超球面上への分布エンコーディングを用いた文書分類

守屋 俊　　町田 秀輔　　柴田 千尋
東京工科大学 バイオ・情報メディア研究科
{c0115334ef, c011624552, shibatachh}@edu.teu.ac.jp

1 はじめに

表現学習とは，教師あり，教師なしを問わず，深層ニューラルネットの内部において，より質の高い埋め込み表現を得ようとする試みである．ここでいう質の高い状態とは，例えば，文書分類を行う際に，各文書に対して，文書の自然なトピックに対応した内部状態のクラスタが存在しているような状態を指す．本研究では，超球面上への分布エンコーディングを用いた文書分類を通して，識別性が高い隠れベクトルを得ることを目標とする．

先行研究としては，ニューラルネットの学習を変分下界の最大化としてとらえ，分布エンコーディングを用いる代表的な表現学習の手法として，Variational Auto Encoder(VAE)[4] や変分情報ボトルネック [1] が挙げられる．一方で，クラスごとに隠れ状態をまとめる研究はいくつか存在する．Center Loss[7] は，より識別性の高い特徴量を得るために，隠れ状態をクラスごとにまとめる手法である．Entropy Penalty[2] では，訓練データとテストデータの分布の違いに頑健な分類モデルを得るために，ニューラルネットの各層の隠れ状態をクラスごとにまとめることによって，隠れベクトルがクラス以外の情報を持つことを制限している．

2 変分情報ボトルネック法

本研究における文書の分布埋め込みとは，文書 x に対して，表現 z を計算が容易な確率分布で定義したものであり，ニューラルネットの重みなどのパラメータを θ_1 として，$P(z|x,\theta_1)$ で表される．ここで，z の分布は，x のみに依存して決定され，ラベル y とは条件付き独立である．文書分類における情報ボトルネック法は，以下の目的関数を最大化することにより，文書分類にとって最適な表現 Z を得る手法である．

$$f_{\mathrm{obj}}(\theta_1) = I(Y,Z) - \beta I(Z,X)$$

ここで，$I(\cdot,\cdot)$ は相互情報量を表し，β は正の定数である．また，X,Y,Z はそれぞれ，文書，ラベル，表現ベクトルを表す確率変数である．上式の第一項では，表現ベクトルとラベルとの間の情報量をより増やす方向に，第二項では，文書が表現ベクトルへ与える情報量を極力減らす方向に，最適化が行われる．θ_1 を決めると，$P(z|x,\theta_1)$ は容易に計算できるが，$P(y|z)$ は計算することは困難である．そこで，VIB では，別の NN(そのパラメータを θ_2 とする) を用いて，容易に計算できる分布 $q(y|z,\theta_2)$ で近似する．また，同様に，$P(z|\theta_1)$ についても，個々の文書データ x によらない，より簡潔な何らかの分布 $r(z)$ で近似することが一般的である．任意の分布間において KL ダイバージェンスが非負であることを用いて計算すると，以下の不等式が成り立つ．

$$\begin{aligned} f_{\mathrm{obj}}(\theta) \quad &\geq E_{x,y}\Big[E_{z\sim P(z|x,\theta_1)}\big[\log q(y|z,\theta_2)\big]\Big] \\ &\quad -\beta\, E_x\Big[\mathrm{KL}\big(\,P(z|x,\theta_1)\,||\,r(z)\,\big)\Big] \end{aligned} \tag{1}$$

上式右辺は一般に変分下界と呼ばれる．第一項は正解ラベルと予測ラベルとのクロスエントロピーを表し，第二項は，z の x によらない分布による正則化を表す．第一項は分布埋め込みを行う点を除くと，通常の決定論的な NN を用いた場合の，多クラス分類におけるクロスエントロピーロスと同等である．変分情報ボトルネック法 (VIB 法) は，この変分下界を最大化するような θ_1, θ_2 を求める方法である．

2.1 文書分類への適用における具体形

まず，本研究では， 文書 x が，BERT[3] などすでに何らかの手法で得られた M 次元のベクトル表現として入力されるものとする ($x \in \mathbb{R}^M$)．また，表現 z も，N 次元のベクトル表現であるとする ($z \in \mathbb{R}^N$)．

次に，[1] と同様に，$P(z|x,\theta_1)$ は，平均 $\mu(x,\theta_1)$，分散 $\sigma^2(x,\theta_1)$ をもつ N 次元の正規分布で与えられるものとする．また，$q(y|z,\theta_2)$ は，softmax 関数などを通

して得られる通常のカテゴリカル分布とする ($f(z,\theta_2)$ とおく). また，$r(z)$ は，デフォルトでは，[1] と同様，正規分布 $\mathcal{N}(0,1)$ とする．

式 1 右辺第一項は，x,y をデータよりサンプルし，z を再パラメータ化トリックを用いてサンプルすると，

$$-\mathrm{CrossEnt}\Big(\, f(\mu(x,\theta_1)+\sigma(x,\theta_1)\varepsilon\,,\,\theta_2)\,,\,y\,\Big) \quad (2)$$

となる．ここで，$\varepsilon\sim\mathcal{N}(0,1)$ である．一方で，第二項は，$\mathcal{N}(\mu,\sigma^2)$ と $\mathcal{N}(0,1)$ との間の KL ダイバージェンスになるから，

$$-\beta\,\mathrm{KL}(P||r) = -\frac{\beta}{2}\big(|\mu|^2 + N(\sigma^2-\log\sigma^2-1)\big) \quad (3)$$

となる．

3 提案手法

本研究では，次の 2 つの手法を提案し，その効果を検証する．一つは，表現ベクトルの分布を正規化することで，分布を超球面付近または超球面上に拘束する方法，もう一つは，$r(z)$ を，ラベルにごとに異なる分布とする方法である．

3.1 超球面上での分布エンコーディング

変分下限の第二項 (式 3) は，目的関数の最大化の観点から考えると，μ を 0 に近づける，および，σ^2 を 1 に近づける[1]，という 2 つの効果を持つ．β はその効果を影響力を表す定数といえる．一般的に，β の値は，分類精度が高くなるように，実験的に決定されるが，結果的に，先行研究では，対象となる問題は異なるものの，β をかなり小さい値にした場合が最適であることが知られている．一方で，β を 0 に近づけると，$|\mu|$ を原点付近に留める力が弱まるため，分布エンコードによる撹乱 (式 2 中の ε) の効果が弱まり，結果として過学習を引き起こしやすくなり，予測精度が安定しなくなると考えられる．そこで，本研究では，$|\mu|$ または $|z|$ を 1 に正規化，つまり超球面 (S^N) 上へ射影し，分布エンコードとすることを提案する．具体的には，次の 3 つの手法に対し，実験を通してその効果を検証する．

(a) μ の正規化： $z=\mu/|\mu|+\sigma\varepsilon$.

(b) z の正規化： $z=\tilde{z}/|\tilde{z}|$，ただし， $\tilde{z}=\mu+\sigma\varepsilon$.

(c) μ の接平面における正規分布の正規化 : $z=\tilde{z}/|\tilde{z}|$，ただし，

$$\tilde{z} \;=\; (1-\sigma\varepsilon\cdot\tilde{\mu})\tilde{\mu}+\sigma\varepsilon, \quad (4)$$

ここで，$\tilde{\mu}=\mu/|\mu|$.

z の分布は，手法 (a) では S^N 上にはないが，(b) および (c) では S^N 上の分布となっている．また，(c) の $\tilde{z}$ は，次のようにして求まる．μ 方向の単位球面上での接平面の法線ベクトルは $\tilde{\mu}$ である．従って，$\mathcal{N}(0,\sigma^2)$ よりサンプルされた点 $\sigma^2\varepsilon$ の接平面への射影は $\sigma^2\varepsilon-(\sigma^2\varepsilon\cdot\tilde{\mu})\tilde{\mu}$ である．ガウス分布の性質より，射影後のベクトルは接平面上の分散 σ^2 の正規分布に従う．これを $\tilde{\mu}$ だけ平行移動させれば良いから，$\sigma^2\varepsilon-(\sigma^2\varepsilon\cdot\tilde{\mu})\tilde{\mu}+\tilde{\mu}$ となり，式 4 を得る．

3.2 ラベル毎の情報ボトルネック

VIB[1] では，文書の表現の分布 $P(z|x,\theta_1)$ の近似分布 $r(z)$ を，$\mathcal{N}(0,1)$ としているが，実際には，多クラス分類においては，z の分布はクラス数分の正規分布の混合分布で近似するほうがより正確である．また，z と y とは，x を介して条件付き独立であるため，$P(z|x,\theta_1)=P(z|x,y,\theta_1)$ であるから，近似分布 r が y に依存してもそのまま P との KL ダイバージェンスを計算することが可能である．そこで，本論文では，近似分布 $r(z)$ を，ラベル y ごとに平均が異なる正規分布の混合分布として定義することを提案する．すなわち，条件付き分布 $r(z|y)$ が $\mathcal{N}(\mu_y,1)$ で与えられるものとする．このとき，式 3 は，$g(\sigma^2)=\sigma^2-\log\sigma^2-1$ とおいて，

$$-\beta\,\mathrm{KL}\left(P||r(z|y)\right) = -\frac{\beta}{2}\left(|\mu-\mu_y|^2+Ng(\sigma^2)\right),$$

と計算できる．また，特に $|\mu|=|\mu_y|=1$ のときは，上式は $\beta\left(\mu\cdot\mu_y-1-Ng(\sigma^2)/2\right)$ となり，その最大化は，μ を μ_y と同じ向きにするような効果があることがわかる．

4 実験

超球面上での分布エンコーディングの比較と，文書分類への分布エンコーディングの効果を検証する 2 つの実験を行った．分類対象のデータセットとしては，20Newsgroups[2]データセットの分類の正解率で比較を行った．

[1] $\sigma^2-log(\sigma^2)$ が $\sigma^2=1$ のとき最小となるため．

[2] http://qwone.com/ jason/20Newsgroups/

超平面上での分布エンコーディングの比較を行うために，節 3.1 の 3 つの手法とベースラインに関して，分類の正解率を比較した．文書分類のモデルには SWEM-concat[6] で用いられている，単語ベクトルの平均と各次元の最大値を結合したベクトルを，節 2.1 における x とする．単語ベクトルの初期化には，GloVe[3] の学習済みベクトルを用いた．

文書分類への分布エンコーディングの効果を検証する実験では，分布エンコーディングの文書分類に対する効果をを確認するために，節 3.1 の手法 b(dist1 と表記) と節 3.2 のラベル毎の情報ボトルネック (distC) とベースラインの 3 つについて，分類の正解率を比較した．また，ベースライン手法と分布エンコーディング手法の隠れベクトルの次元数による性質の違いを確認するために，隠れベクトル z の次元数を 20，100，300 の 3 つの条件で実験を行った．超平面上での分布エンコーディングの比較で用いた手法に加えて，BERT[3] の CLS トークンを x とした場合での実験を行った．

表 2 の実験に使用した BERT では Toronto Book Corpus と Wikipedia を使って学習された BERT-base のアーキテクチャ[4]を使って FineTune する．

超球面上への分布エンコーディングを行う際は，節 3.1 の手法 b を用いた．表 2 の実験を行う際には，全ての枠組みにおいて $\beta = 0.001$ とした．

5 実験結果

超球面上での分布エンコーディングの提案手法 (a)-(c) の効果を表 1 に示す． ベースラインとして，分布エンコードを行わず，ドロップアウトを用いたもの，分布エンコードは行うが，球面上への射影はとらないものを用いた場合の分類精度を，それぞれ，表の 1,2 行目に示す．なお，後者のベースラインで $\beta = 0.1$ となっている理由は，その手法では β の値を調整した結果，その時が最も精度が高かったためである．表からわかるように，超球面上の分布エンコードを用いると，(a)-(c) のどの手法であっても，ベースラインの手法と比較して精度が高くなっていることがわかる．(a)-(c) 内での比較では，z の次元数が 20,100 の時には手法 (b) が最も良く，300 の時には手法 (a) が良い結果となっている．

表 1: エンコーディングの比較；分布エンコードの有無及び超球面の手法 (20news, SWEM)

Distrib.	Space	β	representation dim. 20	100	300
None(dout)	$\mathbb{R}^N$	-	39.70	84.27	84.35
Normal	$\mathbb{R}^N$	0.1	84.61	85.78	85.63
Normal	S^N(a)	0.001	84.75	86.64	**86.76**
Normal	S^N(b)	0.001	**85.63**	**86.71**	86.38
Normal	S^N(c)	0.001	85.08	86.38	86.40

[3]https://nlp.stanford.edu/projects/glove/
[4]https://huggingface.co/transformers/

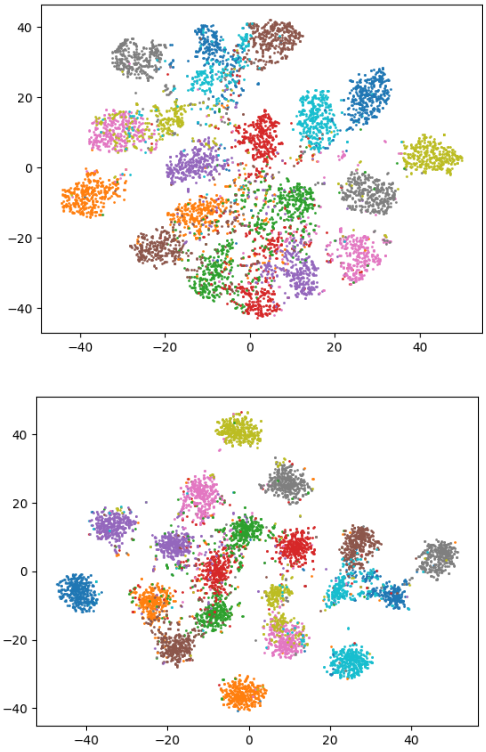

図 1: ドロップアウトを用いた場合 (上) と超球面分布エンコーディング (下) を用いた場合の比較 (300 次元, SWEM)

次に，ボトルネックにおける各手法 (dist1,distC) の効果を検討する．20newsgroups に対する比較結果を表 2 に示す．表 2 より，SWEM の場合，全ての z の次元数に対して，SWEM distC の分類精度が高くなっている．BERT の場合，全ての z の次元数に対して，BERT dist1 の正解率が高くなっている．いずれにせよ，分布埋め込みを行わない場合に比べて，精度が向上していることがわかる．また，特に埋め込み次元数が小さいほど，その差は顕著になっている．

最後に，20newsgroups データセットに対する SWEM base と SWEM distC の 300 次元の z に対して t-SNE[5] を適用し，可視化したものを図 5(上) と図 5(下) に示す．

表 2: 20newsgroups の分類精度

Input	Bottleneck	representation dim. 20	100	300
SWEM	None(dout)	39.70	84.27	84.35
SWEM	dist1	85.63	86.71	**86.38**
SWEM	distC	**86.18**	**86.82**	**86.38**
BERT	None(dout)		84.68	
BERT	dist1	**86.27**	**86.52**	**86.03**
BERT	distC	86.21	85.28	83.44

6 考察

表 1 の結果より，節 3.1 の手法 c はあまり有効ではなかった．手法 c の意図としては，μ が原点付近のときに，ノイズ ε によってベクトルの各次元の要素の符号が変わってしまうと，正規化した後のベクトルの向きが大きく変わってしまうことを抑制する効果を期待していた．したがって，手法 c の効果が薄いということは，μ の向きがノイズ ε によって大きく変わってしまうことがあまり起きていない，もしくは，μ が原点付近に集まるのは，β が大きい場合であり，分散 σ^2 等との兼ね合いで，そもそも分類がうまくいかない場合にしかそのようなことが起きないといったことが考えられる．

表 2 の結果より，z の次元数が少ない (クラス数相当) 時に，SWEM base の dropout の精度が低くなっている．一方で同じ次元数の z でも，分布エンコーディングベースの手法は，比較的高い正解率を維持している．このことから，分布エンコーディングベースの手法は，分類のために必要な情報を低次元のベクトルに効率的に変換できていると考えられる．

図 5(上) と図 5(下) を比較すると，分布エンコーディングによる手法で得られた z の方が，わずかによくまとまっているように見える．このことから，分布エンコーディングによって，より識別性が高い z が得られたということができる．

BERT の CLS トークンを用いて，クラス毎の情報ボトルネック (distC) を適用すると，隠れベクトル z の次元数が多くなるにつれて，正解率が下がっている．dist1 に関しては，z の次元数によらずに，一定程度の正解率を維持している．このことから，CLS トークンの表現力が高いために，クラスごとにまとめる制約が強い枠組みで BERT を学習すると，隠れ状態の次元数を大きく取ってしまった時には，過学習が起きやすくなってしまっていると考えられる．

7 結論

本研究では，超球面に対する分布エンコーディングを用いた文書分類を行った．実験の結果，分布エンコーディングを用いることによって，より識別性の高い隠れベクトル z が得られた．また，ベースライン手法と比較して，隠れベクトルの次元数が少なくなった時に，分類のために必要な情報を保持した変換ができていることを実験により確認した．

参考文献

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *5th International Conference on Learning Representations, ICLR*.

[2] Devansh Arpit, Caiming Xiong, and Richard Socher. Entropy penalty: Towards generalization beyond the iid assumption. *arXiv preprint arXiv:1910.00164*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[4] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR*.

[5] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9, pp. 2579–2605, 2008.

[6] Dinghan Shen, Guoyin Wang, Wenlin Wang, and et. al. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms.

[7] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.