We thank the reviewers for their valuable comments. We are glad that reviewers noted our paper as novel (R1: "idea is interesting .. and hasn't been tested before", R3: "approach to estimate weight of example is new", R4: "novel approach to curriculum learning by introducing new sets of parameters"), and have appreciated our results (R1: "results are extensive, and show significant improvement in several datasets", R3:"outperforms existing curriculum learning based approaches"). Below, we provide clarifications to the points they have raised, and provide additional experiments requested by the reviewers for improvement of rating.

Reviewer 1:

– **Requested Improvement** *"Decouple the effect of capacity increase and curriculum learning"*: We would like to clarify that the temperature parameters for class and instances are not parameters of the model. They are used only during training to modify the loss function. The architecture used for inference in our model and the baseline are identical, therefore the capacity (number of model parameters) is exactly the same. Hence, the gains we obtain on different datasets and tasks are due to curriculum learning. Thanks for pointing out a potential source of confusion; we will clarify this point in revision. We will also move related works section as suggested.

– *"Applying gradient descent to update parameters is not very original"*: Introducing trainable temperature parameters for instances and class in a dataset, and optimizing them through gradient descent is our original contribution.

– *"Comment on the importance of instance level parameters"*: In Table 1 (in paper) we present an ablation study where instance level curriculum provides additional improvement over class level curriculum on ImageNet and CIFAR100. In addition, the improvements on noisy datasets are solely due to instance level curriculum, since the per sample noise can only be mitigated by instance level curriculum. The reason class level curriculum can not help in this case, is because it assumes homogeneous difficulty across samples within a class.

– *"Missing analysis in paper is to track and analyze parameters"*: Please see Figure 3, and Figure 4 (left) in paper, where we have tracked and analyzed temperature parameters. Figure 3: For learning a detector, curriculum learns easier unoccluded instances first, followed by partial occlusion, and finally heavy occlusion. Figure 4 (left): Shows that the temperature of noisy samples keeps increasing over the course of training, hence decaying their contribution to learning process. We agree that this issue is important in the field of curriculum learning. For final version, we will provide more explicit examples demonstrating the learnt curriculum.

Reviewer 3:

– **Requested Improvement 1** *"It could be interesting to show results on the large WebVision Benchmark"*:

	R18	R18 + DCL
Top-1 Acc	66.3	**67.6**

As you suggested, we conducted experiments using ResNet18 with the same hyper-parameters as we have used for ImageNet in the paper. As shown in table (left), **we obtain an absolute improvement of 1.3% in top-1 accuracy on this challenging dataset** which in addition to being a large-dataset, has noisy labels, and is extremely imbalanced.

– **Requested Improvement 2** *"Would proposed curriculum change robustness to adversarial attacks"*:

Metric	R18	R18 + DCL
Top-1 Acc Adv.	44.3	**46.0**

Thanks for pointing us in this direction. As you suggested, we conducted an initial investigation with untargeted FGSM attack (Goodfellow et al., 2014) on ImageNet and found this direction to be promising. As shown in table (left), **model trained with curriculum obtains 1.7% higher accuracy (post adversarial attack)** compared to baseline.

– *"Curriculum based methods is an interesting direction to speed-up convergence"* : While speeding up training of DNNs was not our explicit goal, we did, thanks to your comment, an analysis for experiments reported in paper on ImageNet. We measured the relative reduction in number of epochs for our method to achieve the same accuracy as the baseline at various points during the training. **On average, our method requires 20% fewer epochs.**

Reviewer 4:

– **Requested Improvement** *"Results on larger training sets or datasets with large number of classes"*: In addition to ImageNet, we conducted new experiments on WebVision dataset (2.3 million training images) and obtain significant gains. Please see the first table above. When we analyzed temperature trajectories over the course of training (eg. Figure 1 right in paper), within the first few epochs, temperature of the hard instance (orange curve) peaks, decaying its contribution to learning. Empirically, most of the temperature variation for instances occurs early on during optimization (<30 epochs). Visiting the same data point 30 times (in multiple datasets of the scale of millions of data-points) was sufficient to learn the instance level temperature parameters. Nevertheless, we agree for datasets which contains of billions of training samples, and training loop might visit a data point only once or twice, alternative formulations should be explored.

– *"Why model without temperature parameters for class can not learn the same loss function?"*: Thank you for pointing this out, we can see this as an easy source of confusion. Unlike scaling each logit with temperature of its respective class (which could indeed be absorbed in weights), in our formulation, we scale all the logits of a sample, with temperature of the target class. In other words, in paper's Eq 1, notice that subscript of class temperature parameters is y_i (target label of sample i) in the denominator ($\sum_j \exp(z_j^i/\sigma_{y_i}^{class})$) and not j. This cannot be absorbed by scaling the weights of the model. We will also clarify the difference suggested on page 8.