敵対的生成ネットワークによるリアルタイム行動認識に関する研究

大楠 幸生　　指導教員：小林 邦和

1　はじめに

近年, 受付ロボットや接客ロボットが導入されてきており, ロボットと人間がコミュニケーションをとる機会が増えてきている. 受付ロボットや接客ロボットは人件費削減や 24 時間可動が点でメリットがあり今後も増えていくと予想されている. これらのロボットはより正確でスムーズに人間とコミュニケーションをとるために音声認識などの工夫が施されている. 音声を認識することで人間の求めている行動が把握しやすくなるため, それを元に行動選択ができる. さらにスムーズにコミュニケーションをとるために行動認識が有効だと考える. 音声認識や行動認識の代表的な手法として深層学習モデルを用いたものがある. 計算機の性能向上とともに深層学習の技術が発展してきており, 現在, 深層学習の技術は様々な分野で応用されている. しかし, 深層学習では学習や予測時に計算量が大きくなってしまうという問題がある. そのため, リアルタイムで結果を得たい場合には用いることが困難である. 近年では,Graphics Processing Unit(GPU) を利用することにより高い計算コストの計算も処理が容易になってきている. しかし CPU では高い計算コストがかかるものを計算することは厳しく, メモリ搭載量も少ない. 受付ロボットや接客ロボットは費用がかかるというデメリットがあるため出来る限り安価なものが求められている. それらの非力な CPU が搭載されているロボット上でリアルタイムに実行するためには, 計算コストとメモリ使用量の両方の削減が必要になってくる.

そこで本研究の目的は, 深層学習での計算コストとメモリ使用量の両方の削減を行うことである. 深層学習の畳み込みニューラルネットワークでは, 畳み込み演算に計算時間がかかるため, 畳み込み演算における計算コストとメモリ使用量の削減を行う. 本研究では, 行動を予測するために敵対的生成ネットワーク (GAN)[1] の一種である VideoGAN を使用する. VideoGAN は,CNN をベースにしている動画像生成のモデルである. VideoGAN[3] で動画像を事前に学習させ, その重みを初期値に与えた VideoGAN の識別部分のネットワークを利用して行動の動画像を学習させることで行動を予測する. しかし, これらの処理は 3 次元の畳み込み処理のため計算コスト, メモリ使用量共に大きくなってしまう. そこで, 本研究では VideoGAN に XNOR-Net[2] を導入することで計算コストとメモリ使用量を削減する. XNOR-Net を用いることで, 畳み込み演算を XNOR 演算と bitcount 演算で行うことができるため, 計算コストを削減することができる.

2　提案手法

本研究では VideoGAN に XNOR-Net を組み込み畳み込み層の計算コスト削減を図る. 図 1 に提案手法の概要を示す. 図 1 の右に示してある Discriminator+XNOR-Net が本研究で提案している部分になっている.

2.1　提案する行動予測法

行動予測は連続的な動作を考慮していく必要があるため, 画像ではなく動画像で学習し予測していく. そのため, 動画像生成モデルである VideoGAN を用いる. VideoGAN は Generator と Discriminator から構成されているモデルである. Generator はノイズから動画像を生成し,Discriminator は本物の動画像と生成された動画像を識別していく. Generator と Discriminator を交互に学習していくことでノイズからより本物に近い動画像を生成する Generator と, より正確に本物と偽物を識別する Discriminator が完成する.

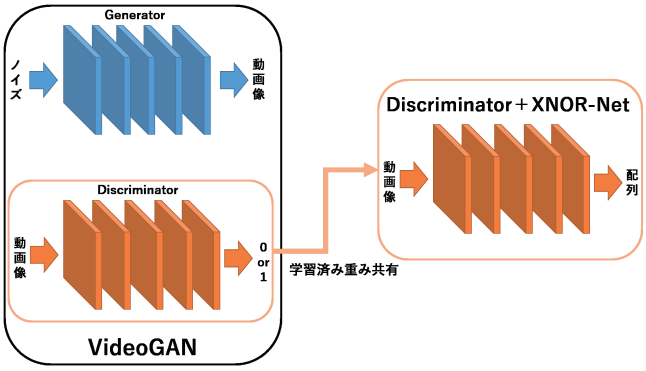

図 1　提案手法の概要

この VideoGAN で様々な動画像を学習し学習した重みを初期値として Discriminator のみを用いて行動予測の学習を行う. Discriminator の出力層を 2 値からラベルの配列にすることで行動予測に合わせた学習を行う. また,XNOR-Net を取り込むことで, 畳み込み層の計算コストを削減する. 図 2 に行動予測に用いる部分のモデル概要を示す. 畳み込み層に XNOR-Net を組み込むことにより, 畳み込み演算が bool 値を用いた演算になるので計算コストを抑え, リアルタイムでの行動予測を目指している.

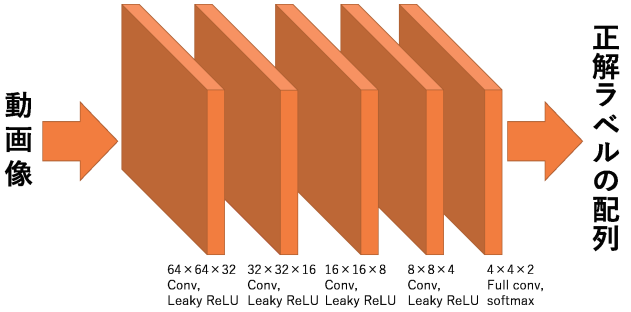

図 2　行動予測のモデル構造 [4]

2.2　XNOR-Net を用いた畳込み

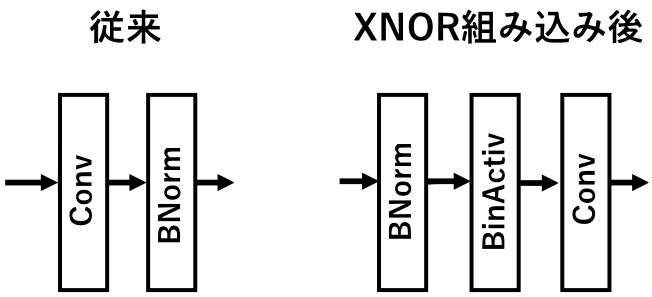

図 3　畳み込み層比較

図 3 に従来の VideoGAN との変更箇所を示す．XNOR-Net を導入した提案手法では先にバッチノーマライゼーションを行い 2 値化による情報損失を抑える．また BinActiv で入力と重みの 2 値化とスケーリング係数を求めている．そして,2 値化後畳み込み演算を行い活性化関数に入力するという流れである．

具体的な 2 値化は図 4 に示したように求める．入力を I, 重みを W,X は 1 回の畳込み演算の入力 (I の一部分) である．入力値と重みを 2 値化し, その値で畳み込み演算を行っていく．α , β はスケーリング係数である．スケーリング係数は図 4 中の式でそれぞれ求められ,2 値化した値にスケーリング係数を乗算することで元の値と近似する役割を担っている．

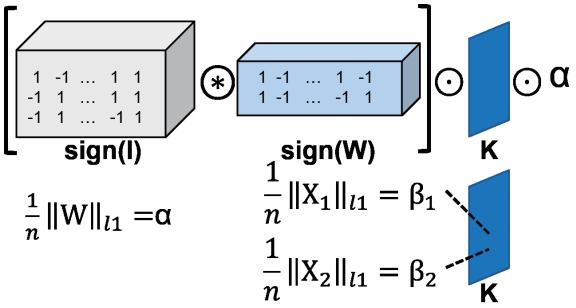

図 4　XNOR-Net 処理

XNOR 演算と bitcount 演算で行う畳み込み層の計算の例を図 5 に示す．XNOR 演算は要素が 0 と 0 または 1 と 1 の時に 1 を返し残りは 0 を返す演算である．また bitcount 演算は要素の中で 1 の数をカウントしていく計算である．

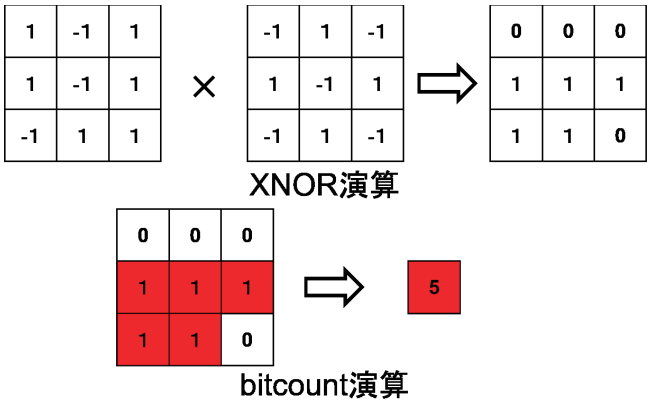

図 5　畳み込み演算 (XNOR-Net)

3　計算機シミュレーション

提案手法と従来手法の計算コスト, メモリ使用量の比較を行う．同一の環境と条件で畳み込み層に XNOR-Net を導入した提案手法と導入していない従来手法を学習させる．シミュレーション手順は, 以下に示す通りである．

- 事前学習した VideoGAN から Discriminator の重みを共有
- 重みを共有した Discriminator で動画像を学習
- 従来手法と提案手法の計算時間, メモリ使用量を比較

今回使用したパラメータを表 1 に示す．学習率や入力サイズなどはすべて VideoGAN の先行研究で用いていたパラメータを使用している．

3.1　結果

比較結果を表 2 に示す．

表 1　設定パラメータ

パラメータ	設定値
入力画像サイズ	64 × 64
入力フレーム数	32
学習率	0.1
バッチサイズ	64
ラベル数	11

表 2　シミュレーション結果

使用モデル	計算時間 [s]	メモリ量 [Mb]
従来手法	132.89	803.29
提案手法	106.46	268.75

3.2　考察

従来手法の Discriminator と比較すると, 計算コストとメモリ使用量共に削減することができている．これらは,XNOR-Net を導入したことにより従来手法では float 型だった変数を bool 型として扱ったことからメモリ使用量が削減できている．計算コストに関しては, 畳み込みの計算を提案手法では XNOR 演算と bitcount 演算で行っているため, 既存の畳み込み演算より削減できていると考えられる．

4　まとめと今後の展望

本研究では,VideoGAN を用いた行動予測の手法を考案した．XNOR-Net を用いることで計算コスト, メモリ使用量共に削減することができた．

今後の展望としては,VideoGAN は GAN から派生した標準的な動画生成モデルなので,XNOR-Net が VideoGAN の派生モデルの DVD-GAN[5] などにも効果的な可能性がある．

参考文献

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio: ”Generative Adversarial Networks”, . arXiv:1406.2661.(2014)

[2] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi: ”XNOR-Net:ImageNet Classification Using Binary Convolutional Neural Networks”, . arXiv preprint arXiv:1603.05279.(2016)

[3] Carl Vondrick, Hamed Pirsiavash, Antonio Torralba: ”Generating Videos with Scene Dynamics”, . arXiv:1609.02612.(2016)

[4] A.L.Maas,A.Y.Hannun,andY.A.Ng: ”Rectifier nonlinearities improve neural network acoustic models”, In Proc. icml (Vol. 30, No. 1, p. 3)(2013, June).

[5] Aidan Clark,Jeff Donahue,Karen Simonyan: ”Adversarial Video Generation on Complex Datasets” arXiv:1907.06571.(2019)