Learned transform compression with optimized entropy encoding

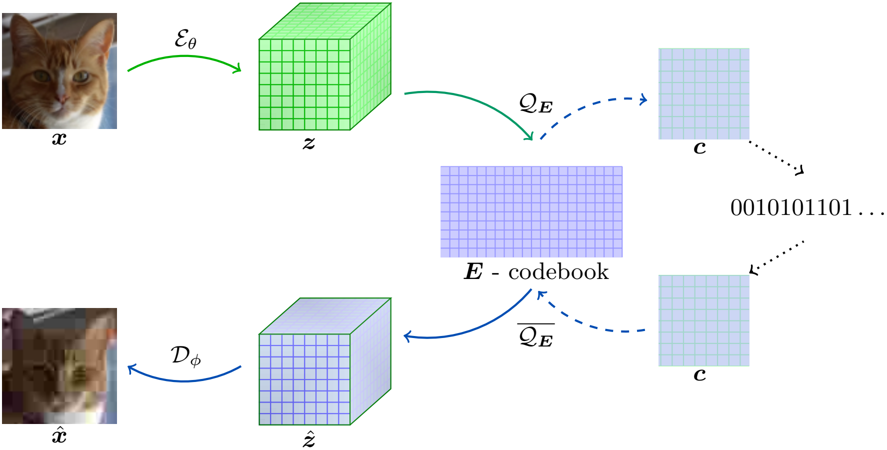
Magda Gregorová

26 March 2021, Würzburg, Germany

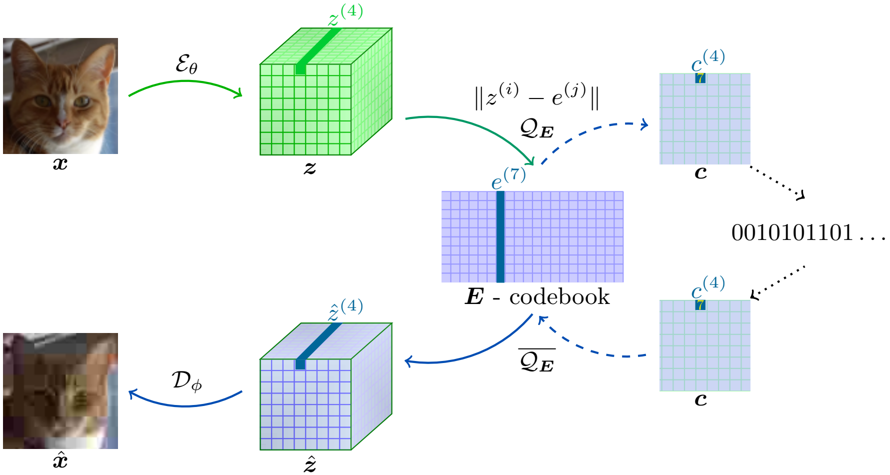
In collaboration with:
Marc Desaules & Alexandros Kalousis

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

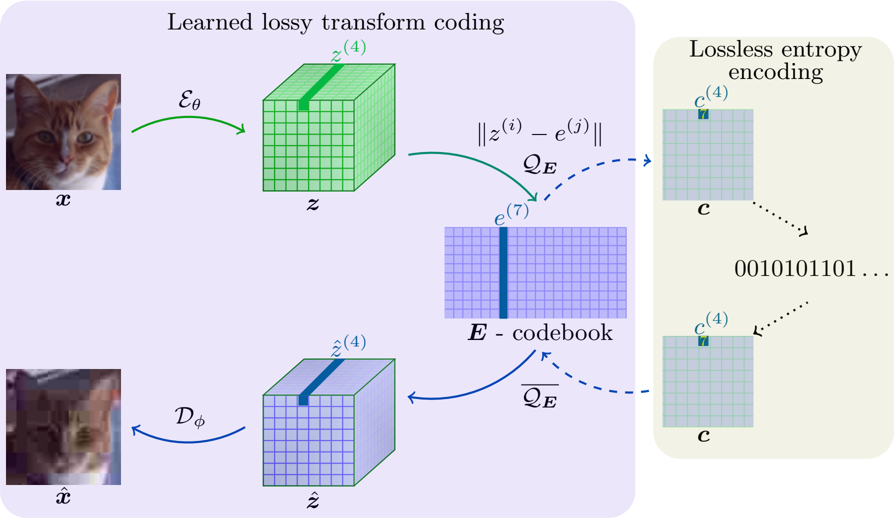
Transform coding with vector quantization



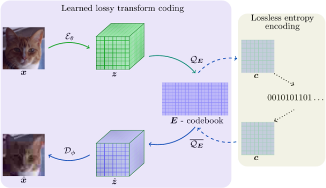
Transform coding with vector quantization



Transform coding with vector quantization



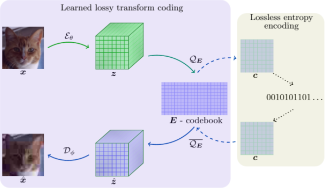
End-to-end optimized compression



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

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

End-to-end optimized compression



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

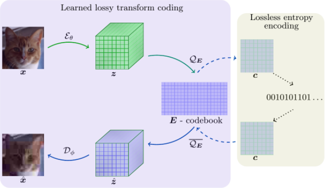
$$\mathcal{L} := \underbrace{\mathbb{E}_{\mu_x} d(\mathtt{x}, \hat{\mathtt{x}})}_{distortion} + \lambda \underbrace{\mathbb{E}_{\mu_c} l(\mathsf{c})}_{rate}$$

Shannon: $l^*(c) = -\log p_c(c) \implies \mathbb{E}_{\mu_c} l^*(c) = -\mathbb{E}_{\mu_c} \log p_c(\mathsf{c}) = \mathbb{H}_{\mu_c}(\mathsf{c})$

$$\int_A p_c \, \mathrm{d}\# = \sum_{\boldsymbol{a} \in \boldsymbol{A}} p_c(\boldsymbol{a}) = \mu_c(\boldsymbol{A}) \qquad \log p_c = ??? \quad \mathbb{H}_{\mu_c}(\mathsf{c}) = ???$$

$$q_c \approx p_c \qquad -\mathbb{E}_{\mu_c} \log q_c(\mathsf{c}) \approx -\mathbb{E}_{\mu_c} \log q_c(\mathsf{c}) \quad \mathbb{H}_{\mu_c|q_c}(\mathsf{c}) \approx \mathbb{H}_{\mu_c}(\mathsf{c})$$

End-to-end optimized compression



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

$$\mathcal{L} := \underbrace{\mathbb{E}_{\mu_x} d(\mathsf{x}, \hat{\mathsf{x}})}_{distortion} + \lambda \underbrace{\mathbb{E}_{\mu_c} l(\mathsf{c})}_{rate}$$

Shannon: $l^*(c) = -\log p_c(c) \implies \mathbb{E}_{\mu_c} l^*(c) = -\mathbb{E}_{\mu_c} \log p_c(\mathsf{c}) = \mathbb{H}_{\mu_c}(\mathsf{c})$

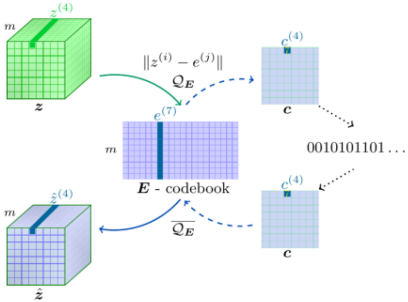
$$\int_A p_c \, \mathrm{d}\# = \sum_{\boldsymbol{a} \in \boldsymbol{A}} p_c(\boldsymbol{a}) = \mu_c(\boldsymbol{A}) \qquad \log p_c = ??? \quad \mathbb{H}_{\mu_c}(\mathsf{c}) = ???$$

$$q_c \approx p_c \qquad -\mathbb{E}_{\mu_c} \log q_c(\mathsf{c}) \approx -\mathbb{E}_{\mu_c} \log q_c(\mathsf{c}) \quad \mathbb{H}_{\mu_c|q_c}(\mathsf{c}) \approx \mathbb{H}_{\mu_c}(\mathsf{c})$$

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

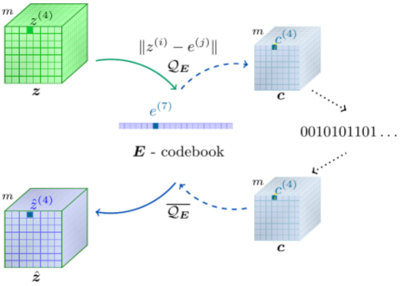
Vector quantization

Vector quantization



message length: d^2

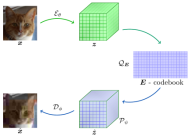
Scalar quantization



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

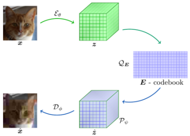


Problem $\mathcal{Q}_{\boldsymbol{E}}$: quantization non-differentiable

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

Model learning - problems



Problem $\mathcal{Q}_{\boldsymbol{E}}$: quantization non-differentiable

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

Problem $\mathbb{H}_{\mu_c}$: entropy not minimized

$q_c \approx p_c: \quad \mathbb{H}_{\mu_c|q_c}(\mathsf{c}) \geq \mathbb{H}_{\mu_c}(\mathsf{c}) \qquad \mathbb{H}_{\mu_c|q_c}(\mathsf{c}) = \overbrace{D_{\mathrm{KL}}(p_c \| q_c)}^{\geq 0} + \mathbb{H}_{\mu_c}(\mathsf{c})$

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

$q_c \to p_c \quad p_c$ fixed

Push-forward measure & soft-quantization

1) $\mu_c[\mathtt{c} \in \boldsymbol{A}] = \mu_c[\mathcal{T}_{\boldsymbol{E},\theta}(\mathtt{x}) \in \boldsymbol{A}] = \mu_x[\mathtt{x} \in \mathcal{T}^{-1}_{\boldsymbol{E},\theta}(\boldsymbol{A})]$ $\quad \mathcal{T}_{\boldsymbol{E},\theta} = \mathcal{Q}_{\boldsymbol{E}} \circ \mathcal{E}_\theta$

$$\mathcal{Q}_{\boldsymbol{E}}, \mathcal{E}_\theta \to \mu_c \to \mathbb{H}_{\mu_c}$$

Push-forward measure & soft-quantization

1) $\mu_c[\mathtt{c} \in \boldsymbol{A}] = \mu_c[\mathcal{T}_{\boldsymbol{E},\theta}(\mathtt{x}) \in \boldsymbol{A}] = \mu_x[\mathtt{x} \in \mathcal{T}^{-1}_{\boldsymbol{E},\theta}(\boldsymbol{A})]$ $\quad \mathcal{T}_{\boldsymbol{E},\theta} = \mathcal{Q}_{\boldsymbol{E}} \circ \mathcal{E}_\theta$

$$\mathcal{Q}_{\boldsymbol{E}}, \mathcal{E}_\theta \to \mu_c \to \mathbb{H}_{\mu_c}$$

2) $p_c(c = j) = \begin{cases} 1 & \text{if } \hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)} \\ 0 & \text{otherwise} \end{cases}$ $\quad h_{ce} = -\frac{1}{n} \sum_i^n \log q_c(c_i), \quad c_i \sim \mu_c$

$$\hat{p}_c(c = j) = \frac{\exp(-\sigma \| \boldsymbol{z} - \boldsymbol{e}^{(j)} \|)}{\sum_j^k \exp(-\sigma \| \boldsymbol{z} - \boldsymbol{e}^{(j)} \|)} \qquad s_{ce} = -\frac{1}{n} \sum_{i,j}^{n,k} \hat{p}_c(c_i = j) \log \mathrm{sg}[q_c(j)]$$

Push-forward measure & soft-quantization

1) $$\mu_c[\mathtt{c} \in \boldsymbol{A}] = \mu_c[\mathcal{T}_{\boldsymbol{E},\theta}(\mathtt{x}) \in \boldsymbol{A}] = \mu_x[\mathtt{x} \in \mathcal{T}^{-1}_{\boldsymbol{E},\theta}(\boldsymbol{A})] \qquad \mathcal{T}_{\boldsymbol{E},\theta} = \mathcal{Q}_{\boldsymbol{E}} \circ \mathcal{E}_\theta$$

$$\mathcal{Q}_{\boldsymbol{E}}, \mathcal{E}_\theta \to \mu_c \to \mathbb{H}_{\mu_c}$$

2) $$p_c(c = j) = \begin{cases} 1 & \text{if } \hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)} \\ 0 & \text{otherwise} \end{cases} \qquad h_{ce} = -\frac{1}{n}\sum_i^n \log q_c(c_i), \quad c_i \sim \mu_c$$

$$\hat{p}_c(c = j) = \frac{\exp(-\sigma\|\boldsymbol{z} - \boldsymbol{e}^{(j)}\|)}{\sum_j^k \exp(-\sigma\|\boldsymbol{z} - \boldsymbol{e}^{(j)}\|)} \qquad s_{ce} = -\frac{1}{n}\sum_{i,j}^{n,k} \hat{p}_c(c_i = j) \log \mathrm{sg}[q_c(j)]$$

3) $$\tilde{\boldsymbol{z}} = \sum_j^k \hat{p}_c(c = j)\, \boldsymbol{e}^{(j)} \qquad \hat{d}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = d(\boldsymbol{x}, \mathcal{D}_\phi[\mathrm{sg}(\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}) + \tilde{\boldsymbol{z}}])$$

Push-forward measure & soft-quantization

1) $\mu_c[\mathsf{c} \in \boldsymbol{A}] = \mu_c[\mathcal{T}_{\boldsymbol{E},\theta}(\mathtt{x}) \in \boldsymbol{A}] = \mu_x[\mathtt{x} \in \mathcal{T}^{-1}_{\boldsymbol{E},\theta}(\boldsymbol{A})] \qquad \mathcal{T}_{\boldsymbol{E},\theta} = \mathcal{Q}_{\boldsymbol{E}} \circ \mathcal{E}_\theta$

$$\mathcal{Q}_{\boldsymbol{E}}, \mathcal{E}_\theta \to \mu_c \to \mathbb{H}_{\mu_c}$$

2) $p_c(c = j) = \begin{cases} 1 & \text{if } \hat{\boldsymbol{z}} = \boldsymbol{e}^{(j)} \\ 0 & \text{otherwise} \end{cases} \qquad h_{ce} = -\frac{1}{n}\sum_i^n \log q_c(c_i), \quad c_i \sim \mu_c$

$$\hat{p}_c(c = j) = \frac{\exp(-\sigma\|\boldsymbol{z} - \boldsymbol{e}^{(j)}\|)}{\sum_j^k \exp(-\sigma\|\boldsymbol{z} - \boldsymbol{e}^{(j)}\|)} \qquad s_{ce} = -\frac{1}{n}\sum_{i,j}^{n,k} \hat{p}_c(c_i = j) \log \mathrm{sg}[q_c(j)]$$

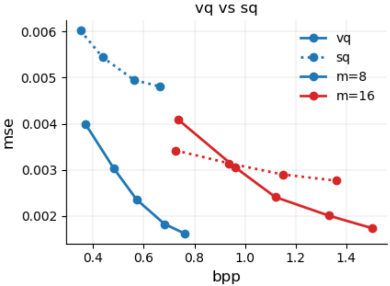
3) $\tilde{\boldsymbol{z}} = \sum_j^k \hat{p}_c(c = j)\, \boldsymbol{e}^{(j)} \qquad \hat{d}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = d(\boldsymbol{x}, \mathcal{D}_\phi[\mathrm{sg}(\hat{\boldsymbol{z}} - \tilde{\boldsymbol{z}}) + \tilde{\boldsymbol{z}}])$

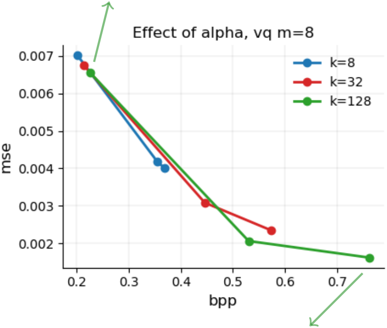
$$\boxed{\arg\min_{\mathcal{E}_\theta, \mathcal{Q}_{\boldsymbol{E}}, \mathcal{D}_\phi, \mathcal{P}_\psi} \frac{1}{n}\sum_i^n \hat{d}(\boldsymbol{x}_i, \hat{\boldsymbol{x}}_i) + \alpha\, s_{ce}(\boldsymbol{c}_i) + \beta\, h_{ce}(\boldsymbol{c}_i), \quad \boldsymbol{x}_i \sim \mu_x}$$

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

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.