Multimodal Machine Learning Lab

Winter Semester 2026/2027

Niklas Deckers and Martin Potthast

Deep Semantic Learning
University of Kassel and hessian.AI

Agenda

- Homework Recap
- (Adversarial) Attacks
- New Homework
- Finding and Reading Papers

©webis 2026

Homework Recap

- Task
- Frontends
- Performance
- Out-of-distribution inputs — symbolic images are difficult
- Attack approaches
- Attack generalization
- Problems? No VCS, AI disclosure incomplete

©webis 2026

(Adversarial) Attacks

- Our definition of attacks: Find some input so that the output is as desired.
- Hidden-box vs. open-box models
- GANs [webis.de]
- Generalization problem of the classifier

©webis 2026

New Homework

- You are given a hidden-box API serving two classification pairs: `class1`/`class2` and `class3`/`class4`
- Our goal will be to find out what they mean
- Use API requests to distill the classifier into a local network (might be a small network on top of CLIP)
- Run an adversarial attack against it (gradient-based)
- Example request and response:

```
POST https://wstud-mmml26-classification.web.webis.de/predict
{"image": "<base64 PNG>", "classes": ["class1", "class2", "class3", "class4"]}

HTTP/2 200
content-type: application/json
{"classes": {"class1": 0.0010988477151840925,
             "class2": 0.9989011287689209,
             "class3": 0.8647602200508118,
             "class4": 0.13523976504802704}}
```

- Submission of ipynbs (please include the output cells) **and of the trained models** in groups until Tuesday, 20:00, via Moodle

©webis 2026

Finding and Reading Papers

- Literature research: webis.de/lecturenotes
- Recommendation: asta.allen.ai
- Reading papers: abstract, introduction, conclusion

©webis 2026