The 2018 Conference on Computational Linguistics and Speech Processing
ROCLING 2018, pp. 111-113

探討鑑別式訓練聲學模型之類神經網路架構及優化方法的改進

Discriminative Training of Acoustic Models Leveraging Improved Neural Network Architecture and Optimization Method

趙偉成 Wei-Cheng Chao, 張修瑞 Hsiu-Jui Chang, 羅天宏 Tien-Hong Lo,
陳柏琳 Berlin Chen
國立臺灣師範大學資訊工程學系
Department of Computer Science and Information Engineering
National Taiwan Normal University
{60647028S, 60647061S, teinhonglo, berlin}@ntnu.edu.tw

摘要

本論文探討聲學模型上的改進對於大詞彙連續中文語音辨識的影響。近幾年來，語音辨識技術已有了長足的進步。其中，隨著深度學習技術以及電腦運算能力的突破性發展，聲學模型化技術已從傳統的高斯混合模型結合隱藏式馬可夫模型(Gaussian Mixture Model-Hidden Markov Model, GMM-HMM)[1][2]，轉變成以使用交互熵(Cross Entropy)作為損失函數的深度類神經網路結合隱藏式馬可夫模型(Deep Neural Network-Hidden Markov Model, DNN-HMM)[3]。DNN-HMM 將以往用 GMM 計算的生成機率透過 DNN 的輸出層所代表的事後機率來近似，輸入特徵使用當前幀還有相鄰的幀，輸出則和 GMM-HMM 常用的 Triphone 共享狀態相同，以得到更低的詞錯誤率(Word Error Rate, WER)或字錯誤率(Character Error Rate, CER)。另一方面，進一步透過鑑別式訓練估測的聲學模型在語音辨識的表現上往往比僅以交互熵做為深度類神經網路損失函數的訓練方式來的好。但由於傳統上進行鑑別式訓練需要使用先進行交互熵訓練的聲學模型來產生詞圖(Word Lattices)，才能再進行下一步聲學模型鑑別式訓練[4][5]。近年來為了減少時間及空間複雜度，有學者對於 Maximum Mutual Information (MMI)訓練，提出了所謂的 Lattice-free 的方式，使產生 Lattice 的步驟能夠在 GPU 上完成[6]，因而讓鑑別式訓

練得以做到端對端的訓練方式[7]，因而大幅縮減了聲學模型訓練所需時間。傳統 DNN-HMM 模型用於語音辨識的缺點在於無法充分利用語音信號之時間依賴性；相對來說時間延遲類神經網路(Time-Delay Neural Network, TDNN)[8]可以包含歷史和未來輸出、對長時間依賴性的語音訊號建模，使 TDNN-HMM 與傳統 DNN-HMM 訓練效率也相仿，因此在使用 LF-MMI 進行鑑別式訓練時，聲學模型的類神經網路部分通常是使用 TDNN。對於 TDNN 而言，增加層數可以說是擷取更長時間的特徵；我們希望加深 TDNN 的網路層數來達到更好的結果，但以往的實驗發現深度的網路常有退化問題，類神經網路的深度之增加準確率反而會下降。因此本篇論文將比較並結合當前先進的聲學模型訓練方法，例如[9]。對網路的矩陣分解訓練可以使網路訓練更穩定，以期達到更佳的語音辨識表現。另一方面，梯度下降是執行優化的最流行的算法之一，也是迄今為止優化類神經網絡的最常用方法。而常見的優化算法有隨機梯度下降法(Stochastic Gradient Descent, SGD)、RMSprop、Adam、Adagrad、Adadelta[10]。等演算法；其中，SGD 算法在語音辨識任務上最被廣為使用。而本論文則採用來回針法(Backstitch)[11]。做為模型優化的演算法；它是一種基於 SGD 上的改進，希望能夠藉由兩步驟的更新 Minibatch，以達到更好的效果。在中文廣播新聞的辨識任務顯示，上述改進方法的結合能讓 TDNN-LF-MMI 的模型在字錯誤率(Character Error Rate, CER)有相當顯著的降低。

關鍵詞：中文大詞彙連續語音辨識、聲學模型、鑑別式訓練、矩陣分解、來回針法

參考文獻

[1] Lawrence R. Rabiner et al., “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE, 1989*

[2] Mark Gales and Steve Yang, “The application of hidden markov models in speech recognition,” *Foundations and Trends® in Signal Processing, 2008*

[3] Geoffrey Hinton et al., “Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups,” *IEEE Signal processing magazine, 2012*

[4] Lalit Bahl et al., “Maximum mutual information estimation of hidden markov model parameters for speech recognition,” *IEEE ICASSP, 1986*

[5] Karel Veselý, et al., “Sequence-discriminative training of deep neural networks,” *Interspeech, 2013*

[6] Daniel Povey et al., “Purely sequence-trained neural networks for ASR based on lattice-free MMI,” *Interspeech, 2016*

[7] Hossein Hadian et al., “End-to-end speech recognition using lattice-free MMI,” *Interspeech, 2018*

[8] Alexander Waibel et al., “Phoneme recognition using time-delay neural networks,” *IEEE Transactions on Acoustics Speech, and Signal Processing, 1989*

[9] Daniel Povey et al., “Semi-orthogonal low-rank matrix factorization for deep neural networks,” *Interspeech, 2018*

[10] Sebastian Ruder, “An overview of gradient descent optimization algorithms,” arXiv , 2016

[11] Yiming Wang et al., “Backstitch: counteracting finite-sample bias via negative steps,” *Interspeech, 2017*