HW4 - Theory

hhp9256 - Hoang Pham

1 Theory

1.1 Energy Based Model Intuition

(a) Energy based model allow for modeling for 1-to-many or 1 to infinite continuum of y by not using single inference from input of $x \to y$, but to infer the level of dependency between (x, y) using the energy function F. Most compatible pair will have the lowest energy F: $\hat{y} = \arg\min_y \{F(x,y), (x,y) \in \mathcal{X} \times \mathcal{Y}\}$.

(b) Energy based model doesn't use softargmax to produce probability distribution, instead it produce the similarity (energy) score. To be more precise, other models produce $p(y|x)$, while the energy model produces $F_{\mathbf{W}}(x, y)$, where $\mathbf{W}$ is the learnable weights.

(c) We can use the Boltzmann-Gibbs distribution to reverse-calculate the probability $p(y|x)$:

$$p(y|x) = \frac{\exp\left(-\beta F_{\mathbf{W}}(x,y)\right)}{\int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x,y')\right)},$$

where the temperature β is a hyperparameter of the model that can be tuned.

(d) Using the Boltzmann-Gibbs distribution above, we can tune $\beta \to \infty$ to make $p(y|x)$ output argmax, while $\beta \to 0^+$ yields a smoother distribution.

(e) The loss functions are used to shape the energy function.

(f) Only pushing down energy of correct inputs will lead the energy to be flat everywhere (0). This does not help in learning as it does not penalize the incorrect input.

(g) Methods to shape the energy function:

- Contrastive method, which push down energy on correct input pairs and push up energy on incorrect input pairs.
- Architectural method, where we build the energy so that the volume of low energy regions is limited or minimized through regularization.

(h) We can use negative log-likelihood loss:

$$\ell_{\text{example}}(\mathbf{x}, \mathbf{y}, \mathbf{W}) = F_{\mathbf{W}}(x,y) + \frac{1}{\beta} \log \int_{y'} \exp\left(-\beta E_{\mathbf{W}}(\mathbf{y}', \mathbf{x})\right)$$

(i) $\hat{y} = \arg\min_y \{F(x,y)\}$ and $F(x,y) = \arg\min_z \{G(x,y,z)\}$.

1.2 Negative log-likelihood loss

(a) Similar to previous part (c):

$$p(y|x) = \frac{\exp\left(-\beta F_{\mathbf{W}}(x,y)\right)}{\int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x,y')\right)}$$

(b) With log defined as natural logarithm,

$$\begin{aligned}\ell_{\mathbf{W}}(x, y) &= -\log \frac{\exp\left(-\beta F_{\mathbf{W}}(x, y)\right)}{\int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)} \\ &= \beta F_{\mathbf{W}}(x, y) + \log \int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)\end{aligned}$$

Normalizing the loss with $\frac{1}{\beta}$, we have

$$\ell_{\mathbf{W}}(x, y) = F_{\mathbf{W}}(x, y) + \frac{1}{\beta} \log \int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right).$$

(c) We have

$$\frac{\partial \ell_{\mathbf{W}}(x, y)}{\partial \mathbf{W}} = \frac{\partial F_{\mathbf{W}}(x, y)}{\partial \mathbf{W}} + \frac{1}{\beta} \frac{\partial \left[\log \int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)\right]}{\partial \mathbf{W}}$$

Using chain rules, we have

$$\begin{aligned}\frac{\partial \left[\log \int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)\right]}{\partial \mathbf{W}} &= \frac{1}{g} \frac{\partial \int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)}{\partial \mathbf{W}} && \left(\text{where } g := g(\mathbf{W}) = \int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)\right) \\ &= \frac{1}{g} \int_{y'} \frac{\partial \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)}{\partial \mathbf{W}} \\ &= \frac{1}{g} \int_{y'} -\beta F_{\mathbf{W}}(x, y') \frac{\partial F_{\mathbf{W}}(x, y')}{\partial \mathbf{W}} \\ &= -\beta \int_{y'} \frac{F_{\mathbf{W}}(x, y')}{g} \frac{F_{\mathbf{W}}(x, y')}{\partial \mathbf{W}} && (\text{since } g \text{ is a scalar}) \\ &= -\beta \int_{y'} \frac{F_{\mathbf{W}}(x, y')}{\int_{y'} \exp\left(-\beta F_{\mathbf{W}}(x, y')\right)} \frac{F_{\mathbf{W}}(x, y')}{\partial \mathbf{W}} \\ &= -\beta \int_{y'} p(y'|x) \frac{\partial F_{\mathbf{W}}(x, y')}{\partial \mathbf{W}}.\end{aligned}$$

Replacing in the original formula, we have

$$\frac{\partial \ell_{\mathbf{W}}(x, y)}{\partial \mathbf{W}} = \frac{\partial F_{\mathbf{W}}(x, y)}{\partial \mathbf{W}} - \int_{y'} p(y'|x) \frac{\partial F_{\mathbf{W}}(x, y')}{\partial \mathbf{W}}.$$

The integral over y' might cause the loss intractable. There are several ways to circumvent this:

- For discrete data, like text sequence, integral is simply sum over all possible input y', so the loss is in fact tractable.
- However, for continuous data like image, we can use Monte Carlos Method to estimate the integral.

(d) This is because NLL pushes down on the energy of the correct answer while pushing up on the energies of all answers in proportion to their probabilities. For compatible pair (x, y) with small loss $-\log p(y|x)$, this means $F_{\mathbf{W}}(x, y)$ is getting arbitrary small. Similarly, for incompatible pair (x, y), $F_{\mathbf{W}}(x, y)$ can get arbitrarily large.

1.3 Comparing contrastive loss functions

(a) We have

$$\frac{\partial \ell_{simple}}{\partial \mathbf{W}} = \frac{\partial F_{\mathbf{W}}(x, y)^{+}}{\partial \mathbf{W}} + \frac{\partial \left[m - F_{\mathbf{W}}(x, \overline{y})\right]^{+}}{\partial \mathbf{W}}$$

where

$$\frac{\partial F_{\mathbf{W}}(x, y)^{+}}{\partial \mathbf{W}} = \begin{cases} 0 & \text{if } F_{\mathbf{W}}(x, y) \leq 0 \\ \frac{\partial F_{\mathbf{W}}(x, y)}{\partial \mathbf{W}} & \text{otherwise,} \end{cases}$$

$$\frac{\partial \left[m - F_{\mathbf{W}}(x, \overline{y})\right]^{+}}{\partial \mathbf{W}} = \begin{cases} 0 & \text{if } m \geq F_{\mathbf{W}}(x, \overline{y}) \\ -\frac{\partial F_{\mathbf{W}}(x, \overline{y})}{\partial \mathbf{W}} & \text{otherwise} \end{cases}$$

(b) For hinge loss, we have

$$\begin{aligned}\frac{\partial \ell_{hinge}}{\partial \mathbf{W}} &= \frac{\partial \left[m + F_{\mathbf{W}}(x,y) - F_{\mathbf{W}}(x,\overline{y})\right]^{+}}{\partial \mathbf{W}} \\ &= \begin{cases} 0 & \text{if } m + F_{\mathbf{W}}(x,y) \leq F_{\mathbf{W}}(x,\overline{y}) \\ \frac{\partial F_{\mathbf{W}}(x,y)}{\partial \mathbf{W}} - \frac{\partial F_{\mathbf{W}}(x,\overline{y})}{\partial \mathbf{W}} & \text{otherwise} \end{cases}\end{aligned}$$

(c) For log loss, we have

$$\begin{aligned}\frac{\partial \ell_{log}}{\partial \mathbf{W}} &= \frac{\partial \log\left(1 + \exp\left(F_{\mathbf{W}}(x,y) - F_{\mathbf{W}}(x,\overline{y})\right)\right)}{\partial \mathbf{W}} \\ &= \frac{\exp\left(F_{\mathbf{W}}(x,y) - F_{\mathbf{W}}(x,\overline{y})\right)}{1 + \exp\left(F_{\mathbf{W}}(x,y) - F_{\mathbf{W}}(x,\overline{y})\right)} \left(\frac{\partial F_{\mathbf{W}}(x,y)}{\partial \mathbf{W}} - \frac{\partial F_{\mathbf{W}}(x,\overline{y})}{\partial \mathbf{W}}\right)\end{aligned}$$

(d) For square-square loss (SS), we have

$$\frac{\partial \ell_{ss}}{\partial \mathbf{W}} = \frac{\partial F_{\mathbf{W}}(x,y)^{+2}}{\partial \mathbf{W}} + \frac{\partial \left[(m - F_{\mathbf{W}}(x,\overline{y})^{+2}\right]}{\partial \mathbf{W}},$$

where

$$\frac{\partial F_{\mathbf{W}}(x,y)^{+2}}{\partial \mathbf{W}} = \begin{cases} 0 & \text{if } F_{\mathbf{W}}(x,y) \leq 0 \\ 2F_{\mathbf{W}}(x,y)\frac{\partial F_{\mathbf{W}}(x,y)}{\partial \mathbf{W}} & \text{otherwise} \end{cases}$$

and

$$\frac{\partial \left[(m - F_{\mathbf{W}}(x,\overline{y})^{+2}\right]}{\partial \mathbf{W}} = \begin{cases} 0 & \text{if } F_{\mathbf{W}}(x,\overline{y}) \leq m \\ -2(m - F_{\mathbf{W}}(x,\overline{y}))\frac{\partial F_{\mathbf{W}}(x,\overline{y})}{\partial \mathbf{W}} & \text{otherwise} \end{cases}.$$

(e) (1) Negative log likelihood log involves all examples. It also needs an integral and can be in tractable in continuous data like images, unlike other log function.

(2) As $m \to \infty$, the term before the gradient in log loss tends to 0, which makes log loss looks like hinge loss in the first case ($F_{\mathbf{W}}(x,y) - F_{\mathbf{W}}(x,\overline{y}) \leq -m$). When $m \to 0$, this makes the term before the gradient in log loss tends to 1, makes log loss looks like hinge in the second case.

(3) For square-square and simple loss, we see how the correct inputs energy goes to 0 and incorrect inputs energy goes to at least m. For log and hinge loss, we don't see how energy in different types of energy are pushed, but we only see the relative difference between correct and incorrect inputs.

(4) Simple loss, like L1 loss, can lead to vanishing gradient, unlike square-square loss (like L2 loss).