Guide de couleur pour les modèles de diffusion et application à la compression d'images très bas débit

Tom Bordin Thomas Maugey

INRIA Bretagne

{tom.bordin, thomas.maugey}@inria.fr

Résumé

Nous présentons de nouvelles équations permettant de conditionner les modèles de diffusion avec un guide de couleur sans entraînement. Cette méthode s'applique à tous modèles de diffusions, y compris ceux fonctionnant dans l'espace latent. Nous montrons comment cette méthode peut être utilisée dans un contexte de compression des images à extrêmement bas débits (<0.01 bpp), en améliorant ainsi à la fois la fidélité à la sémantique, mais aussi la qualité des images décodées.

Mots clefs

diffusion, compression, communication sémantique

1 Introduction

Dans le contexte de compression très bas débit ($\sim$ 0.01bpp), l'optimisation d'un compromis débit-distorsion provoque une perte de l'information sémantique et de la qualité perceptuelle de l'image. Le compromis existant entre distorsion et perception est en effet exacerbé particulièrement par le très bas débit [1]. Ainsi, certaines méthodes plus récentes proposent de s'éloigner du critère de distorsion au profit de nouveaux critères, tels que la qualité perceptuelle et/ou de la préservation de la sémantique de l'image. Par exemple, HiFiC[2], propose de minimiser un compromis entre la qualité perceptuelle et la distorsion. En allant plus loin, PICS[3], propose une description de l'image par une information textuelle et par les contours des objets, abandonnant complètement le critère de distorsion. Entre les deux, d'autres méthodes PerCo[4] ou MS-ILLM[5] utilisent la sémantique de l'image dans des méthodes optimisant le compromis perception distorsion.

De manière similaire, nous choisissons de représenter une image comme : (i) une image très basse résolution décrivant l'aspect général de la couleur dans l'image originale, et (ii) une information sémantique en s'appuyant sur un modèle de fondation : CLIP[6]. Nous proposons ainsi de reprendre le framework de CoCliCo [7] basé sur des modèles de diffusions. Néanmoins, ce framework ne pouvait pas faire une utilisation complète de l'information donnée par l'image très basse résolution.

En effet, les modèles de diffusion se sont rapidement imposés ces dernières années pour des tâches génératives, et sont par conséquent principalement entraînés pour aller du texte vers l'image. Afin de les rendre compatibles avec d'autres formes de conditions (comme la couleur dans notre cas), la plupart des méthodes de compression ci-dessus ont dû procéder au ré-entraînement de ceux-ci, impliquant un coût important en temps et puissance de calcul. L'alternative que nous explorons dans cet article consiste à guider sans entraînement la génération durant l'inférence. Plus précisément, nous réécrivons les équations de guide pour le cas où la condition est donnée par l'image très basse résolution. Cela nous permet ainsi d'améliorer significativement le schéma CoCliCo en termes de fidélité à l'image originale.

2 Contrôles des modèles de diffusion

Nous rappelons brièvement le fonctionnement des modèles de diffusions, ainsi que quelques notations utiles à la compréhension. Puis, nous faisons un court bilan sur les méthodes existantes.

2.1 Principe de la diffusion

Les modèles de diffusion[8] sont des modèles génératifs. Ils fonctionnent d'une manière similaire à du débruitage, transformant itérativement un bruit gaussien en une image. Les modèles les plus récents, pour des questions de mémoire et de temps de calcul, agissent sur l'espace latent d'auto-encodeurs plutôt que sur les images, générant ainsi des représentations latentes.

Le bruit gaussien est ainsi modélisé, par une suite de paramètres (α_t), telle que, le signal $\boldsymbol{z}_t$ suive à l'étape t :

$$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_0 + \sqrt{1-\alpha_t}\epsilon \quad \text{avec} \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}) \tag{1}$$

de telle manière que $\boldsymbol{z}_0$ soit une image et $\boldsymbol{z}_T$ soit un bruit gaussien. Le rôle du modèle de diffusion est alors d'estimer la valeur de ϵ pour chaque t. On peut alors calculer $\boldsymbol{z}_{t-1}$ à partir de $\boldsymbol{z}_t$ à l'aide d'un planificateur. Notamment, les *Denoising Diffusion Implicit Models*(DDIM)[9] montrent qu'il est possible de formuler, à chaque étape, l'image en cours de débruitage comme la somme d'une image inter-

médiaire prédite et d'un bruit résiduel :

$$\boldsymbol{z}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{z}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\boldsymbol{z}_t, t)}{\sqrt{\alpha_t}} \right)}_{\text{"prédiction } \boldsymbol{z}_0 \,:\, \hat{\boldsymbol{z}}_0(\boldsymbol{z}_t, t)\text{"}} + \underbrace{\sqrt{1-\alpha_{t-1}}\epsilon_\theta(\boldsymbol{z}_t, t)}_{\text{"bruit résiduel } \boldsymbol{z}_t\text{"}} \tag{2}$$

2.2 Méthodes existantes

Il existe aujourd'hui plusieurs méthodes permettant de contrôler la génération des modèles de diffusion. Certaines nécessitent un entraînement du modèle, d'autres peuvent être utilisées durant l'inférence, nous nous intéressons ici principalement à ces dernières.
Le contrôle sans entraînement consiste à modifier durant l'inférence le signal débruité, en influençant [11], imposant[14, 10], ou encore corrigeant [12, 15, 13] la prédiction faite par le modèle de diffusion. Nous résumons les différentes méthodes existantes dans le tableau 1.
La correction se fait à chaque étape t par ce que l'on appelle le guidage et nécessite l'estimation du terme G suivant :

$$G(\boldsymbol{z}_t, t, \boldsymbol{c}) = -\sqrt{1-\alpha_t}\nabla_{\boldsymbol{z}_t} \log p_t(\boldsymbol{c}|\boldsymbol{z}_t) \tag{3}$$

où $p_t(\boldsymbol{c}|\boldsymbol{z}_t)$ est la densité de probabilité d'avoir la condition $\boldsymbol{c}$ connaissant le signal bruité $\boldsymbol{z}_t$.
Un obstacle majeur au contrôle des modèles de diffusion est l'utilisation d'un espace latent. En effet, quand la diffusion est effectuée dans l'espace latent d'un auto-encodeur [2], les propriétés que l'on pouvait à priori avoir sur le signal d'une image sont perdues. Notamment, on ne maîtrise plus la comparaison entre le signal bruité et la condition.

3 Guidage de la couleur et application à la compression

3.1 Formulation du guidage

Nous choisissons de guider le modèle sur la condition d'un aspect global de la couleur. Pour cela, nous modélisons la couleur comme une décimation sur fréquences de l'image, ne gardant que les plus basses fréquences. Nous pouvons alors écrire $\boldsymbol{c} = \boldsymbol{A}\boldsymbol{x}$ pour une image $\boldsymbol{x}$.
Nous modélisons l'erreur de prédiction du bruit par le modèle de diffusion par une gaussienne $\mathcal{N}(0, \overline{\lambda}_t)$. Puis dans un second temps, nous estimons la réponse du décodeur D de l'espace latent à faible un bruit gaussien comme une fonction linéaire : $D(\boldsymbol{z} + \delta\epsilon) = \boldsymbol{x} + \overline{a}\delta + \overline{b}\delta\epsilon'$.
Avec ces deux hypothèses, en partant de la distribution du signal estimée au temps t dans l'équation DDIM (2), nous montrons qu'il est aussi possible de conclure sur la distribution de la condition que l'on veut imposer à l'image. Nous obtenons ainsi, une estimation de la distribution $\boldsymbol{c}|\boldsymbol{z}_t$ et pouvons conclure sur la forme de l'équation de guidage dans le domaine latent :

$$\boxed{G(\boldsymbol{z}_t, t, \boldsymbol{c}) = \frac{\alpha_t}{2\overline{b}^2\overline{\lambda}_t{}^2\sqrt{1-\alpha_t}}\nabla_{\boldsymbol{z}_t}\|\boldsymbol{c} - \hat{\boldsymbol{c}}_t - \overline{a}\frac{\overline{\lambda}_t\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}}\boldsymbol{A}\mathbf{1}_n\|^2} \tag{4}$$

avec une équation similaire dans le domaine pixel, le décodeur est alors parfait avec $a, b = (0, 1)$.

3.2 Application à la compression

Nous appliquons la méthode de guidage dans un contexte compression sémantique très bas débit, optimisant ainsi l'utilisation de l'information de couleur. Nous reprenons le schéma de compression de CoCliCo. L'image est représentée par une information sémantique au travers du modèle de fondation CLIP et complémentée par une information sur l'aspect général de couleur. Le décodage génératif est réalisé par un modèle de diffusion conditionnel, CLIP vers image, guider par la couleur. Cette représentation de la sémantique très compacte nous permet d'atteindre de très bas débits.
Nous présentons en Figure.1, les résultats de notre méthode de compression en utilisant une carte de couleur 25×25. Ces images ont été encodées avec un débit moyen de $0.0098bpp$ pour les méthodes utilisant du guidage et un débit légèrement supérieur pour VVC[16] et PICS[3]. Nous pouvons voir que les images décodées avec notre méthode restent fidèles à l'original en termes de sémantique et de couleurs et que notre guidage améliore la fidélité à la couleur. Une évaluation plus détaillée présentant plus de figures, de comparaisons et de métriques est disponible dans l'article complet.

4 Conclusion

Nous proposons dans cet article une méthode permettant de guider les modèles de diffusion sans entraînement avec un conditionnement sur l'aspect général de la couleur de l'image. Nous montrons que cette méthode peut être appliquée à la compression, dans un contexte extrêmement bas débit, pour compléter une représentation sémantique. Nous obtenons ainsi des images à la fois fidèles à l'original en termes de couleur et sémantique, tout en conservant une très haute qualité perceptuelle.

Annexe

Ces travaux ont été financés par l'Agence Nationale de la Recherche (MADARE, Project-ANR-21-CE48-0002). La version complète de l'article fait l'objet d'une soumission à Transactions on Image Processing : https://arxiv.org/pdf/2404.06865

Références

[1] Yochai Blau et Tomer Michaeli. The perception-distortion tradeoff. Dans *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018.

[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, et Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models, Avril 2022. arXiv :2112.10752 [cs].

[3] Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, et Shirin Saeedi Bidokhti. Text+ sketch : Image compression at ultra low rates. *arXiv preprint arXiv :2307.01944*, 2023.

TABLEAU 1 – *Comparaison des méthodes de contrôle*

Méthode	Équation clé	$\boldsymbol{c} =$	Entraînement	Contrôle	Domaine latent
Conditionnement [8]	$\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$	N'importe	Oui	**Bon**	**Oui** (même équation)
Conditionnement forcé[10]	$\emptyset$	Séparable[1]	**Non**	**Bon**	Non
Initialisation [11]	$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{c}\sqrt{1-\alpha_t}\epsilon$	Image/Couleur	**Non**	Moyen	**Oui** (même équation)
Guidage par classificateur[12]	$-\sqrt{1-\alpha_t}\nabla_{\boldsymbol{z}_t} \log p_t(\boldsymbol{c}\|\boldsymbol{z}_t)$	Valeur de la classe	Oui	**Bon**	Non
Guidage universel[13]	$s\sqrt{1-\alpha_t}\nabla_{\boldsymbol{z}_t}\|\boldsymbol{c} - \hat{\boldsymbol{c}}_t\|_2^2$	N'importe	**Non**	Moyen	**Oui** (même équation)
Notre méthode	$\frac{\alpha_t}{2\lambda_t\sqrt{1-\alpha_t}}\nabla_{\boldsymbol{z}_t}\|\boldsymbol{c} - \hat{\boldsymbol{c}}_t\|_2^2$	$\boldsymbol{Ax}$ (i.e couleurs)	**Non**	**Bon**	**Oui** (équation (4))

[1] La condition peut être séparée de l'image (masque, décomposition, etc)

FIGURE 1 – *Comparaison de différentes méthodes de compression, avec une carte de couleurs de taille* 25×25.

[4] Marlene Careil, Matthew J Muckley, Jakob Verbeek, et Stéphane Lathuilière. Towards image compression with perfect realism at ultra-low bitrates. Dans *The Twelfth International Conference on Learning Representations*, 2023.

[5] Matthew J Muckley, Alaaeldin El-Nouby, Karen Ullrich, Hervé Jégou, et Jakob Verbeek. Improving statistical fidelity for neural image compression with implicit local likelihood models. Dans *International Conference on Machine Learning*, pages 25426–25443. PMLR, 2023.

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. Dans *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[7] Tom Bordin, Tom Bachard, et Thomas Maugey. Coclico : Extremely low bitrate image compression based on clip semantic and tiny color map. Dans *2024 IEEE 37th International Workshop of Picture Coding Symposium (PCS)*. IEEE, 2024.

[8] Yang Song et Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

[9] Jiaming Song, Chenlin Meng, et Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv :2010.02502*, 2020.

[10] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, et Luc Van Gool. Repaint : Inpainting using denoising diffusion probabilistic models. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[11] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, et Stefano Ermon. Sdedit : Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv :2108.01073*, 2021.

[12] Prafulla Dhariwal et Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. Dans *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.

[13] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, et Tom Goldstein. Universal guidance for diffusion models. Dans *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[14] Di You, Andreas Floros, et Pier Luigi Dragotti. Indigo : An inn-guided probabilistic diffusion algorithm for inverse problems. *arXiv preprint arXiv :2306.02949*, 2023.

[15] Jonathan Ho et Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv :2207.12598*, 2022.

[16] Adam Wieckowski, Jens Brandenburg, Tobias Hinz, Christian Bartnik, Valeri George, Gabriel Hege, Christian Helmrich, Anastasia Henkel, Christian Lehmann, Christian Stoffers, Ivan Zupancic, Benjamin Bross, et Detlev Marpe. Vvenc : An open and optimized vvc encoder implementation. Dans *Proc. IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, pages 1–2.