SRVAE math

Etienne Bardet

May 2025

1 Introduction

Ce document a pour but de détailler les calculs formulant la loss d'un VAE dans un premier temps puis d'un VAE Conditionnel ("à deux étages") dans un second temps.

2 VAE

2.1 Évidence

Dans un VAE, nous cherchons à maximiser l'évidence, qui est la probabilité de retrouver notre données, conditionné sur les paramètres du modèle, soit :

$$p(x) = \int p(x|z)p(z)dz$$

2.2 Encodeur

Cependant, pour estimer $p(x|z)$, il nous faut connaitre $p(z|x)$ pour utiliser le théorème de Bayes.

Nous allons approcher cette distribution par un réseau de neurones qui fournira $q_\phi(z|x)$. Nous supposons ici que la distribution des données de l'espace latent est gaussienne, soit :

$$q_\phi(z|x) \sim \mathcal{N}(\mu_\phi, \Sigma_\phi)$$

Avec Σ_ϕ une matrice de covariance diagonale (les l coefficients de sa diagonale étant l sorties de l'encodeur). μ_ϕ étant une autre sortie de l'encodeur, de même taille.

2.3 ELBO

Reformulons la formule de l'évidence :

$$p(x) = \int \frac{p(x,z)}{q_\phi(z|x)} q_\phi(z|x)dz$$

L'espérance conditionnelle nous permet d'obtenir l'égalité suivante

$$p(x) = \mathbb{E}_{q_\phi(z|x)}\left[\frac{p(x,z)}{q_\phi(z|x)}\right] \tag{1}$$

En passant, au log, nous avons

$$\log(p(x)) = \log\left(\mathbb{E}_{q_\phi(z|x)}\left[\frac{p(x,z)}{q_\phi(z|x)}\right]\right)$$

Par concavité du logarithme, nous pouvons utiliser l'inégalité de Jensen, qui nous permet de borner la log-évidence

$$\log(p(x)) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p(x,z)}{q_\phi(z|x)}\right)\right] = \mathbb{E}_{q_\phi(z|x)}[\log(p(x,z)) - \log(q_\phi(z|x))] \tag{2}$$

2.3.1 $p(x,z)$

La formule d'une probabilité conditionnelle, nous donne pour $p(x,z)$

$$\log(p(x,z)) = \log(p(x|z)) + \log(p(z))$$

2.3.2 Attache aux données

Nous pouvons développer un terme d'attache aux données dans $\log(p(x|z))$ En effet, en partant sur le principe que notre décodeur est Gaussien, nous avons donc la loi suivante pour $p(x|z)$

$$p(x|z) \sim \mathcal{N}(\mu_\theta, \gamma^2\mathbb{I})$$

Où μ_θ représente la sortie du décodeur.

En utilisant la formule d'une distribution Gaussienne multivariée, nous pouvons donc développer le terme d'attache aux données comme suit

$$\log(p(x|z)) = -\frac{(x-\hat{x})^2}{2\gamma^2} - \log((2\pi)^{\frac{k}{2}}|\gamma^2\mathbb{I}|^{\frac{1}{2}})$$

Nous ne pouvons pas jouer sur le terme d'échelle de la Gaussienne : $\frac{k}{2}\log(2\pi)$, nous allons donc utiliser la proportionnalité pour exprimer

$$\log(p(x|z)) \propto -\frac{(x-\hat{x})^2}{2\gamma^2} - k\log(\gamma) \tag{3}$$

À noter qu'ici, nous décidons de paramétrer la variance du décodeur par un paramètre γ qui est appris durant l'entrainement.

Note : k ici est une constante qui représente la dimension de x.

2.3.3 Reste de l'ELBO

Continuons par développer le reste de l'elbo dans notre équation. Il nous reste donc à développer les termes suivants :

$$\mathbb{E}_{q_\phi(z|x)}[\log(p(z)) - \log(q_\phi(z|x))]$$

Ces deux termes peuvent se regrouper pour donner notamment

$$-\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{q_\phi(z|x)}{p(z)}\right)\right] = -\int q_\phi(z|x)\log\left(\frac{q_\phi(z|x)}{p(z)}\right) = -\mathcal{KL}(q_\phi(z|x)||p(z))$$

Notre ELBO peut donc s'écrire finalement de la façon suivante (on écrit le $-ELBO$ car nous allons l'optimiser en minimisant :

$$-ELBO = \frac{(x-\hat{x})^2}{2\gamma^2} + k\log(\gamma) + \mathcal{KL}(q_\phi(z|x)||p(z)) \tag{4}$$

3 VAE Conditionnel

Dans cette section, nous allons explorer une nouvelle branche qui consiste à utiliser l'information d'une nouvelle image qui est liée d'une façon ou d'une autre à l'image d'origine pour conditionner notre modèle dessus.

3.1 Dépendances

Nous commençons par exprimer le graphe de dépendances suivant :

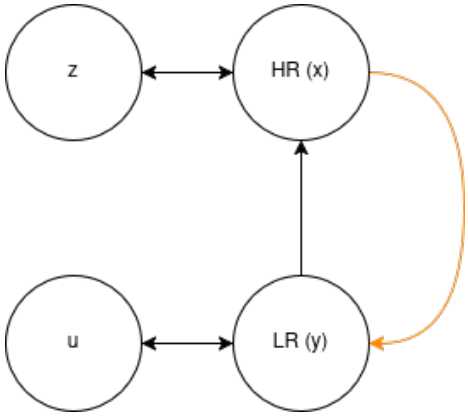

Figure 1: Graphique des dépendances variationnelles

Expliquons les dépendances :

- u est la variable latente de y qui dépend directement d'y.

- z est la variable latente de x qui dépend directement d'x.
- y est une image sous-résolue de x, la dépendance d'y par rapport à x est considérée comme déterministe.
- x dépend de y étant sa version Haute Résolution, la transformation de y à x est "surjective" (plusieurs x pour un y donné).

3.1.1 Simplifications

Étant donné les dépendances vues plus tôt, nous allons nous autoriser plusieurs simplifications, notamment les suivantes :

- $y|x$ déterministe.
- u ne dépend pas de x dans notre modèle.
- $q(z|y,x,u) \approx q(z|x)$ soit z capture les informations contenues dans x qui ne sont pas dans y et u.

Pour le second point, nous considérons que les informations apportées par y sont suffisantes dans sa propre représentation. (Autrement dit, pas besoin de x pour la reconstruction de y, nous nous ramenons alors à un VAE classique). De plus, nous n'avons pas accès à x dans un cas de super-résolution, donc nous ne pourrions pas l'utiliser hors de l'entrainement, donc ce qui poserait un problème pour sampler dans le modèle.

3.1.2 Discussion

Nous allons dans un premier temps développer les équations de la même manière que dans [1].

Il apparait évident que certaines simplifications sont potentiellement "optimistes". Nous pouvons notamment penser à enlever la dépendance en u pour reconstruire x. D'un premier point de vue, l'information est redondante avec y (u étant au mieux un changement de base de y, au pire une version compressée). Nous élaborerons les calculs avec puis sans cette simplification et nous observerons son effet sur les résultats.

3.2 Évidence

De la même manière que dans 2, nous voulons maximiser $p(x)$. Commencons par poser $w = [u, z, y]$. Avec notamment :

- z, la variable latente de x
- y, la version basse résolution de x
- u, la variable latente de y

Nous pouvons donc exprimer $p(x)$ comme la marginalisation de la loi jointe de (x, w).

$$p(x) = \int p(x, w)dw = \int \frac{p(x, w)}{q(w|x)} q(w|x)dw \tag{5}$$

L'espérance conditionnelle est alors

$$p(x) = \mathbb{E}_{q(w|x)} \left[\frac{p(x, w)}{q(w|x)} \right] \tag{6}$$

Commençons par réexprimer la loi jointe :

$$p(x, w) = p(x|y, u, z)p(z|y, u)p(y|u)p(u) \tag{7}$$

Nous pouvons négliger la dépendance en u pour x, car les informations sont redondantes avec y. Cette loi jointe approximée est donc :

$$p(x, w) \approx p(x|y, z)p(z|y, u)p(y|u)p(u) \tag{8}$$

Nous avons également :

$$q(w|x) = q(z|y, x, u)q(u|x, y)q(y|x) \tag{9}$$

Utilisant les simplifications de 3.1.1, nous pouvons donc supprimer les dépendances "inutiles" dans $q(z|y, u, x)$

$$q(w|x) \approx q(z|x)q(u|y)q(y|x) \tag{10}$$

3.3 Expression de la log-évidence

En passant cette équation au log, nous pouvons donc exprimer l'évidence de la façon suivante, commençons par l'évidence :

$$p(x) = \mathbb{E}_{q(z|x)q(u|y)q(y|x)} \left[\frac{p(x|u, z, y)p(z|y, u)p(y|u)p(u)}{q(z|x)q(u|y)q(y|x)} \right]$$

En utilisant la formule de Jensen pour décomposer le terme de droite, nous pouvons obtenir l'inégalité suivante :

$$\log(p(x)) \geq \mathbb{E}_{q(z|x)q(u|y)q(y|x)} \left[\log \left(\frac{p(x|z, y)p(z|y, u)p(y|u)p(u)}{q(z|x)q(u|y)q(y|x)} \right) \right]$$

En développant les différents termes de cette équation, nous obtenons l'équation suivante que nous allons étudier termes par termes.

$$\begin{aligned} \log(p(x)) \geq \mathbb{E}_{q(z|y,x)} \left[\log(p(x|y, z)) \right] + \mathbb{E}_{q(z|y,x)q(u|y)} \left[\log(p(z|y, u)) \right] \\ + \mathbb{E}_{q(u|y)} \left[\log(p(y|u)) \right] + \mathbb{E}_{q(u|y)} \left[\log(p(u)) \right] - \mathbb{E}_{q(z|x)q(u|y)} \left[\log(q(z|y, x)) \right] \\ - \mathbb{E}_{q(u|y)} \left[\log(q(u|y)) \right] \end{aligned} \tag{11}$$

3.4 Attaches aux données

Comme dans la section 2.3.2, nous pouvons sortir plusieurs termes d'attache aux données de cette équation, nous pouvons reconnaitre dès le début que :

$$\mathcal{L}_x = \mathbb{E}_{q(z|y,x)}\left[\log(p(x|y,z))\right] \qquad \mathcal{L}_y = \mathbb{E}_{q(u|y)}\left[\log(p(y|u))\right]$$

Sont les deux termes d'attache aux données. Comme plus tôt, nous utilisons un décodeur à variance diagonale paramétrée par $\gamma_{x,y}$ qui sont deux variances apprises pour les deux décodeurs.

3.4.1 Attache à y

Cette partie est la plus simple car elle s'apparente le plus au VAE classique. En effet, nous avons donc un décodeur suivant la loi suivante :

$$p(y|u) \sim \mathcal{N}(\mu_\theta, \gamma_y^2 \mathbb{I}) \tag{12}$$

Développant cela, nous pouvons développer le terme $\mathcal{L}_y$ de la façon suivante

$$\mathcal{L}_y = \mathbb{E}_{q(y|u)}\left[-\frac{||y-\hat{y}||_2^2}{2\gamma_y^2} - N_y \log(\gamma_y) - \frac{N_y}{2}\log(2\pi)\right]$$

Puisque nous ne pouvons pas optimiser le terme en 2π, nous allons donc utiliser la proportionnalité pour optimiser le terme suivant :

$$\mathcal{L}_y \propto -\mathbb{E}_{q(y|u)}\left[\frac{||y-\hat{y}||_2^2}{2\gamma_y^2} + N_y \log(\gamma_y)\right] \tag{13}$$

Cette espérance, ici, se calcule de façon empirique, nous obtenons alors les termes suivants :

$$\mathcal{L}_y \propto -\mathbb{E}_{q(y|u)}\left[\frac{N_y}{2\gamma_y^2 N_y}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 + N_y \log(\gamma_y)\right]$$

Après factorisation, nous obtenons donc la formule suivante :

$$\mathcal{L}_y \propto -N_y * \mathbb{E}_{q(y|u)}\left[\frac{MSE(y,\hat{y})}{2\gamma_y^2} + log(\gamma_y)\right] \tag{14}$$

3.4.2 Attache à x

De la même façon que dans la sous-section précédente, nous pouvons développer le terme

$$\mathcal{L}_x = \mathbb{E}_{q(z|y,x)}\left[\log(p(x|y,z))\right]$$

Cela nous donne alors :

$$\mathcal{L}_x \propto \mathbb{E}_{q(z|y,x)}\left[-\frac{N_x}{2\gamma_x^2 N_x}\sum_{i=1}^{n}(x_i-\hat{x}_i)^2 - N_x\log(\gamma_x)\right]$$

Qui donne donc :

$$\mathcal{L}_x \propto -N_x * \mathbb{E}_{q(z|y,x)}\left[\frac{MSE(x,\hat{x})}{2\gamma_x^2} + \log(\gamma_x)\right]$$

Un lecteur avisé se demandera, à juste titre, où est passée la dépendance en y dans ce terme. Ce terme en réalité se cache dans la loi du décodeur où nous avons en réalité :

$$p(x|y,z) \sim \mathcal{N}(\mu_{\psi|y}, \gamma_x^2\mathbb{I}) \tag{15}$$

Le décodeur prend également y en entrée pour estimer x.

3.4.3 Espérances

Nous remarquerons que nous avons encore deux espérances dans les attaches aux données dont nous ne pouvons nous défaire. La solution ici étant de passer par une approximation de Monte-Carlo, soit, d'exprimer :

$$\mathcal{L}_x \propto -\frac{N_x}{N_b} * \sum_{i=1}^{N_b}\frac{MSE(x,\hat{x}[z_i])}{2\gamma_x^2} - N_x * \log(\gamma_x)$$

3.5 Distance au priors

De ce qu'il reste de non analysé dans l'ELBO, nous pouvons extraire plusieurs termes, qui sont, nous le verrons, des distances. Ainsi, nous pouvons extraire les deux distances suivantes

$$D_z = \mathbb{E}_{q(z|y,x)q(u|y)}\left[\log(p(z|y,u))\right] - \mathbb{E}_{q(z|y,x)q(u|y)}\left[\log(q(z|y,x))\right]$$

$$D_u = \mathbb{E}_{q(u|y)}\left[\log(p(u))\right] - \mathbb{E}_{q(u|y)}\left[\log(q(u|y))\right]$$

3.5.1 Prior Gaussien

Nous allons commencer par analyser D_u. Ce cas est relativement simple, car il s'agit, tout simplement, du VAE classique développé dans la section 2.3.3.

$$D_u = \mathbb{E}_{q(u|y)}\left[\log\left(\frac{p(u)}{q(u|y)}\right)\right]$$

Par définition de l'espérance conditionnelle, nous avons donc en réalité :

$$D_u = -\int q(u|y)\log\left(\frac{q(u|y)}{p(u)}\right) = -\mathcal{KL}(q(u|y)||p(u))$$

On reconnait la divergence de Kullback-Leibler, la distance dont nous parlions plus tôt.

3.5.2 Prior conditionnel

Passons maintenant à la distance D_z qui est moins évidente. Le calcul se fait de la même façon et nous obtenons rapidement :

$$D_z = -\mathcal{KL}(q(z|x)||p(z|y,u)) \tag{16}$$

La différence est donc que la prior de la branche Haute-Résolution n'est pas une gaussienne, mais une loi qui sera apprise. Ne connaissant pas la loi, $z|y,u$ nous sommes donc obligés d'effectuer l'approximation par un réseau de neurones de la façon suivante :

$$p(z|y,u) \approx p_\phi(z|y,u)$$

3.6 ELBO

La négative ELBO (que l'on minimisera) s'exprime alors de façon très simple :

$$\begin{aligned} -ELBO = N_y \left[\frac{MSE(y,\hat{y})}{2\gamma_y^2} + \log(\gamma_y)\right] + N_x \left[\frac{MSE(x,\hat{x})}{2\gamma_x^2} + \log(\gamma_x)\right] \\ + \mathcal{KL}(q(z|x)||p(z|y,u)) + \mathcal{KL}(q(u|y)||p(u)) \end{aligned} \tag{17}$$

3.7 Modèle plus complexe

Plusieurs simplifications ont été faites dans le modèle précédent. Certaines sont judicieuses, d'autres, peut-être moins.

Dans un cadre de super-résolution ou résolution d'un problème conditionnel, il est judicieux de supprimer la dépendance en x pour l'étage y quand nous voulons inférer x sachant y. C'est notre cas ici. Il peut notamment être intéressant de réinclure la dépendance en u pour la reconstruction de x de cette façon, nous pourrions propager les gradients de MSE dans les blocs de réseaux neuronaux qui s'occupent de la tâche de super-résolution. Cela devrait nous aider à améliorer les performances de la tâche en amont. L'implémentation est un problème d'architecture que nous n'étudierons pas ici.

Il peut également être intéressant d'étudier le rajout des dépendances en y, u dans z.

4 Dé-simplifications

4.1 Présentation des dé-simplifications

Dans la partie 3, nous nous sommes demandé ce qu'il se passerait si nous réintroduisions certaines dépendances ignorées dans le modèle d'origine. Notamment, le graphe de dépendance deviendrait donc :

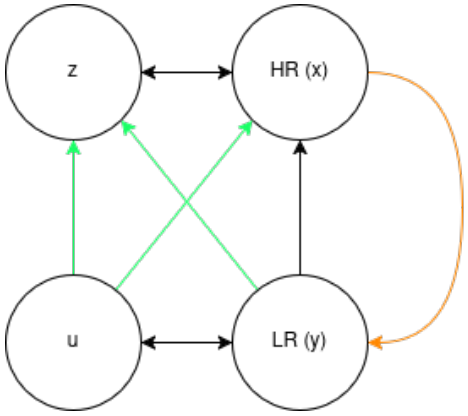

Figure 2: Modèle plus complexe

Par rapport à la section précédente, nous reprenons donc la loi jointe :

$$p(x, w) = p(x|y, u, z)p(z|y, u)p(y|u)p(u)$$

Nous ne modifions pas cette loi jointe, cette fois.

La loi conditionnelle $q(w|x)$, par contre est modifiée :

$$q(w|x) = q(z|y, x, u)q(u|x, y)q(y|x) \tag{18}$$

Nous allons garder ici la dépendance en y, u dans z mais bien enlever x dans u. La justification est que si nous gardions cette dépendance en x, nos dépendances deviendrait cycliques, nous aurions besoin de x pour sampler u qui permettrait alors de sampler z pour retrouver ... x.

Nous avons donc

$$q(w|x) \approx q(z|y, x, u)q(u|y)q(y|x) \tag{19}$$

4.2 Évidence modifiée

Nous allons développer les calculs en réutilisant une grande majorité de la partie 3 En reprenant l'expression de l'évidence vue plus tôt, nous obtenons :

$$p(x) = \mathbb{E}_{q(z|x,y,u)q(u|y)q(y|x)}\left[\frac{p(x|u, z, y)p(z|y, u)p(y|u)p(u)}{q(z|x, y, u)q(u|y)q(y|x)}\right]$$

L'inégalité de Jensen permet de développer les termes suivants :

$$\log(p(x)) \geq \mathbb{E}_{q(z|y,x,u)q(u|y)q(y|x)}\left[\log(p(x|y,z,u))\right]+\mathbb{E}_{q(z|u,y,x)q(u|y)}\left[\log(p(z|y,u))\right] +\mathbb{E}_{q(u|y)}\left[\log(p(y|u))\right]+\mathbb{E}_{q(u|y)}\left[\log(p(u))\right]-\mathbb{E}_{q(z|x,y,u)q(u|y)q(y|x)}\left[\log(q(z|y,x,u))\right] - \mathbb{E}_{q(u|y)}\left[\log(q(u|y))\right] \quad (20)$$

En utilisant les différentes parties vues précédemment, (2.3.2 et 2.3.3), nous développons les termes suivants séparément :

4.3 Attaches

L'attache à y, n'est pas modifiée, nous conservons alors :

$$\mathcal{L}_y \propto -\frac{N_y}{N_b} * \sum_{i=1}^{N_b} \frac{MSE(y, \hat{y}[z_i])}{2\gamma_y^2} - N_y * \log(\gamma_y)$$

4.3.1 Attache à x

L'attache à x cependant est modifiée, car nous introduisons une dépendance en y,u dans le terme $\mathcal{L}_x$.

$$\mathcal{L}_x = \mathbb{E}_{q(z|y,x,u)q(u|y)q(y|u)}\left[\log(p(x|y,z,u))\right]$$

Dans le cas d'un décodeur Gaussien, nous gardons une structure similaire à plus tôt, avec

$$\mathcal{L}_x \propto -\frac{N_x}{N_b} * \sum_{i=1}^{N_b} \frac{MSE(x_i, \hat{x}_i[z_i, u_i, y_i])}{2\gamma_x^2} - N_x * \log(\gamma_x)$$

Avec la loi de $\hat{x}$ modifiée comme suit :

$$x \sim \mathcal{N}(\mu_\eta, \gamma_x^2 \mathbb{I}|y, z, u)$$

4.4 Priors

4.4.1 Prior Gaussien

Nous reprenons les calculs développés dans la partie 3.5

$$D_u = -\int q(u|y) \log\left(\frac{q(u|y)}{p(u)}\right) = -\mathcal{KL}(q(u|y)||p(u))$$

4.4.2 Prior Conditionnel

Ce prior qui était de la forme :

$$D_z = -\mathcal{KL}(q(z|x)||p(z|y,u))$$

Devient donc, en ajoutant les dépendances :

$$D_z = \mathbb{E}_{q(z|u,y,x)q(u|y)q(y|x)}\left[\log(p(z|y,u))\right] - \mathbb{E}_{q(z|x,y,u)q(u|y)q(y|x)}\left[\log(q(z|y,x,u))\right]$$

Nous pouvons utiliser la formule de l'espérance conditionnelle une nouvelle fois pour pouvoir développer ce calcul

$$D_z = -\mathbb{E}_{q(z|u,y,x)q(u|y)q(y|x)}\left[\log\left(\frac{q(z|y,x,u)}{p(z|y,u)}\right)\right]$$

$$D_z = -\mathbb{E}_{q(u|y)q(y|x)}\int q(z|u,y,x)\log\left(\frac{q(z|y,x,u)}{p(z|y,u)}\right)dz$$

$$D_z = -\mathcal{KL}(q(z|x,u,y)||p(z|y,u)) \tag{21}$$

5 MoG

Admettons maintenant un mélange de gaussiennes :

$$p(u) = MoG(w,\mu,\sigma) = \sum_{i=1}^{Comp} w_i\mathcal{N}(\mu_i,\sigma_i^2) \tag{22}$$

Nous ne changeons qu'une chose : le prior de u. Qu'est ce que cela traduit pour l'ELBO ? [2] montre que nous pouvons borner

$$\mathcal{KL}(p||\sum_i w_i q_i) \leq \sum_i w_i\mathcal{KL}(p||q_i)$$

Ainsi, nous avons donc une borne inf de l'ELBO qui est une borne inf de $p(x)$.

Notre ELBO deviendrait alors :

$$-ELBO = N_y\left[\frac{MSE(y,\hat{y})}{2\gamma_y^2} + \log(\gamma_y)\right] + N_x\left[\frac{MSE(x,\hat{x})}{2\gamma_x^2} + \log(\gamma_x)\right] + \mathcal{KL}(q(z|x)||p(z|y,u)) + \sum_{i=1}^{n_w} w_i\mathcal{KL}(q(u|y)||p_i(u)) \tag{23}$$

Avec $p_i(u) \sim \mathcal{N}(\mu_i,\sigma_i^2)$.

References

[1] Ioannis Gatopoulos, Maarten Stol, and Jakub M. Tomczak. *Super-Resolution Variational Auto-Encoders*. June 2020. DOI: 10.48550/arXiv.2006.05218. arXiv: 2006.05218 [cs]. (Visited on 05/15/2025).

[2] John R. Hershey and Peder A. Olsen. "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models". In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Vol. 4. Apr. 2007, pp. IV-317-IV–320. DOI: 10.1109/ICASSP.2007.366913. (Visited on 05/21/2025).