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
- Attack Types (Course Outline)
- Attacks Based on Relative Distances
- New Homework
- Cluster Onboarding

©webis 2026

Homework Recap

- Results of the homework show: Attacks in discrete token space generalize across underlying CLIP model variants
- Comparing the prompts: Dictionary support, token length
- Revealing the groundtruth targets

©webis 2026

Attack Types

- Attack target can be: (1) inputs, (2) CLIP model weights, (3) CLIP training data
- Until now: Focused on manipulating inputs
- Will now slowly move towards working on the model weights
- Later in the course: Manipulating model weights directly (by searching for locations in the model that are suitable for manipulation w.r.t. local behavior)
- Outlook: Can we modify training data to achieve a maximum effect on (certain local) behavior of the CLIP model? What are effects on applications such as generative models (which almost always use CLIP)?
- Can we identify fundamental problems underlying the CLIP training (leading to similar problems for all model variants)?

©webis 2026

Attacks Based on Relative Distances

- 3 spaces in which distances can be computed: (1) Text space, (2) image space, (3) shared embedding space
- Important: Relative distances are not uniform across these spaces
- Three attack scenarios:
  1. Obfuscation (e.g., to mitigate censorship): Find two points with maximum distance in the text space that map to the same point in the image space
  2. SEO (making one's own brand more findable through improved retrievability): Find an image with as many (or distant) as possible points that map from the text space
  3. Bias identification: The concept "doctor" should have the same distance from images of male doctors as from images of female doctors (related work: arxiv.org)

©webis 2026

New Homework

- Objective of the group work: Differentiate according to your interests; more open regarding the results; develop reusable toolsets
- This task will span two weeks
- **No lab next Wednesday (25.11.26)**
- But: **Mandatory** consultations on both Mondays (23.+30.11.26)
- Submission deadline: Tuesday, 01.12.26, 20:00
- Choose one of the attack scenarios on the previous slide
- Until next Monday: Literature research about CLIP weak points (related to your scenario), and implement your attack scenario (through optimization), do some experiments and present them in a frontend that's specific to your scenario and that shows the effect of the attack
- Until the Monday after that: Mitigate your attack (locally?) using LoRA (see arxiv.org and arxiv.org); including experiments and some demonstration
- For the final submission, also make an overview of the general-purpose tools that you developed (e.g., agent markdown instructions, best practices, maybe even models) that you would like to share with the other groups
- Code (including commit history) must be pushed to the Webis Gitlab

©webis 2026

Cluster Onboarding

- See onboarding instructions and knowledgebase
- Important rules:
  - You must not use any of the H100s (hopper)
  - You may use the GPU partitions (gpu:2g.10gb:1 or gpu:3g.20gb:1) or A100s (ampere), but only if required by your model
  - Only one GPU per job
  - Only one job per user at a time
  - No jobs longer than 5 hours (set a time limit)
  - Jobs can be terminated at any time - use checkpointing

©webis 2026