Multimodal Machine Learning Lab

Winter Semester 2024/2025

Niklas Deckers and Martin Potthast

Deep Semantic Learning
University of Kassel and hessian.AI

Agenda

- Prototype Progress Update
- Lab Experiments in IR
- Evaluation Experiments for Interactive Generation

Evaluation Experiments for Interactive Generation

- Steps from classic Cranfield-style IR evaluation: Building a corpus, defining topics, obtaining judgments
- Classic ML evaluation uses pre-defined and annotated evaluation data and has emphasis on avoiding leakage between training and test data
- Challenges in our system come from its interactive, generative and explorative aspects
- Using ideas from Cranfield to streamline open-end user studies (by giving the users a narrative to follow during interaction and evaluation)
- Must differentiate between explorative and descriptive user intent (different topics and different judgments)

Building a Corpus

- Objective: Corpus should be as universal as possible
- Argumentation: Stable Diffusion is very universal w.r.t. the representable images

Defining Topics

- Must include a narrative and maybe an initial prompt
- Must fully describe the user intent, from which the user behavior must be derived in each step
- Users interact with the system while acting on behalf of the user described by the topic
- Improves reproducibility and scalability of the experiments (compared to having users *just interact with the system*)

Ideas for Generating Topics

- Stock image meta data often contains keywords close to narrative for SEO purposes
- Prompt logs (lexica.art etc.) might allow to derive narratives; probably not representative
- Might want to check for a wide coverage of the space spanned by Stable Diffusion by the initial prompts in the narratives

Ideas for Designing the User Studies

- Experts vs. crowd sourcing (cost vs. quality tradeoff)
- Splitting the process among users to make sure that users don’t deviate from the given topic over time (including splitting off evaluation and/or defining the initial prompt)
- Diversity from having multiple users do the same task (on the same topics)
- Might also mix in results (intermediate/final) from other users or from baseline/groundtruth into the preference selection process to check for agreement

Metrics

- Checking whether a target image was reached (probably feasible for descriptive user intent only)
- Similarly to Portrayal: Checking whether the final image fulfills distinct pre-defined criteria
- User experience (via questionnaire)
- Number of iteration needed until convergence/target
- ...?

Ideas for Putting the System to a Test

- Abstract concepts like diligence, tiredness, creativity... (but emotions might be easy to visualize by showing humans)
- Having users generate very specific objects (which might be unsuitable for explorative systems)
- Creating funny images usually does not work through user-guided iterations if there is no specific funny idea in the beginning. This might be a difficult challenge for explorative systems.

Exercise for Next Week

- Topic definitions often follow a certain challenge idea
- Example: Queries submitted to an IR system can be ambiguous, like `pet therapy`, which can mean giving therapy to a pet or a pet giving therapy to a human.
- Your task: Come up with a challenge idea for topics that are particularly challenging (but still solvable) for systems that are interactive, generative and explorative. Describe the idea and why it might be a challenge. Formulate 3-5 topics that demonstrate this challenge (including initial prompts, descriptions and narratives).
- Submission via email (Tuesday evening)
- We will have a brainstorming session together halfway through the deadline