DSP Final project

資工四 b03902125 林映廷

HYBRID SPEECH RECOGNITION WITH DEEP BIDERECTIONAL LSTM

概述

將深度雙向 LSTM(DBLSTM)作為在和 HMM 的混合系統中的聲音模型，並且在 TIMIT 和 WSJ 兩個資料上做測試，看其各表現如何。

模型架構

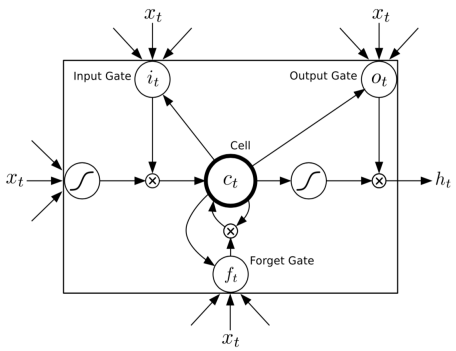

Fig. 1. Long Short-term Memory Cell

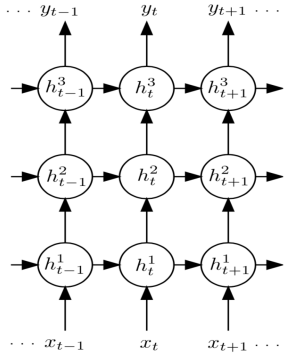

Fig. 3. Deep Recurrent Neural Network

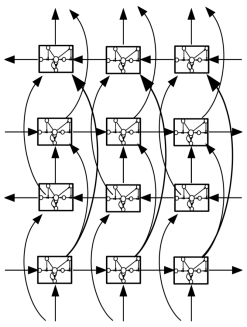

Fig. 4. Deep Bidirectional Long Short-Term Memory Network (DBLSTM)

Fig.4 為基本的主要模型架構，圖中的每一個單元皆為 Fig.1 組成，再從 Fig.3 最原始的架構延伸所得。單向 LSTM 只能捕捉到單一方向的資訊，但如果是雙向 LSTM 前後文的資訊都能捕捉到，亦即雙向 LSTM 的資訊量比較豐富、比較詳細。此外，深度也是關鍵，能在音訊資料上得到更高等級且抽象的樣貌。

模型訓練

以 cross entropy 交叉熵為目標函數，最小化 cross entropy 來得到最佳的

weight。

$$-\log \Pr(\boldsymbol{z}|\boldsymbol{x}) = -\sum_{t=1}^{T} \log y_t^{z_t}$$

其 cross entropy 的表示式如上，是負的 log 機率值，微分後如下。

$$-\frac{\partial \log \Pr(\boldsymbol{z}|\boldsymbol{x})}{\partial \hat{y}_t^k} = y_t^k - \delta_{k,z_t}$$

在 gradient descent 的過程中，由 SGD、backpropagation，以及上式的幫助下，逐步找到最佳的 weight。此外，有些實驗會加上 Gaussian noise，以達到簡化模型的目的，類似於 regularization 的功用。

TIMIT 和 WSJ 的實驗結果

Table 1. TIMIT Results with End-To-End Training.

TRAINING METHOD	DEV PER	TEST PER
CTC	19.05 ± 0.11	21.57 ± 0.25
CTC (NOISE)	16.34 ± 0.07	18.63 ± 0.16
TRANSDUCER	15.97 ± 0.28	**18.07 ± 0.24**

Table 2. TIMIT Results with Hybrid Training.

NETWORK	DEV PER TEST PER	DEV FER TEST FER	DEV CE TEST CE
DBRNN	19.91 ± 0.22 21.92 ± 0.35	30.82 ± 0.31 31.91 ± 0.47	1.07 ± 0.010 1.12 ± 0.014
DBLSTM	17.44 ± 0.156 19.34 ± 0.15	28.43 ± 0.14 29.55 ± 0.31	0.93 ± 0.011 0.98 ± 0.019
DBLSTM (NOISE)	16.11 ± 0.15 **17.99 ± 0.13**	26.64 ± 0.08 **27.88 ± 0.16**	0.88 ± 0.008 **0.93 ± 0.004**

將 CTC 和 Transducer 在 TIMIT 上測試的結果作為對照組，和以 hybrid 訓練的 DBRNN 還有 DBLSTM 作為實驗組，會發現目前所有訓練方法中表現最好的是 Transducer。

Table 3. WSJ Results. All results recorded on the dev93 evaluation set. 'WER' is word error rate, 'FER' is frame error rate and 'CE' is cross entropy error in nats per frame.

SYSTEM	WER	FER	CE
DBLSTM	**11.7**	30.0	1.15
DBLSTM (NOISE)	12.0	**28.2**	**1.12**
DNN	12.3	44.6	1.68
SGMM [20]	13.1	–	–

sGMM 是在 WSJ 上測試的訓練方法中的基準線。DBLSTM 表現得比 DNN 和 sGMM 還要好。

結論與展望

希望未來可以在語言模型影響比較小的資料上訓練。此外，找出為何目標

函數和解碼後的表現會有如此的差距。

心得

可以了解更多 DBLSTM 的應用，還有延伸，對於 DBLSTM 有更多的了解。可以看出筆者在這方面研究顯著，其實還有另外一篇相關的論文，也是同一批筆者所撰寫，希望再找一天把那一篇也看一看，並且嘗試著把論文中提到的方法自己也能實現一遍，更能體會筆者的感受。

REFERENCE

A.Graves, N.Jaitly, A.Mohamed, “HYBRID SPEECH RECOGNITION WITH DEEP BIDERECTIONAL LSTM”