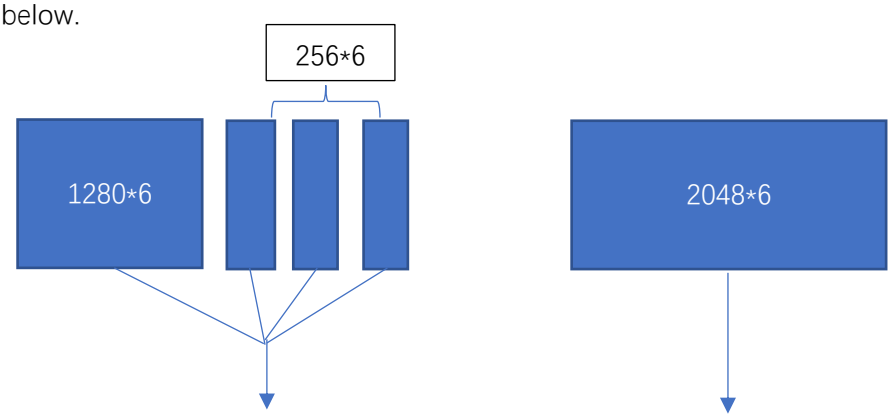
Experiment conclusions

Firstly I have to admit that I made a stupid mistake that when I want to resample the first 256*6 model's training data, I wrongly choose the 1280*6 mdl to evaluate the WER of the training data. I didn't realize it until I've done training all the model. So I have to change my basic thought and have a different Comparison. But there still exists some useful information though this experiment has deviation with my basic thought. The new structure looks like below.



Several results

	2048*6 (1280+256*3)	Ensemble 1+2+3+4	1280*6(1)	256*6 (2)	256*6 (3)	256* 6(4)
bd_tgpr_dev 93	**9.85**	10.53	10.37	17.21	18.80	20.68
bd_tgpr_dev 93_fg	**8.95**	9.34	9.21	15.55	16.88	18.36
bd_tgpr_eval 92	**6.31**	6.63	6.50	12.65	13.72	15.12
bd_tgpr_eval 92_fg	**5.28**	**5.28**	5.46	10.77	12.26	13.22
tg_dev93	**11.63**	11.94	11.85	18.80	20.11	21.62
tg_eval92	**7.67**	8.40	8.22	14.11	15.17	16.39
tgpr_dev93	**12.38**	12.63	12.67	19.19	20.62	21.86
tgpr_eval92	**8.86**	**8.86**	**8.86**	14.74	15.91	17.21

1. It seems that ensemble model cannot do better than first 1280*6 model not to say 2048*6. maybe Adaboost can do better when NN have the same representation ability in other words these models have to have same structure.

2. The more complex data model learns on , the worse result they will have. That's as expected cause maybe harder training data really confused them.

3. When model trains on resampled data , they do better on test data. For example, model-2's WER on training data is 33.82% and model-3's WER on training data is 44.7%.This is

really interesting cause model usually cannot do well on test set as they do on training set. Maybe they really learned some information.

Confusions

1. The weight for each model is 0.86778 , 0.05940 , 0.04139, 0.03143 and they sum up to be 1. I think it's more reasonable to weight the softmax layer before log. However , it do much better to directly weight the log-softmax layer which is unexpected.
2. The really log-likelihood for each model is posterior * prior and the prior is based on the training data. However training data for each model is different, I don't know whether it influence the results.