Re-derivation of Out of Distribution Paper

Arjun Gupta

September 4, 2019

This document is a sheet to do some derivations /explanations for the Out of Distribution Paper (https://arxiv.org/pdf/1907.04572.pdf). The majority of the explanation for NRM is based on the original paper at (https://arxiv.org/abs/1811.02657).

0.1 Latent Variable Likelihood

The Out of Distribution Paper uses the likelihood of latent variables in the Neural Rendering Model to assess whether a given sample is in distribution or out of distribution. The neural rendering model re-creates the image by inverting the process the CNN uses to make the prediction. **An important realization is that the NRM model is not just the decoder. It is the combination of the CNN at the beginning and the decoder at the end.** The NRM model uses the same weights and filters ($W(l)^T$) as the CNN transposed to regenerate the image. The latent variables are the relu activations in each layer $s(l)$ and the maximum region of the max pooling $t(l)$. The optimal latent variables $z^*(l) = \{s^*(l), t^*(l)\}$ are derived from the forward pass of the CNN used to get the class probabilities. The process for generating an image in NRM is:

1. Feed the image x into the CNN, keep track of activations $s^*(l)$ and the locations of max pooling $t^*(l)$. The result is a vector of class probabilities y representing $p(y|x)$.

2. Based on $p(y|x)$, we choose a class template μ_y as our coarsest image $h(L)$.

3. Latent variables $z^*(l) = \{s^*(l), t^*(l)\}$ from the CNN step are used at each layer to iteratively refine the base image using the formula:

$$h(l-1) = \sum_{p \in h(l)} T(t^*(l,p))B(l,p)W^T(l)(s^*(l,p)h(l,p)) \tag{1}$$

   Which uses $s^*(l)$ to select whether a given template is to be rendered or not. $W(l)^T$ is the transposed weight at the corresponding layer of the CNN. $B(l,p)$ pads the template (which is the size of the convolution filter) with zeros to make it the size of the image. $T(t^*l,p)$ is the translation matrix, which translates the rendered template to the area of max pooling if applicable.

4. The final rendered image then has noise added to it:

$$x|\{z,y\} = N(h(0), \sigma^2 I) \tag{2}$$

Therefore, the process for generating the output at each layer requires that we do deconvolutions with each of the filters on the pixels that were activated. The distribution of the latent variables $z = \{t(l), s(l)\}$ can be seen as based on a structured prior. The paper claims this prior is:

$$\begin{aligned} \eta(y,z) &= \sum_{l=1}^{L} \langle b(l,t), s(l) \odot h(l) \rangle \\ p(y,z) &= \frac{\exp\left(\eta(y,z)\right)\pi_y}{\sum_{y_i,z_i} \exp\left(\eta(y_i,z_i)\right)\pi_{y_i}} \end{aligned} \tag{3}$$

Where $b(l)$ are the parameters of the prior distribution, $s(l)$ are the activations, and $t(l)$ are the translations for layer l. All three of these are of the dimensions of $h(l)$ on which they are acting. They construct this prior specifically because it is the conjugate prior for the model likelihood. Because it is a conjugate prior, the parameters $b(l)$ in the prior become the biases in the CNN. **It is important to note that the forward pass of the CNN is step that "samples" the optimal latent variables z^* from the prior distribution.**

0.2 Data Likelihood

The second theorem that they use in the out of distribution paper is used to estimate the probability of the data x_i:

$$\log p(x_i) \leq E_{y_i,z_i}\left[p\left(x_i,(y_i,z_i)\right)\right] \approx -\frac{1}{2\sigma^2}||x_i - h(y_i^*, z_i^*, 0)||^2 + \log \pi_{z_i^*|y_i^*} \tag{4}$$

The first in equality is trivially true. This is effectively just a combination of the Reconstruction loss (first term) and the likelihood of the latent variables (second term)

0.3 Reconstruction Loss