1Q-01

敵対的生成ネットワークを用いた動画予測

畑　諒翼†　　篠澤　佳久††

慶應義塾大学大学院 理工学研究科†　　慶應義塾大学 理工学部††

1　はじめに

近年，多種多様な分野において，深層学習を用いた予測技術と生成技術に関する研究が盛んに行われている．具体的には医療，交通，気象，スポーツ，金融等多岐の分野に渡っている．特に交通分野における自動車の自動運転技術を安全かつ正確に実現するためには，ドライブレコーダー等の動画における高精度な予測技術だけでなく生成技術が必要不可欠である．しかしながら，動画を対象とした予測技術に関して未だ期待されている段階に至っていない[1].

そこで本研究においては，連続したフレームから構成されている動画において，その続きの動画を同じく連続したフレームで生成する敵対的生成ネットワーク（GAN）を提案する．特に提案する GAN においては，Convolutional LSTM Network[2]（以下 ConvLSTM と略す）や Attention 機構の導入により，その有用性を示す．

2　提案

本研究における動画予測は，敵対的生成ネットワークに 1 秒間の動画をフレーム単位で入力し，その続きの 0.3 秒間の動画を同じくフレーム単位で生成することを試みる．自動車のフロントの中心部分に取り付けられたカメラで撮影された A2D2 データセット（30fps）を対象とする[3].

下記の手順にて，動画予測を行う（図 1）．

（1）Generator に 1 秒間の動画を入力し，その続きの 0.3 秒間の動画を生成する（図 1 左図）．

（2）（1）で生成された動画と正解の動画を Discriminator へ入力し，正解であるかどうかを見分ける（図 1 右図）．

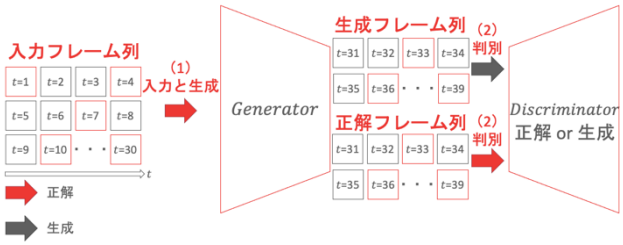

図 1 提案（左：入力と生成，右：判別）

Video Prediction using Generative Adversarial Networks
Ryosuke Hata†, Yoshihisa Shinozawa††
†Graduate School of Science and Technology, Keio University
††Faculty of Science and Technology, Keio University

図 1 において，30 枚の入力フレームの集合と 9 枚の正解フレーム，生成フレームの集合を対象に，フレームを 2 枚だけスキップした処理（以下フレームスキップ処理と表す）を行う．ただし，最初（t=1）と最後（t=30）のフレームは固定とすることから，最後と最後から 2 番目（t=28）のフレーム間は 1 枚だけのフレームスキップ処理となる．従って，モデルに入力する入力フレーム列の長さは 11，正解フレーム列と生成フレーム列の長さはともに 3 となる．

3　提案手法

先行研究[1]においては，Generator に Convolutional GRU Network[4]と ResNetBlock[5], Discriminator に ResNetBlock を用いた動画生成を行っている．そこで本研究においては，下記の 2 種類の手法を用いて動画予測を行う．

手法①：Generator に ConvLSTM と ResNetBlock, Discriminator に Self-Attention GAN[6]（以下 SAGAN と略す）を用いた動画予測

手法②：手法①の Generator に Attention 機構を導入した動画予測

3.1　手法①

ConvLSTM はフレーム間における時系列的な情報を，ResNetBlock はフレームごとの特徴を抽出する手法である．SAGAN では Self-Attention 機構によって，画像中の大域的な特徴を考慮し，判別の精度を従来よりも向上させた手法である．手法①の手順を以下に示す．

（1）動画データセットから連続した 30 枚の入力フレームの集合とその続きの連続した 9 枚の正解フレームの集合を取り出し，フレームスキップ処理を行う．

（2）Generator は（1）で作成した 1 秒間の動画である 11 枚の入力フレームを ConvLSTM に通し，最後の 3 つの ConvLSTMcell から得られる特徴をそれぞれ ResNetBlock に入力することで 3 枚のフレームを生成する．Discriminator は 3 枚の正解フレームと 3 枚の生成されたフレームを SAGAN に入力して判別する．これらを競合させるように学習を進行させる．

（3）学習後の Generator に未知の 1 秒間の動画を入力し，その続きの 0.3 秒間の動画を生成する．

3.2　手法②

手法①では，ConvLSTM によってのみ時系列的な情報を抽出していたが，より正確に抽出することを目的として，Generator に Attention 機構を追加する．具体的には，Query を最後の 3 つの ConvLSTMcell から得られる特徴，Key と Value は 30 枚の入力フレームの集合として算出する．手法②の手順を以下に示す．

（1）動画データセットから連続した 30 枚の入力フレームの集合とその続きの連続した 9 枚の正解フレームの集合を取り出し，フレームスキップ処理を行う．

（2）Generator は（1）で作成した 1 秒間の動画である 11 枚の入力フレームを ConvLSTM に通し，最後の 3 つの ConvLSTMcell から得られる特徴と Attention 機構によって算出される特徴を結合したものをそれぞれ ResNetBlock に入力することで 3 枚のフレームを生成する．Discriminator は 3 枚の正解フレームと 3 枚の生成されたフレームを SAGAN に入力して判別する．これらを競合させるように学習を進行させる．

（3）学習後の Generator に未知の 1 秒間の動画を入力し，その続きの 0.3 秒間の動画を生成する．

4　評価実験

A2D2 データセットを分割して得られた 960 本のテストデータを用いて評価実験を行った．手法①と手法②について，画像の復元度合いを図る指標である平均二乗誤差（以下 MSE と表す）を用いて比較した．

結果を表 1 に示す．表 1 には全てのテストデータから得られる正解フレーム列と生成フレーム列との MSE を算出した．そして全フレームでこれを求め，その平均値を示す．

表 1 評価実験の結果

	生成フレーム 1 枚あたりの MSE
手法①	1903
手法②	344.7

表 1 より，手法②の方が手法①よりも精度の向上が図れたことが分かる．手法②における Attention 機構は手法①と比較して，滑らかで緻密な時系列的な情報を抽出することができたと考える．しかし手法②の場合，計算量が非常に大きくなってしまう点について検討する必要がある．

次に手法②における正解フレーム列と生成フレーム列を図 2 に示す．

図 2 正解フレーム列と生成フレーム列
（上：正解フレーム列，下：生成フレーム列）

図 2 より，左側のフレームから右側のフレームへかけての時系列的な動きが見られる．しかし，車や背景におけるぼやけが多くなっていることに関して改善の余地があると考える．

5　まとめ

本研究においては，ドライブレコーダーの動画を対象として，より精度の高い動画予測を行う手法の考案を試みた．今後は，Query と Key，Value を変更した Attention 機構の提案や物体検出手法と組み合わせることで上述した問題点の改良を行っていく予定である．

参考文献

[1] Aiden Clark, Jeff Donahue and Karen Simonyan.: Adversarial Video Generation on Complex Datasets, In Int. Conf. Learning Representaions（ICLR），（2020）．

[2] Xingjian Shi, Zhourong Chen, Hao Wang and Dit-Yan Yeung.: Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, Proceedings of the 28th International Conference on Neural Information Processing Systems, pp.802-810,（2015）．

[3] Jakob Geyer *et al.*: A2D2: Audi Autonomous Driving Dataset, arXiv:2004.06320,（2020）．

[4] Nicolas Ballas, Li Yao, Chris Pal and Aaron Courville., Delving Deeper into Convolutional Networks for Learning Video Representations, In Int. Conf. Learning Representaions（ICLR），（2016）．

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.: Deep Residual Learning for Image Recognition, Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.770-778,（2016）．

[6] Han Zhang, Ian Goodfellow, Dimitris Metaxas, Augustus Odena.: Self-Attention Generative Adversarial Networks, Proceegings of the 36th International Conference on Machine Learning (PMLR), pp.7354-7363,（2019）．