Separate Vocal from Background Music Using One Dimensional Cycle GAN

JHIH WEI, YU
(于治維, C54076055)
Department of Biomedical Engineering of National Cheng Kung University
Tainan, Taiwan, ROC
c54076055@gmail.com

Abstract

隨著網路社會的發展，雖然資料的取得成本越來越低，但成對的標記資料卻不多。故深度學習模型要如何使用非成對的弱標記資料進行訓練成為非常重要的課題。這篇文章將嘗試使用 cycle GAN 來訓練非成對的歌曲及伴奏音檔，挑戰在於小樣本成對標記資料 (217 對歌曲)、小樣本非成對標記資料 (1526 首歌曲)、不平衡資料、輕量架構模型，不使用 FFT 頻譜圖。

Keywords: 非成對資料、小樣本資料、不平衡資料、輕量架構模型，無頻譜圖

1. Introduction

1.1. Motivation

很多時候想唱歌找不到伴唱帶，於是想說如果可以有自動生成伴奏的魔法就好了，於是就有了這個計畫。

1.2. Related work

多數在語音或歌聲分離的研究多將時間序列訊號轉換成頻譜圖，並且多使用成對資料 [2]。從過去的研究可以看出一維 convolution 組成的 unet 可能有足夠的能力解析聲音訊號 [5]，故先轉換成頻譜圖的可能不是必要的，同時將資料轉換成頻譜圖後資料容量會增加不少，由於目前的電腦資源拮据，若可以減少訓練資料的容量會是很大的幫助。

在這篇論文中 [7] 揭示了 attention 機制如何引導模型的訓練。而在這篇論文中 [3] 說明了如何將 attention 應用在 unet 中。

1.3. Challenge

本研究的困難點在於這邊訓練的資料以及測試的資料類型迥異，樣本稀少，且多為不成對資料。即便是成對資料，因為採樣的偏差，以及音訊的母帶工程，使得歌曲及伴奏相減後也很難得到純人聲。

此外，這裡選擇使用 one dimensional cnn 完成任務，而不使用一般的 STFT 頻譜圖

2. System framework

2.1. 資料集

所有訓練用歌曲資料分為成對資料，以及非成對資料，隨機抽取抽取 0.1 作為驗證集，剩下 0.9 作為訓練集。比特率為 320kbps，採樣率為 44.1kHz，雙聲道。

成對資料為 179 首日本動漫歌曲及其對應伴奏，但無標準人聲資料，資料來源為個人收藏的 88 張日本動漫歌曲專輯。歌曲及其對應伴奏各計 774 分鐘

非成對資料來自 Youtube 自動生成的官方 Topic 頻道，直接下載自他人整理的播放清單，根據類別分為:

K-pop instrumental: 652 首 (2260 分鐘);
K-pop: 518 首 (1764 分鐘);
Lit Rap/Hip-Hop Instrumental: 154 首 (526 分鐘);
Hip Hop: 199 首 (634 分鐘)。
總共 1523 首歌曲，5186 分鐘。

2.2. 前處理

成對資料的部分，音訊的時間序列原則上以 250ms 作為間隔，切分為 500ms 的音樂片段。並以以下算法將歌曲及其伴奏片段對齊後做為 pretrain 的資料預訓練 generator。

通過用查詢 sigmaQ 和對齊候選 sigmaS[j] 的標準差分別除以這些值來歸一化尺度。以下均值和振幅調整被稱為 z-normalization，指的是減去均值、並以方差為單位的正態隨機變量的 z-cores。相應的滾動測量稱為 zdist：[1]

$$\mathrm{zdist}(Q,S)=\min_{0\le j<n-m+1}\left(\sum_{i=0}^{m-1}\left|\frac{q[i]-\mu_Q}{\sigma_Q}-\frac{s[i+j]-\mu_S[j]}{\sigma_S[j]}\right|^2\right)^{\frac{1}{2}}$$

非成對資料的部分，音訊的時間序列也以 250ms 作為間隔，切分為 500ms 的音樂片段。將 instrumental 片段當作真實資料，原始歌曲片段當作虛假資料，兩者做為 pretrain 的資料預訓練 discriminator。

非成對資料中的原始歌曲片段也會作為 generator 的訓練集。

2.3. 模型架構

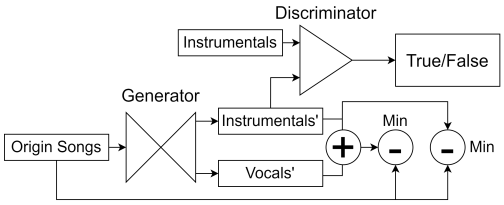

Figure 1. GAN structure

此架構是模仿 Cycle GAN 的架構，由於 generator 的輸出使用簡單求和應該要足以還原輸入，所以不需要反向的 generator。由於資料的限制，僅有 Instrumentals 的 discriminator。

2.4. loss function 設計

因為 Vocals' 沒有被控制，有可能 Vocals' 混和了部分的 Instrumentals'，即便 Instrumentals' 通過檢驗。

為了避免以上情形發生，需要讓 Instrumentals' 盡量接近 Original songs，故以下 loss function 設計考慮了此情形並效果以 alpha 控制。

2.5. 訓練流程

2.5.1 Discriminator pretrianing

由於 genertor 的設計確保了尚未訓練完成的 Instrumental' 會是 Original songs 與 Instrumentals 的線性組合，故將 instrumental 片段當作真實資料，原始歌曲片段當作虛假資料 pretrian discriminator 是合理的。

2.5.2 Generator pretrianing

使用少量成對資料 pretrain Generator。

2.5.3 GAN training

3. Expected results

1. 另外找不同類型流行歌曲做為測試集，其 Instrumental 聽起來無人聲，若有對應之伴奏，兩者之 STOI [6] 應接近 1，PESQ [4] 評分應高於 4 分

2. Vocal 之 STOI [6] 應接近 1，PESQ [4] 評分應高於 4 分

References

[1] Christian Hundt. Aligning time series at the speed of light. https://developer.nvidia.com/blog/aligning-time-series-at-the-speed-of-light/. Accessed: 2021-04-29. 1

[2] Andreas Jansson, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar, and Tillman Weyde. Singing voice separation with deep u-net convolutional networks. 2017. 1

[3] Zhen-Liang Ni, Gui-Bin Bian, Xiao-Hu Zhou, Zeng-Guang Hou, Xiao-Liang Xie, Chen Wang, Yan-Jie Zhou, Rui-Qi Li, and Zhen Li. Raunet: Residual attention u-net for semantic segmentation of cataract surgical instruments. In International Conference on Neural Information Processing, pages 139–149. Springer, 2019. 1

[4] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In 2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221), volume 2, pages 749–752. IEEE, 2001. 2

[5] Daniel Stoller, Mi Tian, Sebastian Ewert, and Simon Dixon. Seq-u-net: A one-dimensional causal u-net for efficient sequence modelling. arXiv preprint arXiv:1911.06393, 2019. 1

[6] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In 2010 IEEE international conference on acoustics, speech and signal processing, pages 4214–4217. IEEE, 2010. 2

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 1