TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Noisy Channel RDAs

Rate-Distortion Autoencoders (RDAs)

We compress a continuous signal y to a bit string $\tilde{z}_\Phi(y)$.

We decompress $\tilde{z}_\Phi(y)$ to $y_\Phi(\tilde{z}_\Phi(y))$.

$$\Phi^* = \operatorname*{argmin}_\Phi \; E_{y\sim\mathrm{Pop}} \quad |\tilde{z}_\Phi(y)| + \lambda \mathrm{Dist}(y, y_\Phi(\tilde{z}_\Phi(y)))$$

Rate-Distortion Autoencoders (RDAs)

Since rounding is not differentiable we train by replace rounding by additive noise.

$$\mathcal{L}(\Phi) = E_{y\sim\mathrm{Pop}}\; E_{\epsilon} \left\{ \begin{array}{c} -\ln p_\Phi(z_\Phi(y)+\epsilon) \\ +\;\lambda \mathrm{Dist}(y, y_\Phi(z_\Phi(y)+\epsilon)) \end{array} \right.$$

A noisy-channel RDA uses the noise version without rounding.

Noisy Channel RDAs

$$z = z_\Phi(y, \epsilon) \quad \epsilon \text{ is fixed (parameter independent) noise}$$

$$\Phi^* = \operatorname*{argmin}_\Phi \; I_\Phi(y, z) + \lambda E_{y,\epsilon} \, \mathrm{Dist}(y, y_\Phi(z_\Phi(y, \epsilon)))$$

By the channel capacity theorem $I(y, z)$ is the **rate** of information transfer from y to z.

Noisy Channel RDAs

$$z = z_\Phi(y,\epsilon) \;\; \epsilon \text{ is fixed (parameter independent) noise}$$

$$\Phi^* = \operatorname*{argmin}_\Phi \; I_\Phi(y,z) + \lambda E_{y,\epsilon} \operatorname{Dist}(y, y_\Phi(z_\Phi(y,\epsilon)))$$

Using parameter-independent noise is called the "reparameterization trick" and allows SGD.

$$\nabla_\Phi \, E_{y,\epsilon} \operatorname{Dist}(y, y_\Phi(z_\Phi(y,\epsilon)))$$

$$= E_{y,\epsilon} \, \nabla_\Phi \operatorname{Dist}(y, y_\Phi(z_\Phi(y,\epsilon)))$$

Mutual Information as a Channel Rate

Typically $z_\Phi(y, \epsilon)$ is simple. For example

$$\epsilon \sim \mathcal{N}(0, I)$$

$$z_\Phi(y, \epsilon) = \mu_\Phi(y) + \sigma_\Phi(y) \odot \epsilon$$

In this example $p_\Phi(z|y)$ is easily computed.

Mutual Information Replaces Rate

$$I_\Phi(y,z) \;=\; E_{y,\epsilon}\;\ln\frac{\mathrm{pop}(y)p_\Phi(z|y)}{\mathrm{pop}(y)p_{\mathrm{pop},\Phi}(z)}$$

$$\;=\; E_{y,\epsilon}\;\ln\frac{p_\Phi(z|y)}{p_{\mathrm{pop},\Phi}(z)}$$

where $\quad p_{\mathrm{pop},\Phi}(z) \;=\; E_{y\sim\mathrm{pop}}\;\;p_\Phi(z|y)$

A Variational Bound

$$p_{\mathrm{pop},\Phi}(z) = E_{y \sim \mathrm{pop}} \;\; p_\Phi(z|y)$$

We cannot compute $p_{\mathrm{pop},\Phi}(z)$.

Instead we will use a model $\hat{p}_\Phi(z)$ to approximate $p_{\mathrm{pop},\Phi}(z)$.

A Variational Bound

$$
\begin{aligned}
I(y,z) &= E_{y,\epsilon}\;\ln\frac{p_\Phi(z|y)}{p_{\mathrm{pop},\Phi}(z)}\\
&= E_{y,\epsilon}\;\ln\frac{p_\Phi(z|y)}{\hat{p}_\Phi(z)} + E_{y,\epsilon}\;\ln\frac{\hat{p}_\Phi(z)}{p_{\mathrm{pop},\Phi}(z)}\\
&= E_{y,\epsilon}\;\ln\frac{p_\Phi(z|y)}{\hat{p}_\Phi(z)} - KL(p_{\mathrm{pop},\Phi}(z),\hat{p}_\Phi(z))\\
&\leq E_{y,\epsilon}\;\ln\frac{p_\Phi(z|y)}{\hat{p}_\Phi(z)}
\end{aligned}
$$

The Noisy Channel RDA

General Noisy Channel RDA:

$$\Phi^* \;=\; \underset{\Phi}{\operatorname{argmin}}\, E_{y,\epsilon}\;\ln\frac{p_\Phi(z_\Phi(y,\epsilon)|y)}{\hat{p}_\Phi(z_\Phi(y,\epsilon))} + \lambda \mathrm{Dist}(y, y_\Phi(z_\Phi(y,\epsilon)))$$

Uniform Box Noise (Rounding) RDA:

$$\Phi^* \;=\; \underset{\Phi}{\operatorname{argmin}}\; E_y\; E_{\epsilon\sim[-1/2,1/2]^d}$$

$$-\ln \hat{p}_\Phi(z_\Phi(y)+\epsilon) + \lambda \mathrm{Dist}(y, y_\Phi(z_\Phi(y)+\epsilon))$$

END