

Number 5: Paper Review

1. [2 points] The paper shows that training with very weak (noisy) labels still leads to robust learning of features that generalize well. Why is deep learning so robust to noise? Provide some conjectures, relating it to what you know about machine/deep learning and optimization.

We discussed the softmax layer before in class. For classification problems, this layer helps factor in the relative scores of each class label. The inter class relationships may be difficult for noise to affect significantly.

Also, deep neural nets are widely trained using some form of gradient descent optimization, which cases where the loss is non-convex, converges to local minima, or an estimation of the true global minima. Estimations can be more invariant to noise. For example, taking the mean of a sample can reduce the effects of noise.

Noise could also be viewed as simply part of the overall problem. The noise just raises the complexity of the problem, thus raising the threshold for the model capacity to sufficiently solve the problem. Deep networks have many non-linear layers, which result in a model with very high capacity. It is possible that modern deep networks are deep enough to have a high enough model complexity to describe problems that incorporate noise.

2. [2 points] Another finding is that learning features using similar label spaces (i.e. on categories that overlap with what you are hoping to generalize to) is more successful than learning features on dissimilar label spaces. Why might that be the case?

Often times, label spaces can be converted into numerical vector spaces. So the problem can be rephrased as why learning features from a “similar” or “overlapping” vector space is more successful than learning on “non-overlapping” vector spaces.

From a linear algebra perspective, the essence of a space is encoded in its basis so the notion of “overlapping” could be represented by intersecting bases. This means that there is a subspace which is spanned by the vectors that belong to both bases. This subspace may contain information that is shared across both vector spaces. Learning features related to this subspace may reveal useful information about both spaces, which includes the target space. Dissimilar label spaces will not share this subspace and useful overlapping information cannot be learned.