Visually Perspective Similarity Metric for T2I Models

Sahana Subramanya Kowshik
skowshik@bu.edu

Anming Gu
agu2002@bu.edu

Sijia Li
scartt@bu.edu

Abstract

Advances in AI research have resulted in deep–learning models capable of generating images from basic text prompts. One limitation of these text-to-image models is that there is no known metric to measure how robust these models are to slight changes in the prompt. We aim to measure these changes in the generated output images by creating a metric for image similarity.

1 Introduction

Text-to-Image (T2I) models such as Dall-E2 and Stable Diffusion can generate photorealistic images in response to a text prompt. However, the images generated by these models are not consistent. The same text prompt can generate diverse output images, or making small modifications to the prompt by changing the location or color of the objects or adding words like "a" and "the" can alter the generated image. No similarity metric exists to measure the variance in the output images generated by these T2I models.

In this project, we aim to develop a good visually perspective similarity metric to quantify these variations in the generated images by utilizing known metrics like L_2 norm, cosine similarity, and inception scores. We also intend to plot the visual similarity between the intermediate frames of the stable diffusion model.

1.1 Related Works

Text-to-image models such as Imagen [1] and Dall-E2 [2] has achieved remarkable performance in text-to-image synthesis. Stable Diffusion [3] is an open-source lightweight latent text-to-image diffusion model similar to Google's Imagen. They achieve a new state of the art for unconditional image generation, semantic scene synthesis, and super-resolution. Since image generation happens in the latent space, it significantly reduces the computational requirements. However, these models do not preserve the identity of the subject across different settings. Given a sample prompt of text and a few images of an object, DreamBooth [4] provides an approach to generate realistic images of the object that match the given prompt while maintaining subject consistency.

All these models create realistic images and art from a description in natural language. However, they don't quantify the similarity or the variance between the generated output images.

1.2 Problem Formulation

On a high level, we aim to assess the similarity of the images generated by a T2I model given the same and distinct text prompts. Finding similarities between the output images based on their pixel values is not an accurate way to compare them since the output images are inconsistent.

Preprint. Under review.

Therefore, we first use pre-trained models like VGG16 and OpenAI CLIP to extract key features and characteristics from these images and then compare how similar the resulting feature vectors are. We then examine the sensitivity of these models to minute perturbations in prompts like changing the color or location of the object. Finally, we plot the similarity between intermediate frames of stable diffusion. [1]

2 Methods

This section describes the data, the pre-trained models utilized, and the various similarity metrics used for comparison.

2.1 Data

ImageNet-v2 [5] dataset is used to test the robustness of the L_2 Norm similarity metric. 100 prompts are used to generate pairs of images for both stable-diffusion and Dall-E2.

2.2 Pre-trained Models

2.2.1 VGG16 Model

The VGG16 pre-trained model with input dimension $(224, 224, 3)$ is used, and the final prediction layer is removed so that the output dimension is $(4096, 1)$. The output of the last fully connected layer is used as the feature vector for an image.

2.2.2 Contrastive Language-Image Pre-Training (CLIP) Model

OpenAI Contrastive Language-Image Pre-Training (CLIP) [6] Model, which is a neural network already trained on a variety of (image, text) pairs, is used with sentence-transformers to compute dense vector representations for the generated images.

2.3 Similarity Metrics

2.3.1 Similarity Metric Based on L_2 Norm

The similarity between two images represented by feature vectors v_1 and v_2 is computed using

$$\text{sim}(v_1, v_2) = 1 - \|\hat{v_1} - \hat{v_2}\|_2,$$

where $\hat{v} = v/\|v\|_2$. Before testing the similarity model, the data is normalized using the mean and standard deviation of ImageNet. The more similar the images are, the closer the value is to 1.

2.3.2 Cosine Similarity

The similarity of two images represented by feature vectors v_1 and v_2 is computed using

$$\text{sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|}$$

The more similar the images are, the closer the value is to 1.

2.3.3 Inception Score

The inception score is calculated by first using a pre-trained Inception v3 model to predict the class probabilities for each generated image. Kullback-Leibler divergence or KL divergence (relative entropy) between the conditional and marginal probability distributions is calculated using the formula

$$KL = p(y|x) * (\log p(y|x) - \log p(y))$$

The KL divergence is then summed over all images and averaged over all classes, and the exponent of the result is calculated to give the final score. It gives a lower score if the images belong to the same class.

[1] The code is available at https://github.com/sahanakowshik/Visually-Perspective-Similarity-Metric-for-T2I-Models

3 Results

3.1 Robustness Result

Robustness results of the L_2 Norm-based similarity metric are summarized in table 1. The similarity of one image in ImageNet-v2 compared with the other images in the dataset gave an average similarity of 0.582 and a standard deviation of 0.066. The VGG16 model is very robust to both uniform and normal perturbations.

Table 1: Similarity metrics of the pre-trained VGG16 model. The robustness of the model is tested using 1000 images from the ImageNet-v2 test set. $\mathcal{N}(\mu, \sigma)$ represents normal noise with mean μ and standard deviation σ, and $\mathcal{U}[\alpha, \beta]$ represents uniform noise on the interval $[\alpha, \beta]$.

Perturbation	Mean Similarity	Variance
$\mathcal{N}(0, 0.01)$	0.992	0.003
$\mathcal{N}(0, 0.05)$	0.949	0.017
$\mathcal{N}(0, 0.1)$	0.914	0.024
$\mathcal{U}[-0.05, 0.05]$	0.972	0.01
$\mathcal{U}[-0.1, 0.1]$	0.942	0.0186
$\mathcal{U}[-0.2, 0.2]$	0.907	0.025

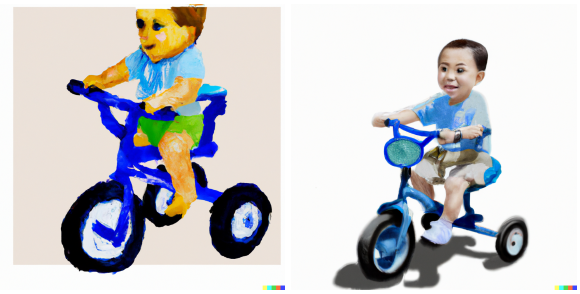
3.2 T2I Results

Stable Diffusion and Dall-E2 models are tested by generating images using two modes: using the same prompt and using a slightly different prompt. We show example outputs of Stable Diffusion in Figure 1 and example outputs of Dall-E2 in Figure 2.

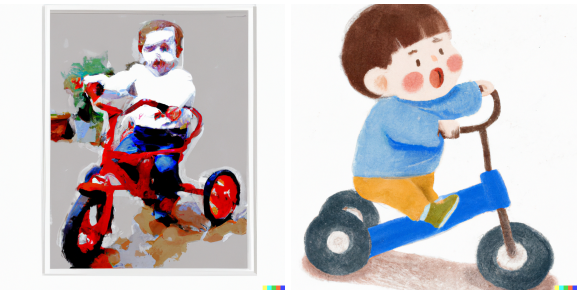
(a) Prompt: "A painting of a baby riding a blue bike."

(b) Prompt: (left) "A painting of a baby riding a *red* bike." (right) "A painting of a baby riding a *blue* bike."

Figure 1: Images generated by Stable Diffusion. (a) using the same prompt. (b) using modified prompts.



(a) Prompt: "A painting of a baby riding a blue bike."



(b) Prompt: (left) "A painting of a baby riding a *red* bike." (right) "A painting of a baby riding a *blue* bike."

Figure 2: Images generated by Dall-E2. (a) using the same prompt. (b) using modified prompts.

3.2.1 Similarity Scores

Table 2: Similarity metrics on Stable Diffusion.

Prompt Type	VGG16 Feature Vector				CLIP Feature Vector			
	L_2 Norm		Cosine similarity		L_2 Norm		Cosine similarity	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Same prompt	0.67	0.005	0.48	0.018	0.38	0.020	0.80	0.007
Modified prompt	0.66	0.004	0.48	0.015	0.38	0.013	0.80	0.005
Different prompt	0.55	0.006	0.25	0.007	0.05	0.006	0.55	0.005

Table 3: Similarity metrics for Dall-E2.

Prompt Type	VGG16 Feature Vector				CLIP Feature Vector			
	L_2 Norm		Cosine similarity		L_2 Norm		Cosine similarity	
	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
Same prompt	0.67	0.005	0.47	0.019	0.35	0.030	0.77	0.015
Modified prompt	0.68	0.002	0.48	0.010	0.40	0.013	0.81	0.005
Different prompt	0.57	0.003	0.29	0.009	0.08	0.007	0.57	0.006

Tables 2 and 3 give similarity metrics for the images generated by a Stable Diffusion model using the methods described in Equations 2.3.1 and 2.3.2. Table 5 gives the mean and variance of the inception scores of 100 pairs of images generated using the Stable Diffusion model and Dall-E2.

Table 4: Inception scores for Stable Diffusion and Dall-E2

Prompt Type	Stable Diffusion		Dall-E2	
	Mean	Variance	Mean	Variance
Same prompt	1.45	0.06	1.47	0.05
Modified prompt	1.44	0.06	1.47	0.05
Different prompt	1.70	0.04	1.68	0.03

3.3 Stable Diffusion Intermediate Frames

Figure 3 shows the intermediate results of a Stable Diffusion model for the prompt "A painting of a baby riding a blue bike" over the course of 100 iterations. Notice that in the early iterations, it appears that the model is defining borders between objects, while in later iterations, the model smooths out the noise.

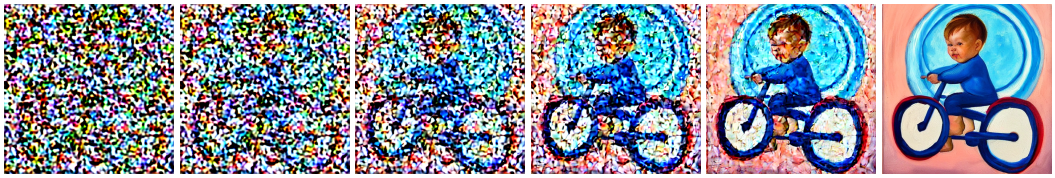
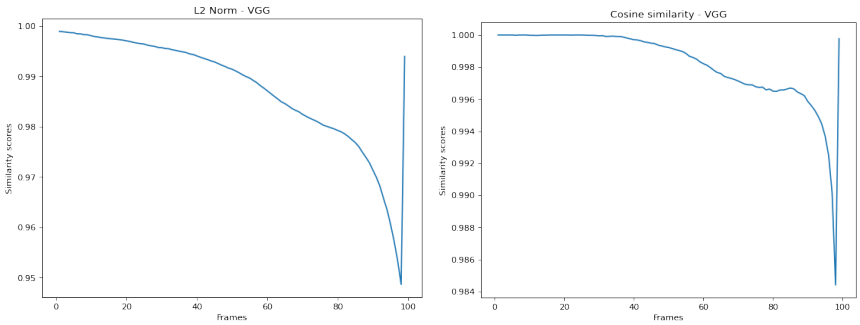
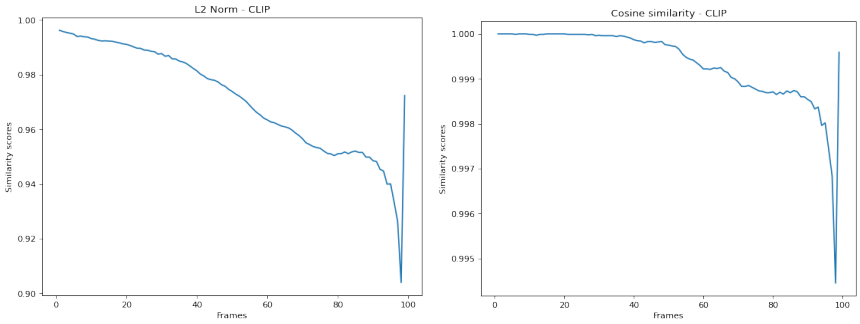


Figure 3: Intermediate results of a Stable Diffusion model for the prompt "A painting of a baby riding a blue bike." The model is run for 100 iterations. The result is shown for iterations 1, 20, 40, 60, 80, and 100.



(a) VGG16 model.



(b) CLIP model.

Figure 4: Plot of similarity between consecutive iterations of Stable Diffusion, averaged over 100 trials. The left plots show mean L_2 norm-based similarity scores while the right plots shows mean cosine similarity scores.

Figures 4 and 5 show the mean similarity plots between consecutive iterations of Stable Diffusion over 100 iterations. We average the data over 100 trials. Through these figures, we aim to better understand how Stable Diffusion generates images.

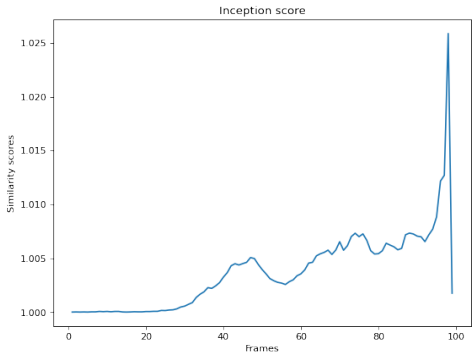


Figure 5: The plot shows the mean inception scores of the successive frames of the Stable Diffusion model run for 100 iterations.

3.4 Sketch-Guided T2I Results

We also test our method on the images from [7], shown in Table 5. The L_2 norm and cosine similarity scores for the images generated using a sketch and different prompts are quite similar to the results in Tables 2 and 3. However, their inception scores are higher compared to results in Table 4.

Table 5: Similarity metrics on images from [7].

VGG16 Feature Vector				CLIP Feature Vector				Inception Score	
L_2 Norm		Cosine similarity		L_2 Norm		Cosine similarity			
Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0.69	0.006	0.50	0.029	0.35	0.023	0.778	0.001	1.72	0.108

4 Conclusion

Our experiments show that, despite their diversity, the images generated by T2I models for a given prompt have very similar features. There is no significant difference in similarity scores between using the same prompt and slightly modifying the prompt for both Stable Diffusion and Dall-E2. Similar prompts result in a high similarity score between their corresponding generated images. Although the images generated by the same prompt differ greatly from those produced by adding random noise (Table 1), they have many similarities.

We also find that the $L2$ norm on the VGG16 feature vector and cosine similarity on the CLIP model's dense vector are useful metrics for comparing the generated images. They are extremely efficient in identifying images generated by similar prompts. Another effective method for identifying similar images without using their feature vectors is the inception score. When we look at the intermediate frames of the Stable Diffusion model, in the early iterations, the similarity between consecutive frames decreases. However, in the last few iterations, the similarity increases drastically.

References

[1] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

[2] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[4] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[5] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400, 2019.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[7] Andrey Voynov, Kfir Abernan, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. 2022.