Evaluation of small dataset training with GSCD & SpecAugment

Student : 陳憲億、胡祐嘉
Advisor : Chingwei Yeh and Tay-Jyi Lin

Outline

Action item：設計實驗證明 SpecAugment 方法是有成效的

- **實驗目的：**使用 SpecAugment 對小型數據做增量，並使用 Hello edge 證明此方法有效
- **實驗方法 (流程)：**
  1. 將 GSCD 12 種關鍵字中的“yes”語料從 4000 筆開始下降至 250 筆 (將現有的語料隨機刪除至需求筆數)，並依序將 GSCD 放入 Hello edge 中做訓練，訓練結束後，使用原始的 GSCD 做測試，最後由產生出來的混淆矩陣來計算每次 yes 的辨識率，藉以找出語料數量為多少時，將不足以訓練出理想值
  2. 找到正確率嚴重下降的點後，使用 SpecAugment 將不足的語料增量到與原始語料相似數量 (≈ 4000)，統計 yes 辨識率變化
- **實驗參數：**

  (Hello edge)
  1. 使用 10 種關鍵字並分為 12 類 (silence, unknow, yes, no, up, down, left, right, on, off, stop, go)
  2. 使用 dnn 模型，大小為 144 x 144 x 144
  3. learning rate： 0.0005, 0.0001, 0.00002
  4. training step： 10000, 10000, 10000

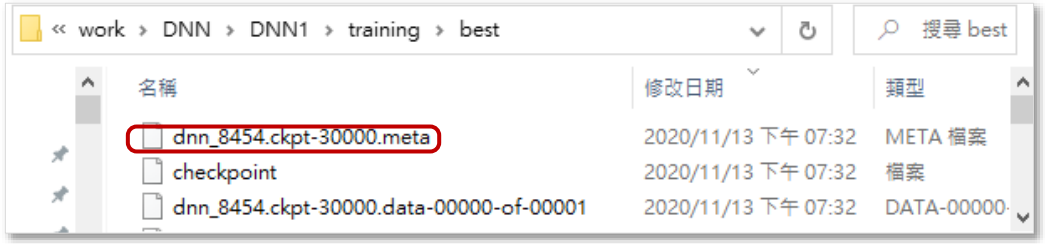
  (SpecAugment)

  時間扭曲參數 = 13, 頻率遮罩參數 = 5, 時間遮罩參數 = 17, 頻率遮罩數量 = 1, 時間遮罩數量 = 1
- **減量物理意義：**yes 語料減量後，因波型與 left 相近，因此多數被錯誤判斷為 left
- **增量物理意義：**yes 語料經 SpecAugment 增量後，可得到與原先波形有一定差異的樣本，藉以得到數據多樣性使神經網路建構更完整
- **實驗猜想：**當訓練語料不足時，辨識率產生嚴重下降，此時使用 SpecAugment 將語料做增量，辨識率可以有效上升
- **實驗結果：**假設只擁有 250 筆語料，正確率為 48% 左右，將其增量至 5000+250 筆，正確率上升為 68% 左右，推論正確率沒有像原先 85% 這麼好，是因為時間扭曲參數太大，過度扭曲變形造成訓練成效沒有很好
- **結論：**目前實驗數據並不理想，需再改變實驗參數，或是再增加訓練語料進行訓練

實驗進行流程

1 輸入指令： (spec) D:\Work_Space\KeyWordSpotting-for-MCU-master>python train.py --model_architecture dnn --model_size_info 144 144 144 --dct_coefficient_count 10 --window_size_ms 40 --window_stride_ms 40 --learning_rate 0.0005,0.0001,0.00002 --how_many_training_steps 10000,10000,10000 --summaries_dir work/DNN/DNN1/retrain_logs --train_dir work/DNN/DNN1/training --data_url="" --data_dir="./speech_dataset/"

使用複製的GSCD語料，保留原始資料做測試



執行完成，訓練數據以 checkpoint 的方式保存

2 輸入指令： (spec) D:\Work_Space\KeyWordSpotting-for-MCU-master>python test.py --model_architecture dnn --model_size_info 144 144 144 --dct_coefficient_count 10 --window_size_ms 40 --window_stride_ms 40 --checkpoint="./work/DNN/DNN1/training/best/dnn_8454.ckpt-30000" --data_url=""

```
I1114 21:14:28.288953 15500 test.py:112] Confusion Matrix:
 [[3121    0    0    0    0    0    0    0    0    0    0    0]
 [   2 2261   46   49   73  126   76  104   84   33   63  132]
 [   2   74 2801   35    3   29  143    5    2   21    7   12]
 [   1   58    9 2764   17  106   15   13    5    2   26  210]
 [   0   38    1   14 2592   18   11   20   35  123  115   37]
 [   0   56   16  104   23 2730   26    6   37    5   25   53]
 [   0   52   63   12   35   18 2709   74   10   16   10   21]
 [   0   93    1    6   15   13   81 2671   18    3    1   10]
 [   3   49    0    3   69   20    5   10 2784   81   25   11]
 [   7   22    1    2  119    4   12    9   63 2751   33   22]
 [   5   32    3    7  102   14   19    3    2   16 2896   21]
 [   2   75    6  314   63  133   20    7   11   19   36 2413]]
INFO:tensorflow:Training accuracy = 88.13% (N=36871)
I1114 21:14:28.293978 15500 test.py:114] Training accuracy = 88.13% (N=36871)
```

執行完成，產生 Training、Test 的混淆矩陣

3 GSCD語料處置： D:\Work_Space\KeyWordSpotting-for-MCU-master\speech_dataset (遞減組，用於訓練的語料)

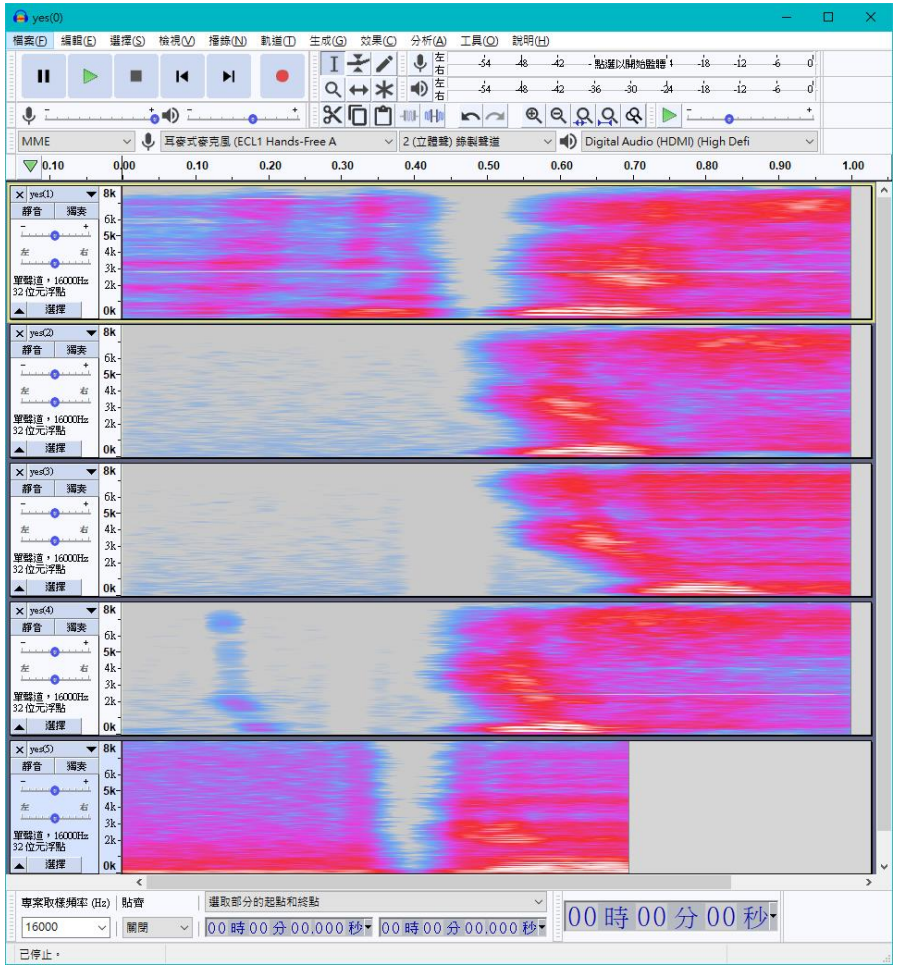
D:\Work_Space\Google_speech_dataset (保持不變組，用於測試的語料)

調整 SpecAugment 參數

```
spec_augment(mel_spectrogram, time_warping_para=80, frequency_masking_para=100,
             time_masking_para=27, frequency_mask_num=1, time_mask_num=1)
```

除6，並取 4 捨 5 入 (參數須為整數)

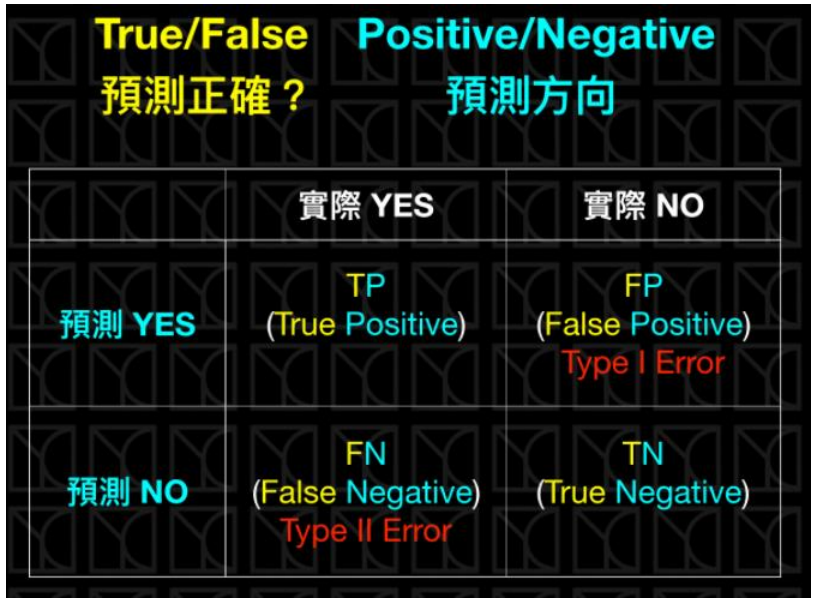
```
spec_augment(mel_spectrogram, time_warping_para=13, frequency_masking_para=5,
             time_masking_para=17, frequency_mask_num=1, time_mask_num=1)
```



使用下方參數產生的 yes 音檔頻譜圖

混淆矩陣(confusion matrix)

混淆矩陣(confusion matrix)



Accuracy = (TP + TN) / total N

計算準確率：

```
INFO:tensorflow:set_size=5032
INFO:tensorflow:Confusion Matrix:
          silence unknow yes  no   up  down left right on  off stop go
silence  [[420    0    0    0    0    0    0    0    0    0    0    0]
unknow   [  0  299   13   11   10   22    4   11   16    8   12   14]
yes      [  0   13  491    7    0    4   17    0    0    2    0    3]
no       [  0    3    8  327    0   20    6    1    0    1    0   39]
up       [  1    9    1    4  363    1    4    5    8   11   12    6]
down     [  0   12    7   30    3  322    4    3    6    1    4   14]
left     [  0   10   32    3    7    3  334   17    0    3    2    1]
right    [  0   18    3    2    3    0   13  347    5    3    0    2]
on       [  0   14    0    1    7   11    1    0  340   16    3    3]
off      [  1    6    1    3   21    0    6    4   16  336    3    5]
stop     [  0    4    1    2   19    4    4    1    2    3  365    6]
go       [  0   15    6   70    9   11    4    0    0    1    1  285]]
INFO:tensorflow:Test accuracy = 84.04% (N=5032)
```

Yes Accuracy = $\frac{491}{13 + 491 + 7 + 4 + 17 + 2 + 3}$ = 91.43%

Total Accuracy = (TP + TN) / total N
= 4229 / 5032
= 84.04%

Type I、II Error

```
INFO:tensorflow:set_size=4864
INFO:tensorflow:Confusion Matrix:
```

	silence	unknow	yes	no	up	down	left	right	on	off	stop	go
silence	406	0	0	0	0	0	0	0	0	0	0	0
unknow	0	280	11	14	4	24	8	19	12	4	10	20
yes	0	18	269	8	0	5	86	3	0	0	3	5
no	0	10	2	312	1	30	6	1	0	1	1	41
up	0	13	2	2	351	5	3	3	13	14	15	4
down	1	13	3	30	4	323	6	3	1	1	5	16
left	0	16	17	5	5	2	345	14	0	4	2	2
right	0	22	1	0	2	1	11	351	1	3	1	3
on	1	22	1	0	4	11	1	0	334	19	2	1
off	0	8	4	3	14	1	7	2	20	333	4	6
stop	0	7	3	3	16	3	1	1	1	3	365	8
go	0	12	3	67	8	13	5	4	1	2	6	281

```
INFO:tensorflow:Test accuracy = 81.21% (N=4864)
```

Error I (實際上錯誤卻被判斷為正確) = False Posite / 所有被判斷為 yes 的數量

(11+269+2+2+3+17+1+1+4+3+3) = 316

= (316-269) / 316

= 14.87%

Error II (實際上正確卻被判斷為錯誤) = False Negative / Total yes

(18+269+8+5+86+3+3+5) = 397

= (397-269) / 397

= 32.24%

物理意義

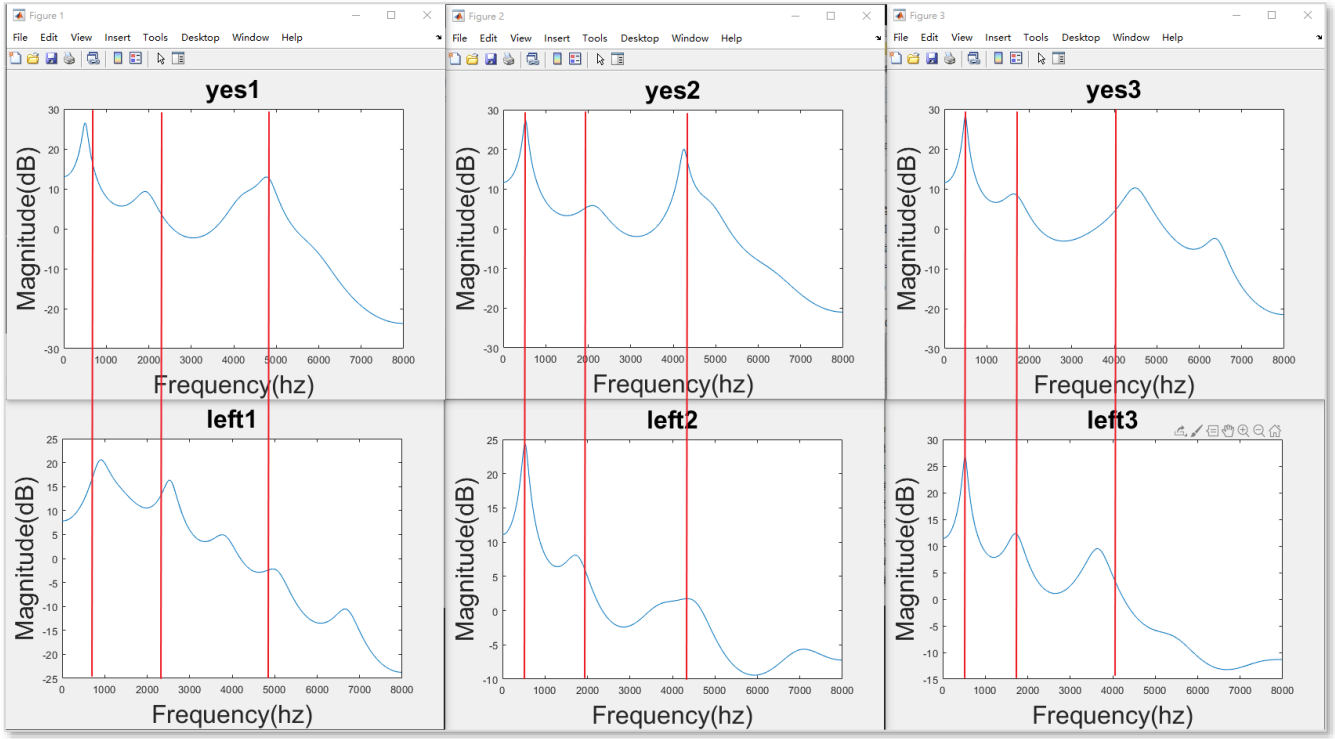
減少語料：

減少 yes 語料後，多數被判斷為 left，觀察 format 後，推論因 yes 與 left 波型相像，因此在訓練資料不足的情況下，容易被判斷為 left

SpecAugment 增量語料：

推論 yes 語料經 SpecAugment 增量後，可以得到與原先波形具有一定差異的樣本，從而得到具有多樣性的訓練資料來建構完整的神經網路模型，以達到判別準確率的上升

Left 與 yes 波形比較



Accuracy 數據分析

GSCD 遞減	4000	2500	2000	1500	1000	500	250
Training Accuracy (yes音檔)	93.45%	85.17%	89.37%	84.86%	80.65%	71.64%	53.99%
Error I (錯的被判斷為正確的)	13.77%	6.36%	6.22%	3.48%	4.94%	1.88%	1.03%
Error II (對的被判斷為錯的)	6.55%	14.83%	10.63%	15.14%	19.35%	28.36%	46.01%
Test Accuracy (yes音檔)	91.06%	84.29%	88.64%	84.05%	80.67%	72.70%	47.82%
Error I (錯的被判斷為正確的)	14.55%	7.71%	10.36%	6.70%	6.18%	2.59%	6.16%
Error II (對的被判斷為錯的)	18.94%	15.71%	11.36%	15.95%	19.33%	27.30%	52.18%

Spec 遞增	250	500 (250+**250**)	750 (250+**500**)	1000 (250+**750**)	4000 (250+**3750**)	5250 (250+**4750**)
Training Accuracy (yes音檔)	53.99%	60.70%	64.37%	67.74%	62.86%	68.33%
Error I (錯的被判斷為正確的)	1.03%	4.70%	4%	4.89%	10.13%	9.20%
Error II (對的被判斷為錯的)	46.01%	39.26%	35.63%	32.26%	37.14%	31.67%
Test Accuracy (yes音檔)	47.82%	59.69%	62.97%	67%	63.22%	67.78%
Error I (錯的被判斷為正確的)	6.16%	5.60%	6.02%	7.60%	11.31%	14.87%
Error II (對的被判斷為錯的)	52.18%	40.30%	37.02%	33%	36.78%	32.24%

	原始數量
silence	0
unknow	67283
yes	3744
no	3941
up	3723
down	3917
left	3801
right	3778
on	3845
off	3745
stop	3872
go	3880