VAE による WSD の半教師あり学習

新納 浩幸
茨城大学工学部情報工学科
hiroyuki.shinnou.0828@vc.ibaraki.ac.jp

1 はじめに

本論文では Kingma らの提案した Variational AutoEncoder (以下 VAE) による半教師あり学習 [5] を語義曖昧性解消 (Word Sense Disambiguation, 以下 WSD) に利用した結果を報告する。

WSD のような分類問題に対しては、教師あり学習と大量のラベル付きデータにより高精度に解決可能である。ただし WSD ではラベル付きデータは対象単語の語義タグ付きの用例に相当し、人手でラベルを付けざるを得ないので、大量のラベル付きデータを用意することは困難である。これは WSD だけではなく、現実のほとんどの分類問題で生じている問題である。一方、ある種のタスクでは、ラベル付きデータは少量しか準備できないが、ラベルのないデータならば大量に存在するという状況がある。例えば WSD であれば対象単語を含む文がラベルなしデータに対応するので、ラベルなしデータを大量に用意することはそれほど困難なことではない。このように少量のラベル付きデータと大量のラベルなしデータから分類器を学習するのが半教師あり学習である。一般に少量のラベル付きデータのみから学習する教師あり学習よりも大量のラベルなしデータを合わせて利用する半教師あり学習の方が精度の高い分類器を構築できる。

半教師あり学習は現実的なアプローチであるため、従来より多くの研究があるが、近年、深層生成モデルを利用した半教師あり学習が提案された。深層生成モデルを利用した半教師あり学習は、生成モデルの実現法にネットワークを利用しており、その表現力の高さから現在の半教師あり学習手法の state of the art となっている。ただし具体的なネットワークの構成は試行錯誤で得るしかなく、実装上の工夫も必要であり、満足いく結果を得られないことも多い。

ここでは深層生成モデルを利用した半教師あり学習として、Kingma らの提案した手法 [5] を WSD に試してみた。利用できることは確認できたが、識別の精度はネットワークの構造やメタパラメータにかなり依存し、効果的に利用するにはある種のコツが必要であった。

2 関連研究

半教師あり学習に対しては多くの研究がある。古典的には Co-training [2] と NBEM [7] がよく知られている。Co-training は 2 つの独立した観点から相互に分類器を改善してゆく手法であり、NBEM はラベルを欠損値と考えて、Naive Bayes のモデルに EM アルゴリズムを用いる手法である。またアイデアの観点からは半教師あり学習は大きく 2 つに大別できる。一つは、ラベル付きデータから得られる分類器を使って、ラベルなしデータに確信度付きのラベルを付けて、それを利用して分類器を改善してゆくタイプの手法である。self-trainig [1] やラベル伝播 [14] などがこのタイプである。もう一つの手法がデータをある空間 [1] へマップするタイプの手法である。まずラベルなしデータをうまく分離できるような空間にマップし、次にラベル付きデータもその空間にマップし、その空間上で分類器の学習と識別を行うタイプである。通常、低次元にマップできれば、クラスを分ける境界を推定するためのラベル付きデータは少量で済むので、半教師あり学習が成立する。多様体論の手法 [10] や生成モデル [3] を利用した手法がこのタイプである。

深層生成モデルを利用した半教師あり学習は、生成モデルを利用した半教師あり学習と枠組み的には同じである。ラベルなしデータをうまく分離するような空間にマップする手法にネットワークを利用していると見なせる [5][8][11]。

3 VAE による半教師あり学習

VAE は深層生成モデルの 1 つであり、画像の生成などに利用されている。VAE を半教師あり学習に利

[1] 一般にもとのデータの次元よりも低次元の空間。

用した研究としては [5] や [9] などがある。ここでは [5] で提案された手法を説明する。

[5] では 3 つの学習手法 M1, M2 及び M1+M2 が提案されている。M1 が教師なし学習の手法、M2 が半教師あり学習の手法、そして M1+M2 が M1 と M2 を組み合わせた半教師あり学習の手法である。MNIST のデータセットで 100 個のラベル付きデータだけを用いて、M2 では約 0.90、M1+M2 では約 0.96 の正解率が得られている。

3.1 M1

$\boldsymbol{x}$ を観測データ、$\boldsymbol{z}$ を潜在変数として、生成モデルを以下で定義する。

$$p(\boldsymbol{z}) = N(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{1}), \quad p_\theta(\boldsymbol{x}|\boldsymbol{z}) = f(\boldsymbol{x};\boldsymbol{z},\boldsymbol{\theta})$$

f は尤度関数であり、ここではベルヌーイ分布を用いた。$\boldsymbol{\theta}$ がニューラルネットのパラメータである。また $\boldsymbol{z}$ の事後分布を以下で近似する。

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = N(\boldsymbol{z}|\boldsymbol{\mu}_\phi(x), diag(\boldsymbol{\sigma}^2_\phi(x)))$$

モデルのパラメータは θ と ϕ である。対数尤度比 $\log p_\theta(\boldsymbol{x})$ を最大化することでこれらパラメータを求める。イェンセンの不等式を利用すると以下の不等式が得られる。

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}) &= \log \int p_\theta(\boldsymbol{x}|\boldsymbol{z}) p_\theta(\boldsymbol{z}) d\boldsymbol{z} \\
&\geq \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{p_\theta(\boldsymbol{x}|\boldsymbol{z}) p_\theta(\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} \\
&= \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z} \\
&\quad - \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})} \\
&= E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] \\
&\quad - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})) \qquad (1)
\end{aligned}$$

式 (1) の第 2 項は $p(\boldsymbol{z})$ も $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ も各次元が独立な正規分布なので、その値は解析的に求まる。パラメータは $\log p_\theta(\boldsymbol{x})$ を最大化することで求まるが、$\log p_\theta(\boldsymbol{x})$ を最大化する代わりにその下限である式 (1) を最大化することで求める。ただし式 (1) の第 1 項は $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ に従って $\boldsymbol{z}$ をサンプリングするが、このままだとパラメータの微分値が求まらない。そこで変数変換のトリックを用いる。まず標準正規分布から $\boldsymbol{\epsilon}$ をサンプリングし、$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\epsilon} \cdot \boldsymbol{\sigma}$ とすることで微分が可能となる。式 (1) を $-L(\boldsymbol{x})$ とおき、$L(\boldsymbol{x})$ を損失関数としてパラメータを求めことで生成モデルが完成する。

3.2 M2

$\boldsymbol{x}$ のラベルを y とする。生成モデルを以下で定義する。

$$p(\boldsymbol{z}) = N(\boldsymbol{z}|\boldsymbol{0},\boldsymbol{1}), \;\; p(y) = \frac{1}{N_c}, \;\; p_\theta(\boldsymbol{x}|\boldsymbol{z},y) = f(\boldsymbol{x};\boldsymbol{z},y,\boldsymbol{\theta})$$

M1 と同様の計算から以下の不等式が得られる。

$$\begin{aligned}
\log p_\theta(\boldsymbol{x},y) &\geq E_{q_\phi(\boldsymbol{z}|\boldsymbol{x},y)}[\log p_\theta(\boldsymbol{x}|y,\boldsymbol{z}) + \log p(\boldsymbol{z}) \\
&\quad - \log q_\phi(\boldsymbol{z}|\boldsymbol{x},y)] = -L(\boldsymbol{x},y)
\end{aligned}$$

更に以下の不等式も得られる。

$$\begin{aligned}
\log p_\theta(\boldsymbol{x}) &\geq E_{q_\phi(y,\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|y,\boldsymbol{z}) + \log p(\boldsymbol{z}) \\
&\quad - \log q_\phi(y,\boldsymbol{z}|\boldsymbol{x})] \\
&= \sum_y q_\phi(y,\boldsymbol{z}|\boldsymbol{x})(-L(\boldsymbol{x},y) + H(q_\phi(y|\boldsymbol{x})) \\
&= -U(\boldsymbol{x})
\end{aligned}$$

つまりラベル付きデータ D_l からは $L(\boldsymbol{x},y)$ を最小化し、ラベルなしデータ D_{ul} からは $U(\boldsymbol{x})$ を最小化することでパラメータが求まる。つまり以下の J を損失関数とすればよい。

$$J = \sum_{D_l} L(\boldsymbol{x},y) + \sum_{D_{ul}} U(\boldsymbol{x})$$

最終的な識別は $q_\phi(y|\boldsymbol{x})$ を利用して行う。ただし $q_\phi(y|\boldsymbol{x})$ は $-U(\boldsymbol{x})$ の中にしか現れないために、分類器の学習ができない。そこで以下の拡張を行う。

$$J^\alpha = J + \alpha \cdot E_{D_l}[-\log q_\phi(y|\boldsymbol{x})]$$

ここで α は総データ数の 1 割の数とする。

3.3 M1+M2

M1+M2 では、M1 を用いて生成モデルを学習し、その結果を利用して各データ $\boldsymbol{x}$ を潜在変数 $\boldsymbol{z}_1$ に変換してから M2 を適用する。

4 WSD のための文脈のベクトル化

WSD の対象単語を w として、w を含む用例をベクトル x として表現すれば、前述した VAE の半教師あり学習をそのまま利用できることは明らかである。ここではベクトル x を w の文脈ベクトルと呼ぶ。

文脈ベクトルの作成方法は様々である。従来は直前の単語や直後の単語などの素性を one-hot-vector で

表し、それらを結合した形で文脈ベクトルを作成していた。ここではこれを基本ベクトルと呼ぶ、近年は word2vec 等から得られた分散表現を用いて文脈ベクトルを作成した方が精度が高くなることが示されている [12][13]。例えば [12] では基本ベクトルに対象単語の前後 2 単語、計 4 単語の分散表現を連結して文脈ベクトルを作成することで、基本ベクトルのみを用いた WSD の識別精度を改善している。[16] は分散表現データ nwjc2vec を用いると、SemEval-2 の日本語辞書タスクでは対象単語の前後 2 単語、計 4 単語の分散表現のみを連結して文脈ベクトルを作った方が識別精度が高いことを示した。ここでは [16] の結果を踏まえて、対象単語の前後 2 単語、計 4 単語の分散表現のみを連結して文脈ベクトルを作成する。

5 実験

実験は SemEval-2 の日本語辞書タスクで設定された対象単語の “子供” のデータを用いる。“子供” の語義は以下の 2 つである。

```
17877-0-0-1-0   幼い子。児童
17877-0-0-2-0   自分のもうけた子
```

ラベル付きデータは 50 用例、テストデータも 50 用例である。また毎日新聞 CD-ROM ’93 年度版から ’99 年度版のコーパスから単語 “子供” を含む 7,380 文を取り出し、それをラベルなしデータとした。各用例は対象単語 “子供” の前後 2 単語の分散表現を nwjc2vec から得ることで、800 次元のベクトルとして表現した。

これらデータを用いた VAE による半教師あり学習の M2 を実行した結果を以下に示す。横軸は学習の epoch 数を示し、縦軸はテストデータに対する正解率を示す。得られた正解率は約 0.70 である。

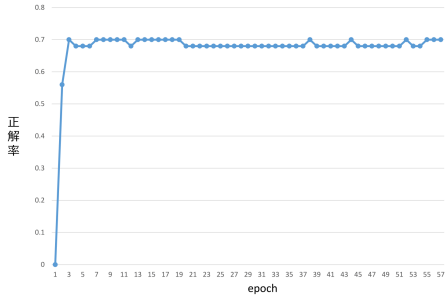

図 1: “子供” の WSD の半教師あり学習結果

また基本ベクトルと SVM を用いた場合の正解率は 0.62 であった。nwjc2vec から得られた 800 次元のベクトルから SVM で分類器を学習した場合の正解率は 0.64 であった。上記の M2 による正解率はそれらよりも高く、ラベルなしデータを用いた効果はあった。

6 考察

6.1 実装上のコツ

VAE による半教師あり学習では $\boldsymbol{x}$ から y、$\boldsymbol{x}$ と y から $\boldsymbol{z}$ および $\boldsymbol{z}$ と y から $\boldsymbol{x}$ の変換をネットワークで表現する（図 2 参照）。そのネットワークの構造に特に制約はなく、自由に設定できる。しかし下手に設定すると全く学習できない。

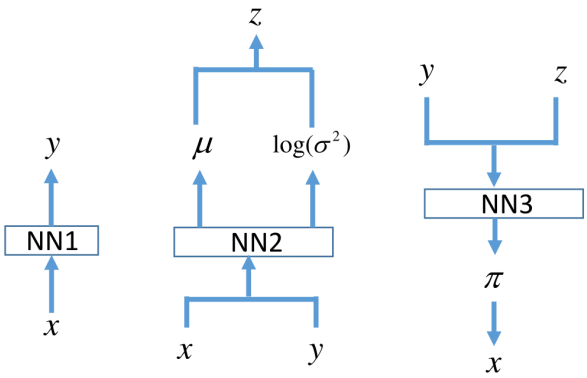

図 2: VAE の半教師あり学習のためのネットワーク

例えば図 2 の NN2 において $\log(\sigma^2)$ を出力しているが、単に σ^2 を出力させるようにしてしまうと負の値が出力される場合に不具合が生じる。$\log(\sigma^2)$ を出力させるのが Tips である。また図 2 の NN2 で μ と $\log(\sigma^2)$ を独立に生成するような形にするとうまく学習できない。また NN1、NN2、NN3 の層はできるだけ少ない層がよい。ここではどれも最小の層で NN1 は 1 層、NN2 は 3 層、NN3 は 2 層になっている。また図 2 の x と y を NN2 の入力とするが、NN2 の中では x と y の出力を合わせて次の層に渡す。この合わせる部分は、2 つのベクトルを連結するのか、足すのか、掛けるのかなどの実現方法があり、何が正解かは分からない。ここでは掛けることで実装した。更に潜在変数 $\boldsymbol{z}$ の次元数もいくつに設定すべきか難しい。次元数に応じて得られる結果がかなり異なった。ここでは次元数は 5 に設定した。また学習時の学習率もかなり結果に影響した。ここでは最適化のアルゴリズムは Adam を使い、学習率は 0.0005 に設定した。

本論文で実装したネットワークの形が最適であるとは思えない。また学習時に利用するメタパラメータ

（バッチのサイズや J^{α} の α）も適切ではなかった可能性がある。つまり VAE による半教師あり学習は理論的には明瞭であるが、その実装にはかなりのコツが必要だと思われる。このコツの習得を今後の課題とする。

6.2 文脈ベクトル

本論文では対象単語の文脈ベクトルとして前後 2 単語の分散表現を利用した。この場合、本質的に学習では前後 2 単語が汎化されるだけであり、WSD の識別のための情報が前後 2 単語の外にある場合には、正しい識別ができる理由がない。そこで対象単語の前後の文脈を双方向 LSTM から得る研究が行われている [4][6]。また順方向 LSTM と前後 2 単語の分散表現から文脈ベクトルを得る研究もある [15]。これらの手法を使えば、対象単語の文脈ベクトルは更に圧縮したベクトルで表現でき、しかもこれは対象単語の前後の文脈をベクトル化したものであり、本質的に WSD で必要とされる情報が全て組み込まれている。今後はこのように作成したベクトルを用いたい。

7 おわりに

本論文では VAE を利用して WSD の半教師あり学習を試みた。小規模の実験から、利用可能な手法であることは確認できた。ただし精度の向上には、より文脈を適切に表した文脈ベクトルを利用する必要がある。またネットワークの構造や実装上の細かな違いが精度に大きく影響しており、使いこなすためのコツが必要な手法だと思われる。潜在的に能力の高い手法なので、今後はこのコツの習得を目指し、共有できる知識として提示することを行いたい。

謝辞

本研究の一部は国立国語研究所の共同研究プロジェクト「all-words WSD システムの構築及び分類語彙表と岩波国語辞典の対応表作成への利用」の研究成果を報告したものである。

参考文献

[1] Steven Abney. *Semisupervised learning for computational linguistics*. CRC Press, 2007.

[2] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 92–100. ACM, 1998.

[3] Fabio G Cozman, Ira Cohen, and Marcelo C Cirelo. Semi-supervised learning of mixture models. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 99–106, 2003.

[4] Kågebäck, Mikael and Salomonsson, Hans. Word Sense Disambiguation using a Bidirectional LSTM. In *5th Workshop on Cognitive Aspects of the Lexicon (CogALex)*. Association for Computational Linguistics, 2016.

[5] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pp. 3581–3589, 2014.

[6] Oren Melamud, Jacob Goldberger, and Ido Dagan. context2vec: Learning Generic Context Embedding with Bidirectional LSTM. In *CoNLL-2016*, pp. 51–61, 2016.

[7] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, Vol. 39, No. 2/3, pp. 103–134, 2000.

[8] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pp. 3546–3554, 2015.

[9] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

[10] Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pp. 2294–2302, 2011.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

[12] Hiromu Sugawara, Hiroya Takamura, Ryohei Sasano, and Manabu Okumura. Context Representation with Word Embeddings for WSD. In *PACLING-2015*, pp. 108–119, 2015.

[13] Shoma Yamaki, Hiroyuki Shinnou, Kanako Komiya, and Minoru Sasaki. Supervised Word Sense Disambiguation with Sentences Similarities from Context Word Embeddings. In *PACLIC-30*, pp. Y16–1010, 2016.

[14] Xiaojin Zhu and Zoubin Ghahramani. *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[15] 新納浩幸, 古宮嘉那子, 佐々木稔. 順方向多層 LSTM と分散表現を用いた教師あり学習による語義曖昧性解消. 情報処理学会第 232 回自然言語処理研究会, pp. NL–232–4, 2017.

[16] 新納浩幸, 浅原正幸, 古宮嘉那子, 佐々木稔. nwjc2vec : 国語研日本語ウェブコーパスから構築した単語の分散表現データ. 自然言語処理, Vol. 24, No. 5, pp. 705–720, 2017.