PFE - References

Etienne BARDET

April 24, 2025

References

[1] C. B. Azizi, C. Guilloteau, G. Roussel, and M. Puigt, "Coupled VAE and Interpolator Approach for Fast Hyperspectral Image Emulation," in *2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. IEEE, 2024, pp. 1–5.

> Cocorico, ce papier parle d'un VAE où l'on remplace l'encodeur par un réseau 'interpolateur' qui vient projeter des variables biophysiques dans l'espace latent après avoir entrainé le VAE pour de la reconstruction. Le décodeur est frozen pour la deuxième partie.

[2] G. Dorta, S. Vicente, L. Agapito, N. D. Campbell, and I. Simpson, "Structured uncertainty prediction networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5477–5485.

> Amélioration des VAE en apprenant une variance plus complexe et creuse en utilisant cholesky. Cela améliore les échantillons mais au prix d'une petite quantité de calculs en plus. Il est aussi proposé une façon plus performante à l'aide d'une distance de mahalanobis et de normalisation spectrale : regarder la distance entre la classe la plus proche et la prédiction (pas pertinent ici)

[3] I. Dumeur, S. Valero, and J. Inglada, "Self-supervised spatio-temporal representation learning of Satellite Image Time Series," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 4350–4367, 2024.

Modèle inspiré de BERT. U-BARN → Encodeur Spatial et Spectral. L'image est divisée en patches qui sont encodés puis reconstruits par un U-NET. Les sorties des Unets sont ensuite embeddées par position pour être ensuite données à un transformer. L'idée est d'ensuite d'ajouter une tâche (classif/segmentation, reconstruction) à la fin du U-BARN pour pouvoir l'entrainer à une tâche précise. Problème : pas un VAE, les représentations sont un espace de petite dimension mais pas probabiliste. Quid des incertitudes ? Autre question : les patches introduisent une limite dans la spatialité de la donnée, on reste local. Le côté temporel traverse les patchs embeddings : chaque instant de prise de vue passe dans un SSE. On reshape les SITS en b*t,c,h,w puis en b x w x h ,t ,d avant le transformer

[4] I. Gatopoulos, M. Stol, and J. M. Tomczak, "Super-resolution variational auto-encoders," *arXiv preprint arXiv:2006.05218*, 2020.

Construit sur \citesohn2015Learning. Utilisation d'une image LR pour sampler un premier étage de variables latentes u, puis samplage de z avec u et y et finalement x | z. Papier qui n'entraine que sur des imagettes de 32*32 | 64*64 car peu de puissance de calcul. Possibilité d'amélioration en 256*256 ??

[5] I. Gatopoulos and J. M. Tomczak, "Self-Supervised Variational Auto-Encoders," *Entropy*, vol. 23, no. 6, p. 747, June 2021.

[6] B. Maud, M. Chabert, F. Genin, C. Latry, and T. Oberlin, "Deep priors for satellite image restoration with accurate uncertainties," *arXiv preprint arXiv:2412.04130*, 2024.

VBLE, utilisation d'une architecture CAE, proche d'un VAE qui estime un hyperprior à partir de \barz. On peut donc sampler z à partir de l'hyperprior puis le décoder pour obtenir la moyenne et la variance (en utilisant deux décodeurs). Voir \citerybkin2021simple concernant la méthode. Comment mesurer incertitude ? Ici, pour un pixel, on sample 100 fois l'image et on regarde le pourcentage de fois où la valeur du pixel est dans la valeur du nouveau sample \pm5% ?

[7] J. Prost, A. Houdard, A. Almansa, and N. Papadakis, "Efficient Posterior Sampling For Diverse Super-Resolution with Hierarchical VAE Prior," in *VISAPP 2024-19th International Conference on Computer Vision Theory and Applications*, 2024.

Dans ce papier, on a un VAE hiérarchique conditionnel en quelque sorte. l'idée est que les variables latentes données par l'encodeur low-res captent les détails basse fréquences et que conditionné selon ca, on peut ensuite sampler les variables latentes pour ensuite sampler une version High-res en la passant dans notre modèle génératif : $$. L'idée est qu'on va contraindre les k premières variables latentes de l'encodeur HR et LR à correspondre en distribution (via une divergence KL) ce qui essaye de contraindre l'espace latent de capturer une représentation BF/LR de l'image. Implem : tu prends un VDVAE et à chaque étage tu viens calculer des moyennes et variances pour les var latentes ?

[8] O. Rybkin, K. Daniilidis, and S. Levine, "Simple and effective VAE training with calibrated decoders," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9179–9189.

Ici, l'idée est de trouver des méthodes plus faciles/efficaces d'utiliser un VAE. La méthode clé proposée est par exemple de calculer la variance selon les données et de sortir la moyenne via un réseau de neurones. NB : La revue par les pairs est relativement mitigée sur l'intérêt du papier.

[9] T. Salimans, A. Karpathy, X. Chen, and D. P. Kingma, "PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications," Jan. 2017.

1 Annotations (24/04/2025 10:32:50)

¡¡ By choosing a simple continuous distribution for modeling ν (like the logistic distribution as done by Kingma et al. (2016)) we obtain a smooth and memory efficient predictive distribution for x. Here, we take this continuous univariate

distribution to be a mixture of logistic distributions which allows us to easily calculate the probability on the observed discretized value x, as shown in equation (2) ¿¿ (Salimans et al., 2017, p. 2) Utilisation de mixture de logistiques pour modéliser la densité de proba sur 0,255

[10] Y. She, C. Atzberger, A. Blake, A. Gualandi, and S. Keshav, "MAGIC: Modular Auto-encoder for Generalisable Model Inversion with Bias Corrections," *arXiv preprint arXiv:2405.18953*, 2024.

Utilisation d'un AE classique où le décodeur est remplacé par un modèle physique en pytorch. Adaptabilité avec un VAE pour faire de la génération ? Problème, on perd la partie p(x|z). Réseau linéaire, pertinence ? Les variables ont plus de sens physique mais même perf qu'un AE classique sur ce problème. Cité une fois à la rédaction de ce document.

[11] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

Conditionnement d'un VAE avec une entrée supplémentaire (typiquement un label) pour conditionner le modèle dans l'espace latent

[12] A. Vazhentsev, G. Kuzmin, A. Shelmanov, A. Tsvigun, E. Tsymbalov, K. Fedyanin, M. Panov, A. Panchenko, G. Gusev, M. Burtsev, *et al.*, "Uncertainty estimation of transformer predictions for misclassification detection," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8237–8252.

Estimation d'incertitudes d'un transformer en utilisant du dropout pendant l'inférence. En droppant quelques neurones, on peut venir sampler la distribution estimée de la sortie et donc avoir une idée de l'incertitude.

[13] Y. Zérah, "Physics-based deep representation learning of vegetation using optical satellite image time series," Ph.D. dissertation, Université de Toulouse, 2024.

Utilisation d'un modèle physique comme décodeur d'un VAE pour donner de l'interprétabilité aux variables latentes. Les

variables latentes sont séparées en deux parties (vecteurs coupés en deux) une partie décodeur 'aléatoire', une partie modèle physique déterministe. Une contrainte de reconstruction, MCRL : \mathcalL_MCRL = -ln(p(x_i|z_i)) permet notamment d'avoir une moyenne proche des x_i et une variance \hat\sigma^2_i qui représente l'incertitude du décodeur.