Supplementary Material
A Unified Surrogate Loss Framework by Re-expression and Interpolation

Lanlan Liu[1,2], Mingzhe Wang[2], and Jia Deng[2]

[1] University of Michigan, Ann Arbor MI 48105, USA
llanlan@umich.edu
[2] Princeton University, Princeton NJ 08544, USA
{mingzhew,jiadeng}@cs.princeton.edu

1 Architecture and Training Details

1.1 Binary Classification

Network Architecture The network takes a 28×28 image as input and flatten it to be one dimension with 784 elements. It then pass the flattened input through two fully-connected layers, one with 500 neurons and one with 300 neurons. Each fully-connected layer has a ReLU activation layer followed. The prediction is then given with an output layer with a 2-way output.

Training Details The baseline loss is a 2-class cross-entropy loss. Our loss is formulated as described in Sec. 4.2. The h function is the sigmoid relaxation of the binary variables. The g function is the interpolation over anchors of the binary variables. For example, the anchor that gives the best performance is that all positive examples have higher scores than negative examples—all $b_{i,j} = 1$. Following Sec. 3.4, three types of anchors are generated randomly. The good anchors and nearby anchors are obtained by flipping one binary bit from the best anchor and the anchor at the current training step respectively.We sample 16 anchors for each type.

We train models with the baseline loss and UniLoss with SGD using the same training schedule: with a fixed learning rate of 0.01 for 30 epochs.

1.2 Pose Estimation

Network Architecture We use the Stacked Hourglass Network architecture. It takes a 224×224 image as input and passes it through one hourglass block as described in [3]. It outputs a heatmap for each joint. A heatmap essentially gives scores measuring how likely is the joint to be there for each pixel or region centered at that pixel.

Training Details The baseline loss is the Mean Squared Error (MSE) between the predicted heatmaps and the manually-designed "ground truth" heatmaps. More specifically, the target "ground truth" heatmap has a 2D Gaussian bump centered on the ground truth joint. The shape of the Gaussian bump is controlled by its variance σ and the bump size. The commonly chosen σ is 1 and the bump size 7.

Our loss is formulated as in Sec. 4.3. The h function is the sigmoid relaxation of the binary variables. The g function is the interpolation over anchors of the binary variables. For example, the anchor that gives the best performance is that all binary values to be 1. Following Sec. 3.4, three types of anchors are generated. For good anchors, we flip a small number of bits from the best. Nearby anchors are flipped from the current configuration by randomly picking a positive/negative pixel in the current output heatmaps and flipping all bits associated with this pixel. We sample 16 anchors for each type.

We train the model both UniLoss and MSE with RMSProp [2] using an initial learning rate 2.5e-4 for 30 epochs and then divided by 4 for every 10 epochs until 50 epochs.

1.3 Classification

Network Architecture We use the ResNet-20 architecture [1]. The network takes a 32×32 image as input. The input image first goes through a 3×3 convolution layer followed by a batch normalization layer and a ReLU layer. It then goes through three ResNet building blocks with down-sampling in between. After an average pooling layer, the output layer is a fully-connected layer with a 10-way output.

Training Details Our baseline loss is a 10-way cross-entropy (CE) loss. Our loss is formulated as in Sec. 4.4. The h function is the sigmoid relaxation of the binary variables. The g function is the interpolation over anchors of the binary variables. For example, the anchor that gives the best performance is that all binary values to be 1, meaning that for each image, the ground truth class has higher score than every other class. The good anchors and nearby anchors are obtained by flipping one binary bit from the best anchor and the anchor at the current training step respectively.We sample 16 anchors for each type.

We use the same data augmentation for both with random cropping with a padding of 4 and random horizontal flipping. We also pre-process the images with per-pixel normalization. We train both models with SGD using an initial learning rate of 0.1, divided by 10 and 100 at the 140 epoch and the 160 epoch, with a total of 200 epochs on CIFAR-10. On CIFAR-100, we train baseline with the same training schedule and UniLoss with 5x training schedule but 20% binary variables at each step.

Table 1. The L2 distances and rank correlation coefficients between the approximated metric value and true metric value pairs for binary configurations with various hamming distances from those encountered during training.

Hamming Distance (of #binaries)	0	$\frac{1}{512}$	$\frac{1}{128}$	$\frac{1}{32}$	$\frac{1}{8}$	$\frac{1}{2}$
L2 Distance	0.19	0.19	0.57	2.83	8.49	12.25
Rank Correlation Coeffcient	0.98	0.97	0.92	0.70	0.35	0.15

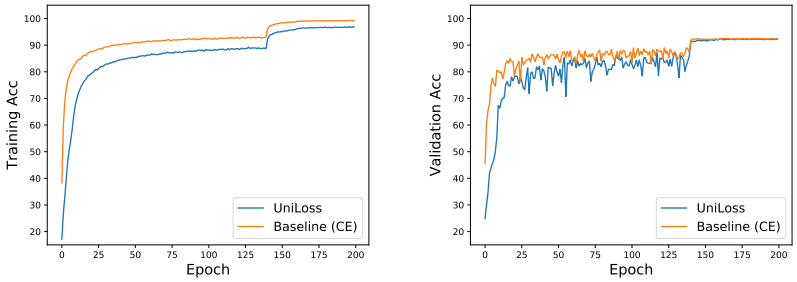

Fig. 1. Training and validation accuracy curves for Cross-Entropy (CE) and UniLoss on CIFAR-10.

2 Additional Experimental Results

2.1 Analysis of the Interpolation of g

To evaluate how well the interpolator approximates the true performance metric, we sample binary configurations with various hamming distances from those encountered during training. The hamming distances ranges from 0, $\frac{1}{512}, \ldots, \frac{1}{2}$ of total number of binary variables (#binaries). We compute the L2 distances and rank correlation coefficients between the approximated metric value and true metric value pairs, as in Table 1. We see that approximation is quite good when distance is small but poor when distance is large.

2.2 Training and Validation Curves

We present the training accuracy and validation accuracy curves over epochs in Fig. 1 and Fig. 2. We see that while UniLoss has similar validation accuracy (PCKh) later in the training with the CE (MSE) baseline, its training accuracy (PCKh) is slightly lower than the baseline over the whole training. One hypothesis is that due to the noise introduced by the randomness in anchor sampling, UniLoss naturally has a regularization effect compared to conventional losses.

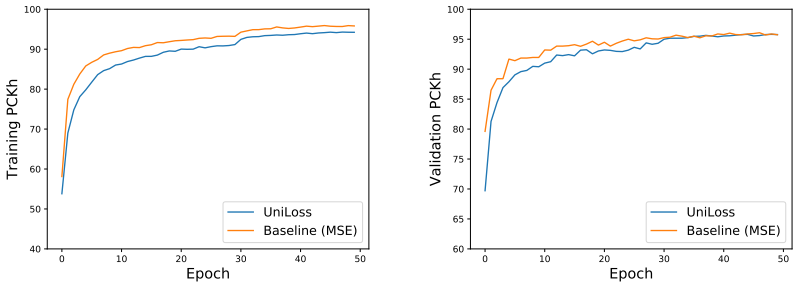

Fig. 2. Training and validation PCKh curves for Mean Square Error (MSE and UniLoss on MPII.

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
2. Hinton, G., Srivastava, N., Swersky, K.: Lecture 6a overview of mini–batch gradient descent. Coursera Lecture slides https://class. coursera. org/neuralnets-2012-001/lecture,[Online (2012)
3. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)