informatiCup
Der GI-Wettbewerb für Studierende

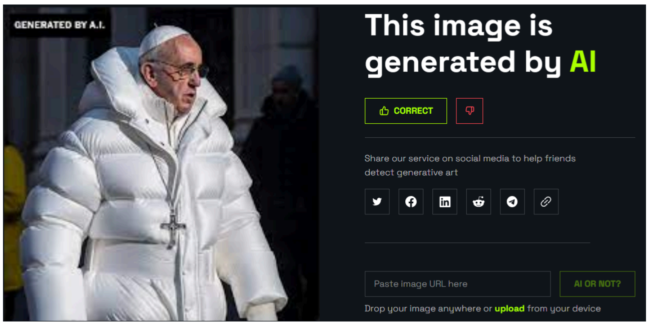
Producer: Gregor Kobsik
Concept: Johannes Nicolai

Fool AI-Detectors - (No) More Fake-News

Images created by text-to-image generators look more and more realistic. Large language models create more and more complex texts and answer questions in a way they are hardly distinguishable from a human. This does not only unlock an enormous potential to create creative content and to automate workflows, but also makes it possible to create fake news with minimal effort. To address this concern, the fictional startup “No More Fake News” (NMFK) set its mission to create a new model that will reliably and safely recognize content generated by AI to fight fake news produced by generative models.

The startup asked you to first conduct a technical market analysis. You are tasked to evaluate generally available AI models that can be used to produce texts and images. Furthermore, you shall evaluate which models would be suitable to detect AI generated content and assess their performance.

In a subsequent step, the startup likes to show that the market for new tools to detect AI generated content is not saturated yet. The associated task for you would be to create an application that can fool existing AI detectors by applying minimal changes to AI generated texts and images. The team from NMFK envisions using that application to augment their own data in order to make their own product as robust as possible against those kinds of modifications. That’s why your application should work fast and reliably on very large images and arbitrarily long texts.

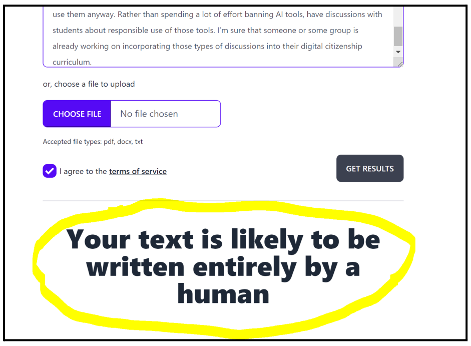


This image is generated by AI

Share our service on social media to help friends detect generative art

Drop your image anywhere or upload from your device

informatiCup 2024 tries to address the current “zeitgeist” around AI and generative models which affect our day to day lives more and more and will become an integral part of many consumer and industrial applications going forward. Even today, smart assistants like auto-correct, automated image editing are already common-place and will tremendously improve with generative models. To master informatiCup 2024’s challenge, participants have to demonstrate their skills to understand AI concepts in general, analyze large models and create creative solutions to fool those models. The broad mix of commercial aspects like market analysis and academic aspects like research on the current state of generative models is intentional to inspire our participants to also think about the impact of computer science on our society.



use them anyway. Rather than spending a lot of effort banning AI tools, have discussions with students about responsible use of those tools. I’m sure that someone or some group is already working on incorporating those types of discussions into their digital citizenship curriculum.

or, choose a file to upload

Accepted file types: pdf, docx, txt

☑ I agree to the terms of service

Your text is likely to be written entirely by a human

What is part of the challenge?

1. Technical market analysis:
    a. Generative models for text and image generation
    b. AI detectors for text and image generation
    c. Evaluation and analysis
2. Application design and implementation:
    a. Augmenting tool for images
    b. Augmenting tool for text
    c. Optional: Own AI detector (bonus points for nice integration / UX)

How does the application input look like?

- An image/text created by AI
- Image size / text length: “arbitrary”
    - minimal: 64 x 64 pixels. maximal: 1920 x 1080 pixels
    - minimal: 3 words, maximal 512 words
- Image-Generators:
    - DALL-E 2: https://openai.com/dall-e-2
    - Midjourney: https://www.midjourney.com/home
    - StableDiffusion: https://dreamstudio.ai/
- Text-Generators:
    - ChatGPT: https://chat.openai.com/
    - Bard: https://bard.google.com/
    - …
- We create a public test data sets for the students
- We create a private evaluation data set for the jury
- generated & real images / texts (internet or created by the jury)

How does output look like?

- A modified text/image that is not recognized by an AI-detector.
    - Text/image should look as much possible like the original
- Image size: same as input
- Text length: approximately like input

How can we evaluate submissions?

- Qualitative:
    - We manually compare Input and Output based on the evaluation and test data set.
- Quantitative Image:
    - AI-Detector Tools:
        - AI or Not: www.aiornot.com/

        - Illuminarty: https://illuminarty.ai/en/
        - Hive Moderation: https://hivemoderation.com/
        - …
    - Visual: RMSE on inputs and outputs
    - Semantic: CLIP Image Embedding, Embedding Vector Similarity (?)
- Quantitative Text:
    - AI-Detector Tool:
        - Hive Moderation: https://hivemoderation.com/
        - ZeroGPT: https://www.zerogpt.com/
        - Copyleaks: https://copyleaks.com/ai-content-detector
        - Writer: https://writer.com/ai-content-detector/
        -

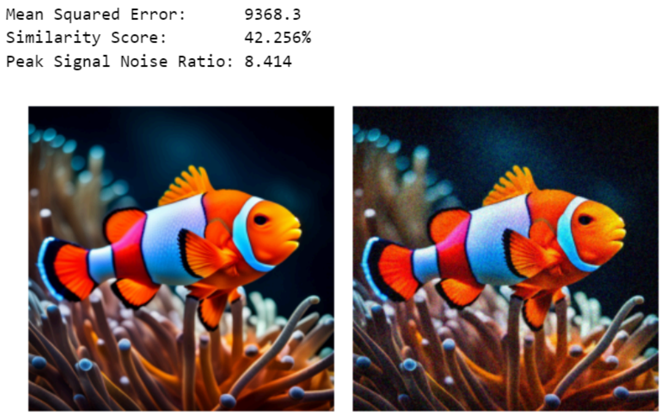
What are the challenges solving the tasks?

- Maximize visual and semantic similarity of the images while still fooling the detector
- meta understanding of generative models and AI detectors (blackbox testing)
- hardware and resource constraints:
    - reference: Intel i5 10Gen, RTX 2070S, 16GB RAM
    - 5 seconds compute time per text / image
    - If an on-prem AI model was used, maximum of 24h training / fine tuning
    - Publicly available models (APIs or weights) can be used without limitations

Add Gaussian Noise to Image:

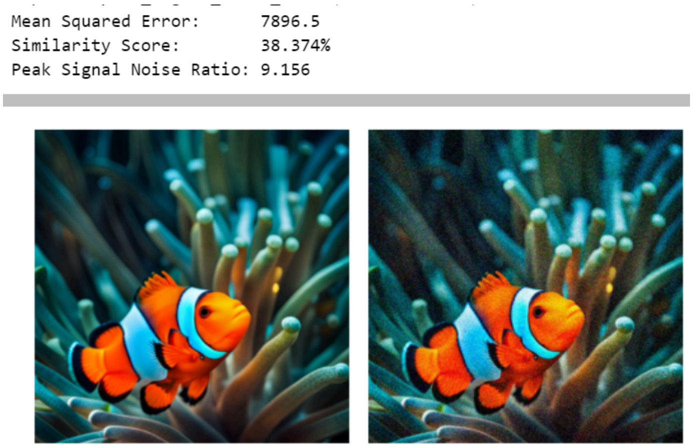
(larger number = higher AI generated probability)

Tool	left (AI Gen.)	right (+ Noise)
AI or Not	AI	AI
Illuminarty	90.7%	30.8%
Hive	100%	99.9%



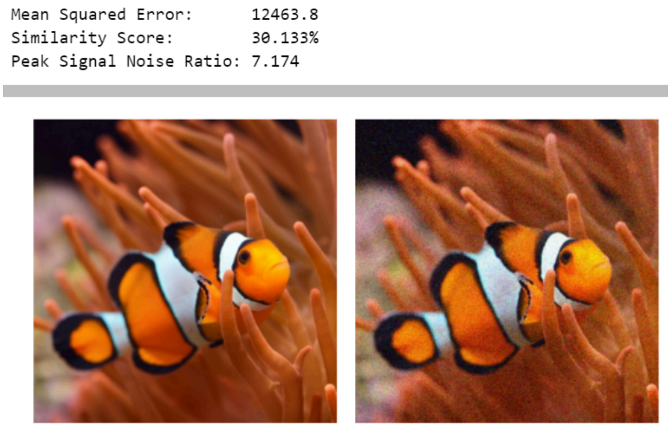
```
Mean Squared Error:       9368.3
Similarity Score:         42.256%
Peak Signal Noise Ratio:  8.414
```

Tool	left (AI Gen.)	right (+ Noise)
AI or Not	AI	Human
Illuminarty	98.9%	77.5%
Hive	100%	85.5%



```
Mean Squared Error:       7896.5
Similarity Score:         38.374%
Peak Signal Noise Ratio:  9.156
```

Tool	left (Human)	right (+ Noise)
AI or Not	Unsure	AI
Illuminarty	63.3%	26.2%
Hive	0.5%	0.4%



```
Mean Squared Error:       12463.8
Similarity Score:         30.133%
Peak Signal Noise Ratio:  7.174
```

Press Releases / Social Media

https://www.spiegel.de/netzwelt/web/quiz-zu-kuenstlicher-intelligenz-koennen-sie-ki-bilder-von-echten-fotos-unterscheiden-a-5b2d7cb4-d26f-4a5b-b2bf-c2a2ea98485d

https://www.nbcnews.com/tech/pope-francis-ai-generated-images-fool-internet-rcna76838

https://www.sueddeutsche.de/kultur/kuenstliche-intelligenz-urheberschaft-1.5741014

https://www.tagesschau.de/ausland/europa/italien-chatgpt-ki-101.html