Precision-weighted predictive (variational autoen-) coding with structured uncertainty distributions

PyTorch implementation

Stefan Brugger

PhD candidate, CUBRIC, Cardiff University

Academic Clinical Fellow, Centre for Academic Mental Health, University of Bristol

Goals

- Aim is to scale precision-weighted predictive coding from low-dimensional toy examples to regular machine learning scale
- Generalized predictive coding approaches developed by Friston & collaborators 2007-2010: filtering (as in Kalman) schemes for VB inversion of hierarchical dynamic models used in DCM & as metaphors for perception & action in the brain

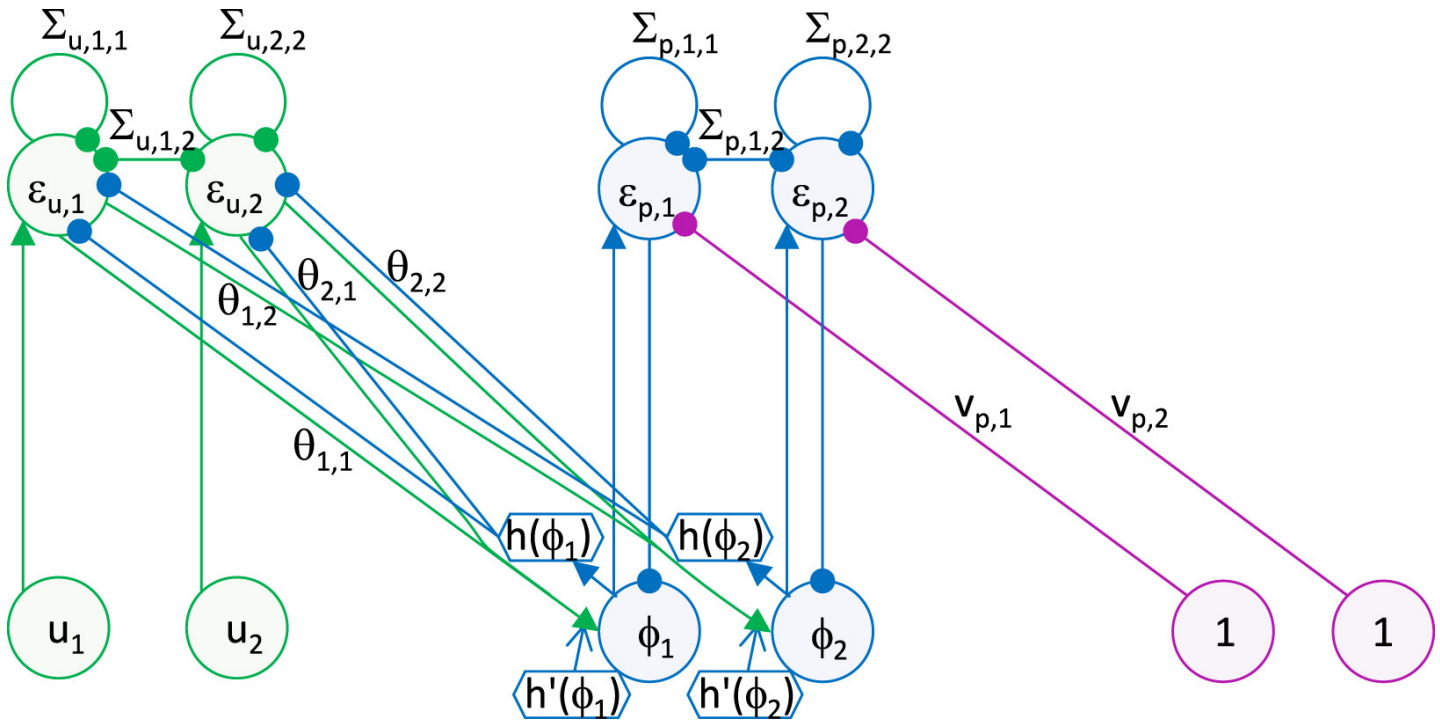
Antecedents

- Predictive coding
- Free energy principle
- Recurrent neural network
- Variational autoencoding

VAE

- Image -> Encoder -> Latent representation -> decoder -> image reconstruction
- Loss = D_{KL}(Latent, Prior) + reconstruction error
- Prior ~ N(0,I)

Precision-weighted Predictive Coding



Bogacz 2017; difference with the present model is addition of noise to descending predictions. Σ now also defines precision of gaussian noise

The current model

- Activations (expectations) ve_{i+1}
- Noise/uncertainty: covariance Σ_{i+1}
- Sample $v_{i+1} \sim N(ve_{i+1}, \Sigma_{i+1})$
- User-set # of transposed convolutions (nn.ConvTranspose2d) to predict v_i from v_{i+1}
- Prediction error $PE = v_i - CT_{i+1}(v_{i+1})$
- Level-wise loss: PE weighted by precision matrix Σ^{-1} plus uncertainty:
  - Free energy $F = 0.5\ (PE\ \Sigma^{-1}\ PE^T + \ln|\Sigma|)$
  - v_0 is the image
- Adjust v_i each iteration; after n iterations parameters of CT_n, Σ^{-1} updated. Don't need to calculate derivatives, update terms etc – PyTorch does this automatically

Inference and Learning

- Synaptic parameters (CN weights, Σ components) change at slower timescale relative to neural activations (v)
- In practice after n inference iterations, m learning iterations take place (Adam updates synaptic parameters; same loss)
- Additional layer-wise loss in form of VAE latent loss
- At lower levels, probabilistic latent spaces organized topologically (pixel-wise), in a convolutional fashion (with weight sharing of any connecting linear layers)
- Provides an alternative dimension reduction technique (cf pooling layers) that does not discard data (see Hinton's criticisms of max-pooling https://mirror2image.wordpress.com/2014/11/11/geoffrey-hinton-on-max-pooling-reddit-ama/)

Scaling covariance matrices

- Want to move beyond diagonal covariance matrix employed in most VAEs
- Full matrix: memory requirements ~ $O(n^2)$ for linear layer as Σ^{-1}
- Σ needs to be +ve definite: Cholesky decomposition
- Sparse Cholesky decomposition (Dorta et al 2018a):
  - L_{ij} is only non-zero if $i \geq j$ and i and j are neighbours in the image plane, where pixels i and j are neighbours if a patch of size f centred at i contains j
- Memory L ~ $O(nf^2)$, $f << n$
- Dorta et al 2018a covariance network:
  - Estimates L from v
  - Predictive coding approach
    - estimate L_i from v_{i+1}?

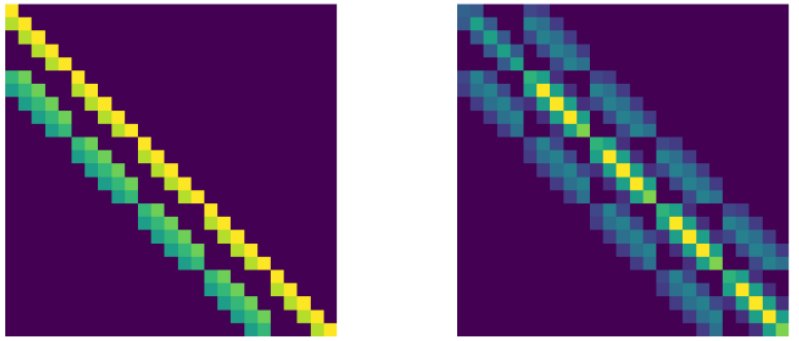


Figure 2: Left, an example of the sparsity patterns in the band-diagonal lower-triangular matrices $\mathbf{L}$, that are estimated by our model. Right, the precision matrix $\mathbf{\Lambda} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$.

Scaling further: Dorta 2018b

- Uses non-diagonal covariance matrix for VAE – relaxes independence assumption
- Improves quality of samples – correlated noise added to reconstruction

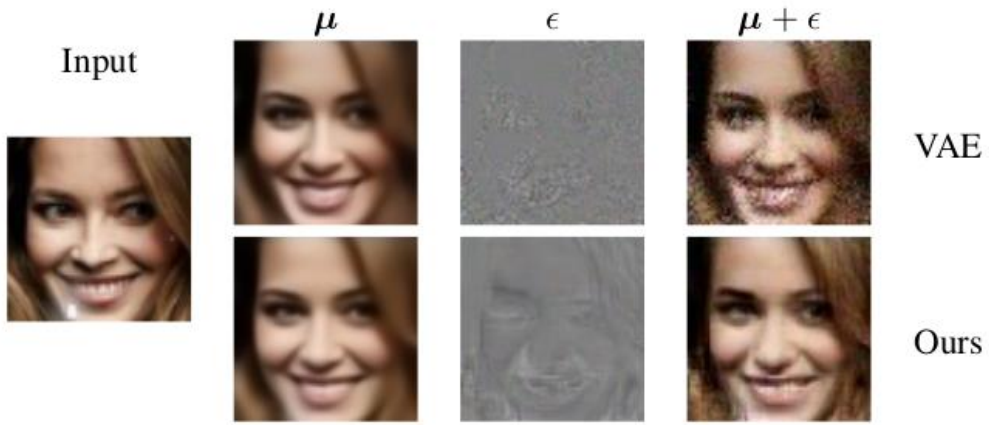


Figure 1: Given an input image, reconstructions from a VAE and our model are shown. The VAE models the output distribution as a factorized Gaussian, while our model uses a structured Gaussian likelihood. We show the means μ and a sample ϵ from the corresponding covariances. The correlated noise sample of our model better captures the structure present in natural images.

Scaling further: Dorta 2018b

- Prior on covariance matrix: needs heavy regularisation:
  - "Intuitively, little variance is desired on the predicted covariance matrix, i.e. the model should be certain about its predictions. Another issue is the prediction of spurious correlations, which are not well supported by the data."
  - Inverse Wishart and Gamma-Gaussian (diag/off-diag): "too strongly diagonal"
- Eventually used custom loss function
- Reduced dimensionality of L further with learnable basis set & weights estimated per-image
- Not dissimilar to approach used for variance components in DEM/GF

Covariances in Generalized Predictive Coding

- "The advantage of the Laplace assumption is that the conditional covariance is a simple function of the modes" ??
- VFELA p224-225 gives formula for free energy for static models in terms of causes, parameters (theta) and hyperparameters (lambda):

$$F = -\frac{1}{2}\varepsilon^T\Sigma^{-1}\varepsilon + \frac{1}{2}\ln|\Sigma^{-1}| - \frac{n}{2}\ln 2\pi \qquad -\frac{1}{2}\varepsilon^{\theta T}\Pi^\theta\varepsilon^\theta + \frac{1}{2}\ln|\Pi^\theta| + \frac{1}{2}\ln|\Sigma^\theta| \qquad -\frac{1}{2}\varepsilon^{\lambda T}\Pi^\lambda\varepsilon^\lambda + \frac{1}{2}\ln|\Pi^\lambda| + \frac{1}{2}\ln|\Sigma^\lambda|$$

- Pi = precisions = Σ^{-1}
- See also equation 3.8, generalised filtering paper – also includes extra logdet terms for 1st and 2nd order parameters
- Theta refers to convolutional layers here – not sure it is desirable to have covariance matrix over these too
  - Shridhar et al: simple method to make Bayesian CNNs with diagonal covariance – probably all that's needed?

Expected precision & second order PEs (notes)

- Complicated
- Kanai et al (2015):
  - Sum of squared PE for each covariance component sent to pulvinar (matrix cells) from cortex (layer 5) – enables expected precision estimation
  - "Neuroanatomical observations of the pulvinar suggest that for every direct connection between two cortical regions, there is a parallel, indirect pathway that goes through the pulvinar. This is called the replication principle... it seems reasonab le to hypothesize that the functional role of the pulvinar is to optimize the gain of cortical prediction errors according to their expected precision. To fulfil this role, the pulvinar needs to encode expected precision and mediate gain modulation."
  - So: expected precision components at level i (hyperparameters, pulvinar) are a function of causes at level i+1
  - "These effects are formally distinct: the first-order predictions (of lower expectations) have a negative (driving) effect on the prediction errors, whereas the second-order predictions (of their precision) have a positive (modulatory) effect.

  - PwPE = $(v_1 - g(v_2)).\Sigma^{-1} = (v_1 - g(v_2)).\pi(v_2)$
  - $F = \ldots + (v_1 - g(v_2)).\pi(v_2)(v_1 - g(v_2))^T + \ldots$

  - $\pi(v_i)$ and it's hyperparameters are in pulvinar. Can see that $dF/d\pi(v_2)$ is sum of squared PEs
  - Paper uses $\pi(v_i) = \exp(-v2)$, and also a function of power of prediction errors about their mean, reflecting texture (bespoke but interesting function as looking for high fluctuations in luminance. As is the dropping of precision? Allowing precise beliefs about cause =0 to become imprecise and be influenced by ascending PEs to become cause =10 (or whatever)
  - But presumably linear dnn can learn appropriate pi functions

Second order PEs - why would these be helpful?

- Used in the (better) FEP papers to model attention effects (Feldman & Friston), identification of objects by texture (Kanai et al 2015)

Applications

- A better model of (representational) dynamics of the visual system than vanilla RNNs (Kietzmann et al 2019)
  - In particular relating to the effect of prior knowledge on image processing
  - MEG of Mooney image viewing
- In-silico 'hallucinations', with computational analogues of
  - Noisier sensory systems
  - Synaptic gain impairments
  - Greater high-level expected precision

Open questions/problems

- Dynamic models with states (eg video prediction) - Maell Cullen has version that does this using LSTM but no precision stuff
  - Fountas et al arXiv (2020): deep active inference agent includes modelling state transitions (does action but more basic 'sensory' system)
- Attention (in ML sense) based probabilistic encodings at higher levels?
  - i.e. not just based on conv grid but on inferred objects
- Simulate magno- & parvo-cellular streams:
  - Magnocellular pathway receives low-pass filtered version of image
  - Rapid processing provides prior for top of high-frequency stream
    - See Kveraga 2007 J Neurosci
- Is this how 'hierarchical' VAEs work?

Relevant recent developments in Active Inference

- Fountas et al "Deep active inference agents using Monte-Carlo methods", arXiv:2006.04176.
- Cullen et al. **A Meta-Bayesian Model of Intentional Visual Search (Arxiv)**

Relevant recent developments in DL models of vision

- Zhuang.. Yamins - Unsupervised Neural Network Models of the Ventral Visual Stream. Arxiv
- Dapello.. DiCarlo - Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations (bioArxiv)
- Instance-level contrastive learning yields human brain-like representation without category-supervision (Konkle; bioArxiv)
- **Learning to Combine Top-Down and Bottom-Up Signals in Recurrent Neural Networks with Attention over Modules (arxiv)**
-

References

- Bogacz R. A Tutorial on the Free-Energy Framework for Modelling Perception and Learning. J Math Psychol. 2017 Feb;76(Pt B):198-211.
- Dorta, Vicente S, Agapito L, Campbell N, Simpson I. Structured Uncertainty Prediction Networks. Arxiv 2018a
- Dorta, Vicente S, Agapito L, Campbell N, Simpson I. Training VAEs under Structured Residuals. Arxiv 2018b
- Zafeirios Fountas, Noor Sajid, Pedro A.M. Mediano, Karl Friston. Deep active inference agents using Monte-Carlo methods. Arxiv 2020
- Kestutis Kveraga, Jasmine Boshyan and Moshe Bar. Magnocellular Projections as the Trigger of Top-Down Facilitation in Recognition. J Neurosci (2007) 27 (48) 13232-13240
- Tim C. Kietzmann, Courtney J. Spoerer, Lynn K. A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. Recurrence is required to capture the representational dynamics of the human visual system. PNAS (2019) 116 (43) 21854-21863
- Ryota Kanai, Yutaka Komura, Stewart Shipp and Karl Friston. Cerebral hierarchies: predictive processing, precision and the pulvinar. Philos Trans R Soc Lond B Biol Sci. 2015 May 19;370(1668):20140169.