7. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks: Supplementary Materials

7.1. Performance Gains Versus α

The α asymmetry hyperparameter, we argued, allows us to accommodate for various different priors on the symmetry between tasks. A low value of α results in gradient norms which are of similar magnitude across tasks, ensuring that each task has approximately equal impact on the training dynamics throughout training. A high value of α will penalize tasks whose losses drop too quickly, instead placing more weight on tasks whose losses are dropping more slowly.

For our NYUv2 experiments, we chose $\alpha = 1.5$ as our optimal value for α, and in Section 5.4 we touched upon how increasing α pushes the task weights $w_i(t)$ farther apart. It is interesting to note, however, that we achieve overall gains in performance for almost all positive values of α for which GradNorm is numerically stable[3]. These results are summarized in Figure 7.

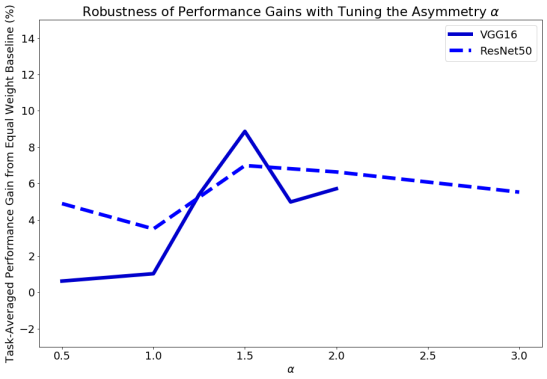

Figure 7. **Performance gains on NYUv2+kpts for various settings of α.** For various values of α, we plot the average performance gain (defined as the mean of the percent change in the test loss compared to the equal weights baseline across all tasks) on NYUv2+kpts. We show results for both the VGG16 backbone (solid line) and the ResNet50 backbone (dotted line). We show performance gains at all values of α tested, although gains appear to peak around $\alpha = 1.5$. No points past $\alpha > 2$ are shown for the VGG16 backbone as GradNorm weights are unstable past this point for this particular architectural backbone.

We see from Figure 7 that we achieve performance gains at almost all values of α. However, for NYUv2+kpts in particular, these performance gains seem to be peaked at $\alpha = 1.5$ for both backbone architectures. Moreover, the ResNet architecture seems more robust to α than the VGG architecture, although both architectures offer a similar level of gains with the proper setting of α. Most importantly, the consistently positive performance gains across all values of α suggest that *any kind of gradient balancing* (even in suboptimal regimes) is healthy for multitask network training.

[3] At large positive values of α, which in the NYUv2 case corresponded to $\alpha \geq 3$, some weights were pushed too close to zero and GradNorm updates became unstable.

7.2. Additional Experiments on a Multitask Facial Landmark Dataset

We perform additional experiments on the Multitask Facial Landmark (MTFL) dataset (Zhang et al., 2014). This dataset contains approximately 13k images of faces, split into a training set of 10k and a test set of 3k. Images are each labeled with (x, y) coordinates of five facial landmarks (left eye, right eye, nose, left lip, and right lip), along with four class labels (gender, smiling, glasses, and pose). Examples of images and labels from the dataset are given in Figure 8.

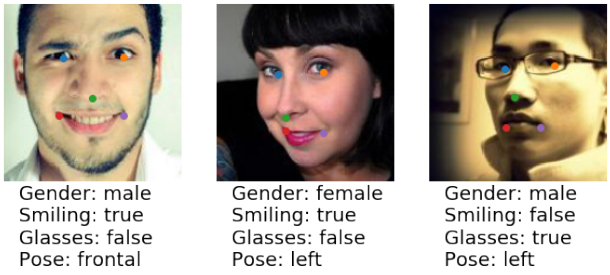

Figure 8. **Examples from the Multi-Task Facial Landmark (MTFL) dataset.**

The MTFL dataset provides a good opportunity to test GradNorm, as it is a rich mixture of classification and regression tasks. We perform experiments at two different input resolutions: 40x40 and 160x160. For our 40x40 experiments we use the same architecture as in (Zhang et al., 2014) to ensure a fair comparison, while for our 160x160 experiments we use a deeper version of the architecture in (Zhang et al., 2014); the deeper model layer stack is [CONV-5-16][POOL-2][CONV-3-32]^2[POOL-2][CONV-3-64]^2[POOL-2][[CONV-3-128]^2[POOL-2]]^2[CONV-3-128]^2[FC-100][FC-18], where CONV-X-F denotes a convolution with filter size X and F output filters, POOL-2 denotes a 2x2 pooling layer with stride 2, and FC-X is a dense layer with X outputs. All networks output 18 values: 10 coordinates for facial landmarks, and 4 pairs of 2 softmax scores for each classifier.

The results on the MTFL dataset are shown in Table 3. Keypoint error is a mean over L_2 distance errors for all five facial landmarks, normalized to the inter-ocular distance, while failure rate is the percent of images for which keypoint error is over 10%. For both resolutions, GradNorm outperforms other methods on all tasks (save for glasses

Table 3. **Test error on the Multi-Task Facial Landmark (MTFL) dataset for GradNorm and various baselines**. Lower values are better and best performance for each task is bolded. Experiments are performed for two different input resolutions, 40x40 and 160x160. In all cases, GradNorm shows superior performance, especially on gender and smiles classification. GradNorm also matches the performance of (Zhang et al., 2014) on keypoint prediction at 40x40 resolution, even though the latter only tries to optimize keypoint accuracy (sacrificing classification accuracy in the process).

Method	Input Resolution	Keypoint Err. (%)	Failure. Rate. (%)	Gender Err. (%)	Smiles Err. (%)	Glasses Err. (%)	Pose Err. (%)
Equal Weights	40x40	8.3	27.4	20.3	19.2	8.1	38.9
(Zhang et al., 2014)	40x40	8.2	**25.0**	-	-	-	-
(Kendall et al., 2017)	40x40	8.3	27.2	20.7	18.5	8.1	38.9
GradNorm $\alpha = 0.3$	40x40	**8.0**	**25.0**	**17.3**	**16.9**	8.1	38.9
Equal Weights	160x160	6.8	15.2	18.6	17.4	8.1	38.9
(Kendall et al., 2017)	160x160	7.2	18.3	38.1	18.4	8.1	38.9
GradNorm $\alpha = 0.2$	160x160	**6.5**	**14.3**	**14.4**	**15.4**	8.1	38.9

and pose prediction, both of which always quickly converge to the majority classifier and refuse to train further). GradNorm also matches the performance of (Zhang et al., 2014) on keypoints, even though the latter did not try to optimize for classifier performance and only stressed keypoint accuracy. It should be noted that the keypoint prediction and failure rate improvements are likely within error bars; a 1% absolute improvement in keypoint error represents a very fine sub-pixel improvement, and thus may not represent a statistically significant gain. Ultimately, we interpret these results as showing that GradNorm significantly improves classification accuracy on gender and smiles, while at least matching all other methods on all other tasks.

We reiterate that both glasses and pose classification always converge to the majority classifier. Such tasks which become "stuck" during training pose a problem for GradNorm, as the GradNorm algorithm would tend to continuously increase the loss weights for these tasks. For future work, we are looking into ways to alleviate this issue, by detecting pathological tasks online and removing them from the GradNorm update equation.

Despite such obstacles, GradNorm still provides superior performance on this dataset and it is instructive to examine why. After all loss weights are initialized to $w_i(0) = 1$, we find that (Kendall et al., 2017) tends to increase the loss weight for keypoints relative to that of the classifier losses, while GradNorm aggressively decreases the relative keypoint loss weights. For GradNorm training runs, we often find that $w_{\text{kpt}}(t)$ converges to a value ≤ 0.01, showing that even with gradients that are smaller by two orders of magnitude compared to (Kendall et al., 2017) or the equal weights method, the keypoint task trains properly with no attenuation of accuracy.

To summarize, GradNorm is the only method that correctly identifies that the classification tasks in the MTFL dataset are relatively undertrained and need to be boosted. In contrast, (Kendall et al., 2017) makes the inverse decision by placing more relative focus on keypoint regression, and often performs quite poorly on classification (especially for higher resolution inputs). These experiments thus highlight GradNorm's ability to identify and benefit tasks which require more attention during training.