Multimodal Machine Learning Lab

Winter Semester 2024/2025

Niklas Deckers and Martin Potthast

Deep Semantic Learning
University of Kassel and hessian.AI

Evaluating Generative Models Using Topics

Evaluating descriptive (i.e., non-creative) tasks trough topics is relatively easy since initial prompt, voting behavior and evaluation can be based on a single description.

What are difficulties of evaluating creative tasks through topics?

- ❑ Branching (it is difficult to exhaustively consider all possible outcomes of the systems when designing the evaluation)
- ❑ Subjectivity (users have individual preferences that should be reflected; annotators have different interpretations of these preferences)
- ❑ Modelling user surprise (difficult since it requires a fine-grained model of the user’s knowledge)

Evaluation through topics depends on what is considered the “final result” of the method:

- The collection of all images shown in the process
- Or the single final result image
- Depending on this, the topics might have to be defined differently

Ideas for challenges that can be expressed through topics:

- ❑ In general: Split the user input into initial prompt (can be probed through topics) and voting behavior (can also be probed through topics, but this is not as straightforward)
- ❑ Ambiguity (homonyms: evaluation should include multiple runs that require different meanings; typos)
- ❑ Coarse initial prompts that leave some aspects out
- ❑ Expert knowledge (that is not part of the training dataset)
- ❑ Variety in formulation: The initial prompt could use imperatives (“draw a ...”) or formulate desires (“I want a ...”) or an abstract description (“show me something funny”)
- ❑ Generating images that trigger the desired emotions (instead of just illustrating them)
- ❑ Discrepancy between the narrative and the initial prompt (what the user writes is different from what the user wants)
- ❑ User changes their voting behavior mid-process
- ❑ Voting behavior varies due to the user exploring what they want to see
- ❑ Subjectivity: Users find different jokes funny; different things spark nostalgia

Modelling user surprise in topic narratives:

- Enumeration of everything that surprises the user
- or of everything that doesn’t surprise them
- or: Description of the user (which implies what is surprising)
- Problem of completeness (exhaustive enumerations are difficult)
- Spanning the space of things that are surprising through implicit descriptions might make annotating difficult
- Must think beyond “lists of objects that should be contained in the images” since the evaluation should also work for future systems (that might be able to handle very complex prompts)

Partitioning the space of generated images into three categories:

- ❑ Concepts that the user finds unsuprising
- ❑ Concepts that the user finds surprising
- ❑ Concepts that are too random/out of scope for the user
- ❑ Maybe have the model explicitly learn the (subjective) border to the too random images?

Taking images as part of the narrative

- A narrative should contain precise descriptions of what is acceptable
- This might be represented in the form of a requirements sheet like they have for passport photos (with positive and negative examples, grouped by categories)
- Mood boards/Pinterest boards might be used to characterize the users as a person within the narrative
- Might include (1) images of things that the user finds interesting (2) images that characterize the user in a more abstract way
- Such mood boards might also be possible as input for the creative system - taken from the everyday life of the user or from related images on e.g. lexica that the user finds interesting

Additional ideas for the metrics:

- Is there a form of abandonment (where users become impatient and give up early or at least would want to give up)?
- How does the gain evolve (i.e., does the method only have marginal use from some point)?
- Annotators could be asked to bid for topics that they would want to evolve further

Further ideas for putting the systems to the test:

- Given a text, find a symbolic image that represents this text. Evaluate how well this image corresponds to the text.

- For visualizing abstract concepts: Test whether another user can correctly match the generated image with the target concept

Ideas for input modalities for the voting system:

- Sliders
- Trello cards (drag and drop to reorder)
- Putting images to an axis
- Thumbs up/down
- Swipe left/right
- Input numbers
- Buttons for numbers
- Comparative binary labels (when shown two images at a time)
- Ternary options (too boring, suprises me, too random)
- Emojis (for surprise, confusion, preference, disgust, boredom) - should also be mapped to keyboard keys. Maybe glue emojis to a physical keyboard?