Lab 6: Group work on projects

The goal of this lab is for you to make progress on your project, together as a group. You'll set goals and work towards them, and report what you got done, challenges you faced, and subsequent plans.

Group name: How Do you turn this on.

Group members present in lab today: Yi-Ting Yeh, Ting-Rui Chiang

1: Plan

1. What is your plan for today, and this week?
   Today: Discuss with lecturers on our progress, and decide our future plan.
   This week: 1) Check the performance of the ASR system without a neural language model. 2) Try different pruning strategies. 3) Finish the first training of the layer-purned model.
2. How will each group member contribute towards this plan?
   Yi-Ting: Layer pruning.
   Ting-Rui: Head pruning.

2: Execution

1. What have you achieved today / this week? Was this more than you had planned to get done? If so, what do you think worked well?

   Today:

   We discussed our progress. Previously, we found that head pruning does not decrease the latency pretty much. The lecturers suggested we check the latency of different head pruning strategies. Specifically, we may try to preserve the locality during head purning. We conjectured that the neural language model may also induce a great portion of latency. Therefore, we also discussed the choice of language models. If we need to use an external language model, the lecturers suggest we compare their perplexity and latency.

   This week:

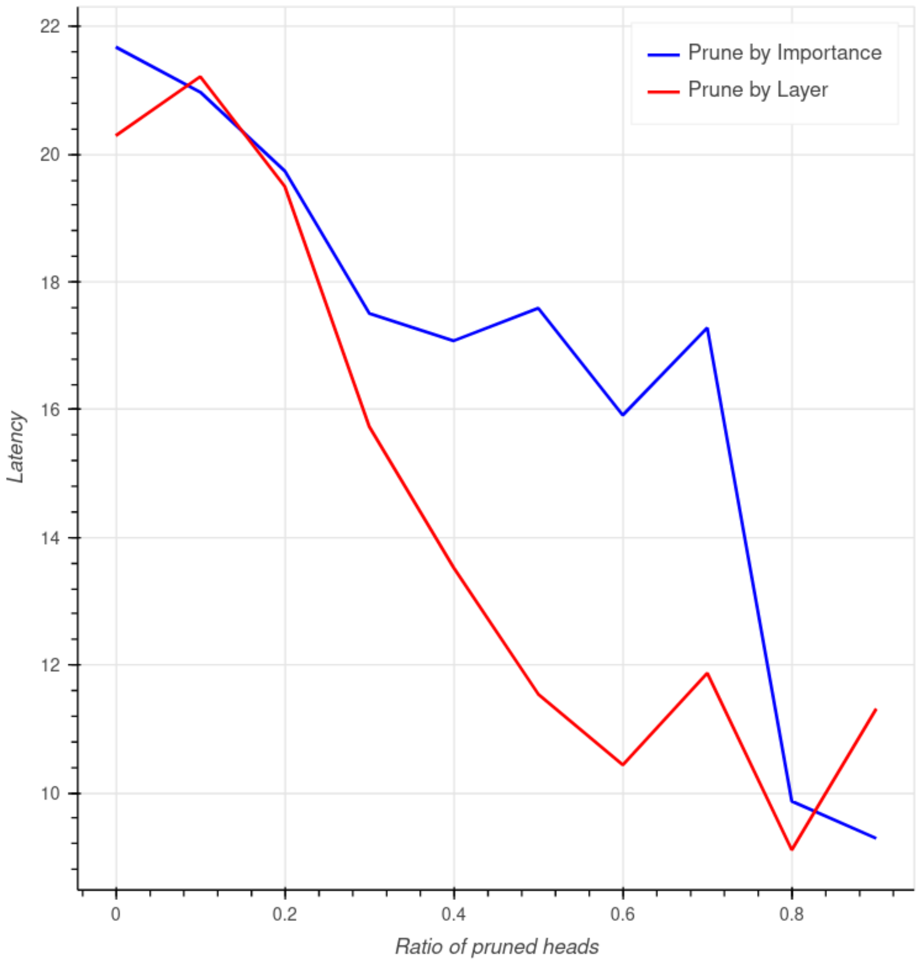
   (1) We measure the WER of the ASR system without a neural language model. We found that not using a neural language model affects the WER by less than 1%.

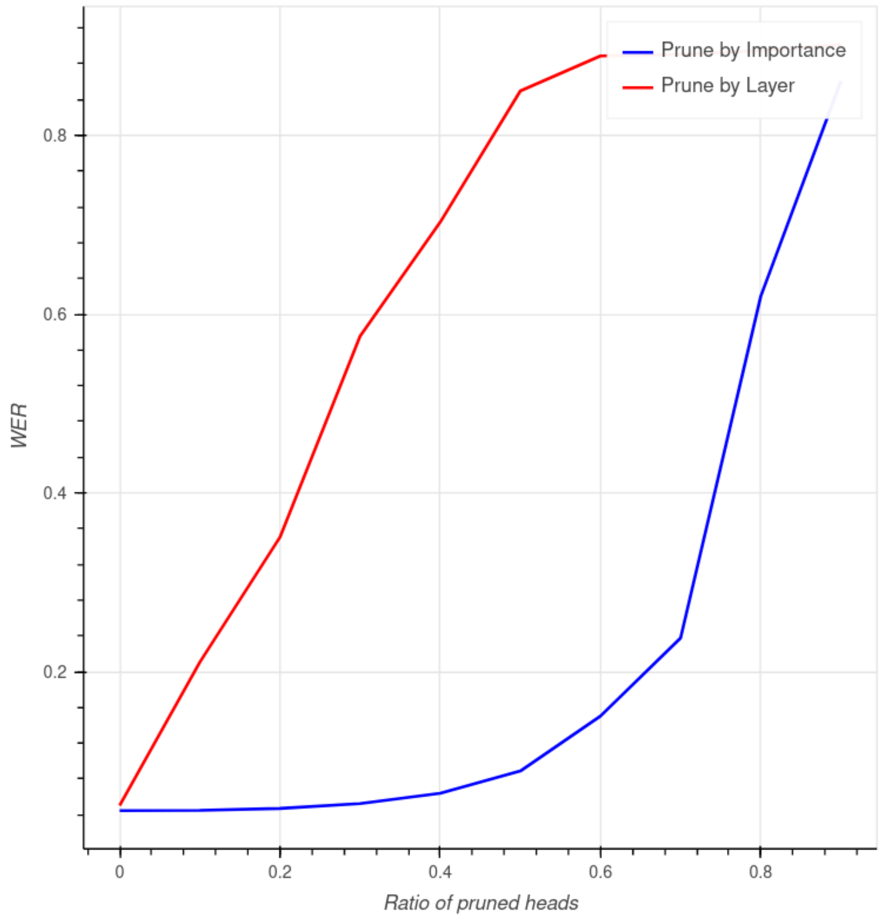
Moreover, it reduces the average latency from 44.42 sec/instance to 21.68 sec/instance. Based on this result, we think we can discard the neural language model.

(2) We tried different pruning strategies. As an initial step, we tried different normalization methods when computing the head importance. However, it does not affect the latency apparently. One possible explanation is that different normalization methods do not make much difference in the locality of pruned heads.

In order to test the hypothesis that locality can affect the latency, we conducted an experiment, where we sort the heads by the layer they belong to, and prune the heads in the lower layer first. This pruning policy keeps the maximum locality. The result is shown in the following figure. It shows that pruning by layer is more efficient, indicating that locality is indeed important. However, this method also hurts the performance pretty much. We may need to think about how to prune a model with high locality while maintaining high WER.

(3) We finish the training of layer pruning models with LayerDrop and intermediate CTC loss. We tried two settings. The first setting is to follow the original paper (https://arxiv.org/pdf/2106.09216.pdf) and train the model from scratch on a smaller subset of LibriSpeech. However, it takes too long to train an Transformer ASR model from scratch and the final WER is poor, due to our change of config. Therefore, we switch to finetune the pretrained model with LayerDrop and intermediate CTC loss. We are now experimenting with the best config for fine-tuning models without hurting performance.


Prune by Importance
Prune by Layer
22
20
18
16
14
12
10
Latency
0
0.2
0.4
0.6
0.8
Ratio of pruned heads



2. Was there anything you had hoped to achieve, but did not? What happened? How did you work to resolve these challenges?

We haven't finished fine-tuning ASR models with LayerDrop and intermediate CTC loss. It is because using intermediate CTC loss to finetune a trained Transformer ASR model drastically increases the loss. We suspect it is because 1) the layer representations of the model will be very different if we use the intermediate CTC loss. 2) the fine-tuning process is sensitive to the choice of hyperparameters, such as which layer should be used as an intermediate layer. We are now experimenting with the choice of hyperparameters and will try to only use LayerDrop or intermediate CTC during the fine-tuning process.

3. What were the contributions of each group member towards all of the above?
   Yi-Ting: Layer pruning.
   Ting-Rui: Pruning strategies

3: Next steps

1. Are you making sufficient progress towards completing your final project? Explain why or why not. If not, please report how you plan to change the scope and/or focus of your project accordingly.

   Yes, we are on the right track. For head-pruning, we found the bottleneck of latency of the model which is LM. Now we can experiment on the tradeoff of the inference speed and WER with different pruning strategies. For layer-pruning, we discover problems of fine-tuning models and come up with possible solutions. Our workaround of fine-tuning pretrained models with layer-pruning techniques should be useful for other researchers.

2. Based on your work today / this week, and your answer to (1), what are your group's planned next steps?

   We plan to experiment with more head pruning strategies and apply head pruning on the layer-pruned model. We want to experiment with more hyperparameter settings of fine-tuning layer-pruned models and find a successful strategy.

3. How will each group member contribute towards those steps?

   Yi-Ting: Layer pruning
   Ting-Rui: Head pruning strategies.