1A1-OS-05a-4in1

深層生成モデルのサンプリングダイナミクスが実現する概念への引き込み

Concept Formation Realized by the Sampling Dynamics in Deep Generative Models

長野祥大 *1　唐木田亮 *2　岡田真人 *1*2*3
Yoshihiro Nagano　Ryo Karakida　Masato Okada

*1東京大学 大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo

*2産総研 人工知能研究センター
AIST The Artificial Intelligence Research Center

*3理化学研究所 脳科学総合研究センター
RIKEN Brain Science Institute

Deep generative models are reported to be effective in wide applications including image generation. These models perform probabilistic inference based on sampling process, but it still remains unclear how the models recognize the input data during sampling. In this study, we numerically analyzed the dynamics of sampling in Variational Auto-Encoder trained with hierarchical data. Our experiments demonstrated that the transient dynamics of the latent variable were first attracted to the "concept", which is the center of the memorized patterns, and then escaped into each memory. The trajectories reflected the hierarchy of the dataset, and this behavior was closely related to the concept formation and retrieval process in recurrent associative memory models. Moreover, as the input's noise got increased, the activity pattern was drawn into a more abstract concept. These results indicate that the inference strategy of the model changes depending on input uncertainty.

1. はじめに

近年，主に画像生成などの分野で深層生成モデルが注目されている [Kingma 13]．これらのモデルはサンプリングを用いた確率的推論を行うが，その過程で入力データをどのように認識するかについては未解明な点が多い．

本研究では，階層的なデータセットで学習した Variational Auto-Encoder (VAE)[Kingma 13] のサンプリングダイナミクスを数値的に検証した．我々は VAE のサンプリングダイナミクスが多数の学習させた入力パターンの中心，"概念" に一度近づいたあとそれぞれの記憶パターンへ遷移することを明らかにした．その軌道はデータセットの階層性を反映しており，再帰的結合を持つ連想記憶モデルにおける記憶想起の過程と深い関連を示唆する結果であった．更に，入力に加えるノイズが大きくなるほど潜在変数の活動パターンはより抽象的な概念に引き込まれた．これらの結果から，VAE の推論の戦略は外部入力の不確実性に依存して変化することが示唆された．

2. モデル

VAE は高次元のデータ空間 $\boldsymbol{x}$ を低次元の潜在変数の空間 $\boldsymbol{z}$ に条件付き確率 $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ と $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ を用いてエンコードする確率的ニューラルネットワークモデルであり，目的関数は

$$\mathcal{L}(\theta,\phi;\boldsymbol{x}^{(i)}) = -D_{\mathrm{KL}}\Big(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})||p_\theta(\boldsymbol{z})\Big)+\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\Big[\log p_\theta(\boldsymbol{x}^i|\boldsymbol{z})\Big], \tag{1}$$

で与えられる [Kingma 13] (図 1)．認識と生成を以下の更新式

$$\boldsymbol{x}(t+1) \sim p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}(t)) \tag{2}$$

$$\boldsymbol{z}(t+1) \sim q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}(t+1)), \tag{3}$$

に従って T ステップ繰り返すことで，VAE は入力 $\boldsymbol{x}_0$ から条件付けでサンプリングを行い，潜在空間での時間発展 $\{\boldsymbol{z}(0),\boldsymbol{z}(1),\ldots,\boldsymbol{z}(T)\}$ を得る．本研究において，我々は 1024 の隠れ素子と 100 の潜在変数を持ち tanh の活性化関数を持つ 3 層ネットワークを構築した．学習及びその後の数値解析には手書き文字数字データセットである MNIST を用いた．

潜在空間の解析のために，数字における "概念" を以下の様に定義する．MNIST データの "0" から "9" までのラベルを num，そのラベル内のデータ番号を i とする．あるデータを入力として与えた際の潜在空間での活動パターンを $\boldsymbol{\xi}^{(i)}_{\mathrm{num}} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)}_{\mathrm{num}})}[\boldsymbol{z}]$ と定義し，数字ラベルが num の概念を $\bar{\boldsymbol{\xi}}_{\mathrm{num}} = \frac{1}{N_{\mathrm{num}}}\sum_i^{N_{\mathrm{num}}}\boldsymbol{\xi}^{(i)}_{\mathrm{num}}$，数字全体の概念を $\bar{\boldsymbol{\xi}}_{\mathrm{all}} = \frac{1}{10}\sum_{\mathrm{num}=0}^{9}\bar{\boldsymbol{\xi}}_{\mathrm{num}}$ とした．

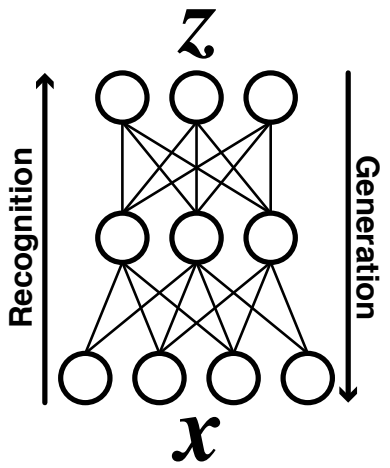

図 1: ネットワークの模式図．確率変数 $\boldsymbol{x}$ と $\boldsymbol{z}$ の対応関係を決定的なニューラルネットワークの写像で定義し，確率的要素は $p(\boldsymbol{z})$ のサンプリングが担う．$\boldsymbol{x}$ から $\boldsymbol{z}$ へのネットワークと $\boldsymbol{z}$ から $\boldsymbol{x}$ へのネットワークは必ずしも一致しない．

連絡先: 岡田真人，東京大学新領域創成科学研究科，
okada@k.u-tokyo.ac.jp

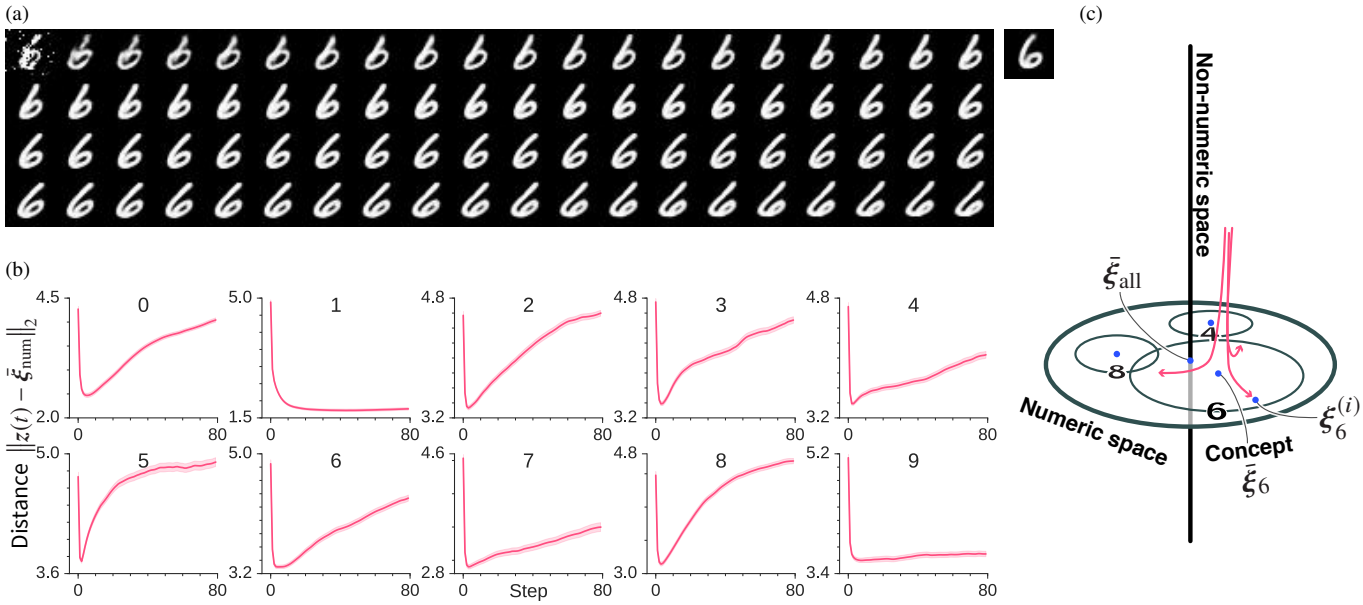

図 2: (a): データ空間におけるサンプリング. 右は概念 $\bar{\xi}_6$ から生成された画像. (b): すべての数字に関する概念 $\bar{\xi}_{\text{num}}$ からの距離の時間発展. 影は ±1 の標準誤差を表す (500 試行). (c): 潜在状態空間における発火パターンの概念図.

3. 結果

図 2aはデータ空間における活動の時間発展である. 我々は “6” の画像にノイズを $p = 0.2$ の割合で印加したものを初期値 $\boldsymbol{x}_0$ として与えた. モデルは壊された入力からノイズのない “6” を再構成した. 出力画像は右に示す潜在空間における様々な “6” の学習サンプルの平均である概念 $\bar{\xi}_6$ に近づいたあと, 歪んだ固有の画像に変化した (図 2a). 図 2aに示す推論のダイナミクスを定量化するために, 我々は発火パターン $\boldsymbol{z}(t)$ と潜在空間における概念 $\bar{\xi}_{\text{num}}$ の間の距離をすべての数字に関して検証した (図 2b). 発火パターンは高速にそれぞれの概念 $\bar{\xi}_{\text{num}}$ に近づいたあと, 低速に離れた (num $= 0, 1, \ldots, 9$). これらの結果はモデルがそれぞれのデータを認識するとき, 活動パターンの軌道が概念を一度経由することを示唆する (図 2c).

更に, 我々は入力のノイズの割合と活動パターンが引き込まれるデータの階層の関係について検証した. 図 3にノイズの割合に対するそれぞれの概念との最小距離を示す. 我々はモデルの挙動に従って推論の戦略を 3 つに分類した. ステージ I では, 少量のノイズの入力に対して活動は元のパターン $\xi_6^{(i)}$ に最も近かった. ステージ II では最も近い概念は中程度のノイズに伴って $\bar{\xi}_6$ に変化し, 最後にステージ III において活動は概念 $\bar{\xi}_{\text{all}}$ に近づいた. 本モデルはノイズが大きい環境下では物体を認識することは難しく, 活動パターンがすべての記憶の中心に最も近づくことを明らかにした. この結果から, 本モデルは入力の不確実性に応じて推論の戦略を変化させているといえる.

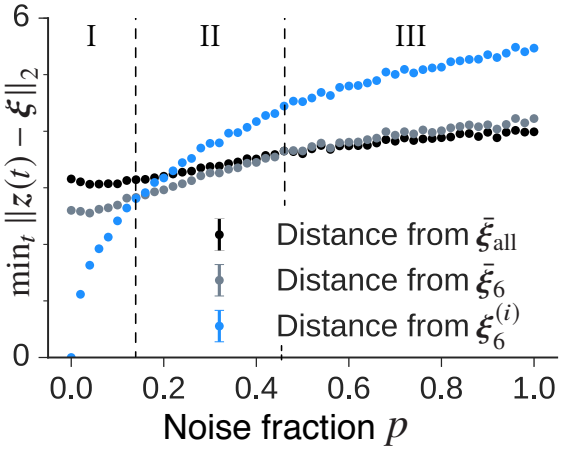

図 3: ノイズの割合に対する概念からの最小距離.

4. 議論

再帰的な結合を持つ連想記憶モデルでは複数の相関したパターンを埋め込むことで概念を形成することが知られており [Amari 77], その発火活動の時間ダイナミクスは一度概念に近づき, その後離脱することが解析されている [Matsumoto 05]. 本研究の結果はこれらの知見と深く関連することが示唆された.

また, 関連研究としてデータの階層性と深層ネットワークの学習過程 [Saxe 13] や層構造 [Bengio 13] との関連が指摘されており, 本研究の結果は学習やモデルの構造にとどまらず, 推論の過程もまたデータの階層性を反映することを示唆した.

参考文献

[Amari 77] Amari, S.-I.: Neural theory of association and concept-formation, *Biological cybernetics*, Vol. 26, No. 3, pp. 175–185 (1977)

[Bengio 13] Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S.: Better mixing via deep representations., in *International conference on machine learning*, pp. 552–560 (2013)

[Kingma 13] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013)

[Matsumoto 05] Matsumoto, N., Okada, M., Sugase-Miyamoto, Y., and Yamane, S.: Neuronal mechanisms encoding global-to-fine information in inferior-temporal cortex, *Journal of computational neuroscience*, Vol. 18, No. 1, pp. 85–103 (2005)

[Saxe 13] Saxe, A. M., McClelland, J. L., and Ganguli, S.: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, *arXiv preprint arXiv:1312.6120* (2013)