TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Rate-Distortion Autoencoders (RDAs)

Noisy Channel RDAs

Gaussian Noisy Channel RDAs

Rate-Distortion Autoencoders
(Image Compression)

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

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

We can then define a rate-distortion loss.

$$\mathcal{L}(\Phi) = E_{y \sim \mathrm{Pop}}\; |\tilde{z}_\Phi(y)| + \lambda \mathrm{Dist}(y, y_\Phi(\tilde{z}_\Phi(y)))$$

where $|\tilde{z}|$ is the number of bits in the bit string $\tilde{z}$.

Common Distortion Functions

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

It is common to take

$$\mathrm{Dist}(y,\hat{y}) = ||y-\hat{y}||^2 \qquad (L_2)$$

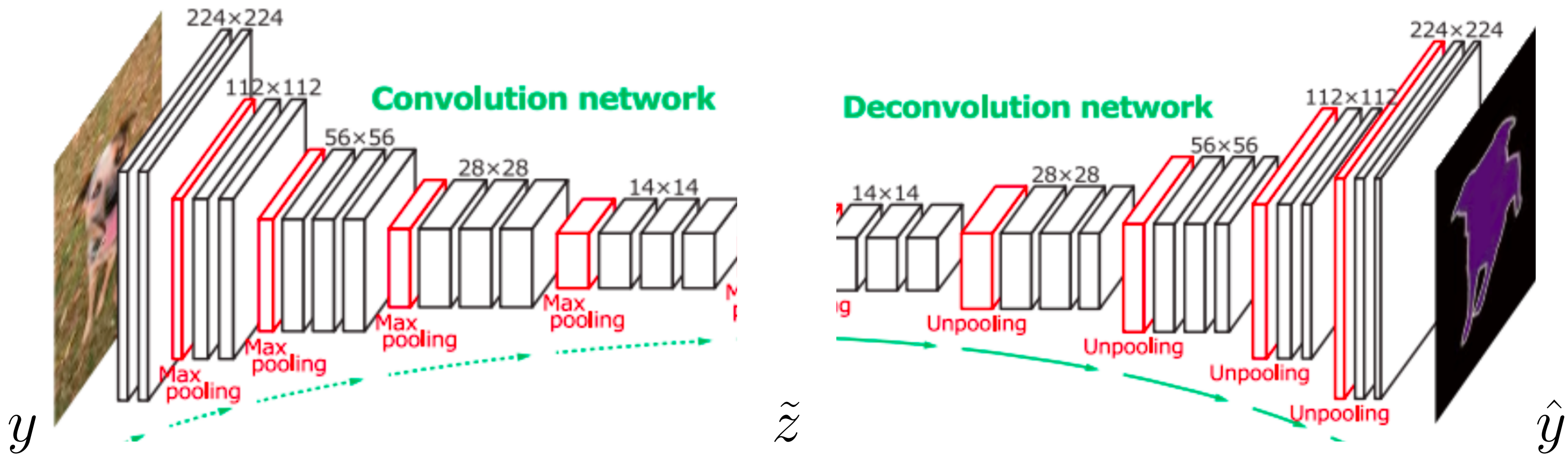
or

$$\mathrm{Dist}(y,\hat{y}) = ||y-\hat{y}||_1 \qquad (L_1)$$

CNN-based Image Compression

These slides are loosely based on

End-to-End Optimized Image Compression, Balle, Laparra, Simoncelli, ICLR 2017.



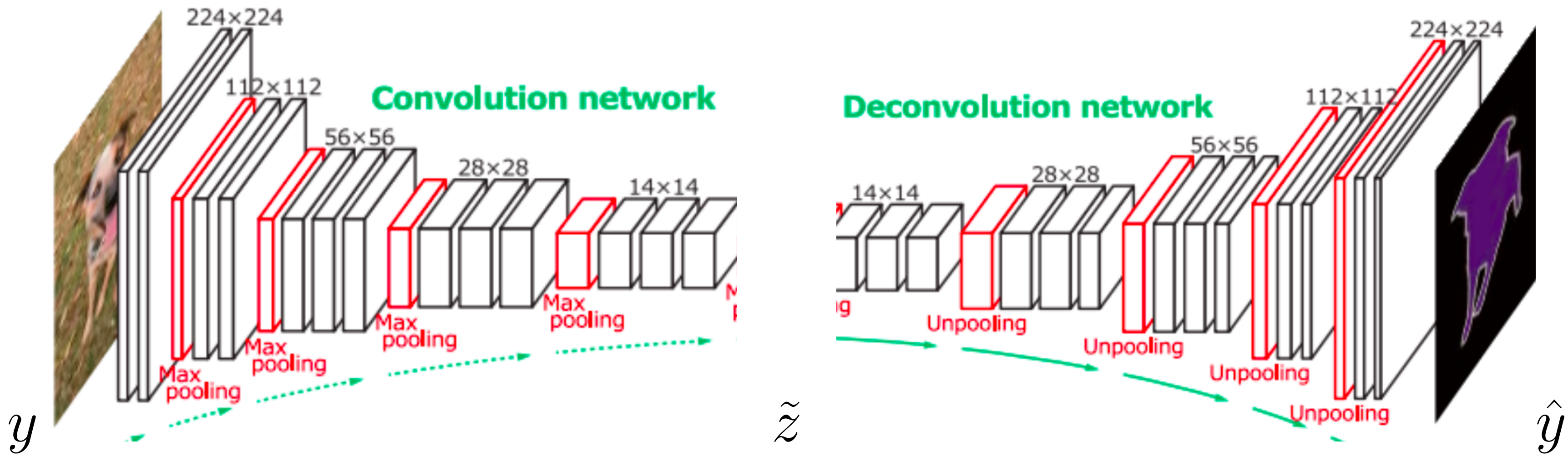
Rounding a Tensor

Take $z_\Phi(y)$ can be a layer in a CNN applied to image y. $z_\Phi(y)$ can have with both spatial and feature dimensions.

Take $\tilde{z}_\Phi(y)$ to be the result of rounding each component of the continuous tensor $z_\Phi(y)$ to the nearest integer.

$$\tilde{z}_\Phi(y)[x, y, i] = \lfloor z_\Phi(y)[x, y, i] + 1/2 \rfloor$$

Increasing Spatial Dimension in Decoding



Increasing Spatial Dimension in Decoding (Deconvolution)

To increase spatial dimension we use 4 times the desired output the features.

$$L'_{\ell+1}[x, y, i] = \sigma\left(W[\Delta X, \Delta Y, J, i]\ L'_{\ell}[x + \Delta X, y + \Delta Y, J]\right)$$

We then reshape $L'_{\ell+1}[X, Y, I]$ to $L'_{\ell+1}[2X, 2Y, I/4]$.

Rounding is not Differentiable

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

Because of rounding, $\tilde{z}_\Phi(y)$ is discrete and the gradients are zero.

We will train using a differentiable approximation.

Rate: Replacing Code Length with Differential Entropy

$$\mathcal{L}_{\text{rate}}(\Phi) = E_{y\sim\text{Pop}}\,|\tilde{z}_\Phi(y)|$$

Recall that $\tilde{z}_\Phi(y)$ is a rounding of a continuous encoding $z_\Phi(y)$.

By using a nontrivial code for integers — say Huffman coding integers — we can approximate the code length of the rounded integer with a continuous probability density.

$$|\tilde{z}_\Phi(y)| \approx \sum_{x,y,i} -\ln p_\Phi(z_\Phi(y)[x,y,i])$$

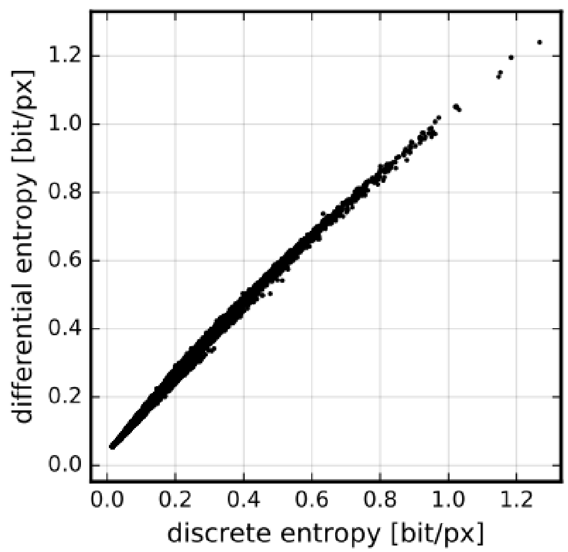
Distortion: Replacing Rounding with Noise

We can make distortion differentiable by modeling rounding as the addition of noise.

$$\begin{aligned}\mathcal{L}_{\mathrm{dist}}(\Phi) &= E_{y\sim\mathrm{Pop}}\,\mathrm{Dist}(y, y_\Phi(\tilde{z}_\Phi(y))) \\ &\approx E_{y,\epsilon}\,\mathrm{Dist}(y,\ y_\Phi(z_\Phi(y)+\epsilon))\end{aligned}$$

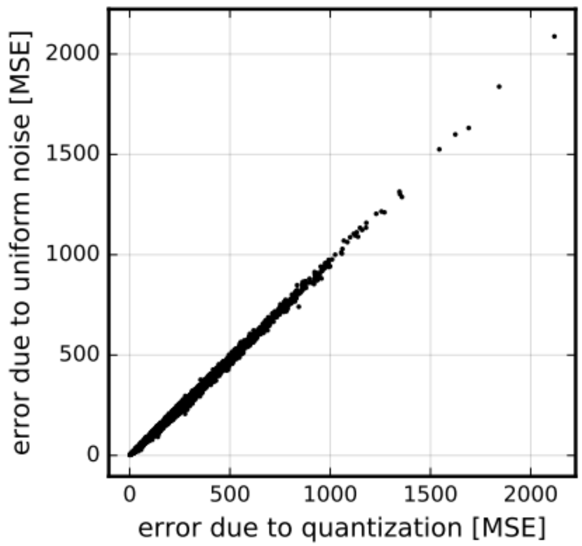
Here ϵ is a noise vector each component of which is drawn uniformly from $(-1/2, 1/2)$.

Rate: Differential Entropy vs. Discrete Entropy



Each point is a rate for an image measured in both differential entropy and discrete entropy. The size of the rate changes as we change the weight λ.

Distortion: Noise vs. Rounding



Each point is a distortion for an image measured in both a rounding model and a noise model. The size of the distortion changes as we change the weight λ.

JPEG at 4283 bytes or .121 bits per pixel

JPEG, 4283 bytes (0.121 bit/px), PSNR: 24.85 dB/29.23 dB, MS-SSIM: 0.8079

JPEG 2000 at 4004 bytes or .113 bits per pixel

JPEG 2000, 4004 bytes (0.113 bit/px), PSNR: 26.61 dB/33.88 dB, MS-SSIM: 0.8860

Deep Autoencoder at 3986 bytes or .113 bits per pixel

Proposed method, 3986 bytes (0.113 bit/px), PSNR: 27.01 dB/34.16 dB, MS-SSIM: 0.9039

Noisy-Channel RDAs

The image compression case study training was based on a differentiable loss

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}} \Big(\; E_y \;\; -\ln p_\Phi(z_\Phi(y))\Big) + \lambda E_{y,\epsilon} \; \mathrm{Dist}(y, y_\Phi(z_\Phi(y)+\epsilon))$$

In a rate-distortion auto-encoder we will replace the rate term with a channel capacity (rate) for a noisy channel on continuous variables.

Mutual Information as a Channel Rate

$$\Phi^* = \underset{\Phi}{\text{argmin}} \left(E_y \; - \ln p_\Phi(z_\Phi(y))\right) + \lambda E_{y,\epsilon} \, \text{Dist}(y, y_\Phi(z_\Phi(y)+\epsilon))$$

is replaced by

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

$$\Phi^* \;=\; \underset{\Phi}{\text{argmin}} \;\; I_\Phi(y,\tilde{z}) + \lambda E_{y,\epsilon} \, \text{Dist}(y, y_\Phi(\tilde{z}))$$

By the channel capacity theorem $I(y,\tilde{z})$ is the **rate** of information transfer from y to $\tilde{z}$. Differential mutual infomation is more meaningful than differential cross entropy.

Mutual Information as a Channel Rate

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

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

Taking the distribution on ϵ to be parameter independent is called the "reparameterization trick" and allows SGD.

$$\nabla_\Phi \; E_{y,\epsilon}\,\mathrm{Dist}(y, y_\Phi(z_\Phi(y) + f_\Phi(\epsilon)))$$

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

Mutual Information as a Channel Rate

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

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

Typically $f_\Phi(y, \epsilon)$ is simple, such as $\sigma_\Phi(y) \odot \epsilon$, so that $p_\Phi(\tilde{z}|y)$ is easily computed.

Mutual Information Replaces Cross Entropy

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

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

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

A Variational Bound

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

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

Instead we will use a variational bound involving a computable model $q_\Phi(\tilde{z})$

A Variational Bound

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

A Fundamental Equation for the Continuous Case

$$\tilde{z} = z_\Phi(y) + f_\Phi(y, \epsilon)$$

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

Gaussian Noisy-Channel RDA

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

$$\tilde{z}[i] \ = \ z_\Phi(y)[i] + \sigma_\Phi(y)\epsilon[i] \quad \epsilon[i] \sim \mathcal{N}(0,1)$$

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

$$q_\Phi(\tilde{z}[i]) \ = \ \mathcal{N}(\mu_q[i], \sigma_q[i])$$

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

Gaussian Noisy-Channel RDA

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

We will show that in the Gaussian case can fix q_Φ

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

$$q_\Phi(\tilde{z}[i]) = \mathcal{N}(0, 1)$$

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

Gaussian Noisy-Channel RDA

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

$$= \underset{\Phi}{\operatorname{argmin}}\ E_{y\sim \mathrm{Pop}} \begin{pmatrix} KL(p_\Phi(\tilde{z}|y), q_\Phi(\tilde{z})) \\ \\ +\lambda\ E_\epsilon\ \mathrm{Dist}(y,\ y_\Phi(\tilde{z})) \end{pmatrix}$$

Closed Form KL-Divergence

$$KL(p_\Phi(\tilde{z}|y), q_\Phi(\tilde{z}))$$

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

Standardizing $p_\Phi(z)$

$$KL(p_\Phi(\tilde{z}|y), p_\Phi(\tilde{z}))$$

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

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

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

Standardizing $p_\Phi(z)$

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

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

Setting Φ' so that

$$\begin{aligned} z_{\Phi'}(y)[i] &= (z_\Phi(y)[i] - \mu_q[i])/\sigma_q[i] \\ \sigma^\epsilon_{\Phi'}(y)[i] &= \sigma_\Phi(y)[i]/\sigma_q[i] \end{aligned}$$

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

Sampling

Sample $\tilde{z} \sim \mathcal{N}(0, I)$ and compute $y_\Phi(\tilde{z})$

[Alec Radford]

Summary: Rate-Distortion

RDA: y continuous, $\tilde{z}$ a bit string,

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

Gaussian RDA: $\tilde{z} = z_\Phi(y) + \sigma_\Phi(y) \odot \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I)$

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}}\, E_{y\sim\mathrm{Pop}} \begin{pmatrix} & KL(p_\Phi(\tilde{z}|y), \mathcal{N}(0,I)) \\ + & \lambda E_\epsilon\, \mathrm{Dist}(y, y_\Phi(\tilde{z})) \end{pmatrix}$$

END