Learned transform compression with optimized entropy encoding

Magda Gregorová

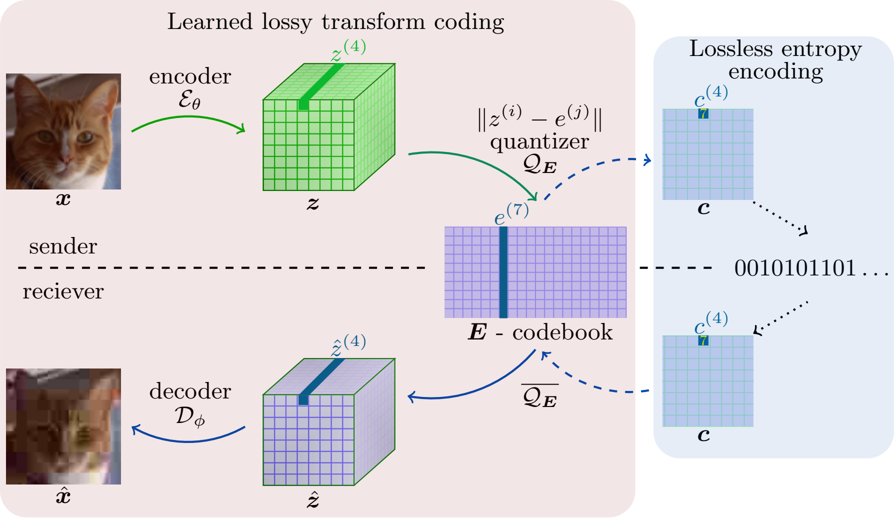
DMML workshop 6 July 2021, Geneva

In collaboration with:
Marc Desaules & Alexandros Kalousis

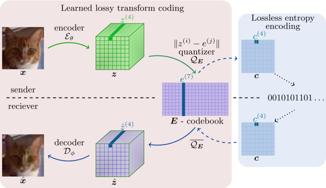
Hes·so
Haute Ecole Spécialisée
de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts
Western Switzerland

dmml group

Transform coding with vector quantization



End-to-end optimized compression



Learn $\mathcal{E}_\theta, \mathcal{D}_\phi, \mathcal{Q}_{\boldsymbol{E}}$ by minimizing

$$\mathcal{L} := \underbrace{\mathbb{E}_{\mu_x} d(\mathtt{x}, \hat{\mathtt{x}})}_{\textit{distortion}} + \lambda \underbrace{\mathbb{E}_{\mu_c} l(\mathtt{c})}_{\textit{rate}} \qquad \text{(trade-off)}$$

data $\mathtt{x} \sim \mu_x$ symbols $\mathtt{c} \sim \mu_c$
(unknown probability measures)

reconstruction error:

$$d(\mathtt{x}, \hat{\mathtt{x}}) = \|\mathtt{x} - \hat{\mathtt{x}}\|_2^2$$

(or ℓ_1, MS-SSIM, . . .)

length of binary encoding:

$$l(c) = -\log p_c(c) \qquad \text{(Shannon)}$$

$$\text{pmf: } \int_A p_c \, \mathrm{d}\# = \sum_{\boldsymbol{a} \in \boldsymbol{A}} p_c(\boldsymbol{a}) = \mu_c(\boldsymbol{A})$$

rate = entropy: $\mathbb{E}_{\mu_c} l(c) = -\mathbb{E}_{\mu_c} \log p_c(\mathtt{c}) = \mathbb{H}_{\mu_c}(\mathtt{c})$

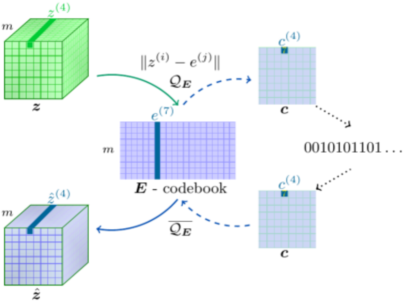
unknown $p_c \Rightarrow$ cannot evaluate $\mathbb{H}_{\mu_c}(\mathtt{c}) \Rightarrow$ replace by estimate $q_c \approx p_c$

rate ≈ cross-entropy: $\mathbb{E}_{\mu_c} l(c) \approx -\mathbb{E}_{\mu_c} \log q_c(\mathtt{c}) = \mathbb{H}_{\mu_c | q_c}(\mathtt{c})$

$$\mathcal{E}_\theta, \mathcal{D}_\phi, \mathcal{Q}_{\boldsymbol{E}}, \mathcal{P}_\psi \qquad \mathcal{L} := \underbrace{\mathbb{E}_{\mu_x} d(\mathtt{x}, \hat{\mathtt{x}})}_{\textit{distortion}} + \lambda \underbrace{\mathbb{H}_{\mu_c | q_c}(\mathtt{c})}_{\textit{rate}}$$

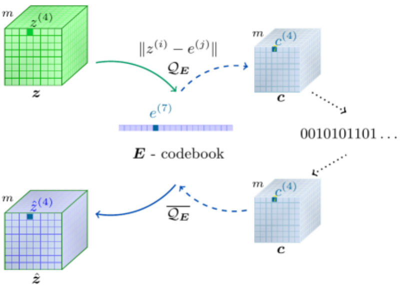
Vector quantization

Vector quantization



message length: d^2

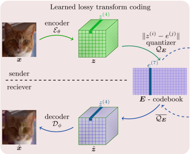
Scalar quantization



message length: $d^2 m$

$$\mathcal{Q}_{\boldsymbol{E}}: \quad \hat{\boldsymbol{z}}^{(i)} = \arg\min_{\boldsymbol{e}^{(j)}} \|\boldsymbol{z}^{(i)} - \boldsymbol{e}^{(j)}\| \qquad c^{(i)} = \{j : \hat{\boldsymbol{z}}^{(i)} = \boldsymbol{e}^{(j)}\}$$

Model learning - problems



1) Non-differentiability of quantization operation

Forward: $\boldsymbol{x} \xrightarrow{\mathcal{E}_\theta} \boldsymbol{z} \xrightarrow{\mathcal{Q}_{\boldsymbol{E}}} \hat{\boldsymbol{z}} \xrightarrow{\mathcal{D}_\phi} \hat{\boldsymbol{x}} \longrightarrow d(\boldsymbol{x}, \hat{\boldsymbol{x}})$

Backward: $\boxed{\cancel{\boldsymbol{x} \xleftarrow{\nabla_\theta} \nabla_{\boldsymbol{z}}}} \xleftarrow{\nabla_{\boldsymbol{E}}} \nabla_{\hat{\boldsymbol{z}}} \xleftarrow{\nabla_\phi} \nabla_{\hat{\boldsymbol{x}}} \longleftarrow d(\boldsymbol{x}, \hat{\boldsymbol{x}})$

2) Cross-entropy minimization does not minimize rate

$$\mathbb{H}_{\mu_c|q_c}(\mathsf{c}) = -\mathbb{E}_{\mu_c} \log q_c(\mathsf{c}) = \overbrace{D_{\mathrm{KL}}(p_c \| q_c)}^{\geq 0} + \mathbb{H}_{\mu_c}(\mathsf{c})$$

$$\min_{q_c} \ \mathbb{H}_{\mu_c|q_c}(\mathsf{c}) \ \Leftrightarrow \ \min_{q_c} \ D_{\mathrm{KL}}(p_c \| q_c)$$

$\mathbb{H}_{\mu_c}(\mathsf{c})$ not function of q_c so not optimized $\Rightarrow$ rate not optimized

Solutions - i), ii), iii)

i) soft quantization for backward gradients

forward hard (non-differentiable):

$$p_z(\hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)}|\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{e}^{(j)} = \arg\min_{\boldsymbol{e}^{(i)} \in \boldsymbol{E}} \|\boldsymbol{z}(\boldsymbol{x}) - \boldsymbol{e}^{(i)}\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{\boldsymbol{z}}(\boldsymbol{x}) = \sum_{\boldsymbol{e}^{(j)} \in \boldsymbol{E}} p_z(\hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)}|\boldsymbol{x})\, \boldsymbol{e}^{(j)}$$

backward soft (differentiable):

$$\hat{p}_z(\hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)}|\boldsymbol{x}) = \frac{\exp(-\sigma\|\boldsymbol{z} - \boldsymbol{e}^{(j)}\|)}{\sum_i^k \exp(-\sigma\|\boldsymbol{z} - \boldsymbol{e}^{(i)}\|)}$$

$$\tilde{\boldsymbol{z}}(\boldsymbol{x}) = \sum_{\boldsymbol{e}^{(j)} \in \boldsymbol{E}} \hat{p}_z(\hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)}|\boldsymbol{x})\, \boldsymbol{e}^{(j)}$$

Solutions - i), ii), iii)

ii) pushforward measure

μ_x - unkown & fixed, μ_c pushforward $f_*(\mu_x), f = \mathcal{Q}_{\boldsymbol{E}} \circ \mathcal{E}_\theta$ - unkown & not fixed

change $\mathcal{E}_\theta$ and $\mathcal{Q}_{\boldsymbol{E}}$ to change μ_c and hence rate $\mathbb{H}_{\mu_c}$

iii) soft cross-entropy

$$\mathbb{H}_{\mu_c|q_c}(\mathsf{c}) = -\mathbb{E}_{\mu_c} \log q_c(\mathsf{c}) = -\int_{\mathrm{x}} \sum_j p_c(\mathsf{c} = j|\boldsymbol{x}) \log q_c(\mathsf{c} = j) \, \mathrm{d}\mu_x$$

hard cross-entropy (no gradients to $\mathcal{E}_\theta$ and $\mathcal{Q}_{\boldsymbol{E}}$, no rate effect):

$$\mathbb{H}_{\mu_c|q_c}(\mathsf{c}) = h(\mathsf{c}) \approx -\frac{1}{n} \sum_i^n \sum_j p_c(\mathsf{c} = j|\boldsymbol{x}_i) \log q_c(\mathsf{c} = j), \quad p_c(\mathsf{c} = j|\boldsymbol{x}) = p_z(\hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)}|\boldsymbol{x})$$

soft cross-entropy (gradients to $\mathcal{E}_\theta$ and $\mathcal{Q}_{\boldsymbol{E}}$, rate effect):

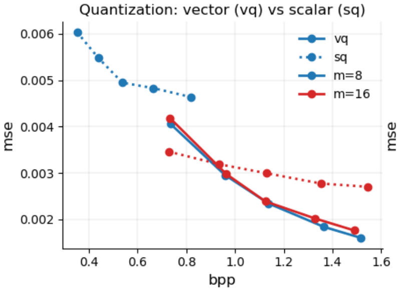
$$\mathbb{H}_{\mu_c|q_c}(\mathsf{c}) = s(\mathsf{c}) \approx -\frac{1}{n} \sum_i^n \sum_j \hat{p}_c(\mathsf{c} = j|\boldsymbol{x}_i) \log \mathrm{sg} q_c(\mathsf{c} = j), \quad \hat{p}_c(\mathsf{c} = j|\boldsymbol{x}) = \hat{p}_z(\hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)}|\boldsymbol{x})$$

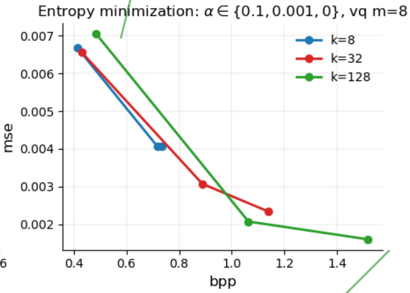
$$\boxed{\mathcal{L}(\mathcal{E}_\theta, \mathcal{D}_\phi, \mathcal{Q}_{\boldsymbol{E}}, \mathcal{P}_\psi) := \sum_i^n d(\mathrm{x}, \mathcal{D}_\phi[\mathrm{sg}(\hat{\boldsymbol{z}}_i - \tilde{\boldsymbol{z}}_i) + \tilde{\boldsymbol{z}}_i]) + \alpha s(\boldsymbol{c}_i) + \beta h(\boldsymbol{c}_i)}$$

Proof of concept - experiments

$\mathcal{E}_\theta, \mathcal{D}_\phi$: CNN, stride-2 down-/up-sampling, 64 kernels size 3-4, 10 residual blocks with skip connections

$$\mathcal{P}_\psi : q_c(\mathsf{c}) = \prod_i^{d^2} q_{c_i}(\mathsf{c}_i), \quad q_{c_i} = q_{c_j}$$





ADAM, one cycle cosine schedule

$\sigma = 1, \ \beta = 1$ Imagenet32

Future work ideas

i) technical improvements

ii) instance specific dictionary

iii) squeeze more from entropy

iv) BB-ANS for VQVAE

v) other random ideas

i) Technical improvements

Better probability model:

Current $q_c(\boldsymbol{c}) = \prod_i^d q_i(c_i), q_i = q_j \quad \Rightarrow \quad$ more complex model e.g. AR or IDF

Challenge: $q_c \approx p_c$ not fixed but evolving during training; training stability?

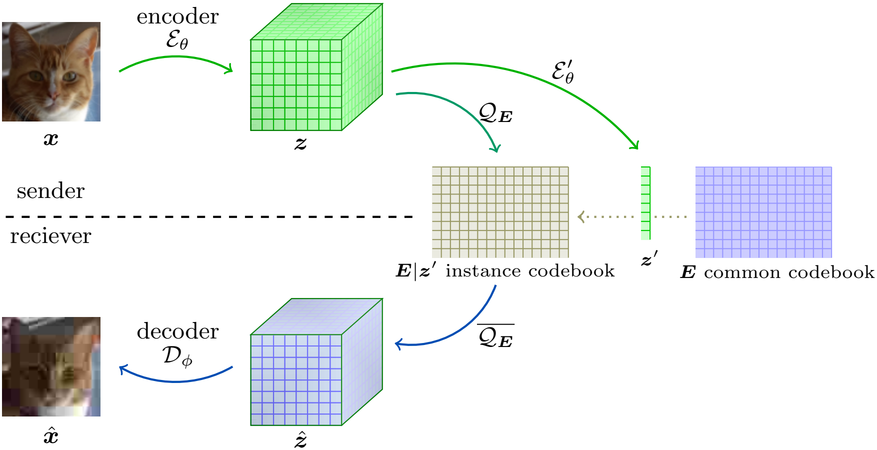
Backward gradient info:

soft relaxation vs streight-through vs. soft relaxation with anealing

Initialization of $\boldsymbol{E}$:

random uniform vs k-means++

ii) Instance specific dictionary



Idea: $\boldsymbol{E}|\boldsymbol{z}'$ better for specific instance $\boldsymbol{x}$ then generic $\boldsymbol{E}$
transmitt: $\boldsymbol{c}, \boldsymbol{z}'$ $\rightarrow$ trade-off size of $\boldsymbol{z}'$ vs $\boldsymbol{E}$

$\mathcal{C}_\xi(\boldsymbol{z}', \boldsymbol{E}) = \boldsymbol{E}|\boldsymbol{z}'$: architecture so that not ignoring $\boldsymbol{z}'$, complex vs constrained transformation (e.g. completely free vs only shuffle columns to improve entropy)

iii) Squeeze more from entropy

link to MAP, ELBO (VAE), more info theory?

iv) BB-ANS for VQVAE

Townsend (2019) BB-ANS: efficient lossless compression using VAE

open question - need to discretize latent z before encoding via ANS

van den Oord (2017) VQVAE: learned discretization via vector quantization

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{c}} p(\boldsymbol{x}|\boldsymbol{c})p(\boldsymbol{c}), \qquad \boldsymbol{c} \in \{0, 1, \ldots, K\}$$

$$\text{deterministic:} \quad q(\boldsymbol{c} = k|\boldsymbol{x}) = \begin{cases} 1 & \text{if } k = \arg\min_i \|\boldsymbol{z}(\boldsymbol{x}) - \boldsymbol{e}^{(i)}\|_2^2 \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow \quad D_{\mathrm{KL}}(q(\boldsymbol{c} = k|\boldsymbol{x})\|p(c)) = \log K \quad \Rightarrow$ can be dropped from loss

streight-through to backprob to $\boldsymbol{z}$ $\quad \Rightarrow$ needs k-means loss for $\boldsymbol{E}$ updates

fully deterministic scheme not amenable to BB

Proposed method:

$$\text{stochastic:} \quad q(\boldsymbol{c} = k|\boldsymbol{x}) = \frac{\exp(\|\boldsymbol{z}(\boldsymbol{x}) - \boldsymbol{e}^{(k)}\|_2^2)}{\sum_i^K \exp(\|\boldsymbol{z}(\boldsymbol{x}) - \boldsymbol{e}^{(i)}\|_2^2)}$$

$c \sim \quad q(\boldsymbol{c} = k|\boldsymbol{x})$ Gumbel soft-max etc. $\quad D_{\mathrm{KL}}(q(\boldsymbol{c} = k|\boldsymbol{x})\|p(c)) = -\mathbb{H}_{q(\boldsymbol{c}|\boldsymbol{x})} + \log K$

stochastic amenable to BB, What does it bring compared to conti latent?

note: $\quad \min -\mathbb{H}_{q(\boldsymbol{c}|\boldsymbol{x})}$ $\quad$ good for BB

v) Other random ideas

Side info:

Use side info (e.g. ABB meta-data to generate data and compress only the differences

$\Rightarrow$ major patterns covered by generations, compress only the irregularities (surprises)

Variable representation power:

sender has acces to full model so can evaluate the transmission error $\rightarrow$ if too big, compress less and vice versa

$\Rightarrow$ train multiple models (hierarchical, composable) with different rates and apply these selectively to different instances (e.g driven by $||\boldsymbol{z}_i - \mathcal{Q}(\boldsymbol{z}_i)||$)

Autoregressive dictionary:

current quantizer $\mathcal{Q}$ uses single $\boldsymbol{E}$ for quantizing all latent vectors $\boldsymbol{z}$

$\Rightarrow$ learn $\boldsymbol{E}^{(1)}$ to be used for $\boldsymbol{z}^{(1)}$ and $f : (\boldsymbol{E}^{(i)}, \mathcal{Q}(\boldsymbol{z}^{(i)}) \rightarrow \boldsymbol{E}^{(i+1)}$ to be used for $\boldsymbol{z}^{(i)}$

References

Agustsson, E., Mentzer, F., Tschannen, M., Cavigelli, L., Timofte, R., Benini, L., & Van Gool, L. (2017). *"Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations."* arXiv:1704.00648.

Balle, J., Laparra, V. & Simoncelli, E. P. (2017). *"End-to-end Optimized Image Compression."* ICLR.

Cover, T. M. & Thomas, T. M. (2006). *"Elements of Information Theory."* Wiley.

Habibian, A., van Rozendaal, T. Tomczak, J. M., & Cohen, T. S. (2019). *"Video Compression With Rate-Distortion Autoencoders."* ICCV.

Mentzer, F., Agustsson, F., Tschannen, M., Timofte, R., Van Gool, L. (2018). *"Conditional Probability Models for Deep Image Compression."* CVPR.

Sayood, K. (2012). *"Introduction to Data Compression."* Elsevier

Theis, L., Shi, W., Cunningham, A. & Huszar, F. (2017). *"Lossy Image Compression with Compressive Autoencoders."* ICLR.

van den Oord, A., Vinyals, O. & Kavukcuoglu, K. (2017). *"Neural Discrete Representation Learning."* NeurIPS.

Williams, W., Ringer, S., Ash, T., Hughes, J., MacLeod, D. & Dougherty, J. (2020). *"Hierarchical Quantized Autoencoders."* NeurIPS.

Townsend, J., Bird, T. & Barber, D. (2019) *"Practical Lossless Compression with Latent Variables using Bits Back Coding"* ICLR.