VARIATIONAL NEURAL CELLULAR AUTOMATA

Anonymous authors
Paper under double-blind review

ABSTRACT

In nature, the process of cellular growth and differentiation has lead to an amazing diversity of organisms — algae, starfish, giant sequoia, tardigrades, and orcas are all created by the same generative process. Inspired by the incredible diversity of this biological generative process, we propose a generative model, the Variational Neural Cellular Automata (VNCA), which is loosely inspired by the biological processes of cellular growth and differentiation. Unlike previous related works, the VNCA is a proper probabilistic generative model, and we evaluate it according to best practices. We find that the VNCA learns to reconstruct samples well and that despite its relatively few parameters and simple local-only communication, the VNCA can learn to generate a large variety of output from information encoded in a common vector format. While there is a significant gap to the current state-of-the-art in terms of generative modeling performance, we show that the VNCA can learn a purely self-organizing generative process of data. Additionally, the self-organizing nature bestows the VNCA with some inherent robustness against perturbations in the early stages of growth.

1 INTRODUCTION

The process of cellular growth and differentiation is capable of creating an astonishing breadth of form and function, filling every conceivable niche in our ecosystem. Organisms range from complex multi-cellular organisms like trees, birds, and humans to tiny microorganisms living near hydrothermal vents in hot, toxic water at immense pressure. All these incredibly diverse organisms are the result of the same generative process of cellular growth and differentiation.

Cellular Automata (CA) are computational systems inspired by this process of cellular growth and differentiation, where "cells" iteratively update their state based on the state of their neighbor cells and the rules of the CA. Even with simple rules, CA can exhibit complex and surprising behavior — with just four simple rules, Conway's Game of Life is Turing complete. Neural Cellular Automata (NCA) are CA where the cell states are vectors, and the rules of the CA are parameterized and learned by neural networks (Mordvintsev et al., 2020; Nichele et al., 2017; Stanley & Miikkulainen, 2003). NCAs have been shown to learn to generate images, 3D structures, and even functional artifacts that are capable of regenerating when damaged (Mordvintsev et al., 2020; Sudhakaran et al., 2021). While these results are impressive, the NCAs can only generate (and regenerate) the single artifact it is trained on, lacking the diverse generative properties of current probabilistic generative models. In this paper, we introduce the VNCA, an NCA based architecture that addresses these limitation while only relying on local communication and self-organization.

Generative modeling is a fundamental task in machine learning, and several methods have been proposed. Probabilistic generative models are an attractive class of models that directly define a distribution over the data $p(x)$, which enable sampling and computing (at least approximately) likelihoods of observed data. The average likelihood on test data can then be compared between models, which enables fair comparison between generative models. The Variational Auto-Encoder (VAE) is a seminal probabilistic generative model, which models the data using a latent variable model, such that $p(x) = \int_z p_\theta(x|z)p(z)$, where $p(z)$ is a fixed prior and $p_\theta(x|z)$ is a learned decoder (Kingma & Welling, 2013). The parameters of the model is learned by maximizing the Evidence Lower Bound (ELBO), a lower bound on $\log p(x)$, which can be efficiently computed using amortized variational inference (Kingma & Welling, 2013).

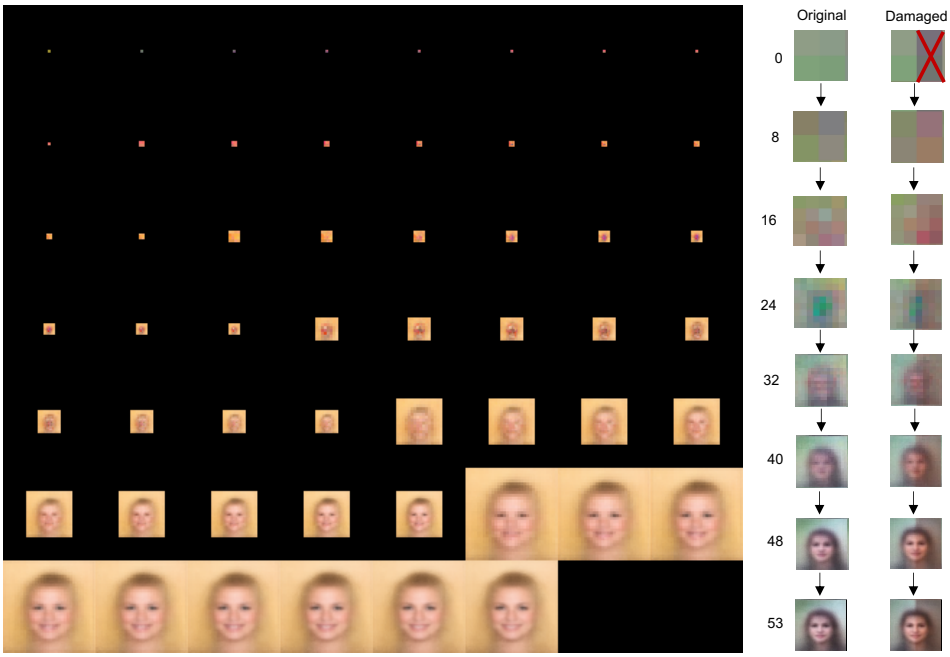

Figure 1: **VNCA self-organized growth and damage recovery.** Left: The VNCA has learned a self-organising generative process that generates faces, from an initial random seed of cells from $\mathcal{N}(0, I)$. Time goes top-down and left-to-right. Every eight steps all the cells duplicate, which can be seen on the diagonal. Right: A growth sequence of selected steps is shown with and without setting some of the cell states to zero. The VNCA shows resilience to damage early in growth.

On a high level, our proposed generative model combines NCAs with VAEs, by using a NCA as the decoder in a VAE. We propose a novel NCA architecture, which duplicates all cells every N steps, inspired by cell mitosis. This has the computational benefit that the VNCA only computes cell updates for the living cells at each step, which are relatively few early in growth. While we found it to not be important for our VNCA experiments, this growth-by-doubling does prevent cells from dying; we can imagine that in other application this might may be an important property. While our model is inspired by the biological process of cellular growth and differentiation, it is not intended to be an accurate model of this process and is naturally constrained by what we can efficiently compute.

Other works have explored similar directions. The VAE-NCA proposed in Chen & Wang (2020), is actually, despite the name, not a VAE since it does not define a generative model or use variational inference. Rather it is an auto-encoder that decodes a set of weights, which parameterize a NCA that reconstructs images. NCAM (Ruiz et al., 2020) similarly uses an auto-encoder to decode weights for a NCA, which then generates images. In contrast to both of these, our method *is* a generative model, uses a single learned NCA, and samples the initial state of the seed cells from a prior $p(z)$. StampCA (Frans, 2021) also learns an auto-encoder but similarly encodes the initial hidden NCA state of a central cell. None of these are generative models, in the sense that it's not possible to sample data from them, nor do they assign likelihoods to samples. While previous works all show impressive reconstructions, it is impossible to evaluate how good generative models they really are. In contrast, we aim to offer a fair and honest evaluation of the VNCA as a generative model.

Graph Neural Networks (GNN) can be seen as a generalization of the local-only communication in NCA, where the 2D grid neighborhood is relaxed into an arbitrary graph neighborhood. GNNs compute vector-valued messages along the edges of the graph that are accumulated at the nodes, using a shared function among all nodes and edges. GNNs have been successfully used to model molecular properties (Gilmer et al., 2017), reasoning about objects and their interactions (Santoro et al., 2017), and learning to control self-assembling morphologies (Pathak et al., 2019).

Powerful generative image models have been proposed based on the variational auto-encoder framework. A common distinction is whether the decoder is auto-regressive or not, i.e., whether the de-

coder conditions its outputs on previous outputs, usually those above and to the left of the current pixel (Van Oord et al., 2016; Salimans et al., 2017). Auto-regressive decoders are so effective at modelling $p(x)$ directly that they have tendency to ignore the latent codes thus failing to reconstruct images well (Alemi et al., 2018). Additionally, they are expensive to sample from since sampling must be done one pixel at a time. Our decoder is a not auto-regressive, so is cheaper to sample from, and is more similar to e.g., BIVA (Maaløe et al., 2019) or NVAE (Vahdat & Kautz, 2020). Contrary to BIVA, NVAE and other similar deep convolution based VAEs, our VNCA defines a relatively simple decoder function that only consider the immediate neighborhood of a cell (pixel) at each step and is applied iteratively over a number of steps. The VNCA thus aims to learns a self-organising generative process.

NCAs and self-organizing systems in general, represent an exciting research direction. The repeated application of a simple function that only relies on local communication lends itself well to a distributed leaderless computation paradigm, useful in, for instance, swarm robotics (Rubenstein et al., 2012). Additionally, such self-organizing systems are often inherently robust to perturbations (Mordvintsev et al., 2020; Najarro & Risi, 2020; Tang & Ha, 2021; Sudhakaran et al., 2021). In the VNCA, we show that this robustness allows us to set some of the cells' states to zero during the growth process and – without being trained to do so – the system is able to partially recover (Fig. 1,right). By introducing an NCA that is a proper generative probabilistic model, we hope to further spur the development of methods at the intersection of artificial life and generative models.

2 Variational Neural Cellular Automata

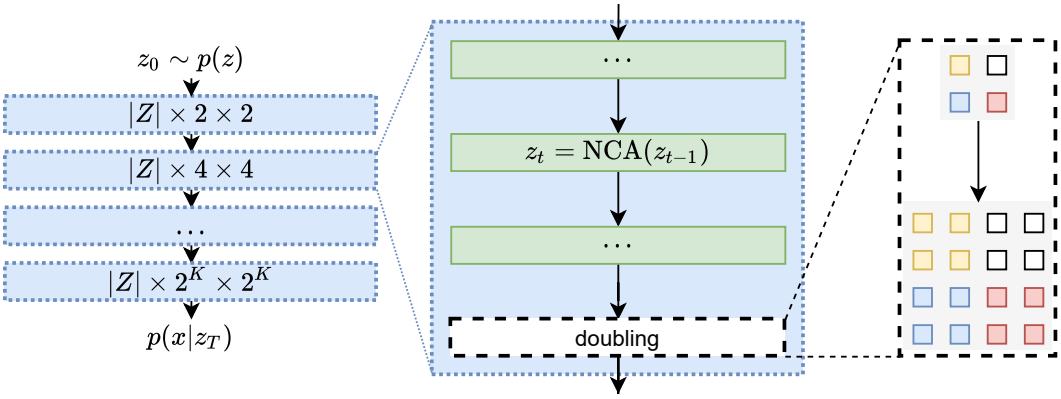

Figure 2: VNCA overview. Left: Generative process. z_0 is sampled from $p(z)$ and acts as an initial 2×2 seed of cells, which grows through a series of K doubling NCA steps. $|Z|$ denotes the size of the latent space, which is identical to the size of the cell states. Finally, the last cell hidden state conditions the parameters of the likelihood distribution $p(x|z)$. Middle: Each doubling NCA step consists of a number of NCA steps followed by a doubling operation. At each NCA step, the cells can only communicate with their immediate neighbors. Right: The doubling operator doubles the cell grid as depicted, where the color indicates the state vector of each cell.

The VNCA defines a generative latent variable model,

$$z \sim p(z)\,,$$
$$x \sim p_\theta(x|z)\,,$$

where $p(z) = \mathcal{N}(z|\mu = 0, \Sigma = I)$ is a Gaussian with zero mean and diagonal co-variance and $p_\theta(x|z)$ is a learned decoder, with parameters θ. The model is trained by maximizing the evidence lower bound (ELBO),

$$\log p(x) - D_{\mathrm{KL}}(q_\phi(z|x) \parallel p(z|x)) = -D_{\mathrm{KL}}(q_\phi(z|x) \parallel p(z)) + \mathbb{E}_{z\sim q_\phi(z|x)} \log p_\theta(x|z)\,,$$

where, $q_\phi(z|x)$ is a learned encoder, with parameters ϕ, which learns to perform amortized variational inference. Since $D_{\mathrm{KL}} \geq 0$, the right hand side is a lower bound on $\log p(x)$ and increasing it

either increases $\log p(x)$ or decreases $D_{\mathrm{KL}}(q_\phi(z|x) \parallel p(z|x))$, which is how closely the amortized variational inference matches the true unknown posterior $p(z|x)$. By using the reparameterization trick, the sampling, $z \sim q_\phi(z|x)$, is differentiable, and both ϕ and θ can be learned efficiently with stochastic gradient ascent (Kingma & Welling, 2013). The D_{KL} term on the right hand side can be weighted with a parameter β to control whether the VAE should prioritize better reconstructions ($\beta < 1$), or better samples ($\beta > 1$) (Higgins et al., 2016). For $\beta > 1$ this is still a (looser) lower bound on $\log p(x)$.

The decoder of the generative process, $p_\theta(x|z)$, is based on a NCA (Fig. 2). The NCA defines a recurrent computation over vector-valued "cells" in a grid. At step t of the NCA, the cells compute an additive update to their state, based only on the state of their neighbor cells in a 3×3 neighborhood,

$$z_{i,t} = z_{i,t-1} + u_\theta(\{z_{j,t-1}\}_{j \in N(i)}),$$

where u_θ is a learned update function, index i denote a single cell and $N(i)$ denotes the indexes of the 3×3 neighborhood of cell i, including i. All cells are updated in parallel. In practice, u_θ is efficiently implemented using a Convolutional Neural Network (CNN) (Gilpin, 2019) with a single 3×3 filter followed by a number of 1×1 filters and non-linearities. The initial grid z_0 consists of a single sample from the latent space repeated in a 2×2 grid.

A traditional NCA defines whether a cell is alive or dead based on the cell state and its neighbors' cell states (Mordvintsev et al., 2020). This allows the NCA to "grow" from a small initial seed of alive cells. Our NCA uses a different approach inspired by cell mitosis. Every M steps of regular NCA updates all the cells duplicate. The cells duplicate by creating new cells initialized to their own current state to the right and below themselves, "pushing" the existing cells out of the way as they do so. This is efficiently implemented using a "nearest" type image rescaling operation. See the right panel of fig. 2 for an illustration.

After K doubling and $M(K+1)$ steps of NCA updates, the final cell states z_T condition the parameters of the likelihood distribution. For RGB images, we use a discrete logistic likelihood with L mixtures (Salimans et al., 2017), such that the $10L$ first channels are the parameters of the discrete logistic mixture[1]. For binary images, we use a Bernoulli likelihood with the first channel of each cell being $\log p$.

The update function is a single 3×3 convolution, followed by four residual blocks (He et al., 2016). Each block consists of a 1×1 convolution, Exponential Linear Unit (ELU) non-linearity (Clevert et al., 2015) and another 1×1 convolution. Finally, a single 1×1 convolution maps to the cell state size. This last convolution layers' weights and biases are initialized to zero to prevent the recurrent computation from overflowing. The number of channels for all convolutional layers, the size of the latent space, and the cell state are all 256. The total number of parameters in the NCA decoder is approximately 1.2M.

Since we are primarily interested in examining the properties of the NCA decoder, the encoder $q_\phi(z|x)$ is a fairly standard deep convolutional network. It consists of a single layer of 5×5 convolution with 32 channels followed by four layers of 5×5 convolution with stride 2, where each of these layers has twice the number of channels as the previous layer. Finally, the down-scaled image is flattened and passed through a dense layer which maps to twice the latent space dimensionality. The first half encodes the mean and the second half the log-variance of a Gaussian. All convolutional layers are followed by ELU non-linearities. For a detailed model definition, see the appendix.

3 EXPERIMENTS

For all experiments we use a batch size of 32, Adam optimizer (Kingma & Ba, 2014), 10^{-4} learning rate, $L = 1$ logistic mixture component, clip the gradient norm to 10 (Pascanu et al., 2013) and train the VNCA for 100.000 gradient updates. We use a single sample to compute the ELBO when training and measure final log-likelihoods on the test set using 128 importance weighted samples, which gives a tighter lower bound than the ELBO (Burda et al., 2015).

[1] We use the implementation from `https://github.com/vlievin/biva-pytorch`

3.1 MNIST

For our first experiment we chose MNIST, since it is widely used in the generative modeling literature and a relatively easy dataset. Since the VNCA generates images that have height and width that are powers of 2, we pad the MNIST images to 32×32 with zeros and use $K = 4$ doublings. We use the statically binarized MNIST from Larochelle & Murray (2011).

Figure 3: MNIST results. Left: Test set reconstructions. Right: Unconditional samples from the prior. The VNCA achieves $\log p(x) \geq -84.23$ evaluated with 128 importance weighted samples.

The results show that the VNCA learns a generative model of MNIST digits and also is capable of accurately reconstructing digits given a latent code by the encoder (Fig. 3). It achieves $\log p(x) \geq -84.23$ evaluated with 128 importance weighted samples. While not state-of-the-art, which is around -78.6 (Maaløe et al., 2019), it is a decent generative model. In comparison, the complex auto-regressive PixelCNN decoder achieves -81.3 (Van Oord et al., 2016).

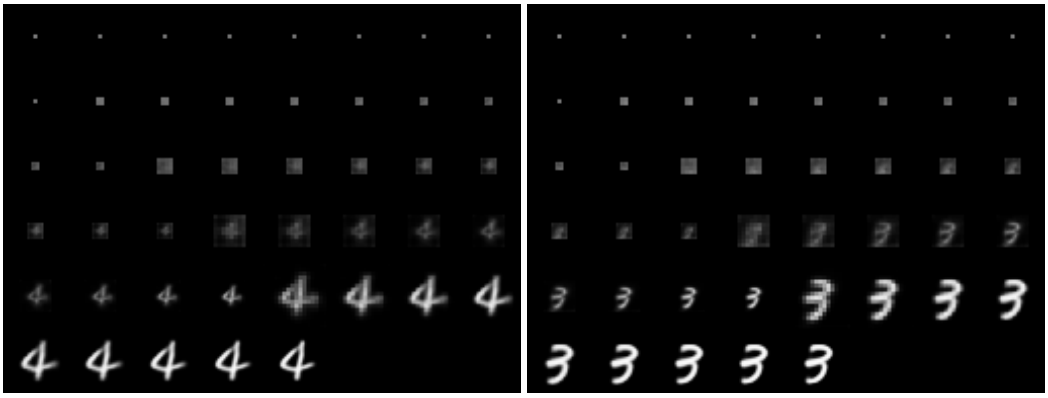

Figure 4: Self-organising MNIST digits, with the seed cell states sampled from the prior. Each 32×32 sub-image shows the digit after 1 step of NCA or doubling. Time flows top-to-bottom and left-to-right. The doubling steps can be seen on the diagonal. Note: this shows the likelihood probabilities instead of samples from the likelihood for clarity.

Fig. 4 shows the self-organization and growth of the digits. Initially, they are just uniform squares, but after doubling twice to a size of 8×8, it appears that the process is iterating towards a pattern. After doubling, the NCA steps shows signs of "enhancing" the image, moving towards a smoother, more natural-looking digit. The VNCA has learned a self-organizing process that converges to diverse naturally looking hand-written digits.

Since the VNCA defines a recurrent computation, we can run it for more steps than it was trained for, thus potentially increasing the resolution of the image. Fig. 5 shows the growth of an MNIST digit for two more doublings. The final step the VNCA was trained for is highlighted with a gray square. After the first doubling to 64×64, the NCA again sharpens the image for a few steps, but after some time it degrades and starts producing artifacts. Interestingly, the VNCA appears to have some limited super-resolution capabilities up to twice the original resolution, if carefully used.

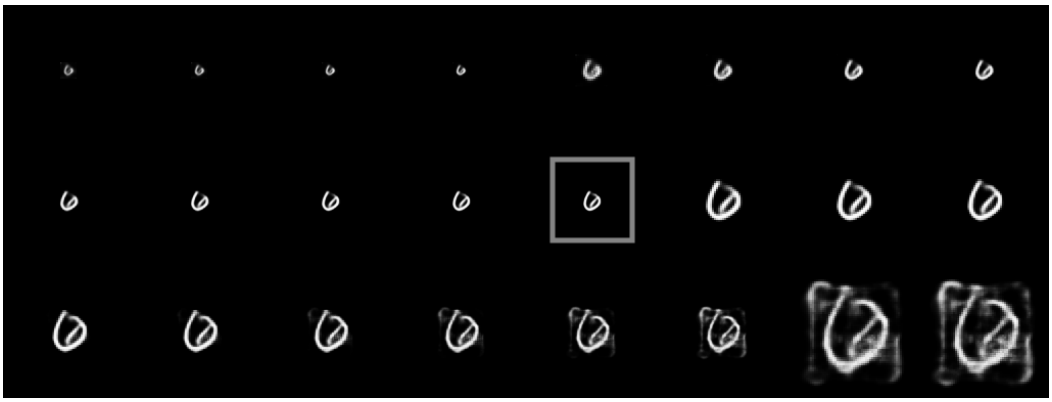

Figure 5: The growth of an MNIST digit for two extra doublings to a size of 128×128. The first 32 steps are not shown. The 32×32 digit the network was trained to generate is highlighted with a gray square. The VNCA shows some capabilities in producing super-resolution images but produces artefacts and deteriorate, if run for too long.

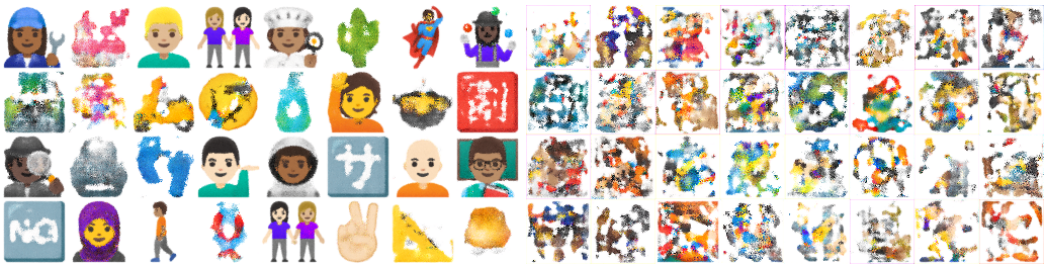

Figure 6: Noto Emoji results. Left: Test set reconstructions. Right: Unconditional samples from the prior. The VNCA achieves 3.09 bits per dimension on the test set evaluated with 128 importance weighted samples.

3.2 NOTO EMOJI

The Noto Emoji font contains 2656 vector graphic emojis[2]. This dataset allows for comparison with previous NCA auto-encoder work, which uses it (Frans, 2021; Chen & Wang, 2020; Ruiz et al., 2020; Mordvintsev et al., 2020). Also, the emoji are diverse, which tests the ability of the VNCA to generate a large variety of outputs encoded in a single latent vector format. We train a VNCA on 64×64 emoji examples using $K = 5$ doublings and using 20% (531) of the images as a test set.

The VNCA learns to reconstruct the inputs well — similarly to the auto-encoder NCAs (Frans, 2021; Chen & Wang, 2020; Ruiz et al., 2020) — as can be seen on the left but fails to generate realistic-looking samples, as can be seen on the right (Fig. 6). It does particularly well on the faces and human emojis, probably because they make up a large part of the dataset and are the most regular. While it reconstructs most images well, it fails on some, in particular the signs and symbols that are unique, not resembling any other emojis in the dataset, e.g., a highway, a safety pin, and a microscope. While the VNCA here fails to learn a good generative model, we are impressed with its ability to generate such diverse outputs from a single common vector encoding, using a simple self-organising process.

3.3 CELEBA

Finally, we train a VNCA on the CelebA dataset, which contains 202,599 images of celebrity faces (Liu et al., 2015). CelebA has been extensively used in the generative modeling literature, which allows us to fairly compare the VNCA against state-of-the-art generative methods.

[2]https://github.com/googlefonts/noto-emoji

Figure 7: CelebA results. Left: Test set reconstructions. Right: Unconditional samples from the prior. The VNCA achieves 4.42 bits per dimension on the test set evaluated with 128 importance weighted samples.

The VNCA learns to reconstruct the images well, but the unconditional samples do not look like faces, although there are some faint face-like structures in some of the samples (Fig. 7). The VNCA achieves 4.42 bits per dimension evaluated with 128 importance weighted samples. This is far from state-of-the-art which is around 2 bits per dimension, using diffusion based models and deep hierarchical VAEs (Kim et al., 2021; Vahdat & Kautz, 2020).

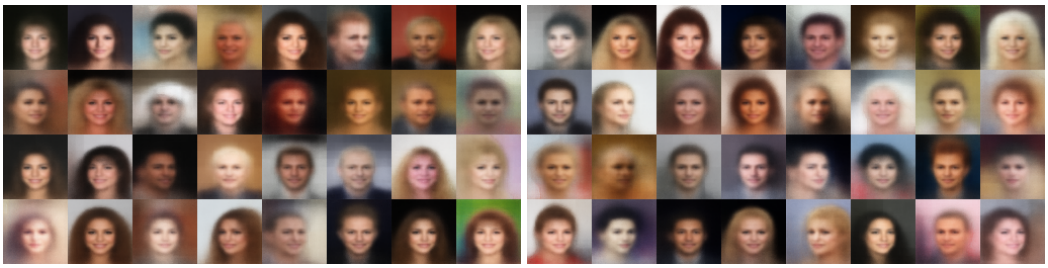

Figure 8: VNCA CelebA results when trained with $\beta = 100$. Left: Test set reconstructions. Right: Unconditional samples from the prior. Note: this is showing an average over 1000 samples from $p(x|z)$, since individual samples are noisy. This version achieves much worse bits per dimension ≥ 5.40, but visually the samples are much better.

Seeing that the reconstructions are good, but the samples are poor, we train a VNCA with $\beta = 100$, which strongly encourages the VNCA to produce better samples at the cost of the reconstructions. See fig. 8 for the results. This VNCA is technically a much worse generative model with bits per dimension ≥ 5.40, but visually the samples are much better. Individually the samples are noisy due to the large variance of the logistic distribution, so we visualize the average sample. For individual samples from the likelihood see the appendix.

With the improved visual samples, we can visualize the growth (Fig. 9). Here we see that the VNCA has learned a self-organizing generative process that generates faces from random initial seeds. Similar to the MNIST results, the images go through alternating steps of doubling and NCA steps that refine the pixellated images after a doubling.

Resilience to damage. Due to the iterative self-organising nature of the computation, the original NCA architectures were shown to be resilient in the face of damage, especially if trained for it (Mordvintsev et al., 2020). To examine the VNCAs resilience to damage, we perform an experiment in which we set the right half of all cell states to zero, at varying stages of growth (Fig. 10). Despite not being trained for it, the VNCA is resilient to damage in the early stages of growth, especially to z_0. Damage at later stages of growth simply leave the right half of the image blank, while the left half still look decent. It seems the early steps focus on creating a globally consistent image, and later steps focus on more local features. This is consistent with the increasing resolution along the steps, which determine how large part of an image the VNCA can act on, through its local-only communication.

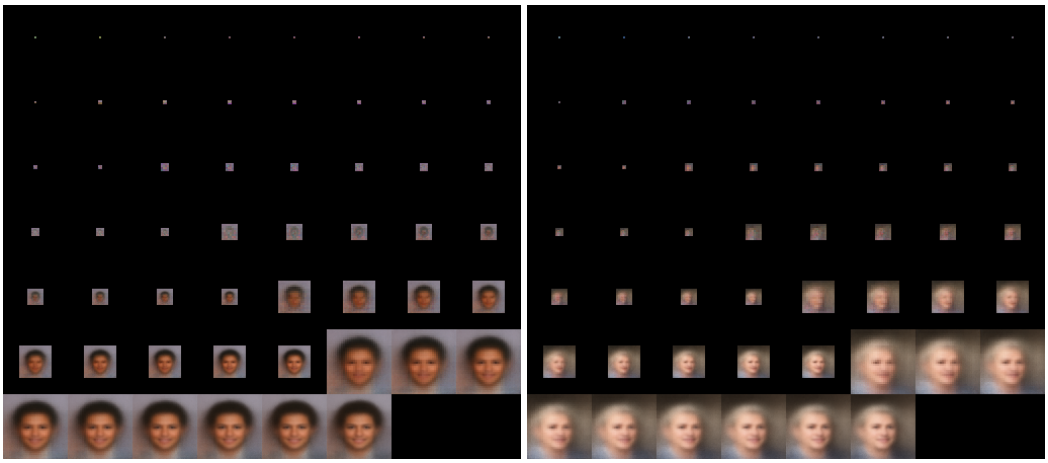

Figure 9: Growth of CelebA samples from the prior from a VNCA trained with $\beta = 100$. Note: this is showing an average over 1000 samples from $p(x|z)$, since individual samples are noisy.

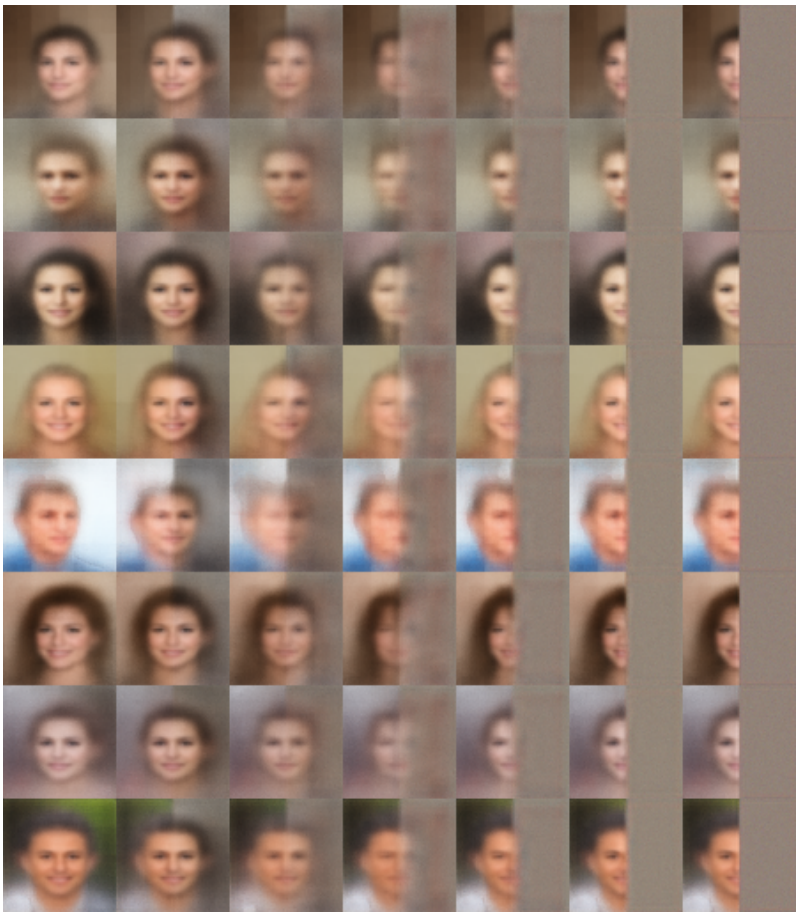

Figure 10: Unconditional samples, after being exposed to damage at varying stages of growth. The leftmost image is not exposed to any damage. The next six images represent damage applied at steps $t = [0, 8, 16, 24, 32, 40]$ corresponding to the initial state, and the steps just before each doubling.

4 DISCUSSION

Inspired by the biological generative process of cellular division and differentiation, we introduced the VNCA, a generative model based on a self-organising & iterative computation, relying only on local cellular communication. We showed that, while far from state-of-the-art, it is capable of learning a self-organising generative model of data that exhibits some resilience to damage, especially early in the stages of growth, and some rudimentary super-resolution capability.

One challenging aspect of the VNCA is that all the parameters of the decoder are in the update function, which is iteratively applied $M(K+1)$ times (K doublings and M NCA steps per doubling). This means that adding more parameters, and thus more computation, to the decoder is $M(K+1)$ times more computationally costly than in a standard non-iterative decoder. In other words, we could have a decoder for CelebA with $M(K+1) = 48$ times more parameters at the same computational cost. For this reason, the update function we use has few parameters, just 1.2M, compared to state-of-the-art decoders which routinely have hundreds of millions, allowing much more powerful decoders (Maaløe et al., 2019; Salimans et al., 2017; Vahdat & Kautz, 2020). The VNCA showed some resilience to damage, despite not being trained explicitly for this. It would be interesting to try to train it explicitly to recover from damage, akin to the NCA in Mordvintsev et al. (2020). Similarly it would be interesting to train it more directly to perform super-resolution. However, we would need to modify the generative model, and it is not immediately clear what a generative model that takes damage and super-resolution into account would look like.

Another set of generative models that generates and reconstructs data in an iterative process are diffusion models (Sohl-Dickstein et al., 2015), which have recently achieved state-of-the-art results in datasets like CelebA (Ho et al., 2020; Kim et al., 2021). While the VNCA decoding processes involves applying a function repeatedly, the decoding processes in a diffusion models involves iteratively sampling from a Markov chain of approximate distribution. In the future, the connection to diffusion models should be explored in more detail.

REFERENCES

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168. PMLR, 2018.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Mingxiang Chen and Zhecheng Wang. Image generation with neural cellular automatas. *arXiv preprint arXiv:2010.04949*, 2020.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

Kevin Frans. Stampca: Growing emoji with conditional neural cellular automata. *kvfrans.com*, 2021. URL `http://kvfrans.com/stampca-conditional-neural-cellular-automata/`.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pp. 1263–1272. PMLR, 2017.

William Gilpin. Cellular automata as convolutional neural networks. *Physical Review E*, 100(3): 032402, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Score matching model for unbounded data score. *arXiv preprint arXiv:2106.05527*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 29–37. JMLR Workshop and Conference Proceedings, 2011.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Lars Maaløe, Marco Fraccaro, Valentin Liévin, and Ole Winther. Biva: A very deep hierarchy of latent variables for generative modeling. *arXiv preprint arXiv:1902.02102*, 2019.

Alexander Mordvintsev, Ettore Randazzo, Eyvind Niklasson, and Michael Levin. Growing neural cellular automata. *Distill*, 2020. doi: 10.23915/distill.00023. https://distill.pub/2020/growing-ca.

Elias Najarro and Sebastian Risi. Meta-Learning through Hebbian Plasticity in Random Networks. In *Advances in Neural Information Processing Systems*, 2020. URL `https://arxiv.org/abs/2007.02686`.

Stefano Nichele, Mathias Berild Ose, Sebastian Risi, and Gunnar Tufte. Ca-neat: evolved compositional pattern producing networks for cellular automata morphogenesis and replication. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3):687–700, 2017.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR, 2013.

Deepak Pathak, Chris Lu, Trevor Darrell, Phillip Isola, and Alexei A Efros. Learning to control self-assembling morphologies: a study of generalization via modularity. *arXiv preprint arXiv:1902.05546*, 2019.

Michael Rubenstein, Christian Ahler, and Radhika Nagpal. Kilobot: A low cost scalable robot system for collective behaviors. In *2012 IEEE international conference on robotics and automation*, pp. 3293–3298. IEEE, 2012.

Alejandro Hernandez Ruiz, Armand Vilalta, and Francesc Moreno-Noguer. Neural cellular automata manifold. *arXiv preprint arXiv:2006.12155*, 2020.

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

Adam Santoro, David Raposo, David GT Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427*, 2017.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/sohl-dickstein15.html`.

Kenneth O Stanley and Risto Miikkulainen. A taxonomy for artificial embryogeny. *Artificial life*, 9 (2):93–130, 2003.

Shyam Sudhakaran, Djordje Grbic, Siyan Li, Adam Katona, Elias Najarro, Claire Glanois, and Sebastian Risi. Growing 3d artefacts and functional machines with neural cellular automata. *arXiv preprint arXiv:2103.08737*, 2021.

Yujin Tang and David Ha. The sensory neuron as a transformer: Permutation-invariant neural networks for reinforcement learning. *arXiv preprint arXiv:2109.02869*, 2021.

Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *arXiv preprint arXiv:2007.03898*, 2020.

Aaron Van Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pp. 1747–1756. PMLR, 2016.

A Appendix

A.1 Model specification

The detailed model specification can be seen below in the standard PyTorch Module format.

```
VNCA(
  (encoder): Sequential(
    (0): Conv2d(3, 32, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
    (1): ELU(alpha=1.0)
    (2): Conv2d(32, 64, kernel_size=(5, 5), stride=(2, 2), padding=(2, 2))
    (3): ELU(alpha=1.0)
    (4): Conv2d(64, 128, kernel_size=(5, 5), stride=(2, 2), padding=(2, 2))
    (5): ELU(alpha=1.0)
    (6): Conv2d(128, 256, kernel_size=(5, 5), stride=(2, 2), padding=(2, 2))
    (7): ELU(alpha=1.0)
    (8): Conv2d(256, 512, kernel_size=(5, 5), stride=(2, 2), padding=(2, 2))
    (9): ELU(alpha=1.0)
    (10): Flatten(start_dim=1, end_dim=-1)
    (11): Linear(in_features=8192, out_features=512, bias=True)
  )
  (nca): DataParallel(
    (module): NCA(
      (update_net): Sequential(
        (0): Conv2d(256, 256, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
        (1): Residual(
          (delegate): Sequential(
            (0): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
            (1): ELU(alpha=1.0)
            (2): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
          )
        )
        (2): Residual(
          (delegate): Sequential(
            (0): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
            (1): ELU(alpha=1.0)
            (2): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
          )
        )
        (3): Residual(
          (delegate): Sequential(
            (0): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
            (1): ELU(alpha=1.0)
            (2): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
          )
        )
        (4): Residual(
          (delegate): Sequential(
            (0): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
            (1): ELU(alpha=1.0)
```

```
            (2): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
          )
        )
        (5): Conv2d(256, 256, kernel_size=(1, 1), stride=(1, 1))
      )
    )
  )
)
```

A.2 CELEBA UNCONDITIONAL SAMPLES

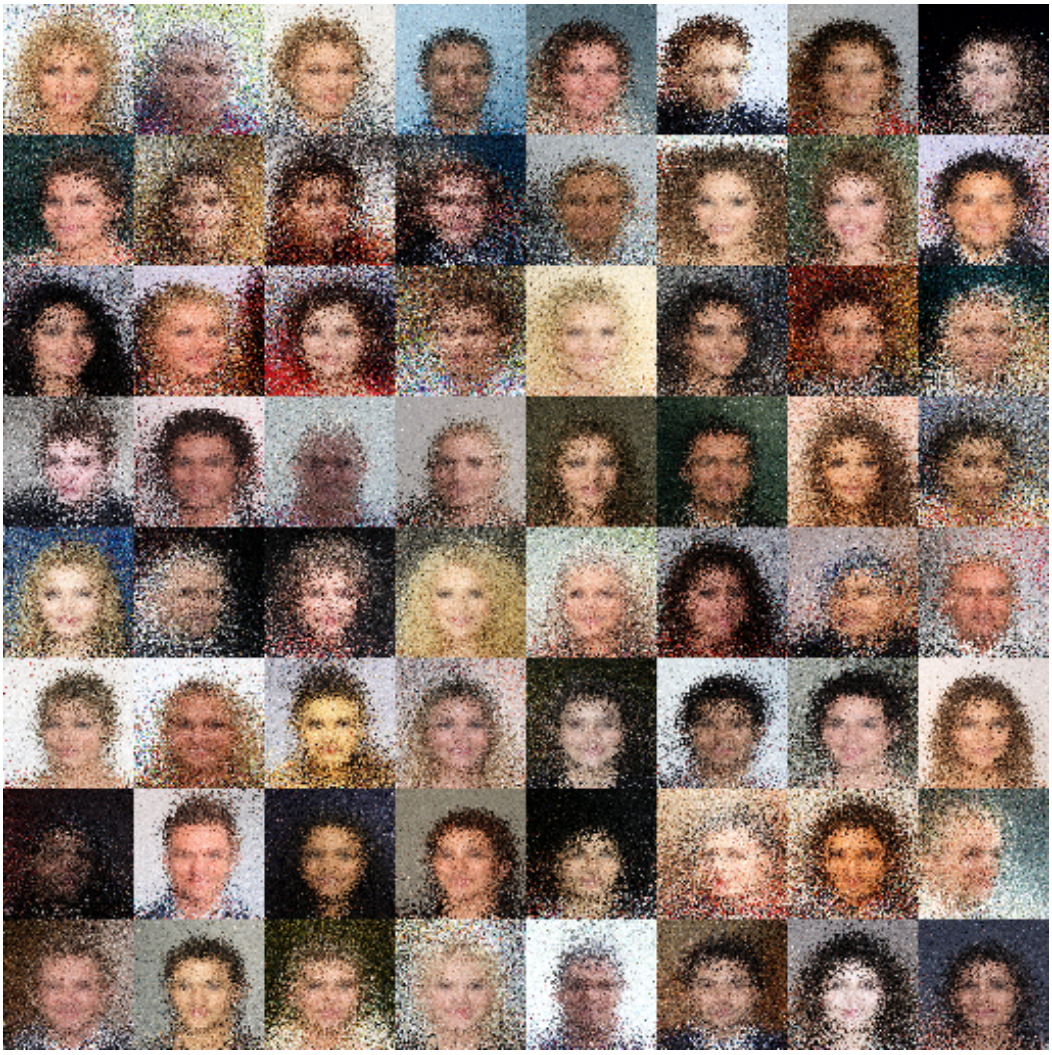

Figure 11: CelebA unconditional individual samples from a VNCA trained with $\beta = 100$.

A.3 ADDITIONAL EXPERIMENTS

We tried several variations of the VNCA that, perhaps surprisingly, were not better. We report them here so that future research can take them into account.

Size We tried deeper (up to 8 residual layers) and wider (512) update function networks, with no effect. A bigger latent space of 512 didn't seem to matter. A bigger NCA hidden state of 512 also didn't seem to matter. We tried $L = \{1, 4\}$ logistic mixtures, and $L = 1$ was better.

Non-strict NCA architectures. We tried replacing the 1×1 convolution layers with 3×3, which means the NCA is no longer relying only on local communication, but this also wasn't much better. We also tried a version that had a separate update network per scale, i.e., K networks. Surprisingly this was only slightly better.

Batch size Batch size did, surprisingly, seem to matter. Bigger batches consistently trained a lot better. With batch size 4 it was impossible to train. Batch size 32 seemed sufficient and was chosen as a trade-off between speed and quality.

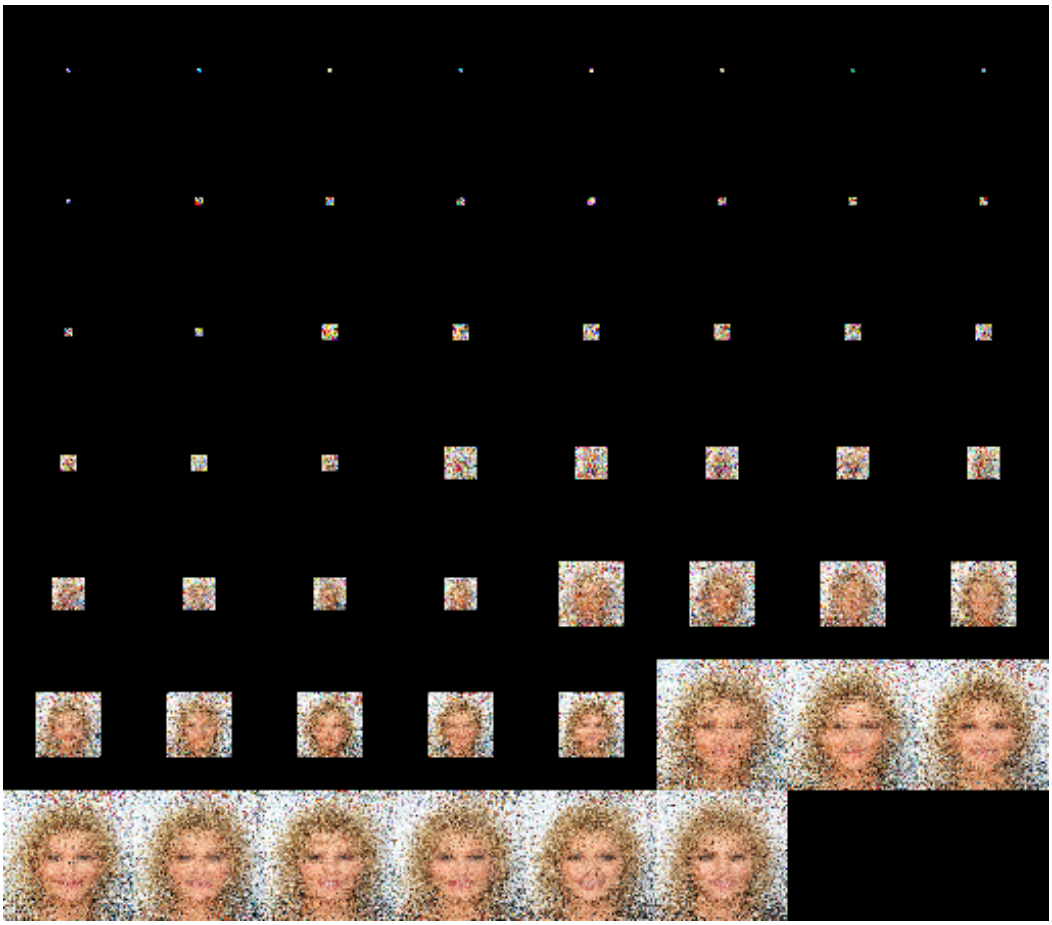

Figure 12: CelebA growth of unconditional sample from a VNCA trained with $\beta = 100$.

Update function: We tried 1) A GRU update function, 2) $z_{t+1} = z_t + \sigma(b)u(z_t)$, where b was a learned scalar and σ is the sigmoid and 3) a version which held half of the cell hidden state constant throughout all updates. None were better.