Predictive Coding, Variational Autoencoders, and Biological Connections

Anonymous Author(s)
Affiliation
Address
email

Abstract

Predictive coding, within theoretical neuroscience, and variational autoencoders, within machine learning, both involve latent Gaussian models and variational inference. We outline connections and contrasts between these areas, then highlight two frontiers for cross-pollination: backpropagation and normalizing flows.

1 Introduction

Predictive coding [35, 12] and variational autoencoders (VAEs) [21, 37] both frame perception as a generative process, modeling observations, $\mathbf{x}$, using latent variables, $\mathbf{z}$, through a probabilistic model, $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})$. Both areas also use variational inference, introducing an approximate posterior, $q(\mathbf{z}|\mathbf{x})$, to infer $\mathbf{z}$ and learn the model parameters, θ. These similarities are the result of a common origin, with Mumford [33], Dayan et al. [7], and others formalizing earlier ideas [43]. Since their inception, these areas have developed their own techniques largely independently. We explore these relationships (see also [42, 26]) and highlight two opportunities for the transfer of ideas.

2 Background

2.1 Predictive Coding

Predictive coding is a theory of thalamocortical function, formulated with probabilistic models. Under this theory, the brain constructs a generative model of sensory inputs. Top-down projections convey predictions of lower-level activity, while bottom-up projections convert the prediction error at each level into an updated perceptual state estimate. Such models are often formulated with Gaussian distributions, with analytical non-linear (e.g. polynomial) functions parameterizing the generative mappings [35, 12]. Variational inference is performed using gradient-based optimization on the distribution parameters of $q(\mathbf{z}|\mathbf{x})$, where the gradient is a linear combination of (prediction) errors for each variable. Learning can also be performed using gradient-based optimization.

2.2 Variational Autoencoders

VAEs are a class of Bayesian machine learning models, combining probabilistic models with deep neural networks. They consist of an *encoder* network (with parameters ϕ), parameterizing $q_\phi(\mathbf{z}|\mathbf{x})$, and a *decoder* network, parameterizing $p_\theta(\mathbf{x}|\mathbf{z})$. Thus, rather than performing gradient-based inference, VAEs *amortize* inference optimization with a learned network [14], improving the computational efficiency. The gradients for the inference network parameters are obtained by reparameterizing stochastic samples, $\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})$ [21, 37], however, other gradient estimators are also possible [45, 32]. Learning the parameters, θ and ϕ, of the deep networks is performed using gradient-based optimization, with gradients obtained using backpropagation [39].

Submitted to the Real Neurons & Hidden Units Workshop @ NeurIPS 2019. Do not distribute.

3 Connections & Contrasts

Model Predictive coding tends to use analytical functions to parameterize conditional mappings [12], whereas VAEs use deep networks. If these functions are analogous to neural projections, then deep networks would correspond to dendrites. In sequential models, predictive coding models temporal changes explicitly [13], whereas VAEs typically use recurrent networks, e.g. [6]. Predictive coding implements prediction covariance matrices with lateral connections, which are optimized as a form of "attention" [10]. VAEs have a similar mechanism, called normalizing flows [36] (Section 4).

Inference Predictive coding and VAEs both use variational inference, often setting $q(\mathbf{z}|\mathbf{x})$ as Gaussian. Predictive coding uses errors ($\boldsymbol{\varepsilon}_\mathbf{x}$ and $\boldsymbol{\varepsilon}_\mathbf{z}$) to perform gradient-based inference (Fig. 1), whereas VAEs use amortized inference, encoding $\mathbf{x}$. However, amortized inference can also be performed by iteratively encoding errors [29, 28]. This resembles predictive coding, but learns the inference model weights.

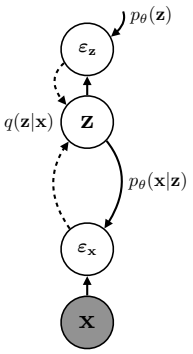


Figure 1: **Model & Inference**. Basic computation graph for performing variational inference.

4 Frontiers

Backpropagation & Learning The biological plausibility of backpropagation [39] remains an open question. Backprop requires non-local information, whereas biology relies on local learning rules [17, 30, 5]. Yet, biologically-plausible formulations of backprop have been proposed [40, 23, 47, 18, 25], with recent formulations in generative models [4, 24, 2] and predictive coding [44]. In these set-ups, errors at each latent level provide a local signal to drive learning of inference and generative weights. However, these formulations place biological neurons in correspondence with artificial neurons, while we have drawn this correspondence with deep networks. This suggests a larger role for 1) non-linear dendritic computation [31] and segmented dendrites [3, 16], as well as 2) backpropagating action potentials within neurons (Fig. 2a) [46, 22]. Thus, backprop may be analogous to learning *within* particular neuron types, while signaling across synapses may be handled via alternative techniques [45, 37, 21].

Normalizing Flows We often consider parametric distributions, as they enable efficient evaluation and sampling. However, simple distributions are limiting. *Normalizing flows* (NFs) [38, 8, 36] provide more complexity while maintaining tractable evaluation and sampling. They consist of a tractable *base* distribution and one or more invertible *transforms*. With the base distribution as $p_\theta(\mathbf{u})$ and the transforms as $\mathbf{v} = f_\theta(\mathbf{u})$, the probability $p_\theta(\mathbf{v})$ is given by the *change of variables* formula:

$$p_\theta(\mathbf{v}) = p_\theta(\mathbf{u}) \left|\det\left(d\mathbf{v}/d\mathbf{u}\right)\right|^{-1}, \tag{1}$$

where $\det(\cdot)$ denotes the determinant and $|\cdot|$ denotes absolute value. The determinant term is the local scaling of space when moving from $\mathbf{u}$ to $\mathbf{v}$. A popular family of transforms is that of autoregressive affine transforms [20, 34]. One example is given by

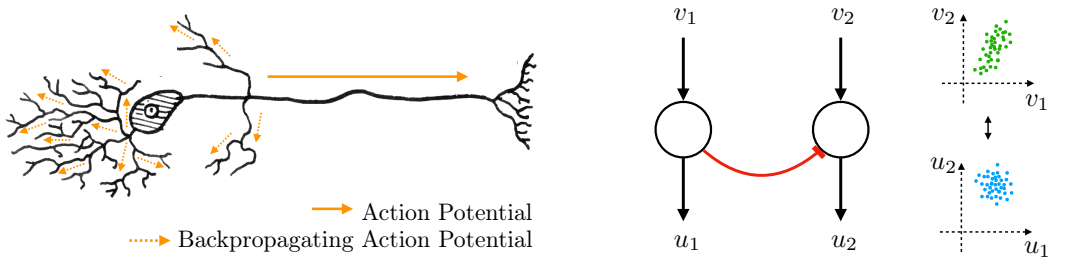
$$v_i = \alpha_\theta(\mathbf{v}_{<i}) + \beta_\theta(\mathbf{v}_{<i}) \cdot u_i, \tag{2}$$

where v_i is the ith dimension of $\mathbf{v}$ and α_θ and β_θ are functions. The inverse transform (Fig. 2b) is

$$u_i = (v_i - \alpha_\theta(\mathbf{v}_{<i}))/\beta_\theta(\mathbf{v}_{<i}), \tag{3}$$

a normalization (whitening) operation. Thus, we can sample from complex distributions by starting with simple distributions and applying local affine transforms. Conversely, we can evaluate inputs from complex distributions by applying normalization transforms, then evaluating in a simpler space.

Local lateral inhibition is ubiquitous in neural systems, thought to implement normalization [11]. These circuits give rise to decorrelation in retina [15], LGN [9], and cortex [19], as well as correlation in central pattern generators [27]. NFs may describe these circuits and help justify design choices in predictive coding. Evaluating flow-based conditional likelihoods [1] involves whitening the observations, as in [35]. When applied across time, NFs can resemble temporal derivatives [13]. Likewise, a prior covariance with lateral weights [12] corresponds to a NF with linear dependencies [20]. NFs have also been explored in the context of action [41], providing correlated motor outputs.



(a) Backpropagation within Neurons

(b) Normalizing Flow via Lateral Inhibition

Figure 2: **Frontiers**. **(a)** Backpropagation may be more analogous to backpropagating action potentials *within* neurons rather than across networks of neurons. **(b)** Affine normalizing flows can be implemented with lateral inhibitory connections. In the figure, a correlated vector, $\mathbf{v}$, is inversely transformed into a decorrelated vector, $\mathbf{u}$, simplifying the evaluation of predictions.

References

[1] S. Agrawal and A. Dukkipati. Deep variational inference without pixel-wise reconstruction. *arXiv preprint arXiv:1611.05209*, 2016.

[2] S. Bartunov, A. Santoro, B. Richards, L. Marris, G. E. Hinton, and T. Lillicrap. Assessing the scalability of biologically-motivated deep learning algorithms and architectures. In *Advances in Neural Information Processing Systems*, 2018.

[3] J. M. Bekkers. Pyramidal neurons. *Current Biology*, 2011.

[4] Y. Bengio. How auto-encoders could provide credit assignment in deep networks via target propagation. *arXiv preprint arXiv:1407.7906*, 2014.

[5] G.-q. Bi and M.-m. Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of neuroscience*, 1998.

[6] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, 2015.

[7] P. Dayan, G. E. Hinton, R. M. Neal, and R. S. Zemel. The helmholtz machine. *Neural computation*, 1995.

[8] L. Dinh, D. Krueger, and Y. Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[9] D. W. Dong and J. J. Atick. Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 1995.

[10] H. Feldman and K. J. Friston. Attention, uncertainty, and free-energy. *Frontiers in human neuroscience*, 2010.

[11] P. Földiak. Forming sparse representations by local anti-hebbian learning. *Biological cybernetics*, 1990.

[12] K. Friston. A theory of cortical responses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 2005.

[13] K. Friston. Hierarchical models in the brain. *PLoS computational biology*, 2008.

[14] S. Gershman and N. Goodman. Amortized inference in probabilistic reasoning. In *Cognitive Science Society*, 2014.

[15] D. J. Graham, D. M. Chandler, and D. J. Field. Can the theory of "whitening" explain the center-surround properties of retinal ganglion cell receptive fields? *Vision research*, 46(18):2901–2913, 2006.

[16] J. Guergiuev, T. P. Lillicrap, and B. A. Richards. Biologically feasible deep learning with segregated dendrites. *arXiv preprint arXiv:1610.00161*, 2016.

[17] D. O. Hebb. The organization of behavior; a neuropsycholocigal theory. *A Wiley Book in Clinical Psychology.*, 1949.

[18] G. E. Hinton. How to do backpropagation in a brain. *NeurIPS Deep Learning Workshop*, 2007.

[19] P. D. King, J. Zylberberg, and M. R. DeWeese. Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of v1. *Journal of Neuroscience*, 2013.

[20] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information*

processing systems, 2016.

[21] D. P. Kingma and M. Welling. Stochastic gradient vb and the variational auto-encoder. In *Proceedings of the International Conference on Learning Representations*, 2014.

[22] K. P. Körding and P. König. A learning rule for dynamic recruitment and decorrelation. *Neural Networks*, 2000.

[23] K. P. Körding and P. König. Supervised and unsupervised learning with two sites of synaptic integration. *Journal of computational neuroscience*, 2001.

[24] D.-H. Lee, S. Zhang, A. Fischer, and Y. Bengio. Difference target propagation. In *Joint european conference on machine learning and knowledge discovery in databases*, 2015.

[25] T. P. Lillicrap, D. Cownden, D. B. Tweed, and C. J. Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature communications*, 2016.

[26] W. Lotter, G. Kreiman, and D. Cox. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. *arXiv preprint arXiv:1805.10734*, 2018.

[27] E. Marder and D. Bucher. Central pattern generators and the control of rhythmic movements. *Current biology*, 2001.

[28] J. Marino, M. Cvitkovic, and Y. Yue. A general method for amortizing variational filtering. In *Advances in Neural Information Processing Systems*, 2018.

[29] J. Marino, Y. Yue, and S. Mandt. Iterative amortized inference. In *International Conference on Machine Learning*, 2018.

[30] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 1997.

[31] B. W. Mel. The clusteron: toward a simple abstraction for a complex neuron. In *Advances in neural information processing systems*, 1992.

[32] A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, 2014.

[33] D. Mumford. On the computational architecture of the neocortex. *Biological cybernetics*, 1992.

[34] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, 2017.

[35] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 1999.

[36] D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

[37] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, 2014.

[38] O. Rippel and R. P. Adams. High-dimensional probability estimation with deep density models. *arXiv preprint arXiv:1302.5125*, 2013.

[39] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 1986.

[40] D. G. Stork. Is backpropagation biologically plausible. In *International Joint Conference on Neural Networks*, 1989.

[41] Y. Tang and S. Agrawal. Boosting trust region policy optimization by normalizing flows policy. *arXiv preprint arXiv:1809.10326*, 2018.

[42] G. van den Broeke. What auto-encoders could learn from brains. Master's thesis, 2016.

[43] H. Von Helmholtz. *Handbuch der physiologischen Optik*. 1867.

[44] J. C. Whittington and R. Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural computation*, 2017.

[45] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992.

[46] S. R. Williams and G. J. Stuart. Backpropagation of physiological spike trains in neocortical pyramidal neurons: implications for temporal coding in dendrites. *Journal of Neuroscience*, 2000.

[47] X. Xie and H. S. Seung. Equivalence of backpropagation and contrastive hebbian learning in a layered network. *Neural computation*, 2003.