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
- Gradient-Based Optimization in Discrete Spaces
- New Homework
- Reading Papers

©webis 2026

Homework Recap

- Task
- Different approaches
- Results (groups of images)
- Example approach: *Practical Black-Box Attacks against Machine Learning* arxiv.org

©webis 2026

Gradient-Based Optimization in Discrete Spaces

- Many of your approaches were based on searching some solution spaces
- Problem: Some vulnerabilities of CLIP are so hidden that we'll get to the point where search alone will not be sufficient
- I would like you to learn about gradient-based attacks
- How do gradients work on discrete input spaces?
- *Hard Prompts Made Easy: Gradient-Based Discrete Optimization for Prompt Tuning and Discovery* arxiv.org

©webis 2026

New Homework

- Objective: Find optimal prompts for the classes of the distilled models
- Optimality could mean: High for one class; or high for one class *and* low for the other
- There are some search-based approaches for this (popular: CLIP Interrogator)
- However, I would like you to use some gradient-based method (as shown in *Hard Prompts Made Easy*)
- Submission of ipynbs (please include the output cells) in groups until Tuesday, 20:00, via Moodle

©webis 2026

Reading Papers

- Literature research: webis.de/lecturenotes

©webis 2026