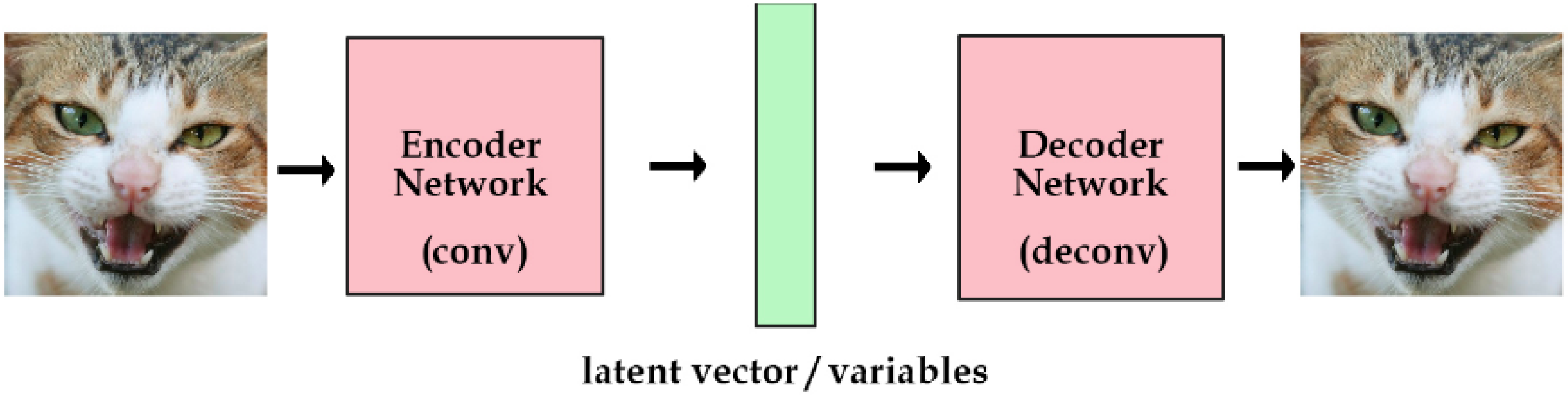
TTIC 31230, Fundamentals of Deep Learning

David McAllester, April 2017

A Case Study in Rate-Distortion Autoencoding

A Case Study in Image Compression

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



[Kevin Frans]

Rate-Distortion Autoencoders

We consider lossy compression. Here we assume:

- An deterministic encoder network $z_\Phi(x)$.
- A "coding distribution" $P_\Theta^{\mathrm{code}}(z)$ defining lossless coding and decoding for z.
- A decoder network $\hat{x}_\Psi(z)$
- A distortion function $L(x, \hat{x})$

$$\Phi^*, \Psi^*, \Theta^* = \operatorname*{argmin}_{\Phi,\Psi,\Theta} \; \mathrm{E}_{x \sim D} \left[\log \frac{1}{P_\Theta^{\mathrm{code}}(z_\Phi(x))} + \lambda L(x, \hat{x}_\Psi(z_\Phi(x))) \right]$$

A formal comparison with variational autoencoders is given at the end of these slides.

The Encoder

This paper uses a three layer CNN for the encoder.

The first layer is computed stride 4.

The last two layers are computed stride 2.

They use a normalization layer rather than an activation function.

$$v_i = \frac{u_i}{\left(\beta_i + \sum_j \gamma_{i,j}\, u_j^2\right)^{1/2}}$$

β_i and $\gamma_{i,j} = \gamma_{j,i}$ are trained.

The number of numbers

Final image dimension is reduced by a factor of 16 with 192 channels per pixel (192 channels is for color images).

$$192 < 16 \times 16 \times 3 = 768$$

These 192 numbers are rounded to integers.

The 192 integers are coded losslessly using $P_{\Theta}^{\mathrm{code}}$.

The Decoder

This is a deconvolution network of the same architecture with independent parameters.

There is a special parameterization of the “inverter” for the normalization layer.

Rounding the Numbers

We let $z_\Phi(x)$ be the unrounded numerical representation and $\hat{x}_\Phi(x)$ be the result of rounding.

$$\hat{z}_\Phi(x)_i = \mathrm{round}(z_\Phi(x)_i) = \lfloor z_\Phi(x)_i + 1/2 \rfloor$$

Each integer channel of the final layer is coded independently.

Context-based adaptive binary arithmetic coding framework (CABAC; Marpe, Schwarz, and Wiegand, 2003).

Training

We now have the optimization problem

$$\begin{aligned} &\Phi^*,\ \Theta^*,\ \Psi^* \\ =\ &\operatorname*{argmin}_{\Phi,\Theta,\Psi} \mathrm{E}_x \left[\left(\log_2 \frac{1}{P_\Theta(\hat{z}_\Phi(x))} \right) + \lambda ||x - \hat{x}_\Psi(\hat{z}_\Phi(x))||^2 \right] \end{aligned}$$

Issue: The rounding causes the gradients for Φ to be zero.

Modeling Rounding with Noise

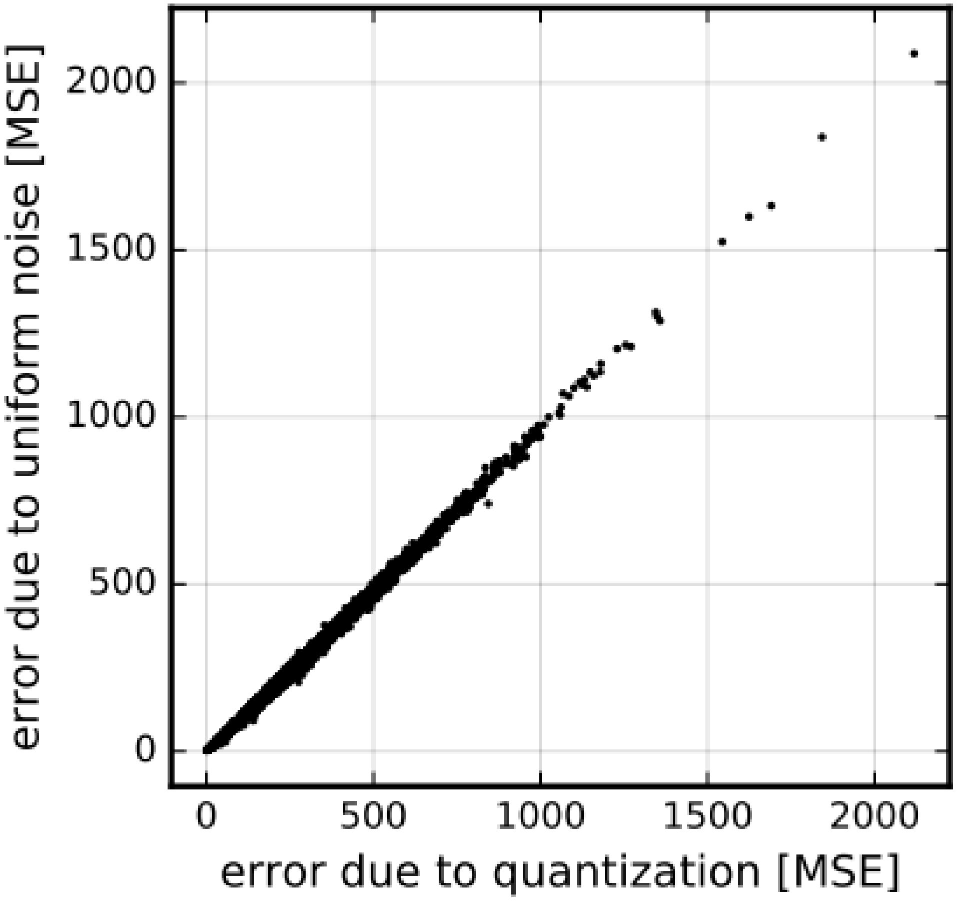
At train time (but not test time) the rounding is replaced with additive noise.

$$\begin{aligned} &\Phi^*,\ \Theta^*,\ \Psi^* \\ =\ &\operatorname*{argmin}_{\Phi,\Theta,\Psi} \mathrm{E}_{x,\epsilon} \left[\left(\log_2 \frac{1}{P_\Theta(z_\Phi(x)+\epsilon)}\right) + \lambda||x - \hat{x}_\Psi(z_\Phi(x)+\epsilon)||^2\right] \end{aligned}$$

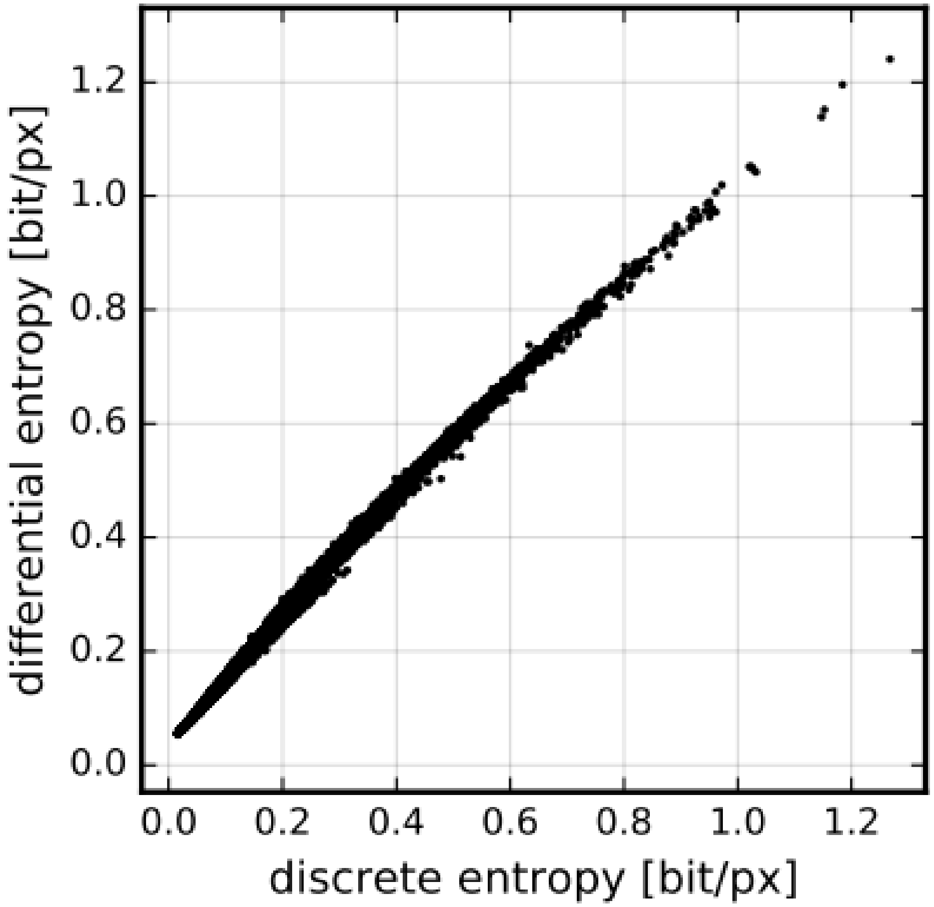
ϵ_i drawn uniformly from $[-1/2,\ 1/2]$

P_Θ defines a piecewise linear density for each coordinate of z.

Noise vs. Rounding



Differential Entropy vs. Discrete Entropy



Varying the Level Of Compression

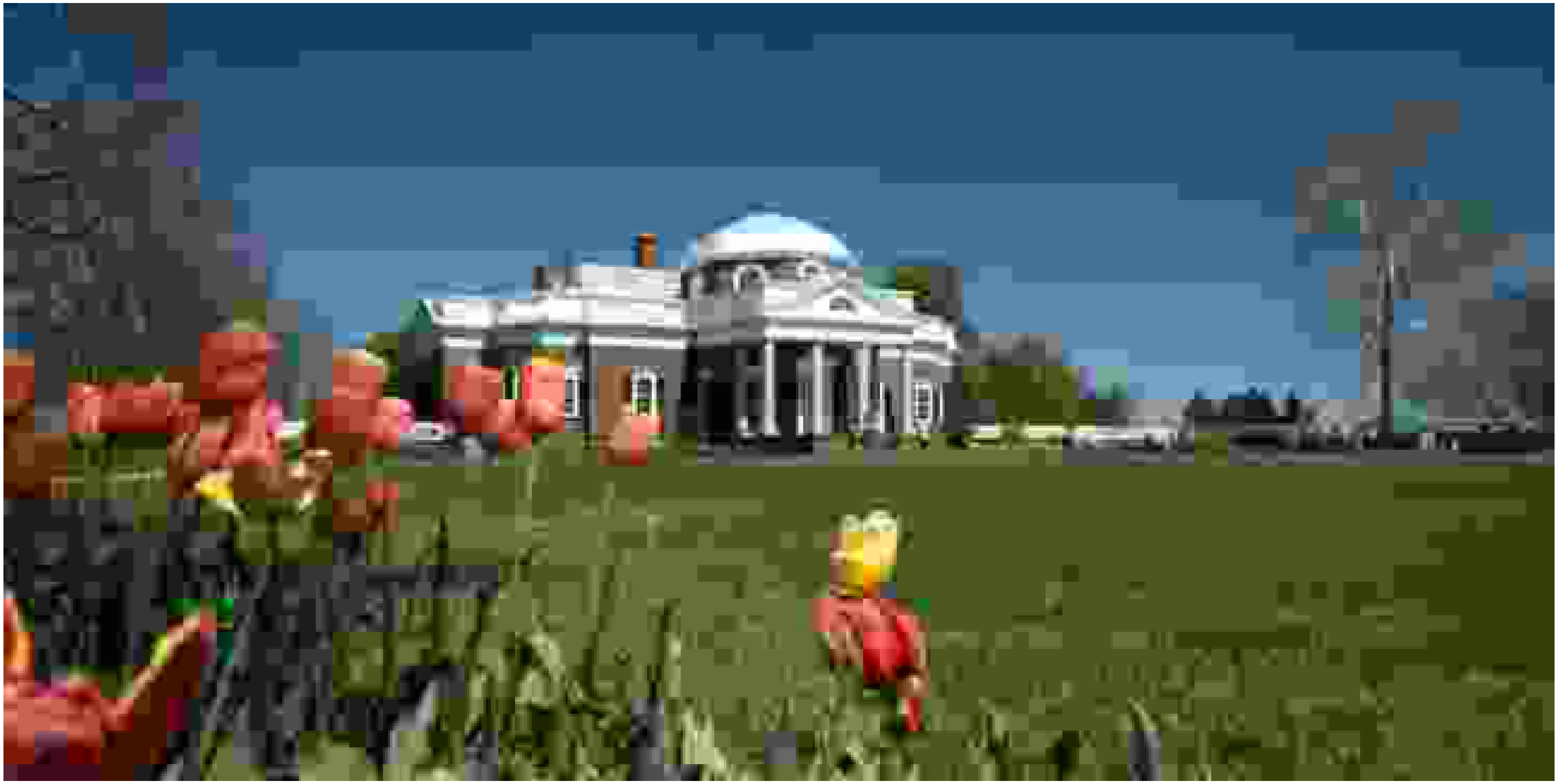
$$\begin{aligned}&\Phi^*,\ \Theta^*,\ \Psi^*\\ =\ &\mathop{\mathrm{argmin}}_{\Phi,\Theta,\Psi} \mathrm{E}_{x,\epsilon}\left[\left(\log_2 \frac{1}{P_\Theta(z_\Phi(x)+\epsilon)}\right) + \lambda||x - \hat{x}_\Psi(z_\Phi(x)+\epsilon)||^2\right]\end{aligned}$$

Different levels of compression correspond to different values of λ.

In all levels of compression we replace 768 numbers by 192 numbers.

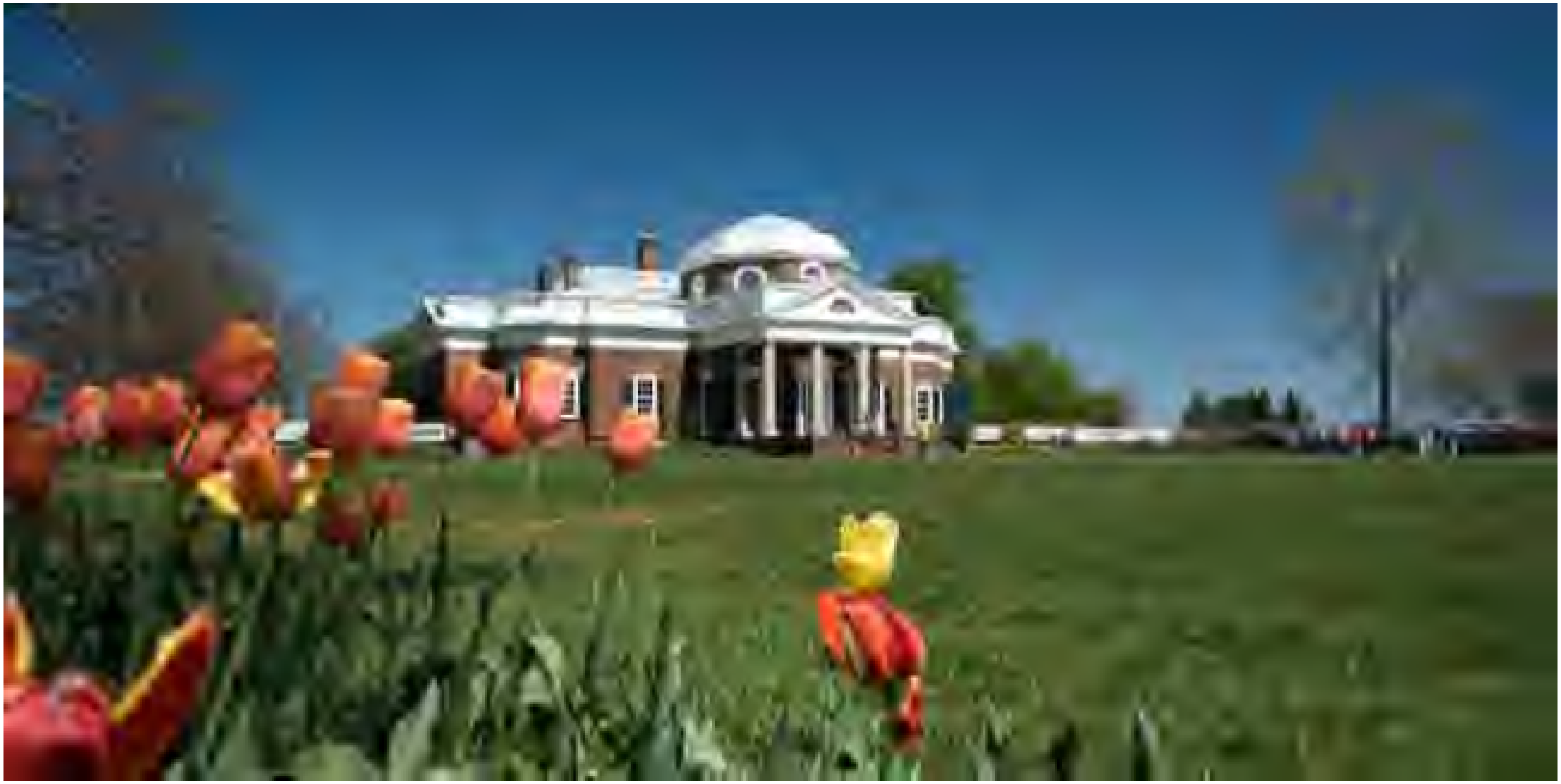
Higher levels of compression result in more compact distributions on the 192 numbers.

JPEG at 4283 bytes or .121 bits per pixel



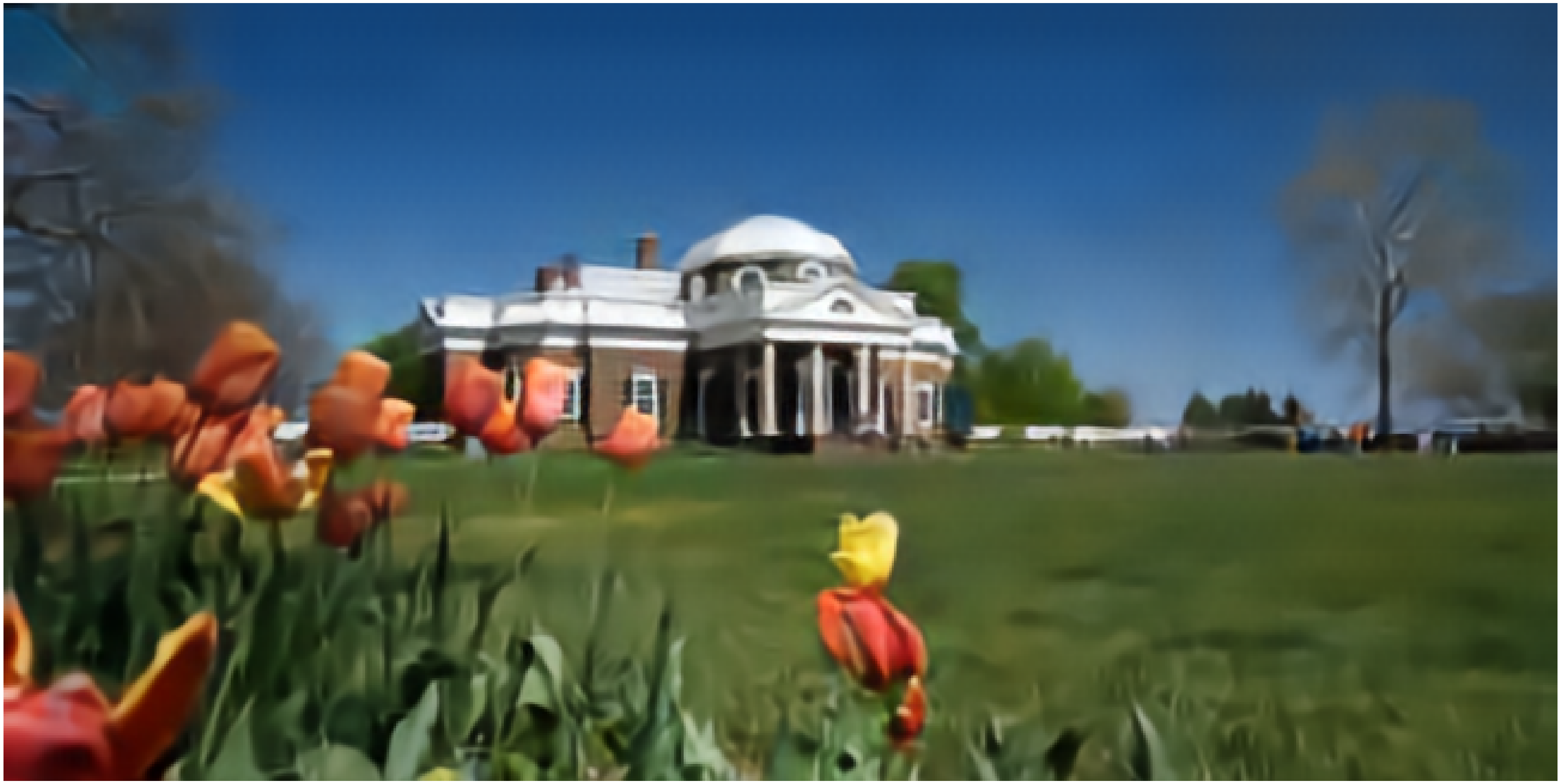
JPEG, 4283 bytes (0.121 bit/px), PSNR: 24.85 dB/29.23 dB, MS-SSIM: 0.8079

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

Proposed Method 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

Rate-Distortion vs. Variational Autoencoders

$$
\begin{aligned}
& \Phi^*,\ \Psi^*,\ \Theta^* \\
= \ & \operatorname*{argmin}_{\Phi,\Psi,\Theta} \mathrm{E}_x \left[\left(\log \frac{1}{P_\Theta^{\mathrm{code}}(z_\Phi(x))}\right) + \lambda L(x, \hat{x}_\Psi(z_\Phi(x)))\right] \\
= \ & \operatorname*{argmin}_{\Phi,\Psi,\Theta} \mathrm{E}_x \left[\left(\log \frac{1}{P_\Theta^{\mathrm{gen}}(z_\Phi(x))}\right) + \log \frac{1}{P_\Psi^{\mathrm{dec}}(x|z_\Phi(x))}\right] \\
= \ & \operatorname*{argmin}_{\Phi,\Psi,\Theta} \mathrm{E}_x \left[\left(\log \frac{1}{P_\Theta^{\mathrm{gen}}(z_\Phi(x))P_\Psi^{\mathrm{dec}}(x|z_\Phi(x))}\right)\right] \\
= \ & \operatorname*{argmin}_{\Phi,\Psi,\Theta} \mathrm{E}_{x\sim D, z\sim P_\Phi^{\mathrm{enc}}(\cdot|x)} \left[\left(\log \frac{1}{P_\Theta^{\mathrm{gen}}(z)P_\Psi^{\mathrm{dec}}(x|z)}\right)\right] - H(P_\Phi^{\mathrm{enc}}(\cdot|x))
\end{aligned}
$$

Rate-Distortion vs. Variational Autoencoders

$$
\begin{aligned}
& \operatorname*{argmin}_{\Phi,\Psi,\Theta} \mathrm{E}_{x\sim D,z\sim P_\Phi^{\mathrm{enc}}(\cdot|x)} \left[\left(\log \frac{1}{P_\Theta^{\mathrm{gen}}(z)P_\Psi^{\mathrm{dec}}(x|z)}\right)\right] - H(P_\Phi^{\mathrm{enc}}(\cdot|x)) \\
= \; & \operatorname*{argmin}_{\Phi,\Psi,\Theta} \mathrm{E}_x \left[\log \; \frac{1}{P_{\Theta,\Psi}(x)} + KL(P_\Phi^{\mathrm{enc}}(z|x), P_{\Theta,\Psi}(z|x))\right]
\end{aligned}
$$

In the Rate-distortion autoencoder $P_\Phi^{\mathrm{enc}}(z|x)$ is deterministic and the KL divergence term cannot be driven to zero for rates less then $H(x)$.

We should avoid interpreting the distortion term as $\log(1/P_\Psi^{\mathrm{dec}}(x|z))$.

END