Conditionnement sémantique latent pour la compression multi-image

Tom BACHARD[1, 2] Thomas MAUGEY[1]

[1]Inria Bretagne, Campus de Beaulieu, 263 Av. Général Leclerc, 35042 Rennes, France

[2]IRISA, Campus de Beaulieun 263 Av. Général Leclerc, 35000 Rennes, France

Résumé – Classiquement, les algorithmes de codage compressent de manière indépendante les images d'une même collection, et ce même si des corrélations sémantiques existent entre ces images. Dans ce article, nous explorons une solution pour prendre en compte cette redondance haut niveau et l'exploiter afin de réduire le coût de stockage de cette collection d'images. Dans un premier temps, nous définissons le cadre de la compression multi-image. Ensuite, nous dérivons une fonction de coût pour conditionner l'espace latent d'un auto-encodeur variationnel afin d'aligner les images de même nature sémantique. Enfin, nous démontrons expérimentalement l'intérêt de ce conditionnement en montrant qu'il amène à une représentation plus compacte de la collection d'images.

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 condition 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.

1 Introduction

La quantité de données créée, stockée et échangée dans le monde augmente de jour en jour, et ce de manière exponentielle. Afin de faire face à ce « Tsunami de données » (2,5 trilliards de bytes sont quotidiennement créés), de nouveaux algorithmes de compression sont développés et améliorés : algorithmes standards, tels que VVC [4], en allant jusqu'à des méthodes basées *deep learning* [5]. Cependant, malgré des taux de compression impressionnants, ces méthodes se cantonnent à de la compression mono-image. Cette limitation est en particulier restrictive lorsque l'on compresse une collection d'images, où notamment sont présentes des redondances, non exploitées. Dans ce travail, nous proposons un nouveau cadre de compression, la compression multi-image (*multi-item compression*, *MIC*), dont le but est de coder une collection d'images tout en prenant en compte les redondances présentes dans la collection.

Dans le cadre de la *MIC*, nous définissons deux types de redondances ; la redondance basée pixel, et la redondance basée sémantique, voir figure 1. La première représente les groupes de pixels partagés par plusieurs images de la collection. Elle peut par exemple être prise en compte par des approches multi-vues [3], ou de la compression basée *cloud* [2]. La seconde, en revanche, est liée à une interprétation haut niveau des images (*ex.* les objets, les concepts, les sentiments, ...) qui n'est pas forcément représentative des pixels des images. Au mieux de nos connaissances, la redondance sémantique n'a jamais été utilisée pour compresser des collections d'images, contrairement à la redondance basée pixels. Le but de la *MIC* est donc de prendre en compte cette redondance sémantique et de l'exploiter afin de réduire les coûts de compression de collections d'images.

Ce travail fait suite à un travail préliminaire [1], dans lequel un exemple jouet introduisait la compression multi-image.

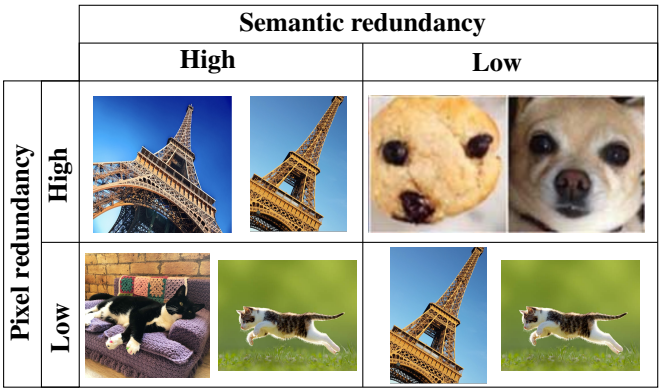


FIGURE 1 : Redondance pixel, redondance sémantique.

Nous proposons ici un schéma de compression basé sur un auto-encodeur variationnel existant [7], que nous adaptons pour être capable d'exploiter les redondance sémantiques. Ensuite, nous discutons différents scénarios permettant de prendre en compte les redondances sémantiques dans l'espace latent. Nous montrons qu'un alignement total des vecteurs latents peut être contre-productif, et qu'alors une méthode plus raffinée est nécessaire : on aligne ensemble les vecteurs sémantiquement cohérents. Une preuve expérimentale est apportée dans la dernière section, laquelle raffine les résultats de [1].

2 Compression multi-image

La compression multi-image (*MIC*) est un cadre de compression visant à coder efficacement une collection de N images $\mathcal{X} = (\mathbf{X}_1, \dots \mathbf{X}_N)$ via l'exploitation de ses redondances. L'efficacité de ce schéma de codage est donc mesurée à la foi en termes de reconstruction des images (via le PSNR) et en termes

de taux compression (ici, via le pseudo-rang, voir section 4.1). La MIC comporte deux phases ; une phase d'apprentissage des redondances présentes et une phase de compression utilisant ces redondances.

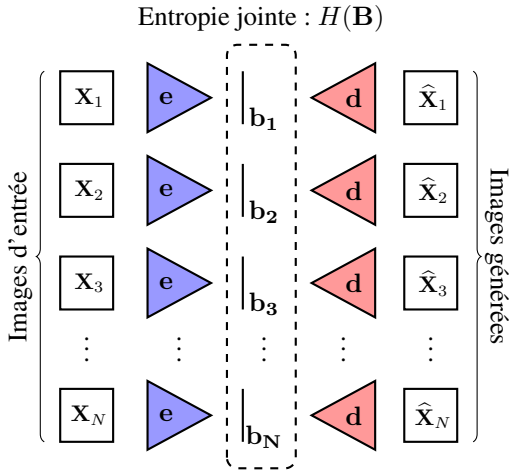


FIGURE 2 : Schéma de compression.

2.1 Schéma de compression

Notre définition de la *MIC* est donnée en figure 2. Nous supposons que les images de la collection $\mathcal{X}$ partagent des redondances, a minima sémantiques mais aussi éventuellement au niveau des pixels. L'encodeur e construit une représentation latente de chaque image, notée $\mathbf{b}_i = e(\mathbf{X}_i)$. Nous choisissons d'encoder les images individuellement, puisqu'il est ainsi facile d'ajouter une image à la base de données sans avoir à refaire toute la partie de compression. On note $\mathbf{B}$ la matrice latente agglomérée par colonnes de tous les vecteurs latents : $\mathbf{B} = [\mathbf{b_1}, \ldots, \mathbf{b}_N]$.

Après avoir compressé chaque image via l'encodeur e, nous nous intéressons à l'information partagée par les vecteurs latents. En effet, plus les vecteurs latents partagent de l'information, plus le taux de compression théorique R est élevé. On voit ici une relation entre R et $H(\mathbf{B})$, l'entropie jointe des $\mathbf{b}_i$.

Finalement, les images sont reconstruites via le décodeur d : $\widehat{\mathbf{X}}_i = d(\mathbf{b}_i)$. Notons que dans ce schéma, nous n'appliquons pas de quantification puisque l'intérêt de ce travail est d'évaluer l'exploitation des redondances dans l'espace latent.

2.2 Obectif de la MIC

L'objectif de notre proposition de la MIC est donné par l'équation (1). Le taux de compression R est associé à H, l'entropie jointe de $\mathbf{B}$ dans l'espace latent. La contrainte de reconstruction classique en compression d'image est donnée par le seuil T sur le PSNR de $\mathbf{X}$ par rapport à $\widehat{\mathbf{X}}$.

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

En pratique, le but est d'utiliser la corrélation inter-image afin d'obtenir une entropie jointe inférieure à la somme des entropies prises individuellement, tel qu'une compression classique le ferait, *i.e.*, $H(\mathbf{b}_1, ..., \mathbf{b}_N) < \sum_i H(\mathbf{b}_i)$. Ainsi, si une transformation dans l'espace latent décorrèle les $\mathbf{b}_i$, le compactage d'énergie de la MIC sera meilleur que le compactage dans le cas classique mono-image. Cette inégalité est obtenue si l'espace latent est conditionné de manière à aligner les vecteurs latents.

3 Réduction de l'entropie latente

3.1 Alignement total

Le principal gain de la MIC provient du fait que la corrélation entre les vecteurs latents est à la fois encouragée, mais aussi exploitée. Dans cette optique, nous discutons ici comment concevoir un encodeur e imposant cette propriété.

Nous supposons que la distribution des vecteurs latents suit une loi gaussienne centrée de matrice de covariance $\Sigma_{\mathbf{B}}$. Cette hypothèse classique nous permet d'avoir une forme explicite de l'entropie jointe de $\mathbf{B}$, donnée en équation (2). Cependant, la vraie covariance $\Sigma_{\mathbf{B}}$ est inconnue. Nous faisons alors le choix de l'approximer par la matrice de Gram normalisée de $\mathbf{B}$: $\mathbf{\Sigma_B} \simeq \mathbf{G_B} = \mathbf{B^T B}$.

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

Ainsi, après normalisation des constantes pour la minimisation, on obtient l'équation (3) à partir de l'équation (1).

$$\min_{e,d} \log\left(|\mathbf{G_B}|\right) \tag{3}$$
$$\text{s.t. } \forall i \in [\![1,N]\!], \text{ PSNR}\big(\mathbf{X}_i, \hat{\mathbf{X}}_i\big) > T$$

Résoudre l'équation (3) revient à aligner tous les vecteurs latents selon la même direction, *i.e.* à n'avoir, à une multiplication matricielle près, qu'un seul vecteur latent.. On se référera à ce problème comme à celui de l'alignement total.

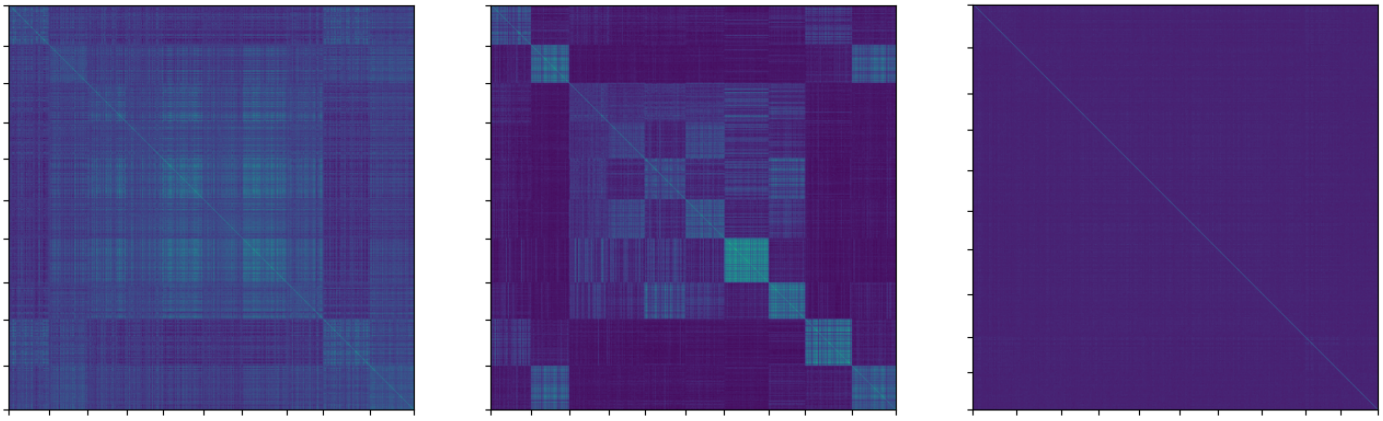
3.2 Alignement sémantique

Bien que dérivant directement de la formulation de la *MIC*, entraîner un encodeur à résoudre l'équation (3) va pénaliser le schéma global de compression, notamment en termes d'expressivité lors de la décompression. Ainsi, au lieu d'aligner tous les vecteurs latents ensemble, nous proposons de réaliser un alignement plus subtil, prenant en compte la sémantique de la base de données. Il s'agit d'exploiter des redondances qui existent bel et bien dans $\mathcal{X}$ mais pas encore transposées dans l'espace latent. Ainsi, on se donne une sémantique d'une base de données sous la forme d'une matrice de Gram $G^*_{\mathbf{B}}$, et notre but va être de réduire la distance entre cette sémantique théorique et notre matrice de Gram courante $\widehat{G_{\mathbf{B}}}$.

Dans le cadre de ce travail, on pose $G^*_{\mathbf{B}}[i,j] = 1$ si les images i et j appartiennent à la même classe, et 0 sinon. Ainsi, les images de même classe doivent être alignées ensemble, tout en étant orthogonales avec les autres images des autres classes.

Puisque nous travaillons avec des matrices de covariance, ou assimilées, nous utilisons la distance de matrices de covariance proposée dans [6] : $d_{cov}(A, B) = 1 - \frac{Tr(AB)}{\|A\|\|B\|}$. Nous obtenons ainsi le problème de minimisation donné par l'équation (4), nommé *alignement sémantique*.

$$\min_{e,d} \text{d}_{cov}\left(\hat{\mathbf{G}}_{\mathbf{B}}, \mathbf{G}^*_{\mathbf{B}}\right) \tag{4}$$
$$\text{s.t. } \forall i \in [\![1,N]\!], \text{ PSNR}\big(\mathbf{X}_i, \hat{\mathbf{X}}_i\big) > T$$



(a) $\mathbf{G}_B$ pour *CAE*. (b) $\mathbf{G}_B$ pour *s-CAE*. (c) $\mathbf{G}_B$ pour *r-CAE*.

FIGURE 3 : Gramiennes par classes (gradations noires) pour les différents modèles d'alignement. Plus clair signifie plus corrélé.

(a) Projection t-SNE pour *CAE*. (b) Projection t-SNE pour *s-CAE*. (c) Projection t-SNE pour *r-CAE*.

FIGURE 4 : Projections t-SNE par classes pour les différents modèles d'alignement.

On dérive alors la fonction de *loss* donnée en équation (5) pour l'entraînement de notre encodeur.

$$\mathcal{L}(\mathbf{X}) = \lambda \mathrm{d}_{cov}\left(\hat{\mathbf{G}}_{\mathbf{B}}, \mathbf{G}^*_{\mathbf{B}}\right) + \frac{1}{N}\sum_{i=1}^{N} MSE(\mathbf{X}_i, \hat{\mathbf{X}}_i) \quad (5)$$

4 Expériences

4.1 Protocole expérimental

L'architecture que nous utilisons pour nos expériences provient de [7]. Afin d'avoir plus de flexibilité sur le conditionnement de l'espace latent, nous avons supprimé la partie sur les mixtures de gaussienne, ainsi que la quantification binaire. Le compromis entre la distorsion et le conditionnement est modulé par l'hyper-paramètre λ (voir par exemple l'équation (5)).

L'entraînement du réseau de neurones comprend 100 *epochs* sur le jeu de données CIFAR10 (60000 images séparées en 10 classes). Les résultats proposés dans cette section sont obtenus sur une sous partie du jeu test, 2000 images, représentative du jeu de données originel. La sémantique de la collection d'images, $\mathbf{G}^*_{\mathbf{B}}$, est ensuite calculée en fonction de ces 2000 images, avec $\mathbf{G}^*_{\mathbf{B}}[i,j] = 1$ si la classe de la $i^{\text{ème}}$ image et $j^{\text{ème}}$ image sont les mêmes, et 0 sinon.

En termes de métrique, la distorsion sera évaluée via le PSRN (rgb), et le compactage de la collection via le pseudo-rang de $\mathbf{B}$, *i.e.* le nombre de valeurs propres au dessus d'une valeur seuil, (ici 0.01), donc bien dans l'espace latent. En effet, le rang classique n'est pas suffisant, puisque toujours quasiment plein même après alignement. On fera ainsi apparaître un compris distorsion-compactage, plutôt qu'un compromis distorsion-taux de compression.

Pour nos expériences, différents modèles seront utilisés afin d'avoir des comparaison objectives de nos contributions :

- **CAE**, le modèle témoin, sans conditionnement ;
- **t-CAE**, le modèle d'alignement **t**otal, voir équation (3) ;
- **s-CAE**, le modèle d'alignement **s**émantique où $\mathbf{G}^*_{\mathbf{B}}$ provient des données, voir équation (4) ;
- **r-CAE**, un modèle d'alignement aléatoire (*random*). Même modèle que s-CAE mais où $\mathbf{G}^*_{\mathbf{B}}$ est tirée aléatoirement au début de l'expérience. Le but de ce modèle est de montrer que seul l'alignement sémantique, et non le fait d'aligner, amène les propriétés désirées.

4.2 Résultats

Deux aspects sont évalués dans ce travail, tout d'abord à quel point le conditionnement – l'alignement sémantique – est réussi lors de l'apprentissage. Ensuite seulement, on observera le résultat du conditionnement sur le compromis distorsion-compactage.

4.2.1 Évaluation de l'alignement sémantique

La figure 3 montre respectivement les gramiennes (dont les colonnes sont classées par classe) de CAE, s-CAE et r-CAE une fois l'entraînement terminé. Plus le point est clair, plus les vecteurs sont corrélés. La figure 4 montre respectivement les projections TSNE [8] des espaces latents de CAE, s-CAE et r-CAE.

De ces figures, plusieurs observations peuvent être réalisées. On voit tout d'abord que l'absence de conditionnement implique l'absence de structure corrélée dans l'espace latent. En effet, puisque l'encodeur de CAE est entraîné à compresser des images de manière individuelle, il n'y a pas de raison de créer et/ou de mettre en lumière une éventuelle corrélation dans l'espace latent. Des figures liées à s-CAE, on observe que l'alignement sémantique, en plus d'être réussi, délimite exactement les classes initiales des données. On observe que les classes qui ne sont pas bien séparées, sont sémantiquement proches pour des êtres humains : les classes « chat » et « chien » sont sémantiquement plus proches l'une de l'autre que n'importe quel autre couple de classes. Enfin, on voit qu'un alignement non sémantique, chez r-CAE, n'induit aucune structure intéressante dans l'espace latent.

4.2.2 Évaluation du compromis

Dans la seconde expérience, nous calculons un compromis entre la qualité de reconstruction (exprimée via le PSRN) et le compactage de la base d'images dans l'espace latent (exprimé par le rang effectif de $\mathbf{B}$). La figure 5 résume les différentes courbes de compromis (*i.e.* différentes valeurs de λ) pour les modèles t-CAE et s-CAE, ainsi qu'un point pour le modèle témoin CAE.

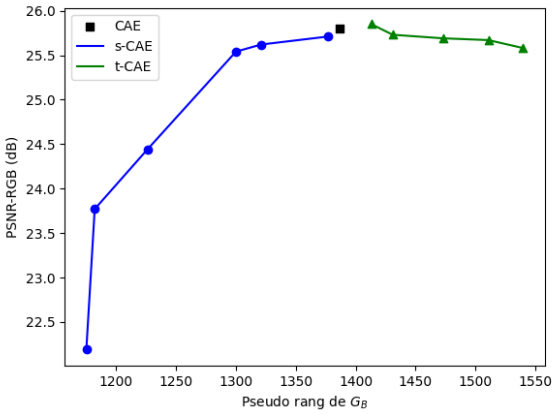


FIGURE 5 : Compromis pour les différents modèles.

Nous observons tout d'abord qu'aucun des modèles non-sémantiques n'arrive à produire une courbe de compromis intéressante. On observe que seul le modèle s-CAE y arrive. Ce résultat montre que prendre en compte les redondances sémantiques aide au conditionnement de l'espace latent. De plus, pour s-CAE, on voit qu'une diminution du rang effectif entraîne une dégradation des images. En effet, un alignement trop fort de concepts encore trop grossiers que sont les classes dégrade nécessairement la reconstruction. Finalement cette expérience met en lumière un réel compromis entre la reconstruction et le compactage latent, intrinsèquement lié au taux de compression.

5 Conclusion

Dans ce article, nous implantons la compression multi-image dans une architecture de réseaux de neurones de l'état de l'art. Nous confirmons ainsi les résultats préliminaires de [1], à savoir que l'alignement sémantique dans l'espace latent permet un meilleur compactage d'une collection d'images que les schémas de compression individuelle classiques. Un travail futur exploitera ce compactage afin d'atteindre un taux de compression meilleur que celui la compression classique.

Références

[1] Tom BACHARD, Anju Jose TOM et Thomas MAUGEY : Semantic alignment for multi-item compression. *In 2022 IEEE International Conference on Image Processing (ICIP)*, pages 2841–2845, 2022.

[2] Jean BÉGAINT, Dominique THOREAU, Philippe GUILLOTEL et Christine GUILLEMOT : Region-based prediction for image compression in the cloud. *IEEE Transactions on Image Processing*, 27(4):1835–1846, 2017.

[3] Jill M BOYCE, Renaud DORÉ, Adrian DZIEMBOWSKI, Julien FLEUREAU, Joel JUNG, Bart KROON, Basel SALAHIEH, Vinod Kumar Malamal VADAKITAL et Lu YU : Mpeg immersive video coding standard. *Proceedings of the IEEE*, 109(9):1521–1536, 2021.

[4] Benjamin BROSS, Ye-Kui WANG, Yan YE, Shan LIU, Jianle CHEN, Gary J SULLIVAN et Jens-Rainer OHM : Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

[5] Dandan DING, Zhan MA, Di CHEN, Qingshuang CHEN, Zoe LIU et Fengqing ZHU : Advances in video compression system using deep neural network : A review and case studies. *Proceedings of the IEEE*, 2021.

[6] Markus HERDIN, Nicolai CZINK, Hüseyin OZCELIK et Ernst BONEK : Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels. *In 2005 IEEE 61st Vehicular Technology Conference*, volume 1, pages 136–140. IEEE, 2005.

[7] Lucas THEIS, Wenzhe SHI, Andrew CUNNINGHAM et Ferenc HUSZÁR : Lossy image compression with compressive autoencoders. *arXiv preprint arXiv :1703.00395*, 2017.

[8] Laurens Van der MAATEN et Geoffrey HINTON : Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.