3L1-OS-12a-3

DCGANの作る仮想世界と現実空間のMap

Virtual Space and Cognitive Map organized from DCGAN

小島大樹 *1 土井樹 *1 池上高志 *1
Hiroki Kojima Itsuki Doi Takashi Ikegami

*1東京大学総合文化研究科
Graduate School of Arts and Sciences, University of Tokyo

DCGAN takes image dataset as input and gradually learns to generate realistic images. The generator maps the continuous vectors to corresponding images, so this system does not simply remember all the image in the dataset, but reconstruct the images in the continuous space. Through this process, structures in the image space emerge and generated images include some novel images. We investigated this structure and the novel images in detail, using DCGAN and VAEGAN with the dataset from first-person perspective images and sound spectrogram images.

1. はじめに

深層学習はもともと、画像の分類において高い成績を発揮していたが、最近では画像の生成についても良い成果をあげている。例えば、2015 年に発表された DCGAN[Radford 2015] は、画像のデータセットを学習することで、そこに含まれていそうな画像を生成できるようになる。ここで出来上がる生成器は、連続値の z ベクトルを入力として受けて画像を出力するため、ただデータセットに含まれる画像を覚えるのではなく、連続的な空間の中に画像を再構成することになる。それ故に、学習が進むなかで z ベクトルの空間上にどう画像が連続的に配置されるかという構造が自然に出来上がり、また再構成するため、元々のデータセットにないものも出力に含まれる。

一方、人の脳も同様のことをしているといえる。すなわち、日々受け取る感覚入力をすべてそのままの形で覚えるのではなく、自分の頭で再構成することによって、例えば場所の記憶が形作られるし、元々の感覚経験に含まれないものが生成できるからこそ、想像や夢が可能になっている。

そこでこの類似を念頭におきながら、画像生成の学習をおこなったときに付随する構造や、元々データセットに含まれていないような出力にどのようなものがあるかを調べることとした。これらは性能評価のための解析では基本的にとりあげられることがないため、十分な知見がこれまで得られてきていない。ここでは、DCGAN/VAEGAN という特定の生成系について、一人称視点の画像と音の学習をおこなった場合についての結果について、これらの観点から詳しく解析をおこなうこととした。

2. Model

ここでは DCGAN と VAEGAN という 2 種類の画像生成系を用いて実験をおこなった。

DCGAN[Radford 2015] は、GAN[Goodfellow 2014] の一つであり、画像の生成器 (Gen:Generator) と、画像が元々のデータセットに含まれていそうかどうかを判定する判別器 (Dis:Discriminator) を競い合わせて学習させるものである。生成器は連続値の z ベクトルを受けとり、画像を出力するため、出力画像は連続空間 (以下 z 空間と呼ぶこととする) に再構成されることになる。

また VAEGAN[Larsen 2015] は、DCGAN と類似しているが、生成器と判別器に加え、画像を入力として受け取り、その画像に近い画像を生成する z ベクトルを出力する、変換器 (Enc:Encoder) を含んでいる。他の違いとしては、生成器が基本的には VAE に基づいていることが挙げられる。ただし通常の VAE と違い、画像の類似度としてピクセルの値の比較ではなく、判別器の中間層の値の類似度を用いている。

2.1 データセット

z 空間は学習データセットの性質に強く依存すると考えられるため、どのようなデータセットを使うかということが重要である。通常の DCGAN では一定の基準を設けて、その範囲内で無差別に画像を収集したものを学習データセットとして用いることが一般的である。しかし、通常の状況で人が受け取る視覚情報は、一人称的な視点に基づくものである。

そこで本研究では、このデータセットの性質の違いが本質的なものであると考え、一人称的な視点に基づく画像のデータセットを使用して DCGAN の学習を行い、これにより自然に外界の内部モデルが z 空間上に構成されるのではないかと考えて実験を行った。

このデータセットとしてここでは、一人称視点の動画を撮影し、このフレームを画像として取り出したものを用いることとした。

3. 結果

3.1 一人称視点画像の学習

まず、人の場合との比較を考え、一人称視点の画像のデータセットを用いて学習をおこなった。これは、キャンパス内を歩き回って得られたビデオのフレームによって構成されるデータセットである。

学習結果の一例が図 1 であり、実際のキャンパスの風景と類似した画像を生成するようになった。さらに、生成器への入力ベクトルを連続的に変化させていくことで、画像が連続的に変化し、もともとの画像セットは離散的なものであったが、それが連続的な空間に再構成されていることがわかる（図 2)。

ここでさらに仮想空間の構造を調べるために VAEGAN を用いた。これを使うと画像の入力から z 空間への変換をすることができため、各画像の仮想空間内での位置を知ることができ

連絡先: 小島大樹: kojima@sacral.c.u-tokyo.ac.jp

図 1: input 画像の例（左）と output 画像の例（右）

図 2: z の値を直線的に変化させた時に得られる出力画像の例

る。ただし、通常 z 空間は高次元のため、この位置だけから構造を捉えることは難しい。そこで z 空間の構造を調べる別のやり方として、仮想空間を閉じて解析する closed loop という方法についてここでは述べる。これは、ある z ベクトルを生成器により画像にし、それを変換器に入れて z ベクトルに変換する、ということを繰り返すもので、これにより z 空間を力学系として取り扱うことができる。$(z(t+1) = Enc(Gen(z(t))))$

通常の VAEGAN の場合、この loop を繰り返すことによって、ほとんどの条件において一つないしは複数のアトラクターに収束した。

一方で、予測モデルを使った場合には、学習の条件によって、アトラクターに収束するだけでなく、limit cycle や、chaos 的な振る舞いが見られた。（図 3）

図 3: closed loop の結果の例：違う条件で学習した 2 つの結果を示す。上は limit cycle、下は chaos 的になっている。数字は loop を回した回数。

3.2 音の学習

また、同じ学習システムを用いて音の学習を行うことができる。これは音をスペクトログラムに変換することで画像にし、それを同様に学習させることで実現できる。ただし、周波数方向を log scale にして、なおかつ音に戻せるようにここでは invertible constant-Q transform [Holighaus 2013] というやり方を使用した。人でも蝸牛において対数スケールの周波数情報に変換している [Irino 1993] ため、自然な変換法であると考えている。

ここでは sin 波だけを学習した場合を示す。単音だけを学習しても、ポリフォニーや、音程の変化といったものが付け加わっている。

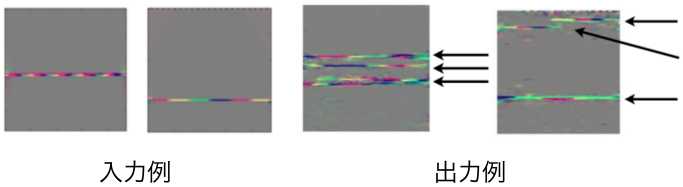

図 4: 音の学習の結果：左は学習データセットの例。右の 2 つは出力の例で、そのうち左のものはポリフォニーが、右のものはポリフォニーに加え音程の変化が見られる。

4. 議論

本研究では、生成系の学習の際にそれに付随してできる構造や、付け加わるものについて調べた。これらのものは機械学習としては通常失敗とみなされるため、積極的に調べられてこなかったが、このような汎化のクラスの違いによって調べることで、例えば脳の記憶システムに関して洞察が得られるかもしれない。今後は他のデータセットをより体系的に変えて調べることで、入力するデータセットと出来上がる構造の関係について明らかにしたいと考えている。

参考文献

[Goodfellow 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... and Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2672-2680).

[Radford 2015] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.

[Larsen 2015] Larsen, A. B. L., Sønderby, S. K., and Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300.

[Holighaus 2013] Holighaus, Nicki, et al. "A framework for invertible, real-time constant-Q transforms." Audio, Speech, and Language Processing, IEEE Transactions on 21.4 (2013): 775-785.

[Irino 1993] Irino, Toshio, and Hideki Kawahara. "Signal reconstruction from modified auditory wavelet transform." IEEE Transactions on Signal Processing 41.12 (1993): 3549-3554.