

# Deep Learning for Computer Vision

NTU, Fall 2024, Homework2

電機所碩一 李彥璋 R13921093

## Problem 1: Diffusion Models (15%)

1. (5%) Describe your implementation details and the difficulties you encountered.

### Implementation details

#### I. 模型架構設計 : ContextUnet

- ContextUnet 是基於U-Net的架構，用來處理影像數據以及額外的條件，如數字標籤(digit\_labels)和資料集標籤(dataset\_labels)。
- ResidualConvBlock: 一個基本的殘差卷積塊，實現殘差連接。
- UnetDown與UnetUp: 實現U-Net架構中的下採樣和上採樣部分。
- EmbedFC: 此嵌入層用於將數字標籤和資料集標籤轉換為高維嵌入向量。
- 時間嵌入(Time Embeddings): 使用EmbedFC將時間步(t)轉換為嵌入向量，讓模型在不同的擾動時間步中進行有效的學習。
- 上下文嵌入(Context Embeddings): 數字標籤和資料集標籤經過嵌入層轉換後，與時間嵌入結合，並在上採樣過程中逐元素相乘，實現條件生成。

#### II. DDPM調度設計 : ddpm\_schedules

- ddpm\_schedules 函式計算DDPM所需的各種參數，如 $\alpha_t$ 、 $\beta_t$ 、 $\alpha^{-1}t$ 等。這些參數能在擾動(forward process)和反向過程(reverse process)中確保模型能夠學習從噪聲中恢復原始圖像。

#### III. 訓練過程

- 隨機遮蔽條件資訊(context\_mask): 在訓練過程中，隨機遮蔽部分條件資訊(數字標籤和資料集標籤)，這種策略稱為Classifier-free Guidance，旨在提升模型的泛化能力，讓模型學會在部分條件缺失的情況下仍能生成高質量的影像。

#### IV. 取樣過程 : sample 函式

- 條件生成: sample函式允許指定數字標籤和資料集標籤，根據這些條件生成新的影像。這是通過將條件嵌入傳遞給ContextUnet來實現的。
- Classifier-free Guidance: 在取樣過程中，通過調整預測噪聲的方式，結合有條件和無條件的預測，從而提升生成影像的品質和條件的一致性。具體實現為將有條件預測的噪聲乘以 $1+guide_w$ ，無條件預測的噪聲乘以 $-guide_w$ ，並進行線性組合。

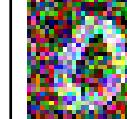
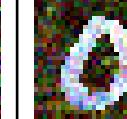
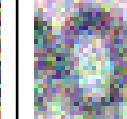
## Difficulties

起初，我誤以為在生成圖片時不需要使用 Classifier-free Guidance，結果導致模型的準確度僅在93%左右徘徊，且生成圖片的品質也不理想。此外，在學習率的調整上，我嘗試了很長時間，最終發現將後期的學習率略微降低有助於模型更好地收斂。另一方面，我也花了很多時間研究如何修改 UNet 的架構，以適應兩個不同的資料集。

2. (5%) Please show 10 generated images for each digit (0-9) from both MNIST-M & SVHN dataset in your report. You can put all 100 outputs in one image with columns indicating different noise inputs and rows indicating different digits.

MNIST-M	SVHN
	

3. (5%) Visualize a total of six images from both MNIST-M & SVHN datasets in the reverse process of the first “0” in your outputs in (2) and with different time steps.

Time step	T = 1	T = 40	T = 120	T = 200	T = 320	T = 500
MNIST-M						
SVHN						

## Problem 2: DDIM (15%)

1. (7.5%) Please generate face images of noise 00.pt ~ 03.pt with different eta in one grid. Report and explain your observation in this experiment.

	00.pt	01.pt	02.pt	03.pt
eta = 0.0				
eta = 0.25				
eta = 0.50				
eta = 0.75				
eta = 1.0				

### My observation

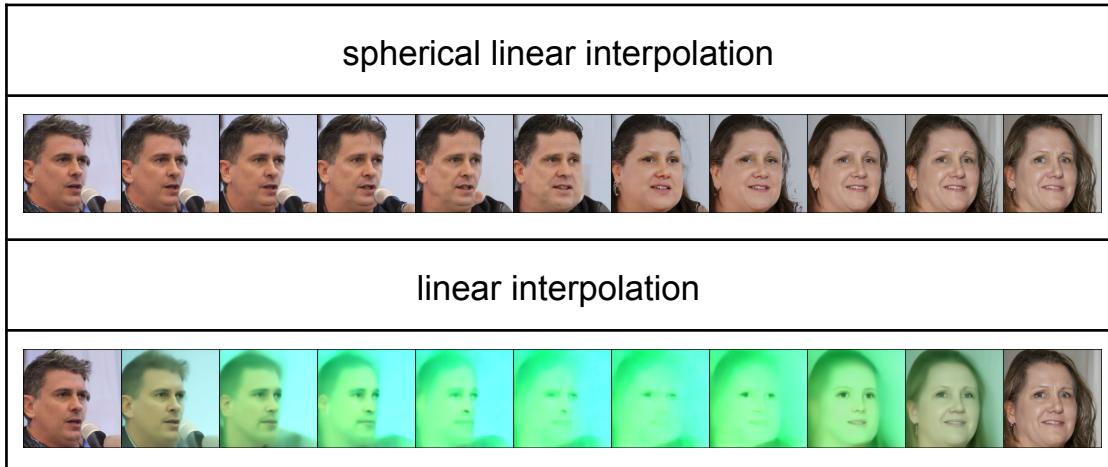
在這個實驗中，使用不同的  $\eta$  值生成臉部圖像。隨著  $\eta$  值增加，生成圖像的隨機性和多樣性逐漸增強，而穩定性和清晰度則有所減弱（頭髮和臉部紋理）。當  $\eta = 0.0$  時，生成的圖像最穩定、細節最清晰，缺乏隨機性；隨著  $\eta$  增加至 0.25 和 0.5，圖像變得更加多樣，保持了較好的平衡；當  $\eta$  增至 0.75 和 1.0 時，圖像的隨機性進一步增加，但特徵變得更加不穩定且部分細節模糊。這表明  $\eta$  值控制了生成過程中的隨機性和變異性，合適的  $\eta$  值可以在圖像品質與多樣性之間取得平衡。

與 DDPM(Denoising Diffusion Probabilistic Models)相比, DDIM(Denoising Diffusion Implicit Models)具有以下優勢:

- I. 生成速度更快:DDIM 可以通過非隨機的推理過程實現更少步驟的生成, 從而顯著加快生成速度, 而不損失太多圖像品質。
- II. 更高的控制性:DDIM 的推理過程允許通過調整  $\eta$  值來控制生成過程中的隨機性和多樣性, 從而使得生成圖像的品質和多樣性之間可以更靈活地進行權衡。
- III. 確定性生成:當  $\eta = 0$  時, DDIM 的生成過程是確定性的, 這意味著相同的輸入噪聲會得到相同的輸出, 這對於需要可重現結果的應用非常有用。

2. (7.5%) Please generate the face images of the interpolation of noise 00.pt ~ 01.pt.

The interpolation formula is spherical linear interpolation, which is also known as slerp. What will happen if we simply use linear interpolation? Explain and report your observation.



**My observation**

如果使用spherical linear interpolation實作, 看起來像是一個從臉 A 平滑地變成臉 B 的過程;如果使用linear interpolation實作, 中間過渡的部分會有一些奇怪的綠色光暈籠罩, 看起來不太自然。

在生成模型中, 噪聲向量通常被假設服從標準正態分佈, 即均值為 0, 方差為 1。然而, 使用linear interpolation進行插值時, 結果向量可能不再符合這一分佈特性。具體而言, 向量的長度可能會變大或變小, 導致方差偏離 1。這種方差的改變會對生成圖像的品質產生負面影響, 例如出現模糊或失焦、噪點增多、特徵不連續等現象。

為了解決這些問題, 通常使用spherical linear interpolation在高維球面上進行插值, 從而確保向量的方差始終保持為 1, 以獲得更平滑、連續且高品質的插值結果。

**Problem 3: Personalization (15%)**

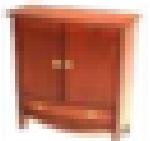
1. (7.5%) Conduct the CLIP-based zero shot classification on the `hw2_data/clip_zeroshot/val`, explain how CLIP do this, report the accuracy and 5 successful/failed cases.

#### How CLIP do this

CLIP (Contrastive Language-Image Pretraining) 利用一個同時處理圖像和文本的雙模態模型進行零樣本分類。具體而言，CLIP首先將預定的類別名稱(`id2label.json`中的標籤)嵌入到一個文本模板中，如「A photo of {object}」，生成一組文本描述。這些文本描述通過文本編碼器轉換為文本特徵向量。

同時，待分類的圖像經過圖像編碼器提取出圖像特徵向量。CLIP將圖像特徵向量與所有文本特徵向量計算餘弦相似度，衡量圖像與每個文本描述的匹配程度。最終，選擇相似度最高的文本描述所對應的類別作為預測結果。

這種方法利用了CLIP在大規模圖文對數據集上的預訓練能力，無需在特定任務上進行微調就能實現零樣本學習，對未見過的類別進行有效的分類。

Accuracy and 5 successful/failed cases					
<pre>(ldm) ycli219@lab:~/Desktop/H2O\$ python3 p3-0shot-classification.py Accuracy: 56.52%</pre>					
<p>Successful Cases:</p> <pre>Image: 32_457.png, True Label: bus, Predicted Label: bus Image: 26_490.png, True Label: rose, Predicted Label: rose Image: 21_469.png, True Label: wardrobe, Predicted Label: wardrobe Image: 44_471.png, True Label: elephant, Predicted Label: elephant Image: 13_485.png, True Label: willow_tree, Predicted Label: willow_tree</pre>					
<p>Failed Cases:</p> <pre>Image: 20_463.png, True Label: beetle, Predicted Label: worm Image: 15_493.png, True Label: raccoon, Predicted Label: willow_tree Image: 7_457.png, True Label: fox, Predicted Label: kangaroo Image: 27_464.png, True Label: bee, Predicted Label: willow_tree Image: 25_493.png, True Label: wolf, Predicted Label: willow_tree</pre>					
Accuracy : 56.52%					
Successful					
Failed					

2. (7.5%) What will happen if you simply generate an image containing multiple

concepts (e.g., a <new1> next to a <new2>)? You can use your own objects or the provided cat images in the dataset. Share your findings and survey a related paper that works on multiple concepts personalization, and share their method.

<new1>狗 ; <new3>貓, 皆使用助教提供的圖片訓練	
a <new3> next to a <new1>	a <new1> next to a <new3>
	

### My findings

在這兩張圖中，我們可以看到生成模型在處理多重概念時面臨的挑戰。左圖中，貓的外觀正常，但狗的特徵卻受到貓的影響，表現出類似貓的眼睛和身形；右圖中，狗的外觀正常，但貓的臉部形狀卻顯得有些像狗。

這些結果顯示模型在處理多個概念時，無法有效保持兩個物種的特徵獨立，而是產生了特徵混合的現象，突顯出其在多重概念處理方面的局限性。

### Survey a related paper

在《Multi-Concept Customization of Text-to-Image Diffusion》一文中，研究者提出了名為「Custom Diffusion」的方法，用於將多個個人化的概念同時融入於文本生成圖像模型中。該方法主要針對兩個挑戰：

- I. 概念的快速嵌入：即使用少量圖片將新的個人化概念引入模型，如寵物或個人收藏品，並確保模型不會因為新概念的加入而遺忘既有知識。
- II. 多概念合成：將多個新概念同時引入，並能在生成的圖像中準確地展現出多概念的組合。

### 方法概述

Custom Diffusion方法透過優化預訓練模型中少量的參數來實現個性化定制。具體而

言，該方法僅針對文本到圖像的交叉注意力層中的「key」和「value」映射進行微調，從而更新模型的內部表示，而非全面微調模型。這種精簡的微調大大縮短了訓練時間（約6分鐘）並減少了存儲需求。此外，為了防止模型忘記已有知識，Custom Diffusion 還在訓練中使用了正則化數據集。

### 多概念組合技術

Custom Diffusion 支持兩種多概念組合方式：

- I. 聯合訓練：將多個概念的訓練數據合併，並使用不同的標籤（如V1和V2）表示不同的概念，使模型能在同一場景中生成多個個性化概念的圖像。
- II. 限制優化法：針對每個新概念分別訓練並保存關鍵參數，然後通過一種閉式優化方法將這些參數合併，從而在生成階段實現多概念組合。

研究結果顯示，Custom Diffusion 在生成包含多個新概念的圖像時，無論是文本對齊度還是圖像相似度均優於現有方法如 DreamBooth 和 Textual Inversion，並在訓練效率與內存需求方面表現出顯著的優勢。

## Reference

[1] chatgpt

<https://chatgpt.com>

[2] Classifier-free guidance

<https://arxiv.org/abs/2207.12598>

[3] Conditional\_Diffusion\_MNIST

[https://github.com/TeaPearce/Conditional\\_Diffusion\\_MNIST](https://github.com/TeaPearce/Conditional_Diffusion_MNIST)

[4] DDIM 簡明講解與 PyTorch 實現：加速擴散模型采樣的通用方法

<https://zhouyifan.net/2023/07/07/20230702-DDIM>

[5] Textual-inversion fine-tuning for Stable Diffusion using d<sup>ffusers</sup>

[https://github.com/huggingface/notebooks/blob/main/diffusers/sd\\_textual\\_inversion\\_training.ipynb](https://github.com/huggingface/notebooks/blob/main/diffusers/sd_textual_inversion_training.ipynb)

[6] CLIP

<https://github.com/openai/CLIP>

[7] Multi-Concept Customization of Text-to-Image Diffusion

<https://arxiv.org/abs/2212.04488>