如何生成逼真影片？了解 Sora 模型訓練關鍵

Sora 是 OpenAI 於日前推出的文字轉影片 AI，可以生成最長一分鐘、不同長寬比、不同解析度的影片，因 OpenAI 釋出的生成影片極為逼真，Sora 成為了最近熱門話題之一。接著就讓我們來看看 Sora 模型背後的原理吧！

diffusion transformer

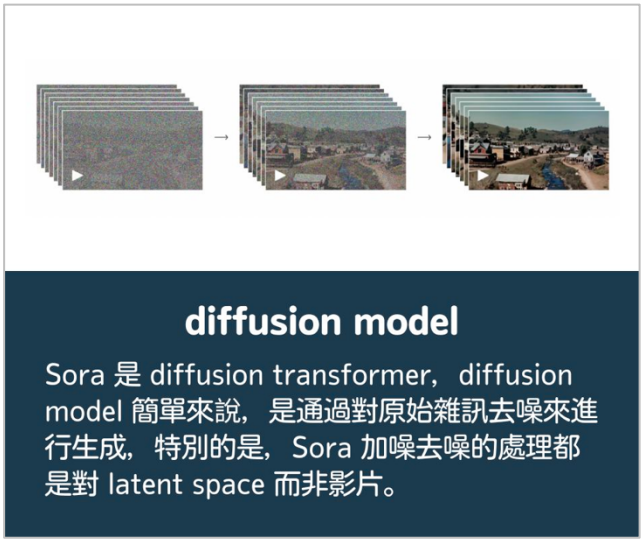
Sora 是 diffusion transformer，執行流程為輸入雜訊 patches 與條件資料 (比如 prompts)，接著預測「乾淨」patches。diffusion model 擴散模型簡單來說，是通過對原始雜訊去噪來進行生成，特別的是，Sora 加噪去噪的處理都是對 latent space 而非影片。

開發團隊訓練了一個對視覺資料做降維的網絡，input 為原始影像（由於圖片是一幀影片，所以模型的訓練資料與生成內容可以是圖片），output 為被壓縮的 latent representation。此外還有訓練一個相應的 decode 模型，來將被生成的 latents 拼湊還原為 pixel space。

transformer：將影像變為 patches 補丁們

許多大型語言模型 (LLM) 訓練大量資料都是基於 transformer，transformer 的特點就是可以在處理序列資料時，考慮序列中所有的元素 token 並依其重要性給予不同的權重。以 LLM 為例，transformer 可以評估一個句子中所有單詞的重要性並賦予權重，藉此可以處理各式文本資料如程式碼、數字、其他自然語言。

而 Sora 開發團隊參考了這樣的概念，相對於 LLM 的 token 之於文本，Sora 有 patches 補丁們之於影像。OpenAI 表示，無論是訓練影片還是圖片的生成模型，patches 化的影像資料都可以更有效地被模型理解，而且 patches 有高度的可重構性。



Sora 模型訓練關鍵

Sora 的資料處理方法是先 encode 壓縮影片資料到低維 latent space，再將其分解為 spacetime patches，接著平面化 spacetime patches，然後開始進行 transformer 訓練。Sora 可以輸入和輸出不同解析度、時長與長寬比影片的關鍵，就在於不同序列長度的 spacetime patches。

當然，極為大量的資金、人力、資源等也是必不可少的。

參考資料 | OpenAI Sora Research: Video generation models as world simulators