Quantifying image realism via language model reasoning

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 interpretable 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

Given unknown *real* and *weird* probability distributions

$$\mathrm{P}_{\mathsf{real}}(\mathbf{x}) : \mathbb{R}^{n \times n} \to [0, 1] \quad \mathrm{P}_{\mathsf{weird}}(\mathbf{x}) : \mathbb{R}^{n \times n} \to [0, 1]$$

and samples from the distributions

$$\mathcal{D}_{\mathsf{r}} = \{\mathbf{x}_{\mathsf{r}}^{i} \mid \mathbf{x}_{\mathsf{r}}^{i} \sim \mathrm{P}_{\mathsf{real}}(\mathbf{x})\}_{i=1}^{N}$$

$$\mathcal{D}_{\mathsf{w}} = \{\mathbf{x}_{\mathsf{w}}^{i} \mid \mathbf{x}_{\mathsf{w}}^{i} \sim \mathrm{P}_{\mathsf{weird}}(\mathbf{x})\}_{i=1}^{N}$$

Problem statement

We need to find a *reality-check* function

$$f_{\text{weird}} : \mathbb{R}^{n \times n} \to \mathbb{R}_+$$

that defines the realism score, that is for *real* image $\mathbf{x}_\text{r}$ and *weird* image $\mathbf{x}_\text{w}$, provided that they are close in a sense of similarity measure $\langle \cdot, \cdot \rangle$:

$$\langle \mathbf{x}_\text{r}, \mathbf{x}_\text{w} \rangle \leq \varepsilon,$$

the following holds true

$$f_{\text{weird}}(\mathbf{x}_\text{r}) < f_{\text{weird}}(\mathbf{x}_\text{w}).$$

Existing methods

1. **Probability**
   Realistic objects have high probability under distribution P.
2. **Weak typicality**

$$\mathbf{x}^N = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \overset{\text{iid}}{\sim} \mathrm{P}$$

$$\mathbb{P}\big[\lim_{N\to\infty} -\frac{1}{N} \log \mathrm{P}(\mathbf{x}^N) = H[\mathbf{x}_n]\big] = 1$$

Typical set is

$$A_\delta^N = \{\mathbf{x} : | - \frac{1}{N} \log \mathrm{P}(\mathbf{x}^N) - H[\mathbf{x}_n]| < \delta\},$$

its elements are *weakly typical*.

Existing methods

3. **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

4. **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.

5. **Universal critics** Measure of randomness to decide whether $\mathbf{x}$ was drawn from P:

$$U(\mathbf{x}) = -\log \mathrm{P}(\mathbf{x}) - K(\mathbf{x})$$

$K(\mathbf{x})$ is *Kolmogorov complexity* of $\mathbf{x}$.

Proposed method

Extracting atomic facts

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

$$\mathrm{f_{cap}}(\mathbf{x}) = \mathrm{F_A} = \{[t_1^i, \dots, t_L^i] \mid i \in \overline{1, m}\}$$

Pairwise natural language inference

For each ordered pair of facts $(f_i, f_j) \in \mathrm{F_A} \times \mathrm{F_A}$ we calculate entailment score via $\mathrm{f_{nli}} : \mathrm{T}^L \times \mathrm{T}^L \to [-1, 1]$. The results are presented in the form of a matrix

$$S_{ij} = \mathrm{f_{nli}}(f_i, f_j).$$

Aggregating pairwise scores

We take the sum of matrix elements, if both pairs (f_i, f_j) and (f_j, f_i) are contradictory and average it by the number of pairs:

$$\mathrm{f_{agg}}(S) = -\frac{1}{m^2} \sum_{i<j} (S_{ij} + S_{ji}) \mathbb{I}[S_{ij}, S_{ji} < 0]$$

Proposed method

Resulting metric

The final formula for *reality-check* function is

$$\mathrm{f}_{\mathrm{weird}} = \mathrm{f}_{\mathrm{agg}} \circ \mathrm{f}_{\mathrm{nli}} \circ \mathrm{f}_{\mathrm{cap}}$$

Hypothesis

Resulting reality scores $R = \{f_{weird}(\mathbf{x})\}$ *will correlate with probability densities* $P = \{\mathrm{P}_{real}(\mathbf{x})\}$*:*

$$r_s = \rho_{\mathrm{R}(f_{weird}(\mathbf{x})),\mathrm{R}(\mathrm{P}_{real}(\mathbf{x}))} \leq -0.5$$

Computational experiments

Data

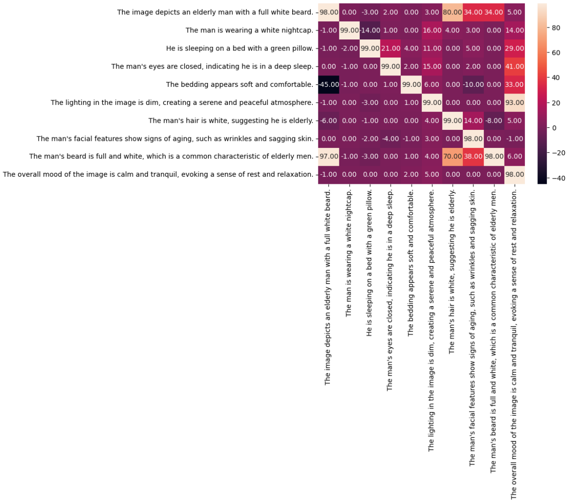
Figure: Examples of *real* and *weird* images.

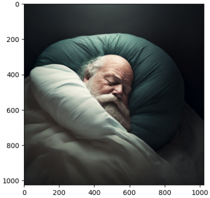
Metrics

1. $$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}[\mathrm{f}_{\text{weird}}(\mathbf{x}_r^i) < \mathrm{f}_{\text{weird}}(\mathbf{x}_w^i)]$$
2. Spearman's rank correlation coefficient

Results

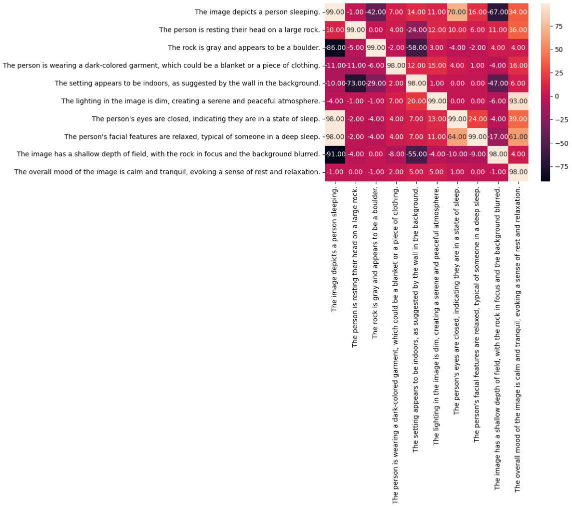
$f_{\text{weird}} = 0.31$

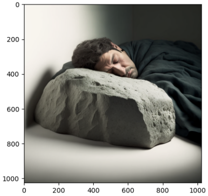




Results

$f_{\text{weird}} = 6.56$





Results

f_{nli} / f_{cap}	sileod	MoritzLaurer	t5-true
LLaVa	**0.68**	0.42	0.63
BLIP	0.53	**0.68**	0.53
GPT-2	0.37	0.32	0.37
GPT-4o	0.63	**0.68**	0.37

Table: Accuracy of realistic images detection using various functions for f_{cap} and f_{nli}.

Results

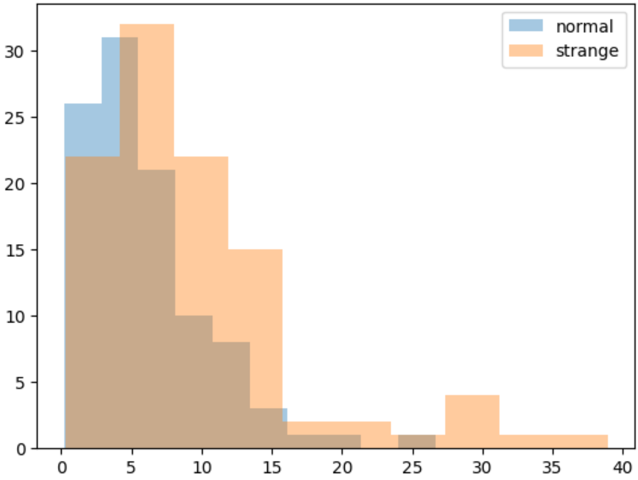


Figure: Reality scores for the whole dataset obtained using *LLaVa* model for captioning and *sileod* model for contradiction detection.
$p - value = 4e - 5$ for Kolmogorov–Smirnov test.

Conclusion

- Explored existing approaches in detecting image realism.
- Developed new interpretable method of quantifying image realism.
- Computational experiments on detecting *weird* images with different configurations.
- Hypotheses about properties of the *reality-check* function.

Future work:

- Conduct experiments with measuring correlation with other methods.
- Analyse the interpretability of the proposed method in more details.
- Carry on comprehensive study of ablation.

Literature

1. Lucas Theis. 2024. What makes an image realistic?. Proceedings of the 41st International Conference on Machine Learning.
   URL: `https://arxiv.org/abs/2403.04493`
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
   URL: `https://arxiv.org/abs/2303.07274`