# 夏季レポート 2

情報科学類 3 年 江畑 拓哉 (201611350)

## 1 Sequence to Better Sequence: Continuous Revision of Combinatorial Structures　の翻訳

Sequence to Better Sequence: 組み合わせ構造の連続的な変形

### 1.1 要約

シーケンスなデータのペアを観測する学習をおこない、新しく入力されるシーケンスを連続的に修正するモデルを提示し、その関連についての結果を報告する。作成したフレームワークは、修正後の例も必要もせず提案された修正結果を追加評価する必要もない。シーケンスなデータの要素に対する ”組み合わせについての検索 (Combinatorial Search) ” を避けるため、連続的な潜在因子を有する生成モデルを使用することを選択し、つまり VAE 及び結果予測のニューラルネットワークモジュールを用いて ”結合した近似推論”(joint approximate interface) を行うことで学習する。このモデル下では、勾配法は推定された結果の連続的な潜在因子を最適化するために効率的な手法として用いられます。この最適化に適切な制約をかけ、VAE モデルのデコーダを利用して変形されたシーケンスを生成することで、変形は元のシーケンスに対して根本的には近いものの、より良い結果を得ることが出来、そしてそれは自然な結果に見えます。自然言語の文章を変形するためのタスクでは、このアプローチが先述の要件を高い確率で満たしていることが経験的に証明されている。

### 1.2 導入

シーケンシャルなデータを扱う機械翻訳や音声合成などの複雑なタスクにおいて、RNN モデルは良い成果を収めていることが知られいる。以下では例えば語彙のような離散値のシンボル $s \in \boldsymbol{S}$ を扱った可変長の長さ T のシーケンス $x = (s_1, \ldots, s_T) \in \boldsymbol{X}$ について議論する。

短文 (sentence) はその典型的な例であり、$s_j$ はその言語の単語を表現することになる。多くの場合では、 $\boldsymbol{X}$ (その言語が持つ単語を用いた、考えられるすべての組み合わせの集合) のうち、ほんの一部のみが自然な (現実的な) 短文として成立する。

例えば、ランダムな単語列は自然に読むことのできる一貫性のある文章をほとんど生成することはなく、同様にランダムなアミノ酸の配列が生物学的に活性なタンパク質を形成

することは非常に困難である。

本研究では、それぞれのシーケンス x が対応する結果 $y \in \mathbb{R}$ に関連付けられているアプリケーションを検討している。例えば、ニュース記事のタイトルや Twitter の投稿はその後ネット上でなされた共有数と関連付けることができる。また、合成タンパク質のアミノ酸配列は、その臨床的有効性と関連付けることができる。我々は教師付き学習の標準的な設定を考えて、シーケンスのペアの集合であるデータセット $\boldsymbol{D}_n = \{(x_i, y_i)\}_{i=1}^{n} \overset{iid}{sim} p_{XY}$ を想定している。周辺分布 $p_X$ は自然なシーケンスの生成モデルとして仮定されており、$\boldsymbol{X}$ の小さな部分空間に集中していると考えることができる。この文章中において、 p は参照している変数に依存した密度関数と分布関数の両方を意味している。

モデルを $\boldsymbol{D}_n$ へ最適化 (fit) させた次に、新しいシーケンス $x_0 \in \boldsymbol{X}$ (この出力 $y$ はわかっていないものとする) を用意する。我々の目的は、このシーケンスに対して変形されたパターンのシーケンス ($\hat{y}$) を迅速に特定することにある。つまり以下のような式を考えることになる。

$$x^\star = \underset{x \in \boldsymbol{C}_{x_0}}{\operatorname{argmax}} \mathbb{E}[\boldsymbol{Y} | \boldsymbol{X} = x] \tag{1}$$

ここで、集合 $\boldsymbol{C}_{x_0}$ は適用可能な変形された自然なシーケンスの集合であり、それらは単に元のシーケンス $x$ に小さな修正を施したものであると保証されていて欲しい。

生成モデルの観点から考えると、これらの 2 つの目的を達成するためには以下のような必要要件がある。

$p_{\boldsymbol{X}(x^\star)}$ は小さすぎず、 $x^\star$ と $x_0$ は似たような潜在特性を持っている。例えば短文を変形する場合には、変形した短文が自然に読むことが出来て (現実的な短文の分布の元ではそうなることが自然であると考えられる)、元の短文の意味を保持していることが不可欠である。

この最適化は難しいと考えられる。なぜなら制約集合と目的関数は非常に複雑で、いずれも未知のものであるからだ (これらはデータから学習しなければならない)。短文のような多くのシーケンスのデータにおいて、その空間 $\boldsymbol{X}$ あるいは $\boldsymbol{S}$ に対して直接標準的な距離関数 (例えばレーベンシュタイン処理や TF-IDF 近似) を適用することは、意味のある類似点を補足するには不十分であるが、これらは連続的な潜在要素の適切に学習された空間を、単純なメトリックによって忠実に反映することができる。

本研究では、我々はニューラルネットワークを用いて学習された連続値の潜在表現を利用して (1) を簡単な微分可能な最適化に変換する生成モデルのフレームワークを導入する。生成モデルが適合したならば、我々が提案する手順によって先述の必要条件を満たす新しいシーケンスを (高い確率で) 生成することができる。

## 1.3 関連手法

模倣学習とは異なり、我々の設定では特定のシーケンスの改良されたバージョンが利用可能である必要はない。これにより、Seq2Seq モデルの直接適用が妨げられる。似たようなアプローチとして、ベイズ最適化を利用して化学物質の構成を提案するために AutoEncoder を利用した研究がある。(`https://arxiv.org/abs/1610.02415`) しかし逐次的なバンディット/強化学習の設定と異なり、訓練データ以外の結果を見ることは出来

ず、修正を施した新しいシーケンスを生成してる結果が出ているものを見つけることはできなかった。

我々の提案している手法はシーケンスとそのペアとなるシーケンスの組み合わせのデータセットのみを必要としており、つまり広い分野に適用することができる。

組み合わせの構造はしばしば複雑なヒューリスティックな探索によって最適化される。その例としては、遺伝的プログラミングを挙げることができる。しかしサーチは各イテレーションで隔離された変更を評価することに依存するが、順序の良い変化はしばしば大きなコンテクストに渡って行われる。(例えば文章のあるフレーズを変化させること) 膨大な数の可能性から、そのような変形は検索 (サーチ) 手順によっては見出されにくく、そのような方法は高次元の連続的な値について議論する勾配ベースの最適化のほうが優れていることが一般的に言われている。

組み合わせ最適化とは異なり、我々のフレームワークはテスト時に効率の良い変形を見つけるために勾配を活用しています。Simonyan らと Nguyen らはまた、ニューラル予測に関して入力に対する勾配ベースの最適化手法を提案しているが、このような作業は (あるソースの変形ではなく) 条件付き生成に集中しており、主に連続画像に関する問題に限定されている。

## 1.4 手法

良い変形を見つけるにあたって、我々は最初に貪欲な組み合わせ問題を、目的と制約がより単純な形式となる連続空間に持ち込むことを考えた。つまり我々は Figure 1A のような確率的なモデルによってデータが生成されていると考えた。ここにおいて、潜在要因である $\boldsymbol{Z} \in \mathbb{R}^d$ は $\boldsymbol{X}$ $\boldsymbol{Y}$ (入力と出力) を生成するための連続値を取るパラメータであり、我々は事前にこれを $p_{\boldsymbol{Z}} = N(0, \boldsymbol{I})$ とする。(we adopt the prior pz = N(0,I)) これらの値の関係は $\boldsymbol{F}, \boldsymbol{E}, \boldsymbol{D}$ によって要約される。これらは、このモデルで効率的な近似推論を可能にするために、それぞれ 1 つのニューラルネットワークを用いて訓練されるパラメータである。

最初のステップでは、まずパラメータを調節してモデルを $\boldsymbol{D}_n$ に近づけることである。

Encoder を $\mathcal{E}$ 、Decoder を $\mathcal{D}$ とし、そして予測の出力を $\mathcal{F}$ とする。高品質の変形を行う良いモデルは、以下の特性を持っている。

- $\boldsymbol{Y}$ は $\boldsymbol{Z}$ から効率的に推論することが出来、この関係は滑らかな関数形 (functional form) に従う。
- $\boldsymbol{D}$ は合理的な事前確率を有する任意の $z$ を与えられた現実的なシーケンス $x$ を生成する。
- 自然なシーケンスの分布は潜在空間 $\boldsymbol{Z}$ において幾何学的に単純である。

これらの特性を持つために、我々は以下の方策を取る。

- $\mathcal{F}$ として単純な feed-forward network を選択する。

- $\boldsymbol{D}$ を z が与えられたときの最もそれらしい x として定義する。(defining D as the most-likely x given z)

- $\boldsymbol{Z}$ に単純な仮定 $N(0, \boldsymbol{I})$ を事前に付与する。

我々の $\boldsymbol{Z}$ の表現の望ましい別の特徴としては、隣接する z 値から基本的に類似したシーケンスが生成されるように意味のあるシーケンスの特徴をエンコードすることである。画像データに適用する場合、VAE モデルは我々のそれと同様に、スケール、回転、及び他の独立した視覚的概念などの顕著な特徴を解消する潜在的な表現を学ぶことがわかっています。文章の場合における学習された再帰的な機構の潜在表現 (ここで使用されているモデルに類似している) は、文章間の潜在空間における距離と人間が判断した類似性に強い相関があることがわかっている。

このように簡略化されたジオメトリを利用することによって、潜在的なベクトル空間における基本的なシフトは、シーケンスの要素の組み合わせ空間を直接操作する試みよりも高品質の変形を生成することができる。

これらの要求を満たすことのできるモデルのフィッティングが終わったならば、与えられたシーケンス $x_0 \in \boldsymbol{X}$ を変形する戦略を適用する。その概略は Figure 1B に示されているものである。まず、我々は学習した Encoding Map から入力の潜在表現 $z_0 = E(x_0)$ を計算します。潜在変数 $z$ は連続値であるため、効率的な勾配ベースの最適化を使用して、 $F(z)$ の近くの局所最適 $z^\star$ を見つけることができる。($z_0$ の周りに設定された後に定義される単純な制約内に見つけられる) $z^\star$ に対して、変形されたシーケンス $x^\star$ を得るために、我々は単純な Decoding map $\boldsymbol{D}$ (学習済みモデルに対して定義されている) を適用する。仮定されているモデル下で、潜在表現の最適化は ($\boldsymbol{F}$ を介して推測される)Y の大きな値を生成する生成構成を識別しようと試みる。そして次の復号化ステップでは、潜在因子の最適化された設定によって生成される最も可能性の高いシーケンスを求める。

## 1.5 Variational Autoencoder

X, Z の関係における近似推論のために、Variational Autoencoder モデルを利用する。我々の用いる VAE において、シーケンスの生成モデルは、尤度関数 $p_{\boldsymbol{D}}(x|z)$ と組み合わされた潜在値 $z$ に対する我々の事前確率に対して指定される。この尤度関数は、$z$ における任意のシーケンス $x$ の尤度を評価するための decoder network $\mathcal{D}$ が出力するものである。任意のシーケンス $x$ が与えられたとき、encoder network $\mathcal{E}$ は、潜在値 $p(z|x) \propto p_{\boldsymbol{D}}(x|z)p_{\boldsymbol{Z}}(z)$ の真の事後確率の変分近似 (Variational approximation) $q_{\boldsymbol{E}}(z|x)$ を出力する。なおこの変分近似には、Kingma & Welling よ Bowman らによって提唱されているように、対角共分散 (diagonal coviarance) を持つ変分族 (variational family) $q_{\boldsymbol{E}} = N(\mu_{z|x}, \Sigma_{z|x})$ を利用する。

我々の変形の手法では、シーケンスを潜在値 $z$ の最大事後 (MAP) 構成 (the maximum a posteriori configuration) にマッピングする符号化手順を採用している。(これは encoder

network $\mathcal{E}$ によって推定される)

$\mathcal{E}, \mathcal{D}$ のパラメータは、訓練データにおけるそれぞれの観測に対する周辺尤度の下限を最大化する確率変分推論 (stochastic variational inference) を用いて学習される。

$$\begin{aligned} \log p_{\boldsymbol{X}} &\geqslant -[\mathcal{L}_{rec}(x) + \mathcal{L}_{pri}(x)] \\ \mathcal{L}_{rec}(x) &= -\mathbb{E}_{q_{\boldsymbol{E}}(z|x)}[\log p_{\boldsymbol{D}}(x|z)] \\ \mathcal{L}_{pri}(x) &= KL(q_{\boldsymbol{E}}(z|x)||p_{\boldsymbol{Z}}) \end{aligned} \tag{2}$$

$\sigma_{z|x} = diag(\Sigma_{z|x})$ と定義すると、$q_{\boldsymbol{E}}, q_{\boldsymbol{Z}}$ が対角ガウス分布であるとき、事前強制 (prior-enforcing)KL ダイバージェンスは異なる閉形表現 (closed form expression)(see. `https://minus9d.hatenablog.com/entry/20130624/1372084229` ) を持つ。 $\mathcal{L}_{rec}$ 項 (すなわち decoder モデルの元での対数尤度) を再構築したものは、ただ一つの取り出されたモンテカルロ標本 $z \sim q_{\boldsymbol{E}}(z|x)$ より効率的に近似することができる。ニューラルネットワーク $\mathcal{E}, \mathcal{D}$ のパラメータに関して、我々のデータ $\mathcal{D}_n$ に対して変分加減を最適化するために、我々は誤差逆伝搬法と Kingma & Welling による再パラメータ化のトリック (see. `https://arxiv.org/pdf/1312.6114.pdf`) を用いて得られた、(2) の確率的勾配を用いる。

全体を通して、我々の encoder/decoder モデル $\mathcal{E}, \mathcal{D}$ は RNN である。RNN は各時間のステップ $t \in \{1, ;T\}$ において固定サイズの隠れ層の状態を示すベクトル $h_t \in \mathbb{R}^d$ が入力シーケンスの次の要素に基づいて更新されていくという、シーケンシャルなデータ $x = (s_1, \ldots, s_T)$ に対するニューラルネットワークである。与えられた x に対して近似的な事後確率を生成するために、我々の encoder network $\mathcal{E}$ には RNN の最終的な隠れ層の状態を表すベクトルに対して以下の層を追加している。(パラメータとして、$\boldsymbol{W}, b$ を取っている)

$$\begin{aligned} \mu_{z|x} &= \boldsymbol{W}_\mu h_T + b_\mu \in \mathbb{R}^d \\ \sigma z|x &= exp(-|\boldsymbol{W}_\sigma v + b_\sigma) \\ v &= ReLU(\boldsymbol{W}_v h_T + b_v) \end{aligned} \tag{3}$$

$\sigma_{z|x} \in \mathbb{R}^d$ の (二乗された) 要素は、我々の近似された事後共分散 (approximate-posterior coviarance) $\Sigma_{z|x}$ の対角要素を形成する。

$\mathcal{L}_{pri}$ は $\sigma_{z|x} = \vec{1}$ で最小化され、encoding の分散が更に増えるとこれが悪化する可能性がある (我々の事後近似は Unimodal(単峰) である)) ため、我々の変分族 (variational family) の 1 を超える $\sigma_{z|x}$ の値を単純に考えることが出来ない。この制限を加えることは、より安定的な学習を促し、また、真の事後確率が分散 $\leqslant 1$ で単峰性 (see. 正規分布) に近づくように encoder, decoder が共進化することを助ける (encourage)。

シーケンスの尤度を評価するため、RNN である $\mathcal{D}$ を隠れ層の状態を表すベクトル $h_t$ のみならず、以下の追加された出力も考慮する。

$$\pi_t = softmax(\boldsymbol{W}_\pi h_t + b_\pi) \tag{4}$$

それぞれのポジション $t$ において、$h_t$ を要約することで、$p(s_t|s_1,\ldots,s_{t-1})$ を予測する。 $p(s_1,\ldots s_T) = \Pi_{t=1}^{T} p(s_t|s_{t-1},\ldots,s_1)$ という因数分解を用いることで、 $p_D(x|z) = \Pi_{t=1}^{T}\pi_t[s_t]$ を得ることができる。これは最初の隠れ層の状態を表すベクトル $h_0 = z$ と、$x = (s_1,\ldots,s_T)$ を $\mathcal{D}$ に与えることで計算される。与えられた潜在設定 $z$ より、我々の変形は以下に示されるよりもっともらしい観測を通してでシーケンスを復号することで得られる。

$$D(z) = \underset{x\in\boldsymbol{X}}{\operatorname{argmax}}\, p_{\boldsymbol{D}}(x|z) \tag{5}$$

(5) を用いたよりもっともらしい復号は、組み合わせ問題それ自身であるが、 $p(x|z)$ の逐次因数分解を利用するビームサーチを用いることでより効率的に見つけることができる。$x^\star = \boldsymbol{D}(z) \in \boldsymbol{X}$ において、この $p_{\boldsymbol{X}}(x^\star)$ でも $p(z|x^\star)$ でもない探索を用いた復号戦略はとても小さいものである。

## 2 MEMO 1

元論文：http://www.mit.edu/~jonasm/info/Seq2betterSeq.pdf
ソースコード：https://bitbucket.org/jwmueller/sequence-to-better-sequence/
参考資料:https://www.slideshare.net/KazukiInamura/ai-lab-sequence-to-better-sequence-cont
TODO:最適化手法の理解
TODO:ソースコードの解読

類似研究 1 ：https://arxiv.org/pdf/1705.09655.pdf
ソースコード：https://github.com/shentianxiao/language-style-transfer
参考資料:https://www.slideshare.net/yuyasoneoka/dlstyle-transfer-from-nonparallel-text-by
TODO:Cross-Alignment Autoencoder の理解
TODO:ソースコードの解読

類似研究 3 ：http://proceedings.mlr.press/v70/hu17e/hu17e.pdf
ソースコード:https://github.com/GBLin5566/toward-controlled-generation-of-text-pytorch
TODO:論文の理解

## 3 MEMO 2

Seq2Seq の新しいモデル Pervative attention ：https://arxiv.org/abs/1808.03867
ソースコード：https://github.com/elbayadm/attn2d
TODO:Masked Convolution の動作の調査
TODO:ソースコード解読

# 4 MEMO3

NN 以外の手法を用いたチャットシステムの既存研究

遺伝的アルゴリズムを用いた会話型ご当地キャラクタによる地域活性化手法の提案 (http://www.hakodate-ct.ac.jp/~tokai/tokai/research/paper/ga2014.pdf)

遺伝的アルゴリズムを用いた文脈処理による質疑応答処理 (http://www.anlp.jp/proceedings/annual_meeting/2006/pdf_dir/P8-1.pdf)

A deep reinforcement learning chatbot(https://arxiv.org/pdf/1709.02349.pdf)

A deep reinforcement learning chatbot implementation (https://github.com/pochih/RL-Chatbot)