

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2021

Noisy Channel RDAs

The KL term as Channel Capacity

$$\begin{aligned}\Phi^* &= \operatorname{argmin}_{\Phi} E_{y,z} \ln \frac{p_{\psi}(z \mid y)}{p_{\Phi}(z)} - \ln p_{\Phi}(y|z) \\ &= \operatorname{argmin}_{\Phi} I_{\Psi,\Phi}(y, z) + E_{y,z} - \ln p_{\Phi}(y|z)\end{aligned}$$

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

L_2 Distortion

$$\mathcal{L}(\Phi) = E_{y \sim P_{\text{op}}} - \ln P_{\Phi}(\tilde{z}_{\Phi}(y)) + \lambda \text{Dist}(y, y_{\Phi}(\tilde{z}_{\Phi}(y)))$$

It is common to take

$$\begin{aligned} \text{Dist}(y, \hat{y}) &= ||y - \hat{y}||^2 \quad (L_2) \\ &= -\frac{1}{\lambda} \ln p(y|\hat{y}) + C \quad \text{for } p(y|\hat{y}) \propto \exp(-\lambda ||y - \hat{y}||^2) \end{aligned}$$

We will ignore the log density interpretation and just call this L_2 distortion.

L_1 Distortion

$$\mathcal{L}(\Phi) = E_{y \sim P_{\text{op}}} - \ln P_{\Phi}(\tilde{z}_{\Phi}(y)) + \lambda \text{Dist}(y, y_{\Phi}(\tilde{z}_{\Phi}(y)))$$

Alternatively we have

$$\begin{aligned} \text{Dist}(y, \hat{y}) &= \|y - \hat{y}\|_1 && (L_1) \\ &= -\frac{1}{\lambda} \ln p(y|\hat{y}) + C \quad \text{for } p(y|\hat{y}) \propto \exp(-\lambda \|y - \hat{y}\|_1) \end{aligned}$$

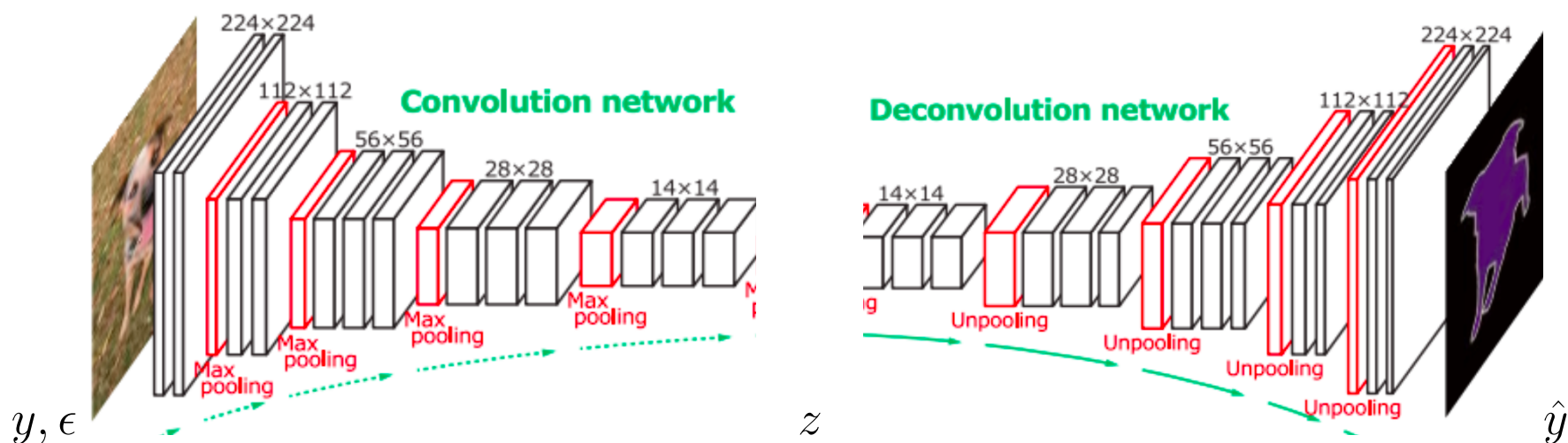
Again, we will ignore the log density interpretation and just call this L_1 distortion.

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

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



VAE = RDA

$$\text{VAE: } \Phi^* = \underset{\Phi}{\operatorname{argmin}} E_{y \sim \text{Pop}, z \sim \hat{P}_{\Phi}(z|y)} \ln \frac{\hat{P}_{\Phi}(z|y)}{P_{\Phi}(z)} - \ln P_{\Phi}(y|z)$$

$P_{\Phi}(z)$, $P_{\Phi}(y|z)$ and $\hat{P}_{\Phi}(z|y)$ are model components and we can switch the notation to $\hat{P}_{\Phi}(z)$ $\hat{P}_{\Phi}(y|z)$ and $P_{\Phi}(z|y)$ with no change in the model.

$$\text{RDA: } \Phi^* = \underset{\Phi}{\operatorname{argmin}} E_{y \sim \text{Pop}, z \sim P_{\Phi}(z|y)} \ln \frac{P_{\Phi}(z|y)}{\hat{P}_{\Phi}(z)} - \ln \hat{P}_{\Phi}(y|z)$$

In an RDA we take $P_{\Phi}(y, z)$ to be $\text{Pop}(y)P_{\Phi}(z|y)$ so that the rate term is an upper bound on $I_{\Phi}(y, z)$.

Sampling

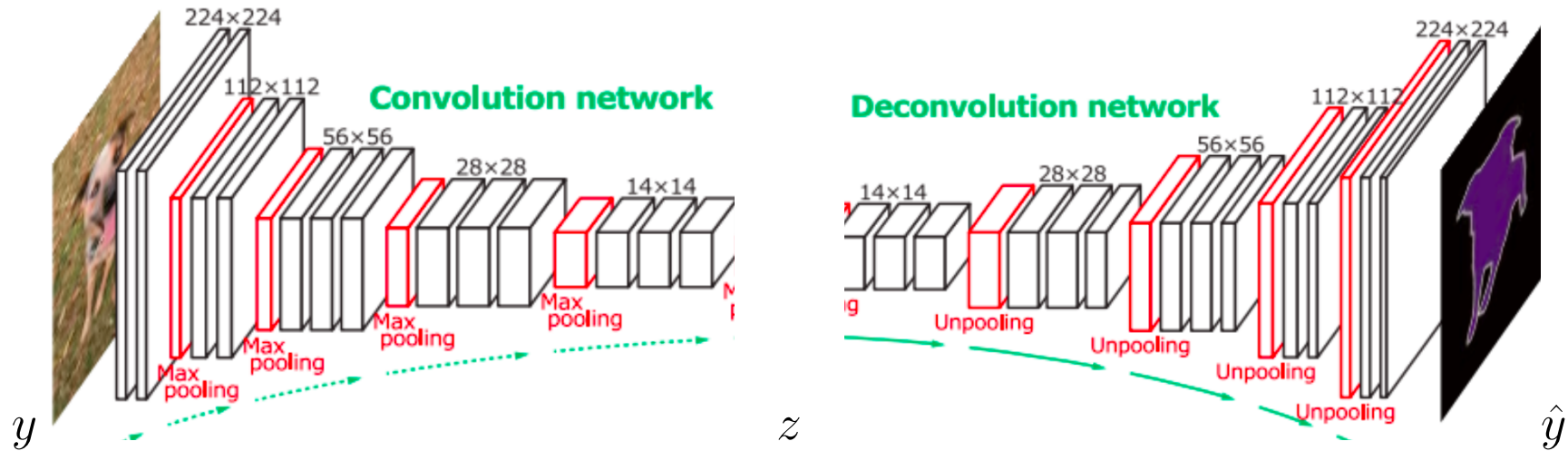
We can require $\hat{p}_\Phi(z)$ be Gaussian. In that case we can sample z from $\hat{p}_\Phi(z)$ and generate images (as in a GAN).



[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.

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

$$z_\Phi(y, \epsilon) = \mu_\Phi(y) + \sigma_\Phi(y) \odot \epsilon \quad \epsilon \sim \mathcal{N}(0, I)$$

$$p_\Phi(z[i]|y) = \mathcal{N}(\mu_\Phi(y)[i], \sigma_\Phi(y)[i])$$

$$\hat{p}_\Phi(z[i]) = \mathcal{N}(\hat{\mu}_z[i], \hat{\sigma}_z[i])$$

$$\operatorname{Dist}(y, \hat{y}) = ||y - \hat{y}||^2$$

Gaussian Noisy-Channel RDA

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

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

$$p_{\Phi}(z[i]|y) = \mathcal{N}(\mu_{\Phi}(y)[i], \sigma_{\Phi}(y)[i])$$

$$\hat{p}_{\Phi}(z[i]) = \mathcal{N}(0, 1)$$

$$\operatorname{Dist}(y, \hat{y}) = ||y - \hat{y}||^2$$

Gaussian Noisy-Channel RDA

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

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

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

$$\begin{aligned} & KL(p_{\Phi'}(z|y), \mathcal{N}(0, I)) \\ &= \sum_i \frac{\sigma_{\Phi'}(y)[i]^2 + \mu_{\Phi'}(y)[i]^2}{2} + \ln \frac{1}{\sigma_{\Phi'}(y)[i]} - \frac{1}{2} \end{aligned}$$

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

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

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

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