Topological defects propagate information in deep neural networks

Nabil Iqbal
Dept of Mathematical Sciences, Durham University
AMLab, University of Amsterdam
nabil.iqbal@durham.ac.uk

Max Welling
CuspAI
AMLab, University of Amsterdam
m.welling@uva.nl

Abstract

We study the effect of spontaneously broken discrete symmetries on the flow of information through deep neural networks. In physical systems, a spontaneously broken symmetry allows for the formation of *topological defects*, stable localized excitations that arise from the existence of distinct degenerate ground states. We demonstrate a similar phenomenon in deep learning. In particular, we study a image manipulation task solved by a simple toy model with recurrent dynamics inspired by Kuramoto oscillators. We show that a spontaneously broken $\mathbb{Z}_2$ symmetry in the internal representation space allows for the formation of a $\mathbb{Z}_2$ topological defect, or *domain wall*. These defects carry information in a novel manner through the layers of the network, providing potential advantages during training. We also demonstrate that under the addition of noise at each layer of the trained network, the spontaneous symmetry breaking is destroyed at a continuous phase transition, at which the RMSE loss displays indications of finite-size scaling laws that are familiar from critical phenomena in statistical physics. We speculate on further applications.

1 Introduction

In this work we will study the interplay of *symmetry* and the propagation of information through a deep neural network. This propagation of information is well-studied; training a very deep network can be intricate as it requires information about inputs and gradients to flow through the layers with neither significant decay nor divergence. It is interesting to make an analogy with physical systems, considering the direction of layer propagation as "time". From this point of view the machine learning problem of propagating information coherently through layers is somewhat similar to the physics problem of creating structures that are stable under time evolution in a potentially noisy and chaotic environment. It is thus interesting to study physical solutions to this problem and explore their deep learning analogues.

One mechanism for the creation of stable structures arises in the presence of discrete symmetries. When a physical system exhibits dynamics that are invariant under a symmetry, it can sometimes happen that the equilibrium or low-energy states are themselves not invariant under the symmetry. This situation – called *spontaneous symmetry breaking* – often allows for the creation of stable objects called *topological defects*; in the simplest case these defects are *domain walls* that connect one low-energy state with another in space, as shown in Figure 2.

In this work we will briefly review this concept. We will then study a simplified model of deep learning, inspired by [1] and structurally similar to well-known models in statistical physics. In this model of deep learning – where internal layers can be chosen to be equivariant under a $\mathbb{Z}_2$ discrete symmetry – similar topological defects exist and propagate coherently through the layers of the network. The existence of these stable objects is outside the usual "edge of chaos" paradigm. We

Accepted to the AI for Science Workshop (NeurIPS 2025).

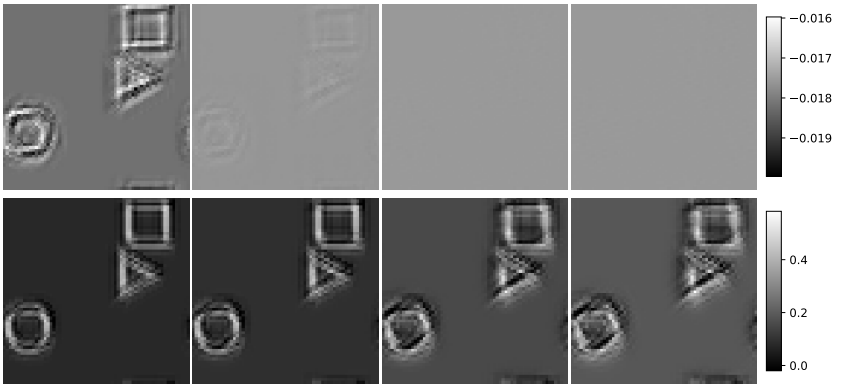

Figure 1: Example of information propagation through two networks at initialization. *Top:* output from model defined in (8) with no discrete symmetry. *Bottom:* output from our model defined in (4) with preserved $\mathbb{Z}_2$ symmetry. Plots shown are with $T = 5, 7, 15, 20$ from left. Information is rapidly washed out if symmetry is broken whereas it propagates through the layers – encoded in topological defects – if symmetry is present.

argue that they enhance trainability, in that the neural network learns to manipulate these objects and uses them to perform the required calculation. This provides an interesting viewpoint on the nature of computation in this model, as well as motivating architectural innovations that could have practical applications.

1.1 Previous work

This propagation of information through layers of a network is a well-studied problem (e.g. [2–5]; see [6] for a review) with many practical applications, though to our knowledge the specific role of symmetry and the mechanisms we discuss have not been studied before. The precise model we study is a simplified version of [1], which fits into a larger body of literature where physically-motivated architectures perform computations [7–9].

1.2 Brief review of spontaneous symmetry breaking and domain walls

Here for completeness we provide a brief review of the relevant concepts in spontaneous symmetry breaking and domain walls using the canonical example of the 2d Ising model on a square lattice [1] The reader who is already sufficiently familiar with these concepts is encouraged to skip to Section 2.

The Ising model is described by a set of binary variables $\sigma_i = \pm 1$ (usually called "spins") on a square lattice, with sites labeled by i. The energy E of a given configuration of spins is defined to be

$$E[\sigma] = -\sum_{\langle ij \rangle} \sigma_i \sigma_j \tag{1}$$

where the notation $\langle ij \rangle$ indicates a sum over nearest neighbours. Importantly this system has a $\mathbb{Z}_2$ symmetry: the operation

$$\sigma_i \to -\sigma_i \tag{2}$$

leaves the energy invariant. We see that this energy penalizes the disagreement of two spins who are connected by a link $\langle ij \rangle$.

The probability of a configuration of spins at a temperature $T = \beta^{-1}$ is given by the Boltzmann weight

$$p[\sigma] = \frac{1}{Z(\beta)} \exp(-\beta E[\sigma]) \tag{3}$$

where the partition function $Z(\beta)$ provides a normalizing constant. Note that at low temperatures (large β) most of the probability mass will accumulate on configurations of low energy. The states

[1] This material is standard: see e.g. [10, 11] for reviews of topological defects, or [12] for a textbook review of the 2d Ising model.

which minimize the energy – the *ground states* – can easily seen to be those with all spins aligned, i.e. *either* $\sigma_i = +1$ for all i *or* $\sigma_i = -1$ for all i. Each individual ground state is *not* invariant under the $\mathbb{Z}_2$ symmetry (2); rather the symmetry operation takes one of the ground states to another. This is an example of *spontaneous symmetry breaking*.

It is possible to consider a configuration of spins which is in one of the ground states $\sigma_i = +1$ in a localized region of space, and in the other ground state $\sigma_i = -1$ in the rest of space, as shown in Figure 2. The adjoining 2d regions are connected by a 1d junction which interpolates between the two ground states; this junction is called a *topological defect* or a *domain wall*. Note that the energy of this configuration does not depend on the area enclosed by the domain wall (as both the interior and exterior of the domain wall are separately in their ground state), but only on its perimeter. It thus costs "less energy" than one might have expected for a large object in a local system. Furthermore there is a low-energy spectrum of states formed by deforming it while leaving its length constant. The existence of such an object and its associated low-energy deformations is a direct consequence of the pattern of symmetry breaking and spatial locality.

Next, it is interesting to consider *dynamics* in time. One usually chooses a Markovian and local update rule for the spins $\sigma_{i,t} \to \sigma_{i,t+1}$ such that the late-time dynamics of the resulting Markov chain samples from the equilibrium distribution (3). Examples of such update rules (e.g. Glauber, or Metropolis-Hastings single-spin flips) can be found in standard textboks, e.g. [13]. Under such an update rule the domain wall can slowly change its shape as spins at the boundary are flipped, but it cannot suddenly cease to exist, as that would require the coordinated flipping of *all* of the spins in the interior, an exponentially unlikely occurrence in a local system.

Thus domain walls are approximately stable in time, and if one ever wanted to encode information stably in the 2d Ising model they would be a useful way to do so.

As one increases the temperature, the spontaneous symmetry breaking is lost at a phase transition at a critical temperature $\beta = \beta_c = \frac{1}{2}\log(1+\sqrt{2})$. The 2d Ising phase transition transition is very well understood [14]. We will return to this later in Section 3 to explore potential analogues in deep learning.

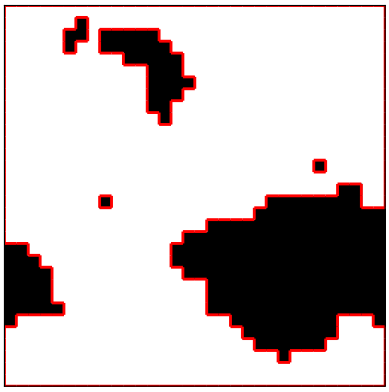

Figure 2: A typical configuration of the 2d Ising model at low temperature. Note existence of macrosopic regions of spins with $\sigma_i = +1$ (black) and $\sigma_i = -1$ (white). Domain walls are the boundary of these regions, highlighted in red.

2 Kuramoto oscillators for information processing

We now introduce the basic model architecture we will use for our problem; it is designed to be similar to well-studied models (i.e. the Heisenberg model; see e.g. [15] for a review) in statistical physics, albeit with less symmetry.

2.1 Model background

The model we use is a simplified version of the Kuramoto model studied in [1]. The basic representation space is a set of N dimensional oscillators arranged on a square lattice $\mathbf{x}_i \in \mathbb{R}^{N\times H\times W}$, where the sites are labeled by i. Each oscillator is normalized to have length 1 in the internal direction, $||\mathbf{x}_i|| = 1$, so the true representation space is given by $(S^{N-1})^{H\times W}$.

Each oscillator is time-dependent, and its time evolution is given by

$$\dot{\mathbf{x}}_i(t) = \Pi_{\mathbf{x}_i(t)}\left(\sum_j J_{ij}\mathbf{x}_j(t)\right) \tag{4}$$

where J_{ij} is a trainable and finite convolutional kernel with width k that is fully connected in the N-dimensional internal space, and $\Pi_{\mathbf{x}_i}$ is a projector which guarantees that the time derivative lies

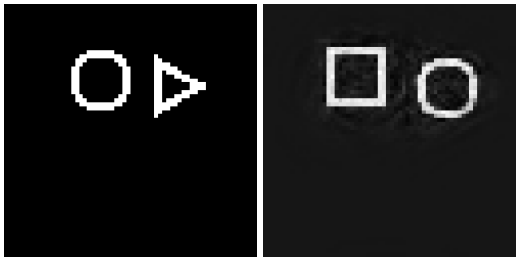

Figure 3: Example of task. *Left:* Input image is a series of randomly distributed shapes, each of which is a circle, triangle or square. Training data consists of examples where images have been transformed according to the rule (circle to square), (square to triangle) and (triangle to circle) *Right:* Output from trained network.

on the tangent plane to the sphere i.e.

$$\Pi_{\mathbf{x}}(\mathbf{y}) = \mathbf{y} - (\mathbf{x} \cdot \mathbf{y})\mathbf{x} \, . \tag{5}$$

The computation is done by supplying the input data as initial conditions at $\mathbf{x}_i(t = 0)$ and then solving (a discrete approximation of) the time evolution equation (4) to a later time $t = T$; the output to the calculation is then read off in the values of the rotors $\mathbf{x}_i(t = T)$, as shown in Figure 4. It was shown in [1] that a similar framework (with some extra features – a nontrivial natural frequency and with data supplied as an external sourcing term) provides strong performance – superior to many methods with conventional threshold units – on a variety of computer vision and reasoning tasks.

For simplicity of analysis we will consider a further restriction: we will take J_{ij} to be *symmetric* under the interchange of its internal and spatial indices. The dynamics (4) is then Lyapunov and the following energy functional monotonically decreases:

$$E(t) \equiv -\frac{1}{2} \sum_{i,j} (\mathbf{x}_i)^T J_{ij} \mathbf{x}_j \qquad \frac{dE}{dt} < 0 \, . \tag{6}$$

When performing a computation we expect the final state of the system at $t = T$ to be in an approximate minimum of the energy, where the choice of minimum depends on the initial conditions at $t = 0$ (i.e. the input). A successfully trained neural network has appropriately adjusted the trainable J_{ij} to obtain an an energy landscape such that the resulting minimum of the energy performs the desired computation.

Symmetries: In the absence of any further restrictions on J_{ij}, this model has a $\mathbb{Z}_2$ symmetry which acts as

$$\mathbf{x}_i \to -\mathbf{x}_i \tag{7}$$

This symmetry action preserves the equations of motion (in that if $\mathbf{x}_i(t)$ is a solution, then so is $-\mathbf{x}_i(t)$). Equivalently, it leaves the energy functional (6) invariant. As $||\mathbf{x}_i|| = 1$ the symmetry has no fixed point, and there is no way for a configuration to be invariant under the symmetry. Thus minimum energy configurations always come in equal-energy pairs related by (7). This symmetry will play a crucial role in what follows.

Outlook: If T is large, (6) may lead one to be surprised that this model can perform any task at all; translational eqivariance implies that the true minima of the energy will be constant (or possibly with a simple repeating unit cell) in the spatial directions, and thus one might expect the model to have a tendency to provide outputs that are approximately independent of the input. We will revisit this shortly.

Finally, this recurrent architecture can be viewed as representing a deep neural network, where weights are shared across layers and where the number of discrete time evolution steps T is a proxy for the depth. In the end we discuss important differences between our toy model and more conventional networks.

2.2 Domain walls and their role

We orient the rest of the discussion around empirical observations. For illustrative purposes we consider a toy "transmogrification" task in computer vision: as input we supply an image with several

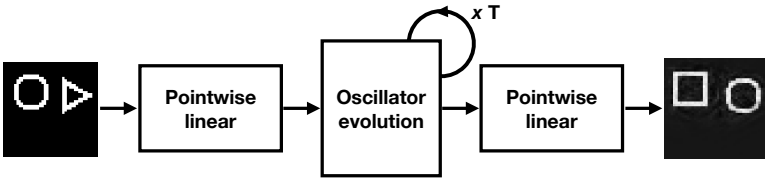

Figure 4: Network architecture; a linear layer acting pointwise is used to lift the input image into the initial conditions on the $\mathbf{x}_i(t=0)$. A discretization of the ODE (4) is then evolved for T steps and then a further linear layer acting pointwise on $\mathbf{x}_i(t=T)$ is used to create a single-channel output image.

shapes, each of which is either a circle, a triangle, or a square. The task is to convert each circle to a square, each square to a triangle, and each triangle to a circle, as shown in Figure 3.

The input data is a single channel image where each pixel is either black or white, encoded as $p_i^{\text{input}} = \pm 0.5$. We use a pointwise linear layer (with no bias) to embed this input image into $\mathbf{x}_i(t=0)$ and then evolve with T discrete time evolution steps, and at the end we use another pointwise linear layer (with bias) to read out the output p_i^{output}. Details (of which some are important for appropriate $\mathbb{Z}_2$ equivariance) are given in Section A.1 of the Supplementary Material.

To illustrate the important role played by the domain walls and $\mathbb{Z}_2$ symmetry, we will compare the model defined by (4) with a *symmetry-broken* variant of the model that does not have a $\mathbb{Z}_2$ symmetry:

$$\dot{\mathbf{x}}_i(t) = \Pi_{\mathbf{x}_i(t)} \left(\sum_j J_{ij} \mathbf{x}_j(t) + \mathbf{c} \right) \tag{8}$$

where $\mathbf{c}$ is a trainable spatially constant bias vector which breaks the symmetry $\mathbf{x}_i \to -\mathbf{x}_i$. Note that if one was considering the most general network without consideration of symmetry one would use this model and not (4).

Initialization. We first consider both networks at initialization, i.e. with the J_{ij} (and $\mathbf{c}$ if it exists) randomly chosen. In Figure 1 we show the output of both networks as we consider steadily increasing T.

Note that as T becomes large the output of the symmetry-broken network (8) steadily loses its memory of the input data as it approaches the (homogenous) global minimum of the energy. In the language of the "edge of chaos" this network is deeply in the ordered phase: the final configuration of the $\mathbf{x}_i(t)$ always tends to the minimum energy configuration, independent of the input.

For the symmetry-preserving network (4) this is not the case: interestingly, no matter how deep the network the information propagates all the way through at initialization despite being in an "ordered" phase. This happens due to the $\mathbb{Z}_2$ symmetry, which implies that there are at least two equilibrium configurations related by $\mathbf{x}_i \to -\mathbf{x}_i$. The local convolutional dynamics means that these two minima end up being glued together by domain walls to stably encode the initial conditions, as we show more explicitly in Figure 5.

Training: We now use an MSE loss and backpropagation to train the network to perform the task shown in Figure 3.

In Figure 6 we show that there is an advantage at early stages in training in having the $\mathbb{Z}_2$ symmetry. We attribute this to the information-propagating abilities of the domain walls described above. This task is simple enough that both types of models generally perform almost perfectly to the eye after 20 epochs of training (see e.g. Figure 3).

In Figure 9 in the Supplementary Material we show the existence of $\mathbb{Z}_2$ domain walls in the *trained* network; this is a direct analogue of Figure 5 for the network at initialization, and supports the idea that the trained network continues to use the $\mathbb{Z}_2$ symmetry to store information.

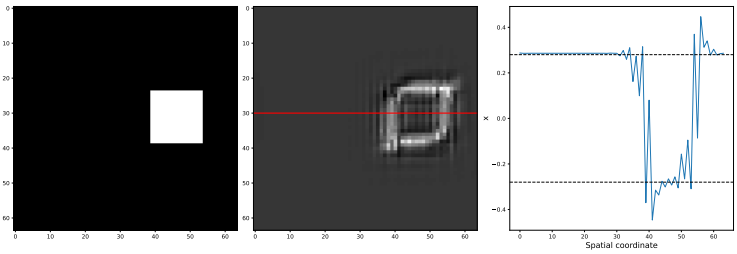

Figure 5: Existence of domain walls at initialization. *Left:* Sample input image, where we are using a filled in square to obtain larger area of constant pixel value, which is helpful for seeing the distinct ground states in the other two panels. This image is not in the training data. *Middle:* Energy density at $T = 20$ of randomly initialized network with $\mathbb{Z}_2$ symmetry. *Right:* A slice through the image along the red line showing one of the components of the oscillators $x = \mathbf{x}^0$. Note that energy density is mostly localized on the boundaries and is roughly the same inside and out, and that $\mathbf{x}$ inside the shape is approximately related by $\mathbf{x} \to -\mathbf{x}$ to that outside the shape.

3 Phase transitions and noise

We have seen that the model seems to use domain walls in the $\mathbb{Z}_2$ symmetry to encode information. The existence of domain walls depends on the fact that the symmetry is *spontaneously broken*, which in this context simply means that the states that minimize the energy (i.e. the attractor states at late T) are not invariant under the symmetry. They cannot be, as the action $\mathbf{x} \to -\mathbf{x}$ has no fixed points if $||\mathbf{x}|| = 1$.

3.1 Review of finite-temperature restoration of symmetry

Nevertheless there is a situation where we can consider destroying the domain walls through the addition of *noise*. Consider modifying the evolution equation (4) with a noise term

$$\dot{\mathbf{x}}_i(t) = \Pi_{\mathbf{x}_i(t)} \left(\sum_j J_{ij} \mathbf{x}_j(t) + \eta_i(t) \right) \tag{9}$$

where we add uncorrelated noise $\langle \eta_i(t) \eta_j(t) \rangle = \eta_0^2 \delta_{ij} \delta(t - t')$ at each timestep. Now the resulting Langevin dynamics means that $\mathbf{x}(\mathbf{t})$ will no longer move in a manner that strictly minimizes its energy. Instead the resulting late-time configurations will be drawn from a Boltzmann distribution $p(\mathbf{x}) \sim \exp(-\beta E(\mathbf{x}))$ where the value of the inverse temperature $\beta^{-1} \sim \eta_0^2$.

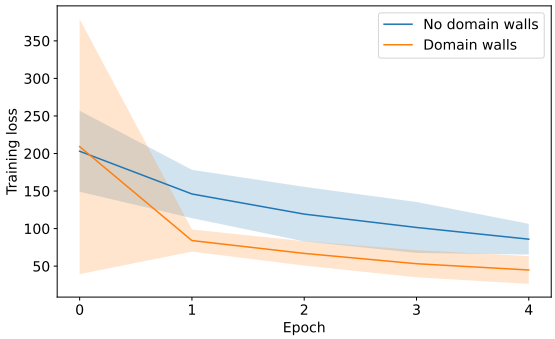

Figure 6: Advantage in training from the existence of domain walls. First 5 epochs of training are shown with $T = 30$. Shaded area is standard deviation over training 5 models in each class. Blue curve is given by model defined in (8) and orange curve by that defined in (4).

As we increase the noise level η_0, fluctuations in $\mathbf{x}(t)$ will become stronger and stronger, until eventually we expect that the symmetry will no longer be spontaneously broken. Rather than the system settling into one equilibrium $\mathbf{x}$ that minimizes the energy (and thus spontaneously breaks the symmetry), it will move from configuration to configuration in a way that is invariant under the symmetry[2], destroying any domain walls that might be present.

For conventional systems (such as the 2d Ising model discussed above) such a symmetry-restoring transition happens at a critical point which involves fluctuations that happen at all length scales,

[2]A canonical example of this is the destruction of magnetism as the temperature is increased; at low temperatures all the electron spins in the system are aligned, picking a preferred direction in the *ferromagnetic* phase. At high temperatures they fluctuate rapidly, and we find the *paramagnetic* phase.

and can usually be described by a continuum *conformal field theory* that does not depend on microscopic details [14, 16]. One consequence of this is that near the critical point the dependence of thermodynamic observables on the system size L and the temperature β is constrained. If one considers a system with a critical point at $\beta = \beta_c$, then dimensionless observables will not depend on the system size L and $\beta - \beta_c$ independently, but rather on the combination

$$f(L, \beta) = f\left(L^{\alpha}(\beta - \beta_c)\right) \tag{10}$$

where α is a pure number called a "critical exponent" that characterizes the phase transition in question. In mean field theory $\alpha = 2$, and for the 2d Ising model $\alpha = 1$. The non-trivial functional form (10) can be seen in simulations and is a clean signature of a phase transition; see e.g. [17] for an example of the relevant plots for the 2d Ising model.

3.2 Noisy neural network

We now consider adding noise to our system as in (9); if spontaneous symmetry breaking is important for the functioning of our neural network, then its performance will be degraded at a sufficiently high level of the noise. Furthermore, the local convolutional dynamics in our construction suggests that perhaps this destruction will happen at a critical point that obeys the usual rules of statistical physics.

To explore this hypothesis, in Figure 7 we compute the RMSE loss of the trained network for various sizes of input image. The maximum number of shapes included in an image is taken to scale with the size of the image so as to maintain a notion of local translational equivariance.

The RMSE loss indeed seems to *approximately* obey the form dictated by scale invariance (10), i.e.

$$\mathrm{RMSE}(L, \eta) = f(L^{\alpha}(\eta - \eta_c)). \tag{11}$$

In particular – nontrivially – there is a critical value of the noise $\eta = \eta_c \approx 0.0046$ at which all curves of different system size L intersect, as required by (11).

What is the value of the critical exponent α? If α is picked correctly then (11) says that all the curves at different L should collapse into one function of $L^{\alpha}(\eta - \eta_c)$ near the critical point. The best data collapse – which is still somewhat imperfect in our opinion, working well in the disordered phase but not in the other – appears to be around $\alpha \approx 0.3$. This value is somewhat surprising. The simplest model that fits our symmetries – $\mathbb{Z}_2$ and two spatial dimensions – is the 2d Ising model, for which we would have found $\alpha = 1$. (In the Supplementary Material we show that this value of α leads to a much worse data collapse). A conservative reading is that this transition is not in the 2d Ising class, and that its nature deserves further study.

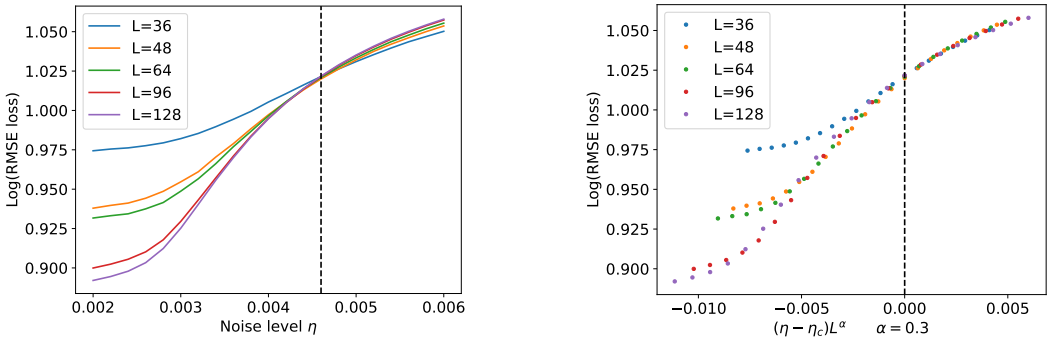

Figure 7: *Left:* We demonstrate that upon the addition of noise the RMSE loss curves appear to intersect at a critical point $\eta = \eta_c$ (dashed line). *Right:* The best data collapse that we could find in the vicinity of the critical point is at $\alpha \approx 0.3$.

For reference we show the output from the network with various noise levels in Figure 8. We find it curious that even at the supposed critical point the output still looks fairly reasonable to the eye. We also discuss various caveats from our analysis in Section A.3 in the Supplementary Material.

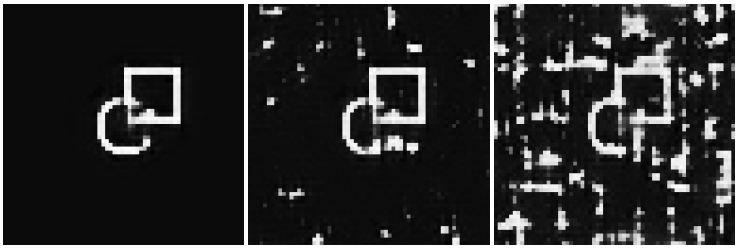

Figure 8: Adding noise to neural network evolution to show destruction of spontaneous symmetry breaking at critical point. *Left:* No noise. *Middle:* Noise supplied at putative critical point $\eta = 0.0046 \approx \eta_c$. *Right:* Noise level at $\eta = 0.009 \gg \eta_c$

4 Conclusion

In this work we have shown empirically that the spontaneous breaking of internal symmetries – and the resulting topological defects – plays an interesting role in information propagation through the recurrent layers of a deep neural network. In Figure 1 we saw that these objects permit information to be propagated through layers despite the network naively expected to be in an "ordered" phase in which all inputs are washed out. Activation of these objects is very simple: one must simply constrain the layers to preserve a discrete symmetry.

This observation fits into an existing body of knowledge. It is well-known in statistical physics that the presence of topological defects can dramatically change the long-distance and late-time properties of a system. Celebrated examples include the Berezinskii-Kosterlitz-Thouless transition in thin films [18, 19],[3], Polyakov confinement in three-dimensional electrodynamics [21, 22], the melting of crystals due to defect proliferation [23] and many others. The phenomena discussed in this work seems to be an example in a different context.

In our eyes there are many directions for future research. In this work we studied the simplest possible physically-motivated model that could still perform a non-trivial computation. It is important to study more realistic machine-learning models and layer structures; e.g. one important direction is to understand whether similar conclusions apply when each layer is *different* (i.e. without weight sharing in layers) but still preserves the symmetry.

We have not yet attempted a theoretical understanding. This should be possible by importing the calculational tools used to describe the physical phenomena above into the learning theory of [2–5]. There is also much to be understood about the phase transition described in Section 3; we performed a first attempt at understanding this using the standard tools of static critical phenomena, but this is actually a non-equilibrium phenomenon.

Zooming out slightly, the central ideas that made this possible were symmetry and locality, where the latter arose from the finite depth of the convolutional kernel in (4). One could also consider other symmetry groups and dynamical degrees of freedom, in which case the dimensionality and nature of the defects can change. If the symmetry in question is continuous then one will find in addition Goldstone modes [24] which one might also expect to propagate information coherently through layers. In the case of a $U(1)$ symmetry a possible topological defect is a *vortex*; it is interesting to note that such structures have been observed and argued to play an important role in real-time brain imaging [25].

Finally, one could consider other notions of "locality" than convolutional, e.g. those arising from a graph or attention-based architecture. To our knowledge there is little study of the dynamics of topological defects in such situations. We are optimistic that the ideas discussed in this work might eventually lead not only to theoretical insights but also operationally useful principles for new neural architectures.

[3]This is the topic of one-half of the 2016 Nobel Prize in physics, and indeed the prize announcement [20] provides an accessible introduction to the subject.

Acknowledgments

We are grateful to T. Miyato for helpful discussions at the initial stages of this work. This work was supported by a grant from the Simons Foundation (PD-Pivot Fellow-00004147, NI). NI is supported in part by the STFC under grant number ST/T000708/1.

References

[1] T. Miyato, S. Löwe, A. Geiger, and M. Welling, "Artificial kuramoto oscillatory neurons," *arXiv preprint arXiv:2410.13821* (2024) .

[2] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep information propagation," in *International Conference on Learning Representations*. 2017. `https://openreview.net/forum?id=H1W1UN9gg`.

[3] G. Yang and S. Schoenholz, "Mean field residual networks: On the edge of chaos," *Advances in neural information processing systems* **30** (2017) .

[4] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. Schoenholz, and J. Pennington, "Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks," in *International conference on machine learning*, pp. 5393–5402, PMLR. 2018.

[5] M. Chen, J. Pennington, and S. Schoenholz, "Dynamical isometry and a mean field theory of rnns: Gating enables signal propagation in recurrent neural networks," in *International Conference on Machine Learning*, pp. 873–882, PMLR. 2018.

[6] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, "Statistical mechanics of deep learning," *Annual review of condensed matter physics* **11** no. 1, (2020) 501–528.

[7] L. H. Liboni, R. C. Budzinski, A. N. Busch, S. Löwe, T. A. Keller, M. Welling, and L. E. Muller, "Image segmentation with traveling waves in an exactly solvable recurrent neural network," *arXiv preprint arXiv:2311.16943* (2023) .

[8] T. A. Keller, L. Muller, T. Sejnowski, and M. Welling, "Traveling waves encode the recent past and enhance sequence learning," *arXiv preprint arXiv:2309.08045* (2023) .

[9] M. Jacobs, R. C. Budzinski, L. Muller, D. Ba, and T. A. Keller, "Traveling waves integrate spatial information through time," *arXiv preprint arXiv:2502.06034* (2025) .

[10] S. Coleman, *Aspects of symmetry: selected Erice lectures*. Cambridge University Press, 1988.

[11] R. Rajaraman, "Solitons and instantons. an introduction to solitons and instantons in quantum field theory,".

[12] M. Kardar, *Statistical physics of fields*. Cambridge University Press, 2007.

[13] M. Newman and G. Barkema, *Monte Carlo methods in statistical physics*. Oxford University Press, 1999.

[14] P. Francesco, P. Mathieu, and D. Sénéchal, *Conformal field theory*. Springer Science & Business Media, 2012.

[15] S. Sachdev, "Quantum phase transitions," *Physics world* **12** no. 4, (1999) 33.

[16] J. Cardy, *Scaling and renormalization in statistical physics*, vol. 5. Cambridge university press, 1996.

[17] H. G. Katzgraber, "Introduction to Monte Carlo Methods," *arXiv e-prints* (May, 2009) arXiv:0905.1629, `arXiv:0905.1629 [cond-mat.stat-mech]`.

[18] V. L. Berezinskii, "Destruction of Long-range Order in One-dimensional and Two-dimensional Systems having a Continuous Symmetry Group I. Classical Systems," *Sov. Phys. JETP* **32** (1971) 493–500.

[19] J. M. Kosterlitz and D. J. Thouless, "Ordering, metastability and phase transitions in two-dimensional systems," *Journal of Physics C: Solid State Physics* **6** no. 7, (1973) 1181.

[20] Nobel Prize Outreach, "Advanced information." `https://www.nobelprize.org/prizes/physics/2016/advanced-information/`, 2025. Accessed: 2025-08-27.

[21] A. M. Polyakov, "Compact Gauge Fields and the Infrared Catastrophe," *Phys. Lett. B* **59** (1975) 82–84.

[22] A. M. Polyakov, "Quark Confinement and Topology of Gauge Groups," *Nucl. Phys. B* **120** (1977) 429–458.

[23] B. Halperin and D. R. Nelson, "Theory of two-dimensional melting," *Physical Review Letters* **41** no. 2, (1978) 121.

[24] J. Goldstone, A. Salam, and S. Weinberg, "Broken symmetries," *Physical Review* **127** no. 3, (1962) 965.

[25] Y. Xu, X. Long, J. Feng, and P. Gong, "Interacting spiral wave patterns underlie complex brain dynamics and are related to cognitive processing," *Nature human behaviour* **7** no. 7, (2023) 1196–1215.

A Supplementary Material

A.1 Experimental details

In our experiments the training set is 28K pairs of 64×64 images. Each pair consists of an input image, which is a set of randomly positioned and sampled shapes drawn from (circle, square, triangle), and an output image, which has the same shapes transformed as (circle to square), (square to triangle) and (triangle to circle), as shown in Figure 3. We store white pixels with the pixel value $+0.5$ and black pixels with -0.5.

Our network architecture is shown in Figure 4. and consists of an initial linear $1 \to N$ linear layer (with zero bias) acting on each spatial point, time evolution for T steps of (4), and a final output $N \to 1$ linear layer (with bias). It is important that the initial linear layer have zero bias – this means that the transformation on pixel space (black $\to$ white) is represented as the $\mathbb{Z}_2$ $\mathbf{x} \to -\mathbf{x}$ on the representation space. (Without this, there is a great deal more variability in model performance, as the model sometimes fails to use the domain walls).

We take $N = 16$ and a kernel size of 5 for the J_{ij}. We discretize the time-evolution as

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \gamma \dot{\mathbf{x}}_i(t) \tag{12}$$

with $\gamma = 0.25$. To obtain a fully trained network we train for 20 epochs.

A.2 Supplementary images

We present some further plots to support our conclusions.

In Figure 9 we present the analogue of Figure 5 for a trained network, which shows the existence of domain walls in the latent variables of the network.

In Figure 10 we demonstrate that attempting to fit the RMSE loss with the Ising value of the critical exponent $\alpha = 1$ leads to a visibly worse data collapse than the value suggested in the text $\alpha \approx 0.3$. Though we remain confused about the nature of the critical point, it is definitely not 2d Ising.

A.3 Possible caveats with finite-size scaling analysis

We find the phase transition described in the main text extremely interesting. For completeness we would here like to discuss all of the potential caveats with that analysis, which we plan to tackle in future work:

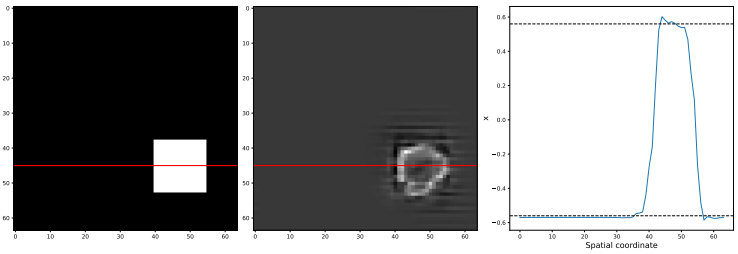

Figure 9: This is the analogue for a trained network of Figure 5 (which was for a random network). *Left:* Sample input image. *Middle:* Energy density at $T = 30$ of a trained network with $\mathbb{Z}_2$ symmetry. *Right:* A slice through the image showing one of the components of the oscillators $x = \mathbf{x}^0$. Note that energy density is mostly localized on the boundaries and is roughly the same inside and out, and that $\mathbf{x}$ inside the shape is approximately related by $\mathbf{x} \to -\mathbf{x}$ to that outside the shape.

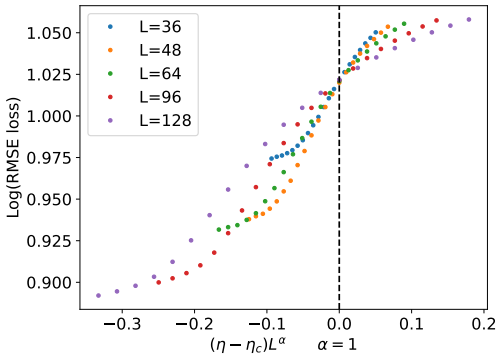

Figure 10: Attempt to fit the data collapse of the RMSE loss using the 2d Ising exponent $\alpha = 1$; this is a much worse fit than the right panel of Figure 7.

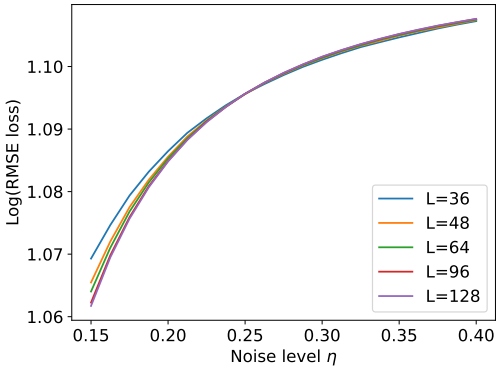

Figure 11: Attempt to search for crossing in symmetry-broken model.

1. Above, the RMSE loss is computed on an output $p(x)$ that assigns the background (black) to 0 and the foreground (white) to 1. We do not find similar scaling for other choices (e.g. ± 0.5). We also do not find the scaling for the MSE loss. In standard critical phenomena it is important that one compute an observable that is dimensionless, or at a more technical level has a vanishing scaling dimension (e.g. the so-called "Binder cumulant" [17] is a usual choice). We believe that only for the choice of pixel values above does the RMSE loss have vanishing scaling dimension, but we do not have a precise understanding as to why.
2. Different trained networks appear to display some variability of the ideal α, which is in conflict with general principles of universality. One possible explanation is that the "equilibrium" configuration for different network have different repeating unit cells which break different subgroups of the *translational* symmetry. We leave a careful study of this to future work. Another possibility is that this is simply an out-of-equilibrium observable that does not obey usual equilibrium statistical physics rules.
3. Finally, as a useful sanity check one should perform the same analysis on a symmetry *explicitly broken* network. We have done so in Figure 11. here we don't see any sharp evidence of crossing (note there is no "fan out" on the right hand side of Figure 11, unlike in Figure 7). Nevertheless all the curves do begin to coincide. We are not sure at the moment whether this describes a transition in the thermodynamic sense, and we believe this deserves further study.