Appendices

A. Adaptive Asymmetry dynamics

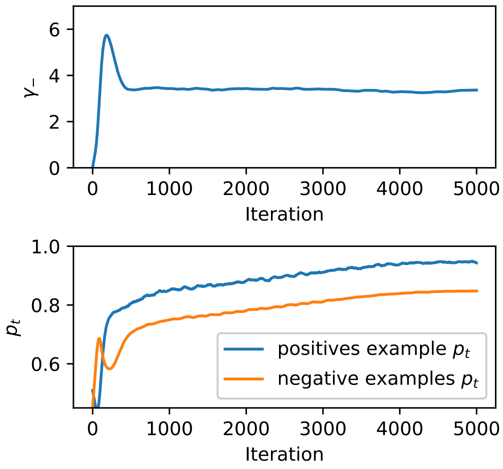

Figure 9: **Adaptive Asymmetry Dynamics.** Values of γ_- and Δp throughout the training, for $\Delta p_{\text{target}} = 0.1$. γ_+ is set to 0, m is set to 0.05.

B. Multi-Label General Training Details

Unless stated explicitly otherwise, we used the following training procedure: We trained the model for 60 epochs using Adam optimizer and 1-cycle policy [26], with maximal learning rate of 2e-4. For regularization, we used standard augmentation techniques [8]. We found that the common ImageNet statistics normalization [14, 8, 27] does not improve results, and instead used a simpler normalization - scaling all the RGB channels to be between 0 and 1. Following the experiments in section 3, for ASL we used $\gamma_- = 4$, $\gamma_+ = 0$ and $m = 0.05$, and for focal loss we used $\gamma = 2$. Our default and recommended backbone for multi-label training is TResNet-L. However, for fair comparison to previous works we also added ResNet101 backbone results on some datasets (TResNet-L and ResNet101 are equivalent in runtime).

C. Comparing MS-COCO On All Common Metrics

In Table 7 we compare ASL results, to known state-of-the-art methods, on all common metrics for MS-COCO dataset.

D. Comparing Loss Function on Pascal-VOC Dataset

In Table 8 we compare ASL results to other loss functions on Pascal-VOC dataset.

E. NUS-WIDE

NUS-WIDE [7] dataset originally contained 269,648 images from Flicker, that have been manually annotated with 81 visual concepts. Since some urls have been deleted, we were able to download only 220,000 images, similar to [10]. We can find in previous works [30, 21] other variants of NUS-WIDE dataset, and its hard to do a one-to-one comparison. We recommend using our publicly available variant[1] for standardization and a completely fair comparison in future works. We used the standard 70-30 train-test split [10, 30, 21]. Our training settings were identical to the ones used for MS-COCO. We can see from Table 9 that ASL improves the known state-of-the-art results on NUS-WIDE by a large margin. In Table 10 we compare ASL results to other loss functions on NUS-WIDE dataset, again showing that ASL outperform cross-entropy and focal-loss..

F. Open Images

Open Images (v6) [17] is a large scale dataset, which consists of 9 million training images and $125,436$ test images. It is partially annotated with human labels and machine-generated labels. The scale of Open Images is much larger than previous multi-label datasets such as NUS-WIDE, Pascal-VOC and MS-COCO. Also, it contains a considerable amount of unannotated labels. That allows us to test ASL on extreme classification [39], and high mislabeleing scenarios. Due to missing links on flicker, we were able to download only $114,648$ test images from Open Images dataset, which contain about $5,400$ unique tagged classes. For dealing with the partial labeling methodology of Open Images dataset, we set all untagged labels as negative, with reduced weights. Due to the large the number of images, we trained our network for 30 epochs on input resolution of 224, and finetuned it for 5 epochs on input resolution of 448. Since the level of positive-negative imbalancing is significantly higher than MS-COCO, we increased the level of loss asymmetry: For ASL, we trained with $\gamma_- = 7, \gamma_+ = 0$. For Focal loss, we trained with $\gamma = 4$. Other training details are similar to the ones used for MS-COCO.

To the best of our knowledge, no results for other methods were published yet for v6 variant of Open Images. Hence, we compare ASL only to the other common loss

[1] Our NUS-WIDE variant can be download from: https://drive.google.com/file/d/0B7IzDz-4yH_HMFdiSE44R1lselE/view

Method	All							Top 3					
	mAP	CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CADM [5]	82.3	82.5	72.2	77.0	84.0	75.6	79.6	87.1	63.6	73.5	89.4	66.0	76.0
ML-GCN [6]	83.0	85.1	72.0	78.0	85.8	75.4	80.3	87.2	64.6	74.2	89.1	66.7	76.3
KSSNet [21]	83.7	84.6	73.2	77.2	87.8	76.2	81.5	-	-	-	-	-	-
MS-CMA [36]	83.8	82.9	74.4	78.4	84.4	77.9	81.0	86.7	64.9	74.3	90.9	**67.2**	77.2
MCAR [12]	83.8	85.0	72.1	78.0	88.0	73.9	80.3	88.1	**65.5**	**75.1**	91.0	66.3	76.7
ASL (TResNet-L)	**86.6**	**87.2**	**76.4**	**81.4**	**88.2**	**79.2**	**81.8**	**91.8**	63.4	**75.1**	**92.9**	66.4	**77.4**

Table 7: **Comparison of ASL to known state-of-the-art models on MS-COCO dataset.** All metrics are in %. Results are reported for input resolution 448.

Method	mAP (ImageNet Only Pretrain)	mAP (Extra Pretrain Data)
CE	93.2	95.0
Focal Loss	93.8	95.4
ASL	**94.6**	**95.8**

Table 8: **Comparison of ASL to other loss functions on Pascal-VOC dataset.** Metrics are in %.

Method	mAP	CF1	OF1
S-CLs [21]	60.1	58.7	73.7
MS-CMA [36]	61.4	60.5	73.8
SRN [40]	62.0	58.5	73.4
ICME [5]	62.8	60.7	74.1
ASL (ResNet101)	**63.9**	**62.7**	**74.6**
ASL (TResNet-L)	**65.2**	**63.6**	**75.0**

Table 9: **Comparison of ASL to known state-of-the-art models on NUS-WIDE dataset.** All metrics are in %.

Method	mAP	CF1	OF1
CE (Ours)	63.1	61.7	74.6
Focal loss (Ours)	64.0	62.9	74.7
ASL (Ours)	**65.2**	**63.6**	**75.0**

Table 10: **Comparison of ASL to known other loss functions on NUS-WIDE dataset.** All metrics are in %.

functions in multi-label classification. Yet we hope that our result can serve as a benchmark for future comparisons.

Open Images Results appear in Table 11. We can see from Table 11 that ASL significantly outperforms focal loss and cross-entropy on Open Images, demonstrating that ASL is suitable for large datasets and extreme classification cases.

Method	micro mAP[%]	macro mAP[%]
CE	84.8	92.0
Focal Loss	84.9	92.2
ASL	**86.3**	**92.8**

Table 11: **Comparison of ASL to focal loss and cross-entropy on Open Images V6 dataset.**

G. Fine-Grain Single-Label Classification Results

For testing ASL on fine-grain single-label classification, we chose to work on the competitive Herbarium 2020 FGVC7 Challenge [16]. The goal of Herbarium 2020 is to identify vascular plant species from a large, long-tailed collection Herbarium specimens provided by the New York Botanical Garden (NYBG). The dataset contains over 1M images representing over 32,000 plant species. This is a dataset with a long tail; there are a minimum of 3 specimens per species, however, some species are represented by more than a hundred specimens. The metric chosen for the competition is macro F1 score. For Focal loss, we trained with $\gamma = 2$. For ASL, we trained with $\gamma_- = 4, \gamma_+ = 0$. The metric chosen for the competition is macro F1 score. In Table 12 we bring results of ASL on Herbarium dataset, and compare it to regular focal loss. We can see from Table 12 that ASL outperforms focal loss on this fine-grain single-label classification dataset by a large margin. Note that Herbarium 2020 was a CVPR-Kaggle classification competition. Our ASL test-set score would achieve the 3rd place in the competition, among 153 teams.

Method	macro F1 [%]
Focal Loss	76.1
ASL	**77.6**

Table 12: **Comparison of ASL to focal loss on Herbarium dataset**. Macro-F1 is the competition official metrics. All results are on an unseen private-set.

H. Object Detection Results

For testing ASL on object detection, we used the MS-COCO [20] dataset (object detection task), which contains a training set of 118k images, and an evaluation set of 5k images. For training, we used the popular mm-detection [3] package, with the enhancements discussed in ATSS [38] and FCOS [28] as the object detection method. We trained a TResNet-M [25] model with SGD optimizer for 70 epochs, with momentum of 0.9 , weight decay of 0.0001 and batch size of 48. We used learning rate warm up, initial learning rate of 0.01 and 10x reduction at epochs 40, 60. For ASL we used $\gamma_+ = 1, \gamma_- = 2$. For focal loss we used the common value, $\gamma = 2$ [19]. Note that unlike multi-label and fine-grain single-label classification datasets, for object detection $\gamma_+ = 0$ was not the optimal solution. The reason for this might be the need to balance the contribution from the 3 losses used in object detection (classification, bounding box and centerness). We should further investigate this issue in the future.

Our object detection method, FCOS [28], uses 3 different types of losses: classification (focal loss), bounding box (IoU loss) and centerness (plain cross-entropy). The only component which is effected by the large presence of background samples is the classification loss. Hence, for testing we replaced only the classification focal loss with ASL.

In Table 13 we compare the mAP score obtained from ASL training to the score obtained with standard focal loss. We can see from Table 13 that ASL outscores regular focal loss, yielding an 0.4% improvement to the mAP score.

Method	mAP [%]
Focal Loss	44.0
ASL	**44.4**

Table 13: **Comparison of ASL to focal loss on MS-COCO detection dataset.**

A. Handling Mislabeled Samples

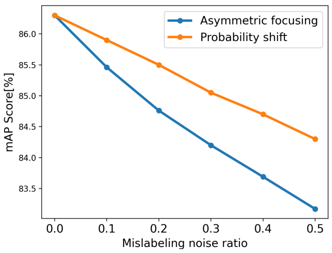

Figure 10: **Handling mislabeled annotation.**

We conducted an experiment on MS-COCO dataset, where we simulated a noisy dataset by randomly flipping a known ratio of positive samples to negative ones, and measure the accuracy. We compare in Figure 1 results of our two proposed asymmetric mechanisms: asymmetric focusing and probability margin. We see that the probability margin mechanism indeed enables to handle the mislabeling noise well, and the relative improvement is higher as the noise ratio increases.

B. Probability Analysis on Inference

Following Figure 4, we present in Figure 11 the averaged probability relations during inference, on the validation set. As can be seen, ASL reduces the gap between the positive and negative averaged probabilities also on the validation set.

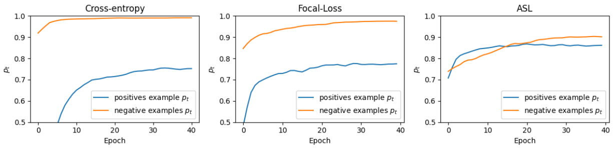

Figure 11: **Probability analysis at inference time.**