# Re-derivation of Out of Distribution Paper

Arjun Gupta

August 23, 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 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. The NRM model uses the same weights and filters ($W^T(l)$) 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 generation step for NRM is best summarized as:

$$\begin{aligned} h(l-1) &= \sum p \in h(l)T(t(l,p))B(l)W^T(l)(s(l,p)h(l,p)) \\ x|z,y &= N(h(0),\sigma^2 I) \end{aligned} \tag{1}$$

Where the first equation is used to successively generate each layer of the NRM encoding from layer $N$ to layer 0, and the final output $x$ is the **last layer** (0) with gaussian noise added.

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:

$$\pi_{x|y} = Softmax\left(\frac{1}{\sigma^2}\sum_{l=1}^{L}\langle b(l), s(l) \odot t(l)\rangle\right) \tag{2}$$

Where $b(l)$ is the bias and also 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. Why is this the prior? I am having trouble understanding why the latent variables $z$ must come from a distribution considering they directly correspond to $s(l), t(l)$ derived from the CNN.