動画中のオブジェクトの運動を捉える敵対的生成ネットワーク

Motion Disentangled Bidirectional GAN

内海佑麻 [1] 阿部佑樹 [2] 妹尾卓磨 [2] 今井倫太 [1,2]
Yuma Uchiumi[1] Yuki Abe[2] Takuma Seno[2] Michita Imai[1,2]

[1] 慶應義塾大学理工学部情報工学科
[1] Department of Information and Computer Science, Keio University
[2] 慶應義塾大学理工学研究科
[2] Graduate School of Science and Technology, Keio University

Abstract: Video data contains information on the movement of objects or the sequential pattern. In general, when extracting motion information of the object from the video, it is necessary to capture not the consistent pattern for the object recognition but the sequential pattern for the motion recognition. To tackle this problem, we propose Motion Disentangled Bidirectional GAN (MDBiGAN) that decomposes the latent variable of Bidirectional GAN into two different variables. One represents the sequential patterns, and the other represents the consistent patterns. Then, the exeperiments show that MDBiGAN clearly extracts the motion feature from the video.

1 序論

動画データの特性は，動画中の各画像フレームが時間連続性をもつことであり，物体の運動やパターンの変化に関する情報が含まれている．フレーム間に内在された運動や変化を陽に表現・獲得する手法は，動画分類や動画予測にとって有用である．

動画中の時間的に連続した画像に一貫して登場するオブジェクトの運動をコンピュータで捉えるためには，系列データに対する特徴量抽出が必要である．つまり，動画からオブジェクトの運動情報を抽出する場合，オブジェクトの識別に用いるパターン情報ではなく，オブジェクトの運動に現れる系列的パターン情報を捉える必要がある．たとえば，犬が走る描写を捉えた動画と猫が走る描写を捉えた動画について，それぞれ二つの動画がもつオブジェクト情報は「犬」「猫」であり異なるが，オブジェクトの運動情報は「走る動作」であり同じである．また，特徴抽出法として教師あり学習を用いる場合，オブジェクトの運動に対するクラスラベルを予め作成する必要があり作業コストが大きい．よって，検討すべき特徴抽出法としては，動画データ全体に対する情報圧縮ではなく，系列データのもつ時間連続的な情報のみを分離的に抽出できる教師なし学習手法が望ましい．

本論文では，動画データからオブジェクトの運動情報を抽出する問題に対して，解釈性が高く制御可能な潜在変数によって特徴量を表現する GAN[1] ベースの教師なし学習手法 MDBiGAN(Motion Disentangled Bidirectional GAN) を提案する．MDBiGAN では，MoCoGAN[2] と同様に，動画データが与えられたとき，オブジェクトの情報 (Content) を表現する潜在変数に関しては,「潜在変数の値が連続フレーム内の各画像で値が不変である」という仮定をおき，オブジェクトの運動情報 (Motion) を表現する潜在変数に関しては,「潜在変数の値が連続フレーム内の各画像で値が変化する」という仮定をおく．さらに，Bidirectional GAN[3] の学習フレームワークを適用させることで，2 つの潜在変数と動画データとの双方向変換を獲得するエンドツーエンドな学習方法を構築し，特徴量抽出器とデータ生成器の同時獲得を実現する．また実験として，特徴抽出器を用いて動画のクラスタリングを行い，提案手法によって獲得された特徴量がオブジェクトの運動を適切に捉えているかどうかを検証し，獲得されたデータ生成器を用いて動画生成を行い，元動画と生成動画との比較を行う．

2 関連研究

2.1 GAN

敵対的生成ネットワーク (Generative Adversarial Networks, GAN)[1] は，データ $X \in \mathcal{X}$ の真の分布 $p_r(x)$ に一致するような生成分布 $p_g(x)$ を獲得するモデルである．GAN では，Generator と Discriminator と呼ばれる 2 つのニューラルネットワークを用意し，与えられたデータサンプルから，これらを教師なし学習させる．

具体的には，事前に定義された確率変数 $z \sim p(z)$ を用意して，$p_g(x|z)$ を表現するデータ生成器 $G : Z \to X$ と，与えられたデータ X に対して真偽の 2 値分類を行う識別器 $D : X \to \{1, 0\}$ を同時学習する．GAN の目的関数は次式のようになる．

$$\begin{aligned}&\min_G \max_D \mathcal{L}_{GAN}(D, G)\\&= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)))]\\&= \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x))] \quad (1)\end{aligned}$$

GAN によって獲得された生成分布 $p_g(x)$ を用いれば，陽に定義された確率変数 $z \sim p(z)$ を用いて，データ X' をサンプル（生成）することが可能である．

$$X' \sim p_g(x) = p_g(x|z)p(z) \approx p_r(x) \quad (2)$$
$$x' = G(z), \quad z \sim p(z) \quad (3)$$

2.2 MoCoGAN

MoCoGAN[2] は，動画データを扱う GAN の拡張モデルであり，各画像に対応する潜在変数 $\mathbf{z}$ を，オブジェクト (Content) に相当する量 $\mathbf{z_c}$ とオブジェクトの動き (Motion) に相当する量 $\mathbf{z_m}$ に分ける．たとえば，T フレームの画像を持つ動画データ $\hat{\mathbf{x}}$ を MoCoGAN の Generator G が生成する場合，各画像に対応する潜在変数 $\left[\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)}\right]$ は，$\mathbf{z_c}$ と $\mathbf{z_m}$ を用いて構成される．なお，MoCoGAN では動画中に連続して登場するオブジェクトに相当する量 $\mathbf{z_c}$ が各画像間で不変であるという仮定をおくため，T フレームすべての生成画像に対して，$\mathbf{z_c}$ の値は不変となる．

$$\hat{\mathbf{x}} = \left[\hat{\mathbf{x}}^{(1)}, \ldots, \hat{\mathbf{x}}^{(T)}\right] = \left[G(\mathbf{z}^{(1)}), \ldots, G(\mathbf{z}^{(T)})\right] \quad (4)$$

$$where, \left[\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)}\right] = \left[\left[\begin{array}{c}\mathbf{z_c}\\ \mathbf{z_m}^{(1)}\end{array}\right], \ldots, \left[\begin{array}{c}\mathbf{z_c}\\ \mathbf{z_m}^{(T)}\end{array}\right]\right] \quad (5)$$

MoCoGAN では，画像単位の discriminator D_I と動画単位の Discriminator D_V を用意し，それぞれの出力に GAN の目的関数を適用する．したがって，MoCoGAN の目的関数は，次式のようになる．

$$\begin{aligned}&\mathcal{L}_{MoCoGAN}(D, G)\\&= \mathbb{E}_{x\sim p_r(x)}[\log D_I(x)]\\&+ \mathbb{E}_{z_c\sim p(z_c), z_m\sim p(z_m)}[\log(1 - D_I(G(z_c, z_m)))]\\&+ \mathbb{E}_{x\sim p_r(x)}[\log D_V(x)]\\&+ \mathbb{E}_{z_c\sim p(z_c), z_m\sim p(z_m)}[\log(1 - D_V(G(z_c, z_m)))] \quad (6)\end{aligned}$$

2.3 BiGAN

GAN は，データ生成時に与える潜在変数 Z の解釈が難しいという問題をもつ．双方向敵対的生成ネットワーク (Bidirectional GAN, BiGAN)[3][4] やその派生モデル [5][6] では，GAN の Generator $G : Z \to X$ に与える潜在変数 Z をデータ X から獲得する Encoder $E : X \to Z$ を新たに導入することで，データ X と潜在変数 Z との入出力関係を双方向に学習させることで，特徴抽出器 E とデータ生成器 G を同時獲得する学習手法を提案している．BiGAN の目的関数は次式のようになる．

$$\begin{aligned}&\mathcal{L}_{BiGAN}(D, E, G)\\&= \mathbb{E}_{x\sim p_r(x)}[\log D(x, E(x))]\\&+ \mathbb{E}_{z\sim p_r(z)}[\log(1 - D(G(z), z))] \quad (7)\\&= \mathbb{E}_{x\sim p_r(x)}[\mathbb{E}_{z\sim p_g(z)}[\log D(x, z)]]\\&+ \mathbb{E}_{z\sim p_r(z)}[\mathbb{E}_{x\sim p_g(x)}[\log(1 - D(x, z))]] \quad (8)\end{aligned}$$

3 MDBiGAN

MDBiGAN(Motion Disentangled Bidirectional GAN) は，動画からオブジェクトの運動情報を表現する潜在変数への変換を行う Encoder E，オブジェクトを表す潜在変数とオブジェクトの運動情報を表現する潜在変数から動画の生成を行う Generator G，E が抽出した潜在変数と G が生成した動画に対して，真偽判定を行う 2 種類の Discriminator D_I, D_V からなる．学習では BiGAN と同様に，E の出力と G の出力を結合させて D_I, D_V への入力とし，GAN の目的関数を求める．概要を図 1 に示す．以降，動画を X，動画のフレームサイズを T，オブジェクトを表す潜在変数を Z_C，オブジェクトの運動を表す潜在変数を Z_M と表記する，

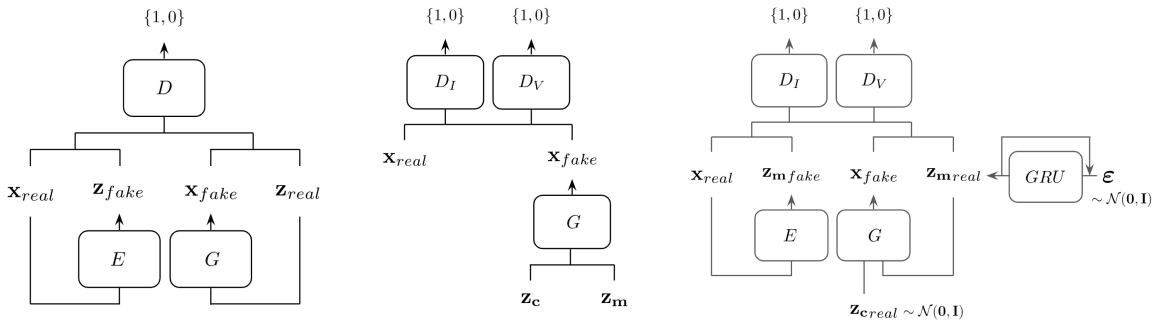
3.1 モデルの設計

3.1.1 Encoder

画像単位で潜在変数の抽出を行う Encoder E は，真の動画データ $\mathbf{x} := \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(T)}\}$ からオブジェクトの運動情報を表す潜在変数の推定量 $\hat{\mathbf{z}}_M := \{\hat{\mathbf{z}}_M^{(1)}, \ldots, \hat{\mathbf{z}}_M^{(T)}\}$ を抽出する．$E : \mathbf{x}^{(i)} \mapsto \mathbf{z}_M^{(i)}$ には，2 次元の畳み込み層を重ねたニューラルネットワークを使う．

$$\hat{\mathbf{z}}_M^{(i)} = E(\mathbf{x}^{(i)}), \quad \forall i \in [1, T] \quad (9)$$

E の入力ベクトル $\mathbf{x}$ と出力ベクトル $\hat{\mathbf{z}}_M$ は，Discriminator D_I, D_V への入力となり，GAN の目的関数からの誤差逆伝播により，E のパラメータ θ_E が更新される．



(a) BiGAN (b) MoCoGAN (c) MDBiGAN

図 1: 各モデルの概要図

3.1.2 Generator

画像単位でデータ生成を行う Generator G は，オブジェクトを表す潜在変数 $\mathbf{z}_C$ とオブジェクトの運動情報を表現する潜在変数 $\mathbf{z}_M := \{\mathbf{z}_M^{(1)}, \ldots, \mathbf{z}_M^{(T)}\}$ から動画データの推定量 $\hat{\mathbf{x}} := \{\hat{\mathbf{x}}^{(1)}, \ldots, \hat{\mathbf{x}}^{(T)}\}$ を生成する．$G : (\mathbf{z}_C, \mathbf{z}_M^{(i)}) \mapsto \mathbf{x}^{(i)}$ には，2 次元の転置畳み込み層を重ねたニューラルネットワークを使う．

$$\hat{\mathbf{x}}^{(i)} = G(\mathbf{z}_C, \mathbf{z}_M^{(i)}), \quad \forall i \in [1, T] \tag{10}$$

$\mathbf{z}_C$ には，標準正規分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ に従う確率変数を用いる．

$$\mathbf{z}_C \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{11}$$

$\mathbf{z}_M$ の各成分 $\{\mathbf{z}_M^{(i)}\}_{i=1}^T$ には，GRU の出力ベクトルを用いる．GRU の入力ベクトル $\{\boldsymbol{\varepsilon}^{(i)}\}_{i=1}^T$ について，$\boldsymbol{\varepsilon}^{(1)}$ は標準正規分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ に従う確率変数とし，$\boldsymbol{\varepsilon}^{(i)}$ $(\forall i \in [2, T])$ には，1 つ前の GRU の出力値 $\mathbf{z}_M^{(i-1)}$ を重み行列 W_M で線形変換させた値を用いる．

$$\forall i \in [1, T], \quad \mathbf{z}_M^{(i)} = GRU(\mathbf{h}^{(i-1)}, \boldsymbol{\varepsilon}^{(i)}) \tag{12}$$

$$\boldsymbol{\varepsilon}^{(i)} = \mathrm{W}_M \mathbf{z}_M^{(i-1)} \tag{13}$$

$$\mathbf{h}^{(0)} = \mathbf{0}, \quad \boldsymbol{\varepsilon}^{(1)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{14}$$

3.1.3 Discriminator

画像単位で入力データの真偽判定を行う Discriminator D_I は，入力データの真偽を画像単位で 2 値分類する．D_I の入力変数は動画 X と潜在変数 Z_M の各成分の結合ベクトル $[X^{(i)}, Z_M^{(i)}]$ $(\forall i \in [1, T])$，出力は $X^{(i)}$ $(\forall i \in [1, T])$ の真偽を表す 2 次元の One-Hot ベクトル $Y^{(i)} \in [0, 1]^2$ の推定量 $\hat{Y}^{(i)} \in [0, 1]^2$ となる．$D_I : X^{(i)}, Z_M^{(i)} \mapsto \hat{Y}^{(i)}$ には 2 次元の畳み込み層を重ねたニューラルネットワークを使う．

$$\hat{Y}^{(i)} = D_I(X^{(i)}, Z_M^{(i)}), \quad \forall i \in [1, T] \tag{15}$$

動画単位で入力データの真偽判定を行う Discriminator D_V は，入力データの真偽を動画単位で 2 値分類する．D_V の入力変数は動画 X と潜在変数 Z_M の結合ベクトル $[X, Z_M]$，出力は X の真偽を表す 2 次元の One-Hot ベクトル $Y \in [0, 1]^2$ の推定量 $\hat{Y} \in [0, 1]^2$ となる．$D_V : (X, Z_M) \mapsto \hat{Y}$ には 3 次元の畳み込み層を重ねたニューラルネットワークを使う．

$$\hat{Y} = D_V(X, Z_M) \tag{16}$$

ここで，D_I, D_V の入力変数 X, Z_M について，Z_M の値は，GRU の出力ベクトルである真値 $\mathbf{z}_M$ と，Encoder モジュールの出力ベクトルである推定値 $\hat{\mathbf{z}}_M$ のいずれかとなり，X の値は，真の動画データ $\mathbf{x}$ と，Generator の出力である推定値 $\hat{\mathbf{x}}$ のいずれかとなることに注意する．

3.2 モデルの学習

MDBiGAN に登場するニューラルネットワーク D_I, D_V, E, G に関して，それぞれの入出力変数は図 2 に示すように計算グラフによって合成微分可能な形で接続されるため，それぞれのモデルパラメータの最適化 (最尤推定) は，BiGAN と同様に目的関数 $\mathcal{L}$ に対する最大最小化問題として定式化できる．よって，MDBiGAN

の目的関数は，次式のようになる．

$$
\begin{aligned}
&\mathcal{L}(D_I, D_V, E, G) \\
&= \mathbb{E}_{x\sim p_r(x)}[\log D_I(x, E(x))] \\
&+ \mathbb{E}_{z_c\sim p(z_c), z_m\sim p(z_m)}[\log(1 - D_I(G(z_c, z_m), z_m))] \\
&+ \mathbb{E}_{x\sim p_r(x)}[\log D_V(x, E(x))] \\
&+ \mathbb{E}_{z_c\sim p(z_c), z_m\sim p(z_m)}[\log(1 - D_V(G(z_c, z_m), z_m))]
\end{aligned} \tag{17}
$$

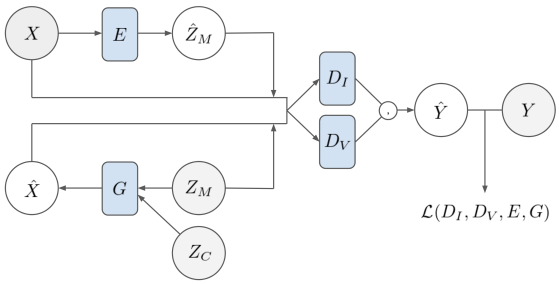


図 2: MDBiGAN のグラフィカルモデル．観測値が与えられる変数を塗りつぶしている．各変数が合成微分可能な形で目的関数 $\mathcal{L}(D_I, D_V, E, G)$ と接続されており，D_I, D_V, E, G の同時学習が可能．

4 実験

4.1 データセット

MDBiGAN が扱うデータとしては，人のジェスチャーや物体の運動など，同一オブジェクトが連続画像間で運動する動画データが適している．そこで，MNIST[7] と MovingMNIST[8] を元にして動画データセット MotionMNIST を作成した．MotionMNIST は，MNIST の手書き数字 ($0 \sim 9$) が運動する動画と，動画中の手書き数字の運動に対応する教師ラベル ($0 \sim 7$) からなる．なお，MDBiGAN は教師なし学習手法であるため，オブジェクトの運動情報に関する教師ラベルは，MDBiGAN の学習時には用いない．それぞれの動画は 20 フレームのグレースケール画像をもち，各画像は 64×64 のサイズで構成される．また，動画中の手書き数字の運動に対するラベルとして，運動パターンを 8 種類用意した．MotionMNIST として，MNIST のテストデータにある 10,000 枚の手書き数字を使い，それぞれの手書き数字に 8 種類のモーションを与えることで，80,000 個の動画データと対応する運動ラベルをもつデータセット用意した．MotionMNIST のサンプルデータを図 3 に示す．

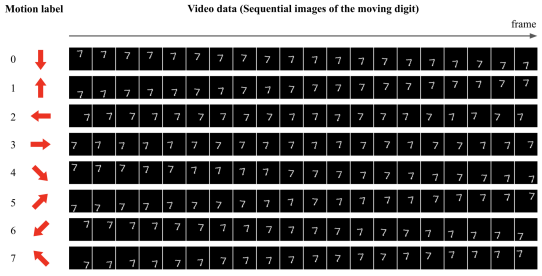


図 3: MotionMNIST データセットのサンプル．MNIST の画像 1 枚に対して 8 種類の運動ラベルを与えて動画を作成する．

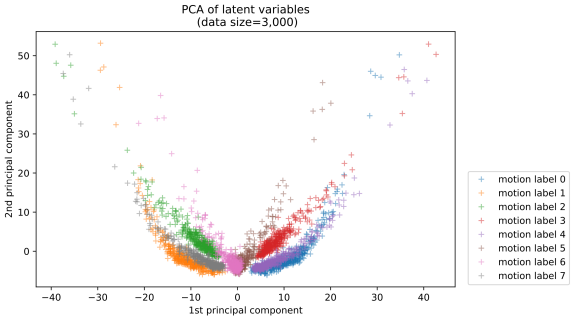
4.2 実験方法

MDBiGAN によって動画データから獲得された潜在変数 Z_M が，オブジェクトの運動特徴量を適切に捉えているかを検証するために，2 つの実験を行う．実験 1 では，Encoder によって推定された潜在変数 $\hat{Z}_M$ を用いて，動画 X のクラスタリングを行い，データセットに設定された動画 X の運動ラベルに応じて潜在変数 $\hat{Z}_M$ がクラスタを形成するかを検証する．実験 2 では，Encoder によって推定された動画 X の潜在変数 $\hat{Z}_M$ を用いて，Generator の出力である生成動画 $\hat{X}$ を獲得し，元動画 X と生成動画 $\hat{X}$ を比較する．また，生成動画 $\hat{X}$ を観察し，潜在変数 Z_C, Z_M の特性について考察する．

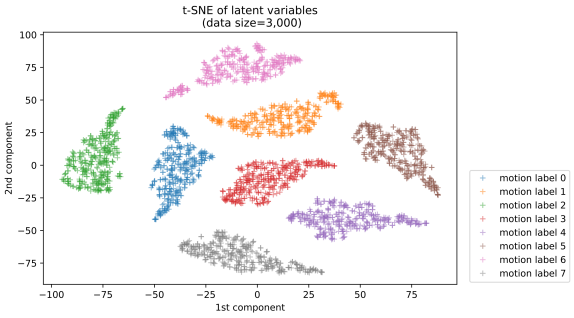
4.3 実験結果

4.3.1 実験 1: 潜在変数を用いた動画のクラスタリング

学習済みの MDBiGAN に含まれる Encoder: $\mathbb{R}^{64\times64} \ni \mathbf{x}^{(i)} \mapsto \mathbf{z}_M^{(i)} \in \mathbb{R}^{16}$ $(\forall i \in [1, 20])$ を用いて，MotionMNIST に含まれる任意の動画データ $\mathbf{x} \in \mathbb{R}^{20\times64\times64}$ からオブジェクトの運動に相当する潜在変数の値 $\mathbf{z}_M := [\mathbf{z}_M^{(1)\mathrm{T}}, \ldots, \mathbf{z}_M^{(20)\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{20\times16}$ が取得できる．$\mathbf{z}_M$ に対して，主成分分析 (Principal Component Analysis, PCA) と t 分布型確率的近傍埋め込み (t-distributed Stochastic Neighbor Embedding, t-SNE)[9] をそれぞれ行なった結果を図 4 に示す．図 4 をみると，動画データ $\mathbf{x}$ に与えられたオブジェクトの 8 種類の運動ラベルに応じて，潜在変数 $\mathbf{z}_M$ は区別可能な 8 つのクラスタを形成している．すなわち，潜在変数 $\mathbf{z}_M$ によって動画データ $\mathbf{x}$ をクラスタリングすることで，オブジェクトの動きによる動画分類が可能となる，



(a) PCA



(b) t-SNE

図 4: 潜在変数 Z_M による動画のクラスタリング

4.3.2 実験 2: 潜在変数を用いてオブジェクトの運動を制御した動画生成

MDBiGAN には，Generator の入力変数として，オブジェクトの情報を表す潜在変数 Z_C とオブジェクトの運動情報を表す潜在変数 Z_M が含まれている．よって，潜在変数 Z_C, Z_M を制御することで，Generator: $(Z_C, Z_M) \mapsto X$ によって生成される動画データを制御することが可能である．MotionMNIST に含まれる任意の動画データ $\mathbf{x} \in \mathbb{R}^{20\times64\times64}$ に対して，Encoder: $X \mapsto Z_M$ によって抽出された潜在変数 $\mathbf{z}_M \in \mathbb{R}^{20\times16}$ と正規乱数 $\mathbf{z}_C \in \mathbb{R}^{20\times64}$ を，Generator: $(Z_C, Z_M) \mapsto X$ に与えることで生成された動画を図 5 に示す．

図から元動画と生成動画でオブジェクトの動きが一致していることが観察されるため，Encoder によって抽出された Z_M は，動画データ中のオブジェクトの運動情報を適切に捉えているとわかる．また，Generator が生成した動画に含まれる手書き数字のパターンがフレーム間で大きく変わらないことは，Z_C の値はフレームに対して不変であることに対応している．生成動画の質について，手書き数字のパターンの生成には成功していない．理由としては，実験に用いた MDBiGAN の学習時に，画像単位の Discriminator に対する目的関数の値が即座に最小化されてしまい．その後の Generator と Encoder のパラメータ更新時に，画像単位の真偽判定の情報がうまく伝達されなかったことが考えられる．

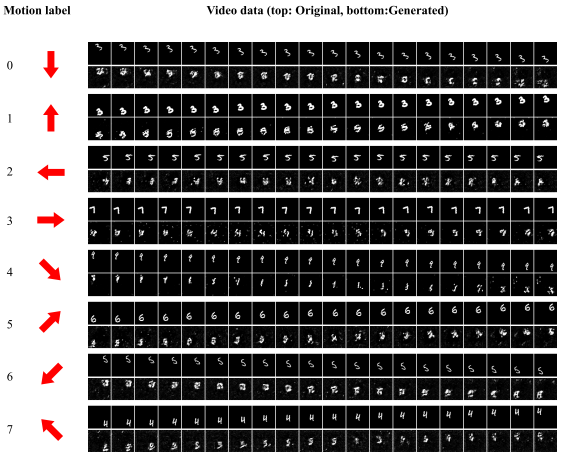


図 5: 元動画と Generator の出力した生成動画との比較．潜在変数 Z_M には元動画に対する Encoder の出力を用いている．

5 将来研究

潜在変数 Z_C の解釈性　本論文の提案モデル MDBiGAN では，動画中のオブジェクトの運動を捉えるため，動画データを Z_C, Z_M という 2 つの潜在変数で表現した．しかし，動画データから潜在変数 Z_C を抽出する Encoder がモデルに含まれないため，学習済みの MDBiGAN を使って Z_C を明示的に制御して動画生成を行うことはできない．すなわち，Z_C の確率分布として与えた多次元正規分布 $\mathcal{N}(\mathbf{0}, \mathbf{I})$ の解釈が難しい．潜在変数 Z_C の解釈性を上げる場合，MDBiGAN に対して，InfoGAN [10] と同様に近似分布 $q(z_c|x)$ を導入することでこれを実現可能である．

系列データに対する表現能力　MDBiGAN では，潜在変数 Z_M が連続画像間の系列性を表現するために，再帰的ニューラルネットワークとして GRU[11] を用いている．しかし，人のハンドジェスチャーや多関節ロボットの歩行動作などの複雑な運動情報を捉えるためには，GRU の表現能力を上げる必要がある．

6 結論

本論文では，オブジェクトの運動情報を表す潜在変数を動画データから教師なし学習により獲得すると同時に，潜在変数を制御することで動画データの生成を行うモデルとして MDBiGAN(Motion Disentangled Bidirectional GAN) を提案した．MDBiGAN では，MoCoGAN を応用して，Generator，Discriminator，Encoder の構成を行い，BiGAN の目的関数を学習に適用することで，動画データと潜在変数との双方向変換の獲得を

実現した．実験では，抽出された潜在変数がオブジェクトの運動情報を適切に捉えているかを検証するため，潜在変数を用いた動画のクラスタリングと，潜在変数を制御した動画生成を行なった．将来研究としては，潜在変数の解釈性を重視した MDBiGAN の拡張と，系列データに対する表現能力向上を目的とした時系列モジュールの検討が挙げられる．

参考文献

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

[2] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *CoRR*, Vol. abs/1707.04993, , 2017.

[3] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[4] Vincent Dumoulin, Mohamed Ishmael Diwan Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron Courville. Adversarially learned inference. 2017.

[5] David Berthelot, Tom Schumm, and Luke Metz. BEGAN: boundary equilibrium generative adversarial networks. *CoRR*, Vol. abs/1703.10717, , 2017.

[6] Mihaela Rosca, Balaji Lakshminarayanan, David Warde-Farley, and Shakir Mohamed. Variational approaches for auto-encoding generative adversarial networks. *CoRR*, Vol. abs/1706.04987, , 2017.

[7] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2278–2324, 1998.

[8] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 843–852, Lille, France, 07–09 Jul 2015. PMLR.

[9] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, Vol. 9, pp. 2579–2605, 2008.

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pp. 2172–2180. Curran Associates, Inc., 2016.

[11] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, Vol. abs/1412.3555, , 2014.