TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Noisy Channel RDAs

Review of Rate-Distortion Autoencoders (RDAs)

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

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

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

The loss is "legitimate" in that , unlike differential cross entropy, the loss terms are guaranteed to be non-negative.

But the discrete cross entropy term is not differentiable.

Noisy channel RDAs use a legitimate yet differentiable loss.

Rate as Channel Capacity

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

$$p_\Phi(z) = \int \mathrm{pop}(y) p_\Phi(z|y) dy \; = \; E_y \;\, p_\Phi(z|y)$$

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

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

The mutual information $I_\Phi(y, z)$ is the channel capacity giving the **rate** of information transfer from y to z.

Mutual Information as a Channel Rate

Typically we have $\epsilon \sim \mathcal{N}(0, I)$ and

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

Here $p_\Phi(z|y)$ is a Gaussian with mean $\mu_\Phi(y)$ and a diagonal covariance matrix with diagonal entries $\sigma_\Phi(y)$.

A Variational Bound on Mutual Information

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; E_{y,\epsilon} \; \ln \frac{p_\Phi(z \mid y)}{p_\Phi(z)} \; + \; \lambda \mathrm{Dist}(y, y_\Phi(z))$$

Here $p_\Phi(z)$ is the marginal of z under the distribution defined by y and ϵ.

$$p_\Phi(z) \; = \; \int \mathrm{pop}(y) p_\Phi(z|y) dy \; = \; E_y \; p_\Phi(z|y)$$

We cannot compute $p_\Phi(z)$.

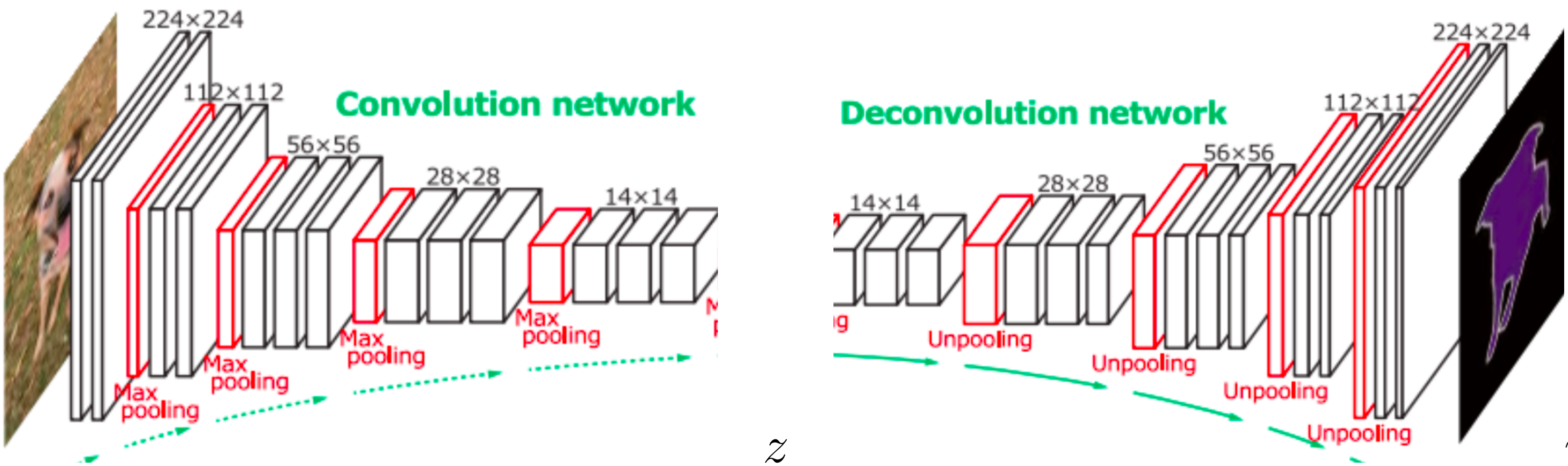
Instead we will use a model $\hat{p}_\Phi(z)$ to approximate $p_\Phi(z)$.

A Variational Bound on Mutual Information

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

The Noisy Channel RDA

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



y, ϵ $\quad z \quad$ $\hat{y}$

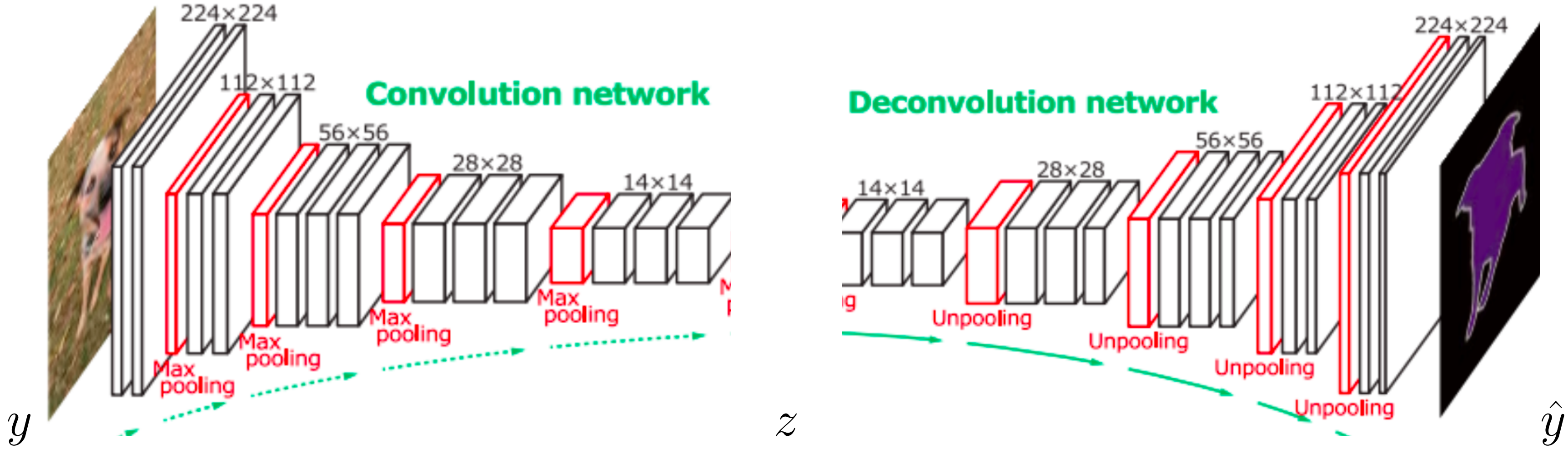
Sampling

We can require $\hat{p}_\Phi(z)$ be Gaussian. In that case we can sample z from $\hat{p}_\Phi(z)$ and generate images (as in a GAN).

[Alec Radford]

This is **sampling** — not compression. We are decompressing noise.

A General Autoencoder



We show below that for $p_\Phi(z|y)$ and $\hat{p}_\Phi(z)$ both required to be Gaussian we can assume without loss of generality that

$$\hat{p}_\Phi(z) = \mathcal{N}(0, I)$$

Gaussian Noisy-Channel RDA

We now show that a reparameterization can always convert $\hat{p}_\Phi(z)$ to a zero-mean identity-covariance Gaussian.

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

$$\begin{aligned}
z_\Phi(y,\epsilon) &= \mu_\Phi(y) + \sigma_\Phi(y) \odot \epsilon \quad \epsilon \sim \mathcal{N}(0, I) \\
p_\Phi(z[i]|y) &= \mathcal{N}(\mu_\Phi(y)[i], \sigma_\Phi(y)[i])) \\
\hat{p}_\Phi(z[i]) &= \mathcal{N}(\hat{\mu}_z[i], \hat{\sigma}_z[i]) \\
\mathrm{Dist}(y, \hat{y}) &= ||y - \hat{y}||^2
\end{aligned}$$

Gaussian Noisy-Channel RDA

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

We will show that we can fix $\hat{p}_\Phi(z)$ to $\mathcal{N}(0, I)$.

$$\begin{aligned}
p_\Phi(z[i]|y) &= \mathcal{N}(\mu_\Phi(y)[i], \sigma_\Phi(y)[i]) \\
\hat{p}_\Phi(z[i]) &= \mathcal{N}(0, 1) \\
\mathrm{Dist}(y, \hat{y}) &= ||y - \hat{y}||^2
\end{aligned}$$

Gaussian Noisy-Channel RDA

$$
\begin{aligned}
\Phi^* \;&=\; \operatorname*{argmin}_{\Phi} E_{y,\epsilon}\; \ln \frac{p_\Phi(z_\Phi(y,\epsilon)|y)}{\hat{p}_\Phi(z_\Phi(y,\epsilon))} + \lambda \mathrm{Dist}(y, y_\Phi(z_\Phi(y,\epsilon))) \\[2em]
&=\; \operatorname*{argmin}_{\Phi}\; E_{y\sim \mathrm{Pop}} \left(\begin{array}{c} KL(p_\Phi(z|y), \hat{p}_\Phi(z)) \\[2em] +\lambda\; E_\epsilon\; \mathrm{Dist}(y,\; y_\Phi(z_\Phi(y,\epsilon))) \end{array} \right)
\end{aligned}
$$

Closed Form KL-Divergence

$$KL(p_\Phi(z|y), \hat{p}_\Phi(z))$$

$$= \sum_i \frac{\sigma_\Phi(y)[i]^2 + (\mu_\Phi(y)[i] - \mu_z[i])^2}{2\sigma_z[i]^2} + \ln\frac{\sigma_z[i]}{\sigma_\Phi(y)[i]} - \frac{1}{2}$$

Standardizing $\hat{p}_\Phi(z)$

$$KL(p_\Phi(z|y), p_\Phi(z))$$

$$= \sum_i \frac{\sigma_\Phi(y)[i]^2 + (\mu_\Phi(y)[i] - \mu_z[i])^2}{2\sigma_z[i]^2} + \ln \frac{\sigma_z[i]}{\sigma_\Phi(y)[i]} - \frac{1}{2}$$

$$KL(p_{\Phi'}(z|y), \mathcal{N}(0, I))$$

$$= \sum_i \frac{\sigma_{\Phi'}(y)[i]^2 + \mu_{\Phi'}(y)[i]^2}{2} + \ln \frac{1}{\sigma_{\Phi'}(y)[i]} - \frac{1}{2}$$

Standardizing $\hat{p}_\Phi(z)$

$$
\begin{aligned}
KL_\Phi &= \sum_i \frac{\sigma_\Phi(y)[i]^2 + (\mu_\Phi(y)[i] - \mu_z[i])^2}{2\sigma_z[i]^2} + \ln \frac{\sigma_z[i]}{\sigma_\Phi(y)[i]} - \frac{1}{2} \\
KL_{\Phi'} &= \sum_i \frac{\sigma_{\Phi'}(y)[i]^2 + \mu_{\Phi'}(y)[i]^2}{2} + \ln \frac{1}{\sigma_{\Phi'}(y)[i]} - \frac{1}{2}
\end{aligned}
$$

Setting Φ' so that

$$
\begin{aligned}
\mu_{\Phi'}(y)[i] &= (\mu_\Phi(y)[i] - \mu_z[i])/\sigma_z[i] \\
\sigma_{\Phi'}(y)[i] &= \sigma_\Phi(y)[i]/\sigma_z[i]
\end{aligned}
$$

gives $KL(p_\Phi(z|y), \hat{p}_\Phi(z)) = KL(p_{\Phi'}(z|y), \mathcal{N}(0, I))$.

END