2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897363

SEMANTIC ALIGNMENT FOR MULTI-ITEM COMPRESSION

Tom Bachard [†] [⋆] *Anju Jose Tom* [⋆] *Thomas Maugey* [⋆]

[†] Univ Rennes, France
[⋆] INRIA, France

ABSTRACT

Coding algorithms usually compress independently the images of a collection, in particular when the correlation between them only resides at the semantic level, *i.e.,* information related to the high-level image content. In this work, we propose a coding solution able to exploit this semantic redundancy to decrease the storage cost of data collections. First we introduce the multi-item compression framework. Then we derive a loss term to shape the latent space of a variational auto-encoder so that the latent vectors of semantically identical images can be aligned. Finally, we experimentally demonstrate that this alignment leads to a more compact representation of the data collection.

***Index Terms*—** Multi image compression, Semantic, Variational AutoEncoder, Representation Learning

1. INTRODUCTION

The amount of data gathered, stored, and exchanged worldwide is getting bigger and bigger, with speed increasing every day. To face this “Tsunami of Data” – 2.5 quintillion bytes are created daily nowadays [1] –, efforts have been directed towards compression algorithms. Standard algorithms, such as VVC [2], achieve better and better compression rates thanks to efficient processing algorithms. Recent breakthroughs have also been achieved with deep-learning based approaches [3–5]. However, while achieving impressive gains, these algorithms are only designed to compress the data one by one, which can be severely limited when compressing entire collections of correlated images. In this work, we propose a new paradigm, called *multi-item compression (MIC)*, that aims at efficiently coding multiple items simultaneously by taking into account similarities among them.

We define and distinguish two types of similarity: *pixel based redundancy* and *semantic redundancy*. The former can be taken into account using, for example, multi-view approaches [6] or cloud-based image compression techniques [7, 8]. However, this kind of similarity is not necessarily present in most databases. The latter, on the other hand, is a higher level similarity that is not present at the pixel level. This kind of similarity is more related to the content (objects, feeling, concept, action, ...) of an image, that we, humans, would categorize the same even though the pixels do not match at all. To the best of authors' knowledge, this semantic similarity has never been taken into account for multi-item compression, whereas there exist more and more applications in which such type of inter-item similarity exists within a picture set (*e.g.,* social networks, personal image database). The proposed *MIC* precisely aims at taking into account this inter-item semantic redundancy in order to reduce the storage cost of such image collections.

For that purpose, we investigate a coding scheme based on variational auto-encoder (VAE) adapted to semantic correlation tracking. We first propose an original problem formulation, where the image data collection rate is measured as the joint entropy of the VAE latent vectors. We then prove that minimizing this joint entropy is equivalent to aligning the latent vectors. After arguing that this solution might not be efficient for reconstruction error, we then propose to take into account the semantic information given by the class to which the image pictures belong. More concretely, we propose to align only the semantically coherent items. Finally, we experimentally demonstrate that this enables to describe the image set in a more compact shape than classical independent coding.

2. MULTI-ITEM COMPRESSION

2.1. Definition

Multi-item compression (*MIC*) is a coding framework that aims at compressing $\mathcal{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)$, a collection of pictures, exploiting the redundancy between them. The efficiency of such coding scheme is measured both in terms of compression rate and reconstruction error.

Our *MIC* formulation is described in Figure 1. Suppose we have a collection of N images sharing common information, at a pixel or higher level. The encoder e builds a representation of the inputs in a latent space. We choose to encode each item individually so that new images can be added to the database without having to recompress every other image again. For each item of index i, we denote the intermediary latent vectors by $\mathbf{b}_i = e(\mathbf{X}_i)$. We additionally denote the column agglomerated latent matrix (for all items) by $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_N]$.

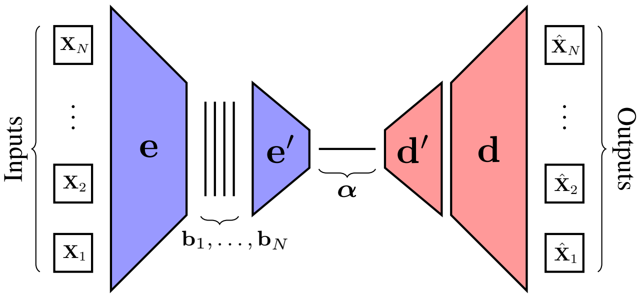


Fig. 1: Compression flow of our *MIC*.

Once each image has been compressed through the encoder e, we then use a classical inter-item lossless entropy encoder e' to achieve the final compressed bitstream. Its role is to compress the different item's latent vector into a bit stream $\boldsymbol{\alpha}$. The compression rate R for the model is then the length of the bit stream $R = |\boldsymbol{\alpha}|$. This entropy coding can be done using, simple coders (*e.g.,* arithmetic [9], Huffman [10]) or more evolved ones (*e.g.,* CABAC [11]).

Finally, the last step proposed in our *MIC* proposition is the decompression stage. First, the inter-item entropy decoder d' is applied. Then, each latent vector (corresponding to each item) is decoded with d. We denote the i^{th} decompressed images by $\hat{\mathbf{X}}_i = d \circ d' \circ e' \circ e(\mathbf{X}_i)$. At this point, we compute the PSNR values of the outputs to assess the quality of the codec.

2.2. MIC objective

The objective for our *MIC* formulation is given in Equation (1), where the reconstruction error is handled by an image-wise PSNR constraint associated with a threshold T. We explicit the rate R as the Shannon joint entropy of the latent vectors $H(\mathbf{b}_1, \ldots, \mathbf{b}_N)$. This further enables to remove e' and d' from the formulation.

$$\min_{e,d} \; H\big(\mathbf{b}_1, \ldots, \mathbf{b}_N\big) \;\; \text{s.t.} \qquad (1)$$
$$\forall i \in [\![1, N]\!], \text{PSNR}\big(\mathbf{X}_i, \hat{\mathbf{X}}_i\big) > T$$

Naturally, the aim is to track the inter-item correlation such that the achieved joint entropy is smaller than the sum of each entropy, as classical independent compression would reach as a rate, *i.e.,* $H(\mathbf{b}_1, \ldots, \mathbf{b}_N) < \sum_i H(\mathbf{b}_i)$. This is possible if the latent space is shaped such that semantically correlated items have aligned latent vectors. In the next Section, we explain how we propose to design such a latent space.

3. REDUCING THE LATENT SPACE JOINT ENTROPY

The gain of a *MIC* framework resides in the fact that correlation between latent vectors is present, which is generally not the case if the latent space is not specifically shaped for that. To achieve such a result, we design an encoder e that learns how to minimize this joint entropy. We now proceed to derive a loss function for neural networks optimizing the *MIC* criterion expressed in Equation (1).

3.1. Aligning in the latent Space

We first suppose that the distribution of the latent vectors follows a centered-multivariate Gaussian distribution $\mathcal{N} \sim (\mathbf{0}_{\mathcal{R}^N}, \mathbf{\Sigma_B})$, where $\mathbf{\Sigma_B}$ is the covariance matrix of the latent vectors. This is a classical assumption from which we explicitly know the differential entropy, and thus derive Equation (2).

$$H\left(\mathbf{b_1}, \ldots, \mathbf{b}_N\right) = \frac{1}{2}\log\left((2e\pi)^D|\mathbf{\Sigma_B}|\right) \qquad (2)$$

Because the true covariance of the multi-variate Gaussian distribution is not known, we approximate it with the Gram matrix, as in Equation (3).

$$\mathbf{\Sigma_B} \simeq \mathbf{G_B} \qquad (3)$$

Where $\mathbf{G_B} = \mathbf{B}^\top\mathbf{B}$, which gives, for each (i, j), $\mathbf{G_B}[i, j] = \langle \mathbf{b_i}, \mathbf{b_j} \rangle$. The optimization problem of Equation (1) becomes Equation (4).

$$\min_{e,d} \log\left(|\mathbf{G_B}|\right) \qquad (4)$$
$$\text{s.t.} \;\; \forall i \in [\![1,N]\!], \; \text{PSNR}\big(\mathbf{X}_i, \hat{\mathbf{X}}_i\big) > T$$

Solving Equation (4) directly would tend to align all the image representations in the latent space, *i.e.*, increasing the off-diagonal term of the Gram matrix $\mathbf{G_B}$ and simultaneously decreasing its determinant. Whereas the rate would indeed decrease, the expressiveness of the encoder e may however suffer, penalizing the reconstruction error.

3.2. Covariance matrix distance

Instead of aligning all the latent vectors (which is what minimizing the $\log\det$ does), we propose to align only those that are semantically coherent. For that purpose, we fix a target Gram matrix, $\mathbf{G}^*_\mathbf{B}$, whose $(i, j)^{th}$ entry is 1 if i^{th} and j^{th} items belong to the same semantic class and are 0 otherwise. Then, we define the loss as the distance between this Gram matrix and the one formed by the actual latent vectors $\hat{\mathbf{G}}_\mathbf{B}$. This way, we expect the latent space to be organized accordingly the semantics present in $\mathbf{G}^*_\mathbf{B}$. Said differently, items of the same class should have aligned latent vectors and items of a different class should have orthogonal latent vectors.

The distance between the two covariance matrices, d_{cov} is the distance proposed in [12], $\mathrm{d}_{cov}(A, B) = 1 - \frac{Tr(AB)}{\|A\|\|B\|}$, where Tr is the trace and $\|.\|$ the Frobenius norm. This is the distance giving the best empirical results for our model. Our final *MIC* formulation is given in Equation (5).

$$\min_{e,d} \ \mathrm{d}_{cov}\left(\hat{\mathbf{G}}_{\mathbf{B}}, \mathbf{G}^*_{\mathbf{B}}\right) \qquad (5)$$
$$\text{s.t. } \forall i \in [\![1,N]\!], \ \mathrm{PSNR}\big(\mathbf{X}_i, \hat{\mathbf{X}}_i\big) > T$$

Finally, from Equation (5) we derive a loss, Equation (6), that we use to train a neural network in order to perform *MIC*. The PSNR term is evaluated through a classical mean squared error (MSE) term. Note that in the loss, we introduce a λ parameter. This is because we later study the impact of the semantic part for a fixed value of the PSNR.

$$\mathcal{L}(\mathbf{X}) = \lambda \mathrm{d}_{cov}\left(\hat{\mathbf{G}}_{\mathbf{B}}, \mathbf{G}^*_{\mathbf{B}}\right) + \frac{1}{N}\sum_{i=1}^{N} MSE(\mathbf{X}_i, \hat{\mathbf{X}}_i) \qquad (6)$$

4. EXPERIMENTS

4.1. Experimental set-up

The architecture we use for the experiments corresponds to a *Variational Auto-Encoder* (VAE) proposed in [13] and is composed of three convolutional layers for the encoder e, and three de-convolutional layers for the decoder d, each layer has a 4×4 kernel, a stride of 2, and a padding of 1. The architecture is deliberately not as expressive as today's state of the art learning-based compression algorithms because on the one hand, the goal is to highlight the tendencies of the *MIC* framework and, on the other hand, to be sure that these tendencies come from our proposal and not from a complex architecture learning this on its own. Future work will focus on how to integrate our proposed intuitions in the most recent and performing architectures.

We test our VAE on the CIFAR dataset as a toy example. This dataset is made of 60 000 sample images homogeneously scattered in 10 classes. That way, the oracle matrix our loss will try to minimize is the one-hot matrix acknowledging that the i^{th} images are in the same class as the j^{th}. When showing experimental results, we restrain ourselves to a *toy database* of 2000 images randomly taken from the validation set of the data set. The class distributions in our toy database respect the ones from the original validation set.

In our simulations, we approximate the compression rate as the rank of the Gram matrix $\mathbf{G}_{\mathbf{B}}$ spanned by the database. this can be justified by the fact that the rank gives a good information on how compactly the $\mathbf{B}$ matrix could be represented. Computing the true compression rate is not relevant in the scope of this work as we are trying to prove that latent alignment is relevant for *MIC*.

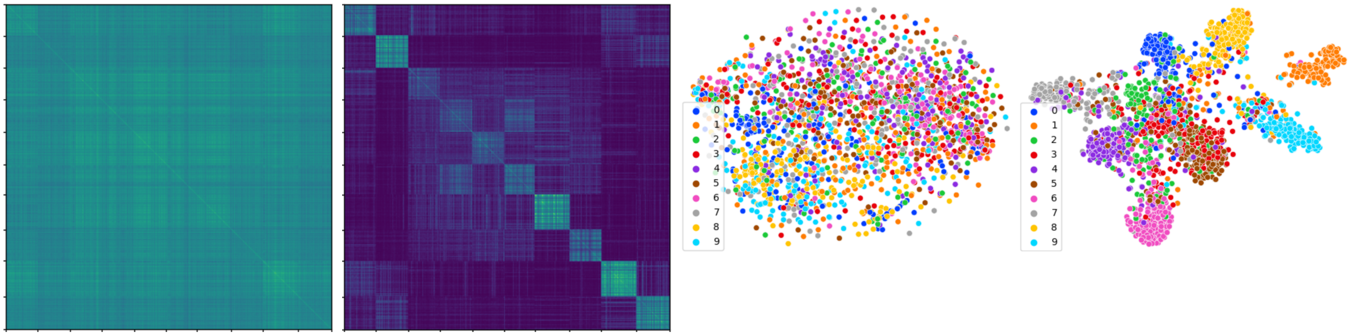
For comparison purposes, we train multiple VAE with different losses. First we train a classical VAE, *C-VAE*, with a classical rate-distortion-based loss (*i.e.* no d_{cov} term is used). This model serves as a baseline. The VAE trained with the semantic loss in Equation (6) is named *sMIC-VAE*. Also, to ensure that our results are due to a semantic alignment, and not to the fact of simply aligning latent vectors inside the latent space, we also train a VAE for which the oracle Gram matrix is random. We refer this model as *rMIC-VAE*. Finally, to prove that the sole minimization of Equation (4) is not sufficient to ensure a good rate-distortion cost, we also train a VAE that will learn to align all of the latent vectors. This model is *gMIC-VAE*, where "g" stands for "global" alignment.

4.2. Alignment in the latent space

We show in Figures 2a and 2b the Gram matrix (darker means less correlation and lighter more) for, respectively, the *C-VAE* and the *sMIC-VAE* on our toy database. We can clearly see that the *sMIC-VAE* Gram matrix has a more "block-diagonal" shape, which demonstrates the ability of the proposed loss term to align latent vectors within a given class. To confirm these observations, we also propose to project the vectors of $\mathbf{B}$ to a 2D dimensional space to observe the impact of our loss term on the latent space's shape. This projection is realized with the *t-SNE* algorithm [14]. Figures 2c and 2d are the *t-SNE* projection of the $\mathbf{B}$ matrix of *C-VAE* and the *sMIC-VAE*. Note that the figures from this section have been achieved with a value of λ of 1. it is clear that the proposed loss has managed to organize the latent space by class.

Several remarks can be made from these results. First, when we look at Figures 2a and 2c, we observe that when it is not forced to do so, a classical VAE does not organize its latent space according to semantics. Said differently, it means that when a VAE is only trained to compress individual items, no high-level information is necessary to represent an image compactly. On the contrary, *C-VAE* tries to spread as much as possible the items of a database, as the absolute cosine angles between two different latent vectors are roughly centered around 0.5 for the Gram matrix. This demonstrates the interest for the introduction of a dedicated loss term when dealing with multi-item compression.

A second observation is the ability of our proposed method to align the latent vectors by class (see Figures 2b and 2d). Even without a dedicated classification network (*e.g.*, no fully connected layer is applied in our VAE), the proposed loss term, based on covariance matrix distance enables us to "group" the items of a class. It leads to more correlated latent vectors within a class, from which we can expect a reduced compression rate. This is what we propose



(a) $\mathbf{G}_B$ for *C-VAE*.

(b) $\mathbf{G}_B$ for *sMIC-VAE*.

(c) t-SNE space for *C-VAE*.

(d) t-SNE space for *sMIC-VAE*.

Fig. 2: Latent space exploration of *C-VAE* and *sMIC-VAE*.

Model	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
gMIC-VAE	1	$\frac{1}{5}$	$\frac{1}{20}$	$\frac{1}{50}$	$\frac{1}{100}$	$\frac{1}{200}$	$\frac{1}{500}$	$\frac{1}{1000}$
sMIC-VAE	1	$\frac{2}{3}$	$\frac{1}{3}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{30}$	$\frac{1}{50}$
rMIC-VAE	1	$\frac{1}{2}$	$\frac{1}{10}$	$\frac{1}{20}$	$\frac{1}{33}$	$\frac{1}{50}$	$\frac{1}{200}$	$\frac{1}{1000}$

Table 1: Values of λ for Figure 3 for the different models.

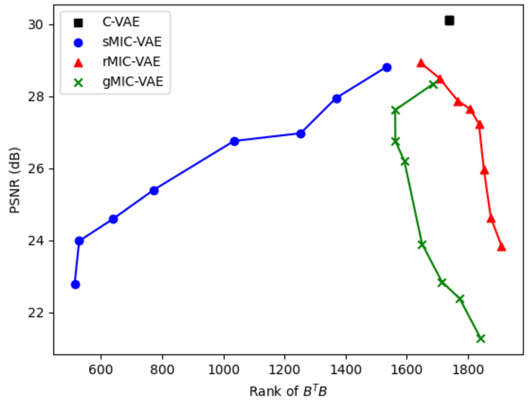


Fig. 3: Rate-distortion costs for *C-VAE*, *sMIC-VAE*, *rMIC-VAE* and *gMIC-VAE* on our toy database.

to test in the following.

4.3. Rate-distortion trade-off

In this experiment, we first compute the rate-distortion cost of the *C-VAE*. We also estimate the *sMIC-VAE*, the *rMIC-VAE* and the *gMIC-VAE* costs for different values of λ, given in Table 1. The results are presented in Figure 3.

We can observe that none of the *gMIC-VAE* and the *rMIC-VAE* manage decrease the rank. In other words, trying to align all items at the same time, or aligning items that are not semantically coherent cannot be done using the VAE. This demonstrates the importance of taking true semantic correlation into account. This statement is even stronger when looking at the behavior of the proposed *sMIC-VAE*. We can see that the proposed loss term enables to decrease the rank (as it could be expected from the alignment observed in the previous Section). Naturally, this reshaping of the latent space decreases the reconstruction loss as it penalizes the individual item compression. However, this experiment highlights a trade-off between the reconstruction error and the inter-item alignment (*i.e.*, the global database compression rate).

5. CONCLUSION

In this paper, we propose a first study to demonstrate the interest in taking into account the semantics for image database compression. We also propose the first solution to achieve a trade-off between the inter-item compression rate and the reconstruction error. Finally, this work also highlights a trade-off between inter-image redundancy and intra-image redundancy, typical of video coding. Future work will focus on the integration of such intuitions in a complete state-of-the-art compression solution, and especially comparing the true bit-rate of our model to the state of the art.

6. REFERENCES

[1] "https://www.domo.com/learn/infographic/data-never-sleeps-9," .

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm, "Overview of the versatile video coding (vvc) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[3] Jean Bégaint, Fabien Racapé, Simon Feltman, and Akshay Pushparaja, "Compressai: a pytorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.

[4] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson, "High-fidelity generative image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 11913–11924, 2020.

[5] Dandan Ding, Zhan Ma, Di Chen, Qingshuang Chen, Zoe Liu, and Fengqing Zhu, "Advances in video compression system using deep neural network: A review and case studies," *Proceedings of the IEEE*, 2021.

[6] Jill M Boyce, Renaud Doré, Adrian Dziembowski, Julien Fleureau, Joel Jung, Bart Kroon, Basel Salahieh, Vinod Kumar Malamal Vadakital, and Lu Yu, "Mpeg immersive video coding standard," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1521–1536, 2021.

[7] Jean Bégaint, Dominique Thoreau, Philippe Guillotel, and Christine Guillemot, "Region-based prediction for image compression in the cloud," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1835–1846, 2017.

[8] Huanjing Yue, Xiaoyan Sun, Jingyu Yang, and Feng Wu, "Cloud-based image coding for mobile devices—toward thousands to one compression," *IEEE transactions on multimedia*, vol. 15, no. 4, pp. 845–857, 2013.

[9] Ian H Witten, Radford M Neal, and John G Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, 1987.

[10] Alistair Moffat, "Huffman coding," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.

[11] Vivienne Sze and Madhukar Budagavi, "High throughput cabac entropy coding in hevc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1778–1791, 2012.

[12] Markus Herdin, Nicolai Czink, Hüseyin Ozcelik, and Ernst Bonek, "Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels," in *2005 IEEE 61st Vehicular Technology Conference*. IEEE, 2005, vol. 1, pp. 136–140.

[13] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *2nd International Conferenceon Learning Representations (ICLR)*, 2013.

[14] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.