2T-03

# β-VAEGANを用いた異常検知のための疑似異常データの生成

秋田 行輝　田口 亮
名古屋工業大学 大学院工学研究科

## 1. まえがき

工業製品等の全数検査には AI による自動外観検査が有効であるが，精度向上には大量の学習データが必要となり，収集やアノテーションにコストを要する．この問題の解決策の1つとして，正常データや少量の異常データのみから疑似異常データを生成し学習データとして利用するというものがある．しかし，多くの場合，生成データの異常度合い等の特徴を詳細に制御することは難しい．β- 変分オートエンコーダー(β-VAE)等の Disentangle された表現の学習手法では，潜在変数の 1 つの次元が画像上の 1 つの特徴・因子に対応するように学習されるため，潜在変数の解釈が可能になるだけでなく，潜在変数を介して生成画像の様々な特徴を詳細に制御することが可能となる．しかし教師なしのみで学習された場合，特徴と潜在次元の明示的な紐づけは行われず，狙った特徴が潜在表現として表れる保証はない．さらに，学習済みのモデルに対して新たな特徴を持つ画像を用いて追加学習を行おうとすると，既存の潜在表現の構造が破壊されてしまう恐れがあるため，追加学習が困難という課題がある．そこで本研究では，潜在変数への条件制約と知識蒸留的なアプローチを導入した β-VAEGAN を用いて，VAE をベースとした Disentangle された表現学習モデルに対して，既存の潜在表現を維持したまま追加学習を行う手法を提案する．

## 2. 提案手法

提案手法のモデル構造を図 1 に示す．本モデルは β-VAE と GAN を組み合わせた β-VAEGAN[1]を基本としている．[1]の β-VAEGAN との差異として，事前学習済みの VAE 系モデルの Encoder を教師 Encoder として追加している．また，Discriminator は Conditional GAN と同様にクラスラベル y を入力に持ち，Decoder の入力の一部には Conditional Filtered GAN[2]に倣った y によるフィルタリング構造を導入している．生徒 Encoder は追加画像の特徴を表現するための次元として，教師 Encoder よりも出力の潜在次元数を増やしている．本稿ではこれを追加次元と呼ぶ．

目的関数を式(1)に示す．生徒 Encoder，教師 Encoder，Decoder，Discriminaor をそれぞれ$E_s$，$E_t$，$G$，$D$とする．式(1)の第 1 項$L_{vae}$は，β-VAE の損失関数と同様に入力画像$x$と再構成画像$G(z)$間の再構成誤差と，q(z|x)と$p(z)$の KL ダイバージェンスで構成されており，式(2)で表される．再構成誤差には$x$と$G(z)$を$D$に入力した際の中間層の出力の平均二乗誤差を用いる．以下では$E_s$，$E_t$の出力をそれぞれ$z_s$，$z_t$と表す．

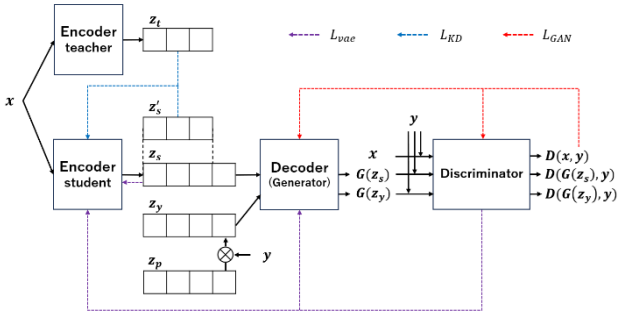



図 1 提案手法のモデル構造

$$\min_{E_s,G} \max_{D} \quad L_{vae} + L_{KD} + L_{GAN} \tag{1}$$

$$L_{vae} = -E_{q_s(z_s|x)}[\log p(x|z_s)] + \beta * KL[q_s(z_s|x) \parallel p(z_s)] \tag{2}$$

式(1)の第 2 項$L_{KD}$は，知識蒸留的なアプローチにより，$E_t$が持つ潜在表現を$E_s$へ継承することを目的とした損失であり，式(3)で表される．$E_s$，$E_t$へ同じ入力$x$を与えた際に得られる$z_s'$と$z_t$間の KL ダイバージェンスを損失とする．ただし$z_s'$は$z_s$から追加次元を除いたものである．

$$L_{KD} = KL[q_s(z_s'|x) \parallel q_t(z_t|x)] \tag{3}$$

式(1)の第 3 項$L_{GAN}$は，VAEGAN の GAN 部分の損失関数に，追加画像の特徴と追加次元の対応を促すための条件制約を加えたものとなっており，式(4)で表される．$D$ は入力$x$ が真の画像であり，かつ付与されたクラスラベル$y$ が正しい場合に 1 を，それ以外の場合に 0 を出力するように学習する．対して，$G$は再構成画像$G(z_s)$と，潜在変数 $z_y$からの生成画像$G(z_y)$が$D$を騙せるように敵対的に学習する．ただし，$z_y$は乱数$z_p$の追加次元部分$z_{p_add}$に対して式(5)に示す関数$f_y$でフィルタリングを行ったものである．クラスラベル$y$ は，追加データには 1，事前学習でも使用したデータには 0 として与え，乱数$z_p$には 0 か 1 をランダムに与える．

$$\begin{aligned} L_{GAN} = {} & E_{x,y}[\log D(x,y)] \\ & + E_{q_s(z_s|x),y}[\log(1 - D(G(z_s),y))] \\ & + \gamma\, E_{z_y,y}[\log(1 - D(G(z_y),y))] \end{aligned} \tag{4}$$

Generating Pseudo-anomaly Data using β-VAEGAN for Anomaly Detection

Yukiteru Akita, Ryo Taguchi
Nagoya Institute of Technology

$$f_y(z_{p\_add}) = \begin{cases} |z_{p\_add}| & (y=1) \\ -|z_{p\_add}| & (y=0) \end{cases} \quad (5)$$
$$z_{p\_add} \sim z_p \ , z_p \sim Unif(-1,1)$$

## 3. 実験

### 3.1 実験条件

提案手法によって追加学習を行ったモデルが，教師モデルの Disentangle された表現を継承できているか，また追加データの特徴を任意の潜在次元に対応付けできているかを確認する．

教師モデルには，Dsprites データセット[3]を用いて学習を行った β-VAE を用いる．Dsprites データセットは形状，大きさ，角度，x 座標，y 座標の 5 つの潜在因子から手続き的に生成された計 737280 枚の画像群である．追加学習時には，データセットから大きさ，角度の因子を固定した 3072 枚を選択し，これらに対して白から赤にかけての5段階の色に変更する処理を加えて作成した$3072 \times 5$ 枚の異常画像群を追加データとして用いる．これにより追加因子として色が付与される．教師 Encoder の出力$z_t$の次元数を 5，生徒 Encoder の出力$z_s$の次元数を 6 とする．

追加因子以外の 5 つの因子についての Disentanglement 精度（以下 D 精度と呼ぶ），全因子についての D 精度，追加因子と追加次元の対応度の 3 つについて，損失関数の条件制約項の係数$\gamma$を 0，1,5,10 と変化させた場合で比較する．D 精度の評価には，Kim らが提案した Disentanglement Metric[4]を用いる．この指標の算出工程は以下である．

1. 1つの因子を固定して生成した L 枚の画像を学習済みの Encoder に入力し，L 個の出力$z$ を得る．
2. $z$ の各次元の分散を計算し，分散が最小の次元を固定した因子の依存次元であると推測する．
3. 1，2 の処理を全因子について M 回繰り返し，推測結果のばらつきから全体の Disentanglement 度合いを評価する．

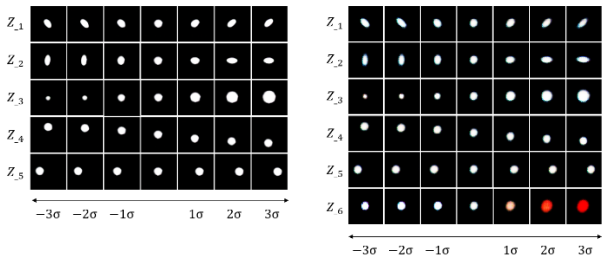
追加因子と追加次元の対応度の評価には，追加因子について上記 1，2 の処理を 100 回繰り返した際に，追加次元が依存次元であると正しく推測された割合を評価指標として用いる．

### 3.2 実験結果と考察

実験結果を表 1 に示す．追加因子以外の D 精度は，$\gamma$=0，1，5 の場合には教師モデルと同等以上となった．また，図 2(a)，(b)にそれぞれ示す教師モデルと追加学習モデル($\gamma$=1)の latent traversal を比較すると，$z_1$から$z_5$の各次元の変化の様子が一致していることが確認できる．以上より知的蒸留的アプローチによる潜在表現の引継ぎが有効に作用していると考えられる．しかし，$\gamma$=10 の場合には精度が低下してしまっている．追加データの大きさ，角度の因子を一定にしているため Discriminator がそれらも追加データの特徴であると学習し，強すぎる条件制約により Disentangle された潜在表現を破壊してそれらを追加次元にも対応づけてしまっていることが原因として考えられる．

表 1 実験結果

| | | Disentangle 精度 | | 追加因子の対応度 |
|---|---|---|---|---|
| | | 追加以外 | 全体 | |
| 教師モデル | | 0.804 | - | - |
| 追加学習モデル | $\gamma$=0 | 0.826 | 0.667 | 0.00 |
| | $\gamma$=1 | 0.808 | 0.840 | 1.00 |
| | $\gamma$=5 | 0.816 | 0.847 | 1.00 |
| | $\gamma$=10 | 0.766 | 0.735 | 1.00 |




(a) 教師モデル　　(b) 追加学習モデル

図 2　latent traversal

追加因子と追加次元の対応度は$\gamma$=1，5，10 の場合には 1.00 となっており完全な対応付けができているが，条件制約を加えない$\gamma$=0 の場合には 0.00 となり全く対応付けできていないことがわかる．これに伴い全因子についての D 精度も，追加因子以外の D 精度と比べて$\gamma$=0 では悪化し，$\gamma$=1，5 では向上している．このことから条件制約が有効に作用していると考えられる．

また，図 2(b)の追加次元$z_6$を見ると，値が正の範囲で白から赤にかけて滑らかに変化していることが確認でき，潜在変数の追加次元の値によって追加データ特有の特徴を制御して画像生成が行えていると言える．

## 4. まとめ

本稿では，VAE ベースの Disentangle された表現獲得モデルへの追加学習手法を提案した．提案手法では，元の潜在表現の構造を維持したまま新たな特徴を任意の潜在次元へ紐づけて追加学習が行えることを確認した．今後は追加データがより少ない場合やトイデータではない複雑な画像についても有効性を検証し，異常検知タスクへの実利用を目指す．

### 参考文献

[1] F. Lüer, et.al.: Adversarial Anomaly Detection using Gaussian Priors and Nonlinear Anomaly Scores, The IEEE International Conference on Data Mining Workshops,(2023).
[2] T. kaneko, et.al.: Generative Attribute Controller with Conditional Filtered Generative Adversarial Networks, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),(2017).
[3] L. Matthey, et.al. : dSprites: Disentanglement testing Sprites dataset,https://github.com/deepmind/dsprit es-dataset/, (2017).
[4] H. Kim, et.al.: Disentangling by Factorising,, The 35th International Conference on Machine Learning, PMLR 80:2649-2658, (2018).