Synthetic Data From Diffusion Models Improves ImageNet Classification

https://arxiv.org/pdf/2304.08466

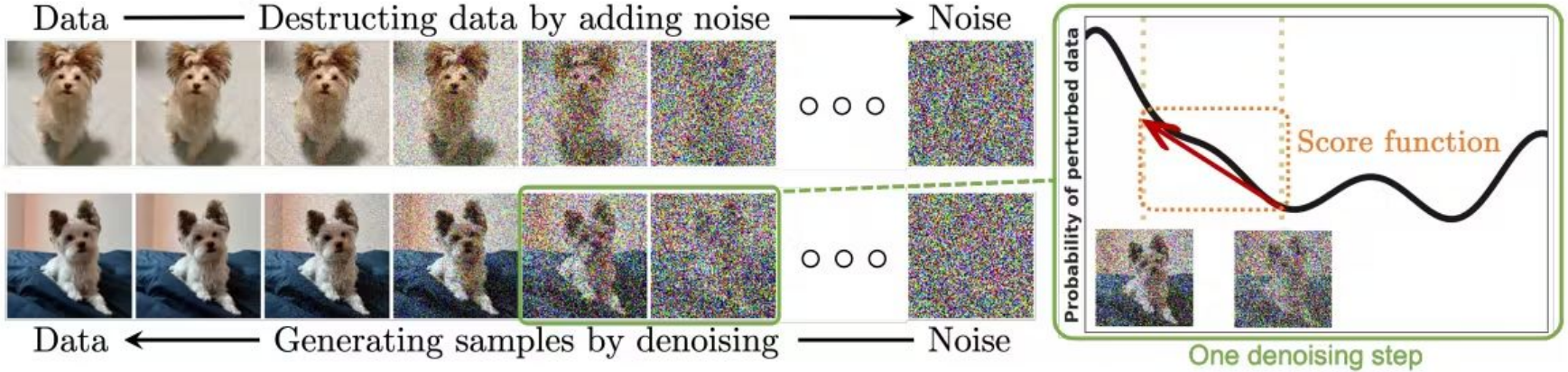
Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, David J. Fleet

Presented by: Alex Zeng & Isaac Picov

Background

Definitions

Denoising Diffusion Probabilistic Models (DDPMs): Diffusion models create images by learning to remove blur from noise, training on images blurred to pure noise, then reversing the process to generate new images by "unblurring" noise to match desired features.



Definitions

- **Class Conditional Model:** A model trained on several classes of labelled data, when given an input matching the label it outputs similar data.

- **Fréchet Inception Distance (FID):**Measures how close the generated images are to real images by comparing their distributions; a lower score means the generated images are more similar to real ones.

- **Inception Score:** Measures both quality and diversity of images that are generated. A higher score means better quality images over a diverse range of classes

Definitions

- **Synthetic Data:** Artificially generated or processed data, unlike authentic data which was collected in the real-world applications.

- **Generative Data Augmentation:** Uses synthetic data to expand existing datasets

- **Representation Learning:** A process in where the model automatically learns important features automatically, without manual input. Makes it useful for downstream tasks, which is when another model uses the learned features from a previous model. Similar to transfer learning.

Research Questions:

- Can diffusion models generate images of high enough quality to improve classification on benchmark datasets?

- Can text-to-image models enhance representation learning and be used in downstream tasks?

Methods

Process

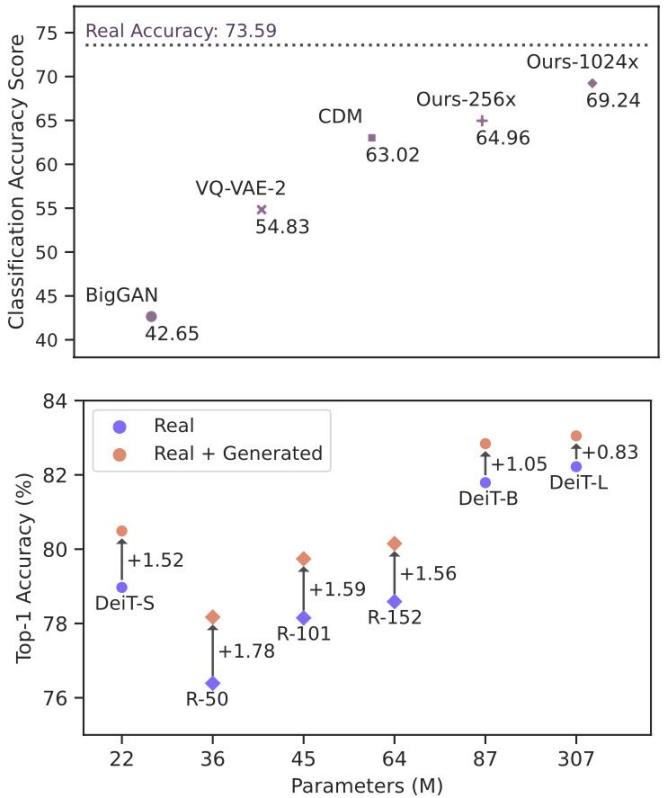
- **<u>Generative Data Augmentation</u>:** Researchers **fine-tuned diffusion models** to achieve state-of-the-art (**SOTA**) class-conditional image generation on the **ImageNet dataset**.

- **<u>High-Resolution Image Quality</u>:** These fine-tuned models reached **SOTA image quality across multiple resolutions**, with scores of **FID 1.76 and IS 239 on 256x256 images**.

- **<u>Evaluation on Classification Accuracy</u>:** By **training ResNet-50 models** on generated images, they demonstrated **SOTA classification accuracy, even when tested on real data**, highlighting the effectiveness of synthetic data.

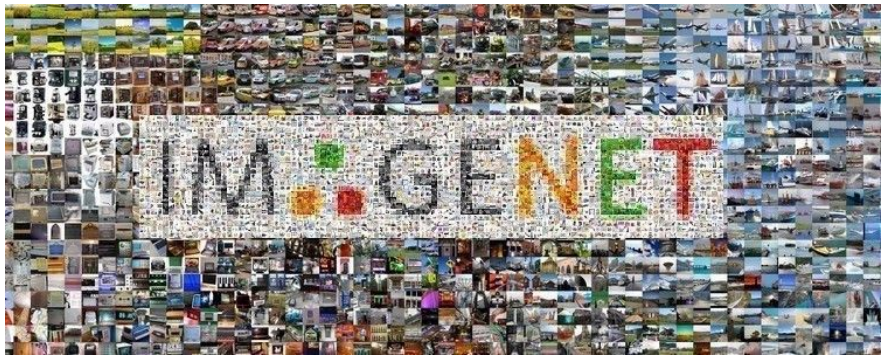
Process

- **Real Accuracy**: baseline accuracy trained on labelled real data, as compared to generated or synthetic data
- **Top-1 Accuracy**: the accuracy of highest-ranked prediction made by a model
- **Classification Accuracy Score**: checks if the top prediction is correct



Fine-Tuning

- **Trained on Imagenet 1K,**
  - Contains **1000 labeled classes** and **1.2 million** training images
  - Evaluated at **64x64**, **256x256**, and **1024x1024**
- **Text to Image Generation Approach**
  - Used a **text-to-image generator**, with **text prompts as input** and **images as output**
  - Used **1-2 word class descriptors** and f**ine-tuned model weights and parameters** for greater precision



Fine Tuning

Real vs. AI?

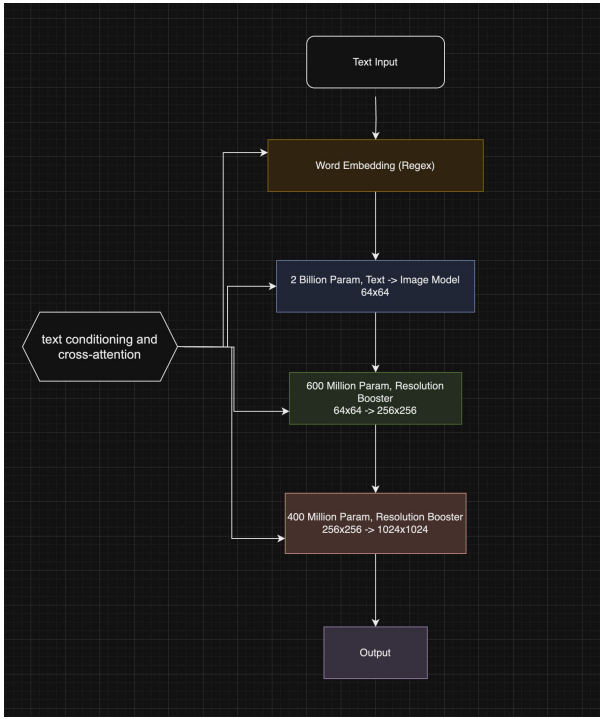
Fine-Tuning

Limited High-Resolution Data: Due to the scarcity of 1024x1024 images in ImageNet, only the f**irst two layers were fine-tuned**, while the **word embedding and final layer were left unchanged**.

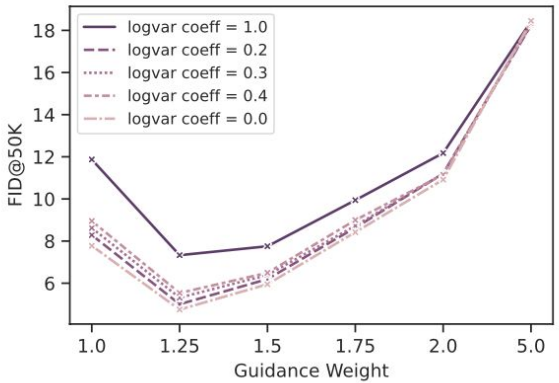
Training Steps: The first layer was **trained over 210,000 steps** and the **second over 490,000 steps**, each with a **batch size of 2048 images**.

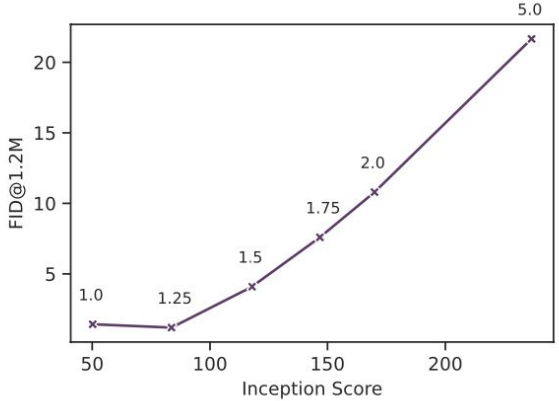
Optimizers Used: Adafactor, a **more computationally efficient optimizer than ADAM**, was used for the first layer; **ADAM was used for the second layer for better performance**.

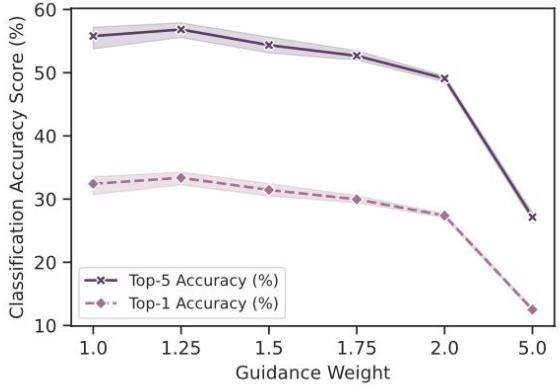
Model Pipeline



Fine Tuning







- **Guidance Weight:** Similar to temperature, determines how faithful the generated outputs are to the training images. The higher the guidance weight the more similar however at the cost of image diversity.
- **Logvar Coeff:** Determines how much noise we take away, in each denoising step.
- **Denoising Steps**: Amount of times we denoise.

Logvar Coeff: 0 & 0.1, Guidance Weight: 1.25, Denoising Steps: 1000

Results

Fine-Tuned Imagen Sample Quality

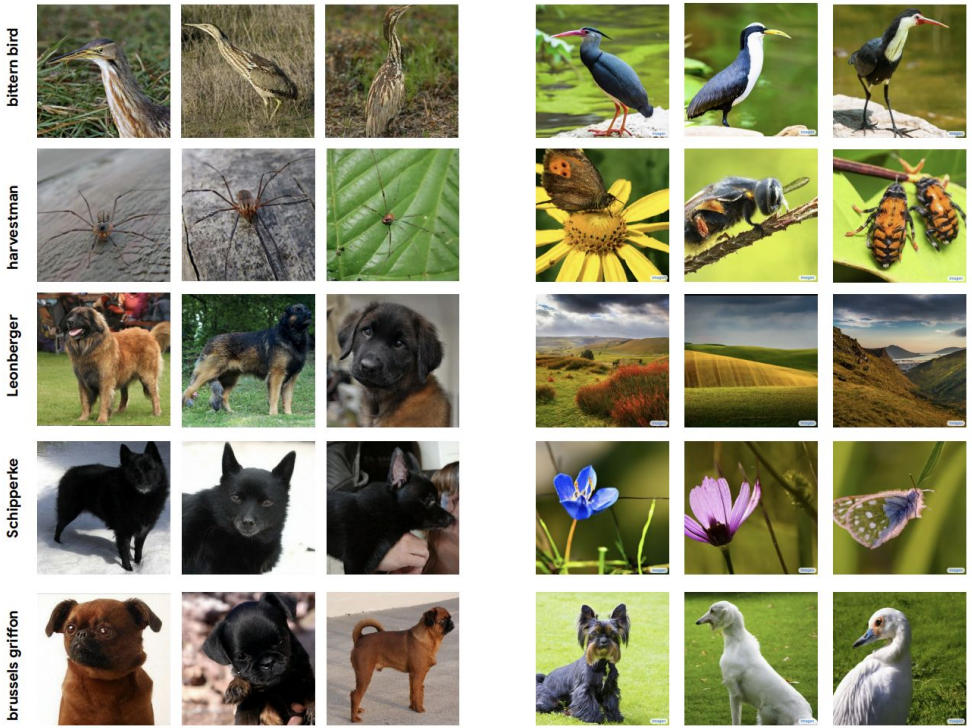
Outperforms other state-of-the-art models in both **FID** and **IS**

Applies for both generated resolutions of **64x64** & **256x256**

Strong evidence that fine-tuning model weights and adjusting sample parameters can **significantly increase** diffusion model performance

Model	FID train	FID validation	IS
64x64 resolution			
BigGAN-deep (Dhariwal & Nichol, 2021) [12]	4.06	-	-
Improved DDPM (Nichol & Dhariwal, 2021) [41]	2.92	-	-
ADM (Dhariwal & Nichol, 2021) [12]	2.07	-	-
CDM (Ho et al, 2022) [26]	1.48	2.48	67.95 ± 1.97
RIN (Jabri et al., 2022) [30]	1.23	-	66.5
RIN + noise schedule (Chen, 2023) [9]	2.04	-	55.8
Ours (Fine-tuned Imagen)	1.21	2.51	85.77 ± 0.06
256x256 resolution			
BigGAN-deep (Brock et al., 2019) [6]	6.9	-	171.4 ± 2.00
VQ-VAE-2 (Razavi et al., 2019) [46]	31.11	-	-
SR3 (Saharia et al., 2021) [49]	11.30	-	-
LDM-4 (Rombach et al., 2022) [47]	10.56	-	103.49
DiT-XL/2 (Peebles & Xie, 2022) [42]	9.62	-	121.5
ADM (Dhariwal & Nichol, 2021) [12]	10.94	-	100.98
ADM+upsampling (Dhariwal & Nichol, 2021) [12]	7.49	-	127.49
CDM (Ho et al, 2022) [26]	4.88	3.76	158.71 ± 2.26
RIN (Jabri et al., 2022) [30]	4.51	4.51	161.0
RIN + noise schedule (Chen, 2023) [9]	3.52	-	186.2
Simple Diffusion (U-Net) (Hoogeboom et al., 2023) [28]	3.76	2.88	171.6 ± 3.07
Simple Diffusion (U-ViT L) (Hoogeboom et al., 2023) [28]	2.77	3.23	211.8 ± 2.93
Ours (Fine-tuned Imagen)	1.76	2.81	239.18 ± 1.14

Before and After



Fine-tuned vs vanilla Imagen generated 1024x1024 images

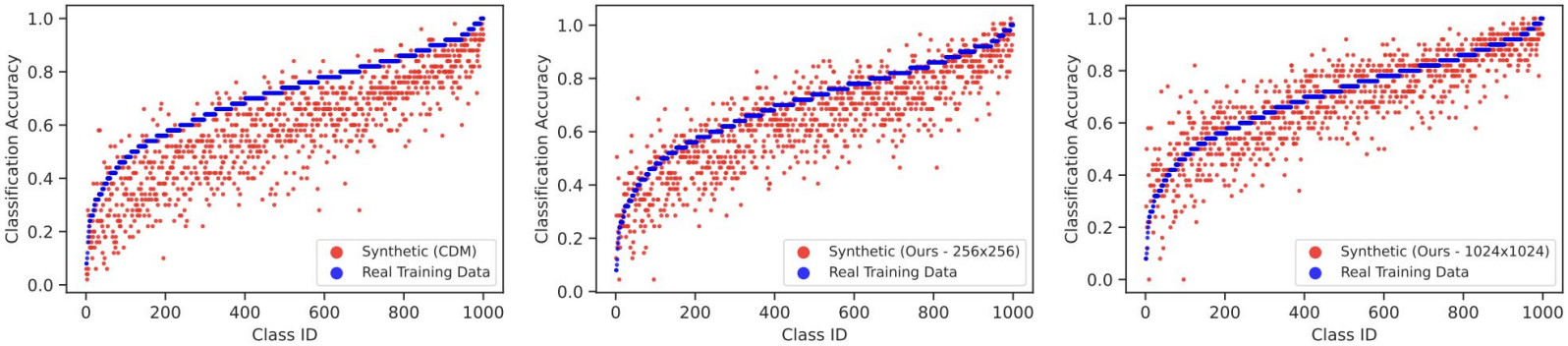
Classification Accuracy Score (ResNet-50)

All images resized to **256×256**, center-cropped to **224×224** for training and evaluation by a **ResNet-50** model

Model	Top-1 Accuracy (%)	Top-5 Accuracy(%)
Real	73.09	91.47
BigGAN-deep (Brock et al., 2019) [6]	42.65	65.92
VQ-VAE-2 (Razavi et al, 2019) [46]	54.83	77.59
CDM (Ho et al, 2022) [26]	63.02	84.06
Ours (256×256 resolution)	64.96	85.66
Ours (1024×1024 resolution)	69.24	88.10

- Outperform previous methods at **256×256** resolution (both **Top-1** and **Top-5** accuracy)
- **1024x1024** marked improvement in **CAS** despite down-sampling to **256×256** for training
- Fine-tuned models achieve state-of-the-art performance with synthetic data, closing the gap to real data-trained models.

Classification Accuracy Score (ResNet-50)



Comparison of Generative Data (red) vs. Real Data (blue)

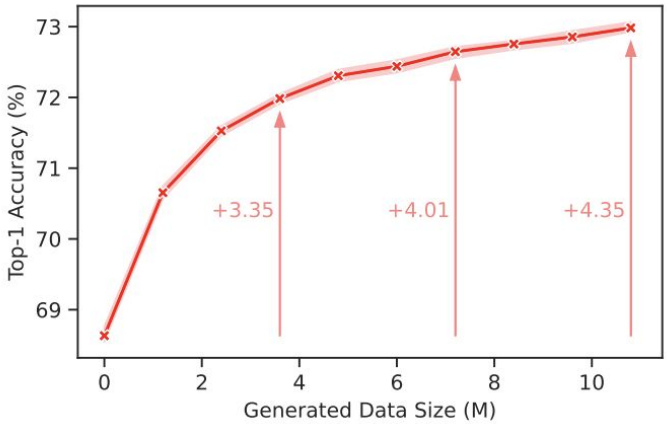
- Accuracy is shown for each of the 1000 ImageNet classes
- Fine-tuned models trained on **generated data** show significant progress, with generative models outperforming real data models in many ImageNet classes, especially at **1024x1024** resolutions.

Classification Accuracy Score (Other Models)

Model	Input Size	Params (M)	Real Only	Generated Only	Real + Generated	Performance Δ
ConvNets						
ResNet-50	224×224	36	76.39	69.24	78.17	+1.78
ResNet-101	224×224	45	78.15	71.31	79.74	+1.59
ResNet-152	224×224	64	78.59	72.38	80.15	+1.56
ResNet-RS-50	160×160	36	79.10	70.72	79.97	+0.87
ResNet-RS-101	160×160	64	80.11	72.73	80.89	+0.78
ResNet-RS-101	190×190	64	81.29	73.63	81.80	+0.51
ResNet-RS-152	224×224	87	82.81	74.46	83.10	+0.29
Transformers						
ViT-S/16	224×224	22	79.89	71.88	81.00	+1.11
DeiT-S	224×224	22	78.97	72.26	80.49	+1.52
DeiT-B	224×224	87	81.79	74.55	82.84	+1.04
DeiT-B	384×384	87	83.16	75.45	83.75	+0.59
DeiT-L	224×224	307	82.22	74.60	83.05	+0.83

The Common Theme: Pure synthetic training data still underperforms compared to using real data, but a combination of both outperforms just using real data

Merging Real and Synthetic Data at Scale



Train Set (M)	256×256	1024×1024
1.2	76.39 ± 0.21	76.39 ± 0.21
2.4	77.61 ± 0.08 (+1.22)	78.12 ± 0.05 (+1.73)
3.6	77.16 ± 0.04 (+0.77)	77.48 ± 0.04 (+1.09)
4.8	76.52 ± 0.04 (+0.13)	76.75 ± 0.07 (+0.36)
6.0	76.09 ± 0.08 (-0.30)	76.34 ± 0.13 (-0.05)
7.2	75.81 ± 0.08 (-0.58)	75.87 ± 0.09 (-0.52)
8.4	75.44 ± 0.06 (-0.95)	75.49 ± 0.07 (-0.90)
9.6	75.28 ± 0.10 (-1.11)	74.72 ± 0.20 (-1.67)
10.8	75.11 ± 0.12 (-1.28)	74.14 ± 0.13 (-2.25)
12.0	75.04 ± 0.05 (-1.35)	73.70 ± 0.09 (-2.69)

Top-1 Accuracy score for base **1.2M** real training data and different amounts of injected synthetic data

- **64x64** resolution show continuous gain in accuracy with growing levels of synthetic data
- **256x256** & **1024x1024** show initial gain followed by a dip with growing levels of synthetic data

Conclusion

Key Findings

There is strong evidence that suggests generative data augmentation is effective with current diffusion models

Using Imagen, a fine-tuned version of it outperforms state-of-the-art models in measurements such as **FID** and **IS**

The fine-tuned model generations used as synthetic training data also surpasses other models when classifying with **ResNet-50**

Although pure synthetic data as training data still underperforms compared to real training data, a combination of the two suggests higher accuracy as compared to using only real data

Limitations and Future Work

Despite **64x64** images obtaining higher classification scores with the increase of synthetic training data, higher resolution performances drop under the same condition

2nd Research Question: Can text-to-image models enhance representation learning and be used in downstream tasks? (**Not Answered**)

This is an issue to be addressed with ongoing research

Thanks for listening!