Traditional Classification Neural Networks are Good Generators:

They are Competitive with DDPMs and GANs

Motivation

Compare to the generative models,

- neural network classifiers are easier to learn.
- neural network classifiers can better model the data's distribution.

Are they ready for image generation?

Neural Network Classifier

The objective of training a neural network classifier is as follows:

$$\min_{f} \mathcal{L}_{\mathsf{cls}}\left(f(x), c\right)$$

- f: neural network
- x: input image
- c: class label for x
- $\mathcal{L}_{\mathsf{cls}}$: classification loss (e.g. cross-entropy loss)

More generally, a neural network classifier can be a cross-model for a text-to-image modeling task such as CLIP.

Overview of the Sampling Process

Starting from a random tensor x_0, by exploiting the knowledge of the classifier, generate an image x_T.

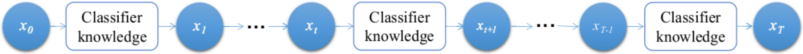

Figure 1: Sampling process

How do we exploit the knowledge of the classifier?

Initial Idea

Method (Directly optimize the input image)

$$x_{t+1} = x_t - \arg\min_{\Delta x_t} \mathcal{L}_{\text{cls}} \left(f \left(x_t + \Delta x_t \right), c \right) \tag{1}$$

- t: time sequence of optimization
- x_0: initial random tensor
- c: target class

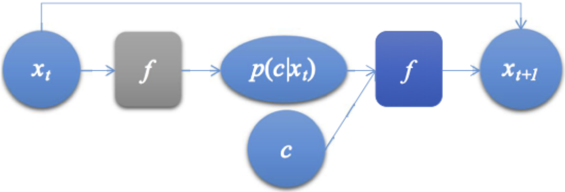

Figure 2: Initial idea

However this objective is actually (almost) equivalent to the *targeted adversarial attack*.

Adversarial Attack

The objective of adversarial attack is to "mislead" the neural networks by making "little" modification to an input.

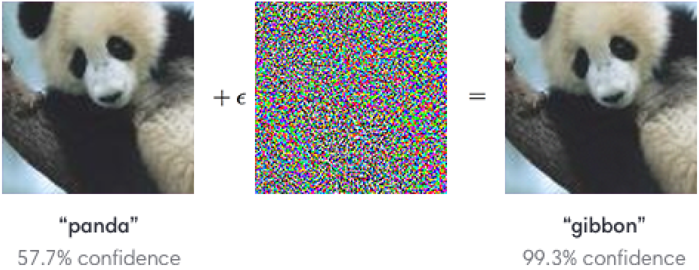

Figure 3: An example of adversarial attack

Adversarial Attack

- **Untargeted Adversarial Attack:** mislead the model to provide *any wrong* answer, i.e.

$$\max_{x^\star} \mathcal{L}_{\mathsf{cls}}\left(f\left(x^\star\right), c\right), \quad \text{s.t.} \quad d\left(x, x^\star\right) < B,$$

  where c is the correct label of x.
- **Targeted Adversarial Attack:** mislead the model to provide the *targeted wrong* answer, i.e.

$$\min_{x^\star} \mathcal{L}_{\mathsf{cls}}\left(f\left(x^\star\right), c^\star\right), \quad \text{s.t.} \quad d\left(x, x^\star\right) < B, \tag{2}$$

  where $c^\star \neq c$ is a specific class assigned by the adversary.

Note equations (1) and (2) are equivalent (apart from the constraint).

Limitation

Equation (1) optimize the high-dimensional input. Hence there could be many *semantic-agnostic* solutions. To address this issue, the authors propose *mask-based stochastic reconstruction model* to make gradients *semantic-aware*.

Similar Limitation in Autoencoder

Q. Why is generative models (specifically, autoencoders) not as effective as discriminative models (such as contrastive learning) in pretraining foundation models for downstream tasks?

A. Autoencoder waste its capability to overfit semantic-agnostic high-frequency details.

Masked Autoencoder

- Masking: Random sampling with *high masking ratio*
- Encoder: ViT, only applied to visible patches.
- Decoder: Light-weight compared to encoder. Takes (i) encoded visible patches (ii) mask tokens.

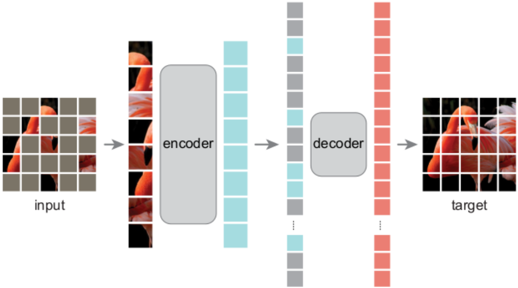

Figure 4: Masked autoencoder architecture

Masked Autoencoder

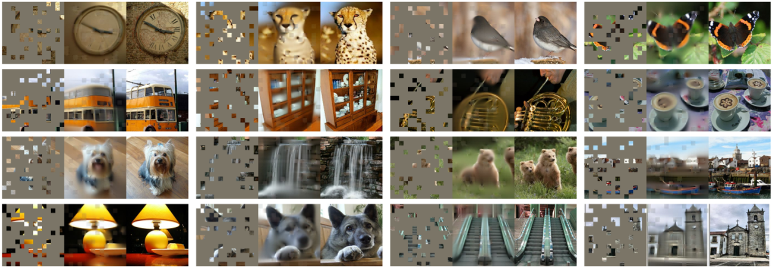

Figure 5: Reconstruction of MAE (80% masking ratio)

Masked Autoencoder

method	pre-train data	ViT-B	ViT-L	ViT-H	ViT-H_{448}
scratch, our impl.	-	82.3	82.6	83.1	-
DINO [5]	IN1K	82.8	-	-	-
MoCo v3 [9]	IN1K	83.2	84.1	-	-
BEiT [2]	IN1K+DALLE	83.2	85.2	-	-
MAE	IN1K	83.6	85.9	86.9	**87.8**

Figure 6: MAE vs self-supervised methods

Mask-Based Stochastic Reconstruction Module

By adding a mask-based stochastic reconstruction module (specifically a masked autoencoder) g, we can rewrite the initial objective (1) as

$$x_{t+1} = x_t - \underset{\Delta x_t}{\arg\min}\, \mathcal{L}_{\mathsf{cls}}\left(f(g(x_t + \Delta x_t)), c\right) \tag{3}$$

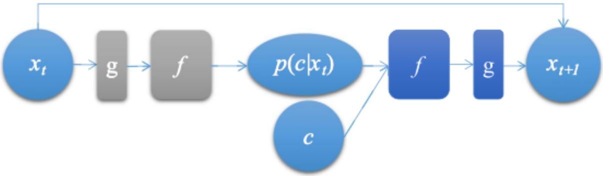

Figure 7: Adding masked-based stochastic reconstruction module

Dilemma of Image Generation

Eimpirically,

$$\begin{array}{ccc} \text{Image resolution} \uparrow & \Longrightarrow & \text{Diversity} \uparrow, \text{ Fidelity} \downarrow \\ & \Updownarrow & \\ \text{Image resolution} \downarrow & \Longrightarrow & \text{Diversity} \downarrow, \text{ Fidelity} \uparrow \end{array}$$

Progressive-Resolution Generation Technique

Start by producing images of low resolution, and gradually increase the resolution of the resulting images exponentially. For instance, sequentially generate images of $64 \times 64, 128 \times 128, 256 \times 256$ as follows:

$$x_0^{64\times64} \xrightarrow{\text{optimize}} x_{\text{opt}}^{64\times64} \xrightarrow{\text{upsample}} x_0^{128\times128}$$

$$\xrightarrow{\text{optimize}} x_{\text{opt}}^{128\times128} \xrightarrow{\text{upsample}} x_0^{256\times256} \xrightarrow{\text{optimize}} x_{\text{opt}}^{256\times256}$$

State-of-the-Art Image Synthesis

Resolutions	Methods	FID ↓	IS ↑
256 × 256	BigGAN-deep [2]	6.95	
	IDDPM [30]	12.26	
	SR3 [43]	11.30	
	DCTransformer [26]	36.51	
	VQ-VAE-2 [39]	31.11	
	ADM [9] w/o condition, w/ guidance	12.00	95.41
	DeepDream [24]	134.69	22.60
	Ours	6.88 †	326.33

Figure 8: Quantitative comparison on the ImageNet 256×256

State-of-the-Art Image Synthesis

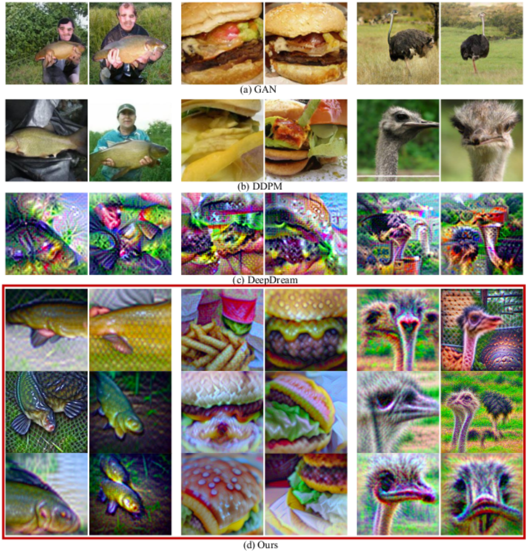

Figure 9: Samples from ImageNet 256×256

State-of-the-Art Image Synthesis

CaG has stronger semantic perception:

- CaG pays more attention to object diversity than background diversity: birds occupy a large area of the picture
- CaG decouples and remove irrelevant object categories: include only Tinca fishes that were not held by people
- Cag appears to be aware of geometric information

Text-to-Image Generation

Text-to-image foundation models as a generalized classifier:

- Extract embeddings for text via the text encoder
- Form the weight of the *classifying layer* with them
- Impose the *classifying layer* on the image encoder

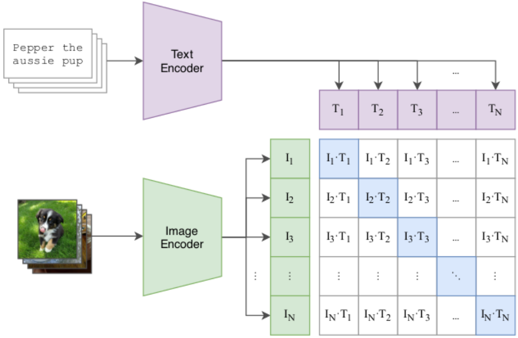

Figure 10: Text-to-image foundation models as a generalized classifier

Text-to-Image Generation

Figure 11: Text-to-image generation