Reading Group, presented by Minseop Park

One-shot Generalization in Deep Generative Model

Danilo J. Rezende, Shakir Mohamed, ICML 2016

Reference Papers
Auto-Encoding Variational Bayes (D.P. Kingma, M. Welling, ICLR 2014)
DRAW: A Recurrent Neural Network For Image Generation (K. Gregor et al, ICML 2015)
Spatial Transformer Networks (M. Jaderberg et al, NIPS 2015)

One-shot Generalization

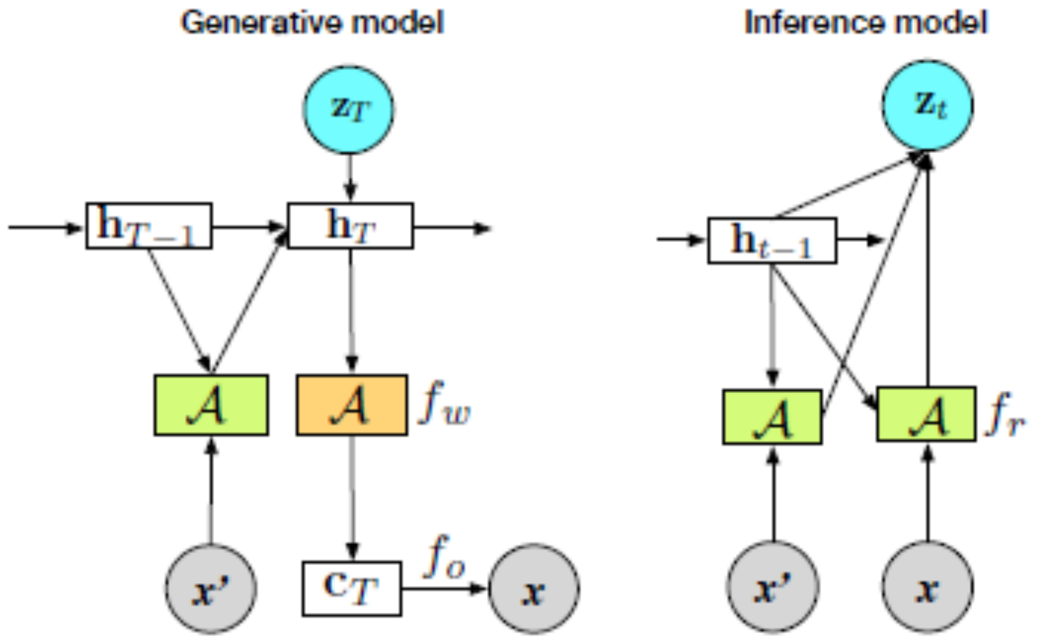
Task

Generation of novel variations of a giv
en exemplar

How?

By conditional, sequential generative
model

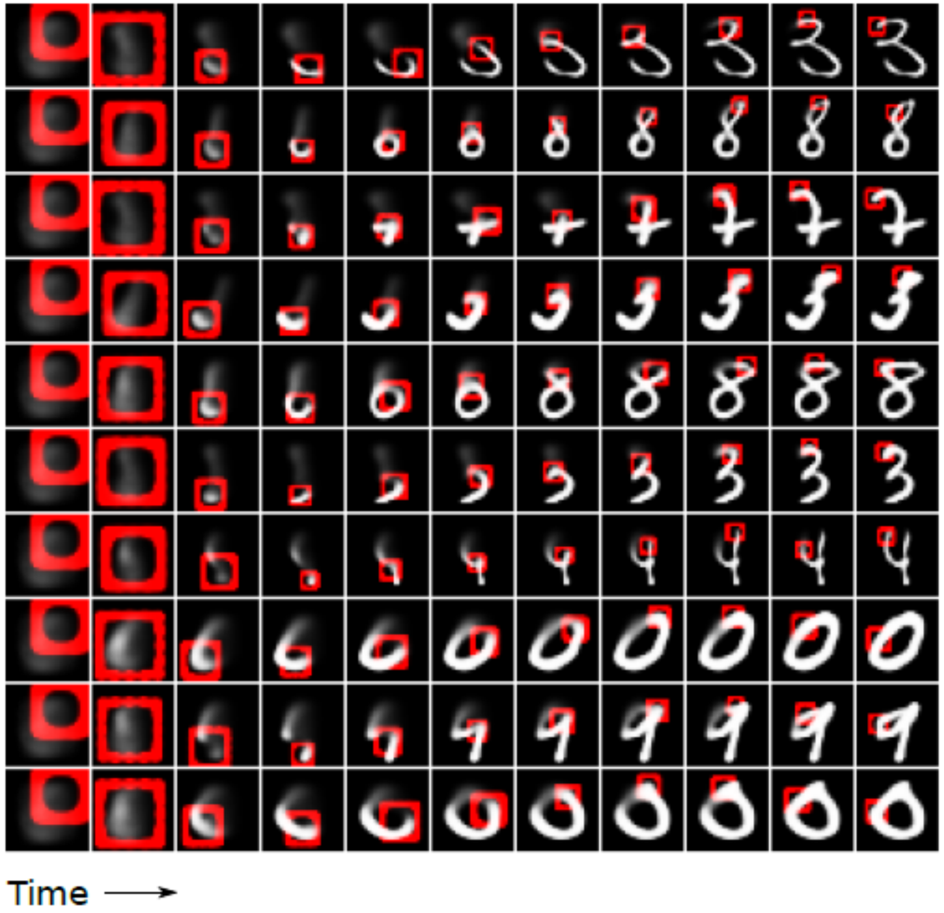
One-shot learning vs. One-shot gener
alization



(b) One-step of the conditional generative model.

DRAW : overview

- DRAW: Deep Recurrent Attentive Writer
- Basic model of sequential generative model
- Sequential VAE + attention
- Idea : Images like MNIST are generated sequ entially
- [https://www.youtube.com/watch?v=Zt-7MI9e KEo](https://www.youtube.com/watch?v=Zt-7MI9eKEo)



Variational Autoencoder

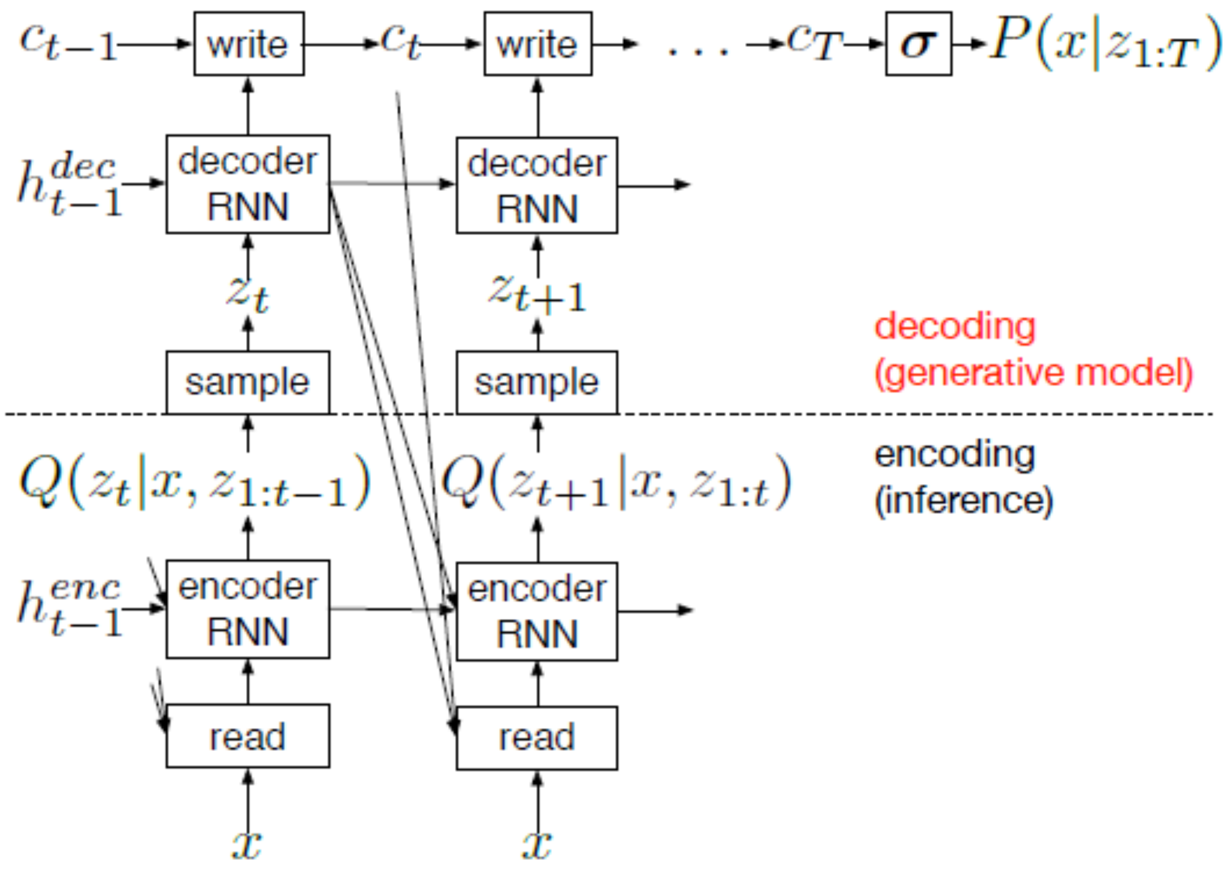
- Optimization of Variation Lower Bound

$$\mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right]$$

Encoder network $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$: $\xrightarrow{\mathbf{x}} \boxed{\text{FNN}(\Phi)} \longrightarrow$ Gaussian Param eters

Decoder network $p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})$: $\xrightarrow{\mathbf{z}} \boxed{\text{FNN}(\theta)} \longrightarrow$ Bernoulli (Gaussian) Parameter

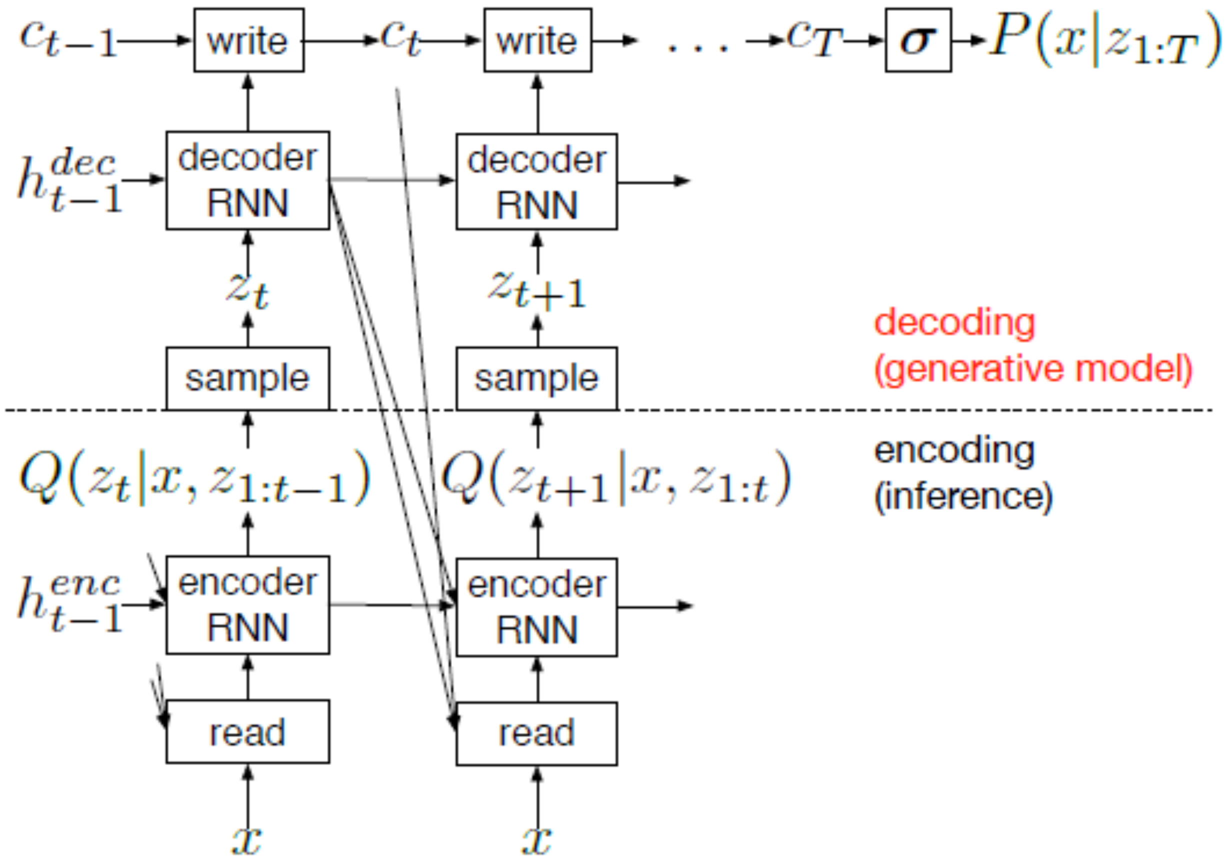
DRAW



Key Features

- Encoder / Decoder Network : LST M
- Additive Canvas
- Attention ; where to read, where/w hat to write

DRAW



$$\hat{x}_t = x - \boldsymbol{\sigma}(c_{t-1})$$

$$r_t = read(x_t, \hat{x}_t, h^{dec}_{t-1})$$

$$h^{enc}_t = RNN^{enc}(h^{enc}_{t-1}, [r_t, h^{dec}_{t-1}])$$

$$z_t \sim Q(Z_t|h^{enc}_t)$$

$$h^{dec}_t = RNN^{dec}(h^{dec}_{t-1}, z_t)$$

$$c_t = c_{t-1} + write(h^{dec}_t)$$

C↓t : canvas matrix
C↓T is used to parameterize $P(x|z)$
read/write : attention mechanism

DRAW

-**Approximate posterior** $Q(z_t \mid h_t^{enc}) = N(z_t \mid \mu_t, \sigma_t)$

where $\mu_t = W(h_t^{enc}), \quad \sigma_t^2 = \exp(W(h_t^{enc}))$

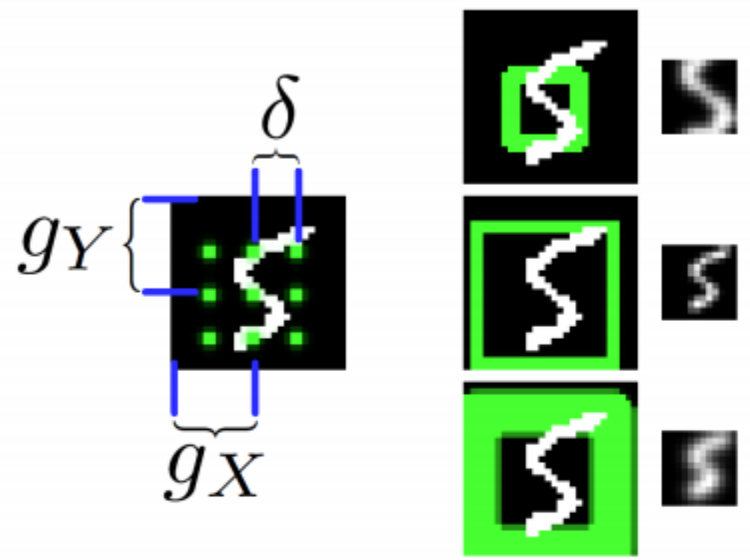
-**Data distribution** $P(x \mid z_{1:T}) = B(x \mid \boldsymbol{\sigma}(c_T))$

-**Data Generation**

$$
\begin{aligned}
\tilde{z}_t &\sim P(Z_t) \\
\tilde{h}_t^{dec} &= RNN^{dec}(\tilde{h}_{t-1}^{dec}, \tilde{z}_t) \\
\tilde{c}_t &= \tilde{c}_{t-1} + write(\tilde{h}_t^{dec}) \\
\tilde{x} &\sim D(X \mid \tilde{c_T})
\end{aligned}
$$

Selective Attention Model

- from the A x B input image, to obtain and N x N attention patch
- Horizontal and vertical filterbank F_X (N x A) and F_Y (N x B)



$$F_X[i,a] = \frac{1}{Z_X} \exp\left(-\frac{(a-\mu_X^i)^2}{2\sigma^2}\right)$$

$$F_Y[j,b] = \frac{1}{Z_Y} \exp\left(-\frac{(b-\mu_Y^j)^2}{2\sigma^2}\right)$$

$$\mu_X^i = g_X + (i - N/2 - 0.5)\delta$$

$$\mu_Y^j = g_Y + (j - N/2 - 0.5)\delta$$

Selective Attention Model

- Attention Parameters are obtained from LSTM output at each time step
- Initial patch covers the whole input image

$$(\tilde{g}_X, \tilde{g}_Y, \log \sigma^2, \log \tilde{\delta}, \log \gamma) = W(h^{dec})$$

$$g_X = \frac{A+1}{2}(\tilde{g}_X + 1)$$
$$g_Y = \frac{B+1}{2}(\tilde{g}_Y + 1)$$
$$\delta = \frac{\max(A, B) - 1}{N - 1}\tilde{\delta}$$

Read / Write Attention

- Read : from the A x B input image, to obtain and N x N attention patch

$$read(x, \hat{x}_t, h_{t-1}^{dec}) = \gamma[F_Y x F_X^T, F_Y \hat{x} F_X^T]$$

- Write : from the N x N attention patch, back to A x B input image

$$w_t = W(h_t^{dec})$$

$$write(h_t^{dec}) = \frac{1}{\hat{\gamma}} \hat{F}_Y^T w_t \hat{F}_X$$

DRAW : results

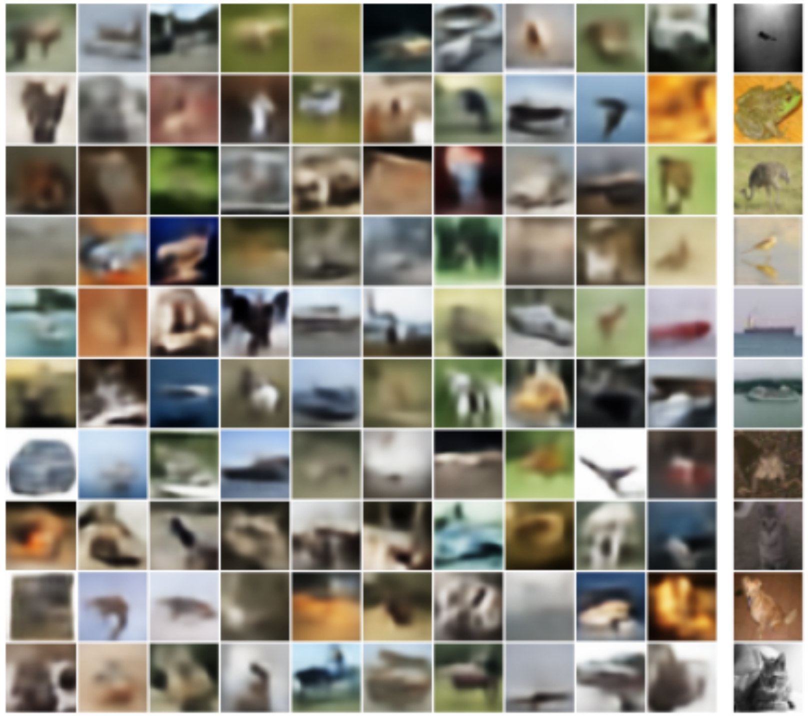


Figure 12. **Generated CIFAR images.** The rightmost column shows the nearest training examples to the column beside it.

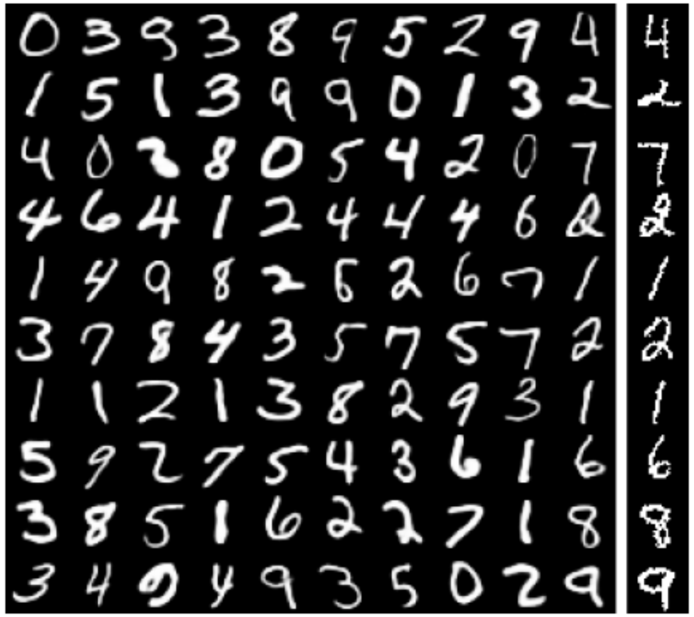
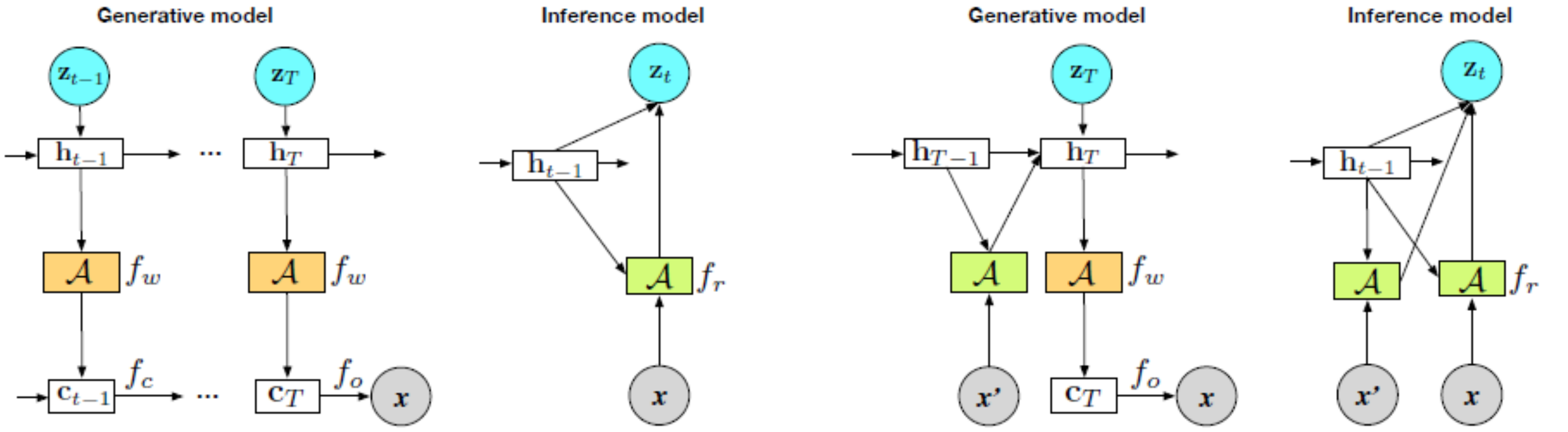


Figure 6. **Generated MNIST images.** All digits were generated by DRAW except those in the rightmost column, which shows the training set images closest to those in the column second to the right (pixelwise L^2 is the distance measure). Note that the network was trained on binary samples, while the generated images are mean probabilities.

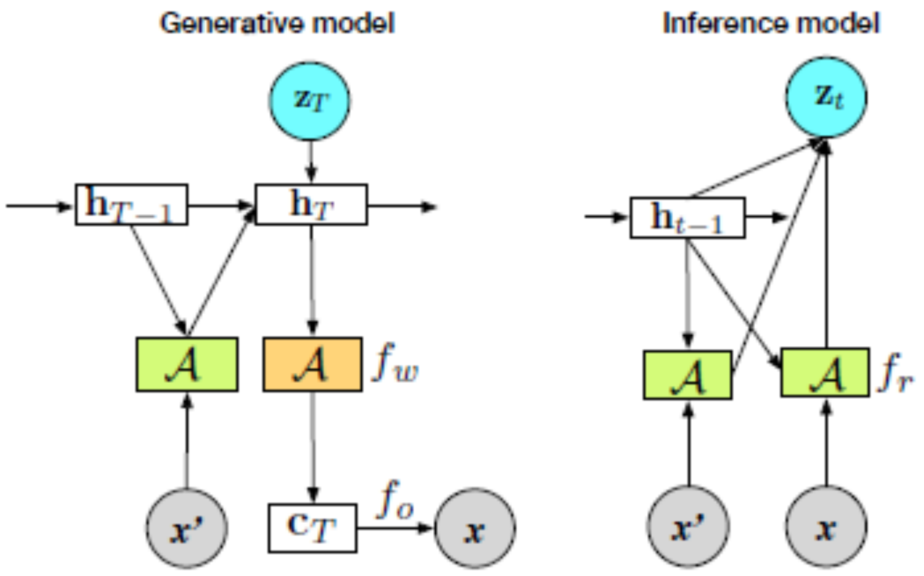
Sequential Generative Model

- Attention model : 2D Gaussian to Spatial Transformer
- Downsize the # of parameters by cutting the connection of canvas to hidden state
- Conditional generative model



(a) Unconditional generative model. (b) One-step of the conditional generative model.

Conditional Generative Model



(b) One-step of the conditional generative model.

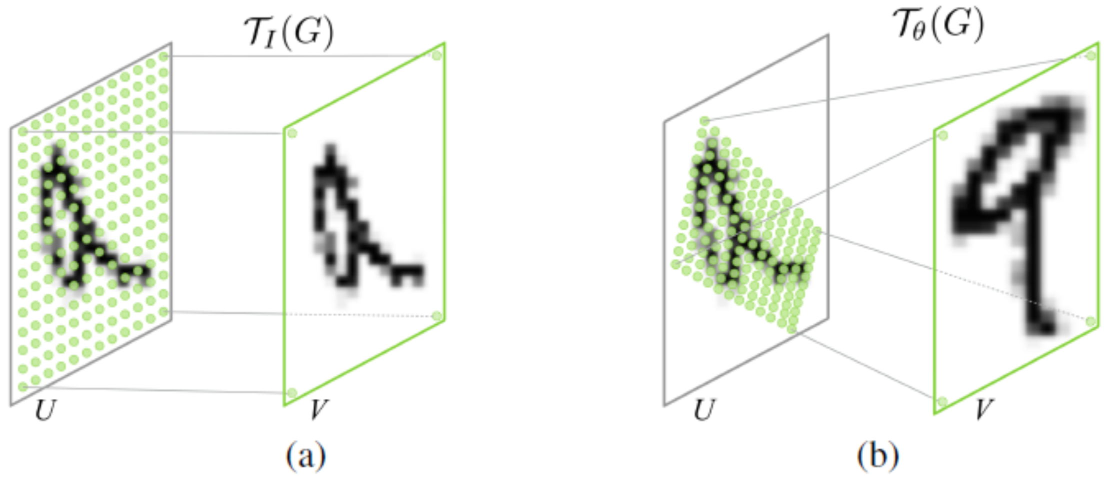
Latent variables	$\mathbf{z}_t \sim$	$\mathcal{N}(\mathbf{z}_t \vert \mathbf{0}, \mathbf{I}) \quad t = 1, \ldots, T$
Context	$\mathbf{v}_t =$	$f_v(\mathbf{h}_{t-1}, \mathbf{x}'; \theta_v)$
Hidden state	$\mathbf{h}_t =$	$f_h(\mathbf{h}_{t-1}, \mathbf{z}_t, \mathbf{v}_t; \theta_h)$
Hidden Canvas	$\mathbf{c}_t =$	$f_c(\mathbf{c}_{t-1}, \mathbf{h}_t; \theta_c)$
Observation	$\mathbf{x} \sim$	$p(\mathbf{x} \vert f_o(\mathbf{c}_T; \theta_o))$

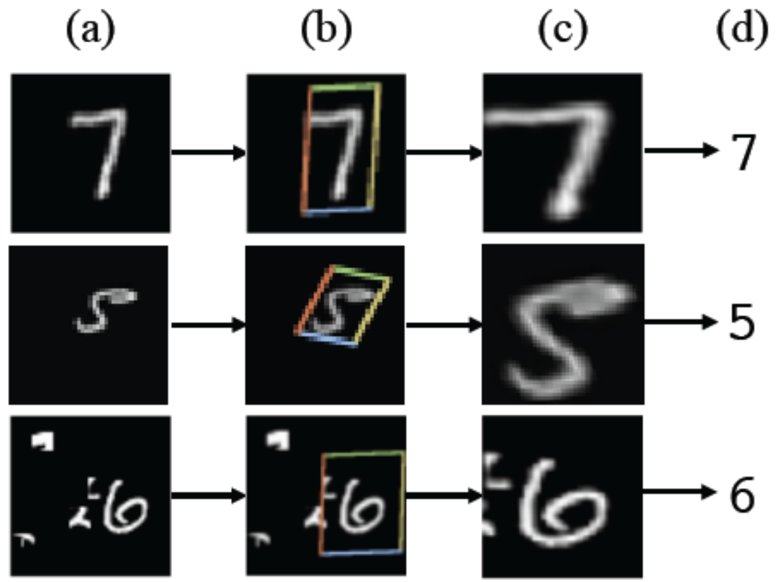
$$f_c(\mathbf{c}_{t-1}, \mathbf{h}_t; \theta_c) = \mathbf{c}_{t-1} + f_w(\mathbf{h}_t; \theta_c),$$

$f\downarrow h$: LSTM, state transitio
n
$f\downarrow c$: Additive Canvas
$f\downarrow v$: read attention
$f\downarrow w$: write attention (Spa
tial Transformer)

Spatial Transformer

- Differentiable attention model with affine transformation
- Learn parameters of affine transform





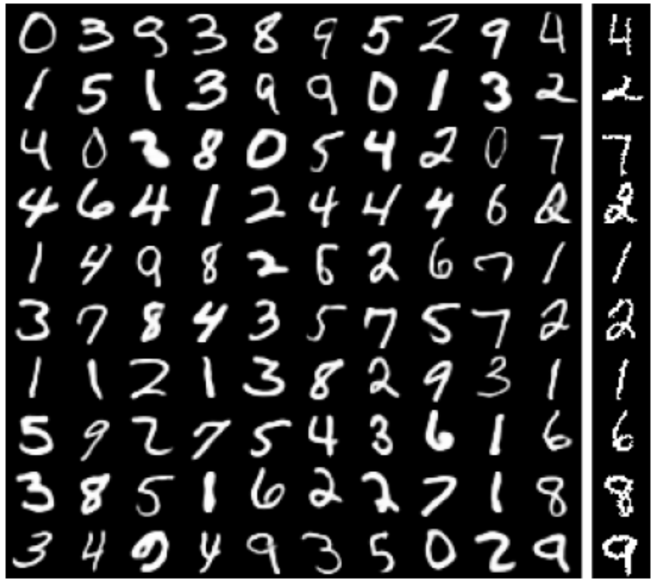
Result comparison

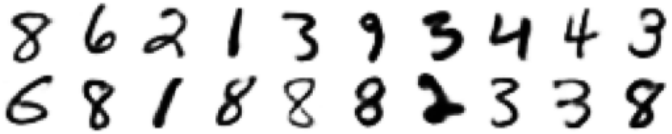
Table 1. Test set negative log-likelihood on MNIST.

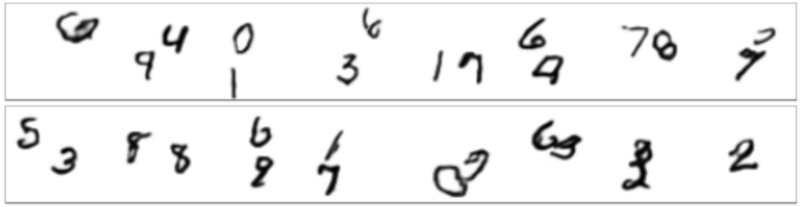
Model	Test NLL
From Gregor et al. (2015) and Burda et al. (20)	
DBM 2hl	≈84.62
DBN 2hl	≈84.55
NADE	88.33
DLGM-VAE	≈ 86.60
VAE + HVI/Norm Flow	≈ 85.10
DARN	≈ 84.13
DRAW (64 steps, no attention)	≤ 87.40
DRAW (64 steps, Gaussian attention)	≤ 80.97
IWAE (2 layers; 50 particles)	≈ 82.90

Sequential generative models

Attention	Canvas	Steps	Train	Test NLL
Spatial tr.	CGRU	80	78.5	≤**80.5(0.3)**
Spatial tr.	Additive	80	80.1	≤81.6(0.4)
Spatial tr.	CGRU	30	80.1	≤81.5(0.4)
Spatial tr.	Additive	30	79.1	≤82.6(0.5)
Fully conn.	CGRU	80	80.0	≤98.7(0.8)







One-shot generalization : 3 tasks

- Task 1 : Unconditional free generation
- Task 2 : Generation of novel variations of a given exemplar
- Task 3 : Generation of representative samples from a novel alphabet

Result for task 1

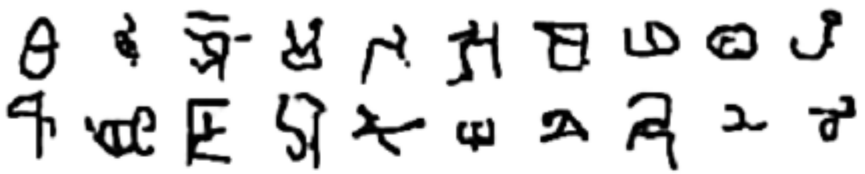


Figure 8. Unconditional samples for 52 $\times$ 52 omniglot (task 1). For a video of the generation process, see https://www.youtube.com/watch?v=HQEI2xfTgm4

One-shot generalization : results

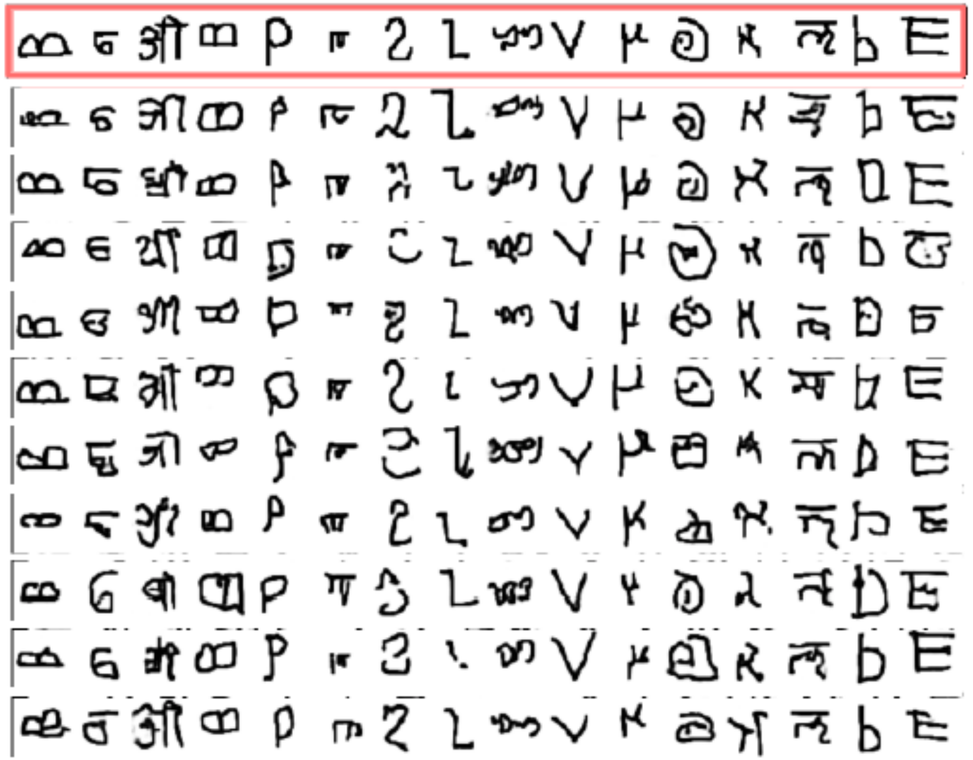


Figure 9. Generating new examplars of a given character for the weak generalization test (task 2*a*). The first row shows the test images and the next 10 are one-shot samples from the model.

Figure 11. Generating new exemplars from a novel alphabet (task 3). The first row shows the test images, and the next 10 rows are one-shot samples generated by the model.

Thank you