Learning to Generate with Memory

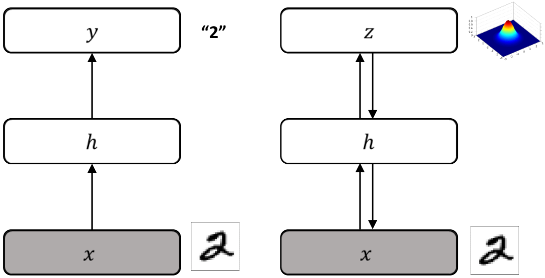
Chongxuan Li, Jun Zhu and Bo Zhang

Department of Computer Science and Technology
Tsinghua University

Jun 20, 2016
ICML at NYC

Deep Generative Models

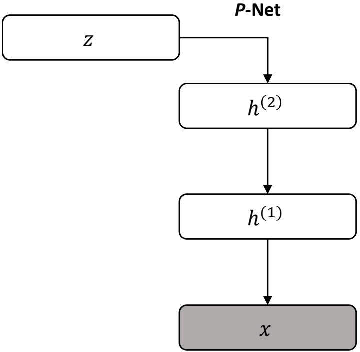
- learn abstract representations from unlabeled data
- handle the uncertainty in data



Feed-forward NN   DGM

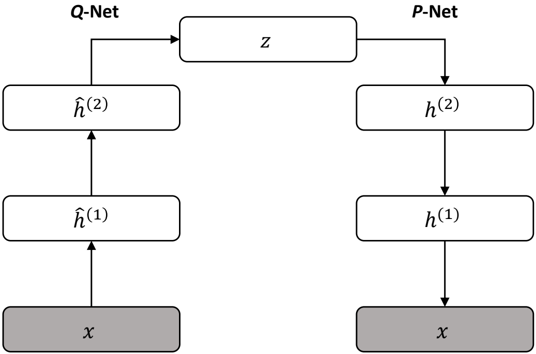
Directed DGMs

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



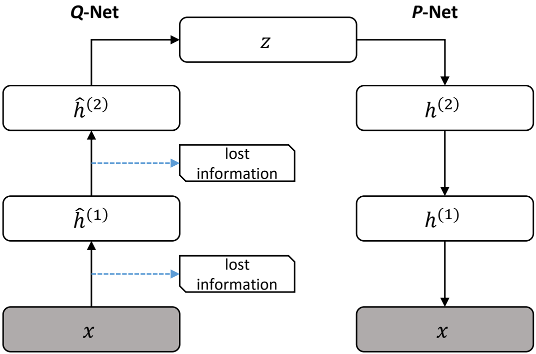
P-net and Q-net

$p(x, z) = p(z)p(x|z), q(z|x) \approx p(z|x)$ (Kingma and Welling [2014], Rezende et al. [2014])



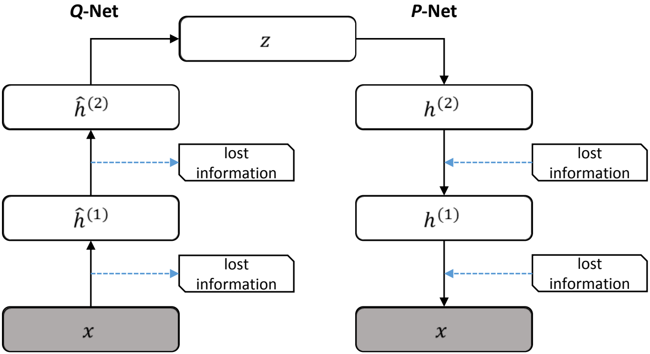
Competition between *P*-net and *Q*-net

Ignore lost information



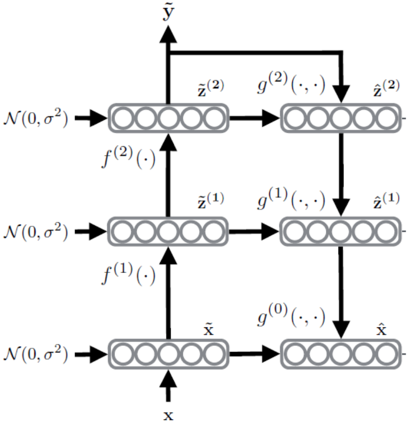
Ideal Case

Know lost information



The Ladder Network

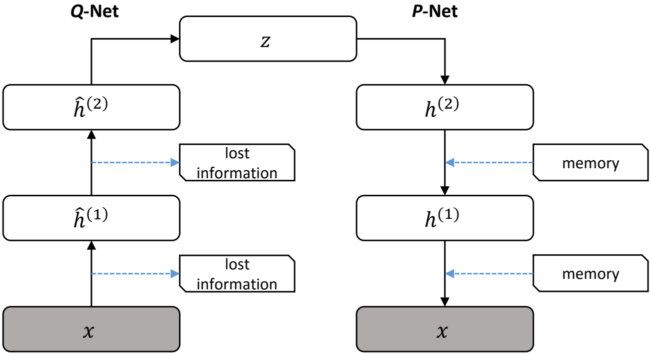
Lateral connections (Rasmus et al. [2015]), invalid in DGMs



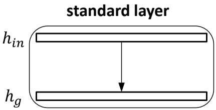
Tsinghua University

P-net with Memory

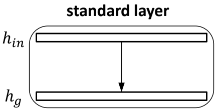
Encode and retrieve lost information

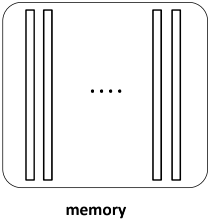


Building Block

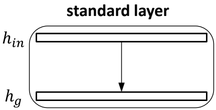


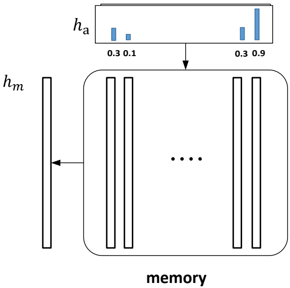
Building Block



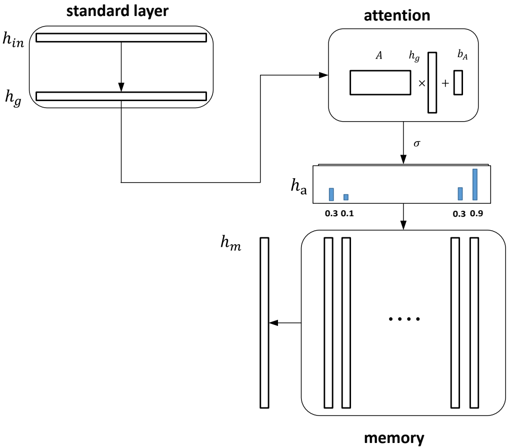


Building Block

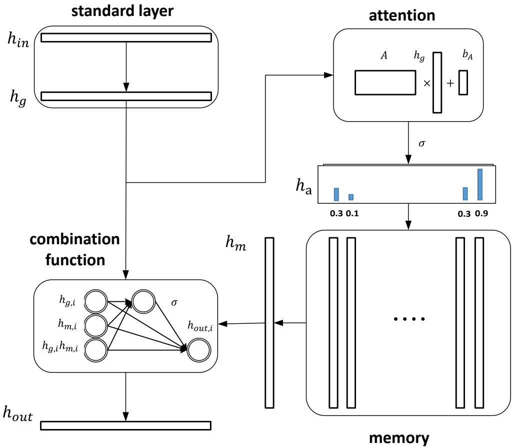




Building Block



Building Block



Mathematic Formulation

- Attention function:

$$h_a = \mathrm{sigmoid}(A^T h_g + b_A) \text{ or } h_a = \mathrm{softmax}(A^T h_g + b_A)$$

- Memory is parameterized as a matrix M,

$$h_m = M h_a$$

- Combination function:

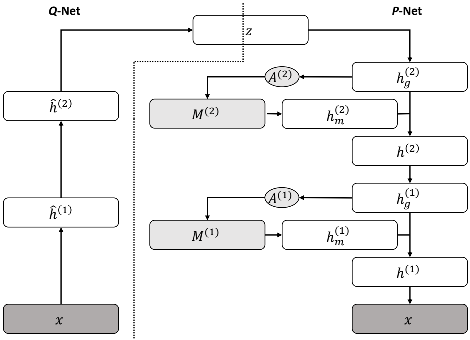
$$h_{out} = h_m + h_g \text{ or } h_{out} = a + b_1 c \text{ where}$$

$$a = a_1 + a_2 \odot h_m + a_3 \odot h_g + a_4 \odot h_g \odot h_m,$$

$$c = \sigma(c_1 + c_2 \odot h_m + c_3 \odot h_g + c_4 \odot h_g \odot h_m).$$

Tsinghua University

Asymmetric Architecture



$$\max_{\theta_g,\theta_r} \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbb{E}_{q(z|x;\theta_r)}[\log p(x, z; \theta_g) - \log q(z|x; \theta_r)]$$

Related Work

Memory units to capture long-term dependencies:

- Algorithm inference (Graves et al. [2014])
- Question answering (Weston et al. [2015], Sukhbaatar et al. [2015])
- Neural language transduction (Grefenstette et al. [2015])

Two main differences:

- Trained in unsupervised manner for generative tasks
- Our memory can't be written directly but updated via optimization

Settings

Baselines:

- VAE (Kingma and Welling [2014]) and IWAE (Burda et al. [2015])

Choice of components:

- MEM-VAE: sigmoid + element-wise MLP
- MEM-VAE-VIS: softmax + element-wise sum

Architecture used on MNIST:

- VAE: 530-530-100 hidden units, 1,550K parameters
- MEM-VAE: 500-500-100 hidden units + 70-30 memory slots, 1,559K parameters

Density Estimation

MODELS	MNIST	OCR-LETTERS
VAE	-85.69	-30.09
MEM-VAE(ours)	**-84.41**	**-29.09**

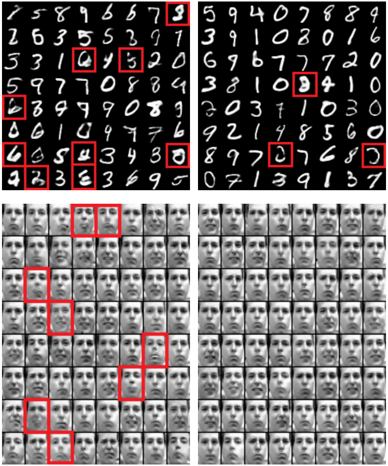
Density Estimation

MODELS	MNIST	OCR-LETTERS
VAE	-85.69	-30.09
MEM-VAE(ours)	**-84.41**	**-29.09**
IWAE-5	-84.43	-28.69
MEM-IWAE-5(ours)	**-83.26**	**-27.65**
IWAE-50	-83.58	-27.60
MEM-IWAE-50(ours)	**-82.84**	**-26.90**

Density Estimation

Models	MNIST	OCR-letters
VAE	-85.69	-30.09
MEM-VAE(ours)	-84.41	-29.09
IWAE-5	-84.43	-28.69
MEM-IWAE-5(ours)	-83.26	-27.65
IWAE-50	-83.58	-27.60
MEM-IWAE-50(ours)	**-82.84**	**-26.90**
DBN	-84.55	-
S2-IWAE-50	**-82.90**	-
*RWS-SBN/SBN**	-85.48	-29.99
*RWS-NADE/NADE**	-85.23	**-26.43**
*NADE**	-88.86	-27.22
*DARN**	**-84.13**	-28.17

Random Generation



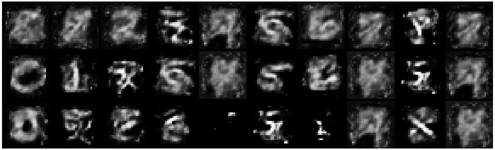
Missing Value Imputation

Visualization of Slots

Averaged preference of top-3 slots for each class:

"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"
0.27	0.82	0.33	0.11	0.34	0.15	0.49	0.27	0.09	0.28
0.24	0.09	0.06	0.11	0.30	0.13	0.12	0.27	0.09	0.21
0.18	0.05	0.06	0.11	0.07	0.07	0.05	0.11	0.09	0.18

Corresponding images:



Conclusion

Contribution:

- Introduce external memory mechanisms to DGMs
- Empirically test MEM-VAE on various tasks

Future work:

- Systematic investigation on other types of memory and attention
- Class-conditional models for better generation and classification
- Extensions in CNN case for high-dimensional image generation

Thank you!

References

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *arXiv:1509.00519*, 2015.

A. Graves, G. Wayne, and I. Danihelka. Neural Turing machines. In *arXiv:1410.5401*, 2014.

E. Grefenstette, K. M. Hermann, M. Suleyman, and P. Blunsom. Learning to transduce with unbounded memory. In *NIPS*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In *NIPS*, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *NIPS*, 2015.

J. Weston, S. Chopra, and A. Bordes. Memory networks. In *ICLR*, 2015.