Math of Predictive Coding

Shobhith Vadlamudi (ED21B069)

July 25, 2024

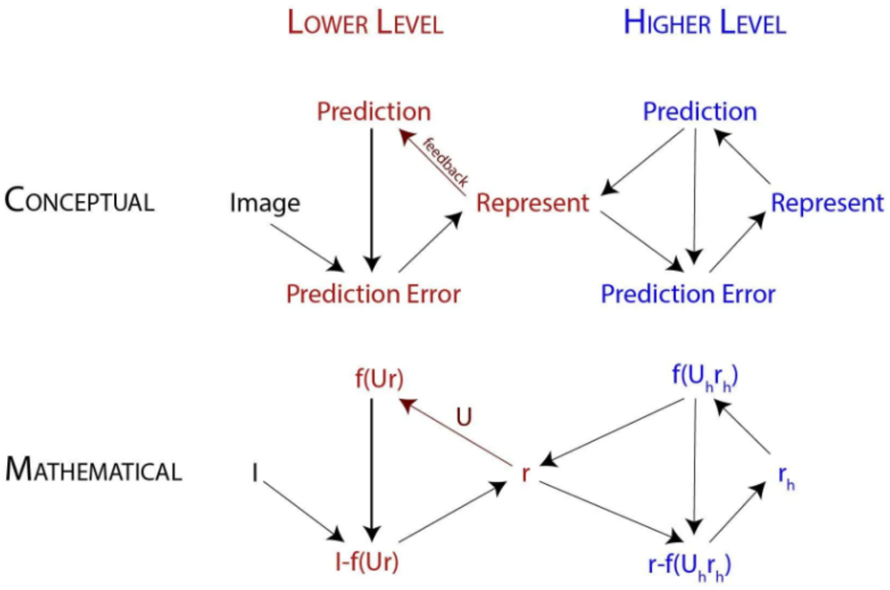

Figure 1: Hierarchical Model

1 Goal of Predictive Coding

- We need a network that takes in images and represents the image in the network's activations. We want this network to represent the causes of the image.
- The causes of the image can be thought of as the underlying patterns and features of the objects that make up the image.
- For example, let's take a banana characterized by distinctive features, such as its curved shape, color, peel texture, etc. So, if the network were trained on images of bananas, we would expect it to learn these basis features. Instead of the network memorizing that this pixel configuration represents a banana, we are trying to encode higher-level semantic information.
- We are trying to infer causes U over the dataset and an r for each image. This is what is meant by representing the image in the network's activations.

2 The Generative model

Let us call the image formed on the retina **I**. This is a vector of pixel values. The causes of retinal images are given by a matrix **U** and a vector r. The columns of **U** are a basis for the causes. We can think of r as the activities or firing rates of neurons. Another way to think about this would be the weights associated with each basis cause. We assume images are related to causes in the following way

$$I = f(Ur) + n$$

Here, n is a noise vector representing the error between the image and the inferred causes. This is the generative model because it describes how causes in the world generate retinal images.

3 Non-Hierarchial Version

- So now we have the generative model. To infer causes U over the dataset and an r for each image, we use Bayesian inference.
- Bayes theorem is given as

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{1}$$

  Where:

  - $P(A|B)$ is the posterior probability of event A given event B.
  - $P(B|A)$ is the likelihood of event B given event A.
  - $P(A)$ is the prior probability of event A.
  - $P(B)$ is the prior probability of event B.

- Here we want to maximize the posterior $p(U, r|I)$ .Using Bayes theorem we can write

$$p(U, r|I) \propto p(I|U, r) \cdot p(U) \cdot p(r) \tag{2}$$

  Where:

  - $p(U)$ and $p(r)$ are priors on the causes.

  If we find the values of the causes that maximize the right-hand side of the equation, we will maximize the left-hand side.

- We assume that n from the generative model is normally distributed with 0 mean and variance σ^2. So we see that the term $p(I|U, r)$ is a normal distribution with mean $f(U, r)$ and variance σ^2.

$$p(I|U, r) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}|I - Ur|^2\right) \tag{3}$$

- We then define the loss function by taking the negative log of the RHS.

$$L = -\log\left(p(I|U, r)p(U)p(r)\right) = -\log\left(p(I|U, r)\right) - \log\left(p(U)\right) - \log\left(p(r)\right) \tag{4}$$

  Expanding this we get

$$\begin{aligned} -\log(p(I|U, r)) &= -\log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}|I - Ur|^2\right)\right) \\ &= \frac{1}{2\sigma^2}|I - Ur|^2 + \frac{1}{2}\log(2\pi\sigma^2) \end{aligned}$$

  2nd term is a constant, so we can ignore it while optimizing.

$$L = \frac{1}{\sigma^2}|I - Ur|^2 + g(r) + h(U) \tag{5}$$

  Where the functions g and h are the negative logarithms of the priors on U and r, respectively. We used Gaussian prior distributions for both these model parameters because this was sufficient to illustrate the properties of the model. This results in $g(r) = \alpha \sum_i r_i^2$ and $h(U) = \lambda \sum_{i,j} U_{i,j}^2$, where α and λ are positive constants related to the variance of the Gaussian prior distributions.

4 Hierarchical Model

- We assumed that the images are caused by causes, but in the hierarchical model, the causes are caused by more abstract, higher-level causes U_h and r_h. So we treat the causes r, of the retinal image as if it were sensory input to a more abstract system.

  $$r = f(U_h r_h) + n_h \tag{6}$$

  Now the posterior will be $p(U, r, U_h, r_h|I)$. Applying Bayes theorem to this, we get

  $$p(U, r, U_h, r_h|I) = \frac{p(I|U, r, U_h, r_h) \cdot p(U, r, U_h, r_h)}{p(I)} \tag{7}$$

- $p(U, r, U_h, r_h)$ can be expanded as $p(r|U, U_h, r_h) \cdot p(U|U_h, r_h) \cdot p(U_h|r_h) \cdot p(r_h)$. On simplification, this becomes $p(r|U_h, r_h) \cdot p(U) \cdot p(U_h) \cdot p(r_h)$

- Using the same argument as before, we get

  $$p(I|U, r, U_h, r_h) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}|I - f(Ur)|^2\right)$$

(8)

  $$p(r|U_h, r_h) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}|r - f(U_h r_h)|^2\right)$$

(9)

- Calculating the negative logarithm, we get the loss function as follows

  $$L = \frac{1}{\sigma^2}|I - f(Ur)|^2 + \frac{1}{\sigma^2}|r - f(U_h r_h)|^2 + g(r) + h(U) + g(r) + h(U)$$

(10)

5 Minimizing L

- To minimize L, we take derivatives of L with respect to r and U and use gradient descent.

- Taking the derivative wrt r we get

  $$\frac{dL}{dr} = -\frac{2}{\sigma^2}U^T(I - f(Ur))f'(Ur)^T + \frac{2}{\sigma^2}(r - f(U_h r_h)) + g'(r) \tag{11}$$

- Applying gradient descent on r, we get

  $$r_t = r_{t-1} - \frac{k_1}{2}\frac{\partial L}{\partial r} \tag{12}$$

  $$r_t = r_{t-1} + \frac{k_1}{\sigma^2}U^T\frac{df^T}{dx_{t-1}}(I - f(Ur_{t-1})) + \frac{k_1}{\sigma^2}(f(Ur^h_{t-1}) - r_{t-1}) - \frac{k_1}{2}g'(r_{t-1}) \tag{13}$$

  where $x = Ur$

- Applying gradient descent on U, we get :

$$U_t = U_{t-1} - \frac{k_2}{2}\frac{\partial L}{\partial U} \tag{14}$$

$$U_t = \frac{k_2}{\sigma^2}\frac{df^T}{dx_{t-1}}\left(I - f(U_{t-1}r)\right)r^T - k_2\lambda U_{t-1} \tag{15}$$

  Where k1 and k2 are learning rates.