VRNN

Shashank Madhusudhan

August 2019

1 Introduction

This write up is based on a paper which claims that including latent random variables within the hidden state of a RNN will help deal with variations in structured sequential data like natural speech. It claims that a VRNN(variable RNN) can perform this task.

The authors suggest the use of Variational Auto Encoder which will map the latent random variable to observed outputs. The authors have chosen to work on **highly structured data**, meaning that which has high signal to noise ratio and a complex relationship between the factors causing the variation and the observed output.

2 Equations and models

$$h_t = f_{theta}(x_t, h_{t-1})$$

is the equation of a RNN to calculate hidden states. Here f is the non-linear transition function with parameter theta.

$$p(x_1, x_2...x_T) = \Pi_1^T p(x_t|x_{<t})$$

is the probability of a sequence in a RNN which is the product of the conditional probabilities. The author argues that the way these conditional probabilities are calculated in a standard RNN is a hindrance to find out variations.

$$p(x_t|x_{<t}) = g_x(h_{t-1})$$

is used to calculate the conditional probability. The author uses a **Gaussian Mixture Model** to calculate the conditional probability.

The claim made is that this is the first place where they have introduced a temporal dependency for the latent variables, meaning to find out the latent variables, previous timesteps are used.

2.0.1 Variational Auto Encoders

VAE is a generative model which can generate variations of an input. When we pass an input,a VAE passes it through a neural network to create two vectors, a mean vector and a standard deviation vector. These vectors are randomly sampled and this sampling is passed through a decoder to generate variations of the input because it is seeing different samples of the same input(stochastic generation).

We have input x and set of latent variables z which are meant to capture the variations in x. The joint distribution is defined as:

$$p(x, z) = p(x|z)p(z)$$

p(z) is chosen to be a Gaussian distribution which has information about the prior latent variables. The conditional probability is estimated by a neural network as mentioned in the above paragraph.

The **KL divergence** is used in the loss function so that all encoding inputs are as close to each other as possible also while being distinct to allow for smooth interpolation.