6N-05

# GANを用いたマルチモーダル情報の生成

藤元陸† 堀井隆斗† 日永田智絵† 宮澤和貴† 長井隆行†
電気通信大学†

## 1 はじめに

深層学習技術の発展に伴い，画像や音などの高次元データを違和感なく生成可能なモデルが提案されている[2]．一方で，動画 (画像と音声) や絵本 (画像と文章) のような複数のモダリティで構成されるデータの生成に関してはいまだ十分でない．マルチモーダルデータの生成は単一モダリティデータの生成と比べて困難な課題である．なぜなら，モダリティが増えることによりデータが複雑になり，また，モダリティ間の関係性を考慮して生成する必要があるからである．

本稿ではマルチモーダルデータの生成モデルを提案する. 提案モデルは敵対的生成ネットワーク (GAN)[1] を拡張して, 複数データを生成できるようにしたモデルである. そして，このモデルで学習を行い, 実際にマルチモーダルデータを生成できるか確認する. また, マルチモーダルデータはモダリティ間に関係性があるが, 生成モデルが関係性を考慮してデータを生成しているかどうかを検証する.

## 2 Multimodal Generative Adversarial Networks(MGAN)

### 2.1 MGANの定式化

マルチモーダルデータを同時に生成するためのネットワーク，Multimodal Generative Adversarial Networks(MGAN) を提案する．Fig. 1 は，生成データのモダリティが二種類の場合の MGAN のネットワーク図である．MGAN は複数のモダリティのデータを生成するために，複数の GAN で構成される．通常の GAN はノイズと生成データは 1 対 1 で対応するが，MGAN ではノイズは複数モダリティのデータと対応する．また，それぞれの GAN の Generator の入力層と Discriminator の出力層は共有する．これにより，高次の特徴が共有される．

GAN1 の生成データを $g_1(z)$，GAN2 の生成データを $g_2(z)$ とすると，Discriminator の出力は $d(g_1(z), g_2(z))$ と表される．このとき，MGAN の学習は以下のミニマックスゲームで与えられる．

$$\max_{g_1,g_2}\min_{d} V(d,g_1,g_2) = E_{x\sim px}[-\log d(x)] + E_{z\sim pz}[-\log d(1-d(g_1(z),g_2(z)))] \quad (1)$$

ただし，$x$ はデータセットの分布 $px$ からサンプルしたデータ $x$ を表す．$z$ は適当なノイズ $z$ の分布 $pz$ からサンプルしたノイズ $z$ を表す．

### 2.2 モダリティ間のバランス調整

MGAN では複数のモダリティを生成するが，データによって各モダリティの生成しやすさが異なる．その

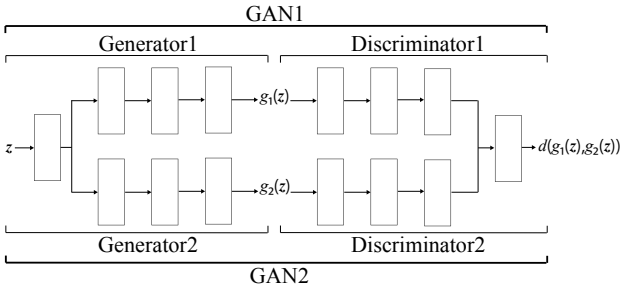


Fig. 1: Multimodal Generative Adversarial Networks

ため，あるモダリティは生成できるが，他のモダリティは生成できないということが発生する．そこで，モダリティごとの生成優先度の調整を Feature matching[3] を用いて行う．Feature matching を利用することで，ネットワーク構造を変化させずにモダリティ間の損失のバランスを調整できる．

Feature matching は GAN の生成を安定化させるテクニックの一つであり，Discriminator に真のデータ $x$ を入力したときの中間層の出力 $f(x)$ と，$g(z)$ を入力したときの中間層の出力 $f(g(z))$ のそれぞれの期待値が一致するように Generator を訓練する．通常の GAN では以下の項を Generator の損失に加える．

$$||E_{x\sim px}f(x) - E_{z\sim pz}f(g(x))||_2^2 \quad (2)$$

MGAN では各モダリティに対し Feature matching を適用する．Discriminator1 の中間層の出力を $f_1(\cdot)$,Discriminator2 の中間層の出力を $f_2(\cdot)$ とすると以下の項を Generator の損失に加える．

$$a||E_{x\sim px}f_1(x) - E_{z\sim pz}f_1(g(x))||_2^2 + b||E_{x\sim px}f_2(x) - E_{z\sim pz}f_2(g(x))||_2^2 \quad (3)$$

$a$,$b$ はパラメータであり，式 (1) に対する Feature matching の重みと各モダリティの重みを調整する．$a$ の重みを大きくするとモダリティ 1 に関する損失が大きくなり，モダリティ 1 の方が優先的にパラメータ更新される．逆に，$b$ の重みを大きくするとモダリティ 2 のパラメータ更新が優先される．$a$ と b の重みはデータやネットワーク構造によって最適な値が変わるため，設計者が生成結果を考慮しつつ調整する必要がある．

## 3 実験

MGAN がモダリティ間の対応関係を考慮してマルチモーダルデータを生成できるかの検証を行う．

### 3.1 データセット

モダリティ間の対応関係を検証するために Fig. 2(a) のロボットの腕をシミュレーション上で動かし，視覚情報と関節角情報を取得した．視覚情報は頭部のカメラから画像を取得し，画像サイズを 80 × 80pixel に変

「Generation of multimodal information using GAN」
†The University of Electro-Communications


Fig. 3: 生成画像と関節角の再構成画像.

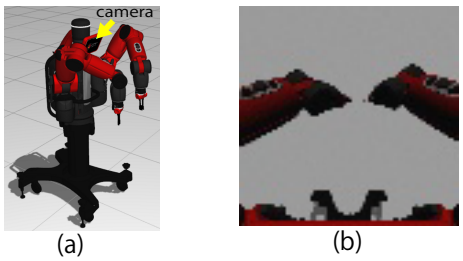


Fig. 2: 実験で用いたロボットと視覚画像

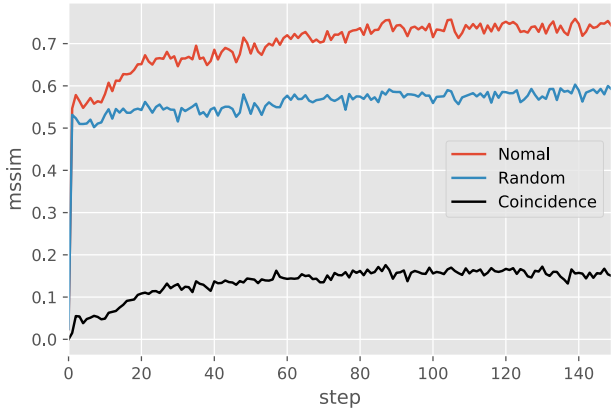


Fig. 4: SSIM result

換した．Fig. 2(b) は変換後の画像の一つである．関節角情報は片腕につき 5 関節を動かし，両腕で計 10 関節のデータを視覚情報とほぼ同時に取得した．

### 3.2 学習結果

MGAN の学習結果を Fig. 3 に示す．上段の画像は MGAN が生成した視覚情報, 下段は MGAN が生成した関節角情報を用いてシミュレーション上のロボットを動かして再現した画像である．視覚情報は綺麗な画像を生成できた．また，再現画像と生成した視覚情報を比較すると全体的な見た目や腕の位置が近いことがわかる．これは，視覚情報と関節角情報の対応関係を正しく学習できたことを示している．

### 3.3 SSIM によるモデルの評価

MGAN がモダリティ間の関係性を考慮してデータを生成しているかを SSIM[4] を用いて検証する．SSIM は画像間の類似度の手法である．生成画像と関節角情報による再現画像を学習中の各 step で 100 組生成し，step ごとに算出した SSIM の平均値 mssim が Fig. 4 の赤線（Nomal）である．赤線は生成画像と再現画像がどれだけ似ているかを表している．しかし，この値はあくまで画像の一致度であり，モダリティ間の一致度を表してるわけではない．モダリティ間の関係性を考慮して生成しているかを評価するためには，背景やロボットの動かない部位の影響を除く必要がある．そこで，各ステップで生成画像と再現画像のランダムな組み合わせを 1000 組作り，SSIM を算出した (Fig. 4 の青線)．この値はモダリティ間の関係性の影響を除いた評価値だと考えられる．この二つの評価値の差によってモダリティ間の関係性を評価した (Fig. 4 の黒線)．80step 程で 0.17 程度に収束しており，MGAN がモダリティ間の関係性を考慮して生成していることを示している．

## 4 結論

本稿では，マルチモーダルデータを生成するネットワーク MGAN を提案した．実際にシミュレータ上でロボットの腕を動かしたときの画像と関節角の情報を用いて学習し，マルチモーダルデータの生成を行った．生成した関節角情報をもとにシミュレータで再現した画像と生成画像を比較すると類似していた．また，SSIM をもちいてモダリティ間の共起性を算出したところ，学習と共に増加していた．このことから，MGAN はモダリティ間の共起性を考慮してマルチモーダルデータの生成ができるネットワークであるといえる．今後の課題として，他の様々なモダリティへの適用，潜在空間との関連性の解析，時系列データへの対応などが挙げられる．

## 謝辞

本研究の一部は、JST CREST（JPMJCR15E3）の支援を受けて実施したものである。

## 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 *Advances in neural information processing systems*, pp. 2672–2680, 2014.

2) Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

3) Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

4) Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, Vol. 13, No. 4, pp. 600–612, 2004.