Milestone 2

Generating Thumbnails for Text Articles Isaiah Williams

Introduction

The adage goes “A picture is worth a thousand words.” In this project, we hope to realize 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.

Related Works

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. [3] 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. [5] 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 [2]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. [1] This framework utilizes the Transformer architecture [3] 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.

Approach

My initial approach was simply to train a diffusion model based off of OpenAI's Improved Diffusion model (https://github.com/openai/improved-diffusion) that reproduces the model references in the landmark paper *Improved Denoising Diffusion Probabilistic Models.* This paper and our initial model uses a Denoising diffusion probabilistic model (DDPM). These models take an image and learn a series of operations that converts the image into Gaussian noise. This noise can be reversed using the learned operations to generate a similar image structure from a randomly sampled point in the Guassian noise. This way we can create different varieties of the same image. Instead, pre-trained weights from a 64x64 unconditional diffusion model were used and fine-tuned upon for our experiment. This is appropriate because many of the features learned from OpenAI's initial dataset will also be present in our dataset.

To utilize the 64x64 diffusion model, we reduced the size of the images in our dataset by first cropping each image into a square base upon its shortest side and then reducing the dimensionality to 64x64 pixels. A range of losses are used in the model most notably MSE loss. After fine-tuning, we are able to sample our model for images and compare them to the real images of our test set. Using this, we can calculate an FID score for our models as a metric of comparison.

Dataset

For my project, I am using the Visual News 2021 dataset obtained from https://github.com/FuxiaoLiu/VisualNews-Repository. 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 I took a slice of 10,000 images to train the baseline model. The final model will use 10,000 images and 10,000 to validate and test on.

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. This created my initial training set.

Method

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. On top of this we have taken the pretrained weights of a 64x64 image diffusion model and fine-tuned it with 10,000 images from our data set as described in the dataset section. These images were pre-processed by centering and square cropping and then reducing the dimensionality to 64x64. Fine-tuning took place on a G4dn-xlarge AWS EC2 instance over the course of ~18 hours.

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

As a baseline comparison, we will test our model against the non fine-tuned diffusion model with both the entire document as input prompt and our BART generated sentence as input prompt.

Experiments

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 base model against the 10,000 samples in our test set. To see examples, see images attached at the end of this document.* The following is our measured FID scores:

FID (Our Fine-tuned Model):	65.47587615584553
FID (Base Model):	216.10920810225895

To calculate these FID scores, 1024 samples were taken at random from each model. These 1024 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.

Future Experiments

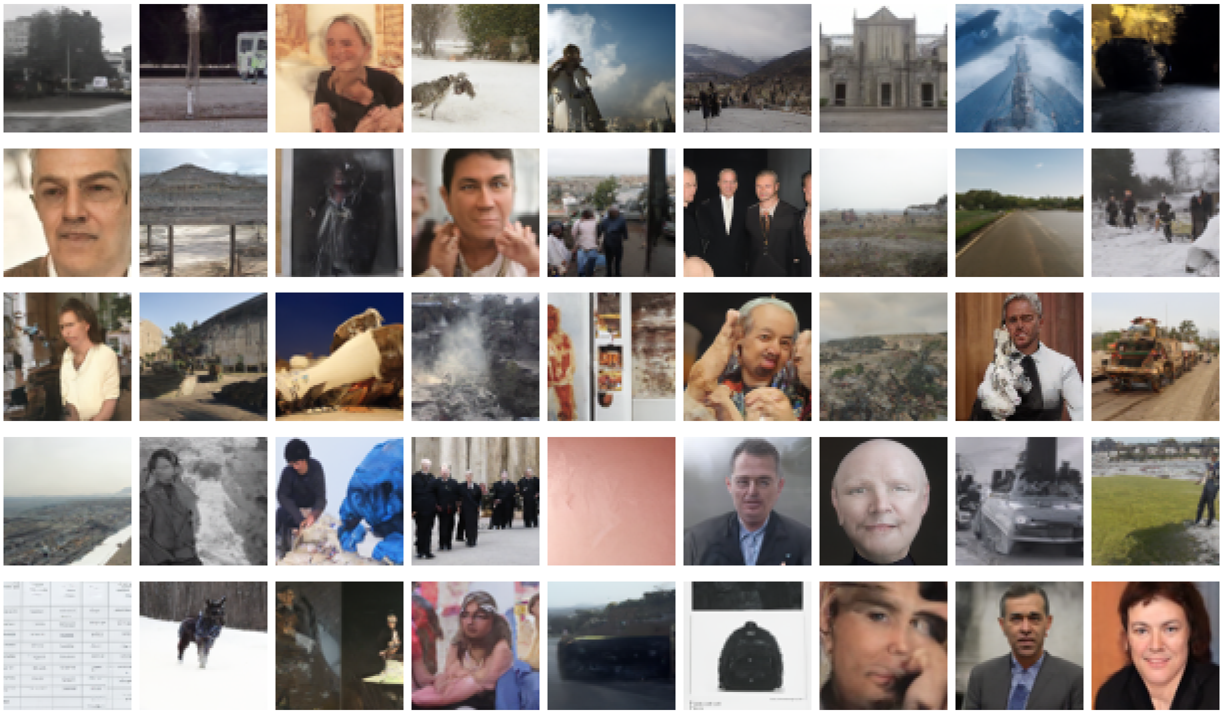
We are going to run 6 experiments:

Experiments	**Models**	**CLIP Text Prompt**
#1	Our Fine Tuned Model	BART Summary Sentence
#2	Our Fine Tuned Model	Article Title
#3	Our Fine Tuned Model	Entire Document
#4	Base Model	BART Summary Sentence
#5	Base Model	Article Title
#6	Base Model	Entire Document

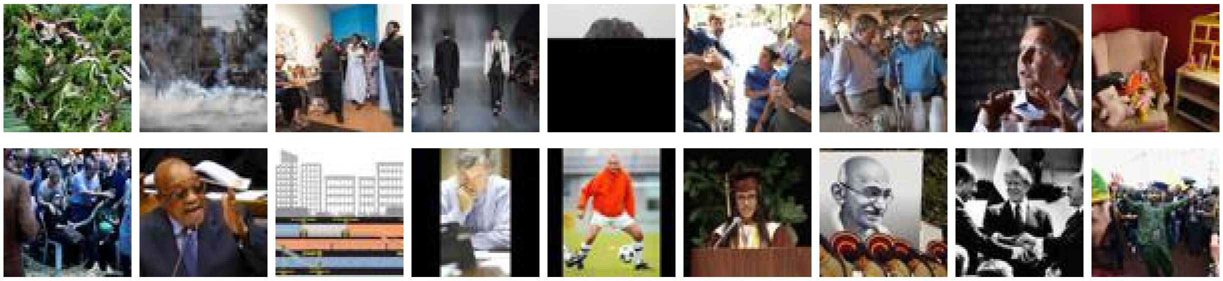
We will then compute the FID of each of these models with 16 samples versus the 1 ground truth image caption in our test set. We will repeat this for 100 different Image caption pairs. We are currently in the process of attaching the gradient signal of the CLIP model to our diffusion model. This component is critical in running our experiments. The results of CLIP guided diffusion will be shown in the final paper. We have successfully used CLIP guidance on our unconditional diffusion model, however, for some reason, after 100 steps, the generated image goes from recognizable to very noisy. We think this may have something to do with the noise scheduling along with the small 64x64 image size we are using in the experiment. Regardless, the results of the CLIP guided diffusion model can be shown relative to one another for proof of concept.

Also, in the final paper, we will analyze the actual title, text, and one sentence summary of the article for use as input to the text prompt. We currently have a BART model ready to be attached to our pipeline once CLIP guidance is functional.

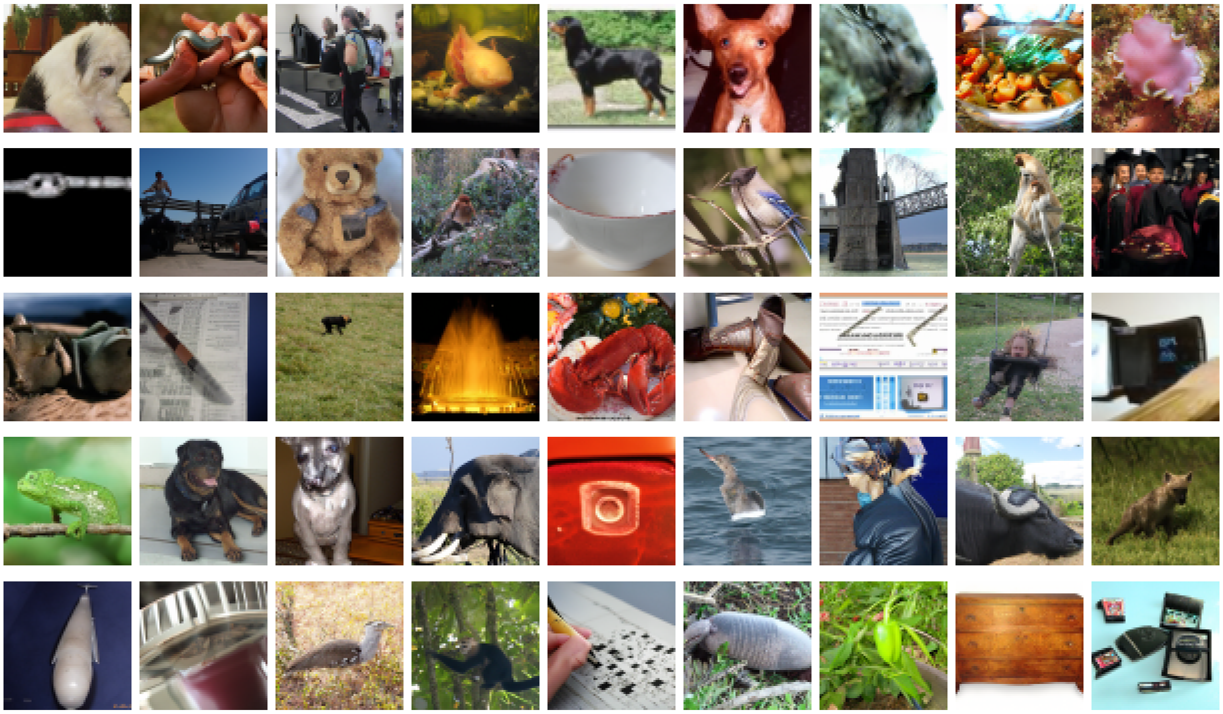
Generated Images sampled from Fine-tuned model: (45 Images - No Prompt)



Real Images Sampled from Test Set: (45 Images)



Generated Images from Base 64x64 model(45 images - No Prompt)



Citations

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

[2]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 fromnatural language supervision,”CoRR, vol. abs/2103.00020, 2021.

[3]J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,”CoRR,vol. abs/2006.11239, 2020.

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

[5]A. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,”CoRR,vol. abs/2102.09672, 2021