Abstract

Ensembling is a popular method used to improve performance as a last resort. However, ensembling multiple models finetuned from a single pretrained model has been not very effective; this could be due to the lack of diversity among ensemble members. This paper proposes Multi-Ticket Ensemble, which finetunes different subnetworks of a single pretrained model and ensembles them. We empirically demonstrated that winning-ticket subnetworks produced more diverse predictions than dense networks, and their ensemble outperformed the standard ensemble on some tasks.

1 Introduction

Ensembling (Levin et al., 1989; Domingos, 1997) has long been an easy and effective approach to improve model performance by averaging the outputs of multiple comparable but independent models. Allen-Zhu and Li (2020) explain that different models obtain different views for judgments, and the ensemble uses complementary views to make more robust decisions. A good ensemble requires diverse member models. However, how to encourage diversity without sacrificing the accuracy of each model is non-trivial (Liu and Yao, 1999; Kirillov et al., 2016; Rame and Cord, 2021).

The pretrain-then-finetune paradigm has become another best practice for achieving state-of-the-art performance on NLP tasks (Devlin et al., 2019). The cost of large-scale pretraining, however, is enormously high (Sharir et al., 2020); This often makes it difficult to independently pretrain multiple models. Therefore, most researchers and practitioners only use a single pretrained model, which is distributed by resource-rich organizations.

This situation brings up a novel question to ensemble learning: Can we make an effective ensemble from only a single pre-trained model? Although ensembles can be combined with the pretrain-then-finetune paradigm, an ensemble of models finetuned from a single pretrained model is much less effective than that using different pretrained models from scratch in many tasks (Raffel et al., 2020). Naïve ensemble offers limited improvements, possibly due to the lack of diversity of finetuning from the same initial parameters.

In this paper, we propose a simple yet effective method called Multi-Ticket Ensemble, ensembling finetuned winning-ticket subnetworks (Frankle and Carbin, 2019) in a single pretrained model. We empirically demonstrate that pruning a single pretrained model can make diverse models, and their ensemble can outperform the naïve dense ensemble if winning-ticket subnetworks are found.

2 Diversity in a Single Pretrained Model

In this paper, we discuss the most standard way of ensemble, which averages the outputs of multiple neural networks; each has the same architecture but different parameters. That is, let \( f(x; \theta) \) be the
output of a model with the parameter vector $\theta$ given the input $x$, the output of an ensemble is $f_M(x) = \sum_{\theta \in M} f(x; \theta) / |M|$, where $M = \{\theta_1, \ldots, \theta_{|M|}\}$ is the member parameters.

2.1 Diversity from Finetuning

As discussed, when constructing an ensemble $f_M$ by finetuning from a single pretrained model multiple times with different random seeds $\{s_1, \ldots, s_{|M|}\}$, the boost in performance tends to be only marginal. In the case of BERT (Devlin et al., 2019) and its variants, three sources of diversities can be considered: random initialization of the task-specific layer, dataset shuffling for stochastic gradient descent (SGD), and dropout. However, empirically, such finetuned parameters tend not to be largely different from the initial parameters, and they do not lead to diverse models (Radiya-Dixit and Wang, 2020). Of course, if one adds significant noise to the parameters, it leads to diversity; however, it would also hurt accuracy.

2.2 Diversity from Pruning

To make models ensuring both accuracy and diversity, we focus on subnetworks in the pretrained model. Different subnetworks employ different subspaces of the pre-trained knowledge (Radiya-Dixit and Wang, 2020; Zhao et al., 2020; Cao et al., 2021); this would help the subnetworks to acquire different views, which can be a source of desired diversity\(^1\). Also, in terms of accuracy, recent studies on the lottery ticket hypothesis (Frankle and Carbin, 2019) suggest that a dense network at initialization contains a subnetwork, called the winning ticket, whose accuracy becomes comparable to that of the dense one after the same training. Interestingly, the pretrained BERT also has a winning ticket for finetuning on downstream tasks (Chen et al., 2020). Thus, if we can find diverse winning tickets, they can be good ensemble members with the two desirable properties: diversity and accuracy.

3 Subnetwork Exploration

We propose a simple yet effective method, multi-ticket ensemble, which finetunes different subnetworks instead of dense networks. Because it could be a key how to find subnetworks, we explore three variants based on iterative magnitude pruning.

---

\(^1\)Some concurrent and recent studies also investigate subnetworks for effective ensemble (Durasov et al., 2021; Havasi et al., 2021) for training-from-scratch settings of image recognition.

![Figure 2: Overview of iterative magnitude pruning (Section 3.1). We can also use regularizers during finetuning to diversify pruning (Section 3.2).](image_url)
**Active Masking** To maximize the diversity of the surviving parameters of member models, it is necessary to prune the surviving parameters of the random seed $s_1$ when building a model with the next random seed $s_2$. Thus, during finetuning with seed $s_2$, we apply the $L_1$ regularizer on the first surviving parameters. Likewise, with the following seeds $s_3, s_4, \ldots, s_i, \ldots, s_{|M|}$, we cumulatively use the average of the surviving masks as the regularizer coefficient mask. Let $m_{s_i} \in \{0, 1\}^{|\theta|}$ be the pruning mask indicating surviving parameters from seed $s_i$, the coefficient mask with seed $s_i$ is $t_{s_i} = \sum_{j<i} m_{s_j} / (i-1)$. We call this affirmative policy as active masking.

**Random Masking** In active masking, each coefficient mask has a sequential dependence on the preceding random seeds. Thus, the training of ensemble members cannot be parallelized. Therefore, we also experiment with a simpler and parallelizable variant, random masking, where a mask is independently and randomly generated from a random seed. With a random seed $s_i$, we generate the seed-dependent random binary mask, i.e., $t_{s_i} = m_{s_i}^{\text{rand}} \in \{0, 1\}^{|\theta|}$, where each element is sampled from Bernoulli distribution and 0’s probability equals to the target pruning ratio.

## 4 Experiments

We evaluate the performance of ensembles using four finetuning schemes: (1) finetuning without pruning (BASELINE), (2) finetuning of lottery-ticket subnetworks found with the naïve iterative magnitude pruning (BASE-LT), and (3) with $L_1$ regularizer by the active masking (ACTIVE-LT) or (4) random masking (RANDOM-LT). We also compare with (5) BAGGING-based ensemble, which trains dense models on different random 90% training subsets. We use the GLUE benchmark (Wang et al., 2018) as tasks. The implementation and settings follow Chen et al. (2020)² using the Transformers library (Wolf et al., 2020) and its bert-base-uncased pretrained model. We report the average performance using twenty different random seeds. Ensembles are evaluated using exhaustive combinations of five members. We also perform Student’s t-test for validating statistical significance³. Note

### Table 1: The performances (single, ens.) and the improvements by ensembling (diff.)

<table>
<thead>
<tr>
<th></th>
<th>MRPC</th>
<th>STS-B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>single</td>
<td>ens.</td>
</tr>
<tr>
<td>BASELINE</td>
<td>83.48</td>
<td>84.34</td>
</tr>
<tr>
<td>BAGGING</td>
<td>82.87</td>
<td>84.19</td>
</tr>
<tr>
<td>BASE-LT</td>
<td>83.84</td>
<td>84.98</td>
</tr>
<tr>
<td>ACTIVE-LT</td>
<td>83.22</td>
<td>84.60</td>
</tr>
<tr>
<td>RANDOM-LT</td>
<td>83.53</td>
<td>85.05</td>
</tr>
</tbody>
</table>

Note that not all evaluation samples satisfy independence assumption.

---

²We found a bug in Chen et al. (2020)’s implementation on GitHub, so we fixed it and experimented with the correct version.

³Note that not all evaluation samples satisfy independence assumption.

4Raffel et al. (2020) reported that the same problem happened on almost all tasks (GLUE (Wang et al., 2018), SuperGLUE (Wang et al., 2019), SQuAD (Rajpurkar et al., 2016), summarization, and machine translation) using the T5 model.

<table>
<thead>
<tr>
<th></th>
<th>BASE-LT</th>
<th>ACTIVE-LT</th>
<th>RANDOM-LT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.85</td>
<td>0.90</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Figure 3: Comparison of the performances and the number of ensemble members on MRPC (left) and STS-B (right). They are represented as the relative gain compared with BASELINE’s accuracy.

### 4.1 Accuracy

We show the results on MRPC (Dolan and Brockett, 2005) and STS-B (Cer et al., 2017) in Table 1. Multi-ticket ensembles (*-LT) outperform BASELINE and BAGGING significantly ($p < 0.001$). This result supports the effectiveness of multi-ticket ensemble. Note that the improvements of *-LT are attributable to ensembling (diff.) rather than to any performance gains of the individual models (single). We also plot the improvements (ens. values relative to BASELINE) as a function of the number of ensemble members on MRPC and STS-B in Figure 3. This also clearly shows that while the single models of *-LT have accuracy similar to BASELINE, the gains appear when ensembling them. While multi-ticket ensemble works well even with the naive pruning method (BASE-LT), RANDOM-LT and ACTIVE-LT achieve the better ensembling effect on average; this suggests the effectiveness of regularizers. Interestingly, RANDOM-LT is simpler but more effective than ACTIVE-LT.
When Winning Tickets are Less Accurate

Does multi-ticket ensemble work well on any tasks? The answer is no. To enjoy the benefit from multi-ticket ensemble, we have to find diverse winning-ticket subnetworks sufficiently comparable to their dense network. When winning tickets are less accurate than the baseline, their ensembles often fail to outperform the baseline’s ensemble. It happened to CoLA (Warstadt et al., 2019), QNLI (Rajpurkar et al., 2016), SST-2 (Socher et al., 2013), MNLI (Williams et al., 2018); the naive iterative magnitude pruning did not find comparable winning-ticket subnetworks (with or sometimes even without regularizers)\(^5\). Note that, even in such a case, RANDOM-LT often yielded a higher effect of ensembling (diff.), while the degradation of single models canceled out the effect in total, and BAGGING also failed to improve. More sophisticated pruning methods (Blalock et al., 2020; Sanh et al., 2020) or tuning will find better winning-ticket subnetworks and maximize the opportunities for multi-ticket ensemble in future work.

### 4.2 Diversity of Predictions

As an auxiliary analysis of behaviors, we show that each subnetwork produces diverse predictions. Because any existing diversity scores do not completely explain or justify the ensemble performance\(^6\), we discuss only rough trends in five popular metrics of classification diversity; Q statistic (Yule, 1900), ratio errors (Aksela, 2003), negative double fault (Giacinto and Roli, 2001), disagreement measure (Skalak, 1996), and correlation coefficient (Kuncheva and Whitaker, 2003). See Kuncheva and Whitaker (2003); Cruz et al. (2020) for their summarized definitions. As shown in Table 2, in all the metrics, winning-ticket subnetworks (*-LT) produced more diverse predictions than the baseline using the dense networks (BASELINE).

### 4.3 Diversity of Subnetwork Structures

We finally revealed the diversity of the subnetwork structures on MRPC. We calculated the overlap ratio of two pruning masks, which is defined as intersection over union, \(\text{IoU} = \frac{|m_i \cap m_j|}{|m_i \cup m_j|}\) (Chen et al., 2020). In Figure 4, we show the overlap ratio between the pruning masks for the five random seeds, i.e., \(\{m_{s_1}, \ldots, m_{s_5}\}\). At first, we can see that ACTIVE-LT and RANDOM-LT using the regularizers resulted in diverse pruning. This higher diversity could lead to the best improvements by ensembling, as discussed in Section 4.1. Secondly, BASE-LT produced surprisingly similar (99%) pruning masks with different random seeds. However, recall that even BASE-LT using the naive iterative magnitude pruning performed better than BASELINE. This result shows that even seemingly small changes in structure can improve the diversity of predictions and the performance of the ensemble.

<table>
<thead>
<tr>
<th></th>
<th>Q</th>
<th>R</th>
<th>ND</th>
<th>D</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>BASE-LT</td>
<td>0.96</td>
<td>0.72</td>
<td>-0.12</td>
<td>0.09</td>
<td>0.69</td>
</tr>
<tr>
<td>ACTIVE-LT</td>
<td>0.94</td>
<td>0.94</td>
<td>-0.11</td>
<td>0.11</td>
<td>0.62</td>
</tr>
<tr>
<td>RANDOM-LT</td>
<td>0.94</td>
<td>0.94</td>
<td>-0.11</td>
<td>0.10</td>
<td>0.63</td>
</tr>
</tbody>
</table>

Table 2: Diversity metrics on MRPC. The signs, \(\downarrow\) and \(\uparrow\), indicate that the metric gets lower and higher when the predictions are diverse. \(Q\) = Q statistic, \(R\) = ratio errors, \(ND\) = negative double fault, \(D\) = disagreement measure, \(C\) = correlation coefficient.

Figure 4: Overlap ratio of pruning masks \(m_{s_i}\) between different seeds on MRPC. The lower (yellowier) the value is, the more dissimilar the two masks are.

---

\(^5\)Although some studies (Prasanna et al., 2020; Chen et al.; Liang et al., 2021) reported that they found winning-ticket subnetworks on these tasks, our finding did not contradict it. Their subnetworks were often actually a little worse than their dense networks, as well as we found. Chen et al. (2020) defined winning tickets as subnetworks with performances within one standard deviation from the dense networks. Prasanna et al. (2020) considered subnetworks with even 90% performance as winning tickets.

\(^6\)For example, comparing BASELINE with RANDOM-LT of pruning ratio 20%, their average values of single/ensemble/difference are 91.38/91.93/+0.55 vs. 91.09/91.90/+0.81 on SST-2.

\(^7\)This also happens to experiments with roberta-base while multi-ticket ensemble still works well on MRPC.

\(^8\)Finding such a convenient diversity metric itself is still a challenge in the research community (Wu et al., 2021).

5 Conclusion

We raised a question on difficulty of ensembling large-scale pretrained models. As an efficient remedy, we explored methods to use subnetworks in a single model. We empirically demonstrated that ensembling winning-ticket subnetworks could outperform the dense ensembles via diversification and indicated a limitation too.
References


Alexander Kirillov, Bogdan Savchynskyy, Carsten Rother, Stefan Lee, and Dhruv Batra. 2016. CVPR tutorial: Diversity meets deep networks - inference, ensemble learning, and applications.


