Deep Learning HW4

Liam Carpenter

November 2021

1.1

a

Energy models are really a inference mechanism that measures the compatibility between an input and an output. Therefore in situations where multiple outputs are possible from a given input, EBM's have the ability to rank and choose the most compatible of these based on the minimization over Y of the energy function.

b

Energy based models output un-normalized scores, whereas probabilistic models have to output normalized scores.

c

You can convert from a energy function $F_W(x, y)$ to a conditional distribution using the Gibb's distribution, $P(y|x) = \frac{e^{-\beta F(x,y)}}{\int_{y'} e^{-\beta F(x,y)} dy'}$

d

The loss function's role is to "push down" on the energy of compatible answers and "push up" on the energy for incompatible answers. The energy function is used for inference, finding the output that minimizes the energy over the space of all outputs.

e

Yes the energy function itself can be used as a Loss Function, though it will only push down on the compatible answers without pushing up on incompatible answers.

f

In some cases where pushing down on good examples does not implicitly push up on bad examples, the energy landscape becomes flat and the model will produce identical outputs no matter the input, ie. joint embedding architectures. This occurs often when the energy function is bounded below by 0.

g

Contrastive: This uses the design of the loss function to push down on the energy function of compatible predictions as expected, but also has a component of the loss that uses "most offending examples", examples that are attributed low energy but are incorrect, to push up on regions where the energy landscape is low but should be high.

Regularization: Explicitly include a term in the loss function that bounds the total low energy space. Then when good examples' energy are being pushed down other regions will necessarily be pushed up.

Architectural: Build a model in such a way that the volume of low energy space is bounded.

h

Hinge Loss: $\ell_{\text{example}}(x, y, W) = [F_W(x, y) - F_W(x, \bar{y}) + m(\bar{y}, y))]^+$ Where $\bar{y}$ is the most offending example.

1.2

a

$P(y|x) = \frac{e^{-\beta F_W(x,y)}}{\sum_{i=1}^n e^{-\beta F_W(x,i)}}$

b

$-log(P(y|x)) = -log(\frac{e^{-\beta F_W(x,y)}}{\sum_{i=1}^n e^{-\beta F_W(x,i)}}) = -log(e^{-\beta F_W(x,y)}) + log(\sum_{i=1}^n e^{-\beta F_W(x,i)})$
$= \beta F_W(x, y) + log(\sum_{i=1}^n e^{-\beta F_W(x,i)})$

c

$\frac{\partial[-\frac{1}{\beta}P(y|x)]}{\partial W} = \frac{\partial F_W(x,y)}{\partial W}\frac{\partial \frac{1}{\beta} - log(\sum_{i=1}^n e^{-\beta F_W(x,i)})}{\partial W} = \frac{\partial F_W(x,y)}{\partial W} - \frac{1}{\sum_{i=1}^n e^{-\beta F_W(x,i)}} \sum_{i=1}^n e^{-\beta F_W(x,i)} \frac{\partial F_W(x,i)}{\partial W}$
$= \frac{\partial F_W(x,y)}{\partial W} - \sum_{i=1}^n \frac{e^{-\beta F_W(x,i)} \frac{\partial F_W(x,i)}{\partial W}}{\sum_{i=1}^n e^{-\beta F_W(x,i)}} = \frac{\partial F_W(x,y)}{\partial W} - \sum_{i=1}^n P(i|x) \frac{\partial F_W(x,i)}{\partial W} = \frac{\partial F_W(x,y)}{\partial W} -$
$\mathbb{E}[\frac{\partial F_W(x,i)}{\partial W}]$

In the continous case this is an integral and y can be of high dimensionality so the integral can be intractable. This can be avoided with monte carlo methods, that sample from the distribution $P(y|x)$ and compute the integrand using this.

Doing this many times the value converges to the expectation of our gradient, which is exactly the integral we are trying to compute.

d

Looking at the gradient step

$$w \leftarrow w - \eta \frac{\partial F_W(x,y)}{\partial W} + \eta \sum_{i=1}^{n} P(i|x) \frac{\partial F_W(x,i)}{\partial W}$$

We can see here in contrast to other contrastive methods, that the energy of all of the examples are going to get pulled up (including the correct one, though not by as much as it will get pushed down), rather than a few well chosen examples.

1 1.3

a

$$\frac{\partial \ell_{simple}}{W} = \begin{cases} \frac{\partial F_W(x,y)}{\partial W} & \text{if } F_W(x,y) > 0, m - F_W(x,\bar{y}) < 0 \\ -\frac{\partial F_W(x,\bar{y})}{\partial W} & \text{if } F_W(x,y) < 0, m - F_W(x,\bar{y}) > 0 \\ \frac{\partial F_W(x,y)}{\partial W} - \frac{\partial F_W(x,\bar{y})}{\partial W} & \text{if } F_W(x,y) > 0, m - F_W(x,\bar{y}) \\ 0 \quad \text{otherwise} \end{cases}$$

b

$$\frac{\partial \ell_{hinge}}{W} = \begin{cases} \frac{\partial F_W(x,y)}{\partial W} - \frac{\partial F_W(x,\hat{y})}{\partial W} + m & \text{if } F_W(x,y) - F_W(x,\bar{y}) + m > 0 \\ 0 \quad \text{otherwise} \end{cases}$$

c

$$\frac{\partial \ell_{square-square}}{W} = \begin{cases} 2F_W(x,y)\frac{\partial F_W(x,y)}{\partial W} & \text{if } F_W(x,y) > 0, m - F_W(x,\bar{y}) < 0 \\ -2F_W(x,\bar{y})\frac{\partial F_W(x,\bar{y})}{\partial W} & \text{if } F_W(x,y) < 0, m - F_W(x,\bar{y}) > 0 \\ 2F_W(x,y)\frac{\partial F_W(x,y)}{\partial W} - 2(m - F_W(x,\bar{y}))\frac{\partial F_W(x,\bar{y})}{\partial W}, & \text{if } F_W(x,y) > 0, m - F_W(x,\bar{y}) > 0 \\ 0 \quad \text{otherwise} \end{cases}$$

d

a

NLL loss doesn't just push up the energy on choice adversarial examples, it pushes the energy up across all examples, making for very steep energy surfaces. It does not discriminate by how bad something is, but only whether or not it is correct. All of these losses above use properly select adversarial examples, specifically chosen to have the minimum energy among incorrect answers. In addition there is a notion of how bad the energy of the adversarial example

has to be before it is penalized, energy is pushed up, by the model through the margin term.

b

The margin specifies how far away from the energy of the correct answer the energy of the most offending answer must be before the model penalizes the difference. Importantly the hinge loss focuses in only on the difference in energies and so does not constrain the specific values that the energies must take. So through the equation we can see that when the energy between the most offending example and the correct example is greater than $-m$ the model will penalize the difference of the two energies, implicitly pushing the correct example down and the most offending example up.

c

These loss treat the correct examples and the most offending examples independently. This means that large values of correct answers and small answers of offending answers, up to the margin, are penalized. These losses will push the correct labels energy down to 0 and are therefore suitable for energy functions that are bounded below by zero, such as energy functions that output some sort of distance measure.