The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

VAEGAN と Attention を活用した異常検出手法

Anomaly Detection with VAEGAN and Attention

木村 大毅 *1 Daiki Kimura　Subhajit Chaudhury *1　成田 穂 *2 Minori Narita　Asim Munawar *1　立花 隆輝 *1 Ryuki Tachibana

*1IBM Research AI　*2東京大学 The University of Tokyo

Visual anomaly detection is common in several applications including medical screening and production quality check. There are various methods that are using the reconstructed image from a generative model which is trained only normal patterns. However, the previous method used a generative adversarial network has a problem of dropping a local minimum and robustness of noise. In this paper, we propose an anomaly detection method using the reconstructed image from a conditional generative model which is called VAEGAN. We also propose introducing an attention mechanism by using the trained model. We conducted experiments on multiple datasets contained noise, we verified the proposed method over-performed the previous methods.

1. 序論

近年，異常検出を実現する手法は，数多く提案されており，不良品検出や，監視カメラからの危険人物の検出，医療の発病検知などの領域に応用されている [1–3]．その中でも，判定結果の根拠を目で確認できるため画像を用いた研究は多い [2–4]．

画像を用いた異常検出として，生成モデルで通常クラスの画像を十分学習させた後，モデルから生成された画像と入力画像との誤差画像を基に異常を検出する手法が代表的である．生成モデルとして Auto-Encoder (AE) を応用した異常検出手法 [5] や，Variational Auto-Encoder (VAE) [6] を応用した手法 [7]，Generative Adversarial Networks (GAN) [8] を活用した手法 [9] が既に提案されている．その中でも，GAN を活用した手法 [9] は，精度が良いものの，学習コストがかかる問題 [10] や，初期値によっては局所解に陥ってしまう問題がある．そこで，VAE を用いて GAN の初期値を決めることで，入力画像に近い生成画像を出力する VAEGAN [11] の活用が必要であると考えられる．

一方で，実世界での利用を想定した場合，入力画像にはノイズが多く含まれる．生成モデルでは平均的な画像を出力するため，入力画像にノイズが混入している場合，モデルはノイズの少ない画像を生成する．そのため，ノイズ部には再構成誤差が生じる．先に上げた手法では，再構成誤差の大きさにより判定を行っているため，異常に起因する誤差と，ノイズに起因する誤差を分ける必要があるが，その機構は備わっていない．すなわち，ノイズ部の誤差と本来の異常領域の誤差の区別ができなく，ノイズへの頑健性が懸念される．そこで我々は以前，異常クラスの一部を教示し，識別に必要となる領域に焦点を当てることで，ノイズに頑健な手法を提案した [12,13]．ところが，異常画像の一部を事前に知り得る必要があり，それらが未知の場合は根本的な問題となる．

そこで本稿では，まず，VAEGAN を用いた異常値判検出手法を提案する．次に VAEGAN の Discriminator から Attention を算出し，ノイズ耐性のある異常検出手法を提案する．そして，VAE，GAN を活用した従来手法と比較実験を実施し，評価する．提案手法の概略図を図 1 に示す．

連絡先: 木村 大毅, IBM Research AI, 東京都中央区日本橋箱崎町 19-21, daiki@jp.ibm.com

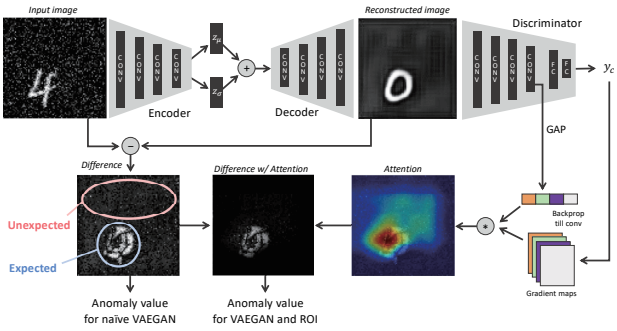


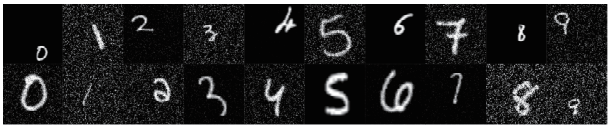
図 1: 提案手法の概要図：‘0’ が正常クラスであり，‘4’ の画像が入力された時の手法による異常度合いの違い．Naïve に VAEGAN を用いた手法では，ピンクの領域がノイズの有無に起因しているため，判別に悪影響を及ぼす可能性が高い．

2. 提案手法

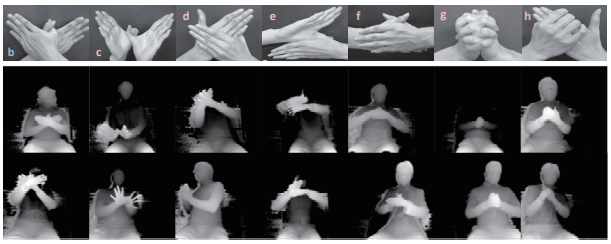
まずは，VAEGAN を応用した異常検出について述べる．VAEGAN [11] は，Encoder と，Decoder，Discriminator から構成されている．Encoder は入力された画像から潜在変数を生成し，Decoder は潜在変数から画像を生成する．Discriminator は，学習画像なのか生成画像なのかを判別する．Encoder を $\mathcal{G}_{enc}$，Decoder を $\mathcal{G}_{dec}$，Discriminator を $\mathcal{D}$ とした時，それぞれは式 (1) の min max を解くことで算出される．$\mathcal{L}_{prior}$ は $q(z|x)$ と $p(z)$ の KL ダイバージェンス，$\mathcal{L}_{feat}$ は入力画像と生成画像の誤差である．$\mathcal{L}_{GAN}$ は式 (2) で算出される．$\mathcal{D}$ は，学習画像であれば 1，生成画像であれば 0 とする．

$$\min_{\mathcal{G}_{enc},\mathcal{G}_{dec}} \max_{\mathcal{D}} \quad \mathcal{L}_{GAN} + \lambda \mathcal{L}_{prior} + \mathcal{L}_{feat} \tag{1}$$

$$\begin{aligned}\mathcal{L}_{GAN} =& \mathbb{E}_{x\sim p_{data}} \log[\mathcal{D}(\boldsymbol{x})] + \mathbb{E}_{z\sim p_z} \log[1 - \mathcal{D}(\mathcal{G}_{dec}(\boldsymbol{z}))] \\ &+ \mathbb{E}_{x\sim p_{data}} \log[1 - \mathcal{D}(\mathcal{G}_{dec}(\mathcal{G}_{enc}(\boldsymbol{x})))]\end{aligned} \tag{2}$$



(a) ノイズを混入させた MNIST



(b) 鳩のポーズのデータセット：上段は手の白黒画像で，下段は実際の深度画像例．一列目の b のみが正常クラスで，他は全て異常クラス．

図 2: データセットの画像例

次に，異常検出のステップを説明する．まず学習データの正常クラスについて前述のステップにて十分に学習した VAEGAN を用意する．そして，テスト画像を VAEGAN に入力し画像を生成する．最後に，生成画像と入力したテスト画像との誤差画像の大きさを求め，その値が一定値 *1 以上であれば “異常” と，一定値以内であれば “正常” と判定する．

ところが，この Naïve な手法では，図 1 に示したとおり，ノイズが混入されている場合にノイズ部が判定結果に悪影響を及ぼす．そこで，Discriminator から Attention を算出して，それを誤差画像に掛け合わせることで，ノイズの影響を削減する．すなわち，学習画像と生成画像の判別の際に注目した領域に対して，重みを付けて異常検出を実施する．本稿では，Attention の算出に　Class Activation Map (CAM) [14] を算出する手法の一つである Grad-CAM [15] を用いた．

3. 実験

本稿では，ノイズを付加した MNIST データセットと，医療の分野で使用されている山口キツネ·ハト模倣テスト [16] を参考にした鳩のポーズの画像データセットを用いて実験を行う．前者は定量的な評価を，後者は実世界のノイズに対する頑健性を確認するために実施する．

MNIST は，本来ノイズがなく，画像全体に対象となる数字が描かれている．ところが，本稿ではノイズ頑健性の検証が目的であるため，画像に無作為な正規ノイズを付与し，また画像サイズを 3 倍，文字の大きさや場所を画像ごとに無作為に変更したデータセットを作成した．鳩のポーズのデータセットは，18 名の参加者に鳩の手のポーズを取るように指示して，その時の様子を写した深度画像である．これらのデータセットの詳細は，既存研究 [12,13] にて，詳細を言及している．ただ後者について，従来研究の条件では，提案手法で識別精度が天井効果になる可能性があったため，問題設定をより難しく且つ現実的である「画像枚数を各ポーズ 10 枚のみ」と変更した．図 2 にデータセットの画像例を示す．

表 1 にそれぞれのデータセットに対して，提案手法と既存手法を用いて異常検出を実施したときの Receiver Operatorating Characteristic (ROC) 曲線の AUC の平均値と標準偏差を記す．Naïve に VAEGAN を用いた手法，及び VAEGAN と Attention を活用した手法のどちらにおいても従来手法以上の精度を，また Attention を活用した手法の方が更なる高精度となることを確認した．

*1　本稿では ROC で評価を実施するため，値の設定はしない

MNIST w/ noise	normal digit		average	Pigeon
	0	1		[16]
VAE [7]	.63±.2	.83±.1	.64±.1	.64±.0
GAN [9]	.45±.0	.75±.0	.54±.1	.49±.0
Naïve VAEGAN	.71±.0	.86±.0	.69±.1	.68±.0
VAEGAN+Attn	**.74±.0**	**.92±.0**	**.73±.1**	**.77±.1**

表 1: それぞれの手法における ROC 曲線の AUC 値：MNIST の average は正常クラスを 0∼9 に変更した 10 通りの平均．正常クラスを ‘0’,‘1’ にしたときの結果のみ個別に記載．

4. 結論

本稿では，VAEGAN を活用した異常検出手法と，更に Attention も活用する手法を提案した．そして，これらの手法は，従来手法よりも高精度に異常検出が可能であることを確認した．今後の展望として，VAEGAN の学習中に Attention 情報を活用する手法の提案などが挙げられる．

参考文献

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM computing surveys*, 2009.

[2] A. Taboada-Crispi, and et al., “Anomaly detection in medical image analysis,” *DIBA*, 2009.

[3] M. Sabokrou, and et al., “Real-time anomaly detection and localization in crowded scenes,” *CVPR workshop*, 2015.

[4] M. Prastawa, and et al., “A brain tumor segmentation framework based on outlier detection,” *MIA*, 2004.

[5] M. Sakurada and T. Yairi, “Anomaly detection using autoencoders with nonlinear dimensionality reduction,” *MLSDA*, 2014.

[6] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *ICLR*, 2014.

[7] J. An and S. Cho, “Variational autoencoder based anomaly detection using reconstruction probability,” *SNU*, 2015.

[8] I. Goodfellow, and et al., “Generative adversarial nets,” *NIPS*, 2014.

[9] T. Schlegl, and et al., “Unsupervised anomaly detection with generative adversarial networks to guide marker discovery,” *IPMI*, 2017.

[10] H. Zenati, and et al., “Efficient gan-based anomaly detection,” *arXiv*, 2018.

[11] A. B. L. Larsen, and et al., “Autoencoding beyond pixels using a learned similarity metric,” *ICML*, 2016.

[12] M. Narita, D. Kimura, R. Tachibana, “Spatially-weighted anomaly detection with regression model,” *SSII*, 2018.

[13] D. Kimura, M. Narita, A. Munawar, and R. Tachibana, “Spatially-weighted anomaly detection,” *MIRU*, 2018.

[14] B. Zhou, and et al., “Learning deep features for discriminative localization,” *CVPR*, 2016.

[15] R. Selvaraju, et al., “Grad-cam: Visual explanations from deep networks via gradient-based localization,” *ICCV*, 2017.

[16] H. Yamaguchi, and et al., “Yamaguchi fox-pigeon imitation test: a rapid test for dementia,” *Dementia*, 2010.