國立台灣海洋大學資訊工程學系專題報告

題目

Image-Guided Image Style Transfer with Diffusion model

作者

00957134	黃永毅	00957134@mail.ntou.edu.tw
00957050	張銀軒	00957050@mail.ntou.edu.tw
00957126	賴柏勛	00957126@mail.ntou.edu.tw

報告編號:NTOUCSE 112 -丁培毅- 競賽組 02

指導教授：丁培毅 博士

中華民國 112 年 12 月 10 日

摘要

擴散模型在圖像風格轉換中的表現一直都很出色，但最大的問題是模型本身速度不快，且擴散模型的隨機性也影響了產出的內容。大部分現有的方法需要對擴散模型進行微調，或者用額外的神經網絡。而我們使用了一種不需要額外訓練，直接使用額外的 loss function 來試圖將預訓練擴散模型的輸出導向到想要的方向。通過這種方法來提高使用者建構的速度，而不用花太多時間來微調擴散模型，並與其他現有不同的圖像風格轉換方法來做比較。

前言

風格轉換指將給定圖像的風格轉換為另一種風格，同時保留其內容。過去幾年有許多基於 GAN 的方法。而最近，使用預先訓練圖像生成器和圖像文本編碼器(encoder)的讓網路本身不需要或只用很少的訓練就能達到文字引導風格轉換。

近年擴散模型在圖像生成、修改的方面展現出了極高的品質，也有許多人使用擴散模型搭配不同的方法達到風格轉換，

文獻回顧

圖片風格

在 Neural style transfer [1] 中提到，透過 CNN 中各層的 fliter 擷取圖片在不同特徵上的 activation value，把這些 activation value (每一個特徵的 activation value 皆為 2 維矩陣)攤平成 1 維矩陣，接著把這些 1 維矩陣合併成 2 維矩陣，再透過與該 2 維矩陣之轉置矩陣的乘積作為該圖片特徵之間的關係，稱為 gram matrix，便能代表圖片的風格。

diffusion

Diffusion model 是一種生成模型，其主要優點為可以生成高質量的樣本，模型在 forward process 逐步往圖片添加隨機高斯雜訊，而在 reverse process 時慢慢去除雜訊進而產生結果，而這個過程需要一個 score function(noise predictor)來引導如何去除雜訊。具體來說，以 DDPM(Denoising Diffusion Probabilistic Models)為例，DDPM 在時間 t ∈ [1, ..., T]依序添加雜訊於原始圖片 X_0 上，而 X_t 可視為添加 t 次高斯雜訊的圖片，其中 q(Xt|Xt-1)為一次 forward process ，代表以 X_(t−1)加上一次雜訊來產生 X_t，而這也是一個 Markov chain，

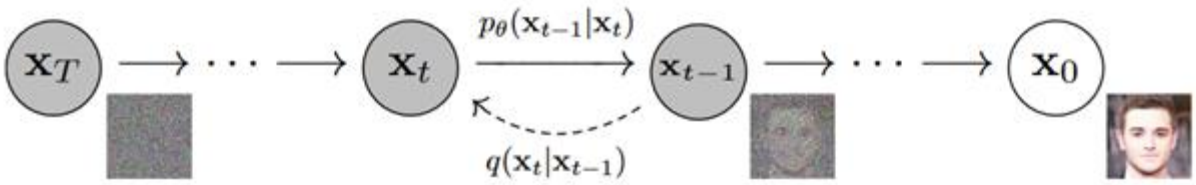


Figure 2: The directed graphical model considered in this work.

而β能影響增加雜訊的步驟，在原文中β是從 0.0001~0.02 的 Linear schedule。而後也有文章使用 cosine schedule 改善了 forward process 中，圖片資訊破壞過快的問題。

而 reverse process 則是需要一個 noise predictor 來預測雜訊並加以刪除，而 reverse process 也是一個 Markov chain。有了 noise predictor，我們就可以在每個 time step 將預測的高斯雜訊從圖片中扣除，最後得到一張乾淨的圖片。

研究方法

我們使用 OpenAI 開發的 guided diffusion，其優點為在 sampling 時對 Diffusion Model 的輸出進行條件約束，而無需為每個具體情境重新訓練網絡，而 loss 可分為保留圖片內容的 content loss 以及確保風格轉換正確性的 style loss

總損失 (Total Loss)：

$$L_{total} = L_{content} + L_{style}$$

content loss 可分為三部分

$I_{original}$ 代表原始圖片

$I_{generated}$ 代表生成圖片

1. 計算原始圖片與生成圖片的 MSE

   $L_{content1} = \frac{1}{N*M}\sum_{i=1}^{N}\sum_{j=1}^{M}(I_{original}[i,j]-I_{generated}[i,j])^2$
2. 計算原始圖片與生成圖片在 VGG feature map 中的 MSE

   $L_{content2} = \frac{1}{K*L}\sum_{k=1}^{K}\sum_{l=1}^{L}(F(I_{original})[k,\ l]-F(I_{generated})[k,\ l])^2$
3. 根據 ZeCon 在文字導向風格轉換裡的主要貢獻

$$L_{Zecon}(\widehat{x_0}, t, x_0) = E_{x_0}\left[\sum_l \sum_s \mathrm{cl}\left(\widehat{z_l^s}, z_l^s, z_l^{S\backslash backslashs}\right)\right]$$

$$L_{content} = L_{content1} + L_{content2} + L_{Zecon}$$

style loss 結合了文字導向風格轉換使用的 CLIP score，改為將兩張圖片放進 CLIP Space 進行比較，以及風格轉換常用的計算兩張圖片的 Gram matrix 再計算相似度的方法，α 和 β 代表兩種做法的權重

$$L_{style} = \alpha \cdot C_{score} + \beta \cdot G_{score}$$

1. 使用 CLIP Score 比較兩張圖片在 CLIP 空間中的相似度

   $C_{score} = C(I_{original}, I_{generated})$
2. 使用 Gram Matrix 計算兩張圖片的風格相似度

   $G_{score} = G(I_{original}) * G(I_{generated})$

實驗結果

資料來源

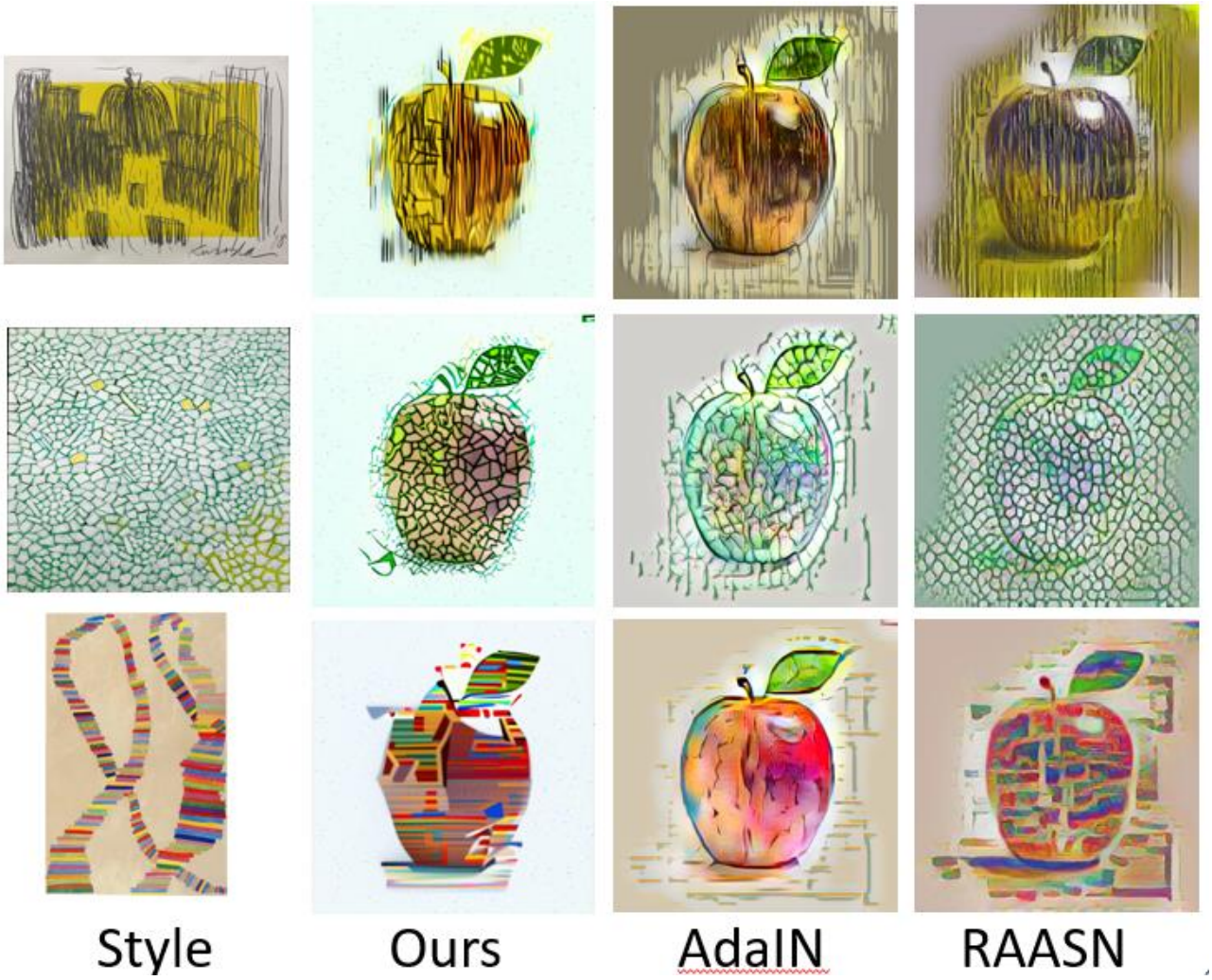
我們從 Places365 隨機挑選 20 張圖片作為 content image，style images 則是從 WikiArt 隨機選擇 4 種風格，每種風格再隨機選擇 100 張做為 style images

跟其他方法的比較

我們選擇下列的方法進行比較

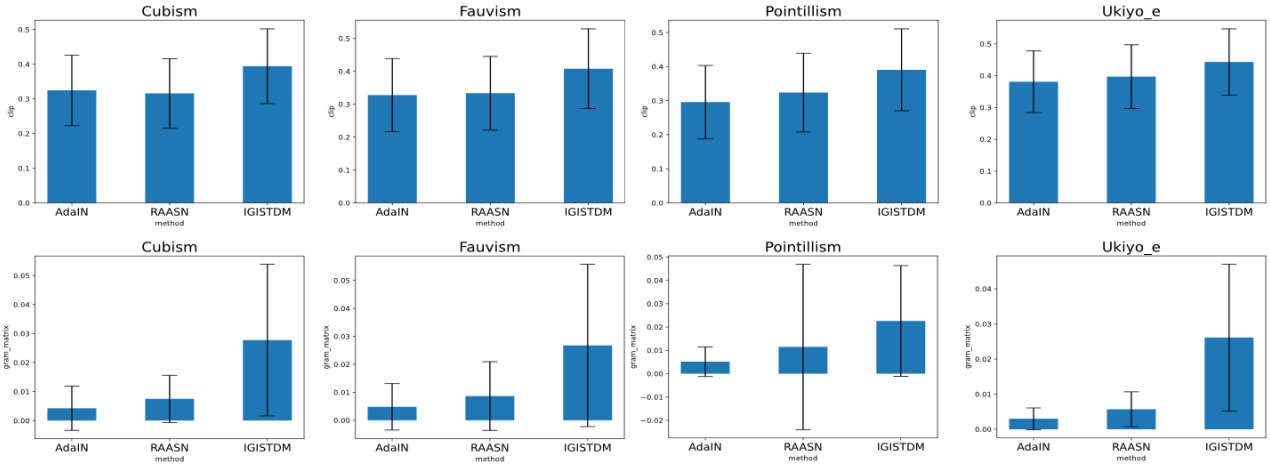
1. AdaIN[3]
2. RAASN[4]

轉換結果



分析

我們分別使用 ZeCon[2] 所使用的 CLIP score 以及 gram matrix 之間的 MSE 計算轉換結果與原風格圖片的 Style loss ，如下圖所示



結論

在本研究中，我們使用了 OpenAI 預訓練的 Guided Diffusion 作為生成模型，在不微調模型的前提下，透過設計 Content loss 以及 Style loss 來影響生成結果，達到風格轉換的目的，節省了以往需要訓練模型所消耗的資源。

在研究進度中，我們遇到了使用 CLIP 作為 Style loss 時的挑戰，由於 CLIP 將整個風格導向圖片 encode 進 CLIP space，導致生成結果可能包含風格圖片的內容或輪廓。此外，CLIP 對顏色訊息的較少理解也是一個問題，使生成結果的顏色相對保守。我們也觀察到 Gram matrix 在這方面有一定的優勢，能夠補足 CLIP 的不足，提供更豐富的顏色訊息。

雖然目前的結果顯示我們的方法在數值上較其他做法差，但這也為未來的優化提供了方向。我們認為對於超參數設計還有改進的空間，在 Content loss 和 Style loss 之間取得更好的平衡，將會對生成的結果有碩大的影響。此外，也能嘗試微調生成模型，以尋找此作法潛在於特定領域的應用。

總體而言，雖然目前的結果不如預期，但這次研究為相關領域提供了一個有價值的參考範例，並且未來也能透過優化相關參數以及 loss 架構的探討來更近此作法。

參考文獻

[1] L. A. Gatys, A. S. Ecker and M. Bethge, “Image Style Transfer Using Convolutional Neural Networks,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2414-2423, doi: 10.1109/CVPR.2016.265.
[2] Zero-Shot Contrastive Loss for Text-Guided Diffusion Image Style Transfer, arXiv:2303.08622v2
[3] Arbitrary style transfer in real-time with adaptive instance normalization, arXiv:1703.06868v2
[4] Exploring the structure of a real-time, arbitrary neural artistic stylization network, arXiv:1705.06830v2

專題分工及貢獻度說明

編號	姓名	主要工作內容	專題貢獻度 (100%)
1	黃永毅	風格轉換實作、報告文章(摘要、前言、文獻回顧-diffusion、研究方法、結論)、風格轉換 700 張圖片	50%
2	張銀軒	報告文章(文獻回顧-圖片風格、實驗結果)、實驗(設計、風格轉換 7300 張圖片、進行其他 2 種方法的實驗(16000 張圖片)、結果圖表分析)、成果網站、海報製作	40%
3	賴柏勛	風格轉換 loss function	10%