Offline Evaluation of Set-Based Text-to-Image Generation

Negar Arabzadeh, Fernando Diaz, Junfeng He
University of Waterloo, Google

GitHub

Paper

TL;DR: This study introduces new offline evaluation metrics for Text-to-Image systems, emphasizing both relevance and diversity in generated image sets. While current evaluation metrics like Fréchet Inception Distance (FID) focus on distribution similarity, our study introduces evaluation methods based on user interaction and browsing behaviors .Human studies validate our approach.

Problem Definition

- **Tasks definition:** Given a prompt or query q, a TTI system s, generate and presents the generated images in a grid view of $m \times n$ images.
- **Task Evaluation:** An evaluation metric μ is a function that, given an arrangement of generated images X, a prompt q, and side information about the image utility u_q(e.g. an example target image), computes a scalar value where a higher value indicates better performance of system.
- Performance of a system over a space of prompts Q :

$$E_{q\sim Q}\left[\mu(s(q), u_q)\right]$$

Design Desiderata

The theory of measuring the quality of an ideation process focuses on:

- **Fluency:** Total number of relevant items generated.
- **Variety:** The number of unique types of relevant items generated.
- **Novelty:** How distinct relevant items are from previously generated items.
- **Quality**: The degree of relevance of generated items.

Goal: Design metrics capturing different dimensions of ideation effectiveness.

Our Proposed Set of Evaluation Metrics

Preliminaries:

- **Trajectories**: A specific sequence of inspected images as a permutation π of $[1, k]$ which we refer to as a trajectory.
- $\boldsymbol{\gamma}$:The position-based models model the probability of a user inspecting the image at rank positioning the trajectory as γ^{i-1}, where γ is a free parameter controlling the depth the user is likely to reach.
- $\boldsymbol{g^*(x)}$ **:** The probability that a users satisfied by an image.

Proposed metrics:

- **Fluency:**
  - For position-based model, rank-biased precision (RBP) and for cascade model, extended expected reciprocal rank (ERR) metric , is defined as:

$$RBP(\pi) = \sum_{i=1}^{k} f^*(\pi_i)\gamma^{i-1}$$

$$ERR(\pi) = \sum_{i=1}^{k} f^*(\pi_i)\gamma^{i-1} \prod_{j=1}^{1} 1 - g^*(\pi_j)$$

- **Quality**:
  - Unlike IR evaluation setting, we cannot prior judge the relevance of all possible images, we assume that we have access to one or more known relevant example(s) to the prompt.
  - Given a generated image x and a relevant image $\hat{x}$, the relevance of x is as $f^*(x) \in [0,1]$ where $\emptyset$is the image embedding function:

$$f^*(x) = \langle \emptyset(\hat{x}), \emptyset(x) \rangle$$

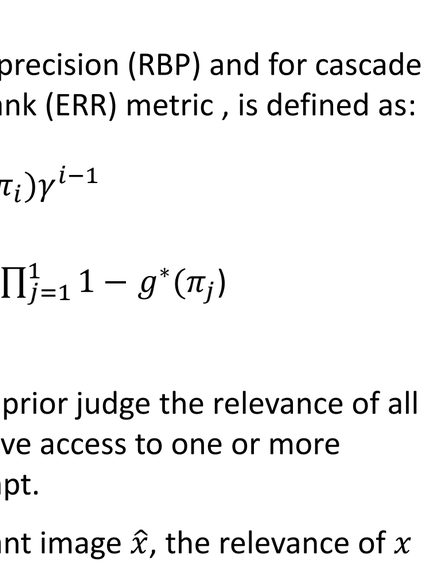
- **Variety and Novelty:** Similar to Maximum Marginal Relevance , we model novelty by discounting the relevance of an image based on previously seen images.

Different variations of our proposed metrics

Metric	User Model	Relevance	Pr(π)	Metric	User Model	Relevance	Pr(π)
RBP	Position-based	$f^*(x)$	Saliency	ERR	Cascade-based	$f^*(x)$	Saliency
RBP^u_η	Position-based	$f^*(x) \times \eta(i,\pi)$	Uniform	ERR^u_η	Cascade-based	$f^*(x) \times \eta(i,\pi)$	Uniform
RBP_η	Position-based	$f^*(x) \times \eta(i,\pi)$	Saliency	ERR_η	Cascade-based	$f^*(x) \times \eta(i,\pi)$	Saliency

Expected Metrics over Trajectories:

- Users may inspect images in arbitrary trajectories based on their position and attractiveness.
- $\Pr(\pi)$: the probability that the user scans the images in the order represented by π;
- Inspired by users tendency to be more attracted to certain images based on their position and visual features, we propose trajectories based on
  1. Visual saliency
  2. Uniform distribution



Experimental Setup

- **COCO captions (MS-COCO)**: A random sample of 500 images from the MS COCO 2017, using one caption per image as the prompt.
  - **Example**: **A herd of cows standing on a grass covered hillside.**
- **Localized Narratives (LM-COCO)**: A more detailed subset of MS-COCO, helps simulating TTI systems with longer prompts.
  - **Example**: **In this picture we can see three cows standing on the grass. There is a tree and few mountains are visible in the background.**
- **Prompts Dataset**:
  - A Collection of 500 prompts from real users of a TTI demo which are more indicative of real-world TTI system . Both implicit and explicit user feedbacks such as thumbs up thumbs down were used to select the generated target images.
  - **Example** : **Origami cow flying over the moon.**

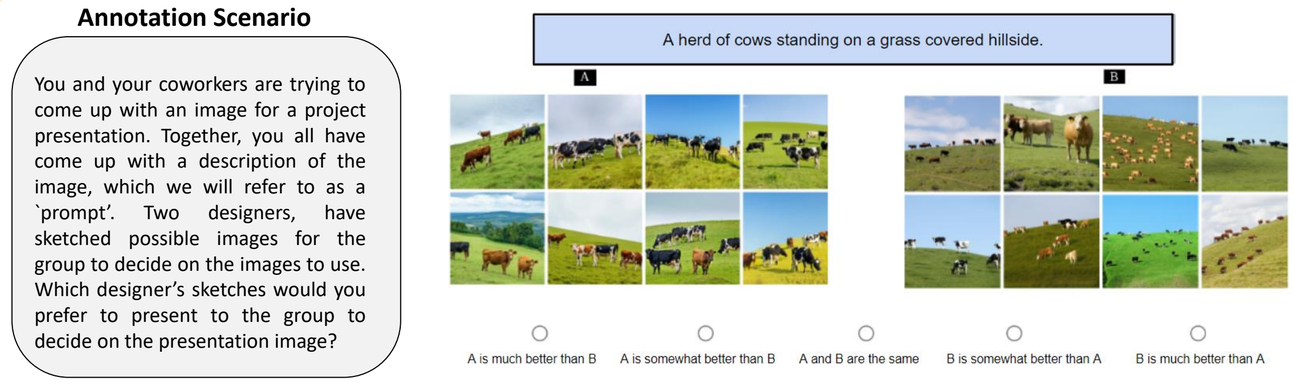
Examples of ground truth Images for COCO datasets (top) and Prompts dataset (Bottom)

TTI systems under experiments:

We consider three TTI systems under experiments and to focus on evaluation perspectives, we refer to these as System S, B and B'.

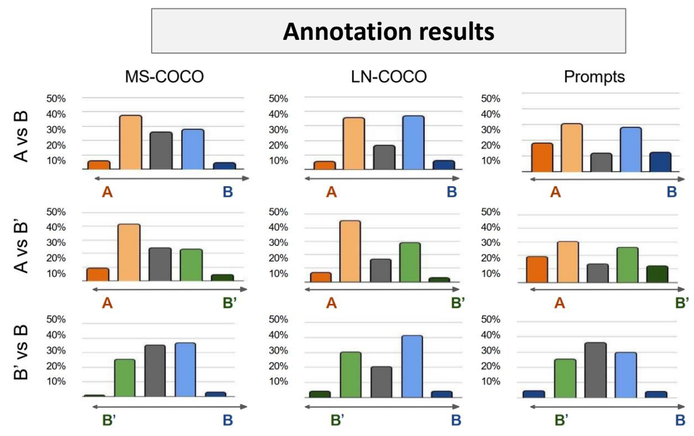
- B' is the smaller version of system B, with fewer number of parameters.

Data Annotation



Findings

- **Dependency on Dataset**: The preference rate between each pair of systems significantly depends on the dataset.
- **Challenging Prompts Distinguish Systems Better**: when prompts are challenging, differences annotators more frequently opted for extreme choices.



Results

Agreement rate of the metrics with human annotations. Statistically significant agreement with Wilcoxon paired test and p-value < 0.05 are shown with * symbol.

	MS-COCO			LN-COCO			Prompts		
	A vs B	A vs B'	B vs B'	A vs B	A vs B'	B vs B'	A vs B	A vs B'	B vs B'
Diversity	47.2%	41.9%	53.1%	45.3%	42.1%	45.7%	45.9%	46.3%	47.1%
RBP	53.7%	55.3*%	55.5%	53.1*%	64.1*%	52.1%	56.9*%	60.2*%	54.2%
RBP_η	**53.8%**	**61.2***%	51.7%	52.5*%	58.8*%	54.2%	58.5*%	60.7*%	54.2%
RBP^u_η	53.3%	58.9*%	52.9%	54.2*%	60.0*%	53.4%	57.7*%	60.2*%	49.5%
ERR	52.2%	55.4*%	51.7%	50.0*%	60.1*%	54.2%	57.7*%	61.5*%	53.5%
ERR_η	52.2%	60.4*%	50.9%	53.1*%	60.8*%	56.2%	**59.3***%	**63.1***%	54.2%
ERR^u_η	52.2%	58.4*%	52.9%	**55.2***%	**62.6***%	52.7%	58.1*%	61.5*%	52.3%

Findings

1.Human Agreement & Metrics:
- Proposed metrics showed a high agreement with human annotators.

2.Impact of Prompt Complexity:
- Simple prompts from MS-COCO led to generating indistinguishably good results.
- More complex prompts, can discern differences in system performance.

3.Superiority of Novelty-Based Metrics:
- Considering variety and novelty in boosts alignment with human annotations.

4.Trajectory Sampling & Challenging Prompts:
- Sampling trajectories from the grid's saliency distribution proved more beneficial than position based trajectories as prompts got more complex.

5.Comparison with Existing FID Metric:
- FID could gauge performance for simpler prompts but it faltered on challenging prompt sets.
- FID metrics sometimes contradicted human preferences, suggesting a need for more refined evaluation metrics.