Speech resynthesis from discrete disentangled self-supervised representations

A. Polyak, et al., “Speech resynthesis from discrete disentangled self-supervised representations,” arXiv:2104.00355 [cs.SD], Jul. 2021

Student : Sian-Yi Chen

Advisor : Tay-Jyi Lin and Chingwei Yeh

Outline

1. Introduction
    - Disentangled representations
2. Architecture

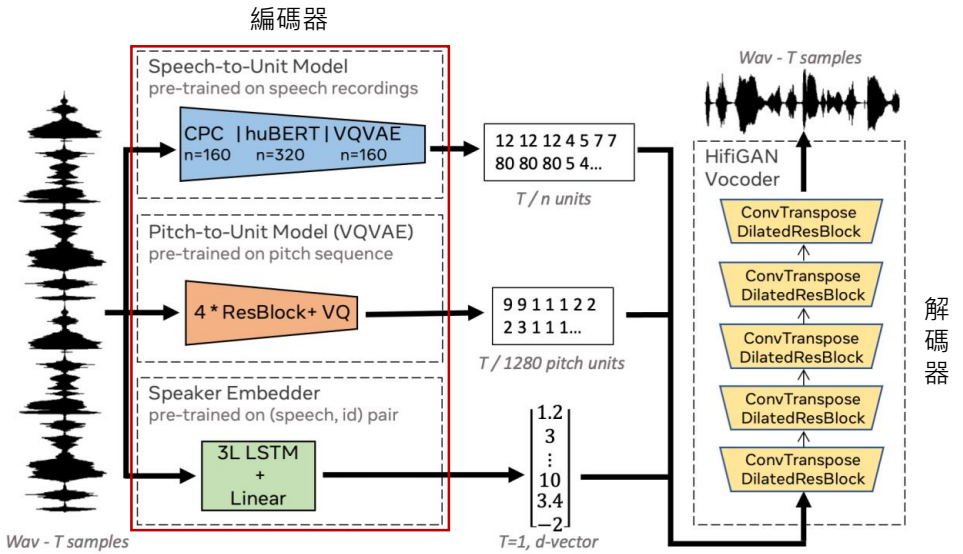
Introduction

論文中的語音轉換使用了自監督學習 (SSL, Self-Supervised Learning) 從原始語音中提取特徵，取代傳統 Mel-頻譜圖的輸入。

語音轉換架構中(圖一)使用了三個編碼器以及一個解碼器，三個編碼器分別從原始語音中提取了[1]語音內容、[2]基本頻率(聲音中頻率最低的音)以及[3]說話者身分，最後再使用解碼器重構訊號。

最後論文表示完成了速率為 365bits/s 的超輕量級語音轉換器。

本篇論文重點則著重在解開編碼器學習到的特徵 (disentangled representations)。



(圖一) 語音再合成架構

Disentangled representations

自監督學習為透過 Model 自動學習，Model 會從輸入資料中學習到 representations (表徵)並透過向量、序列…等數據呈現，而 disentangled representations 是將這些數據分解或解開成可理解的特定特徵，目標是使用 ”高“ 或 “低” 維來模仿人類的快速直覺過程。

舉例來說，在辨識人像照片的預測神經網路中，而模型會透過輸入的圖片生成各種特徵 (e.g. 每個人的身高、手或腳的長度、衣服類型…等)，disentangled representations 會提取其中有用的特徵並使用身高或是衣服這些特定特徵來辨識性別。

Architecture

Wav - T samples

$x = (x_1,..., x_T)$

輸入：原始訊號

$E_c(x) = (v_1,..., v_{T'})$

Speech-to-Unit Model
pre-trained on speech recordings

CPC | huBERT | VQVAE
n=160 n=320 n=160

輸出：以低頻的頻譜採樣序列

12 12 12 4 5 7 7
80 80 80 5 4...

T / n units

目的：尋找輸入的表徵並學習語音內容

CPC
HuBERT
輸出皆為連續的 representation (表徵) → K-means → 轉換成離散表徵 $Z_c = (Z_1,...,Z_L)$

VQ-VAE → 本來就是離散表示

輸入：使用 YAAPT 演算法提取原始訊號

$p = (p_1,..., p_{T'})$

E_{F0} / D_{F0}

Pitch-to-Unit Model (VQVAE)
pre-trained on pitch sequence

4 * ResBlock+ VQ

輸出：低頻基本頻率 (F0) 序列

$Z_{F0} = (z_1,..., z_{L'})$

9 9 11 1 2 2
2 3 1 1 1...

T / 1280 pitch units

目的：尋找輸入的表徵並取出基本頻率

輸入：原始訊號 ($E_{spk}(x)$)

Speaker Embedder
pre-trained on (speech, id) pair

3L LSTM + Linear

輸出：特徵向量 (Z_{spk})

1.2
3
⋮
10
3.4
−2

T=1, d-vector

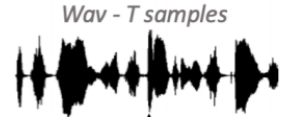
目的：用 Mel-頻譜圖 提取聲音特徵並判斷聲音對象是誰

Z_c、Z_{F0} 透過 embedding (查表的方式)從離散轉回連續表徵並相接

Z_{spk} 透過 embedding 轉換成連續向量與每一 frame 相接

輸入：三個編碼器的輸出

Z_c、Z_{F0}、Z_{spk}



Wav - T samples

輸入：語音再合成

HiFiGAN Vocoder

ConvTranspose DilatedResBlock
ConvTranspose DilatedResBlock
ConvTranspose DilatedResBlock
ConvTranspose DilatedResBlock
ConvTranspose DilatedResBlock

附錄

Unsupervised

論文中使用的三種皆為無監督學習的神經網路

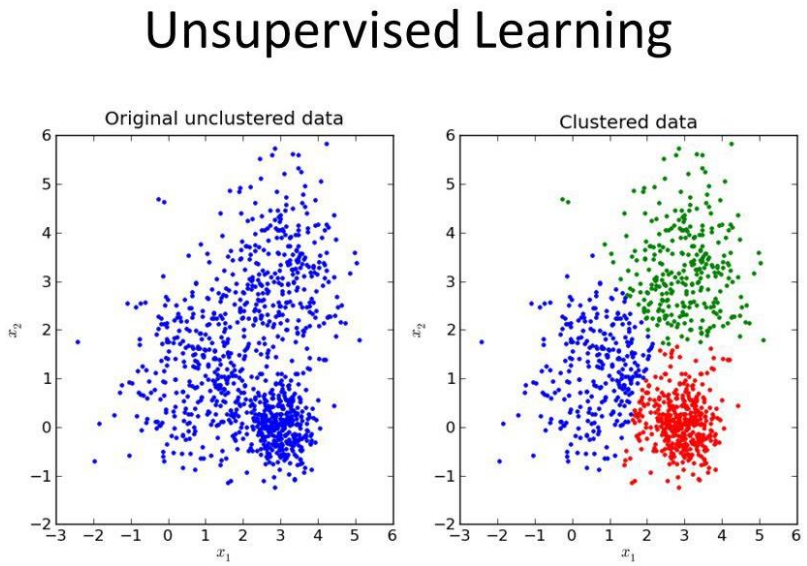
CPC (Contrastive Predictive Coding)

HuBERT (Hidden Unit Bidirectional Encoder Representations from Transformers)

VQ-VAE (Vector-Quantized Variational AutoEncoder)

無監督 (unsupervised) 神經網路特性：模型自行對於數據學習，找到數據間的關係，並將資料分成多群組，稱為聚類。

每一分群無標籤類型。

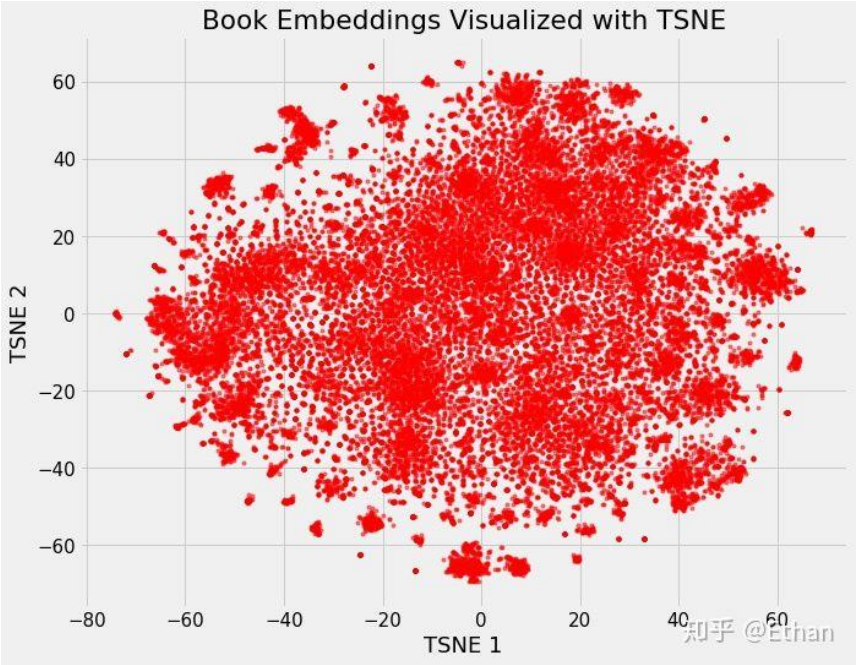


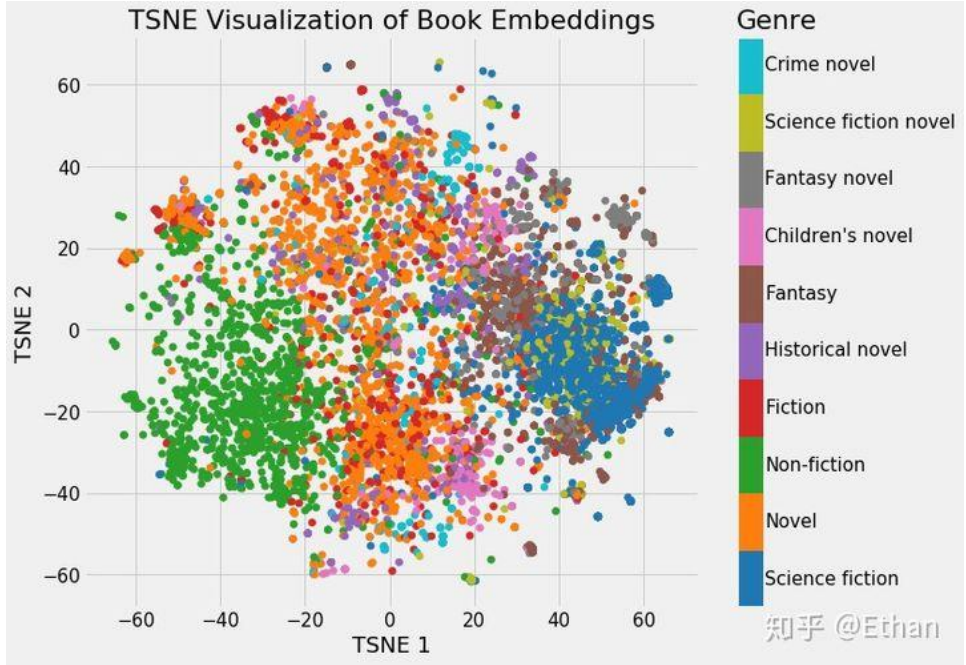
參考資料：

1. Google AI Blog: Evaluating the Unsupervised Learning of Disentangled Representations
2. 【自監督學習】Self-supervised Learning 再次入門_osc_o9qsdhyv - MdEditor

Embedding

Embedding 是一種將離散變量轉換成連續向量的一種方式，在神經網路中 Embedding 不僅可以降低離散空間的維度，還可以使各種變量具有意義。
舉裡來說，假設維基百科上所有書籍原始維 37,000 維，先使用 embedding 將資料映射至 50 維，再使用降維技術 t-Distributed Stochastic Neighbor Embedding (TSNE)，將資料維度降至 2 維，如下圖 (左)
接著將數據中各種特徵填上不同顏色，我們將可以明顯發現效果顯著。





參考資料：Embedding 的理解 - 知乎

Discrete data

離散資料顧名思義就是非連續性資料，連續訊號可以是任何值的數值，像是身高、體重、溫度和長度皆是連續的例子。
而將連續性資料轉換成離散資料有什麼意義？

離散數據的優點：

1. 提高計算效率
2. 分類模型計算需求
3. 圖像處理中的二值化處理
4. 對連續資料數據更方便理解
5. 距離計算模型 (K均值、協同過濾)中降低異常數據對模型的影響，使模型達到更穩定的效果

參考資料：資料離散化及其KMeans演算法實現的理解 - 知乎

Self-Supervised Learning

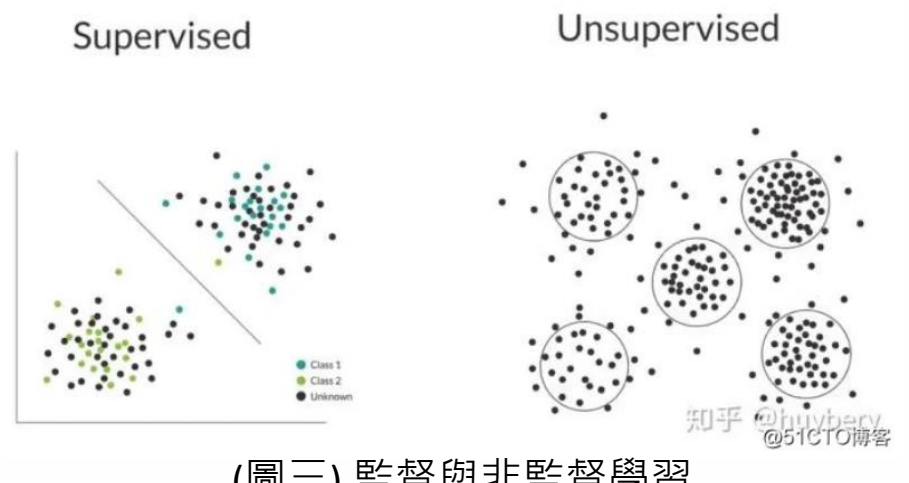
機器學習中的有兩種基本的學習正規化

1. 監督學習
2. 非監督學習

監督學習透過人工大量的標註資料來訓練模型，模型的預測和資料的真實標籤產生後進行反向傳播，通過不斷學習，最後獲得辨識新樣本的能力

非監督學習不依賴任何標籤，自行對資料內特徵挖掘，找到樣本間的關係，如上圖，左邊有三種型態，class1、class2還有未知，而右邊分成5區卻沒有標示任何訊息

兩種最主要的差別在於訓練時是否需要標記標籤資訊



(圖三) 監督與非監督學習

那什麼是自監督學習？

自監督學習主要利用**輔助任務**從大規模非監督學習的資料中自己去挖掘監督資訊。
在非監督學習下，會產生多筆集合，而輔助任務會將這些沒有標籤的集合，選出需要的集合並標上標籤。

也就是自監督學習具有傳統手動取得特徵的優點，又具有非監督學習透過模型自行學習的優點。

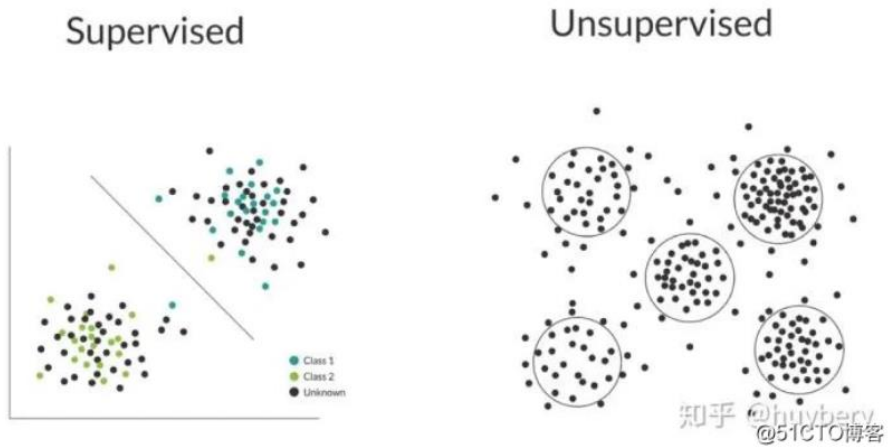
(圖四) 標記蘭花

分類？聚類？

分類與聚類，可以簡單的使用屬於監督學習，或是非監督學習來區分。

分類 (classification)：
根據現有資料的特徵或是屬性去做區分，也就是說這些類別是已知的，因此如果是在監督學習下，就屬於分類。

聚類 (clustering)：
在非監督學習下，不知道模型會將資料分成多少類，分類的標準也是未知的，但每一聚類都是由物理性質或抽象性質相似的類別所組成



【自監督學習】Self-supervised Learning 再次入門