Outline (Cont. from part 1 covered in Lec#17)

Part 1: Basics

- Denoising diffusion probabilistic models (DDPMs)
- Conditional diffusion models
- Large-scale models: DALL-E 2, Stable Diffusion, Imagen

Part 2: Recent Advances

- Denoising diffusion implicit models (DDIMs)
- Stable Diffusion XL, Stable Diffusion 3
- Progressive Distillation
Progressive Distillation

Salimans et al, Progressive distillation for fast sampling of diffusion models
Progressive Distillation: Results

Figure 10: Random samples from our distilled LSUN bedrooms models, for fixed random seed and for varying number of sampling steps.
Outline

Part 1: Basics
• Denoising diffusion probabilistic models (DDPMs)
• Conditional diffusion models
• Large-scale models: DALL-E 2, Stable Diffusion, Imagen

Part 2: Recent Advances
• Denoising diffusion implicit models (DDIMs)
• Stable Diffusion XL, Stable Diffusion 3
• Model Distillation
• Latent Consistency Models (LCM)
Latent Consistency Models

Latent Consistency Models: combine the above idea with Latent Diffusion Models

Song et al., Consistency Models

Luo et al, Latent Consistency Models: Synthesizing High-Resolution Images with Few-step Inference
Latent Consistency Models: Results

![Graph showing FID vs Inference Time for DPM-Solver++ and LCM (Ours) with images of generated samples for 1-Step, 2-Steps, 4-Steps, and 8-Steps Inference.]
Outline

Part 1: Basics
• Denoising diffusion probabilistic models (DDPMs)
• Conditional diffusion models
• Large-scale models: DALL-E 2, Stable Diffusion, Imagen

Part 2: Recent Advances
• Denoising diffusion implicit models (DDIMs)
• Stable Diffusion XL, Stable Diffusion 3
• Progressive Distillation
• Latent Consistency Models (LCM)

(next class) Part 3: Applications and Implementation, Ethical issues
Outline

Part 3: Applications and Implementation; Ethical Issues

• Customizing Diffusion Models
  • Textual Inversion
  • DreamBooth
  • Low Rank Approximation (LoRA)
  • ZipLoRA

• ControlNet

• Prompt-to-Prompt

• InstructPix2Pix

• DreamFusion

• Working with Diffusion Models: Implementation aspects

• Societal, ethical, and legal issues
Outline

Part 3: Applications and Implementation; Ethical Issues

• Customizing Diffusion Models
  • Textual Inversion
  • DreamBooth
  • Low Rank Approximation (LoRA)
  • ZipLoRA
Customizing DMs: Textual inversion

Figure 1: (left) We find new pseudo-words in the embedding space of pre-trained text-to-image models which describe specific concepts. (right) These pseudo-words are composed into new sentences, placing our targets in new scenes, changing their style or ingraining them into new products.

R. Gal et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR 2023
Customizing DMs: Textual inversion

Figure 2: Outline of the text-embedding and inversion process. A string containing our placeholder word is first converted into tokens (i.e. word or sub-word indices in a dictionary). These tokens are converted to continuous vector representations (the “embeddings”, \( v \)). Finally, the embedding vectors are transformed into a conditioning code \( c_\theta(y) \) that guides the generation. We optimize the embedding vector \( v_* \) associated with our pseudo-word \( S_* \), using a reconstruction objective.

R. Gal et al. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. ICLR 2023
Textual inversion: Results

Input samples

“Watercolor painting of $S$ on a branch”
“A house in the style of $S$”
“Grainy photo of $S$ in angry birds”
$S$ made of chocolate
“A $S$ dragon”

“A mosaic depicting $S$”
“Death metal album cover featuring $S$”
“Masterful oil painting of $S$ hanging on the wall”
“An artist drawing a $S$”
“A $S$ dancing ballet”

“A photo of $S$ full of cashew nuts”
“A mouse using $S$ as a boat”
“A photo of a $S$ mask”
“Ramen soup served in $S$”
“Cave mural depicting $S$”
Textual inversion: Results

Input samples → “The streets of Paris in the style of $S_*$” → “Adorable corgi in the style of $S_*$” → “Painting of a black hole in the style of $S_*$” → “Times square in the style of $S_*$” → “Edo period pagoda in the style of $S_*$”
Textual inversion: Comparisons

Skull Mug

Input samples

Ours

“A photo of $S_*$”

DALLE-2
(Image Inputs)

DALLE-2
(Long Captions)

LDM
(Long Captions)

Teapot

“A photo of $S_*$”

A mug having many skulls at the bottom and sculpture of a man at the top of it.

A tea pot, with green, red, blue, yellow, an apple with leaves and a lemon with leaves.
Customizing DMs: DreamBooth

Customizing DMs: DreamBooth

Input Images

Image-guided, DALL-E2

Fidelity
New contexts

Prompt: "retro style yellow alarm clock with a white clock face and a yellow number three on the right part of the clock face in the jungle"

Text-guided, Imagen

Fidelity
New contexts

Ours

Figure 3. **Fine-tuning.** Given ~ 3 – 5 images of a subject we fine-tune a text-to-image diffusion model with the input images paired with a text prompt containing a unique identifier and the name of the class the subject belongs to (e.g., “A [V] dog”), in parallel, we apply a class-specific prior preservation loss, which leverages the semantic prior that the model has on the class and encourages it to generate diverse instances belong to the subject’s class using the class name in a text prompt (e.g., “A dog”).
Customizing DMs: DreamBooth

Reconstruction Loss

"A [V] dog"

Shared Weights

Input images (~3-5)

"A dog"

Class-Specific Prior Preservation Loss

Text → Image

Text → Image

"A dog"

Figure 6. Encouraging diversity with prior-preservation loss. Naive fine-tuning can result in overfitting to input image context and subject appearance (e.g., pose). PPL acts as a regularizer that alleviates overfitting and encourages diversity, allowing for more pose variability and appearance diversity.

DreamBooth: Results

Input images
A [V] backpack in the Grand Canyon
A wet [V] backpack in water
A [V] backpack in Boston
A [V] backpack with the night sky

Input images
A [V] teapot floating in milk
A transparent [V] teapot with milk inside
A [V] teapot pouring tea
A [V] teapot floating in the sea
DreamBooth: Results

**Text-guided view synthesis**

Input images

Top view 🔫 Bottom view 🔴 Back view 🔺

**Art Renditions**

Van Gogh  Michelangelo  Vermeer

“a [painting/sculpture] of a [V] [class noun] in the style of [famous artist]”

**Property Modification**

Panda  Lion  Hippo

“a cross of a [V] dog and a [target species]”
DreamBooth: Comparison with textual inversion

<table>
<thead>
<tr>
<th>Method</th>
<th>Subject Fidelity ↑</th>
<th>Prompt Fidelity ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DreamBooth (Stable Diffusion)</td>
<td>68%</td>
<td>81%</td>
</tr>
<tr>
<td>Textual Inversion (Stable Diffusion)</td>
<td>22%</td>
<td>12%</td>
</tr>
<tr>
<td>Undecided</td>
<td>10%</td>
<td>7%</td>
</tr>
</tbody>
</table>

Input Images

DreamBooth (Imagen)

DreamBooth (Stable Diffusion)

Textual Inversion (Stable Diffusion)

“a [V] vase in the snow” “a [V] vase on the beach” “a [V] vase in the jungle” “a [V] vase with Eiffel Tower in the background”
DreamBooth: Limitations

Figure 9. **Failure modes.** Given a rare prompted context the model might fail at generating the correct environment (a). It is possible for context and subject appearance to become entangled (b). Finally, it is possible for the model to overfit and generate images similar to the training set, especially if prompts reflect the original environment of the training set (c).
Efficient Customization: DreamBooth with LoRA

Weight update in regular finetuning

Inputs $d$ \rightarrow Pretrained weights $W$ \rightarrow Weight update $\Delta W$ \rightarrow Outputs

Weight update in LoRA

Inputs $d$ \rightarrow Pretrained weights $W$ \rightarrow LoRA matrices $A$ and $B$ approximate the weight update matrix $\Delta W$ \rightarrow Outputs

$A$ is initialized with standard normal; $B$ is initialized with zeros

Hu et al., LoRA: Low-Rank Adaptation of Large Language Models, ICLR'22; https://magazine.sebastianraschka.com/p/practical-tips-for-fineturning-llms
LoRA DreamBooth: Results

Input Images

LoRA DreamBooth (r=4)

DreamBooth
# LoRA DreamBooth for Stylizations (on SDXL)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cartoon line drawing</td>
<td><img src="image1" alt="Bicycle" /></td>
<td><img src="image2" alt="Golden Gate Bridge" /></td>
<td><img src="image3" alt="Bird" /></td>
<td><img src="image4" alt="Boat" /></td>
<td><img src="image5" alt="Hat" /></td>
<td><img src="image6" alt="Piano" /></td>
</tr>
<tr>
<td>watercolor painting</td>
<td><img src="image7" alt="Bicycle" /></td>
<td><img src="image8" alt="Golden Gate Bridge" /></td>
<td><img src="image9" alt="Bird" /></td>
<td><img src="image10" alt="Boat" /></td>
<td><img src="image11" alt="Hat" /></td>
<td><img src="image12" alt="Piano" /></td>
</tr>
</tbody>
</table>

Stylizations obtained using DreamBooth on SDXL with LoRA

Shah et al., *ZipLoRA: Any Subject in Any Style by Effectively Merging LoRAs*, arxiv'23
Can we Merge Content and Style LoRAs?

Content Images

“A [V] dog”

Content LoRA $L_c$

[V] with angel wings [V] in wizard outfit [V] as nurse

Style Image

“Flowers in [S] style”

Style LoRA $L_s$

penguin in [S] style bike in [S] style
Can we Merge Content and Style LoRAs?

Content Images
“[V] dog”

Style Image
“Flowers in [S] style”

ZipLoRA

Content LoRA
$L_c$

Style LoRA
$L_s$

“[V] dog in [S] style”
“[V] dog playing with a ball in [S] style”
“Sleeping [V] dog in [S] style”
“[V] dog wearing a crown in [S] style”
ZipLoRA produces successful recontextualizations

Subject and Style Refereneces

Recontextualizations using our method
Anti-DreamBooth

Figure 1: A malicious attacker can collect a user’s images to train a personalized text-to-image generator for malicious purposes. Our system, called Anti-DreamBooth, applies imperceptible perturbations to the user’s images before releasing, making any personalized generator trained on these images fail to produce usable images, protecting the user from that threat.

Part 3: Applications and Implementation; Ethical Issues

- Customizing Diffusion Models
  - Textual Inversion
  - DreamBooth
  - Low Rank Approximation (LoRA)
  - ZipLoRA

- ControlNet
ControlNet

- Add a trainable “wrapper” around a pre-trained DM to fine-tune it for pix2pix tasks

ControlNet

ControlNet

ControlNet

Part 3: Applications and Implementation; Ethical Issues

- Customizing Diffusion Models
  - Textual Inversion
  - DreamBooth
  - Low Rank Approximation (LoRA)
  - ZipLoRA
- ControlNet
- Prompt-to-Prompt and InstructPix2Pix
Prompt-to-Prompt Image Editing

“A. Hertz et al. Prompt-to-Prompt Image Editing with Cross Attention Control. arXiv 2022”
Prompt-to-Prompt Image Editing

Word Swap

Adding a New Phrase

Attention Re-weighting
Prompt-to-Prompt Image Editing

Fixed attention maps and random seed

Fixed random seed
InstructPix2Pix

“Swap sunflowers with roses”  “Add fireworks to the sky”  “Replace the fruits with cake”

“What would it look like if it were snowing?”  “Turn it into a still from a western”  “Make his jacket out of leather”

Figure 1. Given an image and an instruction for how to edit that image, our model performs the appropriate edit. Our model does not require full descriptions for the input or output image, and edits images in the forward pass without per-example inversion or fine-tuning.

InstructPix2Pix

Figure 2. Our method consists of two parts: generating an image editing dataset, and training a diffusion model on that dataset. (a) We first use a finetuned GPT-3 to generate instructions and edited captions. (b) We then use StableDiffusion [52] in combination with Prompt-to-Prompt [17] to generate pairs of images from pairs of captions. We use this procedure to create a dataset (c) of over 450,000 training examples. (d) Finally, our InstructPix2Pix diffusion model is trained on our generated data to edit images from instructions. At inference time, our model generalizes to edit real images from human-written instructions.

### InstructPix2Pix

- **Fine-tuning GPT-3:**

<table>
<thead>
<tr>
<th>Human-written (700 edits)</th>
<th>Input LAION caption</th>
<th>Edit instruction</th>
<th>Edited caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yefim Volkov, Misty Morning girl with horse at sunset painting-of-forest-and-pond</td>
<td>make it afternoon</td>
<td>change the background to a city Without the water.</td>
<td>Yefim Volkov, Misty Afternoon girl with horse at sunset in front of city painting-of-forest</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GPT-3 generated (&gt;450,000 edits)</th>
<th>Input LAION caption</th>
<th>Edit instruction</th>
<th>Edited caption</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alex Hill, Original oil painting on canvas, Moonlight Bay The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it Kate Hudson arriving at the Golden Globes 2015</td>
<td>in the style of a coloring book Add a giant red dragon make her look like a zombie</td>
<td>Alex Hill, Original coloring book illustration, Moonlight Bay The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead Zombie Kate Hudson arriving at the Golden Globes 2015</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Table 1. We label a small text dataset, finetune GPT-3, and use that finetuned model to generate a large dataset of text triplets. As the input caption for both the labeled and generated examples, we use real image captions from LAION. **Highlighted text** is generated by GPT-3.
InstructPix2Pix

- Generating input-output image pairs:

(a) Without Prompt-to-Prompt.  
(b) With Prompt-to-Prompt.

Figure 3. Pair of images generated using StableDiffusion [52] with and without Prompt-to-Prompt [17]. For both, the corresponding captions are “photograph of a girl riding a horse” and “photograph of a girl riding a dragon”.
InstructPix2Pix

• Fine-tuning a DM for image-to-image translation:

“Have her ride a dragon”
Time $t$
InstructPix2Pix: Results

Figure 5. *Mona Lisa* transformed into various artistic mediums.

Figure 6. *The Creation of Adam* with new context and subjects (generated at 768 resolution).
InstructPix2Pix: Results

Input
“Apply face paint”
“What would she look like as a bearded man?”
“Put on a pair of sunglasses”
“She should look 100 years old”

“What if she were in an anime?”
“Make her terrifying”
“Make her more sad”
“Make her James Bond”
“Turn her into Dwayne The Rock Johnson”
InstructPix2Pix: Results

“Make it Paris”
“Make it Hong Kong”
“Make it Manhattan”
“Make it Prague”

“Make it evening”
“Put them on roller skates”
“Turn this into 1900s”
“Make it underwater”

“Make it Minecraft”
“Turn this into the space age”
“Make them into Alexander Calder sculpture”
“Make it a Chihuly”
Figure 13. Failure cases. Left to right: our model is not capable of performing viewpoint changes, can make undesired excessive changes to the image, can sometimes fail to isolate the specified object, and has difficulty reorganizing or swapping objects with each other.
Outline

Part 3: Applications and Implementation; Ethical Issues

- Customizing Diffusion Models
  - Textual Inversion
  - DreamBooth
  - Low Rank Approximation (LoRA)
  - ZipLoRA
- ControlNet
- Prompt-to-Prompt
- InstructPix2Pix
- DreamFusion
Connecting 2D to 3D: DreamFusion

B. Poole, A. Jain, J. Barron, B. Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. arXiv 2022
Connecting 2D to 3D: DreamFusion

B. Poole, A. Jain, J. Barron, B. Mildenhall. DreamFusion: Text-to-3D using 2D Diffusion. arXiv 2022
Outline

Part 3: Applications and Implementation; Ethical Issues

- Customizing Diffusion Models
  - Textual Inversion
  - DreamBooth
  - Low Rank Approximation (LoRA)
  - ZipLoRA
- ControlNet
- Prompt-to-Prompt
- InstructPix2Pix
- DreamFusion
- Working with Diffusion Models: Implementation aspects
Working with Diffusion Models

- Fast moving area with new research coming in everyday
- Variety of pre-trained, open-source models available
- Working directly with the implementations and codebases helps!
- Starting point: diffusers library by huggingface
- Popular open-source models:
  - Stable Diffusion
  - SDXL
  - SD3 (recently announced)
  - Deep-Floyd IF (open-source alternative of Imagen)
- Some useful websites
  - Huggingface, Reddit threads on Stable Diffusion; Civit.ai; publicprompts.art
Part 3: Applications and Implementation; Ethical Issues

- Customizing Diffusion Models
  - Textual Inversion
  - DreamBooth
  - Low Rank Approximation (LoRA)
  - ZipLoRA
- ControlNet
- Prompt-to-Prompt
- InstructPix2Pix
- DreamFusion
- Working with Diffusion Models: Implementation aspects
- Societal, ethical, and legal issues
Societal, ethical, and legal issues

• Closed or open?
• Safe or unsafe?
• Potential for generating DeepFakes and misinformation
• Dataset image rights
• Artists’ rights
• The nature of creativity
In the news

ARTIFICIAL INTELLIGENCE / TECH / LAW

Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content

 Getty Images claims Stability AI ‘unlawfully’ scraped millions of images from its site. It’s a significant escalation in the developing legal battles between generative AI firms and content creators.

By JAMES VINCENT
Jan 17, 2023, 4:30 AM CST | □ 18 Comments / 18 New

In the news

https://www.newyorker.com/culture/infinite-scroll/is-ai-art-stealing-from-artists
In the news

Fake Trump arrest photos: How to spot an AI-generated image

Midjourney Bans AI Images of Chinese President Xi Jinping

The phrase president x is banned. Circumventing this filter to violate our rules may result in your access being revoked.