A Note on Weight Initialization

Tyler R. Scott, Karl Ridgeway, Michael C. Mozer
{tysc7237, karl.ridgeway, mozer}@colorado.edu

3 March 2019

In further explorations of our models, we noticed that the weight-initialization procedure played a larger role in performance than we had expected. We assumed that because we were essentially comparing different loss functions with the same architecture, the weight-initialization procedure would affect all conditions in the same manner. This turned out to be false. Our original scheme used a standard normal distribution, $\mathcal{N}(0, 1)$, to initialize both the weights and biases in the fully-connected and convolutional layers across all methods. However, using a Xavier uniform initialization for the weights and setting the biases to zero led to significantly improved classification accuracy for both the baseline and weight adaptation models, while harming performance for the non-adapted embedding models. Upon further experimentation, we found that the Xavier uniform and Xavier normal initializations led to similar results and furthermore, initializing the biases to zero was not the critical factor in the success of the Xavier initialization. We hypothesized that initializing the biases to zero and using a standard normal distribution for the weights would lead to similar results as seen with the Xavier initializer, but this was not the case. Thus, to our surprise, the weight-initialization procedure interacts with the choice of loss function, and different initializations favor different methods for few-shot learning. Regardless, the qualitative results of our paper still hold:

- Adapted histogram loss and adapted prototypical networks beat out all other methods (non-adapted embeddings, weight adaptation, and baseline).
- Weight adaptation is minimally better than the baseline.
- For small k, the embedding methods outperform weight adaptation and the baseline model.