We sincerely thank the three reviewers for their constructive comments and supports.

Response to Reviewer #1:

Q1: Compressing on the user part. **A:** GPUs are essential to doing effective deep learning. Compared with setting up their own servers, many users tend to spin up cloud instances with GPUs by balancing the flexibility and the investment, especially when the GPUs are only needed for several hours. In addition, not every user is a deep learning expert, and thus a cloud service would be expected to produce efficient deep neural networks according to users' needs.

Q2: Cloud platform is used, no compression is required. **A:** The compressed networks are often deployed on low-end computing devices, e.g. digital cameras and mobile phones. By doing this, the cloud service latency can be avoided and the application can be well executed even there is no Internet connection, which will improve the user experience.

Q3: Novelty. **A:** The novelty of this paper is twofold. Firstly, compressing network with little labeled data is a challenging task that has rarely been investigated by existing works in the field. To compensate the lack of labeled data, we introduce PU learning to select the most related data from a large pool of unlabeled data for the compression task. Secondly, we enhance the robustness of knowledge distillation to deal with data imbalance problem and noise. An attention based multi-scaled feature extractor is developed to cope with PU data better.

Q4: Comparison on using the attention-based feature extraction or not. **A:** The results are shown in Figure 2. The blue line is the result on using attention-based feature extractor. The red line is the result without attention-based feature extractor. The superiority of the proposed architecture leads to the accuracy improvement from 91.39% to 91.57%.

Q5: Comparison on part of, or all the unlabeled data set with KD method. Using robust KD or not. **A:** We include new experiments on CIFAR-10. Randomly choosing 50,000 samples leads to a 87.02% accuracy (91.56% for PU method choosing 50,000 samples). The result is bad since many negative data are selected in this way and is used to train the student network. Using all $1.2M$ unlabeled data leads to a 94.01% accuracy since all the positive data are guaranteed to be selected, but is about 12 times slower than using PU data (93.75% for PU method choosing $0.1M$ samples). Robust KD leads to 0.6% increase.

Q6: Minor problems. **A:** Thanks for this nice concern. All these typos will be corrected in the final version.

Response to Reviewer #2:

Q1: Figure 4. **A:** The first row represents the data in the original dataset. The second is the positive data selected by PU classifier. The third is the negative data selected by PU classifier. Spider is in both selected data and negative data, since the PU classifier does not classify unlabeled data with 100% accuracy, and spider represents the noise in positive data.

Response to Reviewer #3:

Q1: Advantages over related works. **A:** Theoretically, we utilize the strength of PU learning in augmenting data, and the strength of KD method in compressing neural networks, which is suitable for solving compression problem with little labeled data. Besides, we propose multi-feature network with attention and robust KD method to better solve the problem. Experimentally, we compare with the state-of-the-art methods in Table 3, and the accuracy results show the priority of the proposed method. Generally speaking, the proposed method is robust on the number of positive samples, and performs better. In other methods, the student network is compressed by pruning [1] or re-normalization [2] the giant teacher network with unlabeled data, which means that the detail architecture of the teacher network is required. In the proposed method, we only need the input and output interface of the teacher network instead of the whole architecture, which is more flexible for users to protect their own privacy.

Q2: The motivation to tackle imbalanced data problem is unclear. **A:** PU classifier treats all samples in the training set as 'positive', the specific category is undistinguishable. PU may select lots of data for some positive categories, while little for others. This causes data imbalanced problem. Eq.(7) and (8) are used to tackle the imbalanced data problem. We utilize the output of teacher network, and assign larger weights to categories with fewer samples. The weighted surrogate KD loss (Eq.(8)) can alleviate the imbalanced data problem.

Q3: Class prior π_p. **A:** In CIFAR-10 experiments, it is set equal to the ratio of manually selected data. In MNIST experiments, it is set to the real ratio of numbers in EMNIST. In ImageNet experiments, which is much like the reality settings, π_p is estimated by the prior estimation methods, such as [3]. In Figure 2, we analysis the relationship between classification accuracy and class prior. And it shows that our method is robust to the choice of class prior when it is not far from the true class prior.

[1] Tang Y, You S, Xu C, et al. Bringing Giant Neural Networks Down to Earth with Unlabeled Data. arXiv, 2019.

[2] He X, Cheng J. Learning Compression from Limited Unlabeled Data. ECCV, 2018.

[3] M. C. du Plessis and M. Sugiyama. Class prior estimation from positive and unlabeled data. NIPS, 2014.