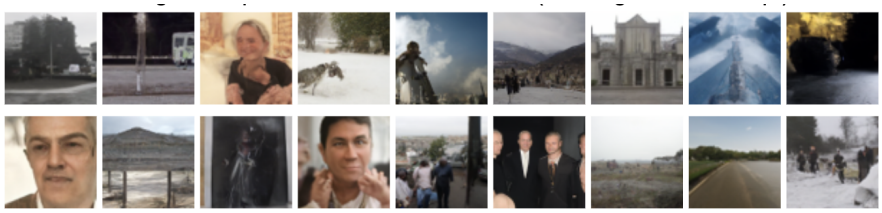
Thumbnail Generation for Online News Content

Isaiah Williams
Department of Computer Science
Stanford University
byronw@stanford.edu



1 Abstract

In this paper, we present a novel pipeline for image generation in the task of thumbnail creation. We also create a newly fine-tuned diffusion model on the VisualNews2021 dataset, and attach CLIP guidance to the 64x64 diffusion model architecture. We run two experiments where we first show that our fine-tuned model outperforms a baseline class-conditional diffusion model in generating news article related images using relative FID measurements. In a 2nd riskier experiment, we attach a CLIP model to our fine-tuned diffusion model for text prompt image generation guidance. Instead of manually enetering the text prompt, we use a BART Neural Network to summarize the article text into one sentence and submit that as our text prompt. Using ground truth captions as an upper bound, we show that This BART -> Diffusion + CLIP pipeline is a powerful and promising architecture for automated thumbnail image generation.

2 Introduction

The adage goes “A picture is worth a thousand words.” In this paper, we demonstrate the beginnings to a new model pipeline that realizes that adage by generating news article thumbnails for articles that have no images. This is important because images act as a catalyst on whether an article is read or not. Previously, a writer would have to either collect relevant images for their post or find one that is tangentially related. With this Diffusion based model, an image could be generated automatically and capture the meaning and sentiment of the article, decreasing the work required and increasing the conversion rate of readers.

A diffusion model has been adapted from thermodynamics. It is trained to add noise to real images until they are completely noisy samples. Then reverse that noisy sample back into the real image.

CS236G: Generative Adversarial Networks, Winter 2022, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

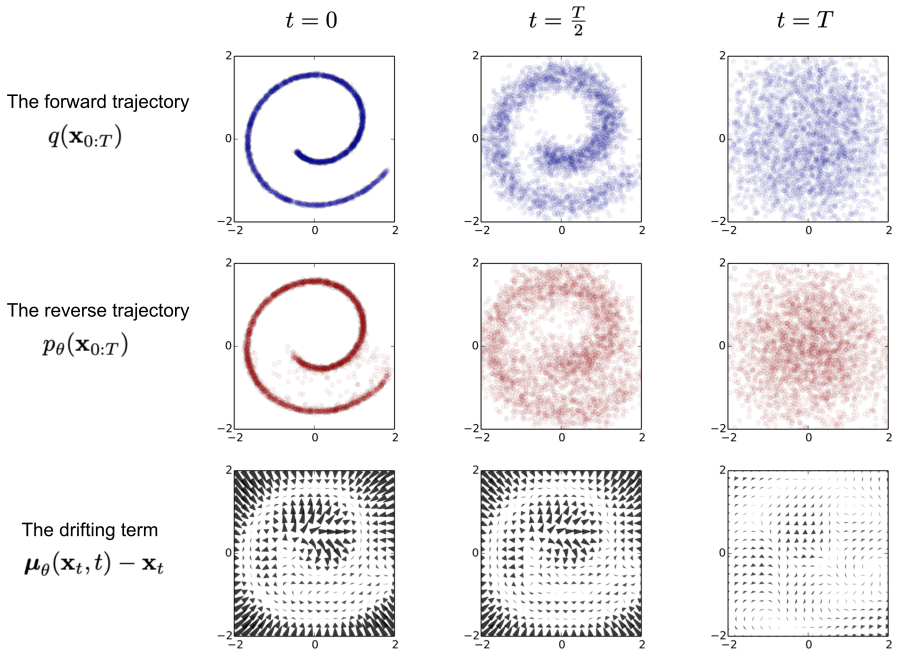


Figure 1. Forward and Reverse Trajectory [1]

3 Background

Denoising diffusion probabilistic models (DDPMs) is a recent development in deep learning that has shown promising results on a variety of datasets. DDPMs are based on the idea of diffusing noise through a latent space in order to learn image patterns. This is done by training models to learn a lossy decompression of gaussian noise in order to generate images. Ho et al. [2] was the first to demonstrate the use of DDPMs and achieved a new state-of-the-art on the CIFAR10 dataset. OpenAI later improved on this work by augmenting the architecture for speed and reasonable computation. [3] By learning the variances in the reverse diffusion process, they were able to drastically reduce the number of steps necessary to achieve similar log-likelihoods of the previous models.

OpenAI's paper "Diffusion Models Beat GANs on Image Synthesis" [4] shifted the focus of image generation from GANs models to diffusion models in the image generation literature. The diffusion model architecture was supplemented with increased depth versus width, more attention heads, and different attention resolutions. This architecture alone outperformed the state of the art GANs models in FID score on multiple datasets. The paper also showed that these models could be guided by the gradient of a classifier trading off diversity for fidelity. This classifier guidance alongside a pipeline for image upsampling produced an image generation model that greatly improved the state of the art that was previously held by BigGAN

When it comes to practical applications of image synthesis, fidelity is often much more valued than diversity. This naturally led to a focus on the improvement of classifier guidance of DDPMs. OpenAI's CLIP model [5] in the paper Learning Transferable Visual Models From Natural Language Supervision proved to be more effective at creating vector representations of text prompts for said guidance. Much of this accuracy is due to multi-class prediction capabilities of the model and its massive training regime of 400 million image/text pairs. Using CLIP as the classifier in the guided diffusion process produces almost indistinguishable from reality image generation. The architecture in this paper utilizes the diffusion + CLIP architecture.

In this paper, we explore a neural network architecture termed BART. [6] This framework utilizes the Transformer architecture and has been demonstrated to effectively generate summarized text. Since much of the diversity of the image is controlled by the inputted text prompt, much focus has been

given to understanding and synthesizing the best prompt for image generation. In this paper, the neural network BART is proposed to generate such a text prompt. Lewis et al introduced BART which is a denoising autoencoder that uses the standard Transformer architecture. BART is particularly effective at text generation which can be specified to text summarization in our context. BART operates by taking an input text and generating a much shorter summarization of that text.

4 Dataset

For this experiment, we are using the Visual News 2021 dataset [7]. The dataset consists of 1 million images from news articles, along with its corresponding article text and auto-labeled short image-captions generated from an encoder decoder model. This dataset is one of the most expansive datasets of its type in both size and diversity. It covers a wide range of diverse news articles coming from BBC, the Guardian, NewYork Times, and Washington Post. For my experiment we used 100,000 image to train the baseline model. The final model will use 10,000 images as a test set.

To obtain the data we contacted the creator of the repo at fl3es@virginia.edu and she emailed me the link to the repo which contained a 100 gb .tar file. We downloaded that file and unpackaged it over a 24 hour period. We then used the data.json file to locate the position of 10,000 images and placed them in a separate folder.
After exploring the dataset, it was observed that the images contained no consistency in size, and some of the images were corrupt. We filtered out all images that could not be opened from our train/test set and also center-square-cropped all of the images. After this, another pass over the train/test set was made to reduce the pixel dimensionality to 64x64 to align with our diffusion model. The small image size made it faster and practical to transfer all of the images from the local machine to the cloud and made paying for the compute needed to to sufficiently train the model more practical.

5 Methods

Diffusion models work by sampling a random point x_0 from a real distribution (a.k.a. picking a real and random photo). Guassian noise is added to this sample in $\boldsymbol{T}$ steps producing noisy samples $x_0, x_1, ..., x_T$. The noise is controlled by a scheduler Beta $\{\beta_t \in (0,1)\}_{t=1}^{t}$. [8]

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

To train our model, we have taken the pretrained weights of a 64x64 image diffusion model and fine-tuned it with 100,000 images from our data set as described in the dataset section. These images were pre-processed by centering and square cropping and then the dimensionality was reduced to 64x64. Fine-tuning took place on a G4dn-xlarge AWS EC2 instance over the course of 36 hours. We then run the training regime using our data learning the parameters to reverse random noise into an image. After training, we sample our model randomly to run experiment 1. Pseudocode for that process can be seen below:

Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Take gradient descent step on
    $\nabla_\theta \left\| \boldsymbol{\epsilon} - \mathbf{z}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$
until converged

Algorithm 2 Sampling

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \mathbf{z}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
end for
return $\mathbf{x}_0$

Figure 2. Training and Sampling Pseudo-code [2]

The following table shows hyper parameters used in this run:

Image Size	Num Channels	Num Res Blocks	Diffusion Steps	Noise Schedule	Learning Rate	Batch Size	Learn Sigma
64	128	3	4000	Linear	1e-4	32	True

Parameters for Training Routine. [3]

Our implementation uses BART to summarize the entire document into one sentence. Our hypothesis is that this single sentence summarization will encapsulate the most important point of the document. The output of BART will then be fed into the diffusion model as a text prompt input. As a upper-bound comparison, we will use the labeled captions associated with the summarize article.

Full Text	Summary
For four long years members of the US Congress had to smile or scowl as a TV star played the role of president. Donald Trump became infamous for the art of lying. On Wednesday another TV performer turned national leader came before Congress. But this one captivated his viewers with truth telling. The Ukrainian president, Volodymyr Zelenskiy, a former actor and ….	Volodymyr Zelenskiy has been on a virtual tour of western capitals over the past three weeks, tailoring his speeches to each nation.

Example Summary

In our dataset *VisualNews2021*, both the article text and an image caption were provided alongside the ground truth image. With this in mind, we test both the BART summarization of the article text as well as the image caption. Intuitively, image captions will be the most accurate description of the image as they are essentially human-labeled image descriptions. For the sake of this experiment, we can use the image captions as an upper bound for CLIP prompt input. The goal of the experiment then becomes to get CLIP guided generation with BART summaries to be as close as possible to that of the CLIP guided generation with image captions.
So we selected 112 different article/caption/image triples from our test set. The article was fed into BART for a generated summary. We then used each of the article summaries as input into CLIP text prompt.

To use FID as a metric of measurement, we must create a distribution of images as ground truth and a distribution of images as the test set. In our first experiment, we created the ground truth set by using our entire test set. This was then compared to 128 randomly sampled generated images from our fine-tuned model. In the second experiment, utilizing FID effectively proved to be a bit more challenging. Our goal is to compare the quality of images produced given 112 distinct article summaries. To reduce noise, we generate 4 images per summary prompt resulting in 448 images. We then take those 448 images as a distribution and compare that to the distribution of the 112 images from the ground-truth.

6 Experiments

Two experiments were ran. The first experiment aims to determine the effectiveness of the fine-tuned model in comparison to a baseline model on the specific data domain. The second experiment aims to determine the effectiveness of the BART summarized input prompt compared to the true caption of the image. In both experiments, FID is used as the metric to measure similarity between generated images and ground-truth images. However, the way FID is used is quite different from experiment 1 to experiment 2. This will be explored in more detail in their respective subsections.

6.1 Experiment 1

The goal of this experiment is to analyze the effectiveness of our fine-tuned model within the domain of news-article thumbnail generation. To calculate FID, we need to randomly sample our fine-tuned model. Calculating the FID score for our model with respect to the ground-truth labels is an obvious demonstration of accuracy, however, given that we have no formalized benchmark for thumbnail generation on this dataset, we need some baseline FID to compare our model's FID score with. To

achieve this, a 64x64 class-conditioned diffusion model was taken from the same Improved Diffusion paper as our pretrained weights were derived from and used as our baseline model.
We calculated the FID of our current fine-tuned model against the 10,000 samples in our test set. We compared this to the FID of the 64x64 conditional baseline model against the 10,000 samples in our test set. To see examples, see images attached at the end of this document.* The FID of the baseline model was 216.109 and the FID of our model was 65.476.

To calculate these FID scores, 128 samples were taken at random from each model. These 128 samples were compared to 10,000 samples from our test set. From the images, we notice that the fine-tuned samples are more on topic with images that tend to come from news articles. However, the images from the base model look to have higher fidelity. From the FID score, we can see that the fine-tuning of the model greatly improves its ability to generate new-article-like images in general.

6.2 Experiment 2

This experiment aimed to determine the effectiveness of using BART-generated summaries as input prompts for the CLIP-guided diffusion process. As discussed in the methods section, we used 4 distinct configurations of (model, prompt method). For the sake of time, these models were respaced to 250 steps. This decreases the quality, but increased the amount of different samples we could generate. For each of the 112 different prompts, 4 images with differing intial noise were generated resulting in 448 images per configuration. The following table shows the FID scores calculated for each configuration:

Config	Models	CLIP Text Prompt	FID
#1	Our Fine Tuned Model	Caption	175.018
#2	Our Fine Tuned Model	BART Summary	179.106
#3	Base Model	Caption	311.190
#4	Base Model	BART Summary	307.8334

As expected, the fine-tuned models perform significantly better than the base model. We also see that the BART summary model performs very close to the Caption model in FID. At the end of this document, you can view some of the generated images for the fine-tuned models.

7 Analysis

The results from experiment 1 are as expected. Our fine-tuned unconditional model greatly outperforms the class-conditional model on the test set. Our model has learned to produce news-like thumbnail images through training. The class conditional model, while perhaps appearing more realistic due to its class conditioning, is not capable of producing the type of images commonly found on news articles. We can safely proceed with our fine-tuned model as an effective model for this diffusion architecture.

The results from experiment suggest that our hypothesis is correct. BART summaries of news articles act as a substantial alternative to human-labeled image captions when it comes to generating appropriate news article thumbnail images. With an FID score difference of only 4, the quality of produced images is very similar. What is interesting, however, is that through subjective analysis we can see the concept of the images is not particularly the same. This makes sense because often the

ground truth image and as a result the caption are not necessarily what the article is about. Yet, the FID score is still so close to that of the caption image. This suggest that the article summaries still remain in phase with the type of images used as news article images. This is significant because in practical application, most often there will not be a particular caption that is requested from image generation. With BART summarization we can take advantage of the implicit signal that is the article text to generate a strong prompt for CLIP guidance.

Our Base Models in the 2nd experiment are the checkpoint weights from the improved diffusion architecture [3]. Unlike the fine-tuned models, these models have not been trained on any images within the news article domain. It is suspicious that the FID score for the summarized prompt is higher than that of the caption prompt using the base model. Upon further examination of the images generated, many of the images resulted in a style as seen in the figure below

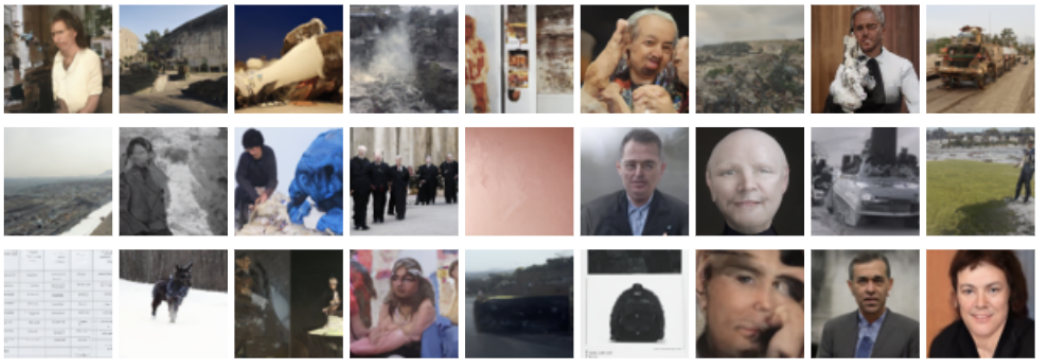
Figure 3. Suspicious Images from the Base Model. [3]

It seems as if the base model weights are either inaccurate as listed in the Improved Diffusion github repo, or that the weights as is are not compatible with CLIP guidance in the text domain of news articles. Further exploration and debugging must be done to remedy this mysterious image generation

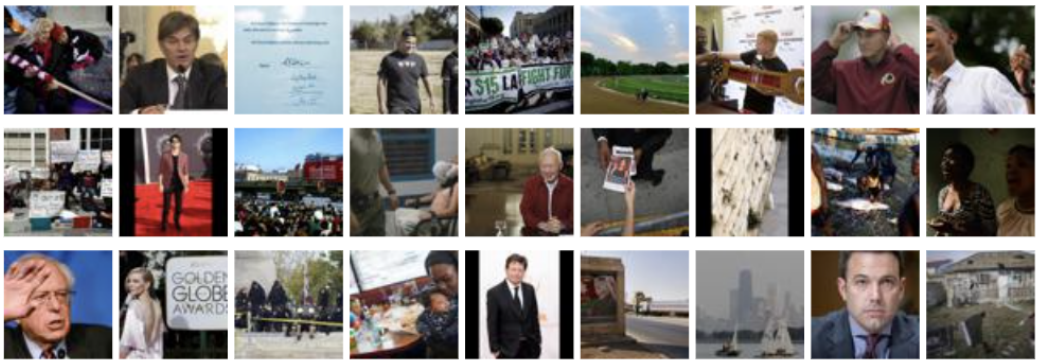
8 Conclusion

In this paper, we have presented a novel way to generate news article thumbnails given nothing more than the articles text. Our primary hypothesis was that BART summaries of the article would prove to be a satisfactory substitution for image captions. The very similar calculated FID scores suggest that this substitution is satisfactory. On top of this, we produced a 64 by 64 dimension diffusion model fine-tuned on one of the latest and largest image/caption datasets available. This fine-tuned model proves to be effective at generating new article themed images from both of quantitative and qualitative perspective.
To get this neural network pipeline ready for practical applications, we plan to train the same model for much longer (on the scale of two weeks instead of two days) and on the entire dataset (a million images vs a hundred thousand). Given the surprisingly realistic results of other diffusion models [4] it seems likely that extremely realistic image generations can be achieved under these new conditions. Also, image resolution can be increased to 256x256 or 512x512 in future iterations for use in other applications besides thumbnails.

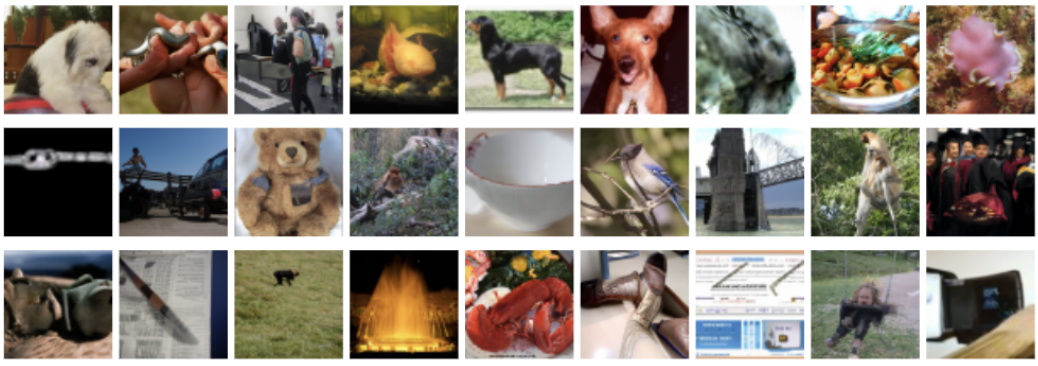
Generated Images sampled from Fine-tuned model:



Real Images Sampled from Test Set:

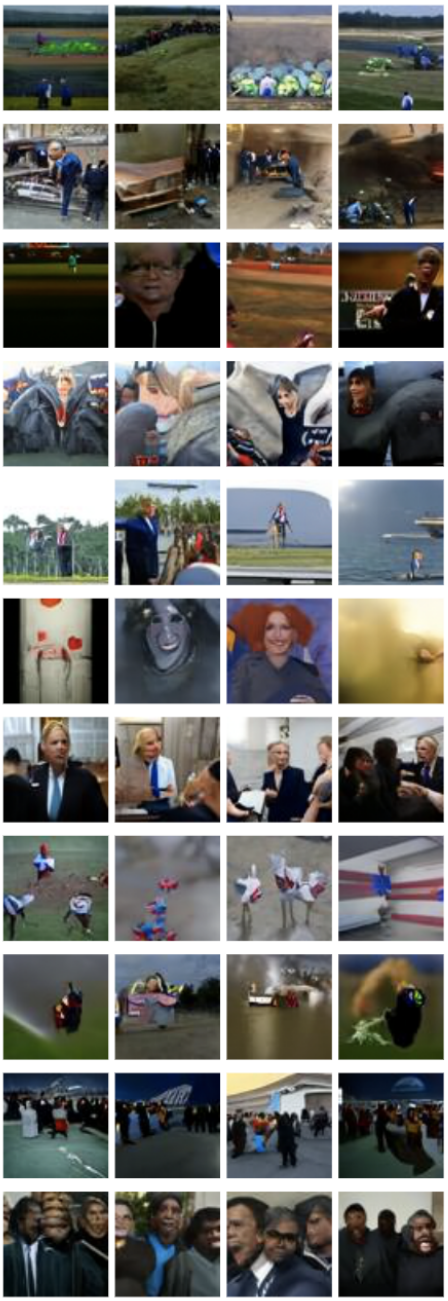


Generated Images from Baseline Conditional 64x64 model

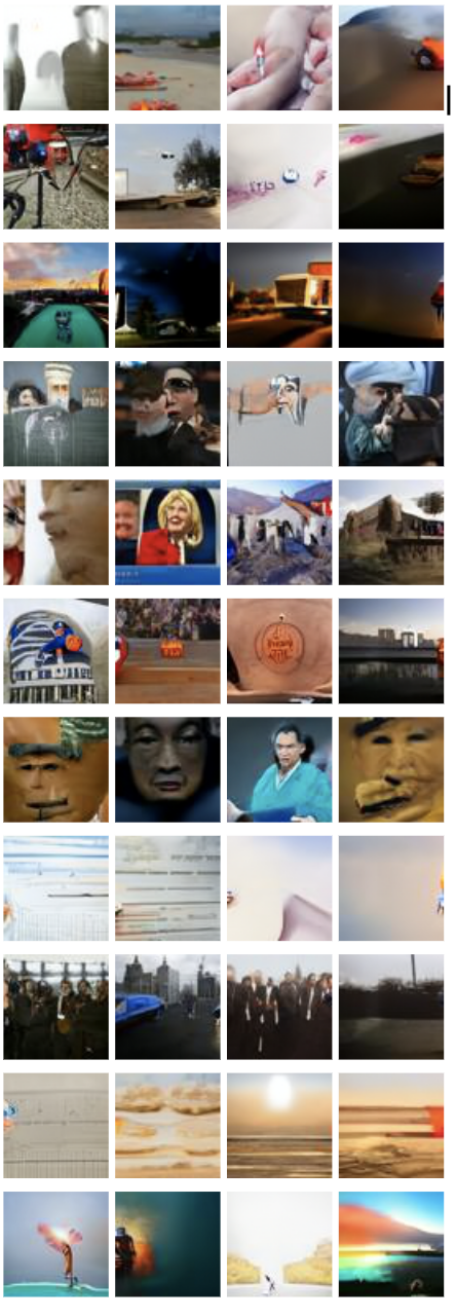


4 Samples of Generated Images from Finetuned Model with Captions vs BART Summary

Captions



Bart Summary



References

[1] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2256–2265, PMLR, 07–09 Jul 2015.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *CoRR*, vol. abs/2006.11239, 2020.

[3] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *CoRR*, vol. abs/2102.09672, 2021.

[4] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *CoRR*, vol. abs/2105.05233, 2021.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *CoRR*, vol. abs/1910.13461, 2019.

[7] F. Liu, Y. Wang, T. Wang, and V. Ordonez, "Visualnews : Benchmark and challenges in entity-aware image captioning," 2020.

[8] L. Weng, "What are diffusion models?," *lilianweng.github.io*, 2021.