1 Complementary Experimental Results

1.1 Language Modeling Results of Different Sparsity

Table 1. Perplexity obtained by different pruning methods using different sparsity on LLaMA and LLaMA-2 models

			LLaMA		LLaMA-2	
Method	Weight Update	Sparsity	7B	13B	7B	13B
Dense	-	0%	5.68	5.09	5.12	4.57
Magnitude	×	40%	8.60	8.42	7.31	5.26
SparseGPT	✓	40%	6.33	5.55	5.71	5.02
Wanda	×	40%	6.39	5.51	5.67	5.01
Pruner-Zero	×	40%	6.18	**5.42**	**5.55**	**4.89**
Our method	×	40%	**6.17**	5.43	**5.55**	4.91
Magnitude	×	60%	559.65	229.44	3e4	11.22
SparseGPT	✓	60%	10.32	8.47	9.59	7.78
Wanda	×	60%	10.72	8.76	10.05	7.92
Pruner-Zero	×	60%	10.03	7.58	9.55	6.89
Our method	×	60%	**9.29**	**7.21**	**8.60**	**6.70**

When pruning the model as a whole using a sparsity of 40%, our method is not effective in maintaining the performance of the pruned model, and even reduces the performance of the pruned model. However, when the model is pruned using a sparsity of 60%, our method is able to effectively maintain the performance of the model. In particular, after pruning by our method, the perplexity of the LLaMA-7B model is 9.29, which is a decrease of 0.74 compared to the Pruner-Zero method. For the LLaMA-13B model, pruning by our method can reduce the perplexity of the model on the test set by 0.37. Compared to the previous method, the performance of the LLaMA-2-7B model is most significantly improved after pruning by our method. The perplexity decreases from 9.55 to 8.60, and the performance of the pruned model is improved by 9.95% compared with that of the model obtained by the existing method. Although in the LLaMA-2-13B model, the present method also appears to have a performance improvement that is not as good as the rest of the models, it still has a perplexity decrease of 0.19

At the same time, our method is able to reduce the impact of pruning on the model by adjusting the sparsity used in each layer of the model under different sparsity, thus improving the performance of the pruned model. Although the method is not able to maintain the performance of the pruned model when the sparsity of the pruning is low, under the condition of increasing the sparsity, our method is able to show more robustness than the existing methods, which can effectively maintain the performance of the model and reduce the effect of pruning on the model. We believe that this is because at lower sparsity, the

layers of the model that are affected by pruning are mostly redundant parameters that contribute little to the model performance, and even if the sparsity of these layers is reduced during pruning, most of the retained parameters are non-critical parameters that do not contribute much to the final performance of the model. However, as the sparsity used in pruning increases, a large number of critical parameters in these layers are discarded, causing great damage to the model performance. Our method reduces the sparsity used in these layers and allows some of the critical parameters to be retained, thus reducing the performance loss caused by pruning and maintaining the original performance of the model in a more effective way than other methods.

1.2 Zero-shot Task Results of Different Sparsity

Table 2. Mean zero-shot accuracies obtained by different pruning methods using different sparsity on LLaMA and LLaMA-2 models

			LLaMA		LLaMA-2	
Method	Weight Update	Sparsity	7B	13B	7B	13B
Dense	-	0%	59.99	62.59	59.71	63.03
Magnitude	×	40%	58.57	58.11	60.97	64.33
SparseGPT	✓	40%	**62.65**	64.96	**63.48**	66.05
Wanda	×	40%	62.48	**65.46**	63.23	**67.36**
Pruner-Zero	×	40%	61.84	64.90	62.28	66.65
Our method	×	40%	62.60	65.24	62.31	65.80
Magnitude	×	50%	46.94	47.61	51.14	52.85
SparseGPT	✓	50%	54.94	58.61	56.24	60.72
Wanda	×	50%	54.21	59.33	56.24	60.83
Pruner-Zero	×	50%	**59.56**	62.67	58.87	64.83
Our method	×	50%	59.49	**62.99**	**59.99**	**64.92**
Magnitude	×	60%	39.34	44.15	43.14	48.59
SparseGPT	✓	60%	**55.59**	58.44	**55.64**	**60.97**
Wanda	×	60%	54.81	58.39	54.41	59.60
Pruner-Zero	×	60%	53.00	57.91	52.71	59.44
Our method	×	60%	54.27	**58.55**	53.50	60.25

When pruning the models using a sparsity of 40%, the average accuracy obtained by our method on all four models is not as good as the accuracy of the models pruned by Wanda method. When the models are pruned using an overall sparsity of 60%, our method is only able to obtain a model with an accuracy that exceeds the accuracy of existing pruning methods on the LLaMA-13B model. However, for the remaining three models, our method yields models with less accuracy than the Wanda and SparseGPT methods.

Such experimental results seem to indicate that for the zero-shot task, our method does not play a great role in reducing the effect of pruning on the model

and maintaining the model performance at 40% and 60% sparsity. However, considering that the layer-wise pruning method based on PI proposed in this paper needs to be combined with the unstructured pruning method, the pruning metric proposed by the Pruner-Zero method is used in this paper to compute the importance scores of the parameters in the model so that the parameters can be filtered and discarded. The performance of the Pruner-Zero method is worse than that of the SparseGPT method as well as the Wanda method. Therefore, such experimental results do not indicate that the present method is ineffective at 40% and 60% sparsity. On the contrary, both in 40% and 60% sparsity, the accuracy of the most models after pruning of our method exceeds that of the Pruner-Zero method, and also exceeds that of the SparseGPT and Wanda methods in some models, which fully indicates that our method is able to effectively reduce the degradation of the performance of the model caused by pruning and maintain the performance of the model after pruning. performance after pruning.

In this paper, it is found that when using 40% sparsity for pruning, although the average accuracy of the model after layer-wise pruning is higher than that of the model after pruning by the Pruner-Zero method, such an improvement is not particularly obvious, and the LLaMA-7B model is able to achieve the greatest accuracy improvement, with a specific improvement value of 0.76%. However, when pruning is performed using the sparsity of 60%, the average accuracy of the models pruned by the method proposed in this paper has a more significant improvement than that of the sub-models pruned by the Pruner-Zero method, in which the LLaMA-7B model is able to obtain an average accuracy improvement of 1.27%, and even the LLaMA-13B model, which is the one with the smallest improvement, is also able to improve its accuracy by 0.64%. Such a phenomenon illustrates that as the sparsity used in the pruning process increases, the effectiveness of the present method in maintaining the model performance also increases, showing good robustness under large sparsity rate conditions.

1.3 Analysis of Two Layers

In the large language model, the first and the last layers serve as the data input and output layers. Thus, we consider that these two layers play a more important role than the other layers. It seems that they should not be involved in the adjustment of the pruning rate. However, the experimental results in the previous subsection show that these two layers do not seem to be affected by pruning that much. Therefore, in this section, we design experiments to compare the impact of these two layers undergoing pruning rate adjustment or not. We first let these two layers not participate in the pruning rate adjustment, keeping the original 50% sparsity rate, and calculate the perplexity of the model after pruning. Subsequently, we let these two layers participate in the pruning rate adjustment along with other layers, and then calculate the perplexity of the model. The experimental results are shown in table3, and it can be seen that not adjusting the pruning rate for these two layers has a modest performance improvement on the model with a larger number of parameters. However, the perplexity on the LLaMA-7B and LLaMA-2-7B models increases by 0.02. Therefore, we believe

Table 3. The effect of whether or not to adjust the sparsity of the first and last layers on model performance

		LLaMA		LLaMA-2	
Whether adjust	Sparsity	7B	13B	7B	13B
Adjust	50%	6.84	5.9	6.19	5.35
Not Adjust	50%	6.86	5.89	6.2	5.34

that these two layers should be subjected to the pruning rate adjustment along with the other layers as well.

1.4 Analysis of Hyperparameter

In our experiment, the sparsity parameter to be adjusted that we used is 4%. This is because we believe that if the adjusted sparsity rate is too large, then even those layers that are less affected by pruning will be affected by the final performance of the pruning model due to the fact that too many parameters are removed during the pruning process resulting in the absence of some key parameters. If the adjusted sparsity rate is too small, then those layers that are more affected by pruning will still be more seriously affected by pruning because too few parameters are retained, thus making the final performance of the model decrease. Therefore, we adjust the sparsity rate of the required modifications in the range of 1% to 10% and conduct the experiments accordingly. The results

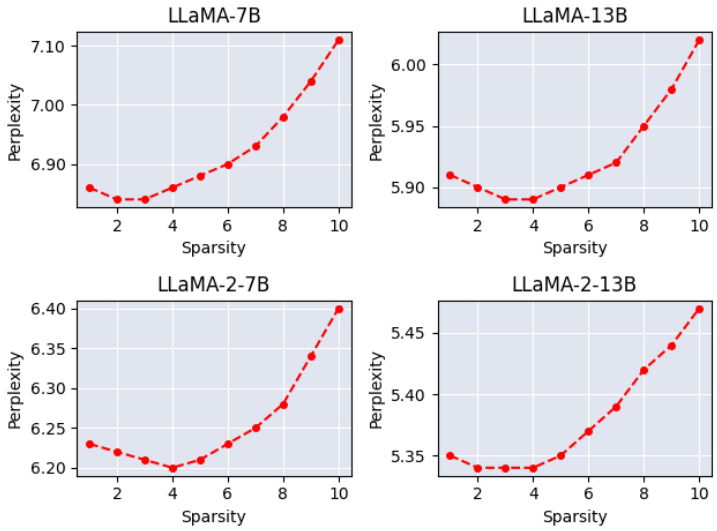

Fig. 1. With different adjusted sparsity, perplexity on four LLaMA models

of the experiments are shown in figure1, in which the performance of most of the models rises with the increase of the adjusted sparsity rate and reaches the highest performance point when the adjusted sparsity rate is 4%, and then declines with the continued increase of the adjusted sparsity rate. However, in the LLaMA-7B model, although the performance of the model reaches the highest point of performance at the adjusted sparsity of 2% and 3%, the performance of the model is still close to the optimum at the adjusted sparsity of 4%. Therefore, we finally determined 4% as our adjusted sparsity rate.

1.5 Robustness

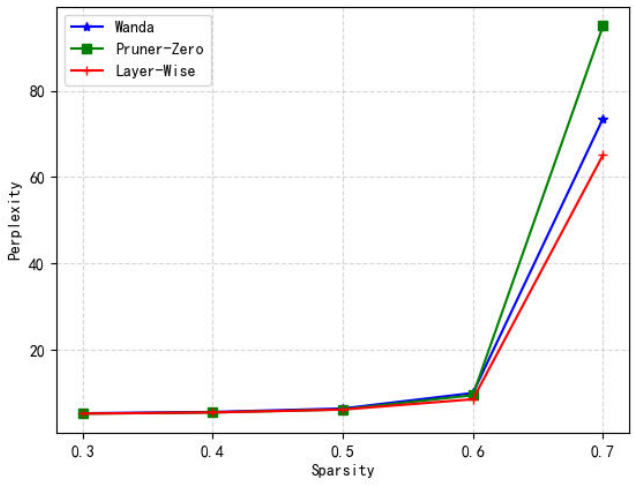

Fig. 2. Perplexity curve of LLaMA-2-7B model performance with sparsity

The perplexity of the sub-models obtained by pruning the LLaMA-2-7B model using three different pruning methods, namely Wanda, Pruner-Zero, and the method proposed in this paper, on the WikiText2 test set at sparsity rates of 30%-70% is illustrated in Figure 2. During the experiment, as the sparsity of the pruning used for the model increases, a large number of parameters are discarded, causing a decrease in the model performance, but when the sparsity is increased to 60%, the model obtained by the layer-wise pruning method based on the PI possesses a lower perplexity than the model pruned by the other methods. This suggests that with a sparsity of 60%, the method has already demonstrated better performance than the other methods. And when the sparsity is set to 70%, our method shows strong robustness and is able to maintain the performance of the pruned model more effectively compared to the other two methods. Such

experimental results illustrate the better robustness of the layer-wise pruning method based on the PI, which can maintain the performance of the model better than the other methods when the sparsity of pruning needs to be increased for the extreme case of obtaining a quantum model with fewer parameters.