

**PhD**

**Do Deep Nets Really Need to be Deep ax1410 nips14**

2018/02/22 11:07 AM

we need to train on the logits because some information is lost by the soft max operation;

for instance values of 10, 20 and 30 can end up giving almost 0 for the first two and almost 1 for the last;

in addition, different sets of logits can produce the same probability values of the soft max so we definitely do lose some important information if we train on the soft max output Instead of the logits directly;

11:14 AM

The loss function used is simply the L2 loss between the logits;

it seems that normalizing the logits by subtracting the mean and dividing by the standard deviation can also provide a slight improvement in performance but is not necessary for training;