Multimodal Machine Learning Lab

Winter Semester 2024/2025

Niklas Deckers and Martin Potthast

Deep Semantic Learning
University of Kassel and hessian.AI

Designing a Small Initial User Study

Objective: Finding out whether the directed methods are actually better than just giving random images to the user

Methods That Should Be Compared

- EMA
- Random
- Function-based
- Baselines: Search on Lexica, normal prompting?

Initial Idea

- Compare final results for different users (assigning each user a randomly chosen system)
- Problem: Results depend on the used prompt
- $\rightarrow$ Use the same prompt for different systems, or use many more different prompts

Configuring the Number of Test Subjects

- When asking a test subject about multiple systems/prompt
  - The order of systems might affect the ratings
  - A learning effect might kick in across the presented systems
  - Impressions might dilute between the different systems, which requires targeted interviews and a presentation that explicitly shows that different systems are used
- When asking each test subject about a single system/prompt (requires more test subjects)
  - Effects that are specific to a single test subject might become less visible
  - The mean over more users might bring more validity
- Also consider practicability arguments restricting the number of test subjects
- Initial experiment to determine the number of iterations needed for an effect (might help estimating reasonable numbers of test subjects)

Dimensions for Comparison

- Final result
- Satisfaction
- Targeting the creative component:
  - Before using the system: Users describe the creative idea they have to create a picture from the prompt
  - After using the system: Users describe the creative idea(s?) that the system has come up with
  - Relative rating?
  - This evaluation mode would require targeted interviewing or giving the users example responses

Practical Considerations

- Limit the number of iterations?
- Allow early stopping (by user request)?
- Termination due to frustration or satisfaction? Should allow users to talk about their experience

Comparing the Systems

- Indirectly via a score
- Directly via pairwise annotations (should include comparing a method with itself to assess deviation)
- Comparison using identical prompts?
- When repeating the systems: Tell the users that there are no repetitions (i.e., that every prompt uses a different system)

Strategies for Assigning Methods to the Prompts

- Something related to Latin hypercube sampling?
- See `greedy_permutations.py` for a selection involving the permutations' Hamming distance

Solution for Now

- Given a list of prompts
- Each test subject goes through the full list
- For each prompt, a system is assigned
- Assignment mappings differ between the test subjects (see `greedy_permutations.py`)

Tasks

- Describe the experiments (fill the given Overleaf document)
- Prepare system for the experiments (creating logs incl. images, blind mode, filling in prompts, etc.)
- Come up with prompts (and for each prompt detail why they where chosen - what is important here?)