Abstract

This paper introduces the Deep Recurrent Attentive Writer (DRAW) neural network architecture for image generation. DRAW networks combine a novel spatial attention mechanism that mimics the foveation of the human eye, with a sequential variational auto-encoding framework that allows for the iterative construction of complex images. The system substantially improves on the state of the art for generative models on MNIST, and, when trained on the Street View House Numbers dataset, it generates images that cannot be distinguished from real data with the naked eye.

1. Introduction

A person asked to draw, paint or otherwise recreate a visual scene will naturally do so in a sequential, iterative fashion, reassessing their handiwork after each modification. Rough outlines are gradually replaced by precise forms, lines are sharpened, darkened or erased, shapes are altered, and the final picture emerges. Most approaches to automatic image generation, however, aim to generate entire scenes at once. In the context of generative neural networks, this typically means that all the pixels are conditioned on a single latent distribution (Dayan et al., 1995; Hinton & Salakhutdinov, 2006; Larochelle & Murray, 2011). As well as precluding the possibility of iterative self-correction, the “one shot” approach is fundamentally difficult to scale to large images. The Deep Recurrent Attentive Writer (DRAW) architecture represents a shift towards a more natural form of image construction, in which parts of a scene are created independently from others, and approximate sketches are successively refined.

The core of the DRAW architecture is a pair of recurrent neural networks: an encoder network that compresses the real images presented during training, and a decoder that reconstitutes images after receiving codes. The combined system is trained end-to-end with stochastic gradient descent, where the loss function is a variational upper bound on the log-likelihood of the data. It therefore belongs to the family of variational auto-encoders, a recently emerged hybrid of deep learning and variational inference that has led to significant advances in generative modelling (Gregor et al., 2014; Kingma & Welling, 2014; Rezende et al., 2014; Mnih & Gregor, 2014; Salimans et al., 2014). Where DRAW differs from its siblings is that, rather than generat-
ing images in a single pass, it iteratively constructs scenes through an accumulation of modifications emitted by the decoder, each of which is observed by the encoder.

An obvious correlate of generating images step by step is the ability to selectively attend to parts of the scene while ignoring others. A wealth of results in the past few years suggest that visual structure can be better captured by a sequence of partial glimpses, or foveations, than by a single sweep through the entire image (Larochelle & Hinton, 2010; Denil et al., 2012; Tang et al., 2013; Ranzato, 2014; Zheng et al., 2014; Mnih et al., 2014; Ba et al., 2014; Sermanet et al., 2014). The main challenge faced by sequential attention models is learning where to look, which can be addressed with reinforcement learning techniques such as policy gradients (Mnih et al., 2014). The attention model in DRAW, however, is fully differentiable, making it possible to train with standard backpropagation. In this sense it resembles the selective read and write operations developed for the Neural Turing Machine (Graves et al., 2014).

The following section defines the DRAW architecture, along with the loss function used for training and the procedure for image generation. Section 3 presents the selective attention model and shows how it is applied to reading and modifying images. Section 4 provides experimental results on the MNIST, Street View House Numbers and CIFAR-10 datasets, with examples of generated images; and concluding remarks are given in Section 5. Lastly, we would like to direct the reader to the video accompanying this paper (https://www.youtube.com/watch?v=Zt-7Mi9eKEo) which contains examples of DRAW networks reading and generating images.

2. The DRAW Network

The basic structure of a DRAW network is similar to that of other variational auto-encoders: an encoder network determines a distribution over latent codes that capture salient information about the input data; a decoder network receives samples from the code distribution and uses them to condition its own distribution over images. However there are three key differences. Firstly, both the encoder and decoder are recurrent networks in DRAW, so that a sequence of code samples is exchanged between them; moreover the encoder is privy to the decoder’s previous outputs, allowing it to tailor the codes it sends according to the decoder’s behaviour so far. Secondly, the decoder’s outputs are successively added to the distribution that will ultimately generate the data, as opposed to emitting this distribution in a single step. And thirdly, a dynamically updated attention mechanism is used to restrict both the input region observed by the encoder, and the output region modified by the decoder. In simple terms, the network decides at each time-step “where to read” and “where to write” as well as “what to write”. The architecture is sketched in Fig. 2, alongside a feedforward variational auto-encoder.

2.1. Network Architecture

Let $RNN^{enc}$ be the function enacted by the encoder network at a single time-step. The output of $RNN^{enc}$ at time $t$ is the encoder hidden vector $h^{enc}_t$. Similarly the output of the decoder $RNN^{dec}$ at time $t$ is the hidden vector $h^{dec}_t$. In general the encoder and decoder may be implemented by any recurrent neural network. In our experiments we used the Long Short-Term Memory architecture (LSTM; Hochreiter & Schmidhuber (1997)) for both, in the extended form with forget gates (Gers et al., 2000). We favour LSTM due to its proven track record for handling long-range dependencies in real sequential data (Graves, 2013; Sutskever et al., 2014). Throughout the paper, we use the notation $b = W(a)$ to denote a linear weight matrix with bias from the vector $a$ to the vector $b$.

At each time-step $t$, the encoder receives input from both the image $x$ and from the previous decoder hidden vector $h^{dec}_{t-1}$. The precise form of the encoder input depends on a read operation, which will be defined in the next section. The output $h^{enc}_t$ of the encoder is used to parameterise a distribution $Q(Z_t| h^{enc}_t)$ over the latent vector $z_t$. In our
experiments the latent distribution is a diagonal Gaussian
\( \mathcal{N}(Z_t|\mu_t, \sigma_t) \):

\[
\mu_t = W(h_t^{enc}) \\
\sigma_t = \exp (W(h_t^{enc}))
\]  

(1) (2)

Bernoulli distributions are more common than Gaussians for latent variables in auto-encoders (Dayan et al., 1995; Gregor et al., 2014); however a great advantage of Gaussian latents is that the gradient of a function of the samples with respect to the distribution parameters can be easily obtained using the so-called reparameterization trick (Kingma & Welling, 2014; Rezende et al., 2014). This makes it straightforward to back-propagate unbiased, low variance stochastic gradients of the loss function through the latent distribution.

At each time-step a sample \( z_t \sim Q(Z_t|h_t^{enc}) \) drawn from the latent distribution is passed as input to the decoder. The output \( h_t^{dec} \) of the decoder is added (via a write operation, defined in the sequel) to a cumulative canvas matrix \( c_t \), which is ultimately used to reconstruct the image. The total number of time-steps \( T \) consumed by the network before performing the reconstruction is a free parameter that must be specified in advance.

For each image \( x \) presented to the network, \( c_0, h_0^{enc}, h_0^{dec} \) are initialised to learned biases, and the DRAW network iteratively computes the following equations for \( t = 1, \ldots, T \):

\[
\hat{x}_t = x - \sigma(c_{t-1}) \\
r_t = \text{read}(x_t, \hat{x}_t, h_t^{dec}) \\
h_t^{enc} = \text{RNN}^{enc}(h_{t-1}^{enc}, [r_t, h_t^{dec}]) \\
z_t \sim Q(Z_t|h_t^{enc}) \\
h_t^{dec} = \text{RNN}^{dec}(h_{t-1}^{dec}, z_t) \\
c_t = c_{t-1} + \text{write}(h_t^{dec})
\]

(3) (4) (5) (6) (7) (8)

where \( \hat{x}_t \) is the error image, \([v, w]\) is the concatenation of vectors \( v \) and \( w \) into a single vector, and \( \sigma \) denotes the logistic sigmoid function: \( \sigma(x) = \frac{1}{1+\exp(-x)} \). Note that \( h_t^{enc} \), and hence \( Q(Z_t|h_t^{enc}) \), depends on both \( x \) and the history \( z_{1:t-1} \) of previous latent samples. We will sometimes make this dependency explicit by writing \( Q(Z_t|x, z_{1:t-1}) \), as shown in Fig. 2. \( h_t^{enc} \) can also be passed as input to the read operation; however we did not find that this helped performance and therefore omitted it.

2.2. Loss Function

The final canvas matrix \( c_T \) is used to parameterise a model \( D(X|c_T) \) of the input data. If the input is binary, the natural choice for \( D \) is a Bernoulli distribution with means given by \( \sigma(c_T) \). The reconstruction loss \( \mathcal{L}^x \) is defined as the negative log probability of \( x \) under \( D \):

\[ \mathcal{L}^x = -\log D(x|c_T) \]  

(9)

The latent loss \( \mathcal{L}^z \) for a sequence of latent distributions \( Q(Z_t|h_t^{enc}) \) is defined as the summed Kullback-Leibler divergence of some latent prior \( P(Z_t) \) from \( Q(Z_t|h_t^{enc}) \):

\[ \mathcal{L}^z = \sum_{t=1}^{T} KL(Q(Z_t|h_t^{enc})||P(Z_t)) \]  

(10)

Note that this loss depends upon the latent samples \( z_t \) drawn from \( Q(Z_t|h_t^{enc}) \), which depend in turn on the input \( x \). If the latent distribution is a diagonal Gaussian with \( \mu_t, \sigma_t \) as defined in Eqs 1 and 2, a simple choice for \( P(Z_t) \) is a standard Gaussian with mean zero and standard deviation one, in which case Eq. 10 becomes

\[ \mathcal{L}^z = \frac{1}{2} \left( \sum_{t=1}^{T} \mu_t^2 + \sigma_t^2 - \log \sigma_t^2 \right) - T/2 \]  

(11)

The total loss \( \mathcal{L} \) for the network is the expectation of the sum of the reconstruction and latent losses:

\[ \mathcal{L} = \mathbb{E}[\mathcal{L}^x + \mathcal{L}^z] \]  

(12)

which we optimise using a single sample of \( z \) for each stochastic gradient descent step.

\( \mathcal{L}^z \) can be interpreted as the number of nats required to transmit the latent sample sequence \( z_{1:T} \) to the decoder from the prior, and (if \( x \) is discrete) \( \mathcal{L}^x \) is the number of nats required for the decoder to reconstruct \( x \) given \( z_{1:T} \). The total loss is therefore equivalent to the expected compression of the data by the decoder and prior.

2.3. Stochastic Data Generation

An image \( \hat{x} \) can be generated by a DRAW network by iteratively picking latent samples \( \tilde{z}_t \) from the prior \( P \), then running the decoder to update the canvas matrix \( \tilde{c}_t \). After \( T \) repetitions of this process the generated image is a sample from \( D(X|\tilde{c}_T) \):

\[
\tilde{z}_t \sim P(Z_t) \\
\tilde{h}_t^{dec} = \text{RNN}^{dec}(\tilde{h}_{t-1}^{dec}, \tilde{z}_t) \\
\tilde{c}_t = \tilde{c}_{t-1} + \text{write}(\tilde{h}_t^{dec}) \\
\hat{x} \sim D(X|\tilde{c}_T)
\]

(13) (14) (15) (16)

Note that the encoder is not involved in image generation.

3. Read and Write Operations

The DRAW network described in the previous section is not complete until the read and write operations in Eqs. 4 and 8 have been defined. This section describes two ways to do so, one with selective attention and one without.
3.1. Reading and Writing Without Attention

In the simplest instantiation of DRAW the entire input image is passed to the encoder at every time-step, and the decoder modifies the entire canvas matrix at every time-step. In this case the read and write operations reduce to

\[
\begin{align*}
\text{read}(x, \hat{x}_t, h_{i-1}^{\text{dec}}) &= [x, \hat{x}_t] \quad (17) \\
\text{write}(h_t^{\text{dec}}) &= W(h_t^{\text{dec}}) \quad (18)
\end{align*}
\]

However this approach does not allow the encoder to focus on only part of the input when creating the latent distribution; nor does it allow the decoder to modify only a part of the canvas vector. In other words it does not provide the network with an explicit selective attention mechanism, which we believe to be crucial to large scale image generation. We refer to the above configuration as “DRAW without attention”.

3.2. Selective Attention Model

To endow the network with selective attention without sacrificing the benefits of gradient descent training, we take inspiration from the differentiable attention mechanisms recently used in handwriting synthesis (Graves, 2013) and Neural Turing Machines (Graves et al., 2014). Unlike the aforementioned works, we consider an explicitly two-dimensional form of attention, where an array of 2D Gaussian filters is applied to the image, yielding an image ‘patch’ of smoothly varying location and zoom. This configuration, which we refer to simply as “DRAW”, somewhat resembles the affine transformations used in computer graphics-based autoencoders (Tieleman, 2014).

As illustrated in Fig. 3, the \( N \times N \) grid of Gaussian filters is positioned on the image by specifying the co-ordinates of the grid centre and the stride distance between adjacent filters. The stride controls the ‘zoom’ of the patch; that is, the larger the stride, the larger an area of the original image will be visible in the attention patch, but the lower the effective resolution of the patch will be. The grid centre \( (g_X, g_Y) \) and stride \( \delta \) (both of which are real-valued) determine the mean location \( \mu^X_i, \mu^Y_j \) of the filter at row \( i \), column \( j \) in the patch as follows:

\[
\begin{align*}
\mu^X_i &= g_X + (i - N/2 - 0.5) \delta \quad (19) \\
\mu^Y_j &= g_Y + (j - N/2 - 0.5) \delta \quad (20)
\end{align*}
\]

Two more parameters are required to fully specify the attention model: the isotropic variance \( \sigma^2 \) of the Gaussian filters, and a scalar intensity \( \gamma \) that multiplies the filter response. Given an \( A \times B \) input image \( x \), all five attention parameters are dynamically determined at each time step via a linear transformation of the decoder output \( h^{\text{dec}} \):

\[
\begin{align*}
(\hat{g}_X, \hat{g}_Y, \log \sigma^2, \log \delta, \log \gamma) &= W(h^{\text{dec}}) \quad (21) \\
g_X &= \frac{A + 1}{2} (\hat{g}_X + 1) \quad (22) \\
g_Y &= \frac{B + 1}{2} (\hat{g}_Y + 1) \quad (23) \\
\delta &= \frac{\max(A, B) - 1}{N - 1} \quad (24)
\end{align*}
\]

where the variance, stride and intensity are emitted in the log-scale to ensure positivity. The scaling of \( g_X, g_Y \) and \( \delta \) is chosen to ensure that the initial patch (with a randomly initialised network) roughly covers the whole input image.

Given the attention parameters emitted by the decoder, the horizontal and vertical filterbank matrices \( F_X \) and \( F_Y \) (dimensions \( N \times A \) and \( N \times B \) respectively) are defined as follows:

\[
\begin{align*}
F_X[i, a] &= \frac{1}{Z_X} \exp \left( -\frac{(a - \mu^X_i)^2}{2\sigma^2} \right) \quad (25) \\
F_Y[j, b] &= \frac{1}{Z_Y} \exp \left( -\frac{(b - \mu^Y_j)^2}{2\sigma^2} \right) \quad (26)
\end{align*}
\]

where \((i, j)\) is a point in the attention patch, \((a, b)\) is a point in the input image, and \(Z_X, Z_Y\) are normalisation constants that ensure that \( \sum_a F_X[i, a] = 1 \) and \( \sum_b F_Y[j, b] = 1 \).
We assess the ability of DRAW to generate realistic-looking images by training on three datasets of progressively increasing visual complexity: MNIST (LeCun et al., 1998), Street View House Numbers (SVHN) (Netzer et al., 2011) and CIFAR-10 (Krizhevsky, 2009). The images generated by the network are always novel (not simply copies of training examples), and are virtually indistinguishable from real data for MNIST and SVHN; the generated CIFAR images are somewhat blurry, but still contain recognisable structure from natural scenes. The binarized MNIST results substantially improve on the state of the art. As a preliminary exercise, we also evaluate the 2D attention module of the DRAW network on cluttered MNIST classification.

For all experiments, the model \( D(X|c_T) \) of the input data was a Bernoulli distribution with means given by \( \sigma(c_T) \). For the MNIST experiments, the reconstruction loss from Eq 9 was the usual binary cross-entropy term. For the SVHN and CIFAR-10 experiments, the red, green and blue pixel intensities were represented as numbers between 0 and 1, which were then interpreted as independent colour emission probabilities. The reconstruction loss was therefore the cross-entropy between the pixel intensities and the model probabilities. Although this approach worked well in practice, it means that the training loss did not correspond to the true compression cost of RGB images.

Network hyper-parameters for all the experiments are presented in Table 3. The Adam optimisation algorithm (Kingma & Ba, 2014) was used throughout. Examples of generation sequences for MNIST and SVHN are provided in the accompanying video (https://www.youtube.com/watch?v=Zt-7Mi9eKEo).

### 4.1. Cluttered MNIST Classification

To test the classification efficacy of the DRAW attention mechanism (as opposed to its ability to aid in image generation), we evaluate its performance on the 100 × 100 cluttered translated MNIST task (Mnih et al., 2014). Each image in cluttered MNIST contains many digit-like fragments of visual clutter that the network must distinguish from the true digit to be classified. As illustrated in Fig. 5, having an iterative attention model allows the network to progressively zoom in on the relevant region of the image, and ignore the clutter outside it.

Our model consists of an LSTM recurrent network that receives a 12 × 12 ‘glimpse’ from the input image at each time-step, using the selective read operation defined in Section 3.2. After a fixed number of glimpses the network uses a softmax layer to classify the MNIST digit. The network is similar to the recently introduced Recurrent Attention Model (RAM) (Mnih et al., 2014), except that our attention method is differentiable; we therefore refer to it as “Differentiable RAM”.

The results in Table 1 demonstrate a significant improvement in test error over the original RAM network. Moreover our model had only a single attention patch at each
time-step, whereas RAM used four, at different zooms.

4.2. MNIST Generation

We trained the full DRAW network as a generative model on the binarized MNIST dataset (Salakhutdinov & Murray, 2008). This dataset has been widely studied in the literature, allowing us to compare the numerical performance (measured in average nats per image on the test set) of DRAW with existing methods. Table 2 shows that DRAW without selective attention performs comparably to other recent generative models such as DARN, NADE and DBMs, and that DRAW with attention considerably improves on the state of the art.

Once the DRAW network was trained, we generated MNIST digits following the method in Section 2.3, examples of which are presented in Fig. 6. Fig. 7 illustrates the image generation sequence for a DRAW network without selective attention (see Section 3.1). It is interesting to compare this with the generation sequence for DRAW with attention, as depicted in Fig. 1. Whereas without attention it progressively sharpens a blurred image in a global way,
with attention it constructs the digit by tracing the lines—much like a person with a pen.

4.3. MNIST Generation with Two Digits

The main motivation for using an attention-based generative model is that large images can be built up iteratively, by adding to a small part of the image at a time. To test this capability in a controlled fashion, we trained DRAW to generate images with two $28 \times 28$ MNIST images chosen at random and placed at random locations in a $60 \times 60$ black background. In cases where the two digits overlap, the pixel intensities were added together at each point and clipped to be no greater than one. Examples of generated data are shown in Fig. 8. The network typically generates one digit and then the other, suggesting an ability to recreate composite scenes from simple pieces.

4.4. Street View House Number Generation

MNIST digits are very simplistic in terms of visual structure, and we were keen to see how well DRAW performed on natural images. Our first natural image generation experiment used the multi-digit Street View House Numbers dataset (Netzer et al., 2011). We used the same preprocessing as (Goodfellow et al., 2013), yielding a $64 \times 64$ house number image for each training example. The network was then trained using $54 \times 54$ patches extracted at random locations from the preprocessed images. The SVHN training set contains 231,053 images, and the validation set contains 4,701 images.

The house number images generated by the network are highly realistic, as shown in Figs. 9 and 10. Fig. 11 reveals that, despite the long training time, the DRAW network underfit the SVHN training data.

4.5. Generating CIFAR Images

The most challenging dataset we applied DRAW to was the CIFAR-10 collection of natural images (Krizhevsky,
Table 3. Experimental Hyper-Parameters.

<table>
<thead>
<tr>
<th>Task</th>
<th>glimpses</th>
<th>LSTM #h</th>
<th>#z</th>
<th>Read Size</th>
<th>Write Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 × 100 MNIST Classification</td>
<td>8</td>
<td>256</td>
<td>-</td>
<td>12 × 12</td>
<td>-</td>
</tr>
<tr>
<td>MNIST Model</td>
<td>64</td>
<td>256</td>
<td>100</td>
<td>2 × 2</td>
<td>5 × 5</td>
</tr>
<tr>
<td>SVHN Model</td>
<td>32</td>
<td>800</td>
<td>100</td>
<td>12 × 12</td>
<td>12 × 12</td>
</tr>
<tr>
<td>CIFAR Model</td>
<td>64</td>
<td>400</td>
<td>200</td>
<td>5 × 5</td>
<td>5 × 5</td>
</tr>
</tbody>
</table>

Figure 10. SVHN Generation Sequences. The red rectangle indicates the attention patch. Notice how the network draws the digits one at a time, and how it moves and scales the writing patch to produce numbers with different slopes and sizes.

Figure 11. Training and validation cost on SVHN. The validation cost is consistently lower because the validation set patches were extracted from the image centre (rather than from random locations, as in the training set). The network was never able to overfit on the training data.

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

2009). CIFAR-10 is very diverse, and with only 50,000 training examples it is very difficult to generate realistic-looking objects without overfitting (in other words, without copying from the training set). Nonetheless the images in Fig. 12 demonstrate that DRAW is able to capture much of the shape, colour and composition of real photographs.

5. Conclusion

This paper introduced the Deep Recurrent Attentive Writer (DRAW) neural network architecture, and demonstrated its ability to generate highly realistic natural images such as photographs of house numbers, as well as improving on the best known results for binarized MNIST generation. We also established that the two-dimensional differentiable attention mechanism embedded in DRAW is beneficial not only to image generation, but also to image classification.

Acknowledgments

Of the many who assisted in creating this paper, we are especially thankful to Koray Kavukcuoglu, Volodymyr Mnih, Jimmy Ba, Yaroslav Bulatov, Greg Wayne, Andrei Rusu and Shakir Mohamed.
References


Netzer, Yuval, Wang, Tao, Coates, Adam, Bissacco, Alessandro, Wu, Bo, and Ng, Andrew Y. Reading digits in natural images with unsupervised feature learning. 2011.


