Autoencoder based image compression: can the learning be quantization independent?
Thierry Dumas, Aline Roumy, Christine Guillemot

To cite this version:
Thierry Dumas, Aline Roumy, Christine Guillemot. Autoencoder based image compression: can the learning be quantization independent?. ICASSP 2018 - International Conference on Acoustics, Speech and Signal Processing ICASSP, Apr 2018, Calgary, Canada. pp.1188-1192, 10.1109/ICASSP.2018.8462263. hal-01713644

HAL Id: hal-01713644
https://hal.archives-ouvertes.fr/hal-01713644
Submitted on 22 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
AUTOENCODER BASED IMAGE COMPRESSION: CAN THE LEARNING BE QUANTIZATION INDEPENDENT?

Thierry Dumas, Aline, Roumy and Christine Guillemot

INRIA Rennes Bretagne-Atlantique
thierry.dumas@inria.fr, aline.roumy@inria.fr, christine.guillemot@inria.fr

ABSTRACT
This paper explores the problem of learning transforms for image compression via autoencoders. Usually, the rate-distortion performances of image compression are tuned by varying the quantization step size. In the case of autoencoders, this in principle would require learning one transform per rate-distortion point at a given quantization step size. Here, we show that comparable performances can be obtained with a unique learned transform. The different rate-distortion points are then reached by varying the quantization step size at test time. This approach saves a lot of training time.

Index Terms— Image compression, deep autoencoders, quantization.

1. INTRODUCTION
Image coding standards all use linear and invertible transforms to convert an image into coefficients with low statistical dependencies, i.e. suited for scalar quantization. Notably, the discrete cosine transform (DCT) is the most commonly used for two reasons: 
(i) it is image-independent, implying that the DCT does not need to be transmitted, 
(ii) it approaches the optimal orthogonal transform in terms of rate-distortion, assuming that natural images can be modeled by zero-mean Gaussian-Markov processes with high correlation [1]. Deep autoencoders have been shown as promising tools for finding alternative transforms [2, 3, 4]. Autoencoders learn the encoder-decoder non-linear transform from natural images.

In the best image compression algorithms based on autoencoders [5, 6, 7], one transform is learned per rate-distortion point at a given quantization step size. Then, the quantization step size remains unchanged at test time so that the training and test conditions are identical. By contrast, image coding standards implement adaptive quantizations [8, 9]. Should the quantization be imposed during the training? To answer this, we propose an approach where the transform and the quantization are learned jointly. Then, we investigate whether, at test time, the compression falls apart when the co-efficients obtained with the learned transform are quantized using quantization step sizes which differ from those in the training stage. The code to reproduce our numerical results and train the autoencoders is available online¹.

Matrices and tensors are denoted by bold letters. $\|X\|_F$ is the Frobenius norm of $X$. $X \odot Z$ is the elementwise multiplication between $X$ and $Z$.

2. JOINT LEARNING OF THE TRANSFORM AND THE QUANTIZATION
Section 2 introduces an efficient autoencoder for image compression. Then, it details our proposal for learning jointly this autoencoder transform and the quantization.

2.1. Autoencoder for image compression
An autoencoder is a neural network with an encoder $g_e$, parametrized by $\theta$, that computes a representation $\hat{Y}$ from the data $X$, and a decoder $g_d$, parametrized by $\phi$, that gives a reconstruction $\hat{X}$ of $X$, see Figure 1. Autoencoders can be used for denoising or dimensionality reduction. When it is used for compression, the representation is also quantized, leading to the new quantized representation $Y = Q(\hat{Y})$. If an autoencoder has fully-connected layers [10, 11, 12], the number of parameters depends on the image size. This implies that one autoencoder has to be trained per image size. To avoid this, an architecture without fully-connected layer is chosen. It exclusively comprises convolutional layers and non-linear operators. In this case, $Y \in \mathbb{R}^{h \times w \times m}$ is a set of $m$ feature maps of size $n = h \times w$, see Figure 1.

The basic autoencoder training minimizes the image reconstruction error [13]. In order to create a rate-distortion optimization, the authors in [6] add the minimization of the entropy of the quantized representation. Moreover, a bit allocation is performed by learning a normalization for each feature map of $Y$. The encoder followed by the normalizations at the encoder side, parametrized by $\varphi_e$, are denoted $\overline{g}_e(\cdot; \theta, \varphi_e)$. Similarly, the normalizations at the decoder side, parametrized by $\varphi_d$, followed by the decoder are denoted $\overline{g}_d(\cdot; \varphi_d, \phi)$. Finally, this leads to (1).

¹www.irisa.fr/temics/demos/visualization_ae/visualizationAE.htm
We address the problem of optimizing the quantization step size for each feature map of $\mathbf{Y}$. Because of the quantization, the function to be minimized is an implicit function of the quantization step size $\{\delta_i\}_{i=1,...,m}$. The target is to make it an explicit function of $\{\delta_i\}_{i=1,...,m}$. For $q \in \{\ldots,-\delta_i,0,\delta_i,\ldots\}$,

$$\hat{p}_i(q) = \frac{1}{n} \sum_{j=1}^{n} \log_2(\hat{p}_i(\hat{y}_{ij})) + \gamma \in \mathbb{R}^+$$  \hspace{1cm} (2)

$$\hat{p}_i = p_i * l_i$$  \hspace{1cm} (3)

The $i$th matrix of $\mathcal{E} \in \mathbb{R}^{h \times w \times m}$ contains $n$ realizations $\{\varepsilon_{ij}\}_{j=1,...,n}$ of $\mathcal{E}_i$, $\mathcal{E}_i$ being a continuous random variable of probability density function $l_i$. In (3), the function to be minimized is differentiable with respect to $\theta$. $\theta$ can thus be learned via gradient-based methods. However, $\{\delta_i\}_{i=1,...,m}$ cannot yet be learned as the function to be minimized in (3) is not differentiable with respect to $\{\delta_i\}_{i=1,...,m}$. This is resolved using the change of variable $\mathcal{E}_i = \delta_i \mathcal{T}$ where $\mathcal{T}$ is a random variable following the uniform distribution of support $[-0.5, 0.5]$. Now, the minimization over $\{\delta_i\}_{i=1,...,m}$ is feasible, see (4).

$$\min_{\theta, \phi} \mathbb{E} \left[ \| \mathbf{X} - g_d(\mathcal{Q}_e(\mathbf{X}; \theta, \varphi_e)); \varphi_d, \phi \|_F^2 + \gamma \sum_{i=1}^{m} \mathbf{h}_i \right]$$  \hspace{1cm} (1)

$$\mathbf{h}_i = -\log_2(\hat{p}_i) - \frac{1}{n} \sum_{j=1}^{n} \log_2(\hat{p}_i(y_{ij} + \varepsilon_{ij}))$$  \hspace{1cm} (2)

$\mathbf{h}_i$ is feasible.

In the end, we end up with three groups of parameters: $\{\theta, \phi\}$, $\{\delta_i\}_{i=1,...,m}$ and $\{\psi_i\}_{i=1,...,m}$. These three groups are learned by alternating three different stochastic gradient descents. All the training heuristics are detailed in the code1.

Section 2 has developed an approach for learning explicitly the transform and a quantization step size for each feature map of $\mathbf{Y}$. Before evaluating this approach in Section 4, Section 3 studies what would happen if, at test time, the coefficients in $\mathbf{Y}$ are quantized using quantization step sizes that differ from those in the training stage. This first requires understanding the internal structure of $\mathbf{Y}$ after the training.

3. INSIDE THE LEARNED REPRESENTATION

This section studies the different feature maps of $\mathbf{Y}$ after the training. To this end, a deep convolutional autoencoder must first be built and trained. $\mathcal{Q}_e$ is the composition of a convolutional layer, a generalized divisive normalization (GDN) [15], a convolutional layer, a GDN and a convolutional layer.
After the training, a test set of 3,120 luminance images of size $256 \times 256$ is created from the Kodak suite. Here, $\mathbf{X}$ refers to a test luminance image. Figure 2 shows the normed histogram of the 50th feature map of $\mathbf{Y} = g_e (\mathbf{X}; \theta)$ and that of its 125th feature map, averaged over the test set. Every feature map of $\mathbf{Y}$, except the 90th, has a normed histogram similar to those displayed. To be more precise, let’s write the probability density function of the Laplace distribution with mean $\mu \in \mathbb{R}$ and scale $\lambda \in \mathbb{R}^*_+$, denoted $f(.; \mu, \lambda)$.

$g_d$ is the reverse composition, replacing each GDN with an inverse generalized divisive normalization (IGDN) [15] and each convolutional layer with a transpose convolutional layer [16]. It is important to stress that $m = 128$, $\mathbf{X}$ has one channel and the convolutional strides and paddings are chosen such that $h$ and $w$ are 16 times smaller than respectively the height and the width of $\mathbf{X}$. Therefore, the number of pixels in $\mathbf{X}$ is twice the number of coefficients in $\mathbf{Y}$. The training set contains 24000 luminance images of size $256 \times 256$ that are extracted from ImageNet [17]. The minimization is (4), $\gamma = 10000.0$. Note that, if a GDN was placed immediately after $g_e$, a IGDN was placed immediately before $g_d$ and, $\forall i \in [1, m]$, $\delta_i = 1.0$ was not learned, the autoencoder architecture and the training would correspond to [6].

### 3.1. Distribution of the learned representation

After the training, a test set of 24 luminance images of size $512 \times 768$ is created from the Kodak suite. Here, $\mathbf{X}$ refers to a test luminance image. Figure 2 shows the normed histogram of the 50th feature map of $\mathbf{Y} = g_e (\mathbf{X}; \theta)$ and that of its 125th feature map, averaged over the test set. Every feature map of $\mathbf{Y}$, except the 90th, has a normed histogram similar to those displayed. To be more precise, let’s write the probability density function of the Laplace distribution with mean $\mu \in \mathbb{R}$ and scale $\lambda \in \mathbb{R}^*_+$, denoted $f(.; \mu, \lambda)$.

$\forall i \in [1, m]$, $i \neq 90$, there exists $\mu_i \in \mathbb{R}$ and $\lambda_i \in \mathbb{R}^*_+$ such that $f(.; \mu_i, \lambda_i)$ fits well the normed histogram of the $i$th feature map of $\mathbf{Y}$. Note that most of the $m - 1$ scales belong to $[0.5, 2.0]$, see Figure 3. For transformed coefficients having a zero-mean Laplace distribution, [18] proves that a uniform reconstruction quantizer (URQ) with constant decision offsets approaches the optimal scalar quantizer in terms of squared-error distortion for any quantization step size. Yet, in our case, (i) the $m - 1$ Laplace probability density functions are not zero-mean, (ii) uniform scalar quantizers are used instead of this URQ. The point (i) is not problematic as an extra set of luminance images is used to compute an approximation $\tilde{\pi}_i \in \mathbb{R}$ of the mean of the $i$th feature map of $\mathbf{Y}$, then, at test time, the $i$th feature map of $\mathbf{Y}$ is centered via $\tilde{\pi}_i$ before being quantized. Note that $\{\tilde{\pi}_i\}_{i=1...m}$ does not depend on the test luminance images, thus incurring no transmission cost. Regarding the point (ii), it must be noted that the decoder mapping of the URQ is exactly the decoder mapping of the uniform scalar quantizer with same quantization step size. Since our case comes close to the requirements of the proof in [18], at test time, the rate-distortion trade-off should not collapse as the quantization step sizes deviate from the learned values. This will be verified in Section 4.

#### 3.2. Internal structure of the learned representation

The shortcoming of the previous fitting is that it does not reveal what information each matrix of $\mathbf{Y}$ encodes. To discover it, further visualizations are needed. The most common way of exploring a deep convolutional neural network (CNN) trained for image recognition is to look at the image, at the CNN input, resulting from the maximization over its pixels of a given neural activation in the CNN [19, 20, 21]. Precisely, [19, 20] maximize over the image pixels a given neural activation at the CNN output, i.e. a class probability. This shows what image features characterize this class according to the CNN. In our case, the maximization over the image pixels of a given coefficient in $\mathbf{Y}$ does not yield interpretable images. Indeed, the coefficients in $\mathbf{Y}$ are not bounded. This may explain why the maximization often returns saturated images.
Alternatively, the information the \( j \)th feature map of \( Y \) encodes, \( j \in [1,m] \), can be seen as follows. \( \forall i \in [1,m] \), all the coefficients in the \( i \)th feature map of \( Y \) are set to \( p_i \). This way, the feature maps of \( Y \) contain no significant information. Then, a single coefficient in the \( j \)th feature map of \( Y \) is set to \( \alpha \in \mathbb{R} \) and \( \hat{X} = g_d(Q(Y):\phi) \) is displayed. \( \alpha \) is selected such that it is near one of the two tails of the Laplace distribution of the \( j \)th feature map of \( Y \). Figure 4 shows the \( 64 \times 64 \) crop at the top-left of \( X \) when the single coefficient is located at the top-left corner of the \( j \)th feature map of \( Y \), \( j \in \{50,125\} \). We see that the 50th feature map of \( Y \) encodes a spatially localized image feature whereas its 250th feature map encodes a spatially extended image feature. Moreover, the image feature is turned into its symmetrical feature, with respect to the mean pixel intensity, by moving \( \alpha \) from the right tail of the Laplace distribution of the \( j \)th feature map of \( Y \) to the left tail. This linear behaviour is observed for each feature map of \( Y \).

It is interesting to see that, given the fitting in Section 3.1, \( Y \) is similar to the DCT coefficients for blocks of prediction error samples in H.265 [9] in terms of distribution. However, when looking at the information each feature map of \( Y \) encodes, \( Y \) has nothing to do with these DCT coefficients.

\section{4. EXPERIMENTS}

We now evaluate in terms of rate-distortion performances: (i) whether the way of learning the quantization matters, (ii) whether, at test time, it is efficient to quantize the coefficients obtained with the learned transform using quantization step sizes which differ from those in the training stage. This is done by comparing three cases.

The 1st case follows the approach in [6]. One transform is learned per rate-distortion point, the bit allocation being learned via the normalizations. In details, an autoencoder is trained for each \( \gamma \in S = \{10000.0, 12000.0, 16000.0, 24000.0, 40000.0, 72000.0, 96000.0\} \). During the training and at test time, the quantization step size is fixed to 1.0.

In the 2nd case, a unique transform is learned, the bit allocation being done by learning a quantization step size per feature map. More precisely, a single autoencoder is trained for \( \gamma = 10000.0 \) and \( \{d_i\}_{i=1,m} \) is learned, see Section 2. At test time, the rate varies as the quantization step sizes are equal to the learned quantization step sizes multiplied by \( \beta \in B = \{1.0, 1.25, 1.5, 2.0, 3.0, 4.0, 6.0, 8.0, 10.0\} \).

In the 3rd case, a unique transform is learned, the bit allocation being learned via the normalizations. In details, a single autoencoder is trained for \( \gamma = 10000.0 \) and, during the training, the quantization step size is 1.0. At test time, the rate varies as the quantization step size spans \( B \).

In the 2nd case, the autoencoder has the architecture described at the beginning of Section 3. In the 1st and 3rd case, a GDN is also placed after \( g_e \) and an IGDN is placed before \( g_d \). The autoencoders are trained on 24000 luminance images of size \( 256 \times 256 \) that are extracted from ImageNet. Then, at test time, the 24 luminance images from the Kodak suite are inserted into the autoencoders. The rate is estimated via the empirical entropy of the quantized coefficients, assuming that the quantized coefficients are i.i.d. Note that, for the 2nd and the 3rd case, we have also implemented a binarizer and a binary arithmetic coder to compress the quantized coefficients losslessly, see the code\(^1\). The difference between the estimated rate and the exact rate via the lossless coding is always smaller than 0.04 bpb. Figure 5 shows the rate-distortion curves averaged over the 24 luminance images. The JPEG2000 curve is obtained using ImageMagick. The H.265 [22] curve is computed via the version HM-16.15. There is hardly any difference between the 2nd and the 3rd case.

The 2nd and the 3rd case, we have also implemented a binarizer and a binary arithmetic coder to compress the quantized coefficients losslessly, see the code\(^1\). The difference between the estimated rate and the exact rate via the lossless coding is always smaller than 0.04 bpb. Figure 5 shows the rate-distortion curves averaged over the 24 luminance images. The JPEG2000 curve is obtained using ImageMagick. The H.265 [22] curve is computed via the version HM-16.15. There is hardly any difference between the 2nd and the 3rd case.

\section{5. CONCLUSION}

Using a unique transform learned via autoencoders and various quantization step sizes at test time, it is possible to compress as well as when learning one transform per rate-distortion point at a given quantization step size. Moreover, the learned transformed outperform other image compression algorithms based on transforms.
6. REFERENCES


