Jade Sanchez

ITAI 2376

4/9/2025

Midterm Report

Understanding Diffusion

What is forward diffusion?
During forward diffusion a clean image receives increasing levels of noise at regular time intervals. The gradual increase of image corruption leads to a complete conversion into noise. The training process enables the system to develop the ability to undo drift steps one by one.

Why is noise added gradually?
The model trains through numerous small denoising steps because of gradual noise addition. Adding complete noise at one time creates too much difficulty for recovering the initial image. When changes occur progressively during training this process enables students to more easily understand the steps while enhancing model stability.

When do images become recognizable?
Visualization of digits first emerges when reverse diffusion reaches between 40% to 60% through the denoising phase. Narrow and coarse noise particles vanish from images at initial steps followed by fine enhanced image structures emerging at the conclusion of the process.

Model Architecture

U-Net demonstrates strong characteristics for working with diffusion models.
Through its encoder-decoder design U-Net includes skip connection features that enable operation efficiency. The implementation of this method permits preserving spatial elements during reconstruction which produces clean images by denoising contamination. The image generation task benefiting the most from U-Net is complex detail-focused work.

What do skip connections do?
The encoder layers of U-Net transmit their feature maps directly to the matching decoder layers through these skip connections. By utilizing skip connections U-Net keeps vital information that would typically get compromised during downsampling operations.

How does class conditioning work?

Class conditioning applies an embedding layer for vectorizing class label values (such as “7”). At different network stages the vectors get added to feature maps in order to orient the generation toward particular classes.

Training Analysis

What does the loss mean?
The loss function serves to determine the accuracy of predictions from model noise when compared to the initial added noise in forward diffusion. Lower loss = better noise prediction = better image recovery.

How did image quality evolve?
During early phases of the experiment images appeared unrecognizable while being blurred. The images gradually became easier to understand after the completion of multiple epochs. The training and validation loss data demonstrated consistent improvement according to the shown graph results.

What is the reason behind implementing time embeddings?
Embedded times in the model enable it to determine the amount of added noise. These components enable the model to adjust its clean-up operations based on which step it currently performs in the reverse diffusion time span.

CLIP Evaluation (Bonus)

What do CLIP scores tell us?
CLIP provides a system to check the quality of generated images' adherence to their designated classifications. The visual-textual similarity improves as the score increases. Through this validation process the model functions as the second validating system for our output.

Which numbers does the model find simpler to produce?
The model can easily understand and learn simple digits such as “1” and “0” because their distinctive shapes provide clear distinctions from other numbers. In the case of complex digits such as "5" or "8" network confusion arises from features that overlap or create curves during generation.

How could CLIP improve generation?

The CLIP scoring method could function during training by incorporating it as an additional training mechanism. Directing training towards producing outcomes that match semantic expectations would be possible through the incorporation of the CLIP scores.

Practical Applications

Real-world uses of diffusion models:

- AI art and image generation (e.g., DALL·E)
- Data augmentation in healthcare (e.g., MRI synthesis)
- Image restoration or inpainting
- Text-to-image synthesis

Limitations of current model:

- Slow generation (many steps)
- Blurry output if undertrained
- Limited resolution (28x28 for MNIST)

Three improvements I'd make:

The training process should include the CIFAR-10 dataset which offers higher resolution than current implementations.

Classifier-free guidance provides users with more flexibility when producing expressive outputs during the generation process.

Attention-based models should be adopted to improve detail precision during modeling processes.