Autoencodeurs variationnels à registre de vecteurs pour la détection d'anomalies

Hugo Gangloff, Minh-Tan Pham, Luc Courtrai, Sébastien Lefèvre

Autoencodeurs variationnels à registre de vecteurs pour la détection d'anomalies

Hugo Gangloff Minh-Tan Pham Luc Courtrai Sébastien Lefèvre

IRISA, Université Bretagne Sud, UMR 6074, 56000 Vannes, France

hugo.gangloff@irisa.fr

Résumé

La Détection d'Anomalies (AD) en contexte non-supervisé est une thématique de recherche importante. En pratique, les anomalies ne sont pas connues à l'avance. Dans ce cadre, les Autoencodeurs Variationnels (VAE) constituent une approche très populaire et ont donné naissance aux Autoencodeurs Variationnels à registre de vecteurs (VQ-VAE). Nous présentons une nouvelle approche tirant profit de métriques inhérentes aux VQ-VAE pour l'AD. Dans notre approche, nous calculons une distance entre la sortie de l'encodeur et les vecteurs du registre qui définit l'espace latent. Cette distance vient compléter l'information apportée par une seconde métrique exploitant les capacités de reconstruction du VQ-VAE. Nous évaluons notre modèle sur trois jeux de données et montrons qu'il est compétitif avec d'autres modèles issus de l'état de l'art.

Mots Clef

détection d'anomalies, autoencodeurs variationnels à registre de vecteurs, traitement de l'image

Abstract

Unsupervised Anomaly Detection (AD) is an important research topic. In many practical cases, the anomalies are unknown in advance. In this context, Variational Autoencoders (VAEs) are widely used and have been extended to Vector-Quantized VAEs (VQ-VAEs). We present for the first time a robust approach which takes advantage of the inner metrics of VQ-VAEs for AD. In our approach, the distance between the output of the encoder and the codebook vectors of a VQ-VAE complements a reconstruction-based metric to improve AD results. We compare our model with state-of-the-art AD models on three standard datasets. Experiments show that the proposed method yields high competitive results.

Ce travail a été effectué dans le cadre du projet Game of Trawls. Nous remercions le Fonds européen pour les affaires maritimes et la pêche (contrat numéro 18/2216442) et France Filière Pêche (contrat numéro 19/1000544) pour le financement. Ce travail a aussi été effectué dans le cadre du projet SEMMACAPE, financé par l'Agence de la transition écologique lors de l'appel à projet "Sustainable Energies" (2018–2019).

Keywords

anomaly detection, vector-quantized variational autoencoders, image processing

1 Introduction

1.1 La détection d'anomalies

La Détection d'Anomalies (AD) est un domaine de recherche historique avec un large spectre d'applications telles que la détection de défauts industriels, le diagnostic médical, la détection de fraude, la détection d'intrusion, etc. [20] Une anomalie est une observation qui diffère des autres observations de manière importante, au point de faire croire qu'elle a été créée par un autre processus [6]. L'observation anormale diffère donc d'une normalité qui est induite par les autres observations. L'AD a été révolutionnée par les approches d'apprentissage profond qui ont permis de capturer et modéliser la normalité avec une précision inégalée jusqu'alors [24].

L'AD dans un cadre non-supervisé ou faiblement supervisé est l'approche la plus courante en pratique. En effet, les anomalies ne sont pas connues à l'avance ; il n'est donc pas possible de réunir et d'étiqueter des données anormales comme dans une approche supervisée classique. Nous nous concentrons dans cet article sur l'AD non-supervisée dans les images. Elle se divise en trois grandes approches pouvant se recouper en pratique [28]. L'AD peut être basée sur :

- des caractéristiques extraites des données, que l'on classifie par la suite [26, 4] ;
- une approche probabiliste (tests de vraisemblance, tests statistiques, etc.) [18, 17] ;
- une reconstruction issue du modèle, que l'on compare à l'image d'entrée [1, 31].

Les méthodes fondées sur une reconstruction sont les plus populaires dans la littérature ; elles mettent souvent en jeu des modèles à variables latentes, appelés Autoencodeurs Variationnels (VAE), que nous présentons maintenant. Notons que la nouvelle approche présentée dans cet article peut être vue comme la combinaison d'une approche de type reconstruction et d'une approche reposant sur des caractéristiques extraites.

Remarque : l'AD pour les images regroupe deux grandes catégories. D'une part, l'AD au niveau de l'image, où l'on cherche à séparer l'image anormale d'autres images anormales. D'autre part, l'AD au niveau du pixel, où l'on cherche à localiser les pixels anormaux de chaque image. Les deux catégories d'AD sont étudiées dans cet article.

1.2 La Détection d'Anomalies avec les VAE

Dans les cadres de l'AD non-supervisée ou faiblement supervisée, les modèles génératifs profonds sont populaires [3]. Parmi ces modèles, les Autoencodeurs Variationnels (VAE) sont très utilisés [8]. Les VAE sont introduits dans un cadre probabiliste qui est détaillé dans [7]. En résumé, les VAE transforment une image d'entrée $\boldsymbol{x}$ en une représentation latente et compressée $\boldsymbol{z}$ à l'aide d'un encodeur stochastique $q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})$. L'image est alors reconstruite en $\hat{\boldsymbol{x}}$ à l'aide d'un décodeur stochastique $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. L'entrainement du modèle maximise une borne inférieure de la vraisemblance,

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\varphi}}(\boldsymbol{x}) = \underbrace{\mathbb{E}_{q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]}_{\text{terme de reconstruction}} - \underbrace{\mathbb{KL}(q_{\boldsymbol{\varphi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}))}_{\text{terme de régularisation}}, \tag{1}$$

via l'optimisation alternée sur les ensembles de paramètres $\boldsymbol{\varphi}$ et $\boldsymbol{\theta}$. Intuitivement, le premier terme de l'Éq. (1) favorise les reconstructions $\hat{\boldsymbol{x}}$ qui sont similaires aux entrées $\boldsymbol{x}$ sous la contrainte d'un terme de régularisation (la divergence de Kullback Leibler (KL)). Dans un VAE, plusieurs métriques peuvent alors être calculées à partir de l'espace latent ou de la reconstruction. Ces métriques sont utilisées dans plusieurs approches pour l'AD au niveau de l'image et au niveau du pixel. Notons par exemple l'utilisation de l'image résiduelle entre $\boldsymbol{x}$ et $\hat{\boldsymbol{x}}$, de la valeur du terme de reconstruction ou du terme de divergence de KL, de la dérivée de l'Éq. (1) par rapport à $\boldsymbol{x}$, etc. Appliquées à l'AD faiblement supervisée, ces métriques peuvent être retrouvées, par exemple, dans [1, 31, 12, 5, 27].

Plus récemment, les Autoencodeurs Variationnels à registre de vecteurs (VQ-VAE) [19, 22] ont été introduits. Ils seront vus en détail dans la Section 2.1. Quelques articles font état de l'utilisation des VQ-VAE pour l'AD (par exemple [28, 17, 21]). Cependant, ces articles développent des modèles plus complexes qui reposent sur un apprentissage en deux étapes avec des modèles autorégressifs profonds entrainés sur les espaces latents discrets. Nous ne nous comparerons pas avec cette famille d'approches. En effet, notre contribution principale est de montrer que les VQ-VAE peuvent, à eux seuls, apprendre des représentations de la normalité engendrant des métriques robustes et compétitives pour l'AD. Nous suivons une démarche similaire aux meilleurs résultats d'AD obtenus avec les VAE. Dans le contexte de l'AD, les métriques inhérentes aux VAE ont été très étudiées à l'inverse de celles des VQ-VAE.

1.3 Organisation de l'article

Dans la section suivante, nous introduisons les VQ-VAE et leur utilisation pour l'AD. Nous présentons ensuite une approche nouvelle et robuste fondée sur les métriques inhérentes aux VQ-VAE pour l'AD aux niveaux de l'image et du pixel. Dans la dernière section, notre approche est comparée à d'autres modèles de l'état de l'art sur des jeux de données classiques : MVTec [2], UCSD-Ped1 [15] et CIFAR-10 [9].

2 Autoencodeurs Variationnels à registre de vecteurs

2.1 Le modèle

Les VQ-VAE, introduits dans [19], sont des modèles dont l'espace latent est discret et relié à un registre de vecteurs. Des études ont montré qu'ils peuvent apprendre des représentations riches et compressées, tout en produisant des reconstructions plus détaillées que les VAE classiques [22, 23]. L'utilisation des VQ-VAE pour la détection d'anomalies semble alors intéressante, en particulier, nous pouvons nous attendre à une image résiduelle moins bruitée dans les approches par reconstruction.

L'espace latent discret et le registre de vecteurs impose une procédure d'entrainement différente pour les VQ-VAE par rapport aux VAE. L'encodeur, noté $\mathrm{Enc}_{\boldsymbol{\varphi}}$, devient ici déterministe. Soit M le nombre d'états possibles pour les variables latentes $z_k, \forall k \in \{1, \dots, K\}$, où K est la dimension de l'espace latent. Les vecteurs, $(e_1, \dots, e_M)$, du registre sont à valeurs dans $\mathbb{R}^D$, où $D \in \mathbb{N}^*$. À partir d'une sortie de l'encodeur, $\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}$, nous choisissons le vecteur du registre le plus proche, $z_k, \forall k$:

$$z_k = \underset{m \in \{1,\dots,M\}}{\arg\min} \|(z_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})})_k - e_m\|_2. \tag{2}$$

Cela peut être interprété dans le cadre des VAE classiques par l'introduction d'une distribution variationnelle discrète et déterministe

$$q_{\boldsymbol{\varphi}}(z_k = m|\boldsymbol{x}) = \begin{cases} 1 \text{ si } m = \underset{m \in \{1,\dots,M\}}{\arg\min} \|(z_{\mathrm{Enc}_{\boldsymbol{\varphi}}})_k - e_m\|_2, \\ 0 \text{ sinon.} \end{cases} \tag{3}$$

L'entrée du décodeur, notée $\boldsymbol{z}_{\mathrm{Dec}_{\boldsymbol{\theta}}}$, est elle aussi déterministiquement définie par $(z_{\mathrm{Dec}_{\boldsymbol{\theta}}})_k = e_{z_k}, \forall k$. Dans le cas où des images sont traitées par un encodeur et un décodeur convolutionnels, comme dans le cas de cet article, l'espace latent $\boldsymbol{z}$ peut lui aussi être convolutionnel ; il peut alors être vu comme une image latente.

La fonction de coût d'un VQ-VAE inclut le même terme de reconstruction que le VAE. Elle inclut aussi un terme spécifique qui permet l'apprentissage des paramètres du registre de vecteurs. En effet, le gradient ne peut se propager à travers les opérations déterministes citées plus haut : il est alors copié automatiquement évitant le registre de vecteurs qui n'est alors pas mis à jour, d'où l'introduction d'un second terme appelé *terme d'alignement*. Un troisième terme de régularisation est aussi ajouté pour la stabilité de l'apprentissage. Finalement, pour une entrée $\boldsymbol{x}$, la fonction de

coût du VQ-VAE est :

$$\mathcal{L}^{VQ-VAE}_{\boldsymbol{\theta},\boldsymbol{\varphi},\boldsymbol{e}}(\boldsymbol{x}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}_{\mathrm{Dec}_{\boldsymbol{\theta}}(\boldsymbol{x})}) + \underbrace{\|\mathrm{sg}[\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}] - \boldsymbol{e}\|_2^2}_{\text{terme d'alignement}} + \beta\|\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})} - \mathrm{sg}[\boldsymbol{e}]\|_2^2, \quad (4)$$

où β est un scalaire et sg est un opérateur qui stoppe le gradient.

2.2 Détection d'anomalies avec les VQ-VAE

Nous définissons maintenant des métriques pour l'AD avec les VQ-VAE. Nous nous intéressons principalement à l'AD au niveau du pixel. Ainsi, nous définissons d'abord une métrique de reconstruction, appelée Carte de Similarité (SM), qui utilise de manière sous-jacente la métrique de similarité des structures (SSIM) [29]. Pour un pixel i :

$$\mathrm{SM}(x_i) = \mathrm{SSIM}(\boldsymbol{p}_i, \boldsymbol{q}_i) = \frac{(2\mu_{\boldsymbol{p}}\mu_{\boldsymbol{q}} + c_1)(2\sigma_{\boldsymbol{pq}} + c_2)}{(\mu^2_{\boldsymbol{p}} + \mu^2_{\boldsymbol{q}} + c_1)(\sigma^2_{\boldsymbol{p}} + \sigma^2_{\boldsymbol{q}} + c_2)}, \quad (5)$$

où $\boldsymbol{p}_i$ (respectivement $\boldsymbol{q}_i$) est une zone autour du pixel i de $\boldsymbol{x}$ (respectivement $\hat{\boldsymbol{x}}$). μ_p, σ_p et σ_{pq} représentent, respectivement, la moyenne, la variance et la covariance de la zone. Les valeurs des scalaires sont fixées à $c_1 = 0.01$ et $c_2 = 0.03$ [29].

Nous définissons aussi une nouvelle métrique sur l'espace latent qui engendre une carte d'anomalies, appelée Carte d'Alignement (AM) :

$$\mathrm{AM}(\boldsymbol{x}) = \|\mathrm{sg}[\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}] - \boldsymbol{e}\|_2^2. \quad (6)$$

Le raisonnement motivant l'AM est le suivant. Lors de l'entrainement, les vecteurs du registre se rapprochent spatialement de la sortie de l'encodeur et réciproquement, en vertu des deux derniers termes de l'Éq. (4). C'est pourquoi, lors du test, les anomalies (encodées dans $\boldsymbol{z}_{\mathrm{Enc}_{\boldsymbol{\varphi}}(\boldsymbol{x})}$), qui n'ont jamais été vues par le modèle, seront plus éloignées spatialement des vecteurs du registre, que les caractéristiques normales vues lors de l'entrainement.

Notons que cette carte d'anomalies est définie dans l'espace latent et n'est pas directement utilisable pour l'AD au niveau du pixel. Dans la section suivante, nous proposons une approche efficace pour la segmentation d'anomalies avec l'AM.

Remarque : L'AM peut être reliée aux approches utilisant le terme de divergence de KL pour la segmentation d'anomalies dans les VAE [31]. Cependant, comme nous le montrerons dans les expériences, l'AM d'un VQ-VAE semble être plus robuste et interprétable car nous obtenons, avec les VQ-VAE, des résultats supérieurs aux VAE.

2.3 Utilisation de la Carte d'Alignement pour améliorer l'AD

L'AM peut être vue comme une petite image de même dimension que l'espace latent, où quelques pixels ont des valeurs plus intenses. Ces pixels peuvent être vus comme des marqueurs. Ils correspondent aux variables latentes pour lesquelles le terme d'alignement est élevé, c'est-à-dire, selon notre approche, des anomalies. Pour être utilisée avec la SM, l'AM est d'abord suréchantillonnée. Nous lui appliquons ensuite une dilatation morphologique à niveaux de gris qui a pour but de mettre en évidence les marqueurs. Cependant, aucun marqueur ne représente fidèlement les anomalies car ils sont simplement issus d'un suréchantillonnage. C'est pourquoi nous proposons la multiplication de l'AM avec la SM. La Figure 1 résume toutes les étapes de notre approche que nous appelons VQ-VAE SSIM+AM.

L'un des avantages de cette approche est qu'elle ne repose pas simplement sur la reconstruction du VQ-VAE. En effet, dans les approches traditionnelles des VAE, une hypothèse forte est faite : les anomalies qui n'ont pas été vues à l'entrainement vont disparaître à la reconstruction. Sous cette hypothèse, il sera alors facile de les isoler via une image résiduelle par exemple [1]. Cependant, à cause du flou intrinsèque présent dans les reconstructions, l'image résiduelle ne permet pas de facilement localiser les anomalies. C'est pourquoi, utiliser l'AM comme source d'information pour localiser les anomalies a de l'intérêt dans la mesure où l'on évite d'utiliser la reconstruction.

3 Expériences et Résultats

3.1 Architecture du modèle

Nous considérons la même architecture pour le VQ-VAE dans toutes les expériences qui suivent. Elle s'appuie sur l'architecture du VQ-VAE proposée dans [19] :

- L'encodeur comporte trois couches convolutionnelles (4/2/1), chacune suivie d'une activation de type ReLU et d'une *Batch Normalization*. Trois couches résiduelles suivent ensuite (composée d'une activation ReLU, d'une couche convolutionnelle (3/1/1), d'une activation ReLU et d'une couche convolutionnelle (1/1/0)). Toutes les couches ont une profondeur 256, mis à part les entrées qui ont une profondeur 1 ou 3.
- L'espace latent a les dimensions de hauteur et largeur des entrées divisées par 8 (lorsqu'il y a 3 convolutions dans l'encodeur). Le registre de vecteur a comme grandeurs caractéristiques $M = 512$ et $D = 256$.
- Le décodeur est construit comme miroir de l'encodeur.

Notons que la taille du registre de vecteur a un rôle crucial sur les capacités de compression et généralisation du modèle [30]. Nous reprenons ici les valeurs de M et D issues de [19] et [28] qui donnent lieu à très peu de compression au niveau de l'espace latent (registre de vecteur de grande taille). Cela améliore les capacités de reconstruction mais diminue par ailleurs les capacités de généralisation. Ce contexte semble favorable au cas de la détection d'anomalies : les anomalies auront du mal à être reconstruites et seront plus facilement détectées.

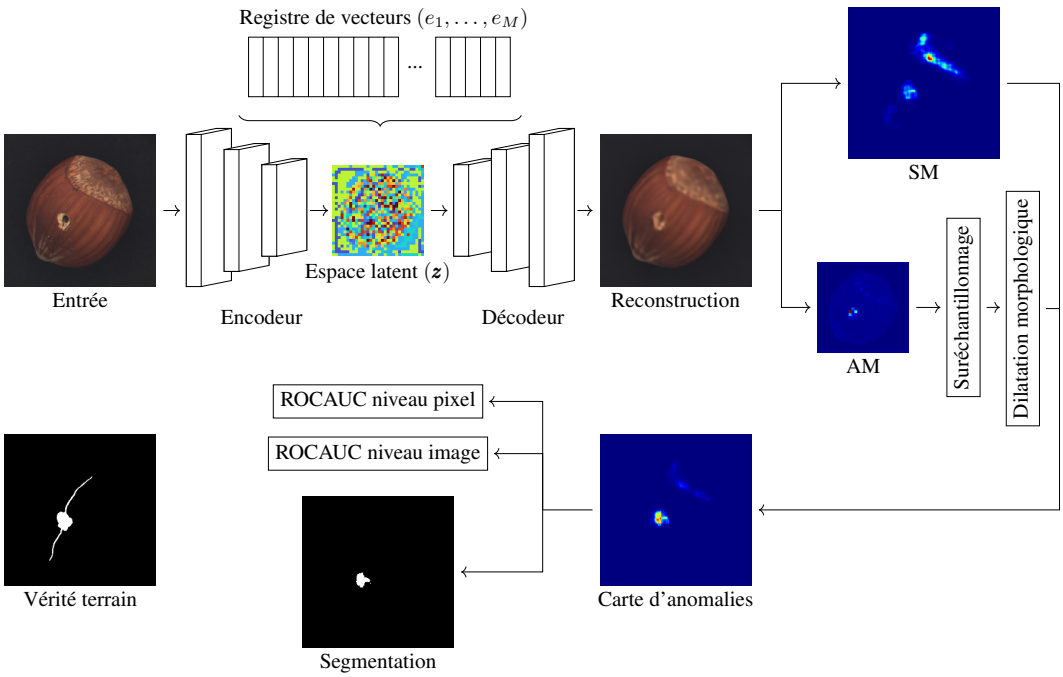


FIGURE 1 – L'approche proposée pour améliorer l'AD avec les VQ-VAE (Sections 2.2 et 2.3). L'architecture du modèle est décrite en Section 3.1. Nous appelons l'approche VQ-VAE SSIM+AM. La SM peut directement être utilisée pour les calculs de ROCAUC ou pour segmenter les anomalies. Dans ce cas l'approche est appelée VQ-VAE SSIM.

Notons également que nos choix d'architectures pour l'AD au niveau du pixel dans les sections suivantes sont le fruit d'un compromis entre l'architecture standard de VQ-VAE proposée par [19] et la volonté de conserver un espace latent convolutionnel de hauteur et largeur 32, comme proposé par [28]. Pour l'AD au niveau de l'image 3.4, la taille de l'espace latent sera nécessairement plus réduite car les images d'entrées sont de basse résolution.
De plus, nous recalons les images encodées sur 255 valeurs dans l'intervalle $[0, 1]$, et modélisons la sortie du décodeur $\boldsymbol{x}$ comme la réalisation d'une variable aléatoire de Bernoulli continue [14]. En effet, nous notons que cela menait à une convergence plus stable et rapide de l'étape d'estimation des paramètres, par rapport au cas où une distribution gaussienne modélise le décodeur $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$.

Dans la suite, le VQ-VAE est évalué contre des modèles comparables. Ainsi, les approches sélectionnées comportent elles aussi uniquement une architecture encodeur-décodeur (sans module complémentaire) et font un usage limité des pré- et post-traitements.

Remarque : Dans les expériences qui suivent, les modèles n'utilisent pas de données étiquetées. Cependant, les modèles ne sont entrainés qu'avec des données normales, dépourvues d'anomalies. En ce sens, il ne s'agit pas d'apprentissage non-supervisé mais faiblement supervisé.

3.2 Le jeu de données MVTec

Le jeu de données MVTec [2] est un jeu de données standard pour l'AD sur les images. Il contient des images RGB normales et défectueuses de 15 types d'objets manufacturés. Plusieurs types de défauts sont présents pour chaque type d'objet (5354 images au total, dont 1888 défectueuses). En accord avec la littérature, nous reportons les résultats en terme de ROCAUC au niveau du pixel. Cette grandeur est calculée à partir de la carte d'anomalies générée par les modèles et de la vérité terrain. L'approche complète pour l'AD avec le VQ-VAE, VQ-VAE SSIM+AM, a été décrite en Section 2.3. Nous reportons aussi les résultats de l'approche VQ-VAE SSIM, qui n'utilise pas l'AM. Nous comparons nos approches avec : un Autoencodeur classique avec SSIM (AE SSIM) [2], une approche récente *Visually Explained VAE* (VEVAE) [12] et une autre approche récente *Fully Convolutional Data Description* (FCDD) [13]. Nous n'utilisons pour l'entrai-

Catégorie	AE SSIM [2]	VEVAE [12]	FCDD [13]	*VQ-VAE SSIM*	*VQ-VAE SSIM+AM*
Tapis	0.87	0.78	**0.96**	*0.92*	*0.94*
Grille	0.94	0.73	0.91	***0.99***	***0.99***
Cuir	0.78	0.95	**0.98**	***0.98***	***0.98***
Carrelage	0.59	0.80	**0.91**	*0.70*	*0.75*
Bois	0.73	0.77	**0.88**	*0.82*	*0.84*
Bouteille	0.93	0.87	**0.97**	*0.94*	*0.95*
Câble	0.82	**0.90**	**0.90**	*0.87*	*0.87*
Capsule	**0.94**	0.74	0.93	*0.93*	***0.94***
Noisette	0.97	0.98	0.95	*0.98*	***0.99***
Écrou	0.89	0.90	**0.94**	*0.89*	*0.90*
Pillule	**0.91**	0.83	0.81	*0.86*	*0.90*
Vis	0.96	0.97	0.86	***0.98***	***0.98***
Brosse	0.92	0.94	0.94	*0.96*	***0.97***
Transistor	0.90	**0.93**	0.88	*0.77*	*0.78*
Fermeture	0.88	0.78	0.92	*0.97*	***0.98***
Moyenne	0.86	0.86	**0.92**	*0.90*	***0.92***

TABLE 1 – Scores de ROCAUC sur le jeu de données MVTec.

nement que des images dépourvues d'anomalies. Chaque modèle est entrainé sur chaque catégorie d'images et des techniques d'augmentation de données sont utilisées pour atteindre un dataset de taille 1024 images (rotation aléatoire et effet miroir aléatoire). Cette procédure est commune aux approches auxquelles nous nous comparons. De plus, à l'instar de [12] et [2], nous redimensionnons les images à la taille 256×256, ce qui produit un espace latent de taille 32×32.

Le Tableau 1 montre les scores pour chaque modèle pour chaque catégorie du jeu de données. Les scores pour les autres modèles sont tirés de leurs publications respectives. Nous pouvons voir que notre approche VQ-VAE AM+SSIM obtient des résultats similaires à l'approche FCDD qui constitue l'état de l'art pour cette famille de modèles. Notre approche donne des résultats meilleurs que AE SSIM et VEVAE. Cela pourrait de suggérer que les VQ-VAE constituent des architectures préférables aux Autoencodeurs et aux VAE pour ce type de tâches. Finalement, on note que l'utilisation de l'AM améliore la carte d'anomalies finale car l'approche VQ-VAE SSIM+AM donne un score meilleur que l'approche VQ-VAE SSIM.

Nous remarquons que les résultats les plus mauvais de l'approche VQ-VAE semblent être liés à des défauts relativement gros qui sont bien reconstruits (par exemple, les points noirs dans la catégorie *Carrelage* ou une partie d'objet manquante dans la catégorie *Transistor*). Dans ces cas, les métriques ne permettent pas la localisation des anomalies. À l'inverse, la force de cette approche semble résider dans sa grande capacité à détecter les défauts les plus subtils grâce à la qualité des reconstructions (par exemple, les trous dans la catégorie *Noisette* ou les petits défauts dans la catégorie *Vis*). La Figure 2 donne quelques illustrations de l'expérience.

	VAE L2 [1]	VEVAE [12]	*VQ-VAE SSIM*	*VQ-VAE SSIM+AM*
ROCAUC	0.86	0.92	***0.95***	***0.95***

TABLE 2 – Scores de ROCAUC sur le jeu de données UCSD-Ped1.

3.3 Le jeu de données UCSD-Ped1

Dans cette seconde expérience, nous considérons un autre jeu de données standard de l'AD pour les images et les vidéos : le jeu de données UCSD-Ped1 [15]. Ce dernier est composé d'images en nuance de gris de séquences vidéo de piétons dans un parc. Nous comptabilisons 6400 images pour l'entrainement et 2000 images pour le test. Les images sont redimensionnées à la taille 128×128 comme dans [12]. La métrique utilisée est à nouveau la ROCAUC au niveau des pixels. La tâche d'AD consiste ici à repérer tous les objects en mouvement qui ne sont pas des piétons (voitures, vélos, skateboards, etc.).

Nous comparons nos approches VQ-VAE avec une architecture de VAE classique [1] et le VEVAE [12]. À cause de la plus faible résolution des images de départ, nous réduisons l'encodeur et le décodeur présentés en Section 3.1 à deux couches convolutionnelles. L'espace latent est alors de dimension 32×32. À nouveau, pour tous les modèles comparés, seules des images dépourvues d'anomalies sont utilisées pour l'entraînement. De plus, nous avons constaté que des résultats bien meilleurs étaient obtenus par les approches VQ-VAE en effectuant une dilatation morphologique sur la SM. Nous expliquons cette adaptation de l'approche spécifique à cette expérience par la faible résolution des images du jeu de données.. Cette opération affecte visuellement la carte d'anomalies (voir les résultats finaux de la Figure 3).

Le Tableau 2 donne les scores de ROCAUC pour chaque modèle, sur toutes les images du jeu de données de test. Les scores des autres modèles sont tirés de la publication [12]. Nous pouvons voir que notre modèle obtient un meilleur score ce qui suggère que les cartes d'anomalies produites par les approches VQ-VAE sont plus pertinentes. Dans le cadre de cette expérience, l'apport de l'AM semble limité. Des illustrations de l'expérience sont données en Figure 3.

3.4 Le jeu de données CIFAR-10

Dans cette dernière expérience nous abordons le problème de l'AD au niveau des images entières et considérons le jeu de données CIFAR-10 [9]. Ce jeu de données regroupe 50000 images d'entrainement et 10000 images de test équitablement réparties sur 10 catégories. Les images de taille 32×32 sont utilisées sans redimensionnement. Les modèles sont évalués selon la logique *un-contre-tous*. C'est-à-dire qu'un modèle est entrainé sur les images d'une catégorie seulement et qu'au moment de l'évaluation, il s'agit de différencier les images de la classe normale, sur laquelle le modèle a été entrainé, et les images de la classe anormale,

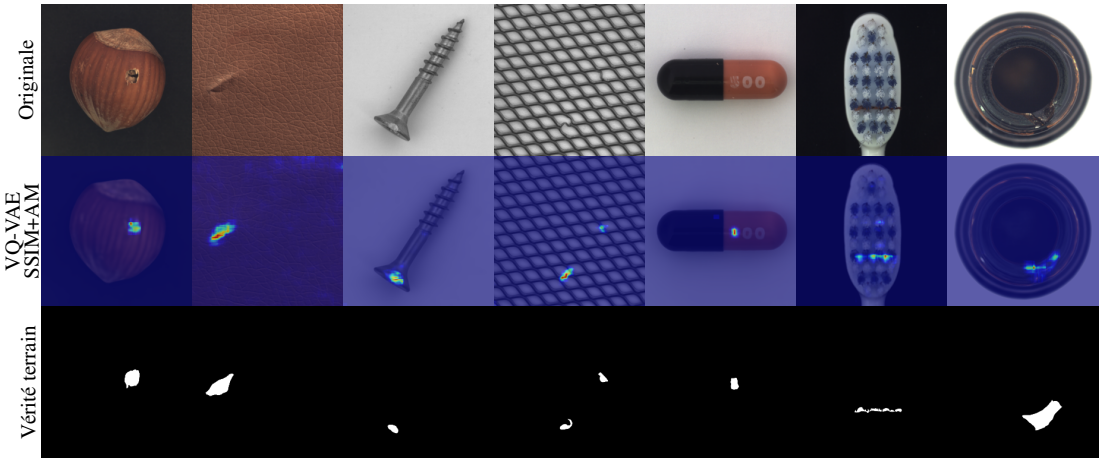


FIGURE 2 – Sélection d'illustrations pour l'expérience MVTec. *Haut :* images originales. *Milieu :* les cartes d'anomalies superposées aux reconstructions du VQ-VAE. *Bas :* La segmentation des anomalies.

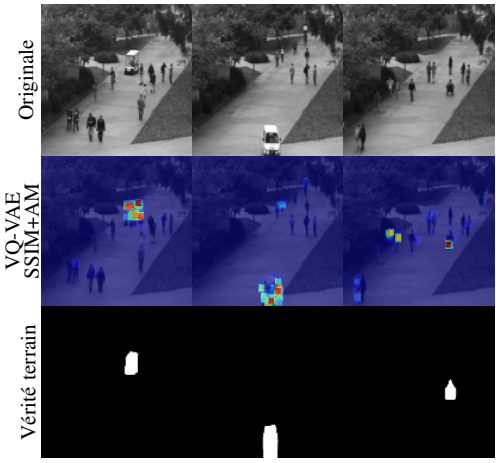


FIGURE 3 – Sélection d'illustrations du modèles VQ-VAE SSIM+AM pour l'expérience UCSD-Ped1.

qui regroupe toutes les autres classes. La métrique utilisée est alors la ROCAUC au niveau des images entières.

Nous comparons nos approches basées sur les VQ-VAE avec un Autoencodeur [16] et l'approche *Deep Support Vector Data Description* [25], une approche classique pour l'AD sur des images entières. Afin d'obtenir un score d'anomalie pour une image entière dans les approches VQ-VAE, nous prenons la valeur moyenne de la carte d'anomalies finale. Notons que, par rapport à la description faite en Section 3.1, les architectures ne sont ici composées que de deux couches convolutionnelles. L'espace latent est alors de taille 8×8.

Le Tableau 3 donne les scores pour tous les modèles pour toutes les catégories du jeu de données. Les scores pour les autres modèles sont tirés de la publication [25]. Nous pouvons voir que l'approche VQ-VAE améliore les scores des autres modèles et semble donc également produire des cartes d'anomalies pertinentes pour de l'AD sur des

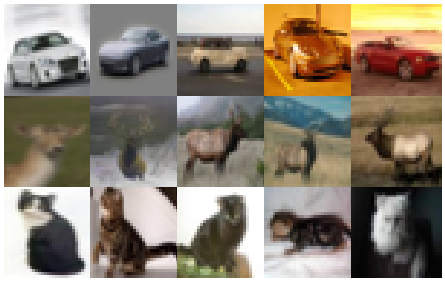
Catégorie	AE L2 [16]	DSVDD [25]	*VQ-VAE SSIM*	*VQ-VAE SSIM+AM*
Avion	0.59	0.62	***0.71***	*0.69*
Voiture	0.57	**0.66**	*0.63*	*0.65*
Oiseau	0.49	0.51	*0.63*	***0.65***
Chat	0.58	0.59	*0.62*	***0.63***
Biche	0.54	0.61	*0.60*	***0.64***
Chien	0.62	0.66	***0.67***	***0.67***
Grenouille	0.51	**0.68**	*0.61*	*0.63*
Cheval	0.59	**0.67**	*0.63*	*0.63*
Bâteau	**0.77**	0.76	*0.74*	*0.74*
Camion	0.67	**0.73**	*0.64*	*0.65*
Moyenne	0.59	0.65	*0.65*	***0.66***

TABLE 3 – Scores de ROCAUC sur le jeu de données CIFAR-10.

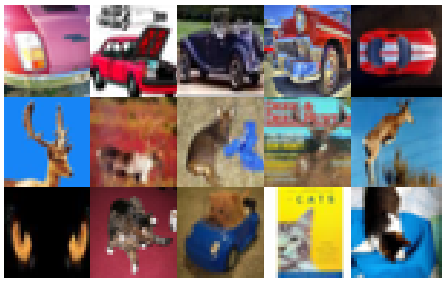
images entières. La Figure 4 illustre l'expérience en montrant les images les plus normales et les plus anormales selon la métrique issue du VQ-VAE. Le VQ-VAE semble alors capable de correctement donner un score de normalité élevé à des images relativement diverses au sein d'une même classe (diversité des couleurs, des arrière-plans, du paysage, etc.). Cela corrobore le fait que le VQ-VAE est capable d'extraire des caractéristiques pertinentes au sein d'images d'une même catégorie.

4 Conclusion

Dans cet article, nous avons illustré le potentiel des VQ-VAE pour l'AD. Nous avons mis en avant, pour la première fois, le fait que les métriques issues des VQ-VAE sont assez robustes et riches pour détecter les anomalies dans les images. En effet, après avoir développé une approche intuitive pour la construction d'une carte d'anomalies à l'aide des VQ-VAE, nous obtenons des résultats compétitifs avec d'autres approches de l'état de l'art pour l'AD au niveau des pixels (jeux de données MVTec et UCSD-Ped1) et pour l'AD au niveau des images (jeu de données CIFAR-10).



(a) Images les plus normales.



(b) Images les plus anormales.

FIGURE 4 – Sélection d'illustations pour l'expérience CIFAR-10 avec le modèle VQ-VAE SSIM+AM : échantillons des images les plus normales (a) et anormales (b) des catégories *Voiture* (haut), *Biche* (milieu) et *Chat* (bas), lorsque le modèle est entraîné sur cette même catégorie.

Les résultats montrent que les métriques inhérentes aux VQ-VAE donnent de meilleurs résultats que celles des VAE, sans introduire de complexité supplémentaire dans le modèle. Ces résultats sont donc en accord avec l'intérêt croissant de la communauté pour les VQ-VAE.

Dans le but d'une amélioration des performances du modèle à partir de nouvelles métriques sur l'espace latent des VQ-VAE, des futures recherches pourraient considérer tirer profit d'autres résultats récents, notamment l'utilisation de plusieurs registres de vecteurs [30] [10] ou de procédures d'entrainement particulières [11].

Remerciements

Ce travail a été effectué dans le cadre du projet Game of Trawls. Nous remercions le Fonds européen pour les affaires maritimes et la pêche (contrat numéro 18/2216442) et France Filière Pêche (contrat numéro 19/1000544) pour le financement. Ce travail a aussi été effectué dans le cadre du projet SEMMACAPE, financé par l'Agence de la transition écologique lors de l'appel à projet "Sustainable Energies" (2018–2019).

Références

[1] C. BAUR et al. "Autoencoders for unsupervised anomaly segmentation in brain MR images : a comparative study". In : *Medical Image Analysis* (2021), p. 101952.

[2] P. BERGMANN et al. "MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection". In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, p. 9592-9600.

[3] S. BOND-TAYLOR et al. "Deep Generative Modelling : A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models". In : *IEEE transactions on pattern analysis and machine intelligence* (2021). In press.

[4] T. DEFARD et al. "Padim : a patch distribution modeling framework for anomaly detection and localization". In : *International Conference on Pattern Recognition*. Springer. 2021, p. 475-489.

[5] D. DEHAENE et al. "Iterative energy-based projection on a normal data manifold for anomaly localization". In : *8th International Conference on Learning Representations, ICLR*. 2020.

[6] D. M. HAWKINS. *Identification of Outliers*. Monographs on applied probability and statistics. Chapman et Hall, 1980. ISBN : 9780412219009.

[7] D. P. KINGMA et M. WELLING. "An Introduction to Variational Autoencoders". In : *Foundations and Trends® in Machine Learning* 12.4 (2019), p. 307-392. ISSN : 1935-8237.

[8] D. P. KINGMA et M. WELLING. "Auto-encoding variational Bayes". In : *2nd International Conference on Learning Representations, ICLR*. 2014.

[9] A. KRIZHEVSKY, Geoffrey HINTON et al. *Learning multiple layers of features from tiny images*. Rapp. tech. 2009.

[10] Chieh-Hsin LAI, Dongmian ZOU et Gilad LERMAN. "Robust Vector Quantized-Variational Autoencoder". In : *arXiv preprint arXiv :2202.01987* (2022).

[11] A. ŁAŃCUCKI et al. "Robust Training of Vector Quantized Bottleneck Models". In : *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, p. 1-7.

[12] W. LIU et al. "Towards visually explaining variational autoencoders". In : *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, p. 8642-8651.

[13] P. LIZNERSKI et al. "Explainable deep one-class classification". In : *9th International Conference on Learning Representations, ICLR*. 2021.

[14] G. LOAIZA-GANEM et J. P. CUNNINGHAM. "The continuous Bernoulli : fixing a pervasive error in variational autoencoders". In : *Advances in Neural Information Processing Systems*. T. 32. Curran Associates, Inc., 2019.

[15] V. MAHADEVAN et al. “Anomaly detection in crowded scenes”. In : *IEEE Society Conference on Computer Vision and Pattern Recognition*. 2010, p. 1975-1981.

[16] A. MAKHZANI et B. J FREY. “Winner-Take-All Autoencoders”. In : *Advances in Neural Information Processing Systems*. T. 28. Curran Associates, Inc., 2015.

[17] S. N. MARIMONT et G. TARRONI. “Anomaly Detection Through Latent Space Restoration Using Vector Quantized Variational Autoencoders”. In : *18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, p. 1764-1767.

[18] E. T NALISNICK et al. “Detecting Out-of-Distribution Inputs to Deep Generative Models Using a Test for Typicality”. In : *4th Workshop on Bayesian deep learning (NIPS)*. 2019.

[19] A. van den OORD, O. VINYALS et K. KAVUKCUOGLU. “Neural Discrete Representation Learning”. In : *Advances in Neural Information Processing Systems*. T. 30. Curran Associates, Inc., 2017.

[20] M. A. F. PIMENTEL et al. “A review of novelty detection”. In : *Signal Processing* 99 (2014), p. 215-249.

[21] W. H. Lopez PINAYA et al. “Unsupervised Brain Anomaly Detection and Segmentation with Transformers”. In : *Medical Imaging with Deep Learning*. T. 143. Proceedings of Machine Learning Research. PMLR, 2021, p. 596-617.

[22] A. RAMESH et al. “Zero-Shot Text-to-Image Generation”. In : *Proceedings of the 38th International Conference on Machine Learning*. T. 139. Proceedings of Machine Learning Research. PMLR, 2021, p. 8821-8831.

[23] A. RAZAVI, A. van den OORD et O. VINYALS. “Generating diverse high-fidelity images with VQ-VAE-2”. In : *Advances in neural information processing systems*. 2019, p. 14866-14876.

[24] L. RUFF et al. “A unifying review of deep and shallow anomaly detection”. In : *Proceedings of the IEEE* (2021).

[25] L. RUFF et al. “Deep One-Class Classification”. In : *Proceedings of the 35th International Conference on Machine Learning*. Sous la dir. de Jennifer DY et Andreas KRAUSE. T. 80. Proceedings of Machine Learning Research. PMLR, oct. 2018, p. 4393-4402.

[26] L. RUFF et al. “Deep Semi-Supervised Anomaly Detection”. In : *8th International Conference on Learning Representations, ICLR*. 2020.

[27] S. VENKATARAMANAN et al. “Attention guided anomaly localization in images”. In : *European Conference on Computer Vision*. Springer. 2020, p. 485-503.

[28] L. WANG et al. “Image Anomaly Detection Using Normal Data Only by Latent Space Resampling”. In : *Applied Sciences* 10.23 (2020), p. 8660.

[29] Z. WANG et al. “Image quality assessment : from error visibility to structural similarity”. In : *IEEE transactions on image processing* 13.4 (2004), p. 600-612.

[30] Hanwei WU et Markus FLIERL. “Learning product codebooks using vector-quantized autoencoders for image retrieval”. In : *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2019, p. 1-5.

[31] D. ZIMMERER et al. “Unsupervised anomaly localization using variational auto-encoders”. In : *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, p. 289-297.