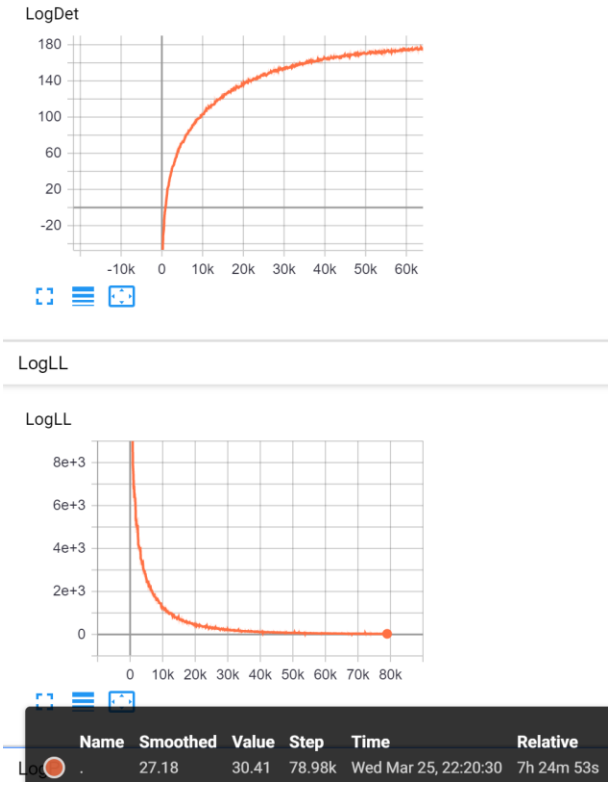
Catalogue

DNF

实验 1　　2020.3.24

Realnvp　　block=10　　hidden=512　　input=120　　cond_dim=120　var=1 & 1k



训练集在隐空间高斯性

skew = 0.0031436711142305285 kurt = -1.1428269232454675

测试集在隐空间高斯性

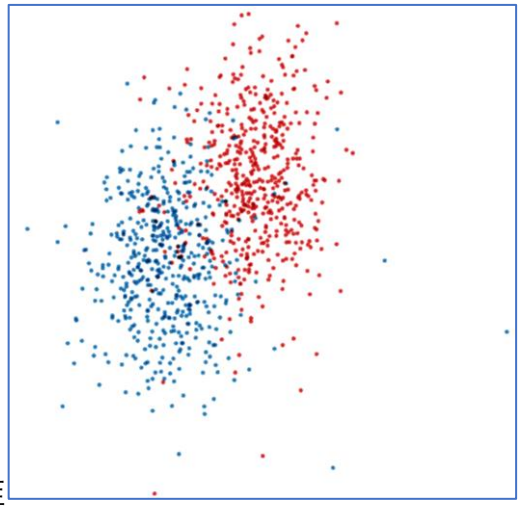
wsj+ce:

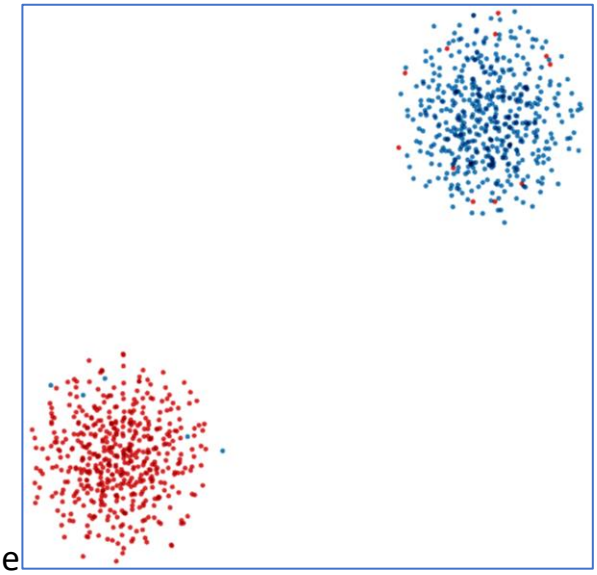
skew = 0.02316198189194741 kurt = -0.7585778409715496

wsj+ce+je:

skew = 0.02062119066443605 kurt = -0.8155078098045876

训练集可视化：

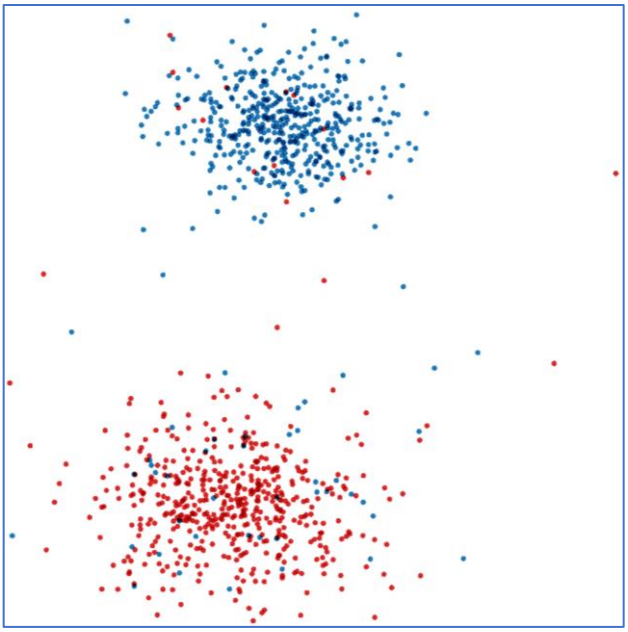
直接取两维 

T-sne 

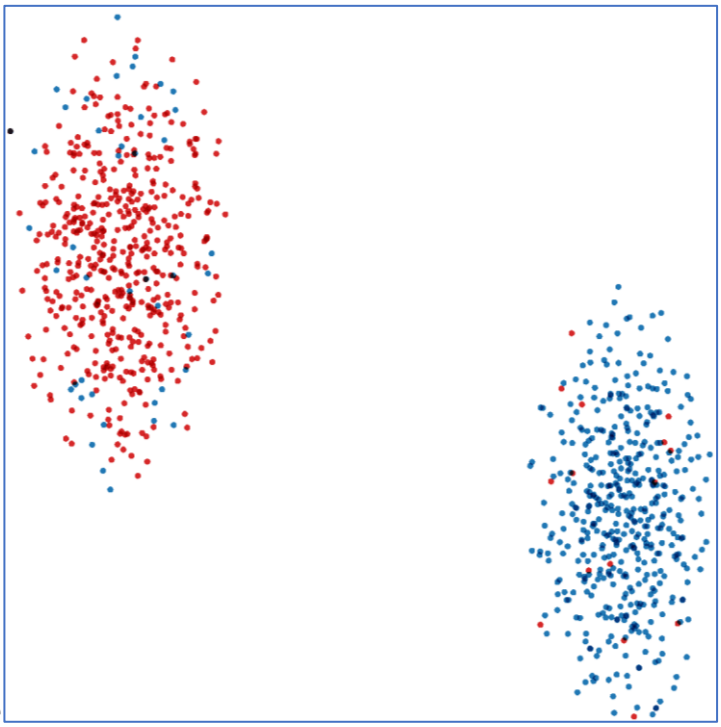
测试集可视化：

用纯英语和中国英语（取自训练集外的同数据集）

两维

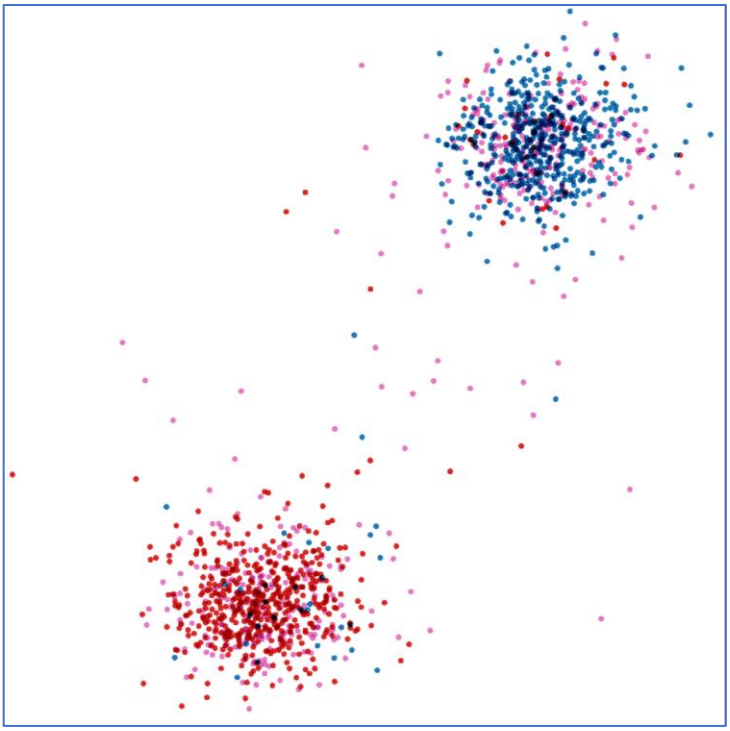


T-sne

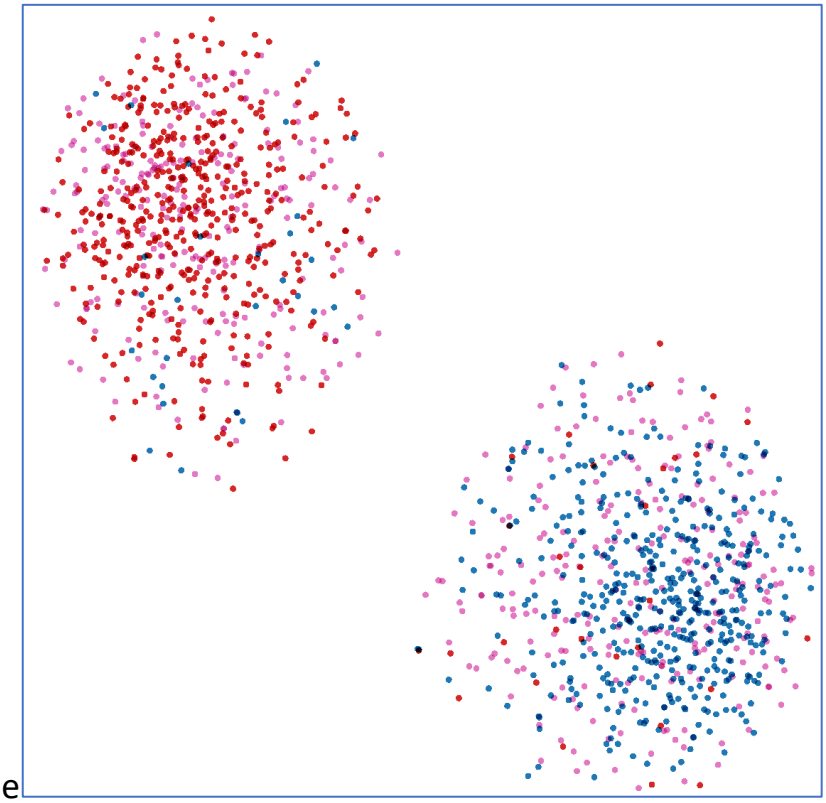


使用纯英语、中国英语、日本英语； 蓝色 wsj，红色 ce，粉色 je

两维



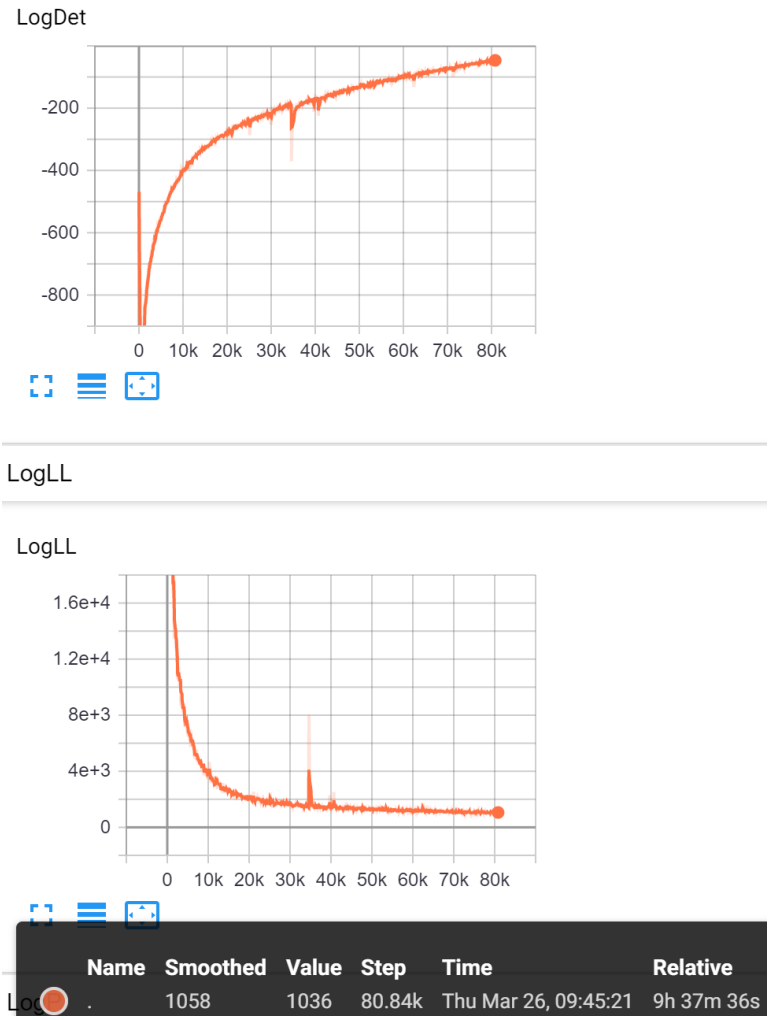
t-sne



可视化结果表明，模型对于集内（wsj 和 ce）区分性较好

实验 2　　2020.3.25

Realnvp　　block=10　　hidden=512　　input=440　　cond_dim=400　var=1 & 1k



有一个明显的、奇怪的波动。

训练集在隐空间高斯性：

c(dim=400): *skew = 0.0006283868383616209 kurt = -1.5424493272780615*

r(dim=40): *skew = 0.008583634172100573 kurt = 0.7271008120156788*

测试集在隐空间高斯性：

wsj+ce:

c(dim=400): *skew = 0.0001360115630782843 kurt = -1.6097791542651472*

r(dim=40): *skew = 0.0037121913279406725 kurt = 0.7380078159736696*

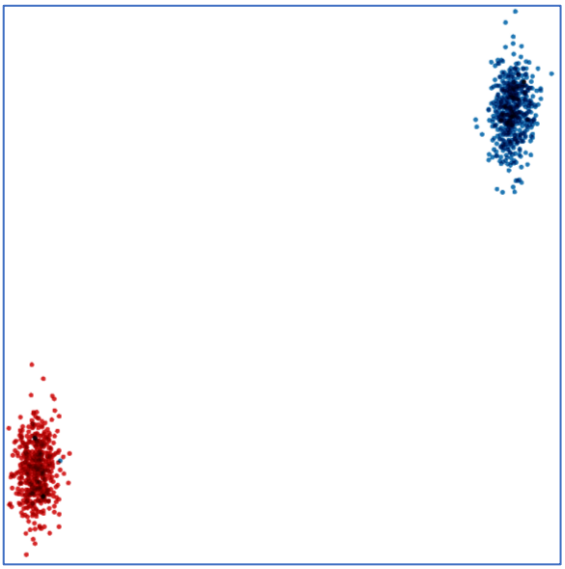
wsj+ce+je:

c(dim=400): *skew = 7.981912058312446e-05 kurt = -1.5953344626309138*

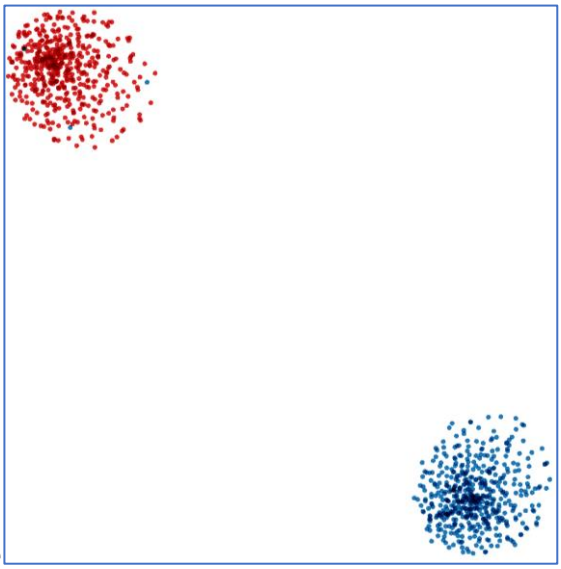
r(dim=40): *skew = 0.00020940345712006092 kurt = 0.7430761053304892*

训练集分布

两维



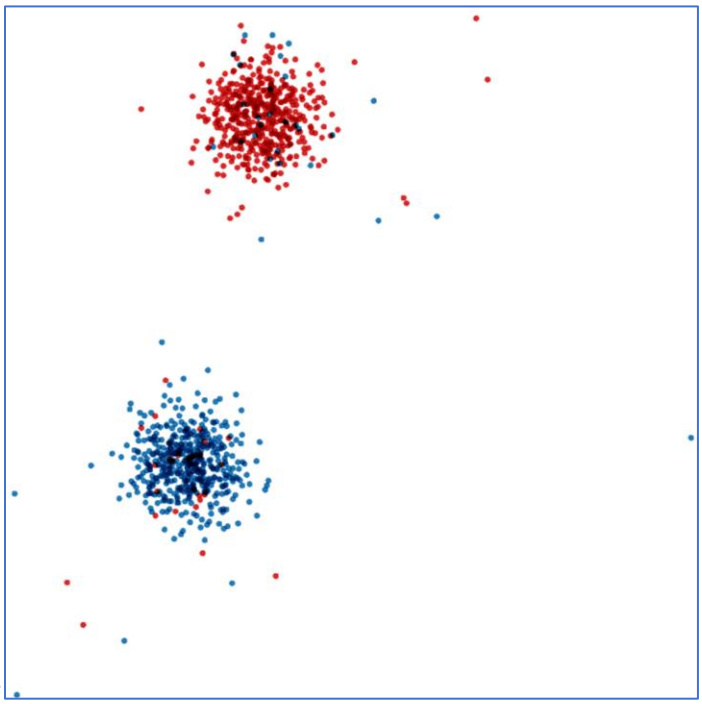
T-sne



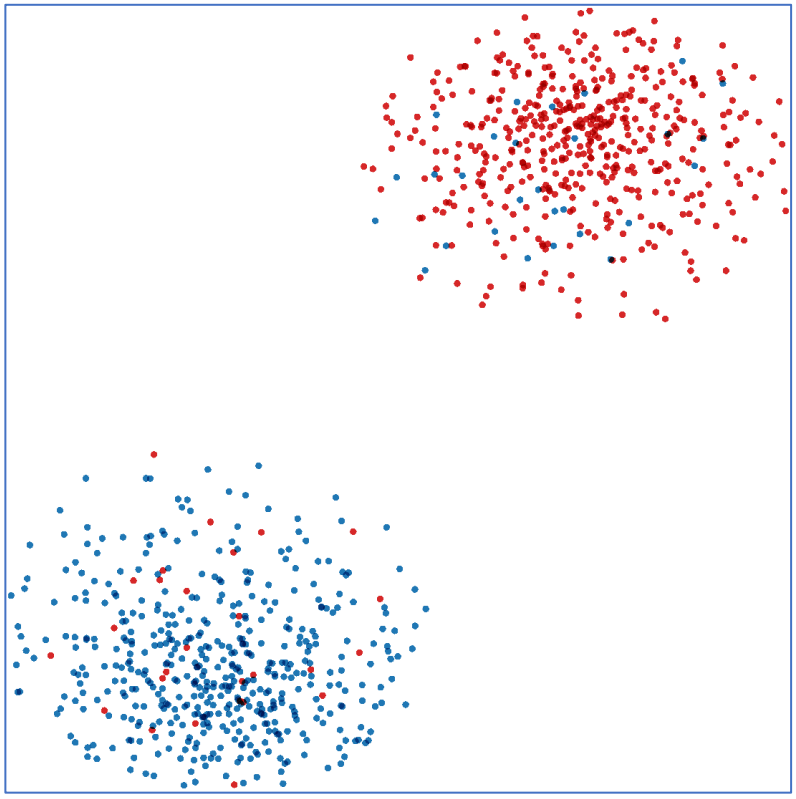
测试集分布

wsj+ce

两维

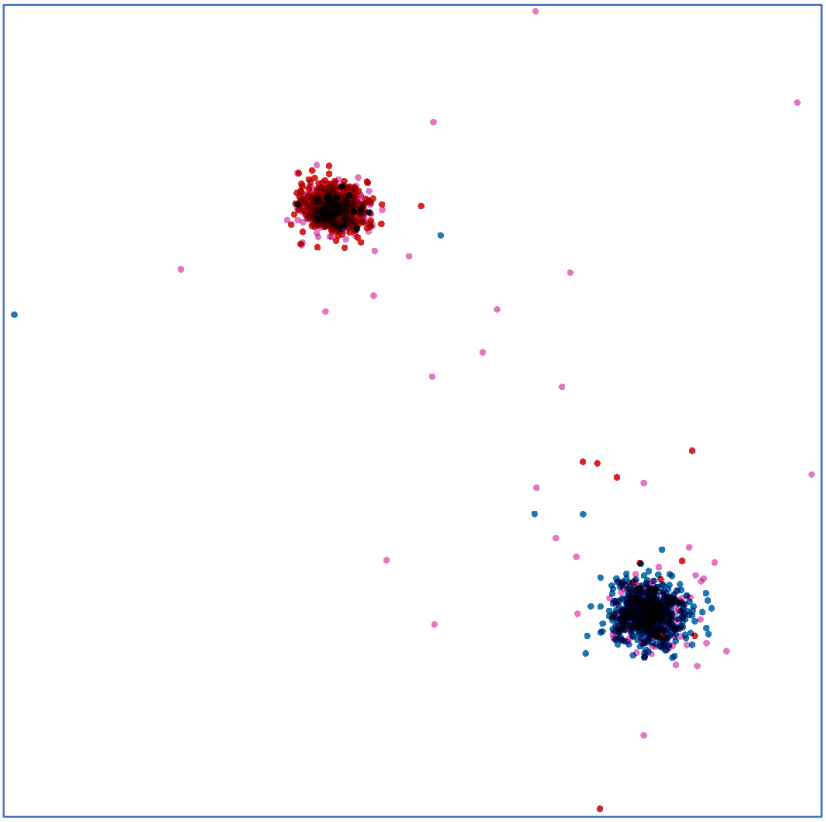


t-sne

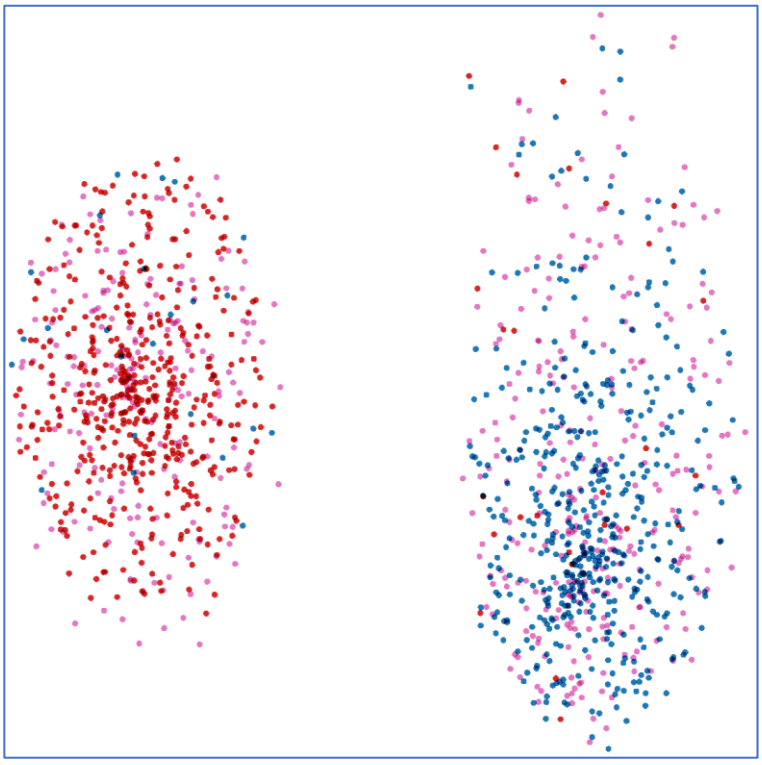


wsj+ce+je (wsj&ce 来自同集）蓝色 wsj，红色 ce，粉色 je

两维

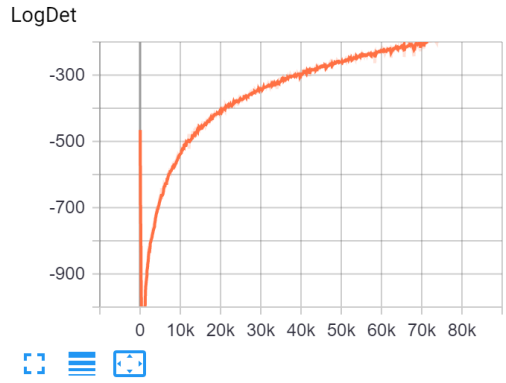


t-sne

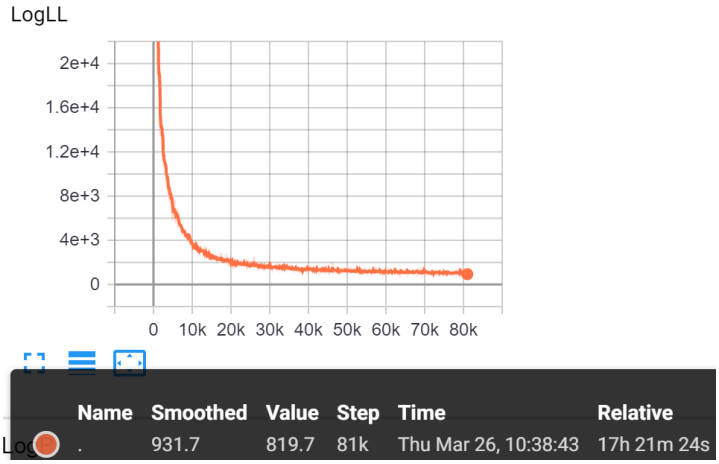


实验 3　　2020.3.25

Realnvp　　block=10　　hidden=512　　input=440　　cond_dim=440　var=1 & 1k



LogLL



同样的数据，这次没有出现奇怪的波动。

训练集在隐空间高斯性：

c(dim=440): *skew = -0.014642360260371457 kurt = -1.4761772624802512*

测试集在隐空间高斯性：

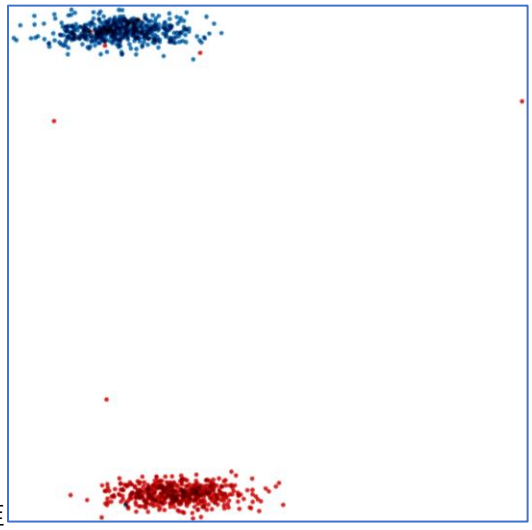
wsj+ce 两分类：

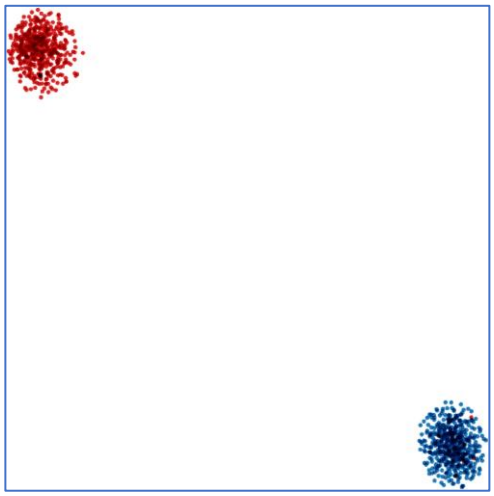
c(dim=440): *skew = -0.005829388468092392 kurt = -1.5395667708074203*

wsj+ce+je 三分类:

c(dim=440): *skew = -0.005821182986123445 kurt = -1.5287577271416113*

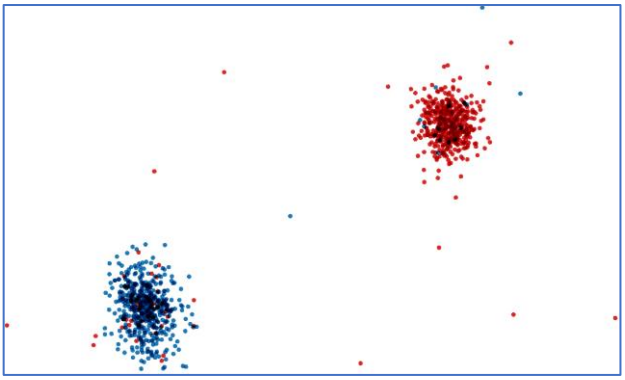
训练集分布：

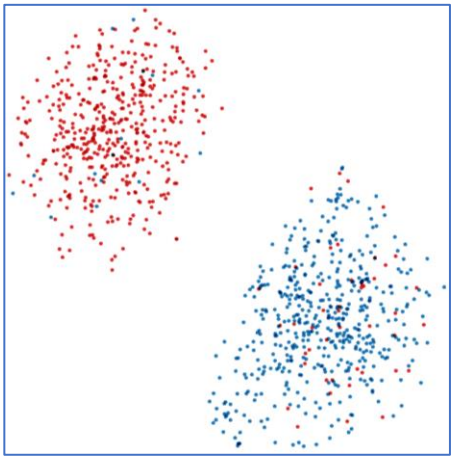
两维 

t-sne 

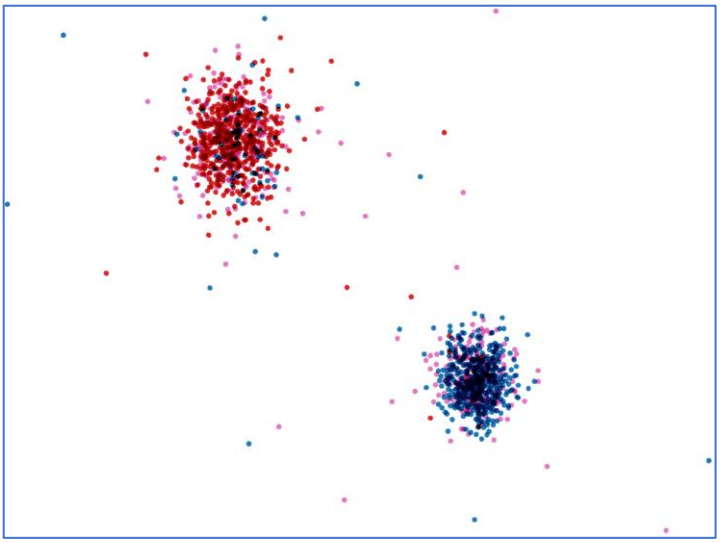
测试集分布：

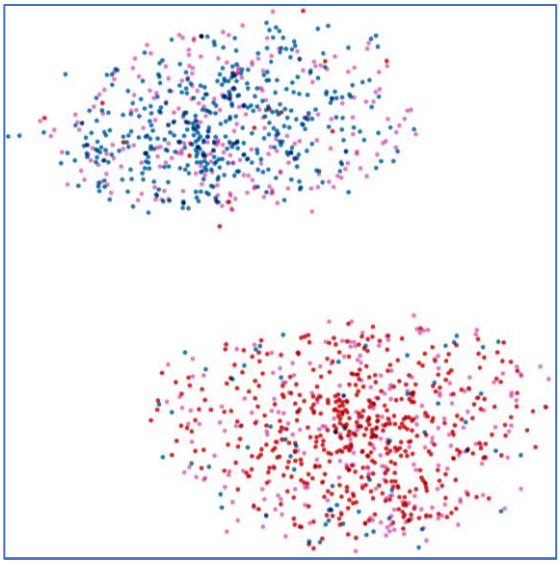
Wsj+ce：

两维 

t-sne 

wsj+ce+je；蓝色 wsj，红色 ce，粉色 je

两维 

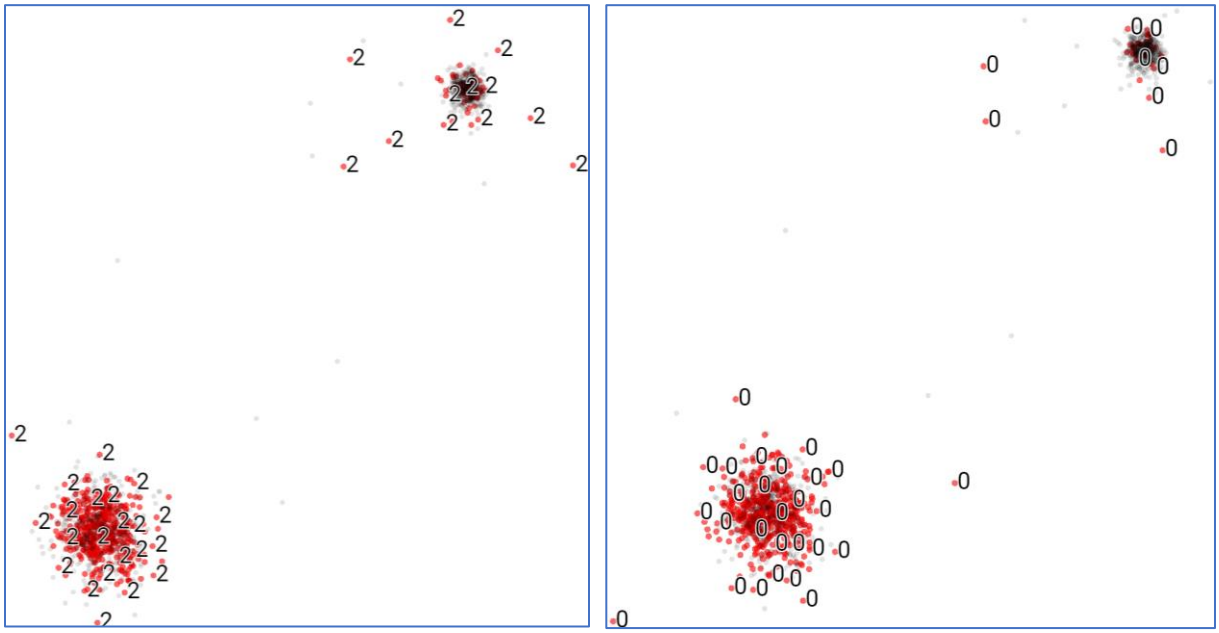
t-sne 

实验 1、2、3 说明，在可视化的层面上，对于 wsj 和 ce 分类性较好。

对于没见过的 je，依然不会分布到极其外围的位置，说明可能学到了英文发音的信息。

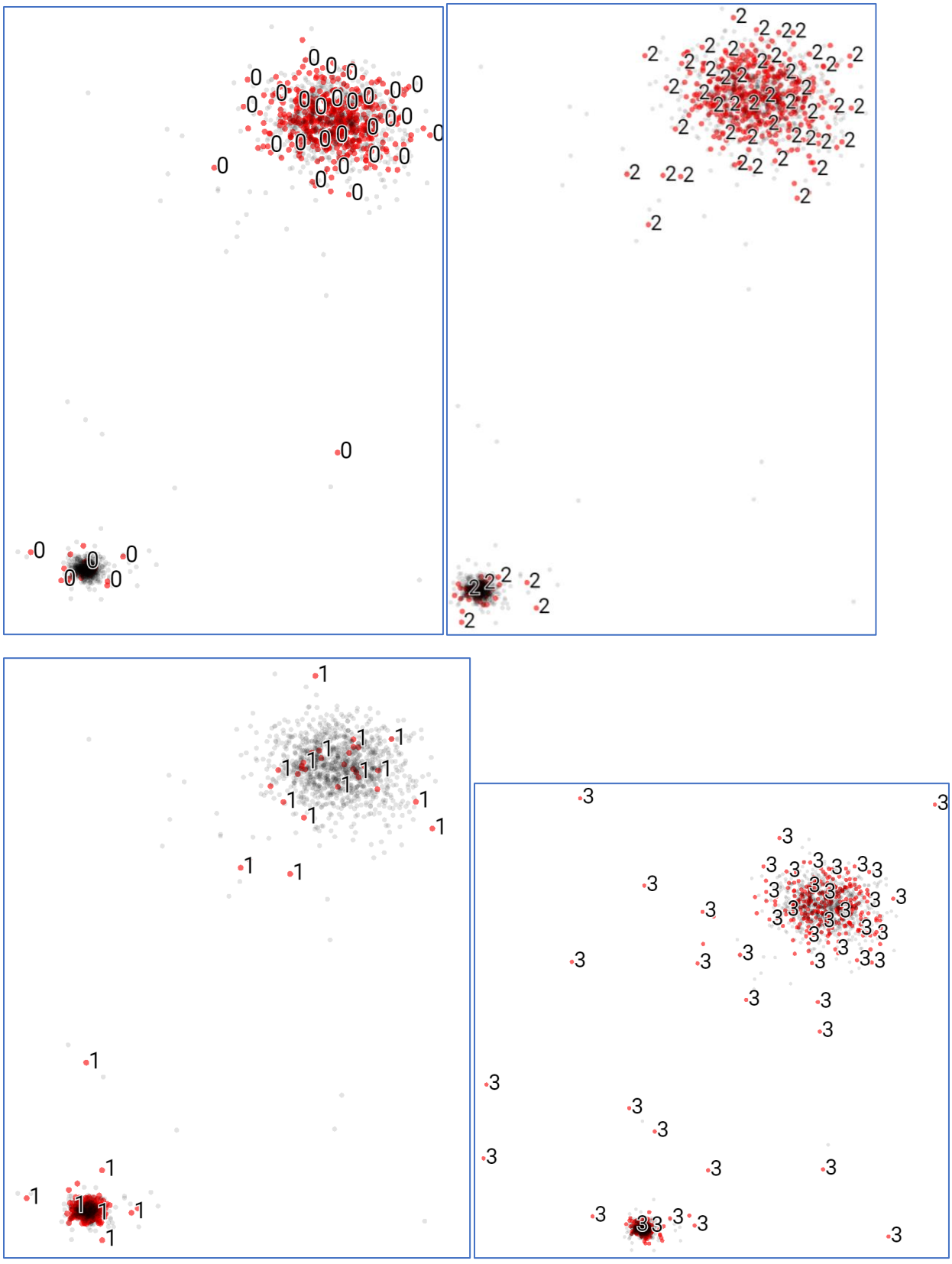
I. 使用 wsj+ce+timit 测试

直接取两维



0 是 wsj 测试集，2 是 timit 集。可以排除数据集信道影响

II. 使用 wsj+ce+timit+je 测试



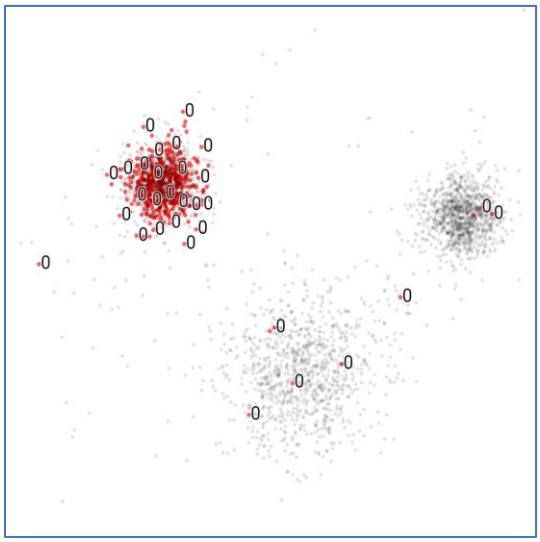
0、2 分别是 wsj 和 timit 测试集，说明在该两维上，较好英语发音分布于右上角。

1 是 Chinese English 测试集，说明在该两维上，较差/ce 分布于左下角。

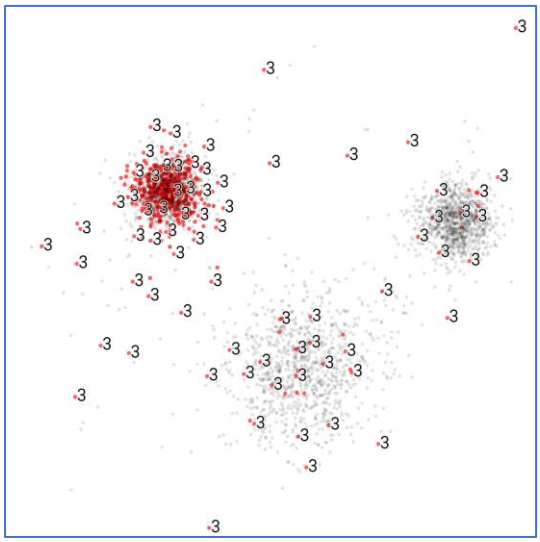
3 是 Japanese English 测试集，对于该模型没有见过的“方言”，无法用可视化找到它的分布规律（图中 je 在两类分布上的比例都是 50%左右）。

此处选择的 je 都是得分较低的，但从图上来说，它们在 English class 上依然有较高的分布。需要进一步的计算。

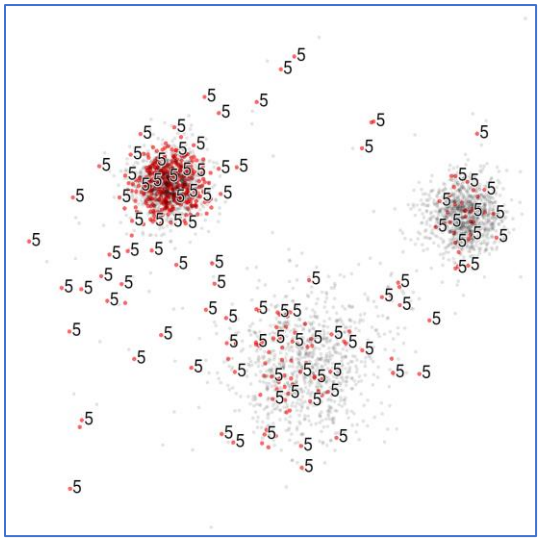
使用三分类训练 training set: wsj+ce+je



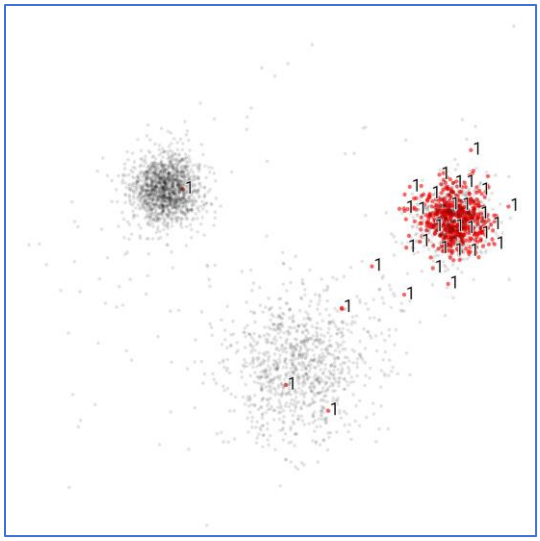
wsj



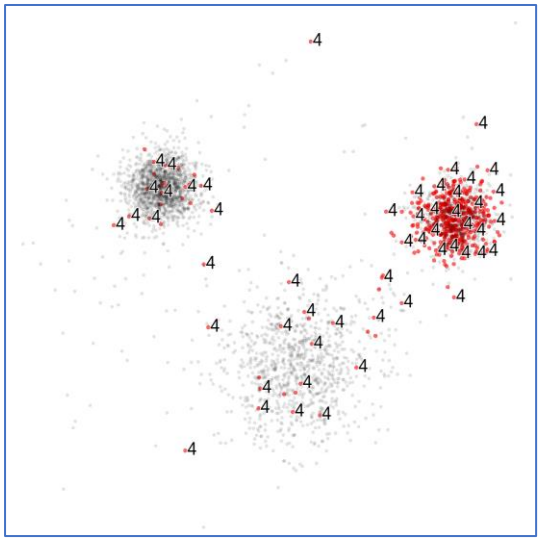
wsj-test



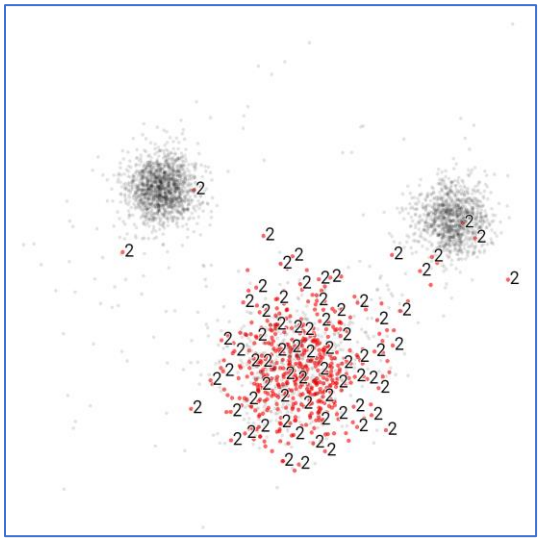
timit



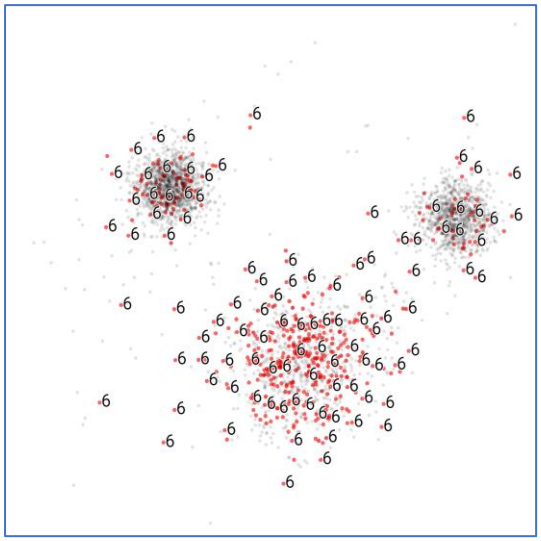
ce



ce-test



je



je-test(not only 1&2 score)

wsj+ce+je 训练	frame	utt
English(wsj+timit)	0.7999	0.9431
Chinese English	0.8837	0.9764
Japanese English	0.7111	0.9405
voxceleb	0.3827	0.1769

Vox 中只有 38%正确分到了 English，45%被错误分到了 Chinese English 中

wsj+ce 训练	Frame	utt
English(wsj+timit)	0.9199	0.9969
Chinese English	0.9409	0.9979
Japanese English	0.4068	0.2979
voxceleb	0.5067	0.5460

训练集中并不包含 je 的音频

使用 voxceleb 和 cnceleb 训练	Frame	utt
voxceleb	0.7219	0.9140
cnceleb	0.8649	0.9016
wsj(将其分到 EN 的比例)	0.7220	0.9106
cn_1s	0.5214	0.5421
cn_3s	0.5736	0.6292
je(将其分到 EN 的比例)	0.5813	0.6226
ce(将其分类到 CN 的比例)	0.5344	0.6080

对于模型没见过的 Chinese English, 模型更倾向于将其分类到 Chinese 类中

但同样陌生的 Japanese English 更倾向于被分类到 English 中

而 wsj 确实被分类到了 English 中，且有很高的比例

奇怪的是 cn_1s 和 cn_3s 的结果，对打分进行观察后发现这些音频在两类上的 logP 很相似，比如：-506.3572998046875　　-508.2145690917969。但这些很小的差值造成了大量的错误分类。

Normal flow model

用 wsj-50h 中的 9/10 训练

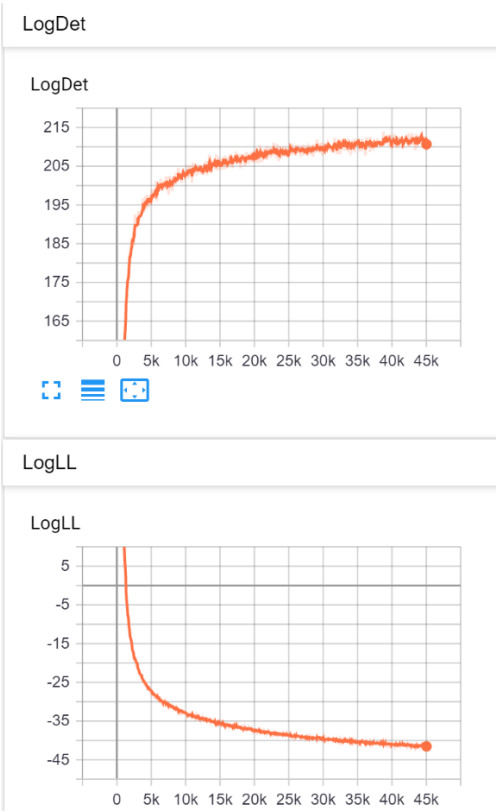
--epochs 100 --batch-size 10000 --lr 0.0001 --num-blocks 10

real_nvp / relu / num_hidden = 256

以下 test 均用 epoch=95 的 model

a. 训练结果：

LogLL (loss) **= -41.869190** LogP = -170.205643 LogDet = 212.074829



b. 训练集（约 45h）在 latent 空间：

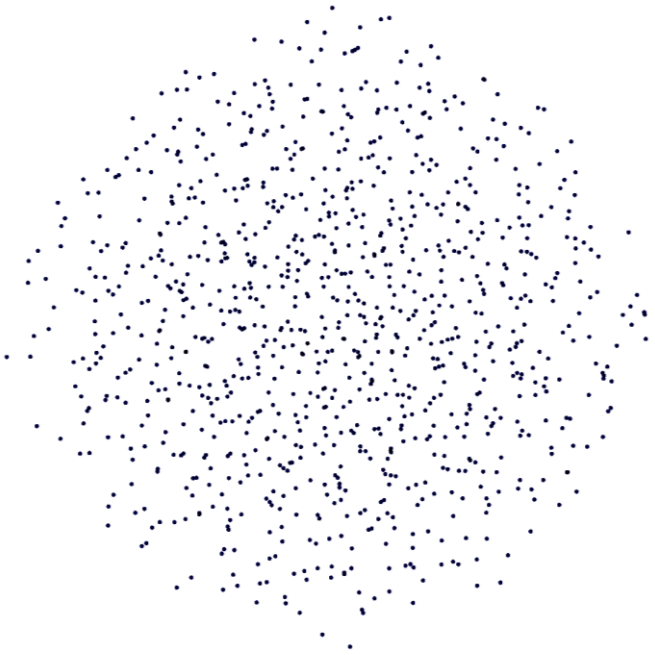
LogLL = 1479.559040 LogP = -1443.457739 LogDet = -36.101301

456 万帧，120 维

c. wsj-50h 训练集外的 1/10（约 5h）在 latent 空间：

LogLL = 1466.184247 LogP = -1429.757313 LogDet = -36.426934

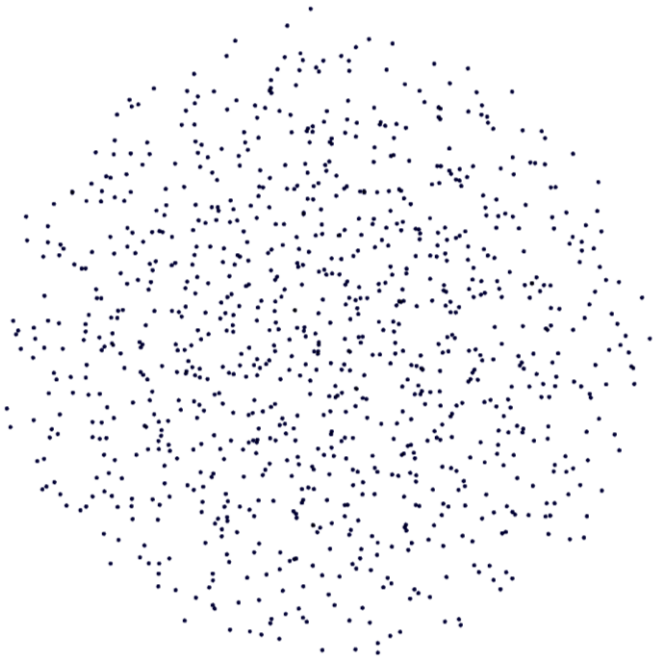
46 万帧，120 维



d. L2（Chinglish 集，约 30h）在 latent 空间:

LogLL = 1655.813159 LogP = -1601.595372 LogDet = -54.217786

422 万帧，120 维



e. JE (Japanese English, about 2h. Include score 1~5)

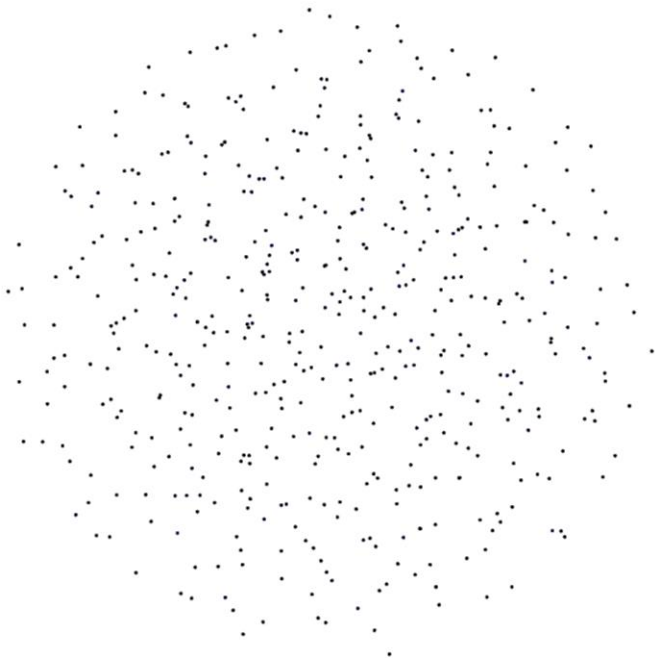
LogLL = 1573.815700 LogP = -1529.885478 LogDet = -43.930221

20 万帧，120 维

f. JE(Score 1&2)

LogLL = 1568.640994 LogP = -1525.555617 LogDet = -43.085376

75000 帧，120 维



g. wsj-test ce je(1&2)在 latent 空间中分布