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. What is disentangled representation？
   - Disentangled 的物理意義
   - What is latent factor (潛在因素)？
   - 目前對於 disentangled 的非正式定義
   - 應用實作框架、總結
2. What is self-supervised？
   - Supervised learning vs. unsupervised learning
   - Self-supervised task
   - Take Word2Vec for example
3. The relations between disentangled representation and self-supervised
   - 兩者在神經網路中的關係

[1] Y. Bengio, A. Courville and P. Vincent, “Representation learning: a review and new perspectives,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Aug. 2013, pp.1798-1828
[2] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations,“ in *ICML*, 2019, pp.4114-4124

Disentangled representation (1/3)

Disentangled representation 是在模仿人類認知的過程，希望學到輸入資料的高維抽象表示
目前 disentangled representation 尚未有正式的定義
最早提及 disentangled representation 在 2013 年的 [1]

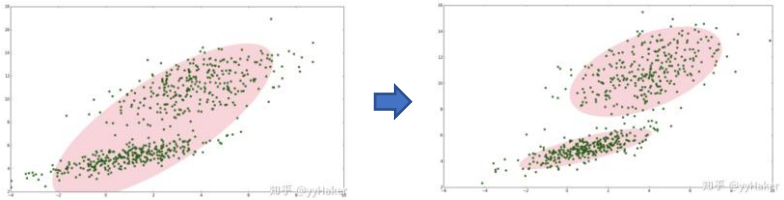
> Single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors.

描述在一直變化的資料中，具有有一些不變的 latent factor，而這些 latent factor 能表示這些一直在變化的資料分布

潛在的因素 (latent factor)：指的是一個抽象的概念，觀察不到的數值，e.g. 狀態、行為...等
舉例來說，今天有 5 部電影、4 位受訪者，評分 1~5，可以繪製出一 5x4 矩陣
Latent Factor 可以理解成每一個受試者對於電影喜歡的潛在因素，像是某一段浪漫情節、武打動作的呈現或是是由某一位演員或是某一位導演所編，而 Latent Factor 影響著受試者對電影的評分。

真正的資料分布是由這些 latent factor 所組成，有點類似混合高斯分布，目的是將這些 latent factor 都分離出來

高斯混合模型 (GMM)：高斯分布，又稱常態分布，而混合模型由多個高斯分布函數組成(圖一)，然後透過 GMM 可以分離多個高斯分布 (圖二)



圖一　　圖二

[1] Y. Bengio, A. Courville and P. Vincent, “Representation learning: a review and new perspectives,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Aug. 2013, pp.1798-1828
[2] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations,” in *ICML*, 2019, pp.4114-4124

Disentangled representation (2/3)

並在 [2] 中重新描述了 [1] 中的內容

> A disentangled representation should separate the distinct, informative factors of variations in the data.

強調不同 latent factor 之間是互相獨立，互不影響
在生成高維資料時中對某些維度會隨與之對應的 latent factor 變化而變化，不會隨其他 factor 變化而變化

> A change in a single underlying factor of variation should lead to a change in a single factor in the learned representation

接著 [2] 描述不同維度與 latent factor之間具有映射關係， latent factor 改變會導致 representation 中對應的維度取值改變

接著從生成資料的建構過程，解讀 disentangled representation：
在 representation 中，通常將 資料 x 產生過程分為兩部分：

1. 從先驗分布 P(z) 中採樣取得 latent factor (z)
2. 從條件資料生成分布 $p(x|z)$ 中採樣取得的資料觀測值為 x

並假設資料 x 由一系列有物理意義的 factor {v_1, v_2,..., v_n} 組成，通過未知的非線性映射函數 $Unknown(x)$ 作用相互耦合生成，即 $x = Unknown(v_1, v_2,..., v_n)$
而條件資料生成分布 $p(x|z)$ 即是未知的非線性映射函數 $Unknown(x)$ 的近似函數
最後再由取得的 latent factor (z) 還原資料 x

[1] Y. Bengio, A. Courville and P. Vincent, “Representation learning: a review and new perspectives,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Aug. 2013, pp.1798-1828
[2] F. Locatello et al., "Challenging common assumptions in the unsupervised learning of disentangled representations,” in *ICML*, 2019, pp.4114-4124

Disentangled representation (3/3)

實作框架：

目前幾乎所有 disentangled representation 都是使用 VAE、GAN 神經網路實作

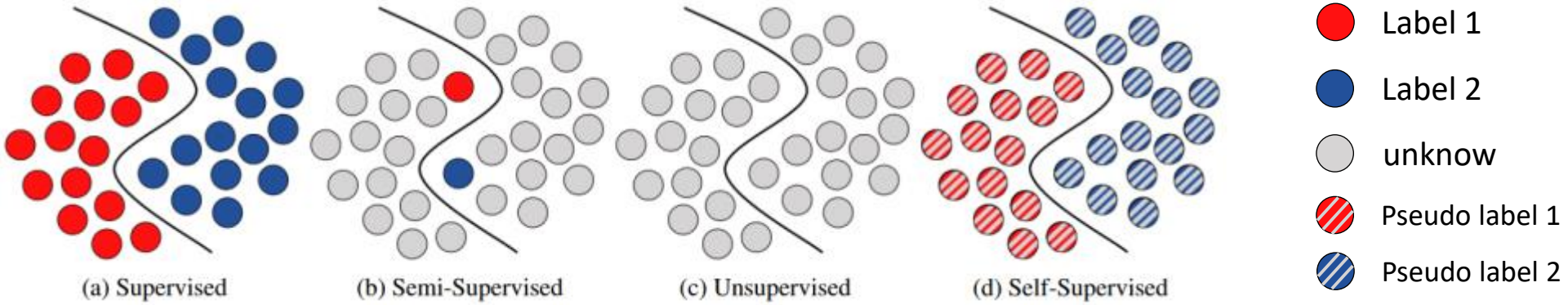
總結：

Disentangled representation 是模仿人類認知的過程，希望學習到輸入資料的高維抽象表示，在模型可解釋性、生成對象產生和控制以及零成本學習等問題上具有巨大的優勢。

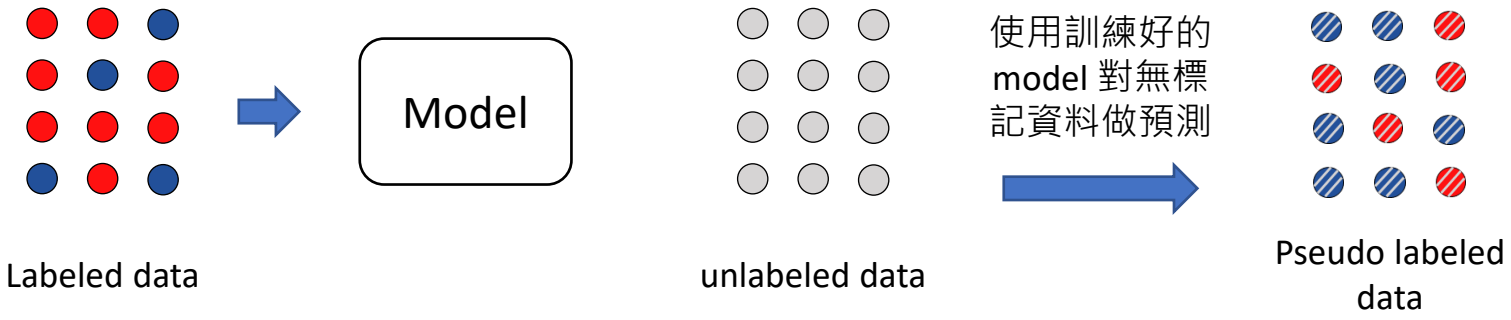
Disentangled representation 實際上是將機器學習與傳統數學模型互相結合的方法

- 機器學習：可解釋性低、資料需求大、難以零成本學習、難以對生成對象控制，並且缺乏理論基礎
- 傳統數學模型：可解釋強、資料需求低、可用微積分來解決零成本學習問題，對於生成對象控制容易完成

What is self-supervised

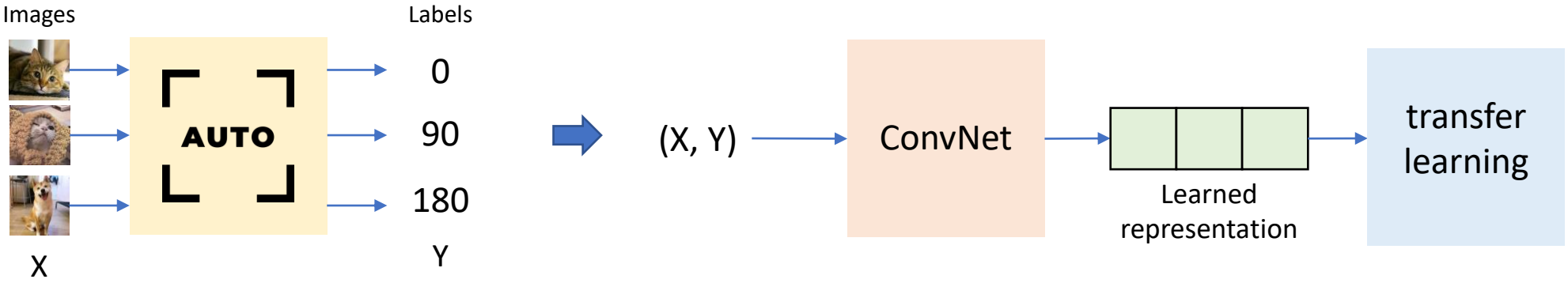


(a) 監督學習：需人工將大量的資料標上標籤。
(b) 半監督學習：標註少量資料標籤。
(c) 無監督學習：將大量的資料直接丟進神經網路，任其隨意分類。
(d) 自監督學習：透過 pretext task 自動取得 pseudo label，並透過這些 pseudo label 進行訓練；pretext task 可透過下圖理解 。



Self-supervised task

要在在深度學習中應用監督學習，我們需要有足夠多有標記的資料，但人工標記資料既耗時又而昂貴，所以將低成本的無監督學習轉變成有監督的方法就是自監督學習。



自監督學習任務設計模式：

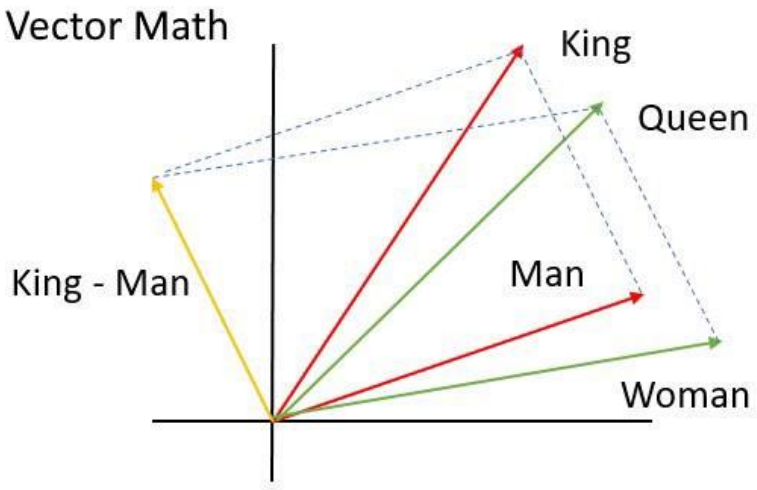
1. 根據所有待預測部分之外的資訊預測任意一部分資訊。
2. 根據過去預測未來。
3. 根據過去最近的情況預測未來。
4. 根據現在預測過去。
5. 根據底層資訊預測頂層資訊。
6. 根據可見的資訊預測不可見的資訊。
7. 假設有一部分輸入資料未知，並且對其進行預測。

常見的自監督任務：

1. 預測單詞
2. 預測上下文
3. 語句顛倒
4. Rotation (圖片旋轉)
5. Colorization (圖片上色)
6. Inpainting (圖片修補)
7. Jigsaw puzzle (關係預測)

Take Word2Vec for example

我們使用 word2vec 作為一個簡單的例子說明， word2vec 模型透過學習大量的文本資料，將字詞用數學向量的方式來代表他們的語意，也就是語意相似的單詞會有較近的距離 (圖三)，其中主要有兩種訓練模式，給定上下文，預測輸入的字詞 (圖四)，另外一種則是輸入字詞後，預測上下文 (圖五) 。



國王 (king) 減掉男人 (man) 加上女人 (woman) 會變皇后 (queen) 的話 - Word Representations in Vector Space (圖三)

通過上下文預測字詞

Google 是一個 __ __ __ 搜尋引擎

(圖四)

通過輸入字詞預測上下文

__ __ __ __ 強大的 __ __ __ __

(圖五)

The relations between disentangled representation and self-supervised

在大部分的情況下，深度學習需要在訓練資料與測試資料足夠多的形況下才能獲得較好的訓練效果，
然後很多時間我們的資料具有維度高，但 label 相對少的性質，所以在 label 少的時候，我們希望透過自監督學習幫助我們學到高維度資料中的資訊，而希望這些資訊對於下游任務有很大的幫助。

- 下游任務：在自然語言中幫助監督學習完成的任務，像是句子分類任務(語句語意上是否有效)、問答任務(預測問句答案)、專名識別任務(識別文本中具有特定意義的實體)

而透過自監督學習到的特徵，可能對於下游任務來說無法起到效果，此時就要透過 disentangled representation 來達成此目的。
藉由 disentangled representation 取得一系列與自監督學習相似的特徵，再利用這些特徵到下游任務進行訓練。