable fixed vector q_{cls} instead of the input X.

$$Q_{cls} = q_{cls} W^Q$$

$$f_{\mathrm{att}}(X) = \mathrm{softmax}\left(\frac{Q_{cls} K^T}{\sqrt{d_k}}\right) V,$$

where $K = XW^K, V = XW^V$.

Proposed method

Resulting function

The final formula for base method is

$$f_{\text{base}} = f_{\text{agg}} \circ f_{\text{nli}} \circ f_{\text{cap}}$$

$$f_{\text{reality}} = f_{\text{att}} \circ f_{\text{enc}} \circ f_{\text{cap}}$$

Optimization objective

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \left[\ell(0, f_{\text{reality}}(\mathbf{x}_{\text{r}}^{i})) + \ell(1, f_{\text{reality}}(\mathbf{x}_{\text{w}}^{i}))\right] \to \min_{W^Q, W^K, W^V, q_{\text{cls}}},$$

where

$$\ell(y, \hat{y}) = -\left[y \log(\sigma(-\hat{y})) + (1-y)\log(1-\sigma(-\hat{y}))\right]$$

Hypothesis

Correlation between resulting reality scores $R = \{f_{reality}(\mathbf{x})\}$ and presence of hallucinations in generated facts F_A is statistically significant on significance level 0.05.

Computational experiments

Data

Figure: Examples of *real* and *weird* images.

Metric

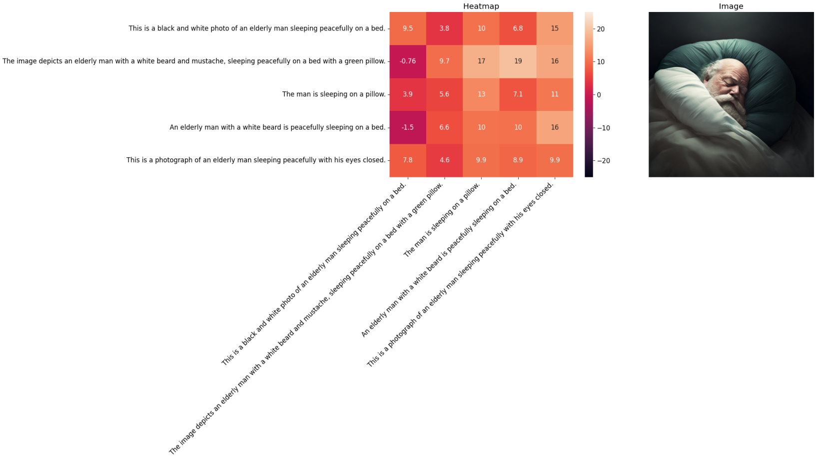
$$\text{Accuracy} = \frac{1}{2N}\sum_{i=1}^{N}\left[\mathbb{I}[f_{\text{reality}}(\mathbf{x}_r^i) \geq \tau] + \mathbb{I}[f_{\text{reality}}(\mathbf{x}_w^i) < \tau]\right]$$

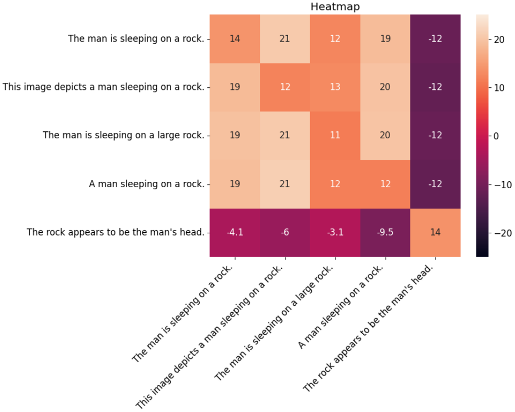
Results

Model	#	Mode	Acc ↑
BLIP2 FlanT5-XL	3.94B (188M)	ft	60.00
BLIP2 FlanT5-XXL	12.4B (188M)	ft	73.00
Attention-pooling	7.9B (2K)	ft	**73.54**
LLaVA 1.6 Mistral 7B	7.57B	zs	52.45
LLaVA 1.6 Vicuna 13B	13.4B	zs	56.37
InstructBLIP	7B	zs	61.27
InstructBLIP	13B	zs	62.25
Base method (NLI w/ clust agg)	7.9B	zs	**72.55**

Table: Accuracy of different approaches on the WHOOPS! benchmark.

Results


Heatmap
Image
This is a black and white photo of an elderly man sleeping peacefully on a bed.
The image depicts an elderly man with a white beard and mustache, sleeping peacefully on a bed with a green pillow.
The man is sleeping on a pillow.
An elderly man with a white beard is peacefully sleeping on a bed.
This is a photograph of an elderly man sleeping peacefully with his eyes closed.
9.5 3.8 10 6.8 15
-0.76 9.7 17 19 16
3.9 5.6 13 7.1 11
-1.5 6.6 10 10 16
7.8 4.6 9.9 8.9 9.9
20
10
0
−10
−20


Heatmap
The man is sleeping on a rock.
This image depicts a man sleeping on a rock.
The man is sleeping on a large rock.
A man sleeping on a rock.
The rock appears to be the man's head.
14 21 12 19 -12
19 12 13 20 -12
19 21 11 20 -12
19 21 12 12 -12
-4.1 -6 -3.1 -9.5 14
20
10
0
−10
−20


Image

Results

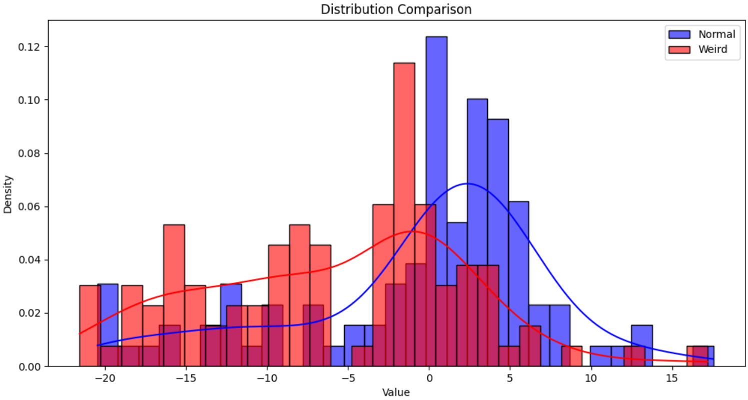


Figure: Reality scores for the whole dataset. p-value$=10^{-9}$ for Kolmogorov–Smirnov test.

Analysis

Measure	Value
$\mathbb{P}$(weird | digital)	0.76
$\mathbb{P}$(weird | hallucination)	0.81
$\mathbb{P}$(weird | hallucination & digital)	**0.93**

Table: The conditional probability of model prediction being weird given the occurrence of the hallucination or the marker from the corresponding set of words.

χ^2 test for contingency table analysis

		Hallucination		
		No	Yes	Total
Model prediction	Normal	78	10	88
	Weird	74	42	116
	Total	152	52	204

ϕ-coefficient $= 0.27$, p-value$=10^{-3}$

Conclusion

- Explored existing approaches in detecting image realism.
- Developed new method of quantifying image realism.
- Computational experiments on detecting *weird* images with different configurations.
- Hypotheses about properties of the *reality-check* function confirmed.
- Article submitted to The 39th Annual AAAI Conference on Artificial Intelligence.

Contribution: Developed research idea; conducted experiments with NLI method, linear probing, baseline models; analyzed obtained results and emerging phenomena.

Future work:

- Check transferability to other datasets.
- Conduct experiments on the dispersion of methods.

Literature

1. Lucas Theis. 2024. What makes an image realistic?. Proceedings of the 41st International Conference on Machine Learning.
URL:
`https://proceedings.mlr.press/v235/theis24a.html`
2. Sewon Min et al. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. Association for Computational Linguistics.
URL: `https://aclanthology.org/2023.emnlp-main.741`
3. Nitzan Bitton-Guetta et al. 2023. Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images. In IEEE/CVF International Conference on Computer Vision.
URL: `https://openaccess.thecvf.com/content/ICCV2023/papers/Bitton-Guetta_Breaking_Common_Sense_WHOOPS_A_Vision-and-Language_Benchmark_of_Synthetic_and_ICCV_2023_paper.pdf`