DeepBayes Summer School - Paper Assignment

Karolina Stosio

April 10, 2019

4. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles

a.

- Instead of outputting a single value, the NN is modified to output predicted μ and σ^2 of a Normal distribution. Consequently, instead of using the mean square error,the objective function is the negative log-likelihood of the true value of y under the distribution $\mathcal{N}(\mu(x), \sigma(x))$.
- The architecture and the loss define a Gaussian distribution on the outputs.
- The ensemble of networks induce a mixture of Gaussian distributions.

b.

- Adversarial examples are such modifications of examples (e.g. images) that are close to the originals (e.g. undistinguished from original images by humans) but generate different predictions.
- Adversarial examples provide samples from the area where the prediction distribution changes faster than the input distribution. The logic of including them in the training is to learn a smoother prediction distribution (by forcing same classification for sufficiently small variations of the inputs).
- Not in the light of the definition mentioned above.

c.

The most straightforward thing is to follow a procedure from the section 3.5 of the paper. An ensemble of network trained on the data set (with different random initialisation) should learn the real distribution of inputs and outputs, provided there are enough good examples. Then, an entropy of the prediction vector can be used to check the degree of agreement within the ensemble. If the entropy is high, the ensemble is uncertain about this example and it can be considered corrupted. If the entropy is low, the data point is likely to be a good example.

Alternatively, one could train an ensemble of networks on bootstrapped subsets of the original data set. This could have an advantage over the previous approach, since while the distribution of the original data should remain unchanged between botstrapped subsets, the same may not be true for the distribution of the corrupted examples. If that is the case, using the bootstrap for the training could enhance the difference between the entropies of the prediction vector.

Practically, an iterative approach could be used, in which the following steps are repeated until the ensemble reaches a desirable degree of agreement for all examples: 1. the networks are trained on the data set, 2. each example is evaluated by an ensemble, 3. a new data set is created by removing examples for which the entropy of the prediction vector exceeded the selected threshold. The remaining data set should consist of good images.