Are we making the assumption that, given $x_{\leq T}$, $z_{\leq T}$ and $x_{>T}$ are conditionally independent? Without this assumption, equation (10) in the paper should have been

$$q(z_{\leq T}|x_{\leq T}) = \prod_{t=1}^{T} q(z_t|x_{\leq T}, z_{<t})$$

instead of

$$q(z_{\leq T}|x_{\leq T}) = \prod_{t=1}^{T} q(z_t|x_{\leq t}, z_{<t})$$

And the objective in equation (11) wouldn't hold as well.

- A Recurrent Latent Variable Model for Sequential Data
- Junyoung Chung, Yoshua Bengio

Motivation

Traditional RNNs models $p(x_t|x_{<t})$ directly. This limits the form $p(x_t|x_{<t})$ can take. However, in practice, $p(x_t|x_{<t})$ maybe extremely multi-model and exhibit extreme variability. In this case, directly modeling $p(x_t|x_{<t})$ is insufficient.

Sequence modeling with RNN

RNN models the joint distribution $x = (x_1, x_2, \ldots, x_T)$ be modeling $p(x_t|x_{<t})$ recursively. A hidden state h_t is used to remember $x_1, \ldots, x_t$, and this is recursively defined as

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

Given this, we can define the conditional distribution $p(x_t|x_{<t})$ as

$$p(x_t|x_{<t}) = g_\tau(x_t, h_{t-1})$$

since h_{t-1} is a deterministic function of x_{t-1}, this makes sense.

The main representational power of an RNN comes from g_τ. This determines how complex the distribution can be. Typically, g_τ is defined in terms of a function that gives the parameter of a parametric distribution, like a mixture of gaussian, or multinomial distribution.

However, since we can only use a relatively simple g_τ, the model's modeling ability is significantly limited. When modelling sequences that are highly variable and highly structured, this is inadequate.

Variational Recurrent Neural Network

Preview. Instead of modelling $p(x_{\leq t})$, we will introduce a number of latent variables $p(z_{\leq t})$. And we assume the process of generating x_t given $z_{<t}$ and $x_{<t}$:

1. z_t is drawn from $p(z_t|x_{<t}, z_{<t})$
2. x_t is drawn from $p(x_t|z_{\leq t}, x_{<t})$

This is a typical VAE formulation. With this formulation, $p(x_t|x_{<t}, z_{<t})$ can be highly complex yet structured.

Generation The VRNN contains a VAE at every timestep. However, these VAEs are conditioned on the state variable h_{t-1} of an RNN. To define $p(x_t, z_t|x_{<t}, z_{<t})$, we will first define h_{t-1} to be deterministic function of $x_{<t}, z_{<t}$ as

$$h_t = f_\theta(\varphi_\tau^x(x), \varphi_\tau^x(z), h_{t-1})$$

Given this, we define

$$z_t \sim \mathcal{N}(\mu_{z,t}, diag(\sigma_{z,t}^2)) \quad [\mu_{z,t}, \sigma_{z,t}] = \varphi_\tau^{prior}(h_{t-1})$$

and

$$x_t|z_t \sim \mathcal{N}(\mu_{x,t}, diag(\sigma_{x,t}^2)) \quad [\mu_{x,t}, \sigma_{x,t}] = \varphi_\tau^{dec}(\varphi_\tau^z(z), h_{t-1})$$

Note the above two distributions are actually condition on $x_{<t}, z_{<t}$.

Given these, the join distribution $p(x_{\leq T}, z_{\leq T})$ is then given be

$$p(x_{\leq T}, z_{\leq T}) = \prod_{t=1}^{T} p(x_t|z_{\leq t}, x_{<t})p(z_t|z_{<t}, x_{<t})$$

Inference. Given $x_{<t}, z_{<t}$, we try to approximate z_t given x_t, namely $q(z_t|x_{\leq t}, z_{<t})$. We then define

$$z_t|x_t \sim \mathcal{N}(\mu_{z,t}, diag(\sigma_{z,t}^2)) \quad [\mu_{z,t}, \sigma_{z,t}] = \varphi_\tau^{dec}(\varphi_\tau^x(x), h_{t-1})$$

Given this, the approximate posterior over the whole sequence is then

$$q(z_{\leq T}|x_{\leq T}) = \prod_{t=1}^{T} q(z_t|x_{\leq T}, z_{<t})$$

It seems that we are assuming z_t and $x_{>t}$ are conditionally independent given $x_{\leq T}, z_{<t}$. So this is

$$q(z_{\leq T}|x_{\leq T}) = \prod_{t=1}^{T} q(z_t|x_{\leq t}, z_{<t})$$

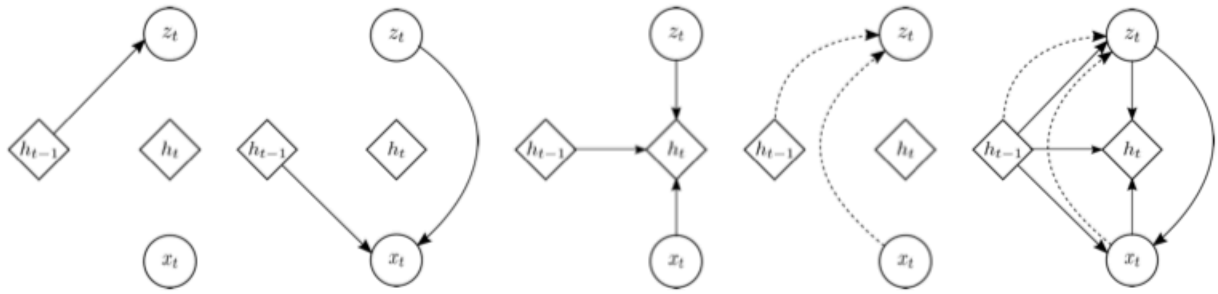
Learning. The training objective is given by

$$\mathbb{E}_{q(z_{\leq T}|x_{\leq T})}[\log \frac{p(z_{\leq T}, x_{\leq T})}{q(z_{\leq T}|x_{\leq T})}]$$

With the above three equations, and the assumption that $z_{\leq T}$ and $x_{>T}$ and conditionally independent given $x_{\leq T}$, we can derive the following objective:

$$\mathbb{E}_{q(\mathbf{z} \leq T|\mathbf{x} \leq T)}\left[\sum_{t=1}^{T}\left(-\mathrm{KL}\left(q\left(\mathbf{z}_t|\mathbf{x}_{\leq t}, \mathbf{z}_{<t}\right)\|p\left(\mathbf{z}_t|\mathbf{x}_{<t}, \mathbf{z}_{<t}\right)\right)+\log p\left(\mathbf{x}_t|\mathbf{z}_{\leq t}, \mathbf{x}_{<t}\right)\right)\right]$$

This is a good graph



(a) Prior (b) Generation (c) Recurrence (d) Inference (e) Overall

Figure 1: Graphical illustrations of each operation of the VRNN: (a) computing the conditional prior using Eq. (5); (b) generating function using Eq. (6); (c) updating the RNN hidden state using Eq. (7); (d) inference of the approximate posterior using Eq. (9); (e) overall computational paths of the VRNN.