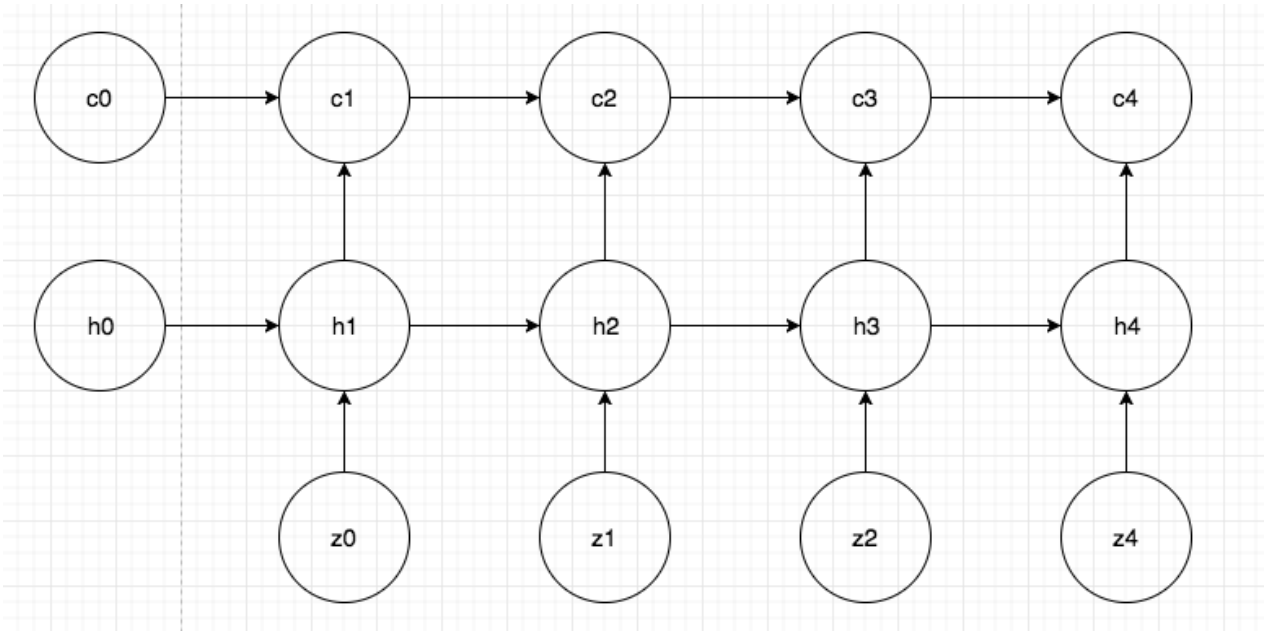
- DRAW: A Recurrent Neural Network For Image Generation
- Karol Gregor

Generation and Inference

Generation. Though implicitly specified, the paper tries to make asssumption on the data generation process:



Here, $z_1, \dots, z_T$ are sampled **independently** from some prior. After that,

- A delta w_1 is computed from z_1, and $c_1 = c_0 + w_1$
- A delta w_2 is computed from z_1, z_2, and $c_2 = c_1 + w_2$
- etc...

Here we are basically describing a way of generating an image from several latent variables, iteratively. **Intuitively, the ordering of z_i's shoud reflect the importance of each variable.**

Inference. Now given x, we need to infer $z_1, \dots, z_T$. Typically, z_1 should encode the most important information of x, so it is natural to infer that first. After that, given x, z_1, we infer z_2, etc.

From the above description, it is natural to design to seperate neural networks to perform generation and inference.

Generation.

For $i = 1, \dots, T$

1. $z_t \sim p(z)$
2. $h_t^{dec} = \text{RNN}^{dec}(h_{t-1}^{dec}, z_t)$
3. $c_t = c_{t-1} + write(h_t^{dec})$

And finally, $x \sim p(x|c_T)$.

Inference

The idea is to infer z_t given $x, z_{t-1}, \dots, z_1$. We will encode this using h_t^{dec}. Since h_{t-1}^{dec} encodes $z_1, \dots, z_{t-1}$, we can naturally do this model h_t^{enc} as a function of x and h_{t-1}^{dec}.

In theory that should be suffcient. But nonetheless the author do it the following way

$$\begin{aligned}
\hat{x}_t &= x - \sigma(c_{t-1}) \\
r_t &= read(x_t, \hat{x}_t, h_{t-1}^{dec}) \\
h_t^{enc} &= \text{RNN}^{enc}(h_{t-1}^{enc}, [r_t, h_{t-1}^{dec}]) \\
z_t &\sim q(z_t | h_t^{enc})
\end{aligned}$$

Where $\hat{x}_t$ is called the **error image**. This makes sense since givne $z_1, \dots, z_{t-1}$, c_{t-1} has been determined, so conditioned on x is equivalent to be conditioned on $\hat{x}$.

Loss. The loss is just the variational lower bound.

Read and Write

No attention. In the simplest case, no attention is used. It is then just

$$\begin{aligned}
read(x, \hat{x}, h_{t-1}^{dec}) &= [x, \hat{x}] \\
write(h_{t-1}^{dec}) &= W(h_t^{dec})
\end{aligned}$$

where W is some linear NN.

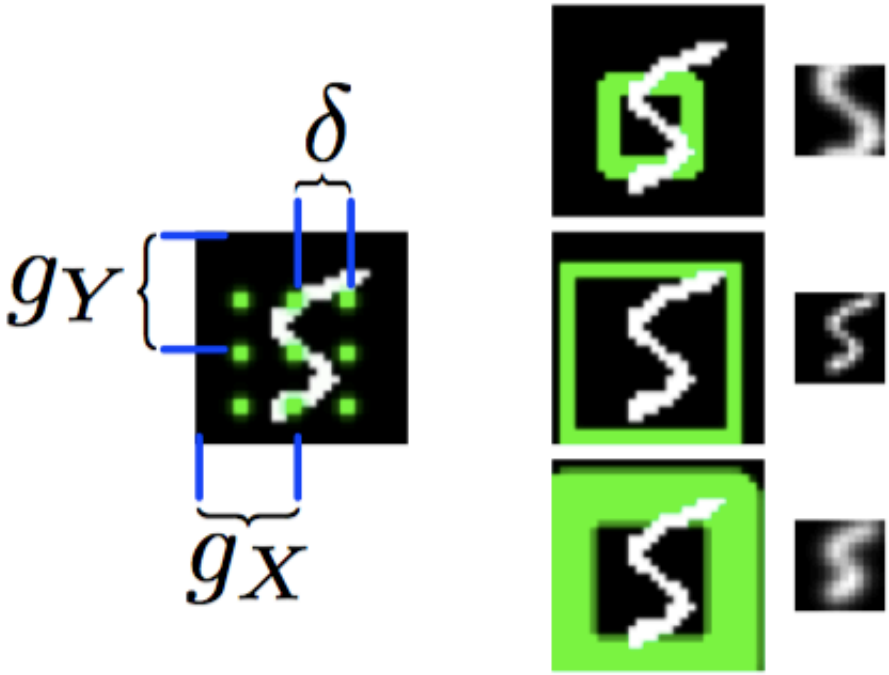


Figure 3. **Left:** A 3×3 grid of filters superimposed on an image. The stride (δ) and centre location (g_X, g_Y) are indicated. **Right:** Three $N \times N$ patches extracted from the image ($N = 12$). The green rectangles on the left indicate the boundary and precision (σ) of the patches, while the patches themselves are shown to the right. The top patch has a small δ and high σ, giving a zoomed-in but blurry view of the centre of the digit; the middle patch has large δ and low σ, effectively downsampling the whole image; and the bottom patch has high δ and σ.

Selective attention. The key is the form of attention. Here, we want the output of $read$ and $write$ to be fixed-sized. So it is natural to use a $N \times N$ grid to sample from the input, in $read$. Specifically, this grid is specified as

- g_X, g_Y: grid center
- δ: stride
- σ^2: filter size
- γ: response intensity.

And these will be obtained via a linear transformation of h^{dec}. I won't address the details of these process, but finally, the read operation can be written as

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

and the write operation will be a inverse process as

$$\begin{aligned} w_t &= W\left(h_t^{dec}\right) \\ \text{write}\left(h_t^{dec}\right) &= \frac{1}{\hat{\gamma}} \hat{F}_Y^T w_t \hat{F}_X \end{aligned}$$

Results

The most important thing to note is how the network with and without attention works:

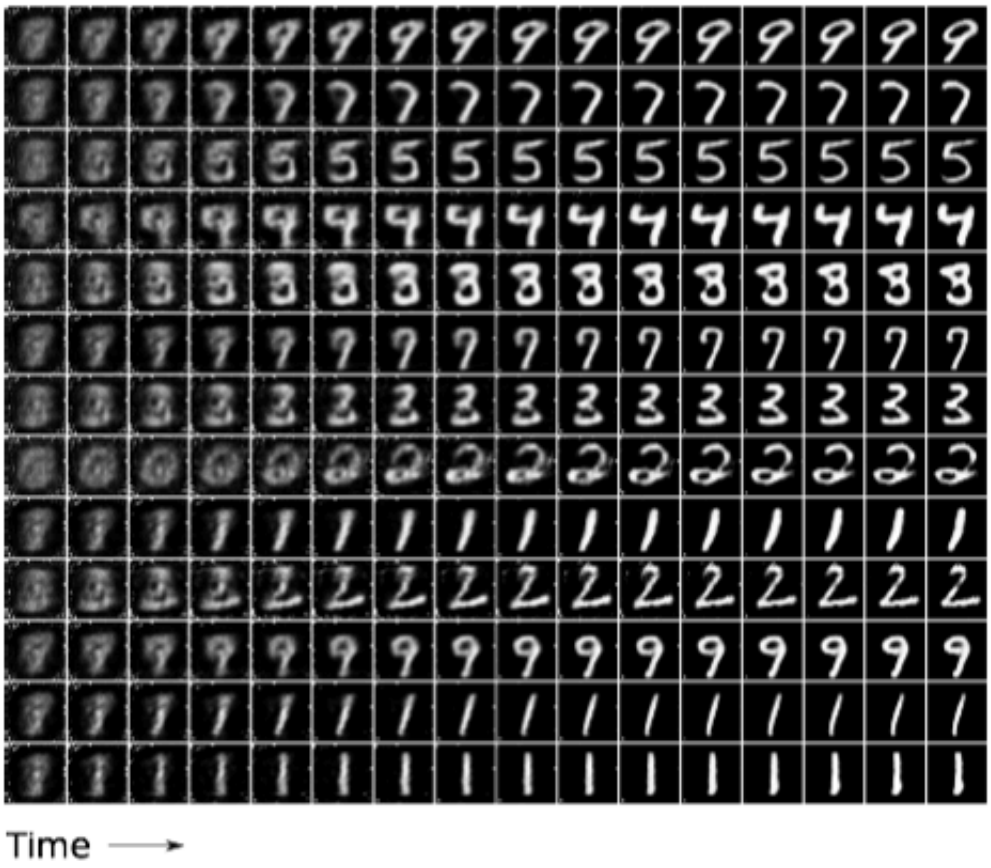


Figure 7. **MNIST generation sequences for DRAW without attention.** Notice how the network first generates a very blurry image that is subsequently refined.

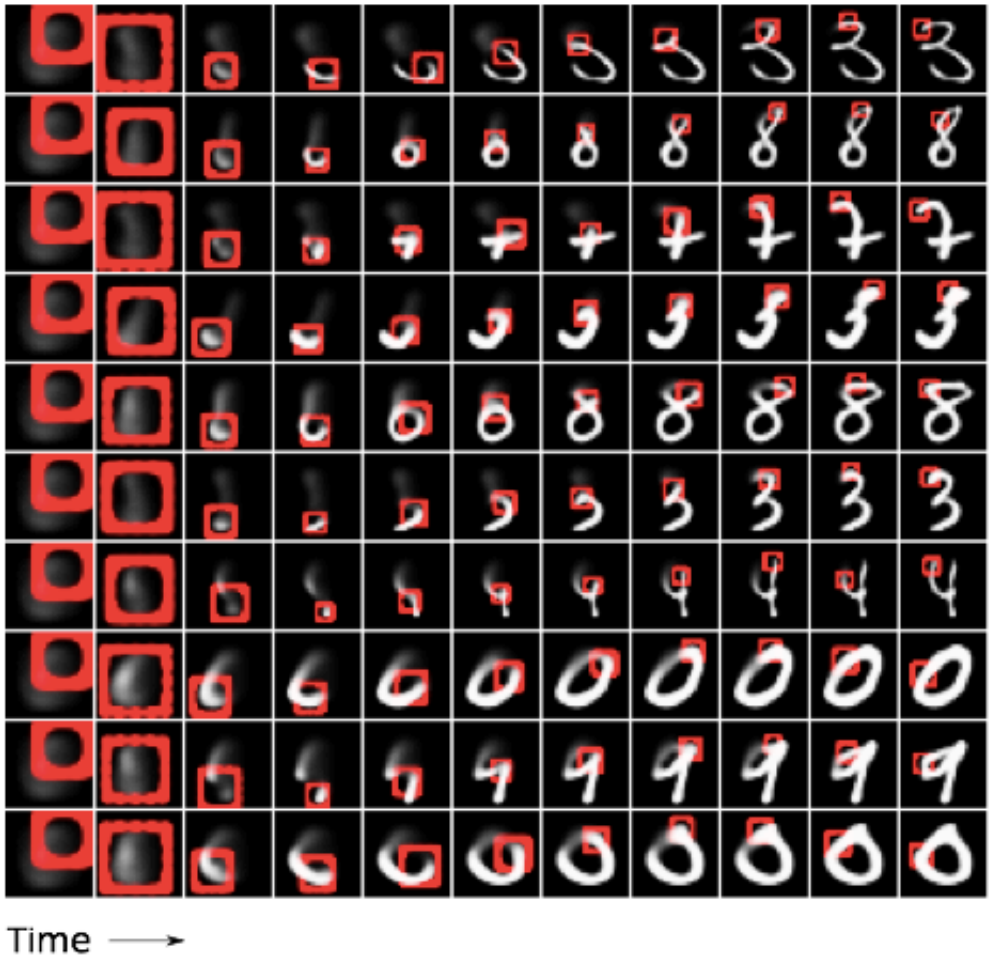


Figure 1. **A trained DRAW network generating MNIST digits.** Each row shows successive stages in the generation of a single digit. Note how the lines composing the digits appear to be "drawn" by the network. The red rectangle delimits the area attended to by the network at each time-step, with the focal precision indicated by the width of the rectangle border.

The author does not given an explanation of this.