VAE に基づく分子システムシミュレーションの向上

山﨑瑛梨佳　(指導教員：Aubert-Kato Nathanael)

1　はじめに

近年, 深層学習の分野で様々な生成モデルが考案されコンピュータがデータを新たに生成することが可能になった. AutoEncoder[1] や Variational Autoencoder[2], Generative Adversarial Networks[3] などの生成ネットワーク発表されてきた. これらのモデルは画像データを対象に用いた研究が多く報告されている [4][5]. しかし, これらのモデルは汎用性が高いため画像以外のデータについて用いることによる結果も期待される。

そこで本研究では, 数値データにより表現される分子システムシミュレーションに生成モデルを適用し, 分子システムシミュレーションでより現実的な結果が得られるようなモデルを構築する過程で必要となる Variational Autoencoder を作成している.

2　関連研究

2.1　VAE(Variational Autoencoder)

まず, AutoEncoder のモデルを次で与える (図 1 参照).

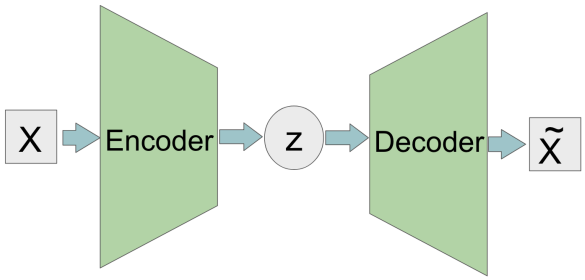

図 1: AutoEncoder のモデル概略

ここで，X は入力 $\tilde{X}$ は出力, z は潜在変数.

Encoder が入力されたベクトルの次元圧縮を行い, 潜在変数を介して Decoder が元の入力を再現しようとする. 入力と出力が近くなるようにネットワークを学習するため, 潜在変数は入力データの特徴を保持する.

この AutoEncoder の潜在変数に確率分布を導入したモデルが VAE である (図 2 参照).

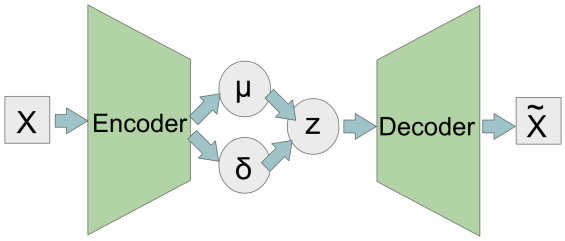

図 2: VAE のモデル概略

入力 X に近い出力 $\tilde{X}$ を得るためこれらの復元誤差をもとに Encoder と Decoder の重みを調整し潜在変数を求めていく. VAE では, Encoder が多変量ガウス分布の平均ベクトル μ と分散ベクトル σ を求め, これらを元に潜在変数 z を得る. VAE はガウス分布に従う乱数を学習時に用いるため, 潜在変数 z がガウス分布に従う. よって, 入力の類似度を潜在変数に反映しやすいという特徴がある.

2.2　分子システムシミュレーション

DNA 分子のように小さなシステムをプログラムし DNA 濃度のパターンを得るような PEN DNA toolbox によるシミュレーション [6] が行われている. 図 3 は, 実験結果の例で, sigmoid 関数や sin 関数のような特徴が見られる.

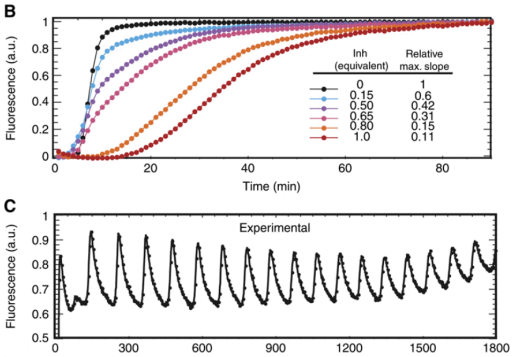

図 3: 実験結果の例

3　実験

3.1　実験概要

数値データを対象にした VAE を Keras を用いて実装し, 入力データに対する出力データの再現度や潜在変数の分布, 入力データに含まれていないようなデータを生成する際の潜在変数と対応の観測を行っている. なお, 分子システムシミュレーションにおいて sin 関数や sigmoid 関数と類似した, ないしそれらを組み合わせたような結果が得られることが多いことから, 入力データとして振動数や初期値, 正負の異なる sin 関数や sigmoid 関数を用いている.

3.2　実験結果

まず，VAE の入力データに対する出力データの再現度を観察した. 図 4 は, sin 関数を入力データとし出力されたデータと元データを重ね合わせている (左), sigmoid 関数についても同様 (右).

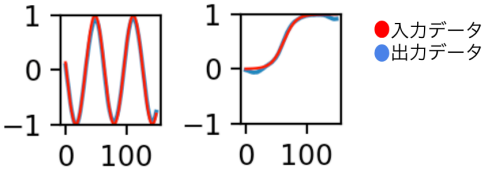

図 4: 作成した VAE による再現

sin カーブの振幅や振動数, 初期値など入力データの特徴を再現できている. sigmoid 関数についてはやや sin カーブの干渉を受けているものの概ね再現できていることがわかる.

次に, 図 5 は潜在変数の分布の例である. 以下は潜在変数を 10 次元としたときに, 2 次元として切り出してマッピングを行なった結果である.

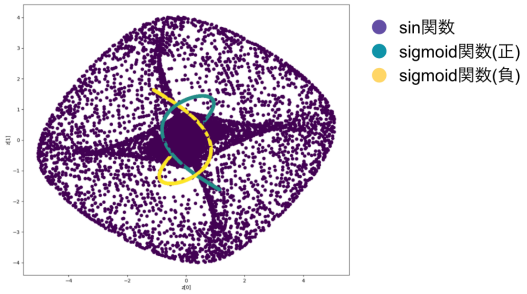

図 5: 潜在変数の散布図

入力データの特徴に応じて潜在変数が散布・集合していることがわかる. また, sigmoid 関数は連続的で単調な変化をするのに対し, sin 関数が周期を持つという特徴が散布図にも表れている.

さらに，入力データに含まれていないようなデータを生成する際の潜在変数と対応の観測を行なっている. 図 6 は，任意の入力データに対応する潜在変数を抽出し, そこからもう一つの任意の入力データに対応する潜在変数まで連続的に変化させた時の出力データを追ったもののである. (中略含む)　なお, 赤線が元の入力データ青線が潜在変数から得られた出力データである.

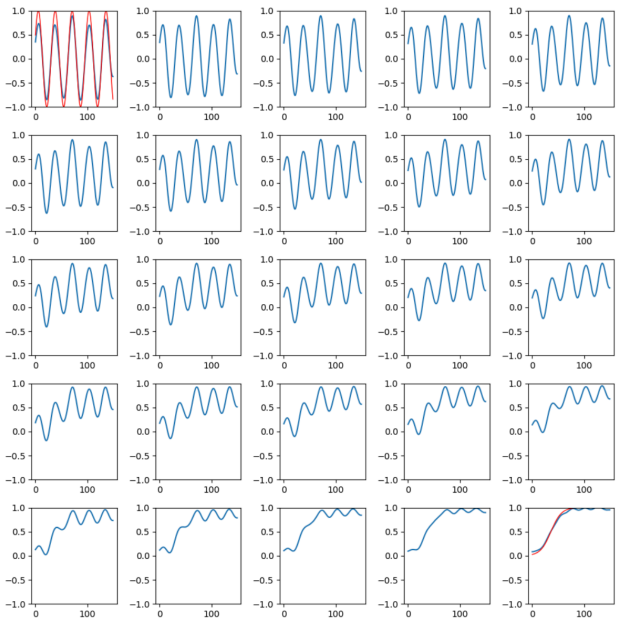

図 6: sin 曲線から sigmoid 曲線へ生成の変化

始点となる sin 関数の特徴に徐々に終点となる sigmoid 関数の特徴が加わっていく. 振幅は始点付近では終点に比べてかなり大きく, 段階を追って小さくなっていき, 端点も同様に段階的に変化していくが, 振動数については終点付近まで特徴を保持していることが分かる. したがって, 潜在変数によって特徴ごとの抽出が実現していると考えることができる.

4　まとめと今後の課題

生成モデル VAE を作成し, 多角的に観測を行いこのモデルの妥当性を調査した. 入力によって出力の再現度にブレが見られたため改善の余地はまだ見られるが, この VAE を GAN のような他の生成モデルに組み込んだような新たなモデルを導入することで性能向上を図りたい. その新たなモデルと比較することで本研究に置ける生成モデルの問題点が明確になる可能性に期待したい.

また, 本研究では実際の実験データに近い sin 関数と sigmoid 関数を対象に実験を行なったが分子システムシミュレーションに用いられているような実際のデータを対象にした実験を行う必要がある.

参考文献

[1] R. R. Salakhutdinov G. E. Hinton. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[2] CARL DOERSCH. Tutorial on variational autoencoders. 2016.

[3] Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Yoshua Bengio Ian J. Goodfellow, Jean Pouget-Abadie. Generative adversarial networks. 2014.

[4] Soumith Chintala Alec Radford, Luke Metz. Unsupervised representation learning with deep convolutional generative adversarial networks. *ICLR*, 2016.

[5] Haoran Xie Raymond Y.K. Lau Zhen Wang Stephen Paul Smolley Xudong Mao, Qing Li. Least squares generative adversarial networks. *The IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017.

[6] Kevin Montagne, Raphael Plasson, Yasuyuki Sakai, Teruo Fujii, and Yannick Rondelez. Programming an in vitro dna oscillator using a molecular networking strategy. *Molecular systems biology*, 7(1):466, 2011.