Predictive Coding Notes

Adam Kosiorek

September 12, 2017

1 Predictive Coding

Modern sequential predictive models tend to update its hidden state at every time-step. It can be argued, however, that if a model is able to predict the world perfectly, it should not update its state. On the contrary, if a perfect prediction is available, the model should also be capable of evolving its hidden state so as to reflect the change of the state of the world w. r. t. the prediction. Predictive coding formalises this behaviour by using only prediction errors to update the world.

The idea dates back to the Kalman filter and beyond . Recent advances in neural networks make it possible to frame the predictive coding problem as an auto-regressive neural network. Indeed, it has been addressed in a number of papers (Canziani and Culurciello, 2017; Lotter, Kreiman, and Cox, 2016). These models use the mean-squared error between the prediction and the corresponding input to update their hidden state. It leads to an ill-posed problem since there are many futures possible. The prediction, which is a maximum-likelihood estimate of what might happen, is only a single instantiation thereof. It would be theoretically more sound to normalise prediction errors by the covariance matrix of errors, an approach adopted by Kalman filtering. Friston, 2009 argues this is also the approach taken by the brain, and it explains a lot of phenomena happening there. He suggests that the computational architecture of the brain forms a hierarchical system, where each layer constantly predicts the output of the lower levels of hierarchy in a fully Bayesian fashion. It gives rise to surprise, which is the negative log-likelihood of the inputs under the predictive distribution of the model. Friston argues that surprise leads to error normalisation w. r. t. prediction uncertainty, which can be interpreted as attention.

Add reference(s).

In for neural networks, predictive coding has the advantage of minimising the covariance shift problem. If we assume that the expected value of predictions is the same as the expected value of inputs, then computing the surprise of inputs (under Gaussian predictive distribution) can have zero mean and

unit variance. To be determined experimentally. It's similar to Klambauer et al., 2017, but the name "self-normalizing" is somewhat more appropriate here.

Free-energy minimisation problems can be often written as variational inference, which is also true in this case. In what follows, we specify the model as a variational autoencoder, show its computational graph and derive the variational lower bound.

2 Variational Distribution

Let $\mathbf{x}_{1:T}$ be a timeseries of observations e. g., image sequence or a series of joint angles. The probability distribution over input sequences $p(\mathbf{x}_{1:T})$ can be expressed by the product law as $p(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p(\mathbf{x}_t \mid \mathbf{x}_{1:t-1})$ and is in general intractable. We introduce a hierarchy of latent variables $\mathbf{y}_{1:T}$, where $\mathbf{y}_t = \{\mathbf{z}_{1:T}^l\}_{l=1}^{L}$, such that the joint probability distribution $p_\theta(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}) = p_\theta(\mathbf{x}_{1:T} \mid \mathbf{y}_{1:T}) p_\theta(\mathbf{y}_{1:T})$ lives in the space of parametric probability distributions parametrised by parameters θ and is tractable. We further assume the following factorisation:

prior The prior distribution is a dynamical model of the latent space. It specifies the probability distribution over the initial states $\mathbf{y}_0$ and the transition probabilities between states. It may or may not be stationary.

$$p_\theta(\mathbf{y}_{0:T-1}) = p_\theta(\mathbf{y}_0) \prod_{t=1}^{T-1} p_\theta(\mathbf{y}_t \mid \mathbf{y}_{t-1}) \tag{1}$$

generating distribution The generating distribution decodes the samples from the latent space back into the observation space. It uses the learnt dynamics model internally to propagate the state forward.

$$\begin{aligned} p_\theta(\mathbf{x}_{1:T} \mid \mathbf{y}_{1:T}) &= \int p_\theta(\mathbf{y}_0) \prod_{t=0}^{T} p_\theta(\mathbf{x}_{t+1} \mid \mathbf{y}_t) \, d\mathbf{y}_0 \\ &= \int p_\theta(\mathbf{y}_0) \prod_{t=0}^{T} \int p_\theta(\mathbf{x}_{t+1} \mid \widehat{\mathbf{y}}_{t+1}) p_\theta(\widehat{\mathbf{y}}_{t+1} \mid \mathbf{y}_t) \, d\widehat{\mathbf{y}}_{t+1} \, d\mathbf{y}_0 \\ &= p_\theta(\mathbf{x}_{1:T} \mid \mathbf{y}_{1:T-1}) \end{aligned} \tag{2}$$

posterior distribution The posterior distribution is intractable, although its general form reads as follows:

$$p(\mathbf{y}_{1:T-1} \mid \mathbf{x}_{1:T-1}) = \int \prod_{t=1}^{T-1} p(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{y}_{t-1}) p_\theta(\mathbf{y}_0) \, d\mathbf{y}_0 \tag{3}$$

approximate posterior We approximate the posterior with a tractable parametric distribution q_ϕ by assuming that the hidden state $\mathbf{y}_t$ is Markovian.

$$q_\phi(\mathbf{y}_{1:T-1} \mid \mathbf{x}_{1:T-1}, \mathbf{y}_0) = \prod_{t=1}^{T-1} q_\phi(\mathbf{y}_t \mid \mathbf{x}_t, \mathbf{y}_{t-1}) \tag{4}$$

Moreover, since we learn the latent dynamics model, the generating model predicts the input at the next time-step as opposed to standard VAE practice, where it merely reconstructs the input from the current time-step.

2.1 Multimodal Distributions for Predictive Coding

The expressive power of VAEs is often constrained by too simple a form of the approximate posterior distribution and increasing the complexity of the approximate posterior typically leads to better results (Karl et al., 2016; Kingma et al., 2016; Rezende and Mohamed, 2015). Complexity-increasing methods used so far typically lead to increased correlation between latent variables and/or multimodal distribution. While correlated latent variables may be desirable, they are against the principle coding principle , since the latent variables should be orthogonal in order to maximum the maximum amount of information possible. Multimodal distributions don't have this caveat. In case of predictive-coding, however, they lead to scenarios where the computed surprise on the input is ambiguous.

Add reference(s).
- is that even true?

Assume a bimodal distribution predictive $p(\widehat{z}_{t+1} \mid z_t) = a\mathcal{N}(-1, 1) + b\mathcal{N}(1, 1)$ and an input z_{t+1}. The value of the surprise is $s_{t+1} = -\log p(\widehat{z}_{t+1} \mid z_t)|_{\widehat{z}_{t+1}=z_{t+1}}$. Even in case of a unimodal distribution, there exist an infinite number of points giving rise to the same value of surprise, but at least that value exclusively determines the distance of the input from that mode. In case of a multimodal distribution, the value of surprise indicates the distance, but we don't know from which mode. That leads to encoding ambiguity and is disturbing. This in itself could be an argument of why all distributions used in predictive coding should be unimodal. Is that the case in the brain?

3 ELBO

It's looking pretty straightforward, but to be derived...

4 Hierarchical Surprise Minimisation

Let $\mathbf{z}_t^l$ be an encoding produced by layer l at time t and let $p_\theta(\widehat{\mathbf{z}}_t^l \mid \mathbf{y}_{t-1})$ be a corresponding predictive distribution. Surprise then reads as $\mathbf{s}_t = -\log p_\theta(\widehat{\mathbf{z}}_t^l \mid \mathbf{y}_{t-1})|_{\widehat{\mathbf{z}}_t^l=\mathbf{z}_t^l}$.

We would like to minimise surprise at every layer in order to force the system to decompose the state into a hierarchical representation, where higher levels in the hierarchy correspond to more abstract concepts. Naively penalizing surprise, however, might lead to a scenario where $\mathbf{z}_t^l$ converges to a single value independent of the input $\mathbf{x}_t$ and the predictive distribution becomes sharply peaked at that value. In other words, the entirety of information is channelled through the lowest layer $l = 1$, while all layers for $l > 2$ collapse to a degenerate state. The first layer does not collapse due to prediction loss in the observation space.

In the above setup inputs to the first layer (the observation $\mathbf{x}_t$) and inputs to subsequent layers (encodings $\mathbf{z}_t^l$) are treated differently w. r. t. backpropagated gradients. Namely, the backpropagated errors do not alter the inputs $\mathbf{x}$, but they have every liberty to alter the encoding $\mathbf{z}$. A simple solution is to cut the gradient paths from surprise $\mathbf{s}_t^l$ to layers $k < l$. This way, surprise minimisation acts on the parameters of the predictive distribution, without altering the parameter that lead to generating the input.

5 Architecture

We need L encoders, L decoders, L dynamic models and $L - 1$ combinators. Each of the components could in principle be either stochastic or deterministic. To contain the model within the VAE framework, we take stochastic encoders and decoders, stochastic dynamic models, which we use as latent space priors, and deterministic combinators. Computation graph is given in fig. 1.

encoder Given the previous latent state and a prediction error, it computes the parameters of the approximate posterior $q_\phi\left(\mathbf{z}_t^l \mid \mathbf{z}_{t-1}^l, \Delta\mathbf{z}_t^l\right)$.

dynamic model Computes the probability distribution over the new state estimate given a a previous state estimate, it's part of the prior as $p_\theta(\mathbf{z}_{t+1} \mid \mathbf{z}_{t-1})$.

combinator State is predicted based on more abstract representation. This module refines the prediction by using lower-level information available in the previous state estimate. It is a part of the generating model.

decoder The final part of the generating model, transforms the refined state estimate computed by the combinator and computes the probability distribution over the lower level state estimate $p_\theta\left(\mathbf{z}_{t+1}^l \mid \mathbf{z}_t^{l+1}, \widehat{\mathbf{z}}_{t+1}^{l+1}\right)$.

References

Canziani, Alfredo and Eugenio Culurciello (2017). "CortexNet: a Generic Network Family for Robust Visual Temporal Representations". In: arXiv: 1706.02735. URL: http://arxiv.org/abs/1706.02735.

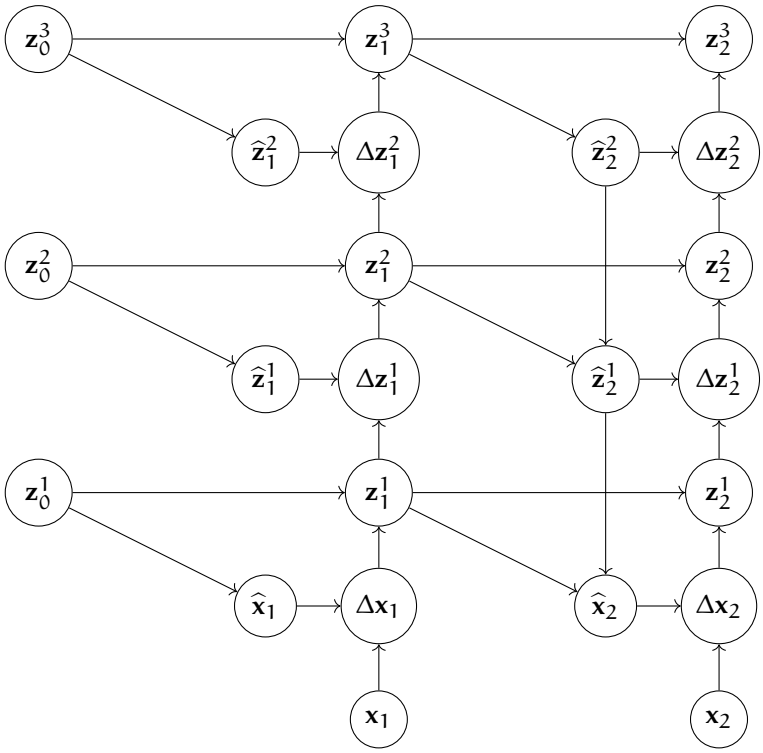

Figure 1: Computation graph of the predictive coding framework.

Friston, Karl (2009). "The free-energy principle: a rough guide to the brain?" In: *Trends in Cognitive Sciences* 13.7, pp. 293–301. ISSN: 13646613. DOI: 10.1016/j.tics.2009.04.005. URL: http://www.fil.ion.ucl.ac.uk/%7B~%7Dkarl/The%20free-energy%20principle%20-%20a%20rough%20guide%20to%20the%20brain.pdf.

Karl, Maximilian, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt (2016). "Deep Variational Bayes Filters: Unsupervised Learning of State Space Models from Raw Data". In: *Proceedings of the Intuitive Physics Workshop at NIPS 2016* i, pp. 2–5. arXiv: 1605.06432. URL: http://arxiv.org/abs/1605.06432%20https://arxiv.org/pdf/1605.06432.pdf.

Kingma, Diederik P. et al. (2016). "Improving Variational Inference with Inverse Autoregressive Flow". In: arXiv: 1606.04934. URL: http://arxiv.org/abs/1606.04934.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter (2017). "Self-Normalizing Neural Networks". In: DOI: 1706.02515. arXiv: 1706.02515. URL: http://arxiv.org/abs/1706.02515.

Lotter, William, Gabriel Kreiman, and David Cox (2016). "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning". In: arXiv: 1605.08104. URL: http://arxiv.org/abs/1605.08104.

Rezende, Danilo Jimenez and Shakir Mohamed (2015). "Variational Inference with Normalizing Flows". In: ISSN: 1938-7228. arXiv: 1505.05770. URL: http://arxiv.org/abs/1505.05770.