変分自動符号化モデルと符号ー複合モデルによる短文対話モデルの比較

Comparison Between Variational Autoencoder and Encodder-Decoder Models for Short Conversation

浅川 伸一
Shin asakawa
東京女子大学
Tokyo women's Christian university
asakawa@ieee.org

Abstract

We examined the dialogue generation models. Specifically, vanilla recurrent neural network, sequence-to-sequence, and variational autoencoding models were taken into account. Since these models are frequently employed in this areas, we tried to consider the possibility all of them as cognitive models.

Keywords — Neural Network Language Models, Variational Autoencoder, Discourse models

1. はじめに

会話，あるいは対話には文の理解と生成との両者が含まれる。したがって会話モデルには自然言語処理 (NLP) と機械学習の知識との両者が必要なると考えられる。従来研究では，領域特殊性，領域特化した特定の領域に限定されたモデルが多かった。従来研究とは異なり，Vinyals らはニューラルネットワーク言語モデル (RNNLM) に基づく短文対話モデルを提案した ([19], 図 1).

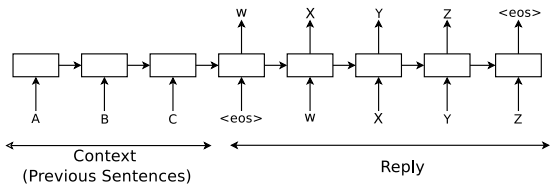

図 1　sequece to sequence モデルの概念図 [19] を改変

このモデルは本質的に翻訳モデル Sequence-to-sequence[18] の枠組みである。したがって翻訳モデルと同程度には領域に依存しない対話モデルを作成しうる端緒が開かられたと見做しうる。

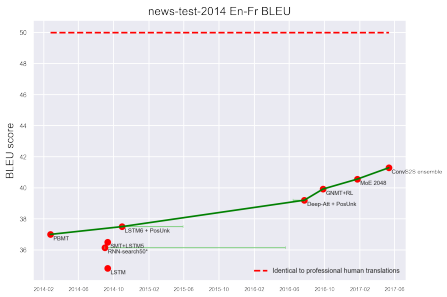

図 2　最近の翻訳モデルの性能の発展 https://github.com/AI-metrics/AI-metrics を改変

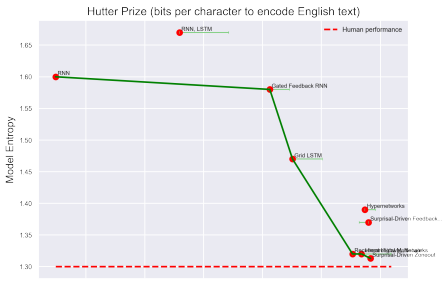

図 3　Hutter Prize データセットにおけるモデル性能の発展 https://github.com/AI-metrics/AI-metrics を改変。赤点線は人間の成績を示している

図 23 に近年のモデル性能の進展を示した。ビデオゲームにおいては人間の成績を遥かに凌駕するモデルが出現している [15] のと同じく，言語理解においても人間の成績に近づいていることが分かる。そこで以下では，対話モデルに用いられる基本モデルの幾つかを概観し整理することを試みる。

2. RNN 参照モデル

Mikolov の RNNLM [12] は NLP における性能評価の参照モデルとなっている。RNNLM は教師あり学習モデルであり，機械翻訳 [18, 1]. および画像脚注付

け課題 [20, 5] における言語材料を符号化可能である。RNNLM においては，一時刻前の単語と隠れ層の時間発展による条件付き確率を最大化するように学習が行われる。モデルは次に来る単語を予測することを学習する。バニラ RNN 参照モデルによって一定の性能が期待できるが，このモデルは文章ののベクトル表現を獲得しているわけではない。このモデルは単語の逐次入力表現から，中間層の時間発展に基づく分散表現を獲得し文章生成を行う。RNNLM は事前知識の作り込みを必要とせず内部表象の確率モデルを獲得し，長期依存などの複雑な系列のモデル分布を学習することができる。しかし，RNNLM ではトピックモデルや高次の統語特徴などの大域的で解釈可能な特徴を表現することはできない。

3. ベクトル埋め込みモデル

文章の分散表象を獲得するためのベクトル空間モデル (vector space model) には，少なくとも以下に示す 4 つのモデルが提案されている。

1. Paragraph Vector[9]
2. Skip-Thought[8]
3. Variational Autoencoder[7]
4. Sequence Autoencoders[10]

Paragraph Vector モデルは word2vec[14, 13, 11] の枠組みを用い，単語の意味表象の替わりに，段落の意味表象を得るモデルである。word2vec と同様 RNN モデルではない [9]。Skip-Thought モデルも教師なし学習モデルであり，ターゲットの隣接文による条件付き文章生成を行うモデルである [8]。通常の自動符号化モデルは大域的な意味特徴を効率的に抽出することは難しい。変分自動符号化モデルを用いて潜在変数 z を変化させた場合に解釈可能な中間的表現が得られるか否かを評価の難しい問題でもある。Sequence autoencoders は，教師あり事前訓練を用いる [4] ことで文章生成が実現されている [10]。このモデルでは，符号化関数 φ_{enc} と確率モデル $p(x|z=\varphi(x))$ とを用いて所与の例文 x についての尤度 z を最大化するように訓練が行われる。このモデルでは encoder も decoder には RNN が用いられた。

4. 変分自動符号化モデル

変分自動符号化モデル (VAE) は生成モデルである [7]。事前確率を定める潜在変数 z を用いることで，適切な分布が生成されるように学習がなされる。VAE は自動符号化モデルにおける関数 $\varphi(x)$ を学習可能の事後認識モデル $q(z|x)$ で置き換えたモデルである。通常 z は正規分布に従い，分散行列が対角行列であること，すなわち各潜在変数が独立であることを仮定する。VAE では z がニューラルネットワークによって定まると仮定する。VAE は文章を固定点として表現することはせず潜在空間上で任意の分布に従うと考える。VAE は $q(z|x)$ で与えられる変動を小さくするように学習が進行する [16]。事後確率分布を事前確率 $p(z) \propto G(\mu=0, \sigma=1)$ を近似するよう，変分下限を目標関数として以下の式 (2) を用いる。

$$
\begin{aligned}
L(\theta; x) &= -\mathrm{KL}\left(q_\theta(z|x) \| p(z)\right) \\
&\quad + \mathbb{E}_{q_\theta(z|x)}\left[\log p_\theta(x|z)\right] \qquad (1) \\
&\leq \log p(x). \qquad (2)
\end{aligned}
$$

モデルが事前確率の下で妥当な確率を有する潜在空間内のあらゆる点から適切な文章を符号化しうるようにする。

VAE モデルを使用した以下の実験では，Kingma らは [7] の正規分布のリパラメトライゼーションを提案した。彼らは確率勾配降下法でモデルを訓練し，各勾配ステップで、単一のサンプルを用いて再構成コストを推定した [7] の事後分布のコスト関数のカルバックライブラーダイバージェンス項を計算する。

[3] は大域的特徴を連続的な潜在変数に明示的に取り込む RNNLM の拡張モデルとしての VAE を提案した。VAE を用いれば，複雑な推論問題を解く必要がある。

Bowmman のモデルは変分自動符号化モデルのアーキテクチャーを使ってこの問題を回避し，変分推論の技法を用いた。[7, 6, 17]。潜在変数を持つ強力なニューラルネットワーク生成モデルのための実践的な訓練手法が用いられた。

5. 変分自動符号化モデルによる推論

N 個の独立同一分布に従うデータセット $\boldsymbol{X} = f\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N}$ をなしていると仮定する。$\boldsymbol{x}$ は連続分布でも離散分布でも良い。$\boldsymbol{x}$ は観測不可能な潜在確率変数 $\boldsymbol{z}$ から確率的に生成されたものとする。生成過程は以下の 2 段階からなる:

1. $z(i)$ は事前分布 $p_\theta(\boldsymbol{z})$ から生成される
2. $\boldsymbol{x}^{(i)}$ は同一の条件付き分布 $p_{\theta^*}(\boldsymbol{z})$ より生成される

事前分布 $p_{\theta^*}(\boldsymbol{z})$ と尤度 $p^*_\theta(\boldsymbol{x}|\boldsymbol{z})$ はパラメトリックな分布族であり $p_\theta(\boldsymbol{z})$ と $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ と確率密度関数は殆ど至る所で θ と $\boldsymbol{z}$ に関して微分可能であるとする。

生成過程は未知であると仮定する。従って θ^* と潜在変数 $\boldsymbol{z}^{(i)}$ は未知である。

簡単のため，周辺分布と事後分布とは以下のように求めることが可能であると考える。

1. *Intractability*: 周辺尤度の積分 $p_\theta(\boldsymbol{x}) = \int p_\theta(\boldsymbol{z}) p_\theta(\boldsymbol{x}|\boldsymbol{z})\, d\boldsymbol{z}$ は intractable なので，周辺尤度に関する微分が不可能である。事後分布 $p_\theta(\boldsymbol{z}|\boldsymbol{x}) p_\theta(\boldsymbol{z}) / p_\theta(\boldsymbol{x})$ も intractable であるので EM アルゴリズムを適用不可能である。積分できないので平均場変分ベイズのアルゴリズムも適用不可能である。これら intractabilities は複雑な尤度関数 $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ を持つ非線形項を持つ多層パーセプトロンなどのような場合は一般的である。
2. 大規模データ *t*: 確率的勾配降下法を適用するには大規模なデータセットにおいては小規模なミニバッチに基づくパラメータの更新には不向きである。モンテカルロ EM のようなサンプリング手法では収束に要する時間が長く，計算コストがかかる。

これを解決するために以下の 3 点を考える。

1. パラメータ θ の効率的な最尤推定 (ML) または最大事後推定 (MAP) は，システムの隠れ状態から生成される確率過程を推定し実データを模したデータを生成する data that resembles the real data.
2. パラメータ θ の選択と $\boldsymbol{x}$ 所与の元で潜在変数 $\boldsymbol{z}$ の効率的な事後推定。データ表象と符号化のために有益
3. 変数 $\boldsymbol{x}$ の効率的な周辺化量の推定。$\boldsymbol{x}$ の事前分布に基づくあらゆる推定問題で効率の良い周辺化推定量が必要となる。コンピュータビジョン，画像処理，画像生成などでも頻繁に用いられる。

上記の問題を解くため認識モデル $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ を導入する。これは事後確率 $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ の変分推論による平均場近似と見做しうる。パラメータ ϕ は明示的に計算することはできないが，再認モデルのパラメータ ϕ とパラメータ θ の同時分布を導入し学習させる手法を提案した。

符号化理論によれば，非観測変数 $\boldsymbol{z}$ は符号の潜在変数による表現であると考えることができる。データ $\boldsymbol{x}$ が所与のとき，コード $\boldsymbol{z}$ の値から任意の分布 (通常は正規分布を仮定する) を生成するので，再認モデル $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ は確率的符号器であると考えることが可能である。同様に $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ はデータ点 $\boldsymbol{x}$ についての確率的復号器と見做しうる。

6. 変分下限

周辺化尤度は個々のデータ点の周辺化尤度の総和 $(\log p_\theta(\boldsymbol{x}^{(1)}, \cdots, \boldsymbol{x}^{(N)}) = \sum_{i-1}^{N} \log p_\theta(\boldsymbol{x}^{(i)}))$ からなり，以下のように表現される:

$$\begin{aligned}\log p_\theta\left(\boldsymbol{x}^{(i)}\right) &= D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}\middle|\boldsymbol{x}^{(i)}\right) \middle\| p_\theta\left(\boldsymbol{z}\middle|\boldsymbol{x}^{(i)}\right)\right) \\ &\quad + \mathcal{L}\left(\theta, \phi; \boldsymbol{x}^{(i)}\right) \qquad (3)\end{aligned}$$

上式右辺第一項はカルバック=ライブラー情報量である。KL 情報量は非負であるので，右辺第二項 $\mathcal{L}\left(\theta, \phi; \boldsymbol{x}^{(i)}\right)$ は変分下界と呼ばれ，以下のように表記される:

$$\begin{aligned}\log p_\theta\left(\boldsymbol{x}^{(i)}\right) &\geq \mathcal{L}\left(\theta, \phi, \boldsymbol{x}^{(i)}\right) \qquad (4) \\ &= \mathbb{E}_{q_\phi}(\boldsymbol{z}|\boldsymbol{x})\left[-\log q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] \\ &\quad + \log p_\theta(\boldsymbol{x}, \boldsymbol{z}), \qquad (5)\end{aligned}$$

上式は以下のように書くことができる。

$$\begin{aligned}\mathcal{L}\left(\theta, \phi, \boldsymbol{x}^{(i)}\right) &= -D_{\mathrm{KL}}\left(q_\phi\left(\boldsymbol{z}\middle|\boldsymbol{x}^{(i)}\right) \| p_\theta(\boldsymbol{z})\right) \\ &\quad \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}\left[\log p_\theta\left(\boldsymbol{x}^{(i)}|\boldsymbol{z}\right)\right] \qquad (6)\end{aligned}$$

下限 $\mathcal{L}\left(\theta, \phi; \boldsymbol{x}^{(i)}\right)$ は変分パラメータ ϕ と生成パラメータ $\boldsymbol{\theta}$ について微分可能であるので最適化することができる。この変分下限を ϕ について微分するには通常のナイーブベイズモンテカルロ法による勾配の推定量は以下のようになる。

$$\begin{aligned}\nabla_\phi \mathbb{E}_{q_\phi(\boldsymbol{z})}\left[f(\boldsymbol{z})\right] &= \mathbb{E}_{\phi(\boldsymbol{z})}\left[f(\boldsymbol{z}) \nabla_{q_\phi(\boldsymbol{z})} \log q_{\phi(\boldsymbol{z})}\right] \\ &\sim \frac{1}{L}\sum_{l=1}^{L} f(\boldsymbol{z}) \nabla_{q(\phi)} \log q_\phi\left(\boldsymbol{z}^{(l)}\right) \qquad (7)\end{aligned}$$

ここで $\boldsymbol{z}^{(l)} \sim q_\phi\left(\boldsymbol{z}\middle|\boldsymbol{x}^{(i)}\right)$ である。この勾配の推定量は分散が大きい [2] ので，計算は簡単ではない。

7. 自動符号化変分ベイズモデル

難しい事後分布を持つ潜在変数と大きなデータセットが存在する場合，有向確率モデルで効率的な推論と学習が可能であろうか。[7] は大規模データセットに対応する確率的変分推論の学習アルゴリズムを提案した。軽度の微分可能な条件下では，難しいタスクでも動作する。 彼らは以下の 2 点を強調している：

1. 変分下限のリパラメトリゼーションが通常の確率的勾配降下法により最適化可能な下限推定量をもたらすこと
2. 独立な標本点ごとに連続潜在変数を持つデータセットに対して事後推論，近似推論モデル（認識モデルとも呼ばれる）の下限推定を用いて，使用

して難解な事後確率を効率良く推定可能であること

図 4 は変分自動符号化モデルのプレート表記法である。

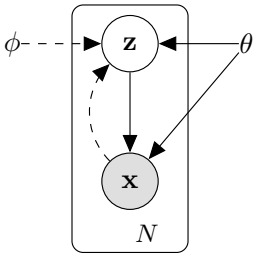

図 4　VAE

通常の階層ベイズモデルと異なり可観測変数 $\boldsymbol{x}$ から潜在変数 $\boldsymbol{x}$ へも有向線が伸びている。これにより変分自動符号化ベイズモデルは推定する手法が限定される。

8.　変分自動符号化言語モデル

[3] はそれまで画像分類に用いられてきた変分自動符号化モデルを文章に適用した。画像と文章との相違から困難であると思われる点を指摘し，標準的な言語モデルにおいては大域変数が必要ではないことを示した。彼らのモデルは大規模コーパスにも適用可能である。再帰的ニューラルネットワークを用いた文章生成モデル (RNNLM) では逐次的に一語づつ単語を生成することが行われるが，Bowman らは変分自動符号化に基づく RNNLM を提案した [3]。このモデルは文章全体を表象する潜在分布を仮定し，以下の 3 点のような文章の性質を表現することが可能である。

1. スタイル
2. トピック
3. 文全体の統語論的特徴

文章の表象についての分布からのサンプリングにより，決定論的で多様な文を生成可能である。潜在空間上で変数を変化させながら，逐次観察すれば，自然が文章の変動が観察可能であった。モデルは文章表現の潜在空間モデルであると見做しうる。加えてこのモデルでは文章の穴埋め問題に対して適切に答えることが可能であることが示された。Bowman らのモデルを変分自動符号化言語モデルと呼ぶ。図 5 に変分自動符号化言語モデルの概念図を示した。

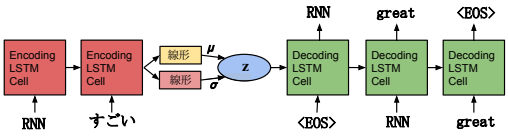

図 5　変分自動符号化言語モデルの概念図

Sequence-to-sequence 翻訳モデル [18] と異なる点は，ソース文とターゲット文との間に潜在変数を示すパラメータ μ と σ が存在し，かつ両者が潜在空間へ射影し $\boldsymbol{z}$ を形成しているコトである。$\boldsymbol{z}$ からターゲット文が生成される。

9.　考察

大域潜在変数を導入することで変分自動符号化モデルはターゲット文を抽象化し，表現が柔軟になることが期待される。この特徴は通常の言語モデルは持たない。定量的な評価を含めて対話文，物語生成に対する貢献を評価していく手法の確立が求められている。

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *Proceedings in the International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015.

[2] David M. Blei, Michael I. Jordan, and John W. Paisley. Variational bayesian inference with stochastic search. In John Langford and Joelle Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1367–1374, New York, NY, USA, 2012. ACM.

[3] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv:1511.06349*, May 2016.

[4] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3079–3087. Curran Associates, Inc., 2015.

[5] Jeff Donahue, Hendricks, Lisa Anne, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan , Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR), IEEE Conference*, Boston, Massachusetts, USA, 2015.

[6] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *Proceedings of the 32 nd International Conference on Machine Learning, JMLR*, volume 37, Lille, France, 2015.

[7] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv:1312.6114v10*, May 2014.

[8]Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *arXiv:1506.06726*, 2015.

[9]Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv:1405.4053v2*, May 2014.

[10]Jiwei Li, Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1106–1115, Beijing, China, July 2015. Association for Computational Linguistics.

[11]Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann Lecun, editors, *Proceedings in the International Conference on Learning Representations (ICLR) Workshop*, Scottsdale, Arizona, USA, May 2013.

[12]Tomaś Mikolov, Stefan Kombrink, Lukaś Burget, Jan "Honza" Černocký, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Prague, Czech Republic, May 2011.

[13]Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[14]Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous spaceword representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL*, Atlanta, WA, USA, June 2013.

[15]Volodymyr Mnih, Korya Kavukchuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstr, Shane Legg, and Demis Hassbis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

[16]Tapani Raiko and Mathias Berglund. Techniques for learning binary stochastic feedforward neural networks. *arXiv:1406.2989v3*, Apr 2015.

[17]Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32 nd International Conference on Machine Learning, JMLR*, volume 37, page 15301538, Lille, France, 2015.

[18]Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 3104–3112, Montreal, BC, Canada, 2014.

[19]Oriol Vinyals and Quoc V. Le. A neural conversational model. *arXiv:1506.05869*, 2015.

[20]Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015.