What is the Effect of Importance Weighting in Deep Learning?
(2018)

Jonathon Byrd, Zachary C. Lipton

Summary

Contributions

Importance sampling is a fundamental tool in statistics and machine learning often used when we want to estimate a quantity on some target distribution, but can only sample from a different source distribution. Concretely, given n samples $x_1, \ldots, x_n \sim p(x)$, and the task of estimating some function of the data, say $f(x)$, under the target distribution q, importance sampling produces an unbiased estimate by weighting each sample x according to the likelihood ratio $q(x)/p(x)$:

$$\mathbb{E}_p\left[\frac{q(x)}{p(x)}f(x)\right] = \int_x f(x)\frac{q(x)}{p(x)}p(x)dx = \int_x f(x)q(x)dx = \mathbb{E}_q[f(x)]$$

where in domain adaption, the weighting can either depend on x in the case of covariate shift: $q(x)/p(x)$ or y in the case of label shift $q(y)/p(y)$.

In this paper, the authors investigate the effect of importance weighting on deep neural network and show that given the optimizer used, regularization and normalization of choice, the effect varies. Calling into question the standard application of importance weighting tin difference deep learning tasks.

Results

The authors investigate the effects of importance weighting on neural networks on two-dimensional toy datasets, the CIFAR-10 image dataset, and the Microsoft Research Paraphrase Corpus (MRPC) text dataset.

- The effects of importance weighting vanish as training progresses. After many epochs of training, there is no clear correlation between the class-based importance weights and the classification ratios on either test set images, out-of-domain images, or random vectors.
- Sub-sampling the training set instead of down-weighting the loss function,does have a noticeable effect on classification ratios. Models assign more CIFAR-10 test images from all classes (both in-domain and out-of-domain) as well as more random noise images to the majority class. Notably, weighting the loss function to counteract this imbalance during training does not balance the classification ratios.
- Models with more extreme weighting converge more slowly in decision boundary, and convergence in classification ratio begins to occur long after perfect training accuracy is achieved.
- An effect of importance weighting on classification ratios is present after training ResNet models for 1000 epochs.However, when batch normalization is removed from the model, classification ratios during training resemble those of the ordinary convolutional network
- In the CIFAR experiments, L2 regularization slows the convergence in classification ratios of all models.

Conclusion

The effects from importance weighting on deep networks may only occur in conjunction with early stopping, disappearing asymptotically. For example, when correcting for label shift in test data,test accuracy may deteriorate over training epochs even as the classifier improves owing to the diminishing effect of importance weighting, models with different importance weightings also have high agreement even on out-of-domain images, providing further evidence that the learned decision boundaries are similar.

While importance weighting does appear to have some effect when applied with residual networks we ob-serve that these effects vanish when batch normalization is removed.

Some effect of importance weighting can be realized when applied in combination with L2 regularization. The L2 penalty prevents SGD from reaching the large norm solutions whose loss is dominated by the support vectors, thus preventing convergence to max-margin-like solutions.

Weighting the loss function of deep networks fails to correct for training set class imbalance. However,sub-sampling a class in the training set clearly affects the networks predictions. This finding indicates that perhaps sub-sampling can be an alternative to importance weighting for deep networks on sufficiently large training sets.