Seeing Through the Facade: Understanding the Realism, Expressivity, and Limitations of Diffusion Models

Christopher Pondoc [*1] **Joseph C. O'Brien** [*1] **Joseph Guman** [*1]

Abstract

While text-to-image generation models such as DALLE-2 and Stable Diffusion 2.0 have captured the public psyche with the ability to create photorealistic images, just how "fake" are their outputs? To better understand this question, we present a three-prong process for extracting insights from diffusion models. First, we show strong results in classifying real vs. fake images by using transfer learning with a nearly decade-old model, setting an initial benchmark of realism not yet achieved. After visualizing the classifier's inference decisions, we conclude that concrete, singular subject objects – like buildings and hands – helped distinguish real from fake images. However, we found no consensus on which features were distinct to each of DALLE-2 and Stable Diffusion. Finally, after dissecting the prompts used to generate fake images, we found that prompts that failed to trick our classifier contained similar types of nouns while prompts that succeeded in this task differed for each model. We believe our work can serve as the first step in an iterative process that continuously establishes increasingly difficult benchmarks of realism for diffusion models to overcome. The code for our project is open source: https://github.com/cpondoc/diffusion-model-analysis.

1. Introduction

Recent research in natural language processing and computer vision has shown progress in image generation from text. Projects such as DALLE-2 (Ramesh et al., 2022) and Stable Diffusion 2.0 (Rombach et al., 2022) are examples of diffusion models trained on large language datasets. These models can create photorealistic images, transfer styles from one image to another, and process complex queries. For some, generative models amplify human creativity by serving as a tool that supercharges creative processes (Hoffman, 2022). However, many others feel that developments will only lead to more problems, such as troubling deepfakes, the replacement of human artists by artificial intelligence (AI), and disinformation that can be more quickly created and disseminated (Hancock & Bailenson, 2021; Wayne, 2022).

In the context of these potentially malicious applications, we explore the realism, expressivity, and limitations of diffusion models. Our process aims to answer three questions:

1. How can we **achieve high test accuracy on classifying between real and fake images**? What is our limit with a baseline convolutional neural network (CNN)? How can we leverage a well-established architecture out-of-the-box to push the bound higher?
2. When making classifications, **what features distinguish real images from AI-generated images, and DALLE-2 from Stable Diffusion 2.0**?
3. Finally, **which types of prompts generate the most photorealistic results and successfully trick the highest performing classifier**? **Which types of prompts are unsuccessful at this task**? Does this information connect back to the expressivity of each diffusion model?

With this approach, we hope to offer insights into the limitations of existing diffusion models and potential areas for improvement in photorealistic image generation.

2. Related Work

In the past, image generation research has been centrally focused on more traditional techniques, most notably generative adversarial networks (GANs) (Li et al., 2022; Marra et al., 2018). The majority of the work regarding GAN image identification has employed some variation of CNN architecture (Hulzebosch et al., 2020). A simple CNN has proven effective in this domain because the architecture of

[*]Equal contribution [1]Department of Computer Science, Stanford University, Stanford, CA. Correspondence to: Christopher Pondoc <clpondoc@stanford.edu>, Joseph C. O'Brien <jobrien3@stanford.edu>, Joseph Guman <joeytg@stanford.edu>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023.

a GAN consists of a discriminator half, which focuses on classification. Promising training methods for this approach have involved parity between an original image and a GAN-modified image (Jain et al., 2018). We translate this idea to diffusion models by creating a prompt-based dataset with thematic parity between real and AI-generated images.

As of late, several papers have emerged that focus on distinguishing between real and AI-generated images. Recent literature by Sha, et al. analyzed this task for diffusion models through two modalities: a visual modality and a linguistic modality (Sha et al., 2022). In the visual modality, the team used universal detection to first differentiate between real and fake images. Then, they used source attribution to assign the image to either DALLE-2Pytorch (an unofficial version of DALLE-2) or Stable Diffusion 1.0. In the linguistic modality, the team performed prompt analysis in a similar vein. Even more recent research pre-printed by Corvi, et al. adapted classifiers used for GANs (Corvi et al., 2022). Their specific niche was analyzing scenarios used in social networks, such as during resizing and compression.

Aside from taking inspiration from existing GAN-focused literature, there is no consensus "state-of-the-art" method for classifying real and fake diffusion model images. From the current discourse, we consider the scaffolded approach of both visual and linguistic methods to be strong, one which we embody in our own work. However, we believe that our contribution lies in making the limitations of diffusion models more concrete, whether through specific features of the generated images or the prompts fed into the networks.

3. Dataset

3.1. Using the Flickr30k Dataset

Our central focus when building the prompt-based dataset revolved around establishing parity between real and fake images. Without this equality, the potential for overfitting increases, as there is a higher possibility of our models focusing on the broader subject matter of such images. In turn, these models would conceivably learn to classify based on the overarching topic of an image rather than features associated with each of the two class types.

After experimenting with datasets such as Google's Open Images Dataset V7 – which utilizes concise image labels focused extensively on singular details – we decided to steer away from simpler prompts in favor of full-length sentence descriptions (Benenson & Ferrari, 2022). When fed into diffusion models, these labels' short words and phrases often produced images that were overly concentrated with minor details, unlike their expansive real-life counterparts. After continuing to play around with DALLE-2 and Stable Diffusion, we found that single-sentence prompts that included multiple key aspects created a more complete and accurate result. This led to our decision to use the Flickr30k dataset from the University of Illinois Urbana-Champaign (Young et al., 2014). This dataset contains a corpus of images from Flickr that are each assigned a sentence-long, human-generated description. Supplying a few examples of these descriptions manually to both diffusion models generated results that actually mirrored their real image counterparts, an outcome likely brought about thanks to the human involvement in the description creations.

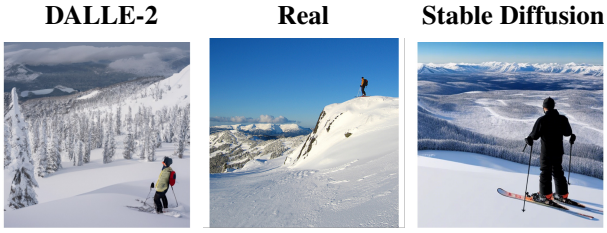

Figure 1. Images corresponding to the prompt: "A skier is overlooking the beautiful white snow covered landscape."

3.2. Generating the Dataset

In total, the Flickr30k dataset has 31782 images. With these images and descriptions in hand, we created our DALLE-2 and Stable Diffusion datasets. We generated 2884 total DALLE-2 images and 8609 total Stable Diffusion images (Rombach et al., 2022). We generated fewer DALLE-2 images due to both costs and OpenAI's content moderation policy, which rejected several of the prompts from Flickr30k (OpenAI, 2022). For both datasets, we only stored one AI-generated image for each real image to neutralize the effects of class imbalance.

4. Methods

4.1. Classifier #1: Baseline CNN

First, we made a two-convolutional-layer neural network. Our model works by twice running a convolutional layer to max pooling layer pair. Afterward, we flatten the features and run a two-layer ReLU neural network. We employed a simple model to gauge the simplicity of the task, effectively setting a "lower bound" on accuracy.

4.2. Classifier #2: ResNet-18

We then employed transfer learning with the 18-layer pre-trained model ResNet-18 (He et al., 2016). Transfer learning works by taking a model that has been trained on one dataset and finetuning the downstream parameters for a different dataset. In doing so, the upstream convolutional layers become trained to recognize features that will also be helpful in similar tasks. This method is particularly useful in low-data settings where training a high-variance CNN model from scratch would not be able to generalize well. Since

DALLE-2 images are expensive to create and there were no pre-existing fake image datasets with the same properties as the Flickr30k dataset, we experimented with applying transfer learning to overcome our low-data conditions. At first, we trained the model without loading in pre-trained weights, as we wanted to investigate the effect of a larger number of layers on classification accuracy independently. Then, we loaded in the pre-trained weights. Despite being a seven-year-old architecture, we specifically used ResNet-18 as the network was trained on real pictures of people with the ImageNet dataset, a task that we thought would be similar to our task of recognizing DALLE-2 images prompted to look like everyday scenes (Deng et al., 2009; He et al., 2016).

4.3. Image Pre-Processing

To reduce computational learning requirements, we converted all of our input images to grayscale, cropped the images around the center to 224 by 224 pixels, and normalized all images across each RGB channel.

4.4. Training Process

Using the Adam optimizer, we trained both models for a minimum of 8 epochs and a maximum of 15 epochs to give the network time to learn (Kingma & Ba, 2014). When updating our weights, we used mini-batch stochastic gradient descent, with batch size equal to 200 and learning rate $\alpha = 0.001$. We also added momentum to our mini-batch stochastic gradient descent, where $\mu = 0.9$. Finally, we used cross-entropy loss for both models to stay consistent with the ResNet-18 architecture. We defined convergence during training to be when the difference in the loss between two consecutive epochs was less than 0.005.

Due to the limited amount of DALLE-2 data compared to Stable Diffusion data – and fake data as a whole relative to real data – we wanted to analyze bias and variance to see if training on a low amount of data would inhibit our results. Thus, while training, we set aside 10% of data as a test set and experimented with using proportions of $50\%, 60\%, 70\%, 80\%$, and 90% of the remaining data for training. By plotting the graphs of both our training loss and accuracies for each proportion, we hoped to determine whether or not there was a sweet spot for the amount of training data or whether the network would generalize with only a small amount of data.

4.5. Tasks

First, we evaluated each model on the tasks of classifying between real vs. DALLE-2 images and real vs. Stable Diffusion images. Then, we generated heatmaps to better visualize the differences between real and fake images that our networks identified. After understanding what distinguished real images from fake images, we then sought to differentiate DALLE-2 images from Stable Diffusion images to understand the key features of each model's output. Finally, we analyzed which prompts produced more effective fake images.

5. Results

5.1. Results of Baseline CNN

Using our two-layer CNN, we saw relatively good results, achieving a test accuracy of 78% on the real vs. DALLE-2 task as well as 85% test accuracy on the real vs. Stable Diffusion task.

Overall, our baseline CNN seems to improve with more data. We did not observe a falling training accuracy as we increased the size of the training set. However, we observe unstable performance peaks and troughs, suggesting a high variance from training all model levels without the implicit bias from pre-training.

5.2. Employing Transfer Learning

We saw large gains with our transfer learning model: with no pre-trained weights, we achieved a test accuracy of 93.67% on the DALLE-2 vs. real task and 93.57% on the Stable Diffusion vs. real task. With the pre-trained weights, we came to a test accuracy of 97.5% on both the real vs. DALLE-2 and real vs. Stable Diffusion tasks.

We observe that transfer learning demonstrates stable behavior regardless of the amount of data during training, showing low variance in stability. Though upstream weights are not trained, we nevertheless predicted a low variance because the original task the ResNet-18 network was trained on – multi-class classification of real images – is similar to the task at hand. As such, we believe there is no need to acquire more data for DALLE-2 and Stable Diffusion to achieve better results.

Table 1. Results from Real vs. Fake Classification Tasks

CLASSIFICATION TASKS		
	DALLE-2	*Stable Diffusion*
2-LAYER CNN	78%	85%
RESNET-18	93.67%	93.57%
RESNET-18 + WEIGHTS	97.5%	97.5%

5.3. Visualizing Real vs. Fake

After our classification tasks, we turned to Smooth Gradient Based Class Activation Mapping++ (Smooth-GRAD CAM++) using the TorchCAM Python library (Fernandez, 2020). This method generates heatmaps to visualize what a given convolutional layer deems imperative for image classification (Omeiza et al., 2019).

After analyzing the heatmaps, we see the model generally

chose to hone in on a small subset of elements of the image. Coupled with our results from training ResNet-18, we thus can infer that diffusion models have a difficult time producing high-level generalizable features, which are features a CNN can easily detect. One primary observation was that the model would often identify hands when they were present in both the real and generated images to motivate its decision. Particularly surprising was the way that the model was able to detect a variety of objects. For instance, we found that the network was able to identify real animals from fake animals, such as birds and dogs. We also found that the model could focus on objects that were more distant, such as buildings, to make its judgment rather than exclusively focusing on closer ones. We believe this attention to specific objects and ability to recognize a wide variety of details comes from ResNet-18 being originally created as a general-purpose classifier for single object images.

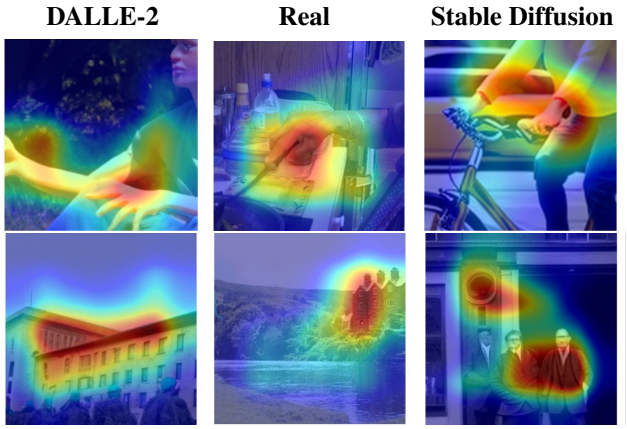

Figure 2. Heatmaps of DALLE-2, Flickr30k, and Stable Diffusion.

5.4. Distinguishing between Generative Models

Our heatmaps and high classification scores led us to hypothesize that neither diffusion model was expressive enough to generate realistic images. However, our network may have also been unable to pick up on distinct features. If untouched upstream weights are better at pulling features from the more expressive real images, then our current model would be evaluating the task of "real vs. diffusion model" rather than the tasks of "real vs. DALLE-2" or "real vs. Stable Diffusion." To investigate this hypothesis, we tasked our transfer learning model with classifying between DALLE-2 and Stable Diffusion. The purpose of this experiment was to see how well the model could pick up on the unique quirks between different model generations as well as the possibility of these features existing.

From performing this experiment, we found that the network achieved strong results: training on just 60% of DALLE-2 data led to an overall test accuracy of 94%. The generated heatmaps for each class remained relatively unchanged.

5.5. Linguistically Decomposing Prompts

The last step in our research process was to determine what distinguished an ineffective prompt from an effective prompt. We define "ineffective prompts" as prompts that created fake images that were correctly classified by the ReNet-18 transfer learning network. We define "effective prompts" as prompts that generate images misclassified as real by the same network. To do so, we calculated a confusion matrix and analyzed the prompts that led to both classifications. In doing so, we identified that the ResNet-18 model misclassified 9 DALLE-2 and 11 Stable Diffusion images while correctly classifying 280 DALLE-2 and 274 Stable Diffusion images.

When analyzing ineffective prompts, we first looked at the frequency of nouns within each. In doing so, we found three types of common prompt elements: people-related elements, clothing-related elements, and setting-related elements. Alongside our visual analysis of the heatmaps, such prompt elements suggest that these nouns are more likely to be malformed when generated by DALLE-2 and Stable Diffusion.

Table 2. Taxonomy of Ineffective Prompt Elements

PROMPT ELEMENTS	
Type	*Examples*
PEOPLE	"MAN", "WOMAN", "PEOPLE", "MEN"
CLOTHING	"SHIRT", "SHORTS", "JEANS", "T-SHIRT"
SETTING	"BACKGROUND", "STREET", "FIELD"

Then, we tried to identify effective prompt characteristics. After looking at the raw nouns in each misclassified prompt, there was no difference in the distribution of the type of subject: there were equally many nouns corresponding to people, animals, and buildings. Thus, we decided to focus on two metrics: the number of nouns and the total number of words in a prompt.

On the networks trained on the real vs. Stable Diffusion task, prompts misclassified as real had both a higher average number of nouns and a larger number of words in the prompt overall. Applying bootstrapping to our data, we derived corresponding p-values of 0.1544 and 0.2897, respectively. Both statistics also showed low variance. However, we saw a different trend with our real vs. DALLE-2 task. While the variance of both statistics remained small, the average number of nouns and words in the message was smaller than the average prompt with p-values 0.3108 and 0.6060, respectively. Though these values are not statistically significant given the size of each dataset, they suggest the possibility of an underlying difference in the composition of "effective prompts" for each of the two models.

When looking at the corresponding images, misclassified Stable Diffusion images were often malformed or busy. This suggests that DALLE-2 may be more expressive than Stable Diffusion. With shorter prompts, DALLE-2 was able to

create more in-depth singular subject images, effectively fooling the network.

6. Conclusion and Future Work

By leveraging CNN architecture and transfer learning to compensate for our data-restricted environment, we achieved a maximum test accuracy of 97.5% for both the real vs. DALLE-2 and the real vs. Stable Diffusion tasks.

Our high accuracy with an almost decades-old architecture suggests that while diffusion models have an incredible propensity for understanding language, both the human eye and our networks can distinguish between real and the most hyper-realistic fake outputs. We also showed that our networks identify concrete objects for each diffusion model during classification. From our prompt analysis, we see that the types of prompts that generate realistic images from each model are different. Overall, much work is left to be done on creating more expressive and general-purpose diffusion models.

Since diffusion models are trained by denoising images, we believe there might be quirks in the heuristics that these models learn. Rather than trying to perfectly reform an image from a prompt, it may be more reliable for the denoiser to simplify some details. This perspective suggests that the future of training diffusion models may involve integrating GAN techniques due to their established ability to generate photorealistic images.

In the future, we plan to expand our results by exploring different state-of-the-art classification architectures, such as vision transformers (Kolesnikov et al., 2021); newer diffusion models, such as Midjourney (Borji, 2023); and other prompt structure variations, such as shorter prompts. Moving forward, this work can act as the start for an iterative process for benchmarking the realism of diffusion models. Specifically, we can continuously use newer, more modern architectures as the diffusion models outperform a given classifier. As the exponential improvement of these models introduces the prospect of more widespread disinformation, we believe that our methods can help achieve the end goal of creating a single classification model that can classify any fake image from any real image.

References

Benenson, R. and Ferrari, V. From colouring-in to pointillism: revisiting semantic segmentation supervision. 2022.

Borji, A. Generated faces in the wild: Quantitative comparison of stable diffusion, midjourney and dall-e 2, 2023.

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., and Verdoliva, L. On the detection of synthetic images generated by diffusion models. 2022. doi: 10.48550/arXiv.2211.00680.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Fernandez, F.-G. Torchcam: class activation explorer. https://github.com/frgfm/torch-cam, March 2020.

Hancock, J. T. and Bailenson, J. N. The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking*, 24(3):149–152, 2021. doi: 10.1089/cyber.2021.29208.jth. PMID: 33760669.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Hoffman, R. Blitzscaling creativity with dall-e, Jul 2022. URL https://www.linkedin.com/pulse/blitzscaling-creativity-dall-e-reid-hoffman/.

Hulzebosch, N., Ibrahimi, S., and Worring, M. Detecting cnn-generated facial images in real-world scenarios. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2729–2738, 2020. doi: 10.1109/CVPRW50498.2020.00329.

Jain, A., Singh, R., and Vatsa, M. On detecting gans and retouching based synthetic alterations. pp. 1–7, 10 2018. doi: 10.1109/BTAS.2018.8698545.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Kolesnikov, A., Dosovitskiy, A., Weissenborn, D., Heigold, G., Uszkoreit, J., Beyer, L., Minderer, M., Dehghani, M., Houlsby, N., Gelly, S., Unterthiner, T., and Zhai, X. An image is worth 16x16 words: Transformers for image recognition at scale. 2021.

Li, Z., Yu, N., Salem, A., Backes, M., Fritz, M., and Zhang, Y. Unganable: Defending against gan-based face manipulation. 2022. doi: 10.48550/arXiv.2210.00957.

Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 384–389, 2018. doi: 10.1109/MIPR.2018.00084.

Omeiza, D., Speakman, S., Cintas, C., and Weldermariam, K. Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. 2019. doi: 10.48550/arXiv.1908.01224.

OpenAI. Usage policies, Nov 2022. URL https://beta.openai.com/docs/usage-policies/content-policy.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. 2022. doi: 10.48550/arXiv.2204.06125.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Sha, Z., Li, Z., Yu, N., and Zhang, Y. De-fake: Detection and attribution of fake images generated by text-to-image diffusion models. 2022. doi: 10.48550/arXiv.2210.06998.

Wayne, E. Will ai replace human artists?, Aug 2022. URL https://artofericwayne.com/2022/05/31/will-ai-replace-human-artists/.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166.