Predictive Coding with Topographic Variational Autoencoders

Anonymous ICCV submission

Paper ID ****

Abstract

Predictive coding is a model of visual processing which suggests that the brain is a generative model of input, with prediction error serving as a signal for both learning and attention. In this work, we show how the equivariant capsules learned by a Topographic Variational Autoencoder can be extended to fit within the predictive coding framework by treating the slow rolling of capsule activations as the forward prediction operator. We demonstrate quantitatively that such an extension leads to improved sequence modeling compared with both topographic and non-topographic baselines, and that the resulting forward predictions are qualitatively more coherent with the provided partial input transformations.

1. Introduction

Topographic organization in the brain describes the observation that nearby neurons on the cortical surface tend to have more strongly correlated activations than spatially distant neurons. From the simple orientation of lines [22] to the complex semantics of natural language [23], organization of cortical activity is observed for a diversity of stimuli and across a range of species. Given such strong and ubiquitous observations, it seems only natural to wonder about the computational benefits of such organization, and if the machine learning community can take advantage of such design principles to develop better inductive priors for deep neural network architectures.

One inductive prior which has gained popularity in recent years is that of equivariance. At a high level, a representation is said to be equivariant if it transforms in a known predictable manner for a given transformation of the input. A fundamental method for constructing equivariant representations is through structured parameter sharing, constrained by the underlying desired transformation group [10, 39, 17, 18]. The most well known example of an equivariant map is the convolution operation, which is equivariant to translation. One can think of a convolutional layer as a function which shares the same feature extractor

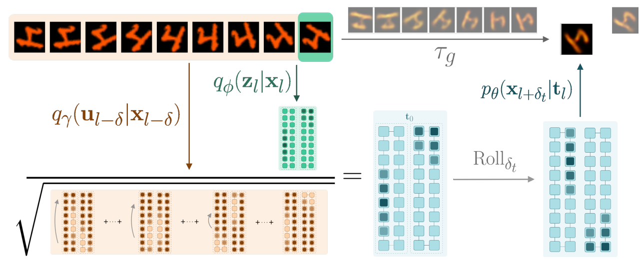

Figure 1. Overview of the Predictive Coding Topographic VAE. The transformation in input space τ_g becomes encoded as a Roll within the equivariant capsule dimension. The model is thus able to forward predict the continuation of the sequence by encoding a partial sequence and rolling activations within the capsules.

parameters over all elements of the translation group, i.e. all spatial locations. Similarly, a model which is equivariant to rotation is one which shares parameters across all rotations. Existing group equivariant neural networks [10] therefore propose to maintain 'capsules' of tied-parameters which are correlated by the action of the group. By reducing the number of trainable parameters while simultaneously increasing the information contained in the representation, equivariant neural networks have demonstrated significant improvements to generalization and data efficiency [10, 42, 51].

These sets of transformed weights, which we refer to as 'equivariant capsules', are reminiscent of a type of topographic organization observed in the primary visual cortex (V1), namely orientation columns [22]. This insight encouraged the development of the Topographic Variational Autoencoder (TVAE) [2], linking equivariance and topographic organization in a single framework. In the original work, the TVAE was introduced and demonstrated to learn topographic equivariant capsules in an entirely unsupervised manner from observed transformation sequences. Further, the inductive priors of equivariance and 'slowness' integrated into the TVAE were demonstrated to be beneficial for modeling sequence transformations, ultimately resulting in higher log-likelihood on held-out data when compared with VAE baselines.

In this work, we propose to extend the TVAE with an additional inductive prior – that of predictive coding [19]. At a high level, predictive coding suggests that one significant goal of the brain is to predict future input and use the forward prediction error as a learning signal. In the context of the TVAE, we observe that the existence of topographically organized capsules, combined with a slowness prior, permit efficient forward prediction through simple forward rolling of capsule activations. We demonstrate empirically that such a model is able to more accurately predict the immediate future, while simultaneously retaining the learned equivariance properties afforded by the original TVAE.

2. Background

The Topographic VAE [2] relies on a combination of fundamentally related inductive priors including *Equivariance*, *Topographic Organization*, and *Slowness*. In this section we will give a brief description of these concepts, and further introduce predictive coding as it relates to this work.

2.1. Equivariance

Equivariance is the mathematical notion of symmetry for functions. A function is said to be an equivariant map if the the result of transforming the input and then computing the function is the same as first computing the function and then transforming the output. In other words, the function and the transformation commute. Formally, $f(\tau_\rho[\mathbf{x}]) = \Gamma_\rho[f(\mathbf{x})]$, where τ and Γ denote the (potentially different) operators on the domain and co-domain respectively, but are indexed by the same element ρ. The introduction of the Group-convolution [10] and ensuing work [11, 51, 50, 18, 47] allowed for the development of analytically equivariant neural network architectures to a broad range of group transformations, demonstrating measurable benefits in domains such as medical imaging [46, 3] and molecular generation [41]. Recently, more work has begun to explore the possibility of 'learned' equivariance guided by the data itself [5, 14, 20]. The TVAE and the extension presented in this paper are another promising step in this direction.

2.2. Topographic Organization

Topographic generative models can be seen as a class of generative models where the latent variables have an underlying topographic organization which determines their correlation structure. As opposed to common generative models such as Independant Component Analysis (ICA) [4, 27] or VAEs [29, 40], the latent variables in a topographic generative model are not assumed to be entirely independant, but instead are more correlated when they are spatially 'close' in a predetermined topology. Typically, simple topologies such as 1 or 2 dimensional grids are used, often with circular boundaries to avoid edge effects.

One way a topographic generative model can be described, as in [25], is as a hierarchical generative model where there exist a set of higher level independant 'variance generating' variables $\mathbf{V}$ which are combined locally along the topology to generate the variances of the lower level topographic variables $\mathbf{T}$. Formally, for an adjacency matrix $\mathbf{W}$, and an appropriate non-linearity ϕ, the variances are computed as $\boldsymbol{\sigma} = \phi(\mathbf{WV})$. In the second stage, the lower level variables are sampled independently, but with their scale determined by the now topographically organized variable $\boldsymbol{\sigma}$: $\mathbf{T} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2\mathbf{I})$. In later work [26], Hyvärinen *et al*. further showed this framework to be a generalization Independant Subspace Analysis (ISA) [24] and some variants of Slow Feature Analysis (SFA) [43, 49, 44] by careful choice of topography $\mathbf{W}$. The Topographic VAE takes advantage of both this framework and these connections to construct slow-transforming capsules which learn to become equivariant to observed sequence transformations.

2.3. Predictive Coding

In the machine learning literature, one of the most intuitive and common frameworks for unsupervised learning relies on predicting unseen or missing contextual data from a given input. This idea, informally called predictive coding, has led to some of the most well known advances in the field across a range of domains including: natural language processing (word2vec [36], GPT3 [6], Bert [13]), vision (CPC [45], SimCLR [7], GreedyInfoMax [34]), speech (APC [8]), and more [21, 32]. In the theoretical neuroscience literature, predictive coding denotes a framework by which the cortex is a generative model of sensory inputs [16, 38, 19, 9], and has been linked to probabilistic latent variable models such as VAEs [35]. Substantial evidence has been gathered supporting the existence of some form of predictive coding in the brain [1, 12, 15], and numerous computational models have been proposed which replicate empirical observations [38, 33, 28]. Given these computational successes, and the mounting support for such a mechanism underlying biological intelligence, we strive to formalize the relationship between predictive coding and TVAEs in this work.

3. Predictive Coding with Topographic VAEs

In this section we introduce the generative model underlying the Predictive Coding Topographic VAE (PCTVAE) and highlight the differences with the original model – including making the conditional generative distribution forward predictive, and limiting the temporal coherence window to only include past variables.

3.1. The Forward Predictive Generative Model

We assume that the observed sequence data is generated from a joint distribution over observed and latent variables $\mathbf{x}_l$ and $\mathbf{t}_l$ which factorizes over timesteps l, and further factorizes into the product of a forward predictive conditional and the prior:

$$p_{\{\mathbf{X}_{l+1}, \mathbf{T}_l\}_l}(\{\mathbf{x}_{l+1}, \mathbf{t}_l\}_l) = \prod_l p_{\mathbf{X}_{l+1}|\mathbf{T}_l}(\mathbf{x}_{l+1}|\mathbf{t}_l) p_{\mathbf{T}_l}(\mathbf{t}_l) \quad (1)$$

ICCV #****

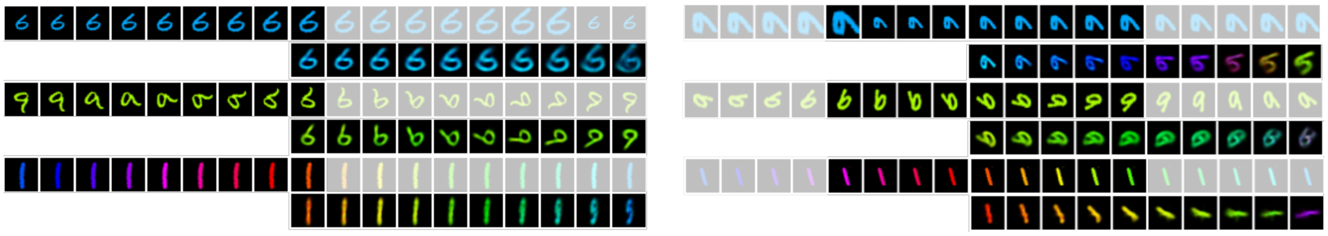

Figure 2. Forward predicted trajectories from the Predictive Coding TVAE (left) and the original TVAE (right). The images in the top row show the true input transformation, with greyed out images being held out. The lower row then shows the reconstruction, constructed by starting at $\mathbf{t}_0$, and progressively rolling the capsules forward to decode the remainder of the sequence. We see the PCTVAE is able to predict sequence transformations accurately, while the TVAE forward predictions slowly lose coherence with the input sequence.

The prior distribution is assumed to be a Topographic Product of Student's-t (TPoT) distribution [48, 37], i.e. $p_{\mathbf{T}_l}(\mathbf{t}_l) = \mathrm{TPoT}(\mathbf{t}_l; \nu)$, and we parameterize the conditional distribution with a flexible function approximator:

$$p_{\mathbf{X}_{l+1}|\mathbf{T}_l}(\mathbf{x}_{l+1}|\mathbf{t}_l) = p_\theta(\mathbf{x}_{l+1}|g_\theta(\mathbf{t}_l)) \tag{2}$$

The goal of training is thus to learn the parameters θ such that the marginal distribution of the model $p_\theta(\mathbf{x}_l)$ matches that of the observed data.

To allow for efficient training, we follow the construction outlined in [2], whereby we construct a TPoT random variable from simpler independant normal random variables $\mathbf{Z}_l$ and $\mathbf{U}_l$ which are amenable to variational inference:

$$\mathbf{T}_l = \frac{\mathbf{Z}_l - \mu}{\sqrt{\mathbf{W}\mathbf{U}_l^2}} \qquad \mathbf{Z}_l, \mathbf{U}_l \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3}$$

where $\mathbf{W}$ defines the chosen topology, and μ is learned.

Past Temporal Coherence As mentioned in the Section 2.2, the Topographic VAE takes advantage of the generalized framework of topographic generative models to induce structured correlations of activations over time – thereby achieving equivariance. In this work, this is achieved by making $\mathbf{T}_l$ a function of a sequence $\{\mathbf{U}_{l-\delta}\}_{\delta=0}^{L}$, and defining $\mathbf{W}$ to connect sequentially rolled copies of past $\mathbf{U}_l$:

$$\mathbf{T}_l = \frac{\mathbf{Z}_l - \mu}{\sqrt{\mathbf{W}\left[\mathbf{U}_l^2; \cdots; \mathbf{U}_{l-L}^2\right]}} \tag{4}$$

where $\left[\mathbf{U}_l^2; \cdots; \mathbf{U}_{l-L}^2\right]$ denotes vertical concatenation of the column vectors $\mathbf{U}_l$, and L can be seen as the past window size. Then, by careful definition of $\mathbf{W}$, we can achieve the 'shifting temporal coherence', defined in [2], yielding equivariant capsules. Explicitly, $\mathbf{W}$ is thus given by:

$$\mathbf{W}\left[\mathbf{U}_l^2; \cdots; \mathbf{U}_{l-L}^2\right] = \sum_{\delta=0}^{L} \mathbf{W}_\delta \mathrm{Roll}_\delta(\mathbf{U}_{l-\delta}^2) \tag{5}$$

where $\mathbf{W}_\delta$ defines a set of disjoint 'capsule' topologies for each time-step, and $\mathrm{Roll}_\delta(\mathbf{U}_{l-\delta}^2)$ denotes a cyclic permutation of δ steps along the capsule dimension (see [2] for exact implementation details).

3.2. The Predictive Coding TVAE

To train the parameters of the generative model θ, we use equation 4 to parameterize an approximate posterior for $\mathbf{t}_l$ in terms of a deterministic transformation of approximate posteriors over simpler Gaussian latent variables $\mathbf{z}_l$ and $\mathbf{u}_l$:

$$q_\phi(\mathbf{z}_l|\mathbf{x}_l) = \mathcal{N}\big(\mathbf{z}_l; \mu_\phi(\mathbf{x}_l), \sigma_\phi(\mathbf{x}_l)\mathbf{I}\big) \tag{6}$$

$$q_\gamma(\mathbf{u}_l|\mathbf{x}_l) = \mathcal{N}\big(\mathbf{u}_l; \mu_\gamma(\mathbf{x}_l), \sigma_\gamma(\mathbf{x}_l)\mathbf{I}\big) \tag{7}$$

$$\mathbf{t}_l = \frac{\mathbf{z}_l - \mu}{\sqrt{\mathbf{W}\left[\mathbf{u}_l^2; \cdots; \mathbf{u}_{l-L}^2\right]}} \tag{8}$$

Additionally, to further encourage the capsule Roll as the forward prediction operator, we integrate a capsule Roll of $\mathbf{t}_l$ by one unit as the first step of the generative model, before decoding $\mathbf{x}_{l+1}$:

$$p_\theta(\mathbf{x}_{l+1}|g_\theta(\mathbf{t}_l)) = p_\theta(\mathbf{x}_{l+1}|\hat{g}_\theta(\mathrm{Roll}_1[\mathbf{t}_l])) \tag{9}$$

We denote this model the Predictive Coding Topographic VAE (PCTVAE) and present an overview of forward prediction in Figure 1. We optimize the parameters θ, ϕ, γ (and μ) through the ELBO, summed over the sequence length S:

$$\begin{aligned} \sum_{l=1}^{S} \mathbb{E}_{Q_{\phi,\gamma}(\mathbf{z}_l,\mathbf{u}_l|\{\mathbf{x}\})} \Big(& \log p_\theta(\mathbf{x}_{l+1}|\hat{g}_\theta(\mathrm{Roll}_1[\mathbf{t}_l])) \\ & - D_{KL}[q_\phi(\mathbf{z}_l|\mathbf{x}_l)||p_{\mathbf{Z}}(\mathbf{z}_l)] \\ & - D_{KL}[q_\gamma(\mathbf{u}_l|\mathbf{x}_l)||p_{\mathbf{U}}(\mathbf{u}_l)]\Big) \end{aligned} \tag{10}$$

where $Q_{\phi,\gamma}(\mathbf{z}_l, \mathbf{u}_l|\{\mathbf{x}\}) = q_\phi(\mathbf{z}_l|\mathbf{x}_l) \prod_{\delta=0}^{L} q_\gamma(\mathbf{u}_{l-\delta}|\mathbf{x}_{l-\delta})$. The fundamental differences of this model with the TVAE are that this model is trained to to maximize the likelihood of *future* inputs through the Roll operation present in the ELBO, and that the construction of $\mathbf{t}_l$ is now only a function of past inputs. As we will demonstrate in the next section, these extensions yields significant improvements to sequence modeling, while simultaneously increasing flexibility by allowing for online training and inference.

4. Experiments

In this section we measure the performance of our model, compared with non-predictive coding baselines, on the transforming color MNIST dataset from [2]. The dataset is composed of MNIST digits [31] sequentially transformed by one of three randomly chosen transformations: spatial rotation, rotation in color (hue) space, or scaling. For each training example, the starting pose (color, angle, scale) is randomly set, and a cyclic sequence of 18 examples is generated according to the chosen transform. The same model architecture as [2] (a 3-layer MLP with ReLU activations) is used for all encoders and decoders of all models presented. For topographic models, the latent space is structured as 18 1-dimensional circular capsules, each of 18 dimensions. Further training details can be found at `https://github.com/anoniccv2021/PCTVAE`.

4.1. Forward Prediction Likelihood

To quantitatively measure the ability of the PCTVAE to predictively model sequences, we train the model to maximize Equation 10 with stochastic gradient descent, and measure the likelihood of held-out test sequences, with only partial sequences as input. Explicitly, for the both the TVAE and PCTVAE, a window size of 9 observations are provided as input and used to generate a capsule representation $\mathbf{t}_0$. The likelihood of the remaining 9 sequence elements is then measured by sequentially rolling the capsule activations forward, and measuring $p_\theta(\mathbf{x}_{\delta_t}|g_\theta(\mathrm{Roll}_{\delta_t}(\mathbf{t_0})))$ for $\delta_t \in \{0, ..., 9\}$. The final reported likelihood values are computed by importance sampling with 10 samples. In Table 4.1 we report the average log-likelihood over this forward predicted sequence for both the original TVAE and PCTVAE, in addition to the log-likelihood at $\delta_t = 0$ (no forward prediction) with a standard VAE. We see the PCTVAE achieves a significantly lower average negative likelihood in the forward prediction task, while maintaining a similar level of approximate equivariance as measured by the equivariance error $\mathcal{E}_{eq}$ (see [2] for a definition). We omit the baseline VAE for the sequence likelihood measurements since it has no defined forward prediction operation.

In Figure 3, we plot the likelihood of future sequence elemets as a function of the forward time offset δ_t. As can be seen, the TVAE model has a marginally higher likelihood for $\delta_t = 0$, but its forward predictive performance rapidly deteriorates as the capsule is rolled forward. Conversely, the PCTVAE is observed to obtain consistently high likelihoods on forward prediction up to 8 steps into the future of the sequence, implying it has learned to capture the transformation sequence structure more accurately. Interestingly, despite the TVAE actually being provided with an input window extending to $\delta_t \leq 4$ (as seen in Figure 2 right), the PCTVAE yields significantly higher likelihoods even for these immediate-future observations.

	NLL @ $\delta_t = 0$	NLL Avg. Seq.	$\mathcal{E}_{eq}$ Avg. Seq.
VAE	190 ± 1	N/A	13274 ± 0
TVAE	$\mathbf{187} \pm 1$	452 ± 16	2122 ± 21
PCTVAE	207 ± 1	$\mathbf{232} \pm 1$	2201 ± 9

Table 1. Neg. log-likelihood (NLL in nats) without forward prediction ($\delta_t = 0$), NLL averaged over the forward predicted sequence, and equivariance error $\mathcal{E}_{eq}$ for a non-topographic VAE, TVAE, and PCTVAE. The PCTVAE achieves the lowest average NLL over the forward predicted sequence while also maintaining low equivariance error. Mean $\pm$ std. over 3 random initalizations.

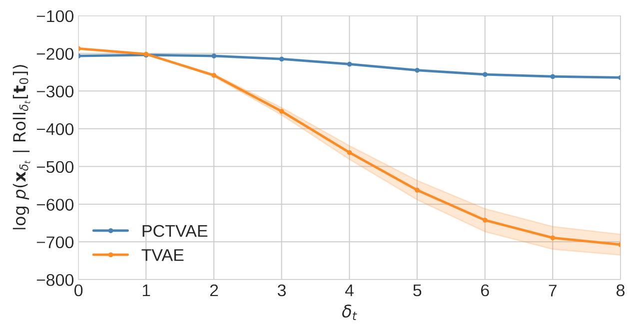

Figure 3. Forward prediction log-likelihood vs. future time offset δ_t. We see that the PCTVAE has consistently high likelihood for sequence elements into the future whereas the likelihood of the TVAE model drops off rapidly. Shading denotes $\pm$ 1 std.

4.2. Sequence Generation

As a qualitative evaluation of the PCTVAE's sequence modeling capacity, we show forward predicted sequences generated by both models in Figure 2. The top row shows the input sequence with grey images held out, and the lower row shows the forward predicted sequence, generated by sequentially rolling the representation $\mathbf{t}_0$ forward, and decoding at each step. As can be seen, the PCTVAE (left) appears to generate sequences which are more coherent with the provided input sequence, while the TVAE (right) is observed to quickly diverge from the true transformation, in agreement with likelihood values of Figure 3.

5. Discussion

In this paper we have proposed an extension of the Topographic VAE to the framework of predictive coding, and have demonstrated an improved ability to model the immediate future both qualitatively and quantitatively. This work is inherently preliminary and limited by the fact that the model is only tested on a single artificial dataset. In future work, we intend to explore the ability of such a model to learn more realistic transformations from natural data, such as from the Natural Sprites dataset [30], and additionally further investigate the downstream computational benefits gained from the learned equivariant capsule representation.

References

[1] Arjen Alink, Caspar M. Schwiedrzik, Axel Kohler, Wolf Singer, and Lars Muckli. Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, 30(8):2960–2966, 2010.

[2] Anonymous Authors. Topographic vaes learn equivariant capsules. *Advances in Neural Information Processing Systems*, 2021. In Submission. Supplied as additional material `https://github.com/anoniccv2021/PCTVAE/blob/master/TVAE.pdf`.

[3] Erik J Bekkers, Maxime W Lafarge, Mitko Veta, Koen AJ Eppenhof, Josien PW Pluim, and Remco Duits. Roto-translation covariant convolutional networks for medical image analysis, 2018.

[4] Anthony J. Bell and Terrence J. Sejnowski. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. *Neural Computation*, 7(6):1129–1159, 11 1995.

[5] Gregory W. Benton, Marc Finzi, Pavel Izmailov, and Andrew Gordon Wilson. Learning invariances in neural networks. *CoRR*, abs/2010.11882, 2020.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

[8] Yu-An Chung and James Glass. Generative pre-training for speech with autoregressive predictive coding, 2020.

[9] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

[10] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016.

[11] Taco Cohen and M. Welling. Steerable cnns. *ArXiv*, abs/1612.08498, 2017.

[12] Hanneke E. M. den Ouden, Jean Daunizeau, Jonathan Roiser, Karl J. Friston, and Klaas E. Stephan. Striatal prediction error modulates cortical coupling. *Journal of Neuroscience*, 30(9):3210–3219, 2010.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[14] Nichita Diaconu and Daniel E. Worrall. Learning to convolve: A generalized weight-tying approach. *CoRR*, abs/1905.04663, 2019.

[15] Tobias Egner, Jim Monti, and Christopher Summerfield. Expectation and surprise determine neural population responses in the ventral visual stream. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 30:16601–8, 12 2010.

[16] P. Elias. Predictive coding-i. *IRE Trans. Inf. Theory*, 1:16–24, 1955.

[17] Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to lie groups on arbitrary continuous data. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3165–3176. PMLR, 13–18 Jul 2020.

[18] Marc Finzi, M. Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. *ArXiv*, abs/2104.09459, 2021.

[19] Karl Friston. A theory of cortical responses. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 360:815–36, 05 2005.

[20] Peter Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 06 1991.

[21] David Ha and Jürgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018.

[22] David H. Hubel and Torsten N. Wiesel. Sequence regularity and geometry of orientation columns in the monkey striate cortex. *Journal of Comparative Neurology*, 158(3):267–293.

[23] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[24] Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

[25] Aapo Hyvärinen, Patrik O Hoyer, and Mika Inki. Topographic independent component analysis. *Neural computation*, 13(7):1527–1558, 2001.

[26] A. Hyvärinen, J. Hurri, and Jaakko J. Väyrynen. A unifying framework for natural image statistics: spatiotemporal activity bubbles. *Neurocomputing*, 58-60:801–806, 2004.

[27] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[28] Georg B. Keller and Thomas D. Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435, 2018.

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

[30] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding, 2021.

[31] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[32] Pengpeng Liu, Michael Lyu, Irwin King, and Jia Xu. Selflow: Self-supervised learning of optical flow. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[33] William Lotter, Gabriel Kreiman, and David Cox. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception, 2018.

[34] Sindy Löwe, Peter O'Connor, and Bastiaan S. Veeling. Putting an end to end-to-end: Gradient-isolated learning of representations, 2020.

[35] Joseph Marino. Predictive coding, variational autoencoders, and biological connections, 2020.

[36] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[37] Simon Osindero, Max Welling, and Geoffrey E. Hinton. Topographic Product Models Applied to Natural Scene Statistics. *Neural Computation*, 18(2):381–414, 02 2006.

[38] Rajesh Rao and Dana Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2:79–87, 02 1999.

[39] Siamak Ravanbakhsh, Jeff Schneider, and Barnabas Poczos. Equivariance through parameter-sharing. *ArXiv*, abs/1702.08389, 2017.

[40] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *ArXiv*, abs/1401.4082, 2014.

[41] Victor Garcia Satorras, Emiel Hoogeboom, Fabian B. Fuchs, Ingmar Posner, and Max Welling. E(n) equivariant normalizing flows, 2021.

[42] Victor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. *CoRR*, abs/2102.09844, 2021.

[43] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19:1022–38, 05 2007.

[44] Richard Turner and Maneesh Sahani. A maximum-likelihood interpretation for slow feature analysis. *Neural computation*, 19(4):1022–1038, 2007.

[45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[46] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. *CoRR*, abs/1806.03962, 2018.

[47] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *ArXiv*, abs/1807.02547, 2018.

[48] Max Welling, Simon Osindero, and Geoffrey E Hinton. Learning sparse topographic representations with products of student-t distributions. In *Advances in neural information processing systems*, pages 1383–1390, 2003.

[49] Laurenz Wiskott and Terrence J Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural computation*, 14(4):715–770, 2002.

[50] Daniel Worrall and Max Welling. Deep scale-spaces: Equivariance over scale. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[51] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: Deep translation and rotation equivariance. *ArXiv*, abs/1612.04642, 2017.