Z-Forcing: Training Stochastic Recurrent Networks

A. Goyal, A. Sordoni, M. Côte,
N. R. Ke, Y. Bengio

Baek Gyuseung

February 2, 2018

Introduction

- Propose new stochastic recurrent model based by unifying successful ideas from recently proposed architectures.
- At training step, add an **auxiliary cost** that *forces* the latent variable to encode useful information.
- Perform better than alternative approaches.

Pipeline

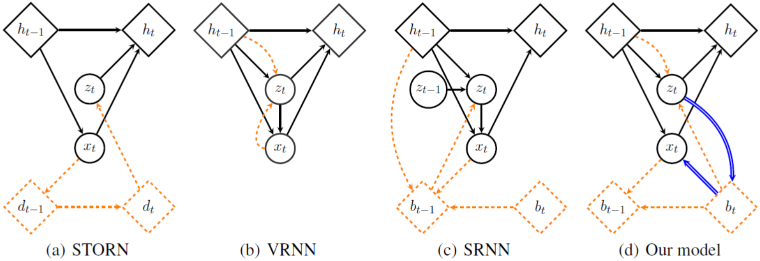

Figure: Diamonds and circles respectively represent deterministic and stochastic states. Dashed lines represent the computation that is part of the inference model. Double lines indicate auxiliary predictions implied by the proposed auxiliary cost.

Variational Inference

- Goal : Find MLE

$$\theta* = \underset{\theta}{\text{argmax}}\, \mathcal{L}(\theta; \mathcal{D}) = \sum_{i=1}^{n} \log \int_z p_\theta(x^i, z)dz$$

- By variational inference, we get

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z))$$

- For a sequential data $x = (x_1, \cdots, x_T)$ and latent variable $z = (z_1, \cdots, z_T)$, $p_\theta(x|z) = \prod_t p_\theta(x_t|z_{1:t-1}, x_{1:t-1})$

Model

- Decoder
  - $h_t = \overrightarrow{f}(x_t, h_{t-1}, z_t)$: LSTM model
  - $p_\theta(x_{t+1}|x_{1:t}, z_{1:t})$ is computed by $f^{(o)}(h_t)$.

- Prior
  - $p_\theta(z_t|x_{1:t}, z_{1:t-1}) = \mathcal{N}(z_t; \mu_t^{(p)}, \sigma_t^{(p)})$ where $[\mu_t^{(p)}, \log \sigma_t^{(p)}] = f^{(p)}(h_{t-1})$

- Inference Model
  - $b_t = \overleftarrow{f}(x_{t+1}, b_{t+1})$: LSTM model
  - $q_\phi(z_t|x) = \mathcal{N}(z_t; \mu_t^{(q)}, \sigma_t^{(q)})$ where $[\mu_t^{(q)}, \log \sigma_t^{(q)}] = f^{(q)}(h_{t-1}, b_t)$

Auxiliary Cost and Learning

- Auxiliary cost
  - $p_\xi(b_t|z_t) = \mathcal{N}(b_t; \mu_t^{(a)}, \sigma_t^{(a)})$ where $[\mu_t^{(a)}, \log \sigma_t^{(a)}] = f^{(a)}(z_t)$
- Loss function

$$\mathcal{L}(x) \geq \sum_t \mathbb{E}_{q_\phi(z_t|x)} [\log p_\theta(x_{t+1}|x_{1:t}, z_{1:t}) + \alpha \log p_\xi(b_t|z_t)]$$

$$- D_{KL}(q_\phi(z_t|x_{1:T})||p_\theta(z_t, x_{1:t}, z_{1:t-1})) + \beta \log p_\xi(x_t|b_t)$$

Experiments

Model	Blizzard	TIMIT
RNN-Gauss	3539	-1900
RNN-GMM	7413	26643
VRNN-I-Gauss	$\geq$ 8933	$\geq$ 28340
VRNN-Gauss	$\geq$ 9223	$\geq$ 28805
VRNN-GMM	$\geq$ 9392	$\geq$ 28982
SRNN (smooth+res$_q$)	$\geq$ 11991	$\geq$ 60550
Ours	$\geq$ 14435	$\geq$ 68132
Ours + kla	$\geq$ 14226	$\geq$ 68903
Ours + aux	$\geq$ **15430**	$\geq$ 69530
Ours + kla, aux	$\geq$ 15024	$\geq$ **70469**

Models	MNIST
DBN 2hl (Germain et al., 2015)	$\approx$ 84.55
NADE (Uria et al., 2016)	88.33
EoNADE-5 2hl (Raiko et al., 2014)	84.68
DLGM 8 (Salimans et al., 2014)	$\approx$ 85.51
DARN 1hl (Gregor et al., 2015)	$\approx$ 84.13
DRAW (Gregor et al., 2015)	$\leq$ 80.97
PixelVAE (Gulrajani et al., 2016)	$\approx$ **79.02**▼
P-Forcing$_{(3\text{-layer})}$ (Goyal et al., 2016)	79.58▼
PixelRNN$_{(1\text{-layer})}$ (Oord et al., 2016)	80.75
PixelRNN$_{(7\text{-layer})}$ (Oord et al., 2016)	79.20▼
MatNets (Bachman, 2016)	78.50▼
Ours$_{(1\text{ layer})}$	$\leq$ 80.60
Ours + aux$_{(1\text{ layer})}$	$\leq$ **80.09**