1 Overall Response

We sincerely thank all the reviewers for their helpful comments and constructive suggestions.

Regarding comments on writing of this paper. We will carefully revise the manuscript and fix all the missing citations (R4), missing explanations (R2), confusing sentences (R3 and R4) and organization (R2 and R3).

Regarding the significance of our approach. The main goal is to propose a framework for dynamic inference with a common backbone. Although both distillation and attention are well-developed, to the best of our knowledge, it is the first attempt for combing them successfully with noticeable accuracy and speed gain. These improvements might open up new research opportunities for exploring the resource efficient ML including mathematical theory and algorithm. Actually, the motivation of using attention and distillation differs from their origins. More specifically, in proposed method, attention is utilized to separate features of different classifiers rather than to find the most informative pixels; and distillation is to transfer knowledge among classifiers in the common backbone, instead of among different models.

2 Response to Reviewer 2

Regarding the comparison with related work. Fig. 1 shows the comparison with Feedback Network[C1], FractalNet[C2] and NestedNet[C3] (results taken from their papers). Proposed method shows a large margin improvement of our proposed method. Compared with [C1,C2,C3], our method is different in two folds: (i) computation of shallow classifiers in the proposed method can be reused by the deeper classifiers, leading to possibility for dynamic inference; (ii) the accuracy of the early predictions in these work is lower than their baselines. In contrast, two of three shallow classifiers in the proposed method achieve higher accuracy than the baseline.

Regarding the specific questions of writing. (i)F_i and F_c denotes features of i_{th} and the deepest classifier. They are outputs of the convolutional layers, also named as "activation", which is unlearnable. (ii) The ensemble prediction is obtained by weighted sum of softmax layers outputs from all the classifiers.

Regarding the sensitivity of hyper-parameters α and λ. The proposed method is robust to hyper-parameters. All the reported accuracy in the paper shares the same hyper-parameters settings: $\alpha = 0.5$, $\lambda = 5 \times 10^{-7}$. In addition, we supplement the accuracy of ResNet18 with various α and λ on CIFAR100 in Figure 2 and 3. The observed accuracy varying range (less than 0.5%) is negligible compared with the improvements (3.37%) from baselines (77.09%).

Regarding the results in Table 2. The main goal of the proposed method is to accelerate the DNN inference with multi-classifiers in a common backbone. Compared with many related work which achieve acceleration at the expense of accuracy loss, the 1.26% accuracy increment on ImageNet in Table 2 may not be trivial with acceptable computation increment (no exceeding 5%, while ResNet50->101 brings another 98% computation for 1.5% accuracy improvement).

3 Response to Reviewer 3

Regarding the hardware friendliness. The hardware friendliness of the proposed method lies in the comparison with scalable neural networks such as SkipNet, as introduced in line69-74. The hardware friendliness of lightweight designs such as MobileNet, will be destroyed when they are utilized as backbones of SkipNet, because SkipNets added complex gated control units on every layer. In contrast, the thresholds based control method in SCAN can maintain the hardware friendliness of backbones for its simpleness.

4 Response to Reviewer 4

Regarding the ensemble for every level (Q1 and Q2). We implement your suggestion and the experiments on ResNet18-CIFAR100 show that every level ensemble leads to 1.64%, 1.48% and 1.67% accuracy increment on the 2_{nd}, 3_{rd} and 4_{th} depth prediction, respectively. With the improved ensemble classifiers on all the levels, scalable inference with new thresholds achieves 1.54% accuracy increment, compared with the origin model with same acceleration. We will further explore this as future works.

Regarding line192-193 (Q3). Observation (iv): the ensemble accuracy is 1.11% higher than that of the deepest classifier on average with almost no computation penalty. Observation (v): compared with classifiers trained individually, the proposed method can bring more accuracy increment on shallow classifiers than the deepest one.

Regarding the missing experiments (Q4). The details of parameter count and total memory usage will be added in revision. An average acceleration ($2.17\times$) and compression ($3.20\times$) ratio is given in line190-line191.

Regarding training time (Q5). Taking a pretrained model, ResNet50 on ImageNet takes 8.75 hours for 25 epoches trained on two 2080Ti GPU devices.

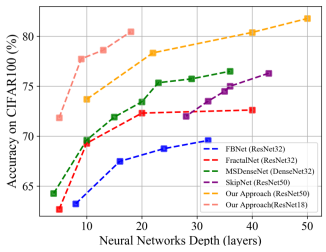

Figure 1: Comparisons.

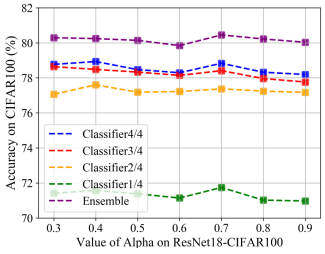

Figure 2: Sensitivity of α.

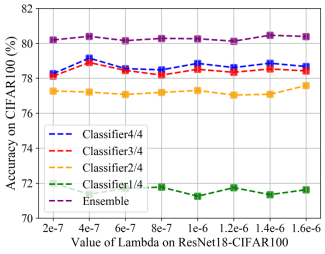

Figure 3: Sensitivity of λ.